One-Shot Medical Landmark Localization by Edge-Guided Transform and Noisy Landmark Refinement

Zihao Yin[1], Ping Gong[2], Chunyu Wang[3], Yizhou Yu[4], and Yizhou Wang[5,6](✉)

[1] Center for Data Science, Peking University, Beijing, China
[2] Deepwise AI Lab, Beijing, China
[3] Microsoft Research Asia, Beijing, China
[4] The University of Hong Kong, Hong Kong
[5] Center on Frontiers of Computing Studies, School of Computer Science, Peking University, Beijing, China
[6] Inst. for Artificial Intelligence, Peking University, Beijing, China
{silvermouse, yizhou.wang}@pku.edu.cn, gongping@deepwise.com, chnuwa@microsoft.com, yizhouy@acm.org

Abstract. As an important upstream task for many medical applications, supervised landmark localization still requires non-negligible annotation costs to achieve desirable performance. Besides, due to cumbersome collection procedures, the limited size of medical landmark datasets impacts the effectiveness of large-scale self-supervised pre-training methods. To address these challenges, we propose a two-stage framework for one-shot medical landmark localization, which first infers landmarks by unsupervised registration from the labeled exemplar to unlabeled targets, and then utilizes these noisy pseudo labels to train robust detectors. To handle the significant structure variations, we learn an end-to-end cascade of global alignment and local deformations, under the guidance of novel loss functions which incorporate edge information. In stage II, we explore self-consistency for selecting reliable pseudo labels and cross-consistency for semi-supervised learning. Our method achieves state-of-the-art performances on public datasets of different body parts, which demonstrates its general applicability. Code is available at https://github.com/GoldExcalibur/EdgeTrans4Mark.

Keywords: Medical Landmark Localization, One-Shot Learning

1 Introduction

Landmark localization is an essential step for many medical image applications, such as dental radiography [1,2], bone age assessment [3,4], vertebra labeling [5] and per-operative measurements [6]. Although fully supervised methods [7,8,9] achieve the state-of-the-art results, the required manual annotations take considerable cost and time. In contrast, given few exemplars and proper instructions, experts are ready to generalize these landmark concepts to unseen targets and

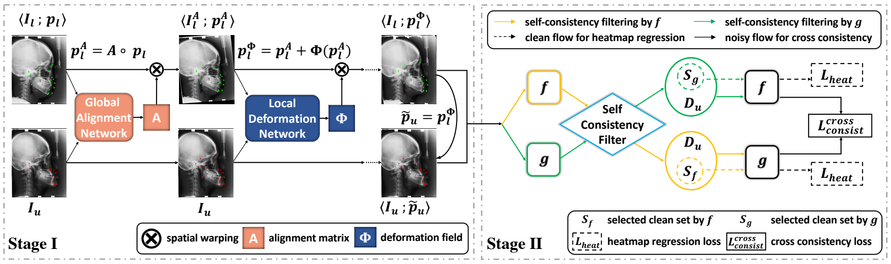
$\langle I_l\,;\,p_l\rangle$
$p_l^A = A \circ p_l$
Global Alignment Network
A
$\langle I_l^A\,;\,p_l^A\rangle$
$p_l^\Phi = p_l^A + \Phi(p_l^A)$
Local Deformation Network
Φ
$\langle I_l\,;\,p_l^\Phi\rangle$
$\tilde{p}_u = p_l^\Phi$
I_u
I_u
$\langle I_u\,;\,\tilde{p}_u\rangle$
Stage I
spatial warping
A alignment matrix
Φ deformation field
self-consistency filtering by f
self-consistency filtering by g
clean flow for heatmap regression
noisy flow for cross consistency
f
g
Self Consistency Filter
S_g
D_u
D_u
S_f
f
g
L_{heat}
$L_{consist}^{cross}$
L_{heat}
Stage II
S_f selected clean set by f
S_g selected clean set by g
L_{heat} heatmap regression loss
$L_{consist}^{cross}$ cross consistency loss

Fig. 1. Overview of the proposed framework. In stage I, unsupervised registration is learned through an end-to-end cascade of global alignment and subsequent local deformations, which aims to predict pseudo landmarks $\tilde{p}_u$ by registering the labeled exemplar I_l to unlabeled targets I_u. Inferred noisy landmarks are further refined in stage II, where two robust landmark detectors f, g are co-trained, by utilizing both self-consistency for sample selection and cross-consistency for semi-supervised learning

annotate them accurately. This motivates us to explore the challenging task of one-shot medical landmark localization.

Besides the scarce supervision, significant differences in spatial structures between images for landmark datasets also increase the difficulty of this task. While they can be quite different in scale, orientation, or intensity due to patient positioning or imaging quality as in hand radiography [3], there are also substantial variations in local structures, such as the front teeth in dental radiography [1]. Furthermore, because of cumbersome acquisition procedures, medical landmark datasets are too expensive to collect in large numbers. Thus, the amount of unlabeled data available is often limited.

Considering these challenges, landmark localization in the low data regime is in urgent need and explored by [10,11,12]. 3FabRec [10] is a method for few-shot face alignment. They first train an autoencoder for face reconstruction and then retask the decoder to heatmap prediction through fine-tuning on labeled sets. However, when having access to only one exemplar and hundreds of data as in our work, qualities of reconstructed images are too poor to perform landmark localization. Motivated by the recent success of contrastive learning, CC2D [11] proposes to detect target landmarks by first solving a self-supervised patch matching task and uses these pseudo labels to retrain new detectors. However, their method overlooks the global spatial relationships of landmarks and is prone to yield inaccurate predictions once overfitting to the local appearance of specific instances. Conversely, DAG [8] employs graph convolution network (GCN) to capture topological constraints of landmarks. Few-shot DAG [12] extends DAG to the few-shot (e.g., five-shot) setting. They report impressive results on several datasets but fail to converge under the extreme one-shot setting.

