Dynamic Strip Convolution and Adaptive Morphology Perception Plugin for Medical Anatomy Segmentation

Guyue Hu, Yukun Kang, Gangming Zhao, Zhe Jin, Chenglong Li, and Jin Tang

***Abstract*— Medical anatomy segmentation is essential for computer-aided diagnosis and lesion localization in medical images. For example, segmenting individual ribs benefits localizing the lung lesions and providing vital medical measurements (such as rib spacing) for generating medical reports. Existing methods segment shape-different anatomies (such as striped ribs, bulky lungs, and angular scapula) with the same network architecture, the morphology heterogeneity is heavily overlooked. Although some shape-aware operators like deformable convolution and dynamic snake convolution have been introduced to cater to specific object morphology, they still struggle with orientation-varying strip structures, such as 24 ribs and 2 clavicles. In this paper, we propose a novel convolutional plugin (DSC-AMP) for medical anatomy segmentation, which is comprised of a dynamic strip convolution (DSC) operator and an adaptive morphology perception (AMP) strategy. Specifically, the dynamic strip convolution customizes gradually varying directions and offsets for each local region, achieving dynamic striped receptive fields. Additionally, the adaptive morphology perception strategy incorporates insights from various shape-aware convolutional kernels, enabling the model to discern and integrate crucial representations corresponding to heterogeneous anatomies. Extensive experiments on two large-scale datasets demonstrate the effectiveness and superiority of the proposed approach for tackling heterogeneous medical anatomy segmentation.**

***Index Terms*— Dynamic Strip Convolution, Adaptive Morphology Perception, Medical Anatomy Segmentation**

I. Introduction

This work was supported by the National Natural Science Foundation of China (No. 62376004), Anhui Provincial Natural Science Foundation (No. 2408085QF201, No. 2208085J18), Natural Science Foundation of Anhui Higher Education Institution (No. 2022AH040014), and Open Project of Anhui Provincial Key Laboratory of Security Artificial Intelligence (No. SAI2024003). (Corresponding author: Chenglong Li)

Guyue Hu and Chenglong Li are with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Security Artificial Intelligence, School of Artificial Intelligence, Anhui University, Hefei, 230601, China (e-mail: guyue.hu@ahu.edu.cn; lcl1314@foxmail.com)

Zhe Jin is with the School of Artificial Intelligence, Anhui University, Hefei, 230601, China (e-mail: jinzhe@ahu.edu.cn)

Yukun Kang and Jin Tang are with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, 230601, China (e-mail: kyk30@stu.ahu.edu.cn; tangjin@ahu.edu.cn)

Gangming Zhao is with the Department of Computer Science, The University of Hong Kong, Hong Kong (e-mail: gmzhao@connect.hku.hk)

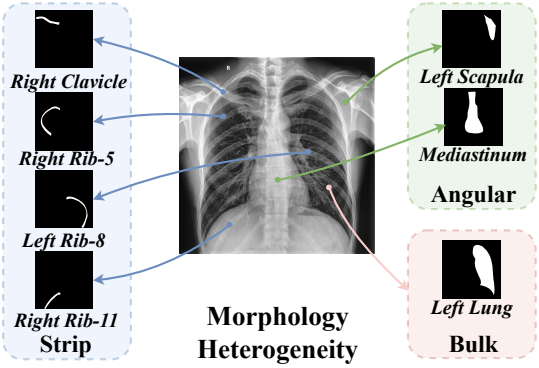

Fig. 1: Illustration of morphology-heterogeneous anatomies in medical chest X-ray images.

MEDICAL anatomy segmentation in X-ray images plays an important role in computer-aided diagnosis and pinpointing abnormalities of various diseases, such as rib fractures, pneumothoraces, and pulmonary infections [1]–[3]. As a crucial foundation for early disease detection and localization, medical anatomy segmentation significantly reduces the workload of healthcare professionals. This enables them to prioritize patient care and treatment, thus greatly enhancing the efficiency of the medical community.

Benefiting from the boom of deep neural networks, medical image segmentation has achieved huge success. The pioneer U-Net [4] utilizes an encoder-decoder structure for shape-agnostic representation learning and mask prediction. Since then, its CNN-based [4]–[9], transformer-like [10], [11], and hybridized variants [12], [13] have dominated the field of medical image segmentation. For example, nnUNet [5] automatically configures its architecture and parameters to adapt different medical imaging datasets. TransUnet [14] integrates Vision Transformer [15] into classical U-Net to capture global information and long-term dependency. The latest VM-UNet [16] combines Vision-Mamba [17] with U-Net to exploit the great capability of establishing long-distance dependencies while upholding linear complexity from the State Space Model, achieving powerful segmentation performance.

However, conventional medical image segmentation approaches suffer from significant performance degradation when dealing with medical anatomy segmentation since the morphology heterogeneity has been heavily overlooked in ex-

isting methods. Although some previous works have incorporated specific shape priors or regularization into conventional segmentation models [18], [19], they have not tackled the issue of heavy morphological heterogeneity.As qualitatively illustrated in Fig. 1, each rib or clavicle exhibits an orientation-varying striped structure, each lung displays a bulky shape, each scapula appears flat with angular protrusions, while the mediastinum presents an irregular angular form. Furthermore, the matrix of Procruste Disparity [20] computed from medical anatomies (see Fig. 5) also quantitatively indicates the serious heterogeneity in classical medical anatomy segmentation datasets. Therefore, different morphology-aware receptive fields are urgently required to fit these morphology-heterogeneous anatomies properly.

Besides, some pioneer works have explored classical shape-aware convolution to accommodate specific shapes in the conventional field of computer vision. Deformable convolution [6] adjusts its sampling grid with learnable offsets suitable for roughly isotropic morphology, such as bulky lungs in Fig. 1. AKConv [21] gives an arbitrary number of parameters and arbitrary sampling shapes to convolution kernels providing richer options for the trade-off between computational overhead and shape-fitting performance. DSConv [22] focuses on slender and tortuous structures stretching along a specific axis. Unfortunately, there is still no customized convolutional operator for dynamic striped morphology with varying stretch orientation and length-width ratio, such as 24 ribs and 2 clavicles in Fig. 1.

To move beyond such limitations, we introduce a novel convolution plugin to tackle heterogeneous medical anatomy segmentation (referred to as DSC-AMP), comprising dynamic strip convolution (DSC) operator and adaptive morphology perception (AMP) strategy. The DSC operator is explored to fit varying morphology with intrinsic dynamic striped receptive fields. Furthermore, the AMP strategy adaptive integrates representations from diverse shape-aware kernels to attend to specific heterogeneous anatomy.

In summary, the main contributions of this paper could be summarized as follows:

1) We propose a novel convolution plugin DSC-AMP for heterogeneous medical anatomy segmentation, which could reform off-the-shelf CNN-based, CNN-hybridized image segmentation network to be morphology-aware by simple layer replacing.
2) We design a dynamic strip convolution that adaptive focuses on elongated structures with gradually varying stretch orientation and length-width ratio, realizing precise segmentation of challenging striped anatomies (such as 24 ribs and 2 clavicles).
3) We introduce an adaptive morphology perception strategy that matches different local morphological structures with appropriate kernels and adaptive integrates these diverse representations, significantly improving the representation diversity and robustness.
4) The proposed DSC-AMP achieves state-of-the-art performance on two large-scale datasets for medical anatomy segmentation, demonstrating its effectiveness in tackling morphological heterogeneity.

II. Related Work

A. Medical Anatomy Segmentation

Medical anatomy segmentation is a critical technique for healthcare that involves assigning an anatomy class label to each pixel of an anatomical structure in medical images [1]–[3], [23]–[25]. Human medical anatomies usually exhibit two intrinsic characteristics: locality and compositionality [26]. The locality means heavy diversity of local morphology, size, and orientation across or within different anatomies. The compositionality means the large structures consisting of smaller structures, such as the ribs consisting of 24 individuals. As a result, medical anatomy segmentation is somewhat more challenging than conventional medical image anatomy segmentation tasks. From the perspective of the network structures, the medical image segmentation methodologies primarily encompass three paradigms including the CNN-based methods [4]–[7], [27], transformer-based approaches [10], [11], and hybridized variants [12], [13]. The popular U-Net [4] and its CNN-based variants [5], [7], [27] usually adopt a U-shaped encoder-decoder network structure and exploit some skip connections to preserve detailed information on human anatomies. They have demonstrated remarkable effectiveness on small-scale datasets for medical anatomy segmentation. Since the groundbreaking work TransUnet [14], the transformer-based segmentation approaches [10]–[13], [28] has achieved rapid development and significantly enhanced the representation generalization. They harness the formidable long-range information acquisition capabilities from Vision Transformers (ViT) [15]. For example, the latest VM-UNet [16] combines the Vision-Mamba [17] with classical U-Net [4] that establishes long-distance dependencies while still upholding linear complexity.

