IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 August 2024, accepted 23 August 2024, date of publication 2 September 2024, date of current version 10 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3452769

RESEARCH ARTICLE

DeformableFormer for Classification of the Presence or Absence of Pancreatic Tissue Fragments in Pancreatic Disease

TAIJI KURAMI[1], TAKUYA ISHIKAWA[2], AND KAZUHIRO HOTTA[1]
[1]Meijo University, Nagoya 468-0073, Japan
[2]Nagoya University, Nagoya 464-8601, Japan
Corresponding author: Taiji Kurami (190442060@ccalumni.meijo-u.ac.jp)

ABSTRACT The purpose of this paper is to classify from an unstained image whether it is available for examination or not, and to exceed the accuracy of visual classification by specialist physicians by machine learning. Currently, Vision Transformer(ViT) and MetaFormer based PoolFormer have shown high accuracy in the image classification task. However, the pancreatic tissue fragment is a part of the image and has a complex shape, so the Vision Transformer, which processes the entire image, and the PoolFormer, which uses localized but fixed Convolution and Pooling, cannot classify it well. To address the problem, we require localized and image-specific feature extraction depending on the shape of a target. Therefore, we propose DeformableFormer, which enables local and dynamic feature extraction depending on the shape of the classification target in each image. To evaluate our method, we classify two categories of pancreatic tissue fragments; available and unavailable for examination. We demonstrated that our method outperformed the accuracy by specialist physicians and conventional methods such as ViT, Poolformer and the method using contrastive learning.

INDEX TERMS Deformable convolution, image classification, MetaFormer.

I. INTRODUCTION

Currently, Endoscopic Ultrasound-Fine Needle Aspiration (EUS-FNA) is used to examine pancreatic cancer. It is an examination using EUS to insert a thin needle into the tumor and collect pancreatic tissue fragments. Then collected pancreatic tissue fragments are then stained to classify whether they are pancreatic cancer. However, staining and visual inspection are time consuming. After staining, if it is determined that the pancreatic tissue fragment cannot be examined because it has not been acquired sufficiently, the acquisition must be done on the other day. This is a big problem that increases the burden on the patient. Therefore, it is desirable to be able to classify them in the state before staining, but it is difficult even for medical specialists to classify them whether pancreatic tissue fragments have been acquired sufficiently before staining or not. Therefore, the objective of this study is to exceed the classification accuracy of medical specialists using machine learning on before staining images.

In recent years, Transformer [4], [5], [21], [27] has shown high accuracy in computer vision. Since Vision Transformer (ViT) [6] which uses a simple Transformer for image classification was introduced, various models [19], [26], [33] have been developed to achieve higher accuracy in image classification [24], [28], object detection [10], [18], [22], [23], image generation [11], [13], [20], and various other tasks [29], [30]. MetaFormer [31], a model that generalizes ViT [8], has been proposed. The reason why Transformer-based methods achieved high accuracy has been attributed to mix the information among tokens called TokenMixer as shown in Fig 1. However, the PoolFormer [31] which changed the TokenMixer part to Pooling and MLP-Mixer [25] which changed the TokenMixer part to

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamad Forouzanfar.

