UOD: Universal One-Shot Detection of Anatomical Landmarks

Heqin Zhu[1,2,3], Quan Quan[3], Qingsong Yao[3], Zaiyi Liu[4], S. Kevin Zhou[1,2]✉

1. School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, 230026, P.R.China
2. Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu, 215123, P.R.China
3. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
4. Department of Radiology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China

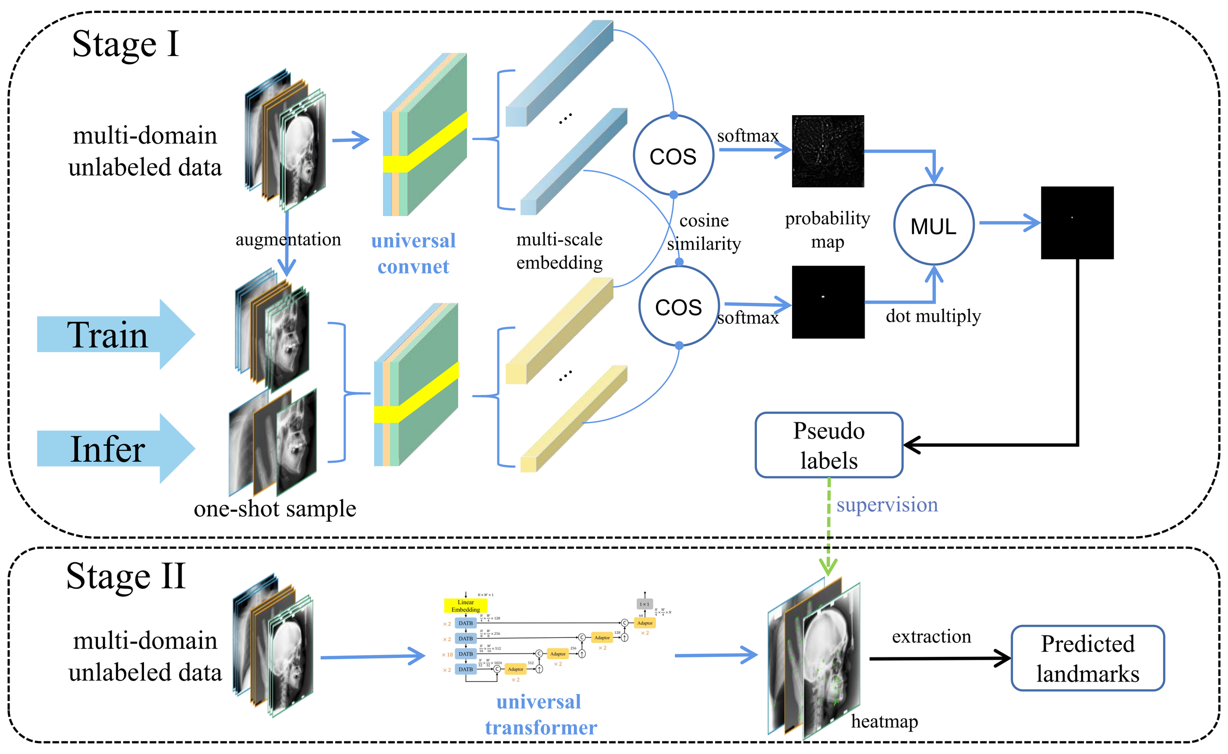
Abstract

One-shot learning is not robust that the performances vary a lot when annotating different sample image. Existing one-shot methods are single-domain specialized and suffer domain preference when handling muti-domain unlabeled data. To tackle these issues, we developed **UOD, a Universal One-shot landmark Detection framework**. UOD consists of two stages and two corresponding universal models. In stage I, a domain-adaptive convolution model is self-supervised learned to generate pseudo landmark labels. In stage II, we design a domain-adaptive transformer to eliminate domain preference and build the global context for multi-domain data. UOD is evaluated on three X-ray datasets in different anatomical domains (i.e., head, hand, chest) and obtained state-of-the-art performances in each domain. The code is available at https://github.com/heqin-zhu/UOD_universal_oneshot_detection

