IdFOPNet: Integrating Identity Attention and Fairness Optimization in Anatomical Landmark Detection

Dexin Zhao[1,2] DEXINZHAO@MAIL.USTC.EDU.CN
Hongbo Ye*[1,2] HONGBOYE@MAIL.USTC.EDU.CN
Minghao Bian*[1,2] BIANMINGHAO@MAIL.USTC.EDU.CN
Yuxin Wu[1,2] WU_YUXIN@MAIL.USTC.EDU.CN
S. Kevin Zhou[1,2,3,4] SKEVINZHOU@USTC.EDU.CN

[1] *School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China (USTC), Hefei Anhui, 230026, China*

[2] *Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu, 215123, P.R. China*

[3] *State Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, China*

[4] *Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China*

Editors: Under Review for MIDL 2025

Abstract

Fairness is the key to addressing bias in anatomical landmark point detection. Existing methods tend to ignore individual identity information, resulting in significant bias exhibited between different age groups and genders. Moreover, current studies mainly focus on improving the accuracy of models and lack the dynamic optimisation mechanism of fairness, resulting in the bias problem not being solved effectively. To this end, we propose an anatomical landmark detection network that integrates **Id**entity features and **F**airness **OP**timization (IdFOPNet). This method leverages a prototype network to detect landmarks by comparing image features with a set of global landmark prototypes. To enhance model fairness, we introduce the Identity Attention mechanism, incorporating identity information as prior knowledge into the detection process. Additionally, we design a penalty-based gradient modulation strategy to dynamically suppress the model's over-reliance on specific biased information during training. We evaluate the IdFOPNet on the CephAdoAdu and Hand X-Rays datasets. Extensive experimental results demonstrate that our method outperforms SOTA approaches in anatomical landmark detection across different ages and genders, and stays fair as well.

Keywords: Anatomical landmark detection, Identity attention mechanism, Fairness optimization, Penalty-based gradient modulation.

1. Introduction

Anatomical landmark detection (ALD), a fundamental task in medical image analysis widely used in clinical diagnosis, surgical planning, and treatment evaluation (Zhou et al., 2019; Chiras et al., 1997), provides reliable anatomical references for clinicians, facilitating automated diagnosis and personalized treatment (Wang et al., 2015).Recent advances in deep

* These authors contributed equally as co-second authors

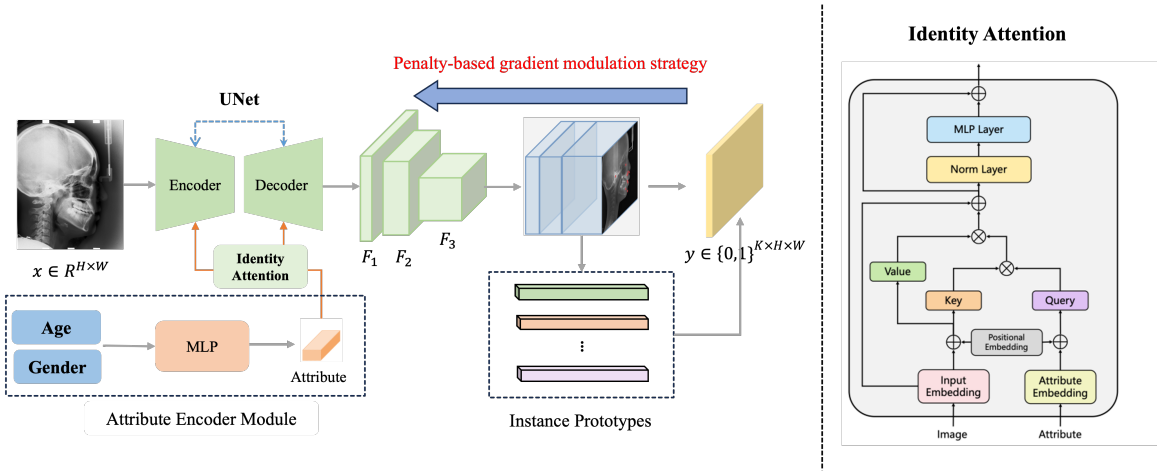

Figure 1: The architecture of IdFOPNet. The network integrates identity features and fairness optimization to address biases in anatomical landmark detection.