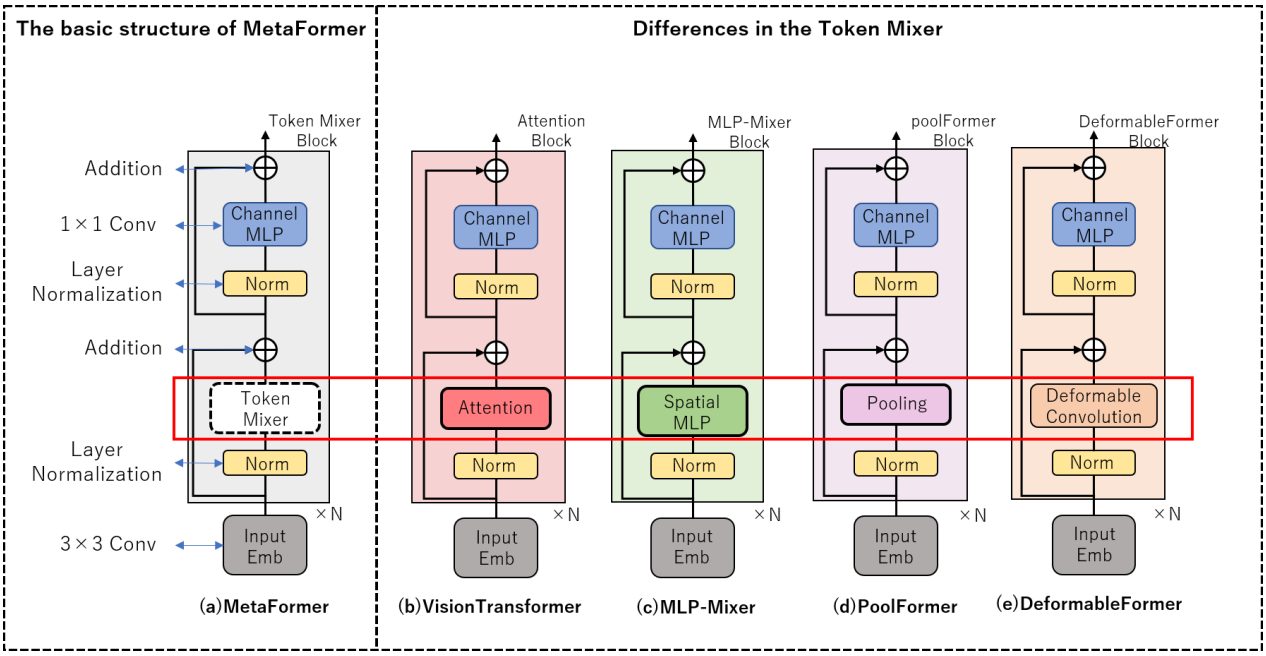


FIGURE 1. (a)MetaFormer [31]: A model that generalizes Vision Transformer [8]. (b) Vision Transformer [8]: Token Mixing part is replaced by Attention. (c) MLP-Mixer [25]: Token Mixing part is replaced by MLP. (d) PoolFormer: Token Mixing is replaced by Pooling. (e) The proposed DeformableFormer: The Token Mixing part is replaced by Deformable Convolution [7].

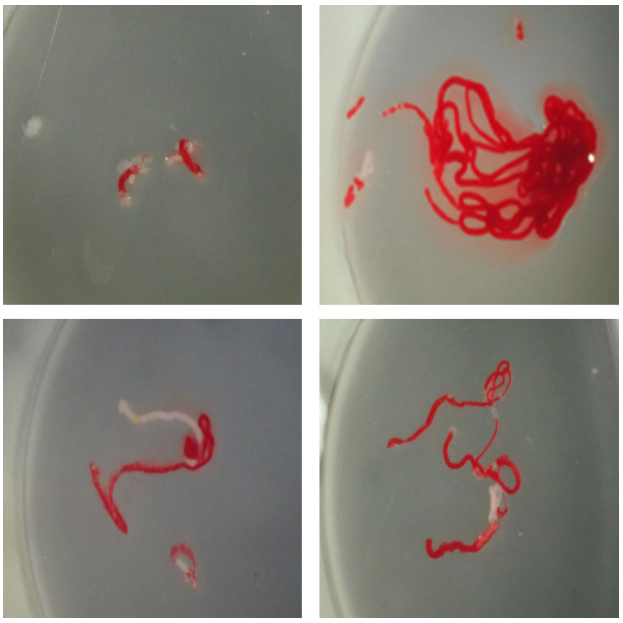


FIGURE 2. Example images of pancreatic tissue fragments used in this study.

MLP showed the similar performance. This suggests that the Transformer structure itself, not the TokenMixer part, was responsible for the performance.

