

MCPNet:Masked Coordinate Pooling-based Attention Network for Medical Landmark Detection

Gyu-Sung Ham¹, Gi Hyun Lim^{1,2}, and Kanghan Oh^{1,2†}

¹ AI Convergence Research Institute, Wonkwang University, Iksan 54538, Republic of Korea

ham1231@wku.ac.kr

² Department of Computer and Software Engineering, Wonkwang University, Iksan 54538, Republic of Korea

ghlim71, khoh888@wku.ac.kr

Abstract. This study introduces a masked coordinate pooling-based attention network (MCPNet), which is a novel framework designed to improve medical landmark detection. Coordinate features are essential in accurately localizing medical landmarks. Recently, many studies have used coordinate information in final regression heads; however, its direct incorporation into the attention mechanism remains largely unexplored. To address this issue, we propose MCPNet, which directly extracts positional features from CNN feature maps and analyzes them in the coordinate domain. Specifically, MCPNet employs learnable masks and coordinate maps to effectively encode diverse and richer anatomical positional cues of the landmarks. Moreover, the proposed mask maps are utilized when transforming the analyzed coordinate features back into the original CNN feature space, ensuring smooth reconstruction within the network. Extensive experiments on four public benchmark datasets demonstrate that the proposed MCPNet outperforms state-of-the-art methods, offering improved accuracy and efficiency.

Keywords: Medical landmark detection · Attention network · Masked coordinate pooling

1 Introduction

The precise localization of medical landmarks is essential for accurately analyzing medical images in radiology, as these landmarks help delineate anatomical structures and provide crucial guidance for diagnosis, treatment planning, and surgical procedures [12, 36]. Recent advancements have driven the development of automated landmark detection methods, which have significantly advanced clinical applications such as image-based diagnosis [9]. Specifically, convolutional neural network (CNN)-based heatmap regression approaches [1, 2, 5, 10, 17, 18, 20,

† Corresponding author.

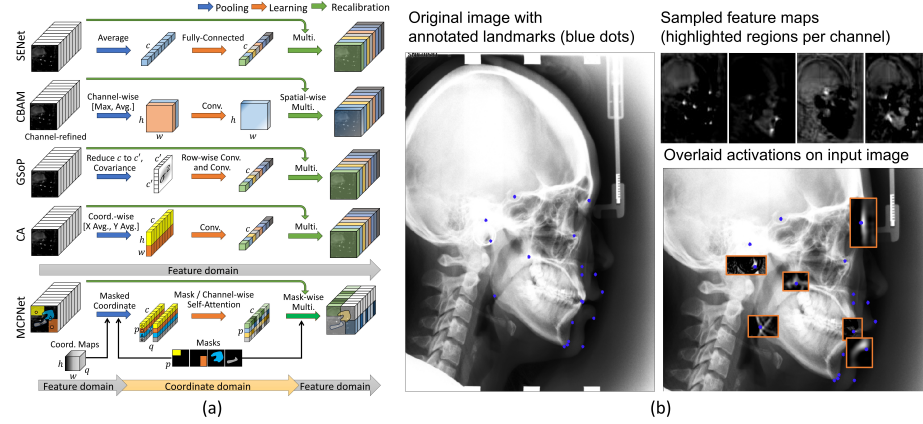


Fig. 1: (a) Conceptual comparison between existing attention mechanisms and MCPNet. Unlike conventional methods that lose spatial details through pooling, MCPNet analyzes spatial relationships directly in the coordinate domain and returns them to the CNN feature domain, preserving fine-grained cues for precise localization. (b) Visualization of feature maps from a trained heatmap regression network. Activated regions highlight that CNNs implicitly encode positional cues around anatomical landmarks, and multiple landmark clues can appear simultaneously within a single channel.

27,28,38,49] have demonstrated rapid and accurate results and improved clinical utility.

In clinical practice, physicians adjust landmark coordinates and examine their geometric relationships, such as distances and angles. This highlights the need for models that can learn coordinate-aware features for precise landmark detection. However, most existing studies [10,24,31,34] have incorporated coordinate features only in the final regression heads without handling them within the attention modules that can be embedded in intermediate layers.

As illustrated in Fig. 1 (a), existing attention modules typically follow a *pooling-learning-recalibration* process. Before feature learning and recalibration, they employ pooling operations such as global average pooling (GAP) [14,44], global max pooling (GMP) [44], covariance pooling [8], and unidirectional pooling [41]. However, these pooling operations are fundamentally unsuitable for medical landmark detection, because they cause a loss of positional information. Figure 1 (b) presents feature map visualizations obtained from a trained heatmap regression network. The results showed that the landmarks and their surrounding areas were distinctly activated and these salient regions serve as valuable abstracted features for directly learning landmark information. As shown at the top right of Fig. 1 (b), multiple landmark clues often exist in a single-channel map. These observations indicate that traditional attention methods based on global pooling [14,41,42,44] overlook critical spatial cues, lose positional information, and are ineffective for fine-grained feature abstraction [10,18].

To address this limitation, we introduce the Masked Coordinate Pooling-

based attention Network (MCPNet), which explicitly models spatial positions. Our core module, masked coordinate pooling (MCP), employs the soft-argmax technique [25] to transform CNN features directly into the coordinate domain, enabling intuitive and differentiable handling of spatial information. In our MCPNet, mask maps perform two essential functions. First, they capture various positional features distributed over each feature channel, which preserves spatial information that might otherwise be lost. Second, they assist during reconstruction by converting the coordinate features into forms compatible with the convolutional layers, thus ensuring smooth integration with the CNN framework. Furthermore, our method incorporates Cartesian coordinate maps and learnable adaptive polar-like coordinate maps, which enhance discriminability and enrich spatial cues. Finally, the extracted positional features were passed into the self-attention layer, where the anatomical contextual relationships were analyzed. The resulting attention maps were then combined with the applied masks to restore them to the original CNN feature space. The principal contributions of this study are summarized as follows:

- For medical landmark detection, we propose MCPNet, an attention framework that analyzes landmark spatial relationships. To the best of our knowledge, this is the first attention framework that performs reasoning directly in the coordinate domain.
- The proposed mask maps form a key contribution of MCPNet. They capture diverse positional features while preserving spatial details, and reconstruct coordinate representations into forms compatible with convolutional operations, ensuring smooth integration within CNN architectures.
- Beyond static Cartesian maps, we introduce learnable adaptive polar-like coordinate maps that improve spatial discriminability and provide a richer geometric interpretation.
- Extensive experiments on six public benchmarks demonstrate that MCPNet outperforms both conventional attention mechanisms and existing landmark detection methods.

