3D Shuffle-Mixer: An Efficient Context-Aware Vision Learner of Transformer-MLP Paradigm for Dense Prediction in Medical Volume

Jianye Pang*, Cheng Jiang*, Yihao Chen, Jianbo Chang, Ming Feng, Renzhi Wang
and Jianhua Yao†, *Senior Member, IEEE*

Abstract— **Dense prediction in medical volume provides enriched guidance for clinical analysis. CNN backbones have met bottleneck due to lack of long-range dependencies and global context modeling power. Recent works proposed to combine vision transformer with CNN, due to its strong global capture ability and learning capability. However, most works are limited to simply applying pure transformer with several fatal flaws (i.e., lack of inductive bias, heavy computation and little consideration for 3D data). Therefore, designing an elegant and efficient vision transformer learner for dense prediction in medical volume is promising and challenging. In this paper, we propose a novel 3D Shuffle-Mixer network of a new Local Vision Transformer-MLP paradigm for medical dense prediction. In our network, a local vision transformer block is utilized to shuffle and learn spatial context from full-view slices of rearranged volume, a residual axial-MLP is designed to mix and capture remaining volume context in a slice-aware manner, and a MLP view aggregator is employed to project the learned full-view rich context to the volume feature in a view-aware manner. Moreover, an Adaptive Scaled Enhanced Shortcut is proposed for local vision transformer to enhance feature along spatial and channel dimensions adaptively, and a CrossMerge is proposed to skip-connect the multi-scale feature appropriately in the pyramid architecture. Extensive experiments demonstrate the proposed model outperforms other state-of-the-art medical dense prediction methods.**

Index Terms— **Dense prediction in medical volume, context-aware, window-based multi-head self-attention, Local Vision Transformer-MLP, Adaptive Scaled Shortcut.**

I. INTRODUCTION

DENSE prediction plays an important role in medical image analysis since it connects with a variety of practical clinical applications [1]–[4] throughout the entire medical imaging community. Compared with classification or recognition, dense prediction (e.g. segmentation, registration, synthesis, reconstruction, etc) aims to extract contextual information and perform pixel-level classification or regression

on the whole image relying on high-resolution and multi-scale features, which calls for higher requirements of model design. Besides, 3D medical volumetric data from like Computed tomography (CT) and Magnetic resonance imaging (MRI) are harder to deal with and far beyond 2D images. Accurate dense prediction in medical volume provides enriched guidance as a fundamental enabler for automated clinical diagnosis and analysis. Thus, many works [5]–[7] have focused on various dense prediction tasks in medical volume. Although the overall frameworks of most tasks are different, the backbones used in them are 3D U-Net [8] and its variants [9] which commonly adopt the pyramid structure based on CNN.

Since the volumetric data has one more dimension than 2D image, it contains richer semantic context in the overall volume. Besides, MRI or CT sometimes comes in multiple modalities with complex composition form. Fully learning context is necessary for dense prediction and raises great challenges to pure CNN models where CNN is not context-aware and lacks the ability to model long distance dependencies on global context. Despite 3D CNN models have achieved great success in medical volume, due to the fact that the amount of 3D volume is much less than 2D scan, 3D CNN models sometimes perform poorly on small datasets, even worse than 2D CNN. Moreover, the computational complexity of 3D CNN is much higher, which leads to limited freedom in structural design. It is noteworthy that the number of applications [10]–[12] on vision transformer or vision MLP in general vision tasks have been exploded very recently. To get rid of the above dilemma of pure CNN for dense prediction, how to use these remarkable architectures to break the bottleneck of CNN backbones has became a hot topic in medical image analysis.

Taking advantage of strong learning capability of vision transformer, a large portion of medical dense prediction [13]–[16] works on adding pure transformer to pyramid CNN recently. However, due to the fatal flaws of pure transformer from large dataset requirement, lack of inductive bias, heavy computation and over-smoothing degradation, pure vision transformer is underwhelming for dense prediction. CeiT [17] proposed using depth-wise convolution into transformer block to fuse information from different patches and increase locality. DeepViT [18] and Refiner [19] proposed re-attention linear transformations to add diversity and alleviate degradation. One of the most notable improvement is the local vision

J. Pang is with the Department of Computer Science, Xi'an Jiaotong University, Xi'an, China (e-mail: jianye.pang97@gmail.com).

C. Jiang and †J. Yao are with the Tencent AI Lab, Shenzhen, China (corresponding author: J. Yao, e-mail: jianhuayao@tencent.com).

Y. Chen, J. Chang, M. Feng and R. Wang are with the Department of Neurosurgery, Peking Union Medical College Hospital, Peking Union.

*J. Pang and C. Jiang contributed equally to this work.

transformer [20] which proposed window-based multi-head self-attention (W-MSA) with less computation burden and more inductive bias. It dramatically improved the performance on dense prediction and applicability to small datasets which further taps transformer's potentials. However, currently there are very few studies [21], [22] considering local vision transformer, at best, but are only simple to introduce existing Swin module [23] into the pyramid structure. Furthermore, most works of local vision transformer are only proposed for 2D images. How to elegantly design an efficient vision learner specifically suitable for 3D medical image volume leveraging local vision transformer to solve the dense prediction task is challenging and hasn't been addressed.

To this end, we propose a novel 3D Shuffle-Mixer of a new Local Vision Transformer-MLP paradigm for dense prediction in medical volume which efficiently learns volume context via adequate information shuffle and mixing. Concretely, we first rearrange volume into different view slices and learn the rich spatial context by W-MSA with transpose shuffle to build cross-window spatial connections. Then the remaining context along third dimension is captured by a slice-aware MLP mixing, which is proposed with an residual axial-MLP and absolute position embedding (APE) on each slice feature to enable the ability to distinguish between slice context. Finally volume context from all views is projected to the complete volume feature through a view-aware aggregator with APE for each view to distinguish between view context. For one 3D Shuffle-Mixer block, both W-MSA and MLP share parameters across views. We also propose an Adaptive Scaled Enhanced Shortcut (ASES) for window-based local vision transformer to adaptively rectificate and refine decoupled features on spatial and channel dimensions. Furthermore, a CrossMerge is proposed to skip-connect and merge view contexts from encoder to decoder elegantly. The effectiveness of our method is validated by comparing it with various existing state-of-the-art (SOTA) methods on three medical dense prediction tasks.

The main contributions of this paper are summarized as follows.

- Different from most existing CNN-Pure Transformer hybrid approaches, we present a novel context-aware vision learner of a new Local Vision Transformer-MLP paradigm for dense prediction in medical volume.
- We present an ASES shortcut for window-based local vision transformer to capture more precise and augmented features enhancing along spatial and channel dimensions adaptively, as well as a CrossMerge skip-connection suitable for pyramid transformer model.
- We conduct extensive validations on one private shape prediction dataset and two pubilc popular segmentation datasets. Our model achieves a great performance gain, surpassing the SOTA methods by a large margin.

The rest of this paper is organized as follows. We introduce the related works in Section II. Then, we elaborate the framework of our 3D Shuffle-Mixer with ASES shortcut and CrossMerge skip-connection in Section III. We further present the analysis and discussions of experimental results in Section IV. Finally, we conclude the paper in Section V.

II. RELATED WORKS

In this section, we first discuss some local vision transformer architectures in general vision tasks and then take the 3D segmentation as an example to review some related works of the recent popular CNN-Transformer (CNN-TR) hybrid backbones for 3D medical image analysis since they're commonly used architectures among various medical downstream dense tasks.

Local Vision Transformers in General Vision Tasks: The first work to show vanilla transformer can compete with SOTA CNNs on large datasets in general vision tasks is ViT [10] which has become the standard benchmark for vision transformer. Local vision transformers improve a lot over pure transformers. They reduce complexity to linear by computing self-attention only within each local window. Moreover, various methods of cross-window connections make up for prior knowledge of locality and translation invariance which contributes to the success of local vision transformers in dense prediction tasks.

Swin Transformer [23] (Swin-TR) first proposed shifted windows to bridge window information to encourage inductive bias, and introduced relative position bias into window-based local vision transformer. Twins [24] combined local attention and global attention alternately in each stage to enhance modeling power. Shuffle Transformer [25] (Shuffle-TR) shuffled fixed windows to make information flow across windows. PVTv1 [26] proposed using pyramid structure in transformer and PVTv2 [27] proposed overlapping patch embedding to enlarge patch window. MSG-Transformer [28] shuffled the message token added on each window for message passing across windows. The most recent SOTA in general vision tasks is CSWin Transformer [29] (CSWin-TR) which proposed cross-shaped window self-attention to enlarge receptive field. Besides, there're several works [12], [30], [31] on vision MLP which used concise structures, but trained on larger datasets to achieve the similar performance. However, above models are all only designed for 2D natural images and there are few researches for 3D input, e.g. Video Swin Transformer [32] (Video Swin-TR) which extends to 3D shifted windows.

CNN-Transformer Backbones in 3D Medical Image Analysis: 3D segmentation is one of the most general and representative tasks in medical image analysis. Well-performed backbones for 3D medical image analysis consist of several newly proposed CNN-TR hybrid models.

Segtran [33] introduced squeeze-and-expansion into transformer and a new learnable sinusoidal position encoding for imposing continuity inductive bias. U-Transformer [34] utilized self- and cross-attention into U-Net. MedT [14] proposed using gated axial attention in transformer. TransUNet [13], CoTr [35] and TransBTS [36] replaced CNN encoder in U-Net with hybrid CNN-TR which connected to pure transformer at the end of CNN to make a strong encoder. UNETR [37] directly replaced whole CNN encoder with cascaded pure transformer encoder to capture multi-scale features. Swin-Unet [21] simply brought Swin block from Swin-TR to build a U-shaped encoder-decoder symmetric architecture to improve feature representation. Apart from the segmentation, there're

also similar CNN-TR backbones [38], [39] in 3D medical detection, reconstruction, etc. However, above models simply used pure transformer or directly copied existing module in the pyramid implementation. Neither proposed a specifically designed local vision transformer block for medical images. Besides, only a few models took 3D volume into account, e.g. Segtran, TransBTS and UNETR, while most models are simply based on 2D scan processing and therefore generally perform less well.