In order to fully exploit the exemplar and available data, we propose to propagate the landmarks from exemplar to unlabeled images through registration, which not only considers global anatomical constraints, but also performs precise matching of local structures. In fact, the dense correspondence between instances learned by registration can be directly leveraged by landmark localization. Besides, registration can be learned efficiently without the need of a large amount of data to produce reasonable results. As shown in Fig. 1, our novel two-stage framework first learns unsupervised registration from the labeled exemplar to unlabeled targets for inferring pseudo landmarks, and then trains robust landmark detectors by exploiting consistency between clean annotations and noisy pseudo labels. For better adaptation to landmark localization, we make several non-trivial contributions to solve the following challenges.

First, there might be a chicken-and-egg issue if you want to infer landmarks for unseen targets through registration, since registration itself usually requires detected landmarks, either for alignment in pre-processing steps [13], or to guide the learning process as extra structural information [14]. To avoid this dilemma, we decompose the total spatial transform into an end-to-end cascade of global alignment and local deformations. To facilitate the registration learning, powerful attention blocks [15] are employed, including self-attention for capturing long-range dependencies and cross-attention for fusing multi-resolution features.

Second, classical reconstruction terms for registration, which mainly consider image similarity based on the distribution of pixel values, are insufficient to constrain the structural consistency of landmarks. Hence we propose novel loss functions that incorporate edge information and landmark locations. Specifically, our reconstruction term further involves the masked similarities of edge structures around interested landmarks. Besides, since different anatomical parts tend to have different displacements, it is beneficial to relax the smoothness constraints of deformation field around boundaries.

Last, under the interference of certain nuisances (e.g., background, abnormal appearance), deformation learned by the low-level registration task can not perfectly capture the high-level semantic correspondence of landmarks. Thus, an essential second stage is introduced to refine noisy predictions with large biases. We train robust landmark detectors, utilizing self-consistency between different views of the same model for sample selection and cross-consistency between different views of different models for semi-supervised learning.

We conduct experiments on public medical landmark datasets of different body parts, including head, hand and chest. Our method consistently outperforms other baselines with notable margins and further narrows the gap with the supervised upper bound. To summarize, our contributions are three-fold:

1. We propose an unsupervised training strategy for inferring pseudo landmarks through registration, which learns an end-to-end cascade of global alignment and local deformations, with the guidance of novel loss functions incorporating edge information.
2. We introduce an effective scheme for training robust landmark detectors with noisy labels, which utilizes self-consistency for selecting reliable pseudo labels and cross-consistency for semi-supervised learning.
3. We conduct experiments on public medical landmark datasets of different body parts. Results show our method stably advances the state-of-the-art for all three applications.

2 Related Work

We briefly review most related works, including one-shot, few-shot and semi-supervised methods for landmark localization, and medical image registration.

2.1 One-shot and few-shot landmark localization

As mentioned, CC2D [11] is motivated by contrastive learning: features for the original patch and its randomly augmented counterparts at the same location are matched using cosine similarity. With the learned matching network and template patches, pseudo-labels are inferred for retraining a multi-task UNet [16] from scratch. Despite their promising results on the cephalometric dataset [1], CC2D overlooks valuable global structure constraints [8], making it difficult to handle multiple similar local structures, such as fingertips. Another interesting work, 3FabRec [10] achieves impressive performance for few-shot face alignment. They first train an adversarial autoencoder for unsupervised face reconstruction, then fine-tune with interleaved layers to the landmark detection task with few labels. 3FabRec demonstrates great benefits of dense pixel-level pre-training for landmark localization. However, for medical applications where the amount of data is orders of magnitude less, it is rather difficult to achieve satisfactory results through pre-training of image reconstruction.

2.2 Semi-supervised landmark localization

Recent advances can be divided into two streams: consistency-based approaches [17,18] and synthetic image based approaches [19,20]. The equivariant landmark transformation (ELT) constraint [17] is built on the intuition that, given a transformed image, the model should produce similarly transformed landmarks. Semantic Consistency [18] encourages learning similar features for landmarks with the same semantics across images. Synthetic image approaches focus on generating desired training data. StyleAlign [19] first disentangles face images to style and structure space, then transfers randomly sampled styles to images with known structures, greatly enriching training space.

2.3 Learning-based deformable image registration

Our work is closely related to learning-based image registration [21,13,22,14]. [14] proposes to utilize detected landmarks as extra structural information to guide the training of registration. VoxelMorph [13] proposes to use a convolutional neural network (CNN) g to learn the registration field Φ unsupervisedly, by optimizing the image similarity between the fixed and moving images, and the smoothness of Φ. [13] requires image pairs to be affinely aligned in the preprocessing step and then focus on the nonlinear correspondence. [13] can further be extended as in DataAug [23] for one-shot medical segmentation, by learning independent spatial and appearance transform for data augmentation.

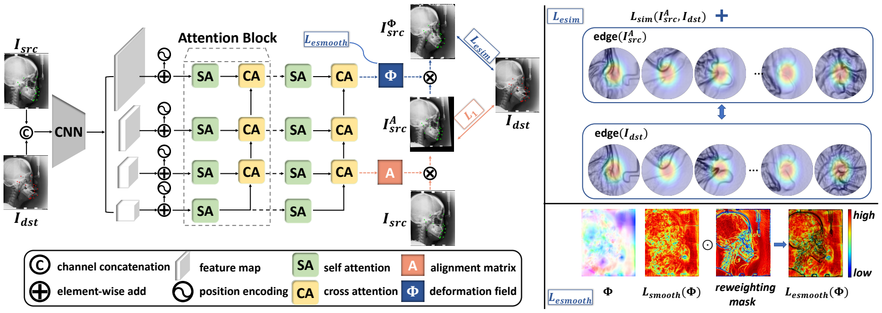

Fig. 2. The architecture of registration model consists of a CNN backbone and cascades of attention blocks. Red lines denote the flow for global alignment, while blue lines denote the flow for local deformation. Besides the image similarity, $\mathcal{L}_{esim}$ is introduced to further penalize local similarities of edge structures around inferred pseudo landmarks, masked by their gaussian heatmaps. $\mathcal{L}_{esmooth}$ relaxes the smoothness constraint around boundaries, to avoid mutual interference between different regions (e.g., anatomical parts, background & foreground)

3 Method

Let I_l, I_u be two images defined over a 2-D spatial domain Ω. $I_l \in \mathbb{R}^{H\times W}$ is the labeled exemplar with N annotated landmarks $p_l = \{(x_l^j, y_l^j) | j = 1, \cdots, N\}$, and I_u is the unlabeled target. I_l and I_u share similar appearance distribution, and landmarks p_l are defined at locations with particular anatomical structures. Our goal is to learn a spatial transformation from I_l to I_u through registration, so that we can infer reliable landmarks for I_u based on p_l, according to the equivariant property of landmarks.

