

CE-TransUnet: A Convolutional Enhanced Model for Pulmonary Alveolus Pathology Image Segmentation

Yongkun Chen^{*1}, Yu Qiu^{*}, Jierui Liu, Shiming Zha, Huayi He, Zheng Li^(✉2)

School of Computer Science and Technology, Sichuan University, Chengdu 610065, P.R. China

{2021141460040, rainqiu, liujierui,
2021141460148, hehuayi}@stu.scu.edu.cn
lizheng@scu.edu.cn

Abstract. Pulmonary alveolus segmentation plays an important role in the diagnosis of alveolar emphysema and lobar pneumonia. Besides, if the alveoli could be accurately segmented in pathological images, the area of tumor beds in non-small-cell lung cancers could be readily calculated and hence can help determine the severity of one's cancer. These factors render the segmentation of alveolar pathological images highly meaningful. However, we have not identified any existing publicly available dataset of alveolar pathological images and no existing methods focused on the segmentation of alveolus. Therefore, we first collected a substantial amount of data and curated it into a dataset called Pulmonary Alveolus Pathology Image(PAPI). Then, we adopt several traditional segmentation methods on PAPI and found them performing poorly either on the whole slide or on edge details. Therefore, we innovate our method Convolutional Enhanced Transformer-based U-net (abbreviated as CE-TransUnet), which is a combination of improved U-net structure and our innovative CE-Transformer block. We circumspectly detect salient characteristics of the pulmonary alveolus and make counterpart improvements in both CE-Transformer blocks and Unet structure. Our experimental results have shown that these adjustments has made our model surpass the current common segmentation models in performance on PAPI and reach a Dice score of 95.31. We are also exploring the robustness of our model to adapt it to a wider range of scenarios. Dataset(currently under the process of going public, will be published once the process is over) and code are available at: <https://github.com/DemonRain7/CE-TransUnet>.

Keywords: semantic segmentation, computer vision, pulmonary alveolus, medical digital pathology image, transformer.

1.1 Introduction

While modern medical imaging infrastructures provide efficient assistance for clinical medical diagnosis, the generation of a 'massive' amount of medical image resources still poses challenges to precise diagnosis. For example, the vast number of lung scans directly leads to an increase in the workload of medical professionals due to lack of

¹ * denotes that these authors contribute equally to this work.

² ✉ Corresponding Author.

help from current machines, thereby increasing the likelihood of missed diagnoses and misdiagnoses. Furthermore, detecting malignant tumors relies on pathological diagnosis, and traditional pathological diagnosis requires pathologists to analyze pathology slides one by one under a microscope, hence putting the diagnostic process under a certain degree of subjectivity. Eventually, the accuracy of diagnosis is directly related to the expertise of the pathologist, and a slight deviation in thought process may potentially lead to diagnostic errors.

In recent years, with the development of artificial intelligence and digital pathology, the significant potential of artificial intelligence in the diagnosis of various diseases has gradually emerged. For example, in the diagnosis of pulmonary diseases, cytopathology is a convenient and rapid method. It can be used for lung cancer screening or general examination. Artificial intelligence can integrate and analyze a large amount of information in a short time, effectively improving the efficiency of pulmonary disease diagnosis. Obviously, artificial intelligence has become a powerful auxiliary diagnostic tool for pathology experts.

The accurate identification and segmentation of pulmonary alveolus (abbreviated as alveolus) play a crucial role in assisting the diagnosis and treatment of common lung diseases. For instance, calculating the size and area of alveolus after segmentation can help determine the presence of alveolar emphysema [1]. Simultaneously, preliminary diagnosis based on the morphology and distribution of alveolus can be made for specific pathological subtypes, such as lobar pneumonia [1]. And the observation of alveolus also proves beneficial in diagnosing rare diseases, like bronchopulmonary isolation in children, where assessing characteristics such as alveolar enlargement or elongation serves as one of the criteria.

Furthermore, the accurate identification of alveolus also provides a novel approach to identifying the tumor bed area in lung cancer. Traditionally, the tumor bed comprises interstitial reactions involving tumor cells, fibrosis, necrosis, and inflammation. However, due to the incomplete nature of lung sections in clinical settings, there may be residual normal lung tissue. Given that normal lung tissue is predominantly composed of alveolus, a precise alveolus identification allows the exact extraction of normal lung tissue areas in sections, which in turn helps determine the remaining tumor bed area [1].

However, no previous medical concerned model has focused on the segmentation of alveolus. Besides, we found that no existing widely-used medical segmentation models can reach the optimal level when identifying alveolus. Therefore, we thoroughly studied the characteristics of alveolar pathology images and selected appropriate modules based on those features to enhance their recognition capabilities.

Research Contribution. Due to the relative poor segmentation capabilities on alveolus-related tasks of existing models in medical image segmentation and inspiration from a previous work [2], we have developed a segmentation network called Convolutional Enhanced Transformer-based Unet (abbreviated as CE-TransUnet), which can capture more of the features from alveolus. Our experimental results demonstrate that CE-TransUnet excels in alveolar segmentation tasks with improved segmentation accuracy and training efficiency compared to state-of-the-arts like nnUnet [3] and SwinUnet [4]. We believe that CE-TransUnet will bring thus new breakthroughs to alveolar

pathology image segmentation tasks as a powerful tool for medical diagnosis and pathological research.

Besides, since there is no extant publicly available datasets for alveoli, we have created one called Pulmonary Alveolus Pathology Image Dataset (abbreviated as PAPI). The dataset originates from lung slices obtained from West China Hospital, which boasts the largest and earliest medical testing center in China to have received accreditation from the College of American Pathologists (CAP), with instructions from the professors there.

2 Related Work

2.1 CNN-Based Medical Image Segmentation

Convolutional Neural Networks (CNNs) [5] have proven to be highly effective in medical image segmentation, with variants like FCN [6] and U-Net [7] leading the field. Advanced architectures, such as UNet++ [8] with nested and dense skip connections, Attention U-Net [9] featuring attention gates for target focus, and Res-UNet [10] incorporating weighted attention mechanisms and ResNet-based skip connections, have shown significant improvements. R2U-Net [11], KiU-Net [12], DoubleU-Net [13], and FANet [14] further demonstrate the versatility of CNNs in addressing specific challenges in medical image segmentation. However, despite their success, CNN-based methods face limitations in modeling long-range dependencies and establishing global context connections.

