

# SC-UneXt: Nested UNeXt Architecture based on Medical Image Segmentation

Lei Wen

SouthWest Minzu University, Chengdu, China

---

**Abstract:** UNet and its various variants are commonly used methods in medical image segmentation tasks; however, many network parameters, complex calculations, and slow usage are problems that need to be overcome. These problems hinder the specific application of fast image segmentation in real-time tasks. At the same time, the lesion area has problems such as small size, irregular shape, and blurred edges, which makes the network feature extraction difficult and the segmentation accuracy needs to be improved. At the same time, medical image segmentation provides a variety of effective methods for the accuracy and robustness of organ segmentation, lesion detection, and classification. Medical images have fixed structures, simple semantics, and diverse details, so integrating rich multi-scale features can improve segmentation accuracy. Given that the density of diseased tissue may be comparable to that of surrounding normal tissue, both global and local information are crucial to segmentation results. To this end, we propose an image segmentation method (SC -UNe X t) based on edge feature extraction and multi-scale feature fusion of convolutional multi-layer perceptron (MLP). The network is a deeply supervised encoder-decoder network, in which the encoder and decoder pass through a series of nested, multiple jump paths to reduce the semantic gap between the feature maps of the encoder and decoder sub-networks.; Multi - scale feature fusion is introduced based on the UNe Finally, we evaluate our model approach on the LIDC dataset public dataset. Experiments have proven the effectiveness of this method. Our model's similarity coefficient and intersection ratio reached 86.44% and 90.86% respectively. Compared with UNet and UNe X t, the network proposed in this article has improved in accuracy, intersection ratio of real values and predicted values, similarity coefficient, and segmentation effect.

**Keywords:** Lightweight; Edge Feature Fusion; Multi-branch Feature Fusion; Medical Image Segmentation.

---

## 1. Introduction

Skin diseases often occur in daily life and cause great trouble to patients. Doctors need to diagnose the diseased areas during their work to help patients recover as soon as possible. At the same time, lung inflammation is a serious disease that may be caused by long-term living habits and genetic mutations. Lung lesions initially manifest as inflammation, but over time, the accumulation of many diseases in the lungs may manifest as inflammation, so the initial diagnosis of pulmonary nodules is very important. When diagnosing pulmonary nodules in the early stage, the large number of patients usually leads to excessive fatigue of doctors. Therefore, doctors are required to repeatedly check CT images to better determine the area of pulmonary nodules and ensure that the diagnosis results are correct. In recent years, with the continuous development of deep learning, this technology has been widely used in clinical diagnosis to assist doctors in quickly detecting CT images and reducing misdiagnosis rates. However, due to the small size and irregular shape of the lesion area, traditional methods have certain limitations when segmenting images. They may not be able to accurately locate the nodule position, resulting in low model segmentation accuracy, and may even increase with the network model layer. Deepening loses some information, so that more features cannot be effectively obtained.

To this end, this paper proposes a lesion image segmentation method based on UNe X t combined with multi-scale feature fusion and attention mechanism. First, we introduce a multi - scale feature fusion mechanism based on the UNe An attention mechanism is introduced based on X t; finally, we use a joint loss function to improve the sensitivity of the network model to edge features.

## 2. Related Work

Previous researchers have tried to build an extremely accurate, effective, and automatic lesion area segmentation system that can help doctors detect lesion areas. These attempts are divided into two broad categories: image processing -based models and deep learning-based models. Image processing models including morphological operations, region growing algorithms and energy optimization techniques are the most commonly used methods. In the morphology-based approach, researchers used morphological opening operations and connection component selection methods to remove vessels attached to nodules. However, it is challenging to use fixed-size morphological templates to isolate lesions that have extensive contact areas with other lesion regions. As a result, more complex morphological processes incorporating shape assumptions are introduced. Kuhnigk et al. 8 found that vessel radius decreased as it evolved toward the lung periphery. Furthermore, they recommended the combined use of rolling ball filters and rule-based pleural nodule analysis. Selection of morphological template size is a significant challenge for morphological methods because it is difficult to identify appropriate templates for morphology of different nodule sizes. Zhou et al12 introduced a fully automated lung segmentation method for parapleural nodules. Nonlinear anisotropic diffusion filtering method is used to reduce image noise. The chest region is extracted using thresholding, 2D hole filling and maximum connected component search methods. Lung parenchyma was separated using fuzzy c-means algorithm, region growing algorithm and dynamic programming method. Abbas et al21 proposed a method to segment pulmonary nodules in CT images. Preprocessing