3.1 Reformulation of Classical Registration Framework

First, in order to adapt to the downstream landmark localization, we reformulate the classical framework for unsupervised registration as follows:

$$\hat{A}, \hat{\Phi} = \arg\min_{A,\phi} \mathcal{L}_{total}(I_{src}, I_{dst} \mid \tilde{p}_{src}), A = g_\theta(I_{src}, I_{dst}), \Phi = g_\theta(I_{src}^{A}, I_{dst}) \quad (1)$$

$$\mathcal{L}_{total} = \mathcal{L}_{global}(I_{src}^{A}, I_{dst}) + \mathcal{L}_{local}(I_{src}^{\Phi}, I_{dst} \mid \tilde{p}_{src}^{\Phi}) + \mathcal{L}_{esmooth}(\Phi \mid I_{src}^{\Phi}) \quad (2)$$

$$I_{src}^{A} = A \otimes I_{src},\ I_{src}^{\Phi} = \Phi \otimes I_{src}^{A},\ \tilde{p}_{src}^{A} = A \circ \tilde{p}_{src},\ \tilde{p}_{src}^{\Phi} = \tilde{p}_{src}^{A} + \Phi(\tilde{p}_{src}^{A}) \quad (3)$$

where we learn an end-to-end cascade of global affine alignment $A \in \mathbb{R}^{2\times 3}$ and local deformation $\Phi \in \mathbb{R}^{H\times W\times 2}$. During training, $\langle I_{src}, I_{dst}\rangle$ is a pair of images randomly sampled from the mini-batch. For inference, I_{src} is fixed as the exemplar I_l and we set I_{dst} as the unlabeled target I_u to predict pseudo landmarks $\tilde{p}_u$. The local similarity $\mathcal{L}_{local}$ is conditioned on $\tilde{p}_{src}^{\Phi}$ (Eq. 3), where $\tilde{p}_{src}$ is an exponential moving average of e-th epoch prediction $\tilde{p}_{src}^{e}$ after certain epoch M:

$$\tilde{p}_{src} = \tau\tilde{p}_{src} + (1-\tau)\tilde{p}_{src}^{e},\ \ e >= M \quad (4)$$

The smoothness constraint $\mathcal{L}_{esmooth}$ is extraly conditioned on I^{Φ}_{src}, to utilize its edge information, which will be discussed in Eq. 16.

3.2 Edge-Guided Global & Local Transform

As in Fig. 2, our registration model g_θ consists of a CNN backbone for feature extraction and several attention blocks for feature fusion.

We adopt the HRNet18 [24] as backbone, which takes a channel-wise concatenation of I_{src} and I_{dst} as input, and outputs four resolutions of feature maps $F_i \in \mathbb{R}^{C\times H_i\times W_i}$ with stride S_i ($H_i = \frac{H}{S_i}, W_i = \frac{W}{S_i}, S_i \in \{32, 16, 8, 4\}, i \in \{1, 2, 3, 4\}$). For each F_i, we use a 1×1 convolution to change its channel dimension into a unified value d_{model} and then reshape it to $\mathbb{R}^{d_{model}\times H_iW_i}$, since the attention blocks expect a sequence as input. Following [25,26], we adopt a concatenation of two 1D learned positional encodings as the recovered order information.

Because of the significant structural differences between I_{src} and I_{dst}, distance between their corresponding pixels could be large. To capture these long-range dependencies, we furnish our model with powerful transformer layers [15]:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{5}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{H}_1, \cdots, \text{H}_h)W^O \tag{6}$$

$$\text{H}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{7}$$

where Q, K, V are embeddings of the same feature map, projected into different spaces for the self-attention module (SA). It is also crucial to fully exploit features from multiple resolutions for registration learning. Thus, we further utilize the cross-attention module (CA) for cross-resolution feature fusion, where K, V are embeddings of the feature map from branch of lower-resolution than Q. For each layer of attention blocks, we stack both one SA module and one CA module following feature maps from each resolution, which progressively incorporates the information from lower resolutions to higher resolutions, and enables the model to gradually figure out the optimal transformation from I_{src} to I_{dst}.

Global Transform Since high-level features are sufficient to capture global relationships, features of lower resolutions (i.e., F_1, F_2) are passed into transformer layers and we use the aggregated representation of the last output state for affine estimation. It is then passed to a two-layer MLP and tanh function to obtain $o \in \mathbb{R}^6$, each element of which denotes the relative changes in translation (t_x, t_y), scale (s_x, s_y), rotation α, and shear β:

$$o = \tanh(\text{MLP}(H)) \tag{8}$$

$$t_x = o_1,\ s_x = 1 + o_3 * sf_x,\ \alpha = o_5 * rot \tag{9}$$

$$t_y = o_2,\ s_y = 1 + o_4 * sf_y,\ \beta = o_6 * sh \tag{10}$$

where sf_x, sf_y, rot, sh are four hyper parameters controlling transformation intensities. Then the affine matrix $A \in \mathbb{R}^{2\times 3}$ can be computed as follows:

$$\begin{pmatrix} s_x \cos\alpha & s_x(\cos\alpha \tan\beta + \sin\alpha) & t_x \\ -s_y \sin\alpha & s_y(-\sin\alpha \tan\beta + \cos\alpha) & t_y \end{pmatrix}$$

which is a composite of several basic transforms. Then I_{src}^{A} is obtained through differentiable bilinear interpolation based on the spatial transformer module [27].

