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.

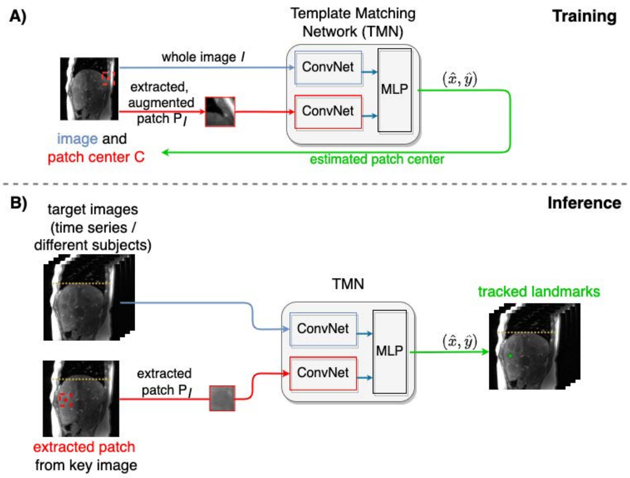


FIGURE 1. The proposed landmark detection and tracking framework consists of a Template Matching Network that combines two encoders (ConvNet) whose outputs are fed into a shared multi-layer perceptron (MLP). One encoder is used to process the full source image (global information) whereas the second 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:

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

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, 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.

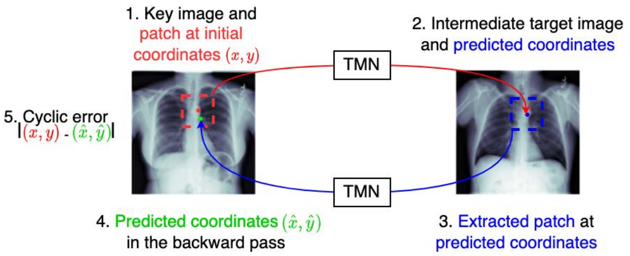

1. Key image and patch at initial coordinates (x, y)
2. Intermediate target image and predicted coordinates
TMN
TMN
5. Cyclic error $|(x, y) - (\hat{x}, \hat{y})|$
4. Predicted coordinates $(\hat{x}, \hat{y})$ in the backward pass
3. Extracted patch at predicted coordinates

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).

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:

$$\hat{x} = \frac{\sum_{i=1}^{N} \hat{x}_i \cdot \frac{1}{\hat{\sigma}_{x_i}}}{\sum_{i=1}^{N} \frac{1}{\hat{\sigma}_{x_i}}}, \tag{3}$$

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

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

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 $\mathbf{P}_{\mathbf{T_i}}$ 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 ± 11.52 years, matrix size = 352 × 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 = $2 \times 2\text{mm}^2$; matrix size = 192 × 176; acquisition time/image = 0.4s) yielding 2D sagittal motion-resolved MR images of the body trunk [43]. 36 patients (60±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.

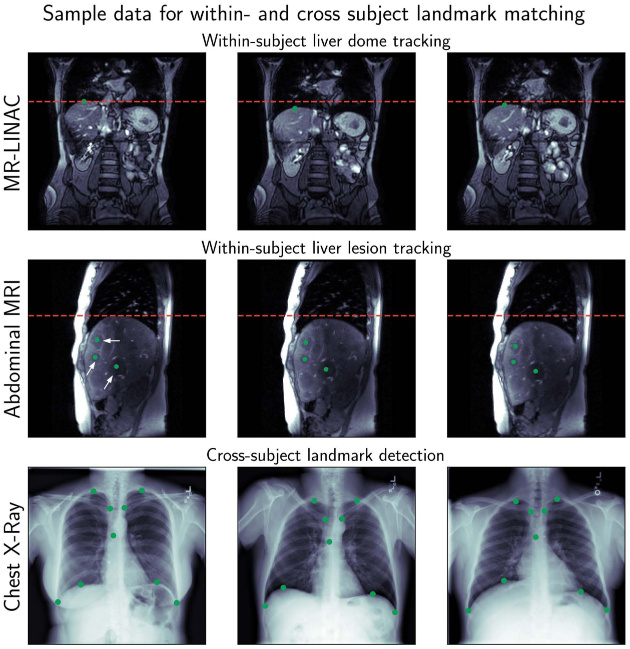

Sample data for within- and cross subject landmark matching
Within-subject liver dome tracking
MR-LINAC
Within-subject liver lesion tracking
Abdominal MRI
Cross-subject landmark detection
Chest X-Ray

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 .