learning have significantly improved ALD performance, with methods such as convolutional neural networks for landmark regression (Oh et al., 2020), two-stage networks for enhanced detection (Jiang et al., 2022), and models incorporating anatomical prior knowledge (Zhou et al., 2024). However, these methods often exhibit inherent biases when applied to diverse populations, such as children or female patients, due to significant anatomical variations (Tian et al., 2025). Addressing these biases, that is, ensuring fairness in ALD, is crucial for reliable clinical applications.

In recent years, researchers have developed methods (Luo et al., 2024a,b) to improve the fairness of deep learning models and mitigate algorithmic bias in medical imaging. These methods largely rely on an implicit assumption that training and testing data distributions remain consistent. However, this assumption often fails in real medical scenarios. For example, when models trained on one type of imaging data are directly deployed on another, performance and fairness often deteriorate significantly due to domain-shift issues. To address these domain shift challenges, domain adaptation (Zhao et al., 2018) and domain generalization (Qiao et al., 2020) have emerged as two core methodological approaches in recent research. Domain adaptation, particularly unsupervised domain adaptation, improves model generalization to new domains by using labeled source domain data and unlabeled target domain data. In contrast, domain generalization assumes that target domain data are unavailable during training and relies solely on source domain data to enhance model applicability in unseen domains (Li et al., 2020). Although these methods have significantly improved cross-domain accuracy, they largely overlook a critical fairness issue: ensuring fair predictions across different population groups during domain transfer to avoid potential discrimination.

In the field of medical imaging landmark detection, we are the first to conduct research on fairness and bias mitigation. Therefore, we analyze the main problems regarding fairness in ALD: ***Problem 1:*** *Individual identity information.* Existing methods typically overlook individual identity information, which can provide powerful prior knowledge for

model learning. For example, in cephalometric analysis, adult subjects generally exhibit clear cranial structures and regular tooth alignment. However, adolescents' anatomical structures often show complex and dynamic morphological changes due to unerupted teeth and mixed dentition states. These significant individual differences can lead to substantial landmark detection bias, especially in cross-population scenarios. Moreover, ignoring identity information makes it difficult to effectively capture these subtle but important feature differences, limiting the model's generalization capability. ***Problem 2****: Fairness optimization mechanism.* An intuitive prior is that when models exhibit bias across different populations, they should be able to identify and intervene in a timely manner, thus guiding the model back to a reliability track. However, current research focuses more on improving overall model accuracy while neglecting the dynamic optimization mechanism for fairness. This amplifies bias issues during the training phase, compromising fairness.

To this end, we propose an ALD method that integrates **Id**entity features and **F**airness **OP**timization (IdFOPNet). The overall approach is illustrated in Figure 1. This method employs a prototype network that performs landmark detection by comparing image features with a set of global landmark prototypes. The network precisely aligns landmark prototypes with input features through an attention mechanism and extracts global landmark prototypes from numerous training samples. Additionally, a mask-based mining strategy is adopted to explore anatomical relationships between landmarks, further improving detection accuracy. To enhance model fairness, we introduce an Identity Attention mechanism (***Address Problem 1***) that injects identity information as prior knowledge into the landmark detection process. This module effectively models feature differences across diverse populations, thereby improving the model's applicability to diverse groups and mitigating bias issues stemming from individual differences. During the learning process, with our preposed penalty-based gradient modulation strategy (***Address Problem 2***) to dynamically suppress the model's over-reliance on specific bias information, the model can actively correct potential biases during the training phase. We evaluate our proposed method on two datasets: CephAdoAdu (Wu et al., 2024) and Hand X-Rays (Payer et al., 2019). Furthermore, we introduce a novel fairness metric specifically tailored to ALD. Extensive experimental results demonstrate that IdFOPNet outperforms existing state-of-the-art (SOTA) approaches in ALD across different age groups and genders, achieving fairer predictive performance across different populations and domains.

