Label Efficient Deep Learning in Medical Imaging

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M. Sc. Marcel Früh

aus Tübingen

Tübingen

2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	06.10.2023
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Andreas Schilling
2. Berichterstatter:	Prof. Dr. Hendrik Lensch

Abstract

Recent state-of-the-art deep learning frameworks require large, fully annotated training datasets that are, depending on the objective, time-consuming to generate. While in most fields, these labelling tasks can be parallelized massively or even outsourced, this is not the case for medical images. Usually, only a highly trained expert is able to generate these datasets. However, since additional manual annotation, especially for the purpose of segmentation or tracking, is typically not part of a radiologist's workflow, large and fully annotated datasets are a rare and scarce good. In this context, a variety of frameworks are proposed in this work to solve the problems that arise due to the lack of annotated training data across different medical imaging tasks and modalities.

The first contribution as part of this thesis was to investigate weakly supervised learning on PET/CT data for the task of lesion segmentation. Using only class labels (tumor vs. no tumor), a classifier was first trained and subsequently used to generate Class Activation Maps highlighting regions with lesions. Based on these region proposals, final tumor segmentation could be performed with high accuracy in clinically relevant metrics. This drastically simplifies the process of training data generation, as only class labels have to be assigned to each slice of a scan instead of a full pixel-wise segmentation.

To further reduce the time required to prepare training data, two self-supervised methods were investigated for the task of anatomical tissue segmentation and landmark detection. To this end, as a second contribution, a state-of-the-art tracking framework based on contrastive random walks was transferred, adapted and extended to the medical imaging domain. As contrastive learning often lacks real-time capability, a self-supervised template matching network was developed to address the task of real-time anatomical tissue tracking, yielding the third contribution of this work. Both of these methods have in common that only during inference the object or region of interest is defined, reducing the number of required labels to as few as one and allowing adaptation to different tasks without having to re-train or access the original training data. Despite the limited amount of labelled data, good results could be achieved for both tracking of organs across subjects as well as tissue tracking within time-series.

State-of-the-art self-supervised learning in medical imaging is usually performed on 2D slices due to the lack of training data and limited computational resources. To exploit the three-dimensional structure of this type of data, self-supervised contrastive learning was performed on entire volumes using over 40,000 whole-body MRI scans forming the fourth

contribution. Due to this pre-training, a large number of downstream tasks could be successfully addressed using only limited labelled data. Furthermore, the learned representations allows to visualize the entire dataset in a two-dimensional view.
To encourage research in the field of automated lesion segmentation in PET/CT image data, the autoPET challenge was organized, which represents the fifth contribution.

Kurzzusammenfassung

Moderne Deep-Learning-Frameworks erfordern große, vollständig annotierte Trainingsdatensätze, deren Erstellung je nach Zielsetzung sehr zeitaufwändig ist. Während der Annotationsvorgang in den meisten Fächern bzw. Bereichen massiv parallelisiert oder sogar ausgelagert werden kann, ist dies bei medizinischen Bildern nicht der Fall. In der Regel ist nur ein gut ausgebildeter Experte in der Lage, diese Datensätze zu erstellen. Da jedoch ein manuelles Labeln, insbesondere zum Zwecke der Segmentierung oder des Trackings, in der Regel nicht Teil des Arbeitsablaufs eines Radiologen ist, sind große und vollständig beschriftete Datensätze ein seltenes und rares Gut. Um dem zu entgegnen, werden in dieser Arbeit eine Reihe von Frameworks aufgezeigt, welche die Probleme, die durch den Mangel an annotierten Trainingsdaten im Bereich der medizinischen Bildgebung entstehen, adressieren.

Der erste Beitrag im Rahmen dieser Arbeit war die Untersuchung von weakly-supervised learning auf PET/CT-Daten für das Problem der Tumorsegmentierung. Unter ausschließlicher Verwendung binärer Klassenlabels (Tumor vs. Kein Tumor) wurde zunächst ein Klassifikator trainiert, der anschließend verwendet wurde, um Class Activation Maps zu erzeugen, welche Regionen mit Läsionen anzeigen. Mit Hilfe dieser Regionsvorschläge konnte eine finale Tumorsegmentierung durchgeführt werden, die eine hohe Genauigkeit bei den klinisch relevanten Metriken erreichte. Dieser Vorgang vereinfacht den Prozess der Trainingsdatenerzeugung drastisch, da anstatt einer vollständigen Tumorsegmentierung nur Klassenlabels für jede Schicht eines Scans erzeugt werden müssen.

Um den Zeitaufwand für die Vorbereitung von Trainingsdaten weiter zu reduzieren, wurden zwei unterschiedliche self-supervised Methoden für die anatomische Gewebesegmentierung und Landmark Detektion untersucht. Hierfür wurde als zweiter Beitrag dieser Arbeit ein state-of-the-art Tracking Framework, welches auf contrastive random walks basiert, auf die medizinische Bildgebung übertragen, angepasst und erweitert. Da contrastive learning oft noch nicht echtzeitfähig eingesetzt werden kann, wurde eine self-supervised Template Matching Architektur entwickelt, um die Aufgabe der Echtzeitverfolgung von anatomischem Gewebe zu ermöglichen. Dies ist der dritte Beitrag dieser Thesis. Beide Methoden haben gemeinsam, dass erst während der Inferenz definiert wird, welches Objekt verfolg werden soll. Hierdurch lässt sich die Anzahl der erforderlichen Labels auf bis zu Eins reduzieren. Ebenso erlaubt dies eine Anpassung an sich ändernde Aufgaben, ohne dass ein erneutes Training oder Zugriff auf die Trainingsdaten notwendig ist. Trotz

der begrenzten Menge an gelabelten Daten konnten, sowohl für das Organtracking zwischen unterschiedlichen Patienten als auch innerhalb einer Zeitreihe, gute Ergebnisse erzielt werden.
Self-supervised learning wurde bisher aufgrund begrenzter Trainingsdaten und Rechenkapazität in der Regel nur auf zwei-dimensionalen Schichten angewendet. Um sich die dreidimensionale Struktur der Daten zunutze zu machen, wurde als vierter Beitrag dieser Arbeit self-supervised contrastive learning auf vollen Volumen und über 40.000 Ganzkörper MRTs durchgeführt. Aufgrund dieses Vortrainings konnten viele Downstream Tasks mit nur sehr begrenztem Labelling durchgeführt werden. Darüber hinaus ermöglicht die gelernte Repräsentation die Visualisierung des gesamten Datensatzes in einer zwei-dimensionalen Ansicht.
Um die Forschung auf dem Gebiet der automatischen Tumorsegmentierung in PET/CT Daten zu fördern, wurde die autoPET Challenge organisiert, was den fünften Beitrag darstellt.

Acknowledgments

I would like to express my deepest gratitude and appreciation to all those who have supported me during this journey.
First and foremost, I would like to sincerely thank my supervisor Andreas Schilling for his support and guidance during these years and especially for encouraging me to start this journey in the very first place. I highly enjoyed our meetings and discussions, which were always a fountain of new ideas.
I am immensely grateful to Sergios Gatidis for all his support over the last years. There are few things that I have enjoyed and learnt so much from as our weekly meetings, with or without Berliners (mostly with). His mentoring, feedback and guidance have been invaluable to me and my research.
I would also like to express my deepest thanks to Thomas Küstner for his ongoing advice and guidance and, of course, for our joint conference travels around the world.
During this work, I had the honor to work together with a group of great people including Tobias Hepp, Marc Fischer, Louisa Fay, Aya Ghoul, Siying Xu and Andreas Daul. Thank you for all our discussions, activities and your support.
I would also like to thank some of my dearest friends, Jonas, Andreas, Dat and Kevin. Thank you for your unwavering friendship and for the best years of my life in Tübingen.
Finally, I am deeply grateful to my family, most especially my parents, as well as my partner Isabelle, for their unwavering support, without which none of this would have been possible. Thank you for always being there for me and making me the person I am today.

Thank you all for your support, encouragement and understanding. This work would not have been possible without your help.

Contents

1. Publications

Accepted Peer-Reviewed Journal Papers

1. **Marcel Früh**, Marc Fischer, Andreas Schilling, Sergios Gatidis, Tobias Hepp: Weakly supervised segmentation of tumor lesions in PET-CT hybrid imaging *In* Journal of Medical Imaging (2021)

2. **Marcel Früh**, Thomas Küstner, Marcel Nachbar, Daniela Thorwarth, Andreas Schilling, Sergios Gatidis: Self-supervised learning for automated anatomical tracking in medical image data with minimal human labeling effort *In* Computer Methods and Programs in Biomedicine (2022)

3. **Marcel Früh**, Andreas Schilling, Sergios Gatidis, Thomas Küstner: Real Time Landmark Detection for Within- and Cross Subject Tracking With Minimal Human Supervision *In* IEEE Access (2022)

4. Sergios Gatidis, Tobias Hepp, **Marcel Früh**, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, Daniel Rubin: A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions *In* Scientific Data (2022)

Accepted Peer-Reviewed Conference Abstracts

5. **Marcel Früh**, Tobias Hepp, Andreas Schilling, Sergios Gatidis, Thomas Küstner: Self-supervised Training for Single-Shot Tumor Tracking in the Presence of Respiratory Motion *In* Proceedings of the International Society for Magnetic Resonance in Medicine (ISMRM) (2022)

6. Siying Xu, **Marcel Früh**, Kerstin Hammernik, Sergios Gatidis, Thomas Küstner: Self-supervised contrastive learning for MR image reconstruction of cardiac CINE on accelerated cohorts *In* Proceedings of the International Society for Magnetic Resonance in Medicine (ISMRM) (2023)

7. Thomas Küstner, Jan Borst, Dominik Nickel, Fabian Bamberg, **Marcel Früh**, Sergios Gatidis: Self-supervised contrastive learning for motion artifact detection in whole-body MRI: Quality assessment across multiple cohorts *In* Proceedings of the International Society for Magnetic Resonance in Medicine (ISMRM) (2023)

Under Review

8. Sergios Gatidis, **Marcel Früh**, Matthias Fabritius, Konstantin Nikolaou, Christian La Fougère, Jin Ye, Junjun He, Yige Peng, Lei Bi, Jun Ma, Bo Wang, Jia Zhang, Yukun Huang, Lars Heiliger, Zdravko Marinov, Jens Kleesiek, Ludovic Sibille, Lei Xiang, Simone Bendazzoli, Mehdi Astaraki, Michael Ingrisch, Clemens Cyran, Bernhard Schölkopf, Thomas Küstner: The autoPET challenge: Towards fully automated lesion segmentation in oncologic PET/CT imaging *submitted to* Nature Machine Intelligence (02/2023)

Not Submitted

9. **Marcel Früh**, Andreas Schilling, Thomas Küstner, Sergios Gatidis: Large Scale, entire-volume, 3D Contrastive Learning on whole-body MRI data

Other Contributions

- Co-Organizer of the *Automated Lesion Segmentation in Whole-Body FDG-PET/CT challenge* (autoPET 2022) at the 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI).

- Co-Organizer of the *Automated Lesion Segmentation in Whole-Body FDG-PET/CT challenge* (autoPET 2023) at the 26th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI).

1.1 Personal Contributions

The following list contains the contributions of the individual authors to all collaborative manuscripts. Manuscript 4 already contains this statement, hence it has been adapted and reused here. The personal contribution (§ 6 Abs. 2 Satz 3 Promotionsordnung) is represented in bold percentages.

Accepted Peer-Reviewed Journal Papers

1. **Marcel Früh**, Marc Fischer, Andreas Schilling, Sergios Gatidis, Tobias Hepp: Weakly supervised segmentation of tumor lesions in PET-CT hybrid imaging:

    - ***M.F.**, A.S., S.G. T.H.*: scientific ideas, design and conception of the work **(75 %)**
    - ***M.F.**, T.H.*: data processing, implementation of the algorithms and evaluation routines, experiments **(85 %)**
    - ***M.F.**, S.G., T.H.*: detailed analysis and interpretation of the results **(90 %)**
    - ***M.F.**, M.Fi., A.S, S.G., T.H.*: manuscript drafting and revision **(70 %)**

2. **Marcel Früh**, Thomas Küstner, Marcel Nachbar, Daniela Thorwarth, Andreas Schilling, Sergios Gatidis: Self-supervised learning for automated anatomical tracking in medical image data with minimal human labeling effort

    - ***M.F.**, A.S. S.G.*: scientific ideas, design and conception of the work **(85 %)**
    - ***M.F.**, M.N., D.T.*: data processing, implementation of the algorithms and evaluation routines, experiments **(80 %)**
    - ***M.F.**, T.K., S.G.*: detailed analysis and interpretation of the results **(90 %)**
    - ***M.F.**, T.K., M.N, D.T., A.S., S.G.*: manuscript drafting and revision **(75 %)**

3. **Marcel Früh**, Andreas Schilling, Sergios Gatidis, Thomas Küstner: Real Time Landmark Detection for Within- and Cross Subject Tracking With Minimal Human Supervision

    - ***M.F.**, A.S., S.G., T.K.*: scientific ideas, design and conception of the work **(90 %)**
    - ***M.F.**, T.K.*: data processing, implementation of the algorithms and evaluation routines, experiments **(95 %)**
    - ***M.F.**, A.S., T.K., S.G.*: detailed analysis and interpretation of the results **(90 %)**
    - ***M.F.**, A.S., T.K., S.G.*: manuscript drafting and revision **(70 %)**

4. Sergios Gatidis, Tobias Hepp, **Marcel Früh**, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, Daniel Rubin: A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions:

    - *S.G., D.R., T.K., T.H, **M.F.***: scientific ideas, design and conception of the work **(10 %)**
    - *S.G., D.R., T.K., T.H., **M.F.***: data processing, implementation of the algorithms and evaluation routines, experiments **(10 %)**
    - *K.N., C.L.F., **M.F.**, C.P., B.S., C.C.*: detailed analysis and interpretation of the results **(30 %)**
    - *S.G., D.R., T.K., T.H., K.N., C.L.F., **M.F.**, C.P., B.S., C.C.*: manuscript drafting and revision **(10 %)**

Accepted Peer-Reviewed Conference Abstracts

5. **Marcel Früh**, Tobias Hepp, Andreas Schilling, Sergios Gatidis, Thomas Küstner: Self-supervised Training for Single-Shot Tumor Tracking in the Presence of Respiratory Motion

    - ***M.F.**, A.S., S.G., T.K.*: scientific ideas, design and conception of the work **(90 %)**
    - ***M.F.**, T.H., T.K.*: data processing, implementation of the algorithms and evaluation routines, experiments **(90 %)**
    - ***M.F.**, T.H., A.S., T.K., S.G.*: detailed analysis and interpretation of the results **(90 %)**
    - ***M.F.**, A.S., T.K., S.G.*: manuscript drafting and revision **(85 %)**

6. Siying Xu, **Marcel Früh**, Kerstin Hammernik, Sergios Gatidis, Thomas Küstner: Self-supervised contrastive learning for MR image reconstruction of cardiac CINE on accelerated cohorts

    - *S.X., **M.F.**, K.H., S.G., T.K*: scientific ideas, design and conception of the work **(15 %)**
    - *S.X., T.K.*: data processing, implementation of the algorithms and evaluation routines, experiments **(0 %)**
    - *S.X., **M.F.**, K.H., S.G., T.K.*: detailed analysis and interpretation of the results **(10 %)**
    - *S.X., **M.F.**, K.H., S.G., T.K.*: manuscript drafting and revision **(5 %)**

7. Thomas Küstner, Jan Borst, Dominik Nickel, Fabian Bamberg, **Marcel Früh**, Sergios Gatidis: Self-supervised contrastive learning for motion artifact detection in whole-body MRI: Quality assessment across multiple cohorts

    - *T.K., **M.F.**, S.G.*: scientific ideas, design and conception of the work **(25 %)**

- *T.K., J.B., D.N.*: data processing, implementation of the algorithms and evaluation routines, experiments **(0 %)**
- *T.K., F.B.,* ***M.F.****, S.G.*: detailed analysis and interpretation of the results **(20 %)**
- *T.K., F.B.,* ***M.F.****, S.G.*: manuscript drafting and revision **(5 %)**

Under Review

8. Sergios Gatidis, **Marcel Früh** et al.: The autoPET challenge: Towards fully automated lesion segmentation in oncologic PET/CT imaging

    - *S.G.,* ***M.F.****, M.F., M.I., C.C., T.K.* : challenge organization, data preparation, software contribution, data analysis, manuscript drafting **(40 %)**

Not Submitted

9. **Marcel Früh**, Andreas Schilling, Thomas Küstner, Sergios Gatidis: Large Scale, entire-volume, 3D Contrastive Learning on whole-body MRI data

    - ***M.F.****, A.S., T.K., S.G.*: scientific ideas, design and conception of the work **(95 %)**
    - ***M.F.****. T.K.*: data processing, implementation of the algorithms and evaluation routines, experiments **(90 %)**
    - ***M.F.****, T.K., S.G.*: detailed analysis and interpretation of the results **(90 %)**
    - ***M.F.****, A.S., T.K., S.G.*: manuscript drafting and revision **(90 %)**

    This work is primarily based on manuscripts 1-4 and 8-9 which are included in the appendix.

2. Introduction

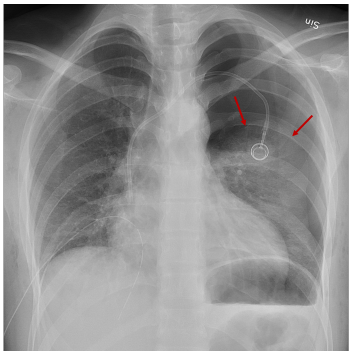
Over the past decade, machine learning systems have transformed many areas of industry and personal life. A variety of highly accurate classification, segmentation and tracking frameworks now exist for a vast number of tasks and problems. From classification of coffee capsules [1] over (semi-) self-driving cars [2] to autonomous drones delivering medical samples [3], machine learning is here to stay. One significant field, however, is not proceeding as fast as the others, namely the medical imaging domain.

2.1 Medical Imaging - A Field that matters

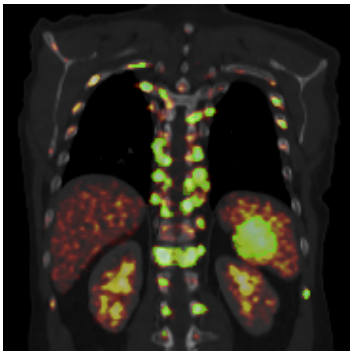
More than 3.6 billion medical imaging procedures are performed worldwide each year, demonstrating the diagnostic importance of these techniques [4]. Depending on the modality (e.g X-Ray, CT, MRI, PET, Echo) and the underlying disease, different tasks or necessities arise. Often, the goal is to classify a disease or medical condition, which typically does not require time-consuming steps. Some medical fields, however, require specific metrics to evaluate the severity of a disease. In cardiology, for example, the volume of the left ventricle is crucial, while in oncology the tumor volume is of particular importance. To obtain these metrics, manual segmentation of the relevant anatomical tissue is required which often results in an enormous expenditure of time, especially in three- or four-dimensional volumes.

Recent technological innovations in radiation oncology, namely the introduction of the MR-LINAC, an MRI/linear accelerator hybrid that allows real-time acquisition of MR images during radiotherapy, have opened up an entirely new need beyond what we humans are capable of: Tracking of anatomical structures in real time (and simultaneously passing this information to the scanner).

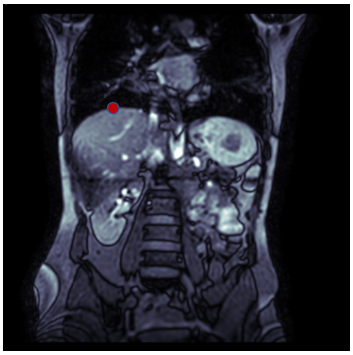
The ability to track organs or lesions in real-time enables image-guided radiotherapy, which could substantially improve outcomes by preventing irradiation of healthy tissue.



(a) Fast Disease Classification: Pneumothorax of the left lung



(b) Laborious Tumor Segmentation: Lymphoma



(c) Beyond Human: Real-time tracking during radiotherapy

Figure 2.1: Comparison between disease classification in radiographs (a) [5], tumor segmentation in FDG-PET/CT data (b) [6], and liver dome tracking during radiotherapy in MR-LINAC data (c) [7]. Classification, regardless of modality, can usually be performed by the radiologist without much manual effort. Tumor/Tissue Segmentation, in contrast, is always a time-consuming and costly task, especially for highly metastatic cancers such as lymphoma or melanoma. Tracking of organs or even anatomical landmarks in real time is not feasible for humans due to the very low latency requirement.

An example representative for each task is visualized in Figure 2.1.
In recent years, a variety of machine learning-based systems have appeared to tackle these above mentioned tasks. First, highly specific classification and regression problems were addressed [8]. Prominent examples include the classification of skin-cancer [9], detection of diabetic retinopathy[10], brain tumor classification [11] and recognition of pneumonia in infant X-Ray images [12]. These frameworks can be used in a supportive way but an expert still needs to review every case, thus their benefit is limited.
In contrast to plain classification, automatic segmentation of anatomical structures is a real support for the physician, as it drastically simplifies the previously time-consuming manual annotation. Several frameworks have been proposed to tackle this problem, most notably the U-Net [13] and nn-U-Net [14]. Many tasks, including tumor segmentation for various cancers and modalities [15, 16, 17, 18], as well as organ segmentation for single organs, multiple organs or task-specific structures like the cardiac chambers [19, 20, 21, 22] have been successfully addressed. Tracking of anatomical tissue also is a crucial but either time-consuming - if done retrospectively - or impossible task due to the real-time constraint. Thus, plenty of frameworks have been introduced, often based on classical Optical Flow

estimation [23, 24, 25]. Although deep learning based Optical Flow techniques exist [26, 27, 28], their application to medical images is scarce as they cannot be applied without retraining due to the domain shift.
This problem is, in fact, common for all mentioned works. All frameworks, except for classical Optical Flow algorithms which work directly at the pixel level, are based on architectures trained in a supervised manner on a fully annotated / segmented dataset that required enormous expert effort beforehand and drastically limits their application to other tasks within the same modality. As a result, more efficient learning routines should be used.

2.1.1 Learning Strategies

Over the years, several learning strategies have emerged, varying in the amount and type of labels required, as well as in their inference routines.

Supervised Learning

Up to this date, supervised learning is the most commonly used approach to train any type of machine learning system. After specification of a task, e.g tumor segmentation in PET/CT images, a team needs to be assembled to manually collect and annotate a large training dataset which is then leveraged to train the (tumor-segmentation) model. The training and inference process is depicted in Fig. 2.2.
Manual annotation, especially segmentation, is always associated with an enormous amount of work. In the field of medical imaging, however, only highly trained experts are able to provide the necessary ground truth, which further complicates training data generation as their time is scarce good.
Another very important drawback is the adaptability to new tasks. Once trained, a supervised model cannot simply change its output and requires retraining to match the new task. To overcome the challenge of needing fully labelled datasets, one approach is to use **Weakly-Supervised Learning**.

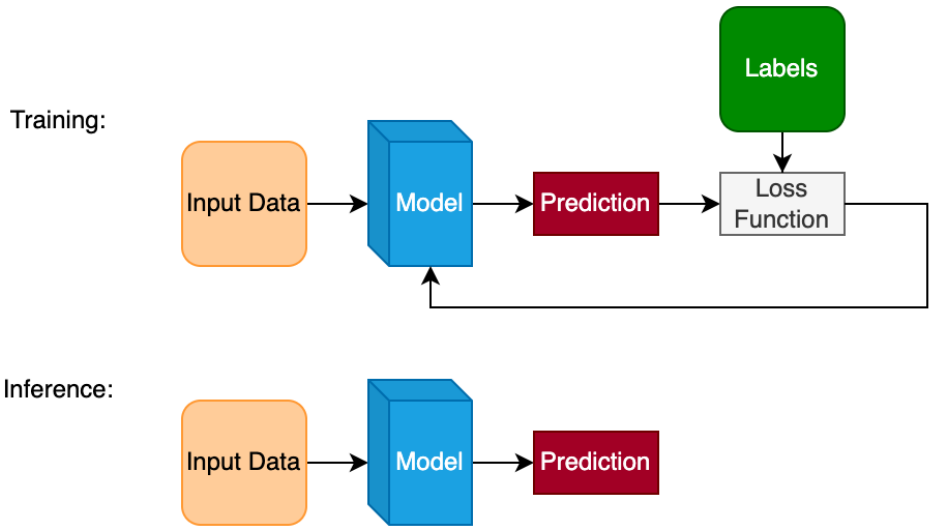


Figure 2.2: Supervised Learning: The model is trained on the task directly and can be used for inference without any subsequent steps.

Weakly-Supervised Learning

Weakly-supervised learning strategies aim to reduce the labelling effort by using weaker annotations such as class labels instead of full segmentations [29, 30, 31]. In a first step, the model is trained using the provided weak labels. Subsequently, some form of post-processing is implemented to extract the knowledge or insights from the model, with the goal of obtaining the desired final prediction. For segmentation or localization tasks, this is often achieved leveraging Class Activation Maps [32, 33]. Fig. 2.3 visualizes the training routine of a weakly-supervised framework.

Weakly-supervised strategies, while capable of drastically reducing the labelling effort, still require some expert effort before training. Their biggest drawback, however, is that they still require retraining to adapt to new tasks.

Recently, a new learning paradigm has evolved that is capable of addressing these two challenges simultaneously: **Self-Supervised Learning**.

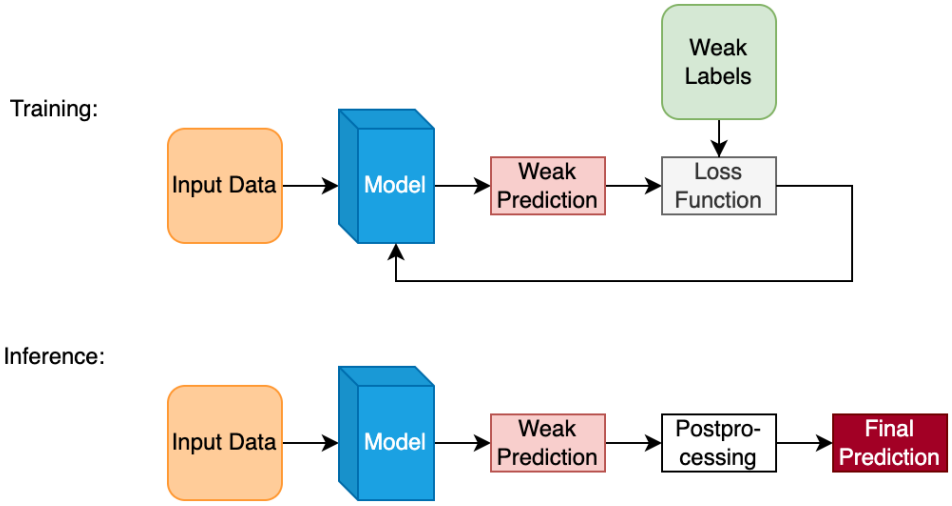


Figure 2.3: Weakly-Supervised Learning: Training is performed using weaker labels (e.g. classification labels instead of full segmentations). During inference, task-specific post-processing is applied to the weak prediction to yield the final estimate.

Self-Supervised Learning

In contrast to supervised- and weakly-supervised strategies, self-supervised frameworks are trained without any labels on some pretext task, often by exploiting specific properties / structures of the training dataset, e.g. by predicting image rotations [34] or by solving jigsaw puzzles [35]. These networks are subsequently finetuned on a small labelled subset using approximately 1-10% of the data (semi-supervised finetuning), yielding comparable results compared to their fully-supervised trained baselines [36]. A high-level overview of a self-supervised system is depicted in Fig. 2.4.

The use of self-supervised pretraining eliminates the dependence on having a large, densely labelled dataset prior to training. Only after the initial training process, during the fine-tuning step, a labelled dataset is required - however, a small fraction compared to fully-supervised training is sufficient. This allows for easy adaptation to new tasks within the same domain, as only few labelled data points are required to fit the model. If the pretext task can be formulated in a way that it directly represents the task of interest, such as in Frueh et al. [7], no fine-tuning is necessary.

More recently, self-supervised learning is usually performed in a **contrastive** setting to embed training data into meaningful feature representations.

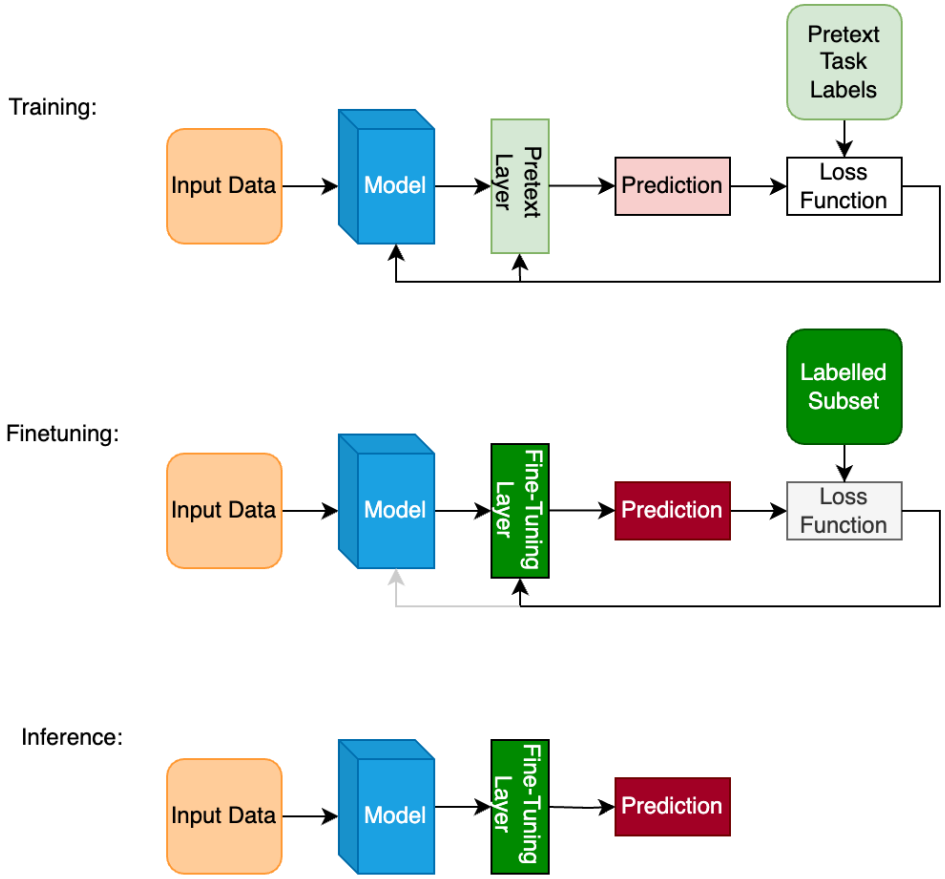


Figure 2.4: Self-Supervised Learning: The model is pretrained on some pretext task that is usually derived from the data itself. After pretraining, the fully connected pretext layer is replaced by the fine-tuning layer to match the final task and then fine-tuned (either the whole-model or only the fine-tuning layer) on a small labelled subset. Inference is subsequently performed without any post-processing.

Self-Supervised Contrastive Learning

Contrastive learning aims to learn an embedding strategy that maps input data to feature vectors, with the intention of obtaining similar feature vectors for similar inputs and dissimilar features for dissimilar inputs, typically using the cosine similarity as a measure of similarity between the feature vectors.

These frameworks, however, are not generally self-supervised as an anchor image paired with positive and negative examples is required [37]. During training, the model learns to pull together the feature vectors of the anchor image and positive example, whilst repelling the negative example.

Chen et al. could prove that contrastive learning without the necessity of any labels is possible using a simple framework called SimCLR [38]. Two different views of the same image are created using extensive, targeted data augmentation corresponding to the anchor and positive examples. All other image pairs within the current batch act as negative examples. Obtaining a representative feature embedding model allows for both fine-tuning combined with direct inference (Fig. 2.4) as well as few-shot inference.

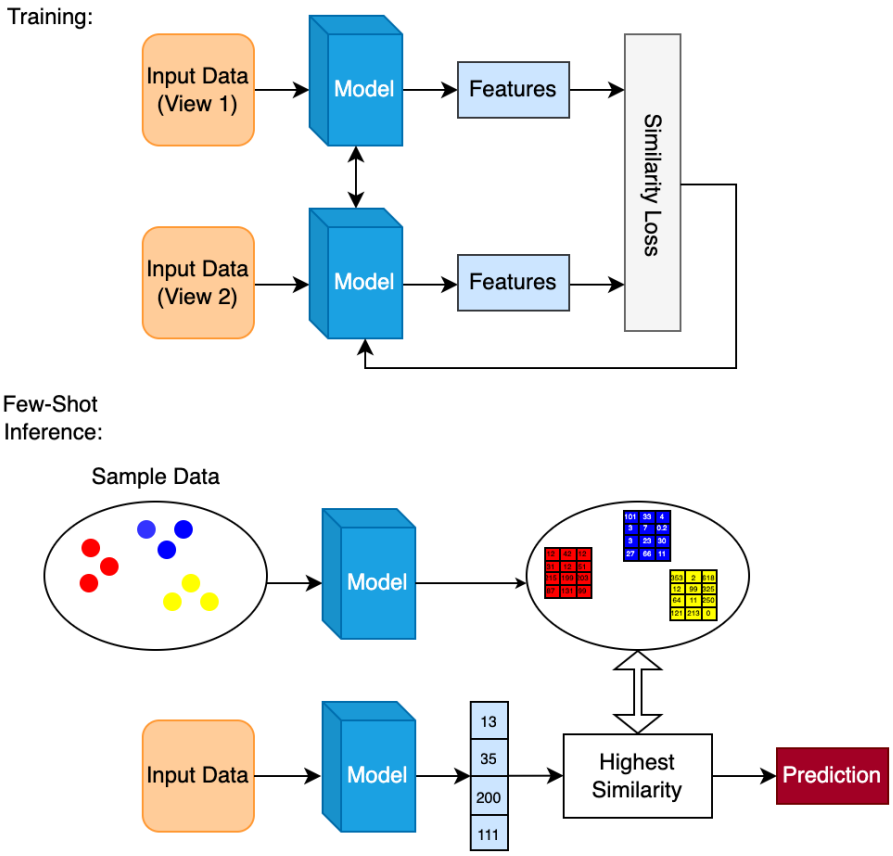


Figure 2.5: Self-Supervised Contrastive Learning: Two different augmentations of the same image are created using data augmentation (as proposed in SimCLR, simplified) which are then embedded into their respective feature vectors and pushed to be close according to some similarity loss. Both direct inference by finetuning (refer Fig. 2.4) as well as few-shot inference are possible.

Few-shot inference can be performed as follows:

1. Select 5-10 samples for each class
2. Store the feature vectors for all samples

3. Map a test image x to its feature representation $\hat{x}$
4. Compute the cosine-similarity between $\hat{x}$ and all stored feature vectors
5. Assign the class with the highest overall similarity to x

However, due to the exhaustive feature comparison process that comes with few-shot inference, this approach is not yet real-time capable.
Fig. 2.5 visualizes the contrastive learning setting using a simplified version of SimCLR. Since 2020, a plethora of self-supervised contrastive frameworks have been released and it is currently regarded as state of the art for most self-supervised learning tasks [39, 36, 40, 41].

2.1.2 Challenges

While all of these learning strategies have been successfully applied to numerous tasks [42, 43, 44], application of weakly-supervised or self-supervised (contrastive) frameworks to medical images is still scarce and often focused on highly preprocessed and simple datasets such as chest X-Rays or dermatological images [45, 46, 47]. Only very recently self-supervised learning has been utilized for more challenging tasks such as tumor segmentation [48].
The main reason for this lag compared to other areas is the high entry barrier due to low availability of training data, combined with its often difficult structure.
In summary, the following challenges occur in the context of medical image data:

- Lack of annotated training data, especially outside the usual domains. Furthermore, for more complex tasks such as real-time tracking during radiotherapy, unlabelled data is also rare.
- Unlike RGB images, which are completely standardized, medical images vary greatly between hospitals and scanners, especially for the field of MRI. Second to that, image quality can be significantly affected by motion, artifacts and noise.
- Due to their three- or four-dimensional structure, handling of medical images is exceptionally computationally demanding. Some applications, such as tracking during radiotherapy, furthermore need to be able to run in real time.
- Trade-off between data privacy and research.

Therefore, the performance of systems that show promising results on everyday images does not necessarily translate well to the medical imaging context due to the different nature of medical image data.

2.2 Research Goal and Outline

The primary objective of this work is to overcome some of these challenges and advance the field of deep learning in medical imaging with a focus on anatomical tissue segmentation and tracking, as these are among the most useful for physicians. One particular goal is to improve image-guided radiotherapy using the MR-LINAC system. To achieve this, real-time tracking of lesions or organs is required, which ultimately allows for patient or beam adjustment, resulting in increased efficacy and reduced radiation damage at the same time. Since annotated training data, especially for tracking tasks, is scarce, all frameworks need to be able to work without full supervision by being either label-efficient or, much better, label-free during training.
To this end, several label efficient and label free frameworks have been either developed or transferred and adapted, implemented and evaluated on a wide range of medical imaging tasks, including lesion segmentation in PET-CT images [49], anatomical tissue tracking [50], real-time landmark detection and tracking [7], as well as large-scale 3D contrastive learning on over 40,000 whole-body MR scans C.1.
In the following section, the outline of this work is presented. A brief background is provided for each paper.

Weakly supervised segmentation of tumor lesions in PET-CT hybrid imaging

Metabolic Tumor Volume (MTV) and Total Lesion Glycolysis (TLG) are two of the most important measures to assess the severity of cancer in a patient [51]. To estimate these metrics, all lesions have to be manually segmented by an expert in PET/CT images. This laborious task can be tackled with great success using deep learning based segmentation architectures, such as the nn-U-Net [1]. However, these supervised methods require a large and fully segmented training dataset in advance, which is expensive and arduous to create.

[1] `https://autopet.grand-challenge.org/`

To circumvent this problem, a weakly supervised strategy was implemented that requires only class labels instead of full segmentations, drastically reducing the labelling effort [49].

Self-supervised learning for automated anatomical tracking in medical image data with minimal human labelling effort

Many tasks require the tracking of organs and other anatomical tissues in time-resolved image data. Examples include measuring the volume change of the left ventricle in CINE-MRI, or tracking the liver during radiotherapy to avoid irradiation of healthy tissue. Theoretically, all these tasks could be solved using standard segmentation networks trained under full supervision. Labelled datasets of sufficient size, however, do not exist, thus exhaustive preparation would be required. Moreover, a supervised trained network is rendered useless if the task changes slightly, e.g. tracking the right ventricle instead of the left ventricle.
These tasks were addressed by adaptation and extension of a self-supervised contrastive framework [52] to the medical domain, enabling automated anatomical tracking using only a single labelled example [50].

Real Time Landmark Detection for Within- and Cross Subject Tracking With Minimal Human Supervision

Contrastive learning has one major drawback: Unless finetuned on a specific task, exhaustive feature comparison is necessary to conduct inference. While this is fast for classification, as only one feature vector per image is compared to the class example vectors, it becomes a major bottleneck for segmentation. If an image of shape (256×256) is transformed into a feature map of shape $(64 \times 64 \times 512)$, 4096 feature vectors, each consisting of 512 elements, must be compared between the first and second image, which is not feasible in real-time with current hardware. Thus, to address this challenge, a self-supervised template matching architecture was developed to estimate the position of anatomical landmarks based on a single labelled example for both time-series and cross-subject tracking in real time [7].

Large Scale, entire-volume, 3D Contrastive Learning on whole-body MRI data

Due to advances in compute power coupled with the availability of a large whole-body MRI dataset (50,000 scans), contrastive learning was performed to map entire-volume whole-body MRI scans to a single feature vector. As the dataset was derived from a large biomedical database (UK Biobank [53]), additional metadata was available (e.g. sex, age, weight, health score), rendering the possibility to evaluate the capacity of the trained framework on medical imaging data. (Appendix C.1)

Advancing the Field through Challenges

Over the years, task-specific challenges have become the primary tool for comparing and advancing machine learning algorithms, especially in less common areas such as medical imaging. By providing a fully annotated dataset [6], infrastructure for automated inference using the grand-challenge platform, as well as a prize pool of €15,000, participants were encouraged to develop and submit frameworks that address lesion segmentation in PET/CT data [54].

3. Results

3.1 Weakly supervised segmentation of tumor lesions in PET-CT hybrid imaging

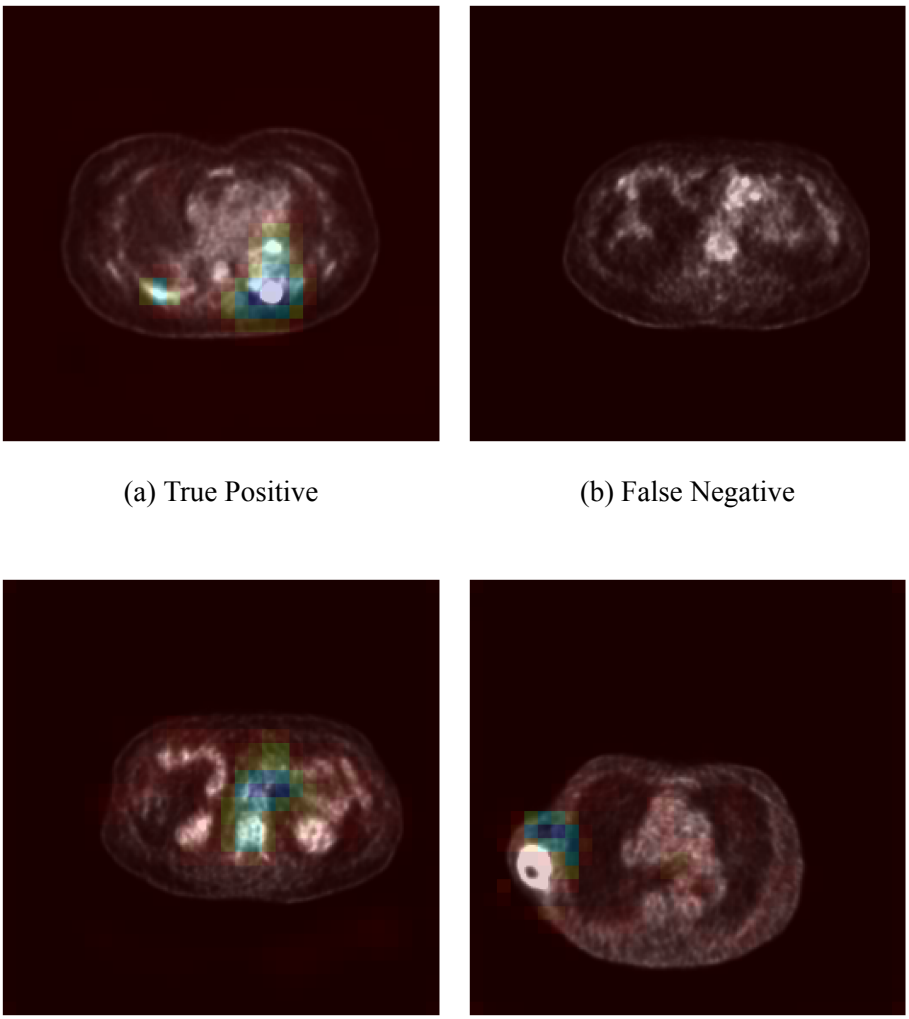
This section summarizes the method and results presented in [49]. The main contribution lies in the development of a weakly supervised framework for lesion segmentation in FDG-PET/CT images.