2 Related Work

Medical Landmark Detection Models. Several studies have introduced multistream networks to jointly learn the local and global features. Payer et al. [28] proposed a spatial configuration network (SCN) that integrates local and spatial configuration heat maps to improve the robustness of limited datasets. Similarly, Kwon et al. [20] and Feng et al. [7] adopted multistage localization strategies that divided the process into local and global stages, and refined the landmark positions through feature fusion. Chen et al. [1] introduced an attention feature pyramid fusion (AFPF) module that merges multiscale CNN features via a voting strategy. In addition to these hierarchical methods, context-aware approaches aim to capture anatomical relationships among landmarks. Oh et al. [18] and Ham et al. [10] incorporated anatomical constraints by learning landmark correlations from heatmap coordinates, whereas Chen et al. [3] proposed region

attention loss to emphasize anatomical structures. Tan et al. [33] further optimized vascular landmark detection using global spatial constraints. More recently, transformer-based models [15, 38, 48] have improved landmark detection by modeling long-range dependencies. For instance, Zhao et al. [48] combined HRNet with a transformer encoder to detect spinal landmarks, whereas Kumaralingam et al. [15] introduced a transformer network with an edge feature encoder and spatial squeeze attention to improve the accuracy. Moreover, Viriyasaranon et al. [38] presented a hybrid transformer (HTC) framework with multiresolution heatmap learning to unify local and global representation learning.

Attention Models. Attention mechanisms are effective in enhancing feature representations by emphasizing informative responses. Hu et al. [14] introduced the squeeze-and-excitation (SE) method, which uses GAP and fully connected layers to recalibrate channel-wise features. CBAM [44] adds spatial attention by combining GAP and GMP to improve feature abstraction. However, these methods can cause the loss of spatial information, which is a limitation in tasks such as medical landmark detection, where spatial relationships are crucial. Higher-order attention mechanisms such as GSOP [8] and AiA [6] capture richer dependencies by modeling pairwise channel correlations, but this incurs high computational costs. By introducing the transformer architecture [37], self-attention has become popular in modeling long-range dependencies. Following this, efficient variants such as the neighborhood attention transformer (NAT) [11] reduce computation by restricting attention to local neighborhoods and inherently encoding spatial proximity. However, this locality also limits the ability to capture nonlocal geometric relationships and anatomical topology, as these models lack explicit coordinate reasoning. These limitations led to the development of our MCP, which explicitly encodes spatial coordinates to capture the structural relationships among landmarks.

Positional Awareness Learning Methods. Previous studies attempted to help networks learn positional information. A common approach introduces positional information into networks using fixed coordinate maps that are either concatenated with or added to feature maps as positional embeddings [22, 37]. However, these methods provide only static coordinate priors and are not adaptive to the underlying feature maps. Unidirectional pooling along a single spatial axis, such as in coordinate attention (CA) [41] and Project & Excite [29], emphasizes directional dependencies. Grid-based regression methods [13] divide feature maps into fixed local cells and perform regressions within each grid. Despite these efforts, existing methods rely on implicitly capturing spatial information. By contrast, our method explicitly extracts coordinate information from feature maps via mask- and coordinate-based operations, allowing the network to learn positional awareness more directly.

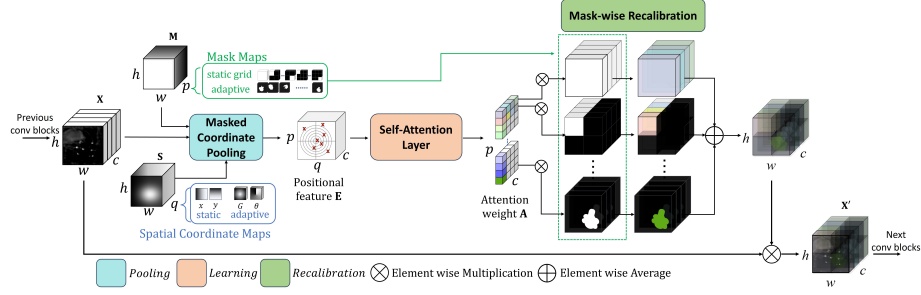


Fig. 2: Overview of the proposed attention block. The input features along with the coordinate and mask maps are processed by MCP to yield transformed positional features. Subsequently, the self-attention layer models contextual relationships among landmarks using these positional features, resulting in weight vectors. Finally, each weight vector is multiplied by its corresponding mask, and the weighted masks are averaged to obtain the final attention map.

3 Proposed Approach

A schematic architecture of the proposed MCPNet is illustrated in Fig. 2. The MCPNet is designed to jointly model positional features and anatomical contextual relationships, enabling precise landmark localization.

3.1 Masked Coordinate Pooling

The MCP module aims to transform feature maps into a coordinate domain. To ensure differentiability, it is designed based on predefined coordinate maps and the soft-argmax technique [21, 32, 46], and employs a mask-based pooling strategy to emphasize meaningful local regions.

Let's consider input features $\mathbf{X} \in \mathbb{R}^{c \times h \times w}$, where c , h , and w denote the channel, height, and width, respectively. Given mask maps $\mathbf{M} \in \mathbb{R}^{p \times h \times w}$ and spatial coordinate maps $\mathbf{S} \in \mathbb{R}^{q \times h \times w}$, the proposed MCP converts the input

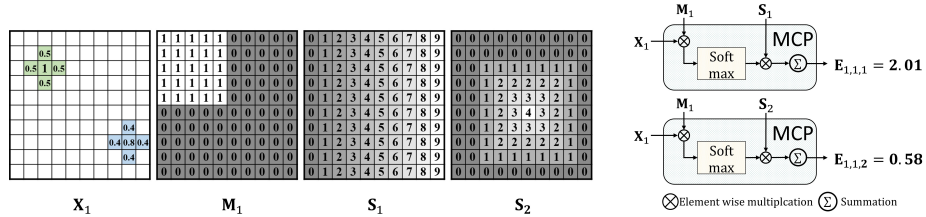


Fig. 3: Illustration of the MCP operation with a simple example. MCP applies mask maps to focus on local regions within a feature map, and performs a soft-argmax operation with spatial coordinate maps to project these activations into the coordinate domain (e.g., $\mathbf{E}_{1,1,1} = 2.01$, $\mathbf{E}_{1,1,2} = 0.58$.)

features $\mathbf{X}$ into positional features $\mathbf{E} \in \mathbb{R}^{c \times p \times q}$, where p and q represent the number of mask maps and coordinate maps, respectively. We denote the k -th feature map $\mathbf{X}_k \in \mathbb{R}^{h \times w}$, the u -th mask map $\mathbf{M}_u \in \mathbb{R}^{h \times w}$, and the v -th coordinate map $\mathbf{S}_v \in \mathbb{R}^{h \times w}$. Given $\mathbf{X}_k$, $\mathbf{M}_u$, and $\mathbf{S}_v$, the soft-argmax based MCP can be defined as:

$$\mathbf{E}_{k,u,v} = \sum \mathbf{S}_v \otimes \text{softmax}(\mathbf{X}_k \otimes \mathbf{M}_u), \quad (1)$$

where $\otimes$ denotes the element-wise multiplication (Hadamard product). The mask map $\mathbf{M}_u$ activates the local region of $\mathbf{X}_k$ while suppressing others. When considering both mask maps $\mathbf{M} \in \mathbb{R}^{p \times h \times w}$ and coordinate maps $\mathbf{S} \in \mathbb{R}^{q \times h \times w}$, the transformed positional features $\mathbf{E} \in \mathbb{R}^{c \times p \times q}$ can be represented as:

$$\mathbf{E} = \begin{bmatrix} \mathbf{E}_{1,1,1} & \cdots & \mathbf{E}_{1,p,1} \\ \vdots & \ddots & \vdots \\ \mathbf{E}_{1,1,q} & \cdots & \mathbf{E}_{1,p,q} \end{bmatrix} \cdots \begin{bmatrix} \mathbf{E}_{c,1,1} & \cdots & \mathbf{E}_{c,p,1} \\ \vdots & \ddots & \vdots \\ \mathbf{E}_{c,1,q} & \cdots & \mathbf{E}_{c,p,q} \end{bmatrix}. \quad (2)$$