III. METHODOLOGY

A. Overview of the Proposed Method

Fig. 1 shows the overall framework of the proposed method which is a pyramid architecture. Our model consists of 3D Shuffle-Mixer blocks (Fig. 2) with ASES (Fig. 3), and Cross-Merge (Fig. 4) to skip-connect and merge multi-scale context between each encoder and decoder block. We adopt convolutional stem [40] as patch embedding for non-overlapping token embedding and training stability. We utilize 3D convolution with kernel size 3 in down-sampling, transposed convolution with kernel size 2 in up-sampling and kernel size 1 in target projection to facilitate information flow between non-overlapping windows.
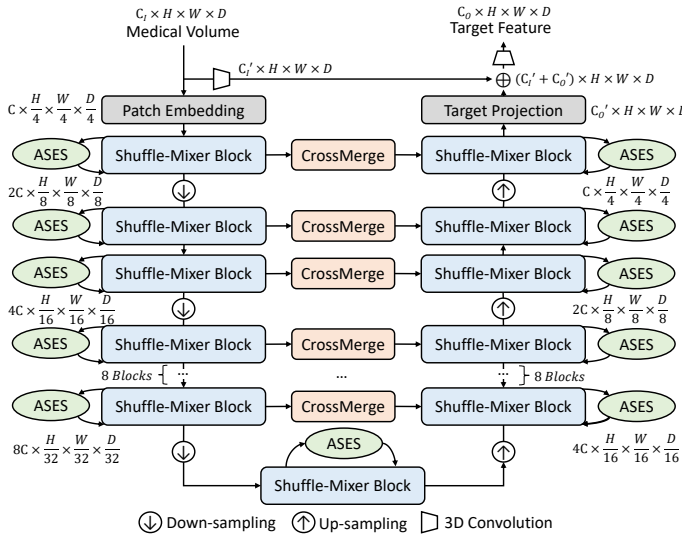

Fig. 1. Overview of the proposed method. The channel of input volume is C_i and output target is C_o.

B. 3D Shuffle-Mixer

1) Full-View Slice Spatial Shuffle Block: We utilize W-MSA to obtain global and local spatial contextual information first. The input feature $\mathbf{Z}$ is partitioned into non-overlapping window feature $\mathbf{Z}^i$. $\mathbf{W}_k^Q$, $\mathbf{W}_k^K$ and $\mathbf{W}_k^V$ are query, key and value matrices for k-th head. In details, computing the k-th head self-attention $\mathbf{H}_k$ and output $\mathbf{W}\text{-}\mathbf{MSA}(\mathbf{Z})$ can be defined as:

$$\mathbf{Z} = \left\{ \mathbf{Z}^1, \mathbf{Z}^2, \dots, \mathbf{Z}^N \right\}, N = HW/M^2$$
$$\mathbf{H}_k^i = \mathbf{Attention}(\mathbf{Z}^i\mathbf{W}_k^Q, \mathbf{Z}^i\mathbf{W}_k^K, \mathbf{Z}^i\mathbf{W}_k^V), \quad i = 1, \dots, N$$
$$\mathbf{H}_k = \left\{ \mathbf{H}_k^1, \mathbf{H}_k^2, \dots, \mathbf{H}_k^N \right\}$$
$$\mathbf{W}\text{-}\mathbf{MSA}(\mathbf{Z}) = \mathbf{Concat}[\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_k]\mathbf{W}^H \tag{1}$$

Where M denotes window size. All heads are concatenated and linearly projected to the output with relative position encoding [23] (RPE) to be window-aware of spatial positions. The attention is defined as:

$$\mathbf{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{SoftMax}(\mathbf{QK}^T/\sqrt{d_k} + \mathbf{RPE})\mathbf{V} \tag{2}$$

However, pure W-MSA lacks window connections, hence we utilize 2D spatial transpose shuffle operation inspired by [12], [25], [28] into W-MSA to fully build long-range cross-window connections. Besides, it can only handle 2D spatial context well, which ignores the context of the third dimension like depth when applied to 3D volume. Therefore, we propose a full-view slice spatial shuffle block to rearrange and extract redundant information residing in full-view spatial slices without complex and time-consuming operation like pure 3D convolutions.

As shown in Fig. 2-(1),(2), we consider slices from each two-dimensional direction as one view of the 3D volume. Thus, a 3D volume (H, W, D) has three kinds of views (H, W), (H, D) and (W, D) which can be directly acquired by rearrange operation. Therefore, we have all slices from 3D volume to learn pure W-MSA and employ transpose shuffle T with rotation restore R on slices in each view to learn shuffle W-MSA respectively. W-MSA learns normalized features through batch norm (BN) first, and a 2D depth-wise convolution (DWConv2D) with 5×5 kernel and 1 stride strengthens spatial neighbor locality on the slice, followed by a multi-layer perceptron (MLP) which maps features to high dimension and back for refining. The computation can be represented as:

$$\mathbf{X}_l^v = \mathbf{W}\text{-}\mathbf{MSA}\left(\mathbf{BN}\left(\mathbf{Z}_l^v\right)\right) + \mathbf{Z}_l^v$$
$$\mathbf{Y}_l^v = \mathbf{DWConv2D}\left(\mathbf{X}_l^v\right) + \mathbf{X}_l^v$$
$$\mathbf{Z}_l^{v'} = \mathbf{MLP}\left(\mathbf{BN}\left(\mathbf{Y}_l^v\right)\right) + \mathbf{Y}_l^v$$
$$\mathbf{X}_l^{v'} = \mathbf{R}(\mathbf{W}\text{-}\mathbf{MSA}(\mathbf{T}(\mathbf{BN}(\mathbf{Z}_l^{v'})))) + \mathbf{Z}_l^{v'} \tag{3}$$
$$\mathbf{Y}_l^{v'} = \mathbf{DWConv2D}(\mathbf{X}_l^{v'}) + \mathbf{X}_l^{v'}$$
$$\hat{\mathbf{Z}}_l^v = \mathbf{MLP}(\mathbf{BN}(\mathbf{Y}_l^{v'})) + \mathbf{Y}_l^{v'}$$
$$v = \{1, 2, 3\}$$

Where $\mathbf{Z}_l^v$ denotes the input volume feature comes from view v in l-th block for W-MSA to learn. Noticed that the parameters of each W-MSA are shared among full-view in each block. On one hand, the computation is drastically reduced. On the other hand, scalability and generalization increase since more inductive bias are brought into vision transformer.

2) Slice-Aware Volume Context Mixing: Full-view information of the 3D volume has been obtained as the output of slice spatial shuffle block. However, in each single view, the output only contains spatial information of 2D slices. To take into account the context among slices, A.K.A along the third dimension, we propose a slice-aware volume context mixing to capture remaining context.

As shown in Fig. 2-(3), the slice-aware volume context mixing contains two branches. In one branch, we add absolute position embedding [41] (APE) which is an automatic learnable parameter on each slice feature to inject positional
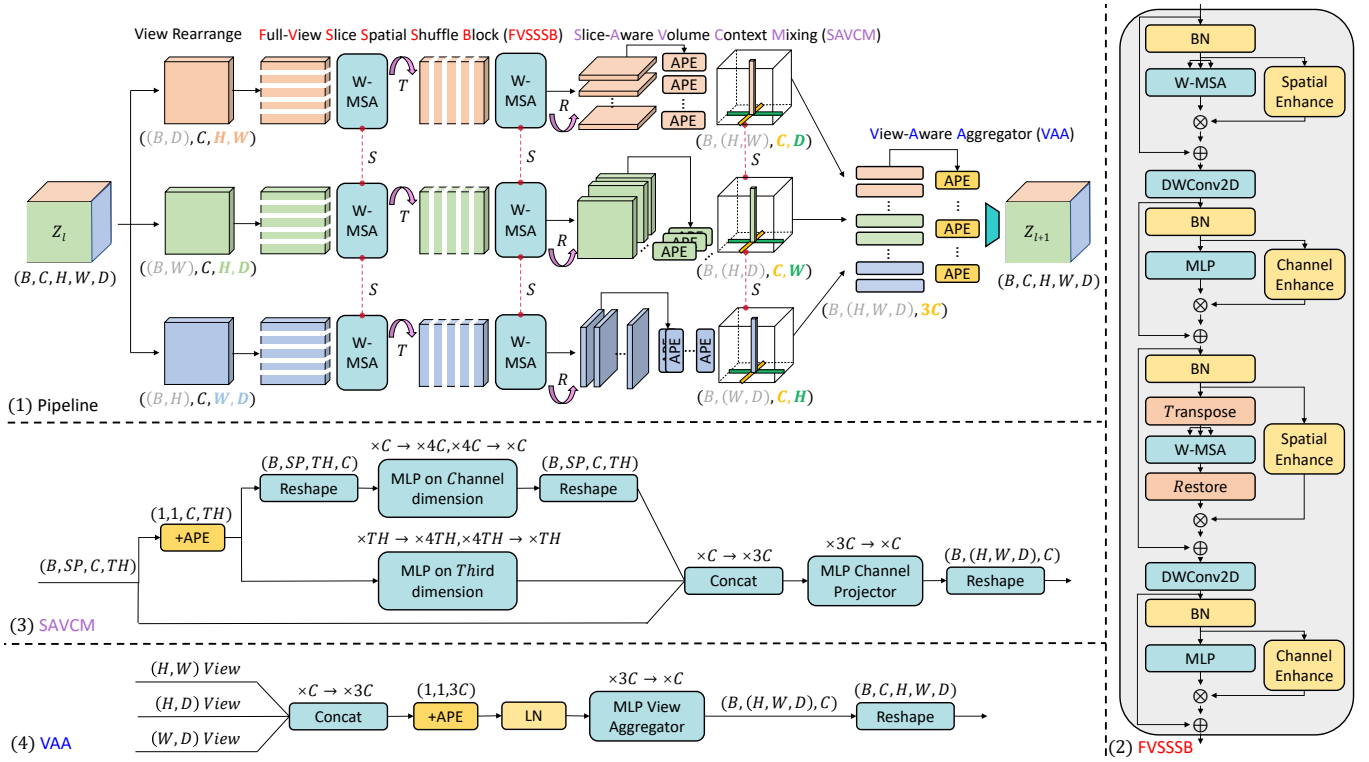
Fig. 2. Architecture of 3D Shuffle-Mixer block. Overall Pipeline of 3D Shuffle-Mixer block is shown in (1), which consists of three sequential parts: Details of Full-View Slice Spatial Shuffle Block with ASES are shown in (2), where ASES is represented as Spatial Enhance and Channel Enhance which will be elaborated in Section III-C; Details of Slice-Aware Volume Context Mixing are shown in (3); Details of View-Aware Aggregator are shown in (4). *In (1), arrow T denotes transpose shuffle operation, arrow R denotes rotation restore operation and red dotted line denotes parameter sharing between connected targets. In (3), SP denotes 2D spatial dimension, e.g. (H, W) and TH denotes third dimension, e.g. D.