A three-step pipeline was developed to obtain the full tumor segmentation using only class labels: First, a binary classifier is trained on two-dimensional PET/CT [6] slices (input shape: $(2 \times 256 \times 256)$) that predicts whether a slice contains a lesion.

Training was conducted using an adapted VGG16 architecture where the first max-pooling layer was removed to increase the feature map size to (32×32), as a large feature map is crucial for the second step. Given the trained lesion classification model and a test subject, each slice of the scan is classified as *tumorous* or *not tumorous*. If the network suspects that a slice contains a lesion, a Class Activation Map (CAM) [32] is computed, highlighting the regions responsible for the decision of the classifier. Subsequently, this map is upscaled to the original image size. If the tumor classifier has been trained properly, this saliency map should only highlight those regions of the slice that contain a lesion. A method specific threshold was then applied to the CAM to decide whether a pixel lies within one of the proposed regions.

Thus, a large feature map is required in order to yield a finer-grained Class Activation Map. A small feature map (e.g., in the extreme case of shape (1×1)) cannot be used to project the estimated lesion area onto the original image because all regional information is lost. Fig. 3.1 depicts examples of accurate and inaccurate CAMs.

In a next step, leveraging the structure of 18F-FDG PET, the standardized uptake value (SUV) distribution within the highlighted regions is estimated. Subsequently, all pixels with an SUV higher than a specific percentile of this SUV distribution are segmented as *tumorous*.



(a) True Positive (b) False Negative

(c) False Positive (d) Feature Map too small

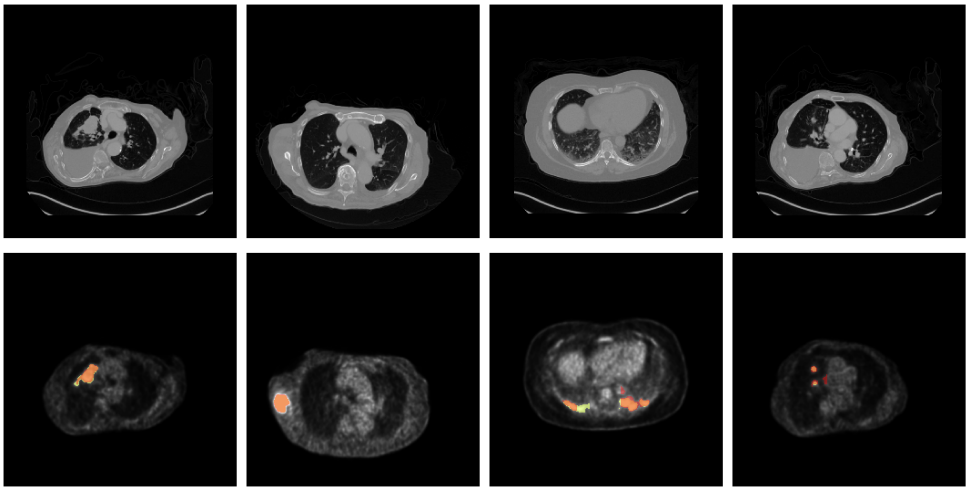
Figure 3.1: Various CAMs: (a) True Positive: The Class Activation Map highlights all lesions. (b) False Negative: The classifier did not recognize the central lesions. (c) False Positive: The classifier erroneously predicted this slice to be tumorous. (d) Class Activation Map too small: Although the classifier was correct, the size of the feature map is too small resulting in a missed lesion after upscaling to the original image size

Percentiles and thresholds for the Class Activation Maps were determined empirically via extensive grid search on the validation data.
Final segmentation performance was assessed across four different Class Activation Map generating methods (CAM [32], GradCAM [55], GradCAM++ [56], ScoreCAM [57]) and evaluated for Dice Score [58], MTV- and TLG deviation on the test dataset.
Application of an empirically determined fixed SUV threshold, as well as a U-Net trained with full supervision, served as lower and upper baselines, respectively.

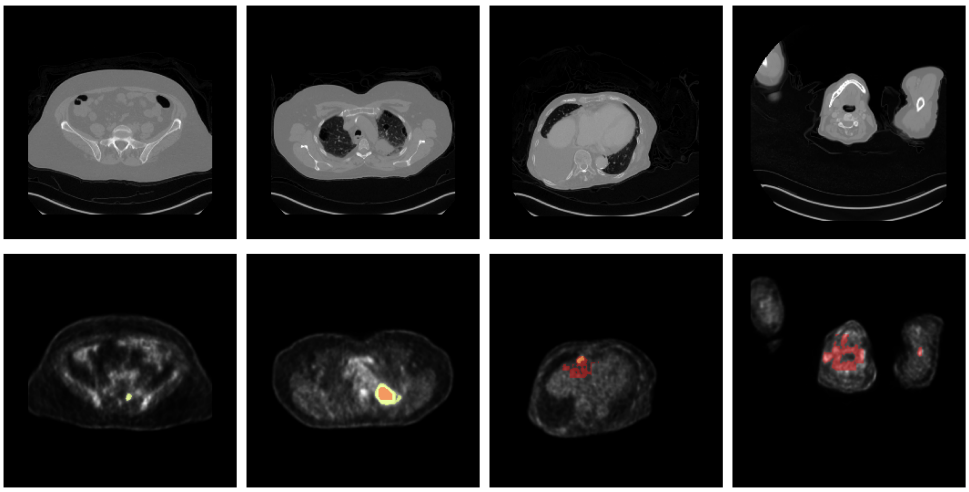
Results

As expected, the supervised U-Net produced the best results in all evaluation metrics (median Dice Score: 0.72, median MTV difference: 17 ml, median TLG deviation: 50 g). Regarding the weakly supervised frameworks, best performance was achieved using CAM and ScoreCAM (median Dice Score: 0.47, median MTV difference: 27/26 ml, median TLG deviation: 101/99 g). GradCAM++ performed slightly worse and GradCAM failed to deliver adequate results. Application of the fixed global threshold produced significantly worse results than the weakly supervised strategies (median Dice Score: 0.29, median MTV difference: 44 ml, median TLG deviation: 167 g).
Metabolic Tumor Volume was overestimated for very small lesions and underestimated for extremely large lesions for all weakly supervised methods. In contrast, more accurate results could be observed for Total Lesion Glycolysis. Overestimation of lesions with low TLG was drastically reduced compared to MTV, while very large TLG still resulted in slight underestimation.
Fig. 3.2 depicts successful and unsuccessful lesion segmentation with corresponding ground truth on PET/CT slices based on four different test subjects.
Overall, weakly supervised segmentation is a viable approach for tumor segmentation in PET/CT data, as clinically relevant metrics could be estimated with acceptable accuracy. The main application, however, is not primarily direct clinical use, but rather fast training data generation in a research setting. Manual segmentation of the entire dataset is circumvented, since only minor corrections to incorrectly estimated segmentations are necessary. Limitations arise primarily from the use of the two thresholds. While this is suitable for FDG PET/CT data, application to other modalities such as CT or MRI may be challenging. Due to computational limitations, training and evaluation were performed on two-dimensional slices instead of full 3D volumes, potentially affecting the performance of the

classifier and thus the extracted segmentation.
Future work should aim to avoid the necessity for thresholds and investigate performance when using more advanced two- or three-dimensional models.



(a) Correct Segmentation



(b) Erroneous Segmentation

Figure 3.2: Good results vs. failures in PET-CT images (top: CT, bottom: PET with corresponding segmentation): (a) In all cases, lesions were segmented (red) with only minor deviations from ground truth (yellow). (b) The weakly supervised strategy did not result in correct segmentations, as either whole lesions were missed, tumor volume was under-/overestimated, or healthy slices were incorrectly classified as tumorous.

3.2 Self-supervised learning for automated anatomical tracking in medical image data with minimal human labeling effort

This section summarizes the method and results presented in [50]. The main contribution lies in the transfer, adaptation and extension of a self-supervised contrastive tracking framework [52] to the medical imaging domain.

Tracking of anatomical structures was implemented using a framework that transforms images into a feature space where similar anatomical structures are close and dissimilar anatomical structures are far apart. Self-supervision was obtained by leveraging the cycle-consistency [59], which proved that tracking of objects in image time series could be achieved without the necessity of any labels during training.

The whole pipeline, as originally proposed in [52], consists of multiple steps, depicted below:

Initially, the image is split into a patch grid with patches of fixed size. A ResNet-18 encoder is then used to transform the individual patches into their respective feature space. Following L2 normalization of these features, the cosine similarity can be computed using the dot product, allowing for similarity comparison between all pairs of patches.

In a second step, the affinity matrix is obtained by comparing the features of all patches of two images within the time series. A patch grid of e.g. shape (7×7) results in an affinity matrix of shape (49×49). Application of the (temperature-scaled) softmax function maps these similarity scores to probabilities which can be subsequently used for loss computation.

To calculate the loss, however, some form of ground truth is required. This is where cycle-consistency becomes relevant. By extending an image sequence $(I_1...I_k)$ with its corresponding palindrome sequence $(I_k...I_1)$, the cross-entropy loss can be calculated between all pairs of patches within the source and target images, allowing for model optimization.

After training, anatomical structures can be tracked as follows: Assume X to be an image sequence consisting of two frames X_0 and X_1.

1. Manually segment the structure of interest (i.e. any organ) in X_0.
2. Use the trained ResNet encoder to map X_0 and X_1 to their corresponding downsampled feature maps $\hat{X}_0$ and $\hat{X}_1$

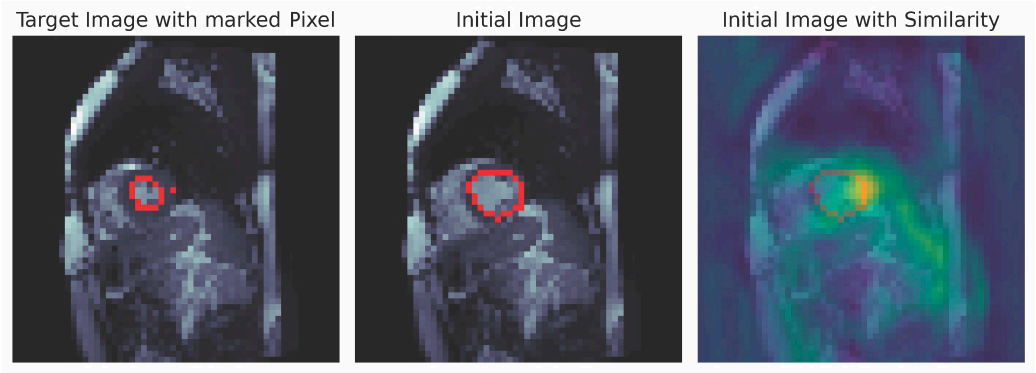
3. Compute the affinity matrix A.

4. Extract the $k = 10$ most similar feature vectors of $\hat{X}_0$ (within a radius $r = 20$) for each feature vector of $\hat{X}_1$. This corresponds to taking the k-largest affinities of each row of A

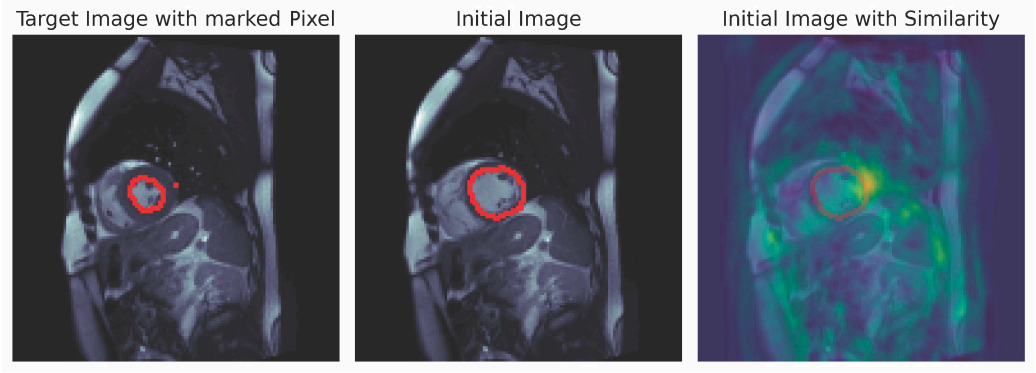
5. Identify the annotations of these k features in X_0 and assign the most frequent one to the corresponding feature vector in $\hat{X}_1$ (and thus to all pixels in the original, not-downsampled image X_1 responsible for this feature vector)

For longer image sequences, the initial segmentation is propagated iteratively by taking the k most similar features based on all previous images instead of just the first one.
To improve performance on medical data, the original framework has been extended with an additional clustering of the predicted segmentation masks. If multiple objects have been segmented, the one with the highest similarity to the provided segmentation mask is taken. As in 3.1, the size of the feature map is crucial to obtain an accurate segmentation result, especially for smaller segmentation masks. Fig. 3.3 visualizes this. To increase the size of the feature maps and thus tackle this problem, the stride of the second layer was reduced to one and images were upsampled up to (784×784) during inference. Retraining using these upscaled images was not required, but will probably result in better performance.
To evaluate and validate this framework, tracking was performed on cardiac CINE MRI [60], cardiac echo [61], and abdominal MRI acquired using the in-house MR-LINAC system.
The organs to track were the left ventricle for both cardiac CINE MRI and cardiac echo, as well as the liver for the abdominal MRI. Evaluation was performed using both the average Dice score between predicted and manual segmentation on all slices, as well as the edge Dice Score where only the final frame was considered. Both forward ($I_0 ... I_k$) and backward ($I_k ... I_0$) mask propagation was evaluated to account for different motion patterns such as contraction or expansion (systole/diastole). Additionally, artificial motion (horizontal image shift of 8% of the image size) was added into the image series to evaluate performance when large displacements occur. Besides tracking within image time series, tracking performance across different subjects was also assessed.
Initial segmentations, as well as ground truth for all intermediate frames have been provided by an experienced radiologist. Comparison was conducted against three Optical Flow baselines, namely Farnebäck Optical Flow [62], PCA-Flow [63] and FlowNet2 [27]. PCA-Flow

and FlowNet2 were applied without any finetuning.



(a) Small Feature Map (64×64)



(b) Large Feature Map (128×128)

Figure 3.3: Affinity between all features in the initial image and the marked pixel in the target image. (a) If the feature map is too small, inaccuracies occur at the tissue boundaries. The features with the highest similarity fall inside the segmented region (left ventricle) and lead to false positive segmentations. (b) With a large feature map, this problem does not occur. The most similar features are outside of the left ventricle yielding a true negative prediction for this pixel. All images were downsized to match the shape of the feature maps.

Results

Quantitative results for mask propagation are depicted in Table 3.1.
Liver tracking in MR-LINAC images was possible with high accuracy in all methods evaluated. Regarding more complex motion, such as tracking of the left ventricle in cardiac CINE MRI, the self-supervised framework (SSL) outperformed the Optical Flow based methods. This is especially true for the forward pass (systole, contraction of the heart)

Table 3.1: Mean average Dice scores (top) and mean average Dice scores with artificial image displacement (bottom)

	Forward Mask Propagation			Backward Mask Propagation		
Methods	Abdominal MRI	Cardiac MRI	Cardiac Echo	Abdominal MRI	Cardiac MRI	Cardiac Echo
SSL	**0.95**	**0.89**	**0.93**	**0.96**	0.90	**0.95**
PCA-Flow	0.94	0.80	0.89	0.93	0.89	0.91
FB-Flow	0.94	0.81	0.91	0.93	**0.92**	0.91
FlowNet2-Flow	**0.95**	0.77	0.91	0.95	0.73	0.92
SSL	**0.95**	**0.89**	**0.93**	**0.96**	**0.90**	**0.96**
PCA-Flow	0.92	0.75	0.77	0.85	0.77	0.84
FB-Flow	0.72	0.52	0.64	0.80	0.60	0.85
FlowNet2-Flow	0.94	0.58	0.81	0.95	0.55	0.81

and confirmed by the edge Dice Score (0.85, 0.61, 0.7, 0.55 for SSL, PCA-Flow, FB-Flow and FlowNet2-Flow, respectively). The backward pass (diastole, expansion of the heart), in contrast, could be accurately tracked by FB-Flow and PCA-Flow as well. FlowNet2 as a supervised trained network failed to capture this motion due to the domain shift to the training data. Tracking of the left ventricle in echocardiography was again possible with all methods.

Injection of artificial motion resulted in a severe performance decline of all Optical Flow based methods, especially FB-Flow. Results produced by SSL remained unchanged.

As expected, tracking of anatomical structures **across** different subjects was significantly superior using SSL compared to the Optical Flow based methods, especially for cardiac MRI. The mean Dice Scores amounted to 0.72 ± 0.25 / 0.33 ± 0.28 / 0.35 ± 0.36 / 0.28 ± 0.30 for SSL, PCA-Flow, FB-Flow and FlowNet2-Flow, respectively. Regarding the other datasets, where the position of the organs to track is more similar, outperformance of SSL was not as pronounced.

Overall, SSL proved to be more robust compared to classical and deep learning-based Optical Flow methods, as it was not affected by large displacements or more complex non-rigid motions which are common in the medical field.

Application of pretrained Optical Flow frameworks trained in a supervised manner is not feasible for most medical imaging tasks due to the domain shift between training and test

data. Even fine-tuning on a labelled subset is not an option, as creation of such a dataset requires a dense motion field for all pixels, which is practically impossible to obtain.
A wide range of time-consuming medical imaging tasks in the fields of diagnostic- as well as interventional radiology can be tackled using this framework, as it removes the necessity to create a large and fully segmented training dataset in advance. It can be used for the purpose of inference (e.g. changes of left ventricular volume) or for training data generation.
The main limitation of this study is that only two-dimensional data was investigated. Many time series in medical imaging are four-dimensional, thus tracking of full volumes instead of individual slices could be beneficial for tracking accuracy. Adaptation of this framework to handle three-dimensional volumes will be part of future work. Furthermore, the creation of the affinity matrix, especially due to the need for a large feature map, requires an enormous amount of computational effort which limits application in fields where real-time performance is required, at least with current computational resources.

3.3 Real Time Landmark Detection for Within- and Cross Subject Tracking With Minimal Human Supervision

This section summarizes the method and results presented in [7]. The main contribution is the development of a self-supervised framework that enables tracking of anatomical structures in real-time.

Real-time, self-supervised tracking was achieved by developing and training a siamese-like CNN for a template matching task that estimates the central coordinates of a patch inside a full image. Two different encoders, one that handles the full image and one that processes the patch, were used. The stacked output of these encoders is then fed to a fully connected MLP yielding the final coordinate estimate (refer Fig. 1 in [7]). In addition to pure patch center regression, uncertainty estimates were added to the network's output, yielding information about the uncertainty of the model's prediction given an input image and patch. High uncertainty can indicate errors during tracking or domain shift of the data. During training, image patches with center coordinates (x, y) are uniformly drawn from source images. Patch and source image are then processed by the network to estimate the coordinates $(\hat{x}, \hat{y})$ of the patch center, paired with the corresponding uncertainty. This enables calculation of the negative log-likelihood between prediction $(\hat{x}, \hat{y})$ and patch centers (x, y) and thus to train the model. As, of course, a pure template matching strategy is not fruitful for cross- or within subject landmark tracking, domain specific data augmentation (rotation, scaling and gamma contrast variation) was applied to the image patches allowing for generalization beyond within-image template matching. Since the patch center coordinates need to remain unchanged, no translation was used.

In contrast to existing template matching based tracking frameworks, such as [64], this framework does not require any previous manual annotation.

During inference, landmarks to track are manually annotated in the initial image. Patches are then extracted around each landmark and fed through the network, paired with either (i) subsequent frames within the time series or (ii) different subjects.

The use of more than one labelled example image might increase performance. To account for that, an uncertainty weighted average over the individual coordinate predictions is computed to yield the final estimate.

Evaluation was carried out on three different datasets and tasks using images of shape (224×224) and squared patches of size (32, 40, 50, 60, 70, 80):

- Within-Subject liver dome tracking in abdominal MRI data from the in-house MR-LINAC
- Within-Subject liver lesion tracking in abdominal CINE MRI [65]
- Cross-Subject landmark detection in chest X-Ray images [45]

To ensure that patches can be drawn from the image borders, all images were padded to $\frac{\text{patch size}}{2}$.

Evaluation was performed using pixel-wise template matching, pixel-wise cross-image matching, and example-based landmark matching. For pixel-wise template matching, the euclidean template matching (patch was directly taken from the target image) errors for all pixels in all test images were computed. Performance of pixel-wise cross-image matching was evaluated using a cyclic error metric [59] which does not require ground truth information: Given two images X and Y (either different subjects or same subjects but different frames), in a first step a patch (x, y) is extracted from X and its location (x', y') is estimated in Y. Subsequently, based on this patch coordinate estimate, a patch is extracted from Y and a prediction about its location in X is made (x'', y''). This allows for application of any regression error metric between (x, y) and (x'', y'').

Formally, the cyclic, label-free evaluation looks as follows [1]:

$$E_{(x,y)} = |f^{\theta}(\mathbf{P_Y}(f^{\theta}(\mathbf{P_X}(x, y), \mathbf{Y})), \mathbf{X}) - (x, y)|,$$

where P_X is the patch extracted from X, P_Y the patch extracted from Y and f^{θ} the trained model.

Using this routine, the cyclic error is computed for every second pixel in the test data. Regarding the task of *within-subject* tracking, cyclic errors were computed between the slice representing the end-inspiration and end-expiration phases. For *cross-subject* detection, cyclic errors were calculated between random pairs of chest X-Ray images.

To assess the performance of example-based landmark detection, liver dome, lesion centers and nine anatomical landmarks were annotated by an experienced radiologist in MR-LINAC data, abdominal CINE MRI and chest X-Rays, respectively, in all test images.

In addition, comprehensive baseline comparison was conducted using the chest X-Ray data:

[1] based on [7], formula 4

- Fully-supervised baseline: A ResNet-50, pretrained on imageNET, was finetuned on the labelled chest X-Ray validation data to estimate the position of the nine anatomical landmarks in the test data.
- Patch-wise feature comparison: A feature embedding network was trained on squared (32×32) image patches using the SimCLR framework to transform these patches into 1024-dimensional feature vectors. Landmark detection can then be performed by first computing the features of the landmark of interest and subsequent comparison of this feature vector with the features of all 32×32 patches in the target image. The patch with the highest feature similarity to the patch describing the landmark of interest represents the final estimate.
- Supervised baseline pretrained with SimCLR: A ResNet-50 encoder was pretrained on the whole chest X-Ray dataset using SimCLR to map whole X-Ray images into 1024-dimensional feature vectors. After pretraining, the network was fine-tuned as described above in the fully-supervised baseline.

On top of to these baseline comparisons, several ablation studies were performed, including the number of available example patches (from 1 to 50), the used encoder (VGG [66], ResNet [67], DenseNet [68], and ConvNext [69]), the size of the MLP after the convolutional encoder, and the distribution used during training for uncertainty estimation.
To the best of my knowledge, this was the first work that used ConvNexts in a self-supervised framework.
All evaluations, baselines and ablation routines were conducted with a focus on runtime, as real-time performance was required.

Results

As expected, pixel-wise template matching showed good results for all datasets. Mean, averaged euclidean errors over all pixels amounted to less than 4 px, with higher errors in the background region. This is supported by the uncertainty, which largely increased towards the image margins (Refer Fig. 4 in [7]).
Cyclic evaluation of cross-image matching resulted in slightly increased errors by about 2-3 px, mainly due to higher errors in the background.
Detection of predefined landmarks yielded excellent results on all datasets. For within-subject tracking of liver dome and liver lesions, mean euclidean errors amounted to less

than 2.2 px. Accuracy of cross-subject anatomical landmark detection was slightly worse, with a mean euclidean error of 5.6 px for one annotated example and 5.0 for 10 available examples.
Among all test subjects, maximal euclidean errors were 3.1, 3.5 and 13.5 pixels for MR-LINAC, abdominal CINE MRI and chest X-Ray data, respectively, with maximal anatomical displacements of 10.1 pixels for the liver dome and 8.9 pixels for liver lesions in MR-LINAC and abdominal CINE MRI, respectively.
Overall, pixel-wise template matching and cross-image matching benefited from larger patches, whereas landmark detection errors decreased with the use of smaller patches.
The baseline comparisons are depicted in Table. 3.2 [2]

Table 3.2: Template Matching Network (TMN) vs baseline methods in chest X-Ray images: Inference times on GPU (CPU) (in ms) are reported. Top: Best euclidean landmark matching errors (mean ± standard deviation). Bottom: Comparison for one labeled example.

	Method	Euclidean Error [px]	Inference Time [ms]
Few Shot	Supervised: Baseline	8.5 ± 5.7	11 (90)
	Feature Comparison	6.2 ± 4.3	1200 (140,000)
	Supervised: Finetuned	6.3 ± 4.0	11 (90)
	TMN (proposed)	5.0 ± 3.4	17 (150)
Single Shot	Supervised: Baseline	15.6 ± 10.4	11 (90)
	Feature Comparison	14.8 ± 6.7	1200 (140,000)
	Supervised: Finetuned	14.9 ± 11.4	11 (90)
	TMN (proposed)	5.6 ± 3.8	6 (75)

Overall, the template matching approach significantly outperformed all baselines, especially when only one annotated datapoint was available. Due to the low inference time, even on CPU, real-time application is feasible.
Ablation studies revealed that at least a two-layer MLP was required for this architecture to work. One layer was not enough to yield accurate results. Replacement of the VGG16 encoder with a ResNet-50 or ConvNext encoder slightly improved results. Changing the distribution during training had only minor impact on the results (Refer Tab. 3 in [7]).

[2] Taken and adapted from [7], TABLE 2

Segmentation

Extension of this framework to segmentation tasks is trivial by patch extraction around all segmented pixels or, at least, the segmentation boundary.
Fig. 3.4 visualizes segmentation results on MR-LINAC and chest X-Ray data.

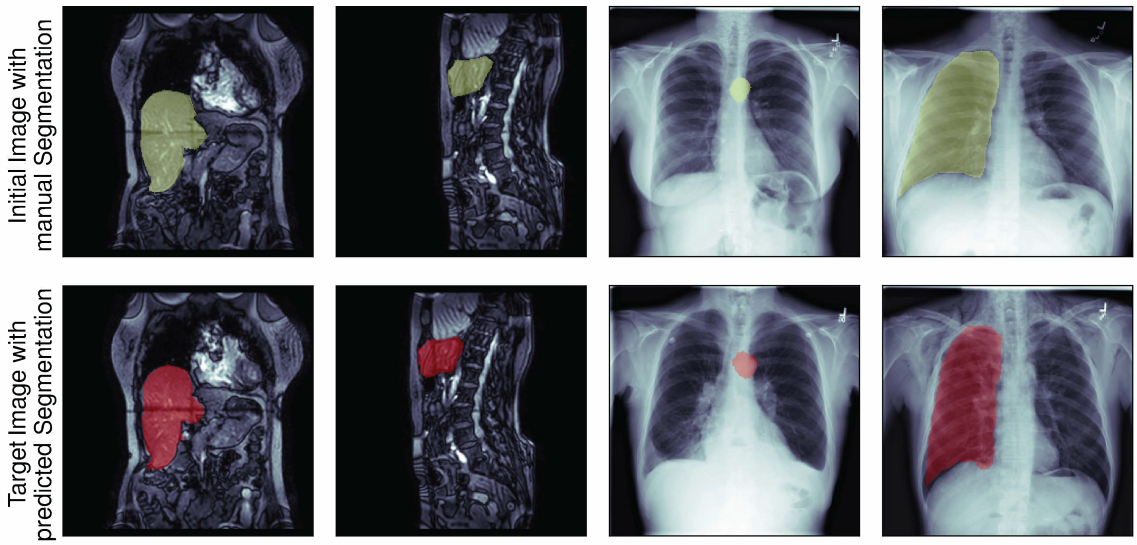


Figure 3.4: Predicted segmentations (bottom) based on one labelled example image (top) for liver, aortic knob and lung segmentation, respectively. Segmentation was performed based on contours only resulting in inference times of less than 30 ms on GPU. Dice Scores amounted to 0.94, 0.95, 0.81 and 0.94, respectively. Addition of artificial horizontal bulk motion (second column) did not affect the result.

In contrast to other tracking and detection works, this framework is capable of both anatomical landmark detection and segmentation in real-time without requiring a single label during training. A single labelled example image is sufficient to perform inference with high accuracy, rendering this approach very useful for image-guided radiotherapy due to the fact, that organ and lesion segmentation is performed prior to each session.
Pixel-wise evaluation of template-matching as well as cross-image matching revealed high performance for the foreground regions. The high uncertainty in the background indicates that this framework learns to recognize relevant structures within a dataset.
The actual tracking of anatomical landmarks, either through time or across subjects, could be performed with high accuracy for all tasks. Due to the higher similarity, within-subject tracking performance was superior compared to cross-subject anatomical landmark detec-

tion within X-Ray images.
As this framework relies on patches to describe what landmark to track, the patch size is the most important hyperparameter. Generally, patches of sizes 32 to 40 px yielded superior results for individual landmark detection compared to larger patch sizes. In contrast, for the pixel-wise evaluations, larger patches led to better performance since the number of patches that contain only background is reduced.
Obtaining an uncertainty score for each prediction is of huge importance for radiotherapy as it allows for instant termination if high uncertainty occurs possibly preventing irradiation of healthy tissue.
First possible applications could involve the definition of an area of interest during radiotherapy. As long as the landmarks to track are within this region, radiation can continue.
Second to that, this framework can be used for training data generation for any landmark detection or segmentation task. Based on only a few examples, a large dataset can be fully segmented and subsequently, after some minor validation, used for supervised learning.
The main limitation of this work is the dependence on the patch size. Future work should focus on implementing pyramidal strategies to combine several patch sizes to increase generalizability and remove the necessity of manual patch size selection.
As this framework can be seamlessly translated to 3D volumes, future work will incorporate evaluation on full volumes instead of individual slices. First results have been promising.

3.4 Large Scale, entire-volume, 3D Contrastive Learning on whole-body MRI data

This section summarizes the method and results presented in C.1. The main contribution lies in the transfer of SimCLR to the three-dimensional domain and its application on entire-volume whole-body MR scans.

Whole-body MRIs were transformed into feature vectors consisting of 512 elements using a three-dimensional ResNet-18 encoder trained in a self-supervised manner using SimCLR. During training, two different augmentations of the initial volume are created and subsequently mapped to their respective feature spaces using the ResNet encoder. Chen et al. [38] could show that adding a subsequent MLP improves performance, thus the feature vectors are once more transformed before computation of the Normalized Temperature-scaled Cross-Entropy loss [70]. Given an image tuple (i,j) sampled from the initial volume (positive pair), the following term is optimized:

$$\ell = -log\frac{\exp(\mathbf{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{k\neq i}\exp(\mathbf{sim}(z_i, z_k)/\tau)},$$

where sim(z_i, z_j) corresponds to the cosine similarity between the resulting feature vectors z_i and z_j, while τ indicates the temperature and z_k contains the feature vectors of the negative samples. This loss is computed for all positive pairs in the batch rendering the final objective.

Training was conducted on 40,000 whole-body MRI volumes ($224 \times 168 \times 163$) provided by the UK Biobank [53] using a batch size of 56 (i.e. 112 volumes are used in a batch) on 8 A100 GPUs (NVIDIA) for 200 epochs. For optimization, ADAM [71] was used together with a linear warmup scheduler up to a learning rate of 1e-3.

Evaluation was performed on 4,174 test subjects (ground truth was provided by the UK Biobank) using the following downstream tasks:

- Sex classification
- Overall health score classification (Poor to Excellent)
- Age regression
- Height regression

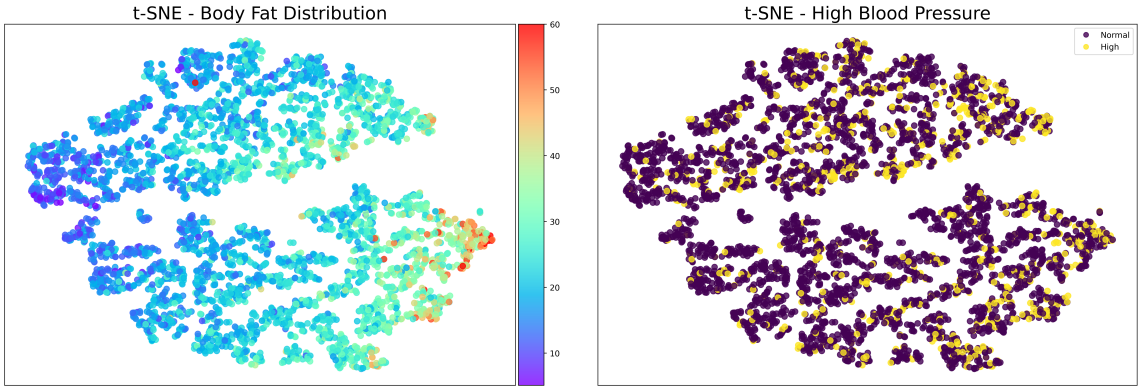


Figure 3.5: t-SNE plots of the feature vectors colorized for body fat (left, in %) and high blood pressure (right).

- Weight regression
- Body fat regression

Classification was performed using an MLP consisting of 100 neurons, whereas standard Ridge regression was used for the regression tasks. Only the feature vectors were used as input, no fine-tuning of the ResNet encoder was performed. To determine the advantage of self-supervised contrastive learning, comparison was conducted using a ResNet-18 that was trained with full supervision on 2, 20, 200 and 400 randomly selected samples.
Second to the evaluation of clinical metrics, feature representations were visualized using t-SNE plots [72].

Results

Overall, as expected, the self-supervised pretrained model significantly outperformed the supervised baseline when only limited training data (2 or 20 samples) were available. More labels led to a convergence of results for all tasks besides the health score classification, where 400 labelled samples were not enough to reach the self-supervised performance. Regarding sex classification, self-supervised pretraining achieved markedly high accuracy using only two labelled samples (one male, one female). This is explained by the t-SNE plot, which shows an almost perfect separation line between the sexes. (Fig. 4, Appendix C.1). Fig. 3.5 visualizes two additional t-SNE plots, one for the body fat distribution and one for the occurrence of high blood pressure.

To the best of my knowledge, this is the first work that performed contrastive learning on entire-volume MRI scans on a very large dataset. Evaluation revealed that contrastive learning can be used to transform MR volumes into useful feature representations without the necessity of a labelled dataset. Subsequent adaptation to a downstream task can be performed in two different ways: Retraining of the whole encoder using the learned weights as starting point, or simply training of a subsequent model using the estimated feature vectors as input. Both methods, however, allow model sharing and retraining without access to the initial dataset.

The main drawback of this framework lies in the efficacy of the labelled subset. If finetuning is performed on only a very small subset, performance may vary drastically depending on the selected data (e.g only male subjects). Evaluation of routines to find the optimal subset will be part of future work.

While compressing a full volume into a feature vector has certain advantages for estimation of high-level attributes such as age or sex, there are certain disadvantages when it comes to prediction of very specific diseases or segmentation tasks. A feature vector that contains 512 elements is just not powerful enough to capture all these properties. Application of contrastive learning on a slice- or pseudo 3D level might result in more detailed feature vectors. Future work should investigate this and conduct corresponding comparisons.

3.5 Advancing the Field through Challenges

This section summarizes the organization and results of the autoPET challenge[3] based on the dataset proposed in [6].
The autoPET - Automated Lesion Segmentation in Whole-Body FDG-PET/CT- challenge was organized to encourage research in the field of lesion segmentation in PET/CT imaging. To this end, a dataset consisting of 1014 fully annotated FDG-PET/CT scans (UKT Tübingen) was prepared and made public [6] that participants were allowed to use. A held-out test set consisting of 150 scans, assembled from two institutions (100 UKT Tübingen / 50 LMU Munich) was used for evaluation. Ranking was performed using Dice Score, false negative volume and false positive volume, with the Dice Score weighted twice.
Training of the models was conducted locally, whereas evaluation was performed on the challenge platform `https://autopet.grand-challenge.org`, to which participants submitted a docker container containing their framework paired with the model weights. The challenge consisted of two different phases. Firstly, the preliminary phase, which started in May 2022 and consisted of 5 test samples, with the goal of allowing the participants to familiarize themselves with the platform. To this end, a pretrained nn-U-Net baseline was provided. The final phase, which started in August 2022, included the 150 test subjects. To encourage participation, a prize pool of 15,000 € was offered.

Results

Throughout the challenge duration (April - September 2022), 359 teams registered from all over the world. 61% were from Asia (41 % China), 20 % from Europe and 16 % from North America. The large majority (75 %) of all participants were working within academic institutions. 253 submissions were made from 37 teams during the preliminary phase. For the final phase, 18 teams submitted their algorithms. Out of these 18 teams, the first 7 were the challenge winners and received their share of the prize pool (6000 €, 3000 €, 2000 €, 4 $\times$ 1000 €).
The winning algorithms were mostly based on a 3D U-Net (nn-U-Net) using the Dice Loss for optimization. Other frameworks were based on transformers or 2D/3D hybrid models. Mean Dice Scores ranged from 0.74 to 0.79, mean false positive volumes from 0.5 to 1.5 ml and mean false negative volumes from 2.1 to 9.5 ml. Results on the out-of-distribution test

[3] `https://autopet.grand-challenge.org/final-leaderboard`

data from Munich were significantly worse (mean Dice Scores from 0.6 to 0.7) compared to Tübingen (mean Dice Scores from 0.8 to 0.88).
Surprisingly, the performance of the winning teams did not vary much, regardless of the model chosen. In fact, the provided nn-U-Net baseline was already among the best algorithms (but of course not included in the final leaderboard), supporting the thesis that for current architectures, the quantity and quality of the training data is much more important than exhaustive model customization.
In the field of medical imaging, challenges have become the best tool to advance the field. With the organization of the autoPET 2022 challenge, a first step was taken towards automated analysis of PET/CT data. Future challenges will address the problem of multi-site, out-of-distribution test data, and will also allow for incorporation of additional training data, possibly enabling self-supervised pretraining.

4. General Discussion & Conclusion

Medical imaging remains one of the medical fields that can benefit most from automated analysis, due to the enormous possible time savings for the radiologists. The bottleneck for the broad deployment of such frameworks, however, lies in the process of training data generation.

For plain classification tasks, training data can be generated without much effort or directly extracted from existing reports. The benefit of such classification systems, however, is limited as the radiologist is usually as fast as the trained model. Of course, supportive applications still have their justification, as they can reduce the likelihood of misdiagnosis. Segmentation tasks, in contrast, require high manual effort, as the physician has to accurately segment all corresponding pixels / voxels, which is an exceptionally tedious task. Automated segmentation of medical images is thus a problem of very high relevance. Unfortunately, large and fully segmented training datasets are still a rare good, affecting the application of supervised methods and therefore delaying the development of useful segmentation models. Since only highly trained experts are able to provide this kind of data, it will probably take a very long time before training datasets are available for a satisfactory number of tasks. Thus, other ways of either data generation or training routines need to be developed and evaluated to advance the field of medical imaging.

In this work, several label-efficient or label-free frameworks were developed or transferred from the normal image domain, implemented and evaluated on several medical imaging datasets ranging from two to four dimensions.

Weakly-supervised segmentation of tumor lesions in FDG-PET/CT images was achieved by training classifier with subsequent region proposals using Class Activation Maps. Leveraging these proposals, an adaptive threshold was applied to the PET images to produce the final segmentation. While this approach proved to be very reliable for small to large tumors, very small or extremely large lesions were typically over- or underestimated, either because the SUV uptake was too small for the adaptive threshold to function properly, or

because the extracted region proposals did not include the whole lesion. Hence, application in clinical practice is feasible to obtain an initial assessment of the tumour burden, but detailed evaluation has to be carried out by an expert. The greater value, however, lies in the potential for training data generation in a research setting. Instead of manual segmentation in all subjects, the radiologist just needs to label the slices and subsequently verify or correct the predicted segmentations, resulting in enormous time savings. Due to the threshold-based routine, this approach is only viable for PET or PET/CT data. Future work should thus focus on avoiding the thresholds to enable transfer to other modalities. However, due to advances in self-supervised learning, it may be more efficient to perform pre-training without any labels with subsequent fine-tuning on a small, fully segmented dataset, rather than using weakly supervised learning.

One core objective of this work was to advance the field of anatomical tissue tracking. Two different frameworks, one based on self-supervised contrastive learning and one based on self-supervised learning with direct-inference were implemented and evaluated on various tasks and modalities. The contrastive framework was transferred from the normal image domain and adapted to the medical field, whereas the direct-inference self-supervised framework was developed directly for the task of anatomical tracking. Both approaches could be trained without the necessity of an annotated dataset, and inference was performed by providing a single (for the contrastive framework) or arbitrarily many (for the self-supervised direct framework) labelled examples. While in theory both approaches are suitable to track landmarks and segmentations, the contrastive framework failed to deliver accurate results when tracking landmarks due to the feature-based comparison. If the exact spot of the landmark does not match the most similar feature vector (but its close surroundings, which are annotated as zero), the landmark will erroneously also be labelled as zero. Even if the euclidean distance between ground truth and location of the most similar feature vector is very small, the evaluation still fails because nothing is annotated in the subsequent image and thus no landmark is tracked. The self-supervised direct framework, in contrast, showed great performance for both tracking of landmarks and segmentations.

The main drawback of the contrastive method lies in its inference time. Due to the expensive feature comparison to obtain the affinity matrix, in combination with the requirement of the large feature map (128^4 operations per frame pair), real-time application is not achievable with current hardware. In contrast, application of the self-supervised, direct-inference framework, was possible in less than 11 ms for landmark tracking and 25-30 ms when tracking segmentations. This allows its use in real-time applications such as image-guided

radiotherapy.

Uncertainty quantification together with inference is of highest importance in medicine, as ”unsecure” predictions can lead to catastrophic failures. In Germany, according to the second AI roadmap published by the *Deutsche Institut für Normung* (DIN) [73], uncertainty quantification should be a requirement for all AI based frameworks in the medical field. While this comes natural with the self-supervised, direct-inference framework by adding the corresponding uncertainty neurons in the output layer, it is not as trivial for contrastive learning. One strategy that typically works well is to train an ensemble and calculate the variance of the predicted results. This, however, scales linearly with inference time and is thus not feasible for real-time applications. First approaches, not relying on ensembling, have been proposed very recently [74] and show promising results.

