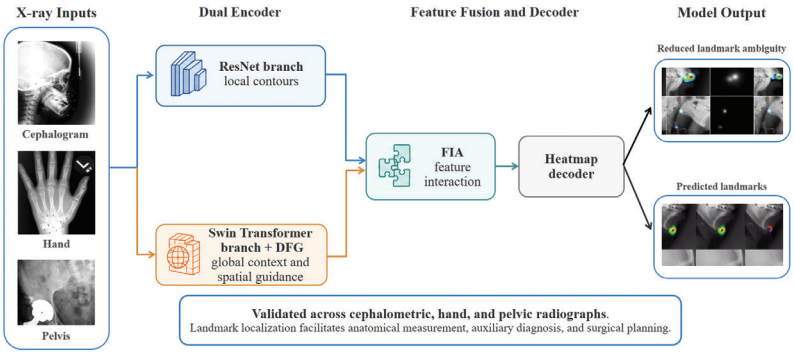
BIO Integration

A Deep Learning System for Automatic Localization of Anatomical Landmarks in X-rays to Assist in Diagnosis and Surgical Planning

Graphical abstract



Highlights

- Res-SwinFusion uses a dual-branch encoder to combine local and global image representations.
- DFG and FIA decrease landmark ambiguity in anatomically similar regions.
- The method achieved strong performance on cephalometric, hand, and pelvic X-ray datasets.
- Further validation will be required before transfer of the model to CT, MRI, or other modalities.

Authors

Hui Zhang, Tengfei Li, Ahmad Alenezi, Xinguo Wang, Qinsheng Hu, Zekun Jiang and Quan Wei

Correspondence

huqs@scu.edu.cn (Q. Hu);
zekun_jiang@163.com (Z. Jiang)

In brief

Res-SwinFusion localizes anatomical landmarks in X-ray images by combining Swin Transformer-derived global context with ResNet-derived local structural detail.

Original Article

BIO Integration

A Deep Learning System for Automatic Localization of Anatomical Landmarks in X-rays to Assist in Diagnosis and Surgical Planning

Hui Zhang[1,2], Tengfei Li[3], Ahmad Alenezi[4], Xinguo Wang[5], Qinsheng Hu[1,6,*], Zekun Jiang[3,*] and Quan Wei[2]

Abstract

Background: Accurate localization of anatomical landmarks is crucial for clinical diagnosis and treatment assessment. However, existing convolutional neural network (CNN)-based methods may result in global spatial information loss and consequent localization failures in the presence of complex anatomical structures or parenchymal abnormalities. Therefore, a method capable of modeling global context while preserving local information is needed.

Methods: Leveraging the Transformer's ability to capture long-range dependencies, we propose a novel landmark localization framework, Res-SwinFusion, which integrates a Swin Transformer and a classical CNN backbone in parallel. To effectively merge their complementary features, we designed a feature interactive aggregation module that fuses semantic representations from both branches. Additionally, we introduced a discrimination feature guidance module to provide pixel-level cues and disambiguate landmark locations. We further analyzed the effects of various Gaussian heatmap settings on convergence.

Results: Res-SwinFusion achieved strong performance across three anatomical landmark localization datasets. The mean radial errors were 1.04 mm and 1.37 mm on two public cephalogram test sets, 0.63 mm on a public hand X-ray dataset, and 1.44 mm on an internal pelvic X-ray dataset. Ablation studies indicated that Transformer-based global modeling, feature interactive aggregation, and discrimination feature guidance each contributed to improved localization accuracy.

Conclusion: The proposed Res-SwinFusion framework offers a solution for anatomical landmark localization with enhanced robustness and precision by combining global contextual modeling and local feature preservation. Code is publicly available at https://github.com/JZK00/Res-SwinFusion.

Keywords

Anatomical landmarks, automatic localization, deep learning, X-ray.

Introduction

Anatomical landmark localization is essential in numerous medical image analysis and applications [1], such as image registration [2], auxiliary diagnosis [3, 4], surgery planning [5], clinical parameter measurements [6], and quantification of various anatomical abnormalities [7]. To achieve clinical significance, a considerable number of landmarks must be positioned. However, manual annotation is time-consuming even for experienced practitioners, and discrepancies among radiologists with different experience levels and backgrounds (inter-observer variation) must also be considered [8]. The main challenge hindering clinical guidance is accurate identification of landmarks, despite varying equipment imaging quality and interindividual anatomical differences [9]. An automatic landmark localization system that is accurate, robust, and rapid would help overcome these issues. In recent years, CNN-based deep learning methods have markedly advanced the development of anatomical landmark automatic detection. In contrast to conventional methods based on prior knowledge [10], template matching [11], or random forests [12], deep learning methods are more accurate and robust [13].

However, current methods have several drawbacks. First, because the large sizes of medical images may exceed memory limitations, performing direct calculations on original images is impractical. Although some approaches have proposed a two-stage framework [14–17], wherein the global stage generates candidate regions,

[1]Department of Orthopedic Surgery and Orthopedic Research Institute, West China Hospital, Sichuan University, Chengdu, Sichuan Province, China

[2]Rehabilitation Medicine Center and Institute of Rehabilitation Medicine, West China Hospital, Sichuan University, Chengdu, Sichuan Province, China

[3]West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, Sichuan Province, China

[4]Radiologic Sciences Department, Allied Health Sciences, Kuwait University, Kuwait City, Kuwait

[5]Winland Intelligent Imaging (Chengdu) Technology Co., Ltd. Chengdu, Sichuan, China

[6]Department of Orthopedic Surgery, Ya'an People's Hospital, Ya'an, China

*Correspondence to: Qinsheng Hu, Sichuan University, West China Hospital, 37# Wainan Guoxue Road, Chengdu, Sichuan Province, China. E-mail: huqs@scu.edu.cn; Zekun Jiang, Sichuan University, West China Hospital, 37# Wainan Guoxue Road, Chengdu, Sichuan Province, China. E-mail: zekun_jiang@163.com

Received: May 4 2026
Revised: June 22 2026
Accepted: July 10 2026
Published Online: August 17 2026

Available at: https://bio-integration.org/

and the refine stage locates landmarks in the cropped high-resolution patches, these methods increase the workflow complexity, because additional training on each separate landmark is necessary. The characteristics of anatomical landmarks, such as generally small scales, high similarity biometrics, and specific spatial position information, pose additional challenges. CNN-based models extract deep features by applying downsampling to facilitate modeling of global information, but they may result in spatial information loss [18]; consequently, local fine-grained features of landmark structures may be discarded. Because structural features at different locations may share similar radian, size, density, and surrounding tissue information, they might be difficult for the model to distinguish. To address these issues, we sought to establish an end-to-end model with global contextual modeling and better local spatial feature representations.

Transformer [19], a sequence transduction model, has achieved widespread successes in natural language processing and can calculate the correlation between any positions in a sequence [20]. The core concept of Transformer is modeling global dependencies between tokens in a sequence through a multi-head self-attention (MSA) mechanism; this process can be interpreted as learning the attention weight distribution and updating feature maps according to these learned weights. Many Transformer-based approaches [20–25] have been proposed for numerous medical image analysis tasks and have achieved superior performance to other CNN-based methods. Therefore, these approaches have excellent potential in medical image applications. TransUNet [21], using Transformer to establish long-range dependence on the deep features extracted by the decoder, has demonstrated feasibility in medical image segmentation. SwinUNet [22], using a symmetric hierarchical structure with Swin Transformer blocks, has demonstrated the superiority of the pure Transformer-based U-shaped network. These studies have shown that the constructed long-term dependencies are beneficial for medical image segmentation [26, 27]. However, the effects of long-term dependencies for anatomical landmark localization have not been well studied.

To address these limitations, we developed ResSwinFusion, a hybrid encoder-decoder framework combining residual CNN blocks with Swin Transformer blocks. The network was designed to preserve local structural detail while modeling long-range spatial dependencies across high-resolution X-ray images. The framework contains two task-specific modules: the discrimination feature guidance (DFG) module, which passes discriminative spatial information across Swin Transformer blocks, and the feature interactive aggregation (FIA) module, which integrates global location-related features with local structural features before heatmap decoding. The main contributions of this framework are threefold. First, the dual encoder captures complementary local and global representations for landmark localization. Second, DFG decreases ambiguity among anatomically similar structures. Third, FIA strengthens feature fusion and increases localization robustness across cephalogram, hand, and pelvic X-ray datasets.

Related work

CNN-based methods for landmark localization

Deep learning methods have achieved substantial advances in the study of anatomical landmark detection over the past decade, and can be categorized primarily into direct coordinate estimation and heatmap regression. Several representative methods based on CNN architectures are reviewed below.

Coordinate-based approaches use CNN-based regression models to directly predict coordinates of landmarks, but they have intrinsic visual ambiguity and highly nonlinear mapping drawbacks [28]. Arik et al. [14] have used CNNs to regress landmark positions, then refined them with a shape-based model. By treating each landmark position as two coordinate variables, Lee et al. [29] have built 38 CNNs to predict 19 landmarks in cephalometric images; however, this framework markedly increases model complexity and is time-consuming. Noothout et al. [30] have used a global-to-local approach, in which the global FCNN regresses the initial locations and the specialized FCNNs refine the landmark coordinates. Despite progress in coordinate regression methods, intrinsic visual ambiguity issues and spatial information loss remain major challenges.

Heatmap-based approaches use the spatial heatmaps to express the probability of a landmark being in a certain location, which can be encoded by a Gaussian function with fixed variance. Pixel labels in heatmaps represent pseudoprobabilities, wherein high responses are considered to provide the location information for the target landmark. Payer et al. [31] have used an end-to-end CNN-based method to regress heatmaps for multiple landmark localization, and additionally have combined a UNet and their Spatial Configuration Net [32] to optimize the method. Thaler et al. [33] have proposed modeling the annotation uncertainty by learning the shape parameters of the landmark heatmaps, and have demonstrated that the location distribution can be modeled through heatmap regression. Zhong et al. [15] have embedded the Attention-Guide mechanism into their two-stage network, which regressed landmark heatmaps from coarse to fine. Chen et al. [34] have fused semantic features with the proposed attentive feature pyramid fusion module at various levels, and combined offset maps in heatmap regression to predict the landmark locations.

Vision transformer

Given the notable success of Transformer [17] in a variety of natural language processing tasks, Transformer-based models have received substantial interest in computer vision and medical image analysis [35]. Dosovitskiy et al. [36] first attempted to split and flatten images into a sequence of patches, and have achieved outstanding performance after pre-training with large-scale data. Swin Transformer [37] was proposed as a hierarchical vision transformer based on the shifted windowing scheme, which effectively decreases

memory consumption and maintains information transmission ability by limiting self-attention computation to non-overlapping windows. Therefore, it can serve as a general backbone encoder, and it has made remarkable achievements in downstream tasks, such as semantic segmentation, object detection, and image classification.