Apart from the supervised paradigm, the segmentation approaches based on foundation segmentation models and parameter-efficient tuning (PET) technique [29], [30] have greatly advanced the image segmentation field. The recent debut of the Segment Anything Model (SAM) [31] marks a significant milestone that extends the prompt-driven paradigm into image segmentation. Then, the visual foundation models also quickly demonstrated their prowess in the domain of medical image segmentation as well. For example, the MedSAM [32] achieves highly accurate segmentation across diverse modalities and targets via meticulously curating a multi-million level dataset and carefully refining SAM on this large-scale dataset. The nnSAM [33] further combines the powerful representation learning ability from SAM [31] with the adaptive configuration capability from nnUNet [5], thereby facilitating dataset-tailored representation learning for medical image segmentation. Despite the huge process in medical anatomy segmentation, most of the existing methods have not incorporated morphological knowledge about human anatomies (such as the 24 ribs in X-ray chest images gradually extending along various directions exhibiting dynamic strip shapes, and the 2 lungs appearing as bulk-like structures). The proposed DSC-AMP could utilize abundant morphological knowledge from anatomical structures to effectively facilitate medical anatomy segmentation.

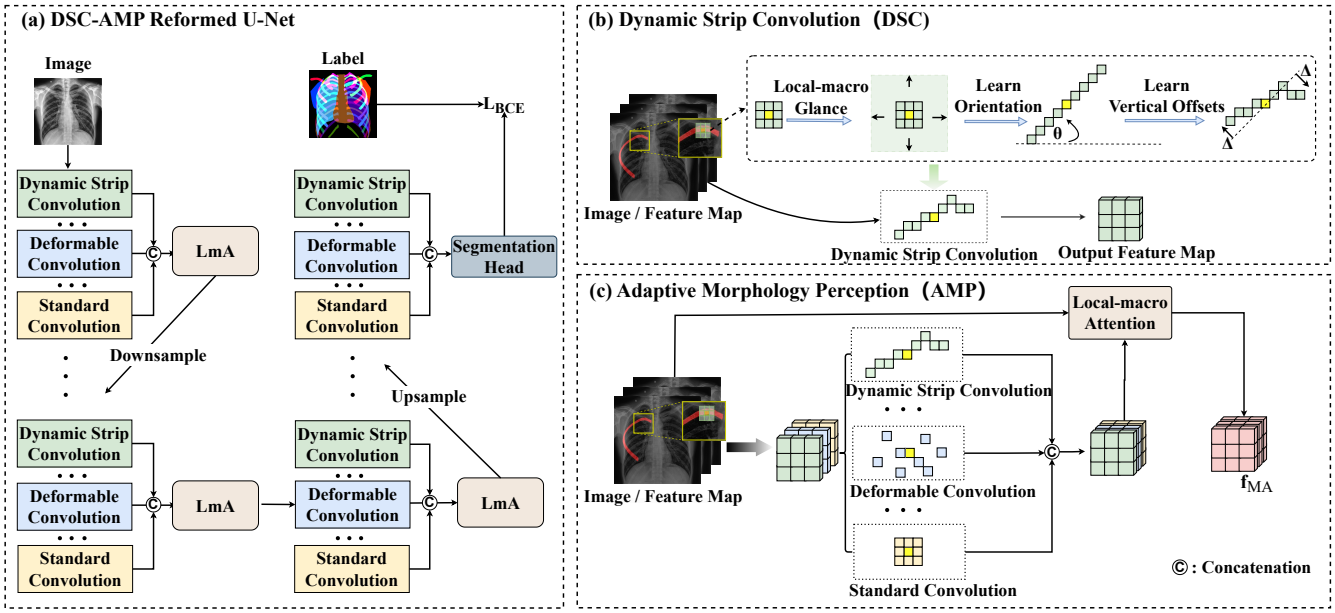

Fig. 2: Overview of the proposed Dynamic Strip Convolution and Adaptive Morphology Perception plugin (DSC-AMP). (a) The illustration of reforming U-Net with the DSC-AMP plugin. (b) The mechanism of Dynamic Strip Convolution (DSC) operator. (c) The mechanism of Adaptive Morphology Perception (AMP) strategy.

B. Shape-aware Convolution

The receptive fields of conventional convolutional kernels are fixed as rectangular shapes, which are insensitive to the geometric morphology in medical images. Some pioneer works adapt the sampling region in convolution to adaptively accommodate the morphological variation of segmentation objects [34]–[40]. To accommodate the changes in object shape and size, the dilated convolution [34] enlarges the receptive field of a convolutional kernel without increasing the number of parameters via a novel dilating operation. In contrast to dilated convolution, which only rigidly expands the receptive field by scaling its original rectangle shape, deformable convolution [35] freely learns offsets for each position in the feature map aiming at adjusting its receptive field in the form of arbitrary shape. DCU-net [39] further combines the deformable convolution with a cascaded U-Net structure and adaptively adjusts the receptive field of its convolution kernels to better fit the scales and shapes of blood vessels. D-LKA Net [41] employs large kernel convolution to enhance the receptive field and combines deformable convolution to focus on shape-related features. The AKConv [21] expands the flexibility of the convolution kernels by accommodating an arbitrary number of parameters and various sampling shapes. This enhancement enriches the option toolboxes for balancing computational overhead and task performance, offering a more nuanced trade-off. DSConv [22] is designed to enhance the perception of thin and tortuous tubular structures. In contrast to the deformable convolution which learns geometric changes freely, the DSConv imposes additional continuity constraints to prevent the perceptual field from wandering off the segmentation target region, especially for thin tubular structures. However, the existing methods still have difficulty in capturing the characteristics of dynamic striped structures with varying stretch orientations and length-width ratios, such as 24 striped ribs. Additionally, morphology heterogeneity has been heavily overlooked in existing methods and urgently requires an adaptive morphology perception strategy that could effectively mine and incorporate diverse representations from heterogeneous medical anatomies.

III. Method

A. Overview Pipeline

We first introduce the overview pipeline of our dynamic strip convolution and adaptive morphology perception plugin (DSC-AMP) tailored for heterogeneous anatomy segmentation in medical images. As illustrated in Fig. 2, the DSC-AMP is composed of two pivotal elements including dynamic strip convolution (DSC) operator and adaptive morphology perception (AMP) strategy. Specifically, the DSC is delicately designed for the intricate dynamic striped anatomy with the morphology of gradually varying stretch orientation and length-width ratios, such as 24 ribs and 2 clavicles. As shown in Fig. 2 (b), the DSC learns an orientation angle and vertical offsets for each local strip structure via a Local-Macro Glance (LmG), then constructs a dynamic strip receptive field with them. As depicted in Fig. 2 (c), the AMP further seamlessly integrates a series of diverse convolutional kernels via local-macro observation and cross-attention mechanism, including standard convolution, deformable convolution, and the proposed dynamic strip convolution. This incorporation enables morphology-aware representation learning across substantial morphology-diverse medical anatomies. Eventually, the proposed DSC-AMP plugin could reform off-the-shelf CNN-based, CNN-hybridized image segmentation network to be more morphology-aware simply by substituting corresponding CNN layers with the proposed DSC-AMP. Without losing

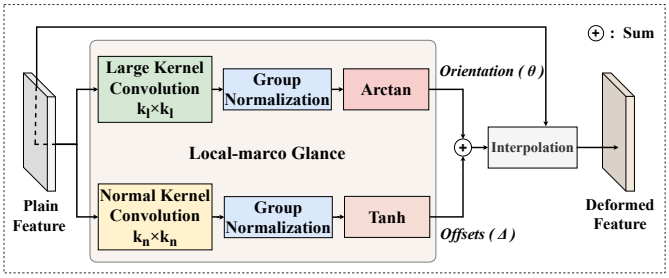

Fig. 3: Detailed structure of the dynamic strip convolution

generality, we take the classical U-Net [4] as an example to illustrate the layer replacement operation, which is shown in Fig. 2 (a). In the subsequent sections, we will dig into the details of the DSC and AMP components separately.