2.2 Vision Transformer

Expanding upon the transformative success of transformers in NLP [15], Dosovitskiy et al. introduced the Vision Transformer (ViT). ViT achieved state-of-the-art performance in image classification tasks by incorporating self-attention mechanisms to capture global information. In an effort to enhance efficiency and reduce dependence on large datasets for generalization, several derivative vision transformers have been proposed [16,17,18]. Vision transformers have demonstrated impressive results in various vision tasks, including providing end-to-end transformer-based models for object detection, as well as for semantic and instance segmentation [19,20,21]. In summary, the success of Vision Transformer and its derivatives extends beyond image classification, showcasing remarkable performance in tasks such as image segmentation. These transformer-based models have proven effective in capturing intricate features and spatial dependencies, highlighting their versatility across a spectrum of computer vision applications.

2.3 SW-MSA in Swin Transformer for Segmentation

Swin-Transformer introduces a pivotal innovation with its window-based multi-head self-attention (W-MSA), demonstrating linear computational complexity. Shifted

window-based MSA (SW-MSA), and it achieves state-of-the-art performance in image recognition and dense prediction tasks such as object detection and semantic segmentation. This design choice proves highly effective in various CV tasks, including image recognition, object detection, and semantic segmentation. In the context of semantic segmentation, SETR [22] adopts transformer as an encoder, showcasing the efficacy of transformers in sequence-to-sequence prediction for segmentation tasks. Segmenter utilizes ViT as the encoder and introduces a mask transformer decoder, highlighting the versatility of transformer-based architectures in segmentation. TransFuse [23], MedT [24], and MCTrans [25] provide different approaches to integrating transformers and CNNs for biomedical segmentation. Unlike most previous transformer-based models, Swin Transformer is flexible to be a general-purpose backbone network by introducing the hierarchical architecture for dense prediction.

3 Method

3.1 Architecture Overview

An overview and module details of the proposed CE-TransUnet are respectively presented in **Fig. 1** and **Fig. 2**. The whole structure is depicted in **Fig. 1**, and the intricate details of the model are elucidated in **Fig. 2**. We retained much of the native Unet encoder layers because our experiment shows the preeminence of original Unet structure (our experiment results are shown in 4.3), and there is empirical evidence suggesting that such improvement permits better results compared to introducing transformer layers early in the encoding layers [26]. Besides, due to the clear edge features in alveolar images (which will be elaborated on in detail in 3.2) and the strong edge detection capability of convolutional layers [27], we have integrated more convolutional elements into the model, which will be further elaborated in 3.2.

As a result, an input $X \in R^{H \times W \times C}$ with a spatial resolution of $H \times W$ and channel C encounters the incorporated CE-Transformer block only after passing several convolutional and convolutional down-sampling modules. After that, we repetitively implement convolutional down-sampling layers as well as CE-Transformer block for several times as the encoder and bottleneck of our model. Eventually, X goes through the Unet decoder and features are derived.

3.2 Convolution-Enhanced Elements

In this subsection, we will first demonstrate the conspicuous edge features of alveolus, and then thoroughly exhibit the convolutional improvements we designed and adopted. As mentioned in 3.1, we pursue convolutional integrations since they could aid our model in better identifying the prominent edge features of pulmonary pathology images [27] and help reduce computational complexity as well [28].

We conducted statistical analysis of Local Binary Patterns (LBP) [29] on the original images of our entire pulmonary alveoli dataset. The results of this analysis are depicted in **Fig. 3**, and you can also go to **Fig. 6** to get an overview of alveolar pathology images.

