Improvement of TransUNet Using Word Patches Created from Different Dataset

Ayato Takama[a], Satoshi Kamiya[b] and Kazuhiro Hotta[c]
Meijo University, 1-501 Shiogamaguchi, Tempaku-ku, Nagoya 468-8502, Japan

Keywords: Semantic Segmentation, TransUNet, Mix Transformer, Word Patches.

Abstract: UNet is widely used in medical image segmentation, but it cannot extract global information sufficiently. On the other hand, TransUNet achieves better accuracy than conventional UNet by combining a CNN, which is good at local features, and a Transformer, which is good at global features. In general, TransUNet requires a large amount of training data, but there are constraints on training images in the medical area. In addition, the encoder of TransUNet uses a pre-trained model on ImageNet consisted of natural images, but the difference between medical images and natural images is a problem. In this paper, we propose a method to learn Word Patches from other medical datasets and effectively utilize them for training TransUNet. Experiments on the ACDC dataset containing 4 classes of 3D MRI images and the Synapse multi-organ segmentation dataset containing 9 classes of CT images show that the proposed method improved the accuracy even with small training data, and we showed that the performance of TransUNet is greatly improved by using Word Patches created from different medical datasets.

1 INTRODUCTION

UNet(Ronneberger et al., 2015) is the most commonly used model for medical image segmentation. UNet based on CNN can extract local features, but it cannot extract global information sufficiently. On the other hand, TransUNet(Chen et al., 2021) combines a CNN, which is good at extracting local features, and a Transformer, which is good at extracting global features, to enable segmentation while extracting more global information than conventional UNets.

In general, Transformers(Vaswani et al., 2017) have the problem of requiring a large number of training images. Since it is difficult to prepare a large number of training samples in the medical field, it is necessary to achieve high accuracy with as few training images as possible. The encoder in TransUNet uses a pre-training model based on ImageNet, which consists of natural images and has a large gap with medical images, and it may be unsuitable as a pre-training model.

Transformer was originally proposed in the field of natural language processing. Unlike natural language processing, the field of image recognition does not have the concept of a word. Therefore, we should create words from medical images and incorporate them into the Transformer to achieve the potential ability of the Transformer.

In this paper, we create Word Patches from a different medical dataset from training images, and the Word Patches are incorporated into TransUNet as cross attention as shown in Figure 1. The information in different medical images that could not be obtained in the pre-training can be used in TransUNet.

Since the similarities between Word Patches and the patterns appearing in the training image of the problem to be solved can be used for learning and classification, the accuracy of TransUNet is expected to be improved even with a small number of training images.

We conducted the experiments on the Automatic Cardiac Diagnosis Challenge (ACDC) dataset(Sakaridis et al., 2021) containing 4 classes of 3D MRI images and the Synapse multi-organ segmentation dataset containing 9 classes of CT images of the lower abdomen. The proposed method was compared to the conventional TransUNet method in the case where only 10% of the training images were used. We incorporated word patches created from various datasets into TransUNet, and confirmed that the proposed method improved the accuracy

[a] https://orcid.org/0000-0001-7255-1328
[b] https://orcid.org/0000-0002-7057-3280
[c] https://orcid.org/0000-0002-5675-8713

Takama, A., Kamiya, S. and Hotta, K.
Improvement of TransUNet Using Word Patches Created from Different Dataset.
DOI: 10.5220/0012419900003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 378-385
ISBN: 978-989-758-684-2; ISSN: 2184-4313