B. Dynamic Strip Convolution

In this section, we will introduce the underlying mechanism and detailed implementation of the proposed dynamic strip convolution. Given a standard 2D convolution kernel, the convolution range of the kernel is confined within a square area, which is shape-agnostic and incapable of capturing the intrinsic representation of dynamic striped medical anatomy. In order to conform strip structures that gradually stretch along arbitrary orientations, inspired by classical deformable convolution [6], we design an angle-offset prediction module in Fig. 3. The module learns an orientation angle θ and corresponding offset Δ via a Local-macro Glance (LmG), which could provide a relative macro perception of local morphology around the kernel center. The angle $\theta \in [-\pi/2, \pi/2]$ denotes the stretch orientation of a strip structure, and the offset $\Delta = \{\delta_j \mid \delta_j \in [-1, 1], j \in \{\lfloor -(k\times k)/2 \rfloor, \cdots, -1, 0, 1, \cdots, \lfloor (k\times k)/2 \rfloor\}\}$ denotes the vertical offset corresponding each location j in a flattened convolution kernel along this direction. For example, the angle-offset prediction module needs to learn one θ and eight δ_j to form a strip convolution kernel of size 3×3 (the offset value at the kernel center is set as 0 directly).

Assume that we have a Cartesian coordinate system $Coord'$: (x', y') in a small local region whose horizontal and vertical directions are respectively paralleled and vertical with the learned orientation angle θ (see Fig. 2 (b)). Given a standard 3×3 convolution kernel in this coordinate system, the positions of its flattened sampling grid could be represented as $C'_{i\pm d} = (x'_{i\pm d}, y'_{i\pm d})$, where $d = \{0, 1, 2, 3, 4\}$ denotes distance from the central point along the x' axis. Inspired by [6], [22], the sampling position of a local strip convolution in this local region is defined as

$$C'_{i\pm d} = \begin{cases} (x'_{i+d}, y'_{i+d}) = (x'_i + d, y'_i + \sum_{j=i}^{i+d} \delta_j), \\ (x'_{i-d}, y'_{i-d}) = (x'_i - d, y'_i + \sum_{j=i-d}^{i} \delta_j). \end{cases} \tag{1}$$

Since the offset $\delta_j \in \Delta$ is successively summed along the x' axis in Eq. 1, it ensures a gradual (non-mutational) change in stretch orientation and length-width ratio of the convolution kernel. Thus, the strip convolution caters to a strip morphology very well. As illustrated in Fig. 3, an angle-offset prediction module further exploits a relative macro glance of local morphology around the kernel center to learn a position-specific orientation angle θ and corresponding offset Δ for each local region, gradually forming a dynamic striped receptive filed in the whole feature map to dynamically modeling varying strip structures, such as 24 ribs and 2 clavicles. The relative macro glance is realized by expanding the local receptive field of the angle-offset prediction module from 3×3 to 5×5, thus providing a relative macro perception of local morphology to enhance its morphology perception ability.

Since the local coordinate system $Coord'$: (x', y') varies with the stretching orientation θ, we exploit a rotating formula (Eq. 2) to it back to the original coordinate system of feature map $Coord$: (x, y), thus simplifying the calculation process and making it convenient to be plugged in existing CNN-based segmentation networks, *i.e.*

$$\begin{cases} x = x^{'} \cos\theta - y^{'} \sin\theta, \\ y = x^{'} \sin\theta + y^{'} \cos\theta, \end{cases} \tag{2}$$

where x and y respectively denote the horizontal and vertical coordinates of a point in the original coordinate system, while $x^{'}$ and $y^{'}$ denote the horizontal and vertical coordinate of the corresponding point in the rotated local coordinate system along the direction of angle θ. Given the receptive field of the proposed dynamic strip convolution in the original coordinate system coordinates at $C_{i\pm d} = (x_{i\pm d}, y_{i\pm d})$, we can derive the following transform formula, *i.e.*

$$C_{i\pm d} = \begin{cases} (x_{i+d}, y_{i+d}) = \big(d\ \cos\theta - \delta_{i+d}\ \sin\theta, \\ \quad d\ \sin\theta + \delta_{i+d}\ \cos\theta\big), \\ (x_{i-d}, y_{i-d}) = \big(-d\ \cos\theta - \delta_{i-d}\ \sin\theta, \\ \quad -d\ \sin\theta + \delta_{i-d}\ \cos\theta\big), \end{cases} \tag{3}$$

where the θ and $\delta_j \in \Delta$ are the dynamic orientation and offset predicted by the angle-offset prediction module in Fig. 3.

In consideration that the predicted offset is typically fractional, the following bilinear interpolation operation is further required during the implementation of our DSC, *i.e.*

$$C = \textstyle\sum_{C''} bilinear\_interpolation(C'', C) \cdot C'' \tag{4}$$

where C denotes a fractional location in Eq. 2 and C'' accounts for every integral spatial positions.

It is worth noting that the classical DSConv [22], which is adept at slender tubular structure along the horizontal or vertical axis, could be treated as a special case of our dynamic strip convolution (DSC) when the stretch orientation θ is fixed on the horizontal or vertical direction. Therefore, it requires additional deliberately-designed multi-view fusion strategy [22] to *implicitly* fit varying structures that don't proceed on the axes. In contrast, the proposed DSC *directly* learns the dynamic orientation and offset for each local region forming a dynamic strip receptive field, which is much better for modeling gradually varying strip morphology, such as 24 ribs and 2 clavicles.

C. Adaptive Morphology Perception

In this section, we dig into the details of the adaptive morphology perception (AMP) strategy which intends to adaptively mine and integrate crucial representations corresponding

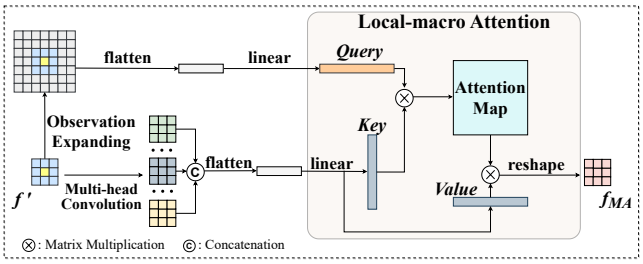

Fig. 4: Implementation details of the proposed Adaptive Morphology Perception strategy

to heterogeneous anatomies during medical image segmentation. We observe that the anatomical structures exhibit substantial morphological variations in both inter-class (*e.g.*, between ribs and lungs) and intra-class (*e.g.*, among different ribs), which is highly heterogeneous (see Fig. 1). Therefore, we apply different morphology-aware and morphology-unaware kernels to capture comprehensive representations for heterogeneous anatomies from different perspectives. Specifically, as shown in Fig. 2 (c), our Dynamic Strip Convolution (DSC) aims at capturing the representation of gradually varying strip-like structures (such as ribs and clavicles), the deformable convolution intends to cater to approximately isotropic bulk-like structures (such as lungs), the standard convolution extracts conventional shape-agnostic representations, etc.

A straightforward manner to integrate these diverse representations is element-wise addition or concatenation. However, they are completely unaware of the local morphology around the kernel center thus impairing the representation discriminativeness. To address this issue, we design a Local-macro Attention (LmA) to guide the representation incorporation process adaptively. As illustrated in Fig. 4, the comprehensive morphology-aware representation $\mathbf{f}_{MA}$ for each convolution layer is obtained via

$$\mathbf{f}_{MA} = \text{LmA}(expand(\mathbf{f}^{'}, \lambda), concat(\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_n)) \quad (5)$$

where the LmA denotes the proposed Local-macro Attention technique, $\mathbf{f}^{'}$ is the local feature map before the above concurrent convolutions, $\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_n$ are n diverse representations perceived from n types of convolution kernels (such as dynamic strip convolution, deformable convolution, etc.). While the magnification factor λ serves as a hyperparameter for expanding the observation range of LmA.

The implementation details are illustrated in Fig. 4. We first expand the observation range of each position in the current convolution layer by λ (*i.e.* magnification factor) times. Then, we utilize it as the query $\boldsymbol{Q}$ and the concatenated representations from diverse types of convolution kernel as the key $\boldsymbol{K}$ and value $\boldsymbol{V}$ for a Local-macro Attention (LmA). Since the magnified feature map in a local area could exhibit more pronounced geometric morphological characteristics, the representations that better conform to local morphology will obtain higher attention scores in the LmA, eventually realizing an adaptive morphology perception and representation incorporation in the corresponding layer. The detailed calculation process of the LmA module is implemented as follows:

$$\begin{cases} \mathbf{Q} = linear(flatten(expand(\mathbf{f}^{'}, \lambda))) \\ \mathbf{K}, \mathbf{V} = linear(flatten(concat(\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_n))) \\ Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\dfrac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \cdot \mathbf{V} \\ \mathbf{f}_{MA} = reshape(Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V})) \end{cases} \quad (6)$$

where the function $expand$ denotes the mentioned operation above for expanding the observation range, and the $linear$ represents a linear layer. The remain functions including $flatten$, $reshape$, $softmax$, $Attention$ are corresponding operations as the name indicates. Finally, the adaptively weighted representation f_{MA} has comprehensive insights from different shape-aware convolutional kernels that concentrate on diverse morphological structures, significantly facilitating the morphology perception capacity of the proposed DSC-AMP.

