# Learning Orthogonal Disentanglement for Domain-Agnostic Medical Image Segmentation

Kai Han, Jiaqi Zhang, Chongwen Lyu, Mengting Li, Jun Chen, Laihua Yang, Guangquan Zhou, *Senior Member, IEEE*, Yang Chen, *Senior Member, IEEE*, Zhe Liu

***Abstract*— Medical image segmentation is vital for clinical diagnosis, lesion localization, treatment planning, and efficacy evaluation. However, traditional models struggle to maintain consistent performance across modalities due to significant variations in textures and imaging principles. To address this challenge, we propose OrthoSeg, a domain-agnostic framework for general medical image segmentation. By orthogonally disentangling anatomical structures from textures, OrthoSeg removes domain-specific interference to capture consistent representations. Specifically, we first design a mutual information-based module to disentangle latent representations, separating domain-agnostic structures from domain-specific textures for effective noise suppression. Second, we enforce spatial geometric consistency via equivariance and invariance penalties to reduce ambiguity and enhance boundaries. Finally, cross-scale topological aggregation is proposed to address lesion scale variations, dynamically adjusting receptive fields and reconstructing target anatomies. OrthoSeg outperforms state-of-the-art methods across seven source and eight unseen datasets in six modalities. It mitigates domain-specific noise and demonstrates promising generalization to unseen domains, taking a step towards robust, domain-agnostic medical image segmentation.**

***Index Terms*— Medical Image Segmentation, Domain Generalization, Domain-Agnostic, Feature Disentanglement**

## I. Introduction

Medical image segmentation plays a crucial role in the diagnosis and treatment of major diseases [1], [2]. By providing precise, pixel-level delineation of target regions, it offers essential support for quantitative clinical analysis and diagnostic decision-making. However, in real-world clinical practice, segmentation models are often applied to unseen domains where the data distribution differs from that of the training set. This discrepancy arises from variations across different hospitals, scanning devices, acquisition protocols, and patient populations. Such distribution mismatch, commonly known as domain shift, significantly impairs model generalization, leading to blurred anatomical boundaries, missing structures, or unstable predictions. Although existing Convolutional Neural Network (CNN)-based [3]–[7] and Transformer-based [8]–[12] segmentation methods have achieved remarkable performance on in-domain datasets, they typically rely heavily on the appearance and texture features of the source domain. Consequently, they tend to overfit to specific imaging characteristics and noise patterns, resulting in severe performance degradation in cross-domain scenarios.

To address these generalization challenges, existing studies primarily explore strategies to simulate or mitigate domain shift. Among these, data augmentation and image style transfer methods [13]–[17] attempt to bridge the domain gap at the image level. However, they often fail to preserve fine

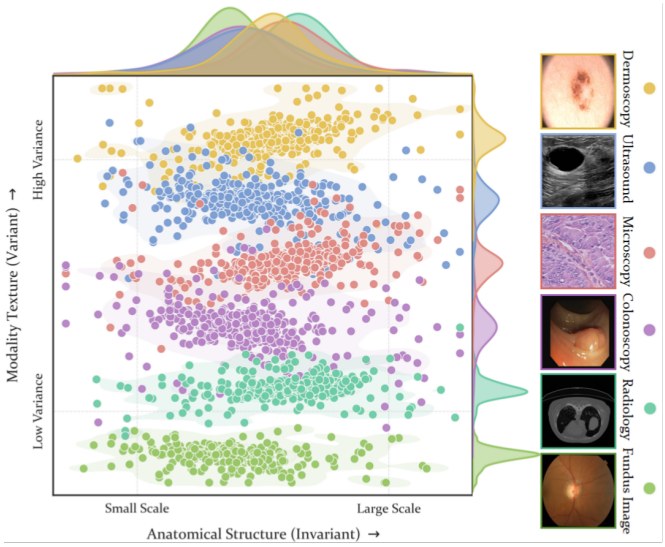


Fig. 1: **Cross-Domain Feature Decoupling.** The marginal probability density distributions indicate that the six imaging modalities exhibit high consistency at the anatomical structure scale. In contrast, the domain-specific physical textures demonstrate significant differences in distribution.

This work was supported in part by the National Natural Science Foundation of China under Grant (62276116). (*Corresponding authors: Zhe Liu.*)

Kai Han, Jiaqi Zhang, Chongwen Lyu, Mengting Li, Jun Chen and Zhe Liu, are with the School of Computer Science and Communication Engineering, Jiangsu University, 212013, China (e-mail: 1000006894@ujs.edu.cn; 3230602065@stmail.ujs.edu.cn; 2212308023@stmail.ujs.edu.cn; 3224401033@stmail.ujs.edu.cn; chenjun@ujs.edu.cn; 1000004088@ujs.edu.cn).

Laihua Yang is with the Department of Imaging, Danyang Traditional Chinese Medicine Hospital, Zhenjiang 274002, China (e-mail: yang15952950577@163.com).

Guangquan Zhou is with the School of Biological Science and Medical Engineering, Southeast University, Nanjing 211102, China (e-mail: guangquan.zhou@seu.edu.cn).

Yang Chen is with the School of Computer Science and Engineering, the Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, and the Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, Southeast University, Nanjing 210096, China (e-mail: chenyang.list@seu.edu.cn).

anatomical details when handling complex intensity variations within the same domain. In contrast, feature alignment and domain generalization methods [18]–[20] inherently rely on global distribution alignment across source domains. This limits their effectiveness in unseen clinical environments with diverse imaging protocols. Therefore, it is necessary to design a segmentation framework that minimizes reliance on site-specific priors and suppresses appearance variations caused by different imaging protocols.

To effectively mitigate cross-domain appearance discrepancies, it is crucial to recognize that medical images encode information across two distinct dimensions: anatomical structure and physical texture. Although the objective morphology of target tissues is physically coupled with imaging devices, these two dimensions encode fundamentally different information. Specifically, domain-specific physical textures exhibit significant domain shifts and high variance. Conversely, structural representations accurately describe objective physical boundaries and possess a highly consistent latent anatomical topology [21], as illustrated in Fig. 1. This observation highlights the necessity to disentangle variable texture noise from stable structural priors. This insight shifts the focus of feature learning from distribution alignment to representation disentanglement. By isolating and discarding volatile texture attributes, the model can rely exclusively on stable anatomical structures to achieve robust, domain-invariant generalization.

Motivated by this, we propose OrthoSeg, a structure-texture orthogonal disentanglement network for robust and generalizable medical image segmentation. Specifically, our framework consists of three core components: Mutual Information Disentanglement (MID), Geometric Consistency Constraint (GCC), and Cross-Scale Topology Aggregation (CSTA). First, the MID strategy is designed to minimize feature correlations, enabling the model to orthogonally disentangle features in the latent space into domain-invariant anatomical structural priors and protocol-specific imaging textures. Second, to address the ambiguity of latent space disentanglement, we introduce a physics-guided GCC module. Under spatial transformations, this module imposes equivariance constraints on structural features to preserve spatial correspondences, and invariance constraints on texture features to capture domain-specific styles, thereby achieving precise boundary reconstruction. Finally, we propose the CSTA module to address the significant scale variations of clinical anatomical structures. By leveraging multi-scale structural pyramids and high-order spatial statistics, it dynamically modulates convolutional receptive fields to achieve precise topological reconstruction. The main contributions of this paper are summarized as follows:

- We propose the OrthoSeg framework, introducing a novel structure-texture orthogonal decoupling paradigm for medical image segmentation. By isolating anatomical structures from domain-specific textures, OrthoSeg fundamentally overcomes the domain gap, enabling highly scalable and efficient cross-domain segmentation across diverse imaging modalities.
- We design the MID and GCC modules to construct a domain-agnostic representation space. By enforcing spatial equivariance and invariance penalties, these modules go beyond simple noise suppression to actively correct deep representation shifts, ensuring robust generalization against varying imaging physics and scanner protocols.
- We present the CSTA module, which dynamically embeds anatomical topology priors into the feature learning process. This allows the network to adaptively reconstruct complex organ boundaries across extreme resolution variations and organ scales, significantly reducing anatomically implausible predictions.
- Extensive experiments across 15 diverse datasets spanning 6 imaging modalities demonstrate that OrthoSeg achieves state-of-the-art performance. Crucially, it breaks the conventional performance-efficiency trade-off: securing superior cross-domain generalization with an ultra-lightweight footprint of only 11.76M parameters.

## II. Related Work

### A. Medical Image Segmentation

Early methods built encoder-decoder frameworks based on CNNs, building upon the skip connection paradigm established by U-Net [5], [6], [22], [23]. ResUNet [22] enhances gradient propagation stability in deep networks by introducing residual connections, while MedNeXt [24] expands the effective receptive field for feature extraction by incorporating large-kernel convolutions and inverted bottleneck structures. However, the inherently local receptive fields of pure CNN architectures limit their capacity to model long-range dependencies in images.