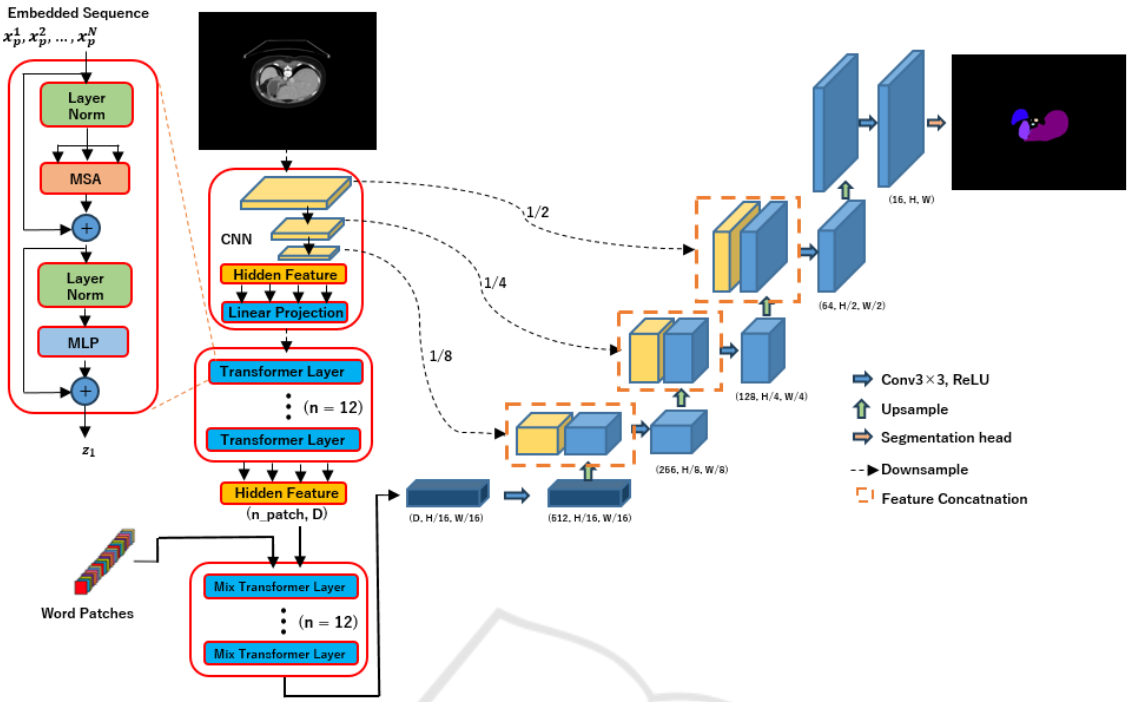

Figure 1: Architecture of the proposed Method.

compared with the original TransUNet with fewer training images.

The paper is organized as follows. Section 2 describes related studies. Section 3 explains the proposed method. Section 4 presents experimental results. Finally, Section 5 describes conclusion and future works.

2 RELATED WORKS

In recent years, many methods for improving the Vision Transformer (ViT)(Dosovitskiy et al., 2020) have been studied after the sucess of ViT in image classification. For example, TrackFormer(Meinhardt et al., 2022), MOTR(Zeng et al., 2022), and TransMOT(Chu et al., 2021) for object tracking, and DETR(Carion et al., 2020) for object detection, SegFormer(Xie et al., 2021), Swin Transformer(Liu et al., 2021), MetaFormer(Yu et al., 2022), and TransUNet are well-known for semantic segmentation.

Among them, TransUNet is an improvement of UNet, the most well-known model for medical image segmentation: the CNN-based UNet can extract local features, but it does not sufficiently extract the global information that is important for segmentation. Therefore, TransUNet combines a CNN, which is good at extracting local features, and a Transformer, which is good at extracting global information, making it possible to perform segmentation while extracting more global information than conventional UNets.

Figure 2 shows the architecture of TransUNet, which is a model that incorporates ViT into the UNet encoder. The encoder performs feature extraction using a CNN to extract local features. Then, the Transformer is used for feature extraction of global information. At the decoder, up-sampling is performed as in UNet. The final output is the segmentation result. The feature maps at encoder are connected to the corresponding layer of the decoder by a skip-connection.