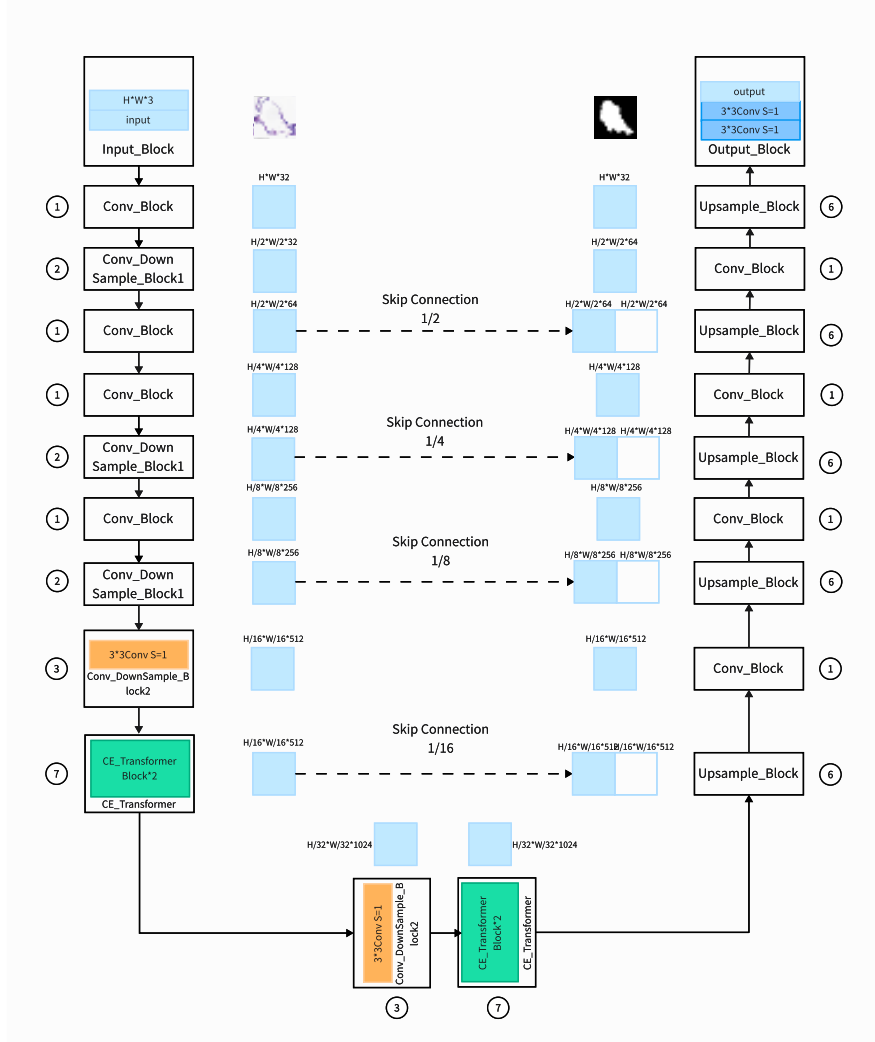


Fig. 1. Overview of CE-TransUnet Structure

The x-axis legend, from LBP-Code1 to LBP-Code9, represents patterns ranging from 00000000, 00000001, 00000011, ... to 11111111 respectively. The y-axis represents the proportion of each pattern. The results from **Fig. 3** indicate that data corresponding to LBP-Code5 pattern accounts for a significant proportion, reaching 13.07%. This suggests that the edge features in pulmonary pathological images are remarkably prominent. The heightened ubiquity of patterns in LBP-Code1 stems from the considerable presence of pulmonary alveoli, our segmentation subject, within the overall image. Moreover, the smoothness inherent to the alveolar interiors aligns with the characteristics of LBP-Code1 patterns.

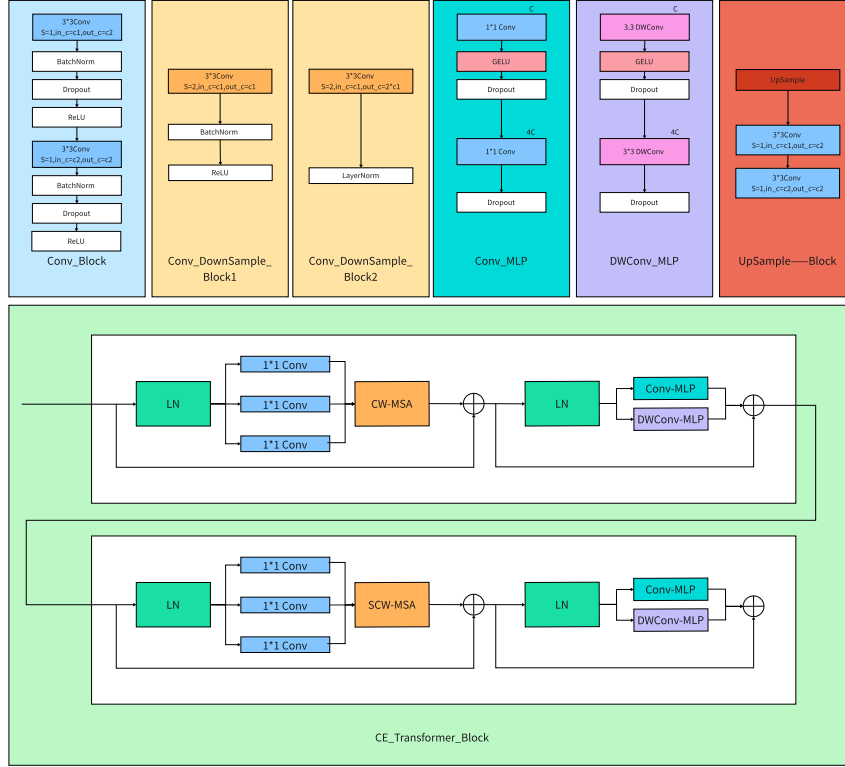


Fig. 2. Module Details

We will then delineate our convolution-enhanced methods in the followings:

(Depthwise) Convolution-Enhanced Multi-Layer Perceptron. We have replace the original two linear layers with convolutional layers in the traditional MLP module in CE-Transformer Block (which will be introduced in 3.3) both to extract features of alveolus more precisely and to reduce more computational burden. Also, depthwise convolutional layers are introduced in MLP. We parallelized it with the non-depthwise convolution-improved MLP in our model, which could provide a different scope from the normal convolution channel and also effectively reduce the number of parameters. Hence, the features of alveoli can be extracted by the model across a broader spectrum of scales, resulting in a more precise output.

Convolution-Enhanced (S)W-MSA. We adopted a convolution-enhanced self-attention mechanism in CE-Transformer Block (which will be introduced in 3.3), with empirical evidence demonstrating its feasibility and superiority, which can tremendously reduce computational complexity and help the attention module to obtain more effective features [28].

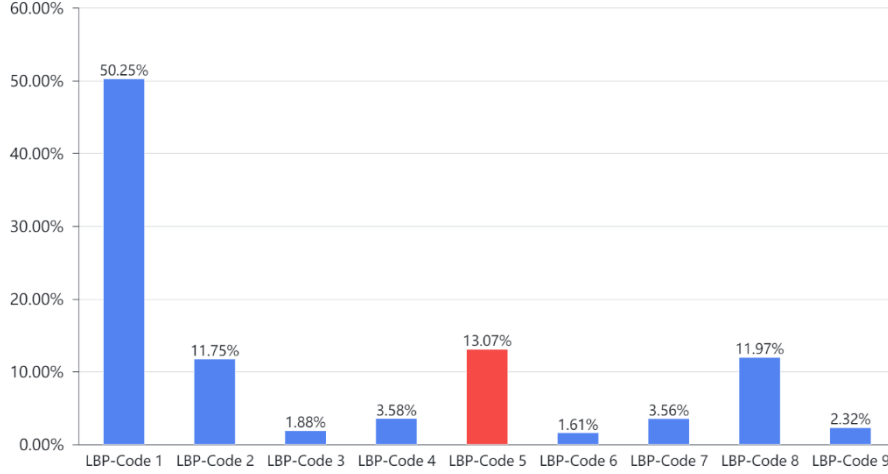


Fig. 3. LBP Statistics of Pulmonary Alveolus Pathology Images

Convolutional Downsampling Module. We adopted convolutional layers instead of pooling layers for down-sampling operations to let the model learn more features and thus obtain an increased receptive field. Long-term researches illustrate that convolutional down-sampling surpasses pooling in image segmentation tasks.