Given the advantages of Transformers in sequence modeling, researchers have increasingly applied them into medical image segmentation [8], [25]. Methods like Swin-UNet [26] construct pure Transformer architectures, replacing convolution operations with self-attention mechanisms to capture global context. Nevertheless, pure Transformer models rely heavily on massive training data and show limitations in processing fine-grained features [27], making hybrid architectures the mainstream. By combining the local feature extraction capability of CNNs with the global modeling advantage of Transformers [28], hybrid architectures improve segmentation accuracy while maintaining computational efficiency. Although these methods have achieved significant progress on specific source domains, their generalization ability remains severely limited. This limitation stems from substantial variations in texture features, intensity distributions, and imaging protocols.

### B. Segmentation Model Domain Generalization

Medical image segmentation models often experience significant performance degradation in unseen domains. This is primarily due to substantial discrepancies in imaging devices, acquisition protocols, and texture distributions across different datasets [29], [30]. Therefore, enhancing generalization across diverse datasets and unseen domains is crucial. It effectively reduces reliance on target domain annotations and additional fine-tuning, thereby facilitating practical clinical deployment.

Existing research primarily addresses this challenge through domain shift mitigation [15]–[17], structure-aware generalization [18], [19], and disentangled or invariant representation

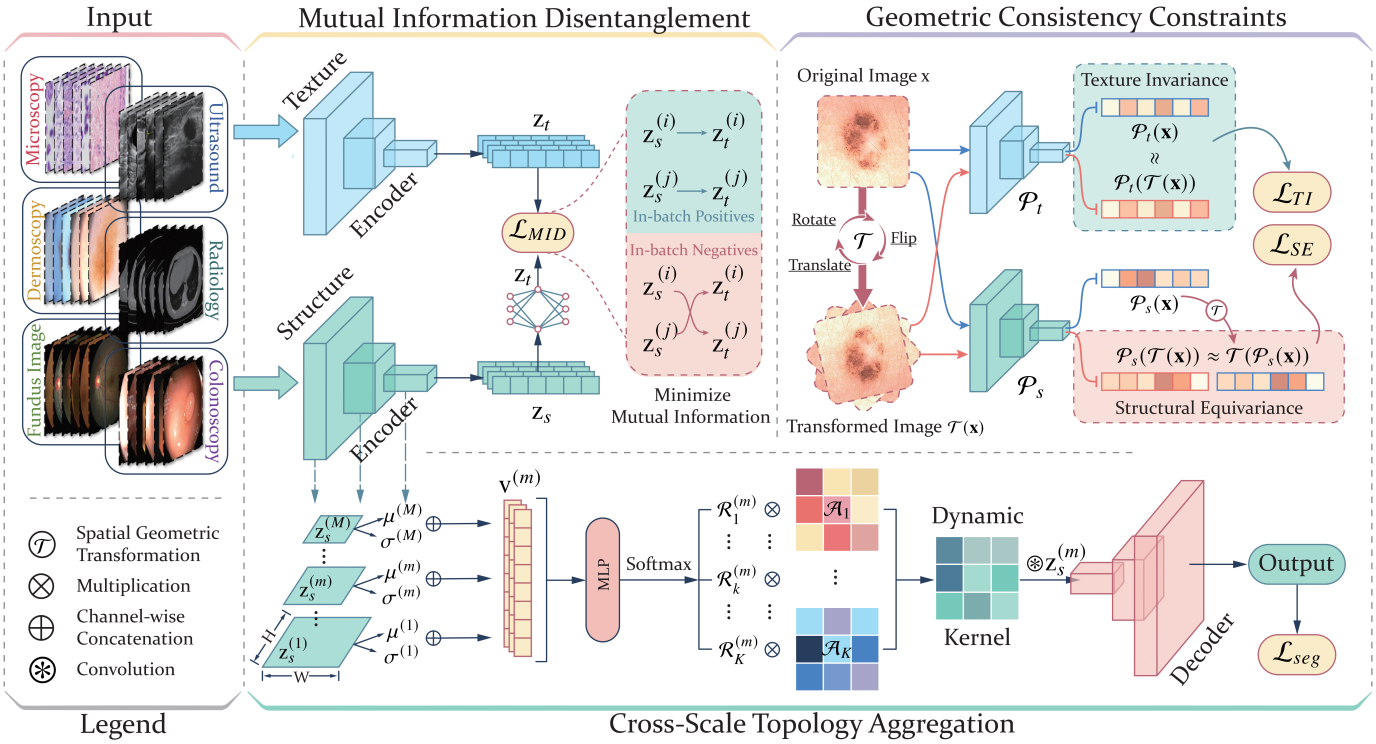


Fig. 2: **Architecture overview.** First, the input images are passed through a dual-branch encoder and orthogonally decoupled into invariant structural features and domain-specific texture features. Subsequently, these features are subjected to spatial transformation constraints (structural equivariance and texture invariance) in the latent space to eliminate representation ambiguity. Finally, the texture-noise-free structural features are fed into a decoder, where dynamic convolutional kernels adaptively fuse multi-scale structural features to output the final high-precision segmentation mask.

learning [20], [31], [32]. Early domain shift mitigation methods, such as ISGAN [15], typically employ style transfer, frequency perturbation, or multi-scale adversarial learning for distribution alignment. However, under significant domain heterogeneity, these methods often struggle to maintain stable semantic consistency. To address this issue, researchers have begun leveraging relatively stable structural information to improve generalization. For example, methods like MADGNet [18] and CGDMNet [19] introduce multi-scale, multi-frequency features and multi-task adaptation mechanisms. Nevertheless, most approaches remain confined to structural enhancement or feature interaction. They lack an explicit disentanglement between stable anatomical structures and variable domain textures.

Beyond domain shift mitigation and structural enhancement, recent studies have also focused on learning domain-invariant representations to improve model generalization. For instance, ConDSeg [20] utilizes contrast-driven feature enhancement and knowledge distillation, effectively boosting the cross-domain segmentation performance of lightweight models. However, these methods mostly learn invariant features only at a coarse-grained level. They lack deep physical semantic constraints and overlook the multi-scale topological variations of organs and lesions. In contrast, our method explicitly disentangles structure from texture. By actively eliminating variable, domain-specific interference, it enables the model to fully rely on stable anatomical priors to combat domain shifts.

## III. Method

The proposed framework is illustrated in Fig. 2. Given an input image, we first extract initial representations via a dual-branch encoder. Building upon this, we design Mutual Information Disentanglement (MID) to strictly separate image features into stable anatomical structures and scanner-specific textures. Subsequently, Geometric Consistency Constraints (GCC) eliminate the representation ambiguity caused by this decoupling, ensuring precise physical boundaries. Finally, Cross-Scale Topology Aggregation (CSTA) leverages the purified structural features for multi-scale adaptive reconstruction, outputting the final segmentation mask.

In this study, datasets from various clinical centers are categorized into a source domain $\mathrm{SD}^s = \{(\mathrm{x}_i^s, \mathrm{y}_i^s)\}_{i=1}^{N_s}$ and an unseen target domain $\mathrm{UD}^t = \{\mathrm{x}_i^t\}_{i=1}^{N_t}$. Here, $\mathrm{x}_i$ denotes the $i$-th input image, and $\mathrm{y}_i$ is its corresponding ground truth. $N_s$ and $N_t$ represent the total number of samples in the source and unseen domains, respectively. We aim to optimize a generalizable parametric segmentation model $\mathcal{F}_\theta : \mathrm{x} \rightarrow \mathrm{y}$ using only the annotated source domain $\mathrm{SD}^s$. The goal is for the model to maintain excellent performance on the source domain while generalizing directly to the unseen domain $\mathrm{UD}^t$.

### A. Mutual Information Disentanglement

To prevent the network from relying on domain-specific textures, this module introduces a Mutual Information Disentanglement (MID) constraint. This constraint strictly separates structural features $\mathrm{z}_s$ from texture features $\mathrm{z}_t$.