In general, Transformer has the problem of requiring a large number of images for training. Since it is difficult to prepare a large number of training samples in the medical field, it is necessary to achieve high accuracy with as few training images as possible. The encoder of TransUNet uses a pre-training model based on ImageNet, which is a natural image and has a large gap with medical images. Therefore, there is a problem that it may not be appropriate as a pre-trained model.

In recent years, there has been a lot of methods (Guibas et al., 2021; Sethi et al., 2021; Tan et al., 2021) that improve token mixing, but there has been no method focusing on the words. We consider that words are important because original Transformer was proposed in natural language processing and used words effectively. Thus, we create Word Patches from some datasets and use them for classification. If the

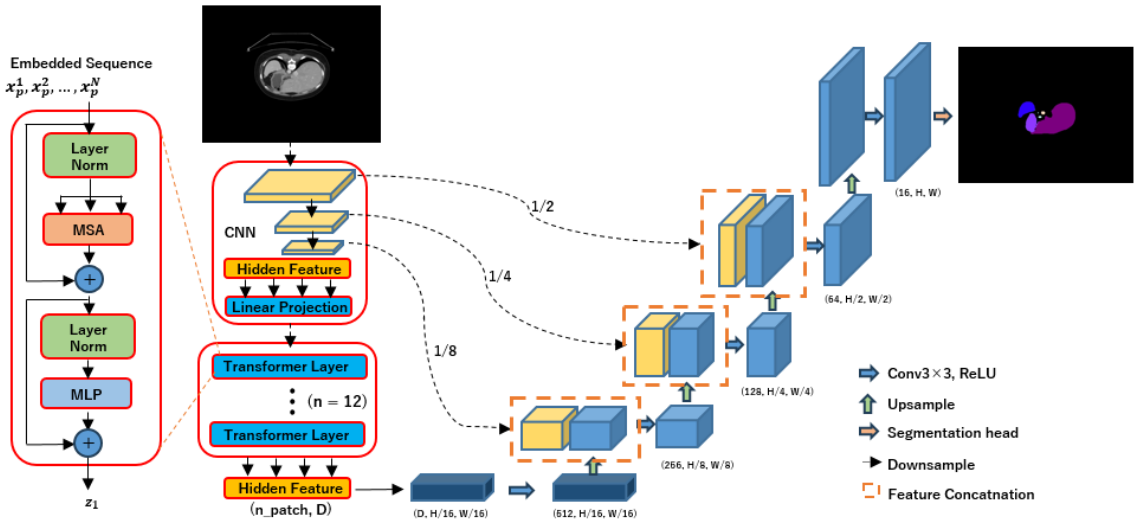

Figure 2: Architecture of TransUNet.

proposed Visual Word Patches are used effectively, we would use the potential performance of the Transformer, and the accuracy would be improved.

PatchCore(Roth et al., 2021) is a method used for anomaly detection. First, we extract features from image patches of only normal images. After that, a memory bank is created by the following sampling method.

1. Randomly sample patch features from a dataset.
2. Select the feature vector that is farthest from the rest of the feature vectors.
3. Among the distances to other feature vectors, choose the shortest distance as the farthest feature vector.
4. Repeat step (3) until the desired number of feature vectors is reached.

By doing the above process, we can obtain patch features that reflect the diversity within the dataset. As a result, the number of patch features stored in the memory bank can be reduced.

In this paper, we extract feature vectors from the layer just before the final layer of a pre-trained ResNet. Word Patches are created from all training images in the dataset by using the above sampling method. We used the Word Patches in TransUNet to use the information of the other medical image dataset effectively.

3 PROPOSED METHOD

In the proposed method, we aim to improve the accuracy of TransUNet by introducing Word Patches created by PatchCore's sampling method into the original TransUNet. By incorporating Word Patches into TransUNet as cross attention, we consider that the information of medical images, that could not be obtained by pre-training with ImageNet, can be used in TransUNet.