Due to the lack of computational resources and available data, 3D contrastive learning on full volumes has not been feasible in the past. Advanced GPUs, paired with a large dataset provided by the UK Biobank, enabled this challenge to be addressed. To this end, a 3D ResNet-18 has been implemented into the SimCLR framework to map the entire volume to a single feature representation. Semi-supervised evaluation revealed that the amount of labels required to fine tune such a pretrained network on a downstream task can be drastically reduced. More interesting than that is the ability to visualize entire populations (or at least the subgroup that had an MRI taken) by application of a t-SNE plot to the predicted feature embeddings. Another useful application is to find the most similar scans to the current one. In the future, this might be used to suggest scans that have been evaluated in the past, improving diagnostic accuracy. Overall, performance and benefits on more complicated and useful tasks still have to be evaluated. Currently, due to the still limited batch size and feature dimension, application on two-dimensional slices might be more promising.

Last but not least, leaving the field of label-efficient learning, the autoPET challenge was planned and organized to motivate deep learning researchers to work on the highly relevant field of lesion segmentation in PET/CT data. By providing a large and fully annotated dataset as well as a platform for automatized inference, 18 teams were encouraged to submit their final results, of which 7 won a part of the prize money. Solutions were mostly based on the nn-U-Net and overall scores did not vary significantly on the held out within-domain test set. In contrast, the out-of-domain data from Munich resulted in a higher variance among the results. Thus, for the second challenge, which is currently in the preliminary phase, a higher focus is set on these kind of problems.

In conclusion, the results of this work are promising that the medical imaging domain benefits from label efficient frameworks, especially in the field of tissue tracking. Using these kind of techniques will allow for more precise and thus more efficient radiotherapy in the future and therefore has the potential to change the outcome of a tumor burden for good.

Bibliography

[1] L. Rubén Braojos et al., School of Engineering, "They weren't sure it could be done – an artificially intelligent coffee machine." https://sti.epfl.ch/they-werent-sure-it-could-be-done-an-artificially-intelligent-coffee-machine/, 2016 (accessed November 7, 2022).

[2] B. für Digitales und Verkehr, "Gesetz zum autonomen fahren tritt in kraft." https://www.bmdv.bund.de/SharedDocs/DE/Artikel/DG/gesetz-zum-autonomen-fahren.html, 2021 (accessed November 7, 2022).

[3] A. e. Klaus Tenning, "Laborproben in luftigen höhen." https://www.alm-ev.de/aktivitaeten/labor-erleben-magazin/neue-wege-gehen/, 2022 (accessed November 7, 2022).

[4] W. H. O. Miss Elizabeth Nabunya Kawooy, "To x-ray or not to x-ray?." https://www.who.int/news-room/feature-stories/detail/to-x-ray-or-not-to-x-ray-, 2016 (accessed November 6, 2022).

[5] M. Häggström, "Pneumothorax." https://commons.wikimedia.org/wiki/File:Expired_X-ray_of_pneumothorax.jpg, 2018 (accessed November 7, 2022).

[6] S. Gatidis, T. Hepp, M. Früh, C. La Fougère, K. Nikolaou, C. Pfannenberg, B. Schölkopf, T. Küstner, C. Cyran, and D. Rubin, "A whole-body fdg-pet/ct dataset with manually annotated tumor lesions," *Scientific Data*, vol. 9, no. 1, pp. 1–7, 2022.

[7] M. Frueh, A. Schilling, S. Gatidis, and T. Kuestner, "Real time landmark detection for within-and cross subject tracking with minimal human supervision," *IEEE Access*, vol. 10, pp. 81192–81202, 2022.

[8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[9] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[10] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[11] H. Mohsen, E.-S. A. El-Dahshan, E.-S. M. El-Horbaty, and A.-B. M. Salem, "Classification using deep learning neural networks for brain tumors," *Future Computing and Informatics Journal*, vol. 3, no. 1, pp. 68–71, 2018.

[12] O. Stephen, M. Sain, U. J. Maduh, and D.-U. Jeong, "An efficient deep learning approach to pneumonia classification in healthcare," *Journal of healthcare engineering*, vol. 2019, 2019.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[14] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[15] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[16] P. F. Christ, F. Ettlinger, F. Grün, M. E. A. Elshaera, J. Lipkova, S. Schlecht, F. Ahmaddy, S. Tatavarty, M. Bickel, P. Bilic, *et al.*, "Automatic liver and tumor segmentation of ct and mri volumes using cascaded fully convolutional neural networks," *arXiv preprint arXiv:1702.05970*, 2017.

[17] X. Zhao, L. Li, W. Lu, and S. Tan, "Tumor co-segmentation in pet/ct using multi-modality fully convolutional neural network," *Physics in Medicine & Biology*, vol. 64, no. 1, p. 015011, 2018.

[18] W. Li *et al.*, "Automatic segmentation of liver tumor in ct images with deep convolutional neural networks," *Journal of Computer and Communications*, vol. 3, no. 11, p. 146, 2015.

[19] P. Hu, F. Wu, J. Peng, P. Liang, and D. Kong, "Automatic 3d liver segmentation based on deep learning and globally optimized surface evolution," *Physics in Medicine & Biology*, vol. 61, no. 24, p. 8676, 2016.

[20] T. Kart, M. Fischer, T. Küstner, T. Hepp, F. Bamberg, S. Winzeck, B. Glocker, D. Rueckert, and S. Gatidis, "Deep learning-based automated abdominal organ segmentation in the uk biobank and german national cohort magnetic resonance imaging studies," *Investigative Radiology*, vol. 56, no. 6, pp. 401–408, 2021.

[21] E. Smistad, A. Østvik, *et al.*, "2d left ventricle segmentation using deep learning," in *2017 IEEE international ultrasonics symposium (IUS)*, pp. 1–4, IEEE, 2017.

[22] M. R. Avendi, A. Kheradvar, and H. Jafarkhani, "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri," *Medical image analysis*, vol. 30, pp. 108–119, 2016.

[23] M. Seregni, C. Paganelli, P. Summers, M. Bellomi, G. Baroni, and M. Riboldi, "A hybrid image registration and matching framework for real-time motion tracking in mri-guided radiotherapy," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 1, pp. 131–139, 2017.

[24] H. Mi, C. Petitjean, B. Dubray, P. Vera, and S. Ruan, "Prediction of lung tumor evolution during radiotherapy in individual patients with pet," *IEEE transactions on medical imaging*, vol. 33, no. 4, pp. 995–1003, 2014.

[25] D. Tenbrinck, S. Schmid, X. Jiang, K. Schäfers, and J. Stypmann, "Histogram-based optical flow for motion estimation in ultrasound imaging," *Journal of mathematical imaging and vision*, vol. 47, no. 1, pp. 138–150, 2013.

[26] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2758–2766, 2015.

[27] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2462–2470, 2017.

[28] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European conference on computer vision*, pp. 402–419, Springer, 2020.

[29] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

[30] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 685–694, 2015.

[31] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, "Weakly-supervised learning of visual relations," in *Proceedings of the ieee international conference on computer vision*, pp. 5179–5188, 2017.

[32] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

[33] K. Sun, H. Shi, Z. Zhang, and Y. Huang, "Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7283–7292, 2021.

[34] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[35] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*, pp. 69–84, Springer, 2016.

[36] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906*, 2021.

[37] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[38] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[39] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2020.

[40] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.

[41] A. Bardes, J. Ponce, and Y. LeCun, "Vicregl: Self-supervised learning of local visual features," *arXiv preprint arXiv:2210.01571*, 2022.

[42] A. Pirinen, E. Gärtner, and C. Sminchisescu, "Domes to drones: Self-supervised active triangulation for 3d human pose reconstruction," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[43] L. Madhuanand, F. Nex, and M. Y. Yang, "Self-supervised monocular depth estimation from oblique uav videos," *ISPRS journal of photogrammetry and remote sensing*, vol. 176, pp. 1–14, 2021.

[44] Z.-Z. Wu, T. Weise, Y. Wang, and Y. Wang, "Convolutional neural network based weakly supervised learning for aircraft detection from remote sensing image," *IEEE Access*, vol. 8, pp. 158097–158106, 2020.

[45] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 590–597, 2019.

[46] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, *et al.*, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Scientific data*, vol. 8, no. 1, pp. 1–8, 2021.

[47] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, *et al.*, "Big self-supervised models advance medical image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3478–3488, 2021.

[48] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20730–20740, 2022.

[49] M. Früh, M. Fischer, A. Schilling, S. Gatidis, and T. Hepp, "Weakly supervised segmentation of tumor lesions in pet-ct hybrid imaging," *Journal of Medical Imaging*, vol. 8, no. 5, p. 054003, 2021.

[50] M. Frueh, T. Kuestner, M. Nachbar, D. Thorwarth, A. Schilling, and S. Gatidis, "Self-supervised learning for automated anatomical tracking in medical image data with minimal human labeling effort," *Computer Methods and Programs in Biomedicine*, vol. 225, p. 107085, 2022.

[51] K. Nie, Y.-X. Zhang, W. Nie, L. Zhu, Y.-N. Chen, Y.-X. Xiao, S.-Y. Liu, and H. Yu, "Prognostic value of metabolic tumour volume and total lesion glycolysis measured by 18f-fluorodeoxyglucose positron emission tomography/computed tomography in small cell lung cancer: a systematic review and meta-analysis," *Journal of Medical Imaging and Radiation Oncology*, vol. 63, no. 1, pp. 84–93, 2019.

[52] A. Jabri, A. Owens, and A. Efros, "Space-time correspondence as a contrastive random walk," *Advances in neural information processing systems*, vol. 33, pp. 19545–19560, 2020.

[53] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, *et al.*, "Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS medicine*, vol. 12, no. 3, p. e1001779, 2015.

[54] S. Gatidis, M. Früh, M. Fabritius, S. Gu, K. Nikolaou, C. La Fougère, J. Ye, J. He, Y. Peng, L. Bi, *et al.*, "The autopet challenge: Towards fully automated lesion segmentation in oncologic pet/ct imaging. preprint at research square (nature portfolio)(2023)."

[55] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[56] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847, IEEE, 2018.

[57] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 24–25, 2020.

[58] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[59] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *2010 20th international conference on pattern recognition*, pp. 2756–2759, IEEE, 2010.

[60] "Data science bowl cardiac challenge data." `https://www.kaggle.com/c/second-annual-data-science-bowl/data`. Accessed: 2010-09-30.

[61] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, *et al.*, "Video-based ai for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, 2020.

[62] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*, pp. 363–370, Springer, 2003.

[63] J. Wulff and M. J. Black, "Efficient sparse-to-dense optical flow estimation using a learned basis and layers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 120–130, 2015.

[64] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *European conference on computer vision*, pp. 850–865, Springer, 2016.

[65] C. Würslin, H. Schmidt, P. Martirosian, C. Brendle, A. Boss, N. F. Schwenzer, and L. Stegger, "Respiratory motion correction in oncologic pet using t1-weighted mr imaging on a simultaneous whole-body pet/mr system," *Journal of nuclear medicine*, vol. 54, no. 3, pp. 464–471, 2013.

[66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[68] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[69] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[70] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," *Advances in neural information processing systems*, vol. 29, 2016.

[71] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[72] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[73] S. Maack, P. Benner, M. Kröll, J. Prager, W. Daum, R. Casperson, T. Heckel, D. Spaltmann, *et al.*, "Deutsche normungsroadmap künstliche intelligenz ausgabe 2," 2022.

[74] T. Wang, J. Lu, Z. Lai, J. Wen, and H. Kong, "Uncertainty-guided pixel contrastive learning for semi-supervised medical image segmentation," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pp. 1444–1450, 2022.

A. Accepted Peer-Reviewed Journal Papers

A.1 Weakly supervised segmentation of tumor lesions in PET-CT hybrid imaging

Authors: *Marcel Früh, Marc Fischer, Andreas Schilling, Sergios Gatidis, Tobias Hepp*

Published in: *Journal of Medical Imaging*

Date of Publication: *October 2021*

Weakly supervised segmentation of tumor lesions in PET-CT hybrid imaging

Marcel Früh,[a,b,*] Marc Fischer,[c] Andreas Schilling,[b] Sergios Gatidis,[a,d] and Tobias Hepp[a,d]

[a]University Hospital Tübingen, Department of Diagnostic and Interventional Radiology, Tübingen, Germany

[b]University of Tübingen, Institute for Visual Computing, Department of Computer Science, Tübingen, Germany

[c]University of Stuttgart, Institute of Signal Processing and System Theory, Stuttgart, Germany

[d]Max Planck Institute for Intelligent Systems, Max Planck Ring 4, Tübingen, Germany

Abstract

Purpose: We introduce and evaluate deep learning methods for weakly supervised segmentation of tumor lesions in whole-body fluorodeoxyglucose-positron emission tomography (FDG-PET) based solely on binary global labels ("tumor" versus "no tumor").

Approach: We propose a three-step approach based on (i) a deep learning framework for image classification, (ii) subsequent generation of class activation maps (CAMs) using different CAM methods (CAM, GradCAM, GradCAM++, ScoreCAM), and (iii) final tumor segmentation based on the aforementioned CAMs. A VGG-based classification neural network was trained to distinguish between PET image slices with and without FDG-avid tumor lesions. Subsequently, the CAMs of this network were used to identify the tumor regions within images. This proposed framework was applied to FDG-PET/CT data of 453 oncological patients with available manually generated ground-truth segmentations. Quantitative segmentation performance was assessed for the different CAM approaches and compared with the manual ground truth segmentation and with supervised segmentation methods. In addition, further biomarkers (MTV and TLG) were extracted from the segmentation masks.

Results: A weakly supervised segmentation of tumor lesions was feasible with satisfactory performance [best median Dice score 0.47, interquartile range (IQR) 0.35] compared with a fully supervised U-Net model (median Dice score 0.72, IQR 0.36) and a simple threshold based segmentation (Dice score 0.29, IQR 0.28). CAM, GradCAM++, and ScoreCAM yielded similar results. However, GradCAM led to inferior results (median Dice score: 0.12, IQR 0.21) and was likely to ignore multiple instances within a given slice. CAM, GradCAM++, and ScoreCAM yielded accurate estimates of metabolic tumor volume (MTV) and tumor lesion glycolysis. Again, worse results were observed for GradCAM.

Conclusions: This work demonstrated the feasibility of weakly supervised segmentation of tumor lesions and accurate estimation of derived metrics such as MTV and tumor lesion glycolysis.

 [DOI: 10.1117/1.JMI.8.5.054003]

Keywords: weakly supervised learning; deep learning; label efficiency; positron emission tomography; computed tomography; oncological imaging.

Paper 21132R received May 25, 2021; accepted for publication Oct. 1, 2021; published online Oct. 13, 2021.

*Address all correspondence to Marcel Früh, Marcel.Frueh@med.uni-tuebingen.de

1 Introduction

Contrast-enhanced computed tomography (CT) remains the backbone for oncological staging, whereas 18-fluordesoxyglucose ([18F]-FDG) positron emission tomography (PET)/CT hybrid imaging plays a central role in the detection of distant metastatic disease.[1] In addition to the detection of tumor spots, FDG-PET provides essential functional information about the tumor metabolism.[2] For instance, the maximum standardized uptake value (SUV) for FDG of primary tumors is a prognostic biomarker for survival in non-small cell lung cancer.[2] In addition to the maximum SUV, state-of-the-art metrics for assessing tumor burden also include the metabolic tumor volume (MTV) and total lesion glycolysis (TLG).[3] Although this information is, in principle, available in routine examinations, the evaluation can imply the manual analysis of a large number of single lesions and thus proves to be problematic in everyday clinical practice and in the exploration of large cohorts. Computer-aided automatic detection and segmentation of tumor lesions is therefore of great importance in PET/CT imaging. In recent years, significant progress has been made in the automatic analysis of medical images, mainly due to the emergence of deep learning methods.[4,5] Deep learning models have already been successfully applied for the detection and segmentation of tumor lesions.[6] Established approaches are mostly based on supervised learning schemes[7] that use a large amount of manually voxel-wise annotated ground-truth data. However, acquiring ground-truth data, in particular for many small tumor lesions, is time consuming and requires an enormous manual labeling effort of an experienced radiologist. Advances in machine learning are pointing to methods that allow learning with a smaller amount of annotated training data.[8] Whereas semi- and self-supervised learning try to boost performance by utilizing unlabeled data, weakly supervised learning reduces the complexity of the label and therefore simplifies the collection of ground-truth annotations. Following the second approach, the location of objects in natural images can be learned to a limited extent from a weaker annotation such as a classification of the imaged object of interest, instead of an actual voxel-wise mask (i.e., the full positional information).[9] Previous studies demonstrate the potential of weakly supervised segmentation based on bounding boxes,[10] scribbles,[11] or image level class labels.[12] In this work, we propose a framework for weakly supervised segmentation of tumor lesions in full-body PET/CT images of patients with cancer. Thus, only a binary slice-by-slice specification of whether malignant tissue is present or not is used as a weak supervision signal. A convolutional neural network (CNN) acts as a classifier. Subsequently, a threshold-based analysis of class activation maps (CAM) is utilized to generate the segmentation mask. We evaluate our proposed approach for different CAM methods and compare its performance in predicting TLG and MTV with supervised segmentation approaches for PET/CT images of oncological patients with lung cancer, lymphoma, and malignant melanoma.

1.1 *Related Work*

The use of CAM for weakly supervised object detection and segmentation has been reported, including in the medical imaging domain. Afshari et al. proposed a FCN architecture for PET lesion segmentation based on bounding boxes and the unsupervised Mumford-Shah segmentation model.[13] Nguyen et al.[14] used GradCAM paired with a ResNet50 to segment uveal melanoma lesions in MRI images. Subsequently, after applying a conditional random field, they trained a U-Net on predicted segmentation masks, which achieved Dice scores similar to the supervised counterpart. Recently, Eyuboglu et al.[15] proposed a weakly supervised method that uses a BERT language model[16] to extract regional abnormality labels from free-text radiology reports of PET/CT examinations. Subsequently, they trained a CNN-based classifier on these labels to automatically detect if there are abnormalities in a certain anatomical region.

2 Materials and Methods

2.1 *Dataset*

In this study, we included full body PET/CT scans of 453 oncological patients (195 females, 258 males) acquired between 2013 and 2016 from an ongoing PET/CT registry study in our

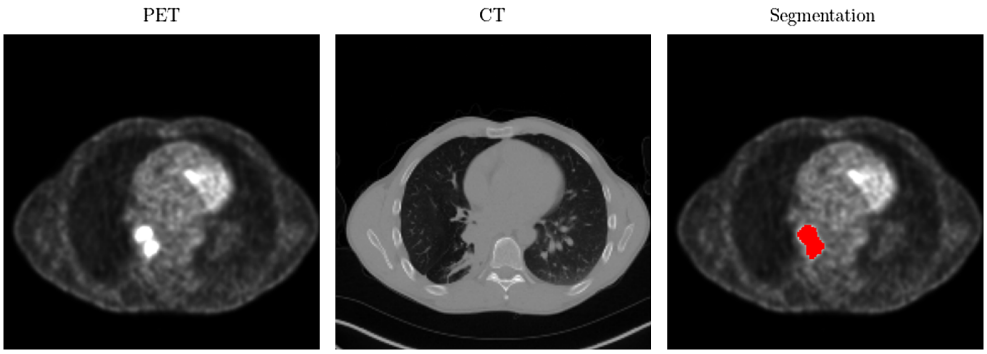


Fig. 1 Exemplary PET/CT slice with high SUV uptake next to the hilum of the right lung. The right image shows the manually annotated segmentation mask as red overlay to the PET image.

hospital.[17] The distribution of oncological diagnoses was as follows: 50% lung cancer, 18% lymphoma, and 32% malignant melanoma. The median age was 64 years (19–95 years). All examinations were performed using standardized protocols including state-of-the-art CT with an intravenous contrast agent (Biograph mCT, Siemens Healthineers, Germany). [18F]-FDG was applied as the PET tracer. The registry study was approved by the Ethics Committee of the University of Tübingen, reference number 064/2013B01.

2.1.1 *Pre-processing*

Voxel-wise SUVs were computed from attenuation corrected PET images.[18] SUV images were pre-aligned and resampled to the resolution of the corresponding CT images by means of linear interpolation (spatial resolution of $2 \times 2 \times 3$ mm, in-slice shape 256×256). To evaluate the performance of the model, a subject level train-validation-test split (60%–20%–20%) was used. All tumor lesions were manually annotated by an experienced radiologist in a slice-by-slice manner (Fig. 1). A slice-wise binary label, which indicates if malignant tissue is present or not, was derived from the segmentation masks as a weak supervision signal.

2.1.2 *Data description*

The median tumor volume was 46.5 ml [interquartile range (IQR) 158.4 ml]. Overall, only 13.5%–14% of the training/test set image slices contained malignant tissue. As shown in Fig. 2, the right skewed distribution of the tumor size within slices reflects a dominance of slices with small tumor proportions.

2.2 *Methods—Weakly Supervised Tumor Segmentation*

First, we describe the proposed method for weakly supervised segmentation. A detailed description of the network architecture as well as the derivation of the utilized CAM methods is given. Finally, we summarize the training routine, the baseline methods and the evaluation methodology.

2.2.1 *Weakly supervised segmentation*

The purpose of weakly supervised segmentation is to achieve a well-performing segmentation model without the need for manually annotated ground-truth segmentation masks. Weak labels (e.g., class labels or bounding boxes) are typically easy to gather and correlate directly with the segmentation mask. Our framework generates a segmentation mask prediction in three separate steps. First, a tumor classification network is trained with the provided slice-level binary labels (tumor/no tumor). Second, CAM methods are used to identify regions that are relevant to the networks decision. An adaptive unsupervised threshold-based image segmentation is applied to the region proposed by the CAM algorithm, yielding the tumor segmentation.

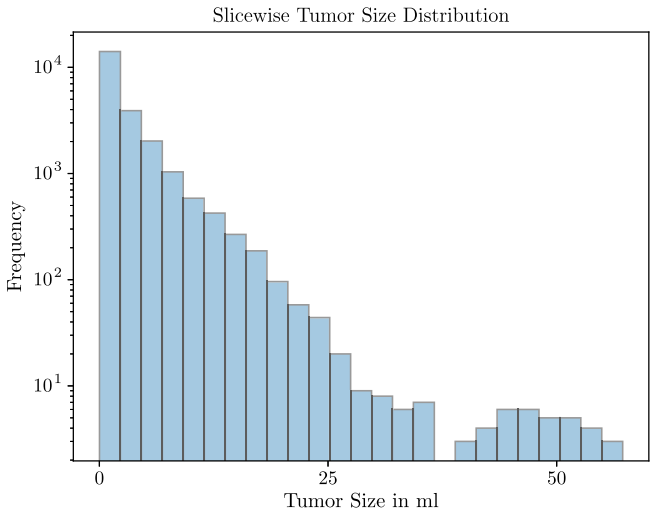


Fig. 2 Distribution of the tumor size for slices with malignant tissue. Slices with small sized tumors are dominating.

Architecture. For the slice-wise classification task, a CNN with VGG-16 base architecture[19] was utilized. The weights of the network were pretrained on Imagenet.[20] By removing the first max-pooling layer of the network, the size of the final feature map was increased to 32×32. Preprocessed PET and CT image slices form the two input channels of the network. The output of the network yields the probability of the slice containing one or more FDG-avid tumor lesions.

Class activation maps. Neural networks form a class of highly non-linear functions, and there is no general recipe for explaining the relevance of input features for the final prediction. One common approach is to visualize the saliency of regions of the input image with respect to the prediction of a CNN. These saliency maps are called CAM[9] (Fig. 3). Four different established methods to derive CAMs were compared in this study.

The classic CAM[9] algorithm requires a specific network architecture with a single fully connected layer following the final global average pooling layer of the convolutional part of the network. The activation map M for class c is computed as the dot product between feature map A_k with k filters of the last convolutional layer of the network and weights w for class c from the fully connected layer:

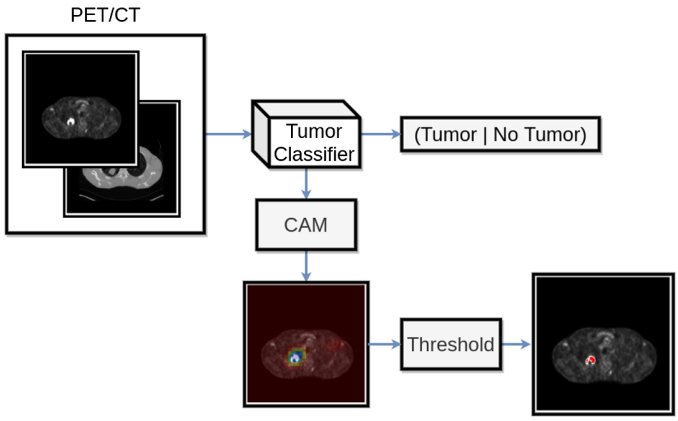


Fig. 3 Proposed processing routine. First, a binary tumor classifier is trained in a supervised manner on PET/CT data. Then a class activation map is computed based on the classifier. Finally, threshold based segmentation is performed on the PET images within the region proposed by the CAM.

$$M_c = \sum_k w_k^c A_k. \tag{1}$$

Compared with CAM, GradCAM[21] shows more flexibility regarding the network architecture. CAM M_c is computed by scaling corresponding feature map A of the last convolutional layer with the gradients of prediction $\hat{y}$ for class c with respect to the elements of A via backpropagation followed by global average pooling:

$$\delta^c = \frac{1}{N} \sum_h \sum_w \frac{\partial \hat{y}^c}{\partial A_{hw}^k}. \tag{2}$$

Subsequently, the linear combination between δ^c and feature map A^k is calculated to compute M_c:

$$M_c = \max\left(\sum_k \delta_k^c A^k, 0\right). \tag{3}$$

GradCAM lacks performance if multiple instances of the same class occur within one image as the focus on one object of class c is enough to yield the corresponding prediction.[22] Often only fragments of the object are considered as these are already sufficient for an accurate classification. This is particularly relevant in tumor segmentation, in which multiple tumor spots regularly appear on a single slice.

GradCAM++[22] tackles this problem by weighting the non-negative gradient of the last convolutional layer with respect to a specific class:

$$M_c = \sum_h \sum_w \alpha_{hw}^{kc} \cdot \max\left(\frac{\partial \hat{y}^c}{\partial A_{hw}^k}, 0\right), \tag{4}$$

where α_{hw}^{kc} is defined as

$$\alpha_{hw}^{kc} = \frac{\frac{\partial^2 \hat{y}^c}{(\partial A_{hw}^k)^2}}{2\frac{\partial^2 \hat{y}^c}{(\partial A_{hw}^k)^2} + \sum_i \sum_j A_{ij}^k \left[\frac{\partial^3 \hat{y}^c}{(\partial A_{hw}^k)^3}\right]}, \tag{5}$$

with i and j indexing over the slice dimensions.

ScoreCAM,[23] just like CAM, does not rely on gradients to derive a CAM M. The input image B is perturbed with the predicted, up-sampled, and normalized feature maps A. For each of these disturbed images, new feature maps A' are computed by forward passes through the network. All A' are subtracted from the original feature map A of the input image B. A subsequent softmax operation yields weights α^c of the following linear combination:

$$M_c = \max\left(\sum_k \alpha_k^c A^k, 0\right). \tag{6}$$

Adaptive threshold. By applying a CAM-method-specific CAM-threshold t_m to the CAMs, a binary regional candidate mask for the tumor area is derived. Thresholded CAMs are upscaled from 32×32 pixels to the original image size by means of nearest neighbour interpolation.

The segmentation mask is subsequently derived by selecting all positions with values larger than a method-specific but fixed percentile q_m of the SUV distribution inside the masked region. Data-specific hyperparameters in the form of CAM-thresholds and intensity percentiles were determined empirically on the training and validation sets. The percentile q_m was empirically determined by performing grid search on the training data with 20 linearly spaced values between 20 and 50. The threshold value t_m was determined in the same manner with ten linear

Algorithm 1 CAM Segmentation

Input: PET/CT slice **X**, Percentile q_m, Adaptive Threshold t_m;

1: Predict class $\hat{y}$ of **X** (Does **X** contain a tumor or not?);

If $\hat{y}$ is tumorous **then**

$\mathbf{H}$ = CAM(**X**)

Else

return Empty segmentation mask

End

2: $H' \leftarrow$ Mask all values $\geq t_m$;

3: Upscale H' from 32x32 pixels (size of the CAM) to the size of X;

4: $H'' \leftarrow X \odot H'$;

5: $t_q \leftarrow$ Calculate the percentile q_m of H'';

6: Segmentation mask = $H \geq t_q$;

Output: Segmentation mask;

spaced values from 0.1 to 0.9. The best values were determined by maximizing the Dice score on the validation data.

Segmentation routine. The complete segmentation routine is presented in the algorithm below.

2.3 *Baselines*

To evaluate the performance of our method, we compared our results with two baselines: a simple global threshold-based segmentation method and a fully supervised U-Net-CNN model.[4]

2.3.1 *Global threshold*

A global threshold based on a fixed SUV percentile was applied to all images in which the classification network predicts a tumor. The percentile was again empirically determined by performing a grid search on the training data with 20 linearly spaced values between 20 and 50 and choosing the one that yielded the highest Dice score.

2.3.2 *Supervised UNET*

We compared our approach with a standard UNET[4] segmentation model trained in a supervised manner on image slices. Our architecture consists of four double convolution layers in both, having the decoder and encoder with skip connections between all levels.

2.3.3 *Training*

As described above, a modified VGG16[19] backbone was used as the tumor classification network. Data augmentation, including slice-wise scaling, rotations, translations, and contrast changes, was applied.[24] The model was implemented using the deep learning framework PyTorch (1.7.1).[25] The network was trained for 50 epochs using a SGD optimizer with a momentum of 0.9,[26] a learning rate of 0.001, and a batch size of 64. To consider class-imbalance, a weighted cross entropy loss ($w = 7.7$) was used.

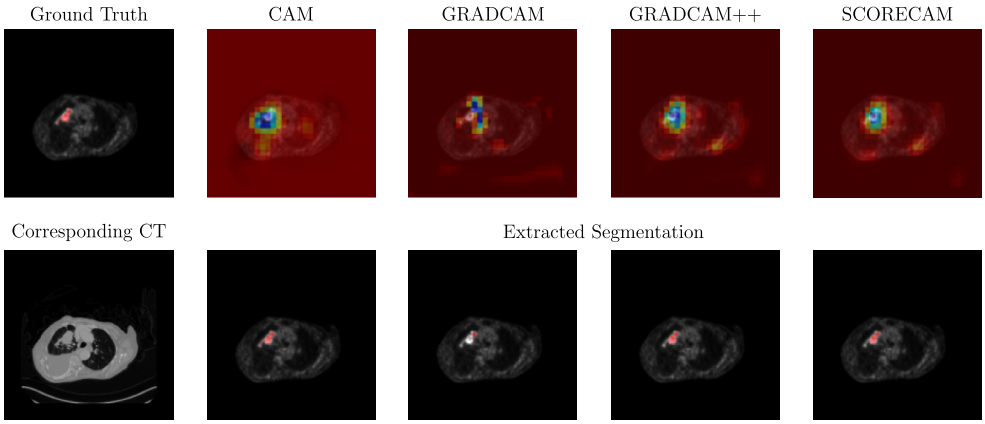


Fig. 4 PET with ground truth segmentation, corresponding activation map based on the four CAM methods, extracted segmentation and corresponding CT for a sample slice with a tumor.

The baseline U-Net model was trained on 2D image slices with a batch size of 64 for 200 epochs using the ADAM optimizer[27] ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with an initial learning rate of $5e-5$. Again, a weighted ($w = 7.7$) cross entropy loss was used. The same data augmentation for the classifier was used.

A dedicated GPU (Tesla V100, NVIDIA, Santa Clara) was used for accelerated computing.

2.3.4 *Statistical analysis*

All results are reported with median and IQR. Additionally for all segmentation methods, intra-class correlations (two-way, agreement) between ground truth annotation and prediction were computed. A global significance level of 0.05 was used.

2.4 *Evaluation*

Our proposed framework and the baselines were evaluated for 90 test subjects. The metrics 3D Dice score (compared with manual ground truth), MTV, and TLG deviation were computed for each patient.

The Dice score is defined as

$$\frac{2|A \cap B|}{|A| + |B|},$$

where A and B are the sets of voxels inside the ground truth and predicted segmentation mask, respectively. The MTV quantifies the volume of tumor regions with high metabolism. TLG is defined as the product of the mean SUV and MTV.[28]

3 Results

3.1 *Weakly Supervised Tumor Segmentation*

The following threshold values (t_m) were derived for CAM, GradCAM, GradCAM++, and ScoreCAM activation maps: 0.3, 0.2, 0.3, and 0.4, respectively. The following SUV percentile thresholds (q_m) were applied: 0.31 for CAM, 0.35 for GradCAM, 0.32 for GradCAM++, and 0.31 for ScoreCAM. Fig. 4 depicts the activation maps based on the four different methods and the corresponding segmentation for a sample slice with a tumor.

3.1.1 *Dice score*

Overall, the supervised U-Net model showed the best performance with a median Dice score of 0.72 (IQR 0.36) (Fig. 5). ScoreCAM and CAM produced the best results of all weakly

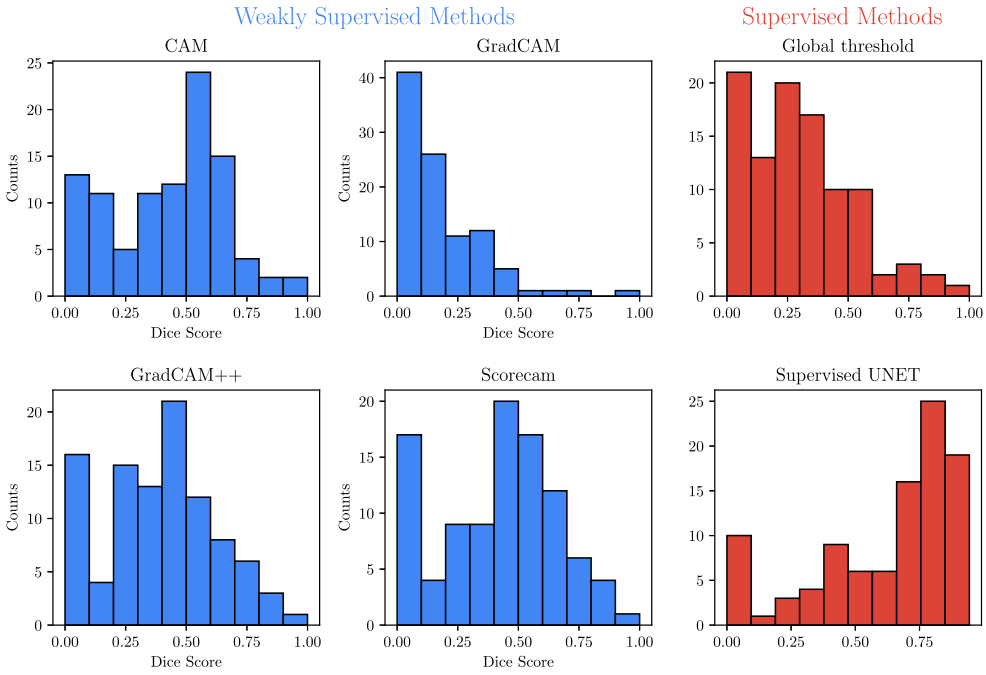


Fig. 5 Per subject Dice scores for the weakly supervised segmentation methods (blue) and the supervised baselines (red).

supervised methods with a median Dice score of 0.47 (IQR 0.35) and 0.46 (IQR 0.35), respectively. GradCAM++ performed slightly worse with a median Dice score of 0.42 (IQR 0.30). GradCAM, which achieved a median Dice score of 0.12 (IQR 0.21), showed significantly worse results. The global threshold method achieved a median Dice of 0.29 (IQR 0.28).

3.1.2 *Evaluation of MTV*

The supervised U-Net again showed the best results for the MTV estimation with a median difference of 17 ml (IQR 27 ml). Small tumors were slightly overestimated (Fig. 6). ScoreCAM (median difference 27 ml, IQR 48 ml), GradCAM++ (median difference 24 ml, IQR 48 ml), and CAM (median difference 26, IQR 68 ml) provided similar results. GradCAM again revealed inferior results with a median difference of 30 ml (IQR 76 ml). For all weakly supervised methods, an overestimation of small tumors and underestimation of large tumors was observed. This characteristic was most prominent in GradCAM and CAM. Using the global threshold baseline method also yielded a strong overestimation of smaller tumors and an underestimation of larger tumors (median difference 44 ml, IQR 92 ml). Those results are further validated by the ICC compared with the manual ground truth segmentation, which showed very similar scores and confidence intervals for CAM, GradCAM++, and ScoreCAM. GradCAM in contrast showed a significantly lower ICC (Table 1). Again, the supervised U-Net showed the highest scores and smallest confidence intervals, whereas the global threshold performed worse than CAM, GradCAM++, and ScoreCAM.

3.1.3 *Evaluation of TLG*

Tumor lesion glycolysis was predicted accurately by all methods except for GradCAM. Again, the supervised U-Net yielded the best results with a median TLG deviation of 50 g (IQR 110 g). No significant over- or underestimation was observed. (Fig. 7) ScoreCAM (median deviation of 99 g, IQR 285 g), GradCAM++ (median deviation 108 g, IQR 267 g), and CAM (median deviation 101 g, IQR 219 g) again achieved closely similar results. GradCAM (median deviation 112 g, IQR 482 g) showed the highest error with overall underestimation of TLG. In general

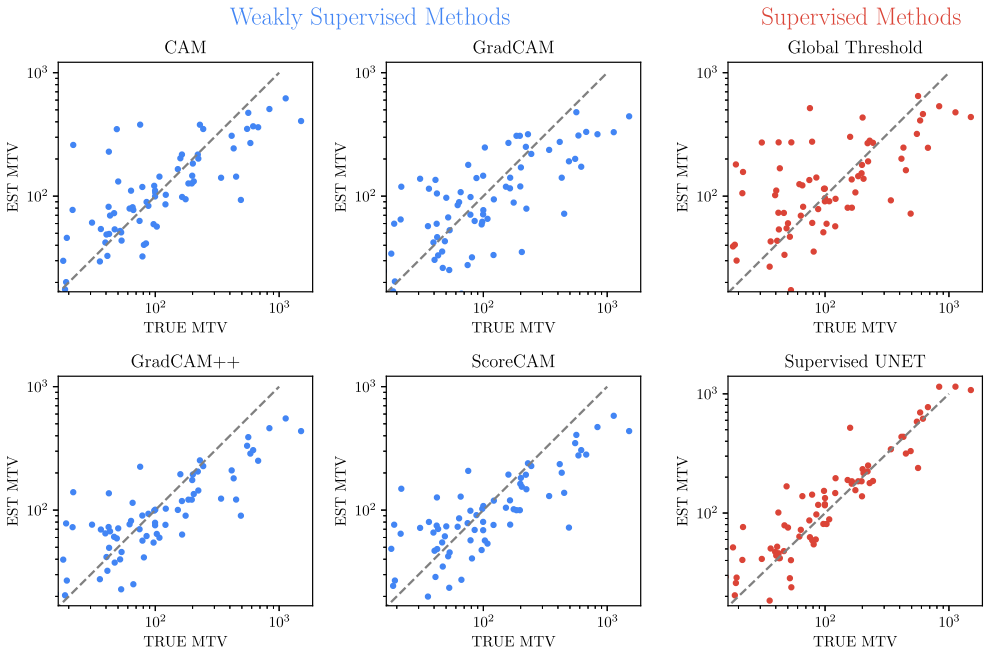


Fig. 6 Comparison between true and estimated MTV. All units in ml.

Table 1 Intra class correlation for estimated and real MTV/TLG.

	MTV (ml)			TLG (g)		
	ICC	95%-CI	p-value	ICC	95%-CI	p-value
CAM	0.64	[0.50, 0.74]	<0.001	0.85	[0.77, 0.90]	<0.001
GradCAM	0.55	[0.39, 0.67]	<0.001	0.40	[0.19, 0.57]	<0.001
GradCAM++	0.64	[0.48, 0.73]	<0.001	0.79	[0.66, 0.86]	<0.001
ScoreCAM	0.64	[0.50, 0.75]	<0.001	0.82	[0.71, 0.88]	<0.001
Threshold	0.59	[0.45, 0.71]	<0.001	0.88	[0.83, 0.92]	<0.001
UNET	0.94	[0.91, 0.96]	<0.001	0.99	[0.98, 0.99]	<0.001

underestimation of TLG of large tumors was observed; no overestimation of the TLG of small tumors occurred. The global threshold showed the largest variance for TLG estimation, again induced by marked overestimation of small lesions (median difference 167 g, IQR 524 g); however, there was less underestimation of larger lesion compared with the weakly supervised methods, which results in a higher ICC score due to less overall systematic error.

4 Discussion

In this study we introduced, evaluated, and compared methods for weakly supervised segmentation of FDG-avid lesions in whole-body FDG-PET images. We established that, using CAMs with subsequent thresholding, weakly supervised segmentation is feasible with satisfactory accuracy. Compared with an upper baseline (a fully supervised UNET) and a lower baseline (a global threshold), we found that CAM, GradCAM++, and ScoreCAM yielded good overall segmentation accuracy whereas the use GradCAM led to inferior results. Overall, image-derived

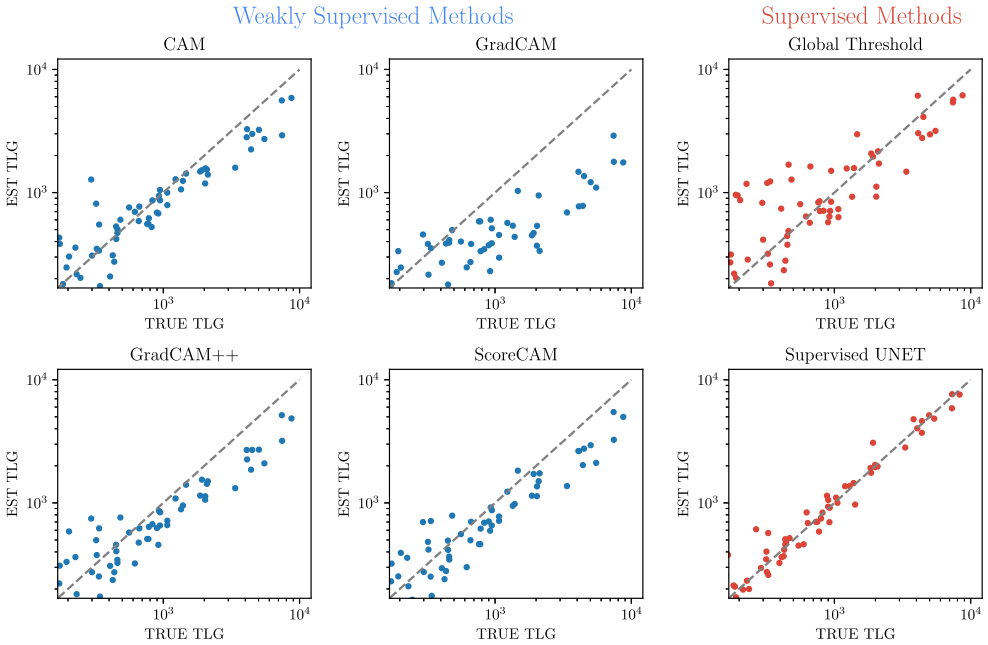


Fig. 7 Comparison between true and estimated TLG. All units in g.

parameters MTV and TLG extracted from these segmentations correlated well with the ground truth values extracted from manual segmentation using CAM, GradCAM++, and ScoreCAM. Again, the use of GradCAM yielded higher deviations.