techniques such as unsharp energy masking and discrete wavelet transform have been used to enhance images. The proposed system has an area overlap measure (AOM) of 95%, a combined equal importance (CEI) of 92%, a Hausdorff distance (HDD) of 91%, and a Hamoud distance (HMD) of 87%. Long et al.22) developed a fully convolutional neural network (FCN) for semantic segmentation. The fully connected layers of the CNN are removed and deconvolutional layers are introduced to make the output dimensions the same as those of the input image. Skip connections are introduced, which combine coarse and fine layers to make dense predictions possible. Ronneberger et al.23 modified and extended FCN and designed a new U Net architecture for biomedical segmentation -. UNet is unique in that the network can perform accurate segmentation with fewer training images.

UNeXt is a paper released by Johns Hopkins University in 2022. It uses convolutions in the early stages and MLP in the latent space stage. The convolutional features are labeled and projected through a tokenized MLP block, and the representation is modeled using the MLP. UNeXt is an important improvement direction for lightweight segmentation. The network adopts overlap-tile strategy, data enhancement, and weighted loss technology, which can be applied to real-time fast image segmentation with simple calculation and fast speed.

The most obvious difference between UNet and UNe to further improve performance, we recommend shifting the input Channel when inputting mlp in order to focus on learning local dependencies. Using tokenized MLP in the latent space reduces the number of parameters and computational complexity while producing better representations to aid segmentation. The network also includes skip connections between encoders and decoders at all levels. Test results show that compared with the current state-of-the-art medical image segmentation architecture, UNeXt reduces the number of parameters by 72x, reduces computational complexity by 68x, increases inference speed by 10x, and also achieves better segmentation performance. Experiments on different data sets have proven that the segmentation effect of the UneXt network needs to be further improved to meet the requirements of instant diagnosis, accuracy and efficiency of medical pathology images.

Although the above methods have improved traditional deep learning methods, satisfactory results cannot be obtained by improving only one method or targeting a certain structure of the model. To this end, this paper proposes an image segmentation method (SC-UNeXt) based on UNeXt that combines edge detection, feature fusion and attention mechanisms to improve the segmentation of lesion areas.

### 3. Method

#### 3.1. SC-UNeXt Network Architecture

In the research, in order to better realize the accurate real-time segmentation task and avoid the problem of losing local low-dimensional features through direct large-area upsampling, which leads to the loss of too many features on the segmentation boundary and the inability to restore complete edge information, only through features The graph superimposes feature information in the channel dimension to retain feature information. This will cause the last few layers of feature maps to be too bloated, causing the model to require a large amount of calculations. Based on these problems, we

adopt a multi-branch feature fusion network for medical image segmentation. We first propagate contextual information to higher-resolution layers through progressive upsampling to obtain preliminary low-level semantic features. We avoid superimposing feature information in the channel dimension of UNet's series of related models, and choose the method of feature map multiplication to fuse features ; therefore, most of the feature information is well preserved, and boundary information can be effectively obtained , effectively reducing the number of failures. The designed skip link uses more detailed low-dimensional feature information as a supplement to feature fusion to ensure that the accuracy is slightly better than UNet (Zhou et al., 2018, 2020), ResUNet++ (Jha et al., 2019) and other networks run much faster than other models; it also has the advantages of high training efficiency and strong generalization ability.