For anomaly detection by PatchCore, feature maps at the 2nd and 3rd blocks in ResNet are used. However, since TransUNet's encoder uses feature maps at the layer just before the final layer of ResNet, we create Word Patches using the feature vectors from the same layer of ResNet by the sampling method in Patchcore.

Because TransUNet can learn and classify using the similarities between training images and Word Patches, we can use the information about what patterns appear in each class and how similar they are to each other. Thus, we expected to improve the segmentation accuracy even with a small number of training images.

Figure 1 illustrates the architecture of the proposed method. Word Patches created from medical images are mixed with the output of TransUNet's Transformer using Mix Transformer. Mix Transformer consists of Multi-Head Attention, Layer Normalization, and MLP. In Multi-Head Attention, the output of TransUNet's encoder is used as Query, while Word Patches are used as Key and Value.

Equation (1) shows the computation of the Mix Transformer.

$$\begin{aligned} Z(Q,K,V) &= Concat(head_1,\ldots,head_{Nh}) \\ head_i &= softmax\left[\frac{Q_i(K_i)^T}{\sqrt{Ch}}\right]V_i \\ &= A_iV_i \end{aligned} \quad (1)$$

Table 1: Accuracy on ACDC dataset while changing the sampling percentage.

ACDC class	Background	Right ventricle	Myocardium	Left ventricle	mIoU
baseline	98.94	71.81	57.96	83.26	77.99
ours (ACDC 10%)	99.00	62.73	62.54	88.70	78.24
ours (ACDC 50%)	98.79	59.57	65.75	88.20	78.08
ours (ACDC 100%)	99.03	61.00	61.85	87.04	77.23

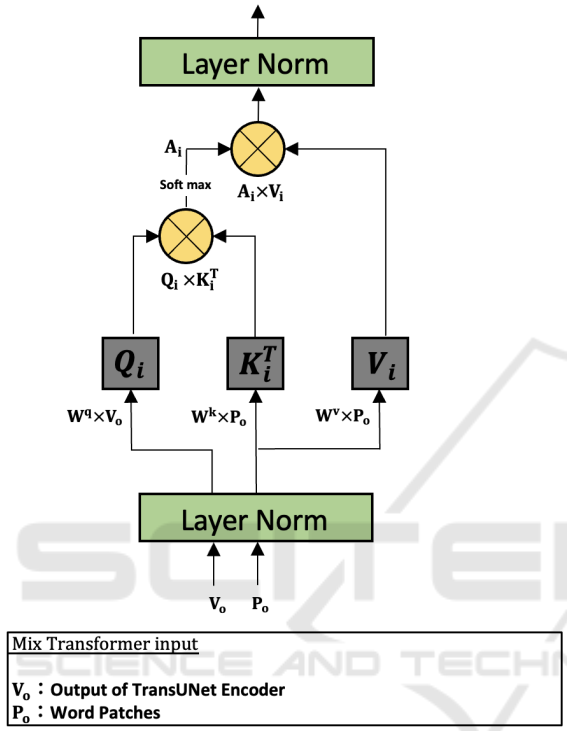

Figure 3: Mix Transformer.

where $Z(Q,K,V)$ is the output of Multi Head Attention in Mix Transformer, A_i is the attention map. Query, Key, Value are computed as

$$\begin{aligned} Q &= W^q V_o \\ K &= W^k P_o \\ V &= W^v P_o \end{aligned} \tag{2}$$

where V_o indicates the output of the conventional TransUNet and P_o indicates the Word Patches created by medical images. W^q, W^k and W^v are the 1×1 convolution.

By using Mix Transformer, we can leverage the similariteis between the feature representations in TransUNet and the Word Patches. We create an attention map from the similarity between the Query (Q) from the output of TransUNet's encoder and the Key (K) from the Word Patches. This allows us to train the model while considering the specific features for medical images and their relationships.