3.3 CE-Transformer Block

Transformer is renowned for its powerful capability to capture long-range dependencies. It means that it could partition images into patches and employs additional positional encodings to model the relative spatial relationships between them, and then capture dependencies even among distant patches. As depicted in **Fig. 3**, the prevalence of the LBP-Code 1 pattern suggests widespread distribution of alveolar cavity areas in pulmonary pathology images. The segmentation objective in pulmonary pathology images is to identify and segment alveolar cavities, and as described in 3.2, the edge features of alveolar cavities also occupy a significant portion of the image. Furthermore, numerous experiments and papers have demonstrated the powerful portability of the Transformer, with precedents showing its adaptability to the Unet architecture [26]. Therefore, the introduction of the Transformer facilitates the recognition of dependencies among high-frequency features such as cavities and edges in the image, thereby aiding in the segmentation of pulmonary alveolar pathology images.

Our CE-Transformer Block first adopts linear layer, and then let normalized data get through (shifted) convolution-enhanced window-based multi-head self-attention ((S)CW-MSA) layer to capture multi-scale features. The (S)CW-MSA module ensures lower computational cost and more focus laid on information communications between patches, hence preventing feature loss [28]. The feature map obtained through (S)CW-MSA are added to the original feature map, and then further input into the LN layer for normalization. After that, our (depthwise) convolution-enhanced multi-layer

perceptron ((DW)Conv-MLP) is converged and further features are extracted. The final result is also added to the feature map obtained before the second LN layer. Based on the structure of our CE-Transformer Block, we can derive the following model workflow:

$$\widehat{z}^1 = \text{CW-MSA} \left(\text{LN}(z^{(l-1)}) \right) + z^{(l-1)}, \quad (1)$$

$$\overline{z}^1 = \text{DWConv-MLP} \left(\text{LN}(\widehat{z}^1) \right) + \widehat{z}^1, \quad (2)$$

$$\dot{z}^1 = \text{Conv-MLP} \left(\text{LN}(\overline{z}^1) \right) + \overline{z}^1, \quad (3)$$

$$z^1 = \widehat{z}^1 + \overline{z}^1 + \dot{z}^1, \quad (4)$$

$$\widehat{z}^{(l+1)} = \text{SCW-MSA} \left(\text{LN}(z^1) \right) + z^{(l)}, \quad (5)$$

$$\overline{z}^{(l+1)} = \text{DWConv-MLP} \left(\text{LN}(\widehat{z}^{(l+1)}) \right) + \widehat{z}^{(l+1)}, \quad (6)$$

$$z^{(l+1)} = \text{Conv-MLP} \left(\text{LN}(\overline{z}^{(l+1)}) \right) + \overline{z}^{(l+1)}, \quad (7)$$

$$z^{(l+1)} = \widehat{z}^{(l+1)} + \overline{z}^{(l+1)} + z^{(l+1)} \quad (8)$$

In the equations above, \widehat{z}^1 , \overline{z}^1 and \dot{z}^1 respectively correspond to the outcome of (S)CW-MSA, DWConv-MLP and Conv-MLP module, while z^1 represents the sum of \widehat{z}^1 , \overline{z}^1 and \dot{z}^1 , the same is for z^{l+1} .

3.4 Encoder

We made some adjustments to the structure of the encoding layer in our model, transitioning from some of the convolutional layers and pooling layers to CE-Transformer blocks and convolutional down-sampling modules respectively. We reserve some convolutional layers as the input part of our model because evidence has shown that introducing transformer blocks early does not yield significant positive effects on feature extraction.

3.5 Bottleneck

We converge two CE-Transformer blocks as the bottleneck of our U-shaped model.

4 Experiment and Results

In this section, we will demonstrate how we get our dataset and our experiment settings. In the former section, we will discuss how we process the original alveolar data to make it a high-quality dataset. After that, we will give our experiment details and compare our CE-TransUnet with the classic and basic medical image segmentation model Unet,

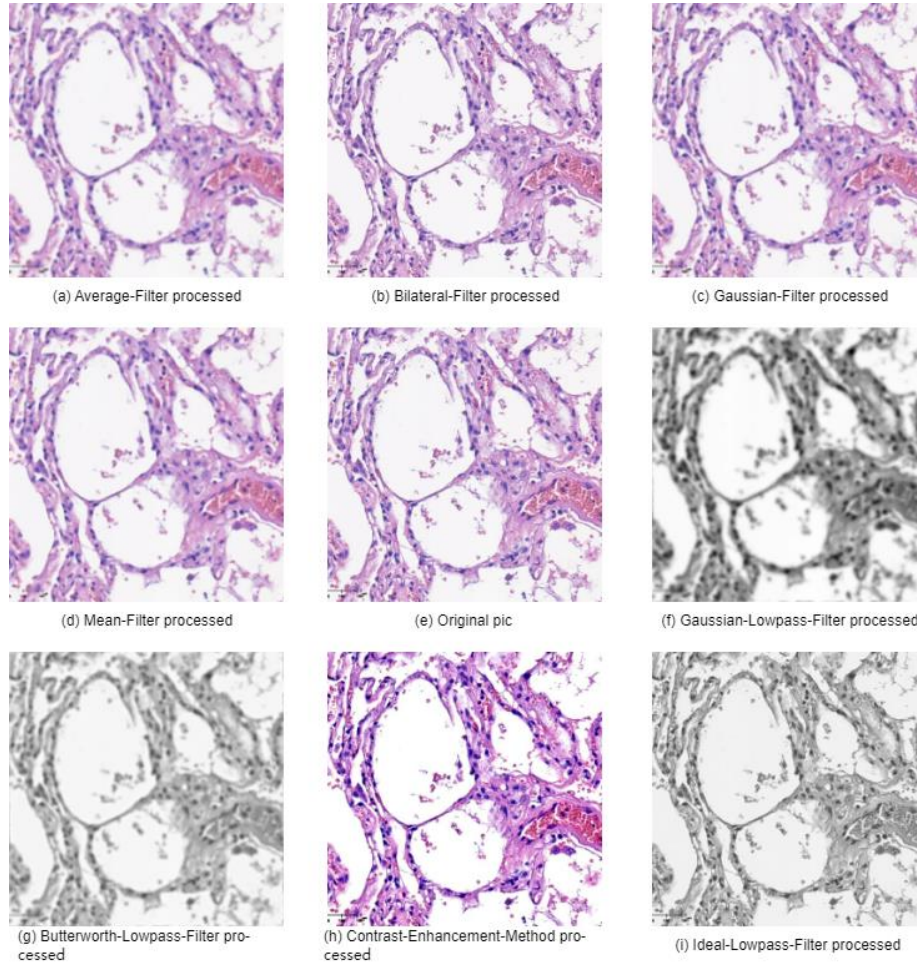


Fig. 4. Demonstration and Comparison of Image Enhancement Methods

some state-of-the-arts model like nnUnet, etc. The results are evaluated through the DICE coefficient to show different models' segmentation accuracy.