Taking the convolutional segmentation networks built on classical stage-block design as an example (*e.g.* U-Net), we illustrated the detailed procedure of our DSC-AMP in Algorithm 1. It consists of a DSC-AMP reformed encoder, a DSC-AMP reformed decoder, and a segmentation head. Finally, the whole segmentation network is end-to-end optimized by the classical Binary Cross-Entropy (BCE) loss [4].

IV. Experiment

A. Datasets

1) CXRS Dataset: The Chest X-ray Segmentation (CXRS) dataset is an in-house dataset curated and annotated jointly by the Anhui University and the Anhui University of Chinese Medicine. It comprises 1254 high-resolution chest X-ray images, each accompanied by pixel-wise annotations of 31 distinct anatomical structures. Most of the X-ray images in the dataset have a resolution higher than $2K \times 2K$, and some samples include challenging conditions such as pulmonary lesions. The anatomical structures annotated in this dataset include 24 ribs, 2 clavicles, 2 scapulae, 2 lungs, and 1 mediastinum. We divided the dataset into training, validation, and test sets in a 7:1:2 distribution, respectively comprising 879 training, 125 validation, and 250 test samples. Moreover, the proportion of normal samples to abnormal samples (including those with pathological lesions) is consistently maintained at a 3:2 ratio within each subset.

As mentioned in the Introduction, the medical anatomies in chest X-ray images qualitatively exhibit significant morphological heterogeneity (refers to Fig. 1). To further quantify this heterogeneity, we employ the Procrustes Analysis technique [20] to calculate the Procruste Disparity between each pair of anatomical structures. A higher Procruste Disparity value indicates larger morphological differences. Fig. 5 quantitatively shows that the anatomical structures in the CXRS dataset exhibit a high degree of morphological heterogeneity, where the anatomy index 1 to 24 represent the ribs, 25 and 26 denote the clavicles, 27 and 28 correspond to the scapulae, 29 and 30 are attributed to the lungs, and 31 signify the mediastinum. The result indicates the CXRS dataset is very suitable for examining our Dynamic Strip Convolution and Adaptive Morphology Perception Plugin (DSC-AMP).

Algorithm 1 Overall procedure of the proposed DSC-AMP

Input: Images $\mathcal{X}$, Labels $\mathcal{Y}$, Number of stages N_{stage}, Number of blocks per stage N_{block}
Output: Updated model
for (x, y) in $(\mathcal{X}, \mathcal{Y})$ **do**
▷ *DSC-AMP Reformed Encoder*
  for $i = 0$ to N_{stage} **do**
    for $j = 0$ to N_{block} **do**
      $\mathbf{f}_1 = standard\_conv(x)$
      $\mathbf{f}_2 = deformable\_conv(x)$
      $\cdots$
      $\mathbf{f}_n = dynamic\_strip\_conv(x)$
      $x \leftarrow \text{AMP}(concat(\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_n))$
    end for
    $f^i_{ma} \leftarrow \text{Downsample}(x)$
  end for
▷ *DSC-AMP Reformed Decoder*
  for $i = N_{stage} - 1$ to -1 **do**
    $x_i \leftarrow \text{Upsample}(x)$
    $x \leftarrow \text{Concat}(f^i_{ma}, x_i)$
    for $j = 0$ to N_{block} **do**
      $\mathbf{f}_1 = standard\_conv(x)$
      $\mathbf{f}_2 = deformable\_conv(x)$
      $\cdots$
      $\mathbf{f}_n = dynamic\_strip\_conv(x)$
      $x \leftarrow \text{AMP}(concat(\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_n))$
    end for
  end for
▷ *Segmentation Head*
  $\hat{y} = segmentation\_head(x)$
  Calculate segmentation loss $\mathcal{L} \leftarrow \mathcal{L}_{BCE}(\hat{y}, y)$
  Calculate gradient and update model via the loss $\mathcal{L}$
end for

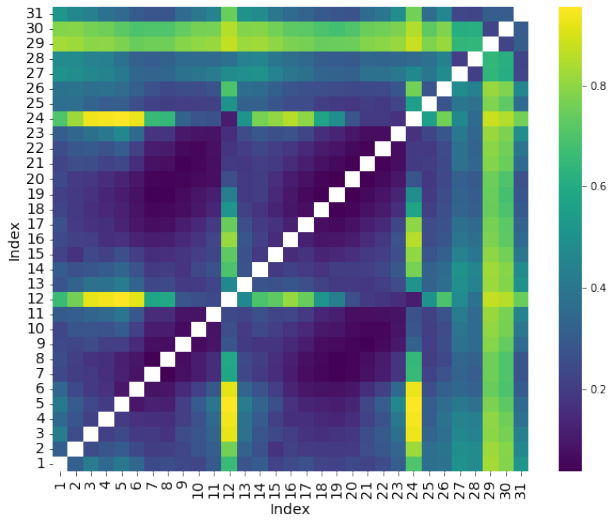

Fig. 5: The Procrustes Disparity between different medical anatomies in the CXRS dataset

2) SMOS Dataset: The Synapse Multi-Organ Segmentation (SMOS) dataset [42] is a heterogeneous multi-organ segmentation dataset in the form of CT scans. The dataset employed in our experiments comprises abdominal CT scans from 30 cases sourced from the Multi-Atlas Abdomen Labeling Challenge in MICCAI 2015 [42]. It contains a total number of 3779 axial abdominal clinical CT images. Each CT volume consists of 85 to 198 slices (512×512 pixels), and the voxel spatial resolution ranges from ([0.54–0.54]×[0.98–0.98]×[2.5–5.0]) mm^3. Following the protocol in [10], [14], we split the dataset into training and testing sets that contain axial slices (images) from 18 cases (2212 samples) and 12 cases, respectively.

B. Evaluation Metrics

Regarding the CXRS dataset, we employ the mean Dice Similarity Coefficient (mDice) and mean Intersection over Union (mIoU) as evaluation metrics. The mIoU is the class-wised mean of the Intersection over Union (IoU) for each segmentation class and mDice is the class-wised mean of the Dice Similarity Coefficient (DSC) for each segmentation class. Formally, the predicted segmentation mask can be categorized into True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) per its relation with the ground-truth mask. Then, the DSC is formulated as follows:

$$\text{DSC} = \frac{2 \times TP}{2 \times TP + FP + FN}, \tag{7}$$

while the IoU is computed via Eq. 8, *i.e.*

$$\text{IoU} = \frac{TP}{TP + FP + FN}, \tag{8}$$

thus mIoU in the CXRS dataset is the average of the IoUs corresponding to 31 medical anatomies.

As for the SMOS dataset, following [14], we utilize the average mean Dice Similarity Coefficient (mDice) and average Hausdorff Distance (HD) as the primary metrics to assess the performance across 8 abdominal organs (including the aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, and stomach). Besides, the mDice is also exclusively employed to evaluate segmentation performance for specific organs in this paper. The HD metric is formulated as follows:

$$\text{HD}(Y, \hat{Y}) = \max\left\{\max_{y \in Y} \min_{\hat{y} \in \hat{Y}} d(y, \hat{y}), \max_{\hat{y} \in \hat{Y}} \min_{y \in Y} d(\hat{y}, y)\right\} \tag{9}$$

where Y and $\hat{Y}$ are the ground-truth mask and predicted segmentation mask, respectively. The term $d(y, \hat{y})$ represents the Euclidean distance between points y and $\hat{y}$.

C. Implementation Details

Regarding the main hyper-parameters in the proposed Dynamic Strip Convolution and Adaptive Morphology Perception Plugin (DSC-AMP), the convolution kernel for Relative-macro Glance (see Fig. 2 (b) and Fig. 3) is empirically set as 5×5, the magnification factor λ in Local-macro Attention module (see Fig. 4) is empirically set as 3. When reforming classical U-Net with the proposed DSC-AMP (see Fig. 2 (a)), the number of output channels corresponding to each stage is set as [64, 128, 256, 512, 1024].