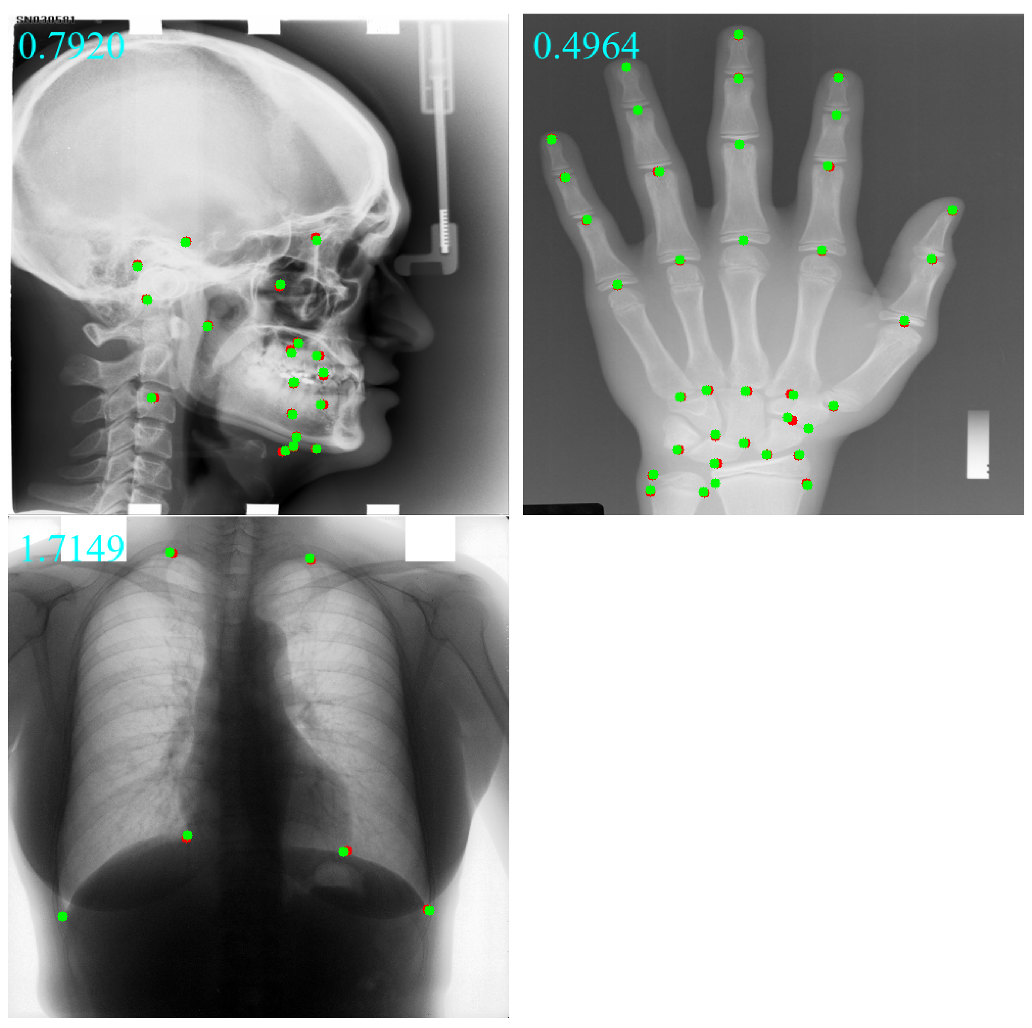
Introduction

UOD framework consists of two stages: 1) Contrastive learning for training a universal model with multi-domain data to generate pseudo landmark labels. 2) Supervised learning for training domain-adaptive transformer (DATR) to avoid domain preference and detect robust and accurate landmarks. A universal model is comprised of domain-specific modules and domain-shared modules, learning the specified features of each domain and common features of all domains to eliminate domain preference and extract representative features for multi-domain data. Moreover, multi-domain one-shot learning reaps benefit from different one-shot samples from various domains, in which cross-domain features are excavated by domain-shared modules.

Our contributions are as follows: **1)** We design **the first universal framework for multi-domain one-shot landmark detection**, which improves detecting accuracy and relieves domain preference on multi-domain data from various anatomical regions. **2)** We design a **domain-adaptive transformer block (DATB)**, which is effective for multi-domain learning and can be used in any other transformer network. **3)** We carry out comprehensive experiments to demonstrate the effectiveness of UOD for obtaining **SOTA performance** on three publicly used X-ray datasets of head, hand, and chest.