The results of this study are relevant for a wide range of segmentation tasks in the medical imaging domain in which the generation of sufficient labeled training data is associated with high effort and cost. Using weak supervision—e.g., as in this study by only providing binary labels on an image level—this effort can be reduced significantly. Our results can thus contribute to more efficient training data generation and thus wider application of machine learning methods in the medical imaging domain.

The contribution of our study beyond existing work is the application to whole body FDG PET data and the detailed comparison of different CAM techniques. We found that CAM, GradCAM++, and ScoreCAM are suitable CAM methods for weakly supervised segmentation as they capture the tumor lesions within PET images, and thus the inferior performance of weakly supervised segmentation using GradCAM can be explained by the known and previously described property of GradCAM to highlight only the few small regions that are relevant for the network output, leading to systematic underestimation of target regions within the image[22]

The main limitation of class activation mapping-based segmentation as implemented in this study is the necessity of two thresholds—one on the CAM to identify the target area and one on the PET image to define the segmentation. Our results show that this works well on FDG-PET data due to the generally higher signal intensity of tumor lesions compared with background tissue. However, generalization to other medical imaging modalities such as CT or MRI, in which lesion intensity is less discriminative, might be limited. Future work will expand the use of class activation mappings to further datasets, including CT or MRI images. To this end, research should focus on methods that avoid the use of thresholds.

In this work, all analyses were performed on 2D slices. However, it would be beneficial to extend the principle of weakly supervised segmentation to 3D image data. This will allow for processing of entire imaging studies of single patients and further decrease the labeling effort. It can be expected, however, that the transition to 3D processing will be associated with a significant increase in computational demand. Although weak supervision saves significant time in creating labels, the precision of a supervised approach could not be reached in our study. If additional manual post-processing efforts are required to achieve sufficient precision for real-world applications, this must be taken into account. However, such corrections are mostly

limited to the exclusion of entire false positive lesions and can therefore be efficiently performed. On the other hand, weak supervision allows a potentially much larger number of subjects to be available as training data. Further studies need to show to what extent this compensates for the poorer accuracy. In particular, this could potentially also provide higher robustness and generalizability than a supervised model with a smaller training sample size.

Finally, the translation of the methodology presented in this paper to other PET tracers should be straightforward and may thus allow for implementing automated segmentation of non-FDG PET data with minimal manual annotation effort.

5 Conclusion

We were able to demonstrate that weakly supervised segmentation of FDG-avid lesions on whole-body FDG-PET is feasible, yielding satisfactory results. Further studies extending the proposed methodology to other PET tracers and medical imaging modalities will be necessary to investigate the transferabilty of the proposed methodology to related segmentation tasks.

Disclosures

There are no conflicts of interest.

Acknowledgments

This project was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), grant No. 438106095, and conducted under Germanys Excellence Strategy - EXC-Number 2064/1 – Project No. 390727645 and EXC-Number 2180 – Project number 390900677.

6 Code, Data, and Materials Availability

Code is publicly available on github: https://github.com/b4shy/weaklySupervisedSegmentation/blob/main/README.md.

References

1. T. C. McLoud, "Staging of lung cancer CT and PET," *Cancer Imaging* **14**, O6 (2014).
2. T. Berghmans et al., "Primary tumor standardized uptake value (SUVmax) measured on fluorodeoxyglucose positron emission tomography (FDG-PET) is of prognostic value for survival in non-small cell lung cancer (NSCLC): a systematic review and meta-analysis (MA) by the European Lung Cancer Working Party for the IASLC Lung Cancer Staging Project," *J. Thorac. Oncol.* **3**, 6–12 (2008).
3. K. Nie et al., "Prognostic value of metabolic tumour volume and total lesion glycolysis measured by 18F-fluorodeoxyglucose positron emission tomography/computed tomography in small cell lung cancer: a systematic review and meta–analysis," *J. Med. Imaging Radiat. Oncol.* **63**, 84–93 (2019).
4. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
5. J. Amin et al., "Big data analysis for brain tumor detection: deep convolutional neural networks," *Future Gener. Comput. Syst.* **87**, 290–297 (2018).
6. X. Zhao et al., "Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network," *Phys. Med. Biol.* **64**, 015011 (2019).
7. S. Jemaa et al., "Tumor segmentation and feature extraction from whole-body FDG-PET/CT using cascaded 2D and 3D convolutional neural networks," *J. Digit. Imaging* **33**, 888–894 (2020).

8. H. Azary and M. Abdoos, "A Semi-supervised method for tumor segmentation in mammogram images," *J. Med. Signals Sens.* **10**(1), 12 (2020).
9. B. Zhou et al., "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2921–2929 (2016).
10. G. Yang et al., "Weakly-supervised convolutional neural networks of renal tumor segmentation in abdominal CTA images," *BMC Med. Imaging* **20**(1), 37 (2020).
11. Z. Ji et al., "Scribble-based hierarchical weakly supervised learning for brain tumor segmentation," *Lect. Notes Comput. Sci.* **11766**, 175–183 (2019).
12. X. Feng et al., "Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules," *Lect. Notes Comput. Sci.* **10435**, 568–576 (2017).
13. S. Afshari et al., "Weakly supervised fully convolutional network for PET lesion segmentation," *Proc. SPIE* **10949**, 109491K (2019).
14. H.-G. Nguyen et al., "A novel segmentation framework for uveal melanoma in magnetic resonance imaging based on class activation maps," tech. rep. (2019).
15. S. Eyuboglu et al., "Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body FDG-PET/CT," *Nat. Commun.* **12**, 1–15 (2021).
16. J. Devlin et al., "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technol.—Proc. Conf.*, Vol. 1, 4171–4186 (2018).
17. C. Pfannenberg et al., "Practice-based evidence for the clinical benefit of PET/CT-results of the first oncologic PET/CT registry in Germany," *Eur. J. Nucl. Med. Mol. Imaging* **46**, 54–64 (2019).
18. M. C. Adams et al., "A systematic review of the factors affecting accuracy of SUV measurements," *Am. J. Roentgenol.* **195**, 310–320 (2010).
19. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556 (2014).
20. O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.* **115**(3), 211–252 (2015).
21. R. R. Selvaraju et al., "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 618–626 (2017).
22. A. Chattopadhay et al., "Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks," in *IEEE Winter Conf. Appl. Comput. Vision* (2018).
23. H. Wang et al., "Score-CAM: score-weighted visual explanations for convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit. Workshops*, pp. 24–25 (2020).
24. A. B. Jung et al., "imgaug," 2020, https://github.com/aleju/imgaug.
25. A. Paszke et al., "PyTorch: an imperative style, high-performance deep learning library," arXiv (2019).
26. I. Sutskever et al., "On the importance of initialization and momentum in deep learning," in *Int. Conf. Mach. Learn.*, pp. 1139–1147, PMLR (2013).
27. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980 (2017).
28. H.-J. Im et al., "Current methods to define metabolic tumor volume in positron emission tomography: which one is better?" *Nucl. Med. Mol. Imaging* **52**, 5–15 (2018).

Marcel Früh is a PhD student at the University Hospital Tübingen. He received his MSc degree in computer science with focus on deep learning from the University of Tübingen in 2020. His main area of interest is machine learning in medical imaging.

Biographies of the other authors are not available.

A.2 Self-supervised learning for automated anatomical tracking in medical image data with minimal human labeling effort

Authors: *Marcel Früh, Thomas Küstner, Marcel Nachbar, Daniela Thorwarth, Andreas Schilling, Sergios Gatidis*

Published in: *Computer Methods and Programs in Biomedicine*

Date of Publication: *October 2022*

Computer Methods and Programs in Biomedicine 225 (2022) 107085

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Self-supervised learning for automated anatomical tracking in medical image data with minimal human labeling effort

Marcel Frueh[a,d], Thomas Kuestner[a], Marcel Nachbar[b], Daniela Thorwarth[b], Andreas Schilling[d], Sergios Gatidis[a,c,*]

[a] University Hospital Tuebingen, Department of Radiology, University of Tuebingen, Hoppe-Seyler-Straße 3 Tuebingen 72076, Germany
[b] Section for Biomedical Physics, Department of Radiation Oncology, University of Tuebingen, Hoppe-Seyler-Straße 3 Tuebingen 72076, Germany
[c] Max Planck Institute for Intelligent Systems, Empirical Inference Department, Max-Planck-Ring 4 Tuebingen 72076, Germany
[d] University of Tuebingen, Institute for Visual Computing, Department of Computer Science, Sand 14 Tuebingen 72076, Germany

ARTICLE INFO

Article history:
Received 27 January 2022
Revised 2 August 2022
Accepted 23 August 2022

Keywords:
Anatomical tracking
Self-supervised learning
MR-LINAC
Image-Guided radiation therapy
Cardiac MRI

ABSTRACT

Background and Objective: Tracking of anatomical structures in time-resolved medical image data plays an important role for various tasks such as volume change estimation or treatment planning. State-of-the-art deep learning techniques for automated tracking, while providing accurate results, require large amounts of human-labeled training data making their wide-spread use time- and resource-intensive. Our contribution in this work is the implementation and adaption of a self-supervised learning (SSL) framework that addresses this bottleneck of training data generation. Methods: To this end we adapted and implemented an SSL framework that allows for automated anatomical tracking without the necessity for human-labeled training data. We evaluated this method by comparison to conventional- and deep learning optical flow (OF)-based tracking methods. We applied all methods on three different time-resolved medical image datasets (abdominal MRI, cardiac MRI, and echocardiography) and assessed their accuracy regarding tracking of pre-defined anatomical structures within and across individuals. Results: We found that SSL-based tracking as well as OF-based methods provide accurate results for simple, rigid and smooth motion patterns. However, regarding more complex motion, e.g. non-rigid or discontinuous motion patterns in the cardiac region, and for cross-subject anatomical matching, SSL-based tracking showed markedly superior performance. Conclusion: We conclude that automated tracking of anatomical structures on time-resolved medical image data with minimal human labeling effort is feasible using SSL and can provide superior results compared to conventional and deep learning OF-based methods.

1. Introduction

Tracking of anatomical structures and organs is a central task in medical image analysis, especially regarding the assessment of time-resolved image data such as CINE MRI or ultrasound images. In such applications, the change in position or volume of certain structures provides insight into the pathophysiology of diseases or allows for planning of therapeutic interventions. For example, the change in volume of the left ventricle as measured in cardiac CINE MRI provides information on cardiac function [1] and the tracking of tumor lesions during respiration enables precise planning and execution of radiation therapy [2–7]. In many situations, tracking of predefined structures is still performed manually or semi-automatically in time-demanding procedures. The conventional approach for automation of tracking tasks in medical imaging uses established registration methods that match different images on pixel-level information [8–11]. Among these methods, optical flow techniques are the most commonly used class of algorithms and have been applied to numerous image processing tasks such as image registration [12], motion correction [13] or lesion tracking [10]. The main drawback of these methods however, lies in the neglect of semantic image content and the necessity for hyperparameter optimization. The introduction of deep learning algorithms has improved automation of numerous image analysis tasks including organ segmentation [14,15] lesion detection [16,17] and optical flow estimation [18–20] which also allows for tracking of anatomical structures in specific applications. How-

* Corresponding author.
E-mail address: Sergios.Gatidis@med.uni-tuebingen.de (S. Gatidis).

https://doi.org/10.1016/j.cmpb.2022.107085

ever, a substantial drawback of most deep-learning approaches is the necessity for large amounts of pixel-wise annotated training datasets which is associated with enormous effort or in some situations even not feasible. To alleviate this challenge, different learning strategies have been proposed including unsupervised learning techniques [21]. Recently, novel machine learning approaches have been introduced that aim to alleviate the problem of large-scale training data labeling by means of semi-, weakly- or self-supervised learning [22–28]. The central concept of these methods lies in exploiting regularities in the underlying data to learn useful representations that can subsequently be used for downstream tasks. Self-supervised learning (SSL) has been proposed for various medical image analysis tasks such as organ segmentation [29], classification [30] or landmark detection [31]. For the specific field of image tracking and registration, SSL has been shown to provide good results without human interaction [32,33]. Specifically, Jabri et al. recently demonstrated the use of a random walk SSL task for object tracking on video data [34]. The purpose of this work was thus to address this bottleneck of training data generation. Our contributions lie in the adaptation and implementation of an SSL framework that allows for automated anatomical tracking without the necessity for human-labeled training data and the comparison of this SSL-based anatomical tracking with Farneback optical flow [35] (FB-Flow), PCA optical flow [36] (PCA-Flow) and FlowNet2 optical flow [37] (FlowNet2-Flow) on three different datasets (abdominal MRI, cardiac CINE MRI, and echocardiography) for the task of organ tracking. To the best of our knowledge this is the first study evaluating and comparing the performance of both, self-supervised deep learning and optical flow based tracking algorithms for tissue tracking on different modalities.

2. Material and methods

2.1. Anatomical tracking

The overarching goal of this study was to implement and evaluate methods that allow for tracking of anatomical structures within image series without human supervision and without the necessity for human-annotated training data.

To this end, four different tracking methods, one based on self-supervised deep learning and three based on Optical Flow were implemented and adapted to a medical imaging context.

2.1.1. Optical flow-based tracking

The purpose of dense optical flow algorithms is to track the movement of all pixels between two frames in an image sequence. For the optical flow methods used in this study, objects are tracked by warping the respective manual segmentation mask according to the estimated flow field.

Farneback optical flow

The Farneback algorithm is a widely used method for object tracking on image sequences [35]. It is based on Polynomial Expansion for neighborhood approximation and was used in this study as a baseline method and a classical representative for Optical Flow estimation. All results produced by the Farneback Optical Flow were obtained using a pyramid scale of 0.5 with 3 levels, a window size of 15, 3 iterations, a pixel neighborhood of 5 to calculate polynomial expansion and 1.2 as standard deviation for derivative smoothing.

PCA optical flow

A more recent approach to Optical Flow estimation, PCA-Flow [36], approximates the Optical Flow by computing the weighted sum over basis flow fields with corresponding weights. The basis flow fields of the original method were derived from Hollywood movies by initial Optical Flow computation using GPUFlow [38] and subsequent calculation of the first 500 principal components of the resulting flow fields. These established basis flow fields were also used in this work without re-training. Default parameters were used.

FlowNet2 optical flow

In contrast to the above methods, FlowNet [18] estimates the Optical Flow by leveraging Convolutional Neural Networks. Two subsequent images are fed to the FlowNet architecture (either stacked as multiple channels or separated through two different encoders) which then outputs an optical flow approximation. The network is trained end-to-end on datasets where ground truth optical flow is available (mostly synthetic datasets, such as Flying Chairs [18], Flying Things3D [39] or MPI Sintel [40]). FlowNet2 [37], the successor of FlowNet, combines multiple FlowNet networks with different architectures to handle both, small and large motion. In this work, a pre-trained (Flying Things3D) FlowNet2 model was deployed directly on the presented use cases. Due to the lack of ground truth motion fields on these datasets, which is typical for medical applications, no fine tuning could be performed.

2.2. Self-Supervised pretraining for deep learning-based object tracking

SSL is based on the concept of using a training task that does not require external labels but instead relies on ground truth that is known by design. A simple example for an SSL task is predicting the rotation of images that were intentionally rotated and where thus the ground truth is directly available [41]. Recently, Jabri et al. proposed a deep learning framework for self-supervised pretraining on videos [34] that can be used for object tracking in image series.

In this framework, self-supervision is achieved by ensuring cycle-consistent mapping of image patches.

In a first step, the individual video frames are divided into patches. Using an encoder ϕ based on the ResNet-18 architecture [42] which maps the patches into their respective feature space, the similarity between two patches p and q is computed using the dot product $d =< \phi(p), \phi(q) >$.

Subsequently, the affinity matrix between two consecutive frames at timepoints t and $t+1$ is computed via:

$$A_t^{t+1}(i,j) = \frac{\exp(d(p_t^i, q_{t+1}^j)/\tau}{\sum_{n=1}^{N}\exp(d(p_t^i, q_{t+1}^n)/\tau} \quad (1)$$

$A_t^{t+1}(i,j)$ contains the stochastic (temperature-scaled) similarity between all pairs of patches (p_t, q_{t+1}).

Affinity between longer frame pairs at timepoints t and $t+k, k>1$ is achieved via multiplication of the intermediate affinity matrices between consecutive frame pairs.

To optimize this framework in a self-supervised way, image sequences are expanded into palindrome sequences (normal and inverse order of images). This enables computation of the cross-entropy loss between initial patches in the first frame and target patches in the final frame (which is identical to the first frame in the palindrome sequence) as each patch acts as its own target label.

Formally, during training, all entries with $i=j$ of the palindrome similarity matrix $A_t^{t+k}A_{t+k}^t$ are maximized while off-diagonal entries are minimized yielding the following loss:

$$L_{CE} = \sum_{i=j}^{N} -\log(A_t^{t+k}A_{t+k}^t(i,j)) \quad (2)$$

To propagate a reference segmentation from the first image through the image sequence during inference, the local features

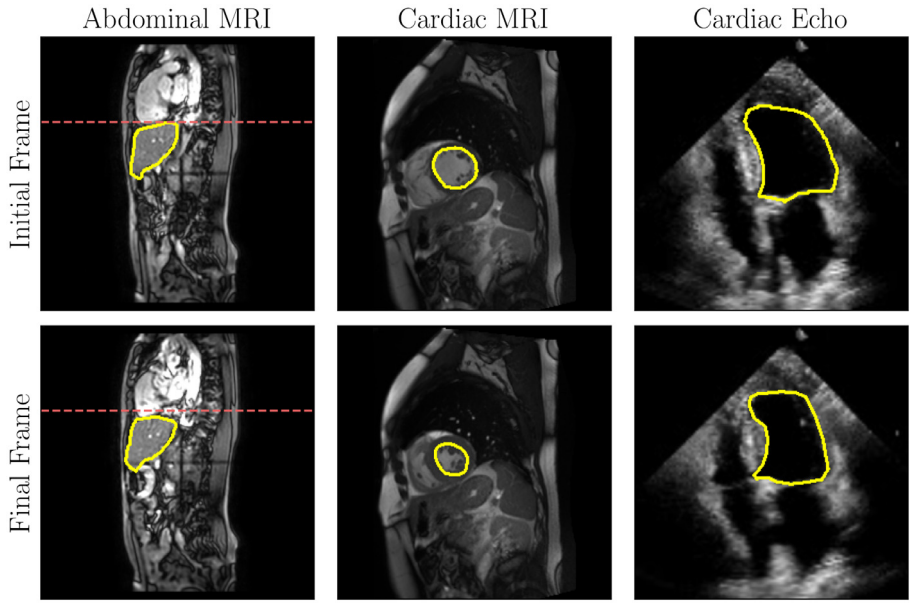


Fig. 1. Exemplary initial and final frame from all three datasets respectively with corresponding manual segmentation of the organ of interest: liver (abdominal MRI, the dotted line serves as a positional reference), left ventricle (cardiac MRI and cardiac Echo). Contours are drawn at liver boundary (abdominal MRI) and endocardium (cardiac CINE MRI and cardiac Echo) by an experienced reader.

are extracted from all images using the trained encoder ϕ in a first step. Subsequently, the affinity matrix is calculated based on these features and the k (k is a hyperparameter) largest affinity values are considered same-class segmentation labels (k-nearest neighbors).

In this study, we extended this method for application on medical image data so that it can be used for tracking in time-resolved data as well as across subjects. We trained the feature extractor from scratch with decreased stride in the second layer to increase accuracy at tissue boundaries on image sequences of 4 images for 500 epochs with a batch size of 6, an initial learning rate of 10^{-4} and softmax temperature of 0.07 for each dataset on an NVIDIA RTX3090 GPU. Using this method, object tracking was performed by initially calculating the feature maps for all images in an image series and subsequently determining the k=10 most similar features within a radius of 20 pixels of all feature maps with respect to an initial manual segmentation. In addition, we performed clustering on the predicted segmentations and, in case multiple objects are recognized in the target image, we selected the object with the highest similarity to the ground truth segmentation. All images were resized to (784 $\times$ 784) for inference yielding feature maps with matrix size of (196 $\times$ 196).

In all following figures, results of SSL-based tracking are displayed in red, of PCA-Flow in blue, of FB-Flow in green and of FlowNet2 in purple.

2.3. Experiments

Three different datasets were used for method evaluation in this work. We chose these datasets as they cover different imaging modalities, anatomical regions and motion patterns (Fig. 1).

1. Cardiac MRI: The anonymized Data Science Bowl Cardiac Challenge Dataset [43] consisting of 1140 cardiac Magnetic Resonance Imaging (MRI) studies of 1140 patients (42 $\pm$ 20 years, 470 female) including 12,121 time-resolved short axis CINE sequences of the heart resulting in 372,810 individual image slices in total with a matrix size of 192 $\times$ 256 and a spatial resolution of 1.4 $\times$ 1.4 mm for an 8 mm slice thickness. Data was acquired in short-axis on both, 1.5 and 3T MRI with a balanced steady-state free precession sequence with varying temporal resolution of approximately 30 frames/heart beat, i.e. 30 frames/second.
2. Cardiac Echo: The EchoNet-Dynamic dataset [44] which consists of 10,030 anonymized dynamic 4-chamber echocardiography (Echo) videos of 10.030 patients (68 $\pm$ 21 years, 4885 female) with a total of 1,770,646 single image frames, a temporal resolution of approximately 50 frames/second and matrix size of 112 $\times$ 112.
3. Abdominal MRI: An anonymized in-house time-resolved MRI of the upper abdomen was acquired on an MR-LINAC system (Unity, Elekta, Stockholm, Sweden). The database includes 230 studies of 50 patients (66 $\pm$ 11 years, 20 female) with three sequences in axial, coronal and sagittal orientation each. A total of 165,264 single image slices was obtained. Data was acquired with a matrix size of 352 $\times$ 352, a spatial resolution of 1.2 $\times$ 1.2 mm, 5 mm slice thickness, and a temporal resolution of 2 frames/second. Patient data were acquired in the context of a clinical phase II trial (NCT04172753).

2.4. Dataset-Specific-Preprocessing

Abdominal MRI

CINE data were acquired over a period of 20–30 min during radiotherapy, yielding around 1000 frames for each image orientation. These long sequences were split into short clips of 20 frames of identical orientation, roughly containing two to three respiratory cycles. All images were divided into a train-validation-test split (60%, 20%, 20%) in a subject-wise manner.

Cardiac CINE MRI Z-normalization was performed on a subject wise manner followed by cropping or padding of the individual two dimensional frames from the whole left ventricular stack to a common shape of (352 $\times$ 352). Again a 60-20-20 subject-wise train-val-test split was used.

Cardiac echo All two-dimensional frames were resized to an input dimension of (352 $\times$ 352), intensity-scaled to the unit range followed by z-normalization and split into short sequences of 30 frames each with subsequent division into a 60-20-20 subject-wise train-val-test split.

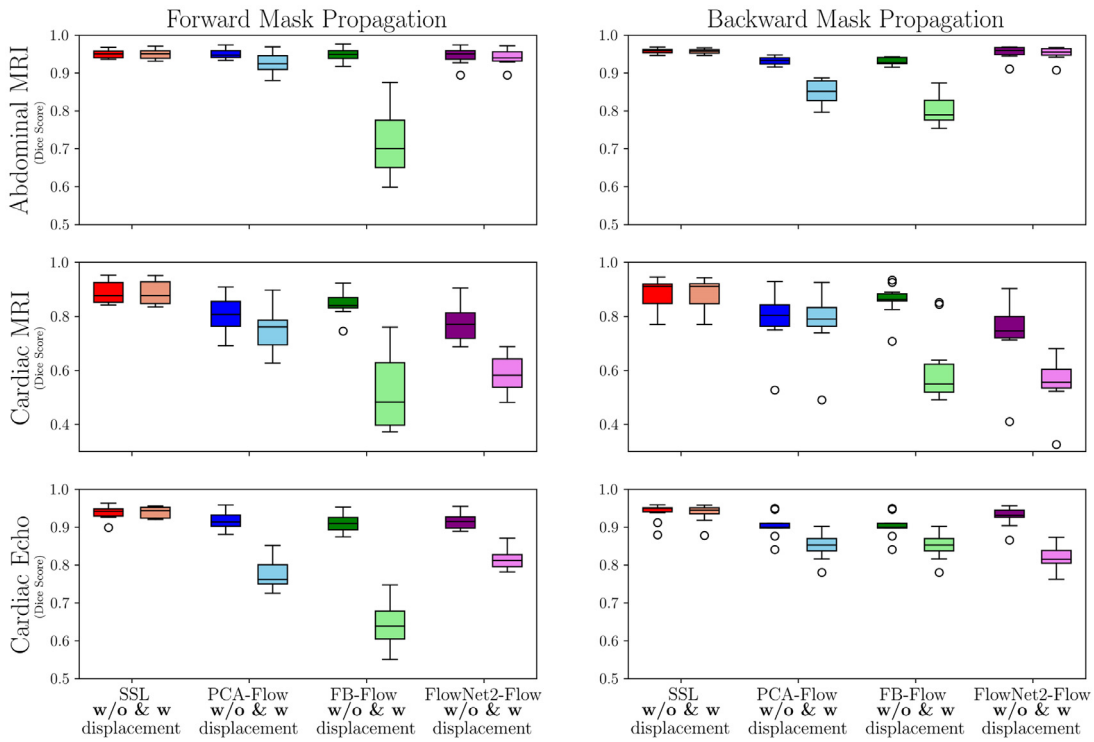


Fig. 2. Averaged Dice scores based on forward and backward passes on both, original (saturated color/left box plot) and artificially displaced data (light colors/right box plot), based on predicted and ground truth segmentations for SSL (red), PCA-Flow (blue), Farneback Flow (green) and FlowNet2 (purple). Overall, tracking based on SSL yields the best performance.

2.5. Evaluation

Accuracy of anatomical tracking using the presented methods was evaluated on two principle tasks: (i) within-subject segmentation mask propagation and (ii) cross-subject segmentation mask propagation.

Mask propagation The first task consisted of propagating a segmentation mask from the initial to the final image (forward pass) and vice versa (backward pass) of a test sequence.

For all datasets, the anatomical target structures to track (liver on abdominal MRI, left ventricle in both cardiac MRI and cardiac Echo) were segmented on all frames of 10 test sequences by an experienced radiologist (SG, 11 years of experience) providing reference data for method evaluation. Regarding the cardiac datasets, manual segmentation was performed for one systolic phase, i.e., from maximal to minimal left ventricular expansion. The liver was manually segmented over one breathing cycle, i.e, from end-inspiration to end-expiration.

Accuracy of mask propagation was quantified as the mean Dice score between the propagated mask and manual reference segmentation on all images within a sequence (averaged dice) by performing both, forward and backward mask propagation. In addition, the Dice score only based on the final target frame (edge dice) was calculated to investigate method performance between the extremes of the image sequences.

To assess the performance of the presented algorithms under bulk motion, we artificially introduced a horizontal image displacement by shifting all images after the central frame of the respective test sequence by 8% of the image width to the right.

Cross-subject anatomical matching In certain clinical situations, identification of anatomical structures across subjects or examination time points can be relevant. In order to assess the capacity of SSL-based and optical flow-based anatomical matching across subjects, we propagated the segmentation masks between all pairs of test images within the respective datasets. Again, the Dice score was used as performance metric.

3. Results

Mask propagation

For the task of liver tracking within MR-LINAC images, we observed a mean averaged forward/backward dice score of 0.95/0.96, 0.94/0.93, 0.94/0.93, 0.95/0.95 for SSL, PCA-Flow, Farneback-Flow and FlowNet2, respectively, as depicted in Fig. 2 top row.

Mean edge dice scores amounted to 0.94/0.96, 0.92/0.90, 0.92/0.91 and 0.94/0.93, respectively.

Tracking of the left ventricle in cardiac MRI yielded the largest difference between the methods. A mean forward/backward averaged dice score of 0.89/0.90 was achieved by the self-supervised learning based method, whereas PCA-Flow yielded 0.80/0.89 and Farneback 0.81/0.92 respectively. FlowNet2 performed worst resulting in a Dice score of 0.77/0.73. Refer to Fig. 2 middle row. This difference is highlighted by the edge dice which amounted to 0.85/0.95, 0.61/0.89, 0.70/0.92 and 0.55/0.54 for all methods, respectively.

In contrast, when performing ventricle tracking in echocardiography we observed similar mean forward/backward averaged dice scores of 0.93/0.95, 0.89/0.91, 0.91/0.91 and 0.91/0.92, respectively as shown in Fig. 2 bottom row. The edge dice amounted to 0.93/0.96, 0.88/0.92, 0.91/0.94 and 0.85/0.88.

Overall, accuracy of mask propagation on the inverse image sequences showed similar results (Fig. 2 right column). Notably, for the cardiac datasets, mask propagation from systole to diastole (backward pass) was markedly more accurate (edge dice) compared to the forward pass (diastole to systole) using PCA-Flow and FB-Flow. In contrast, no relevant differences between forward and backward pass tracking accuracy were observed for the SSL-based method and FlowNet2-Flow.

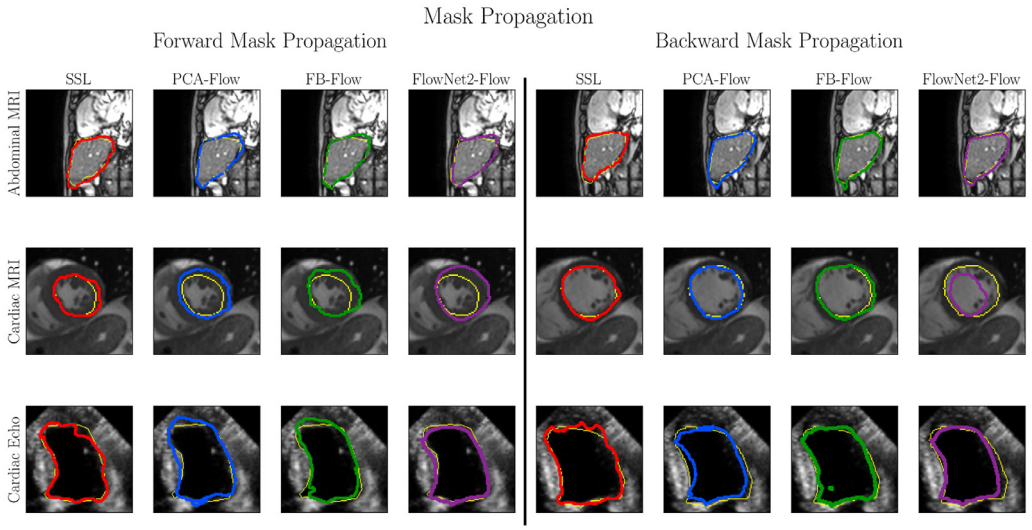


Fig. 3. Visualized mask propagation based on SSL (red, forward/backward edge dice scores of 0.94/0.95, 0.87/0.96 and 0.92/0.96 for the visualized abdominal MRI, cardiac MRI and cardiac Echo, respectively), PCA-Flow (blue, forward/backward edge dice scores of 0.89/0.90, 0.66/0.95 and 0.91/0.92 for the visualized abdominal MRI, cardiac MRI and cardiac Echo respectively), FB-Flow (green, forward/backward edge dice scores of 0.90/0.90, 0.55/0.89 and 0.92/0.95 for the visualized abdominal MRI, cardiac MRI and cardiac Echo, respectively) and FlowNet2-Flow (purple, forward/backward edge dice scores of 0.92/0.91, 0.62/0.68 and 0.84/0.89 for the visualized abdominal MRI, cardiac MRI and cardiac Echo, respectively) for an exemplar validation sequence. The left columns depict the forward propagation whereas the right columns visualize the backward mask propagation. The manual segmentation mask is depicted in yellow.

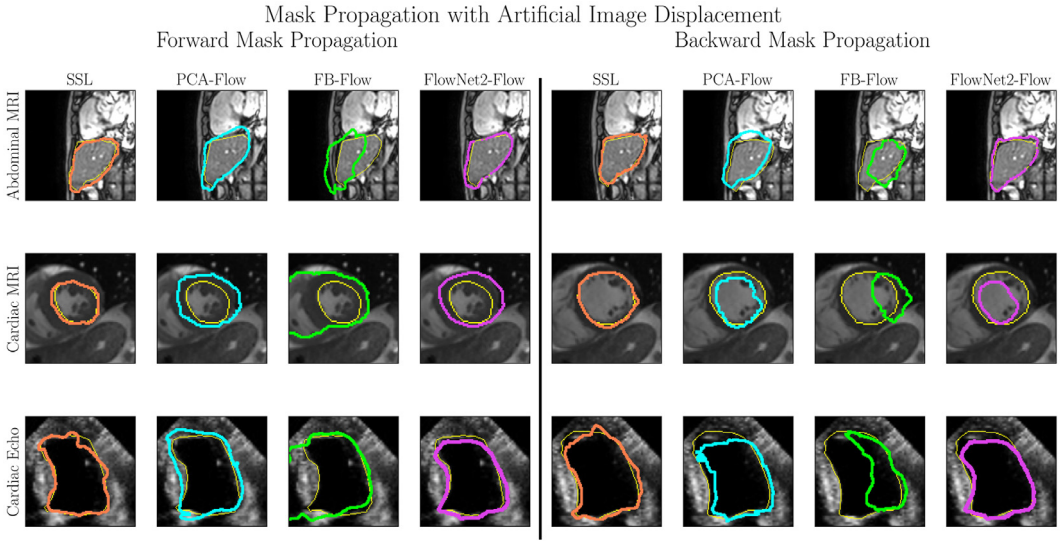


Fig. 4. Mask Propagation with additional image displacement: Visualized mask propagation based on SSL (red, forward/backward edge dice scores of 0.94/0.95, 0.87/0.96 and 0.92/0.94 for the visualized abdominal MRI, cardiac MRI and cardiac Echo respectively), PCA-Flow (blue, forward/backward edge dice scores of 0.81/0.83, 0.58/0.83 and 0.82/0.9 for the visualized abdominal MRI, cardiac MRI and cardiac Echo respectively), FB-Flow (green, forward/backward edge dice scores of 0.54/0.57, 0.14/0.17 and 0.61/0.54 for the visualized abdominal MRI, cardiac MRI and cardiac Echo respectively) and FlowNet2-Flow (purple, forward/backward edge dice scores of 0.92/0.92, 0.57/0.66 and 0.83/0.89 for the visualized abdominal MRI, cardiac MRI and cardiac Echo, respectively) for an exemplar validation sequence. The left columns depict the forward propagation whereas the right columns visualize the backward mask propagation. The manual segmentation mask is depicted in yellow.

Qualitative visual evaluation (Fig. 3, Video 1, supplementary material) further supported the above findings. While liver motion was accurately captured by all methods, systolic cardiac movement was not correctly captured by the optical flow algorithms.

Mask propagation with image displacement

Horizontal image displacement led to a marked performance decrease, for both forward- and backwards mask propagation using optical flow based methods. This drop in accuracy was more pronounced using FB-Flow. In contrast, SSL-based mask propagation showed similarly accurate results compared to the non-displaced version as illustrated in Fig. 2 (light colors). These quantitative results were again confirmed by qualitative inspection of the propagated segmentation masks (Fig. 4).

3.1. Cross-subject anatomical matching

For the task of cross-subject anatomical mapping we observed a markedly better performance of SSL-based tracking compared to the optical flow-based methods. On the cardiac MRI dataset, matching accuracy was 0.72 ± 0.25 using SSL-based tracking compared to 0.33 ± 0.28/0.35 ± 0.36/0.28 ± 0.30 using PCA-Flow/FB-Flow/FlowNet2-Flow respectively. Similar results were observed for

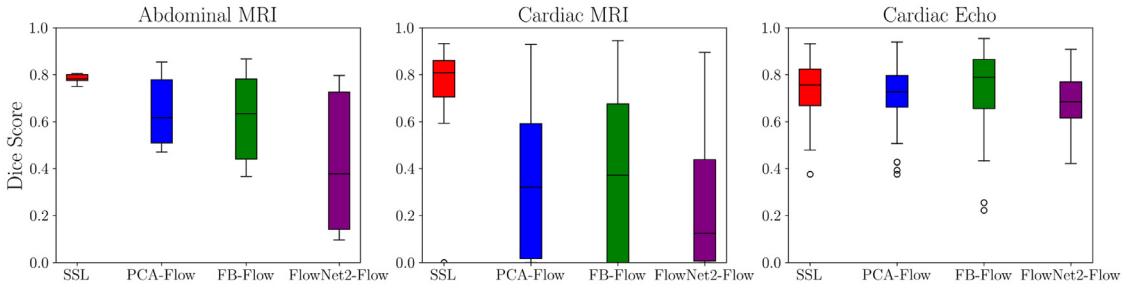


Fig. 5. Dice Scores for cross-subject mask propagation based on predicted and ground truth segmentation for SSL (red), PCA-Flow (blue), FB-Flow (green) and FlowNet2-Flow (purple). Overall, SSL yields much more robust results.

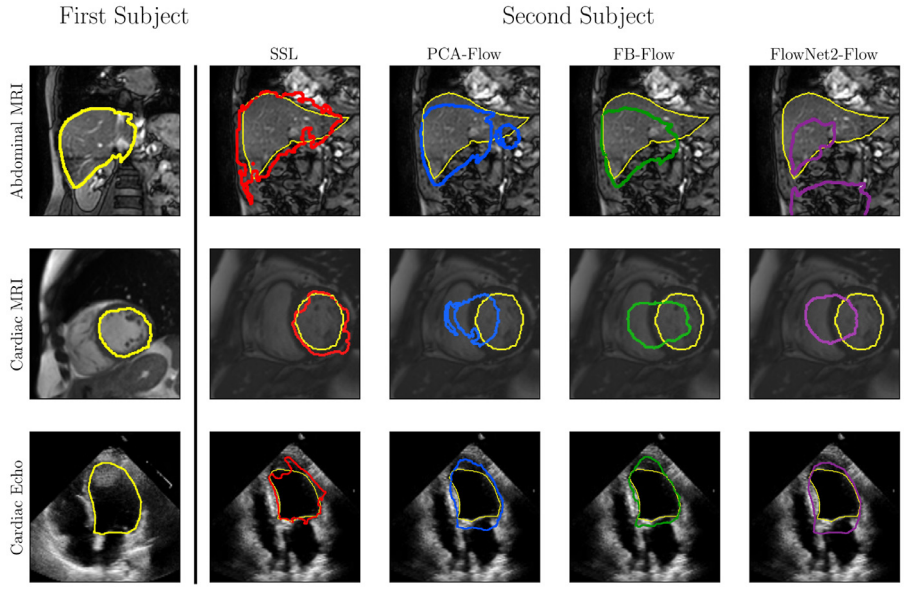


Fig. 6. Visualized cross-subject mask propagation based on SSL (red, dice scores of 0.83, 0.85 and 0.87 for the visualized abdominal MRI, cardiac MRI and cardiac Echo, respectively), PCA-Flow (blue, dice scores of 0.74, 0.38 and 0.79 for the visualized abdominal MRI, cardiac MRI and cardiac Echo, respectively), FB-Flow (green, dice scores of 0.67, 0.47 and 0.85 for the visualized abdominal MRI, cardiac MRI and cardiac Echo, respectively) and FlowNet2-Flow (purple, dice scores of 0.27, 0.29, 0.77 for the visualized abdominal MRI, cardiac MRI and cardiac Echo, respectively) between two sample validation subjects.

the abdominal MRI dataset. In contrast, for the echocardiography dataset, where anatomical positions within the images were similar across subjects, all four methods performed equally well (mean Dice scores of 0.72-0.75) (Fig. 5). The quantitative results were confirmed by qualitative evaluation (Fig. 6).

4. Discussion

In this work, we implemented and evaluated a method for automated anatomical tracking on time-resolved medical image data based on SSL that does not require human labeling effort for algorithm training and compared it to conventional and deep-learning based optical flow-based tracking methods.

We found that these methods in general provide accurate results for rigid and smooth motion patterns and thus can allow for reliable motion tracking for numerous clinical tasks. However, regarding more complex motion, e.g. non-rigid or discontinuous motion patterns, or cross-subject anatomical matching, optical flow-based methods showed markedly inferior performance while tracking accuracy using SSL still provided satisfactory results.

In particular, we found that tracking rigid and non-displaced liver motion during the respiratory cycle yielded similarly good results among all methods. In contrast, contraction and expansion of the left ventricle resulted in strong differences between the four methods. SSL yielded satisfactory tracking performance for cardiac motion, while especially tracking of systolic cardiac motion was inferior using optical flow methods. Overall, PCA-Flow yielded superior results compared to FB-Flow and FlowNet2-Flow.

Similarly, SSL-based motion tracking showed good accuracy for all datasets when artificial image displacement was introduced, while PCA-Flow yielded slightly inferior results and FB-Flow failed to produce meaningful results.

FlowNet2 was able to successfully track the rigid motion of the liver within the abdominal MRI but not the cardiac motion.

These results extended to cross-subject matching where SSL yielded superior results overall. Due to the domain shift in the datasets, supervised optical flow methods are not suitable without retraining.

Numerous potential clinical applications are conceivable using the presented methods for automated anatomical tracking. Regarding diagnostic imaging, tracking of anatomical structures in time-resolved image data can provide accurate information about volume changes and allow for motion-corrected evaluation of dynamic contrast-enhanced imaging. Regarding interventional and therapeutic applications, motion tracking, e.g. in radiation therapy or interventional radiology can allow for tracking of the therapeutic progress and thus enabling more precise treatment delivery. In all these fields of application, it is important to limit the necessity for human-annotated training data which is a significant bottleneck for clinical translation. Thus, using the described methods,

clinical implementation can be achieved with markedly reduced effort compared to e.g. supervised machine learning methods. In general, based on our results, we assess that self-supervised deep learning is the most effective and versatile method to achieve this goal and surpasses the performance of optical flow methods.