In continuous high-dimensional latent spaces, computing the exact mutual information $\mathbb{I}(\mathrm{z}_s;\mathrm{z}_t)$ is often intractable. Therefore, we employ the Contrastive Log-ratio Upper Bound [33] to construct the optimization objective for mutual information minimization. Specifically, we utilize a parametric auxiliary neural network $q_\phi(\mathrm{z}_t|\mathrm{z}_s)$ to predict texture features given structural features. The core idea is to constrain information overlap by contrasting the probability discrepancy between matched feature pairs and randomly shuffled pairs. Its variational upper bound can be expressed as the difference between the expectations of the joint distribution and the product of marginal distributions:

$$\begin{aligned}\mathbb{I}(\mathrm{z}_s;\mathrm{z}_t) \leq \mathbb{E}_{p(\mathrm{z}_s,\mathrm{z}_t)}\left[\log q_\phi(\mathrm{z}_t|\mathrm{z}_s)\right] \\ - \mathbb{E}_{p(\mathrm{z}_s)p(\mathrm{z}_t)}\left[\log q_\phi(\mathrm{z}_t|\mathrm{z}_s)\right].\end{aligned} \tag{1}$$

During training, these two theoretical expectations are approximated via in-batch sampling. Given a mini-batch of $N$ samples, the first expectation (joint distribution) calculates the log-likelihood of paired features $\left(\mathrm{z}_s^{(i)},\mathrm{z}_t^{(i)}\right)$ from the same image. The second expectation (marginal distribution) approximates independent sampling by traversing all possible feature cross-combinations $\left(\mathrm{z}_s^{(i)},\mathrm{z}_t^{(j)}\right)$ within the batch. Based on this sampling strategy, the mutual information disentanglement loss function $\mathcal{L}_{MID}$ can be defined over discrete data as:

$$\begin{aligned}\mathcal{L}_{MID} = \frac{1}{N}\sum_{i=1}^{N}\log q_\phi(\mathrm{z}_t^{(i)}|\mathrm{z}_s^{(i)}) \\ - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\log q_\phi(\mathrm{z}_t^{(j)}|\mathrm{z}_s^{(i)}).\end{aligned} \tag{2}$$

### B. Geometric Consistency Constraints

In medical images, anatomical structures vary synchronously with spatial coordinates and are highly sensitive to spatial locations. In contrast, imaging textures typically manifest as global physical properties unaffected by local geometric transformations. Building upon the orthogonal feature space established by MID module, we propose a geometric consistency constraint based on spatial affine transformations. This constraint enables $\mathrm{z}_s$ to focus precisely on anatomical topology. Meanwhile, it forces $\mathrm{z}_t$ to encode specific domain styles.

Let $\mathcal{T}$ be a spatial geometric transformation sampled from a predefined set (e.g., random rotation, flipping, or translation). Let $\mathcal{P}_s$ and $\mathcal{P}_t$ denote the feature extraction mappings of the structural and texture branches, respectively. We design the following two constraints to incorporate physical priors into the disentangled features:

*1) Structural Equivariance:* Anatomical topological features must reflect the spatial distribution of organs or lesions within the input image. Therefore, if the original image undergoes a spatial transformation, the extracted structural features should exhibit identical geometric deformation. To enable the structural branch to capture this spatially sensitive information, we impose the following constraint:

$$\mathcal{L}_{SE} = \|\mathcal{P}_s(\mathcal{T}(\mathbf{x})) - \mathcal{T}(\mathcal{P}_s(\mathbf{x}))\|_2^2. \tag{3}$$

From an optimization perspective, this equation constrains the commutativity of the feature extraction mapping $\mathcal{P}_s$ and the spatial transformation $\mathcal{T}$. Minimizing the mean squared error between $\mathcal{P}_s(\mathcal{T}(\mathbf{x}))$ and $\mathcal{T}(\mathcal{P}_s(\mathbf{x}))$ compels the network to preserve precise pixel-level relative positional information within $\mathrm{z}_s$.

*2) Texture Invariance:* Unlike structural representations, scanner-specific textures (such as speckle noise in ultrasound or specular highlights in endoscopy) are global physical imaging properties. These inherent attributes of the imaging equipment should remain constant, regardless of changes in sensor or patient positioning (i.e., spatial geometric transformations). Therefore, we impose the following constraint on the texture branch:

$$\mathcal{L}_{TI} = \|\mathcal{P}_t(\mathcal{T}(\mathbf{x})) - \mathcal{P}_t(\mathbf{x})\|_2^2. \tag{4}$$

This constraint minimizes the discrepancy between texture features extracted before and after the transformation, effectively acting as a spatial information filter. It forces the $\mathcal{P}_t$ branch to actively discard all coordinate-dependent geometric information. Consequently, it aggregates only spatially agnostic domain style representations.

### C. Cross-Scale Topology Aggregation

Guided by the aforementioned geometric constraints, the network extracts domain-invariant anatomical features $\mathrm{z}_s$. However, lesions and target organs vary substantially in scale and shape in clinical scenarios. Consequently, conventional static convolutions with fixed receptive fields struggle to capture global topology and fine local details simultaneously. To address this, we design the Cross-Scale Topology Aggregation (CSTA) module. It models $\mathrm{z}_s$ as a multi-scale feature pyramid $\{\mathrm{z}_s^{(m)}\}_{m=1}^{M}$ and performs adaptive reconstruction using a dynamic convolution mechanism.

Specifically, at the $m$-th level of the pyramid, the network first extracts the spatial mean $\mu^{(m)}$ and standard deviation $\sigma^{(m)}$ of the feature map:

$$\begin{aligned}\mu^{(m)} &= \frac{1}{HW}\sum_{i,j}\mathrm{z}_{s,i,j}^{(m)}, \\ \sigma^{(m)} &= \sqrt{\frac{1}{HW}\sum_{i,j}\left(\mathrm{z}_{s,i,j}^{(m)} - \mu^{(m)}\right)^2}\end{aligned} \tag{5}$$

where $H$ and $W$ denote the spatial height and width of the $m$-th level feature map, respectively. To characterize variable lesion morphologies, we leverage the mean to reflect global response intensity and coarse target scale, and the standard deviation to capture boundary complexity and local morphological variations. By concatenating them along the channel dimension, this aggregation of global spatial statistics generates a comprehensive global topological descriptor $\mathrm{v}^{(m)}$ with both scale and morphology awareness:

$$\mathrm{v}^{(m)} = \mu^{(m)} \oplus \sigma^{(m)} \tag{6}$$

where $\oplus$ denotes channel-wise concatenation. We then feed this descriptor into a routing gating function $\mathcal{G}(\cdot)$, parameterized by a Multi-Layer Perceptron (MLP). This function

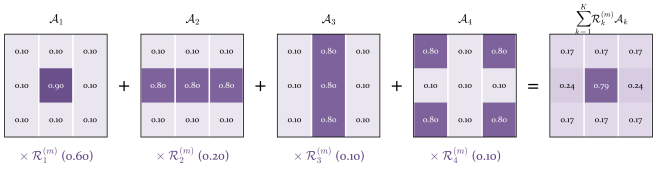

Fig. 3: Visualization of the dynamic kernel synthesis process in the CSTA module ($K = 4$)

dynamically predicts the normalized attention weights $\mathcal{R}_k^{(m)}$ for $K$ predefined convolutional kernel experts $\{\mathcal{A}_k\}_{k=1}^K$. Using linear weighting, the network directly aggregates a customized dynamic convolutional kernel tailored to the current input. This kernel then executes topological reconstruction on the structural features at the same scale:

$$\hat{\mathrm{z}}_s^{(m)} = \left(\sum_{k=1}^{K} \mathcal{R}_k^{(m)} \mathcal{A}_k\right) * \mathrm{z}_s^{(m)} \tag{7}$$

where $*$ denotes the convolution operation, and the routing weight vector is computed as $\mathcal{R}^{(m)} = \mathrm{Softmax}\left(\mathcal{G}(\mathbf{v}^{(m)})\right)$. An illustration of the dynamic fusion of base kernels $(\sum_{k=1}^{K} \mathcal{R}_k^{(m)} \mathcal{A}_k)$ is shown in Fig. 3. Finally, the dynamically reconstructed features at each level, $\{\hat{\mathbf{z}}_s^{(m)}\}_{m=1}^{M}$, are aggregated across scales via skip connections and progressive upsampling to output the segmentation probability map $\hat{\mathbf{y}}$.

### D. Overall Loss Function

When training exclusively on the source domain dataset $\mathrm{SD}^s = \{(\mathrm{x}_i^s, \mathrm{y}_i^s)\}_{i=1}^{N_s}$, the network's overall objective function combines the segmentation task loss and feature disentanglement regularization terms. It is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{Seg} + \lambda\left(t\right)\mathcal{L}_{MID} + \alpha\mathcal{L}_{SE} + \beta\mathcal{L}_{TI} \tag{8}$$

TABLE I: Source Domain Datasets used in our experiments.

| No. | Dataset | Modality | Resolution | Images |
|---|---|---|---|---|
| $\mathrm{SD}^1$ | DSB-2018 [34] | Microscopy | Variable | 670 |
| $\mathrm{SD}^2$ | BUSI [35] | Ultrasound | Variable | 645 |
| $\mathrm{SD}^3$ | ISIC2018 [36] | Dermoscopy | Variable | 2594 |
| $\mathrm{SD}^4$ | COVID19-1 [37] | Radiology | 512 × 512 | 1277 |
| $\mathrm{SD}^5$ | REFUGE [38] | Fundus Image | 2124 × 2056 | 400 |
| $\mathrm{SD}^6$ | CVC-ClinicDB [39] | Colonoscopy | 384 × 288 | 612 |
| $\mathrm{SD}^7$ | Kvasir-SEG [40] | Colonoscopy | Variable | 1000 |

TABLE II: Unseen Domain Datasets used in our experiments.

| No. | Dataset | Modality | Resolution | Images |
|---|---|---|---|---|
| $\mathrm{UD}^1$ | MonuSeg2018 [41] | Microscopy | 256 × 256 | 82 |
| $\mathrm{UD}^2$ | STU [42] | Ultrasound | Variable | 42 |
| $\mathrm{UD}^3$ | $\mathrm{PH}^2$ [43] | Dermoscopy | 767 × 576 | 200 |
| $\mathrm{UD}^4$ | COVID19-2 [44] | Radiology | 512 × 512 | 2535 |
| $\mathrm{UD}^5$ | Drishti-GS [45] | Fundus Image | Variable | 50 |
| $\mathrm{UD}^6$ | CVC-300 [46] | Colonoscopy | 574 × 500 | 60 |
| $\mathrm{UD}^7$ | CVC-ColonDB [47] | Colonoscopy | 574 × 500 | 380 |
| $\mathrm{UD}^8$ | ETIS [48] | Colonoscopy | 1255 × 966 | 196 |

Here, $\mathcal{L}_{Seg}$ is a standard combination of Dice loss and Cross-Entropy loss. The parameters $\alpha$ and $\beta$ are fixed hyperparameters governing the geometric consistency constraints.

