MLP-GAN FOR BRAIN VESSEL IMAGE SEGMENTATION

Bin Xie\textsuperscript{1}, Hao Tang\textsuperscript{2}, Bin Duan\textsuperscript{1}, Dawen Cai\textsuperscript{3}, Yan Yan\textsuperscript{1}

\textsuperscript{1}Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA
\textsuperscript{2}Computer Vision Lab, ETH, Zurich, Switzerland
\textsuperscript{3}Department of Cell and Developmental Biology, University of Michigan, Ann Arbor, MI, USA

ABSTRACT

Brain vessel image segmentation can be used as a promising biomarker for better prevention and treatment of different diseases. One successful approach is to consider the segmentation as an image-to-image translation task and perform a conditional Generative Adversarial Network (cGAN) to learn a transformation between two distributions. In this paper, we present a novel multi-view approach, MLP-GAN, which splits a 3D volumetric brain vessel image into three different dimensional 2D images (i.e., sagittal, coronal, axial) and then feed them into three different 2D cGANs. The proposed MLP-GAN not only alleviates the memory issue which exists in the original 3D neural networks but also retains 3D spatial information. Specifically, we utilize U-Net as the backbone for our generator and redesign the pattern of skip connection integrated with the MLP-Mixer\textsuperscript{1} which has attracted lots of attention recently. Our model obtains the ability to capture cross-patch information to learn global information with the MLP-Mixer. Extensive experiments are performed on the public brain vessel dataset\textsuperscript{2} that show our MLP-GAN outperforms other state-of-the-art methods.

\textbf{Index Terms}— MLP, Multi-View, GANs, Segmentation

1. INTRODUCTION

Medical image analysis has achieved numerous and remarkable progress with deep learning\textsuperscript{3, 4, 5} for the past few years. For example, brain vessel image segmentation\textsuperscript{2} can be used as a promising biomarker for disease diagnosis and prevention via 3D MRI analysis which is quite meaningful to human health studies. However, brain vessel image segmentation is a complicated task due to the interweaving nature and different thicknesses of brain vessels. The image segmentation problem actually can be considered as an image-to-image translation task that transforms original images into segmentation masks, which can be solved by conditional Generative Adversarial Network (cGAN). Therefore, we utilize cGAN to solve the brain vessel image segmentation task.

However, it is quite challenging to build a large 3D medical dataset due to the difficulty of acquisition and annotation. Thus, we used a synthetic brain vessel dataset\textsuperscript{2} that is accurately annotated. Meanwhile, 3D-based deep learning architectures in literature have numerous parameters, which mean significant computation resources are needed. To address these problems, one straightforward approach is to simply use 2D-based deep neural networks by slicing 3D volumetric images into a sequence of 2D plane images along only one dimension. However, this simple strategy will cause the inevitable loss of useful 3D spatial information. Therefore, we propose a novel multi-view approach that splits 3D volumetric images into a sequence of 2D plane images in three different planes for the same arranged indexes to be fed into three different 2D-based cGANs. In this way, we alleviate the memory issue and retain 3D spatial information.

Specifically, we utilize the U-Net\textsuperscript{3} as the backbone for our generator network. Although the skip connections in the original U-Net combine low-level feature maps to semantic feature maps, a simple concatenation cannot maximize their benefits. Recently, MLP-Mixer\textsuperscript{1} has been proposed and obtained lots of attention due to cross-patch and cross-channel communication. Therefore, we redesign our model with the MLP-Mixer as MLP-Mixer enhanced generator. We design different patterns of skip connection by MLP-Mixer to eliminate the semantic gap among different feature maps.

To summarize, we make the following contributions: (i) We propose a novel MLP-GAN for the brain vessel image segmentation task. We involve the MLP-Mixer to learn global information. (ii) To alleviate the memory issue and retain 3D spatial information, we split 3D volumetric images into a sequence of 2D plane images by three different planes and then feed them into three different 2D MLP-based GANs. (iii) We redesign the patterns of skip connection. Each layer in the decoder can learn more semantic information since we utilize the MLP-Mixer layer to mix deeper level layers’ and current level layer’s information from the encoder. (iv) We conduct extensive experiments on a public brain vessels dataset\textsuperscript{2}. Both qualitative and quantitative The results demonstrate that the proposed MLP-GAN achieves new state-of-the-art results.