Local Deformation Field Based on the global transform A, features are again extracted for $\langle I_{src}^{A}, I_{dst} \rangle$. For a more precise pixel-level correspondence, we mainly utilize high-resolution features to predict a deformation field $\varPhi \in \mathbb{R}^{H\times W\times 2}$. Since global transform already roughly aligns two images, the target pixel for a certain location of I_{src}^{A} is more likely to be within a local neighborhood in I_{dst}. Thus, for high-resolution features (i.e., F_3, F_4), we can reduce their spatial scale through reshaping them from $C \times H_i W_i$ to R^2 sequences (e.g., $R = 4$) with size $C \times \frac{H_i W_i}{R^2}$ and perform attention mechanism on each sequence.

The final hidden state from the highest-resolution branch is fed into a linear layer for regressing displacements $(\varDelta x, \varDelta y)$, which implies that $I_{src}^{G}(x + \varDelta x, y + \varDelta y)$ corresponds to $I_{dst}(x, y)$ $(x \in \{1, \cdots, h\}; y \in \{1, \cdots, w\})$. With the deformation field $\varPhi$, we further improve I_{src}^{A} to $I_{src}^{\varPhi}$ through bilinear interpolation.

Loss & Training Strategy For learning of the challenging unsupervised deformation, we apply the local deformation cascade iteratively for $N_{\varPhi}$ times. Thus, the total loss is a weighted combination of one global similarity $\mathcal{L}_{global}$, local similarity $\mathcal{L}_{local}$ and regularization $\mathcal{L}_{reg}$ of $\varPhi_i$ for each step.

$$\mathcal{L}_{stage_I} = \mathcal{L}_{global}(I_{src}^{A}, I_{dst}) + \lambda_1 \sum_{i=1}^{N_{\varPhi}} [\mathcal{L}_{local}(I_{src}^{\varPhi_i}, I_{dst}) + \mathcal{L}_{reg}(\varPhi_i)] \tag{11}$$

$$\mathcal{L}_{reg}(\varPhi_i) = \mathcal{L}_{smooth}(\varPhi_i) + \lambda_2 \mathcal{L}_{inv}(\varPhi_i) + \lambda_3 \mathcal{L}_{syn} \tag{12}$$

Our ultimate goal is to find the discrete semantic correspondence between landmarks, instead of image registration. Similarity terms during the deformation process should gradually weaken impacts of appearance and focus more on structural information. We simply use $\mathcal{L}_1$ for $\mathcal{L}_{global}$. While for $\mathcal{L}_{local}$, we propose to enhance the original image similarity with an additional masked similarity term between their edge maps as follows:

$$\mathcal{L}_{esim} = \mathcal{L}_{sim}(I_{src}^{\varPhi}, I_{dst}) + \mathcal{L}_{sim}(\mathrm{edge}(I_{src}^{\varPhi}), \mathrm{edge}(I_{dst})) \odot \mathrm{Mask}(p_{src}^{\varPhi}) \tag{13}$$

$$\mathrm{Mask}(p) = \frac{1}{N} \sum_{i=1}^{N} \exp(-\frac{1}{2\sigma^2}||u - p_i||^2),\ p \in \mathbb{R}^{N\times 2},\ \forall u \in \varOmega \tag{14}$$

where edge$(\cdot)$ denotes the operator for edge detection (e.g., sobel) and Mask$(\cdot)$ generates the averaged gaussian heatmaps with fixed deviation σ, centered on landmarks p_{src}^{Φ}. $\mathcal{L}_{esim}$ considers the similarity between not only image pairs, but also their local edge maps around landmarks, which is beneficial to enforce the structural consistency of landmarks across images. We adopt the robust structural similarity (SSIM) [28] as $\mathcal{L}_{sim}$ in Eq. 13.

Besides, we observe that different anatomical parts tend to exhibit different displacements. Thus, the original smoothness constraint $\mathcal{L}_{smooth}$ (Eq. 15), which penalizes the approximated spatial gradients of Φ for all pixels, should be relaxed around the boundaries of these parts. We propose to re-weight $\mathcal{L}_{smooth}$ based on the magnitudes of detected edge vectors as in Eq. 16:

$$\mathcal{L}_{smooth}(\Phi) = ||\nabla\Phi(u)||^2, \ \forall u \in \Omega \tag{15}$$

$$\mathcal{L}_{esmooth}(\Phi|I_{src}^{\Phi}) = \mathcal{L}_{smooth}(\Phi) \odot \exp(-||\text{edge}(I_{src}^{\Phi})||^2/T) \tag{16}$$

where T is a hyper parameter to control the sharpness of distribution. In this way, weights for boundaries is lowered by their large magnitudes in edge maps and attenuates the mutual interference between different parts. As in [13], we also adopt $\mathcal{L}_{inv}$ to enforce the invertibility of Φ. For ease of early optimization, we introduce $\mathcal{L}_{syn}$, where we apply random perspective transform to I_{src} and use synthetic pairs with known correspondence to supervise the learning of Φ.

3.3 Stage II: Noisy Landmark Refinement

With the learned spatial correspondence from the exemplar (I_l, p_l) to I_u, we can infer pseudo landmarks $\tilde{p}_u$ for I_u. $h(\cdot)$ computes gaussian heatmaps with fixed standard deviation σ centered on landmark locations. Instead of simply train a new landmark detector with these pseudo-labels, as [11] did, we propose a robust learning scheme for noisy landmarks, coined as *Consistency Co-Teaching* (C2T).

As shown in Alg. 1, C2T builds on the seminal co-teaching framework [29]. Two networks f and g are co-trained to select relatively clean samples S_f and S_g (line 5-6) for the other network (line 7-9), leveraging the small-loss assumption [30,31,29]. C2T's main novelties are two ingredients: self-consistency filtering $\mathcal{L}_{filter}$ (Eq. 17-19) and cross-consistency loss $\mathcal{L}_{con}^{cross}$ (Eq. 20), which not only stabilize training, but also boost performance.

$$\mathcal{L}_{heat}(f, x, h(\tilde{p}_x)) = ||f(x) - h(\tilde{p}_x)||_2 \tag{17}$$

$$\mathcal{L}_{con}^{self}(f, x) = ||f(T_h(x)) - T_{e\rightarrow h}(f(T_e(x)))||_2 \tag{18}$$

$$\mathcal{L}_{filter}(f, x, h) = \mathcal{L}_{heat}(f, x, h) + w\mathcal{L}_{con}^{self}(f, x) \tag{19}$$

As pointed out in [32], in later training epochs, two networks trained by co-teaching could harmfully converge to a consensus. To keep f and g healthily diverged, we introduce $\mathcal{L}_{con}^{self}$(Eq. 18) into $\mathcal{L}_{filter}$. (T_e, T_h) is a pair of easy (weak augmentations) and hard (strong augmentations) views of the original images. Intuitively, if heatmaps predicted by f (g) for x are equivariant to different transformations [17], its pseudo label is more likely to be clean.