In the fields of human pose estimation and facial landmark detection, several recent studies have used Transformer for keypoint localization. TransPose [38] has been used to reveal the spatial dependencies between keypoints by attention layers built in Transformer. TokenPose [39] embeds each keypoint as a token to learn the statistic constraint relationships between keypoints. Li et al. [40] have built cascaded deformable Transformers to directly regress landmarks.

Various Transformer-based frameworks have been applied in medical image analysis. TransFuse [25] combines the two architectures in a parallel manner, thus confirming that global context modeling of Transformer and low-level feature details of CNNs can be integrated and fused. Lin et al. [23] have used two Swin Transformer backbones to encode features from different scale inputs and designed a Transformer Interactive Fusion module to establish interaction between coarse and fine-grained information. Cao et al. [22] have built a U-shaped network based on Swin Transformer blocks and proposed the patch expanding layer to replace convolution-based up-sampling operation. NnFormer [24] extends the window mechanism to 3D and uses skip attention to replace the concatenation operations in typical skip connections. Zhu et al. [41] have developed a domain-adaptive Transformer model to explore the implicit relevance of landmarks. Swin-CE [42] combines a Transformer encoder and convolutional encoder via skip connections to detect cephalometric landmarks; however, their respective advantages have not been explored. Although previous studies have demonstrated the complementarity between CNNs and Transformers, most existing fusion strategies rely on direct concatenation, weighted summation, or attention-guided fusion mechanisms to integrate heterogeneous representations [25, 43]. These approaches focus primarily on enhancing feature representation capability but pay limited attention to the unique spatial discrimination requirements of anatomical landmark localization. In challenging scenarios involving anatomical deformation, tissue overlap, or highly similar structures, conventional fusion strategies might still lead to localization ambiguity. Therefore, herein, we introduced FIA and DFG modules to strengthen the collaboration between global positional modeling and local discriminative representation learning.

In recent years, CNN-Transformer hybrid architectures have undergone extensive development in medical image analysis. Meanwhile, representative studies such as CGD-TraP, MB-TaylorFormer v2 [44], ESTINet [45], DDMSNet [46], and DBLRNet [47] have further highlighted the importance of multi-scale contextual modeling and long-range dependency learning for robust feature representation across diverse vision tasks. Representative examples include TransFuse, which combines parallel CNN and Transformer branches for local-global feature fusion, Swin-UNet, which exploits hierarchical Transformer representations to capture long-range dependencies, and DS-TransUNet, which further investigates multi-scale Transformer feature interactions [22, 25, 48]. However, these methods were designed primarily to improve region-level semantic representation in medical image segmentation. In contrast, anatomical landmark localization requires distinguishing structures with highly similar appearances but different spatial positions. Therefore, directly applying existing fusion strategies might not fully exploit the complementary characteristics of global contextual information and local structural details.

Motivated by this observation, we propose Res-SwinFusion. Unlike existing CNN-Transformer fusion frameworks, our FIA module introduces branch-specific enhancement mechanisms to strengthen positional awareness in Transformer features and local structural representations in CNN features before feature interaction. Furthermore, the DFG module establishes a discriminative feature guidance pathway to reinforce spatial discrimination during Transformer encoding, thereby improving the recognition of anatomically similar structures.

Materials and methods

Below, we briefly introduce our Res-SwinFusion landmark localization framework, then present the details of the network encoder, fusion module, and decoder.

Overview of the architecture

An overview of our approach is illustrated in **Figure 1**. Specifically, inspired by the structures of TransFuse [25] and ST-UNet [18], we integrate dual backbones based on ResNet and Swin Transformer as the feature encoder, which involves one embedding layer, four feature fusion stages, as well as the proposed DFG module to further enhance the performance of Transformer blocks. Symmetrically, the decoder branch is composed of multi-scale features, as well as the last prediction layer for making multi-landmark predictions. Furthermore, the corresponding features of the encoder and decoder are connected in the manner of UNet [49]. To evaluate computational efficiency, we report model parameters, FLOPs, GPU memory consumption, and inference latency to assess the feasibility of our method for clinical deployment (**Supplementary Table 1**).

Encoder

An input 2D X-ray image $X \epsilon \mathbb{R}^{H \times W \times 3}$, where H and W represent the height and width, is first fed into the dual-branch encoder. In the CNN branch, the backbone network adopts an ImageNet-pretrained ResNet, and the features are progressively downsampled in a hierarchical manner at multiple scales. The feature maps are composed of a series of residual block outputs, wherein the resolution of the nth block output can be expressed as $C_n \epsilon \mathbb{R}^{(2^{n-1}K1)\times\frac{H}{2^{n+1}}\times\frac{W}{2^{n+1}}}$. Here, K1 = 256 and varies with the depth of chose ResNet. Then we denote these outputs as layers $C1$, $C2$, $C3$, and $C4$.

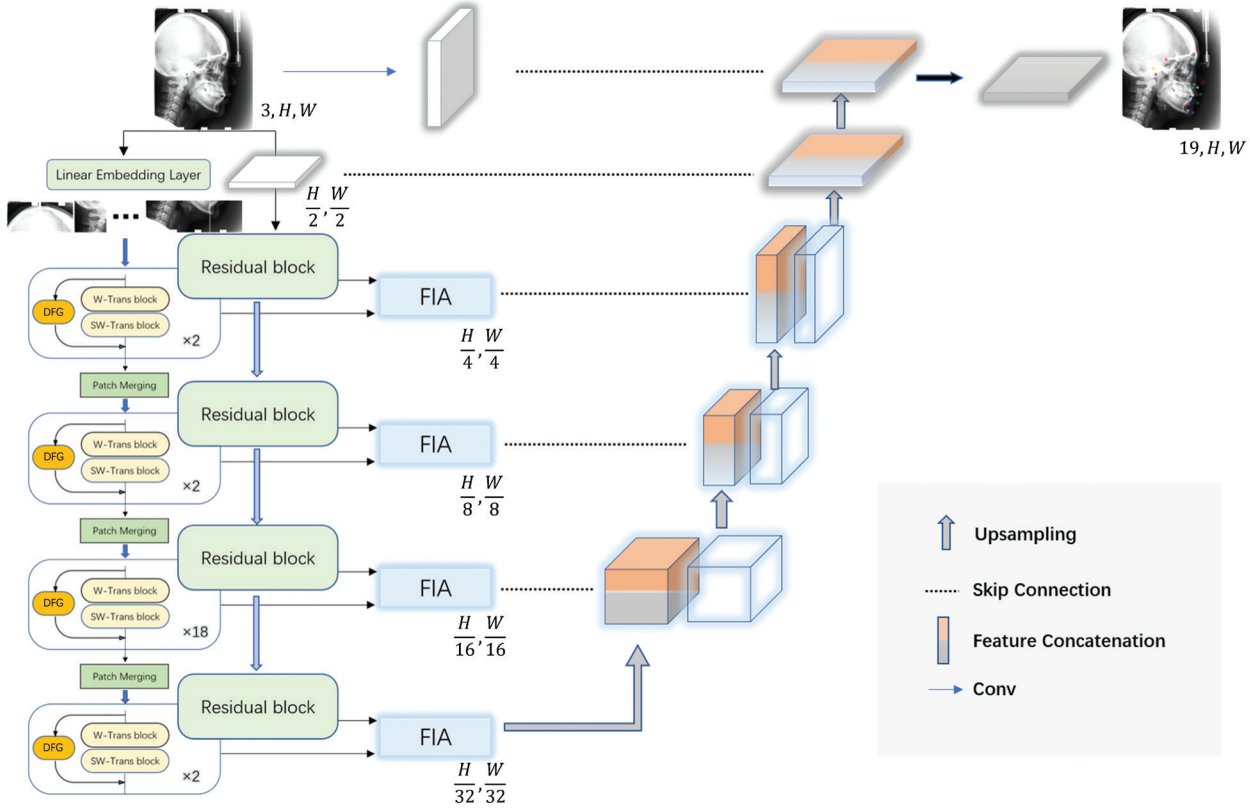


Figure 1 Schematic of Res-SwinFusion (with cephalogram data as the example). The global information and local features are captured from the Swin Transformer encoder and the ResNet backbone, respectively. The encoded information is combined and input into the decoder to produce the localization results.

In addition, according to the design of Swin Transformer architecture, the embedding layer is responsible for dividing the input X into flattened uniform non-overlapping patches $X_p \epsilon \mathbb{R}^{N\times(p^2\times 3)}$, where $N=\frac{HW}{p^2}$ denotes the number of the patch tokens, and p is the patch size. We project these patches into the $K2$ dimensional embedding layer to obtain a high-dimensional tensor $X_e \epsilon \mathbb{R}^{\frac{H}{p}\times\frac{W}{p}\times K2}$. Here, the size of the embedding space $K2$ is set to 128, and p is 4. Subsequently, this tensor is fed into the encoder stacked by Swin Transformer blocks. As in the CNN branch, the hierarchical feature representations with long-term dependencies are captured through four stages, each comprising two blocks. In the last three stages, a patch merging layer is set above the Transformer blocks. We denote the output of each stage as S_n, where n = 1, 2, 3, 4. Specifically, the Swin Transformer block is responsible for feature representation learning, whereas the patch merging layer is used to decrease the feature resolution and increase dimension. Instead of the standard structure, referring to the design of TransUNet [21], we replace the patch merging operation with strided convolution downsampling, which has been demonstrated to aid in modeling object concepts at multiple scales. In addition, inspired by ST-UNet [18], we use a similar residual design named DFG to bridge information between block levels; however, our focus is on how to capture more spatial dependencies to enhance the discriminative representation of landmark structures. Hence, the output size of the four stages is expressed as $\frac{H}{2^{n+1}}\times\frac{W}{2^{n+1}}$, and the dimensions are $2n^{-1}$ $K2$. Below, we elaborate on the described module.