4.1 Dataset

In this subsection, we will first provide an overview of the dataset, then the reasons for creating and subsequently making publicly available the dataset, and finally, methods we adopted for image enhancement and data augmentation.

An Overview of Pulmonary Alveolus Pathology Image Dataset (abbreviated as PAPI). After preprocessing, our pathology alveolus dataset consists of 6056 training samples and 640 testing samples. All 6696 data points have been annotated. You can

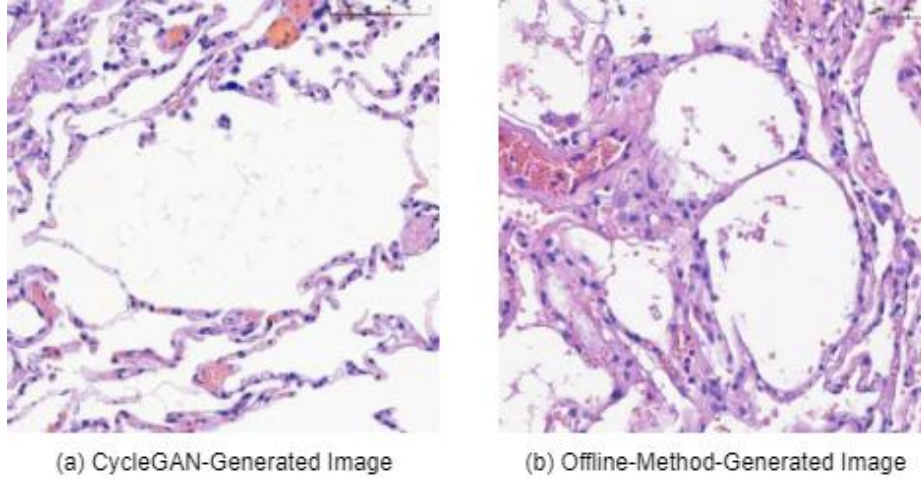


Fig. 5. Demonstration of Data Augmentation Methods

refer to **Fig. 6** to see our raw data along with the corresponding labels.

Backgrounds. Due to factors such as the high cost of annotation, patient privacy, policy regulations, and the relatively niche nature of the field, medical image datasets often exist in a small-sample state. This, in turn, results in a scarcity of medical-related datasets available online, which vary widely in terms of size and quality. Even when obtaining samples directly from specialized medical institutions, it is challenging to obtain a comprehensive dataset with high capacity and quality at a low economic and time cost. Therefore, expanding the dataset and applying data augmentation techniques are necessary operations that enable convolutional neural networks to learn more diverse sample features in subsequent experiments, thereby further improving their overall performance.

Under the guidance of a professor in West China Hospital, our group members conducted an on-site investigation of the pathological image processing procedures. We studied the pathological sections of pulmonary cases and received instruction from professional doctors. After carefully reviewing a large number of slices under their guidance, we identified the required peripheral lung, distant lung, and residual lung slices, which have later been put into experiments.

Image enhancement. Due to the large number of segmentation objects contained in each pathological image obtained from the laboratory for this project, as well as their high resolution, it is not suitable to directly process them using segmentation networks. Also, simply letting machines crop these images made them too vague for our model to identify. Therefore, we manually scanned and selected those available sections (which indeed takes us lots of time), and first ran numerous filters to enhance the pixel quality of the images. Furthermore, when the cropping size at the edges is not large enough, we can overlap images appropriately, meaning there may be larger

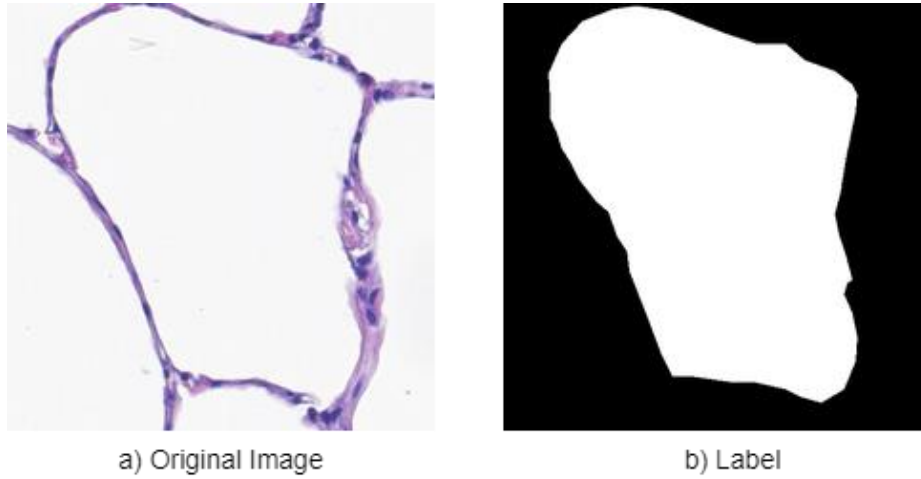


Fig. 6. Demonstration of Pulmonary Alveolus Pathology Image Dataset (PAPI)

overlapping similar domains between the images collected from nearby regions to obtain images that meet the requirements.

In order to eliminate as much noises as possible, we have tested both frequency domain denoising methods like Ideal Low-pass Filter [32], Butterworth Low-pass Filter [33] and Gaussian Low-pass Filter [34], as well as spatial domain denoising methods like Average Filter [35], Gaussian Filter [36], Median Filter [37] and Bilateral Filter [38]. Eventually, we agreed on the Bilateral one as the best option, with the given experiment results(demonstrated in **Fig. 4**).