information and help distinguish slices. Then the slice-aware context is processed by an axial-MLP:

$$
\begin{aligned}
\mathbf{A}_{l+1}^v &= \hat{\mathbf{Z}}_{l+1}^v + \mathbf{APE}_s \\
\mathbf{Axial\text{-}MLP}(\hat{\mathbf{Z}}_{l+1}^v) &= [\mathbf{MLP}_{st}(\mathbf{A}_{l+1}^v); \mathbf{MLP}_{sc}(\mathbf{A}_{l+1}^v)] \\
\hat{\mathbf{A}}_{l+1}^v &= \mathbf{MLP}_{cp}(\mathbf{Concat}[\mathbf{Axial\text{-}MLP}(\hat{\mathbf{Z}}_{l+1}^v); \hat{\mathbf{Z}}_{l+1}^v])
\end{aligned}
\tag{4}
$$

Where s denotes slice-aware, subscript st and sc denote slice-aware along third dimension and channel, and cp denotes channel projector. The axial-MLP is used to linearly map on the third dimension and channel in parallel. In another branch, we use residual shortcut to maintain original feature as identity mapping. The parameters of axial-MLP are shared between among the three views as well.

In contrast to only simple spatial feature [13], [21] learned based on 2D scan, we capture the 3D context with concise computational operations. In addition,

we introduce position information between slices to enable the model to be slice aware.

3) View-Aware Aggregator: Rich 3D volume context has been captured in each single view through slice-aware volume context mixing. To aggregate the full-view together and obtain the final volume feature of 3D Shuffle-Mixer at l+1-th block, we propose a view-aware aggregator to project features of all views along channel dimension in the end of the 3D Shuffle-Mixer block.

As shown in Fig. 2-(4), we first concatenate full-view input along channel dimension and add APE on them, followed by

a layer norm (LN) to normalize view-aware features. Then a MLP view aggregator restore the feature to original channel dimension. The computation is represented as:

$$
\begin{aligned}
\mathbf{V}_{l+1}^v &= \hat{\mathbf{A}}_{l+1}^v + \mathbf{APE}_{vi} \\
\mathbf{Z}_{l+1} &= \mathbf{MLP}_{va}(\mathbf{LN}(\mathbf{Concat}[\mathbf{V}_{l+1}^1; \mathbf{V}_{l+1}^2; \mathbf{V}_{l+1}^3]))
\end{aligned}
\tag{5}
$$

Where vi denotes view-aware and subscript va denotes view aggregator. With addition of APE, MLP view aggregator is aware of the specific view.

C. Adaptive Scaled Enhanced Shortcut for Window-based Local Vision Transformer

To further increase learning capacity of local vision transformer, we propose an ASES for window-based local vision transformer which can enhance feature representation adaptively using its original input meanwhile only costs a little extra computation. ASES is an improvement based on Formula (3), where the computation is represented as:

$$
\begin{aligned}
\mathbf{Z}_l^{v_n} &= \mathbf{BN}(\mathbf{Z}_l^v); \mathbf{Y}_l^{v_n} = \mathbf{BN}(\mathbf{Y}_l^v) \\
\mathbf{W\text{-}MSA}^{\mathbf{E}}(\mathbf{Z}_l^{v_n}) &= \mathbf{E_s}(\mathbf{Z}_l^{v_n}) \otimes \mathbf{W\text{-}MSA}(\mathbf{Z}_l^{v_n}) + \mathbf{Z}_l^v \\
\mathbf{MLP}^{\mathbf{E}}(\mathbf{Y}_l^{v_n}) &= \mathbf{E_c}(\mathbf{Y}_l^{v_n}) \otimes \mathbf{MLP}(\mathbf{Y}_l^{v_n}) + \mathbf{Y}_l^v
\end{aligned}
\tag{6}
$$

Where $\mathbf{Z}_l$ and $\mathbf{Y}_l$ are the original input features for the W-MSA and MLP in the l-th block respectively. $\otimes$ denotes element-wise multiplication. $\mathbf{E_s}$ and $\mathbf{E_c}$ are spatial enhanced
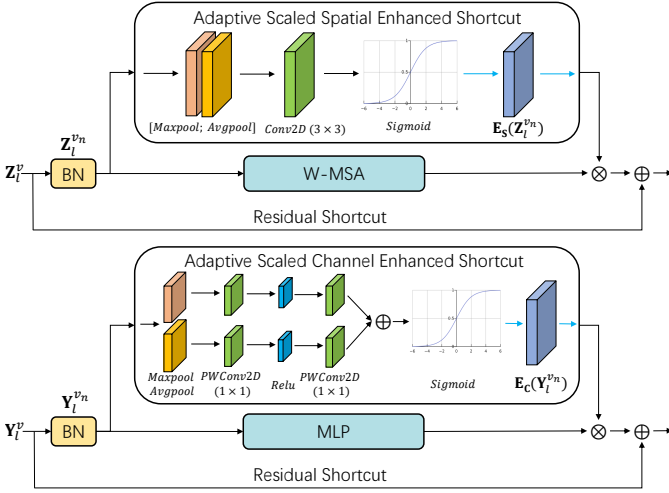
Fig. 3. Architecture of Adaptive Scaled Enhanced Shortcut for Window-based Local Vision Transformer, which consists of spatial enhance and channel enhance. Cyan arrow represents ASES shortcut's output.

Fig. 4. Architecture of CrossMerge Skip-Connection, which is interweaved across from encoder to decoder for merging information individually in each view.

shortcut and channel enhanced shortcut, which are detailed shown in Fig. 3 (also in Fig. 2-(2)) and computed as:

$$\mathbf{E_s}(\mathbf{Z}_l^{v_n}) = \sigma\left(\mathbf{Conv2D}([\mathbf{AvgPool}(\mathbf{Z}_l^{v_n}); \mathbf{MaxPool}(\mathbf{Z}_l^{v_n})])\right)$$
$$= \sigma\left(\mathbf{f^{3\times 3}}\left(\mathbf{Concat}\left[(\mathbf{Z}_l^{v_n})^{\mathrm{avg}}; (\mathbf{Z}_l^{v_n})^{\mathrm{max}}\right]\right)\right) \quad (7)$$

$$\mathbf{E_c}(\mathbf{Y}_l^{v_n}) = \sigma(\mathbf{PWConv2D}(\mathbf{AvgPool}(\mathbf{Y}_l^{v_n})) +$$
$$\mathbf{PWConv2D}(\mathbf{MaxPool}(\mathbf{Y}_l^{v_n}))) \quad (8)$$
$$= \sigma(\mathbf{f^{1\times 1}}((\mathbf{Y}_l^{v_n})^{\mathrm{avg}}) + \mathbf{f^{1\times 1}}((\mathbf{Y}_l^{v_n})^{\mathrm{max}}))$$

Where PWConv2D denotes point-wise convolution with 1×1 kernel. In the window-based local vision transformer equipped with ASES, spatial enhancing on W-MSA learns an adaptive scaled factor $\mathbf{E_s}(\mathbf{Z}_l)$ from input and multiply to the feature learned by W-MSA to rectificate along spatial. Channel enhancing on MLP gives each channel with various intensity $\mathbf{E_c}(\mathbf{Y}_l)$ computed from input and multiply to the feature learned by MLP to rectificate along channel. Therefore, two shortcut enhancements rectificate and augment the visual features using its own original information without bringing extra complicated feature transformations.

The network components are inspired by CBAM [42] which served as a simple kind of attention implicitly, while it serves for a quite different motivation in our ASES here, which is naturally adapt to window-based local vision transformer to explicitly and adaptively enhance the decoupled features. Besides, compared to recent works [18], [19], [43] using additional branch of designed transformation to diverse transformer features as augmented measure, they're usually hard to determine what kind of transformation or how many extra branches are needed and undoubtedly require lots of extra computation. While ASES greatly alleviate these problems.

D. CrossMerge Skip-Connection

To investigate a more appropriate skip-connection form applicable for pyramid vision transformer architecture to transfer multi-scale features from the encoder to the decoder in each block, meanwhile to better cope with merging between query,
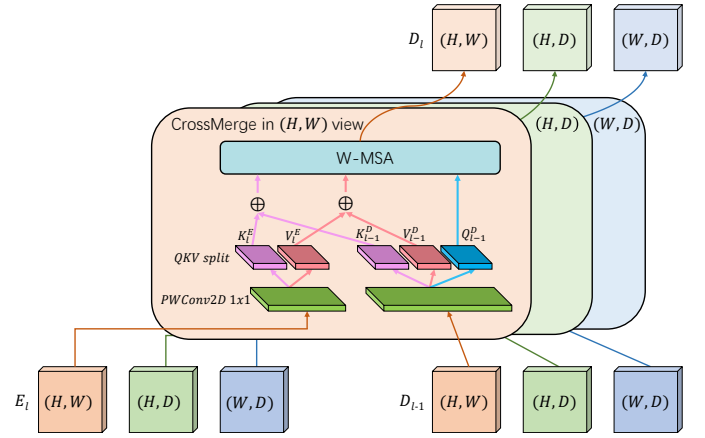
key and value, we design a new skip-connection form named as CrossMerge shown in Fig. 4, which is computed as:

$$\langle \mathbf{K}_l^{\mathbf{E}^v}, \mathbf{V}_l^{\mathbf{E}^v} \rangle = \mathbf{Split}(\mathbf{PWConv2D}(\mathbf{E}_l^v))$$
$$\langle \mathbf{Q}_{l-1}^{\mathbf{D}^v}, \mathbf{K}_{l-1}^{\mathbf{D}^v}, \mathbf{V}_{l-1}^{\mathbf{D}^v} \rangle = \mathbf{Split}(\mathbf{PWConv2D}(\mathbf{D}_{l-1}^v))$$
$$\mathbf{K}_{l-1}^{\mathbf{D}^v{}'} = \mathbf{K}_l^{\mathbf{E}^v} + \mathbf{K}_{l-1}^{\mathbf{D}^v}; \mathbf{V}_{l-1}^{\mathbf{D}^v{}'} = \mathbf{V}_l^{\mathbf{E}^v} + \mathbf{V}_{l-1}^{\mathbf{D}^v}$$
$$\mathbf{CrossMerge}(\mathbf{E}_l^v, \mathbf{D}_{l-1}^v) = \mathbf{W\text{-}MSA}(\mathbf{Q}_{l-1}^{\mathbf{D}^v}, \mathbf{K}_{l-1}^{\mathbf{D}^v{}'}, \mathbf{V}_{l-1}^{\mathbf{D}^v{}'})$$
$$(9)$$

Where $\mathbf{E}_l^v$ and $\mathbf{D}_{l-1}^v$ denotes feature from view v in l-th block of encoder and view v in l-1-th block of decoder. We use PWConv2D to multiply feature channels for stability and split as $\langle \mathbf{Q}, \mathbf{K}, \mathbf{V} \rangle$ respectively in pyramid structure. $\mathbf{K}$ and $\mathbf{V}$ from both encoder and decoder add up while $\mathbf{Q}$ are only from decoder before self-attention. CrossMerge keeps information individually from each view of full-view slice spatial shuffle block's output in encoder and skip-connects with corresponding view in decoder to maintain view consistency.