Swin transformer block

By substituting a module with two partitioning configurations (the window-based MSA and shifted window-based MSA) for the MSA module in the standard Transformer block, Swin Transformer builds two consecutive blocks to decrease the calculation of global self-attention (**Figure 2**). The SW-MSA, unlike the W-MSA, establishes cross-window information interaction. Therefore, self-attention can be performed in the proposed windows without additional computation. On the basis of the shifted window partitioning mechanisms, consecutive calculation procedures can be formulated as follows:

$$\begin{aligned}\hat{s}^{l} &= W-\mathrm{MSA}(\mathrm{LN}(s^{l-1}))+s^{l-1}\\ s^{l} &= \mathrm{MLP}(\mathrm{LN}(\hat{s}^{l}))+\hat{s}^{l}\\ \hat{s}^{l+1} &= \mathrm{SW}-\mathrm{MSA}(\mathrm{LN}(s^{l}))+s^{l}\\ s^{l+1} &= \mathrm{MLP}(\mathrm{LN}(\hat{s}^{l+1}))+\hat{s}^{l+1}\end{aligned} \quad (1)$$

where l denotes the layer index; $\hat{s}l$ and $\hat{s}l^{+1}$ denote the outputs of W-MSA and SW-MSA; MLP denotes the multi-layer perceptron; and LN represents layer normalization. Self-attention is calculated as follows:

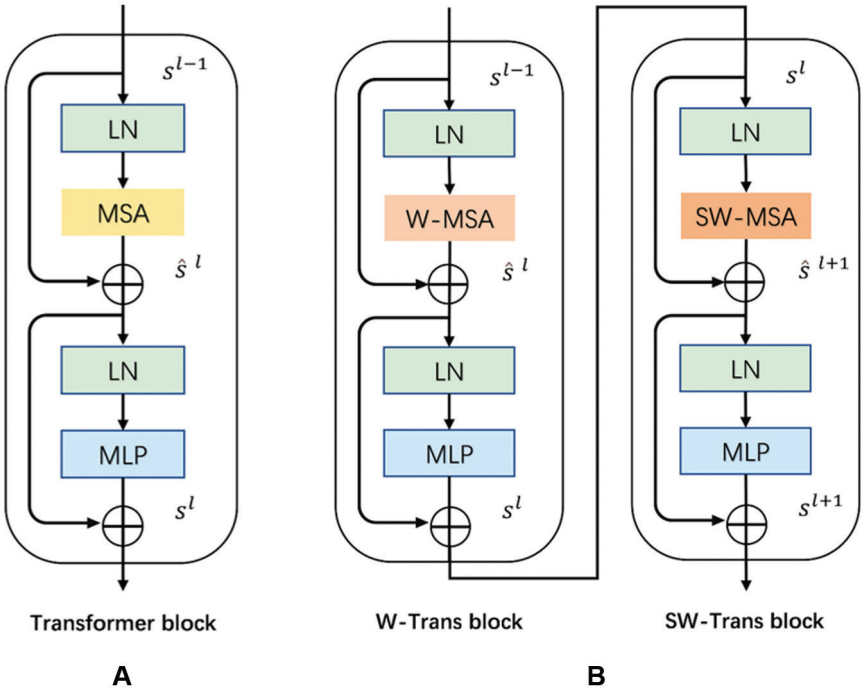


Figure 2 (A) Architecture of the standard Transformer block. (B) Architecture of two cascaded Swin Transformer blocks.

$$\text{Attention}\ (Q,\ K,\ V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{2}$$

where Q, K, and V denote the query, key, and value matrices, respectively; d denotes the size of the query and key; and B denotes the relative position encoding.

Discrimination feature guidance module

Swin Transformer block enables multi-scale feature extraction and effectively decreases memory consumption through the window mechanism. However, it does not fully retain the global modeling abilities of Transformer [18]. For anatomical structures, some specific radians, positions, and shapes must be considered first; however, when viewed at the global level, similar information inevitably leads to ambiguity, and some spatial representation and pixel-level guidance are required to compensate. Therefore, we propose a novel DFG module, which bridges the features before and after computing attention by each Swin Transformer block to deliver the shallow representations to depths, and constructs additional spatial information on the residual edge to further assist global information encoding (**Figure 3**). Consequently, the proportion of target landmarks in the encoding weight is improved, so that the model can establish global attention with more discriminative features and decrease the occurrence of detection ambiguity.

For a given stage $n = (1, 2, 3, 4)$, the input of the W-Trans block $s^{l-1} \epsilon\ \mathbb{R}^{(hw)\times c_2}$ is first reshaped into $s \epsilon\ \mathbb{R}^{c_2\times h\times w}$, where $c_2 = 2n^{-1}\ K2$, $h = \frac{H}{2^{n+1}}$, and $\mathrm{w} = \frac{W}{2^{n+1}}$. We apply a dilated convolution layer to expand the receptive field and compress the channel to $c_2/2$. Subsequently, the global average pooling function is used to obtain the factor of spatial attention. We multiply the above two matrices to balance the sacrifices of spatial resolution. Finally, as in SENet [50], a sigmoid function is used to explicitly model interdependencies among features. Specifically, these operations are denoted as follows:

$$\begin{aligned} s_z &= \text{Conv}^{3\times3}[R(s)] \\ s_v &= R[\text{Avgpool}\ (s_z)] \\ s_k &= \text{Softmax}\ [s_v \otimes R(s_z)] \\ s_a &= \text{Conv}^{1\times1}[s \odot \text{sigmoid}\ [R(s_k)]] \end{aligned} \tag{3}$$

Here, $R(\cdot)$ is the reshape function to facilitate size matching, Avgpool is the global average pooling operation, $\otimes$ denotes matrix multiplication, and $\odot$ denotes the Hadamard product.

Finally, we add the spatial attention features s_a and the output of the second block sl^{+1} to obtain the stage output feature S_n.

$$S_n = R\ [s_a \oplus s^{l+1}] \tag{4}$$

Feature interactive aggregation module

The output information from the dual backbone contains different encoded features. The CNN-based encoder focuses on extracting the local information, and gradually abstracts the features with more advanced semantic information from shallow to deep. However, when objects share similar distribution patterns, confusion during modeling between channel dimensions can lead to large localization errors, because the

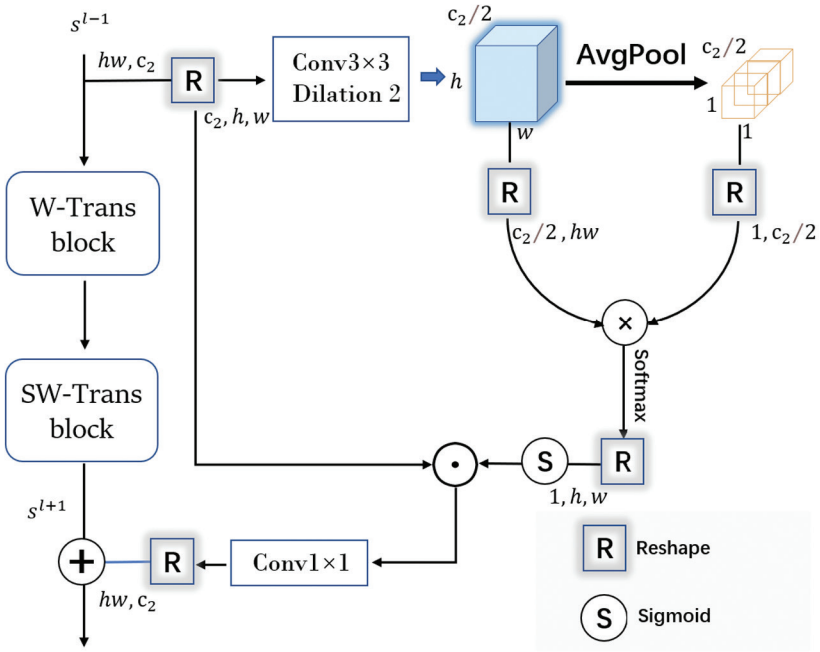


Figure 3 Detailed illustration of our discrimination feature guidance module.

receptive field restricted by the convolution kernel is small in the shallow layer. The Transformer-based encoder focuses on capturing global context connection and can model long-range dependencies; however, the lack of fine-grained information also leads to lower fitting efficiency. As stated above, we sought to establish an end-to-end model combining global context information and local fine-grained features to locate landmarks efficiently and precisely, with effective feature fusion at the core of information exchange. To retain the respective merits of the two branches, we referenced the idea of TransFuse [25] and use an FIA module to promote channel dependence from global information and enhance the spatial coding ability of local features (**Figure 4**). Specifically, the features $C_n \epsilon \mathbb{R}^{c_1 \times h \times w}$ and $S_n \epsilon \mathbb{R}^{c_2 \times h \times w}$ from the dual backbones are fed into the FIA module at the same level. Consequently we obtain multi-level fused feature maps F_n with output resolutions of $\mathbb{R}^{k \times h \times w}$, which are input into the decoder. In detail, the integrated features are obtained through the following steps:

First, for the feature S_n, we use the adaptive average pooling layer to extract the corresponding spatial feature statistics in horizontal and vertical directions, respectively, as shown in the following equation:

$$S_n^{h_i} = \frac{1}{w}\sum_{j=1}^{w} S_n\ (i, j)$$
$$snwi = \frac{1}{h}\sum_{j=1}^{h} S_n\ (i, j) \tag{5}$$

where $S_n^{h_i} \epsilon \mathbb{R}^{c_2 \times h \times w}$, $S_n^{w_i} \epsilon \mathbb{R}^{c_2 \times h \times w}$. Afterward, we multiply these two aggregation tensors to emphasize the spatial position information of landmarks, then perform two convolution operations to model the channel dependence

$$\hat{s}_n = \mathrm{Conv}^{1\times1}(S_n^{h_i} \otimes S_n^{w_i}) \tag{6}$$

Similarly, the spatial attention map generated from the CBAM [51] block is used to promote the spatial coding capability for C_n, which can be denoted as follows:

$$C_n^{spa} = \mathrm{Conv}^{7\times7}\left\langle \mathrm{Avgpool}\ (C_n),\ \mathrm{Maxpool}\ (C_n)\right\rangle$$
$$\hat{C}_n = \mathrm{sigmoid}\ (C_n^{spa}) \odot C_n \tag{7}$$

where Avgpool (C_n), Maxpool $(C_n) \epsilon \mathbb{R}^{1\times h\times w}$ denotes different channel information, and $\langle , \rangle$ is the concatenation operation.