As shown in Fig 2, the images used in this study have many background areas and blood portions that are other than the pancreatic tissue fragment which is the classification target. In addition, the shape of the pancreatic tissue fragments to be classified varies in each image and is complex. When we use ViT [8] for this classification problem, there is the problem that extra areas such as the background and blood are affected, and the accuracy of ViT is not so high because ViT uses attention over entire image. In addition, when we use Poolformer [31], a fixed kernel size does not extract good features due to complex shapes of pancreatic tissue fragments while we extract local features by local average pooling. Therefore, we need to extract features effectively from pancreatic tissue fragments which are different size and complex shape. To extract features well, we propose DeformableFormer which incorporates Deformable Convolution [7] into MetaFormer [31]. This enables the successful learning of the pancreatic tissue fragment to be classified. We realize the classification of before staining images which has been difficult even for medical specialists.

To validate the proposed method, we perform an image classification of two classes of pancreatic tissue fragments; available and unavailable for examination. When we compared our method with the accuracy of a medical specialist, the proposed method was improved 5.20% for the accuracy rate, 2.25% for the precision rate, 4.13% for the recall rate, and 10.72% for the specificity rate. Furthermore, our method outperformed comparison methods such as ViT [8], PoolFormer [31], Resnet [15], and Contrastive Learning based method [14].

Our contributions are as follows:

- Dynamic and localized feature extraction which is appropriate for pancreatic tissue fragment images.

- Pre-stained images were used instead of post-stained images, and classification accuracy was improved even in this condition. Therefore, when our method will be used in real world, it can save time for staining and provide inspection support.
- Existing methods could not outperform the classification accuracy of visual inspection by specialists in the unavailable class for examination. On the other hand, our proposed method could outperform the classification accuracy of visual inspection by specialists in the unavailable class for examination.

The structure of this paper is as follows. Section II describes the details of the proposed method. We show experimental results in Section III. Finally, conclusion and future works are described in Section IV.

II. RELATED WORKS

The previous study [16] using Contrastive Learning used pancreatic tissue fragment images after staining. However, despite the use of post-stained images, the classification accuracy of the unavailable class has not improved compared to the classification accuracy of the medical specialist. Therefore, in this study, we use a Transformer-based method instead of Contrastive Learning. By doing so, we aim to improve the accuracy by using only pre-stained images instead of post-stained pancreatic tissue fragment images.

This study is an image classification study, so the ViT [8] could be used. This method decomposes images into patches and treats them like words, and learns which patches are important based on the relationship between each patch through attention. Therefore, ViT learns global information because it use attention over an entire image. However, as shown in Fig 2, the pancreatic tissue fragment to be classified is only a part of the entire image. Therefore, even if the image is decomposed into patches and the relationship between all patches is learned, it would be difficult to classify the pancreatic tissue fragments that are the classification target because there are too many patches that include background and blood. In other words, local information also needs to be learned.

As shown in Fig 1, MetaFormer [31] is a general architecture that does not specify a Token Mixer and uses the same other configuration as the ViT. Token Mixer is the "Attention" part of ViT, which mixes tokens. First, the input x is processed by an input embedding like the patch embedding in ViT, as in the following equation (1).

$$X = InputEmb(x) \tag{1}$$

where x is the input image. The embedded token X is then fed into a MetaFormer block that contains two remaining sub-blocks that are repeated. The first sub-block contains a Token Mixer to convey information between tokens. This subblock can be represented as in the following equation (2).

$$Y = TokenMixer(Norm(X)) + X \tag{2}$$

where Norm() is the normalization such as Layer Normalization [1], and TokenMixer() indicates a module that mainly mixes token information. Finally, the second sub-block consists of a two-layer MLP with mainly nonlinear activation as shown in the following equation (3).

$$Z = ChannelMLP(Norm(Y)) + Y \tag{3}$$

Therefore, Metaformer [31] is a generalized model of ViT, and high accuracy was achieved for general image classification problem. Thus, we consider that better classification accuracy than specialist physicians should be obtained by using effective Token Mixer for pancreatic tissue fragments while maintaining the structure of the MetaFormer.