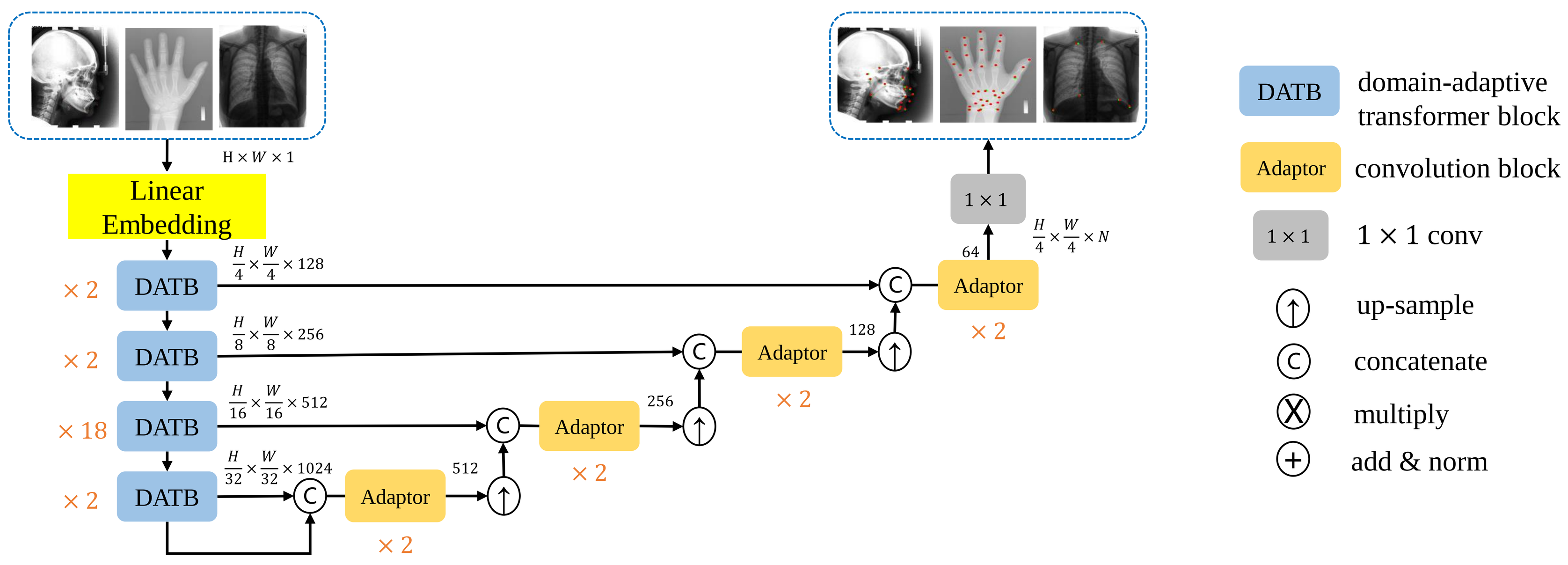
Visualization



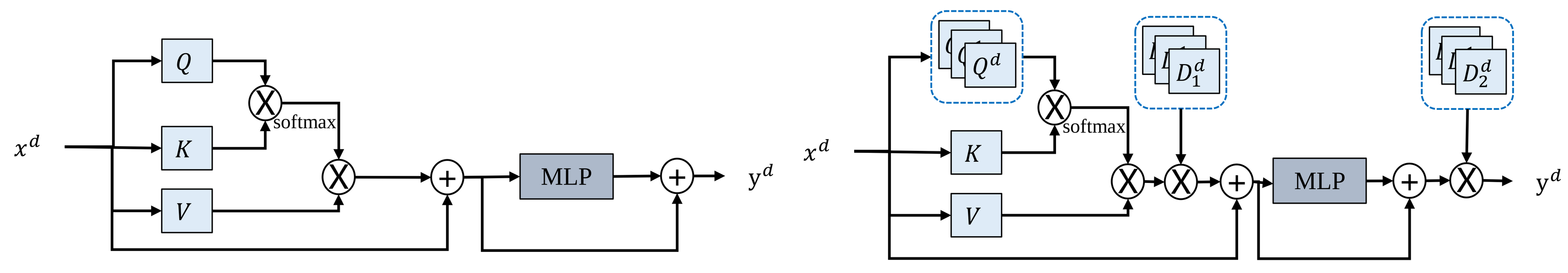
Methodology

Stage I: Contrastive learning Following Yao et al. [1], we employ contrastive learning to train siamese network for matching similar patches of original image and augmented image. Given a multi-domain input image $X^d \in R^{H^d \times W^d \times C^d}$ belongs to domain d from multi-domain data, we randomly select a target point P and crop a half-size patch X_p^d which contains P. After applying data augmentation on X_p^d, the target point is mapped to P_p. Then we feed X^d and X_p^d into the siamese network respectively and obtain the multi-scale feature embeddings. We compute cosine similarity of two feature embeddings from each scale and apply softmax to the cosine similarity map to generate a probability matrix. Finally, we calculate the cross entropy loss of the probability matrix and ground truth map which is produced with the one-hot encoding of P_p^d to optimize the siamese network for learning the latent similarities of patches. At inferring stage, we replace augmented patch X_p^d with the augmented one-shot sample patch X_s^d. We use the annotated one-shot landmarks as target points to formulate the ground truth maps. After obtaining probability matrices, we apply arg max to extract the strongest response points as the pseudo landmarks, which will be used in UOD Stage II.

Stage II: Supervised learning In stage II, we design a universal transformer to capture global relationship of multi-domain data and train it with the pseudo landmarks generated in stage I. The universal transformer has a domain-adaptive transformer encoder and domain-adaptive convolution decoder. The decoder is based on a U-Net [2] decoder with each standard convolution replaced by a domain adaptor [3]. The encoder is based on Swin Transformer [4] with shifted window and limited self-attention within non-overlapping local windows for computation efficiency. Different from Swin