In addition, we employed contrast enhancement techniques to highlight the features of alveoli, thus increasing the differentiation from the surrounding edges. ESR-GAN [30] may also work if you need a dataset of higher resolution.

Data augmentation. At last, we adopted several methods to enlarge our datasets. Offline data augmentation methods will be employed to expand and improve the dataset. In this project, we randomly rotated the images by 90, 180 or 270 degrees. In addition, Cycle-GAN [31] is introduced for help as well, for it is easy for machines to acquire the characteristics of alveolar, which has a nearly-round shape. After initial training on annotated samples, Cycle-GAN will generate alveolar images which, though, have similar structures to the original one, but obtain different and various colour details, hence increasing the robustness of the dataset. This augmentation will further enhance the final recognition accuracy of the subsequent convolutional neural network (CNN). It will allow for more precise extraction of relevant features during alveolar segmentation, making it more robust. Both GAN-generated and Offline-method-generated results are shown in **Fig. 5**.

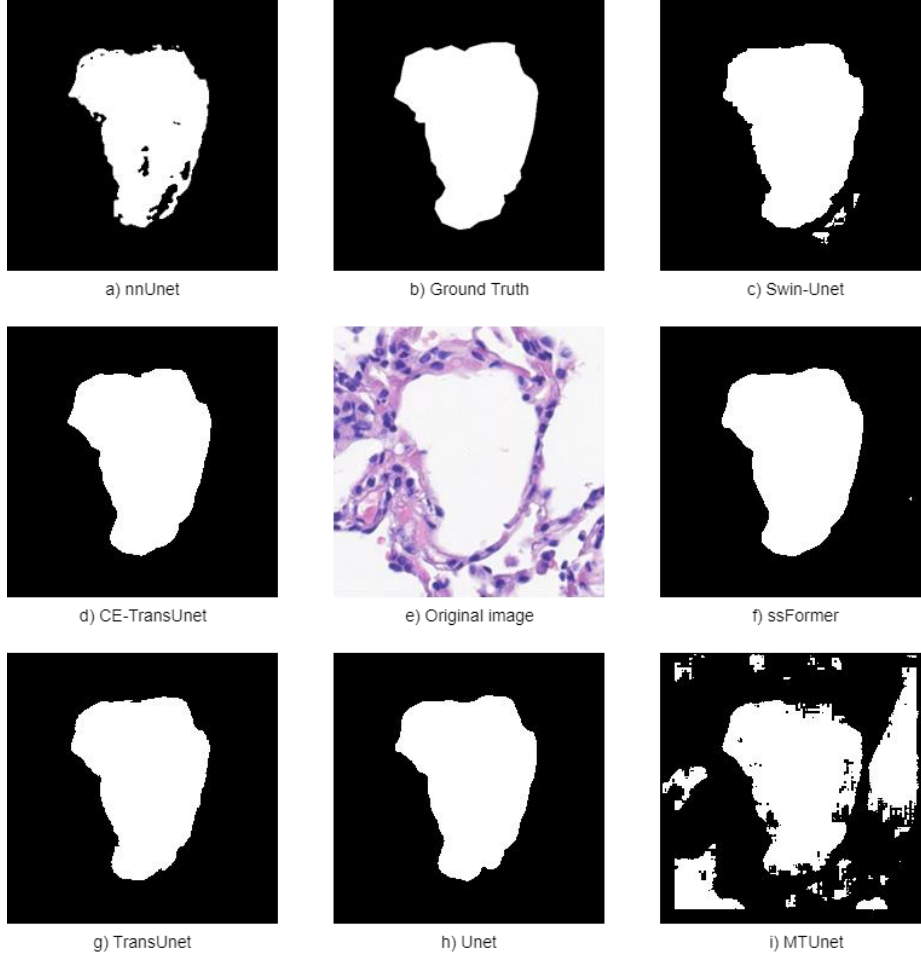


Fig. 7. Segmentation Results of Different Methods

4.2 Implementation Details

We train our CE-TransUnet on a Nvidia 3090 GPU (which has 24 GB memory) with Python 3.8 and Pytorch 2.0 as basic environment setup. To ensure fair comparisons, CE-TransUnet and all the other models are trained with the same training setup. We used the original code for each method, and input size is kept as 224×224 , and learning rate as well as batch size is set to $1e - 2$ and 24 respectively. Additionally, we adopt Adam as the optimizer for training, and IOU as well as Dice coefficient for evaluation. We choose CrossEntropyLoss as our loss function. All these models are trained for 500 epochs, and the best model is selected from the checkpoints.

Table 1. Performance Comparison between Each Model on PAPI

Segmentation Models	Evaluation Metrics		
	Dice Score(%)	Accuracy(%)	IOU(%)
Unet	94.86	97.66	92.05
nnUnet-v2	89.00	89.71	81.34
Swin-Unet	92.10	96.24	87.02
MT-UNet	64.20	74.19	49.67
SSFormer	87.14	93.58	79.66
TransUnet	92.49	96.46	87.66
CE-TransUnet B	93.89	97.07	90.29
CE-TransUnet VT	95.31	97.94	92.87

Table 2. Ablation Study on Input Size

Input Size	Evaluation Metrics		
	Dice Score(%)	Accuracy(%)	IOU(%)
224	95.31	97.94	92.87
448	94.92	97.71	92.16
512	77.97	87.73	66.00

Table 3. Ablation Study on Skip Connections

Number of Skip Connections	Evaluation Metrics		
	Dice Score(%)	Accuracy(%)	IOU(%)
0	67.29	81.15	52.16
1	94.80	97.63	91.91
2	36.58	65.12	23.82
3	95.10	97.81	92.48
4	95.31	97.94	92.87
5	95.06	97.79	92.42