2. Methodology

In this paper, the fairness-based ALD task primarily involves datasets consisting of image-label pairs, denoted as (x, y), where $x \in \mathbb{R}^{H \times W}$ represents an anatomical localization image of size $H \times W$, and $y \in \{0,1\}^{K \times H \times W}$ represents K binary ground truth landmark maps. In each landmark map, only one position is annotated as a landmark point, satisfying $\sum y_{k,:,:} = 1$. Following existing methods, we convert the sparsely distributed landmark maps into K landmark heatmaps $H_k = \text{Gaussian}\,(y_k) \in \mathbb{R}^{H \times W}$ for model training, and apply a Gaussian smoothing strategy to process the heatmaps.

2.1. Overview

The overall pipeline of IdFOPNet is illustrated as shown in Figure 1. For the input anatomical localization image x, we employ U-Net (Ronneberger et al., 2015) as the network backbone f_θ to extract multi-level high-resolution feature maps $\{F_1, F_2, F_3\}$, where $F_1 \in \mathbb{R}^{H/4\times W/4\times D_1}$, $F_2 \in \mathbb{R}^{H/2\times W/2\times D_2}$, and $F_3 \in \mathbb{R}^{H\times W\times D_3}$. To precisely detect sparsely distributed landmarks, these feature maps are upsampled to the original resolution of the input image and concatenated to form a composite feature map $F = \text{concat}\left(\text{up}\left(F_1\right), \text{up}\left(F_2\right), \text{up}\left(F_3\right)\right) \in \mathbb{R}^{H\times W\times D}$, where $D = D_1 + D_2 + D_3$, and $\text{up}(\cdot)$ denotes the upsampling operation. To enhance model fairness, we introduce an Identity Attention mechanism that injects identity information as prior knowledge into the landmark detection process (refer to Section 2.2 for more details).

To capture landmark features more effectively, following (Wang et al., 2023; Wu et al., 2024), we create instance-level landmark prototypes $P_{\text{ins}} = \{p_k^{ins}\}_{k=1}^K$ for each training image x, where $p_k^{ins} \in \mathbb{R}^{1\times 1\times D}$ is computed as:

$$p_k^{ins} = \frac{\sum_{i,j} H_k\left(i,j\right) \cdot F\left(i,j\right)}{\sum_{i,j} H_k\left(i,j\right)}, \tag{1}$$

However, instance-level prototypes only consider single image information(Yao et al., 2021). Therefore, we estimate holistic prototypes $P_{\text{hol}} = \{p_k^{hol}\}_{k=1}^K$ in real-time from numerous training samples by minimizing the differences between instance-level prototypes:

$$\mathcal{L}_{\text{align}} = \frac{1}{K}\sum_{k=1}^{K} \|p_{m,k}^{ins} - p_{n,k}^{ins}\|_2^2, \tag{2}$$

where $p_{m,k}^{ins}$ and $p_{n,k}^{ins}$ represent the k-th instance-level prototypes for images x_m and x_n in mini-batch B, respectively. To utilize anatomical dependencies between landmarks, a prototype relationship mining method based on masked instance prototypes is applied. After generating instance prototypes P_{ins}, some prototypes are randomly masked and replaced with zeros, and landmark position embeddings are introduced as position indicators. The reconstruction process is supervised by the original prototypes:

$$\mathcal{L}_{\text{mine}} = \sum_{k=1}^{K} \|\hat{p}_k^{ins} - p_k^{ins}\|_2^2, \tag{3}$$

where $\hat{p}_k^{ins}$ is a reconstructed prototype in $\hat{P}_{\text{ins}}$, and p_k^{ins} is the corresponding original prototype in P_{ins}. Here, $\hat{P}_{\text{ins}} = \text{MSA}(\text{mask}(P_{\text{ins}}) \bigoplus \text{MLP}(\bar{y}))$, $\bar{y} \in \mathbb{R}^{K\times 2}$. By computing the dot product between feature map F and each prototype p_k^{hol}, we obtain K similarity maps, formulated as: $S_k = p_k^{hol} \cdot F \in \mathbb{R}^{H\times W}$. Landmark detection prediction is achieved by selecting the position with the highest similarity in S_k. During model training, we employ standard regression loss to supervise ALD:

$$\mathcal{L}_{\text{reg}} = \frac{1}{K}\sum_{k=1}^{K} \|S_k - H_k\|_2^2, \tag{4}$$

where H_k is the heatmap of the k-th ground truth landmark. The overall optimization objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reg}} + \lambda_1\mathcal{L}_{\text{align}} + \lambda_2\mathcal{L}_{\text{mine}}, \tag{5}$$

where λ_1 and λ_2 are hyperparameters controlling the weights of different loss terms. Throughout the learning process, we employ a penalty-based gradient modulation process to prevent learning certain biased information (refer to Section Section 2.3 for more details).

2.2. Fair Identity Attention Module

To address bias issues arising from anatomical complexity and individual differences, we propose an identity-based attention mechanism that considers demographic attributes. Figure 1 illustrates the module's architecture. Input images and attribute labels are processed through embedding layers, generating input embeddings $F_i \in \mathbb{R}^{n_i \times d_i}$ and attribute embeddings $F_a \in \mathbb{R}^{n_a \times d_a}$, where n_i, n_a are feature lengths and d_i, d_a are dimensions. Position embeddings P are added to generate attention keys $K = W_k(F_i + P) \in \mathbb{R}^{n_i \times d_i}$ and values $V = W_v(F_i + P) \in \mathbb{R}^{n_i \times d_i}$. Attribute embeddings F_a and P generate queries $Q = W_q(F_i + P) \in \mathbb{R}^{n_a \times d_a}$.

The similarity matrix is computed via dot product of Q and K, then multiplied by V to extract task-relevant features:

$$\text{IDFairAttention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}}\right) \cdot V, \tag{6}$$

where d is a scaling factor. Residual connections preserve input integrity,with normalization and MLP layers applied before the final residual output.

The Fair Identity Attention mechanism improves model performance and fairness by explicitly considering demographic attributes like gender, race, and ethnicity. Its modular design allows seamless integration into existing networks without structural adjustments.

2.3. Penalty-Based Gradient Optimization Strategy

We design a penalty-based gradient modulation strategy to dynamically suppress bias reliance (Winterbottom et al., 2020). This mechanism corrects biases during training by: (1) using UNet to regress encoded features F_i for identity classification; (2) calculating weights for attribute layers based on inconsistency ratios; and (3) adaptively controlling gradient optimization—applying penalties for biased attributes while maintaining conventional gradient descent when contributions are balanced.

For convenience, we denote the training dataset as $\mathcal{D} = \{x_i, y_i\}_{i=1,2,\ldots,N}$. Each x_i contains feature maps from different attributes (e.g., age or gender). We define UNet's encoder as $\varphi(\theta)$, where θ represents the encoder's parameters. Then we let $W \in \mathbb{R}^{1 \times d_\varphi}$ and $b \in \mathbb{R}^1$ represent the parameters of the final linear classifier (regressing to attribute categories). The logits output of the linear classification layer in UNet's encoding layer is as follows:

$$f(x_i) = WF_i + b \in \mathbb{R}^{d_{out}}. \tag{7}$$

A complete gradient estimate $\widetilde{g}(\theta_t^u)$ is defined as:

$$g(\theta_t^m) = \frac{1}{m} \sum_{x \in B_t} \nabla_{\theta^u} L(x; \theta_t^m), \tag{8}$$

where B_t is a mini-batch of size m at step t. Considering gradient propagation for different identity attributes, we design an inconsistency ratio for penalty-based gradient modulation. We define:

$$s_i^{a_j} = \sum_{k=1}^{M} 1_{k=y_i} \cdot softmax\left(W_t^{a_j} \cdot \varphi_t^{a_j}\left(\theta^{a_j}, x_i^{a_j}\right) + \frac{b}{2}\right)_k \tag{9}$$

$$\rho_t^u = \begin{cases} s_i^m / \min(s_i^{a1}, s_i^{a2}) & i \in B_t, s_i^m = \min(s_i^{a1}, s_i^{a2}) \\ 0 & \text{others} \end{cases} \tag{10}$$