Local processing is considered important in this study because of the small size of the pancreatic tissue fragments to be classified. Therefore, PoolFormer is capable of local processing by average pooling. PoolFormer [31] is a simple Average Pooling version of Metaformer's Token Mixer shown in Fig 1. In average pooling, features are extracted by sliding a fixed kernel size from the upper left to the lower right of the image. Therefore, unlike ViT, it is possible to learn local features. However, as shown in Fig 2, the pancreatic tissue fragments to be classified have complex shapes, so it is considered that the fixed kernel size cannot extract good features for classification. Therefore, local feature extraction alone is not sufficient, and dynamic feature extraction depending on on the shape of target in each image is required. In this paper, we propose DeformableFormer [16] to extract effective features for classification from pancreatic tissue fragments of small and complex shape.

III. PROPOSED METHOD

As shown in section II, the first problem is that methods using the processing over the entire image such as ViT [8], are unable to successfully extract features from pancreatic tissue fragments, which are only a part of the image. The second problem is that methods based on local feature extraction with a fixed kernel size such as PoolFormer [31] cannot extract features from pancreatic tissue fragments with complex shape. To address the above two problems, we require localized and image-specific feature extraction depending on the shape of a target. Therefore, the proposed method focuses on Deformable Convolution to change the receptive field according to the scale and shape of the classification target in each image. Since the accuracy of image classification of Metaformer is higher than conventional CNN, we propose the DeformableFormer that we use Deformable convolution as a token mixer in Metaformer in order to extract suitable features for classifying pancreatic tissue fragment images. The details of the DeformableFormer are presented in Sections III-A and III-B.

A. DEFORMABLEFORMER

The architecture of the proposed DeformableFormer is shown in Fig 3. In addition, Table 1 summarizes the image size, patch size, number of channels, kernel size, MLP Ratio, and