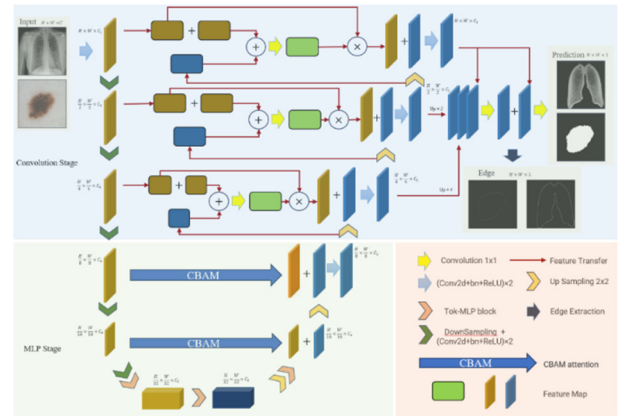


Figure 1. SC-UNeXt structure

Overall framework Our network consists of three parts: encoder, multi-scale cross-skip connection and decoder. In order to better integrate semantic and scale- inconsistent features and further improve the segmentation effect, we propose a cross-joint attention-guided Multi-scale fusion scheme, which solves the problems that arise when fusing features of different scales.

#### 3.2. Convolutional Attention Modules

Convolutional Block Attention Module (CBAM) consists of channel attention and spatial attention. Among them, channel attention strengthens the connection between features in different channels by using maximum pooling and average pooling to pay attention to the information between different channels; spatial attention improves the spatial connection of the network by paying attention to space. Combining the channel and spatial attention mechanisms, adaptive feature extraction can be achieved. The CBAM model structure is shown in Figure 2.

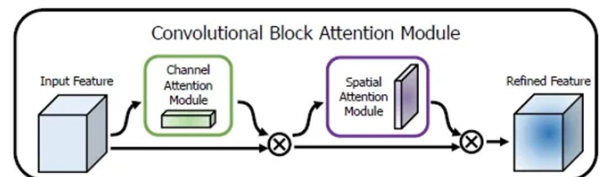


Figure 2. CBAM module

CBAM can be expressed by the following formula:

$$F' = M_C(F) \otimes F \tag{1}$$

$$F'' = M_S(F') \otimes F' \tag{2}$$

Among them,  $F$  is the feature map,  $M_c$  and  $M_s$  represents channel-based and spatial-based attention respectively,  $\otimes$  represents element-wise multiplication,  $F'$  and  $F''$  represents the output feature map after channel attention and spatial attention respectively. Since the input and output sizes of the CBAM module are the same, it can be inserted anywhere in an existing model.

### 3.3. SC Jump Link Module

segmentation network contain more fine-grained information, which facilitates the segmentation of small lesions. Deep segmentation networks can extract more high-level semantic information, thereby improving segmentation accuracy. In addition, rich multi-scale information integrates the characteristics of different receptive fields, which is beneficial to the segmentation of multi-lesion areas. The skip connection is redesigned to aggregate features of different semantic scales on the decoder subnetwork to form a highly flexible feature fusion scheme. The SC skip connection is an operation that connects simple and effective deep and shallow information fusion. In SC - UNeXt; Jump link takes the first layer as an example:

$$z_1 = f([(2x, +z_2) \otimes x_1] + z_2) \quad (3)$$

$Z_3$  feature information fed back by the jump link is fused with the  $Z_4$  advanced semantic features obtained by the convolution operation and the upsampling operation, and then the upsampling operation is performed.

### 3.4. PFM Pyramid Feature Fusion Module

In order to make the network capable of multi-scale detection, the article uses deconvolution to expand feature layers at different levels to the same size, and then adds them at the element level. The fused feature layer has richer multi-scale features.

### 3.5. Joint Loss Function

The joint loss function is composed of a mean square error loss function and an edge loss function. The complex, diverse and unclear edge contours of lesions have a great impact on the segmentation effect of the lesion area. Therefore, increasing the weight of edge feature information samples is of great significance to the segmentation effect, for this we use the edge loss function. The joint loss function is defined as:

Mean squared error loss function (MSE):

$$L(Y|f(x)) = \frac{1}{n} \sum_{i=1}^n (Y_i - f(x_i))^2 \quad (4)$$

Edge feature loss function Edge loss:

$$loss_{edge} = \frac{\sum_{x=1}^W \sum_{y=1}^H E_{i,j} \cdot (|Y_{i,j} - X_{i,j}|)}{WH} \quad (5)$$

Joint loss function Loss (total) :

$$Loss(total) = Loss(MSE) + Loss(edge) \quad (6)$$

Increasing the edge feature semantics in the sample weight in the joint loss function can better optimize edge information.