2. RELATED WORK

GANs for Medical Images Segmentation Generative Adversarial Networks (GANs)\textsuperscript{6} have been widely and suc-
Fig. 1. (a) The proposed MLP-GAN network. (b) One MLP-Mixer enhanced generator. (c) the redesigned MLP-Mixer block

cessfully used for image-to-image translation tasks [7, 8, 9], which translate one domain of images to another domain by learning how to transform the distributions of the two classes. For instance, Isola et al. [7] utilized a conditional GAN to replace the standard noise vector with a source image as input to learn the mapping from input images to target images. Recently, GANs also achieved state-of-the-art results in the medical image segmentation task [10, 11, 12, 13]. For instance, Mondal et al. [13] introduced a novel method based on GANs to train a segmentation model with both labeled and unlabeled images. Different from these methods, in this paper, we propose an MPL-based GAN to tackle the brain vessel image segmentation task. To this end, we present a novel MLP-GAN, which splits a 3D volumetric brain vessel image into three different dimensional 2D images and then feeds them into three different 2D cGANs to obtain the final segmentation. Meanwhile, the proposed MLP-GAN involves the recent MLP-Mixer into skip connections to mix token and channel information to learn global information.

MLP-Mixer in Vision Models

MLP-Mixer [1] has been proved that simply multi-layered perceptrons (MLPs) can achieve high-quality vision results by two types of MLP layers: channel-mixing MLPs and token-mixing MLPs. The architecture of MLP-Mixers accepts a sequence of patches — a "patches × channels" table, generated by rearrangement and linear projection. The channel-mixing MLPs build relationships among different channels by operating on every token independently. The token-mixing MLPs build relationships among different spatial locations with tokens by operating on every channel independently. Recently, the MLP-Mixer achieved state-of-the-art performance in vision tasks such as image generation [14, 15], image classification [16, 17]. For instance, Res-MLP [16] proposed an affine transform layer so that it can achieve a deeper architecture and higher accuracy than the MLP-Mixer.

3. THE PROPOSED MLP-GAN

3.1. Overview

The overview of our proposed MLP-GAN is illustrated in Figure 1 (a). We first split the 3D volumetric image into a sequence of 2D plane images in terms of three dimensions (i.e., coronal, sagittal, axial), respectively. Then, the sequence of 2D plane images is fed into the generator, which consists of three 2D networks with U-Net as the backbone. The generator outputs three 2D plane images with the same shape as the original input images. Next, these generated images are fed into the discriminator along with ground truth segmentation to discriminate whether they are true or false.

3.2. Memory-Friendly Isotropic 2D Input

Most 3D medical models require a 3D stack of images as input, which brings a large burden on hardware. Meanwhile, a medical image typically has a much high resolution. To solve this problem, we decompose the original 3D stack into three separate 2D plane images. For each iteration of training, we only feed three 2D orthogonal plane images into our network. In practice, the shape of our original brain vessel data and ground truth are (325, 304, 600). We crop the original data into several 3D volumetric images shaped (256, 256, 256). Then we split each 3D sample data into a sequence of 2D plane images by the same index on three different dimensions ranging [0, 255] to feed into three generators.

3.3. MLP-Mixer Enhanced Generator

The generator consists of three individual 2D U-Net-based networks, where each 2D network has identical layouts. Figure 1 (b) illustrates a single 2D U-Net-based architecture. We redesign the patterns of skip connection in order to maximize their benefits. Each level in the decoder not only learns information from the previous level in the decoder but also learns information from the skip connection that consists of several MLP-Mixer blocks receiving information from the same level and the deeper level in the encoder. Then we concatenate these information to mix via channel-mixing and token-mixing MLPs to learn more semantic information.

In the 2D U-Net-based generator, we integrate the MLP-Mixers into our generator as shown in Figure 1 (b). We define that the U-Net has 5 levels. For $i$-th level, we define there are $(5-i)$ blocks. The function of each block can be expressed as the equation below,

$$
\mathcal{H}^{i,j}(X) = \begin{cases} 
\text{CNN}^{i,j}(X) & \text{if } j = 0, i + j = 5 \\
\text{MLP}^{i,j}(X) & \text{otherwise}
\end{cases}
$$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
where \( i \) denotes the \( i \)-th layer, and \( j \) denotes the \( j \)-th block in the same level, \( \text{CNN}(\cdot) \) denotes convolution layer function, and \( \text{MLP}(\cdot) \) denotes the MLP-Mixer function. More details for the MLP-Mixer function are shown at Section 3.4 MLP-Mixer block.