Figure 3 presents the MCP operation with a simple example. Note that the computations were performed efficiently by leveraging parallel processing and broadcasting. The pseudo-code and an extended toy example are provided in Appendix A. Our MCP can be regarded as an extended version of the soft-argmax function, which simply applies mask-wise weighting to the feature space. Therefore, it is straightforwardly differentiable [21, 32, 46].

Mask Map. Medical images are extremely large and contain widely dispersed landmarks, where global-only approaches tend to lose local details and hinders precise localization [26, 41]. To address this, MCP employs multi-level regular grid masks with binary patterns as static masks. Each mask activates local regions in a channel map via Hadamard product (see Fig. 3). In addition, we propose an adaptive mask generation method to capture diverse positional features that static masks may overlook:

$$\mathbf{M}_{\text{adapt}} = \text{TReLU}(\text{abs}(\mathbf{X} * \mathbf{W}_k)) * \mathbf{G}_z, \quad (3)$$

where $*$ denotes convolution, $\mathbf{W}_k$ represents learnable 1×1 kernels, and abs is an absolute function. Gaussian smoothing with $\mathbf{G}_z$ ($z=7$) reduces noise and softens mask boundaries, while Threshold ReLU (TReLU) [19] controls sparsity. Figure 4 (a) shows adaptive masks under different thresholds—higher values yield sparser maps. Unlike grid-based masks, adaptive masks have continuous values, enabling a coarse-to-fine hierarchy through multi-threshold design.

Coordinate Map. An overview of the proposed MCP module with adaptive coordinate map generation is illustrated in Fig. 5. We introduce two complementary coordinate maps—static Cartesian and adaptive polar-like—that provide spatial cues from multiple anatomical perspectives. The static coordinate map is defined as follows:

$$\mathbf{S}_{\text{stat}} = \text{IN}([s_x, s_y]), \quad (4)$$

where s_x and s_y denote the x-y coordinate grids and $\text{IN}(\cdot)$ is Instance Normalization (IN). Combining $\mathbf{X}$, mask maps $\mathbf{M}$, and the static coordinate maps $\mathbf{S}_{\text{stat}}$, MCP extracts the static positional features $\mathbf{E}_{\text{stat}} \in \mathbb{R}^{c \times p \times 2}$.

We further generate adaptive coordinate maps inspired by polar coordinates to capture richer and more discriminative spatial representations. These maps consist of Gaussian-based radial and angular components parameterized by learnable centroids and variances. Cartesian coordinates $\mathbf{S}_{\text{stat}}$ serve as the primary geometric reference, while the adaptive coordinates $\mathbf{S}_{\text{adapt}}$ act as learned secondary spatial representations that semantically reinterpret the primary positional structures. Given $\mathbf{E}_{\text{stat}}$, the adaptive coordinate maps are formulated as follows:

$$\mathbf{O}_{\text{ac}} = \Phi(\text{Flat}(\text{GAP}(\mathbf{E}_{\text{stat}}))\mathbf{W}_{\text{a}}), \quad (5)$$

where $\mathbf{W}_{\text{a}}$ denotes the learnable parameters of a FC layer, $\text{GAP}(\cdot)$, $\text{Flat}(\cdot)$, and $\Phi(\cdot)$ represent the GAP, flatten, and sigmoid functions, respectively. To reduce computational cost, $\mathbf{E}_{\text{stat}} \in \mathbb{R}^{c \times p \times 2}$ is summarized to $p \times 2$ using GAP, then projected by $\mathbf{W}_{\text{a}} \in \mathbb{R}^{2p \times 3n}$ to predict n sets of (cx_i, cy_i, σ_i) , corresponding to the centroids and variances for generating the adaptive coordinate maps. For each i -th coordinate, the Gaussian and angular maps are defined as:

$$s_{G_i} = \exp\left(-\frac{(s_x - cx_i)^2 + (s_y - cy_i)^2}{2\sigma_i^2}\right), \quad (6)$$

$$s_{\theta_i} = \tan^{-1}\left(\frac{s_y - cy_i}{s_x - cx_i}\right), \quad (7)$$

where $(s_x, s_y) \in \mathbb{R}^{h \times w}$, and each (cx_i, cy_i, σ_i) is broadcast over the spatial domain. The resulting adaptive coordinate maps $\mathbf{S}_{\text{adapt}} \in \mathbb{R}^{n \times h \times w}$ are given as:

$$\mathbf{S}_{\text{adapt}} = \text{IN}([s_{G_1}, s_{\theta_1}, s_{G_2}, s_{\theta_2}, \dots, s_{G_n}, s_{\theta_n}]), \quad (8)$$

producing spatially localized coordinate bases centered on their learned centroids. The examples are illustrated in Fig. 4 (b).

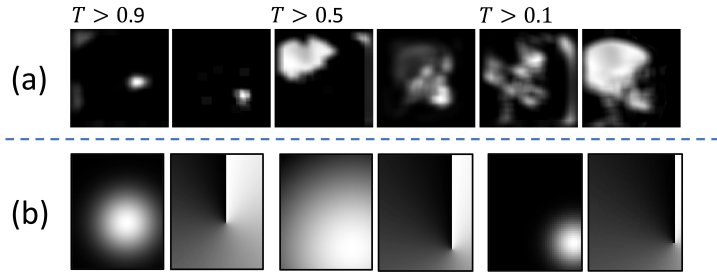


Fig. 4: (a) Schematic of adaptive mask maps with different thresholds. (b) Visualization of adaptive coordinate maps. The left column shows the Gaussian based radial distance maps, while the right column displays angular maps.

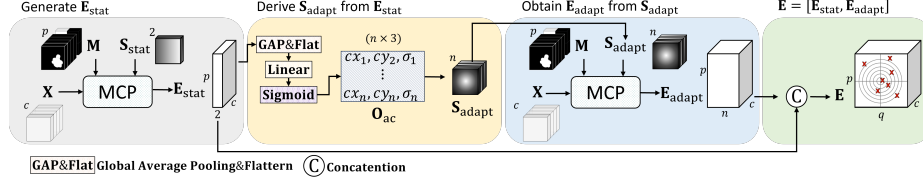


Fig. 5: Overview of the proposed MCP module with adaptive coordinate map generation.