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 ± 17.8 years, 40.6% female). Images were acquired with varying matrix size (resampled to 224 × 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

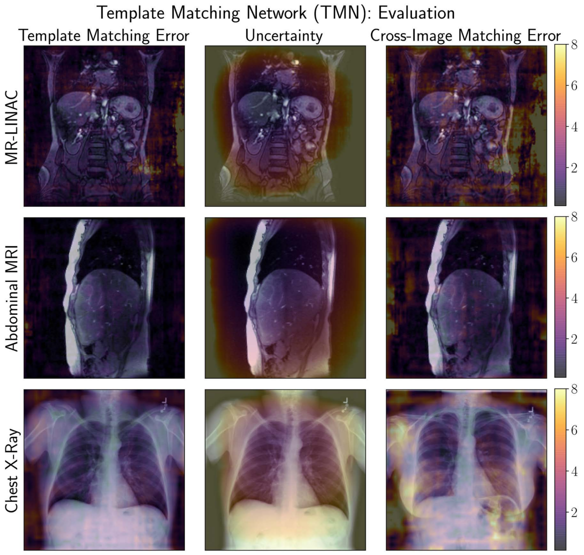


FIGURE 4. Color-coded mean euclidean pixel-wise template matching error (left column, III-A), corresponding estimated uncertainty (central column, III-A) and mean euclidean pixel-wise cross-image matching error (right column, III-B) for each pixel in three representative examples from the MR-LINAC (top 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

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.

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

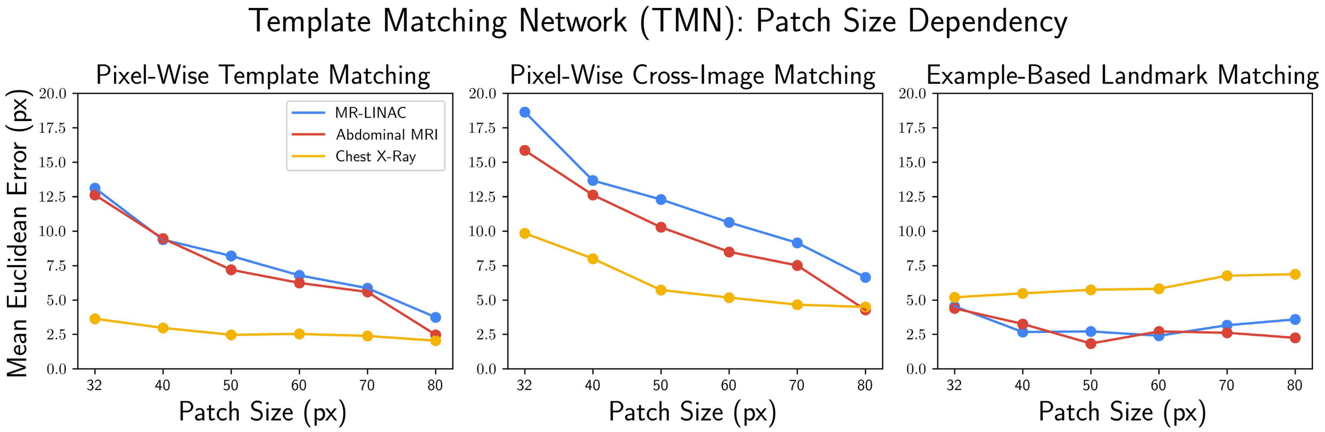


FIGURE 5. Dependency of the mean euclidean errors on various patch sizes in (left) template matching (pixel-wise template matching, III-A), (central) cross-image matching (cyclic error; within-subject tracking for MR-LINAC and abdominal MRI, cross-subject detection for chest X-ray, III-B) and (right)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

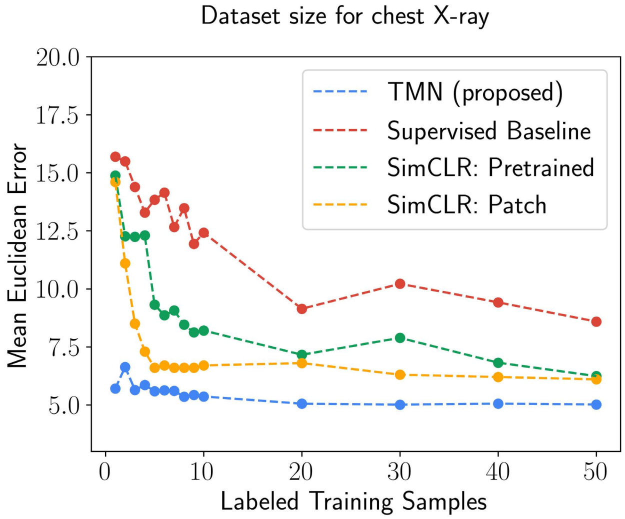


FIGURE 6. Dependency of the average euclidean landmark matching (cross-subject) error on the number of available labeled examples in the 9 chest X-ray landmarks (III-C2) for the proposed template matching network (TMN) (blue), a supervised baseline (red), the patch-wise feature matching baseline (orange)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