Algorithm 1 C2T: Consistency Co-Teaching

```
Input: $L : \{(I_l, H_l)\}, U : \{(I_u, \hat{H}_u)\}_{u=1}^{N_u}$
Parameter: $f, g$: landmark detectors; $\epsilon$: filter rate
Output: $f, g$: landmark detectors
Shuffle $L$ and $U$.
for $T = 1, \cdots, T_{max}$ do
    Fetch mini-batch $D_l, D_u$ from $L, U$ respectively.
    Sample random augmentations $T_e, T_h$ and compute transform $T_{e\to h}$.
    $S_g = \underset{D' \geq \epsilon|D_u|}{\arg\min} \sum_{(x,h)\in D'} L_{filter}(g, x, h), D' \subset D_u$.
    $S_f = \underset{D' \geq \epsilon|D_u|}{\arg\min} \sum_{(x,h)\in D'} L_{filter}(f, x, h), D' \subset D_u$.
    $\mathcal{L}_f^{update} = \sum_{(x,h)\in D_l \cup S_g} \mathcal{L}_{heat}(f, x, h) + \sum_{(x,h)\in D_l \cup D_u} \mathcal{L}_{con}^{cross}(f, g, x)$.
    $\mathcal{L}_g^{update} = \sum_{(x,h)\in D_l \cup S_f} \mathcal{L}_{heat}(g, x, h) + \sum_{(x,h)\in D_l \cup D_u)} \mathcal{L}_{con}^{cross}(g, f, x)$.
    Update $f$ by $\nabla_f \mathcal{L}_f^{update}$ and $g$ by $\nabla_g \mathcal{L}_g^{update}$.
end for
Return $f, g$
```

Instead of simply discarding filtered samples as in [29], we involve them along with selected samples, in $\mathcal{L}_{con}^{cross}$ (Eq. 20) to explore the consistency between

$$\mathcal{L}_{con}^{cross}(f, g, x) = ||f(T_h(x)) - T_{e\to h}(g(T_e(x)))||_2 \tag{20}$$

predictions of different models on different views. Motivated by semi-supervised methods [33], we utilize the confident predictions of each detector on easy views, to enforce consistency of the other detector on hard views. The overall loss function consists of the heatmap regression loss $\mathcal{L}_{heat}$ on the exemplar and selected samples, and the cross-consistency loss $\mathcal{L}_{con}^{cross}$ between the easy-hard pairs.

4 Experiments

We evaluate our method and state-of-the-art methods on multiple public X-ray datasets of different body parts. Furthermore, we conduct ablation studies to demonstrate how different components contribute to our final performance.

4.1 Dataset

Head: This dataset is a widely-used open-source dataset collected for IEEE ISBI 2015 challenge [34,1], which consists of 400 cephalometric radiographs. Two medical experts annotate 19 landmarks manually and we compute their average as the ground truth like [5,11]. The entire dataset is officially split into three parts: the first 150 images are training set, the next 250 images are test set. The original resolution is 2400 × 1935 and the pixel spacing is 0.1 mm.

Hand: This dataset is a public dataset of hand radiographs collected by [35]. [5] further annotate 37 landmarks on fingertips and bone joints. They assume the length between two endpoints of the wrist is 50 mm. The whole dataset is split into a training set of 609 images and a test set of 300 images as in [36].

Chest: This dataset is from a Kaggle challenge. [36] select a subset of 279 images by excluding abnormal cases and annotate six landmarks at the boundaries of the lung. Following [36], we use pixel distance at fixed resolution (512×512) for evaluation, since no pixel spacing information is provided.

4.2 Implementation

For both stages, the input resolution for head and hand datasets is 320×256, while chest images are resized into 256×256 to keep the aspect ratio. Data augmentations used in both stages include random rotation and random scaling. Random horizontal flipping is only applied in stage II. β_1, β_2 for Adam optimizer is set to $0.99, 0.0$ respectively and weight decay is 1e-4.

For stage I, we adopt pretrained HRNet18 [7] as backbone. The channel dimension of extracted feature maps are adjusted to $d_{model} = 64$ through 1×1 convolution, as the input dimension for attention blocks. Self (cross) attention modules are implemented as the encoder (decoder) block designed in [15], with multi-head attention ($N_h = 2$) and dropout $p = 0.1$. For each step of deformation, we stack $N_l = 2$ attention blocks for feature fusion. For alignment estimation, we set sf_x, sf_y, rot, sh to $1, 1, \frac{\pi}{2}, \frac{\pi}{2}$ respectively. We set $\sigma = 3$ for heatmap generation in Eq. 13 and $T = 0.1$ in Eq. 16.

We train the model for 750 epochs. For the first 250 epochs, learning rate lr is fixed as 1e-4 and ramps the weight for local deformation λ_1 (Eq. 11) from 0 to 1. For the remaining epochs, lr is decayed to 5e-5 with the cosine annealing strategy. In Eq. 12, λ_2 is set as 0.25 and λ_3 cosinely ramps down from 5.0 to 0.0. For each mini-batch, half of them are synthesized source-target pairs with known pixel correspondence, while the others are obtained by shuffling images within the current batch. In consideration of computation overhead and memory, all models are trained with one global step and two subsequent local steps ($N_{\Phi} = 2$). We start to infer pseudo labels for the training set after 200 epochs and compute its exponential moving average with a τ of 0.9 in Eq. 4.