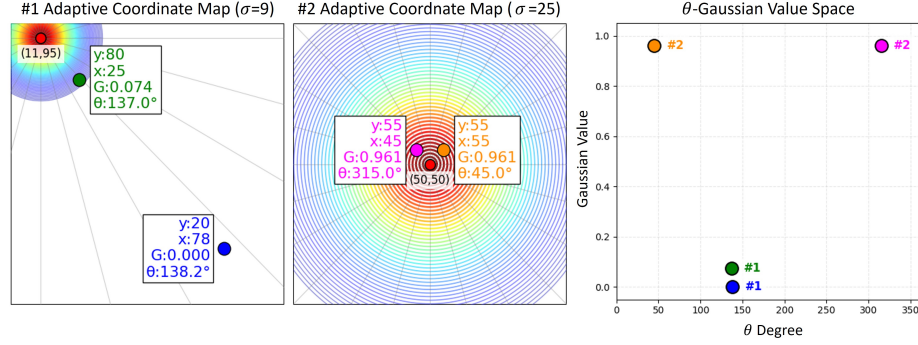


Fig. 6: Illustration of the adaptive coordinate, which enhances spatial discriminability and captures contexture features beyond Cartesian representation.

To better illustrate the effect of our adaptive coordinate map, two contrastive examples are presented in Fig. 6. In the first case (Fig. 6, left), although two features are spatially distant in the Cartesian space, they appear close in the adaptive coordinate space, illustrating context aggregation how contextual information is effectively aggregated. This occurs because the two features are similarly oriented in the θ map, and the small Gaussian variance makes their intensities appear similar. In the second case (Fig. 6, right), two features are close in the Cartesian space, but they appear apart in the adaptive coordinate space by the shift of the center coordinate, showing higher discriminability.

Subsequently, the MCP module processes $\mathbf{X}$, $\mathbf{M}$, and $\mathbf{S}_{\text{adapt}}$ to produce the adaptive positional feature $\mathbf{E}_{\text{adapt}} \in \mathbb{R}^{c \times p \times n}$, which is concatenated with $\mathbf{E}_{\text{stat}}$ to form the final representation $\mathbf{E} = [\mathbf{E}_{\text{stat}}, \mathbf{E}_{\text{adapt}}] \in \mathbb{R}^{c \times p \times q}$.

3.2 Self-Attention Layer

The self-attention layer models anatomical contextual relationships by analyzing the positional feature $\mathbf{E}$ in the coordinate domain, which is derived from diverse masks and coordinate maps. Before computing attention, the spatial feature dimensionality is reduced using a 1×1 convolution followed by layer normalization:

$$\mathbf{E}' = \text{ReLU}(\text{LN}(\mathbf{E} * \mathbf{W}_{\text{rd}})), \quad (9)$$

reducing the size from $c \times p \times q$ to $c \times p \times q'$ ($q' < q$). Afterward, $\mathbf{E}'$ is reshaped into $\hat{\mathbf{E}} \in \mathbb{R}^{c \times pq'}$, where each of the c spatial units is represented as a token with embedding dimension pq' .

$\hat{\mathbf{E}}$ is linearly projected into query $\mathbf{Q}_l$, key $\mathbf{K}_l$, and value $\mathbf{V}_l$ for each head l , using weight matrices of size $pq' \times p$ that align the token dimension with the number of masks (p) to facilitate the subsequent mask-wise feature recalibration. Multi-head attention is performed following the standard Transformer formulation [37]:

$$\mathbf{A}_l = \text{Att}(\mathbf{Q}_l, \mathbf{K}_l, \mathbf{V}_l) = \text{softmax}\left(\frac{\mathbf{Q}_l \mathbf{K}_l^T}{\sqrt{d_k}}\right) \mathbf{V}_l, \quad \mathbf{A} = [\mathbf{A}_1, \dots, \mathbf{A}_h] \mathbf{w}^O. \quad (10)$$

The resulting $\mathbf{A}$ serves as mask-wise attention weights, which are used to recalibrate the corresponding CNN features in the feature domain.

3.3 Mask-Wise Feature Recalibration

We present a masked weighting method that integrates the analyzed coordinate features back into the feature domain, ensuring smooth reconstruction within the network. As shown in Fig. 2 (right part), a visual example is provided to demonstrate how mask-wise feature recalibration is applied. Given the attention weight $\mathbf{A} \in \mathbb{R}^{p \times c}$, mask maps $\mathbf{M} \in \mathbb{R}^{p \times h \times w}$, and input features $\mathbf{X} \in \mathbb{R}^{c \times h \times w}$, the mask wise attention method is defined as:

$$\mathbf{X}' = \mathbf{X} \otimes \left(\frac{1}{p} \sum_{u=1}^p \mathbf{M}_u^{1_c} \otimes \Phi(\mathbf{A}_u^{1_{h \times w}}) \right), \quad (11)$$

where $\mathbf{M}_u^{1_c}$ and $\mathbf{A}_u^{1_{h \times w}}$ are broadcast along channel and spatial dimensions, respectively. We generate p -th attention maps, each highlights distinct local features within spatial and channel domains. These maps are then integrated and normalized to produce a final attention map.

4 Experiments

4.1 Implementation Details

The experiments are performed on a machine with a NVIDIA RTX A6000 GPU. For the optimization, we employ AdamW [23] optimizer with a batch size of 1, and the network is trained for 300 epochs. The initial learning rate is set to 0.0005 and is decreased by a factor of 0.1 at epoch 250 and 280. During training, we employ the Adaptive Wing loss [43] with parameters $\alpha = 2.1$, $\omega = 14.0$, $\epsilon = 1.0$, and $\theta = 1.2$, and perform the data augmentation, including random rotation ($\pm 15^\circ$) and random scaling ($[0.8, 1.1]$). Ground-truth heatmaps are generated

using a Laplacian distribution with $\sigma = 12$. For the proposed MCP, 14 static coordinate maps and 50 adaptive masks are used. MCPNet is built upon a U-shape architecture [30]. Each convolution layer consists of two repeated 3×3 convolutions (with padding) followed by IN and ReLU activation. Down-sampling and up-sampling are performed by 2×2 max-pooling and bilinear interpolation, respectively. The feature map size increases progressively from 64, 128, 256, 512 to 1024 channels in the encoder, and then symmetrically reduces back down to 64 in the decoder. The skip connections are used to concatenate output feature maps of equal spatial dimensions between the encoder and decoder. The proposed attention blocks are placed after the final ReLU activation in each module, where each attention block uses 4 heads in its MHA mechanism.