This study has limitations. First, we only focused on 2D image data. Although it can be expected that our results can be generalized to the 3D case, additional studies will be necessary to provide more insight. Also, we only investigated tracking of physiological anatomical structures and did not assess the tracking performance on pathological data e.g. of tumor lesions. Finally, as a methodological aspect, we did not retrain the feature extractor of PCA-Flow or the FlowNet2-Flow network as our goal was to provide methods that do not require any human labeled data. In principle, performance of these frameworks may be improved by fine tuning on a specific dataset, which however would require additional effort for training data generation.

5. Conclusion

In conclusion, tracking of anatomical structures on time-resolved medical image data with minimal human labeling effort is feasible using SSL and can provide superior results compared to conventional and deep learning OF-based methods. Overall, self-supervised deep learning can be regarded as the more accurate and versatile method and thus potentially be applied to numerous clinical applications.

Declaration of Competing Interest

There are no conflicts of interest.

Acknowledgments

This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), grant number 438106095 and conducted under Germany's Excellence Strategy - EXC-Number 2064/1 - Project number 390727645 and EXC-Number 2180 - Project number 390900677.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.cmpb.2022.107085.

References

[1] V.Y. Wang, H. Lam, D.B. Ennis, B.R. Cowan, A.A. Young, M.P. Nash, Modelling passive diastolic mechanics with quantitative MRI of cardiac structure and function, Med. Image Anal. 13 (5) (2009) 773–784.
[2] J.M. Pollard, Z. Wen, R. Sadagopan, J. Wang, G.S. Ibbott, The future of image-guided radiotherapy will be MR guided, Br. J. Radiol. 90 (1073) (2017) 20160667.
[3] S. Corradini, F. Alongi, N. Andratschke, C. Belka, L. Boldrini, F. Cellini, J. Debus, M. Guckenberger, J. Hörner-Rieber, F. Lagerwaard, et al., MR-guidance in clinical reality: current treatment challenges and future perspectives, Radiat. Oncol. 14 (1) (2019) 1–12.
[4] M.J. Menten, M.F. Fast, A. Wetscherek, C.M. Rank, M. Kachelrieß, D.J. Collins, S. Nill, U. Oelfke, The impact of 2D cine MR imaging parameters on automated tumor and organ localization for MR-guided real-time adaptive radiotherapy, Phys. Med. Biol. 63 (23) (2018) 235005.
[5] S. Al-Ward, M. Wronski, S.B. Ahmad, S. Myrehaug, W. Chu, A. Sahgal, B.M. Keller, The radiobiological impact of motion tracking of liver, pancreas and kidney SBRT tumors in a MR-linac, Phys. Med. Biol. 63 (21) (2018) 215022.
[6] J.S. Witt, S.A. Rosenberg, M.F. Bassetti, MRI-guided adaptive radiotherapy for liver tumours: visualising the future, Lancet Oncol. 21 (2) (2020) e74–e82.
[7] N.R. Huttinga, C.A. van den Berg, P.R. Luijten, A. Sbrizzi, MR-MOTUS: model-based non-rigid motion estimation for MR-guided radiotherapy using a reference image and minimal k-space data, Phys. Med. Biol. 65 (1) (2020) 015004.
[8] C.T. Metz, S. Klein, M. Schaap, T. van Walsum, W.J. Niessen, Nonrigid registration of dynamic medical imaging data using nD+ t B-splines and a groupwise optimization approach, Med. Image Anal. 15 (2) (2011) 238–249.
[9] T.D. Keiper, A. Tai, X. Chen, E. Paulson, F. Lathuilière, S. Bériault, F. Hébert, D.T. Cooper, M. Lachaine, X.A. Li, Feasibility of real-time motion tracking using cine MRI during MR-guided radiation therapy for abdominal targets, Med. Phys. 47 (8) (2020) 3554–3566.
[10] C. Zachiu, N. Papadakis, M. Ries, C. Moonen, B.D. de Senneville, An improved optical flow tracking technique for real-time MR-guided beam therapies in moving organs, Phys. Med. Biol. 60 (23) (2015) 9003.
[11] M.S. Hosseini, M.H. Moradi, M. Tabassian, J. D'hooge, Non-rigid image registration using a modified fuzzy feature-based inference system for 3D cardiac motion estimation, Comput. Methods Programs Biomed. 205 (2021) 106085, doi:10.1016/j.cmpb.2021.106085. https://www.sciencedirect.com/science/article/pii/S0169260721001607.
[12] T. Pock, M. Urschler, C. Zach, R. Beichel, H. Bischof, A duality based algorithm for TV-L 1-optical-flow image registration, in: International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, 2007, pp. 511–518.
[13] M. Dawood, F. Buther, X. Jiang, K.P. Schafers, Respiratory motion correction in 3-D pet data with advanced optical flow algorithms, IEEE Trans. Med. Imaging 27 (8) (2008) 1164–1175.
[14] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[15] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S.P. Pereira, M.J. Clarkson, D.C. Barratt, Automatic multi-organ segmentation on abdominal CT with dense V-networks, IEEE Trans. Med. Imaging 37 (8) (2018) 1822–1834.
[16] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, H. Larochelle, Brain tumor segmentation with deep neural networks, Med. Image Anal. 35 (2017) 18–31.
[17] A. Akselrod-Ballin, L. Karlinsky, S. Alpert, S. Hasoul, R. Ben-Ari, E. Barkan, A Region Based Convolutional Network for Tumor Detection and Classification in Breast Mammography, in: Deep Learning and Data Labeling for Medical Applications, Springer, 2016, pp. 197–205.
[18] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, T. Brox, FlowNet: learning optical flow with convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2758–2766.
[19] Z. Teed, J. Deng, RAFT: recurrent all-pairs field transforms for optical flow, CoRR (2020) arXiv preprint arXiv:2003.12039.
[20] T. Kõstner, J. Pan, H. Qi, G. Cruz, C. Gilliam, T. Blu, B. Yang, S. Gatidis, R. Botnar, C. Prieto, LAPNet: non-rigid registration derived in k-space for magnetic resonance imaging, IEEE Trans. Med. Imaging 40 (12) (2021), doi:10.1109/TMI.2021.3096131. 1–1
[21] X. Bian, X. Luo, C. Wang, W. Liu, X. Lin, DDA-Net: unsupervised cross-modality medical image segmentation via dual domain adaptation, Comput. Methods Programs Biomed. 213 (2022) 106531.
[22] R. Ito, K. Nakae, J. Hata, H. Okano, S. Ishii, Semi-supervised deep learning of brain tissue segmentation, Neural Netw. 116 (2019) 25–34.
[23] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
[24] M. Früh, M. Fischer, A. Schilling, S. Gatidis, T. Hepp, Weakly supervised segmentation of tumor lesions in PET-CT hybrid imaging, J. Med. Imaging 8 (5) (2021) 054003.
[25] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, D. Rueckert, Self-supervised learning for medical image analysis using image context restoration, Med. Image Anal. 58 (2019) 101539.
[26] N. Wang, W. Zhou, Y. Song, C. Ma, W. Liu, H. Li, Unsupervised deep representation learning for real-time tracking, Int. J. Comput. Vis. 129 (2) (2021) 400–418.
[27] X. Li, W. Pei, Z. Zhou, Z. He, H. Lu, M.-H. Yang, Crop-transform-paste: self-supervised learning for visual tracking, arXiv preprint arXiv:2106.10900(2021).
[28] Z. Lai, E. Lu, W. Xie, MAST: a memory-augmented self-supervised tracker, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6479–6488.
[29] M. Chung, J. Lee, M. Lee, J. Lee, Y.-G. Shin, Deeply self-supervised contour embedded neural network applied to liver segmentation, Comput. Methods Programs Biomed. 192 (2020) 105447, doi:10.1016/j.cmpb.2020.105447. https://www.sciencedirect.com/science/article/pii/S0169260719305012
[30] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, et al., Big self-supervised models advance medical image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3478–3488.
[31] M. Frueh, A. Schilling, S. Gatidis, T. Kuestner, Real time landmark detection for within- and cross subject tracking with minimal human supervision, IEEE Access (2022). (Early Access)
[32] H. Li, Y. Fan, Non-rigid image registration using self-supervised fully convolutional networks without training data, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, 2018, pp. 1075–1078.
[33] T. Schmidt, R. Newcombe, D. Fox, Self-supervised visual descriptor learning for dense correspondence, IEEE Rob. Autom. Lett. 2 (2) (2016) 420–427.
[34] A. Jabri, A. Owens, A. Efros, Space-time correspondence as a contrastive random walk, Adv. Neural Inf. Process. Syst. 33 (2020) 19545–19560.
[35] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, in: Scandinavian Conference on Image Analysis, Springer, 2003, pp. 363–370.

[36] J. Wulff, M.J. Black, Efficient sparse-to-dense optical flow estimation using a learned basis and layers, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015, 2015.
[37] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, FlowNet 2.0: evolution of optical flow estimation with deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2462–2470.
[38] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, H. Bischof, Anisotropic Huber-L1 optical flow, in: BMVC, vol. 1, 2009, p. 3.
[39] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, T. Brox, A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4040–4048.
[40] D.J. Butler, J. Wulff, G.B. Stanley, M.J. Black, A naturalistic open source movie for optical flow evaluation, in: A. Fitzgibbon (Ed.), European Conf. on Computer Vision (ECCV), Part IV, LNCS 7577, Springer-Verlag, 2012, pp. 611–625.
[41] S. Gidaris, P. Singh, N. Komodakis, Unsupervised representation learning by predicting image rotations, arXiv preprint arXiv:1803.07728(2018).
[42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[43] Data science bowl cardiac challenge data, (https://www.kaggle.com/c/second-annual-data-science-bowl/data), Accessed: 2010-09-30.
[44] D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C.P. Langlotz, P.A. Heidenreich, R.A. Harrington, D.H. Liang, E.A. Ashley, et al., Video-based ai for beat-to-beat assessment of cardiac function, Nature 580 (7802) (2020) 252–256.

A.3 Real Time Landmark Detection for Within- and Cross Subject Tracking With Minimal Human Supervision

Authors: *Marcel Früh, Andreas Schilling, Sergios Gatidis, Thomas Küstner*

Published in: *IEEE Access*

Date of Publication: *August 2022*

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 July 2022, accepted 27 July 2022, date of publication 1 August 2022, date of current version 8 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3195211

RESEARCH ARTICLE

Real Time Landmark Detection for Within- and Cross Subject Tracking With Minimal Human Supervision

MARCEL FRUEH[1,2], ANDREAS SCHILLING[2], SERGIOS GATIDIS[1,3], AND THOMAS KUESTNER[1], (Member, IEEE)

[1]Medical Image and Data Analysis (MIDAS.lab), Department of Radiology, University Hospital Tübingen, 72076 Tübingen, Germany
[2]Department of Computer Science, Institute for Visual Computing, University of Tübingen, 72076 Tübingen, Germany
[3]Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany

Corresponding author: Marcel Frueh (marcel.frueh@med.uni-tuebingen.de)

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG) through German Research Foundation under Grant 438106095, and in part by the Germany's Excellence Strategy—EXC-2064/1 under Project 390727645.

This work involved human subjects or animals in its research. The author(s) confirm(s) that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Landmark detection plays an important role for a variety of image processing and analysis tasks. Current methods rely on either supervised or semi-supervised learning which often requires large labeled training datasets. Also, retrospective addition of further target landmarks after completion of training is difficult in current methods. In this paper we propose a framework that addresses these limitations and allows for landmark detection based on only few examples and for definition of target landmarks after completed training without retraining. Our proposed approach relies on self-supervised training on a within-image template matching task with regularization by data augmentation. The trained network generalizes to cross-image matching and can thus be extended to example-based landmark detection and tracking. We extensively evaluate the proposed framework on chest X-ray images and abdominal MRI scans and demonstrate high accuracy with only few or even only one labeled example. Additionally we apply it to the task of liver and liver lesion tracking in CINE MRI scans.

INDEX TERMS Landmark detection, magnetic resonance imaging, self-supervised learning, real time motion tracking, x-ray.

I. INTRODUCTION

Automated analysis of image data plays a central role in medical imaging [1], [2]. To this end, anatomical target structures must be reliably recognized in order to enable subsequent processing steps for a wide variety of diagnostic tasks. A typical task of automated image analysis is the detection and tracking of anatomical landmarks within and between images, i.e. the identification of points in images that have structurally similar neighborhoods and similar semantic properties. In applications such as image-guided radiotherapy, the anatomically accurate and low-latency tracking of lesions over time is crucial to administer localized beams for the target lesion. Established methods for automated landmark detection are mostly based on supervised machine learning methods [3]–[5] which rely on large amounts of manually labeled training data. Although these methods provide powerful predictive models, their widespread application to various image data is limited due to the lack of manually annotated training data. Particularly with many landmarks per image, the effort required for manual annotations increases considerably. In addition, supervised methods require landmarks to be defined beforehand; adding additional landmarks usually requires re-training of

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

the model and, importantly, requires access to the initial training dataset as well. Contrastive methods [6]–[8] alleviate these problems but typically violate the real-time constraint due to their extensive feature comparison and are therefore not suitable for tasks where real-time tracking is required, such as image-guided radiotherapy.

In contrast to natural image data, medical images of a particular modality and body region exhibit a high degree of regularity based on a common anatomical structure. We attempt to leverage this regularity to learn a global positional embedding of local image patches in a self-supervised way using targeted data augmentation on a within-image template matching task. This allows to implement a simple yet effective one-shot landmark detection method that requires only a single annotated example per landmark. Additionally, the framework can be extended to an arbitrary number of landmarks without any additional re-training of the model.

In this work, we propose a framework for real-time landmark detection and tracking which is trained self-supervised on minimally labeled data. In contrast to self-supervised methods [9], [10] that rely on similarity measures of image patches (e.g. through contrastive learning), we propose a local-to-global positional embedding which allows for computationally efficient predictions that enable its application in fields where real-time interaction is required. The proposed framework is demonstrated and investigated for automatic real-time liver lesion tracking in time-resolved abdominal magnetic resonance imaging (MRI) and real-time automated liver tracking for image-guided radiotherapy on magnetic resonance linear accelerator (MR-LINAC) data, both of which are subject to respiratory motion. Furthermore we prove the practicability of our method for automated detection of anatomical landmarks in conventional chest X-rays.

A. RELATED WORK

Numerous studies have been published on landmark detection and tracking using a variety of methods and applications [11], [12]. Early work focused on conventional image processing techniques based on hand-crafted features, e.g., for facial feature recognition [13]. More recent papers demonstrate the use of machine learning methods such as regression trees [14] or SVMs [15] and lately mostly Deep learning-based methods using convolutional neural networks in various flavors, e.g. multi-task learning [16], reinforcement learning [17], [18], fully convolutional networks [19], regression networks [20], [21], siamese networks [22], [23] or transformers [24], [25].

While state-of-the-art supervised landmark detection frameworks provide highly accurate predictions, they still rely on large amounts of labeled training data.

Data efficient landmark detection using only a few labeled samples (few- or one-shot learning) has long been an area of scientific interest [26]–[29]: Common approaches for few-shot learning in this context typically rely on semi-supervised [30], [31] or self-supervised learning frameworks consisting of random walk based methods [32], [33], cross-input consistency [34] or neural rendering [35]. Recently, single-shot learning for anatomical landmarks has been introduced by Yan *et al.* [9] which uses contrastive learning to learn local and global embeddings on radiological images for cross-image landmark detection.

II. METHODS

A. CONTRIBUTIONS

We introduce a framework for landmark detection and tracking that

1) does not require labeled data during training and only requires a single labeled example at inference
2) allows for definition of target landmarks at inference time without re-training and without access to the initial training data.
3) directly returns the position of the object combined with to track to enable real-time detection

To this end, we implement a two-step procedure. In the first step, a (siamese-like) neural network [36] is trained on a within-image template matching task [37], [38] using self-supervision and targeted data augmentation. This is in contrast to the existing supervised template matching based tracking methods, which leverage existing positional labels [22], [36]. In the second step, after training, landmarks are identified in a target image by providing a single labeled example patch containing the target landmarks as input to the trained model. This step does not require re-training of the network. We implicitly make the assumption that training data as well as labeled examples are drawn from the same distribution of images that contain a specific object or structure.

B. SELF-SUPERVISED TEMPLATE MATCHING

The template matching task consists of estimating the center position of extracted image patches within source images. The motivation for using this task is that it allows to implicitly learn the distribution of object characteristics within the training data. This in turn should enable subsequent identification of specific landmarks.

In detail, squared image patches $\mathbf{P_I}$ of predefined size are uniformly drawn from the respective source images $\mathbf{I}$.

Both, patch and source image are then fed to a template matching neural network f^θ (with weights θ) to output an estimate $(\hat{x}, \hat{y})$ of the patch center coordinates. We further assume an aleatoric heteroscedastic Gaussian distribution of the samples.

Formally, we thus model this problem as the task to learn the conditional distribution of the center coordinates given the source image and the extracted patch under the assumed Gaussian distribution:

$$\mathbf{P}(\mathbf{C}|\mathbf{P_I}, \mathbf{I}) = \mathcal{N}_2(\boldsymbol{\mu}, \Sigma), \tag{1}$$

where $\mu = (x, y)$ are the patch center coordinates and $\Sigma = \text{diag}(\sigma_x, \sigma_y)$ describes the variance.

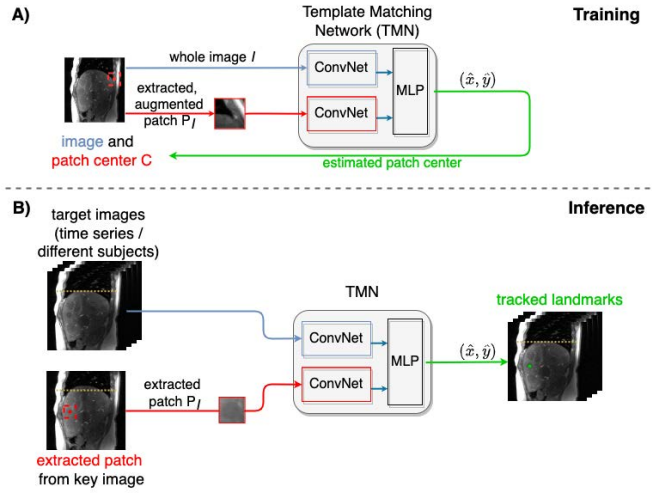


FIGURE 1. The proposed landmark detection and tracking framework consists of a Template Matching Network that combines two encoders (ConvNet) whose outputs are fed into a shared multi-layer perceptron (MLP). One encoder is used to process the full source image (global information) whereas the second encoder processes the extracted (during training augmented) patch (local information). A) During training, the patch is drawn from the same image. B) When performing inference, the patch is manually chosen from an initial key image, extracted and i) tracked in the target image (time series of subsequent images) of the same subject (within-subject tracking), or ii) the same anatomical landmark is identified across the target subjects (cross-subject tracking).

The sampling loss is then given by the negative log-likelihood for x and y, respectively:

$$\mathcal{L}(x, y, \hat{x}, \hat{y}, \hat{\sigma}_x, \hat{\sigma}_y) = \frac{1}{\hat{\sigma}_x^2}(\hat{x} - x)^2 + \ln(\hat{\sigma}_x) + \frac{1}{\hat{\sigma}_y^2}(\hat{y} - y)^2 + \ln(\hat{\sigma}_y) \quad (2)$$

Here, $(\hat{x}, \hat{y}, \hat{\sigma_x}, \hat{\sigma_y}) = f^{\theta}(P_I, I)$ is the network output where $\hat{x}, \hat{y}$ are the estimated patch center positions and $\hat{\sigma_x}, \hat{\sigma_y}$ denote the estimated standard deviation or uncertainty of the given predictions.

The template matching network (TMN) architecture (Fig. 1 A) consists of two separate feature encoders (one for the source image and one for the extracted patch; no weight sharing), each resulting in its own feature vector. These encoders are based on the VGG16 architecture [39] and are pretrained on imageNET [40]. The stacked feature vectors are then fed into three fully connected layers with four linear outputs $(\hat{x}, \hat{y})$ and $(\hat{\sigma_x}, \hat{\sigma_y})$.

C. CROSS-IMAGE MATCHING AND LANDMARK TRACKING

Beyond identifying similar patches within the same image, our goal was to achieve generalization for cross-image landmark detection, i.e. i) tracking a landmark for a given subject over time (series of time-resolved images) or ii) matching/detecting landmarks between different subjects. In the following, we refer to these two cases as i) within-subject tracking and ii) cross-subject detection, respectively. Under the assumption that all training images contain the same or similar objects, we hypothesize that this generalization can be achieved by regularization through data augmentation. Thus,

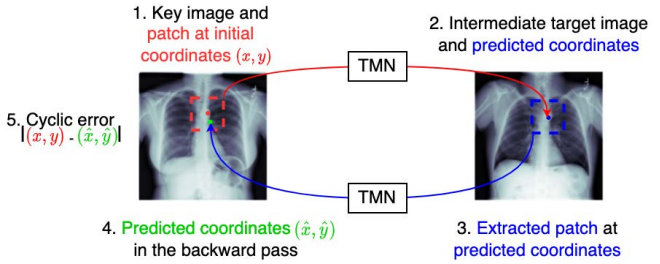


FIGURE 2. Label-free evaluation: Cyclic evaluation routine for cross-image matching. First, a patch is extracted from the source key image (1). The corresponding image patch position is then estimated by the template matching network (TMN) in the intermediate target image (2). In a backward pass, a patch around the predicted coordinates is extracted from the intermediate target image (3) and fed into the TMN to estimate the corresponding position in the original source key image (4). This estimated position is then compared to the initial patch position to compute a cyclic error (5).

we apply domain-specific data augmentation to the whole images (Rotation: -10° to 10°, affine scaling: 0.8 to 1.2 and random resized crops) as well as to the extracted image patches (Rotation: -5° to 5°, affine scaling: 0.9 to 1.3 and gamma contrast variation: (0.5, 2)) [41]. It is important to mention, that the patch center, which represents the target coordinate, is fixed during the augmentation steps. Thus, no translation is used.

The described within-image template matching task can be extended to a cross-image matching task (cross-subject detection and within-subject tracking), where - given an extracted patch from a source key image - the goal is to estimate the semantically corresponding location (coordinates) of this patch in a target image (Fig. 1 B).

Cross-image matching can be naturally extended to example-based landmark detection by feeding both the target image and an extracted patch of the source image containing the desired landmark as its center point to the trained network. The example patch is drawn from an image where the landmark position is known, e.g after manual labeling.

For some databases, we may have access to more (>1) labeled landmarks in the dataset. Thus, to leverage the availability of larger labeled datasets for cross-subject landmark detection, we extend the described procedure to allow for multiple example patches as follows.

Given N example patches, the estimated landmark position $(\hat{x}, \hat{y})$ within the target image is obtained based on these examples by uncertainty-weighted averaging over the single coordinate estimates based on each example patch:

$$\hat{x} = \frac{\sum_{i=1}^{N} \hat{x_i} \cdot \frac{1}{\hat{\sigma}_{x_i}}}{\sum_{i=1}^{N} \frac{1}{\hat{\sigma}_{x_i}}}, \quad (3)$$

where $\hat{x_i}$ is the estimated landmark position based on the i^{th} example patch with corresponding estimated uncertainty $\hat{\sigma_{x_i}}$. $\hat{y}$ is computed in the same fashion.

D. LABEL-FREE EVALUATION

When applying the trained model to cross-image matching, no direct ground truth is available, in contrast to the

initial self-supervised within-image template matching task. This poses a challenge when it comes to the evaluation of algorithm performance for the cross-image matching task. We therefore use a process of label-free evaluation that uses a cyclic estimation of corresponding landmarks between two images [42](Fig. 2). This cyclic evaluation is performed in a two-step procedure: In the forward pass, source patches are extracted for every second pixel within a source key image. For each of these patches, corresponding center coordinates are estimated on N target images using the trained model. In the backward pass, patches are sampled at the estimated coordinates of the target images and used as input to the trained model to estimate the corresponding center positions in the original source key image. The absolute cyclic error $E_{(x,y),i}$ at a given coordinate (x, y) within the source image can be computed for each of the N target images allowing for label-free estimation of model accuracy via

$$E_{(x,y),i} = |f^{\theta}(\mathbf{P}_{\mathbf{T_i}}(f^{\theta}(\mathbf{P}_{\mathbf{S}}(x, y), \mathbf{T_i})), \mathbf{S}) - (x, y)|, \quad i \leq N \tag{4}$$

where S is the source image, T_i the i^{th} target image, $f^{\theta}(\cdot)$ the trained model output (estimated coordinates), $\mathbf{P_S}$ the patch extracted from the source image at position (x, y) and $\mathbf{P_{T_i}}$ the patch extracted from the i^{th} target image at the estimated coordinates in the forward pass. The mean and standard deviation of these errors over all N target images can subsequently be computed for each pixel position in the source image (Fig. 2).

III. EXPERIMENTS

For the purpose of evaluation, we applied the proposed framework in use cases from two medical imaging domains as depicted in Fig. 3. Landmark tracking (MR-LINAC and abdominal MRI) and cross-image matching (chest X-ray) are investigated.

MR-LINAC imaging was performed on a 1.5T MR-LINAC scanner (Philips Healthcare, Best, the Netherlands) in patients undergoing radiotherapy treatment. Images were acquired with a balanced fast field echo sequence yielding time-resolved images of the upper abdomen. The database includes 230 studies of 50 patients (20 female, 66 $\pm$ 11.52 years, matrix size = 352 $\times$ 352; acquisition time/image = 0.5s) with three sequences in axial, coronal and sagittal orientation each, resulting in a total of 165,264 single image slices. Patient data were acquired in the context of a clinical phase II trial (NCT04172753). Data is used for within-subject liver tracking under respiratory motion.

The abdominal MRI data was acquired on a 3T PET/MR (Siemens Biograph mMR, Siemens Healthcare, Erlangen, Germany) in patients with suspected liver or lung metastases for the purpose of respiratory motion correction. In this work, the data is used to track liver lesions under respiratory motion within subjects. Imaging was performed with a spoiled gradient echo sequence (TE/TR = 1.8ms/3.6ms; flip angle = 15°; bandwidth = 670Hz/pixel; resolution = 2 $\times$ 2mm^2; matrix size = 192 $\times$ 176; acquisition time/image = 0.4s)

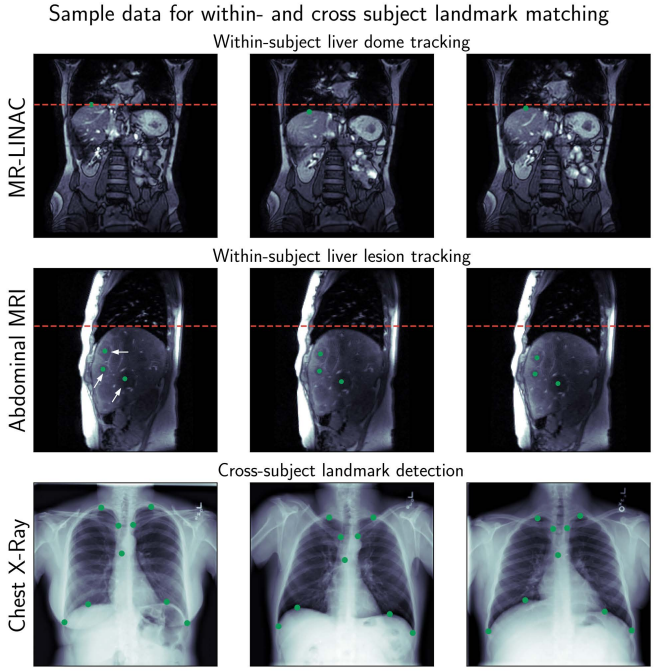


FIGURE 3. Sample data with corresponding ground truth annotations of the landmark to track. The top row visualizes a full respiratory cycle during radiotherapy. The landmark defines the liver dome to track. The central row denotes abdominal MRI at end-expiration, mid-expiration and end-inspiration. Several lesions are visible within the liver of which three example annotated lesions are highlighted by white arrows. The bottom row visualizes three different chest X-ray images with corresponding ground truth annotation as described in III-C .

yielding 2D sagittal motion-resolved MR images of the body trunk [43]. 36 patients (60$\pm$9 years, 20 female) were acquired resulting in 12214 individual slices. The study was approved by the local ethics committee and all patients provided written consent.

The chest X-ray dataset reflects a cross-subject landmark detection task based on Chexpert [44] and contains 224,316 chest X-ray from 65,240 patients (60 $\pm$ 17.8 years, 40.6% female). Images were acquired with varying matrix size (resampled to 224 $\times$ 224) with and without pathological findings.

All images were zero-padded with $\frac{\text{patch size}}{2}$ on each side to ensure that patches could also be sampled from the image margins. Experiments were performed using different patch sizes (32, 40, 50, 60, 70 and 80 pixels) of extracted squared image patches in order to assess the effect of patch size on model performance. For all databases a 60-20-20 train-test-val split with a patient-leave-out approach was used, i.e. unique patients were assigned to each set. A subsequent test dataset was kept separate for all three tasks for final evaluation of the following experiments: Template matching (III-A), cross-image matching (III-B) and example-based landmark detection (III-C).

The proposed template matching network was trained for 1000 epochs with a batch size of 192 using the Adam optimizer [45] ($\beta_1 = 0.9$, $\beta_2 = 0.999$) and an initial learning rate of 1e-4 that is scaled by 0.85 every 80 epochs on a NVIDIA RTX3090 GPU using PyTorch 1.8 [46].

A. PIXEL-WISE TEMPLATE MATCHING

To assess the performance of the training and hence the template matching ability, we compute and report the mean, and standard deviation of the euclidean template matching errors for every pixel in all test images, paired with corresponding uncertainty.

B. PIXEL-WISE CROSS-IMAGE MATCHING

For within-subject tracking (MR-LINAC and abdominal MRI datasets), the cycle errors were computed for every second pixel on 10 image pairs from the test dataset, each pair consisting of the slice of the end-inspiration phase as target image and the slice containing the end-expiration phase as source key image. For cross-subject detection on chest X-ray images, cyclic errors were computed using 10 randomly chosen intermediate target images.

Mean and standard deviation of euclidean cyclic errors were calculated based on all pixels on the corresponding test datasets.

C. EXAMPLE-BASED LANDMARK MATCHING

To evaluate the performance of the proposed framework for identification of predefined landmarks, ground truth data for specific landmarks were generated by an experienced radiologist (S.G., >10 years of experience) for all three datasets. We compute and report the mean, standard deviation, as well as the maximum of the euclidean error between prediction and labeled ground truth over all test subjects and landmarks.

1) WITHIN-SUBJECT MOTION TRACKING

For liver tracking (MR-LINAC), the liver dome was annotated on all slices for 5 subjects for one respiratory cycle in the sagittal and coronal orientation.

For the task of lesion tracking (abdominal MRI), 10 lesions were manually annotated in all slices.

For both tasks, the source slice depicts the state of maximal end-expiration.

2) CROSS-SUBJECT ANATOMICAL LANDMARK DETECTION

On the chest X-ray data, 9 landmarks were manually labeled on 100 images representing the left and right pleural recesses, the left and right diaphragmal domes, the left and right pulmonary apeces, the left and right sternoclavicular joints as well as the carina of the trachea (Fig. 3).

To differentiate between single-shot and few-shot application, up to 50 of these labeled images were used as examples and 50 were used as target images for evaluating the accuracy of example-based landmark detection. Mean euclidean errors, as well as minimal and maximal mean euclidean errors between prediction and ground truth for landmark detection were computed based on all landmarks in the 50 target images.

To assess the performance for the generation of ground truth data based on a single example, we also evaluate the cross-subject landmark detection capability on the model trained on the MR-LINAC dataset. Good performance on this task would allow for efficient creation of large, annotated datasets based on only few labeled samples.

D. COMPARISONS TO BASELINE MODELS

1) COMPARISON TO A SUPERVISED NETWORK BASELINE (SUPERVISED BASELINE)

In order to provide a baseline comparison to fully supervised landmark detection, we used a ResNet-50 [47] CNN pretrained on imageNET with 18 outputs for the chest X-ray images (x and y coordinates for 9 target landmarks) to estimate the coordinates for all landmarks in a single prediction. The same labeled dataset that was used for example-based landmark detection (III-C2) was also used as training data for the supervised network. The network was trained for 20,000 steps using the Adam optimizer with a batch size of 50 and an initial learning rate of 1e-4.

2) COMPARISON TO A PATCH-WISE FEATURE MATCHING BASELINE (SimCLR PATCH)

We trained SimCLR [8] (ResNet-50 backbone) to produce a 1024-dimensional feature vector from squared 32×32 patches on the chest X-ray dataset for cross-subject detection.

The affinity matrix A between an initially selected key patch p_0 and all patches p_{ij} within the next subject is constructed via

$$A_t^{t+1}(i,j) = \langle h_p^\theta(p_0), h_p^\theta(p_{ij}) \rangle, \tag{5}$$

where $h_p^\theta(p)$ is the ℓ_2 normalized feature vector of the respective patch. Patch coordinate estimation is subsequently performed by choosing the patch with maximum affinity to the input patch.

For evaluation, the same labeled dataset as in III-C2 was used.

We trained the patch-wise feature matching baseline for 1000 epochs using proposed SimCLR parameters. Horizontal flips were removed from the data augmentation pipeline.

3) COMPARISON TO A SUPERVISED NETWORK BASELINE PRETRAINED WITH SimCLR (SimCLR PRETRAINED)

In a pre-training setup, we trained SimCLR (ResNet-50 backbone) to produce a 1024-dimensional feature vector from the full image on the chest X-ray dataset for cross-subject detection. Finetuning and inference was subsequently performed in the same setting as in (III-D1)

SimCLR was trained for 1000 epochs using the proposed SimCLR parameters, again without horizontal flips.

For all baseline comparisons, we recorded the mean euclidean landmark estimation error, as well as the inference time. In addition, we track the results against an increasing number of available training examples.

E. ABLATION STUDIES

1) DATASET SIZE

The influence of the available example patches (i.e the number of available labels) on the performance is assessed by computing the mean euclidean landmark errors for [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50] available example patches in each predefined landmark on the chest X-ray data. Squared 32×32 patches were used.

2) ENCODER

To study the impact of the encoder, we evaluate the mean euclidean landmark matching error for different encoders (VGG16 [39], ResNet50 [47], DenseNet121 [48] and ConvNext-Small [49]). Three fully connected layers were used for each architecture.

The training was conducted as described in III. For ConvNext-Small a batch size of 92 was used.

3) SHARED MULTI LAYER PERCEPTRON

To quantify the influence of the subsequent MLP we train the Resnet50 encoder with one to four fully connected layers, each consisting of 4096 neurons with ReLU and Dropout in between.

4) DISTRIBUTION

We investigate the impact that the choice of probability distributions has on training and the associated mean euclidean landmark matching error by comparing the Normal distribution (ℓ_2 loss) to the Laplace distribution (ℓ_1 loss).

IV. RESULTS

A. PIXEL-WISE TEMPLATE MATCHING

Results of pixel-wise template matching are depicted in the left column of Table 1. All results were averaged over all pixels and test subjects and obtained with a patch size of 80×80. In general, lower estimation errors were observed with increasing patch size for all three tasks (Fig. 5 left). Qualitative evaluation shows that higher errors and especially higher uncertainties typically occur in the background region (Fig. 4 left / central columns).

B. PIXEL-WISE CROSS-IMAGE MATCHING

For the task of cross-image matching we observed similar results as for the within-image template matching task. Corresponding results are depicted in the central column of Table 1 and were obtained with a patch size of 80×80. Again, the estimation error generally decreased with increasing patch size (Fig. 5 center), and similar to the template matching task, lower errors were observed in recurrent structures of the abdominal organs and chest regions, whereas higher errors occurred in the periphery and image background (Fig. 4 right column).

C. EXAMPLE-BASED LANDMARK MATCHING

Overall, we observed high accuracy for example-based landmark detection on all tasks using a patch size of 50×50 pixels. Quantitative results for liver dome tracking on MR-LINAC data, liver lesion tracking on the abdominal MRI dataset and estimation of the 9 predefined anatomical landmark positions on chest X-ray images based on one labeled example are depicted in Table 1 (right column). Maximal euclidean errors amounted to 3.5, 3.1 and 13.5 pixels for abdominal MRI, MR-LINAC and chest X-ray, respectively, with maximal motion-induced euclidean displacements of 8.9 and 10.1 pixels on abdominal MRI and MR-LINAC. Generally, smaller patch sizes yielded better results (Fig. 5 right). Qualitative evaluation is depicted in Figure 7.

Cross-subject landmark detection between different subjects within the MR-LINAC dataset resulted in a mean euclidean tracking error of 4.5 pixels. Qualitative evaluation can be found in Fig. 7 (bottom) and visualizes that our method is capable of tracking the liver dome between different subjects.

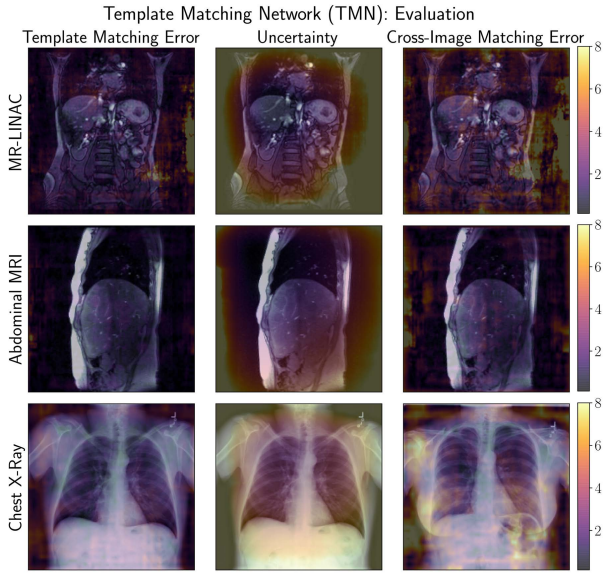


FIGURE 4. Color-coded mean euclidean pixel-wise template matching error (left column, III-A), corresponding estimated uncertainty (central column, III-A) and mean euclidean pixel-wise cross-image matching error (right column, III-B) for each pixel in three representative examples from the MR-LINAC (top row), the abdominal MRI (central row) and the chest X-ray (bottom row) for a patch size of 80. For MR-LINAC and abdominal MRI, a subsequent frame from the same subject was used, whereas the chest X-ray from a different subject was used.

TABLE 1. Evaluation of the proposed Template Matching Network: Mean ± standard deviation of euclidean errors (in pixels) for pixel-wise template matching (III-A), cross-image matching (III-B) and example-based landmark matching (III-C).

Dataset	Template Matching Network (TMN): Evaluation		
	Template Matching [px]	Cross-Image Matching [px]	Landmark Matching [px]
MR-LINAC	3.7 ± 5.4	6.6 ± 8.7	1.8 ± 1.7
Abdominal MRI	2.2 ± 2.8	4.3 ± 4.2	2.1 ± 0.94
Chest X-ray	2.2 ± 1.5	4.4 ± 3.7	5.8 ± 3.9

D. BASELINE COMPARISONS

1) COMPARISON TO A SUPERVISED NETWORK BASELINE (SUPERVISED BASELINE)

The fully supervised landmark detection network yielded a markedly higher landmark estimation error using only few

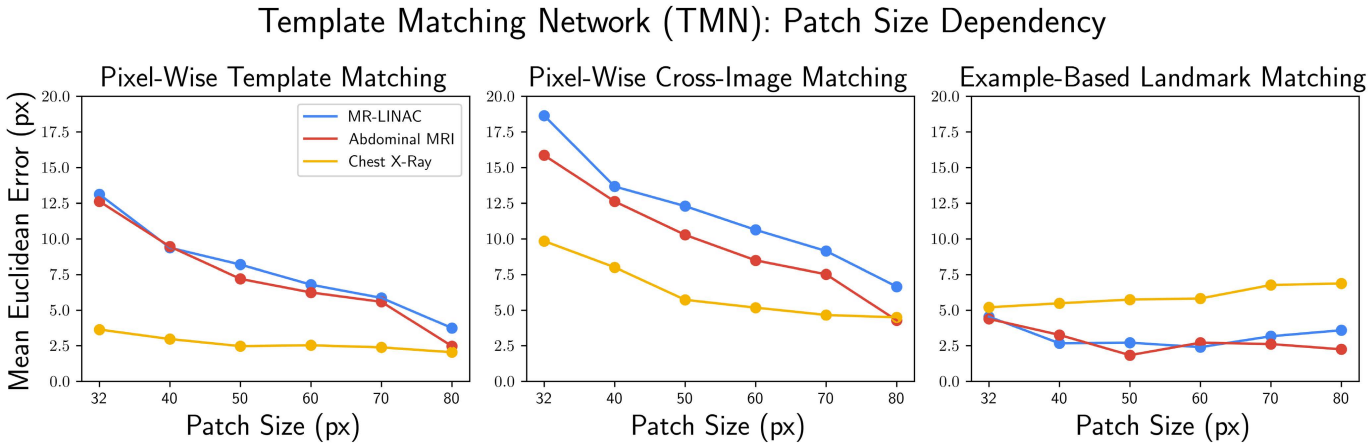


FIGURE 5. Dependency of the mean euclidean errors on various patch sizes in (left) template matching (pixel-wise template matching, III-A), (central) cross-image matching (cyclic error; within-subject tracking for MR-LINAC and abdominal MRI, cross-subject detection for chest X-ray, III-B) and (right) selected example-based landmarks (within-subject motion tracking for MR-LINAC and abdominal MRI, cross-subject anatomical landmark detection for chest X-ray. III-C).

examples and failed to reach the same accuracy as our proposed framework even using 50 labeled training examples (mean euclidean landmark matching error of 8.5 pixels) as shown in Fig. 6 (red).

2) COMPARISON TO A PATCH-WISE FEATURE MATCHING BASELINE (SimCLR PATCH)

Patch-wise feature matching remarkably outperformed the supervised baseline, even with only 5 samples (mean euclidean landmark matching error of 6.6 pixels). Further increase in the number of available examples did not improve performance much. Quantitative evaluation of patch-wise feature comparison is depicted in Fig. 6 (orange).

3) COMPARISON TO A SUPERVISED NETWORK BASELINE PRETRAINED WITH SimCLR (SimCLR PRETRAINED)

In contrast to patch-wise feature comparison, fine-tuning of the SimCLR network benefits from each additional training example. Compared to the supervised baseline without any pretraining, the mean euclidean error is reduced by 30% yielding a mean euclidean landmark matching error of 6.3 (Fig. 6, green).