Transformer [4], we design a domain-adaptive transformer block (DATB) and use it to replace the original transformer block. DATB consists of domain-specific and domain-shared parameters in DATB. We duplicate the query matrix for each domain to learn domain-specific query features and keep key and value matrix domain-shared to learn common knowledge and reduce parameters. We further adopt learnable diagonal matrix after each MSA and MLP module to facilitate the learning of domain-specific features, which costs few parameters ($O(N)$ for $N \times N$ diagonal).



(a) DATR: Domain-adaptive Transformer



(b) Basic transformer block

(c) DATB: Domain-adaptive transformer block

Experimental Results

Method	Label	Head MRE↓ (mm)	Head SDR↑ (%) 2mm	2.5mm	3mm	4mm	Hand MRE↓ (mm)	Hand SDR↑ (%) 2mm	4mm	10mm	Chest MRE↓ (mm)	Chest SDR↑ (%) 2mm	4mm	10mm
YOLO [2]†	all	1.32	81.14	87.85	92.12	96.80	0.85	94.93	99.14	99.67	4.65	31.00	69.00	93.67
YOLO [2]†	25	1.96	62.05	77.68	88.21	97.11	2.88	72.71	92.32	97.65	7.03	19.33	51.67	89.33
YOLO [2]†	10	2.69	47.58	66.47	78.42	90.89	9.70	48.66	76.69	90.52	16.07	11.67	33.67	76.33
YOLO [2]†	5	5.40	26.16	41.32	54.42	73.74	24.35	20.59	48.91	72.94	34.81	4.33	19.00	56.67
CC2D [1]*	1	2.76	42.36	51.82	64.02	78.96	2.65	51.19	82.56	95.62	10.25	11.37	35.73	68.14
Ours†	1	**2.43**	**51.14**	**62.37**	**74.40**	**86.49**	**2.52**	**53.37**	**84.27**	**97.59**	**8.49**	**14.00**	**39.33**	**76.33**

References

[1] Qingsong Yao et al. "One-shot medical landmark detection". In: *MICCAI*. Springer. 2021, pp. 177–188.
[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *MICCAI*. Springer. 2015, pp. 234–241.
[3] Chao Huang et al. "3D U2Net: A 3D Universal U-Net for Multi-domain Medical Image Segmentation". In: *MICCAI*. Springer. 2019, pp. 291–299.
[4] Ze Liu et al. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *ICCV*. 2021, pp. 10012–10022.
[5] Heqin Zhu et al. "You only learn once: Universal anatomical landmark detection". In: *MICCAI*. Springer. 2021, pp. 85–95.