4.2 Comparison with State-of-the Art

We compare the proposed MCPNet with other state-of-the-art methods on four public benchmarks: ISBI2015 [40], Digital Hand Atlas [35], BMPLE [45], and Chest X-ray [49]. These benchmarks encompass challenging conditions such as domain shift across hospitals, abnormal anatomy, and severe deformation, providing a comprehensive evaluation of robustness. We conducted five independent trials for evaluation. For the DHA dataset, we used three-fold cross-validation, as it is released with three predefined folds. Two consistent evaluation metrics are adopted—mean radius error (MRE) and successful detection rate (SDR) under various physical distances (z mm) [39]. The image resolutions are standardized for each benchmark: ISBI2015 (800×640), DHA (1024×800), BMPLE (1152×576), and Chest X-ray (768×768). A detailed overview of these benchmarks and evaluation protocols is provided in Appendix B.

For comparison, we group existing methods into two categories: attention-based and recent medical landmark detection approaches. To validate the effectiveness of our attention module, we compare it with prior attention-based methods under identical training setups and backbone networks, following each method’s own reported hyperparameters. Additionally, we evaluate our model against recent advancements in medical landmark detection for a comprehensive analysis. For methods without standard deviations in our table, we report the best performance values from the original papers.

Table 1 summarizes the quantitative results of MCPNet and other baseline methods, presenting only the key evaluation metrics—SDR under the most challenging threshold (e.g., 2 mm) and MRE. The first seven rows correspond to attention-based models, followed by five recent landmark detection approaches. Due to page limitations, the complete quantitative comparisons and detailed analyses for each benchmark are provided in Appendix C.

On the ISBI2015 benchmark, our proposed MCPNet consistently shows competitive performance on both Test1 (validation) and Test2 (on-site competition) benchmarks. Results on the BMPLE benchmark show that MCPNet consistently outperforms existing methods in SDR@4mm and maintains the lowest MRE. Additionally, across both Test1 and Test2, MCPNet ranks second-highest among all methods, excelling in both MRE and SDR@4mm. For the DHA benchmark,

Table 1: Quantitative comparisons between our MCPNet and other methods on four benchmarks. Bold indicates the best performance, and underline indicates the second-best performance. The symbol " $\uparrow$ " (" $\downarrow$ ") indicates that the higher(lower) is better.

Method	ISBI2015 (Test1, Test2)				BLPLE (Test1, Test2)				DHA		Chest X-ray		#Params
	SDR $\uparrow$ (%, < 2mm)	MRE $\downarrow$ (mm)	SDR $\uparrow$ (%, < 2mm)	MRE $\downarrow$ (mm)	SDR $\uparrow$ (%, < 4mm)	MRE $\downarrow$ (mm)	SDR $\uparrow$ (%, < 4mm)	MRE $\downarrow$ (mm)	SDR $\uparrow$ (%, < 2mm)	MRE $\downarrow$ (mm)	SDR $\uparrow$ (%, < 3px)	MRE $\downarrow$ (mm)	
U-Net [30]	86.22 $\pm$.834	1.14 $\pm$.023	73.58 $\pm$.461	1.46 $\pm$.025	85.55 $\pm$.714	3.11 $\pm$.034	84.58 $\pm$.499	3.15 $\pm$.020	95.91 $\pm$.407	0.65 $\pm$.029	50.85 $\pm$.621	4.30 $\pm$.028	13.19
U-Net † [30]	86.35 $\pm$.822	1.13 $\pm$.023	73.95 $\pm$.501	1.45 $\pm$.024	85.76 $\pm$.613	3.07 $\pm$.033	84.71 $\pm$.516	3.10 $\pm$.028	96.05 $\pm$.414	0.64 $\pm$.026	51.01 $\pm$.543	4.28 $\pm$.025	22.71
SE [14]	86.34 $\pm$.598	1.13 $\pm$.022	74.81 $\pm$.383	1.44 $\pm$.034	85.47 $\pm$.597	3.05 $\pm$.026	84.88 $\pm$.588	3.07 $\pm$.031	96.21 $\pm$.581	0.64 $\pm$.028	51.16 $\pm$.567	4.21 $\pm$.024	13.47
CBAM [44]	86.44 $\pm$.790	1.13 $\pm$.017	75.55 $\pm$.425	1.44 $\pm$.018	85.98 $\pm$.588	3.08 $\pm$.019	84.63 $\pm$.586	3.14 $\pm$.023	96.18 $\pm$.434	0.63 $\pm$.026	50.90 $\pm$.442	4.25 $\pm$.016	13.57
GSOP [8]	86.43 $\pm$.817	1.13 $\pm$.025	75.41 $\pm$.519	1.43 $\pm$.019	86.43 $\pm$.367	2.78 $\pm$.024	85.44 $\pm$.472	2.99 $\pm$.034	96.48 $\pm$.354	0.61 $\pm$.017	51.88 $\pm$.783	4.12 $\pm$.024	24.45
AIA [6]	86.88 $\pm$.625	<u>1.10</u> $\pm$.013	<u>76.33</u> $\pm$.336	1.43 $\pm$.025	87.66 $\pm$.536	2.67 $\pm$.021	84.23 $\pm$.494	2.73 $\pm$.018	96.35 $\pm$.675	0.62 $\pm$.018	52.09 $\pm$.624	4.05 $\pm$.018	22.48
NAT [11]	<u>87.59</u> $\pm$.654	1.11 $\pm$.024	75.15 $\pm$.465	<u>1.41</u> $\pm$.019	<u>87.89</u> $\pm$.374	<u>2.55</u> $\pm$.028	85.44 $\pm$.587	2.68 $\pm$.025	96.76 $\pm$.426	<u>0.60</u> $\pm$.021	<u>52.24</u> $\pm$.561	4.08 $\pm$.027	18.13
CA [13]	86.73 $\pm$.634	1.13 $\pm$.016	75.37 $\pm$.438	1.46 $\pm$.022	86.18 $\pm$.451	2.88 $\pm$.025	85.24 $\pm$.410	<u>2.65</u> $\pm$.022	96.11 $\pm$.468	0.61 $\pm$.026	51.89 $\pm$.756	4.20 $\pm$.019	13.47
D-CeLR [4]	86.04 $\pm$.801	1.15 $\pm$.031	73.54 $\pm$.606	1.45 $\pm$.027	85.10 $\pm$.768	2.88 $\pm$.023	85.01 $\pm$.678	3.11 $\pm$.031	95.55 $\pm$.564	0.64 $\pm$.028	51.73 $\pm$.689	4.33 $\pm$.035	43.47
APFP [1]	86.67	1.17	75.05	1.48	85.25	3.08	84.80	3.23	95.88	0.64	-	-	25.21
DACFL [18]	86.20	1.18	75.90	1.45	86.15	2.89	83.26	3.08	96.07	0.63	51.11	4.11	13.41
NFDP [17]	87.20	1.12	75.11	1.46	-	-	-	-	94.38	0.72	-	-	12.10
HTC [38]	87.15	1.13	75.32	1.43	87.85	2.68	<u>85.58</u>	2.89	<u>96.84</u>	0.56	51.22	<u>3.91</u>	18.24
MCPNet	88.89 $\pm$.623	1.07 $\pm$.021	78.23 $\pm$.448	1.37 $\pm$.023	88.45 $\pm$.586	2.17 $\pm$.026	86.69 $\pm$.434	2.33 $\pm$.025	96.94 $\pm$.347	0.56 $\pm$.024	55.33 $\pm$.548	3.55 $\pm$.026	22.73