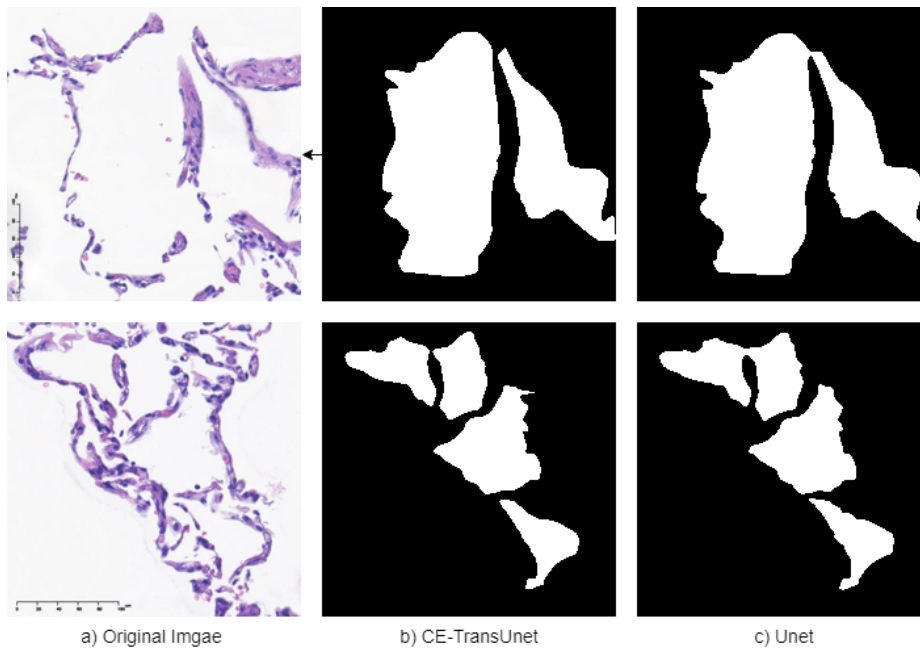
4.3 Main Results

We compare our CE-TransUnet with some classical Unet-architecture models, as well as state-of-the-art methods like nnUnet, Swin-Unet on our alveolar dataset. All these methods are trained with the released codes from their respective authors for fair play. You can refer to the Transformer Set section in 4.4 to get a detailed explanation for CE-TransUnet B and CE-TransUnet VT occurred in **Table 1**.

Quantitative Comparison. From the results in **Table 1**, we notice that CE-TransUnet attains the highest overall evaluation scores, with a Dice score of 95.31%, Accuracy of 97.94%, and IOU score of 92.87%. This suggests its superior segmentation capability compared to other methods in the table. Upon reviewing the table, it is evident that other state-of-the-art methods like Swin-Unet and nnUnet perform even worse than the baseline Unet in the task of alveolar segmentation. We assume the reason Unet works such well is that, as mentioned before in 3.2, convolutional layers

Table 4. Ablation Study on Transformer Block Set

Model Scale	Model Params	Evaluation Metrics		
		Dice Score(%)	Accuracy(%)	IOU(%)
CE-VT	58.46M	95.31	97.94	92.87
CE-T	83.82M	95.09	97.80	92.47
CE-B	65.38M	93.89	97.07	90.29
CE-L	71.78M	93.75	97.04	90.02
CE-VL	109.3M	93.75	97.00	90.06

**Fig. 8.** Comparison of Multiple Instance Segmentation Results Based on Unet and CE-TransUnet

outperforms the transformer modules in a PAPI, where one picture contains only one or few objects. To figure this out, we have done an ablation study (details are presented in 4.4) and found that CE-VT, the one with the most transformer blocks replaced by convolutional layers, performs better, which means that convolutional layers are indeed a well-performed fit for PAPI.

The Dice score of most of the models remains near 90%. Nevertheless, CE-TransUnet achieves a nearly 5-percentage-point enhancement, which indicates that it can better learn the features of the object with convolutional improvements.

Furthermore, despite the minor differences in DICE scores between Unet and our model, the inferior performance of Unet is highlighted in multiple instance tasks, which will be discussed in detail in the subsequent section on Qualitative Comparison.

Qualitative Comparison. We provide visual examples of segmentation results for each model in **Fig. 7**. It is obvious that the segmentation results of our model preserve the utmost edge and inner features, and matches the ground truth most. Meanwhile, Unet, ssFormer, TransUnet and Swin-Unet also reach a nearly-perfect performance, yet minute flaws still exist. The results produced by nnUnet and MTUnet indicate that they are not fit for alveoli segmentation tasks.

As mentioned earlier, although there's little difference in DICE scores between Unet and our model in quantitative comparison, the disadvantage of Unet in multiple instance segmentation tasks is quite evident. As illustrated in **Fig. 8**, Unet fails to segment two closely spaced alveoli without a continuous purple wall separating them in pathological images. Unet tends to connect the mask images of two exclusive alveoli through alveolar cavities. This demonstrates that our model can capture deeper features about the alveoli, rather than just segmenting based on the features presented in pathological images.

4.4 Ablation Study

Extensive experiments are conducted to determine which factor may well influence the segmentation result of CE-TransUnet. Input size, skip connections as well as the number of the inserted transformer blocks will be analyzed below. The dataset for this section is also PAPI (Pulmonary Alveolus Pathology Image Dataset).

Input Size. As shown in **Table 2**, CE-VT has witnessed a sharp decline in its performance in the case where only the input size rises from 224×224 , 448×448 to 512×512 . Since 512 is not a multiple of 7 and the window size in (S)W-MSA is set to a multiple of 7, complex padding issues are hence introduced, resulting in such poor training performance even though more details are revealed through an enhanced input resolution. We eventually adopt 224×224 as input size due to its superior performance.

Skip connections. We demonstrate the effect of the number of skip connections in **Table 3**. The skip connections of our CE-TransUnet are implanted at places of the 1, 1/2, 1/4, 1/8, and 1/16 resolution scales and we gradually delete the skip connections from top to bottom. It is evident that deep skip connections have a substantial impact on result accuracy, as they convey a more abundant and critical feature information. In contrast, shallow skip connections exert a relatively minor influence. We found that retaining the last two skip connections resulted in poor performance, but when only the last one was retained, the performance improved. Based on experimental results, we ultimately chose a model without the topmost skip connection.