## 4. Experimental Settings

### 4.1. Dataset

in the experiment. The NIH Chest X-ray Dataset contains 112,120 disease-labeled X-ray images from 30,805 unique patients. There are 15 categories (14 diseases and "none found"). Images can be grouped into "no findings" or one or more disease categories, showing 14 common chest pathologies. The NIH chest X-ray dataset itself does not contain lung field labels. We randomly selected 2785 samples,

and doctors marked the lung fields of the images. We call this new dataset Haut. The TeHaut dataset contains some severely blurred, occluded, and distorted chest X- rays. The Haut data set contains 1647 normal people and 1138 CXR lung field mask patients, including 193 cases of infiltration, 111 cases of atelectasis, 78 cases of effusion, 65 cases of nodules, 54 cases of masses, 43 cases of pneumothorax, and heart disease. There were 37 cases of enlargement, 37 cases of pleural thickening, 34 cases of pleural thickening with fibrosis, 25 with consolidation, 21 with emphysema, 11 with edema, 10 with pneumonia, 2 417 had a hernia and 417 suffered from multiple medical conditions (including any two or more of the above). To use Efcientnet-b4, as a preprocessing step, the image is downsized to 256×256 pixels.

To bring our experiments as close as possible to point-of-care imaging, we chose the International Skin Imaging Collaboration (ISIC 2018) to benchmark our results. The ISIC dataset contains camera-acquired skin images and corresponding skin lesion region segmentation maps. The ISIC 2018 dataset consists of 2594 images. We resized all images to a resolution of 512 × 512.

## 4.2. Experimental Settings

We use the Adam optimizer with a learning rate of 0.0001 and a momentum of 0.9. We also used a cosine annealing learning rate scheduler with a minimum learning rate up to 0.00001. The batch size is set to 8. We trained UNeXt for a total of 400 epochs. We perform an 80-20 random split three times on the dataset and report the mean and variance. Evaluation indicators, we use IoU, Dice Segmentation index to quantify the segmentation ability of SC-UneXt, Dice Similarity Coefficient (DSC), Dice coefficient is a set similarity measure, I O U is used to evaluate the degree of similarity between predictions and true values. Semantic segmentation can be viewed as pixel-level classification. True Positive (TP): The model prediction is a positive example, that is, a positive example. False positive (FP): The model predicts a positive example, but it is a negative example. False Negative (FN): The model prediction is a negative example, but it is a positive example. True Negative (TN): The model prediction is a counterexample, it is a counterexample.

$$Dice = \frac{2TP}{2TP+FP+FN} \quad (7)$$

$$IOU = \frac{TP}{TP+TN+FP+FN} \quad (8)$$

Dice is usually used to calculate the similarity of two samples. The value range is 0 - 1. The best segmentation result is 1 and the worst value is 0. IOU is calculated as the ratio of the intersection and union of the two sets of real values and predicted values. The larger the ratio, the higher the similarity between the real value and the predicted value, the better the segmentation effect.

## 5. Experiment

### 5.1. Experimental Results

The network frameworks used in our comparative experiments include the most advanced CNN-based networks, such as U-Net, U-NeXt, and SC-UneXt. Below we will quantitatively and qualitatively compare the test results. analyze. Furthermore, the number of parameters in each network is kept to two decimal places. Evaluation of dermatology datasets. Quantitative result analysis, the quantitative comparison results of the data set using different methods are shown in

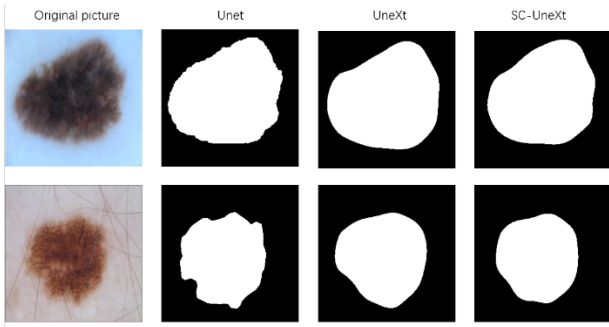


Figure 3. Skin disease segmentation effect

The following table SC-UnexT is the segmentation evaluation indicators Dice, Iou, parameter amount and computing power consumption:

Table 1. Segmentation evaluation index

Method	FLOPS	Parameters	Dice	IOU
U-Net	25.06 GFLOPs	9.04 MB	84.84%	76.37%
ResUNet++	19.31GFLOPs	7.67 MB	85.73%	77.12%
UNext	0.10 GFLOPs	247.62 KB	86.64%	78.71%
SC-UNexT	0.13 GFLOPs	243.63 KB	87.12%	79.13%

We noticed that the IoU and Dice of SC -UNexT are 2.74% and 2.76% higher than U-Net respectively, 0.48% and 0.42%

Table 2. Ablation experiment

Model	DSC evaluation index		ISIC 207 data set Val :24		Data volume train: 9 6		Val :24
	Fold-1	F old-2	F old -3	F old -4	F old -5	mean	variance
5 fold cross validation							
UNexT	82.61%	84.66%	91.47%	87.61%	83.08%	85.89%	0.14%
SC_UNexT( edge_loss )	85.10%	85.53%	88.51%	86.69%	86.84%	86.14%	0.03%
SC_UNexT (skip)	85.73%	85.90%	90.68%	82.46%	83.75%	86.53%	0.13%
SC_UNexT (fuse)	87.17%	85.07%	90.58%	85.59%	84.92%	86.17%	0.10%
SC UNexT(edge loss+skip)	85.90%	87.98%	90.22%	85.79%	89.14%	86.90%	0.05%
SC UNexT (edge loss+fuse)	87.98%	80.14%	90.50%	86.48%	88.03%	86.98%	0.20%
SC_UNexT (skip+fuse)	82.90%	88.62%	88.60%	84.90%	87.30%	85.90%	0.07%
SC_UNexT(edge_loss+skip+fuse)	85.89%	88.11%	89.02%	86.11 %	87.22%	87.29%	0.08%

Through the comparison of ablation experiments, it was found that jump links, feature fusion, and joint loss functions all have an impact on segmentation accuracy. Only under the combined effect of S C-UNexT at the same time, the DSC evaluation index is the highest and the segmentation effect is good.

## 6. Discussion

In this article, our purpose is to make the network better learn effective features and obtain more accurate lesion segmentation results. We propose an improvement that uses feature fusion and increases the proportion of edge features in the total sample features. network architecture. This network is not only better than other methods in terms of evaluation indicators, but is also cost-effective enough to better help

higher than U- NeXt respectively, and 1.39% and 2.01% higher than ResUNet++. We noticed, compared with U-Net's 9.04 M parameters, SC -UNexT's 243.63 KB parameters are also relatively low, roughly the same as UNext. While retaining the lightweight, simple calculation, and fast characteristics of unext, the segmentation effect has been significantly improved from the dice coefficient and IOU coefficient. Regarding the computational power consumption of the model, that is, in terms of flops, it consumes less than Unet, which ensures the lightweight of the model and saves the consumption of computing power.

## 5.2. Ablation Experiment

In actual situations, if a model is to be put online, the model needs to be repeatedly debugged to prevent the model from performing better only on known data sets and performing poorly on unknown data sets. That is to ensure the generalization ability of the model, which refers to the adaptability of machine learning to fresh samples. Only by ensuring the generalization ability of the model can the construction of the model be meaningful. Therefore, cross-validation is particularly important throughout the modeling process. Use training set/test set splitting and cross-validation methods to avoid this situation. As shown in the figure below, split the data set into training set/test set, and perform cross-validation on the training set to obtain the best model parameters., thereby obtaining the score of the model on the test set.

This experiment adopts five-fold cross-validation and conducts cross-validation experiments on SC-UnexT in three directions: jump link, feature fusion, and joint loss function. The experimental results are as follows:

doctors better diagnose the details of these histological images. Future research topics in medical image segmentation will be deep learning to automatically select features from different resolutions, or consider using adversarial training including test images to exploit features in test images without annotations. It achieves high-precision segmentation while also achieving instant feedback and lightweight operations, which plays a role in assisting diagnosis and treatment in the medical process.

## References

- [1] O. C. Eidheim, L. Aurdal, T. Omholt-Jensen, T. Mala, and B. Edwin, "Segmentation of liver vessels as seen in MR and CT images," Int. Congr. Ser., vol. 1268, pp. 201–206, Jun. 2004. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0531513104006132>.