Comparison of our template matching network (patch size 32×32) and all baselines is depicted in Table 2. Both, best results (top, few-shot) and results for only one labeled example (bottom, single-shot) are evaluated in terms of mean euclidean landmark matching error and inference time. For the best scores, 30 labeled example patches were used for patch-wise feature matching, 20 for our framework, and 50 for the two supervised baselines.

E. ABLATION STUDIES

1) DATASET SIZE

Regarding the impact of the number of landmark example patches on mean euclidean landmark matching error, we observed that the landmark estimation errors decreased rapidly from 1 to 20 examples, reaching optimal accuracy at already 20 examples. No further performance gain was observed using 30, 40 or 50 examples (Fig. 6, blue).

TABLE 2. Comparison of Template Matching Network (TMN) to baseline methods for example-based landmark matching (III-C) in the chest X-ray dataset. Inference time on GPU (CPU) (in ms) are reported. Top: Best euclidean landmark matching errors reported as mean ± standard deviation (in pixels). Bottom: Comparison for one labeled example.

	Method	Euclidean Error [px]	Inference Time [ms]
Few Shot	Supervised Baseline	8.5 ± 5.7	11 (90)
	SimCLR: Patch	6.2 ± 4.3	1200 (140,000)
	SimCLR: Pretrained	6.3 ± 4.0	11 (90)
	TMN (proposed)	5.0 ± 3.4	17 (150)
Single Shot	Supervised Baseline	15.6 ± 10.4	11 (90)
	SimCLR: Patch	14.8 ± 6.7	1200 (140,000)
	SimCLR: Pretrained	14.9 ± 11.4	11 (90)
	TMN (proposed)	5.6 ± 3.8	6 (75)

2) ENCODER

Quantitative Analysis of using different encoders is depicted in Table 2 (top) for 20 example images and a patch size of 32×32. Corresponding qualitative analysis is visualized in Fig. 7 (bottom)

No relevant differences between the encoders could be observed, however modern architectures seem to yield slightly superior results compared to VGG-16. All architectures are real-time capable, also on CPU.

3) SHARED MULTI LAYER PERCEPTRON

Results are depicted in Table 2 (bottom) for 20 example images and a patch size of 32×32. A single linear layer was not enough to reliably track landmarks. Increasing the layers gradually increases the performance, reaching its optimum at 3 layers.

4) DISTRIBUTION

When comparing the impact of the distribution (Normal vs Laplace, Table. 3 bottom) we could not observe any relevant performance differences.

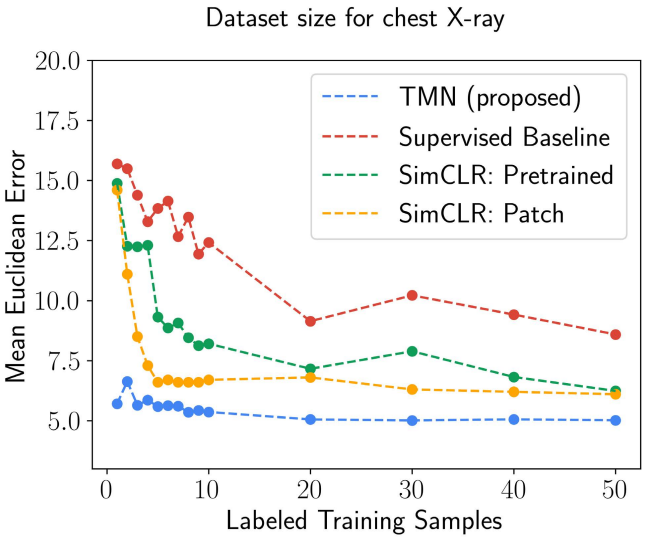


FIGURE 6. Dependency of the average euclidean landmark matching (cross-subject) error on the number of available labeled examples in the 9 chest X-ray landmarks (III-C2) for the proposed template matching network (TMN) (blue), a supervised baseline (red), the patch-wise feature matching baseline (orange) and the supervised baseline pre-trained with SimCLR (green).

TABLE 3. Euclidean landmark matching errors for ablation studies of the template matching network (TMN) in the example-based landmark matching (III-C) of the chest X-ray dataset. Mean ± standard deviation of euclidean errors (in pixels) as well as inference times (in ms) on GPU are reported for different encoder architectures (top) and varying numbers of fully connected layers (bottom).

Encoder	Euclidean Error [px]	Inference Time [ms]
VGG16	5.0 ± 3.5	17 (150)
ResNet50	4.7 ± 3.2	21 (110)
DenseNet121	4.7 ± 3.2	41 (107)
ConvNext-Small	4.8 ± 3.4	20 (316)
Resnet50-1	11.5 ± 8.9	17 (98)
Resnet50-2	5.3 ± 3.8	20 (105)
Resnet50-4	4.9 ± 3.4	24 (120)
ResNet50-Normal	4.7 ± 3.2	21 (110)
ResNet50-Laplace	4.5 ± 3.3	21 (110)

V. DISCUSSION

In this work we introduced a framework for real-time capable landmark detection and tracking on medical images that can be trained on a fully self-supervised basis. Given one or more example images defining the landmarks of interest, our proposed algorithm is able to identify these landmarks on unseen test images. We showed that this approach yields good performance in within-subject (e.g. time series) as well cross-subject landmark detection. In contrast to supervised approaches [50], our proposed framework does not require re-training of the model for detection of specific landmarks but instead relies on the presentation of examples containing target landmarks. Importantly, and in contrast to other self-supervised frameworks [9], [32], [51], due to its high inference speed, it allows for application in real-time critical areas, such as image guided radiotherapy systems where it could potentially allow for tracking of target structures and thus adjustment of treatment parameters.

Evaluation of the template matching capability revealed that our framework successfully learned to map example patches to coordinates. The higher error in background regions indicates that it does not implement a pure template matching strategy but actually learns relevant and recurrent anatomical structures. This is supported by the predicted uncertainty, especially within the MR-LINAC dataset (Fig. 4, top row, central image). Targeted regions of interest (e.g. liver) have a significantly decreased uncertainty compared to background regions or anatomical structures that do not occur regularly within the database (e.g. pelvic region).

Of course, the focus of our framework is not on identifying structures within the same image from which the patch was selected, but across other images (between subjects or in time-series). Evaluation of the pixel-wise tracking capability across different subjects (cross-subject detection) or across different time steps (within-subject tracking) revealed similar results compared to the template matching task effectively showing the ability of our framework to generalize. The error typically increases within background regions or non-recurring structures.

Evaluation of tracking performance of individual anatomical landmarks (9 landmarks for cross-subject chest X-ray images, one landmark for cross-subject liver dome detection, one landmark for within-subject liver dome tracking, multiple landmarks for within-subject lesion tracking) yielded satisfactory results. Overall, motion tracking performance of the within-subject tasks was slightly superior compared to cross-subject performance due to the higher level of similarity. The tracking accuracy is in the lower millimeter range (even maximal displacement errors < 1cm) in contrast to a motion displacement of up to several centimeters (IV-C) which would render acceptable results for any prospective motion tracking or correction strategies.

The patch size is a crucial parameter depending on the underlying task at hand. In general, smaller patch sizes (around 32-50 pixels) tend to yield superior performance for the task of anatomical landmark detection (cross-subject) and tracking (within-subject) compared to larger patches (Fig. 5 right). Contrary to that, when inspecting all pixels within an image, larger patches resulted in better performance due to better background region recognition.

In contrast to motion tracking, where typically only one labeled example patch can be leveraged, multiple patches can be used for cross-subject landmark detection. The use of up to 20 example patches yields a steady improvement in performance. Using additional examples does not improve the result any further on the chest X-ray dataset. We hypothesize that the reason for this pattern resides in the averaging of the individual predictions paired with label noise. The occurrence of this behavior in the patch-wise feature comparison baseline experiment supports this hypothesis.

A single labeled example outperformed all supervised and self-supervised baselines.

The central concept of our proposed approach is the extension of a within-image template matching task to

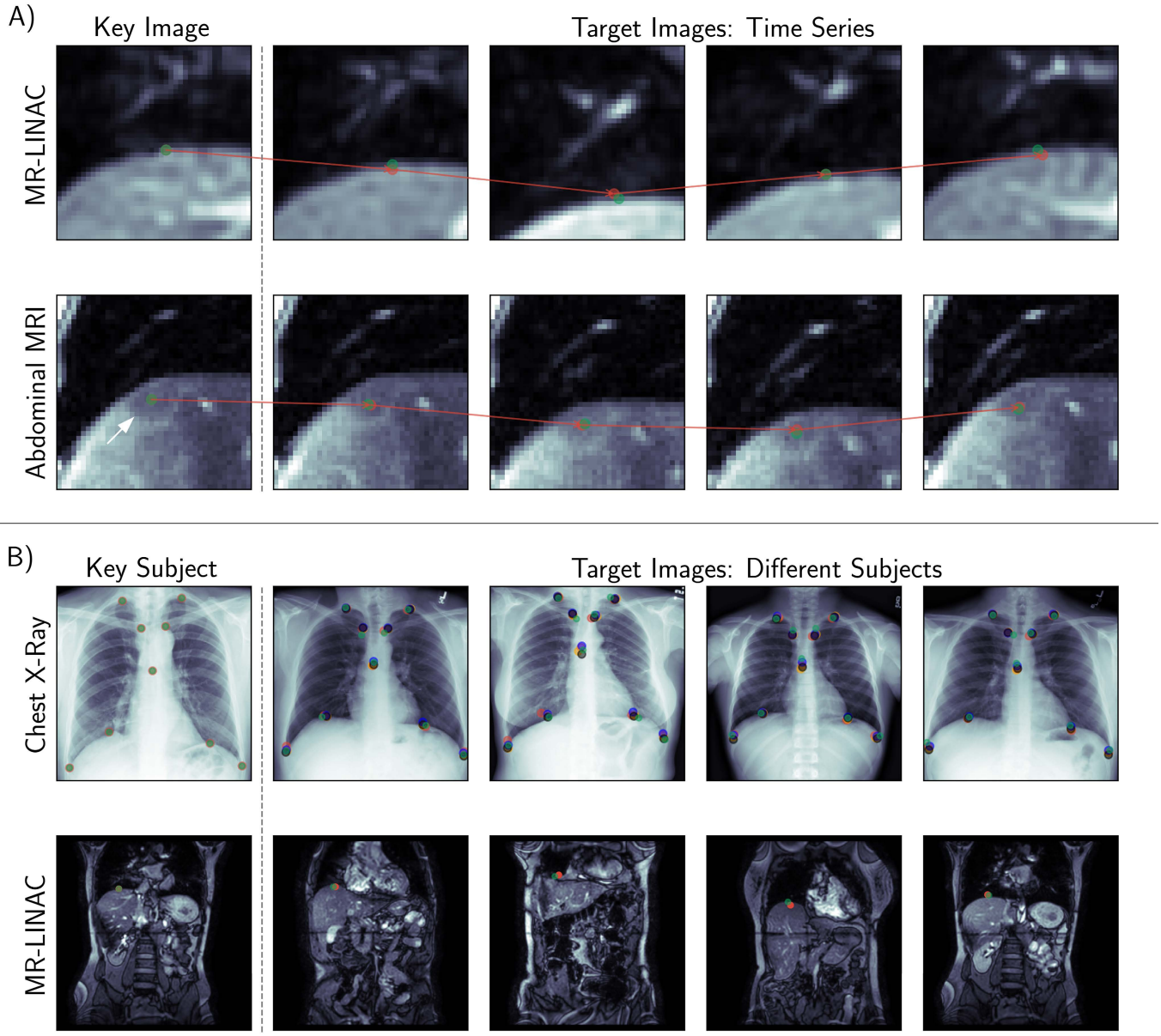


FIGURE 7. Qualitative evaluation of example-based landmark matching (III-C) for (A) within-subject motion tracking and (B) cross-subject anatomical landmark detection. A) Visualized within-subject motion tracking of the liver dome on MR-LINAC data (top) and a liver lesion within the abdominal MRI dataset under respiratory motion (bottom). The red dots indicate the predicted liver dome or tumor lesion center whereas the green ones visualize the ground truth. The red line depicts the displacement of the liver/lesion over time with respect to motion. A complete respiratory cycle of approximately 3s is visualized. B) Example-based cross-subject landmark detection on chest X-ray images (top) with 20 labeled example images and MR-LINAC data (bottom) of one labeled example image. The red dots indicate the predicted landmark whereas the green dots visualize the corresponding ground truth annotation. For the chest X-ray images the predictions are depicted for VGG (red), ResNet (orange), DenseNet (blue) and ConvNext (black).

cross-image matching. This generalization is induced by regularization through data augmentation. Thus, the model focuses on features that are present across all images within the given domain. This becomes evident by our observation that the template matching and cross-image matching tasks yield better performance on foreground regions compared to background regions. Possible applications of our proposed framework include areas where limited amounts of training data are available or where the set of target landmarks needs to be adapted after training without the ability to re-train the model. The proposed framework can potentially be extended to further tasks beyond landmark detection, such as object detection or segmentation and may allow for efficient processing of these tasks based on only few examples.

We acknowledge the following limitations: Coordinate estimation is task specific and might not work well if the image content or its resolution varies substantially. Thus, while being efficient for motion tracking, the performance of cross-subject anatomical landmark detection might suffer from higher variance. Determining the correct patch size depends on the overall goal and image size, thus inductive

bias or costly hyperparameter tuning may be required to determine a good patch size.

A natural extension of our work is application on 3D image data which will be part of future work.

In conclusion, we were able to demonstrate a self-supervised framework for both, cross- and within subject landmark detection and tracking that is capable of running in real-time.

REFERENCES

[1] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, May 2019.

[2] B. Sahiner, A. Pezeshk, L. M. Hadjiiski, X. Wang, K. Drukker, K. H. Cha, R. M. Summers, and M. L. Giger, "Deep learning in medical imaging and radiation therapy," *Med. Phys.*, vol. 46, no. 1, pp. e1–e36, Jan. 2019.

[3] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 918–930, May 2016.

[4] Y. Wu and Q. Ji, "Robust facial landmark detection under significant head poses and occlusion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3658–3666.

[5] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.

[6] F. C. Ghesu, B. Georgescu, A. Mansoor, Y. Yoo, D. Neumann, P. Patel, R. S. Vishwanath, J. M. Balter, Y. Cao, S. Grbic, and D. Comaniciu, "Self-supervised learning from 100 million medical images," 2022, *arXiv:2201.01283*.

[7] A. Bardes, J. Ponce, and Y. LeCun, "VICREG: Variance-invariance-covariance regularization for self-supervised learning," in *Proc. ICLR*, 2022, pp. 1–23.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[9] K. Yan, J. Cai, D. Jin, S. Miao, D. Guo, A. P. Harrison, Y. Tang, J. Xiao, J. Lu, and L. Lu, "SAM: Self-supervised learning of pixel-wise anatomical embeddings in radiological images," *IEEE Trans. Med. Imag.*, early access, Apr. 20, 2022, doi: 10.1109/TMI.2022.3169003.

[10] X. Wang, A. Jabri, and A. A. Efros, "Learning correspondence from the cycle-consistency of time," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2566–2576.

[11] R. Zhang, L. Xu, Z. Yu, Y. Shi, C. Mu, and M. Xu, "Deep-IRTarget: An automatic target detector in infrared imagery using dual-domain feature extraction and allocation," *IEEE Trans. Multimedia*, vol. 24, pp. 1735–1749, 2021.

[12] R. Zhang, L. Wu, Y. Yang, W. Wu, Y. Chen, and M. Xu, "Multi-camera multi-player tracking with deep player identification in sports video," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107260.

[13] R. Lim and T. MJT Reinders, "Facial landmark detection using a Gabor filter representation and a genetic search algorithm," in *Proc. ASCI Conf.*, Lommel, Belgium, 2000.

[14] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1867–1874.

[15] M. Uricár, V. Franc, and V. Hlavác, "Detector of facial landmarks learned by the structured output SVM," in *Proc. Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2012, pp. 547–556.

[16] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 94–108.

[17] F. C. Ghesu, B. Georgescu, T. Mansi, D. Neumann, J. Hornegger, and D. Comaniciu, "An artificial agent for anatomical landmark detection in medical images," in *Proc. Int. Conf. Med. Image Comput.-Assisted Intervent.* 2016, pp. 229–237.

[18] A. Vlontzos, A. Alansary, K. Kamnitsas, D. Rueckert, and B. Kainz, "Multiple landmark detection using multi-agent reinforcement learning," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.* 2019, pp. 262–270.

[19] D. Merget, M. Rock, and G. Rigoll, "Robust facial landmark detection via a fully-convolutional local-global context network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 781–790.

[20] D. Lachinov, A. Getmanskaya, and V. Turlapov, "Cephalometric landmark regression with convolutional neural networks on 3D computed tomography data," *Pattern Recognit. Image Anal.*, vol. 30, no. 3, pp. 512–522, Jul. 2020.

[21] X. Miao, X. Zhen, X. Liu, C. Deng, V. Athitsos, and H. Huang, "Direct shape regression networks for end-to-end face alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5040–5049.

[22] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 850–865.

[23] A. He, C. Luo, X. Tian, and W. Zeng, "A twofold Siamese network for real-time object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4834–4843.

[24] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8126–8135.

[25] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "TrackFormer: Multi-object tracking with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 8844–8854.

[26] A. Kumar and R. Chellappa, "S2LD: Semi-supervised landmark detection in low resolution images and impact on face verification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 758–759.

[27] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee, "Unsupervised discovery of object landmarks as structural representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2694–2703.

[28] S. Yin, S. Wang, X. Chen, and E. Chen, "Exploiting self-supervised and semi-supervised learning for facial landmark tracking with unlabeled data," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2991–2998.

[29] Q. Quan, Q. Yao, J. Li, and S. K. Zhou, "Which images to label for few-shot medical landmark detection?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 20606–20616.

[30] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz, "Improving landmark localization with semi-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1546–1555.

[31] X. Tang, F. Guo, J. Shen, and T. Du, "Facial landmark detection by semi-supervised deep learning," *Neurocomputing*, vol. 297, pp. 22–32, Jul. 2018.

[32] A. Jabri, A. Owens, and A. Efros, "Space-time correspondence as a contrastive random walk," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19545–19560.

[33] M. Frueh, T. Kuestner, M. Nachbar, D. Thorwarth, A. Schilling, and S. Gatidis, "Self-supervised learning for automated anatomical tracking in medical image data with minimal human labeling effort," SSRN, Rochester, NY, USA, Tech. Rep. 21-01816, 2022.

[34] F. Bastani, S. He, and S. Madden, "Self-supervised multi-object tracking with cross-input consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 13695–13706.

[35] W. Yuan, Z. Lv, T. Schmidt, and S. Lovegrove, "STaR: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13144–13152.

[36] Y. Yu, Y. Xiong, W. Huang, and M. R. Scott, "Deformable Siamese attention networks for visual object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6728–6737.

[37] H. Mao, S. Zhu, S. Han, and W. J. Dally, "PatchNet–short-range template matching for efficient video processing," 2021, *arXiv:2103.07371*.

[38] L. Li, L. Han, M. Ding, H. Cao, and H. Hu, "A deep learning semantic template matching framework for remote sensing image registration," *ISPRS J. Photogramm. Remote Sens.*, vol. 181, pp. 205–217, Nov. 2021.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

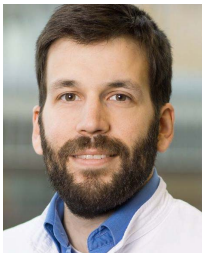
[41] A. B. Jung. (2020). *Imgaug*. Accessed: Feb. 1, 2022. [Online]. Available: https://github.com/aleju/imgaug

[42] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2756–2759.

[43] C. Würslin, H. Schmidt, P. Martirosian, C. Brendle, A. Boss, N. F. Schwenzer, and L. Stegger, "Respiratory motion correction in oncologic PET using T1-weighted MR imaging on a simultaneous whole-body PET/MR system," *J. Nucl. Med.*, vol. 54, no. 3, pp. 464–471, Mar. 2013.
[44] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, C. Chute, R. Ball, J. Seekins, S. S. Halabi, R. Jones, D. B. Larson, C, P. Langlotz, B. N. Patel, and M. P. Lungren, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 590–597.
[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
[46] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[48] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[49] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," 2022, *arXiv:2201.03545*.
[50] B. Bier, M. Unberath, J.-N. Zaech, J. Fotouhi, M. Armand, G. Osgood, N. Navab, and A. and Maier, "X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.*, 2018, pp. 55–63.
[51] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12546–12558.

MARCEL FRUEH received the M.Sc. degree in computer science with focus on deep learning from the University of Tübingen, in 2020. He is currently pursuing the Ph.D. degree with University Hospital Tübingen. His research interest includes machine learning in medical imaging.

ANDREAS SCHILLING received the Diploma degree in physics and the Ph.D. degree in computer science from the University of Tübingen, in 1988 and 1995, respectively. He is currently a Full Professor in visual computing with Eberhard-Karls-Universität Tübingen, Germany. Before 2003, he was a Professor in digital media at Stuttgart Media University. His research interests include machine learning, computer vision, computer graphics and image processing, especially medical image processing and model building.



SERGIOS GATIDIS received the M.D. degree from the University of Tübingen, in 2011, and the M.Sc. degree in mathematics from the University of Hagen, in 2014. In 2017, he was appointed as an Assistant Professor, and in 2020, as an Associate Professor in radiology with the Department of Radiology, University Hospital Tübingen. His research interest includes automated analysis of multiparametric medical image data.

THOMAS KUESTNER (Member, IEEE) received the Ph.D. degree from the University of Stuttgart, Germany, in 2017. From 2018 to 2020, he was with the School of Biomedical Engineering and Imaging Sciences, King's College London, U.K. Since 2020, he has been leading the Group of Medical Imaging and Data Analysis (MIDAS.lab), University Hospital of Tübingen, Germany. He is a Junior Fellow of the International Society for Magnetic Resonance in Medicine (ISMRM). His research interests include multi-parametric and multi-modality deep learning methods in acquisition, reconstruction and analysis for patient-centered workflows and with particular focus on MR-based motion imaging, correction, and reconstruction.

A.4 A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions

Authors: *Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, Daniel Rubin*

Published in: *Nature Scientific Data*

Date of Publication: *October 2022*

scientific data

OPEN

DATA DESCRIPTOR

A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions

Sergios Gatidis[1,2]✉, Tobias Hepp[1,2], Marcel Früh[2], Christian La Fougère[3,4,5], Konstantin Nikolaou[2,4], Christina Pfannenberg[2], Bernhard Schölkopf[1], Thomas Küstner[2], Clemens Cyran[6] & Daniel Rubin[7]

We describe a publicly available dataset of annotated Positron Emission Tomography/Computed Tomography (PET/CT) studies. 1014 whole body Fluorodeoxyglucose (FDG)-PET/CT datasets (501 studies of patients with malignant lymphoma, melanoma and non small cell lung cancer (NSCLC) and 513 studies without PET-positive malignant lesions (negative controls)) acquired between 2014 and 2018 were included. All examinations were acquired on a single, state-of-the-art PET/CT scanner. The imaging protocol consisted of a whole-body FDG-PET acquisition and a corresponding diagnostic CT scan. All FDG-avid lesions identified as malignant based on the clinical PET/CT report were manually segmented on PET images in a slice-per-slice (3D) manner. We provide the anonymized original DICOM files of all studies as well as the corresponding DICOM segmentation masks. In addition, we provide scripts for image processing and conversion to different file formats (NIfTI, mha, hdf5). Primary diagnosis, age and sex are provided as non-imaging information. We demonstrate how this dataset can be used for deep learning-based automated analysis of PET/CT data and provide the trained deep learning model.

Background & Summary

Integrated Positron Emission Tomography/Computed Tomography (PET/CT) has been established as a central diagnostic imaging modality for several mostly oncological indications over the past two decades. The unique strength of this hybrid imaging modality lies in its capability to provide both, highly resolved anatomical information by CT as well as functional and molecular information by PET. With growing numbers of performed examinations, the emergence of novel PET tracers and the increasing clinical demand for quantitative analysis and reporting of PET/CT studies is becoming increasingly complex and time consuming. To overcome this challenge, the implementation of machine learning algorithms for faster, more objective and quantitative medical image analysis has been proposed also for the analysis of PET/CT data. First methodological studies have demonstrated the feasibility of using deep learning frameworks for the detection and segmentation of metabolically active lesions in whole body Fluorodeoxyglucose (FDG)-PET/CT of patients with lung cancer, lymphoma and melanoma[1–4]. Despite these encouraging results, deep learning-based analysis of PET/CT data is still not established in routine clinical settings. Thus, automated medical image analysis, specifically of PET/CT images is an ongoing field of research that requires methodological advances to become clinically applicable. In contrast to the more widely used imaging modalities CT and MRI however, only few datasets of PET/CT studies are publicly accessible to clinical and machine learning scientists who work on automated PET/CT analysis. Even fewer datasets contain image-level ground truth labels to be used for machine learning research[5,6]. This is likely a major obstacle for innovation and clinical translation in this field. Examples of related areas, such as analysis

[1]Max-Planck-Institute for Intelligent Systems, Empirical Inference Department, Tuebingen, 72076, Germany. [2]University Hospital Tübingen, Department of Radiology, Tübingen, 72076, Germany. [3]University Hospital Tübingen, Department of Nuclear Medicine and Clinical Molecular Imaging, Tübingen, 72076, Germany. [4]Cluster of Excellence iFIT (EXC 2180) "Image-Guided and Functionally Instructed Tumor Therapies", Tübingen, 72076, Germany. [5]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ) Partner Site Tübingen, Tübingen, 72076, Germany. [6]Hospital of the Ludwig-Maximilians-University, Department of Radiology, Munich, 81377, Germany. [7]Stanford University, School of Medicine, Department of Biomedical Data Science, Stanford, 94305, USA. ✉e-mail: sergios.gatidis@tuebingen.mpg.de

diagnosis	patient sex	n/o studies	age [mean SD]
Melanoma	female	77	65.0 ± 12.8
	male	111	65.7 ± 13.7
Lymphoma	female	69	45.1 ± 19.7
	male	76	47.3 ± 17.9
Lung Cancer	female	65	64.2 ± 8.7
	male	103	67.0 ± 9.0
Negative	female	233	59.1 ± 14.7
	male	280	58.7 ± 15.1
All	female	444	58.5 ± 16.1
	male	570	60.1 ± 15.9

Table 1. Patient characteristics across the dataset subcategories.

of dermoscopy[7] or retinal images[8], show that the existence of publicly available labeled datasets can serve as a catalyst for method development and validation. The purpose of this project is thus to provide an annotated, publicly available dataset of PET/CT images that enables technical and clinical research in the area of machine learning-based analysis of PET/CT studies and to demonstrate a use case of deep learning-based automated segmentation of tumor lesions. To this end, we composed a dataset of 1,014 oncologic whole-body FDG-PET/CT examinations of patients with lymphoma, lung cancer and malignant melanoma, as well as negative controls together with voxel-wise manual labels of metabolically active tumor lesions. The provided data can be used by researchers of different backgrounds for the development and evaluation of machine learning methods for PET/CT analysis as well as for clinical research regarding the included tumor entities.

Methods

Data collection. Publication of anonymized data was approved by the institutional ethics committee of the Medical Faculty of the University of Tübingen as well as the institutional data security and privacy review board. Data from 1,014 whole-body FDG-PET/CT examinations of 900 patients acquired between 2014 and 2018 as part of a prospective registry study[9] were included in this dataset. Of these 1,014 examinations, 501 are positive samples, meaning they contain at least one FDG-avid tumor lesion and 513 are negative samples, meaning they do not contain FDG-avid tumor lesions. Negative samples stem from patients who were examined by PET/CT with a clinical indication (e.g. follow-up after tumor resection) but did not show any findings of metabolically active malignant disease. The selection criteria for positive samples were: age >18 years, histologically confirmed diagnosis of lung cancer, lymphoma or malignant melanoma, and presence of at least one FDG-avid tumor lesion according to the final clinical report. The selection criteria for negative samples were: age >18 years, no detectable FDG-avid tumor lesion according to the clinical radiology report. Of the 501 positive studies, 168 were acquired in patients with lung cancer, 145 in patients with lymphoma and 188 in patients with melanoma. Patient characteristics are summarized in Table 1.

PET/CT Acquisition. All PET/CT examinations were performed at the University Hospital Tübingen according to a standardized acquisition protocol on a single clinical scanner (Siemens Biograph mCT, Siemens Healthineers, Knoxville, USA) following international guidelines for oncologic PET/CT examinations (Boellaard *et al.* FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0)[10].

Diagnostic whole-body CT was acquired in expiration with arms elevated according to a standardized protocol using the following scan parameters: reference tube current exposure time product, 200 mAs with automated exposure control (CareDose); tube voltage, 120 kV. CT examinations were performed with weight-adapted 90–120 ml intravenous CT contrast agent in a portal-venous phase (Ultravist 370, Bayer Healthcare) or without contrast agent (in case of existing contraindications). CT data were reconstructed in transverse orientation with a slice thickness between 2.0 mm and 3.0 mm with an in-plane voxel edge length between 0.7 and 1.0 mm.

^{18}F-FDG was injected intravenously after at least 6 hours of fasting. PET acquisition was initiated 60 minutes after injection of a weight-adapted dose of approximately 300 MBq 18F-FDG (mean: 314.7 MBq, SD: 22.1 MBq, range: [150, 432] MBq). For the purpose of weight adaptation, target FDG injection acitivities were 250–300 MBq/300–350 MBq/350–400 MBq for patients with a body weight below 60 kg/between 60 and 100 kg/above 100 kg respectively. PET was acquired over four to eight bed positions (usually from the skull base to the mid-thigh level) and reconstructed using a 3D-ordered subset expectation maximization algorithm (two iterations, 21 subsets, Gaussian filter 2.0 mm, matrix size 400 × 400, slice thickness 3.0 mm, voxel size of $2.04 \times 2.04 \times 3\,mm^3$). PET acquisition time was 2 min per bed position. Example PET/CT images are displayed in Fig. 1a.

Data labeling and processing. All examinations were assessed by a radiologist and nuclear medicine specialist in a clinical setting. Based on the report of this clinical assessment, all FDG-avid tumor lesions (primary tumor if present and metastases if present) were segmented by an experienced radiologist (S.G., 10 years of experience in hybrid imaging) using dedicated software (NORA image analysis platform, University of Freiburg, Germany). In case of uncertainty regarding lesion definition, the specific PET/CT studies were reviewed in

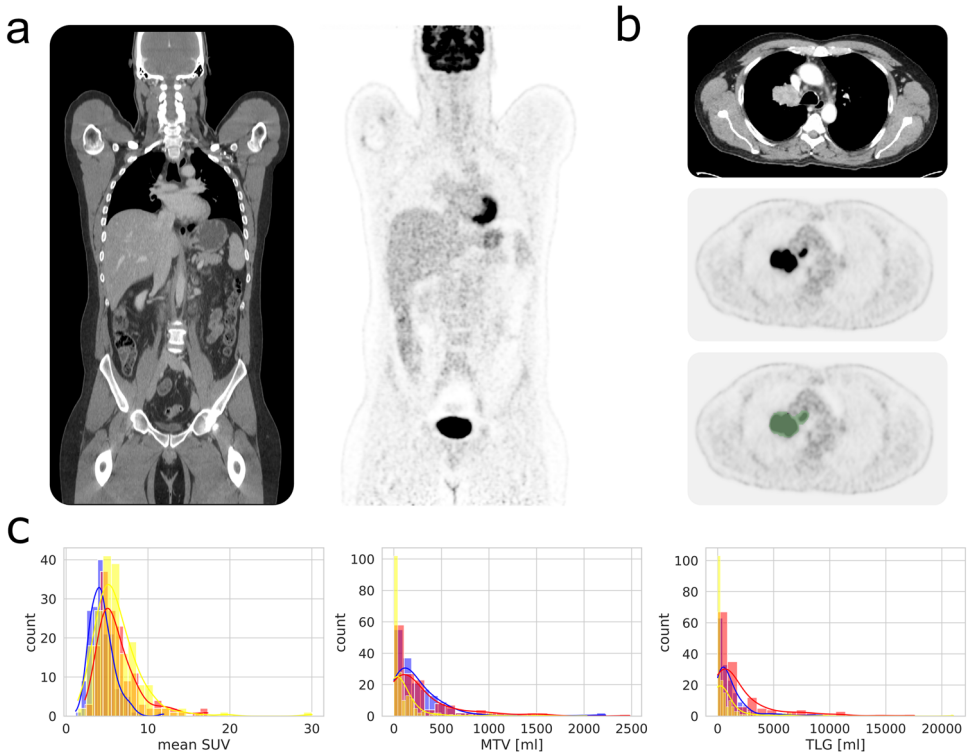


Fig. 1 Dataset properties. (**a**) Coronal views of CT (left) and FDG-PET (right) image volumes without pathologic findings. (**b**) Example of manual tumor segmentation (bottom image, green area) of a lung cancer mass; top: CT, middle: FDG-PET (**c**) Distribution of mean SUV, MTV and TLG of studies in patients with lung cancer (blue), lymphoma (red) and melanoma (yellow).

consensus with the radiologist and nuclear medicine physician who prepared the initial clinical report. To this end CT and corresponding PET volumes were displayed side by side or as an overlay and tumor lesions showing elevated FDG-uptake (visually above blood-pool levels) were segmented in a slice-per-slice manner resulting in 3D binary segmentation masks. An example slice of a segmented tumor lesion is shown in Fig. 1b. DICOM data of PET/CT volumes and corresponding segmentation masks were anonymized upon data upload to The Cancer Imaging Archive[11] using the CTP DICOM anonymizer tool.

Data properties. Of the 1014 studies (900 unique patients) included in this dataset, one study was included of 819 patients, two studies were included of 59 patients, 3 studies of 14 patients, 4 studies of 4 patients and 5 studies of 3 patients. The mean coverage (scan range) of the PET volumes in the longitudinal direction over all datasets was 1050.7 mm (SD: 306.7 mm, min: 600 mm, max: 1983 mm). The three included tumor entities showed similar distributions with respect to metabolic tumor volume (MTV), mean SUV of tumor lesions and total lesion glycolysis (TLG) (Fig. 1c). Overall, in non-negative studies, MTV, mean SUV and TLG amounted to (mean ± SD) 219.9 ± 342.7 ml, 5.6 ± 2.7 and 1330 ± 2296 ml, respectively. For lung cancer studies, these values were 263.6 ± 345.1 ml, 4.4 ± 1.5 and 1234 ± 1622 ml. For lymphoma studies these values 297.5 ± 393.1 ml, 6.3 ± 2.7 and 2042 ± 2941.4 ml. For melanoma studies these values were 121.2 ± 269.4 ml, 6.2 ± 3.1 and 867.3 ± 2113.8 ml.

Data Records

This dataset can be accessed on The Cancer Imaging Archive (TCIA) under the collection name "FDG-PET-CT-Lesions"[12].

DICOM data. Each individual PET/CT dataset consists of three image series stored in the DICOM format: a whole-body CT volume stored as a DICOM image series, a corresponding whole-body FDG-PET volume stored as a DICOM image series and a binary segmentation volume stored in the DICOM segmentation object format. The entire DICOM dataset consists of 1,014 image studies, 3,042 image series and a total of 916,957 single DICOM files (total size of approximately 419 GB). The directory structure of the DICOM dataset is depicted in Fig. 2a. Patients are identified uniquely by their anonymized patient ID.

Conversion to other image formats. To facilitate data usage, we provide Python scripts that allow conversion of DICOM data to other medical image formats (NIfTI and mha) as well as the hdf5 format. (https://github.com/lab-midas/TCIA processing). In addition to file conversion, these scripts generate processed image volumes: a CT volume resampled to the PET volume size and resolution as well as a PET volume with voxel values

a

```
|--- Patient 1
        |--- Study 1
                |--- PET DICOM folder
                |--- CT DICOM folder
                |--- segmentation DICOM folder
        |--- Study 2 (potential 2nd visit of same patient)
                |--- ...
        …
|--- Patient 2
        |--- ...
```

b

```
|--- Patient 1
        |--- Study 1
                |--- PET.nii.gz    (PET image)
                |--- CT.nii.gz     (CT image)
                |--- SEG.nii.gz    (segmentation mask)
                |--- SUV.nii.gz    (PET image translated to SUVs)
                |--- CTres.nii.gz (CT image resampled to PET image resolution and size)
        |--- Study 2 (potential 2nd visit of same patient)
                |--- ...
        ...
|--- Patient 2
        |--- ...
```

Fig. 2 Dataset structure. Patients are identified by a unique, anonymized ID and all studies of a single patient are stored under the respective patient path. (**a**) DICOM data: Each study folder contains three subfolders with DICOM files of the PET volume, the CT volume and the segmentation mask. (**b**) NIfTI data: Using the provided conversion script, DICOM data can be converted to NIfTI files. In addition to NIfTI files of the PET volume (PET.nii.gz), the CT volume (CT.nii.gz) and the segmentation mask (SEG.nii.gz), this script generates NIfTI volumes of the PET image in SUV units (SUV.nii.gz) and a CT volume resample to the PET resolution and shape (CTres.nii.gz).

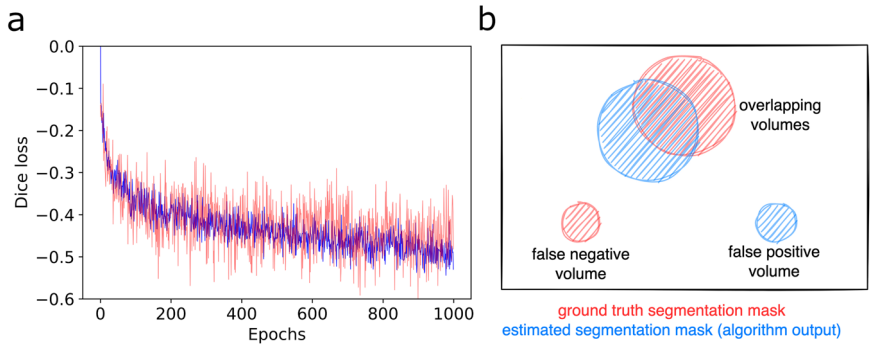


Fig. 3 Training and evaluation. (**a**) Representative loss curve on training data (blue) and validation data (red) from one fold of a 5-fold cross validation. (**b**) Schematic visualization of the proposed evaluation metrics false positive and false negative volumes (in addition to the Dice score).

converted to standardized uptake values (SUV). The data structure of the generated NIfTI files is represented in Fig. 2b. Data in the other formats (mha and hdf5) are generated accordingly.

Metadata. In addition to imaging data, a metadata file in Comma-separated Values (csv) format is provided containing information on study class (lung cancer, melanoma, lymphoma or negative), patient age (in years) and patient sex. In addition, the DICOM header data include information about patient body weight, injected activity and whether CT was contrast-enhanced (in case of non-enhanced CT, the CT series description includes the key word "nativ").

Technical Validation: Deep Learning-based Lesion Segmentation

In order to provide a use case scenario for the provided dataset we trained and evaluated a deep learning model for automated PET lesion segmentation. To this end, we used a standardized and publicly available deep learning framework for medical image segmentation (nnUNet[13]). This framework is based on a 3D U-Net architecture and provides an automated adaptive image processing pipeline. PET volumes converted to SUV units (SUV.nii.gz, Fig. 2b) and corresponding re-sampled CT volumes (CTres.nii.gz, Fig. 2b) were used as model inputs. Training with 5-fold cross validation was performed using the pre-configured model parameters with maximum number of epochs set to 1,000 and an initial learning rate of 1e-4 in a dedicated GPU (NVIDIA A5000). Typical loss and validation curves of a single validation step are depicted in Fig. 3a. For validation of algorithm performance, three metrics were used: Dice score, false positive volume and false negative volume (Fig. 3b). False positive volume was defined as the volume of false positive connected components in the predicted segmentation

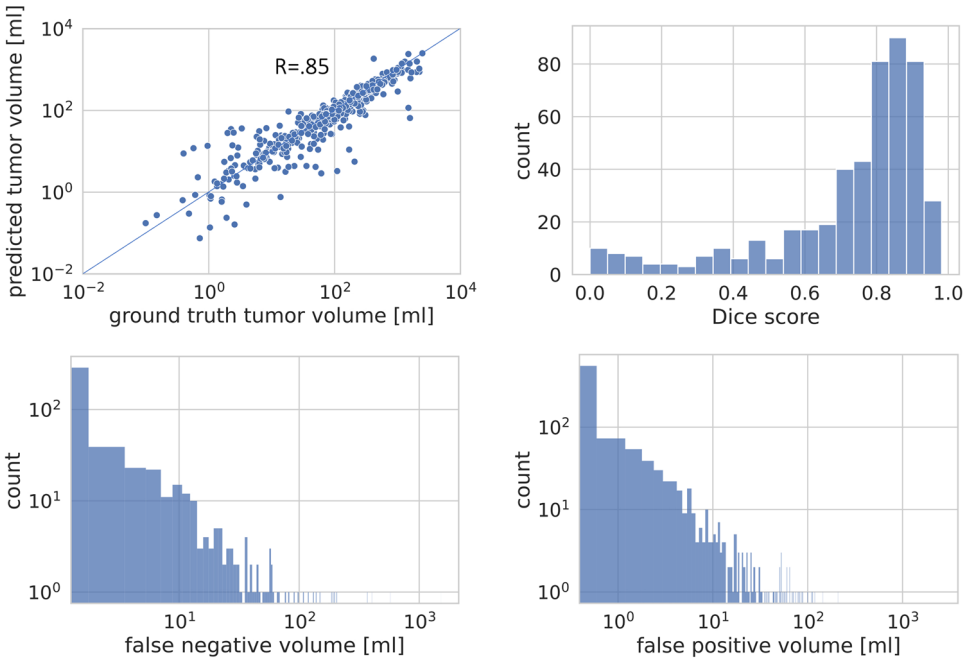


Fig. 4 Quantitative evaluation of automated lesion segmentation. Top left: Correlation of automatically predicted tumor volume with ground truth tumor volumes from manual segmentation in positive studies. Top right: Distribution of Dice coefficients for automated versus manual tumor segmentation in positive studies. Bottom left: Distribution of false negative volumes over all positive studies. Bottom right: Distribution of false positive volumes over all studies.

mask that do not overlap with tumor regions in the ground truth segmentation mask. This can be e.g. areas of physiological FDG-uptake (e.g. brain, heart, kidneys) that are erroneously classified as tumor. False negative volume was defined as the volume of connected components in the ground truth segmentation mask (=tumor lesions) that do not overlap with the estimated segmentation mask. These are tumor lesions that are entirely missed by the segmentation algorithm. In case of negative examples without present tumor lesions in the ground truth segmentation, only false positive volume was applicable as a metric.