For stage II, we retrain two HRNet18 [7] detectors, both initialized from pre-trained weights. We set $\sigma = 3$ for the standard deviation of the gaussian heatmap. Initial lr is 1e-3 and decays by 0.1 at 60 epochs and 80 epochs, until a total of 100 epochs. In the first 30 epochs, filter rate ϵ ramps from 0.0 to 0.8.

4.3 Evaluation

Metrics Mean radial error (MRE) and successful detection rate (SDR) are adopted as evaluation metrics. MRE computes the average Euclidean distances

Table 1. Evaluation on the head, hand and chest dataset. * denotes original reported results in paper. # denotes reproduced performances under the one-shot landmark setting. Supervised method YOLO [36] serves as an universal upper bound. The best results are in **bold** and the second best results are underlined. Strictly using the same exemplar, our result outperforms all other one-shot methods. Performance of stage I is reported for fair comparison with CC2D-SSL

	Head					**Hand**				**Chest**			
Method	MRE↓	SDR↑(%)				MRE↓	SDR↑(%)			MRE↓	SDR↑(%)		
	(mm)	2mm	2.5mm	3mm	4mm	(mm)	2mm	4mm	10mm	(px)	3px	6px	9px
YOLO∗	1.54	77.79	84.65	89.41	94.93	0.84	95.4	99.35	99.75	5.57	57.33	82.67	89.33
3FabRec#	20.12	2.42	3.86	4.98	7.23	9.81	3.98	15.24	60.92	48.67	0.67	2.33	4.67
DataAug#	3.18	32.81	44.42	55.12	73.16	2.51	48.87	85.67	98.91	10.15	15.67	40.67	61.67
CC2D-SSL#	3.41	40.63	49.58	60.31	72.14	2.93	51.59	81.29	95.59	17.37	9.87	27.99	42.11
CC2D-TPL#	2.72	42.59	53.18	66.48	83.22	2.47	54.95	87.16	97.84	12.91	12.67	38.67	57.67
Ours-stage I	2.70	42.78	54.88	65.03	81.01	2.13	60.93	89.43	99.21	10.16	12.33	39.00	60.33
Ours-stage II	**2.13**	**54.69**	**67.47**	**77.85**	**90.02**	**1.82**	**66.39**	**92.93**	**99.97**	**6.89**	**17.33**	**50.33**	**75.33**

between predicted landmarks and ground truth landmarks. Given several thresholds, SDR calculates the proportion of predictions with an error below these thresholds respectively. The unit of MRE is mm if pixel spacing is provided. Otherwise, it is reported as raw pixel distance. We use the same thresholds for SDR as [36], where they developed a fully-supervised universal landmark detector trained on the mix of all three datasets aforementioned. We list their results in Tab. 1 as the supervised upper bound.

4.4 Comparison with Baseline Methods

As shown in Tab. 1, we compare with 3FabRec, DataAug and CC2D on all three datasets. Our results outperform other one-shot methods by notable margins, which demonstrates both effectiveness and general applicability of our method.

As the weak baseline, with only access to hundreds of images, 3FabRec can hardly reconstruct the fine-grained details in X-ray images and thus yield relatively poor results. Our method also shows consistent improvements against the strong baseline DataAug, which relies on the affine alignment in the preprocessing step. While for some cases with drastic changes in spatial structure as in Fig. 3, our end-to-end learned alignment is more beneficial for the subsequent local deformations, since the total transformation learning is extraly guided by our carefully-designed constraints for landmarks.

Compared to the state-of-the-art method CC2D, we achieve superior performance by decreasing the MRE of stage I and stage II by 20.8% (3.41mm→2.70 mm) and 21.7% (2.72mm→2.13mm) respectively. Improvements over CC2D might be attributed to the following two aspects. First, the pretext task of image-patch matching used in [11], is prone to overfitting when having difficulties discriminating local regions centered on those densely-labeled landmarks. Besides, they do not consider the global spatial relationships among these patches. In contrast, our way of inferring landmarks by registration, implicitly takes

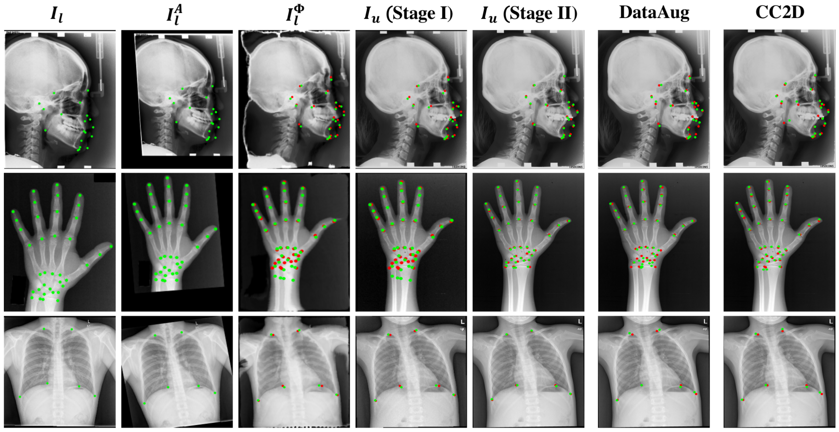

Fig. 3. Visualization of learned transformations and comparison with baseline methods. From left to right, we display the exemplar (I_{src}), intermediate warped results (I_l^A, I_{src}^Φ) and the unlabeled target image I_u. **Green dots** denote locations of the transformed exemplar landmarks. **Red dots** denote the ground truth landmark locations of I_u. Most of the stage I predictions with large biases could be corrected in stage II

such constraints into consideration, and thus naturally avoid abnormal predictions, which can be justified by the obvious increase of SDR@4mm in stage I (72.14% → 81.01%). Second, compared to simply retraining with synthesized samples as in DataAug, or performing majority-voting as in CC2D, our C2T makes better use of the common landmark knowledge contained in pseudo labels. Specifically, reliable labels can be effectively selected to provide more supervision for heatmap regression. On the other hand, remaining instances are not wasted by contributing to the consistency learning. It is especially crucial to handle challenging cases, such as the second row in Fig. 3, which successfully correct the large biases introduced in stage I, for landmarks around wrists.