† U-Net with channel width increased to a comparable parameter count.

although HTC achieves good performance, our MCPNet outperforms all existing methods in the crucial SDR@2mm, highlighting the precision and generalization capabilities of our method. For the Chest X-ray benchmark, our MCPNet models achieve the top two performances among all methods, showing superior SDR results and the lowest MRE. The relatively low SDR@3px is mainly due to the landmarks being located along the liver contours, where local cues are limited, making MRE a more reliable indicator in this dataset. To verify that MCPNet's gains do not simply stem from increased model capacity, we additionally evaluate U-Net † , with channel width increased to match MCPNet's parameter count. U-Net † shows no significant improvement over the original U-Net, confirming that MCPNet's gains arise from the proposed design rather than capacity alone. Further analysis of computational efficiency, including FLOPs, parameters, and inference latency, is provided in Appendix D. Fig. 7 shows qualitative results of our method on two benchmark datasets, ISBI2015 and DHA. In each image, blue dots indicate ground truth landmarks, and green dots denote predicted positions.

4.3 Ablation Study

We conducted extensive ablation experiments on the widely used ISBI2015 test1 and DHA validation sets. Due to page limitations, detailed ablation results are provided in Appendix D.

Table 2 compares three coordinate generation strategies in MCP: (1) using only static coordinates, (2) generating $\mathbf{S}_{\text{adapt}}$ directly from $\mathbf{X}$ via GAP, and (3) generating $\mathbf{S}_{\text{adapt}}$ from the static positional feature $\mathbf{E}_{\text{stat}}$ (ours). Localization accuracy remains limited without adaptive coordinate maps. GAP-based generation offers moderate improvement. In contrast, constructing $\mathbf{S}_{\text{adapt}}$ from $\mathbf{E}_{\text{stat}}$ yields the best performance. It performs secondary reasoning based on meaningful positional features $\mathbf{E}_{\text{stat}}$ and provides informative adaptive spatial cues.

Table 3 illustrates the performance impact of different numbers of adaptive

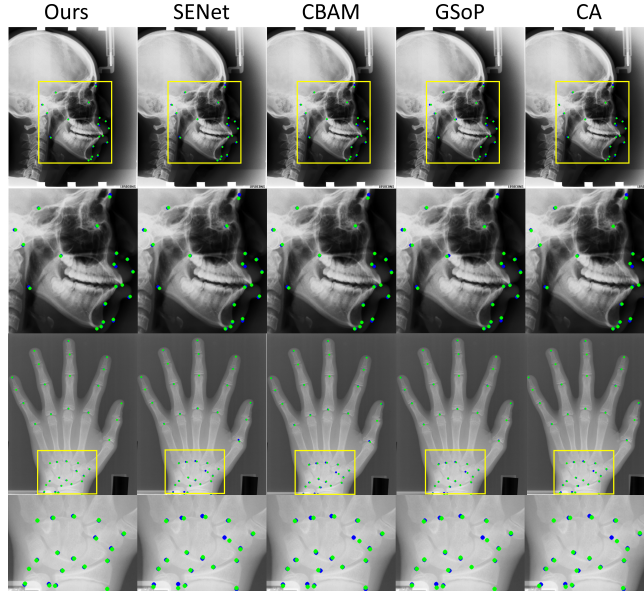


Fig. 7: Visualization results on ISBI2015 and DHA. Blue dots denote ground truth landmarks, and green dots indicate predicted results.

Table 2: Ablation study on adaptive coordinate generation.

Method	ISBI2015		DHA	
	SDR $\uparrow$ (%, < 2mm)	MRE $\downarrow$ (mm)	SDR $\uparrow$ (%, < 2mm)	MRE $\downarrow$ (mm)
No S_{adapt} / E_{adapt}	87.04	1.111	96.33	0.591
S_{adapt} from $\mathbf{X}$ via GAP $\rightarrow E_{\text{adapt}}$	88.35	1.093	96.77	0.566
S_{adapt} from E_{stat} $\rightarrow E_{\text{adapt}}$ (ours)	88.89	1.074	96.94	0.565

Table 3: Impact of number of adaptive coordinate maps.

Number of coordinate map	ISBI2015		DHA		#Params
	SDR $\uparrow$ (%, < 2mm)	MRE $\downarrow$ (mm)	SDR $\uparrow$ (%, < 2mm)	MRE $\downarrow$ (mm)	
10	87.62	1.089	96.79	0.587	22.70M
30	87.61	1.087	96.66	0.576	22.71M
60	88.89	1.074	96.94	0.565	22.73M
100	87.91	1.092	96.51	0.601	22.78M

coordinate maps. The best results were observed with 60 adaptive coordinate maps. Although increasing the number of adaptive coordinate maps adds minimal complexity, raising the count to 100 did not produce substantial performance gains. These findings suggest that while more adaptive coordinate maps can enhance performance, there exists a threshold beyond which additional maps do not contribute to further improvements. Figure 8 shows landmark clusters in Cartesian coordinates and the t-SNE results of our learned adaptive coordinates. The adaptive coordinate space exhibits higher discriminative ability than the Cartesian coordinates on both ISBI and DHA datasets. This result indicates that the adaptive mapping captures richer spatial features.

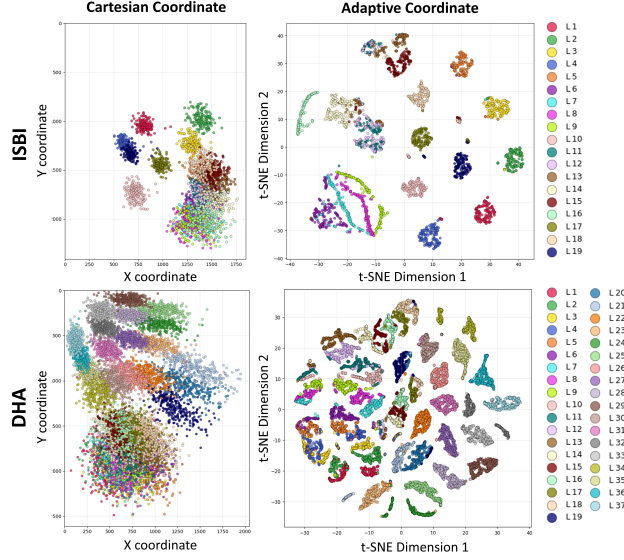


Fig. 8: Landmark distributions on the ISBI and DHA datasets. Left: raw landmark positions in the static Cartesian coordinate space. Right: t-SNE visualization of the trained adaptive coordinate map $\mathbf{S}_{\text{adapt}}$, exhibiting stronger discriminability than the Cartesian space.