TABLE I: Comparing with the state-of-the-art methods in medical anatomy segmentation on the CXRS dataset

Methods	mIoU (%) ↑	mDice (%) ↑	Ribs	Clavicles	Scapulaes	Lungs	Mediastinum
U-Net [4] (baseline)	66.85	77.35	73.57	89.33	86.01	95.06	91.14
Unet++ [43]	65.19	76.04	72.45	87.88	80.79	94.97	91.21
TransUNet [14]	67.12	77.53	74.01	88.70	84.36	94.98	90.99
Swin-Unet [10]	66.37	77.12	73.24	89.57	86.33	95.17	90.76
BRAU Net++ [44]	67.69	78.58	75.05	90.39	86.47	95.03	91.14
nnUNet [5]	68.81	78.97	75.49	90.41	86.42	**95.53**	91.19
MedSAM [32]	54.34	66.36	60.57	81.75	80.55	94.30	90.44
MedSAM-Adapter [45]	61.03	73.27	69.02	83.92	83.73	94.43	90.93
Swin-UMamba [46]	68.22	78.70	75.21	90.22	86.37	95.13	91.07
D-LKA Net [41]	69.53	79.54	76.31	90.14	86.21	95.18	91.20
DSCNet [22]	70.34	80.13	77.39	89.18	85.29	93.92	89.94
DSC-AMP (ours)	**73.03**	**82.36**	**79.90**	**90.50**	**86.49**	95.20	**91.21**

TABLE II: Comparing with the state-of-the-art methods in medical anatomy segmentation on the SMOS segmentation dataset

Methods	mDice (%) ↑	HD (mm) ↓	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
U-Net [4] (baseline)	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
Attention U-Net [7]	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
BRAU-Net [44]	70.27	32.91	78.51	61.69	72.94	67.90	93.14	40.88	84.42	62.68
TransUNet [14]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
Swin-Unet [10]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
TransDeepLab [12]	80.16	21.25	86.04	69.16	84.08	79.88	93.53	61.19	89.00	78.40
HiFormer [47]	80.39	**14.70**	86.21	65.69	85.23	79.77	94.61	59.52	90.99	81.08
PVT-CASCADE [48]	81.06	20.23	83.01	**70.59**	82.23	80.37	94.08	**64.43**	90.10	**83.69**
DSC-AMP (ours)	**81.77**	21.12	**89.61**	65.71	**86.44**	**81.16**	**95.14**	63.14	**91.83**	81.16

As for the training details, we apply classical Binary Cross-Entropy (BCE) loss [4] to optimize segmentation networks in this paper. A range of data augmentation strategies are employed considering the typical scarcity of medical imaging data, including image rotations (90°, 180°, and 270°), image flipping, Gaussian blurring, and Gaussian noise adding. All the models are implemented with PyTorch 1.13 and trained on 4 NVIDIA RTX 3090 graphics cards with 24GB VRAM.

For the CXRS dataset, we resize every sample to the same size of 448×448 and train all models via Adam optimizer [49] with a batch size of 4. A Cosine Annealing learning rate schedule with an initial value of 1e-4 and a momentum of 0.9 is employed during training. All the methodologies on this dataset are trained for 30 epochs for fair comparison.

For the SMOS dataset, we resize all images to the same resolution of 224×224. All models are in total trained for 400 epochs via the stochastic gradient descent (SGD) optimizer, with a batch size of 24, a learning rate of 0.05, a momentum of 0.9, and a weight decay of 1e-4.

D. Experimental Results

To validate the effectiveness of the proposed Dynamic Strip Convolution and Adaptive Morphology Perception plugin (DSC-AMP), we conduct comparison experiments on two large-scale datasets for medical anatomy segmentation, including the CXRS and SMOS datasets. In this section, we will the proposed method with other state-of-the-art methods.

1) Comparison Results of Medical Anatomy Segmentation on the CXRS Dataset: We first plugin the proposed DSC-AMP into a plain U-Net [4] reforming it to be more morphology-aware. Then, we compare it with four categories of state-of-the-art segmentation methods, including the four CNN-based methods [4], [5], [22], [43], four Transformer-based [10], [14], [41], [44], one Mamba-based methods [46], and two large foundation models based methods [32], [45]. For a fair comparison, all the methods were evaluated under the same experimental settings as mentioned in the Implementation Details. We report the mean Intersection over Union (mIoU) and mean Dice Similarity Coefficient (mDice) for all 31 medical anatomies in Table I. Besides, the 31 medical anatomies are then grouped into five categories (including ribs, clavicles, scapulae, lungs, and mediastinum), and the mDice for each category is also reported in Table I for fine-grained comparison. Note that the experimental results of other comparing segmentation methods on the CXRS dataset are reproduced from their official code. The results show that our DSC-AMP achieves state-of-the-art performance on both the mIoU and mDice metrics among all the compared methods. Specifically, our DSC-AMP significantly outperforms the CNN-based methods (*i.e.* [4], [5], [22], [43]). Facilitated by the morphology perception abilities from the proposed DSC-AMP, our methods boost the performance of its baseline (U-Net) with remarkable margins of 6.18% and 5.01% on mIoU and mDice, respectively. In addition, Although only in the form of a simple CNN baseline, our DSC-AMP even outperforms modern Transformer-based approaches such as TransUNet [14] and Swin-Unet [10]. Notably, the famous TransUNet is the first model that applies Transformer in the field of medical image segmentation, our DSC-AMP surpasses it by 5.91% and 4.83% over the metric of mIoU and mDice, respectively. It is worth noting that the large medical foundation model MedSAM [32] and its Adapter [45] also perform worse on the CXRS dataset heterogeneous medical anatomy segmentation. It indicates the limitations of structure-agnostic data-driven large models are not all we need and also underscores the necessity of task-specific methods in the medical field.

According to the class-wise mDice results in Table I, our method achieves the best across various anatomical structures, particularly for ribs. It surpasses its baseline (*i.e*. U-Net) and the SOTA method (*i.e*. DSCNet [22]) by respectively a margin of 6.33% and 2.51% on ribs. The significant performance gain in ribs demonstrates that our Dynamic Strip Convolution has effectively captured the discriminative representation of the varying strip structures such as ribs. Further benefiting from the outstanding perception capability for diverse morphology, our DSC-AMP eventually set state-of-the-art segmentation performances for all anatomical classes in the CXRS dataset.

To conduct intuitive comparisons, we further visualize the medical anatomy segmentation results from different approaches (see Fig. 6) including the classical CNN-based U-Net [4], its Transformer-based variant TransUnet [14], the large kernel convolution method DLK-Net [41], the latest approach DSCNet [22], the large foundation model method MedSAM-Adapter [45], the mamba-based method Swin-UMamba [46] and the proposed DSC-AMP. For a clear result comparison of strip structures, only the 24 ribs are visualized in Fig. 6 (a). The marked results of the first row show that only our DSC-AMP could continuously segment the marked orientation-varying rib, indicating the effectiveness of our dynamic strip convolution. Further benefiting from the adaptive morphology perception strategy, our DSC-AMP distributes more attention to the marked ribs in the second row, thus successfully avoiding serious class errors in other methods. In Fig. 6 (b), we visualize the segmentation results of all 31 classes of heterogeneous anatomies in the CXRS dataset. The results marked in yellow indicate that our DSC-AMP is capable of adaptively perceiving heterogeneous morphology and achieves the best segmentation for different medical anatomies (such as the left scapula and the third rib of the left chest in the first row in Fig. 6 (b)).

2) Comparison Results of Medical Anatomy Segmentation on the SMOS Dataset: To further validate the generalization of the proposed approach, we conduct experiments on another dataset SMOS [42] which is for multi-organ segmentation tasks. Following the experimental settings in BRAUNet++ [44], *i.e*. the mean Dice-Similarity Coefficient (mDice) and Hausdorff Distance (HD) are utilized as the overall evaluation metrics and also the mDice for each individual organ are reported at the same time for fine-grained evaluation. The experimental results are presented in Table II. The results show that our DSC-AMP achieves a mean Dice-Similarity Coefficient (mDice) of 81.77% and a Hausdorff Distance (HD) of 21.12 mm. Specifically, our DSC-AMP reforms a relatively weak baseline (*i.e*. U-Net) to establish a new state-of-the-art result w.r.t the mDice metric and a comparable result with existing methods w.r.t the HD metrics. Although the HD result from the Hiformer [47] is somewhat better than our DSC-AMP, its dedicated Double-Level Fusion Module captures boundary features well that deliberately improve the HD metric but do not improve the mIoU metric significantly. Besides, more fine-grained class-wise comparisons are listed in Table II. Our DSC-AMP consistently outperforms other methods in most organs, and there is a significant performance margin in the aorta. This performance gain is mainly owing to our Dynamic Strip Convolution operator which is good at capturing the discriminative representation from orientation-varying structures. Finally, the proposed Morphology-aware Perception strategy adaptively captures the most crucial morphological information via diverse convolution kernels, yielding superior segmentation performance than existing methods. We also visualize some segmentation results on the SMOS dataset in Fig. 7. We observe that our dynamic strip convolution (DSC) caters better to the orientation-varying structures (such as the pink structure in Fig. 7). Besides, our DSC-AMP also achieves satisfied segmentation on other non-strip organs (such as the cyan structure in Fig. 7). It is because our adaptive morphology perception (AMP) strategy adaptively chooses the more suitable deformable convolution for them, which is good at capturing these rough isotropic shapes.