For existing CatCrossSkip [44] operation, it acts as a concatenation-based cross-attention skip-connection which directly projects all information at once, and is rough to lose feature details. While CrossMerge projects the refined feature in individual views.

IV. EXPERIMENTAL RESULTS

In this section, we describe our experimental settings, including the datasets, evaluation metrics, compared methods and implementation details. We first present results of compared methods, results for the ablation study of key components, and then give discussions of limitations and future work.

A. Datasets and Settings

1) 3D NCCT Hematoma Dataset: The 3-D NCCT hematoma private dataset of hemorrhagic stroke is retrospectively obtained from Chinese Intracranial Hemorrhage Image Database (CICHID). The retrospective study was approved by Institutional Review Board of Peking Union Medical College Hospital (Ethics code:S-K1175). It's for the challenging hematoma expansion shape dense prediction task, which has 3D image

CT_1 and its hematoma lesion segmentation Seg_1 only at initial time in order to predict the expanded lesion Seg_2 at target end time of record in an one-shot setting, where the time interval between initial and target vary widely among different samples. It's a revised dataset based on previous work [45] that some failed cases didn't meet the hematoma expansion standard were excluded. To evaluate Shuffle-Mixer and selected SOTA backbones upon this image-level lesion growth prediction, we replace 3D U-Net with different backbones in the framework [45] based on displacement vector field (DVF). We remove clinical data fusion in DVF to evaluate the performance for a fair comparison. This problem is formulated as a 3D DVF prediction task. We randomly split our dataset into 212, 30 and 60 volumes for train, validation, and test, respectively.

2) Synapse Dataset: The synapse (BTCV) public dataset consists of 30 subjects with abdominal CT scans where 13 organs were annotated, which is used as a small dataset to test for vision transformer. Each CT scan consists of 80 to 225 slices with 512×512 pixels and slice thickness ranging from 1 to 6 mm. Each volume has been pre-processed by normalizing the intensities in the range of [-1000, 1000] HU to [0, 1]. All images are resampled into the isotropic voxel spacing of 1.0 mm during pre-processing. This problem is formulated as a 13 class multi-organ segmentation task. We split synapse dataset into 24 and 6 for train and test respectively, the same as split setting in [37] for a fair comparison.

3) BraTS2019 Dataset: The BraTS2019 public dataset is provided by the Brain Tumor Segmentation 2019 challenge. Each sample is composed of 4 modalities of brain MRI scans. Each modality has a volume of $240\times240\times155$ which has been aligned into the same space. The labels contain 4 classes: background (label 0), necrotic and non-enhancing tumor (label 1), peritumoral edema (label 2) and GD-enhancing tumor (label 4). The segmentation accuracy is measured on 3 classes of compositional regions which are enhancing tumor region (ET, label 1), the tumor core region (TC, labels 1 and 4), and the whole tumor region (WT, labels 1,2 and 4). It contains 335 cases of patients for offline training and 125 cases for online validation by official server.

4) Evaluation Metrics: We use dice similarity coefficient (Dice) score as evaluation metric to measure the predicted region accuracy both on three datasets, represented as: Dice $= \frac{2*TP}{FP+2*TP+FN}$. Additionally, Dice of Change (Doc), Precision, Recall and Jaccard of hematoma expansion region are used to measure predicted region internal filling accuracy more comprehensively on hematoma dataset, which are formulated as: Doc $= \frac{|P_m - M_2| \cap |M_2 - M_1|}{|P_m - M_2| \cup |M_2 - M_1|}$, Precision $= \frac{TP}{FP+TP}$, Recall $= \frac{TP}{TP+FN}$, Jaccard $= \frac{TP}{FP+TP+FN}$. Since shape prediction also concentrates on shape boundary accuracy besides overall region, Hausdorff Distance 95% (HD95) is also used to measure the region boundary accuracy.

B. Compared Methods and Implementation Details

To fully compare different kinds of backbones for dense prediction in medical volume. For hematoma dataset, we use initial hematoma as benchmark, 3D U-Net [8], AHNet [46],

HyperDensenet [47] and Att-Unet [48] as classic simple backbones here. Since it's a one-shot hard task, we also choose cascade ensemble (Cascaded U-Net [49]), multi-task ensemble (segResnetVAE [50]), and model ensemble (nnU-net [51]) to compare. Besides, our motivation is to design a strong 3D vision learner leveraging local vision transformer. Thus we compare with benchmark ViT [10] and SOTA methods of local vision transformer in general vision tasks: Swin-TR [23], Shuffle-TR [25] and CSWin-TR [29] as transformer-based (TR-based) baselines all for three datasets. For synapse dataset, besides simple backbones [52] and TR-based baselines, we compare with CNN-TR: TransUNet [13] and Swin-Unet [21] proposed recently and SOTA UNETR [37]. For BraTS2019 dataset, besides above baselines and previous solutions [53]–[55], we also compare with CNN-TR: Trans-BTS [36] proposed recently and SOTA Segtran [33].

All networks on hematoma dataset are implemented in Pytorch and MONAI[1] framework with Adam optimization for 200 epochs with batch size 4. The learning rate was set to 0.0005. All networks on synpase dataset are implemented and evaluated based on the official code[2] of recent SOTA UNETR. All networks on BraTS2019 dataset are implemented and evaluated based on the official code[3] of recent SOTA Segtran. The input channels are set to 2 for hematoma, 1 for synapse and 4 for BraTS2019 dataset. Besides, the input volume size on three datasets is all set to $128\times128\times128$. In particular, we build our Shuffle-Mixer and other SOTA TR-based methods with the same and common configuration referred in [23], [29], [30] for a fair comparison.

C. Comparison with SOTA upon Hematoma Expansion Shape Prediction Task

TABLE I

PERFORMANCE COMPARISON UPON HEMATOMA EXPANSION SHAPE PREDICTION TASK. (**BOLD** FONT IS FOR THE BEST, <u>*underlined*</u> FONT FOR THE SECOND-BEST AND *wavylines* FONT FOR THE WORST. THESE ARE APPLY TO ALL TABLES IN THIS PAPER.)

Type	Model	Dice↑	Doc↑	Precision↑	Recall↑	Jaccard↑	HD95↓
Benchmark	origin Seg_1	0.5988	0	-	-	-	-
Simple backbone	3D U-Net	0.695	0.4486	0.6097	0.4252	0.2994	12.12
	AHNet	0.6975	0.4565	0.6174	0.4347	0.3065	12.02
	HyperDensenet	0.6906	0.4299	0.6321	0.3837	0.2837	12.8
	Att-Unet	0.6964	0.4495	0.6147	0.4287	0.3	12.44
Ensemble	Cascaded U-Net (share)	0.6938	0.4451	0.6002	0.4263	0.2946	12.71
	Cascaded U-Net (not share)	0.6986	0.4574	0.606	0.4467	0.3082	11.86
	segResnetVAE	0.6993	0.4592	0.617	0.4383	0.3096	11.8
	nnU-net	0.7012	0.4617	0.6171	0.4499	0.3113	11.78
TR-based	pure ViT	- (Not Converge)					
	2D Swin-TR	0.666	0.3395	0.6469	0.2629	0.2105	13.31
	2D Shuffle-TR	0.6942	0.448	0.6117	0.4228	0.2987	12.16
	2D CSWin-TR	0.6898	0.434	0.6164	0.399	0.2866	12.46
	3D video Swin-TR	0.6922	0.4399	0.6165	0.4075	0.2912	12.36
	3D CSWin-TR	0.6969	0.4609	0.5939	0.4513	0.3094	11.9
Ours	3D Shuffle-Mixer	0.7067	0.4992	0.6013	0.5228	0.3456	11.05

We present the performance upon hematoma expansion shape prediction task in Table I. Since the initial hematoma region Seg_1 of CT_1 is directly given in training, it is more important to pay attention to the prediction accuracy of the

[1]https://github.com/Project-MONAI/MONAI

[2]https://github.com/Project-MONAI/research-contributions/tree/master/UNETR/BTCV