5 3D Medical Landmark Detection

For 3D-baseline network, we adopt a 3D U-Net with an encoder-decoder structure and skip connections. The encoder consists of three downsampling stages using 3D max pooling, resulting in four resolution levels with channel widths of 8, 16, 32, and 64 (from shallow to deep). Each stage contains two successive $3 \times 3 \times 3$ convolutions, followed by Instance Normalization and ReLU activation. The decoder mirrors the encoder with three upsampling stages; at each level, the upsampled feature map is concatenated with the corresponding encoder feature through skip connections and refined using the same blocks. A final $1 \times 1 \times 1$ convolution projects the decoder output to the prediction map.

3D Spine. We utilized the volumetric CT scans [28] from the MICCAI CSI 2014 Vertebrae Localization and Identification Challenge. This dataset presents significant challenges, including scoliotic spines, vertebrae fractures, and metal insertions causing severe image artifacts.

Each volume contains up to 26 annotated vertebral landmarks. It consists of 224 CT scans for training and 60 scans for testing. Following Payer et al., we note that the CSI2014 test split contains mislabeled vertebrae in several volumes with localization errors exceeding 100 mm. Therefore, to ensure reliable performance evaluation, we conduct 5-fold cross-validation using only the training set and report the averaged results across folds. The average volume size is

Table 4: Performance comparison on the 3D Spine dataset.

Method	MRE↓	$ID_{rate}(\%)$ ↑	#Params
U-Net [30]	7.3 (-)	82.3	37.50M
SCN [28]	6.2 (-)	86.6	40.77M
SE [14]	7.1 $\pm$.31	82.3	39.89M
CBAM [44]	6.3 $\pm$.22	84.9	40.13M
PE [29]	5.9 $\pm$.44	86.9	42.94M
MCPNet	5.3 $\pm$.37	88.7	41.98M

Table 5: Performance comparison on the 3D Skull dataset.

Method	MRE ↓	SDR (%) ↑				#Param
		2mm	2.5mm	3mm	4mm	
U-Net [30]	1.80 $\pm$.06	75.97	85.41	91.43	96.95	37.50M
SCN [28]	1.62 $\pm$.07	74.45	84.99	92.43	96.36	40.77M
SE [14]	1.97 $\pm$.17	75.33	85.14	93.48	96.13	39.89M
CBAM [44]	1.78 $\pm$.13	75.99	85.81	93.21	96.55	40.13M
SR-Unet [47]	2.01 (-)	76.14	-	92.02	85.84	-
H3DE-Net [16]	1.67 (-)	76.03	86.50	92.31	96.69	-
PE [29]	1.63 $\pm$.12	76.56	88.45	93.33	97.05	42.94M
MCPNet	1.58 $\pm$.08	78.75	89.44	94.38	97.35	41.98M

$512 \times 512 \times 160$ voxels, with an original voxel resolution of $0.34 \times 0.34 \times 2.06$. For our experiments, all input and heatmap volumes were resized to $96 \times 96 \times 192$ voxels. Following the evaluation protocol [28], we report the mean point error (mm) and the identification rate ID_{rate} . A landmark is considered correctly identified if the closest ground-truth landmark corresponds to the correct vertebra and the localization error is within 20 mm .

As shown in Table 4, our MCPNet achieves the lowest localization error with an MRE of 5.3 mm , outperforming all existing methods including SCN and PE. Furthermore, MCPNet obtains the highest identification rate of 88.7%, demonstrating superior robustness in correctly identifying vertebral landmarks. Despite having a comparable number of parameters, MCPNet provides significantly improved accuracy, highlighting the effectiveness of the proposed architecture.

3D Skull. We utilized a publicly available 3D skull CT dataset [16] for anatomical landmark detection. This dataset contains 658 volumetric CT scans, annotated with 14 mandibular molar landmarks, and includes both complete and incomplete cases to simulate realistic clinical conditions. Specifically, the dataset is divided into 458 scans for training, 100 for validation, and 100 for testing. All CT volumes are resized to a fixed spatial resolution of $128 \times 128 \times 64$ voxels, and the corresponding ground-truth heatmaps are resized accordingly to maintain spatial alignment. Following the standard evaluation protocol, we report MRE and SDR, which evaluate the Euclidean distance between predicted and ground-truth landmark coordinates at multiple thresholds.

As reported in Table 5, MCPNet consistently outperforms all competing methods, achieving the lowest MRE of 1.58 mm and the highest SDR at all evaluated thresholds. Compared to the previous state-of-the-art PE and H3DE-Net, MCPNet improves both localization precision and identification reliability. These results validate the effectiveness of the proposed architecture in capturing spatial context for accurate 3D landmark detection. 3D landmark visualizations are shown in Appendix C.

6 Conclusion

In this paper, we proposed a specialized attention network for medical landmark detection focused on directly extracting positional feature from CNN fea-

ture maps and analyses them in the coordinate domain. By introducing MCP along with adaptive mask and coordinate maps, the network efficiently captured diverse and richer anatomical positional cues of the landmark. Moreover, the proposed mask maps were utilized when transforming the analyzed coordinate features back into the original CNN feature space, ensuring smooth reconstruction within the network. Extensive experiments on six public medical landmark datasets demonstrated that the proposed MCPNet outperformed state-of-the-art methods. The proposed framework provides a new direction for integrating coordinate-based features into attention models and could be extended to a wide range of medical imaging tasks that require precise and interpretable localization.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2024-00451472) and this work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2C1008827).

References

1. Chen, R., Ma, Y., Chen, N., Lee, D., Wang, W.: Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 873–881. Springer (2019)
2. Chen, R., Ma, Y., Chen, N., Liu, L., Cui, Z., Lin, Y., Wang, W.: Structure-aware long short-term memory network for 3d cephalometric landmark detection. *IEEE Transactions on Medical Imaging* **41**(7), 1791–1801 (2022)
3. Chen, R., Ma, Y., Liu, L., Chen, N., Cui, Z., Wei, G., Wang, W.: Semi-supervised anatomical landmark detection via shape-regulated self-training. *Neurocomputing* **471**, 335–345 (2022)
4. Dai, C., Wang, Y., Huang, C., Zhou, J., Xu, Q., Xu, M.: A cephalometric landmark regression method based on dual-encoder for high-resolution x-ray image. In: European Conference on Computer Vision. pp. 93–109. Springer (2024)
5. Davison, A.K., Lindner, C., Perry, D.C., Luo, W., Collaborative, M.S.A., Cootes, T.F.: Landmark localisation in radiographs using weighted heatmap displacement voting. In: Computational Methods and Clinical Applications in Musculoskeletal Imaging: 6th International Workshop, MSKI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 6. pp. 73–85. Springer (2019)
6. Fang, P., Zhou, J., Roy, S.K., Ji, P., Petersson, L., Harandi, M.: Attention in attention networks for person retrieval. *IEEE transactions on pattern analysis and machine intelligence* **44**(9), 4626–4641 (2021)
7. Feng, Y., Yang, J., Li, M., Tang, L., Sun, S., Wang, Y.: A bayesian network for simultaneous keyframe and landmark detection in ultrasonic cine. *Medical Image Analysis* p. 103228 (2024)