In the early training stages, the model requires strong penalty signals to decouple highly correlated features. However, as optimization progresses and the feature space tends toward orthogonality, excessive regularization might limit the model's fine-grained fitting capability on the primary segmentation task. Consequently, we introduce a time-decaying weight $\lambda(t)$ for the mutual information loss. This weight dynamically adjusts with the training epoch $t$ and is defined as:

$$\lambda(t) = \tau \cdot \left(1 - \frac{t}{\mathrm{e}_{\max}}\right)^{\gamma} \tag{9}$$

where $\tau$, $\mathrm{e}_{max}$, and $\gamma$ are the initial base weight, the total number of epochs, and the decay exponent controlling the curvature of the weight reduction, respectively.

## IV. EXPERIMENTS

### A. Dataset and Evaluation Metrics

To validate our proposed OrthoSeg model, we evaluated its performance on seven medical image datasets across six modalities: dermoscopy, radiology, ultrasound, microscopy, colonoscopy, and fundus imaging. This evaluation establishes its baseline segmentation performance on the source domains, as summarized in Table I. Notably, fundus imaging involves multi-label segmentation for the optic disc (OD) and optic cup (OC). The other modalities are formulated as binary segmentation tasks. Furthermore, the generalizability of our framework was assessed by selecting corresponding unseen domains for each modality. This comprises a total of eight datasets, detailed in Table II.

To comprehensively evaluate our proposed network against state-of-the-art methods, we employ two widely used metrics: the Dice Similarity Coefficient (Dice) and the 95th percentile Hausdorff Distance (HD95). Dice quantifies the pixel-level regional overlap between the predicted mask and the ground truth. Meanwhile, HD95 is highly sensitive to local boundary errors, effectively assessing the accuracy of segmentation contours.

### B. Implementation Details

All experiments in this study were implemented using PyTorch 1.13.0 and conducted on a single NVIDIA RTX 3090 GPU with 24GB of memory. Medical images from all modalities were uniformly resized to a resolution of $352 \times 352$ prior to being fed into the network. During the training phase, we applied standard data augmentation strategies. These included random rotation $(-15°, 15°)$, random scaling $(0.8, 1.2)$, and random horizontal and vertical flipping. The model was optimized using AdamW with an initial learning rate set to $1 \times 10^{-4}$. We employed a Cosine Annealing learning rate schedule to ensure stable convergence in the later stages of training. All models were trained for 200 epochs with a batch size of 16. We trained the model independently on the seven source datasets. Subsequently, we evaluated it on both these source datasets and the unseen datasets.

TABLE III: Quantitative comparison on universal medical image segmentation (Source Domain). The best and suboptimal results are highlighted. We also provide one-tailed paired t-Test results ($P$-Value) compared to our OrthoSeg and other methods.

| | Methods | Params ↓ | FLOPs ↓ | $SD^1$ | | $SD^2$ | | $SD^3$ | | $SD^4$ | | $SD^5$– OD | | $SD^5$– OC | | $SD^6$ | | $SD^7$ | | $SD^a$ (Avg.) | | $P$-Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | |
| (i) | UNet [3] | 14.80M | 8.43G | 89.69 | 9.51 | 76.90 | 55.47 | 84.26 | 33.37 | 42.56 | 86.14 | 80.72 | 34.51 | 78.10 | 22.35 | 84.76 | 43.33 | 82.97 | 55.88 | 77.50 | 42.57 | 0.00000012 |
| | Att-UNet [4] | 34.88M | 18.30G | 89.61 | 8.78 | 71.15 | 91.29 | 84.65 | 32.14 | 51.28 | 67.41 | 79.65 | 36.90 | 76.85 | 24.10 | 86.79 | 29.85 | 84.03 | 63.93 | 78.00 | 44.30 | 0.00000024 |
| | UNet++ [5] | 36.61M | 16.54G | 90.01 | 8.12 | 76.29 | 61.74 | 86.21 | 38.74 | 64.21 | 54.33 | 81.93 | 31.40 | 79.72 | 21.15 | 87.52 | 38.19 | 84.21 | 54.30 | 81.26 | 38.50 | 0.0000085 |
| | nnUNet [6] | 33.36M | 15.07G | 90.51 | 6.87 | 82.65 | 38.90 | 84.15 | 36.52 | 63.14 | 53.77 | 83.41 | 28.55 | 81.60 | 19.90 | 84.68 | 46.32 | 83.14 | 53.65 | 81.66 | 35.56 | 0.000014 |
| | $M^2$SNet [7] | 36.52M | 11.23G | 90.04 | 7.31 | 83.76 | 33.06 | 87.21 | 26.57 | 76.53 | 34.67 | 84.15 | 26.70 | 82.80 | 18.45 | 90.21 | 28.87 | 88.57 | 31.61 | 85.41 | 25.91 | 0.00032 |
| (ii) | H2Former [9] | 28.41M | 15.65G | 90.41 | 7.22 | 83.85 | 30.14 | 87.88 | 26.55 | 79.75 | 30.82 | 84.76 | 25.95 | 83.42 | 17.85 | 89.18 | 24.65 | 88.92 | 26.80 | 86.02 | 23.75 | 0.00081 |
| | TransUNet [8] | 53.42M | 24.36G | 88.88 | 8.77 | 79.63 | 50.05 | 87.11 | 29.63 | 74.12 | 41.45 | 82.92 | 29.40 | 81.01 | 20.65 | 87.75 | 29.73 | 86.49 | 39.55 | 83.49 | 31.15 | 0.000063 |
| | CCViM [10] | 23.56M | 7.61G | 90.48 | 7.08 | 83.98 | 29.45 | 89.98 | 25.10 | 89.88 | 30.30 | 86.42 | 23.10 | 85.16 | 16.50 | 89.36 | 22.48 | 89.08 | 24.15 | 88.04 | 22.27 | 0.0045 |
| | BRAU-Net++ [11] | 28.38M | 13.46G | 90.68 | 6.55 | 84.45 | 27.95 | 90.22 | 19.35 | 90.15 | 28.40 | 86.61 | 22.70 | 85.92 | 15.98 | 89.66 | 15.88 | 89.95 | 15.06 | 88.45 | 18.98 | 0.013 |
| (iii) | MADGNet [18] | 31.42M | 13.85G | 90.33 | 6.53 | 82.62 | 39.30 | 88.14 | 25.41 | 80.09 | 29.47 | 85.35 | 24.60 | 84.05 | 17.20 | 88.63 | 37.33 | 89.53 | 48.26 | 86.09 | 28.51 | 0.0011 |
| | CGDMNet [19] | 25.06M | 7.47G | 90.72 | 6.45 | 84.62 | 28.15 | 90.45 | 18.89 | 89.92 | 30.56 | 86.65 | 22.80 | 85.84 | 16.10 | 89.72 | 15.24 | 89.84 | 14.92 | 88.47 | 19.14 | 0.015 |
| | ConDSeg [20] | 23.57M | 9.15G | 90.70 | 6.48 | 84.58 | 26.10 | 90.35 | 20.15 | 90.05 | 27.95 | 86.58 | 22.95 | 85.71 | 16.35 | 89.58 | 16.42 | 89.73 | 15.38 | 88.41 | 18.97 | 0.011 |
| | OrthoSeg (Ours) | 11.76M | 8.16G | 91.45 | 5.92 | 85.85 | 27.20 | 91.95 | 15.42 | 91.15 | 26.85 | 87.58 | 18.41 | 86.41 | 10.87 | 90.14 | 13.09 | 90.65 | 12.64 | 89.40 | 16.30 | — |

TABLE IV: Quantitative comparison on universal medical image segmentation (Unseen Domain). The best and suboptimal results are highlighted. We also provide one-tailed paired t-Test results ($P$-Value) compared to our OrthoSeg and other methods.