Inspired by the design of BiFusion [25], we also apply the convolutional layers with small convolution kernels to adjust two different encoding features to the same dimension. We then implement element-level multiplication to integrate both fine-grained features, formulated as follows

$$b_n = \mathrm{Conv}^{3\times3}\ [\mathrm{Conv}^{1\times1}(C_n) \odot \mathrm{Conv}^{1\times1}(S_n)] \tag{8}$$

Finally, the refined features are concatenated, and the output feature F_n is formed via a Residual operation:

$$F_n = \mathrm{Residual}\left\langle \hat{C}_n,\ \hat{s}_n,\ b_n \right\rangle \tag{9}$$

Decoder

As in UNet, we place each F_n into the upsampling layer to expand the resolution by a factor of 2 and merge them with the representations from the encoder via skip connection at the same level. After the last concatenation of the encoder and decoder features, we apply linear interpolation upsampling and obtain the output F with a resolution of $\frac{H}{2}\times\frac{W}{2}$. Notably, to obtain a precise representation of shallow features, we build the cascaded blocks with strided convolution

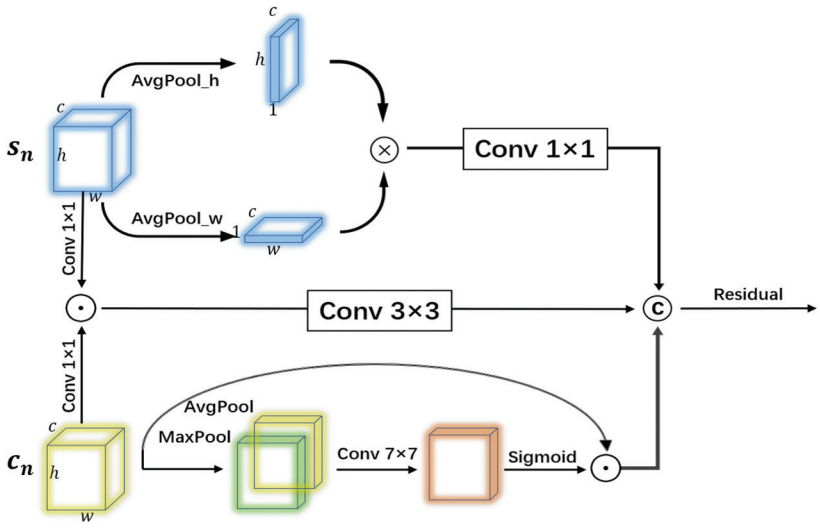


Figure 4 Illustration of our feature interactive aggregation module.

to downsample the input image, then obtain the low-level feature with the same resolution of F. After another skip connection and upsampling, the peer features with the same size of $H \times W$ are calculated by the last feature restoration layer to obtain the final prediction.

Experiments

In this section, we briefly introduce the three benchmark datasets used herein. We then provide detailed experimental settings, post-processing of predicted outputs, and metrics used to evaluate our model's performance.

Datasets

Public cephalogram dataset

A total of 400 cephalometric X-ray images were provided by the ISBI 2015 Grand Challenge [52]. As advised by the organization, the data are divided into three subsets: a training set with 150 images, a Test1 set with 150 images, and a Test2 set with 100 images. Each cephalogram has a resolution of 1935 × 2400 and a pixel spacing of 0.1 mm along both dimensions, and includes 19 landmarks annotated by two experienced specialists. According to the challenge protocol, the meaning of two annotations was used as the ground truth. Furthermore, we rescaled the original images to a size of 768 × 768 and kept the fixed aspect ratio corresponding to its original scale by zero padding.

Public hand dataset

The hand X-ray dataset [31] contains 910 annotated images with an average size of 1563 × 2169. A total of 37 landmarks have been manually annotated on bone joints and fingertips. To unify the pixel distance, we followed the process described by Payer et al. [32], assuming a wrist width of 50 mm, as determined from the two annotated endpoints p and q (the first and fifth points, respectively) at the wrist. Therefore, physical distance can be expressed by multiplication of the pixel spacing by $\frac{50}{\|p-q\|_2}$. We normalized the image size to 512 × 512 and used the setup described by Zhu et al. [53]. We divided the data into 610 training and 300 testing images.

Internal pelvic dataset

This retrospective imaging data acquisition was reviewed and approved by the Ethics Committee of West China Hospital, Sichuan University (No. 2023–1975), with the requirement for written informed consent waived. All patient-related imaging data underwent strict de-identification procedures to remove any personally identifiable information. The original data remained within the hospital's secure server system and were not transferred outside the institution. This dataset consists of 326 postoperative coronal pelvic X-rays from patients who received total hip arthroplasty, and has an average resolution of 4000 × 3200 and a pixel distance of 0.1 mm. Ten landmarks have been annotated by a radiologist and an orthopedic surgeon from the West China Hospital. We used the first 200 images for training and the others as the test set. Notably, some patients' conditions were too complicated (163 cases of osteoarthritis, 54 cases of fracture, and 108 cases of femoral head necrosis) to label and detect. We resized all data to 768 × 768.

Implementation details

Experimental setup

The experiments were conducted on a server with Tesla V100 SXM3-32GB GPUs. We trained all models for 30 epochs by using the Adam optimizer with a batch size of 2. The initial learning rate was set to 0.0001 and was stepped down by a factor of 0.85 every five epochs. No data augmentation was implemented except for resizing, because we

sought to explore the potential of our framework. A total of 45 minutes was required to complete the training procedure. We used Smooth L1 Loss to train the network. The formulation is defined as follows:

$$\text{Loss} = \frac{1}{HWL}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{l=1}^{L}\text{Smooth}L1\ (x_{i,j,l},\ \hat{x}_{i,j,l})$$

$$\text{Smooth}L1\ (x,\ \hat{x}) = \begin{cases} \frac{1}{2}(x-\hat{x})^2, & |x-\hat{x}| < \beta, \\ \beta|x-\hat{x}| - \frac{1}{2}\beta^2, & |x-\hat{x}| \geq \beta \end{cases} \tag{10}$$

where $x_{i,j,l}$ and $\hat{x}_{i,j,l}$ are the pixel intensities of corresponding positions in the ground truth and predicted heatmap, respectively, and β is a threshold value, which we set to 1.5.

Heatmap regression

For landmark localization, heatmap regression has been demonstrated to be a superior method [32]. The coordinate of the target landmark is usually marked by the Gaussian distribution and is formally defined as follows:

$$H(x,\ y) = exp\left(-\frac{(x-\hat{x})^2 + (y-\hat{y})^2}{2\sigma^2}\right) \tag{11}$$

where $(\hat{x}, \hat{y})$ denotes the ground-truth coordinate of a landmark, and σ is used to control the size of the Gaussian distribution. Previous studies have reported the effects of various settings of the parameter σ on experimental results, thus demonstrating that appropriate selection is necessary. Herein, we improved the ground-truth heatmaps from another perspective.

$$G(x,\ y) = \begin{cases} \alpha^{H(x,\ y)}, & H(x,\ y) > 0, \\ 0, & H(x,\ y) \leq 0. \end{cases} \tag{12}$$

Specifically, we added an exponential weight to the Gaussian distribution during the training phase, thus enhancing the importance of landmark pixels. The background pixels were considered invalid, and the exponential increase in target probability values made the network focus on losses from pixels near landmarks. As described in Section V-D.4, the convergence and generalization of the model were improved. Through experimental comparison, we set σ to 12.5 and α to 40 for optimal performance.

Post-processing

During the test phase, we sought to devise a reliable method to identify the most credible coordinate values from the predicted probability maps. To calculate landmark coordinates, we first omitted the heatmap pixels smaller than 0.25 of the maxima, then preserved only the largest connected component and removed the isolated regions, such as noise, thus eliminating some interference. Finally, the averages of positions where the pixel values exceed 0.88 times the maximum value were considered the predicted landmark coordinates.

Evaluation metrics

We used two frequently used metrics to evaluate localization performance: mean radial error (MRE) and successful detection rate (SDR). MRE is defined as the average Euclidean distance between the ground truth and the predicted locations, whereas SDR is the percentage of the predicted landmarks with an MRE less than a given threshold. For the test set with N images and M landmarks, these metrics are calculated as follows:

$$\begin{aligned} \text{MRE} &= \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\left\|x_{n,m} - \hat{x}_{n,m}\right\|_2 \\ \text{SD} &= \frac{1}{\sqrt{NM-1}}\sum_{n=1}^{N}\sum_{m=1}^{M}\left\|x_{n,m} - \text{MRE}\right\|_2 \\ \text{SDR} &= \frac{\left|\{(n,m)\left\|x_{n,m} - \hat{x}_{n,m}\right\|_2 \leq r\}\right|}{NM} \times 100 \end{aligned} \tag{13}$$

where $x_{n,m}$, $\hat{x}_{n,m}$ are the ground truth and estimated location, respectively, and r denotes a certain margin of error, such as 2 mm, 4 mm, or 10 mm.

Results

All experiments were repeated ten times to decrease the influence of training randomness. We report the average performance for quantitative comparisons with current representative methods. For public release of code and checkpoints, we selected the model run with an MRE closest to the repeated-run average. Ablation experiments were then used to quantify the contribution of each component. Representative results from the three datasets are shown in **Figure 5**.

Cephalogram dataset

To verify the efficacy of our algorithm, we present the experimental results on two test datasets and quantitative analysis with the two metrics MRE and SDR. To ensure a fair and unbiased evaluation, we implemented, trained, and tested all comparative baseline models under exactly the same experimental conditions, with identical dataset splits, data preprocessing pipelines, and hardware environments. Our approach clearly outperformed all described methods according to the MRE evaluation criterion and achieved SOTA SDR performance with respect to the other methods (**Table 1**). Specifically, we obtained distance errors of 1.04 mm and 1.37 mm on the two test sets, and corresponding standard deviations of 0.90 and 1.23. Furthermore, we determined the precision ranges with the above thresholds based on the SDR concerning the distance criteria, i.e., 2.0, 2.5, 3.0, and 4.0 mm. On Test1, we achieved the highest SDRs (88.56%, 93.16%, 96.04%, and 98.67%) and obtained equally superior results on Test2 (77.05%, 84.68%, 89.79%, and 95.26%), thereby confirming the reliability of our approach in terms of overall performance. Notably, with