Here, different attributes serve as numerator and denominator respectively. Using ρ_t^u to dynamically monitor contribution differences between identity statistical feature patterns, we can adaptively modulate gradients through k_t^u:

$$k_t^u = \begin{cases} 1 - \tanh(m \cdot \rho_t^u) & \rho_t^u > 1 \\ 1 & \text{others} \end{cases} \tag{11}$$

To escape local minima, a simple but effective generalization enhancement (GE) method is introduced, which adds randomly sampled Gaussian noise $h\left(\theta_t^u\right) \sim \mathcal{N}\left(0, \Sigma^{sgd}\left(\theta_t^u\right)\right)$ to the gradient. Overall, the specific update method is as follows:

$$\theta_{t+1}^u = \theta_t^u - \eta k_t^u (g(\theta_t^u) + h(\theta_t^u)). \tag{12}$$

3. Experiment

3.1. Datasets and Metrics

Datasets: We validate our method using two public datasets: **CephAdoAdu** (Wu et al., 2024) is a cephalometric dataset designed to study the fairness impact of *age* differences. It includes 700 head X-ray images, evenly split into 350 adult and 350 adolescent images, with significant visual differences between the two groups. Each image is annotated with 10 landmarks by dental experts. The dataset is divided into 400 training images and 300 testing images. All images are resized to 1024x1024 for consistency, with an image spacing of 0.1mm. **Hand X-Rays** (Payer et al., 2019) is used to study the fairness impact of *gender* differences. It contains 895 X-ray images, each annotated with 37 hand landmarks. Images are resized to 1024x1216 and split into a 75% training set and a 25% testing set. Due to the lack of spacing information, we assume a distance of 50mm between the two wrist endpoints to estimate actual landmark distances, consistent with previous studies.

Metrics: Building on earlier discussions, we employ two commonly used metrics to evaluate model performance: 1) Mean Radial Error (MRE), which calculates the average Euclidean distance between predicted landmarks and ground truth landmarks; 2) Successful Detection Rate (SDR), defined as the percentage of landmarks accurately detected within a certain distance range from the ground truth landmarks; 3)**Radial Error Noise (REN)**- A novel fairness metric specifically designed to evaluate fairness in the field of ALD. REN measures the variability of model performance across different data distributions by calculating the difference between the MRE of each dataset and the overall average MRE. The specific formula is as follows:

$$REN = \frac{\sum_{i=1}^{n} \omega_i |MRE_i - \frac{\sum_{i=1}^{n} \omega_i MRE_i}{\sum_{i=1}^{n} \omega_i}|}{\sum_{i=1}^{n} \omega_i}, \tag{13}$$

where MRE_i and ω_i denote the MRE value and the weight of the i-th datasets. A smaller REN value indicates more stable performance and higher fairness, while a larger value suggests potential bias towards specific data distributions and lower fairness.

Table 1: Cephalometric ALD results with both adult and adolescent cases, only adult cases, and only adolescent cases, respectively.