We introduce these additional metrics (false positive and false negative volumes) due to the specific requirements of automated lesion segmentation. The specific challenge in automated segmentation of FDG-avid lesions in PET is to avoid false-positive segmentation of anatomical structures that have physiologically high FDG-uptake (e.g. brain, kidney, heart, etc.) while capturing all - even small - tumor lesions. The Dice score alone does not differentiate between false positive or negative segmentation within a correctly detected lesion (e.g. along its borders) and false positive or negative segmentations unconnected to detected lesions (i.e. false positive segmentation of healthy tissue or entirely missed lesions).

Overall, automated lesion segmentation using the described deep learning model showed good agreement with manual ground truth segmentation (Fig. 4). On datasets containing lesions, a high correlation of MTVs was observed between automated and manual segmentation ($r = 0.85$). The mean Dice score of automated compared to manual lesion segmentation was 0.73 (± 0.23) on positive datasets. Mean false positive/false negative volumes were 8.1 (± 81.4) ml/15.1 (± 80.3) ml respectively. Quantitative algorithm performance results on validation data (5-fold cross validation) are summarized in Fig. 4. Figure 5 provides qualitative examples for automated segmentation results.

This presented use case scenario demonstrates how this dataset can be used for the development and validation of algorithms for analysis of PET/CT data. We observed overall high automated segmentation performance that is comparable to previous studies focusing on specific disease entities[4,14]. In combination with methodological advances in the fields of machine learning and computer vision, this dataset can thus contribute to the development of increasingly accurate, robust and clinically useful algorithms for PET/CT analysis.Beyond automated lesion segmentation, this dataset bears the potential to be used for further tasks such as automated organ segmentation or automated lesion tracking. This would require further annotations which can be integrated with relatively low additional effort. For example, the recently published MOOSE framework[15] for automated organ segmentation on PET/CT data can be directly applied to this dataset providing e.g. information about lesions localization.

Usage Notes

For the purpose of visualization, image data can be loaded using freely available medical image data viewers such as the Medical Imaging Interaction Toolkit (https://www.mitk.org/) or 3D Slicer (https://www.slicer.org/). For the purpose of computational data analysis e.g. in Python, 3D image volumes can be read using freely available software such as pydicom (https://pydicom.github.io/) or nibabel (https://nipy.org/packages/nibabel/index.html).

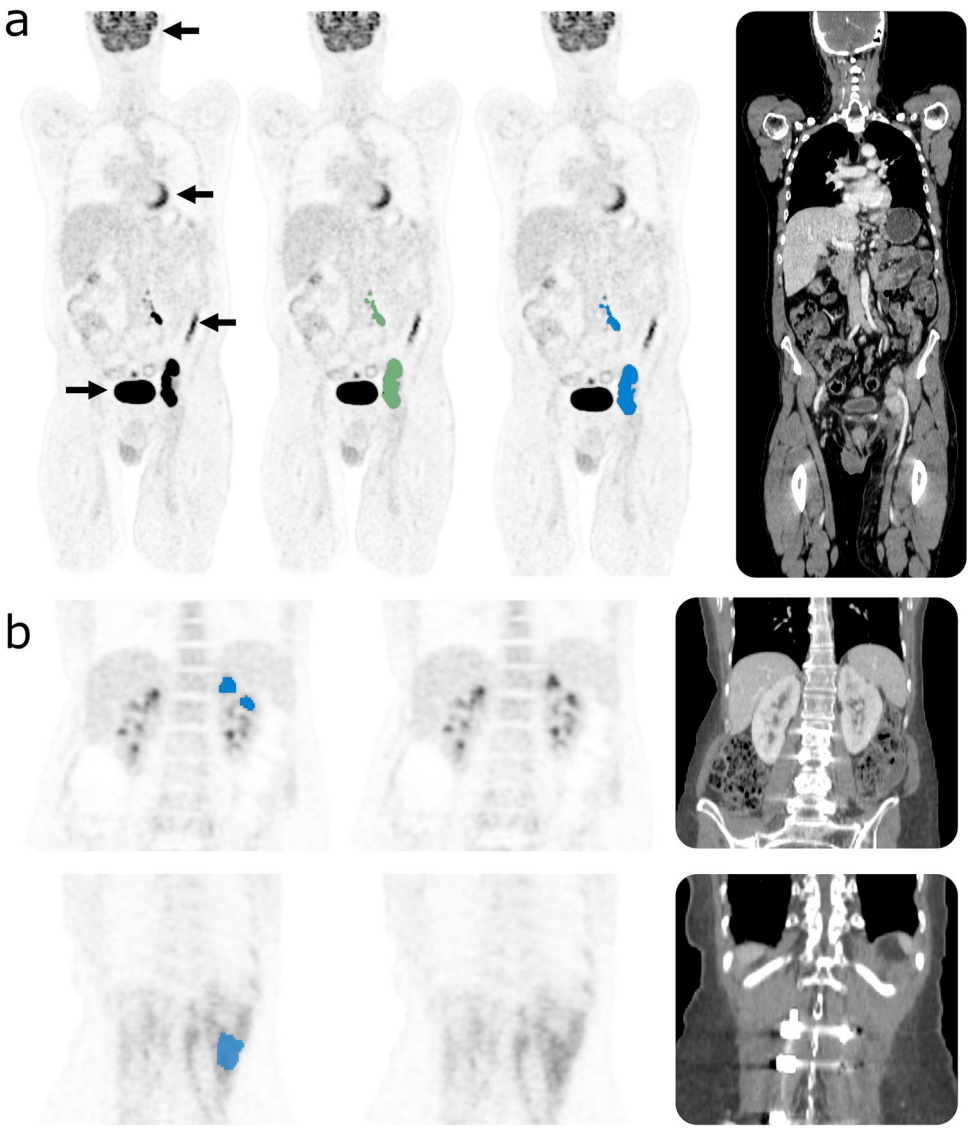


Fig. 5 Examples of automated lesion segmentation. (**a**) Example showing excellent agreement between manual (green) and automated (blue) tumor segmentation in a patient with lymphoma. Black arrows point to physiological FDG-uptake in the brain, heart, bowel and urinary bladder (from top to bottom) that was correctly not segmented. (**b**) Example of false positive segmentation of physiological structures with elevated FDG-uptake. Top: False positive partial segmentation of the left kidney. Bottom: False positive partial segmentation of back muscles.

The data presented in this manuscript is part of the MICCAI autoPET challenge 2022 (https://autopet.grand-challenge.org/).

Code availability

We provide the code of the data conversion and processing pipeline under https://github.com/lab-midas/TCIA processing. The trained PET/CT lesion segmentation model is publicly available under https://github.com/lab-midas/autoPET/tree/master/.

Received: 28 June 2022; Accepted: 23 September 2022;
Published online: 04 October 2022

References

1. Bi, L. *et al.* Recurrent feature fusion learning for multi-modality pet-ct tumor segmentation. *Comput Methods Programs Biomed* **203**, 106043 (2021).
2. Jemaa, S. *et al.* Tumor Segmentation and Feature Extraction from Whole-Body FDG-PET/CT Using Cascaded 2D and 3D Convolutional Neural Networks. *J Digit Imaging* **33**, 888–894 (2020).
3. Capobianco, N. *et al.* F-FDG Uptake Classification Enables Total Metabolic Tumor Volume Estimation in Diffuse Large B-Cell Lymphoma. *J Nucl Med* **62**, 30–36 (2021).

4. Blanc-Durand, P. *et al*. Fully automatic segmentation of diffuse large B cell lymphoma lesions on 3D FDG-PET/CT for total metabolic tumour volume prediction using a convolutional neural network. *Eur J Nucl Med Mol Imaging* **48**, 1362–1370 (2021).
5. Valliéres, M. *et al*. Data from head-neck-pet-ct. *The Cancer Imaging Archive* https://doi.org/10.7937/K9/TCIA.2017.8OJE5Q00 (2017).
6. Li, P. *et al*. A large-scale ct and pet/ct dataset for lung cancer diagnosis. *The Cancer Imaging Archive* https://doi.org/10.7937/TCIA.2020.NNC2-0461 (2020).
7. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data* **5**, 180161 (2018).
8. Pachade, S. *et al*. Retinal fundus multi-disease image dataset (rfmid): A dataset for multi-disease detection research. *Data* **6**, https://doi.org/10.3390/data6020014 (2021).
9. Pfannenberg, C. *et al*. Practice-based evidence for the clinical benefit of PET/CT-results of the first oncologic PET/CT registry in Germany. *Eur J Nucl Med Mol Imaging* **46**, 54–64 (2019).
10. Boellaard, R. *et al*. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur J Nucl Med Mol Imaging* **42**, 328–354 (2015).
11. Clark, K. *et al*. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging* **26**, 1045–1057, https://doi.org/10.1007/s10278-013-9622-7 (2013).
12. Gatidis, S. & Kuestner, T. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions (FDG-PET-CT-Lesions) [dataset]. *The Cancer Imaging Archive* https://doi.org/10.7937/gkr0-xv29 (2022).
13. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* **18**, 203–211 (2021).
14. Oreiller, V. *et al*. Head and neck tumor segmentation in PET/CT: The HECKTOR challenge. *Med Image Anal* **77**, 102336 (2022).
15. Shiyam Sundar, L. K. *et al*. Fully automated, semantic segmentation of whole-body 18F-FDG PET/CT images based on data-centric artificial intelligence. *J Nucl Med* (2022).

Acknowledgements

This project was partly supported by intramural grants of Stanford University and the University of Tübingen. This project was conducted under Germany's Excellence Strategy–EXC-Number 2064/1–390727645 and EXC 2180/1-390900677.

Author contributions

S.G., D.R., T.K., T.H.: conception and design of the work, acquisition, analysis and interpretation of data, creation of new software used in the work, drafting of the manuscript. K.N., C.L.F., M.F., C.P., B.S., C.C.: discussion and interpretation of results, substantial revision of the manuscript. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.G.

B. Under Review

B.1 The autoPET challenge: Towards fully automated lesion segmentation in oncologic PET/CT imaging

Authors: *Sergios Gatidis, Marcel Früh, Matthias Fabritius, Konstantin Nikolaou, Christian La Fougère, Jin Ye, Junjun He, Yige Peng, Lei Bi, Jun Ma, Bo Wang, Jia Zhang, Yukun Huang, Lars Heiliger, Zdravko Marinov, Jens Kleesiek, Ludovic Sibille, Lei Xiang, Simone Bendazzoli, Mehdi Astaraki, Michael Ingrisch, Clemens Cyran, Bernhard Schölkopf, Thomas Küstner*

Submitted to: *Nature Machine Intelligence*

Date of Submission: *January 2023*

The autoPET challenge: Towards fully automated lesion segmentation in oncologic PET/CT imaging

Sergios Gatidis (Sergios.Gatidis@med.uni-tuebingen.de)
University Hospital Tuebingen
Marcel Früh
University Hospital Tübingen
Matthias Fabritius
LMU University Hospital
Sijing Gu
LMU University Hospital
Konstantin Nikolaou
University Hospital Tübingen
Christian La Fougère
University Hospital Tübingen
Jin Ye
Shanghai AI Lab
Junjun He
Shanghai AI Lab
Yige Peng
University of Sydney https://orcid.org/0000-0001-5549-2688
Lei Bi
School of Computer Science
Jun Ma
University of Toronto
Bo Wang
Peter Munk Cardiac Centre
Jia Zhang
United Imaging Healthcare
Yukun Huang
United Imaging Healthcare
Lars Heiliger
University Hospital Essen
Zdravko Marinov
Karlsruhe Institute of Technology https://orcid.org/0000-0003-0373-3958
Rainer Stiefelhagen

Karlsruhe Institute of Technology
Jan Egger
University Hospital Essen
Jens Kleesiek
Institute for AI in Medicine, University Medicine Essen https://orcid.org/0000-0001-8686-0682
Ludovic Sibille
Subtle Medical
Lei Xiang
Subtle Medical
Simone Bendazolli
KTH Royal Institute of Technology
Mehdi Astaraki
KTH Royal Institute of Technology
Bernhard Schölkopf
Max Planck Institute for Intelligent Systmes
Michael Ingrisch
University Hospital, LMU Munich https://orcid.org/0000-0003-0268-9078
Clemens Cyran
University Hospital, LMU Munich
Thomas Küstner
University Hospital Tübingen

Article

Keywords: Machine Learning, Challenge, PET/CT, FDG, oncology, segmentation

Posted Date: June 14th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2572595/v1

Abstract

We describe the results of the autoPET challenge, a biomedical image analysis challenge aimed to motivate and focus research in the field of automated whole-body PET/CT image analysis. The challenge task was the automated segmentation of metabolically active tumor lesions on whole-body FDG-PET/CT. Challenge participants had access to one of the largest publicly available annotated PET/CT data sets for algorithm training. Over 350 teams from all continents registered for the autoPET challenge; the seven best-performing contributions were awarded at the MICCAI annual meeting 2022. Based on the challenge results we conclude that automated tumor lesion segmentation in PET/CT is feasible with high accuracy using state-of-the-art deep learning methods. We observed that algorithm performance in this task may primarily rely on the quality and quantity of input data and less on technical details of the underlying deep learning architecture. Future iterations of the autoPET challenge will focus on clinical translation.

Introduction

Recent advances in computational medical image analysis – in particular the introduction of deep learning methods – have led to substantial progress in numerous medical image analysis tasks including segmentation, regression and classification tasks. As part of this rapid development, medical image analysis challenges have played a crucial role by identifying relevant tasks, motivating and coordinating engagement, defining benchmarks and – perhaps most importantly – providing publicly available labeled data for algorithm development.

Several prominent examples of medical image analysis challenges, such as The Medical Segmentation Decathlon[1], the BRATS challenge[2] or the RSNA Pediatric Bone Age Challenge[3] illustrate the immense impact of such initiatives on their respective fields of research and application.

Most medical imaging challenges focus on the analysis of normal anatomy or the analysis of pathologies in defined anatomic regions, limiting the scope and complexity as well as the amount of required training data. In comparison, computational analysis of whole-body oncologic examinations, as acquired by PET/CT, is associated with higher complexity due to the multimodal nature of the underlying data, the large anatomical coverage, and the high morphological variability of oncologic pathologies. Furthermore, the generation of training labels on oncologic whole-body examinations requires a high level of clinical expertise and can only be performed by experienced medical imaging specialists. These factors contribute to delayed progress in the field of computational whole-body oncologic image processing, specifically regarding whole-body PET/CT imaging. Few studies have reported the development and application of automated PET/CT analysis – specifically automated tumor lesion segmentation – in the past. In these studies, a variety of methodological approaches have been proposed ranging from simple, threshold-based segmentation algorithms[4] to state-of-the-art deep learning methods[5] or combinations thereof[6]. While these studies clearly demonstrate the technical feasibility of automated PET/CT image analysis, the comparison and reproducibility of methods is limited due to the use of proprietary data and algorithms. A recent medical image analysis challenge (HECKTOR

challenge)[7] on automated PET/CT lesion segmentation in the head/neck region demonstrated in an anatomically restricted scenario, how combined efforts by the research community can advance this field in a specified direction.

Automation of the image analysis process in oncologic whole-body PET/CT data is of high interest. Quantitative analysis of PET/CT data requires segmentation of tumor lesions which is time-consuming and labor-intensive, thus associated with high effort and cost. This prevents wide-spread clinical adoption of quantitative image analysis beyond study settings. Automation of this process can thus potentially allow for integration of quantitative PET/CT analysis in routine clinical workflow supporting diagnostic and therapeutic decisions.

To advance the field of automated oncologic PET/CT analysis and to address the existing shortcomings in this area, we conducted the autoPET challenge. The primary challenge task was the automated segmentation of metabolically active tumor lesions in whole-body FDG-PET/CT. To this end, a multi-center database of 1164 oncologic whole-body PET/CT datasets (1014 public training samples and 150 private test samples) with manually segmented tumor lesions was composed. The training dataset is publicly available at The Cancer Imaging Archive (TCIA)[8] and has been previously described in detail[9].

In this work we present the content and results of the autoPET challenge that was conducted as part of MICCAI 2022 aiming to (1) motivate and focus research in the field of automated PET/CT image analysis (2) provide a platform for algorithm comparison and reproduction and (3) document the current state-of-the-art in this field. In addition, we provide analysis on the importance of composition and size of available training data for successful algorithm development.

Results

In the following, we describe the challenge preparation, organization and evaluation following the guidelines for biomedical image analysis challenge reporting (BIAS guidelines)[10]. The public challenge training data set was drawn from the University Hospital Tübingen (UKT). The private challenge test set was partly drawn from the same source (UKT) and partly from the University Hospital of the LMU Munich (LMU).

Challenge Participation

A total number of 359 teams registered for the autoPET challenge including teams from all continents (Fig. 1) with clear geographic concentrations on Asia (61%, mainly China: 41%), Europe (20%) and North America (16%, mainly USA: 13%). As far as disclosed, most participants were affiliated to academic institutions (75%), followed by a smaller group of company employees (12%).

37 teams submitted at least one algorithm to the preliminary challenge phase amounting to 253 total submissions in this phase. In the final challenge phase 18 teams contributed a total of 67 algorithms. The best-performing submission by each team was considered for the challenge leaderboard. The seven

best performing teams were identified as challenge winners – their contributions are described in greater detail as part of this work.

As expected, all final contributions were based on deep learning models. The majority of submitted algorithms relied on a 3D U-Net backbone in combination with a Dice loss. A minority of participants deployed transformer-based architectures or combinations of different architectures (2D and 3D) or used more uncommon loss functions[11] (e.g., focal loss, TopK Dice loss, Lovasz loss, Tversky loss), mostly in combination with a conventional Dice loss. An overview of the technical details is depicted in Fig. 2.

Best performing algorithms

In the following, we provide brief descriptions of the seven best-performing contributions in the order of the final leaderboard followed by individual performance reports. The code for all contributions is publicly available – details are available in the technical papers published by the participating teams and cited below. Overall, the use of a U-Net backbone was a common feature of the best contributions. The additional implementation of rule-based post-processing of algorithm outputs (e. g. threshold-based removal of small connected components from the output segmentation mask) distinguished the top four contributions from the rest of the field. All top-performing teams used both, the PET and CT image volumes as algorithm inputs.

Team Blackbean

The best-performing team chose a deliberately simple approach by using a vanilla U-Net backbone and focusing on ablation studies to identify the best combination of input shape (crop size) and step size during sliding window inference. In addition, a post-processing step was used to minimize the false-positive volume by removing small connected components from the initial algorithm output[12].

Team BDAV

Team BDAV used a combination of self-supervised pre-training (via contrastive learning) and a multi-stage U-Net architecture. The multi-stage U-Net architecture utilized a global segmentation module to conduct coarse tumor segmentation, which was then fed into a local refinement module to reduce the false positives. The multi-stage U-Net model was ensembled with a standard nnUNet model to generate the final prediction[13].

Team FightTumor

This contribution was based on a slightly modified nnUNet model using DiceTopK loss and enhanced data augmentation. In addition, post-processing of the model output was performed by removing small connected components (< 10 voxels) and segmentations in areas with low CT Hounsfield Units (< -1,000 HU)[14].

Team UIH-FL

Team UIH-FL trained a combined 2D and 3D nnUNet model. In addition, they performed post-processing of the model output by removing small connected components (< 4 voxels) and all connected components on the three bottom slices of the predicted pet mask[15].

Team Heiligerl

This contribution was based on an ensemble of an nnUNet-based model and a Swin UNETR. In addition, a classification model was trained to identify negative PET/CT scans without metabolically active lesions, based on maximum intensity projections (MIP), inspired by reading procedures of physicians [16].

Team SM

Team SM proposed a cascaded architecture consisting of a stacked ensemble of low-resolution U-Net models and a subsequent refiner U-Net for high-resolution predictions[17].

Team Flemings

Team Flemings proposed a cascaded architecture consisting of an initial inpainting model to detect and generate lesion-free images, followed by a U-Net-based segmentation, with the residual inpainting image as additional input[18].

nn-Unet (baseline model, out of competition)

To provide a baseline model, the widely used and standardized nn-UNet framework[19] was used with the default settings using PET and CT volumes as input. The trained baseline model is publicly available under https://github.com/lab-midas/autoPET.

Ensemble model (out of competition)

Based on the predictions of these above-described best performing algorithms, including the baseline model, an ensemble model output was computed by pixel-wise majority voting.

Overall, the performance metrics of the best performing teams were slightly different with respect to all three metrics (Fig. 3, A): Dice score (capturing consistency between foreground predictions and manual masks), false negative volume (capturing total volume of missed lesions), and false positive volume (capturing false-positive segmentations of physiologic tracer uptake). The mean Dice score ranged between 0.74 and 0.79, the mean false negative volume between 0.5 and 1.5 ml and the false positive volume between 2.1 and 9.5 ml. When assessing algorithm performance separately for the two data sources (UKT and LMU) of the multicentric training set, we observed that mean Dice scores were overall markedly higher for UKT test data (ranging between 0.8 and 0.88) compared to LMU test data (ranging between 0.6 and 0.7). Mean false negative volumes and false positive volumes were slightly higher for LMU data compared to UKT data (false negative volumes UKT: 0.3 to 1.7 ml, false negative volumes LMU: 0.9 to 2.3 ml; false positive volumes UKT: 1.5 to 5.4 ml, false positive volumes LMU: 3.2 to 20.3 ml).

The best performing team (Blackbean) ranked first regarding the mean dice score and the mean false negative volume and second regarding the mean false positive volume. Interestingly, the provided baseline nnUNet model showed a good overall performance ranking – out of competition – second with respect to the mean Dice score and seventh with respect to the mean false positive and false negative values (Fig. 3, A).

The ensemble prediction (out of competition) based on the top performing contributions and the baseline model showed a superior performance compared to all participating teams with the highest overall mean Dice score (0.81), the second lowest mean false negative volume (0.71 ml) and the lowest mean false positive volume (1.6 ml) (Fig. 3, A).

Typical qualitative examples of model performance and error cases are given in Fig. 4. In general, false positive segmentations mainly occurred in areas of atypical physiological tracer uptake (e. g. unusually large urinary bladder, brown adipose tissue) while tumor lesions adjacent to physiological tracer uptake were more often missed.

Impact of training data composition on algorithm performance

To better understand external factors influencing algorithm performance in general we performed additional ablation studies using the baseline model with different sizes and compositions of training data.

As could be expected, we observed an overall increase in segmentation performance with increasing numbers of training data reflected by increasing Dice scores and decreasing false positive and false negative volumes (Fig. 3, B). Interestingly, in contrast to this overall tendency, Dice scores on LMU test data showed no increase and even slightly decreased with higher numbers of training data, probably as a sign of overfitting to the UKT training data distribution.

In addition, we assessed the impact of the input data composition on algorithm performance. In addition to PET and CT volumes that were also used within the challenge, we added CT-based anatomical organ labels as a potential third input.

Regarding the composition of training data, we observed the highest segmentation performance in terms of Dice scores when using all three inputs (PET, CT and anatomical labels) on both, UKT and LMU data (Fig. 3, B). On UKT data, using all three inputs also gave lower false positive and false negative volumes. On LMU data, the results regarding composition of data and false positive/negative volumes were inconclusive; however, using only PET data resulted in markedly higher false positive volumes on LMU test data.

Figure 5 provides qualitative examples of test data sets and associated segmentation results for test data drawn from UKT and LMU. In agreement with the quantitative results, we qualitatively observed a

larger mismatch between manual and automated tumor lesion segmentation on LMU data (Fig. 5). In general, tumor volumes were locally overestimated on LMU test data explaining the lower Dice scores on LMU test data compared to UKT test data. Also agreeing with the quantitative results, with respect to false positive and false negative volumes, we did not observe any obvious qualitative differences between LMU and UKT test data.

Discussion

In this work we introduce the autoPET challenge on automated PET/CT lesion segmentation – organized as a MICCAI challenge in 2022 – and present its results as well as the results of further analyses to pave the way for a clinical adoption of automated PET/CT image analysis.

The main technical scope of this challenge was the automated segmentation of metabolically active tumor lesions in whole-body FDG PET/CT scans. The best performing contributions demonstrated that this basic and important task can be performed using state-of-the-art deep learning methodology with high overall accuracy.

Interestingly, algorithm performance did not depend in a relevant way on technical details of the deep learning architecture, and the provided nnUNet-based baseline model already performed among the top contributions. Furthermore, an ensemble model of the top-performing algorithms showed the best overall performance. These observations are in line with more general results from machine learning challenges indicating that ensembling of many different algorithms can be superior to optimization of a single algorithm [20].

In contrast, the size and composition of training data had a substantial effect on algorithm performance. First, we observed a slight but relevant increase in lesion segmentation accuracy when using PET and CT data as input compared to a PET-only input indicating that anatomical and morphological information is useful for lesion segmentation. Overall, segmentation accuracy also increased with increasing the size of the training set. However, this effect was not uniform between test data from UKT (same as training distribution) and test data from LMU (different from training distribution): On UKT test data, the increase in training data resulted in marked increase in Dice scores and decrease in false negative volumes with a plateau at around 800-1,000 training samples. On LMU test data however, while false negative volumes also increased with increasing training examples, no improvement and even a slight decline in Dice scores was observed. False positive volumes were relatively low in both test data sets independent of the size of the training set. These results indicate that the generalizability of algorithms trained in a single institution is limited and that a reduction in segmentation performance can be expected when applied to data from different sources, e.g., different scanners of hospitals. This drop in performance is not catastrophic but rather related to different localization of the tumor margins – false positive and false negative volumes were interestingly not worse on the external test data. These results motivate us to place our focus on the topics of robustness and generalization for the next iteration of the autoPET challenge.

The autoPET 2022 is the first, important step towards a long-term goal of fully automated, quantitative oncologic PET/CT image analysis. A number of tasks – beyond lesion segmentation in FDG-PET – need to be addressed. For the near future we identify mainly two: (i) generalization of PET lesion segmentation to different tracers and to different environments (tumor types, scanners, hospitals etc.) and (ii) segmentation of lesions that are only visible on CT due to low or missing tracer uptake. We aim to include these tasks as part of the next iterations of the autoPET challenge.

Materials and Methods

Challenge Mission and Task

The mission of the autoPET challenge is to motivate and focus research in the field of automated PET/CT image analysis, to provide a platform for algorithm comparison and reproduction and to document the current state-of-the-art in this field.

The autoPET challenge task – fully automated segmentation of metabolically active tumor lesions – is a crucial first step towards objective and quantitative oncologic diagnosis, staging and therapy response assessment in whole-body FDG-PET/CT. This task can be performed manually in principle but – depending on tumor spread – can be associated with enormous effort by experts. As a result, lesion segmentation is not performed routinely in clinical settings. Automation of this task will strongly support clinical implementation of PET/CT lesion segmentation and quantitative analysis.

We consider the autoPET challenge 2022 to be the first part of a series of challenges aiming to gradually address increasingly complex aspects of automated PET/CT analysis including detection and segmentation of tumor lesions on CT data, extension to other PET tracers, lesion phenotyping and the analysis of longitudinal imaging studies.

Challenge Organization and Infrastructure

The autoPET Challenge was conducted in 2022 as a MICCAI (Medical Imaging Computing and Computed Assisted Intervention Society) - registered challenge and in cooperation with the European Society of Hybrid, Molecular end Translational Imaging (ESHI-MT). The organizing team consisted of radiologists and medical data scientists from the University Hospital Tübingen (UKT) in Tübingen, Germany and the Ludwig-Maximilian-University (LMU) Hospital in Munich, Germany.

The challenge proposal was submitted to MICCAI in December 2021 and – after undergoing a peer-review process – was approved in February 2022. The final challenge proposal is publicly available [21]. The challenge and its results were presented in a Satellite Event at the 25th International Conference on Medical Imaging Computing and Computed Assisted Intervention in September 2022.

The challenge was opened on April 1st, 2022, with the release of all related information and the public training set. During a first submission phase, starting May 3rd, 2022, participants were able to submit

their algorithms to perform technical sanity checks on a small, preliminary private test data set consisting of 5 representative test samples. The second and final submission phase was launched August 1st, 2022, with the activation of the final, private test data set consisting of 150 test samples (Fig. 1).

The technical realization of autoPET 2022 was conducted on the dedicated Grand Challenge platform (grand-challenge.com, Diagnostic Image Analysis Group, Radboud University Medical Center, The Netherlands) as a type-II challenge (i.e., submission of algorithms by participants to run on a private test set) under the URL https://autopet.grand-challenge.org/. Due to the private nature of the test data set, algorithms were submitted to and deployed on the Grand Challenge computing platform via Docker containers. A time limit of 20 minutes per test sample was set for algorithm execution on the available computation resources (1 TPU with 16 GB GPU memory (NVIDIA, Santa Clara, CA, USA), 8 CPUs and 30 GB of CPU memory).

Technical support was provided to challenge participants in the form of a baseline algorithm example and the associated code base as well as detailed description of the submission process on a public online code repository (https://github.com/lab-midas/autoPET) and the challenge website (https://autopet.grand-challenge.org/).

Participation policies

The use of data for algorithm development was restricted to the provided public training data set. No additional data or machine learning models pre-trained on external data were permitted. Members of the organizer's institutes were allowed to participate in the autoPET challenge but were not eligible for awards. The seven best performing contributions according to the challenge leader board (ranking criteria are described below) were announced publicly awarded with monetary prizes (in €: 6,000 for first place, 3,000 for second place, 2,000 for third place and 1,000 for places four to seven). To be eligible for awards, participating teams were required to publish their code and a technical manuscript describing their methodology and results under an open-source license. Two members of each team were invited to contribute as authors to this manuscript.

Datasets

The public training data consisted of 1,014 anonymized oncologic whole-body FDG-PET/CT data sets together with manually generated segmentation labels of metabolically active tumor lesions drawn from the University Hospital in Tübingen (UKT) (Fig. 6, A). All training data were obtained from a single institution and clinical PET/CT scanner (Biograph mCT, Siemens Healthcare) using a standardized imaging protocol (CT protocol: reference tube current, 200 mAs with automated tube current modulation; tube voltage, 120 kV; i.v. contrast agent injection, 90–120 ml Ultravist 370 (Bayer AG) in the portal-venous phase; slice thickness, 2–3 mm; PET protocol: tracer uptake time, 60 min; injected tracer activity, 300–350 MBq; iterative reconstruction (two iterations, 21 subsets) with Gaussian smoothing (2 mm full width at half-maximum); reconstructed voxel size, 2 x 2 x 3 mm^3). The training data set is publicly available at The Cancer Imaging Archive (TCIA)[8] and has been previously described in detail[22].

The private test data set consisted of 150 anonymized oncologic whole-body FDG-PET/CT data sets together with manually generated segmentation labels of metabolically active tumor lesions (Fig. 6, A). 100 of the 150 training samples were obtained from the same institution (UKT) and acquired with the same imaging protocol as the training data set. The remaining 50 of the 150 training samples were obtained from a different institution (LMU) with variable PET/CT imaging protocols using clinical PET/CT scanners by two vendors (64 TruePoint or Biograph mCT Flow (Siemens Healthineers) or a GE Discovery 690 (GE Healthcare)). These 50 test data sets were acquired using similar protocols (CT protocol: tube current, 100–190 mAs; tube voltage, 120 kV; i.v. contrast agent injection, weight-adapted dose of Ultravist 300 (Bayer AG) or Imeron 350 (Bracco Imaging) in the portal venous phase; slice thickness, 3 mm; PET protocol: tracer uptake time, 60 min; tracer activity, 300–350 MBq; iterative reconstruction (three iterations, 21 subsets) with Gaussian smoothing (2–4 mm full width at half-maximum) reconstructed voxel size, 2.7 x 2.7 x 2–4 x 4 x 5 mm^3).

Only members of the organizing committee had access to the private test data set and its labels.

Both, the training and test data sets included examinations of patients diagnosed with lymphoma, lung cancer or melanoma as well as negative studies (without detectable metabolically active tumor lesion). Training and test populations had a comparable age distribution. 444 of 1,014 training scans (43.7%) and 58 of 150 test scans (38.7%) were of female patients. Regarding the distribution of tumor load, measured by the Metabolic Tumor Volume (MTV) and metabolic tumor activity, measured by the mean Standardized Uptake Value (meanSUV) training and test data showed a good overall agreement (Supplemental Figure S1).

A representative example of a complete data set is provided in the supplemental material (Supplemental Figure S2).

Evaluation and further analyses

Quantitative algorithm performance regarding the challenge task was assessed using three metrics representing different aspects as previously described in [22] (Fig. 6, B). The foreground Dice score was used as an overall metric of agreement between ground truth segmentations and algorithm predictions. In addition, the metrics “false positive volume” and “false negative volume” were used to quantify the erroneous segmentation of healthy tissue and the miss of entire tumor lesions respectively. The false positive volume is defined as the sum of all positive connected components in the prediction that do not overlap with true tumor lesions in the ground truth (i.e., false positive segmentations that are not related to actual tumor lesions). The false negative volume is defined as true positive connected components that do not overlap with positive areas of the prediction (i.e., tumor lesions that were entirely missed).

Challenge submissions were ranked separately for each to these three metrics using their respective means. The final overall rank was derived using the mean of these single rankings (with Dice score being weighted twice) using the Dice score as a tie break. The code used for computation of the challenge metrics is publicly available under https://github.com/lab-midas/autoPET.

To analyze the generalization properties of submitted algorithms, all metrics were also computed separately for the two data sources (UKT and LMU).

To assess the impact of available training data the following additional ablation studies were performed using baseline models based on the standard configuration of the nn-UNet framework[19]: The impact of the number of available training data was assessed by training different versions of the baseline model with 50, 100, 200, 400, 800 or 1,014 randomly drawn training datasets. To assess the impact of additional anatomical information, these baseline models were trained either using only the PET image volumes as model input, or the PET volumes and the corresponding CT volumes or the PET and CT volumes together with CT-based anatomical organ segmentation masks. These organ segmentation masks were derived on using a publicly available pre-trained CT organ segmentation model [23]. This model provides segmentation of 36 anatomical structures including all major organs as well as adipose and lean tissue compartments (Supplemental Fig. 2). The underlying hypothesis for these experiments was that the addition of implicit or explicit anatomical information as input might support the learning process and potentially improve segmentation performance or reduce the number of required training data. It should be noted that the use of anatomical labels was not permitted within the challenge and was only used as part of these additional analyses.

Declarations

Acknowledgements

This project was partly supported by the Leuze Foundation, Owen/Teck, Germany. This project was conducted under Germany's Excellence Strategy–EXC-Number 2064/1–390727645 and EXC 2180/1-390900677. This study is part of the doctoral thesis of Alexandra Kubičkova.

Competing Interests

The Authors declare no Competing Financial or Non-Financial Interests.

Author Contributions

Sergios Gatidis, Marcel Früh, Sijing Gu, Matthias Fabritius, Michael Ingrisch, Clemens Cyran, Thomas Küstner:

Organization of the challenge, Preparation of training and test data, Contribution of software, Data analysis, Drafting of the manuscript

Konstantin Nikolaou, Christian la Fougère, Bernhard Schölkopf:

Scientific and clinical consultation during challenge preparation and data analysis. Critical revision of the manuscript

Jin Ye, Junjun He, Yige Peng, Lei Bi, Jun Ma, Bo Wang, Jia Zhang, Yukun Huang, Lars Heiliger, Zdravko Marinov, Jens Kleesiek, Rainer Stiefelhagen, Jan Egger, Ludovic Sibille, Lei Xiang, Simone Bendazzoli, Mehdi Astaraki:

Members of the best performing participating teams. Contribution of software. Participation in drafting and critical revision of the manuscript.

References

1. Antonelli, M. *et al.* The Medical Segmentation Decathlon. *Nat Commun* **13**, 4128 (2022). https://doi.org:10.1038/s41467-022-30695-9
2. Menze, B. H. *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging* **34**, 1993-2024 (2015). https://doi.org:10.1109/tmi.2014.2377694
3. Halabi, S. S. *et al.* The RSNA Pediatric Bone Age Machine Learning Challenge. *Radiology* **290**, 498-503 (2019). https://doi.org:10.1148/radiol.2018180736
4. Weisman, A. J. *et al.* Comparison of 11 automated PET segmentation methods in lymphoma. *Phys Med Biol* **65**, 235019 (2020). https://doi.org:10.1088/1361-6560/abb6bd
5. Groendahl, A. R. *et al.* A comparison of fully automatic segmentation of tumors and involved nodes in PET/CT of head and neck cancers. *Phys Med Biol* (2021). https://doi.org:10.1088/1361-6560/abe553
6. Capobianco, N. *et al.* Deep-Learning (18)F-FDG Uptake Classification Enables Total Metabolic Tumor Volume Estimation in Diffuse Large B-Cell Lymphoma. *J Nucl Med* **62**, 30-36 (2021). https://doi.org:10.2967/jnumed.120.242412
7. Oreiller, V. *et al.* Head and neck tumor segmentation in PET/CT: The HECKTOR challenge. *Medical Image Analysis* **77**, 102336 (2022). https://doi.org:https://doi.org/10.1016/j.media.2021.102336
8. Gatidis, S. & Kuestner, T. (The Cancer Imaging Archive (TCIA), 2022).
9. Gatidis, S. *et al.* A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions. *Sci Data* **9**, 601 (2022). https://doi.org:10.1038/s41597-022-01718-3
10. Maier-Hein, L. *et al.* BIAS: Transparent reporting of biomedical image analysis challenges. *Medical Image Analysis* **66**, 101796 (2020). https://doi.org:https://doi.org/10.1016/j.media.2020.101796
11. Ma, J. *et al.* Loss odyssey in medical image segmentation. *Medical Image Analysis* **71**, 102035 (2021). https://doi.org:https://doi.org/10.1016/j.media.2021.102035
12. Ye, J. *et al.* Exploring Vanilla U-Net for Lesion Segmentation from Whole-body FDG-PET/CT Scans. arXiv:2210.07490 (2022). <https://ui.adsabs.harvard.edu/abs/2022arXiv221007490Y>.
13. Peng, Y., Kim, J., Feng, D. & Bi, L. Automatic Tumor Segmentation via False Positive Reduction Network for Whole-Body Multi-Modal PET/CT Images. arXiv:2209.07705 (2022).

<https://ui.adsabs.harvard.edu/abs/2022arXiv220907705P>.

14. Ma, J. & Wang, B. *nnU-Net for Automated Lesion Segmentation in Whole-body FDG-PET/CT*, <https://github.com/JunMa11/PETCTSeg/blob/main/technical_report.pdf> (2022).
15. Zhang, J., Huang, Y., Zhang, Z. & Shi, Y. Whole-Body Lesion Segmentation in 18F-FDG PET/CT. arXiv:2209.07851 (2022). <https://ui.adsabs.harvard.edu/abs/2022arXiv220907851Z>.
16. Heiliger, L. *et al.* AutoPET Challenge: Combining nn-Unet with Swin UNETR Augmented by Maximum Intensity Projection Classifier. arXiv:2209.01112 (2022). <https://ui.adsabs.harvard.edu/abs/2022arXiv220901112H>.
17. Sibille, L., Zhan, X. & Xiang, L. Whole-body tumor segmentation of 18F -FDG PET/CT using a cascaded and ensembled convolutional neural networks. arXiv:2210.08068 (2022). <https://ui.adsabs.harvard.edu/abs/2022arXiv221008068S>.
18. Bendazzoli, S. & Astaraki, M. PriorNet: lesion segmentation in PET-CT including prior tumor appearance information. arXiv:2210.02203 (2022). <https://ui.adsabs.harvard.edu/abs/2022arXiv221002203B>.
19. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* **18**, 203-211 (2021). https://doi.org:10.1038/s41592-020-01008-z
20. Erickson, N. *et al. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data*. (2020).
21. Gatidis, S., Küstner, T., Ingrisch, M., Fabritius, M. & Cyran, C. *Automated Lesion Segmentation in Whole-Body FDG- PET/CT*. (Zenodo, 2022).
22. Gatidis, S. *et al.* A whole-body FDG-PET/CT Dataset with manually annotated Tumor Lesions. *Scientific Data* **9**, 601 (2022). https://doi.org:10.1038/s41597-022-01718-3
23. Sundar, L. K. S. *et al.* Fully Automated, Semantic Segmentation of Whole-Body (18)F-FDG PET/CT Images Based on Data-Centric Artificial Intelligence. *J Nucl Med* **63**, 1941-1948 (2022). https://doi.org:10.2967/jnumed.122.264063

Figures

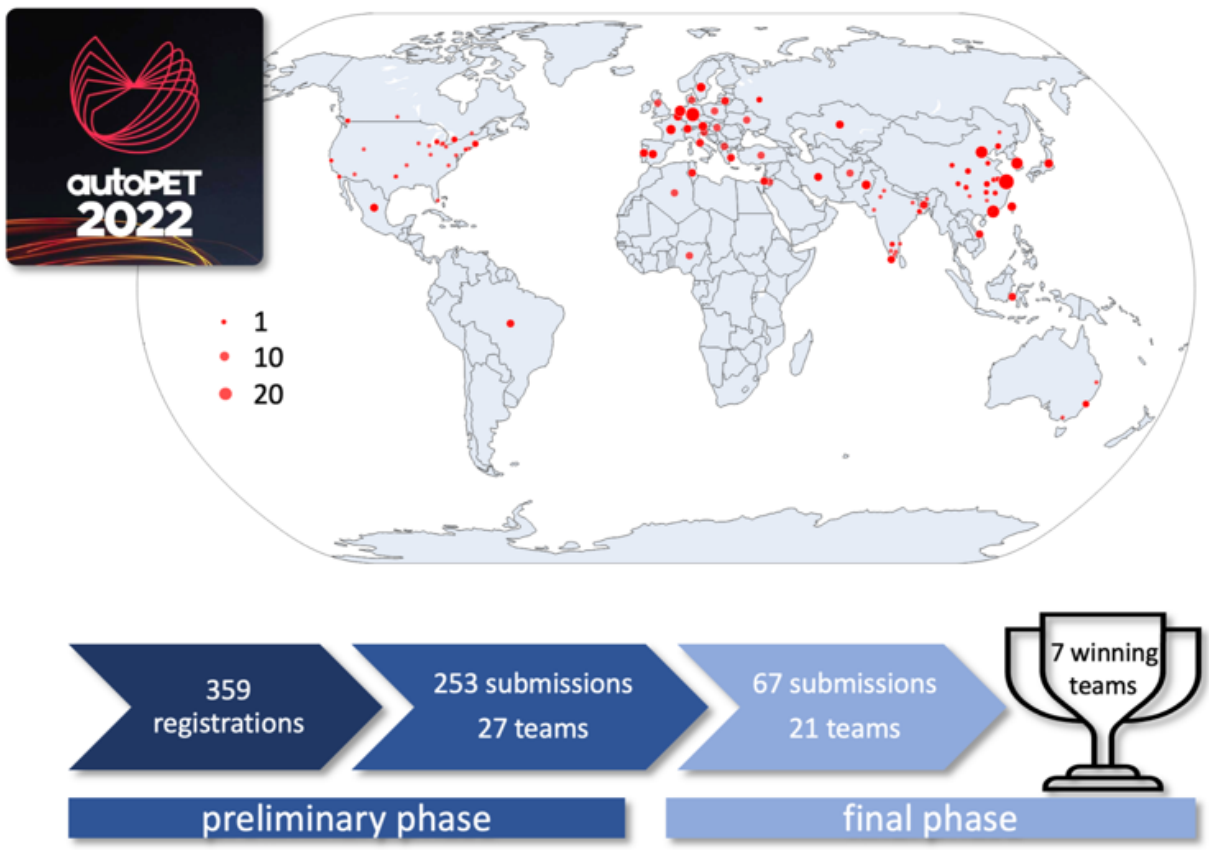


Figure 1

Challenge organization and participation

The autoPET challenge, conducted in 2022, consisted of two phases: a preliminary phase - allowing participants to perform technical validation of their algorithms on a small private test set – and a final phase where participants contributed their algorithms for final evaluation on the entire private test set. The seven best performing contributions were awarded and are described in this paper. In total, 359 teams from all continents participated in this challenge. Top: Geographic distribution of registered teams, bottom: challenge phases and participation.