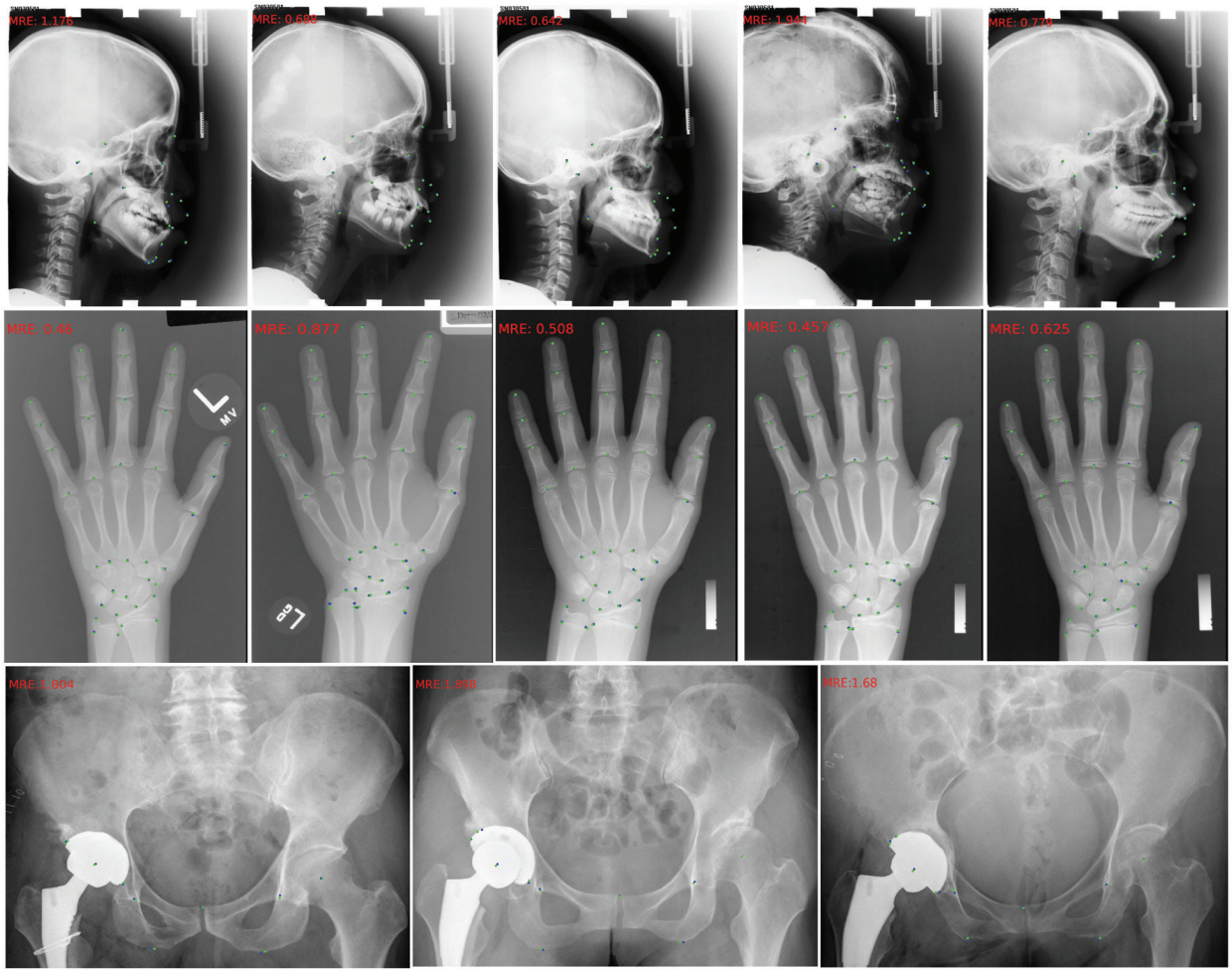


Figure 5 Examples of the qualitative results on cephalogram, hand, and pelvis datasets. Blue points represent the ground truth, and green points represent the predicted landmarks. The MRE of the selected image is marked at the upper left corner.

respect to Thaler et al. [33]'s method, we obtained 1.19% and 1.94% improvements in SDR within 2.0 mm on Test1 and Test2, respectively.

Hand dataset

On this dataset, our method performed well, with an MRE of 0.63 and a 2 mm range SDR of 96.24% and 4 mm of 99.71%. Payer et al. [32] incorporated spatial configuration inside their network and achieved an excellent accuracy of 99.99% within 10 mm and a greatly improved MRE with respect to that of previous methods (**Table 2**). Furthermore, Thaler et al. [33] achieved a best MRE of 0.61 mm through modeling the shape of the target heatmap. With an MRE of 0.63 mm and SDR values of 96.24% (2 mm) and 99.71% (4 mm), our method demonstrates competitive mean error and superior success rates compared with prior approaches.

Internal pelvic dataset

To demonstrate the strengths of our method more objectively, we quantitatively compared it with open-source methods and the typical UNet. Our method achieved an MRE of 1.43 mm and outperformed other methods, according to all metrics (**Table 3**). Notably, with respect to the classic UNet, the improvements in the marked SDRs with our model were 18.84%, 17.96%, 15.53%, and 12.66%, respectively. Moreover, given the inherent complexities of landmark structures, no comparator methods achieved the desired detection performance. In contrast, 96.52% of the landmarks were identified within the error range of 4 mm with our method. Consequently, our method provides advantages even for difficult tasks. However, the localization performance of our approach indicated a margin some distance away from the clinical target. Therefore, in future work, we plan to focus on complex tasks and increase the robustness of the landmark localization system.

Ablation studies

We implemented a set of ablation experiments on the Test1 cephalogram dataset to confirm the contributions of each component. For simplicity, we applied UNet with a ResNet-101 encoder as the baseline network. Furthermore, we used the Swin Transformer backbone with the "base" configuration [37].

Table 1 Comparison of our Method with Prior SOTA Methods on the Cephalometric Dataset

Methods	Test1 Dataset					Test2 Dataset					All Test Datasets				
	MRE ± SD	SDR (%)				MRE ± SD	SDR (%)				MRE	SDR (%)			
		2 mm	2.5 mm	3 mm	4 mm		2 mm	2.5 mm	3 mm	4 mm		2 mm	2.5 mm	3 mm	4 mm
Lindner et al. [9]	1.67	73.68	80.21	85.19	91.74	1.92	66.11	72	77.63	87.43	1.77	70.65	76.93	82.17	89.85
Zhou et al. [24]	1.12 ± 0.88	86.91	91.82	94.88	97.9	1.42 ± 0.84	76	82.9	88.74	94.32	1.24	82.55	88.25	92.42	96.47
Chen et al. [34]	1.17	86.67	92.67	95.54	98.53	1.48	75.5	82.84	88.53	95.05	1.29	82.02	88.74	92.74	97.14
He et al. [16]	1.09 ± 0.31	87.47	92.89	95.15	98.25	1.43 ± 0.22	75.47	83.44	87.84	94.63	1.23	82.67	89.21	92.23	96.80
Payer et al. [32]	-	-	-	-	-	-	-	-	-	-	-	73.33	78.76	83.24	89.75
Zhu et al. [53]	-	-	-	-	-	-	-	-	-	-	1.54	77.79	84.65	89.41	94.93
Thaler et al. [33]	1.07 ± 1.02	87.37	91.86	94.81	97.79	1.38 ± 1.33	75.11	82.53	88.26	94.58	1.19	82.47	88.13	92.19	96.51
Ours	**1.04 ± 0.90**	**88.56**[a]	**93.16**[b]	**96.04**[b]	**98.67**[b]	**1.37 ± 1.23**	**77.05**[b]	**84.68**[a]	**89.79**[a]	**95.26**[b]	**1.17**	**83.96**	**89.77**[a]	**93.54**	**97.31**[a]

- Indicates no corresponding experimental results in the original article.
[a]Indicates P < 0.05.
[b]Indicates P < 0.05.
The best results are marked in bold.

Table 2 Localization Results on the Hand X-ray Dataset and Quantitative Comparison of our Method with Existing Methods

Methods	MRE ± SD	SDR (%)		
		2 mm	4 mm	10 mm
Lindner et al. [9]	0.85 ± 1.01	93.68	98.95	99.94
Urschler et al. [65]	0.80 ± 0.93	92.19	98.46	99.95
Zhu et al. [53][a]	0.84	95.4	99.35	99.75
Payer et al. [32][a]	0.66 ± 0.74	94.99	99.27	99.99
Thaler et al. [33][a]	**0.61 ± 0.67**	95.93	99.54	-
Ours	0.63 ± 0.62	**96.24**[b]	**99.71**	**99.99**

- Indicates no corresponding experimental results in the original article.
[a]Represents open source approaches and indicates no corresponding experimental results in the original article.
[b]Indicates P < 0.05.
The best results are marked in bold.

Effects of the dual backbone encoder

To examine only the effects of introducing the Swin Transformer, we fused the two encoded features via skip connection. The baseline U-Net had the worst performance; however, after aggregation of the global features from Swin Transformer, the performance effectively improved (**Table 4**). Specifically, analysis of the numerical values of all metrics indicated that this structure substantially decreased the MRE, by 0.72. Meanwhile, the SDR within 2.0 mm improved to 72.90%. Therefore, the dual encoder achieved enhanced feature expression by extracting the semantic information with different emphases, even using the simplest connection. For convenience, we use "Swin-U" to denote this network structure.

Effects of the feature interactive aggregation module

Replacing skip connection with our FIA module markedly increased the detection accuracy: the MRE decreased to 1.16 mm, and the SDR in 2.0 mm improved by 13.61%. In addition, we performed a comparison with BiFusion [25], the original motivation underlying this module, to confirm that our fusion model could receive a broader attempt. Our improvements in FIA led to outperformance over BiFusion in all metrics (**Table 4**). Therefore, embedding more global context information and retaining the dual features' respective advantages enabled the characteristic information of anatomical structures to be accurately and effectively learned by the network.

We further examined failure cases from single-backbone models to clarify the roles of the dual encoder and FIA module. When only the CNN backbone was used, similar local structures often distracted the model and produced large errors (**Figure 6A**). This problem became more pronounced in images with poor quality or abnormal anatomy, where the predicted probability map could collapse. When only the Swin Transformer backbone was used, the response was more concentrated but sometimes shifted away from the true landmark (**Figure 6B**). The fusion model decreased both

Table 3 Comparison of Experimental Results on our Internal Pelvic Dataset

Methods	MRE ± SD	SDR (%)			
		2 mm	2.5 mm	3 mm	4 mm
Unet [49]	3.14 ± 2.79	59.55	65.18	74.56	83.86
Zhu et al. [53][a]	2.24 ± 1.89	66.64	70.32	80.18	88.5
Payer et al. [32][a]	2.05 ± 1.77	68.28	72.10	83.07	89.35
Ours	**1.44 ± 1.29**	**78.39**[b]	**83.14**[b]	**90.09**[b]	**96.52**[b]

[a]Represents open source approaches and indicates no corresponding experimental results in the original article.
[b]Indicates $P < 0.05$.
Note: All baseline models listed in this table were evaluated under exactly the same experimental conditions and dataset splits, to ensure a fair comparison.
The best results are marked in bold.

Table 4 Ablation Experiments of Each Module in our Method on the Test1 Dataset

Models	MRE ± SD	SDR (%)			
		2 mm	2.5 mm	3 mm	4 mm
Baseline Unet	2.47 ± 1.90	67.72	74.84	79.43	85.89
Swin-U	1.75 ± 1.59	72.90	78.73	84.29	91.98
Swin-U + BiFusion	1.28 ± 1.44	84.18	90.71	94.01	97.35
Swin-U + FIA	1.16 ± 0.99	86.51	92.15	95.01	98.10
Swin-U + DFG	1.40 ± 1.21	83.27	85.07	90.86	97.34
Swin-U + BiFusion + DFG	1.10 ± 0.97	87.90	92.18	95.42	98.17
Swin-U + FIA + DFG	**1.04 ± 0.90**	**88.56**	**93.16**	**96.04**	**98.67**

The best results are marked in bold.