- [2] A. Bert et al., “An automatic method for colon segmentation in CT colonography,” *Computerized Med. Imag. Graph.*, vol. 33, no. 4, pp. 325–331, Jun. 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0895611109000226>
- [3] T. Klinder, J. Ostermann, M. Ehm, A. Franz, R. Kneser, and C. Lorenz, “Automated model-based vertebra detection, identification, and segmentation in CT images,” *Med. Image Anal.*, vol. 13, no. 3, pp. 471–482, Jun. 2009. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1361841509000085>
- [4] H. Lu, Y. Li, M. Chen, H. Kim, and S. Serikawa, “Brain intelligence: Go beyond artificial intelligence,” *Mobile Netw. Appl.*, vol. 23, no. 2, pp. 368–375, Apr. 2018. doi: 10.1007/s11036-017-0932-8.
- [5] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, “Deep features learning for medical image analysis with convolutional autoencoder neural network,” *IEEE Trans. Big Data*, to be published.
- [6] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, “HealthCPS: Healthcare cyber-physical system assisted by cloud and big data,” *IEEE Syst. J.*, vol. 11, no. 1, pp. 88–95, Mar. 2017.
- [7] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
- [8] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput-Assist. Intervent. (MICCAI)*, Nov. 2015, pp. 234–241.
- [9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
- [10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, Jun. 2012.
- [11] Jha, D., et al. Doubleu-net: A deep convolutional neural network for medical image segmentation. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)* (2020).
- [12] Liu, T. et al. Residual convolutional neural network for cardiac image segmentation and heart disease diagnosis. *IEEE Access*. 8, 82153–82161 (2020).
- [13] Wu, B., Fang, Y. & Lai, X. Left ventricle automatic segmentation in cardiac MRI using a combined CNN and U-net approach. *Comput. Med. Imaging Graph.* 82, 101719 (2020).
- [14] Galati, F. & Zuluaga, M. A. Efficient model monitoring for quality control in cardiac image segmentation. *FIMH 20*, 101–111 (2021).
- [15] Ronneberger, O., Fischer, P., & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI 2015. Part III* 18, 234–241 (2015).
- [16] Milletari, F., Navab, N., & Ahmadi, S. A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*. 565–571 (2016).
- [17] Zhou, Z., et al. UNet++: A nested U-Net architecture for medical image segmentation. *DLMIA/ML-CDS@MICCAI 2018*. 3–11 (2018).
- [18] Zhang, J., et al. MDU-Net: Multi-scale densely connected U-Net for biomedical image segmentation. *arXiv:1812.00352 (arXiv preprint)* (2018).
- [19] Hu, J., Shen, L., & Sun, G. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7132–7141 (2018).
- [20] Hou, Q., Zhou, D., & Feng, J. Coordinate attention for efficient mobile network design. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13708–13717 (2021).
- [21] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018).
- [22] Tu, Z., et al. Maxim: Multi-axis mlp for image processing[C]. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5769–5780 (2022).
- [23] Valanarasu, J. M. J. & Patel, V. M. UNeXt: MLP-based rapid medical image segmentation network. *MICCAI 5*, 23–33 (2022).
- [24] Tyagi, T., Gupta, P., & Singh, P. A hybrid multi-focus image fusion technique using SWT and PCA. In *2020 10th International Conference on Cloud Computing, Data Science and Engineering (Confluence)*. 491–497 (2020).
- [25] Ramlal, S. D., Sachdeva, J., Ahuja, C. K. & Khandelwal, N. An improved multimodal medical image fusion scheme based on hybrid combination of nonsubsampling contourlet transform and stationary wavelet transform. *Int. J. Imaging Syst. Technol.* 29(2), 146–160 (2019).
- [26] Joshi, K., Kirola, M., Chaudhary, S., Diwakar, M., & Joshi, N. K. Multi-focus image fusion using discrete wavelet transform method. In *International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India* (2019).
- [27] Mao, R., et al. Multi-directional laplacian pyramid image fusion algorithm[C]. In *2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*. IEEE, 2018. 568–572 (2018).
- [28] Long, J., Shelhamer, E., & Darrell, T. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3431–3440 (2015).