Mix Transformer used in this experiment was not pre-trained, and when Word Patches were added directly to the encoder output without using the Mix Transformer, the accuracy decreased. This shows that the Mix Transformer allows the similarity between the Word Patches and target data to be taken into account, and this improves the segmentation accuracy.

4 EXPERIMENTS

4.1 Datasets and Parameters

The following experiment utilizes two datasets: the Automatic Cardiac Diagnosis Challenge (ACDC) dataset consisting of 3D MRI images, and the Synapse multi-organ segmentation (SMO) dataset containing CT images of the lower abdomen. The ACDC dataset consists of 4 classes with 690 training images, 95 validation images, and 190 test images. On the other hand, the SMO dataset consists of 9 classes with 2474 training images, 613 validation images, and 692 test images.

We also used the Drosophila dataset(Gerhard et al., 2013) when we create Word Patches. The dataset consists of 20 grayscale cell images of size 1024x1024 pixels. From this dataset, 12 random crops of size 256×256 pixels are used. The cell images are more similar to medical images than the natural images in ImageNet used for pre-training, and we would like to verify the effect of Word Patches using these cell images on TransUNet.

Minibatch size was set to 16, the number of epochs to 300, and Adam and cos scheduler were used as optimization methods. In addition, mean Intersection over Union (mIoU) was used as an evaluation measure. In experiments, 5-fold cross validation was used.

4.2 Sampling Percentage

When we create Word Patches, the percentage of sampling is a hyperparameter. The experimental results on the ACDC dataset while changing the hyperparameter are shown in Table 1. In this experiment,

Table 2: Evaluation result on ACDC dataset.

ACDC class	Background	Right ventricle	Myocardium	Left ventricle	mIoU
baseline	98.94	71.81	57.96	83.26	77.99
ours (Drosophila)	99.13	76.62	59.85	82.73	79.59
ours (ACDC)	99.00	62.73	62.54	88.70	78.24
ours (SMO)	99.20	72.17	61.58	87.05	80.00

Table 3: Evaluation result on SMO dataset.

SMO class	Background	Aorta	Gallbladder	Left kidney	Right kidney	Liver	Pancreas	Spleen	Stomach	mIoU
baseline	96.46	35.01	19.72	55.79	37.77	55.04	30.02	42.61	17.76	43.35
ours (Drosophila)	97.31	38.66	12.39	45.38	44.16	67.99	16.79	52.37	24.98	44.45
ours (ACDC)	97.13	38.72	17.99	49.65	48.83	72.09	7.96	65.07	23.53	46.77
ours (SMO)	96.70	46.07	12.77	51.96	41.60	63.60	18.84	43.13	17.82	43.61

Word Patches were created from the ACDC dataset. In the Table, the baseline refers to the conventional TransUNet without using Word Patches, "ours(ACDC 10%)" is the result when 10% of the original features is selected by sampling, "ours(ACDC 50%)" is the result when 50% of the original features is selected by sampling, and "ours(ACDC 100%)" is the result when no sampling is done and all of the original features are used.

When 50% was used, the accuracy did not improve compared to the case that sampling percentage is 10%. when 100% of original features was used as Word Patches in TransUNet, the accuracy much decreased. This indicates that it is important to select appropriate number of features as Word Patches. As a result, the accuracy was improved the most when 10% of all features were selected as Word Patches. If the sampling percentage is high, similar Word Patches are included. Therefore, 10% is effective, and the sampling percentage was set to 10% in the following all experiments.

4.3 Experimental Results

Table 2 shows the accuracy on the ACDC dataset when Word Patches were created using the Drosophila dataset, ACDC dataset, and SMO dataset. The baseline refers to the conventional TransUNet without using Word Patches, "ours (Drosophila)" indicates the accuracy when Word Patches were created using the Drosophila dataset, "ours (ACDC)" represents the accuracy when Word Patches were created using the ACDC dataset, and "ours (SMO)" indicates the accuracy when Word Patches were created using the SMO dataset. In addition, Table 3 shows the accuracy on the SMO dataset. In this experiment, 10% of randomly selected training images from each dataset were used for training.