The function of the output of each block can be expressed as the following equation:

\[
\mathcal{O}^{i,j}(X) =
\begin{cases} 
\mathcal{H}^{i,j}(\mathcal{I}) & \text{if } j = 0 \\
\mathcal{H}^{i,j}(\mathcal{O}^{i-1,j}(X), \ldots, \mathcal{O}^{i,j-1}(X), \mathcal{O}^{i+1,j-1}(X)) & \text{otherwise,}
\end{cases}
\]

(2)

where \( \mathcal{I} \) is the input images for the first block in each level. In this equation, when \( j = 0 \), the block is the convolution layer, which only receives one input, called \( \mathcal{I} \). The rest of the MLP-Mixers and CNNs will receive several inputs that contain all output of every left block \( (\mathcal{O}^{i,0}(X), \ldots, \mathcal{O}^{i,j-1}(X)) \) in the same level and the left one blocks in bottom one layer \( (\mathcal{O}^{i+1,j-1}(X)) \). All inputs will be concatenated and fed into the MLP-Mixer. In this way, since the pattern of skip connections can mix the spatial information between the current level and bottom level, which can pass more information to deeper layers, we can achieve better performance.

### 3.4. MLP-Mixer Block

Figure 1 (c) illustrates the details of a modified MLP-Mixer block, which takes \( X \in \mathbb{R}^{H \times W \times C} \) with a spatial resolution of \( H \times W \) and \( C \) number of channels as an input. The MLP-Mixer firstly rearranges the original image to several non-overlapping patches and then perform a linear projection to get the shape “patches x channels”. Specifically, the output of the rearrangement denotes \( X_{\text{reshape}} \in \mathbb{R}^{(\frac{H}{S} \times \frac{W}{S}) \times (C\times S \times S)} \) where \( S \) denotes the size of patches. The “patches” and “channels” are equal to \( \frac{H}{S} \times \frac{W}{S} \) and \( C \times S \times S \), respectively.

In order to fit it in our network, we fix the number of neurons of the linear projection as the same as the last dimension of the shape after rearrangement, which can make the shape after rearrangement equalized to the shape after the linear projection. Specifically, the linear project denotes \( W \in \mathbb{R}^{(C\times S\times S) \times (C\times S \times S)} \). The output of the linear project can be expressed as,

\[
X_o = X_{\text{reshape}} W; X_o \in \mathbb{R}^{(\frac{H}{S} \times \frac{W}{S}) \times (C\times S \times S)}
\]

(3)

After the linear projection, it connects two types MLPs with a LayerNorm layerization: token-MLP and channel-MLP, which mix token information and channel information, respectively. There is a matrix transpose between the token-MLP and the channel-MLP for the last two dimensions. For each MLP block, it comprises two fully-connected layers and a GELU activation in the middle. The MLP-Mixer block on MLP(\cdot) is as follows,

\[
\begin{align*}
U_{*,n} &= X_{*,n} + W_2 \sigma(W_1 \text{LayerNorm}(X_o)_{*,n}) \\
\text{MLP}_{m,*} &= X_{m,*} + W_4 \sigma(W_3 \text{LayerNorm}(X_o)_{m,*})
\end{align*}
\]

(4)

where \( n=1 \cdots C \), and \( m=1 \cdots S \), \( \sigma \) is an element-wise nonlinearity (GELU [18]). \( W_1 \) and \( W_2 \) denote two fully-connected layers in the token-MLP module. \( W_3 \) and \( W_4 \) denote two fully-connected layers in the channel-MLP module. The token-MLP module can capture token information that contains spatial information. The channel-MLP module has the ability to capture the channel information. Finally, the final output will be reshaped into the same shape as the input at the very beginning of the MLP-Mixer block.

### 3.5. Multi-Plane Discriminator

The discriminator also consists of three different 2D-base convolution layers’ networks. Each 2D-based convolution layer’s network takes generated images and ground truth as input. And then, their concatenation is fed into 4 times down-sample layers that consist of \( \text{Conv2D} \times 3 \times 3 \), BN, and LeakyReLU. The shape of the output is \( (B, 32, 32, 64) \), which can achieve the functionality of PatchGAN [7].

### 3.6. Overall Optimization Goal

The objective of our proposed MLP-GAN can be expressed as \( \mathcal{L}_{\text{cGAN}}^{i} = \)

\[
\mathbb{E}_{x^{i},y^{i}}[\log D(x^{i},y^{i})] + \mathbb{E}_{x^{i},z}[1 - \log D(x^{i},G(x^{i},z))],
\]

(5)

where \( i \) is sagittal, coronal, axial.
where $G$ means the generator, trying to minimize this objective against an adversarial discriminator $D$ which tries to maximize this objective. This objective consists of three parts which are for sagittal, coronal, and axial, respectively.