8. Gao, Z., Xie, J., Wang, Q., Li, P.: Global second-order pooling convolutional networks. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 3024–3033 (2019)
9. Ghesu, F.C., Georgescu, B., Zheng, Y., Grbic, S., Maier, A., Hornegger, J., Comaniciu, D.: Multi-scale deep reinforcement learning for real-time 3d-landmark detection in ct scans. *IEEE transactions on pattern analysis and machine intelligence* **41**(1), 176–189 (2017)
10. Ham, G.S., Oh, K.: Learning spatial configuration feature for landmark localization in hand x-rays. *Electronics* **12**(19), 4038 (2023)
11. Hassani, A., Walton, S., Li, J., Li, S., Shi, H.: Neighborhood attention transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6185–6194 (2023)
12. Heimann, T., Meinzer, H.P.: Statistical shape models for 3d medical image segmentation: a review. *Medical image analysis* **13**(4), 543–563 (2009)
13. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13713–13722 (2021)
14. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
15. Huang, T., Shi, J., Li, J., Wang, J., Du, J., Shi, J.: Involution transformer based u-net for landmark detection in ultrasound images for diagnosis of infantile ddh. *IEEE Journal of Biomedical and Health Informatics* (2024)
16. Huang, Z., Xu, R., Zhou, X., Wei, Y., Wang, S., Sun, X., Li, H., Yao, Q.: H3denet: Efficient and accurate 3d landmark detection in medical imaging. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2025)
17. Huang, Z., Zhao, R., Leung, F.H., Banerjee, S., Lam, K.M., Zheng, Y.P., Ling, S.H.: Landmark localization from medical images with generative distribution prior. *IEEE Transactions on Medical Imaging* (2024)
18. Kang, J., Oh, K., Oh, I.S.: Accurate landmark localization for medical images using perturbations. *Applied Sciences* **11**(21), 10277 (2021)
19. Konda, K., Memisevic, R., Krueger, D.: Zero-bias autoencoders and the benefits of co-adapting features. *arXiv preprint arXiv:1402.3337* (2014)
20. Kwon, H.J., Koo, H.I., Park, J., Cho, N.I.: Multistage probabilistic approach for the localization of cephalometric landmarks. *IEEE Access* **9**, 21306–21314 (2021)
21. Levine, S., Finn, C., Darrell, T., Abbeel, P.: End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research* **17**(39), 1–40 (2016)
22. Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., Yosinski, J.: An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems* **31** (2018)
23. Loshchilov, I.: Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017)
24. Luvizon, D.C., Tabia, H., Picard, D.: Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics* **85**, 15–22 (2019)
25. Luvizon, D.C., Tabia, H., Picard, D.: Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics* **85**, 15–22 (2019)
26. Noothout, J.M., De Vos, B.D., Wolterink, J.M., Postma, E.M., Smeets, P.A., Takx, R.A., Leiner, T., Viergever, M.A., Išgum, I.: Deep learning-based regression and classification for automatic landmark localization in medical images. *IEEE transactions on medical imaging* **39**(12), 4011–4022 (2020)

27. Park, S.: Cephalometric Landmarks Detection using Fully Convolutional Networks. Ph.D. thesis, Seoul National Univ (2017)
28. Payer, C., Stern, D., Bischof, H., Urschler, M.: Integrating spatial configuration into heatmap regression based cnns for landmark localization. *Medical image analysis* **54**, 207–219 (2019)
29. Rickmann, A.M., Roy, A.G., Sarasua, I., Wachinger, C.: Recalibrating 3d convnets with project & excite. *IEEE transactions on medical imaging* **39**(7), 2461–2471 (2020)
30. Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241. Springer (2015)
31. Sharan, L., Romano, G., Brand, J., Kelm, H., Karck, M., De Simone, R., Engelhardt, S.: Point detection through multi-instance deep heatmap regression for sutures in endoscopy. *International journal of computer assisted radiology and surgery* **16**(12), 2107–2117 (2021)
32. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 529–545 (2018)
33. Tan, Z., Feng, J., Lu, W., Yin, Y., Yang, G., Zhou, J.: Multi-task global optimization-based method for vascular landmark detection. *Computerized Medical Imaging and Graphics* **114**, 102364 (2024)
34. Tiulpin, A., Melekhov, I., Saarakkala, S.: Kneel: Knee anatomical landmark localization using hourglass networks. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. pp. 0–0 (2019)
35. Urschler, M., Ebner, T., Stern, D.: Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization. *Medical image analysis* **43**, 23–36 (2018)
36. Urschler, M., Zach, C., Ditt, H., Bischof, H.: Automatic point landmark matching for regularizing nonlinear intensity registration: Application to thoracic ct images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 710–717. Springer (2006)
37. Vaswani, A.: Attention is all you need. *Advances in Neural Information Processing Systems* (2017)
38. Viriyasaranon, T., Ma, S., Choi, J.H.: 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*. pp. 433–443. Springer (2023)
39. Wang, C.W., Huang, C.T., Hsieh, M.C., Li, C.H., Chang, S.W., Li, W.C., Vandaale, R., Marée, R., Jodogne, S., Geurts, P., 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)
40. Wang, C.W., Huang, C.T., Lee, J.H., Li, C.H., Chang, S.W., Siao, M.J., Lai, T.M., Ibragimov, B., Vrtovec, T., Ronneberger, O., et al.: A benchmark for comparison of dental radiography analysis algorithms. *Medical image analysis* **31**, 63–76 (2016)
41. Wang, D., Zhang, S.: Spatial-aware regression for keypoint localization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 624–633 (2024)
42. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 11534–11542 (2020)

43. Wang, X., Bo, L., Fuxin, L.: Adaptive wing loss for robust face alignment via heatmap regression. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6971–6981 (2019)
44. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
45. Yao, Q., Wang, J., Sun, Y., Quan, Q., Zhu, H., Zhou, S.K.: Relative distance matters for one-shot landmark detection. arXiv preprint arXiv:2203.01687 (2022)
46. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: Lift: Learned invariant feature transform. In: European conference on computer vision. pp. 467–483. Springer (2016)
47. Zhang, R., Mo, H., Hu, W., Jie, B., Xu, L., He, Y., Ke, J., Wang, J.: Super-resolution landmark detection networks for medical images. *Computers in Biology and Medicine* **182**, 109095 (2024)
48. Zhao, M., Meng, N., Cheung, J.P.Y., Yu, C., Lu, P., Zhang, T.: Spinehrformer: A transformer-based deep learning model for automatic spine deformity assessment with prospective validation. *Bioengineering* **10**(11), 1333 (2023)
49. Zhu, H., Yao, Q., Xiao, L., Zhou, S.K.: 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. pp. 85–95. Springer (2021)