TABLE III: Ablation studies of the proposed DSC-AMP on the CXRS and DRIVE datasets.

Primary Components					mDice (%) ↑	
Kernel Types			Fusion Strategies		Dataset	
Standard	Deformable	Dynamic Strip (ours)	Concat	AMP (ours)	CXRS	DRIVE
✓					77.35	80.73
✓	✓		✓		78.56	80.83
✓		✓	✓		79.10	81.65
	✓	✓	✓		80.76	81.70
✓	✓	✓	✓		81.59	81.73
✓	✓	✓		✓	**82.36**	**81.80**

E. Ablation Studies

In order to analyze the effectiveness of every primary component in the proposed approach, extensive ablation experiments have been conducted on the CXRS dataset and the DRIVE dataset. The results on the mean Dice Similarity Coefficient (mDice) metric are reported in Table III. The first row contains the performance of the plain U-Net baseline [4] which only utilizes standard convolution. The 2^{nd} to 5^{th} rows are the experimental results from combination variants consisting of the standard convolution, deformable convolution [6], and the proposed dynamic strip convolution, which are fused by simple multi-head concatenation strategy. These results show that a diverse kernel combination benefits heterogeneous medical anatomy segmentation. At the same time, our dynamic strip convolution performs much better than traditional deformable convolution since the customized capability of modeling abundant varying strip anatomies in the CXRS datasets. Besides, our Adaptive Morphology Perception (AMP) strategy further boosts the segmentation performance from 81.59% to 82.36% on the CXRS dataset(the last and second-to-last rows in Table III), owing that our AMP adaptively perceives heterogeneous medical anatomies via the Local-macro Attention (LmA) and effectively incorporates comprehensive insight from various morphology-aware convolutional kernels. In addition, we visualize the activation maps at a shallow layer from different types of convolution kernels in the AMP module to qualitatively examine their contributions in Fig. 8. The results clearly show that our dynamic strip convolution is more sensitive to orientation-varying strip structures (such as ribs), validating the ability of our DSC to model dynamic strip structures.

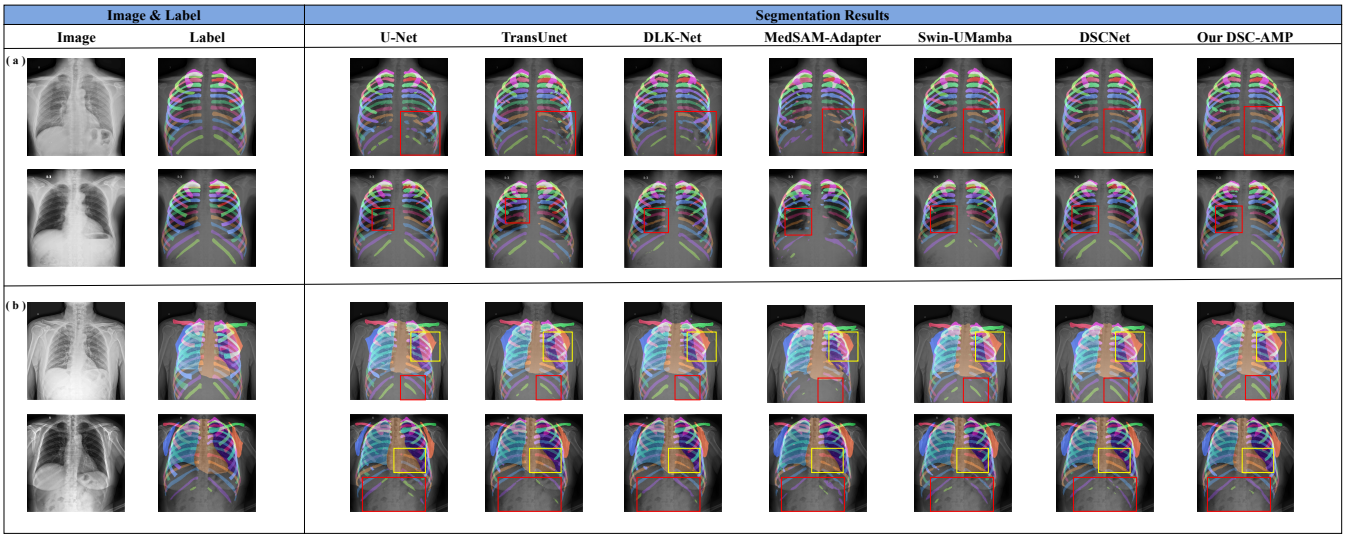

Fig. 6: Visualization comparison of the medical anatomy segmentation results from different approaches on the CXRS dataset. (a) Only the 24 ribs are visualized for a clear comparison, (b) All 31 classes of heterogeneous anatomies are visualized.

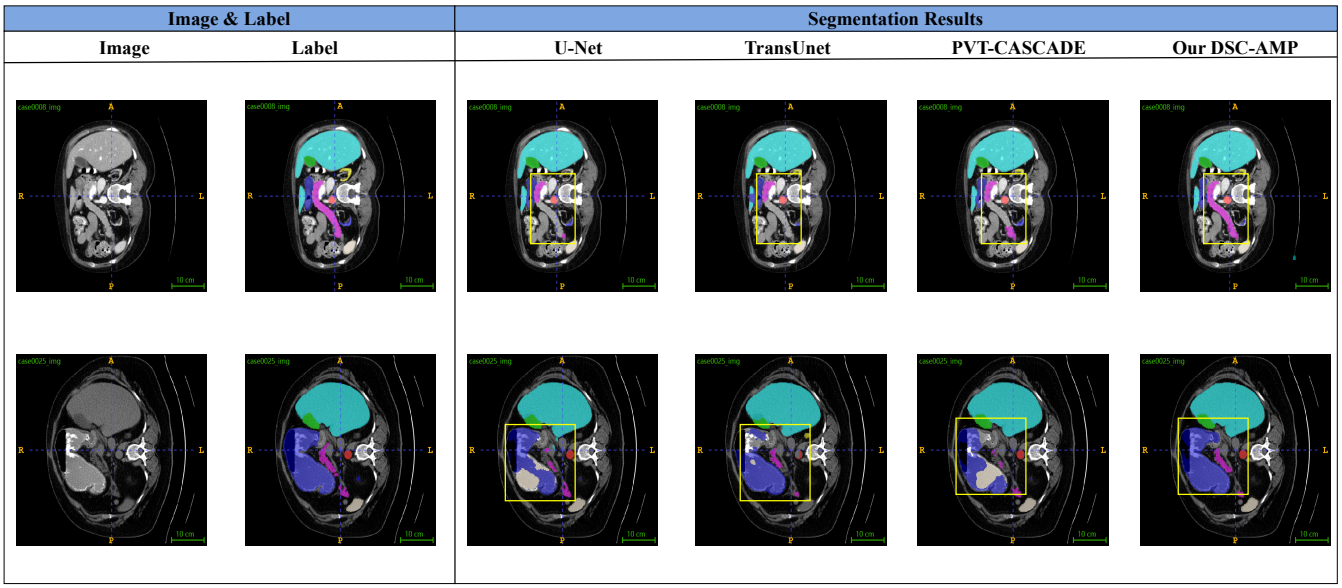

Fig. 7: Visualization comparison of segmentation results from different approaches on the SMOS dataset.