Table 2. Ablation of spatial transform. "L" denotes the alignment is learned

Global	**Local**	**Head**				
Alignment	**Step**	MRE↓	SDR↑(%)			
Type	N_Φ	(mm)	2mm	2.5mm	3mm	4mm
✗	2	3.42	30.78	41.96	52.02	69.36
sift	2	3.10	35.31	47.58	58.23	74.02
affine (L)	1	2.73	39.64	52.19	63.45	80.50
affine (L)	2	2.70	42.78	54.88	65.03	81.01
perspective (L)	2	2.80	40.17	52.48	63.41	79.35

Table 3. Configurations of stage I network

			Head				
backbone	N_l	N_h	MRE↓	SDR↑(%)			
			(mm)	2mm	2.5mm	3mm	4mm
hrnet	0	0	2.93	36.76	48.36	59.58	76.11
	1	1	2.85	37.37	50.08	60.67	78.36
	2	1	2.82	39.87	52.61	63.47	78.25
	2	2	2.70	42.78	54.88	65.03	81.01
unet	2	2	2.95	34.86	47.81	58.63	76.42

4.5 Ablation Study

Here we conduct experiments on the head dataset, to analyze the contributions of different components from multiple aspects including spatial transformation, network architecture, loss function and robustness to exemplar selection.

Spatial Transformation We remove the global alignment in the first row of Tab. 2 and witness a great drop in our performance. Besides, compared to alignment computed by traditional methods (e.g., sift detector), learned alignment is more desirable to reduce the complexity for subsequent local deformations. Compared to single step of local deformation, one more step benefits our performance, decreasing MRE from 2.73mm to 2.70mm. This observation is well-aligned with [37]. We also try to learn perspective transform with more degrees of freedom, but do not observe further improvements as expected. We argue that this might be restricted by the complexity of the head dataset.

Stage I Network Configurations We remove all attention blocks in the first row of Tab. 3 and achieve a MRE of 2.93mm. Furnishing backbone with more layers of attention blocks ($N_l = 0 \rightarrow 2$) and multi-head attention ($N_h = 1 \rightarrow 2$) contributes positively to our final performance. Our method still achieves competitive performance if we switch our backbone to the commonly-used UNet [38].

Table 4. Ablation of loss function for stage I

Loss	**Head**				
	MRE↓	SDR↑(%)			
	(mm)	2mm	2.5mm	3mm	4mm
w/o $\mathcal{L}_{inv}$	3.24	32.67	44.95	55.52	71.81
w/o $\mathcal{L}_{smooth}$	2.97	39.01	50.23	60.46	75.85
w/o $\mathcal{L}_{syn}$	2.86	39.54	51.43	61.68	77.14
$\mathcal{L}_{esim} \rightarrow \mathcal{L}_{sim}$	3.17	33.31	45.64	57.31	73.01
$\mathcal{L}_{esmooth} \rightarrow \mathcal{L}_{smooth}$	2.75	40.27	52.95	64.65	81.12
ours	2.70	42.78	54.88	65.03	81.01

Table 5. Ablation for stage II

$\mathcal{L}_{filter}$	$\mathcal{L}_{con}^{cross}$	**Head**				
		MRE↓	SDR↑(%)			
		(mm)	2mm	2.5mm	3mm	4mm
✗	✗	2.53	44.38	57.37	67.52	83.43
w $\mathcal{L}_{con}^{self}$	✗	2.29	50.80	63.64	74.36	87.53
w/o $\mathcal{L}_{con}^{self}$	✓	2.17	53.81	67.18	77.43	89.70
w $\mathcal{L}_{con}^{self}$	✓	2.13	54.69	67.47	77.85	90.02

Loss Function for StageI In the first three rows of Tab. 4, we remove one regularization term of Φ at a time and find they all contribute to our results. They require the deformation field to be smooth, reversible, and applicable to real images respectively, thus avoiding intermediate abnormal warping results.

To study the contributions of $\mathcal{L}_{esim}$, we replace it with $\mathcal{L}_{sim}$ and find that MRE increases greatly by 17.4%. $\mathcal{L}_{sim}$ only considers image similarity based on pixel values, while our $\mathcal{L}_{esim}$ further incorporates the local structural similarity around landmarks based on edge maps, which is crucial to enforce the consistency across instances and thus adapts better to landmark localization

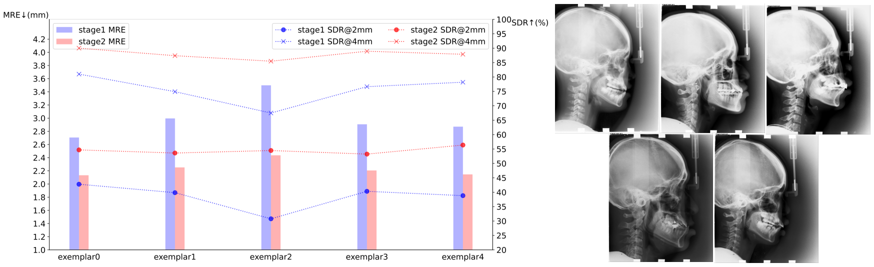

Fig. 4. Experiments for five candidate exemplars on the head dataset

task. Compared to $\mathcal{L}_{esmooth}$, $\mathcal{L}_{smooth}$ also leads to slight degradation in performance. Taking landmarks annotated near boundaries between foreground and background for example, they might be biased towards background due to the strict smoothness enforced by $\mathcal{L}_{smooth}$. $\mathcal{L}_{esmooth}$ is tolerant to discontinuity between different regions and thus mitigates such phenomenon.

Ablation for Stage II As in Tab. 5, we achieve a MRE of 2.53mm by simply retraining with pseudo labels in stage I. Filtering out noisy pseudo labels by heatmap loss $\mathcal{L}_{heat}$ can improve MRE to 2.29mm. And $\mathcal{L}_{con}^{cross}$ involves these filtered samples for semi-supervised learning, which further decreases MRE to 2.17mm. If $\mathcal{L}_{con}^{self}$ is also taken into consideration for filtering, selected pseudo labels can provide more reliable supervision since they are also robust against spatial transformations, which enables us to achieve the final MRE of 2.13mm.