| | Methods | Params ↓ | FLOPs ↓ | $SD^1 \rightarrow UD^1$ | | $SD^2 \rightarrow UD^2$ | | $SD^3 \rightarrow UD^3$ | | $SD^4 \rightarrow UD^4$ | | $SD^5 \rightarrow UD^5$– OD | | $SD^5 \rightarrow UD^5$– OC | | $SD^6 + SD^7 \rightarrow UD^6$ | | $SD^6 + SD^7 \rightarrow UD^7$ | | $SD^6 + SD^7 \rightarrow UD^8$ | | $UD^f$ (Avg.) | | $P$-Value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ | |
| (i) | UNet [3] | 14.80M | 8.43G | 44.21 | 46.34 | 78.81 | 75.82 | 84.14 | 21.26 | 49.58 | 74.89 | 68.53 | 43.21 | 47.18 | 89.30 | 79.50 | 57.09 | 71.11 | 60.59 | 67.75 | 63.04 | 65.65 | 59.06 | 0.0000034 |
| | Att-UNet [4] | 34.88M | 18.30G | 44.49 | 35.19 | 73.74 | 70.18 | 86.58 | 18.42 | 54.73 | 66.52 | 70.12 | 41.95 | 49.35 | 86.75 | 84.87 | 29.38 | 74.13 | 46.69 | 67.26 | 74.47 | 67.25 | 52.17 | 0.00000011 |
| | UNet++ [5] | 36.61M | 16.54G | 45.81 | 36.31 | 75.53 | 66.81 | 87.42 | 16.60 | 61.34 | 55.50 | 72.84 | 39.12 | 52.66 | 84.45 | 82.65 | 36.58 | 74.17 | 45.52 | 69.30 | 55.51 | 69.08 | 48.49 | 0.0000027 |
| | nnUNet [6] | 33.36M | 15.07G | 42.35 | 39.58 | 82.28 | 35.67 | 88.84 | 15.08 | 63.14 | 52.69 | 75.65 | 36.45 | 55.82 | 81.10 | 78.82 | 55.02 | 71.64 | 99.62 | 70.98 | 127.36 | 69.95 | 60.29 | 0.0000042 |
| | $M^2$SNet [7] | 36.52M | 11.23G | 43.81 | 42.96 | 83.37 | 35.14 | 88.14 | 16.75 | 66.83 | 37.41 | 78.10 | 34.88 | 58.94 | 77.25 | 88.27 | 13.94 | 78.72 | 35.78 | 76.48 | 48.96 | 73.63 | 38.12 | 0.000038 |
| (ii) | H2Former [9] | 28.41M | 15.65G | 46.12 | 35.80 | 87.50 | 25.40 | 89.95 | 16.20 | 70.40 | 40.15 | 80.95 | 33.25 | 61.82 | 73.85 | 88.84 | 15.62 | 79.86 | 33.95 | 77.84 | 42.60 | 75.92 | 35.20 | 0.00014 |
| | TransUNet [8] | 53.42M | 24.36G | 45.53 | 37.42 | 85.71 | 43.90 | 86.68 | 20.15 | 68.51 | 55.14 | 82.35 | 31.78 | 64.55 | 70.40 | 84.24 | 34.29 | 75.64 | 75.43 | 70.68 | 89.28 | 73.77 | 50.87 | 0.000035 |
| | CCViM [10] | 23.56M | 7.61G | 46.52 | 35.50 | 85.85 | 29.10 | 89.42 | 22.05 | 81.95 | 39.80 | 84.25 | 29.85 | 68.95 | 65.80 | 89.12 | 14.86 | 81.74 | 29.42 | 79.95 | 35.88 | 78.64 | 33.58 | 0.0018 |
| | BRAU-Net++ [11] | 28.38M | 13.46G | 47.10 | 34.95 | 87.20 | 26.35 | 89.25 | 19.85 | 81.82 | 36.12 | 90.58 | 24.95 | 88.72 | 18.20 | 89.88 | 12.05 | 83.94 | 20.74 | 82.52 | 17.82 | 82.33 | 23.45 | 0.021 |
| (iii) | MADGNet [18] | 31.42M | 13.85G | 45.52 | 60.87 | 86.48 | 27.07 | 89.45 | 14.12 | 78.10 | 32.33 | 83.48 | 30.62 | 66.88 | 67.75 | 89.92 | 41.52 | 76.85 | 69.54 | 78.01 | 91.39 | 77.19 | 48.36 | 0.00052 |
| | CGDMNet [19] | 25.06M | 7.47G | 47.45 | 34.52 | 88.11 | 23.15 | 90.14 | 13.89 | 81.20 | 35.96 | 88.72 | 27.45 | 85.96 | 23.10 | 89.96 | 11.86 | 83.56 | 20.36 | 82.68 | 17.54 | 81.98 | 23.09 | 0.014 |
| | ConDSeg [20] | 23.57M | 9.15G | 46.95 | 35.15 | 86.52 | 27.45 | 89.65 | 20.12 | 81.75 | 35.80 | 89.35 | 26.35 | 87.25 | 20.95 | 89.72 | 12.34 | 84.82 | 21.08 | 82.35 | 18.06 | 82.04 | 24.14 | 0.016 |
| | OrthoSeg (Ours) | 11.76M | 8.16G | 48.15 | 33.85 | 88.92 | 19.85 | 91.35 | 12.42 | 84.22 | 30.55 | 91.45 | 13.21 | 89.61 | 15.66 | 90.75 | 10.93 | 85.42 | 18.15 | 84.65 | 15.80 | 83.84 | 18.94 | — |

### C. Comparison with State-of-the-Art Methods

To comprehensively evaluate OrthoSeg, we compared it against three categories of representative baseline models: *(i) CNN-based methods*, including UNet [3], Att-UNet [4], UNet++ [5], nnUNet [6], and $M^2$SNet [7], *(ii) Transformer-based and hybrid architectures*, such as TransUNet [8], H2Former [9], CCViM [10], and BRAU-Net++ [11], and *(iii) Domain generalization methods*, including MADGNet [18], CGDMNet [19], and ConDSeg [20]. In addition, paired t-tests were conducted across all scenarios for statistical significance analysis.

*1) Quantitative Results:* To evaluate the basic segmentation performance, Table III presents the test results of all models on seven source domains. The significant differences in physical texture and intensity distribution across diverse clinical datasets make it challenging for a single conventional model to maintain stable performance. Overall, whether based on the classical UNet architecture or hybrid Transformer architectures with global modeling capabilities, their performance remains limited when handling complex boundaries, multi-scale targets, or domain-specific appearance features. In contrast, OrthoSeg achieves the best Dice and HD95 scores on nearly all source datasets. Compared with the state-of-the-art baseline BRAU-Net++, OrthoSeg improves the average Dice score by approximately 0.95% and reduces average HD95 by 2.58. This indicates that the structure-texture separation strategy and the dynamic convolution with adaptive receptive field adjustment in the CSTA module can effectively accommodate cross-domain distribution differences and achieve accurate geometric reconstruction of multi-scale lesions.

Table IV further lists the direct inference results on eight unseen target domains without fine-tuning, to assess generalization under severe domain shifts. Similarly, existing methods designed for generalization still exhibit considerable boundary errors or inter-domain performance fluctuations. In contrast, OrthoSeg maintains extremely high consistency in cross-domain evaluations, achieving an average unseen-domain Dice of 83.84% and HD95 of 18.94, significantly outperforming all comparative methods. By suppressing domain noise via the MID module and mitigating decoupling ambiguity with the GCC module, the network explicitly disentangles texture features to reduce cross-domain interference in the latent space. As a result, the model can fully leverage stable anatomical priors to adapt to unseen distributions. Moreover, the paired t-test analysis further confirms that the performance gains achieved by OrthoSeg over major competitors are highly statistically significant and exhibit greater stability

*2) Qualitative Results:* We present a qualitative comparison between OrthoSeg and state-of-the-art baselines on both source and unseen domains to evaluate our method. As shown in Fig. 4, models like CGDMNet [19] and ConDSeg [20] demonstrate reasonable performance on source domains. However, they often struggle with complex boundaries and are prone to structural omissions. In contrast, OrthoSeg accurately delineates smooth and anatomically precise contours. Examples include the scattered cells in $SD^1$ and the irregular lesions in $SD^6$. Furthermore, Fig. 5 illustrates our model's robust generalization against severe domain shifts in unseen domains. Highly challenging scenarios include the noisy microscopic images in $UD^1$ and the complex topologies in $UD^4$. In these challenging scenarios, baseline models suffer from noticeable under-segmentation and artifacts. Meanwhile, OrthoSeg consistently maintains the structural integrity of the lesions. It yields predictions that are highly consistent with the ground truth.