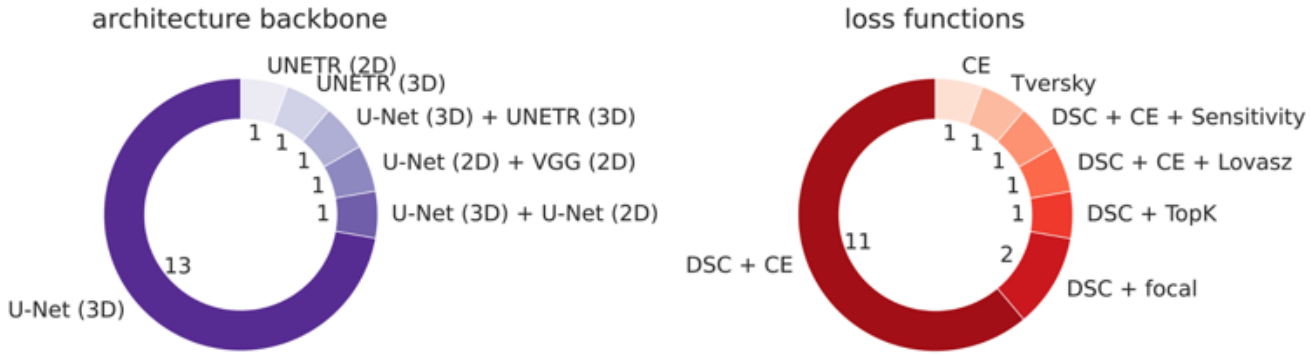


Figure 2

Overview of technical details of the final phase submissions

All participants used deep learning techniques to solve the challenge task. The majority of participating teams used a 3D U-Net architecture (left) together with a combined Dice and cross-entropy (CE) loss (right).

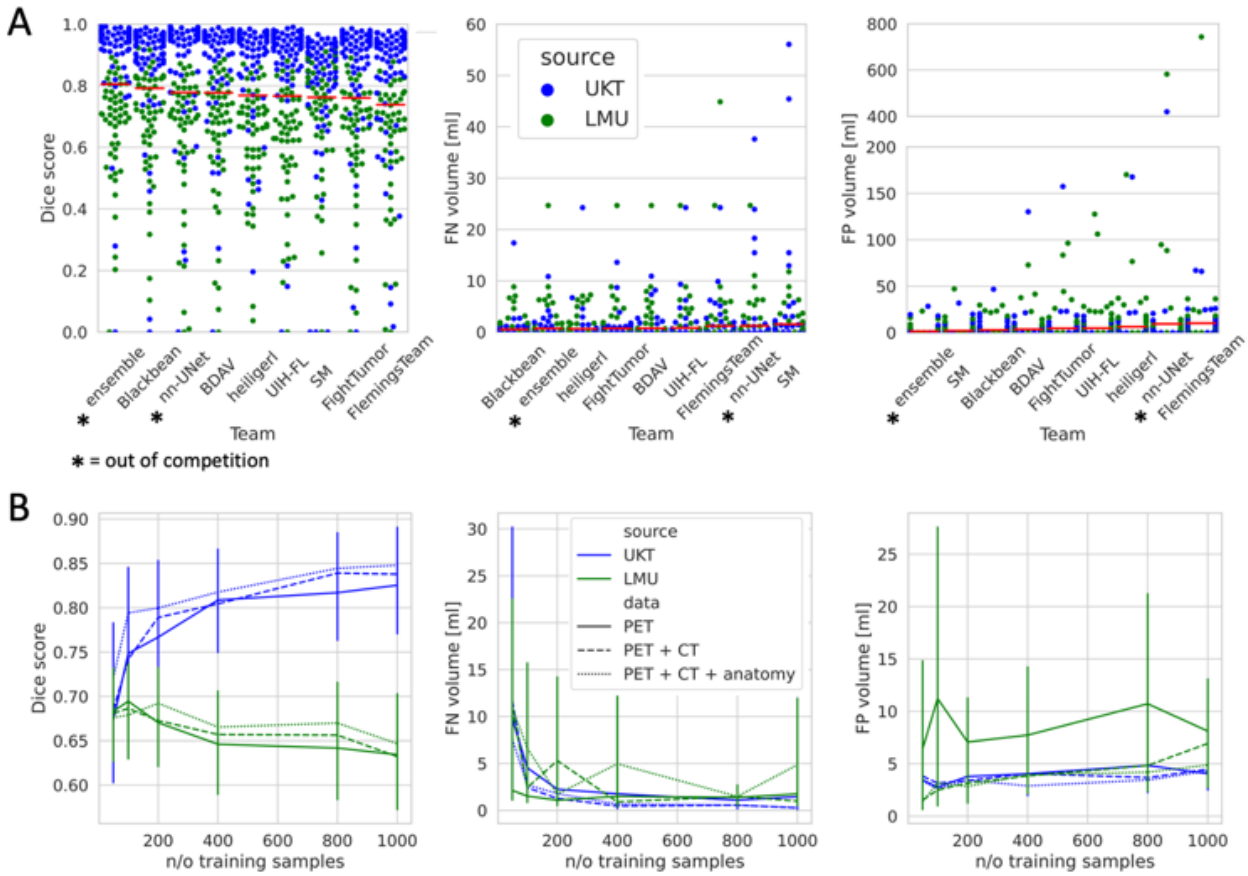


Figure 3

Overview of algorithm performance

A) Challenge results and algorithm performance in terms of Dice score, false negative (FN) volume and false positive (FP) volume. Results are ordered from best (left) to worst (right). Overall, algorithms performed better on the test data drawn from the training distribution (UKT, blue dots) compared to out-of-distribution data (LMU, green dots)

B) Impact of data source (UKT vs. LMU) and number of training samples on the performance of the baseline nn-UNet model in terms of Dice score, false negative (FN) volume and false positive (FP) volume. Overall, algorithm performance was higher on UKT test data compared to LMU test data and improved with increasing number of training samples. Notably, algorithm performance in terms of Dice scores did not improve with increasing numbers of training samples on out-of-distribution data (LMU).

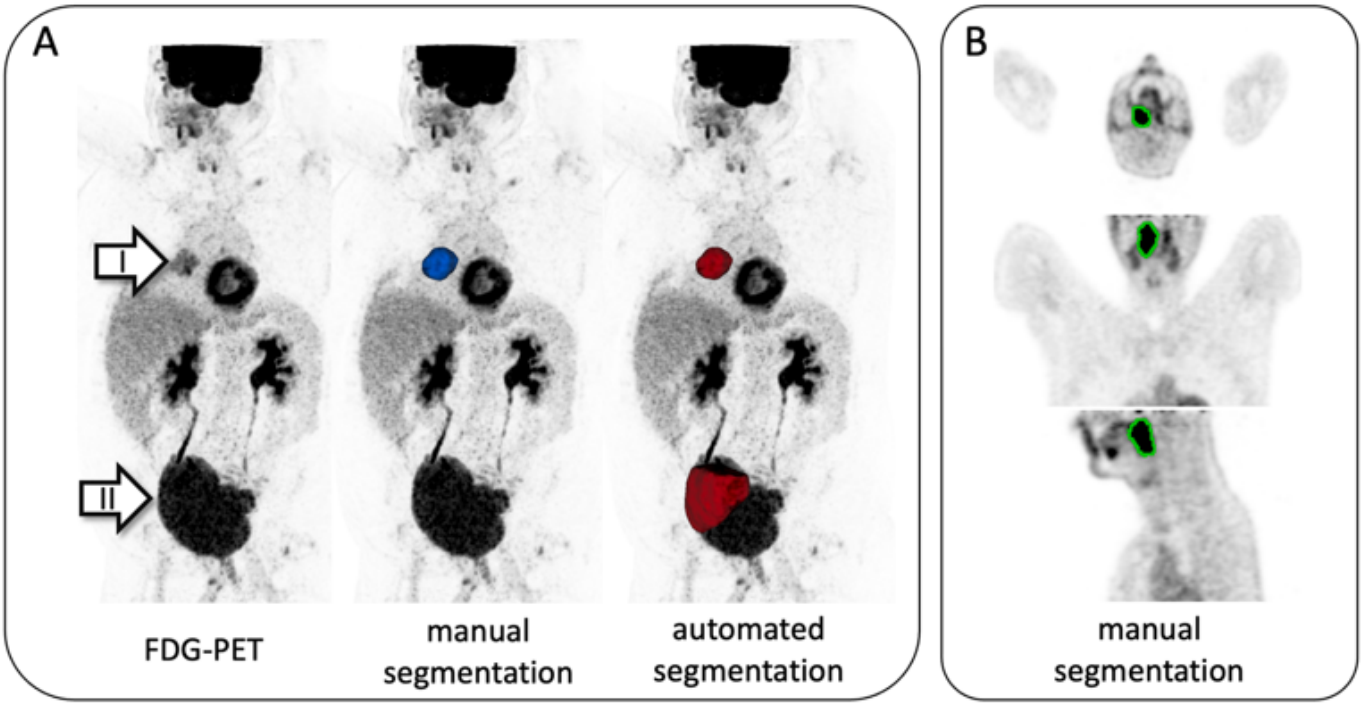


Figure 4

Qualitative examples of false positive and false negative volumes

A) Example of a large false positive volume, drawn from the UKT test data. PET scan of a patient with lung cancer (arrow I). Manual segmentation (in blue) shows the tumor lesion. Automated segmentation (red) using the baseline nn-UNet model accurately captures the tumor volume but in addition includes a large portion of the unusually large urinary bladder (arrow II). This false positive segmentation was observed in the majority of submitted algorithms.

B) Example of a false negative volume, drawn from the LMU test data. PET scan of a patient with Non-Hodgkin-Lymphoma of the right pharyngeal tonsil (manual segmentation outlined in green). This lesion was missed by 5 of the 7 best-performing contributions, probably due to its uncommon location.

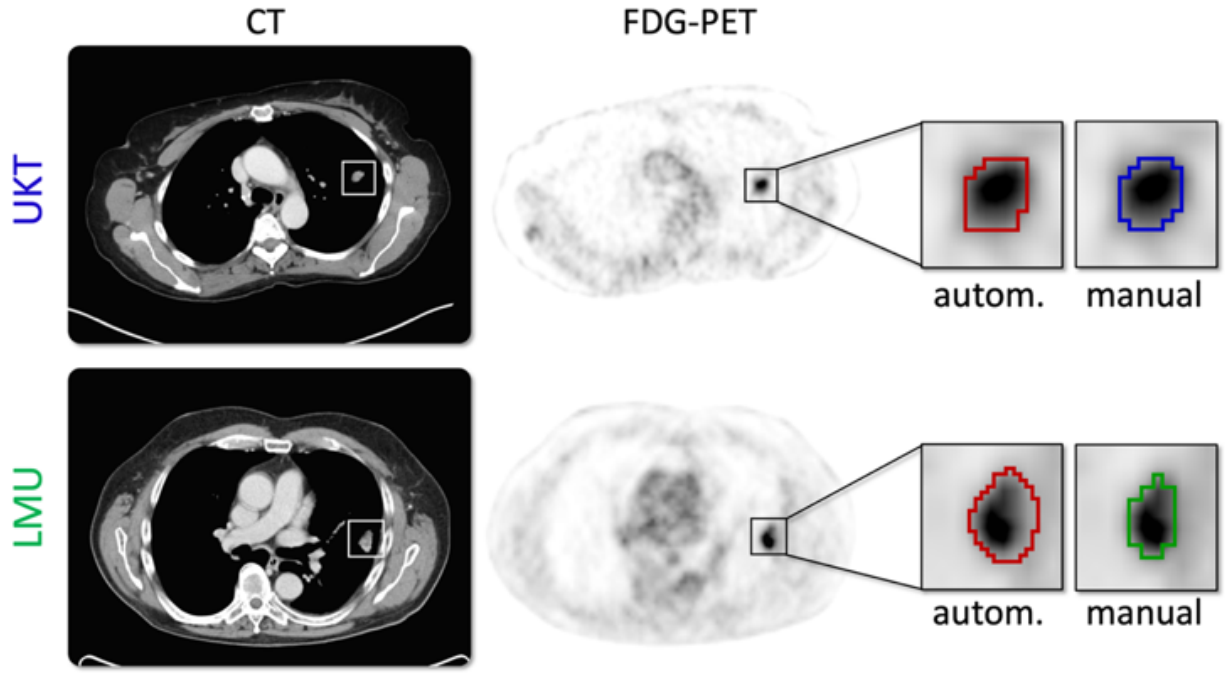


Figure 5

Typical examples of PET/CT data and segmentations from UKT (top) and LMU (bottom)

While CT image appearance is comparable between LMU and UKT data, PET scans have a lower spatial resolution on LMU data. As a result, all algorithms tended to overestimate local tumor volumes on LMU data (right column, outlined in green) compared to the manual ground truth segmentation (outlined in red). On UKT test data, automated tumor segmentations (right column, outlined in blue) showed a better agreement with manual (outlined in red) and segmentations. This is also reflected in the overall lower Dice scores of automated lesion segmentation on LMU test data compared to UKT test data.

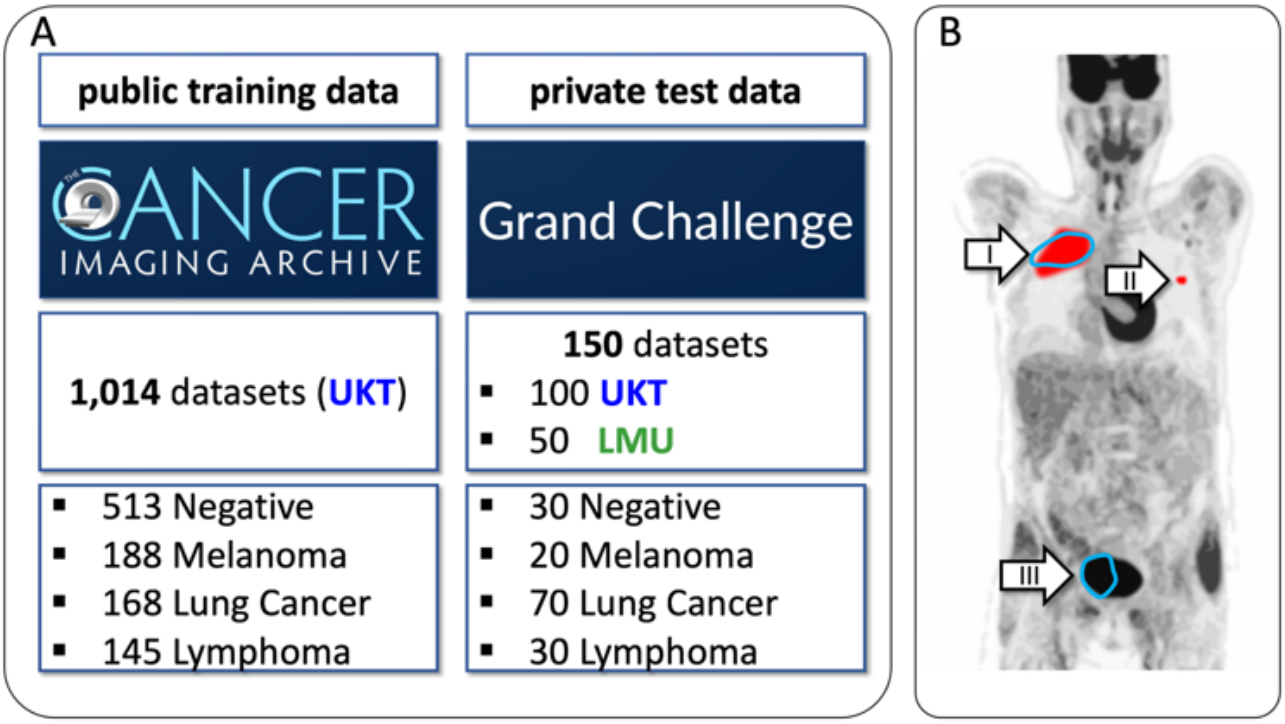


Figure 6

Challenge data and Evaluation Metrics

A) Overview of the composition of training data and test data. Training data were public and drawn from a single institution and scanner (UKT). Test data were private and drawn from two institutions: UKT (same as training data distribution) and LMU (out of training distribution).

B) Schematic illustration of the challenge metrics. I: a primary tumor lesion (red), II: a metastasis (red). Blue: Algorithm output. The Dice score provides a measure of the overall overlap between tumor lesions and algorithm segmentation. In this illustration, the lesion II is a false negative volume, as it is entirely not captured by the algorithm. Segmentation III (partial segmentation of the urinary bladder) is a false positive volume as it is not related to a tumor lesion.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalFigures.docx

C. Not Submitted

C.1 Large Scale, entire-volume, 3D Contrastive Learning on whole-body MRI data

Authors: *Marcel Früh, Andreas Schilling, Thomas Küstner, Sergios Gatidis*

Large Scale, entire-volume, 3D Contrastive Learning on whole-body MRI data

Marcel Frueh, Andreas Schilling, Thomas Küstner, Sergios Gatidis

May 8, 2023

1 Synopsis

Billions of MR images are currently stored in clinical archives world-wide. Manual annotation of all these images for the purpose of supervised machine learning is impossible. To still leverage their potential, novel machine learning strategies are thus required.
Recently, self-supervised learning approaches have been proposed enabling training of deep learning models without requiring human-generated labels. The technical feasibility of these methods has previously demonstrated for patch-wise training.
In this work we propose the use of augmentation-based contrastive learning for the analysis of over 40,000 entire-volume whole-body MRI data from the UK Biobank study and demonstrate its applicability to various down-stream tasks.

2 Summary of Main Findings

3D-Contrastive learning is performed on 40,000 full abdominal MRI volumes. Subsequent application to classification and regression tasks outperformed the supervised baselines significantly, demonstrating the advantage of contrastive learning.

3 Introduction

Deep Learning frameworks have shown remarkable results in many fields of MR image analysis over the past decade.[1–4] One main drawback of supervised approaches is their need for large fully annotated training datasets which are difficult and often impossible to acquire. Additionally, once trained, these frameworks cannot be easily extended to different tasks as re-training and full access to the initial dataset is usually required. Furthermore, it is desirable to learn generic feature representations which can be reused for multiple tasks, instead of task-specific features.
Over the past years, self-supervised learning (SSL) methods have been introduced, aiming to alleviate the above-described challenges. Several self-supervision tasks have been proposed including e.g., prediction of image rotations,[5] image colorization,[6] solving jigsaw puzzles[7] or image inpainting.[8] More recently, self-supervised contrastive frameworks based on image augmentation have been shown to improve on earlier results[9–12] and have also been applied in the field of MRI.[13–15] Previous work focused on patch-wise image processing using 2D or 3D patches[16] as input.

Application of self-supervised contrastive learning on entire 3D MRI volumes however is associated with additional computational challenges and requires a larger number of training data.
Our contributions in this work are: (i) We propose a 3D feature extractor trained in a self-supervised contrastive manner using SimCLR[9] on entire-volume MRI volumes. (ii) we deploy this framework on 40,000 abdominal MRI scans from the UK Biobank (UKB). (iii) We use the learned representations in various downstream tasks yielding high performance for classification and regression based on only few labelled examples and without access to the initial dataset or re-training.

4 Methods

We hypothesize that - using contrastive learning - a feature embedding of an entire volume can be learned which is suitable for various downstream tasks. We therefore train a self-supervised feature encoder based on the SimCLR framework[9] on full whole-body MRI volumes.
During pre-training (Fig. 1A), two different views of the same volume are created using domain specific data augmentation (resized crops: 0.1 to 1, Gaussian noise $\mathcal{N}(0, 0.1)$, gamma contrast variation: -0.5 to 0.5, Gaussian blurring $\sigma = 2$) which are fed to a 3D-ResNet18[17] encoder to produce feature vectors of length 512. To improve performance[9] a subsequent two-layer MLP is used to project the two feature vectors into the loss-space before application of the Normalized Temperature-scaled Cross-Entropy loss.[18]
For the obtained feature vectors z, the cosine similarity (**sim**)is computed yielding the following final optimization term:

$$\ell = -log\frac{\exp(\mathbf{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{k\neq i}\exp(\mathbf{sim}(z_i, z_k)/\tau)} \tag{1}$$

The pre-trained ResNet is used to embed each MR volume into a feature representation which can subsequently be used as input for further downstream tasks.
We trained this framework on 40,000 abdominal MRI volumes from the UK Biobank[19] study based on a dual echo GRE sequence acquired on a clinical 1.5 T scanner (Magnetom Aera, Siemens Healthineers) with the following parameters: matrix size: $224 \times 168 \times 363$, resolution: $2.23 \times 2.23\text{mm}^2$, TE/TR: 2.39 ms, 4.77 ms / 6.69 ms, flip angle: 10°, bandwidth: 440 Hz/pixel). We used the out-of-phase image volumes as input. Training was performed for 200 epochs on 8 dedicated GPUs (A100, NVIDIA) with a batch size of 56 and the ADAM optimization[20] paired with a linear warmup scheduler up to a learning rate of 1e-3.
Experiments:
Evaluation was performed on several classification and regression tasks using 4,174 validation subjects. For classification tasks, a single linear layer consisting of 100 neurons was trained for 10,000 steps on top on the resulting feature embeddings. For regression, plain ridge regression with an α of 4 was used.
Accuracy was used as metric for classification performance for the prediction of sex and an overall health score (1: Poor to 4: Excellent), whereas the mean absolute error (MAE) is reported for the target variables age, weight, height and body fat. The ground truth labels were provided by the UKB. To assess the efficacy of self-supervised training, we trained the single layer / Ridge Regression (see above) on 2, 20, 200 and 400 randomly selected examples and compared these results with fully supervised training using the ResNet18 architecture directly on image data for the same examples.
t-SNE plots were used to visualize the data manifold of feature representations produced by SSL.[21]

5 Results and Discussion

Quantitative evaluations are visualized in Figures 2/3 and qualitative evaluation is depicted in Fig. 5. Accuracy and MAE were significantly better for the SSL-model when using only a small dataset for task-specific training of up to 20 subjects suggesting that self-supervised training indeed produces useful feature representations of data.
The t-SNE plots of feature representations produced by SSL showed a clear separation of sex in accordance with the visual perception but also includes other features such as weight(Fig. 4).
This study has limitations. In this work we mostly focused on simple demographic target variables (age, weight) for prediction tasks. Estimation of more specific properties such as diseases, as well as SSL training for tissue segmentation will be part of future work. Additionally, the impact of the labelled subset will be investigated.

6 Conclusion

In this study we implemented and evaluated large-scale contrastive learning on entire-volume whole-body MRI data using SimCLR generating useful feature representations that can be leveraged for various downstream tasks.

7 Acknowlegments

T.K. and S.G. contributed equally.
This work was carried out under UK Biobank Application 40040. We thank all participants who took part in the UKBB study and the staff in this research program.

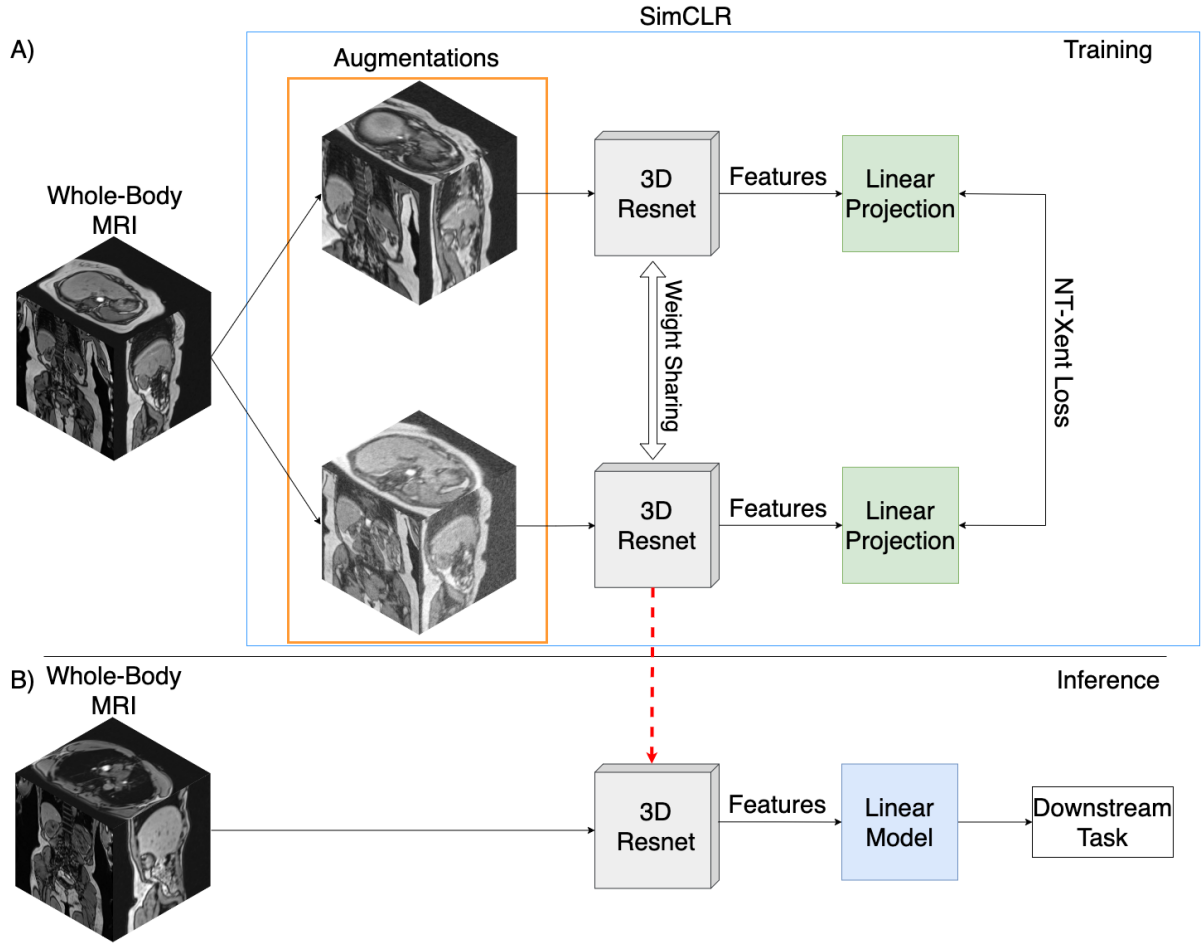


Figure 1: Using SimCLR for self-supervised pre-training of a 3D-ResNet18 to produce feature embeddings of two different augmentations based on the same initial volume. Application of the NT-Xent loss function forces the linear projected features to be similar. During inference (B) the trained ResNet18 is used to embed the full non-augmented volume into a feature vector of length 512 which can be subsequently used as input for various downstream tasks.

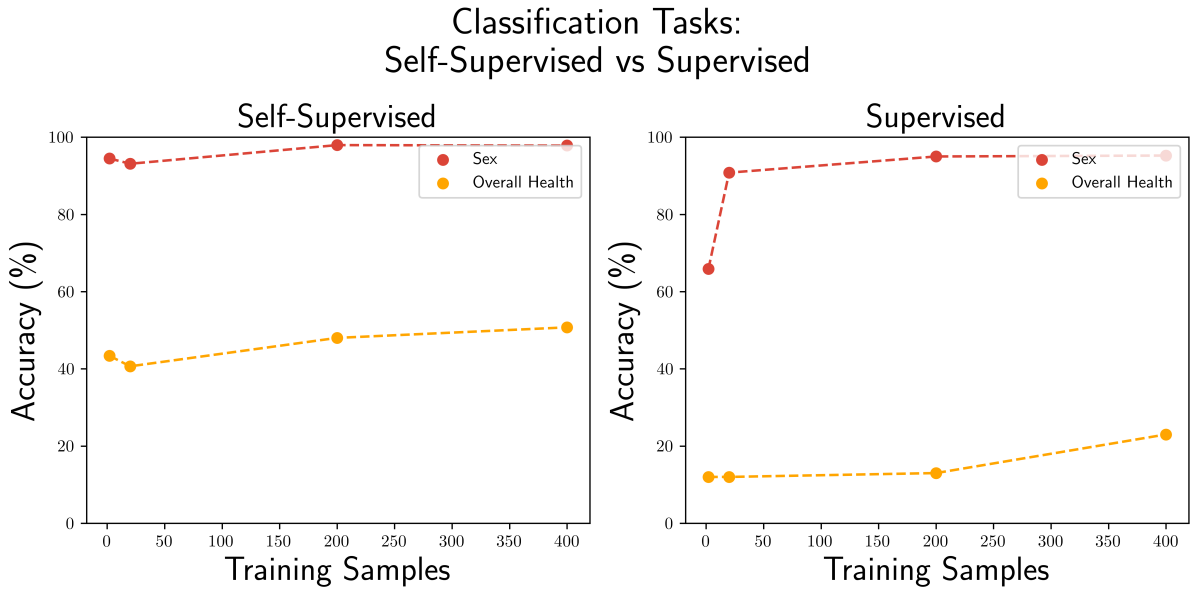


Figure 2: Accuracy dependency on the number of available labelled training samples in the fully supervised baselines (right) and the proposed self-supervised approach (left) for classification of sex and overall health. Using the proposed feature embedding as input improves performance if only limited training data is available.

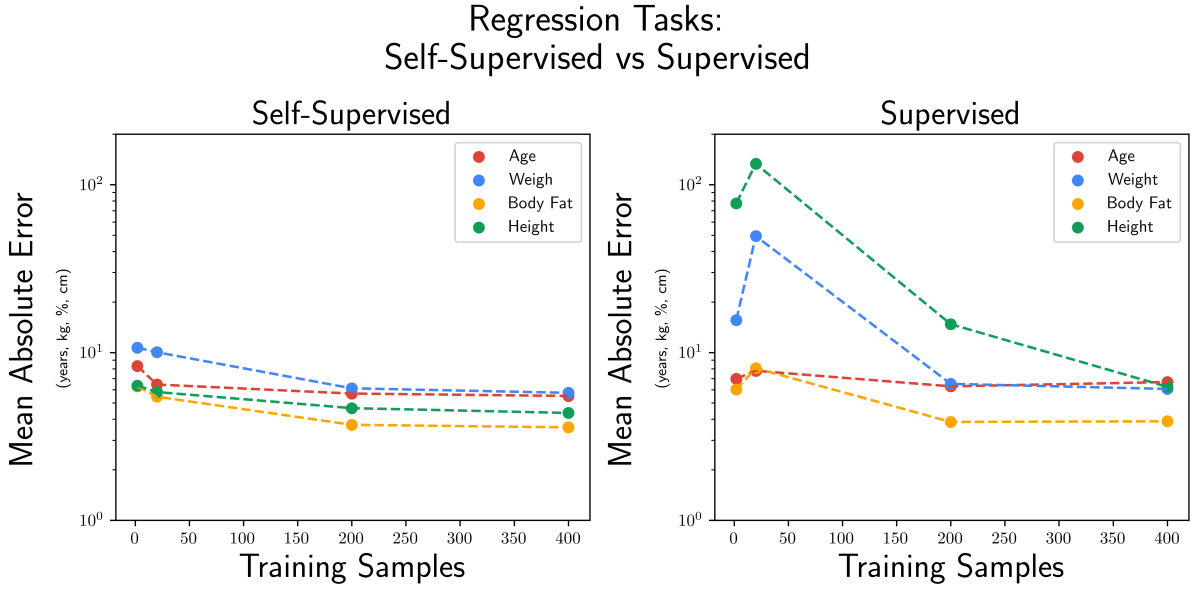


Figure 3: Mean Absolute Error (in years, kg ,% and cm) dependency on the number of available labelled training samples in the fully supervised baselines (right) and the proposed self-supervised approach(left) for regression of age, body fat, weight and height. Using the proposed feature embedding as input yields significant improvement compared to the supervised baselines. Performance gains are considerably higher compared to the classification task.

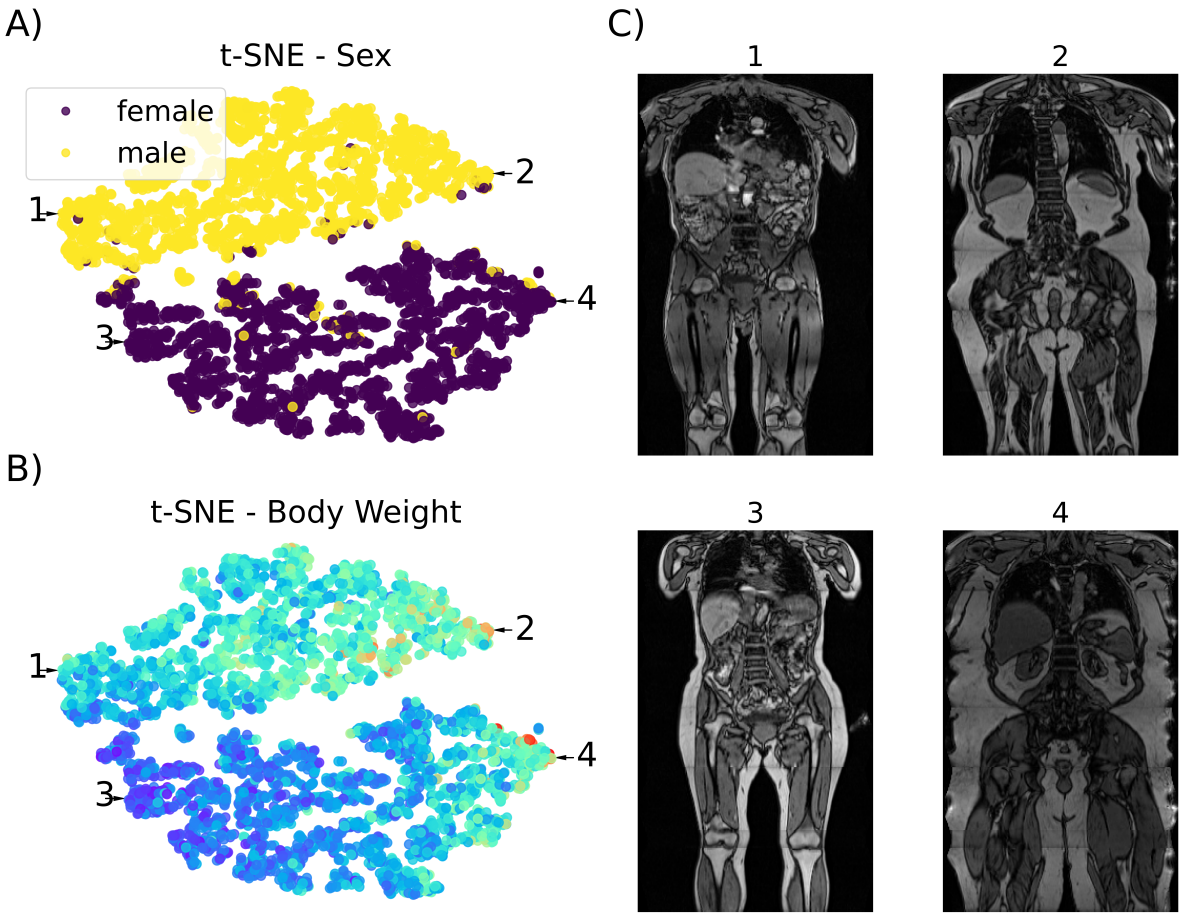


Figure 4: t-SNE visualization of the estimated feature embedding for the proposed self-supervised approach on the validation dataset colorized for A) sex and B) body weight paired with 4 corresponding sample images (C)). The proposed approach clearly learned to use the visible sexual characteristics to distinguish between volumes but also incorporates additional knowledge such as weight.

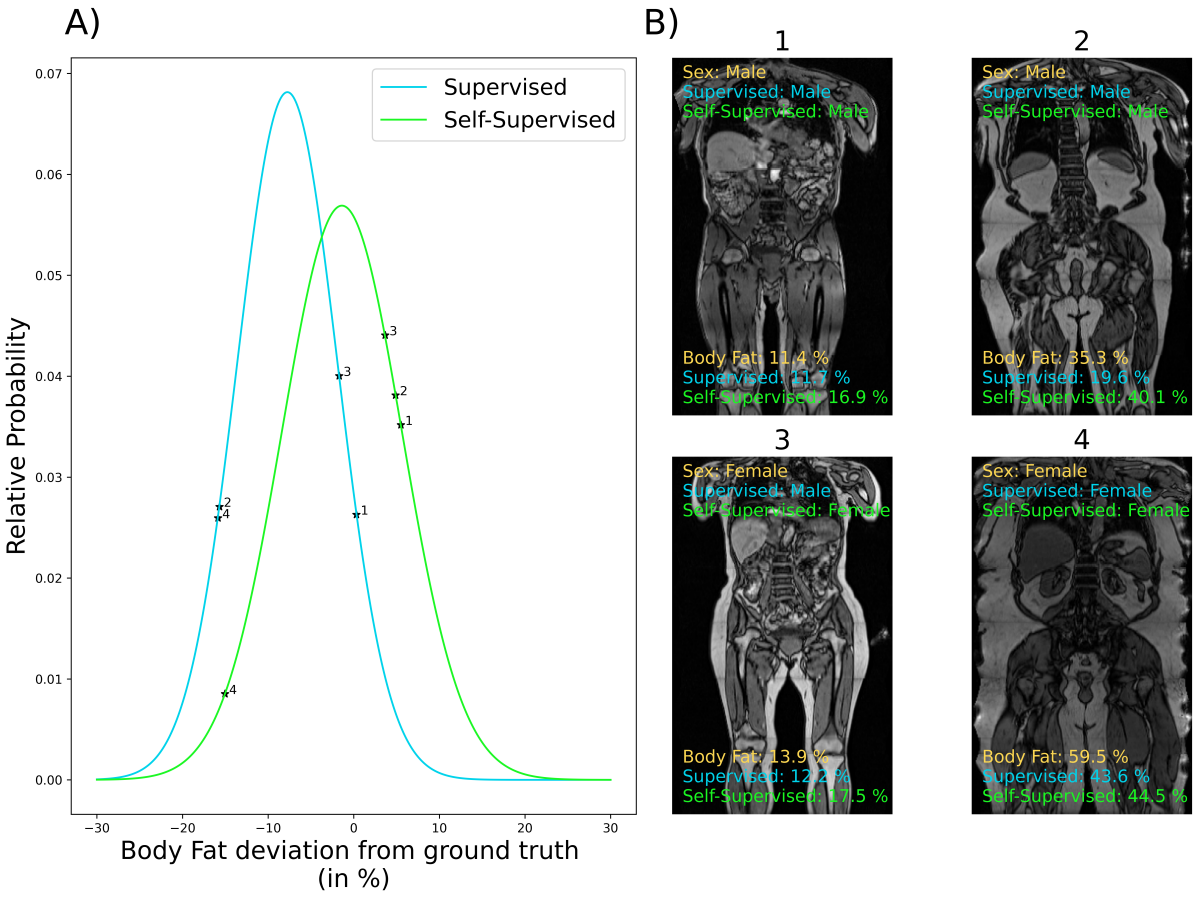


Figure 5: A) Distribution of body fat percentage deviation between prediction and ground truth for the supervised baseline (blue) and self-supervised model (green) using 20 labelled examples. The 4 annotations are highlighted in B) which also depicts the corresponding predictions for body fat and sex paired with ground truth (orange). Overall, the self-supervised model markedly outperforms the supervised baseline.

References

[1] C. M. Hyun, H. P. Kim, S. M. Lee, S. Lee, and J. K. Seo, "Deep learning for undersampled mri reconstruction," *Physics in Medicine & Biology*, vol. 63, no. 13, p. 135007, 2018.

[2] T. Küstner, N. Fuin, K. Hammernik, A. Bustin, H. Qi, R. Hajhosseiny, P. G. Masci, R. Neji, D. Rueckert, R. M. Botnar, *et al.*, "Cinenet: deep learning-based 3d cardiac cine mri reconstruction with multi-coil complex-valued 4d spatio-temporal convolutions," *Scientific reports*, vol. 10, no. 1, pp. 1–13, 2020.

[3] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, "Deep learning for brain mri segmentation: state of the art and future directions," *Journal of digital imaging*, vol. 30, no. 4, pp. 449–459, 2017.

[4] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on mri," *Journal of magnetic resonance imaging*, vol. 49, no. 4, pp. 939–954, 2019.

[5] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[6] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6874–6883, 2017.

[7] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2229–2238, 2019.

[8] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[10] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*, pp. 12310–12320, PMLR, 2021.

[11] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *arXiv preprint arXiv:2105.04906*, 2021.

[12] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2020.

[13] M. Frueh, A. Schilling, S. Gatidis, and T. Kuestner, "Real time landmark detection for within-and cross subject tracking with minimal human supervision," *IEEE Access*, vol. 10, pp. 81192–81202, 2022.

[14] C. Zhao, B. E. Dewey, D. L. Pham, P. A. Calabresi, D. S. Reich, and J. L. Prince, "Smore: a self-supervised anti-aliasing and super-resolution algorithm for mri using deep learning," *IEEE transactions on medical imaging*, vol. 40, no. 3, pp. 805–817, 2020.

[15] M. Frueh, T. Kuestner, M. Nachbar, D. Thorwarth, A. Schilling, and S. Gatidis, "Self-supervised learning for automated anatomical tracking in medical image data with minimal human labeling effort," *Computer Methods and Programs in Biomedicine*, p. 107085, 2022.

[16] Y. Ali, A. Taleb, M. M.-C. Höhne, and C. Lippert, "Self-supervised learning for 3d medical image analysis using 3d simclr and monte carlo dropout," *arXiv preprint arXiv:2109.14288*, 2021.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[18] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," *Advances in neural information processing systems*, vol. 29, 2016.

[19] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, *et al.*, "Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS medicine*, vol. 12, no. 3, p. e1001779, 2015.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.