Methods	Adult + Adolescent					Adult					Adolescent					REN ↓(mm)
	MRE ↓	SDR (%) ↑				MRE ↓	SDR (%) ↑				MRE ↓	SDR (%) ↑				
	(mm, std.)	2mm	2.5mm	3mm	4mm	(mm, std.)	2mm	2.5mm	3mm	4mm	(mm, std.)	2mm	2.5mm	3mm	4mm	
Cascade RCNN	2.31 (0.94)	61.47	73.20	81.13	90.77	2.19 (0.97)	59.93	72.13	80.47	90.80	2.43 (0.94)	63.00	74.27	81.80	90.73	0.080
SCN	1.73 (1.06)	82.97	90.40	93.37	96.57	1.40 (0.48)	82.07	91.20	94.33	97.33	2.05 (1.70)	83.87	89.60	92.40	95.80	0.218
GU2Net	1.69 (0.91)	80.33	88.13	91.47	95.57	1.46 (0.50)	80.27	88.80	92.07	96.33	1.93 (1.35)	80.40	87.47	90.87	94.80	0.158
SR-UNet	1.40 (0.93)	87.17	91.91	94.31	96.70	1.13 (0.89)	86.18	91.25	94.03	97.33	1.55 (1.87)	87.73	91.73	94.13	96.40	0.153
HTC	1.11 (1.08)	88.36	91.94	94.43	97.10	1.11 (1.09)	85.60	91.80	94.31	97.43	1.03 (1.07)	91.10	94.31	96.33	98.73	0.036
CeLDA	1.05 (0.33)	89.13	93.60	96.17	**98.67**	1.10 (0.37)	88.33	92.93	96.20	98.80	1.00 (0.34)	89.93	94.27	96.13	98.53	0.033
HYATT-Net	0.98 (0.33)	88.43	93.00	95.13	98.00	1.00 (0.70)	85.53	91.20	94.00	97.53	0.85 (0.95)	91.33	95.00	96.90	98.33	0.062
Ours	**0.91 (0.30)**	**89.90**	**93.97**	**96.20**	98.40	**0.93 (0.31)**	**89.97**	**93.73**	**96.60**	**98.87**	**0.90 (0.31)**	**91.93**	**95.30**	**96.90**	**98.93**	**0.011**

3.2. Implementation Details

Input images are resized to 512 × 512 and augmented with random brightness, contrast, and Gaussian noise variations. The model is trained for 150 epochs, with an initial learning rate of 0.001, reduced by 0.1 every 50 epochs. In equation (5), λ_1 and λ_2 are set to 1.0 and 3.0, respectively. All implementations use PyTorch and are trained on an NVIDIA H100 GPU with 80GB memory.

3.3. Comparison with SOTA Approaches

As shown in Table 1, IdFOPNet demonstrates outstanding performance on the CephAdoAdu dataset, achieving an average MRE of 0.91mm. This represents reductions of 0.07mm (over 7%) and 0.14mm (over 15%) compared to the previous SOTA methods HYATT-Net (Zhou et al., 2024) and CeLDA (Wu et al., 2024), respectively. CeLDA outperforms earlier general models (Payer et al., 2019; Cai and Vasconcelos, 2018; Zhu et al., 2021; Wu et al., 2023; Ao and Wu, 2023), such as GU2Net, and other SOTA methods like HTC (Viriyasaranon et al., 2023). Furthermore, we report results separately for adolescent and adult data. Across different age groups, our method exhibits excellent performance, achieving MREs of 0.93mm for adolescents and 0.90mm for adults. Additionally, SDR improves across all thresholds, with the SDR within 2mm for adults increasing from 88.33% to 99.97%. In terms of the fairness metric, our approach exhibits a markedly reduced REN value in comparison to prior methods. This underscores a significant mitigation in fairness for ALD, thereby ensuring a more robust and fair performance.This further illustrates that addressing fairness issues with IdFOPNet can substantially enhance overall performance, enabling the model to learn common features across different demographics while minimizing biases towards specific groups.

In the experiments on the hand dataset, we conduct a comprehensive comparison with the current SOTA method, HYATT-Net, as well as other approaches such as HTC (Viriyasaranon et al., 2023), CeLDA (Wu et al., 2024), GU2Net (Zhu et al., 2021), and FARNet (Ao and Wu, 2023). The results in Table 2 indicate that our method outperforms all others across all evaluation metrics. Our MRE is only 0.51mm, representing reductions of 0.02mm compared to HYATT-Net and 0.15mm compared to FARNet. The SDRs at thresholds of 2mm and 4mm reach 97.16% and 99.76%, respectively, showing significant improvements.