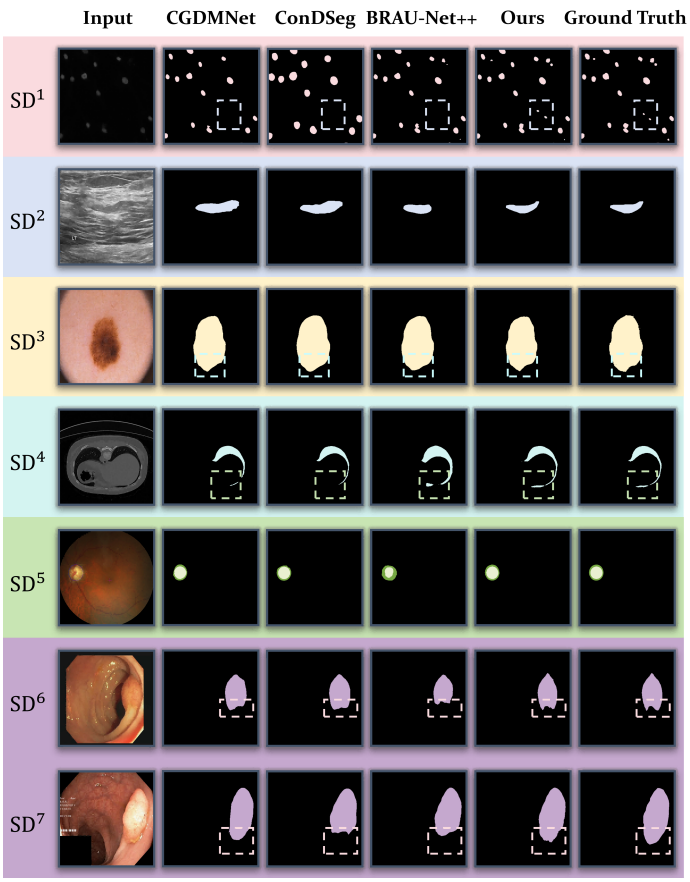


Fig. 4: Qualitative comparison with other methods on source domains.

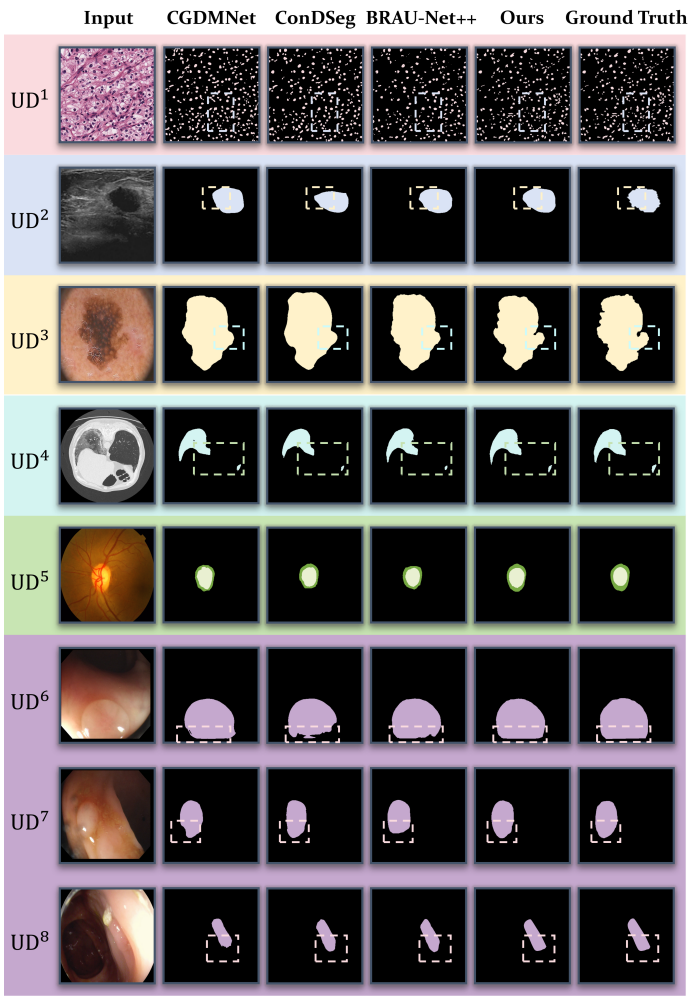


Fig. 5: Qualitative comparison with other methods on unseen domains.

TABLE V: Ablation study results of each component in OrthoSeg.

| MID | GCC | CSTA | Params(M) ↓ | $SD^s$ (Avg.) | | $UD^t$ (Avg.) | |
|---|---|---|---|---|---|---|---|
| | | | | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ |
| ✓ | | | 9.68 | 86.15 | 22.45 | 78.92 | 30.30 |
| | ✓ | | 9.96 | 85.84 | 20.12 | 76.45 | 28.75 |
| | | ✓ | 11.05 | 88.20 | 18.50 | 75.60 | 26.20 |
| ✓ | ✓ | | 10.07 | 87.35 | 19.20 | 81.45 | 21.60 |
| ✓ | | ✓ | 11.32 | 87.95 | 18.30 | 82.10 | 20.85 |
| | ✓ | ✓ | 11.23 | 88.75 | 16.95 | 79.30 | 23.40 |
| ✓ | ✓ | ✓ | 11.76 | **89.40** | **16.30** | **83.84** | **18.94** |

## D. Ablation Studies

We conducted comprehensive ablation studies on the core modules and their internal mechanisms. This validates the individual contributions of OrthoSeg's components to segmentation accuracy and cross-domain generalization.

*1) Ablation Study on Core Modules:* We evaluated the performance impact of each component through incremental integration on both the source domains (SD) and unseen domains (UD). Table V details the model parameters (Params) and the average Dice and HD95 metrics under different module combinations. When only the CSTA module is introduced, the model achieves an average Dice of 88.20% on the SD. This demonstrates its effectiveness in capturing and reconstructing multi-scale anatomical topologies. However, lacking feature decoupling, it suffers from severe texture overfitting, resulting in a UD Dice of only 75.60%. Upon integrating the MID module, the UD Dice substantially increases to 82.10%. This confirms that MID effectively filters out domain-related texture noise in the latent space. Consequently, it forces the network to rely solely on cross-domain consistent structural priors to overcome generalization barriers. Furthermore, pure latent space decoupling easily triggers local boundary representation ambiguity. The GCC module compensates for this deficiency by introducing geometric consistency penalties. Compared to MID+CSTA, adding GCC to the full model substantially reduces the average HD95 on the SD and UD from 18.30 and 20.85 to 16.30 and 18.94, respectively. This indicates that GCC successfully anchors pixel-level spatial coordinates and achieves precise tissue boundary reconstruction.

*2) Ablation Study on Cross-Scale Topology Aggregation (CSTA):* We investigated the impact of the number of base kernels ($K$) in the CSTA module's dynamic convolutions. To assess its effect on model performance and generalization capability, we conducted an in-depth ablation analysis with $K \in \{1, 2, 4, 8\}$. As shown in Table VI, setting $K = 1$ essentially degrades the module into a standard static convolution. This setup lacks topological adaptive capabilities, yielding an average Dice of only 87.23% and 82.12% on the SD and UD, respectively. When $K$ increases to 4, the dynamic

TABLE VI: Ablation study on the number of base kernels $K$ within the CSTA module.

| Datasets | Metric | $K=1$ | $K=2$ | $K=4$ | $K=8$ |
|---|---|---|---|---|---|
| $SD^n$ (Avg.) | Dice ↑ | 87.23 | 88.41 | **89.40** | 88.02 |
| | HD95 ↓ | 18.45 | 19.87 | **16.30** | 17.41 |
| $UD^n$ (Avg.) | Dice ↑ | 82.12 | 82.05 | **83.84** | 82.75 |
| | HD95 ↓ | 19.50 | 19.12 | **18.94** | 20.05 |

network intelligently combines multiple feature extraction experts based on the input image's global statistical priors. This achieves an optimal balance in representation capacity and yields the best performance. However, further increasing $K$ to 8 introduces an excessive number of base kernels. This drastically increases the optimization difficulty of the dynamic routing weights and subsequently triggers overfitting.

TABLE VII: Ablation study results of each component in GCC.

| GCC | | $SD^s$ (Avg.) | | $UD^t$ (Avg.) | |
|---|---|---|---|---|---|
| $\mathcal{L}_{SE}$ | $\mathcal{L}_{TI}$ | Dice ↑ | HD95 ↓ | Dice ↑ | HD95 ↓ |
| ✓ | | 88.42 | 18.04 | 82.55 | 20.03 |
| | ✓ | 88.67 | 17.62 | 82.68 | 19.93 |
| ✓ | ✓ | **89.40** | **16.30** | **83.84** | **18.94** |

### E. Effect of Geometric Consistency Constraints

*1) Ablation Study on GCC:* We analyzed the internal mechanisms of the geometric consistency constraint module to investigate the independent contributions of the structural equivariance ($\mathcal{L}_{SE}$) and texture invariance ($\mathcal{L}_{TI}$) penalties. As shown in Table VII, applying only $\mathcal{L}_{SE}$ anchors the pixel-level spatial coordinates of structural features. This yields an HD95 boundary error of 20.03 on the unseen domain (UD). Conversely, relying solely on $\mathcal{L}_{TI}$ to filter spatial geometric information from the texture representation results in an HD95 of 19.93 on the UD. This demonstrates the value of simultaneously imposing orthogonal geometric and physical constraints on structure and texture within the latent space. It effectively mitigates the representational ambiguity caused by naive feature decoupling. Consequently, it provides a robust physical and semantic guarantee for highly precise tissue boundary reconstruction.