[3]https://github.com/askerlee/segtran

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2022.3191974

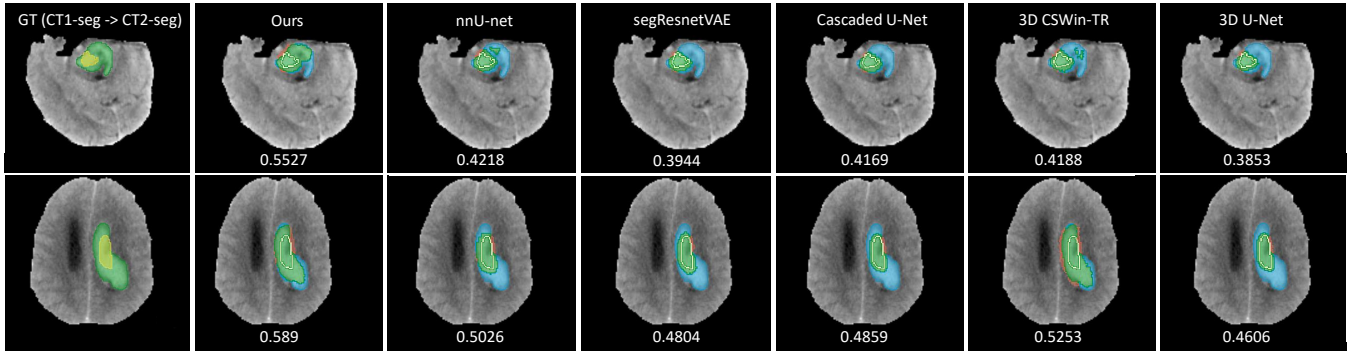PANG *et al.*: PREPARATION OF PAPERS FOR IEEE TRANSACTIONS ON MEDICAL IMAGING 7

Fig. 5. Visualization comparison upon hematoma expansion shape prediction task. The yellow mask in GT denotes Seg_1 and green mask denotes Seg_2. In prediction results, white line denotes the boundary of Seg_1, green mask denotes TP prediction, blue mask denotes FN prediction and orange mask denotes FP prediction. The Doc value of each sample is shown at the bottom of each snapshot.

new expansion region from Seg_1 to Seg_2 when ensuring the accuracy of the overall region. Therefore, we focus on the major metrics Doc and Jaccard. Our model achieves SOTA performance for all metrics except slightly lower precision. Moreover, it outperforms the second-best model nnU-net by 3.75% and 3.43% on two key scores. In particular, our model significantly improved recall by 7.15% at the expense of a little precision sacrificing in contrast to 3D CSWin-TR's second-best recall. In addition, Cascaded U-Net (share) assumes hematoma's linear expansion and Cascaded U-Net (not share) assumes nonlinear expansion. Both mediocre performance indicates that the hematoma doesn't expand regularly with time and our model can make better prediction. All TR-based baselines get poor performance, which indicates that directly applying existing 2D local vision transformer to medical volume doesn't work well, while the Transformer-MLP paradigm designed by us can achieve remarkable results.

We visualize the hematoma expansion shape prediction results in Fig. 5 of Shuffle-Mixer compared with other selected SOTA methods. Ground truth (GT) shows the expansion are highly irregular, where the hematoma regions of Seg_2 and Seg_1 are much different. Except for our model, nnU-net and 3D CSWin-TR, the prediction results of others tend to be conservative, with only a slight expansion along the boundary of Seg_1. Surprisingly, nnU-net correctly predicts only slightly more expansion region, while 3D CSWin-TR and Shuffle-Mixer correctly predict most of the expansion region, where the angle direction and distance of the predicted expansion region are roughly consistent with GT benefiting from the global receptive field of vision transformer. In contrast to 3D CSWin-TR, FP region of our model is much more smaller.

D. Comparison with SOTA upon Popular 3D Image Segmentation Tasks

We evaluate our proposed method upon two popular 3D image segmentation tasks in this subsection. The first one is synapse multi-organ segmentation with CT images and the second one is BraTS2019 tumor segmentation with multi-modal MRI images.

Synapse Multi-organ Segmentation in CT Images: We present the performance upon synapse multi-organ CT segmentation task in Table II. Among all the 12 classes of organ,

Shuffle-Mixer achieves the best performance in 7 classes, second-best performance in 3 classes, and good results in the remaining two classes, with 2.76% better than the second-best model UNETR on average dice score.

We visualize the synapse multi-organ CT segmentation results in Fig. 6 between Shuffle-Mixer and other selected SOTA methods for visualization comparison. From the overall visualization and three zoom-in regions, it can be seen that the prediction results of our model are more precise, refined and closer to GT.

BraTS2019 Tumor Segmentation in MRI Images: We present the performance upon BraTS2019 multi-modal brain tumor MRI segmentation task in Table III. Shuffle-Mixer achieves SOTA results on three classes of tumor regions with 0.92% better than the second-best model Segtran on average dice score. Since the validation samples are not public, so we omit visualization on BraTS2019 here.

E. Effectiveness of Key Component

3D Shuffle-Mixer's full version consists of three key components: 3D Shuffle-Mixer, ASES and CrossMerge. To evaluate the effectiveness and importance of each key component, we conduct the ablation study which compares full version of our model with other three degraded versions as ablation variants: (1) The simplest version which only use pure 3D Shuffle-Mixer as first basic variant; (2) We use simple 3D Shuffle-Mixer equipped with ASES as second variant; (3) We use simple 3D Shuffle-Mixer with CrossMerge as third variant; (4) The full version consists of all three key components.

As shown in Table I, II and III, we can easily see that the overall best baselines are nnU-net (0.4617 on Doc and 0.3113 on Jaccard) on hematoma dataset, UNETR (0.7838 on average dice) on Synapse and Segtran (0.8139 on average dice) on BraTS2019 respectively. Table IV shows the overall ablation results. The simplest version (1) is already better than the best baseline on three datasets significantly, which outperforms by 2.61% and 2.34% on Doc and Jaccard score upon hematoma task, by 1.80% and 0.55% on average dice upon other two segmentation tasks as well. It shows the basic structure of our Shuffle-Mixer has universal superiority among three dense prediction tasks. Ablation (2) equipping with ASES futher

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2022.3191974

8   IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. XX, NO. XX, XXXX 2020

TABLE II
PERFORMANCE COMPARISON UPON SYNAPSE MULTI-ORGAN CT SEGMENTATION TASK.

Type	Model	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	AG	Average Dice↑
Simple backbone	V-Net	0.9454	**0.9354**	0.9269	0.1521	0.2932	0.9633	**0.8805**	0.9004	0.8529	0.2881	0.7377	0.2196	0.6746
	3D U-Net	0.8955	0.8955	0.8937	0.5389	0.6865	0.9424	0.7711	0.8576	0.8108	0.6412	0.5945	0.5674	0.7579
	Att-Unet	0.8875	0.8778	0.8758	**0.5634**	0.6948	0.9005	0.7805	0.837	0.8086	0.6443	0.7064	0.6397	0.768
Paper used	TransUNet	0.9135	0.8374	0.8332	0.445	0.7204	0.9565	0.734	**0.9117**	0.8073	0.67	0.5927	0.5731	0.7496
	Swin-Unet	0.8751	0.908	0.9189	0.5203	0.6028	0.9459	0.7195	0.898	0.7986	**0.6858**	0.7218	0.5939	0.7657
Recent SOTA	UNETR	0.9153	0.9184	0.9071	0.507	0.7094	0.9482	0.8005	0.8653	0.8117	0.6614	0.7138	0.648	0.7838
TR-based	pure ViT	0.8406	0.843	0.7915	0.2799	0.3774	0.9155	0.5476	0.7646	0.603	0.3967	0.312	0.1565	0.569
	2D Swin-TR	0.833	0.8234	0.783	0.349	0.5269	0.936	0.6389	0.7072	0.672	0.3746	0.4817	0.2686	0.6162
	2D shuffle-TR	0.8598	0.8842	0.8827	0.4405	0.667	0.9516	0.7263	0.9003	0.8357	0.6795	0.6221	0.5417	0.7493
	2D CSWin-TR	0.8431	0.8797	0.8725	0.4483	0.6756	0.9439	0.6981	0.8993	0.8363	0.6803	0.6185	0.5409	0.7447
	3D video Swin-TR	0.9102	0.9112	0.8682	0.5381	0.5993	0.9534	0.7117	0.8317	0.7493	0.6068	0.576	0.5139	0.7308
	3D CSWin-TR	0.9168	0.9127	0.8985	0.4143	0.6589	0.9461	0.7684	0.8755	0.8184	0.6083	0.6845	0.5841	0.7572
Ours	3D Shuffle-Mixer	**0.9478**	0.9353	**0.9282**	0.5153	**0.7253**	**0.9667**	0.8314	0.9015	**0.8597**	0.6772	**0.7719**	**0.6763**	**0.8114**

Spleen Right Kidney Left Kidney Gallbladder Esophagus Liver Stomach
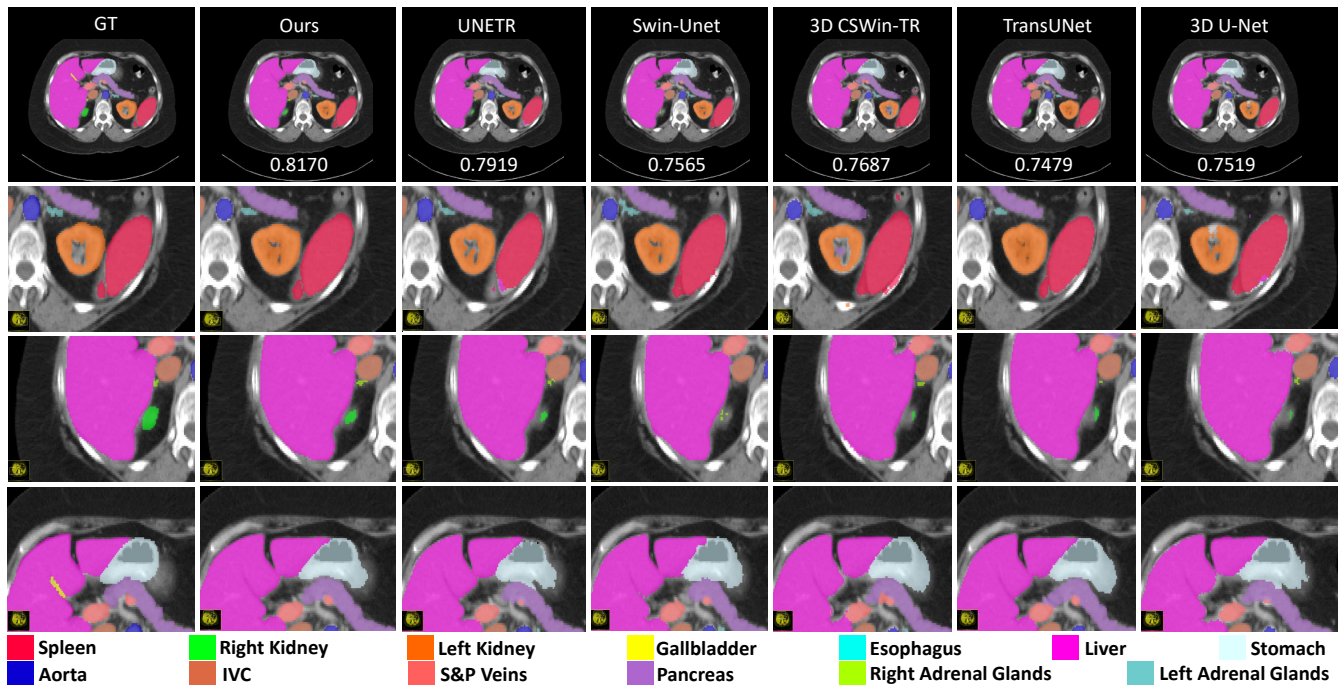Aorta IVC S&P Veins Pancreas Right Adrenal Glands Left Adrenal Glands

Fig. 6. Visualization comparison upon synapse multi-organ CT segmentation task. Snapshots in the first row are overall visualizations and the last three lines are zoom-in region visualizations. The average dice value of each sample is shown at the bottom of each snapshot in the first row.