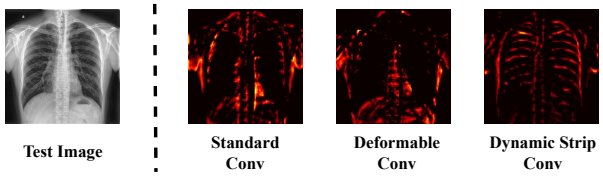

Fig. 8: Visualization of activation maps at a shallow layer for different types of convolution kernels

TABLE IV: Effectiveness evaluation of the proposed Dynamic Strip Convolution (DSC) on the typical CXRS-Rib and DRIVE datasets which contain abundant varying strip structures

Datasets	Convolution Types	mDice (%) ↑	Δ_{mDice}
CXRS-Rib	Standard Convolution	75.45	-
	Deformable Convolution	76.97	+1.52
	Dynamic Snake Convolution	79.13	+3.68
	Dynamic Strip Convolution	**80.97**	**+5.52**
DRIVE	Standard Convolution	80.73	-
	Deformable Convolution	80.83	+0.10
	Dynamic Snake Convolution	81.48	+0.75
	Dynamic Strip Convolution	**81.65**	**+0.92**

F. Effectiveness of the Dynamic Strip Convolution

The proposed Dynamic Strip Convolution plays a pivotal role in our DSC-AMP approach, which is adept at extracting crucial morphological representation by learning the essential geometric information of orientation-varying strip-like structures in medical images. To further individually validate the efficacy of the proposed strip convolution, we conducted segmentation experiments on two datasets that

contain abundant varying stripped structures including the CXRS-Rib and DRIVE [50] datasets. The CXRS-Rib dataset is a selected subset of the CXRS dataset which limits the segmentation targets to only the 24 stripped ribs, and the DRIVE dataset is a public retinal vessel segmentation dataset. To examine the effectiveness and superiority of our dynamic strip convolution operator, we replace every convolution block in the plain U-Net [4] with operators in the form of standard convolution, deformable convolution, dynamic snake convolution, and our strip convolution, respectively. The detailed experimental results are shown in Table IV. We observe the proposed strip convolution significantly enhances the segmentation performance in both datasets which mainly contain elongated structures with gradually varying stretch orientation and length-width ratio. Specifically, replacing standard convolution in plain U-Net with our strip convolution significantly boosts the segmentation performance by large margins of 5.52% and 0.92% on the CXRS-Rib and DRIVE datasets, respectively. The DSConv [22], which is specially designed for slender tubular structures, also makes some improvements over baseline standard convolution. It could be treated as a special case of our dynamic strip convolution along a specific axis. Although it additionally utilized a deliberately designed multi-view fusion strategy [22], our DSC still outperforms it by 1.84% and 0.17% on CXRS-Rib and DRIVE datasets, respectively. Furthermore, we conducted an ablation study that analyzed the impact of applying DSC at different network depths in Table V. The results clearly show that DSC performs most effectively in shallow layers since shallow layers contains more fine-grained strip anatomical structures, such as 24 ribs and 2 clavicles. Its performance gain gradually decreases in deeper layers but still outperforms the standard convolution.

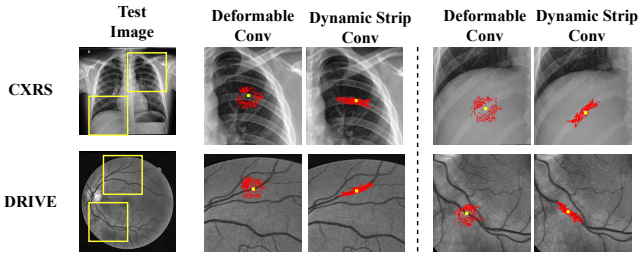

Fig. 9: Visualization comparison of different adaptive kernels.

TABLE V: Effectiveness examination of the DSC along different depths. The DSC operators in a reformed U-Net (Fig. 2a) are replaced by the standard convolution block by block

Block Number					mIOU(%)↑	mDice(%)↑
1	2	3	4	5		
✓	✓	✓	✓	✓	73.03	82.36
×	✓	✓	✓	✓	70.25 (**-2.78**)	79.62 (**-2.74**)
✓	×	✓	✓	✓	71.12 (-1.91)	80.45 (-1.91)
✓	✓	×	✓	✓	72.01 (-1.02)	81.30 (-1.06)
✓	✓	✓	×	✓	72.45 (-0.58)	81.85 (-0.51)
✓	✓	✓	✓	×	72.70 (-0.33)	82.10 (-0.26)

G. Visualization Comparison of Adaptive Kernels

Following [22], we superimpose a total of 729 points (red) of 3 layers onto original test images to visualize the convolutional range and shape associated with a given point (yellow). The results in Fig. 9 indicate that our DSC is more sensitive to strip structures that cater well to strip structures with gradually varying stretch orientation and length-width ratio, such as the marked rib and vessel in Fig. 9.

H. Influence of Hyper-parameters

In this section, we will empirically analyze the influence of some vital Hyper-parameters in the proposed Dynamic Strip Convolution and Adaptive Morphology Perception Plugin (DSC-AMP) and provide some experience in selecting these hyper-parameters.

1) Influence of the Large Kernel Size in the Local-macro Glance (LmG): The Local-macro Glance (LmG) module in Fig. 3 provides the primary relative macro perception of local morphology, thus its large kernel size k_l is a vital parameter to control the perception range for dynamic orientation learning in our DSC. To examine the influence of the large kernel size k_l towards the capability of strip morphology perception in our DSC, we conduct experiments with k_l varies in $\{3, 4, 7\}$ on the CXRS dataset and presented segmentation performance on the metrics of mIoU and mDice in Table VI. We observe the best segmentation performance achieved at $k_l = 5$, where the mIoU and mDice are 73.03% and 82.36%, respectively. The experimental results reveal that appropriately expanding the relative macro perception is beneficial for capturing morphological information in each local area. However, an excessively broad kernel is contrarily detrimental to the local morphological perception, as it might be disturbed by incorrect morphological information from distant areas or different anatomies.

TABLE VI: The influence of the large kernel size in Local-macro Glance (LmG) module

Large Kernel Size ($k_l \times k_l$)	mIoU (%) ↑	mDice (%) ↑
3×3	71.00	80.97
5×5	**73.03**	**82.36**
7×7	71.22	81.28

2) Influence of Magnification Factor λ: The magnification factor λ serves as an important hyper-parameter for expanding the observation range of the Local-micro Attention (LmA) module in Fig. 4 and Eq. 5. Specifically, the observation range is 3×3 when $\lambda = 1$, it will increase to 5×5 when $\lambda = 2$, and so on. To examine the influence of the magnification factor on the proposed Adaptive Morphology Perception strategy, we carry out experiments with a series of $\lambda \in \{1, 2, 3, 4\}$ on the CXRS datasets, the results are shown in Table VII. We can observe that the proposed DSC-AMP achieves its best performance when $\lambda = 3$, indicating a medium-level expansion is beneficial. No expansion (*i.e.* $\lambda = 1$) deteriorates the segmentation outcome where the mDice and mIoU drop from 81.59% to 81.07% and from 71.46 to 71.14%, respectively. Similarly, excessive expansion (such as $\lambda = 4$) also degrades the segmentation performance of medical anatomies.

TABLE VII: Influence of the Magnification Factor

Magnification Factor (λ)	mIoU (%) ↑	mDice (%) ↑
1	71.14	81.07
2	72.39	82.18
3	**73.03**	**82.36**
4	72.80	82.05

TABLE VIII: Comparison of computational complexity and the number of model parameters on the DRIVE Dataset

Methods	#Params	#FLOPs	mDice (%)
U-Net [4]	31.05M	54.86G	80.73
TransUNet [14]	93.23M	32.51G	80.56
CS2Net [51]	8.40M	14.00G	77.53
DCU-net [39]	92.05M	283.20G	80.83
DSCNet [22]	115.76M	332.14G	81.48
DSC-AMP (ours)	104.13M	310.01G	**81.65**

I. Comparison of Model and Computational Complexity

In this section, we reported the number of model parameters (Params) and forward floating-point operations (FLOPs) on the DRIVE dataset for complexity comparison. The size of the input image is fixed as 256×256 for fair comparison when calculating FLOPs. The results in Table VIII indicate that the approaches equipped with deformable kernels (the last three rows) generally require more model parameters and computational resources but achieve significantly superior segmentation performance. Moreover, compared to recent state-of-the-art DSCNet [22] designed for tubular structures, our DSC-AMP achieves better performance with fewer parameters. This superiority is due to our Dynamic Strip Convolution directly learning the angular offsets of strip structures, thus avoiding learning redundant parameters and accommodating a broader range of angular variations.

TABLE IX: Experimental results from various network architectures on the CXRS dataset.

Model	Without DSC-AMP		With DSC-AMP	
	mDice (%)	mIoU (%)	mDice (%)	mIoU (%)
DeepLabV3 [52]	64.04	52.37	**67.16**	**55.23**
UNet [4]	77.35	66.85	**82.36**	**73.03**
TransUNet [14]	77.53	67.12	**78.26**	**67.89**

J. Generalization Analysis of Network Architectures

In this section, we evaluate the generalization ability of our convolutional plugin regarding various network architectures. Besides classical U-Net architecture, we further integrate it into a widely-used CNN segmentation network DeepLabV3 [52] that does not under typical U-Net architecture and a CNN-Transformer hybrid architecture TransUNet [14]. The results on the CXRS dataset in Table IX indicate that the proposed convolutional plugin achieves consistent improvements across various network architectures. Notably, the performance enhancement is more significant in DeepLabV3 and U-Net since they contain pure convolutional layers that could benefit more from the proposed convolutional plugin.