*2) Visualization of Feature Disentanglement:* To visually demonstrate the effectiveness of the GCC module, we projected the latent representations into a 2D space using t-SNE. As shown in Fig. 6 (left), the original features across different domains—along with their structural and textural components—exhibit significant overlap and lack clear separation. In contrast, after applying the GCC module, the feature distribution exhibits distinct patterns, as illustrated in Fig. 6 (right). Our method not only effectively clusters and distinguishes different domain distributions but also strictly separates structural and texture features within a single domain. This demonstrates the effectiveness of the GCC module in enforcing latent space disentanglement and geometric meaningfulness.

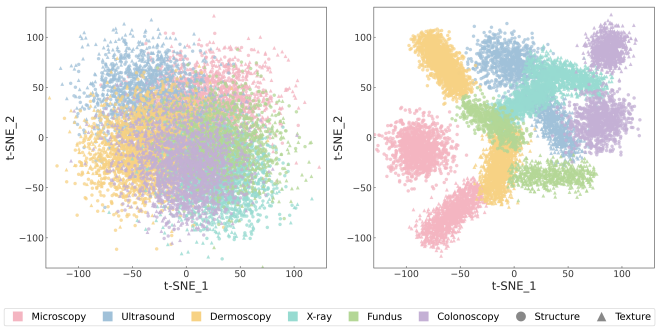


Fig. 6: t-SNE visualization of feature disentanglement.

### F. Hyperparameter Sensitivity Analysis

We conducted an in-depth sensitivity analysis to evaluate the impact of OrthoSeg's parameters on segmentation performance. Specifically, we analyzed $\alpha$ and $\beta$ from the geometric consistency weight in Eq. 8. We also evaluated the decay exponent $\gamma$ and the time-aware scheduling strategy $\lambda(t)$.

**Impact of $\alpha$ and $\beta$:** $\alpha$ and $\beta$ are the core parameters controlling the strengths of $\mathcal{L}_{SE}$ and $\mathcal{L}_{TI}$. As illustrated in Fig. 7, the model achieves superior performance when $\alpha = 0.3$ and $\beta = 1.0$. This indicates that an excessively large $\alpha$ leads to training instability, whereas an excessively small $\beta$ fails to thoroughly filter out scanner-specific noise. The optimal parameter combination effectively balances coordinate-sensitive topological reconstruction with the filtering of spatially invariant properties, thereby ensuring both accuracy and robustness.

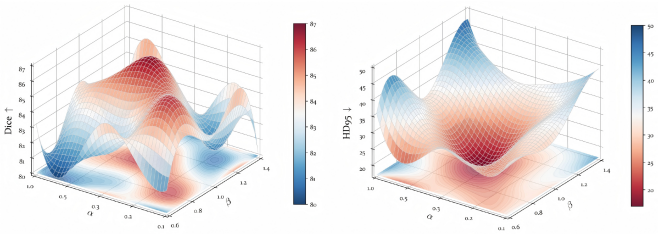


Fig. 7: Parameter sensitivity analysis on $\alpha$ and $\beta$ of OrthoSeg.

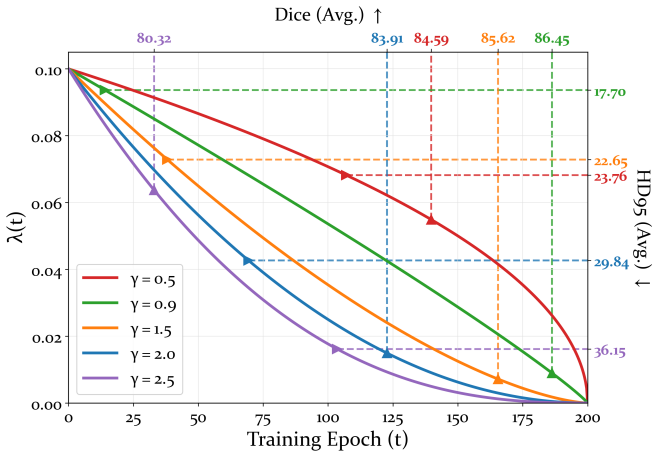


Fig. 8: Function patterns of $\lambda(t)$ under different $\gamma$ values and their corresponding segmentation performance.

**Impact of $\gamma$:** The decay exponent $\gamma$ determines the curvature of the mutual information decoupling intensity across training epochs. As shown in Fig. 8, a smaller $\gamma$ maintains

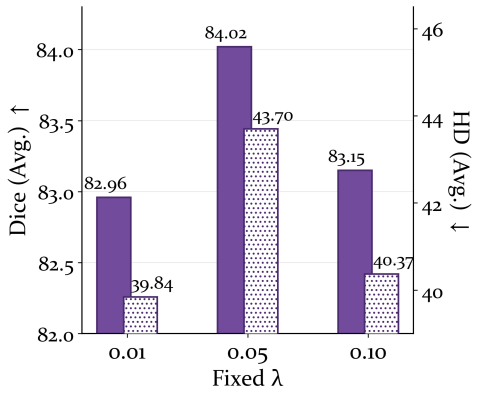


Fig. 9: Performance under different fixed weights $\lambda$

high-intensity decoupling for too long. This extended decoupling benefits feature separation but limits the fine-grained fitting of the segmentation task. Conversely, a larger $\gamma$ causes the decoupling constraint to vanish prematurely.

**Impact of $\lambda(t)$:** To verify the necessity of the time-aware scheduling strategy, we compared a fixed weight against the dynamic scheduling strategy. As shown in Fig. 9, the model's performance under all fixed weight settings is inferior to the dynamic scheduling scheme. This proves the value of imposing a strong decoupling penalty in the early stages of training to construct an orthogonal feature space. Subsequently, a progressive reduction of the regularization weight effectively releases the model's fitting capacity.

## V. Discussion

Existing cross-dataset segmentation is typically limited by the failure of global distribution alignment when facing severe domain heterogeneity. However, the superior performance of OrthoSeg suggests a paradigm shift from global alignment to explicit representation disentanglement. Decoupling through mutual information and actively stripping texture noise effectively addresses the impact of noise across datasets. Additionally, introducing geometric consistency constraints with equivariance and invariance penalties eliminates representation ambiguity and localizes anatomical boundaries. Furthermore, traditional networks heavily rely on static features. In contrast, our cross-scale topological aggregation adapts to multi-scale lesions by dynamically adjusting the receptive field. This further validates the effectiveness of structure-prior-based reconstruction when dealing with unknown domains.

The OrthoSeg architecture introduces several hyperparameters to adjust the feature decoupling and fusion process. However, we consistently used the same unified set of hyperparameters in all experiments across source and unseen domains. This avoided excessive tuning for specific datasets. Consequently, it objectively demonstrates the framework's strong generalization ability across diverse medical imaging scenarios, showcasing its potential for clinical applications.

Despite significant progress, this study mainly focuses on 2D slice-level segmentation. It has not yet applied inter-slice spatial continuity information in 3D imaging. Future work will aim to extend the current structure-texture separation paradigm to the 3D medical domain by incorporating inter-slice continuity. We also plan to explore its application in semi-supervised or unsupervised scenarios. Such extensions will further reduce the model's reliance on high-quality annotated data from the source domain.

## VI. Conclusion

In this work, we proposed OrthoSeg, a medical image segmentation network based on structure–texture orthogonal decoupling. To address severe domain shifts caused by diverse imaging protocols and scanner variations, we introduce mutual information decoupling and geometric consistency constraints. This approach collaboratively eliminates domain-specific texture interference in the latent space while providing physical grounding through spatial equivariance and invariance penalties. Moreover, we design a cross-scale topological aggregation module to dynamically reconstruct multi-scale lesions with low computational overhead. Extensive cross-dataset experiments demonstrate that OrthoSeg achieves competitive performance in both source-domain and cross-domain generalization while maintaining a low parameter count and computational complexity.

## VII. References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[2] A. S. Panayides, H. Chen, N. D. Filipovic, T. Geroski, J. Hou, K. Lekadir, K. Marias, G. K. Matsopoulos, G. Papanastasiou, P. Sarder, G. D. Tourassi, S. A. Tsaftaris, H. Fu, E. C. Kyriacou, C. P. Loizou, M. E. Zervakis, J. H. Saltz, F. E. Shamout, K. C. L. Wong, J. Yao, A. A. Amini, D. I. Fotiadis, C. S. Pattichis, and M. S. Pattichis, "Position paper: Artificial intelligence in medical image analysis: Advances, clinical translation, and emerging frontiers," *IEEE Journal of Biomedical and Health Informatics*, vol. 30, no. 2, pp. 1187–1202, 2026.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[4] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[5] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention*, vol. 11045. Springer, 2018, pp. 3–11.