TABLE III
PERFORMANCE COMPARISON UPON BRATS2019 MULTI-MODAL BRAIN TUMOR MRI SEGMENTATION TASK.

Type	Model	ET	WT	TC	Average Dice↑
Simple backbone	V-Net	0.6362	0.8583	0.653	0.7158
	3D U-Net	0.6165	0.8404	0.6675	0.7082
	Att-Unet	0.7148	0.8885	0.8099	0.8044
Paper used	Extension of nnU-Net	0.737	0.894	0.807	0.8127
	Bag of tricks	0.729	0.904	0.802	0.8117
	KiU-Net	0.734	0.877	0.742	0.7843
	TransBTS	0.735	0.895	0.803	0.811
Recent SOTA	Segtran	0.7381	0.8914	0.8122	0.8139
TR-based	pure ViT	0.583	0.822	0.608	0.671
	2D Swin-TR	0.6392	0.8336	0.7283	0.7337
	2D Shuffle-TR	0.6857	0.8775	0.7477	0.7703
	2D CSWin-TR	0.6723	0.8648	0.7391	0.7587
	3D video Swin-TR	0.6862	0.8621	0.744	0.7641
	3D CSWin-TR	0.6869	0.8611	0.7531	0.767
Ours	3D Shuffle-Mixer	**0.7482**	**0.9064**	**0.8147**	**0.8231**

improves the performance by approximately 0.8% to 0.9% upon hematoma and synapse tasks and 0.3% on BraTS2019. This is because ASES rectifies and enhances the decoupled features on spatial and channel dimensions adaptively, which is also a significant improvement. Ablation (3) adpoting CrossMerge improves the basis by 0.34% and 0.32% upon hematoma task, by approximately 0.1% to 0.2% upon two segmentation tasks. Although CrossMerge improves not much as the first two key components, overall stable improvement also verifies the effectiveness of the skip-connection method proposed by us, which is more suitable for pyramid local vision transformer to add the missing multi-resolution details with full-view structure. The full version (4) of our method achieves the best performance along all ablations and surpasses best baselines on three datasets by a large margin. All ablation study mentioned above demonstrates that each key component plays a role in the overall architecture and improves the

TABLE IV

OVERALL ABLATION STUDY. EFFECTIVENESS OF EACH KEY COMPONENT OF OUR PROPOSED METHOD UPON THREE DATASETS. (BLUE SUPERSCRIPT INDICATES THE PERFORMANCE COMPARED WITH ABLATION (1) BASIC PURE 3D SHUFFLE-MIXER.)

Overall method form			Hematoma		Synapse	BraTS2019			
3D Shuffle-Mixer	ASES	CrossMerge	Doc↑	Jaccard↑	Average Dice↑	ET	WT	TC	Average Dice↑
✓			0.4878	0.3347	0.8018	0.7443	0.9031	0.8109	0.8194
✓	✓		$0.4965^{+0.87\%}$	$0.3431^{+0.84\%}$	$0.8096^{+0.78\%}$	0.7475	0.906	0.8134	$0.8223^{+0.29\%}$
✓		✓	$0.4912^{+0.34\%}$	$0.3379^{+0.32\%}$	$0.8033^{+0.15\%}$	0.7453	0.9037	0.812	$0.8203^{+0.09\%}$
✓	✓	✓	$\mathbf{0.4992}^{+1.14\%}$	$\mathbf{0.3456}^{+1.09\%}$	$\mathbf{0.8114}^{+0.96\%}$	0.7482	0.9064	0.8147	$\mathbf{0.8231}^{+0.37\%}$

performance.

F. Inner-Analysis upon 3D Shuffle-Mixer

TABLE V

INNER ABLATION STUDY. EFFECTIVENESS OF EACH COMPONENT OF 3D SHUFFLE-MIXER ON THREE DATASETS.

3D Shuffle-Mixer form	Hematoma	Synapse	BraTS2019
	Doc↑	Average Dice↑	Average Dice↑
Ours (full version)	**0.4992**	**0.8114**	**0.8231**
W-MSA w/o shuffle	0.4723	0.7868	0.7841
Single view (only H,W)	0.4921	0.7994	0.8179
W/o context mixing	0.4916	0.7945	0.8152
MLP dense context mixing	0.4887	0.7982	0.8167
MSA dense context mixing	- (Out of GPU memory)		
Mixer first	0.489	0.8085	0.8154
View-aware only	0.4881	0.8007	0.8155
Slice-aware only	0.4865	0.8038	0.8173

To have a better understanding and analysis of basic 3D Shuffle-Mixer, we conduct a series of extended ablation experiments on the components of 3D Shuffle-Mixer. To facilitate comparison, we use the full version of our method as the initial structure for all ablation experiments and adjust the setting of 3D Shuffle-Mixer part while keep the other two key components the same.

We conduct four sets of ablation experiments to evaluate 3D Shuffle-Mixer, as shown in Table V: (1) To investigate the effectiveness of the full-view slice spatial shuffle operation, we test two settings, where the first one replaces spatial transpose shuffle with a pure W-MSA without shuffle and the second one only reserves the (H, W) single view. Both ablation settings degrade the performance on all datasets since spatial shuffle helps W-MSA learn from cross-windows to increase receptive field for the transformer, and single-view misses complementary volume information from other views. (2) To investigate the effectiveness of the slice-aware volume context mixing, we compare three settings, where the first one removes the mixing module entirely, and the second and third ones use MLP and MSA dense contest mixing strategy respectively. W/o mixing and MLP dense mixing both degrade performance since mixing brings missing volume context and axial-MLP is sparse which mitigates over-fitting, however, dense MLP introduces extra parameters and increases the risk of MLP overfitting, which leads to poor context processing. MSA dense mixing runs out of GPU memory and does not generate results. (3) To investigate the order of Shuffle and

Mixer operations, we test with the setting where Mixer is run first (Mixer first). Performance reductions prove the logic of our Shuffle-Mixer learning strategy. (4) To investigate the importance of slice-aware and view-aware, we test with two settings to replace APE with slice- or view- aware only. Dramatic reductions indicate that both awareness are necessary to mix, aggregate, and realize the importance of different contexts.

G. Inner-Analysis upon Adaptive Scaled Shortcut

TABLE VI

INNER ABLATION STUDY. EFFECTIVENESS OF EACH COMPONENT OF ADAPTIVE SCALED ENHANCED SHORTCUT ON THREE DATASETS.

ASES form	Hematoma	Synapse	BraTS2019
	Doc↑	Average Dice↑	Average Dice↑
Ours (full version)	**0.4992**	**0.8114**	**0.8231**
W/o ASES	0.4912	0.8033	0.8203
Only spatial enhance on W-MSA	0.4978	0.8082	0.8219
Only channel enhance on MLP	0.4937	0.8065	0.8211
3D CSWin-TR	0.4609	0.7572	0.767
3D CSWin-TR w/ ASES	0.4674	0.7681	0.773
Channel enhance first	0.4913	0.8008	0.8175

To have a better understanding and analysis of ASES, we conduct a series of extended ablation experiments on the components of ASES.

Table VI shows inner ablation study results of ASES: (1) To investigate the effectiveness of the spatial enhance and channel enhance in ASES, we test with three settings, where the first one removes ASES, the second one only has spatial enhance on W-MSA and the third one only has channel enhance on MLP. W/o ASES achieves the lowest performance, adding spatial or channel enhance alone both improves performance to some extent on all three datasets. (2) To investigate the general enhanced capability of ASES, we test with two settings, where one is pure 3D CSWin-TR and another one is 3D CSWin-TR w/ ASES. We find equipping with ASES strongly improves performance on other local vision transformer which demonstrates the general capability and suitability of our ASES. (3) To investigate the order of spatial and channel enhanced shortcuts, we compare with one ablation where channel enhance is run first (channel enhance first). The performance degradation shows that spatial enhance needs to be applied before channel enhance, which is consistent with vision transformer design.

H. Inner-Analysis upon CrossMerge Skip-Connection

TABLE VII

INNER ABLATION STUDY. EFFECTIVENESS OF EACH COMPONENT OF SKIP-CONNECTION ON THREE DATASETS.

Skip-Connection form	Hematoma	Synapse	BraTS2019
	Doc↑	Average Dice↑	Average Dice↑
Ours (full version)	**0.4992**	**0.8114**	**0.8231**
CatLinear (basic U-Net)	0.4878	0.8018	0.8194
CatSkip (Uformer)	0.4945	0.808	0.8202
CrossSkip (Uformer)	0.4917	0.8023	0.8187
CatCrossSkip (Uformer)	0.4938	0.8076	0.821

To find a skip-connection form suitable for vision transformer and have a better understanding of CrossMerge, we conduct extended experiments as five ablations on the components of skip-connections. Table VII shows inner ablation study results of skip-connection: CatLinear is used in basic U-Net. CatSkip, CrossSkip and CatCrossSkip are three variants proposed in Uformer [44]. CrossMerge is proposed in our method. Our CrossMerge achieves the best performance on three datasets, performances of CatLinear and CrossSkip are generally poor since CatLinear is more suitable for backbones of CNN structure and CrossSkip directly converts features from encoder into key, value and converts from decoder into query, missing the merge procedure of information. CatSkip and CatCrossSkip achieve the medium performance. Although these two forms merge features from encoder to decoder, our CrossMerge benefits from refined view-wise skip-connections and performs more stably due to PWConv2D.