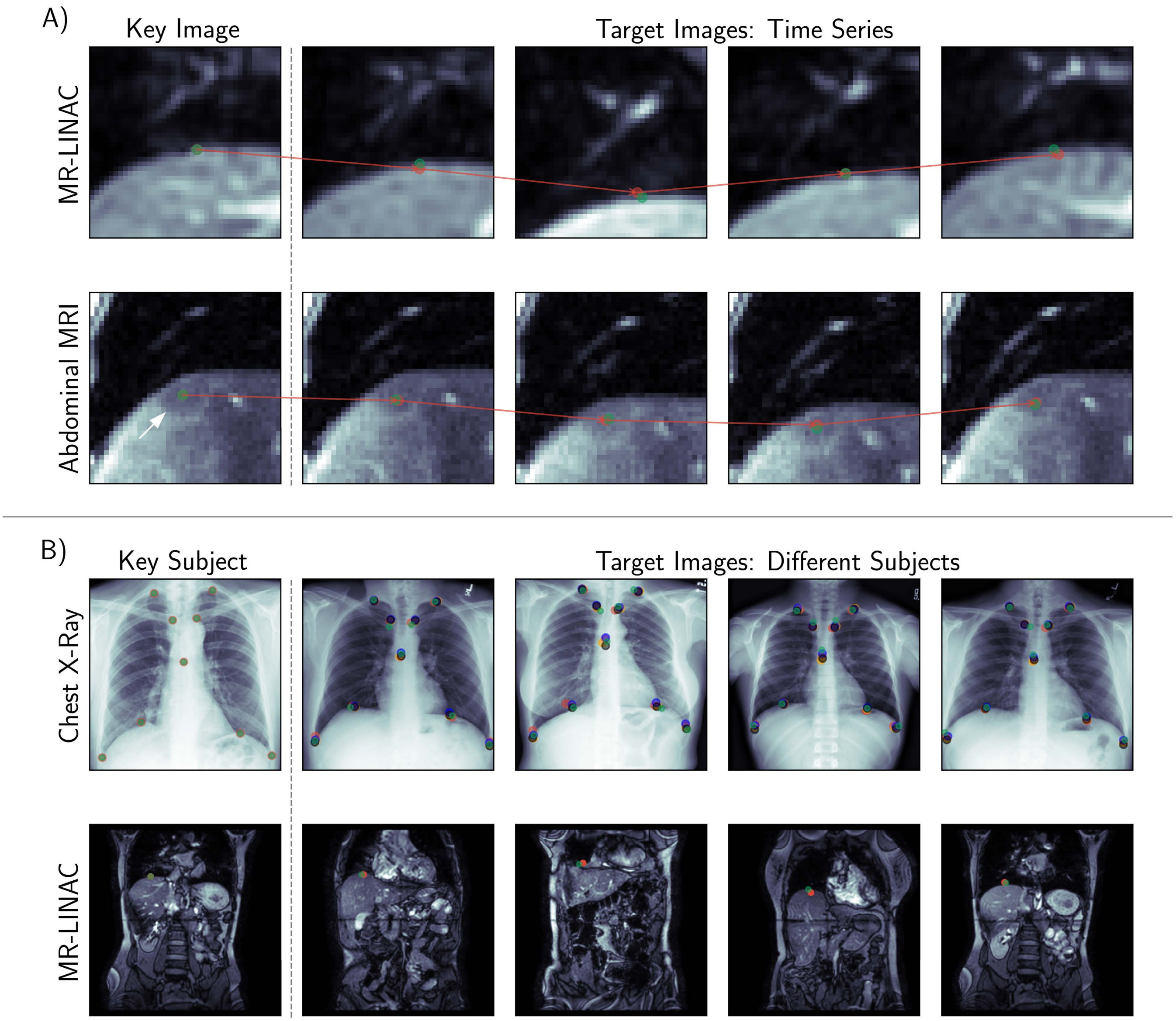


FIGURE 7. Qualitative evaluation of example-based landmark matching (III-C) for (A) within-subject motion tracking and (B) cross-subject anatomical landmark detection. A) Visualized within-subject motion tracking of the liver dome on MR-LINAC data (top) and a liver lesion within the abdominal MRI dataset 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, "Track-Former: 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.

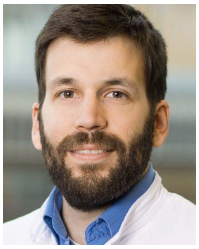
[49] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," 2022, *arXiv:2201.03545*.

[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.