Exemplar Selection In Fig. 4, we show five different candidate exemplars and their corresponding results on the head dataset. Although there might be certain variations in the performances of stage I, our retraining scheme introduced in stage II effectively stabilizes the final performance, with a mean MRE of 2.25mm.

5 Conclusion

We propose a novel two-stage framework for one-shot medical landmark localization. In stage I, an image transform model is learned through unsupervised registration. The total transform is decomposed to an end-to-end cascade of global alignment and local deformations, with the guidance of novel loss functions incorporating edge information. Pseudo landmarks are inferred on unlabeled targets with the exemplar and learned transform model. In stage II, we use these noisy labels to train robust landmark detectors by exploring self-consistency for selecting reliable samples and cross-consistency for semi-supervised learning. Extensive experiments on multiple datasets demonstrate our method surpasses other one-shot methods and further narrows the gap with supervised methods.

Acknowledgements This work is supported in part by MOST-2018AAA010200 4 and NSFC-62061136001.

References

1. Ching-Wei Wang, Cheng-Ta Huang, Jia-Hong Lee, Chung-Hsing Li, Sheng-Wei Chang, Ming-Jhih Siao, Tat-Ming Lai, Bulat Ibragimov, Tomaž Vrtovec, Olaf Ronneberger, et al. A benchmark for comparison of dental radiography analysis algorithms. *Medical image analysis*, 31:63–76, 2016.
2. Runnan Chen, Yuexin Ma, Nenglun Chen, Daniel Lee, and Wenping Wang. Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 873–881. Springer, 2019.
3. María Escobar, Cristina González, Felipe Torres, Laura Daza, Gustavo Triana, and Pablo Arbeláez. Hand pose estimation for pediatric bone age assessment. In *International conference on medical image computing and computer-assisted intervention*, pages 531–539. Springer, 2019.
4. Ping Gong, Zihao Yin, Yizhou Wang, and Yizhou Yu. Towards robust bone age assessment: Rethinking label noise and ambiguity. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 621–630. Springer, 2020.
5. Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Integrating spatial configuration into heatmap regression based cnns for landmark localization. *Medical image analysis*, 54:207–219, 2019.
6. Wei Liu, Yu Wang, Tao Jiang, Ying Chi, Lei Zhang, and Xian-Sheng Hua. Landmarks detection with anatomical constraints for total hip arthroplasty preoperative measurements. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 670–679. Springer, 2020.
7. Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.
8. Weijian Li, Yuhang Lu, Kang Zheng, Haofu Liao, Chihung Lin, Jiebo Luo, Chi-Tung Cheng, Jing Xiao, Le Lu, Chang-Fu Kuo, et al. Structured landmark detection via topology-adapting deep graph learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 266–283. Springer, 2020.
9. Huajun Liu, Fuqiang Liu, Xinyi Fan, and Dong Huang. Polarized self-attention: Towards high-quality pixel-wise regression. *arXiv preprint arXiv:2107.00782*, 2021.
10. Bjorn Browatzki and Christian Wallraven. 3fabrec: Fast few-shot face alignment by reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6110–6120, 2020.
11. Qingsong Yao, Quan Quan, Li Xiao, and S Kevin Zhou. One-shot medical landmark detection. *arXiv preprint arXiv:2103.04527*, 2021.
12. Xiao-Yun Zhou, Bolin Lai, Weijian Li, Yirui Wang, Kang Zheng, Fakai Wang, Chihung Lin, Le Lu, Lingyun Huang, Mei Han, et al. Scalable semi-supervised landmark localization for x-ray images using few-shot deep adaptive graph. *arXiv preprint arXiv:2104.14629*, 2021.
13. Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018.

14. Matthew CH Lee, Ozan Oktay, Andreas Schuh, Michiel Schaap, and Ben Glocker. Image-and-spatial transformer networks for structure-guided image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–345. Springer, 2019.
15. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
16. Qingsong Yao, Zecheng He, Hu Han, and S Kevin Zhou. Miss the point: Targeted adversarial attack on multiple landmark detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 692–702. Springer, 2020.
17. Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2018.
18. Olga Moskvyak, Frederic Maire, Feras Dayoub, and Mahsa Baktashmotlagh. Semi-supervised keypoint localization. *arXiv preprint arXiv:2101.07988*, 2021.
19. Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10153–10163, 2019.
20. Amit Kumar and Rama Chellappa. S2ld: Semi-supervised landmark detection in low-resolution images and impact on face verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 758–759, 2020.
21. Hristina Uzunova, Matthias Wilms, Heinz Handels, and Jan Ehrhardt. Training cnns for image registration from few samples with model-based data augmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 223–231. Springer, 2017.
22. Hongming Li and Yong 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)*, pages 1075–1078. IEEE, 2018.
23. Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8553, 2019.
24. Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.
25. Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018.
26. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.
27. Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.
28. Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

29. Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *arXiv preprint arXiv:1804.06872*, 2018.
30. Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *ArXiv*, abs/1611.03530, 2017.
31. Devansh Arpit, Stanislaw Jastrzkebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pages 233–242. PMLR, 2017.
32. Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pages 7164–7173, 2019.
33. Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.
34. Ching-Wei Wang, Cheng-Ta Huang, Meng-Che Hsieh, Chung-Hsing Li, Sheng-Wei Chang, Wei-Cheng Li, Rémy Vandaele, Raphaël Marée, Sébastien Jodogne, Pierre Geurts, et al. Evaluation and comparison of anatomical landmark detection methods for cephalometric x-ray images: a grand challenge. *IEEE transactions on medical imaging*, 34(9):1890–1900, 2015.
35. Arkadiusz Gertych, Aifeng Zhang, James Sayre, Sylwia Pospiech-Kurkowska, and HK Huang. Bone age assessment of children using a digital hand atlas. *Computerized medical imaging and graphics*, 31(4-5):322–331, 2007.
36. Heqin Zhu, Qingsong Yao, Li Xiao, and S Kevin Zhou. You only learn once: Universal anatomical landmark detection. *arXiv preprint arXiv:2103.04657*, 2021.
37. Shengyu Zhao, Yue Dong, Eric I Chang, Yan Xu, et al. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10600–10610, 2019.
38. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.