I. TR-based Method Build Configurations

We refer to the common configuration in general vision task works [23], [25], [29], [32] to build Shuffle-Mixer and other TR-based compared methods. Table VIII shows the detailed build configurations for TR-based methods implemented in this paper. The depth of pyramid architecture is set as 4, that is, stage-4 is the bottom stage in encoder and setting in decoder is symmetric. Token resolution in stage-1 is $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$ through patch embedding and $2\times$ downsampling of resolution in each stage. The window size is set as 4 in all stages. Block numbers are set as $\{1, 2, 8, 1\}$ and head numbers are set as $\{3, 6, 12, 24\}$ in each stage. Embedding channel dimensions are set as $\{96, 192, 384, 768\}$ which are $32\times$ #head numbers. The expansion ratio of each MLP is set as 4 for all experiments. Besides, the linear layer in vision transformer's MLP of Shuffle-Mixer is implemented by PWConv2D for consideration of better handling 2D spatial slices.

J. Analysis upon Computational Complexity

We use W-MSA in spatial shuffle block for efficient global context modeling with low computational complexity compared with pure MSA. Supposing each window contains $M \times M$ tokens, the computational complexity of pure MSA and W-MSA on a spatial slice of $H \times W$ tokens with channel dimension C is:

$$\Omega(\textbf{Pure MSA}) = 4HWC^2 + 2(HW)^2C$$
$$\Omega(\textbf{W-MSA}) = 4HWC^2 + 2M^2HWC \quad (10)$$

The complexity of pure MSA is quadratic of slice spatial (H, W), yet W-MSA reduces complexity to linear of (H, W) with fixed window size M ($M = 4$, which can be regarded as a constant term) by computing self-attention in each non-overlapping local window. Generally, $M^2 \ll HW$ where HW is huge in dense prediction.

Besides, we use axial-MLP in mixing with sparse computation to capture remaining volume context. We compare the complexity of axial-MLP mixing (**A-MLP-M**) with dense MLP context mixing (**D-MLP-M**) and dense MSA context mixing (**D-MSA-M**) introduced in Section IV-F as follows:

$$\Omega(\textbf{D-MSA-M}) = 8HWDC^2 + 2(HWD)^2C + 2\alpha HWDC^2$$
$$\Omega(\textbf{D-MLP-M}) = 2\alpha HWDC(DC) + 3HWDC^2$$
$$\Omega(\textbf{A-MLP-M}) = 2\alpha HWDC(D + C) + 3HWDC^2$$
$$\quad (11)$$

Noticed that **D-MSA-M** is of pure 3D usage here which is quadratic of 3D volume (H, W, D) with a heavy computation. **A-MLP-M** uses the residual two-branch structure to project context sparsely and individually in order to mitigate overfitting and acquire better performance, which costs a relatively low computation in contrast to **D-MLP-M** where $D + C \ll DC$. Consider the complexity of view MLP aggregator as $3HWDC^2$, which can be negligible as a small computational term over whole pipeline of Shuffle-Mixer block (**SM-B**), therefore the complexity of **SM-B** is computed as:

$$\Omega(\textbf{SM-B}) = \Omega(\textbf{W-MSA}) + \Omega(\textbf{A-MLP-M}) \quad (12)$$

Moreover, the W-MSA and axial-MLP are parameters shared among full-view which further reduces complexity dramatically and enhances locality with inductive bias.

TABLE IX

MODEL COMPLEXITY COMPARISON.

Type	Selected Model	#Param. (10^6)↓	FLOPs (10^{10})↓
2D CNN-TR	Swin-Unet	41.35	36.10
	TransUNet	66.88	104.12
3D CNN-TR	TransBTS	32.99	26.42
	Segtran	**32.97**	65.37
	UNETR	95.83	**17.45**
3D TR block	pure ViT	149.64	993.47
	3D video Swin-TR	60.68	22.87
	3D CSWin-TR	89.56	38.41
	Shuffle-Mixer	85.18 (6th-lowest)	32.30 (4th-lowest)

Table IX shows the results of model complexity. We select some recent SOTA works and classify them into the 2D CNN-TR hybrid model, 3D CNN-TR hybrid model, and 3D TR designed block model for complexity comparison. All models are tested under the setting of hematoma shape prediction task. Although 2D CNN-TRs process 2D convolution on image scans, they use pure 3D MSA directly in the model which leads to a huge computation increase. Therefore, pure 3D ViT has the largest number of parameters and FLOPs, far more than any other models, which makes running pyramid pure ViT out of memory or extremely slow. In addition, Segtran has the optimal number of parameters and UNETR has the optimal FlOPs. Shuffle-Mixer has the 6th-lowest parameters (85.16 million) and the 4th-lowest FLOPs (32.30 billion)

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2022.3191974

PANG *et al.*: PREPARATION OF PAPERS FOR IEEE TRANSACTIONS ON MEDICAL IMAGING
11

TABLE VIII
DETAILED BUILD CONFIGURATIONS FOR TR-BASED METHOD IMPLEMENTATION.

Depth	#Token Resolution	#Window Size	#Embedding Channels	#Block Numbers	#Head Numbers	#MLP Ratio
Stage-1	$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$	4	96	1	3	4
Stage-2	$\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}$	4	192	2	6	4
Stage-3	$\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$	4	384	8	12	4
Stage-4	$\frac{H}{32} \times \frac{W}{32} \times \frac{D}{32}$	4	768	1	24	4

among all nine models which is an acceptable and moderate computational complexity compared to UNETR (95.83 million #Params.) and Swin-Unet (36.10 billion FLOPs). Compared with 3D CSWin-TR, Shuffle-Mixer has less complexity and achieves SOTA performance on 3D medical volume.

K. Robustness and Convergence Speed Analysis

TABLE X
STANDARD DEVIATION COMPARISON OF PERFORMANCE FROM 4 RUNS
FOR EACH MODEL ON TWO DATASETS.

Hematoma		Synapse	
Selected Model	STDEV.(%)↓	Selected Model	STDEV.(%)↓
3D U-Net	0.8292	3D U-Net	1.4917
Cascaded U-Net (not share)	0.4869	TransUNet	1.1031
segResnetVAE	0.7065	Swin-Unet	1.1769
nnU-net	0.4356	UNETR	1.1790
3D CSWin-TR	0.3925	3D CSWin-TR	0.7928
Ours	**0.1717**	Ours	**0.4211**

In addition to the good performance of Shuffle-Mixer on three dense prediction datasets of 3D medical image analysis, we find our model has remarkable robustness and competitive convergence speed. Table X shows the standard deviation (STDEV) performance from each 4 runs of selected models on hematoma and synapse datasets. It can be easily seen that the model with pure CNN or hybrid CNN-TR structure is more unstable than the 3D CSWin-TR and Shuffle-Mixer with 3D designed TR block. 3D U-Net's STDEV reaches at 0.8292% and 1.4917% respectively which is the most unstable. The more vision transformer components are added, the more stable the model will be, which is also consistent with empirical knowledge [56] in general vision tasks. Our Shuffle-Mixer achieves the minimum STDEV at 0.1717% and 0.4211% on both datasets which has remarkable robustness among selected models.

We also compare the model convergence speed during training to analyze robustness. Fig. 7 shows the Doc score of every 5 epochs on validation set upon hematoma expansion shape prediction task. Shuffle-Mixer's Doc is significantly higher than other models, and fluctuates only in a small range of 0.46 to 0.47, where other models fluctuate more severely. Shuffle-Mixer has performed better than other models since the first epoch, and stays ahead throughout the whole training process which shows our model converges steadily almost within the first few epochs.
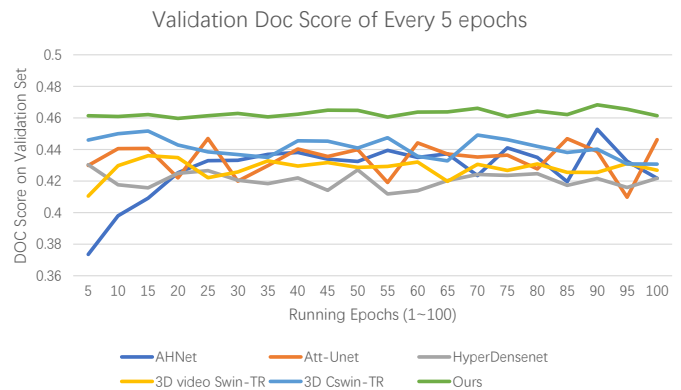

Fig. 7. Doc score of every 5 epochs on validation set upon hematoma expansion shape prediction task.
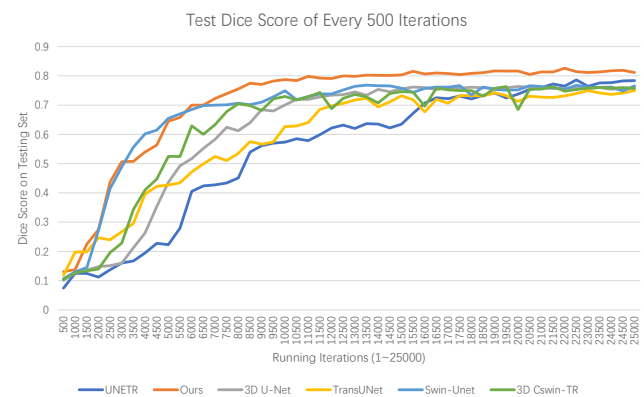

Fig. 8. Dice score of every 500 iterations on testing set upon synapse multi-organ CT segmentation task.

Fig. 8 shows the dice score of every 500 iterations on the test set in the synapse multi-organ CT segmentation task since there is no validation set. It can be seen that only the performance of Shuffle-Mixer and Swin-Unet grows at the fastest speed as training iterations increase, where they quickly converge and the performance tends to be stable. It owes much to the choice of local vision transformer architecture for both models. What is more interesting is that although UNETR is the second-best model in performance, its convergence speed is obviously the slowest, which is probably due to the direct usage of pure MSA resulting in the extra difficulty for model learning. The increasing number of parameters makes UNETR more unstable in the later stages and more prone to over-fitting.