Meanwhile, we import the patch-wise balanced binary cross-entropy to endow the generator abilities near the ground truth output in a binary cross-entropy sense. We utilize $L_{bBCE}^i$ as the extra loss function expressed:

$$\frac{1}{N} \sum_{i=0}^{N} \beta(y^i_j \log(\hat{y}^i_j) + (1 - \beta)(1 - y^i_j) \log(1 - \hat{y}^i_j)),
\quad i = \text{sagittal, coronal, axial}$$

where $\beta = |Y_i|/|Y|$ and $1 - \beta = |Y_i^+|/|Y|$, $|Y|$ and $|Y_i^+|$ denote the mask and non-mask in ground truth sets, respectively. Because the number of mask pixels and non-mask pixels are extremely imbalanced. The balanced binary cross-entropy $L_{bBCE}^i$ can balance the ratio of mask and non-mask. The loss function also consists of three parts that are for sagittal, coronal, and axial, respectively.

Our final objective function can be expressed as:

$$G = \arg \min_G \max_D \sum_i \frac{L_{GAN}^i}{3} + \lambda \times \sum_i \frac{L_{bBCE}^i}{3},
\quad i = \text{sagittal, coronal, axial}$$

where the final objective function consists of conditional GAN loss and balanced binary cross-entropy loss, where two losses average the effect of three different dimensions. Meanwhile, the $\lambda$ coefficient is introduced to balance between $L_{GAN}^i$ and $L_{bBCE}^i$.

4. EXPERIMENTS

4.1. Dataset and Evaluation Metrics

We use the public DeepVesselNet dataset [2] in our experiments. This dataset contains 136 images of brain vessels with shape $(325, 304, 600)$. The dataset is split into 96 training samples, 20 validation samples, and 20 test samples. We employ the Dice coefficient and Intersection over Union (IoU) as our evaluation metrics to measure the performance.

4.2. State-of-the-Art Comparisons

We compare our proposed approach with several leading convolutional-based methods (i.e., 3D-U-Net [19] and 2.5DNet [20]) and several GANs with different generators as the baseline models that include i) 3D-U-Net-GAN; utilizes an original 3D U-Net [19] as the generator, ii) 3D-MLP-TransUnet-GAN [1]: replaces the self-attention block in the TransUnet [21] with the MLP-Mixer for 3D version, iii) 2.5D-U-Net-GAN; utilizes three 2D U-Nets [20] for three different dimensional 2D images as the 2D generator, and iv) 2.5D-MLP-TransUNet-GAN [1]: 2.5D version of method ii). In this part, we crop the original data into several (128, 128, 128) tensors as input.

Figure 2 illustrates the qualitative results of our method compared with other methods. As we can observe, the masks of the generated 2D plane images of the other methods are blurred, and there are some grey points in several masks. Meanwhile, the edges of some masks in generated 2D plane images have tiny trimmings. Compared to the other methods, our approach has better performance in the edges and masks.

The quantitative comparison with state-of-the-art results is illustrated in Table 1. For all metrics, higher means better. The number of Dice and IoU for our model is the highest. It demonstrates our proposed method achieves the best performance in all metrics, which validates the effectiveness of our proposed model. The results demonstrate our multi-view approach outperforms full 3D networks. In this way, our model not only retain 3D spatial information but also alleviate the memory issue compared to the 3D methods.

5. CONCLUSIONS

In this paper, we propose a novel multi-viewed MLP-GAN that decomposes 3D volumetric images into a sequence of 2D plane images in terms of three dimensions. And then feed the sequence of 2D plane images into three different 2D U-Net-based networks as our generator, which not only decreases the complexity and cost but also retains the spatial information. In each 2D generator, we introduce a redesigned pattern of skip connection with the integration of the MLP-Mixer block to obtain the ability to capture global information. In this way, each layer in the decoder can learn more semantic information since MLP-Mixer blocks can mix current and deeper level layers’ information from the encoder to learn overall information, which can maximize the benefits of skip connections. The experimental results demonstrate that our method outperforms the state-of-the-art approaches.