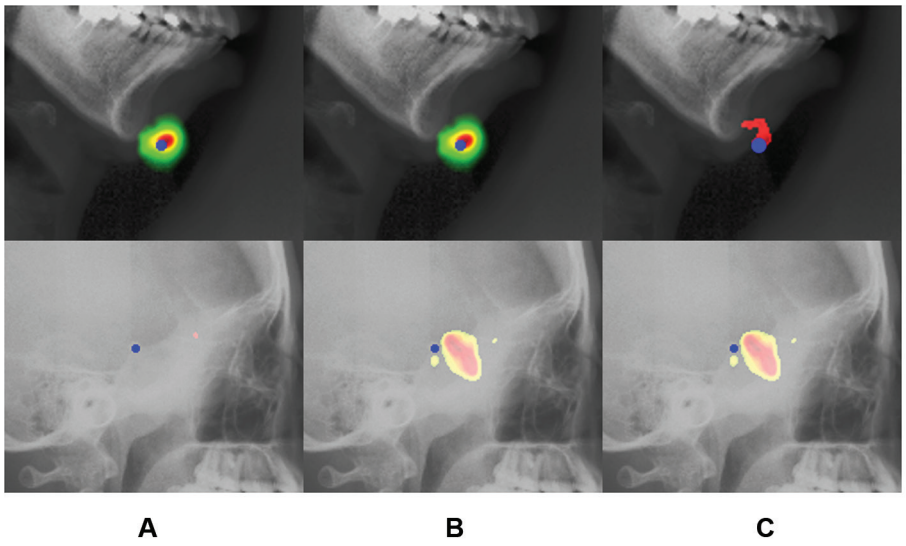


Figure 6 Visualization results of various backbone networks. From left to right: (A) results when the backbone network is only CNNs; (B) results when the backbone network is only Swin Transformer; and (C) results generated by the fusion model.

failure patterns by combining local structural details with global contextual information.

Effects of the discrimination feature guidance module

Similar anatomical structures can mislead the network when full medical images are analyzed. Structures near the target landmark shared similar curvature and surrounding context (**Figure 7A**). Without DFG, these regions had misleading confidence responses, although the target location remained the maximal response (**Figure 7B**). Adding DFG to the Swin Transformer backbone improved the SDR at 2 mm by 2.05% and decreased the MRE to 1.04 mm. Visualization (**Figure 7D**) indicated that DFG suppressed false responses and sharpened the landmark heatmap.

Effects of heatmap ground truth

We conducted an ablation study to examine the heatmap settings. A comparison with the results using the original Gaussian distribution indicated that, when training for only 30 epochs, we achieved enhanced performance (MRE of 1.04; SDR at 2 mm of 88.56%; **Figure 8**). We also analyzed the effect of the

setting of α on the network generalization performance and convergence rate. After use of an exponential weight, the MRE decreased more quickly, and the best results were achieved with α = 40. Therefore, the network achieved accelerated learning of the unique characteristics of anatomical structures by imposing a guide strategy on the landmark pixels.

Case study

Although the model was robust to most complicated anatomic structures, some cases showed large errors. We identified 11 failure cases (radial error of 10 mm) in the cephalogram dataset (**Figure 9**). By analyzing the output probability maps of these cases, we identified several possible sources of error. First, similar radians remained a major cause of ambiguity (e.g., b). Second, some structural changes obscured the landmarks (e.g., a and c); consequently, the network faced difficulty in inferring these special structures from past normal learning. Furthermore, some adjacent landmarks and structures interacted (e.g., the teeth in d). To transcend qualitative observation and rigorously benchmark the model's resilience, we further implemented a quantitative subgroup analysis under these challenging scenarios, categorized into abnormal anatomy, low contrast/quality, and ambiguous landmarks (**Supplementary Table 2**).

Comparison with existing CNN-transformer fusion methods

Unlike TransFuse, which uses the BiFusion module to directly integrate CNN and Transformer representations through attention-guided feature fusion, our FIA module adopts branch-specific enhancement strategies before feature interaction. Specifically, global Transformer features are refined by using positional statistics extracted along horizontal and vertical directions, and CNN features are enhanced through spatial attention to preserve local structural details. This design explicitly addresses the localization ambiguity frequently encountered in anatomical landmark detection.

Furthermore, our DFG module differs from conventional attention mechanisms by introducing an additional discriminative guidance pathway across Swin Transformer blocks. Rather than simply recalibrating feature responses, DFG continually propagates shallow spatial information to deeper representations and supplements Transformer features with enlarged receptive-field information via dilated convolution. As demonstrated in the ablation study, this design increased localization accuracy in regions containing anatomically similar structures and contributed to the overall robustness of the framework.

Discussion

Recent literature has indicated that anatomical landmark localization is moving beyond conventional CNN heatmap regression. Studies have reported high-resolution X-ray Transformer regression networks, direct-coordinate vision Transformer approaches, and broader evidence syntheses for AI-assisted cephalometric analysis [54–56], thus supporting the field's shift toward models that preserve spatial precision while using global context to resolve ambiguous anatomy.

This context clarifies the role of Res-SwinFusion. The model does not simply replace CNNs with a Transformer branch. Instead, it keeps local structural features from ResNet, obtains broader contextual information from Swin Transformer blocks, and then aligns both streams through FIA and DFG. Recent work on explicit anatomical priors and 3D cephalometric localization has also emphasized that anatomical geometry should be built into the model design rather than left to implicit feature learning alone [57, 58]. Our results were consistent with that direction, because the largest gains occurred in cases in which similar local structures otherwise confused the probability maps.

Our model is nevertheless limited by its imaging scope. All experiments used two-dimensional X-ray images, including cephalometric, hand, and postoperative pelvic radiographs. Therefore, the method should not be assumed to be directly transferable to CT, cone-beam CT, MRI, ultrasound, or dynamic fluoroscopy images. CBCT and CT landmark studies require volumetric reasoning, slice-to-volume consistency, anisotropic voxel handling, and different annotation protocols [59–61]. MRI applications add further challenges, because tissue contrast, acquisition planes, motion, and sequence-dependent intensity profiles substantially differ from those of radiographs [62]. Cross-modality transfer is likely to require modality-specific pretraining, 2.5D or 3D backbones, uncertainty estimation, and external validation rather than direct reuse of the current X-ray model.

A second limitation is that the internal pelvic dataset was retrospective and relatively small. Although it included difficult postoperative cases, it came from a single institutional workflow. Future studies should therefore test the model on multi-center datasets with different scanners, acquisition protocols, disease distributions, implants, and image quality levels. Generalist medical foundation models and 2D/3D localization frameworks provide promising tools for such adaptation, but they still require task-specific calibration and clinical validation before deployment [63, 64].

Finally, clinical utility depends on more aspects than landmark error alone. Prospective studies should measure whether automated localization decreases annotation time, improves inter-observer consistency, and changes downstream diagnostic or surgical planning decisions. The model should also report uncertainty for low-quality images, unusual anatomy, and landmarks near overlapping structures. These safeguards will be important before the framework can be used as a decision-support tool in routine practice.

Conclusion

This study presents Res-SwinFusion, a landmark localization network integrating a ResNet branch with a Swin Transformer branch. The model combines local structural detail with long-range contextual modeling, thus improving discrimination between anatomically similar landmarks. The DFG module strengthens spatial guidance

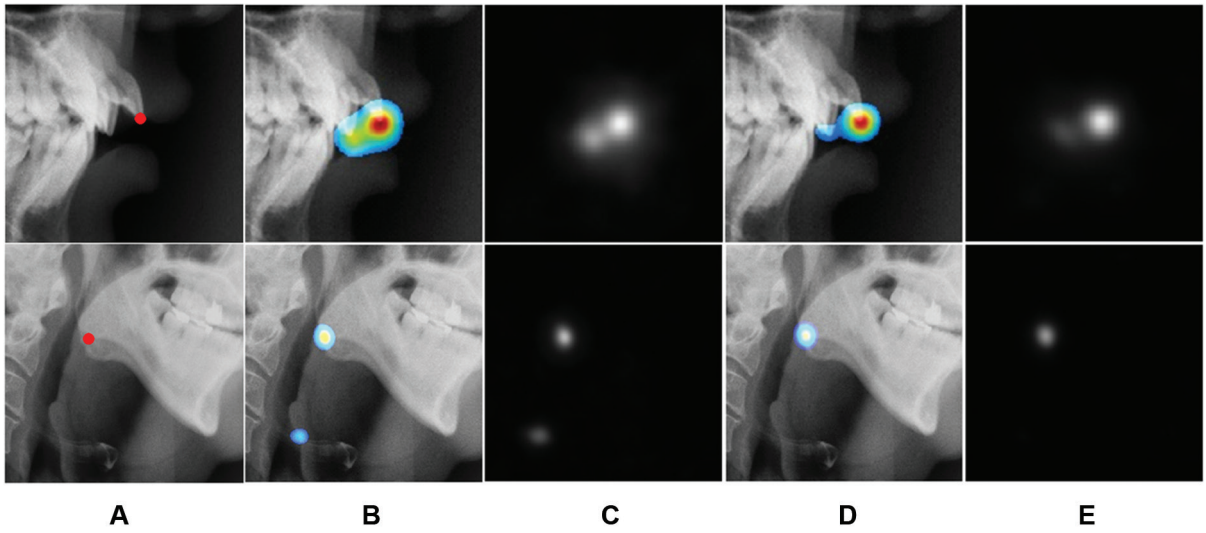


Figure 7 Qualitative analysis and visualization before and after use of DFG. From left to right: (A) the ground truth of the example landmark; (B, C) results without use of DFG; (D, E) results with use of DFG. Additionally, (B) and (D) show the visualization image of the probability maps in the selected region; (C) and (E) present the corresponding output heatmaps.