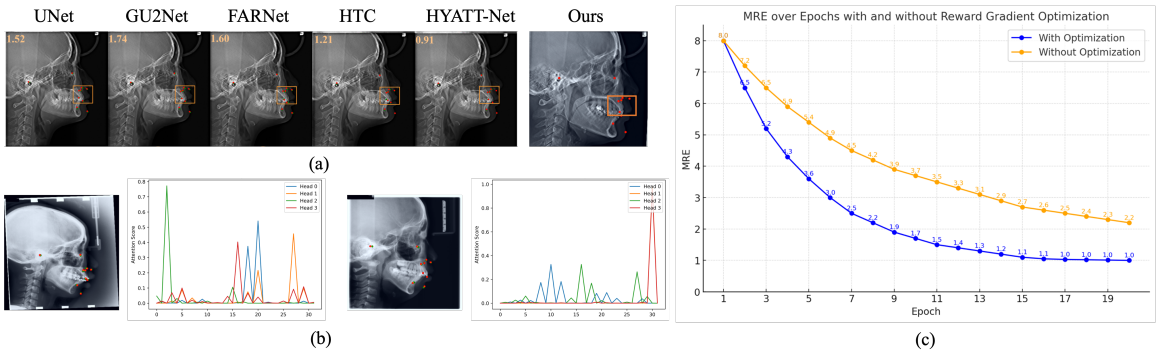

Figure 2: Visualization of ablation experiments. (a) Our method outperforms others in Landmark detection. (b) Neural activations differ for adults and adolescents under the identity attention mechanism, showing its effectiveness. (c) The fairness optimization strategy improves performance and accelerates convergence.

Table 2: MRE and SDR on Hand X-Rays.

Methods	MRE ↓ (mm, std.)	SDR (%) ↑ 2mm	4mm	10mm
GU2Net	0.63 (1.36)	96.01	99.39	99.98
SCN	0.66 (0.74)	94.99	99.27	99.99
FARNet	0.67 (0.74)	95.65	99.58	99.99
CeLDA	1.05 (0.33)	95.26	99.40	99.99
HTC	0.56 (0.58)	96.84	99.63	100.00
HYATT-Net	0.53 (0.56)	97.09	97.70	100.00
Ours	**0.51 (0.54)**	**97.16**	**99.76**	**100.00**

Table 3: Ablation analysis for our IdFOPNet.

Methods	MRE ↓ (mm, std.)	SDR (%) ↑ 2mm	2.5mm	3mm	4mm
CeLDA	1.05 (0.33)	**89.13**	**93.60**	**96.17**	**98.67**
HYATT-Net	**0.98 (0.33)**	88.43	93.00	95.13	98.00
w/o ID Attention	0.95 (0.30)	89.87	93.93	96.19	98.45
w/o Fair OP	0.99 (0.33)	89.18	93.08	96.09	98.48
Ours	**0.91 (0.30)**	**89.90**	**93.97**	**96.20**	98.40

3.4. Analytical Ablation Studies

Ablation studies on the CephAdoAdu dataset (Table 3) show that IdFOPNet outperforms baselines (CeLDA and HYATT-Net) across all metrics. Removing Fair OP reduces performance to CeLDA levels, highlighting its role in balancing accuracy and bias reduction. Without ID Attention, MRE increases from 0.91mm to 0.95mm, and SDR (2mm) drops from 89.90% to 89.87%, confirming its effectiveness in leveraging identity information for consistent performance across age and gender. These results underscore the importance of integrating Fair OP and ID Attention for improved accuracy, fairness, and robustness, further validated by visualizations in Figure 2.

4. Conclusion

This work proposes IdFOPNet, an ALD method that integrates identity features and fairness optimization. The method leverages a prototype network to detect landmarks by comparing image features with global landmark prototypes. To address fairness challenges, it introduces the Identity Attention mechanism, incorporating identity information as prior knowledge, and uses a penalty-based gradient modulation strategy to dynamically reduce model bias during training. Extensive evaluations on the CephAdoAdu and Hand X-Rays datasets show that IdFOPNet outperforms state-of-the-art methods, improving both accuracy and fairness across different age groups and genders.