[6] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[7] X. Zhao, H. Jia, Y. Pang, L. Lv, F. Tian, L. Zhang, W. Sun, and H. Lu, "M$^2$snet: Multi-scale in multi-scale subtraction network for medical image segmentation," *CoRR*, vol. abs/2303.10894, 2023.

[8] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. P. Lungren, S. Zhang, L. Xing, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, vol. 97, p. 103280, 2024.

[9] A. He, K. Wang, T. Li, C. Du, S. Xia, and H. Fu, "H2former: An efficient hierarchical hybrid transformer for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 9, pp. 2763–2775, 2023.

[10] Y. Zhu, D. Zhang, Y. Lin, Y. Feng, and J. Tang, "Merging context clustering with visual state space models for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 44, no. 5, pp. 2131–2142, 2025.

[11] L. Lan, P. Cai, L. Jiang, X. Liu, Y. Li, and Y. Zhang, "Brau-net++: U-shaped hybrid cnn-transformer network for medical image segmentation," *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2026.

[12] M. Wu, T. Liu, X. Dai, C. Ye, J. Wu, S. Funahashi, and T. Yan, "Hmda: A hybrid model with multi-scale deformable attention for medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 2, pp. 1243–1255, 2025.

[13] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision*. IEEE Computer Society, 2017, pp. 2242–2251.

[14] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 9242–9251.

[15] J. Li, Y. Zhang, L. Xu, Y. Yao, and L. Qi, "Isgan: Unsupervised domain adaptation with improved symmetric GAN for cross-modality multi-organ segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 6, pp. 3874–3885, 2025.

[16] W. Zhou, J. Ji, W. Cui, Y. Wang, and Y. Yi, "Unsupervised domain adaptation fundus image segmentation via multi-scale adaptive adversarial learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 10, pp. 5792–5803, 2024.

[17] C. Liu, Y. Cao, and H. Zhu, "The devil is in the frequency: Constrained and adaptive fine-grained domain perturbation for robust medical segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 11, pp. 8306–8319, 2025.

[18] J. Nam, N. S. Syazwany, S. J. Kim, and S. Lee, "Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2024, pp. 11 480–11 491.

[19] J. Cai, H. Li, M. Tan, B. He, W. Lv, and H. Li, "Cross-modal generalizable medical image segmentation with dual-domain deformable transformer and multitask adaptation," *Expert Systems with Applications*, vol. 277, p. 127249, 2025.

[20] M. Lei, H. Wu, X. Lv, and X. Wang, "Condseg: A general medical image segmentation framework via contrast-driven feature enhancement," in *AAAI Conference on Artificial Intelligence*, vol. 39, no. 5, 2025, pp. 4571–4579.

[21] Q. Liu, Q. Dou, and P.-A. Heng, "Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020, pp. 475–485.

[22] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[23] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 1055–1059.

[24] S. Roy, G. Köhler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jäger, and K. H. Maier-Hein, "Mednext: Transformer-driven scaling of convnets for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention*, ser. Lecture Notes in Computer Science. Springer, 2023, pp. 405–415.

[25] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention*, vol. 12901. Springer, 2021, pp. 14–24.

[26] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*. OpenReview.net, 2021.

[28] X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, "Missformer: An effective transformer for 2d medical image segmentation," *IEEE Trans. Medical Imaging*, vol. 42, no. 5, pp. 1484–1494, 2023.

[29] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 3, pp. 1173–1185, 2022.

[30] S. Kumari and P. Singh, "Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives," *Comput. Biol. Medicine*, vol. 170, p. 107912, 2024.

[31] X. Qi, Z. Wu, W. Zou, M. Ren, Y. Gao, M. Sun, S. Zhang, C. Shan, and Z. Sun, "Exploring generalizable distillation for efficient medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 7, pp. 4170–4183, 2024.

[32] L. Zhang, F. Wu, K. Bronik, and B. W. Papiez, "Diffuseg: Domain-driven diffusion for medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 5, pp. 3619–3631, 2025.

[33] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "Club: A contrastive log-ratio upper bound of mutual information," in *International conference on machine learning*. PMLR, 2020, pp. 1779–1788.

[34] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin *et al.*, "Nucleus segmentation across imaging experiments: the 2018 data science bowl," *Nature methods*, vol. 16, no. 12, pp. 1247–1253, 2019.

[35] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.

[36] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1605.01397*, 2016.

[37] J. Ma, C. Ge, Y. Wang, X. An, J. Gao, Z. Yu, M. Zhang, X. Liu, X. Deng, S. Cao, H. Wei, S. Mei, X. Yang, Z. Nie, L. Chen, L. Tian, Y. Zhu, Q. Zhu, G. Dong, and J. He, "Covid-19 ct lung and infection segmentation dataset (version verson 1.0)," 2020.

[38] J. I. Orlando, H. Fu, J. B. Breda, K. Van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee *et al.*, "Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Medical Image Analysis*, vol. 59, p. 101570, 2020.

[39] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.

[40] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*. Springer, 2019, pp. 451–462.

[41] T. L. Dinh, S.-G. Kwon, S.-H. Lee, and K.-R. Kwon, "Breast tumor cell nuclei segmentation in histopathology images using efficientunet++ and multi-organ transfer learning," *Journal of Korea Multimedia Society*, vol. 24, no. 8, pp. 1000–1011, 2021.

[42] Z. Zhuang, N. Li, A. N. Joseph Raj, V. G. Mahesh, and S. Qiu, "An rdau-net model for lesion segmentation in breast ultrasound images," *PloS one*, vol. 14, no. 8, p. e0221535, 2019.

[43] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "Ph 2-a dermoscopic image database for research and benchmarking," in *International Conference of the IEEE Engineering in Medicine and Biology Society*, 2013, pp. 5437–5440.

[44] "Covid19 dataset," https://www.kaggle.com/datasets/piyushsamant11/pidata-new-names.

[45] J. Sivaswamy, S. Krishnadas, A. Chakravarty, G. Joshi, A. S. Tabish *et al.*, "A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis," *JSM Biomedical Imaging Data Papers*, vol. 2, no. 1, p. 1004, 2015.

[46] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of healthcare engineering*, vol. 2017, no. 1, p. 4037190, 2017.

[47] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2015.

[48] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

**Kai Han** received the B.S. degree from Jiangsu University of Science and Technology, Jiangsu, China, in 2019, and the Ph.D. degree in engineering from Jiangsu University, Jiangsu, China, in 2025. He is currently a lecturer with the School of Computer Science and Communication Engineering, Jiangsu University. His research interests include medical image processing and multimodal lesion diagnosis.

**Jiaqi Zhang** is currently pursuing the B.S. degree at the School of Computer Science and Communication Engineering, Jiangsu University. His research interests include medical image processing and computer vision.

**Chongwen Lyu** received his B.S. degree from Xuzhou Medical University in 2020. He is currently pursuing a M.S. degree in the School of Computer Science and Communication Engineering of Jiangsu University. His research interests include medical image processing and text report generation.

**Mengting Li** is currently pursuing the B.S. degree at the School of Computer Science and Communication Engineering, Jiangsu University. Her research interests include medical image processing and multimodal vision.

**Jun Chen** received the M.E. degree from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2019, and the Ph.D. degree from the School of Biomedical Engineering, Sun Yat-sen University, Shenzhen, China, in 2023. He is currently a lecturer with the School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China. His research interests include medical image analysis and computer vision.

**Laihua Yang** received the bachelor's degree in medical imaging from Wannan Medical College in 2003, and the master's degree in imaging medicine and nuclear medicine from Jiangsu University in 2014. He is currently an Associate Professor and a Master's Supervisor. He also serves as a Deputy Chief Physician and the Deputy Director of the Imaging Department at Danyang Traditional Chinese Medicine Hospital. His main research interests focus on abdominal imaging.

**Guangquan Zhou** (Senior Member, IEEE) received the B.Sc. and M.S. degrees from Southeast University, Nanjing, China, in 2000 and 2003, respectively, and the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong, in 2015, all in biomedical engineering. He is currently an Associate Professor with the School of Biological Sciences and Medical Engineering, Southeast University. His research interests include medical image processing and analysis, ultrasound imaging, 3-D ultrasound imaging, pattern recognition, and computer-aided diagnosis.

**Yang Chen** (Senior Member, IEEE) received the MS and PhD degrees in biomedical engineering from first military Medical University, China, in 2004 and 2007, respectively. Since 2008, he has been a faculty member with the Department of Computer Science and Engineering, Southeast University, China. His recent work concentrates on the medical image reconstruction, image analysis, pattern recognition, and computerized-aid diagnosis.

**Zhe Liu** received her bachelor degree in Engineering in Computer Science and Technology from Jilin Normal University in 2004, her master degree in Science in Computer Software and Theory from Jilin University in 2008 and her Ph.D. degree in computer science from Jiangsu University in 2012. She is a visiting scholar at the Department of Radiology at the University of Pittsburgh Medical Center, Pennsylvania, USA, and also a full professor at the School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang. Her research interests include medical image processing, data mining, and pattern recognition.