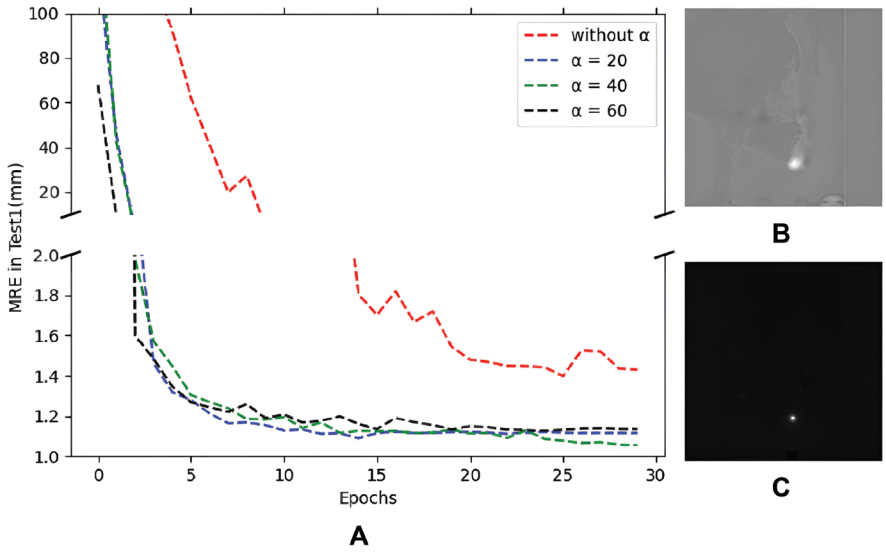


Figure 8 (A) Relationship curves between MRE and epochs on the cephalometric Test1 dataset. Differently colored lines represent different settings of α. (B) Detected probability maps when the training epoch is 5. Top: without α. Bottom: α = 40.

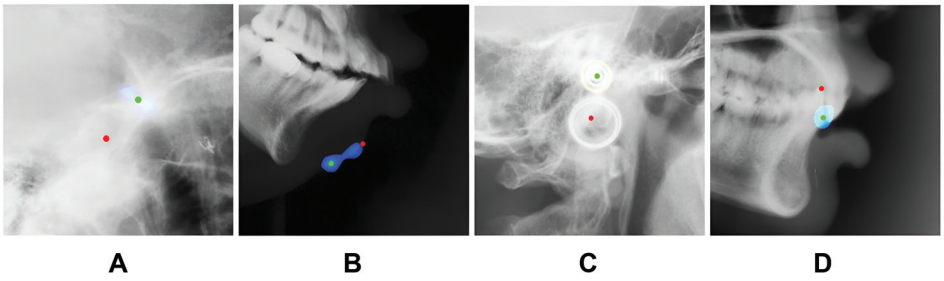


Figure 9 Illustration of detected landmarks with large MREs. Red points indicate the ground truth, and the predicted coordinates are represented by green points. Similarly, the probability maps are shown with the examples.

across Transformer blocks, whereas the FIA module fuses local and global representations before heatmap decoding. Experiments on cephalometric, hand, and pelvic X-ray datasets supported the effectiveness of this design. Future work should validate the method prospectively, test it across additional imaging modalities, and quantify its effects on clinical measurement and planning workflows.

Data availability statement

The datasets generated or analyzed during the study are available in the ISBI 2015 Grand Challenge. Code is publicly available at https://github.com/JZK00/Res-SwinFusion.

Ethical statement

This study used two publicly available, de-identified radiographic datasets and one retrospective institutional pelvic radiograph dataset obtained from West China Hospital, Sichuan University. The retrospective use of the institutional clinical imaging data was reviewed and approved by the Ethics Committee of West China Hospital, Sichuan University (Approval No. 2023-1975). All institutional images were de-identified before analysis. The requirement for written informed consent was waived by the Ethics Committee because of the retrospective nature of the study and the use of de-identified data. The study was conducted in accordance with applicable institutional guidelines and the Declaration of Helsinki.

Author contributions

Conceptualization: Zekun Jiang, Qinsheng Hu, Quan Wei
Data curation: Hui Zhang, Tengfei Li
Formal analysis: Hui Zhang, Tengfei Li
Funding acquisition: Qinsheng Hu, Zekun Jiang, Quan Wei
Investigation: Hui Zhang, Tengfei Li, Ahmad Alenezi
Methodology: Zekun Jiang, Tengfei Li, Xinguo Wang
Project administration: Hui Zhang, Qinsheng Hu
Resources: Hui Zhang, Tengfei Li
Software: Hui Zhang, Tengfei Li
Supervision: Qinsheng Hu, Quan Wei
Validation: Zekun Jiang, Tengfei Li
Visualization: Tengfei Li, Hui Zhang, Ahmad Alenezi
Writing-original draft: Hui Zhang
Writing-review & editing: Hui Zhang, Tengfei Li, Zekun Jiang.

Funding

The research was supported by the Key Research Project of Science & Technology Department of Sichuan Province, China (2024YFFK0041); the Science and Technology Program of Xinjiang Uyghur Autonomous Region (2025E01032); and Chengdu Science and Technology Program (2024-YF05-00632-SN, 2026-XT00-00016-GX).

Conflict of interest

The authors declare no competing interests.

Supplementary materials

Supplementary Material can be downloaded from https://bio-integration.org/wp-content/uploads/2026/08/bioi-20260075_Supplemental.pdf.

References

[1] Liu Y, Song Y, Du Y, Qin C, Xu T. From language to world: bridging LLMs and world models for intelligent surgery. Innov Inform. 2026;2(2):100040. [DOI: 10.59717/j.xinn-inform.2026.100040]

[2] Alam F, Rahman SU, Ullah S, Gulati K. Medical image registration in image guided surgery: issues, challenges and research opportunities. Biocybern Biomed Eng. 2018;38(1):71-89. [DOI: 10.1016/j.bbe.2017.10.001]

[3] Bier B, Goldmann F, Zaech JN, Fotouhi J, Hegeman R, et al. Learning to detect anatomical landmarks of the pelvis in X-rays from arbitrary views. Int J Comput Assist Radiol Surg. 2019;14(9):1463-73. [PMID: 31006106 DOI: 10.1007/s11548-019-01975-5]

[4] Štern D, Payer C, Lepetit V, Urschler M. Automated age estimation from hand MRI volumes using deep learning. In: Ourselin S, Joskowicz L, Sabuncu M, Unal G, Wells W, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Cham: Springer; 2016. pp. 194-202. [DOI: 10.1007/978-3-319-46723-8_23]

[5] Zheng Y, John M, Liao R, Boese J, Kirschstein U, et al. Automatic aorta segmentation and valve landmark detection in C-arm CT: application to aortic valve implantation. In: Jiang T, Navab N, Pluim JPW, Viergever MA, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010. Berlin, Heidelberg: Springer; 2010. pp. 476-83. [DOI: 10.1007/978-3-642-15705-9_58]

[6] Huang Y, Fan F, Syben C, Roser P, Mill L, et al. Cephalogram synthesis and landmark detection in dental cone-beam CT systems. Med Image Anal. 2021;70:102028. [PMID: 33744833 DOI: 10.1016/j.media.2021.102028]

[7] Ibragimov B, Likar B, Pernuš F, Vrtovec T. Shape representation for efficient landmark-based segmentation in 3-D. IEEE Trans Med Imaging. 2014;33(4):861-74. [PMID: 24710155 DOI: 10.1109/TMI.2013.2296976]

[8] Kamoen A, Dermaut L, Verbeeck R. The clinical significance of error measurement in the interpretation of treatment results. Eur J Orthod. 2001;23(5):569-78. [PMID: 11668876 DOI: 10.1093/ejo/23.5.569]

[9] Lindner C, Wang CW, Huang CT, Li CH, Chang SW, et al. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. Sci Rep. 2016;6:33581. [PMID: 27645567 DOI: 10.1038/srep33581]

[10] Grau V, Alcañiz M, Juan MC, Monserrat C, Knoll C. Automatic localization of cephalometric landmarks. J Biomed Inform. 2001;34(3):146-56. [PMID: 11723697 DOI: 10.1006/jbin.2001.1014]

[11] El-Feghi I, Sid-Ahmed MA, Ahmadi M. Automatic localization of craniofacial landmarks for assisted cephalometry. Pattern Recognit. 2004;37(3):609-21. [DOI: 10.1016/j.patcog.2003.09.002]

[12] Mirzaalian H, Hamarneh G. Automatic globally-optimal pictorial structures with random decision forest based likelihoods for cephalometric X-ray landmark detection. In Automatic Cephalometric X-ray Landmark Detection Challenge 2014, in Conjunction with IEEE International Symposium on Biomedical Imaging; 2014. pp. 1-12.

[13] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88. [PMID: 28778026 DOI: 10.1016/j.media.2017.07.005]

[14] Arık SÖ, Ibragimov B, Xing L. Fully automated quantitative cephalometry using convolutional neural networks. J Med Imaging (Bellingham). 2017;4(1):014501. [PMID: 28097213 DOI: 10.1117/1.JMI.4.1.014501]

[15] Zhong Z, Li J, Zhang Z, Jiao Z, Gao X. An attention-guided deep regression model for landmark detection in cephalograms. In: Shen D, et al., editors. Medical Image Computing and Computer Assisted Intervention. Cham: Springer; 2019. pp. 540-8. [DOI: 10.1007/978-3-030-32226-7_60]

[16] He T, Yao J, Tian W, Yi Z, Tang W, et al. Cephalometric landmark detection by considering translational invariance in the two-stage framework. Neurocomput. 2021;464:15-26. [DOI: 10.1016/j.neucom.2021.08.042]

[17] Zeng M, Yan Z, Liu S, Zhou Y, Qiu L. Cascaded convolutional networks for automatic cephalometric landmark detection. Med Image Anal. 2021;68:101904. [PMID: 33290934 DOI: 10.1016/j.media.2020.101904]

[18] He X, Zhou Y, Zhao J, Zhang D, Yao R, et al. Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE Transact Geosci Remote Sens. 2022;60:1-15. [DOI: 10.1109/TGRS.2022.3144165]

[19] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention is all you need. In: von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, editors. Advances in neural information processing systems, Vol. 30. Red Hook, NY: Curran Associates, Inc.; 2017. pp. 3104-14.

[20] Li J, Wang W, Chen C, Zhang T, Zha S, et al. TransBTSV2: Towards better and more efficient volumetric segmentation of medical images. arXiv preprint arXiv:2201.12785. 2022.

[21] Chen J, Lu Y, Yu Q, Luo X, Zhou Y, et al. TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306. 2021.

[22] Cao H, Wang Y, Chen J, Jiang D, Zhang X, et al. Swin-Unet: Unet-like pure transformer for medical image segmentation. In: Karlinsky L, Michaeli T, Nishino K, editors. ECCV 2022: Proceedings of the Computer Vision – ECCV 2022 Workshops, Part III; 2022 Oct 23-27; Tel Aviv, Israel. Cham: Springer; 2023. pp, 205-18. [DOI: 10.1007/978-3-031-25066-8_9]

[23] Lin A, Chen B, Xu J, Zhang Z, Lu G. DS-TransUNet: dual swin transformer U-Net for medical image segmentation. IEEE Trans Instrum Meas. 2022;71:1-15. [DOI: 10.1109/TIM.2022.3178991]

[24] Zhou HY, Guo J, Zhang Y, Yu L, Wang L, et al. nnFormer: interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201. 2021.