From Tables 2, we see that the accuracy was improved by using Word Patches in comparison with baseline. The accuracy is the worst when we use Word Patches from the ACDC dataset which is used for the test. On the other hand, surprisingly, when we use Word Patches created from the SMO dataset which is different from test dataset, the accuracy was the best. Since the ACDC Word Patches are the same as the training images, we consider that the accuracy could not be improved that much. By incorporating Word Patches in cross attention, it is possible to learn and classify each pixel by utilizing the similarities between Word Patches and the patterns that appear in the training images for segmentation. However, if Word Patches are created using the same training images, the advantage of Word Patches, which uses the relationship between different images, cannot be used sufficiently. Therefore, we believe that the highest accuracy was achieved when we create Word Patches from different dataset from training images.

Similarly, in Table 3, the accuracy is the worst when we use Word Patches created from the SMO dataset, while the Word Patches created from the ACDC dataset achieved the best accuracy. This indicates that the usage of Word Patches created from different dataset allows the model to learn features that are not present in its own dataset, and this leads to improve the accuracy. This result shows the effectiveness of our proposed method.

Figure 4 shows the segmentation results. The Figure on the left side shows the segmentation results on the ACDC dataset, while the Figure on the right side shows the segmentation results on the SMO dataset. The grayscale images in the first and third rows represent the test images. From left to right in the second and fourth rows show the ground truth, the baseline (TransUNet without Word Patches), and "ours (Drosophilia)" means that Word Patches are created from Drosophilia dataset, and "ours(ACDC)" means

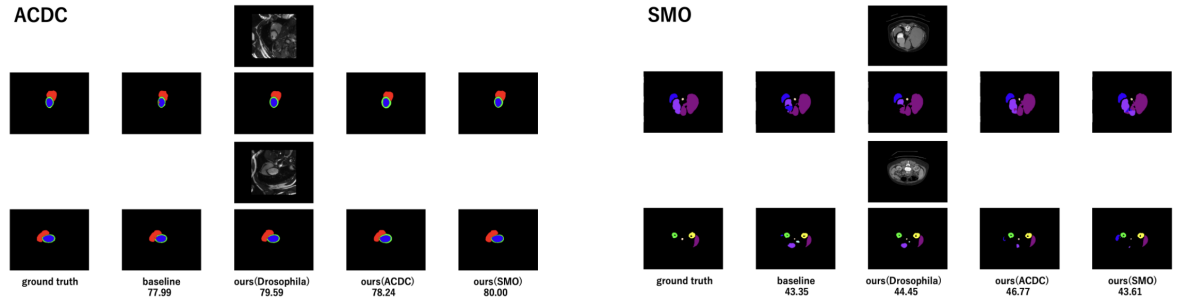

Figure 4: Segmentation Results.

that Word Patches are created from ACDC dataset, and "ours(SMO)" means that Word Patches are created from SMO dataset, respectively. The usage of Word Patches gave better segmentation results compared to the conventional TransUNet without Word Patches. Additionally, we see that Word Patches created from a different dataset lead to improve segmentation quality. This demonstrates the effectiveness of features from different dataset because we can use the features which are not present in training dataset.