Acknowledgments

Supported by Natural Science Foundation of China under Grant 62271465, Suzhou Basic Research Program under Grant SYG202338, and Open Fund Project of Guangdong Academy of Medical Sciences, China (No. YKY-KF202206).

References

Yueyuan Ao and Hong Wu. Feature aggregation and refinement network for 2d anatomical landmark detection. *Journal of Digital Imaging*, 36(2):547–561, 2023.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

J Chiras, C Depriester, A Weill, MT Sola-Martinez, and H Deramond. Percutaneous vertebral surgery. technics and indications. *Journal of neuroradiology= Journal de neuroradiologie*, 24(1):45–59, 1997.

Yankai Jiang, Yiming Li, Xinyue Wang, Yubo Tao, Jun Lin, and Hai Lin. Cephalformer: incorporating global structure constraint into visual features for general cephalometric landmark detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 227–237. Springer, 2022.

Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in neural information processing systems*, 33:3118–3129, 2020.

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12289–12301, 2024a.

Yan Luo, Yu Tian, Min Shi, Louis R Pasquale, Lucy Q Shen, Nazlee Zebardast, Tobias Elze, and Mengyu Wang. Harvard glaucoma fairness: a retinal nerve disease dataset for fairness learning and fair identity normalization. *IEEE Transactions on Medical Imaging*, 2024b.

Kanghan Oh, Il-Seok Oh, Dae-Woo Lee, et al. Deep anatomical context feature learning for cephalometric landmark detection. *IEEE Journal of Biomedical and Health Informatics*, 25(3):806–817, 2020.

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.

Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12556–12565, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Yu Tian, Congcong Wen, Min Shi, Muhammad Muneeb Afzal, Hao Huang, Muhammad Osama Khan, Yan Luo, Yi Fang, and Mengyu Wang. Fairdomain: Achieving fairness in cross-domain medical image segmentation and classification. In *European Conference on Computer Vision*, pages 251–271. Springer, 2025.

Thanaporn Viriyasaranon, Serie Ma, and Jang-Hwan Choi. Anatomical landmark detection using a multiresolution learning approach with a hybrid transformer-cnn model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 433–443. Springer, 2023.

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.

Chong Wang, Yuanhong Chen, Fengbei Liu, Michael Elliott, Chun Fung Kwok, Carlos Pena-Solorzano, Helen Frazer, Davis James McCarthy, and Gustavo Carneiro. An interpretable and accurate deep-learning diagnosis framework modelled with fully and semi-supervised reciprocal learning. *IEEE Transactions on Medical Imaging*, 2023.

Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. On modality bias in the tvqa dataset. *arXiv preprint arXiv:2012.10210*, 2020.

Han Wu, Chong Wang, Lanzhuju Mei, Tong Yang, Min Zhu, Dinggang Shen, and Zhiming Cui. Cephalometric landmark detection across ages with prototypical network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 155–165. Springer, 2024.

Qian Wu, Si Yong Yeo, Yufei Chen, and Jun Liu. Revisiting cephalometric landmark detection from the view of human pose estimation with lightweight super-resolution head. *arXiv preprint arXiv:2309.17143*, 2023.

Qingsong Yao, Quan Quan, Li Xiao, and S Kevin Zhou. One-shot medical landmark detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 177–188. Springer, 2021.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31, 2018.

S Kevin Zhou, Daniel Rueckert, and Gabor Fichtinger. *Handbook of medical image computing and computer assisted intervention*. Academic Press, 2019.

Xiaoqian Zhou, Zhen Huang, Heqin Zhu, Qingsong Yao, and S Kevin Zhou. Hybrid attention network: An efficient approach for anatomy-free landmark detection. *arXiv preprint arXiv:2412.06499*, 2024.

Heqin Zhu, Qingsong Yao, Li Xiao, and S Kevin Zhou. You only learn once: Universal anatomical landmark detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 85–95. Springer, 2021.