Acknowledgments: Y.Y. and D.C. received support by National Institutes of Health (NIH) 1RF1MH124611, 1RF1MH123402. D.C. also received support by National Science Foundation NeuroNex-1707316.

| Method                        | Dice ↑ | IoU ↑ |
|-------------------------------|--------|-------|
| 3D-U-Net [19]                | 0.894  | 0.946 |
| VNet [22]                    | 0.921  | 0.958 |
| 2.5D-U-Net [20]              | 0.891  | 0.935 |
| 3D-U-Net-GAN [19]            | 0.861  | 0.927 |
| 3D-MLP-TransUnet-GAN [21]    | 0.981  | 0.964 |
| 2.5D-U-Net-GAN [20]          | 0.986  | 0.971 |
| 2.5D-MLP-TransUnet-GAN [21]  | 0.982  | 0.965 |
| MLP-GAN (Ours)               | 0.992  | 0.979 |
6. REFERENCES

[1] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, et al., “Mlp-mixer: An all-mlp architecture for vision,” arXiv:2105.01601, 2021.

[2] Giles Tetteh, Velizar Efremov, Nils D Forkert, Matthias Schneider, Jan Kirschke, Bruno Weber, Claus Zimmer, Marie Piraud, and Bjoern H Menze, “Deepvesselnet: Vessel segmentation, centerline prediction, and bifurcation detection in 3-d angiographic volumes,” Frontiers in Neuroscience, vol. 14, 2020.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in MICCAI, 2015.

[4] Zeynettin Akkus, Alfiia Galimzianova, Assaf Hoogi, Daniel L Rubin, and Bradley J Erickson, “Deep learning for brain mri segmentation: state of the art and future directions,” Journal of digital imaging, vol. 30, no. 4, pp. 449–459, 2017.

[5] MR Avendi, Arash Kheradvar, and Hamid Jafarkhani, “A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri,” Medical image analysis, vol. 30, pp. 108–119, 2016.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial nets,” in NeurIPS, 2014.

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, “Image-to-image translation with conditional adversarial networks,” in CVPR, 2017.

[8] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan, “Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation,” in CVPR, 2019.

[9] Mehdi Mirza and Simon Osindero, “Conditional generative adversarial nets,” arXiv:1411.1784, 2014.

[10] Yuan Xue, Tao Xu, Han Zhang, L Rodney Long, and Xiaolei Huang, “Segan: Adversarial network with multi-scale 1 l loss for medical image segmentation.” Neuroinformatics, vol. 16, no. 3, pp. 383–392, 2018.

[11] Pim Moeskops, Mitko Veta, Maxime W Lafarge, Koen AJ Eppenhof, and Josien PW Pluim, “Adversarial training and dilated convolutions for brain mri segmentation,” in Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 56–64. Springer, 2017.

[12] Jaemin Son, Sang Jun Park, and Kyu-Hwan Jung, “Retinal vessel segmentation in fundoscopic images with generative adversarial networks,” arXiv:1706.09318, 2017.

[13] Arnab Kumar Mondal, Jose Dolz, and Christian Desrosiers, “Few-shot 3d multi-modal medical image segmentation using generative adversarial learning,” arXiv:1810.12241, 2018.

[14] George Cazenave and Manuel Ladron De Guevara, “Mixergan: An mlp-based architecture for unpaired image-to-image translation,” arXiv:2105.14110, 2021.

[15] Bin Ren, Hao Tang, and Nicu Sebe, “Cascaded cross mlp-mixer gans for cross-view image translation,” in BMVC, 2021.

[16] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou, “Resmlp: Feedforward networks for image classification with data-efficient training,” arXiv:2105.03404, 2021.

[17] Tan Yu, Xu Li, Yunfeng Cai, Mingming Sun, and Ping Li, “S2-mlp: Spatial-shift mlp architecture for vision,” arXiv:2106.07477, 2021.

[18] Dan Hendrycks and Kevin Gimpel, “Gaussian error linear units (gelus),” arXiv:1606.08415, 2016.

[19] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger, “3d u-net: learning dense volumetric segmentation from sparse annotation,” in MICCAI, 2016.

[20] Christoph Angermann, Markus Haltmeier, Ruth Steiger, Sergiy Pereverzyev, and Elke Gizewski, “Projection-based 2.5 d u-net architecture for fast volumetric segmentation,” in 2019 13th International conference on Sampling Theory and Applications (SampTA). IEEE, 2019, pp. 1–5.

[21] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou, “Transunet: Transformers make strong encoders for medical image segmentation,” arXiv:2102.04306, 2021.

[22] Fausto Milletari, Nassir Navab, and Seyed-Ahmad 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.