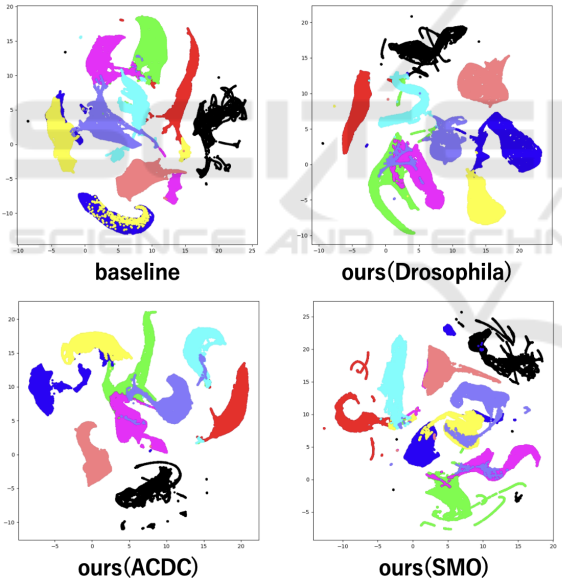

Figure 5: Class distribution per pixel in SMO dataset.

The distribution of classes per pixel in the before layer of final output layers for test images is shown in Figure 5. We used dimensionality reduction by UMAP(McInnes et al., 2018). It shows the distribution of 9 classes in the SMO dataset. Only 90,000 pixels (10,000 pixels of each class) are selected at random because the amount of all pixels is too large. The different classes are represented with different colors. In the baseline, the blue and yellow classes on the left and bottom are heavily overlapping. Some classes in the center also overlap. In contrast, when we use Word Patches, there is no significant overlap between the different classes. Thus, it can be seen that the classification is more successful when Word Patches is used than the baseline without Word Patches. In addition, when Word Patches created from ACDC dataset is used, the same classes are grouped together and the different classes are separated from each other. On the other hand, when SMO dataset was used for creating Word Patches, different classes partially overlapped with each other. This result shows that it is possible to learn and segment each class while utilizing features that are not present in the original dataset when Word Patches is created from different dataset. This enables successful classification.

4.4 Experiments with Different Number of Training Images

Table 4 shows the accuracy on the ACDC dataset when only 10%, 30% and 50% of all training images were used for training the model. In this experiment, Word Patches were created using the Drosophila dataset. Similarly, Table 5 shows the accuracy on the SMO dataset. In both Tables, "10%, 30% and 50%" means that only 10%, 30% and 50% of all training images are used for training the model. From Table 4 and 5, we see that the usage of Word Patches improved the accuracy in both cases. These results demonstrate that Word Patches improved the accuracy with a small number of training images. However, the smallest improvement in accuracy was obtained at 50% for both ACDC and SMO datasets. This indicates that the introduction of Word Patches can improve the accuracy by using information from other data when the amount of training data is small and it is difficult to train with only original data.

4.5 Ablation Study

We conduct ablation study when we mix Word Patches created from different dataset. The experimental results on the ACDC dataset are shown

Table 4: Accuracy on ACDC dataset when we change the number of training images.

ACDC class	Background	Right ventricle	Myocardium	Left ventricle	mIoU
baseline 10%	98.94	71.81	57.96	83.26	77.99
ours 10% (Drosophila)	99.13	76.62	59.85	82.73	79.59
baseline 30%	99.30	71.23	69.85	85.62	81.50
ours 30% (Drosophila)	99.31	77.70	66.88	89.27	83.29
baseline 50%	99.43	84.64	71.47	90.27	86.45
ours 50% (Drosophila)	99.58	86.59	74.64	89.05	87.47

Table 5: Accuracy on SMO dataset when we change the number of training images.

SMO class	Background	Aorta	Gallbladder	Left kidney	Right kidney	Liver	Pancreas	Spleen	Stomach	mIoU
baseline 10%	96.46	35.01	19.72	55.79	37.77	55.04	30.02	42.61	17.76	43.35
ours 10% (Drosophila)	97.31	38.66	12.39	45.38	44.16	67.99	16.79	52.37	24.98	44.45
baseline 30%	98.51	54.54	53.63	44.10	47.40	80.96	35.80	83.53	58.55	61.89
ours 30% (Drosophila)	98.71	60.93	53.85	58.59	46.93	81.08	33.53	75.54	61.47	63.40
baseline 50%	98.77	76.90	56.52	80.08	76.92	80.98	43.14	64.48	69.96	71.97
ours 50% (Drosophila)	99.15	64.62	55.63	67.94	64.00	86.38	44.34	85.27	70.86	72.02

Table 6: Accuracy on ACDC dataset when we mix different Word Patches.