[25] Zhang Y, Liu H, Hu Q. TransFuse: fusing transformers and CNNs for medical image segmentation. In: de Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, et al., editors. MICCAI 2021: Medical Image Computing and Computer Assisted Intervention. Cham: Springer; 2021. pp. 14-24. [DOI: 10.1007/978-3-030-87193-2_2]

[26] Hong W, Kim SM, Choi J, Paeng JY, Mun JH, et al. Deep reinforcement learning using a multi-scale agent with a normalized reward strategy for automatic cephalometric landmark detection. In: 2023 4th International Conference on Big Data Analytics and Practices (IBDAP). Bangkok, Thailand: IEEE; 2023. pp. 1-6. [DOI: 10.1109/IBDAP58581.2023.10271989]

[27] Zhu Z, Gu X, Dong L, Liu Y, Wang Y, et al. An ensemble-based deep learning method through multi-scale cross-attention training for cephalometric landmark localization on lateral X-ray images. 2025. [DOI: 10.21203/rs.3.rs-6105085/v1]

[28] Lu G, Zhang Y, Kong Y, Zhang C, Coatrieux JL, et al. Landmark localization for cephalometric analysis using multiscale image patch-based graph convolutional networks. IEEE J Biomed Health Inform. 2022;26(7):3015-24. [PMID: 35259123 DOI: 10.1109/JBHI.2022.3157722]

[29] Lee H, Park M, Kim J. Cephalometric landmark detection in dental x-ray images using convolutional neural networks. Proc. SPIE 10134, Medical Imaging 2017: Computer-Aided Diagnosis, 101341W; 2017. pp. 494-9. [DOI: 10.1117/12.2255870]

[30] Noothout JMH, De Vos BD, Wolterink JM, Postma EM, Smeets PAM, et al. Deep learning-based regression and classification for automatic landmark localization in medical images. IEEE Trans Med Imaging. 2020;39(12):4011-22. [PMID: 32746142 DOI: 10.1109/TMI.2020.3009002]

[31] Payer C, Štern D, Bischof H, Urschler M. Regressing heatmaps for multiple landmark localization using CNNs. In: Ourselin S, Joskowicz L, Sabuncu M, Unal G, Wells W, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Cham: Springer; 2016. pp. 230-8. [DOI: 10.1007/978-3-319-46723-8_27]

[32] Payer C, Štern D, Bischof H, Urschler M. Integrating spatial configuration into heatmap regression based CNNs for landmark localization. Med Image Anal. 2019;54:207-19. [PMID: 30947144 DOI: 10.1016/j.media.2019.03.007]

[33] Thaler F, Payer C, Urschler M, Štern D. Modeling annotation uncertainty with gaussian heatmaps in landmark localization. J Mach Learn Biomed Imaging. 2021;14:1-27. [DOI: 10.59275/j.melba.2021-77a7]

[34] Chen R, Ma Y, Chen N, Lee D, Wang W. Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting. In: Shen D, et al., editors. International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019. Cham: Springer; 2019. pp. 873-81. [DOI: 10.1007/978-3-030-32248-9_97]

[35] Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, et al. Transformers in medical imaging: a survey. Med Image Anal. 2023;88:102802. [DOI: 10.1016/j.media.2023.102802]

[36] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, et al. An image is worth 16x16 words: transformers for image recognition at scale. International Conference on Learning Representations. arXiv:2010.11929v2. 2021. [DOI: 10.48550/arXiv.2010.11929]

[37] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE; 2021. pp. 9992-10002. [DOI: 10.1109/ICCV48922.2021.00986]

[38] Yang S, Quan Z, Nie M, Yang W. TransPose: keypoint localization via transformer. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE; 2021. pp. 11802-12. [DOI: 10.1109/ICCV48922.2021.01159]

[39] Li Y, Zhang S, Wang Z, Yang S, Zhou E. TokenPose: learning keypoint tokens for human pose estimation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE; 2021, pp. 11313-22. [DOI: 10.1109/ICCV48922.2021.01112]

[40] Li H, Guo Z, Rhee SM, Han S, Han JJ. Towards accurate facial landmark detection via cascaded transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE; 2022. pp. 4176-85. [DOI: 10.1109/CVPR52688.2022.00414]

[41] Zhu H, Yao Q, Zhou SK. DATR: domain-adaptive transformer for multi-domain landmark detection. arXiv:2203.06433v1. 2022. [DOI: 10.48550/arXiv.2203.06433]

[42] Ao Y, Hong W. Swin transformer combined with convolutional encoder for cephalometric landmarks detection. In: 2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). Chengdu, China: IEEE; 2021. pp. 184-7. [DOI: 10.1109/ICCWAMTIP53232.2021.9674147]

[43] Shaker A, Maaz M, Rasheed H, Khan S, Yang MH, et al. UNETR++: delving into efficient and accurate 3D medical image segmentation.

IEEE Trans Med Imaging. 2024;43(9):3377-90. [PMID: 38722726 DOI: 10.1109/TMI.2024.3398728]

[44] Jin Z, Qiu Y, Zhang K, Li H, Luo W. MB-TaylorFormer V2: improved multi-branch linear transformer expanded by Taylor formula for image restoration. IEEE Trans Pattern Anal Mach Intell. 2025;47(7):5990-6005. [PMID: 40208767 DOI: 10.1109/TPAMI.2025.3559891]

[45] Zhang K, Li D, Luo W, Ren W, Liu W. Enhanced spatio-temporal interaction learning for video deraining: faster and better. IEEE Trans Pattern Anal Mach Intell. 2023;45(1):1287-93. [PMID: 35130145 DOI: 10.1109/TPAMI.2022.3148707]

[46] Zhang K, Li R, Yu Y, Luo W, Li C. Deep dense multi-scale network for snow removal using semantic and depth priors. IEEE Trans Image Process. 2021;30:7419-7431. [DOI: 10.1109/TIP.2021.3104166]

[47] Zhang K, Luo W, Zhong Y, Ma L, Liu W, et al. Adversarial spatio-temporal learning for video deblurring. IEEE Trans Image Process. 2019;28:291-301. [DOI: 10.1109/TIP.2018.2867733]

[48] Lin A, Chen B, Xu J, Zhang Z, Lu G, et al. DS-transUNet: dual swin transformer U-net for medical image segmentation. IEEE Trans Instrum Meas. 2022;71:1-15. [DOI: 10.1109/TIM.2022.3178991]

[49] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer; 2015. pp. 234-41. [DOI: 10.1007/978-3-319-24574-4_28]

[50] Hu J, Shen L, Sun G, Albanie S, Wu E. Squeeze-and-excitation networks. Proc. IEEE Conf Comput Vis Pattern Recognit; 2018. pp. 7132-41. [DOI: 10.48550/arXiv.1709.01507]

[51] Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision – ECCV 2018. Cham: Springer; 2018. pp. 3-19. [DOI: 10.1007/978-3-030-01234-2_1]

[52] Wang CW, Huang CT, Lee JH, Li CH, Chang SW, et al. A benchmark for comparison of dental radiography analysis algorithms. Med Image Anal. 2016;31:63-76. [PMID: 26974042 DOI: 10.1016/j.media.2016.02.004]

[53] Zhu H, Yao Q, Xiao L, Zhou SK. You only learn once: universal anatomical landmark detection. In: de Bruijne M, et al. Medical Image Computing and Computer Assisted Intervention. Cham: Springer; 2021. pp. 85-95. [DOI: 10.1007/978-3-030-87240-3_9]

[54] Zhou J, Wang Y, Huang C, Dai C, Tan C. CeLR: a transformer-based regression network for accurate cephalometric landmark detection in high-resolution X-ray imaging. IEEE Trans Med Imaging. 2026;45(5):2283-94. [DOI: 10.1109/TMI.2026.3652170]

[55] Laitenberger F, Scheuer HT, Scheuer HA, Lilienthal E, You S, et al. Cephalometric landmark detection using vision transformers with direct coordinate prediction. J Craniomaxillofac Surg. 2025;53(9):1518-29. [PMID: 40603150 DOI: 10.1016/j.jcms.2025.05.021]

[56] Polizzi A, Leonardi R. Automatic cephalometric landmark identification with artificial intelligence: an umbrella review of systematic reviews. J Dent. 2024;146:105056. [PMID: 38729291 DOI: 10.1016/j.jdent.2024.105056]

[57] Joham SJ, Hadzic A, Urschler M. Implicit is not enough: explicitly enforcing anatomical priors inside landmark localization models. Bioengineering (Basel). 2024;11(9):932. [PMID: 39329674 DOI: 10.3390/bioengineering11090932]

[58] Song C, Jeong Y, Huh H, Park JW, Paeng JY, et al. Multi-scale 3D cephalometric landmark detection based on direct regression with 3D CNN architectures. Diagnostics (Basel). 2024;14(22):2605. [PMID: 39594271 DOI: 10.3390/diagnostics14222605]

[59] Gillot M, Miranda F, Baquero B, Ruellas A, Gurgel M, et al. Automatic landmark identification in cone-beam computed tomography. Orthod Craniofac Res. 2023;26(4):560-7. [PMID: 36811276 DOI: 10.1111/ocr.12642]

[60] Baldini B, Rubiu G, Serafin M, Bologna M, Facchi GM, et al. Automated 3D cephalometry: a lightweight V-net for landmark localization on CBCT. Comput Med Imaging Graph. 2026;128:102700. [PMID: 41519031 DOI: 10.1016/j.compmedimag.2026.102700]

[61] Deitermann M, Pankert T, Jaganathan S, Röhrle O, Hölzle F, et al. Automated detection of mandibular landmarks in CT data using a dual-input approach in a two-stage design. Comput Methods Programs Biomed. 2026;273:109113. [PMID: 41086721 DOI: 10.1016/j.cmpb.2025.109113]

[62] Dhruba DD, Goetz S, Pria OFD, Reith T, Reutzel A, et al. Deep learning-based cardiac MRI planning from localizers to cine views using landmark detection. Acad Radiol. 2026;33(3):924-35. [PMID: 41353071 DOI: 10.1016/j.acra.2025.11.028]

[63] Tao R, Ye K, Zhang W, Sun W, Yu D, et al. X2P-Net: context-aware 2D/3D vertebra localization. Bioengineering (Basel). 2026;13(2):178. [PMID: 41749718 DOI: 10.3390/bioengineering13020178]

[64] Ma J, He Y, Li F, Han L, You C, et al. Segment anything in medical images. Nat Commun. 2024;15(1):654. [PMID: 38253604 DOI: 10.1038/s41467-024-44824-z]

[65] Urschler M, Ebner T, Štern D. Integrating geometric configuration and appearance information into a unified framework for anatomicallandmark localization. Med Image Anal. 2018;43:23-36. [PMID: 28963961 DOI: 10.1016/j.media.2017.09.003.