K. Limitations and Failure cases

Although the proposed DSC-AMP significantly improves the segmentation performance for heterogeneous medical anatomy segmentation, it introduces additional learnable parameters, computational resources, and storage costs compared to standard convolution. Improvements to alleviate these issues wait for future research, such as replacing standard channel operation with depth-wise separable operation and replacing deformable convolution in our AMP with the more efficient DCNv4 [53]. Additionally, our method is designed for 2D medical image segmentation and additional modification is required before applying to 3D medical images.

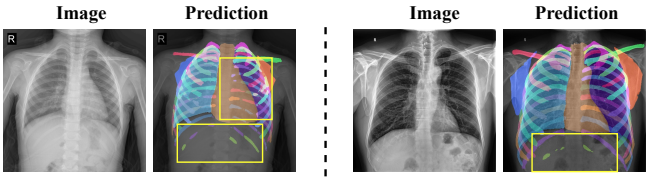

Fig. 10: Failure cases analysis on the CXRS dataset.

In addition, we include some typical failure cases from our DSC-AMP on the test samples in the CXRS dataset, as shown in Fig. 10. Our dynamic strip convolution caters well to most of the strip-like ribs but performs unsatisfying in a few bottom ribs and those overlaps with lungs. This is mainly because these test examples are from patients with certain lesions whose X-ray images have relatively low local contrast. Thus, it is more difficult to distinguish corresponding anatomical organs (such as ribs) from other surrounding organs and additional image enhancement procedures may alleviate this issue.

V. Conclusion

This paper proposes a simple yet effective morphology-aware plugin for medical anatomy segmentation, which can adaptively perceive and model heterogeneous morphological structures. It adaptively adjusts the kernel shape and sampling range of convolutional segmentation layers according to the morphology characteristics of heterogeneous segmentation targets, thereby achieving precise and adaptive segmentation. Specifically, we explore a novel Dynamic Strip Convolution operator customized for slender and orientation-varying strip anatomies such as ribs and clavicles. Furthermore, the proposed Adaptive Morphology Perception strategy matches different local morphological structures with appropriate kernels, and adaptive integrates these diverse representations through local-macro attention. Eventually, our DSC-AMP achieves state-of-the-art performance on two large-scale datasets for anatomy segmentation, including the CXRS and SMOS datasets.

References

[1] G. Du, X. Cao, J. Liang, X. Chen, and Y. Zhan, "Medical image segmentation based on u-net: A review," *Journal of Imaging Science and Technology*, 2020.

[2] R. Azad *et al.*, "Medical image segmentation review: The success of u-net," *arXiv preprint arXiv:2211.14830*, 2022.

[3] D. Zhang *et al.*, "Deep learning for medical image segmentation: tricks, challenges and future directions," *arXiv preprint arXiv:2209.10307*, 2022.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[5] 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.

[6] J. Dai *et al.*, "Deformable convolutional networks," in *ICCV*, 2017.

[7] O. Oktay *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[8] J. Zhang, J. Zhang, G. Hu, Y. Chen, and S. Yu, "Scalenet: a convolutional network to extract multi-scale and fine-grained visual features," *IEEE Access*, vol. 7, pp. 147560–147570, 2019.

[9] B. Cui, G. Hu, and S. Yu, "Rt-net: replay-and-transfer network for class incremental object detection," *Applied Intelligence*, vol. 53, no. 8, pp. 8864–8878, 2023.

[10] H. Cao *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *ECCV*, 2022.

[11] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *MICCAI*, 2021.

[12] R. Azad *et al.*, "Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation," in *MICCAI*, 2022.

[13] R. Azad, R. Arimond, E. K. Aghdam, A. Kazerouni, and D. Merhof, "Dae-former: Dual attention-guided efficient transformer for medical image segmentation," in *MICCAI*, 2023.

[14] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[15] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[16] J. Ruan and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," *arXiv preprint arXiv:2402.02491*, 2024.

[17] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[18] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3d semantic segmentation for medical images," in *MICCAI, 2020*, pp. 552–561, Springer, 2020.

[19] H. Zhang, H. Zhu, J. Jing, P. Li, and Q. Pan, "Curve-like structure detection using multi-sale and boundary assisted segmentation network," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[20] F. Crosilla, A. Beinat, A. Fusiello, E. Maset, and D. Visintini, "Orthogonal procrustes analysis," *Advanced Procrustes Analysis Models in Photogrammetric Computer Vision*, 2019.

[21] X. Zhang *et al.*, "Akconv: Convolutional kernel with arbitrary sampled shapes and arbitrary number of parameters," *arXiv preprint arXiv:2311.11587*, 2023.

[22] Y. Qi, Y. He, X. Qi, Y. Zhang, and G. Yang, "Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation," in *ICCV*, 2023.

[23] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Processing*, vol. 16, no. 5, pp. 1243–1267, 2022.

[24] E. Çallı, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep learning for chest x-ray analysis: A survey," *Medical Image Analysis*, vol. 72, p. 102125, 2021.

[25] S. Lu, J. Liu, X. Wang, and Y. Zhou, "Collaborative multi-metadata fusion to improve the classification of lumbar disc herniation," *IEEE Transactions on Medical Imaging*, 2023.

[26] C. You *et al.*, "Mine your own anatomy: Revisiting medical image segmentation with extremely limited labels," *arXiv preprint arXiv:2209.13476*, 2022.

[27] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3DV*, 2016.

[28] S. Roy *et al.*, "Mednext: transformer-driven scaling of convnets for medical image segmentation," in *MICCAI*, 2023.

[29] V. Lialin, V. Deshpande, and A. Rumshisky, "Scaling down to scale up: A guide to parameter-efficient fine-tuning," *arXiv preprint arXiv:2303.15647*, 2023.

[30] G. Hu, B. He, and H. Zhang, "Compositional prompting video-language models to understand procedure in instructional videos," *Machine Intelligence Research*, vol. 20, no. 2, pp. 249–262, 2023.

[31] A. Kirillov *et al.*, "Segment anything," in *ICCV*, 2023.

[32] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[33] Y. Li *et al.*, "nnsam: Plug-and-play segment anything model improves nnunet performance," *arXiv preprint arXiv:2309.16967*, 2023.

[34] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *CVPR*, 2017.

[35] J. Dai *et al.*, "Deformable convolutional networks," in *ICCV*, 2017.

[36] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "Dunet: A deformable network for retinal vessel segmentation," *Knowledge-Based Systems*, vol. 178, pp. 149–162, 2019.

[37] S. Dong *et al.*, "Deu-net 2.0: Enhanced deformable u-net for 3d cardiac cine mri segmentation," *Medical Image Analysis*, vol. 78, p. 102389, 2022.

[38] C. Zhao, W. Zhu, and S. Feng, "Superpixel guided deformable convolution network for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 31, pp. 3838–3851, 2022.

[39] X. Yang, Z. Li, Y. Guo, and D. Zhou, "Dcu-net: A deformable convolutional neural network based on cascade u-net for retinal vessel segmentation," *Multimedia Tools and Applications*, vol. 81, no. 11, pp. 15593–15607, 2022.

[40] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4003–4012, 2020.

[41] R. Azad *et al.*, "Beyond self-attention: Deformable large kernel attention for medical image segmentation," in *WACV*, 2024.

[42] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, vol. 5, p. 12, 2015.

[43] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang, "A nested u-net architecture for medical image segmentation," in *MICCAI*, 2018.

[44] 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," *arXiv preprint arXiv:2401.00722*, 2024.

[45] J. Wu, W. Ji, Y. Liu, H. Fu, M. Xu, Y. Xu, and Y. Jin, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.12620*, 2023.

[46] J. Liu, H. Yang, H.-Y. Zhou, Y. Xi, L. Yu, Y. Yu, Y. Liang, G. Shi, S. Zhang, H. Zheng, *et al.*, "Swin-umamba: Mamba-based unet with imagenet-based pretraining," *arXiv preprint arXiv:2402.03302*, 2024.

[47] M. Heidari *et al.*, "Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation," in *WACV*, 2023.

[48] M. M. Rahman and R. Marculescu, "Medical image segmentation via cascaded attention decoding," in *WACV*, 2023.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[50] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE transactions on medical imaging*, vol. 23, no. 4, pp. 501–509, 2004.

[51] L. Mou *et al.*, "Cs2-net: Deep learning segmentation of curvilinear structures in medical imaging," *Medical Image Analysis*, vol. 67, p. 101874, 2021.

[52] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[53] Y. Xiong, Z. Li, Y. Chen, F. Wang, X. Zhu, J. Luo, W. Wang, T. Lu, H. Li, Y. Qiao, *et al.*, "Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5652–5661, 2024.