L. Discussion of Limitations and Future Work

3D Shuffle-Mixer is proposed specifically for dense prediction in medical volume. In order to better deal with 3D volumetric data, we reduce computational complexity as much as possible while ensuring performance via concise but efficient computation (e.g. W-MSA and sparse axial-MLP), parameter sharing and full-view 2D computing. Our model not only possess the learning capability of vision transformer on large amounts of data revealed in the general vision tasks, so as to improve the upper limit of performance, but also has greater inductive bias power of the local vision transformer for dense prediction due to the full-view spatial shuffle and context mixing. The data efficient learning ability from small dataset is demonstrated in our experiments on synapse dataset. Although our method achieves excellent performance, it comes with limitations. One limitation is that the data in our training set has similar resolution and scale, and insufficient data scale greatly reduces inductive bias power and affects the learning capability and representational ability of 3D Shuffle-Mixer to some extent. This is a detrimental issue for local vision transformer which leads to deficient context capturing and severe performance degradation. In the future, we will explore more proper design of data efficient learning specifically for the Transformer-MLP paradigm.

Another limitation is that all dimensions of 3D volume are the same size in our setting for convenience, where processing of views in different sizes requires additional pre-processing or presetting hyperparameter in Shuffle-Mixer network to facilitate feature fusion learning from full-view features. This is a strict input pre-condition and the thin-cut slices cannot be handled easily. In the future, we will explore more appropriate manners to handle different sizes of full-view information. The third limitation is that the current model under-performed on cases with very large expansion or multiple lesions merging into one. In order to address this, we need to build a larger training set with more such rare cases.

V. CONCLUSION

In this paper, we have proposed a novel 3D Shuffle-Mixer network exploiting Local Vision Transformer-MLP paradigm for dense prediction in medical volume. In particularly, our model rearranges volume into full-view slices to obtain redundant spatial information, mixes slice-aware features and aggregates view-aware features to fully capture the volume context in a context-aware manner. Besides, we have proposed an Adaptive Scaled Enhanced Shortcut to rectify and augment feature representation of window-based local vision transformer. Furthermore, an appropriate skip-connection form CrossMerge for pyramid transformer is presented to merge feature from encoder to decoder. The experimental results and analysis upon dense prediction tasks have demonstrated that our proposed model outperforms other SOTA compared methods in both quantitative and qualitative measures.

REFERENCES

[1] G. Wang, M. A. Zuluaga, W. Li, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin et al., "Deepigeos: a deep interactive geodesic framework for medical image segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 7, pp. 1559–1572, 2018.

[2] T. Wang, Y. Lei, Y. Fu, J. F. Wynne, W. J. Curran, T. Liu, and X. Yang, "A review on medical imaging synthesis using deep learning and its clinical applications," Journal of Applied Clinical Medical Physics, vol. 22, no. 1, pp. 11–36, 2021.

[3] Q. Huang, Z. Miao, S. Zhou, C. Chang, and X. Li, "Dense prediction and local fusion of superpixels: A framework for breast anatomy segmentation in ultrasound image with scarce data," IEEE Transactions on Instrumentation and Measurement, 2021.

[4] B. Wang, Y. Lei, S. Tian, T. Wang, Y. Liu, P. Patel, A. B. Jani, H. Mao, W. J. Curran, T. Liu et al., "Deeply supervised 3d fully convolutional networks with group dilated convolution for automatic mri prostate segmentation," Medical physics, vol. 46, no. 4, pp. 1707–1718, 2019.

[5] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: a learning framework for deformable medical image registration," IEEE transactions on medical imaging, vol. 38, no. 8, pp. 1788–1800, 2019.

[6] X. Wang, S. Han, Y. Chen, D. Gao, and N. Vasconcelos, "Volumetric attention for 3d medical image segmentation and detection," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 175–184.

[7] M. Ebner, G. Wang, W. Li, M. Aertsen, P. A. Patel, R. Aughwane, A. Melbourne, T. Doel, A. L. David, J. Deprest et al., "An automated localization, segmentation and reconstruction framework for fetal brain mri," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2018, pp. 313–320.

[8] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in International conference on medical image computing and computer-assisted intervention. Springer, 2016, pp. 424–432.

[9] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, 2018, pp. 3–11.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations, 2020.

[11] S. d'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," arXiv preprint arXiv:2103.10697, 2021.

[12] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, D. Keysers, J. Uszkoreit, M. Lucic et al., "Mlp-mixer: An all-mlp architecture for vision," arXiv preprint arXiv:2105.01601, 2021.

[13] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.

[14] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," arXiv preprint arXiv:2102.10662, 2021.

[15] J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du, "Vit-v-net: Vision transformer for unsupervised volumetric medical image registration," arXiv preprint arXiv:2104.06468, 2021.

[16] D. Wang, Z. Wu, and H. Yu, "Ted-net: Convolution-free t2t vision transformer-based encoder-decoder dilation network for low-dose ct denoising," arXiv preprint arXiv:2106.04650, 2021.

[17] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, "Incorporating convolution designs into visual transformers," in Proceedings of the IEEE/CVF International Conference on Computer Vision, October 2021, pp. 579–588.

[18] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "Deepvit: Towards deeper vision transformer," arXiv preprint arXiv:2103.11886, 2021.

[19] D. Zhou, Y. Shi, B. Kang, W. Yu, Z. Jiang, Y. Li, X. Jin, Q. Hou, and J. Feng, "Refiner: Refining self-attention for vision transformers," arXiv preprint arXiv:2106.03714, 2021.

[20] Q. Han, Z. Fan, Q. Dai, L. Sun, M.-M. Cheng, J. Liu, and J. Wang, "Demystifying local vision transformer: Sparse connectivity, weight sharing, and dynamic weight," arXiv preprint arXiv:2106.04263, 2021.

[21] 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," arXiv preprint arXiv:2105.05537, 2021.

This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2022.3191974

PANG *et al.*: PREPARATION OF PAPERS FOR IEEE TRANSACTIONS ON MEDICAL IMAGING
13

[22] M. Wu, Y. Qian, X. Liao, Q. Wang, and P.-A. Heng, "Hepatic vessel segmentation based on 3dswin-transformer with inductive biased multi-head self-attention," *arXiv preprint arXiv:2111.03368*, 2021.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021, pp. 10 012–10 022.

[24] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," *arXiv preprint arXiv:2104.13840*, vol. 1, no. 2, p. 3, 2021.

[25] Z. Huang, Y. Ben, G. Luo, P. Cheng, G. Yu, and B. Fu, "Shuffle transformer: Rethinking spatial shuffle for vision transformer," *arXiv preprint arXiv:2106.03650*, 2021.

[26] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021, pp. 568–578.

[27] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvtv2: Improved baselines with pyramid vision transformer," *arXiv preprint arXiv:2106.13797*, 2021.

[28] J. Fang, L. Xie, X. Wang, X. Zhang, W. Liu, and Q. Tian, "Msg-transformer: Exchanging local spatial information by manipulating messenger tokens," *arXiv preprint arXiv:2105.15168*, 2021.

[29] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," *arXiv preprint arXiv:2107.00652*, 2021.

[30] D. Lian, Z. Yu, X. Sun, and S. Gao, "As-mlp: An axial shifted mlp architecture for vision," *arXiv preprint arXiv:2107.08391*, 2021.

[31] S. Chen, E. Xie, C. Ge, D. Liang, and P. Luo, "Cyclemlp: A mlp-like architecture for dense prediction," *arXiv preprint arXiv:2107.10224*, 2021.

[32] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," *arXiv preprint arXiv:2106.13230*, 2021.

[33] S. Li, X. Sui, X. Luo, X. Xu, L. Yong, and R. S. M. Goh, "Medical image segmentation using squeeze-and-expansion transformers," in *The 30th International Joint Conference on Artificial Intelligence*, 2021.

[34] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, "U-net transformer: self and cross attention for medical image segmentation," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2021, pp. 267–276.

[35] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," *arXiv preprint arXiv:2103.03024*, 2021.

[36] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "Transbts: Multimodal brain tumor segmentation using transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 109–119.

[37] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, January 2022, pp. 574–584.

[38] R. Tao and G. Zheng, "Spine-transformers: Vertebra detection and localization in arbitrary field-of-view spine ct with transformers," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 93–103.

[39] Y. Luo, Y. Wang, C. Zu, B. Zhan, X. Wu, J. Zhou, D. Shen, and L. Zhou, "3d transformer-gan for high-quality pet reconstruction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 276–285.

[40] T. Xiao, P. Dollar, M. Singh, E. Mintun, T. Darrell, and R. Girshick, "Early convolutions help transformers see better," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[41] G. Ke, D. He, and T.-Y. Liu, "Rethinking positional encoding in language pre-training," in *International Conference on Learning Representations*, 2020.

[42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision*, 2018, pp. 3–19.

[43] Y. Tang, K. Han, C. Xu, A. Xiao, Y. Deng, C. Xu, and Y. Wang, "Augmented shortcuts for vision transformers," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[44] Z. Wang, X. Cun, J. Bao, and J. Liu, "Uformer: A general u-shaped transformer for image restoration," *arXiv preprint arXiv:2106.03106*, 2021.

[45] T. Xiao, H. Zheng, X. Wang, X. Chen, J. Chang, J. Yao, H. Shang, and P. Liu, "Intracerebral haemorrhage growth prediction based on displacement vector field and clinical metadata," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 741–751.

[46] S. Liu, D. Xu, S. K. Zhou, O. Pauly, S. Grbic, T. Mertelmeier, J. Wicklein, A. Jerebko, W. Cai, and D. Comaniciu, "3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 851–858.

[47] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation," *IEEE transactions on medical imaging*, vol. 38, no. 5, pp. 1116–1126, 2018.

[48] 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.

[49] S. Zhao, Y. Dong, E. I. Chang, Y. Xu *et al.*, "Recursive cascaded networks for unsupervised medical image registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 600–10 610.

[50] A. Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 311–320.

[51] 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.

[52] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[53] F. Wang, R. Jiang, L. Zheng, C. Meng, and B. Biswal, "3d u-net based brain tumor segmentation and survival days prediction," in *International MICCAI Brainlesion Workshop*. Springer, 2019, pp. 131–141.

[54] Y.-X. Zhao, Y.-M. Zhang, and C.-L. Liu, "Bag of tricks for 3d mri brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2019, pp. 210–220.

[55] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, and V. M. Patel, "Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 363–373.

[56] S. Paul and P.-Y. Chen, "Vision transformers are robust learners," *arXiv preprint arXiv:2105.07581*, 2021.