ACDC class	Background	Right ventricle	Myocardium	Left ventricle	mIoU
baseline	98.94	71.81	57.96	83.26	77.99
ours (Drosophila)	99.13	76.62	59.85	82.73	79.59
ours(ACDC)	99.00	62.73	62.54	88.70	78.24
ours(SMO)	99.20	72.17	61.58	87.05	80.00
ours (Dro_ACDC)	99.12	60.91	62.55	90.23	78.20
ours (Dro_SMO)	99.18	78.84	63.06	85.61	81.67
ours (ACDC_SMO)	99.27	72.48	65.57	85.08	80.60
ours(all)	99.06	65.47	61.60	85.75	77.97

in Table 6. In the Table, ours(Dro_ACDC) is the result with a mixture of Drosophila and ACDC Word Patches, ours(Dro_SMO) is the result with a mixture of Drosophila and SMO Word Patches, ours(ACDC_SMO) is the result with a mixture of ACDC and SMO Word Patches, and ours(all) indicates the result with a mixture of Word Patches of Drosophila, ACDC and SMO datasets.

When some Word Patches were mixed, higher accuracy was obtained than the Word Patches created from only one dataset. When ACDC is the target, Word Patches including SMO dataset achieved higher accuracy. However, when we mix the Word Patches of ACDC, SMO, and Drosophila, the accuracy is higher than that of baseline, but lower than that of the Word Patches created from each dataset. This is because there is too much different information when the attention is used in the Mix Transformer. When the sampling percentage increased, the accuracy decreased as shown in Table 1. Therefore, the similarities between the target image and many Word Patches may not be good for training. These results indicate that it is not enough to simply use Word Patches created from a large number of datasets, but that it is necessary to use Word Patches appropriate for the target.

5 CONCLUSION

We proposed a method to improve the TransUNet by using Word Patches created from the other medical datasets. Through experiments with a small number of training images, we confirmed that the segmentation accuracy was improved. In this paper, we have selected a dataset to create Word Patches manually. In the future, we would like to create Word Patches dynamically according to the property of target data.

ACKNOWLEDGEMENTS

This research is partially supported by JSPS KAKENHI Grant Number 21K11971.

REFERENCES

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

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

Chu, P., Wang, J., You, Q., Ling, H., and Liu, Z. (2021). Transmot: Spatial-temporal graph transformer for multiple object tracking. *arXiv preprint arXiv:2104.00194*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gerhard, S., Funke, J., Martel, J., Cardona, A., and Fetter, R. (2013). Segmented anisotropic sstem dataset of neural tissue. *figshare*, pages 0–0.

Guibas, J., Mardani, M., Li, Z., Tao, A., Anandkumar, A., and Catanzaro, B. (2021). Efficient token mixing for transformers via adaptive fourier neural operators. In *International Conference on Learning Representations*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Meinhardt, T., Kirillov, A., Leal-Taixe, L., and Feichtenhofer, C. (2022). Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8844–8854.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., and Gehler, P. V. (2021). Towards total recall in industrial anomaly detection. *CoRR*, abs/2106.08265.

Sakaridis, C., Dai, D., and Van Gool, L. (2021). Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775.

Sethi, A. et al. (2021). Wavemix: Multi-resolution token mixing for images.

Tan, C.-H., Chen, Q., Wang, W., Zhang, Q., Zheng, S., and Ling, Z.-H. (2021). Ponet: Pooling network for efficient token mixing in long sequences. *arXiv preprint arXiv:2110.02442*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090.

Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., and Yan, S. (2022). Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10819–10829.

Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., and Wei, Y. (2022). Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pages 659–675. Springer.