Transformer Set. The impact of different transformer block sets on test accuracy can be observed from **Table 4**. CE-VeryTiny(CE-VT), CE-Tiny(CE-T), CE-Base(CE-B), CE-Large(CE-L), CE-VeryLarge(CE-VL) correspond to combinations of 2,2, 2,4, 2,2,4,2, 2,2,6,2 and 2,2,18,2, respectively. The results demonstrate that these various combinations have yielded worse impacts on the outcomes with more transformer modules included, and the model complexity generally grows tremendously. The reason why CE-T has such a large number of parameters is because it incorporates four transformer blocks at the bottleneck, where the volume of parameters to be computed

becomes significantly substantial. CE-VT is the lightest version with the best outcome. Therefore, we adopt CE-VT as our final version for alveoli segmentation. We have noticed that the more convolutional layers are employed to substitute transformer blocks, the better the result, which is in consistence to our aforementioned suspect that convolutional layers are extremely fit for the segmentation of objects with clear and salient edge features.

5 Conclusion

In this paper, we have introduced our innovative Pulmonary Alveolus Pathology Image dataset (PAPI). Besides, we have proposed our convolutional enhanced U-shaped method with improved transformer blocks called Convolutional Enhanced Transformer-based Unet (abbreviated as CE-TransUnet), which underwent targeted improvements based on the characteristics of alveoli. Through 4.4 we know that CE-VT with 4 skip connections and 224 as input size performs the best in PAPI. It keeps a simple structure while achieving better performance in PAPI with relatively low computational burden. It also demonstrates outstanding performance in quantitative testing. For future improvements, we plan to make both our method more robust and testify whether CE-TransUnet could perform better in other medical segmentation tasks in the future. We hope our PAPI and CE-TransUnet could make a difference to the pulmonary alveolus-concerned work.

Acknowledgment. We sincerely felt indebted to the professors in West China Hospital, from whom we gained valuable lung slice samples. Without them, this project may not proceed.

References

1. Bu, H., Li, Y., Lai, M., Wang, Y., Wang, G., & Tao, Y. (n.d.). Pathology, 9th edition. Beijing, China: People's Medical Publishing House. ISBN: 978-7-117-26438-9.
2. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012–10022).
3. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2020). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18, 203-211. DOI: 10.1038/s41592-020-01008-z.
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2021). Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. arXiv preprint arXiv:2105.05537 [eess.IV].
5. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.
6. Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3431-3440).

7. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention* (pp. 234-241).
8. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., et al. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings* (pp. 3-11). Springer International Publishing.
9. Siddique, N., Paheding, S., Elkin, C. P., et al. (2021). U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 9, 82031-82057.
10. Xiao, X., Lian, S., Luo, Z., et al. (2018). Weighted res-unet for high-quality retina vessel segmentation. In *2018 9th international conference on information technology in medicine and education (ITME)* (pp. 327-331). IEEE.
11. Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., & Asari, V. K. (2018). Recurrent residual convolutional neural network based on u-net (r2unet) for medical image segmentation. *arXiv preprint arXiv:1802.06955*.
12. Valanarasu, J. M. J., Sindagi, V. A., Hacihaliloglu, I., & Patel, V. M. (2020). Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 363–373). Springer.
13. Jha, D., Riegler, M. A., Johansen, D., Halvorsen, P., & Johansen, H. D. (2020). Doubleunet: A deep convolutional neural network for medical image segmentation. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)* (pp. 558–564). IEEE.
14. Tomar, N. K., Jha, D., Riegler, M. A., Johansen, H. D., Johansen, D., Rittscher, J., Halvorsen, P., & Ali, S. (2021). Fanet: A feedback attention network for improved biomedical image segmentation. *arXiv preprint arXiv:2103.17235*.
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
16. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z-H., Tay, F. E. H., Feng, J., & Yan, S. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 558–567).
17. Wang, W., Xie, E., Li, X., Fan, D-P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 568–578).
18. Chen, C-F. R., Fan, Q., & Panda, R. (2021). Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 357–366).
19. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.
20. Strudel, R., Garcia, R., Laptev, I., & Schmid, C. (2021). Segformer: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 7262–7272).
21. Han, K., et al. (2023). A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1), 87-110. DOI: 10.1109/TPAMI.2022.3152247.

22. Zheng, S., et al. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) (pp. 6881-6890).
23. Zhang, Y., Liu, H., & Hu, Q. (2021). TransFuse: Fusing transformers and CNNs for medical image segmentation. In Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (pp. 14-24).
24. Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (pp. 36-46).
25. Ji, Y., et al. (2021). Multi-compound transformer for accurate biomedical image segmentation. In Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (pp. 326-336).
26. Chen, J., et al. (2021). TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv:2102.04306 [cs.CV].
27. Pu, M., et al. (2022). EDTER: Edge Detection with Transformer. arXiv:2203.08566 [cs.CV].
28. Yang, C., Wang, Y., Zhang, J., Zhang, H., Wei, Z., Lin, Z., & Yuille, A. (2021). Lite Vision Transformer with Enhanced Self-Attention. arXiv:2112.10809 [cs.CV].
29. Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 971-987.
30. Wang, X., et al. (2018). ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops (pp. 63-79).
31. Zhu, J-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 2223-2232).
32. Oppenheim, A. V., & Schaffer, R. W. (1975). *Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, N.J.
33. Butterworth, S. (1930). On the theory of filter amplifiers. *Wireless Engineer*, 7(6), 536-541.
34. Cooley, J. W., & Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90), 297-301.
35. Hoang, T-V., Krumscheid, S., Matthies, H. G., & Tempone, R. (2020). Machine learning-based conditional mean filter: a generalization of the ensemble Kalman filter for nonlinear data assimilation. arXiv preprint arXiv:2001.08073.
36. Wang, M., Zheng, S., Li, X., & Qin, X. (2014). A new image denoising method based on Gaussian filter. In 2014 International Conference on Information Science, Electronics and Electrical Engineering (pp. 286-289). IEEE.
37. Hwang, H. S. M., & Haddad, R. A. (1995). Adaptive median filters: new algorithms and results. *IEEE Transactions on Image Processing*, 4(4), 499-502.
38. Tomasi, C., & Manduchi, R. (1998). Bilateral filtering for gray and color images. In Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271) (pp. 839-846). IEEE.