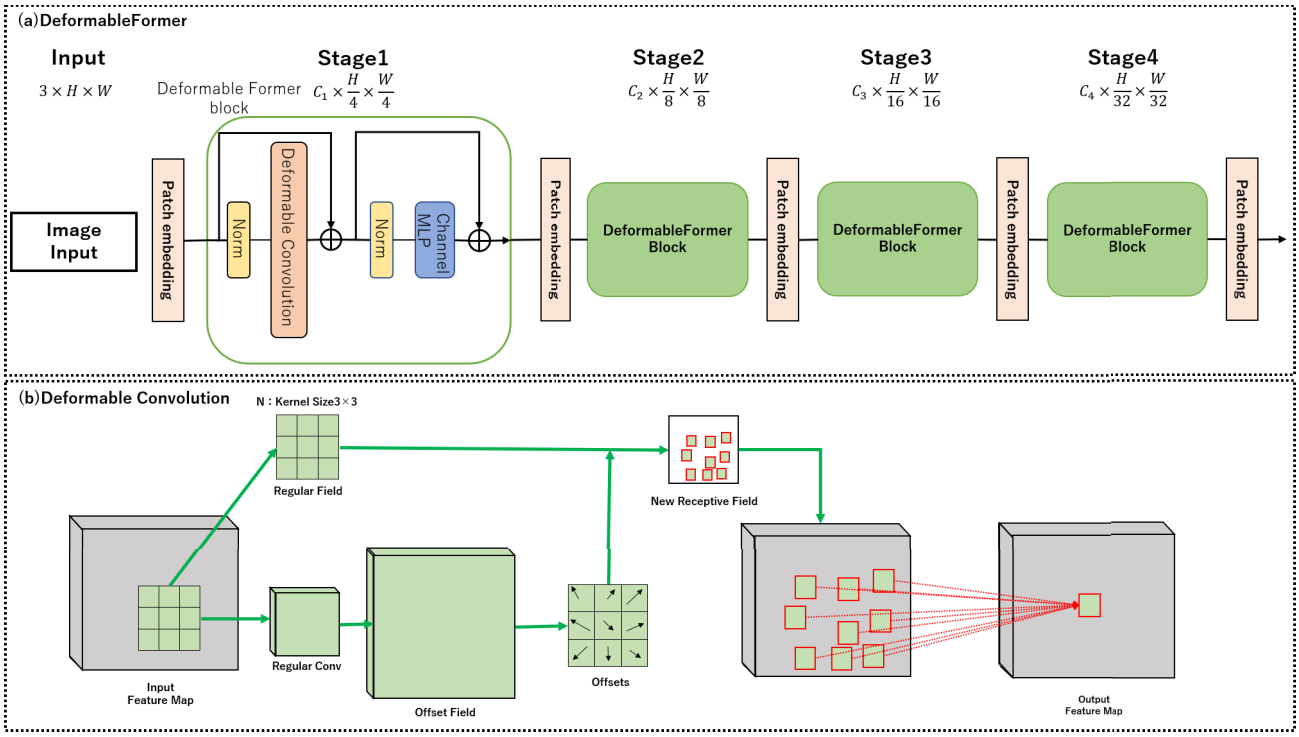


FIGURE 3. (a) Architecture of the DeformableFormer. (b) Architecture of Deformable Convolution [7] in the Token Mixing of DeformableFormer Block.

TABLE 1. Details of DeformableFormer architecture.

Stage	size	Layer		DeformableFormer
1	$\frac{H}{4} \times \frac{W}{4}$	Patch Embedding	Patch Size	7×7, stride 4
			Channel	64
		DeformableFormer Block	Kernel Size	3×3, stride 1
			MLP Ratio	4
			Block	2
2	$\frac{H}{8} \times \frac{W}{8}$	Patch Embedding	Patch Size	3×3, stride 2
			Channel	128
		DeformableFormer Block	Kernel Size	3×3, stride 1
			MLP Ratio	4
			Block	2
3	$\frac{H}{16} \times \frac{W}{16}$	Patch Embedding	Patch Size	3×3, stride 2
			Channel	320
		DeformableFormer Block	Kernel Size	3×3, stride 1
			MLP Ratio	4
			Block	6
4	$\frac{H}{32} \times \frac{W}{32}$	Patch Embedding	Patch Size	3×3, stride 2
			Channel	512
		DeformableFormer Block	Kernel Size	3×3, stride 1
			MLP Ratio	4
			Block	2

number of blocks at each stage in DeformableFormer. From Fig 3 and Table 1, DeformableFormer has a hierarchical structure in which the image size decreases from input to output, similar to CNN [9]. We use group normalization in the normalization layer. DeformableFormer first decomposes the image into patches and performs Deformable Convolution [7] in the DeformableFormer Block at each stage. By using Deformable Convolution [7] as a Token Mixer, it is expected that we can extract features according to the location and shape of a target. We believe that this allows for more localized and dynamic feature extraction, and thus allows for successful learning and classification of even small objects such as pancreatic tissue fragments.

B. DETAILS OF DEFORMABLEFORMER BLOCK

DeformableFormer Block is based on the MetaFormer [31] which is a generalized version of the ViT [8], and performs Deformable Convolution [7] in a Token Mixer to extract features according to the shape of a target. Deformable Convolution in Token Mixer is shown in Equation (4).

$$y(p_0) = \sum_{p_n \in R} \omega(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{4}$$

where x denotes the function to extract pixel values from the coordinate positions. p_0 denotes the pixel that is the center of the kernel in the input image. The p_n denotes the relative position where the kernel is convolved with respect to the input image. Therefore, from Figure 3 and Equation (4), the pixel for convolution is changed dynamically using offset *Deltap* with respect to the input feature. This allows the input data to be cut out irregularly. Therefore, the convolution position is $x(p_0 + p_n + \Delta p_n)$ because it is obtained from the input image. The inner product of the cropped data and the weight ω of convolution is taken. This enables convolution according to the classification target. Therefore, Deformable Convolution uses an additional convolution layer to compute the offset field. This offset field is used to add deformations to the standard filter. Therefore, additional convolution operations are required, which increases the computational complexity. However, the effectiveness of

TABLE 2. Details of the images in the dataset.

Class	Number of original images	After Augmentation	original image size	Image size after cropping
Available	145	290	4608×3456	1600×1600
Unavailable	28	252	4608×3456	1600×1600

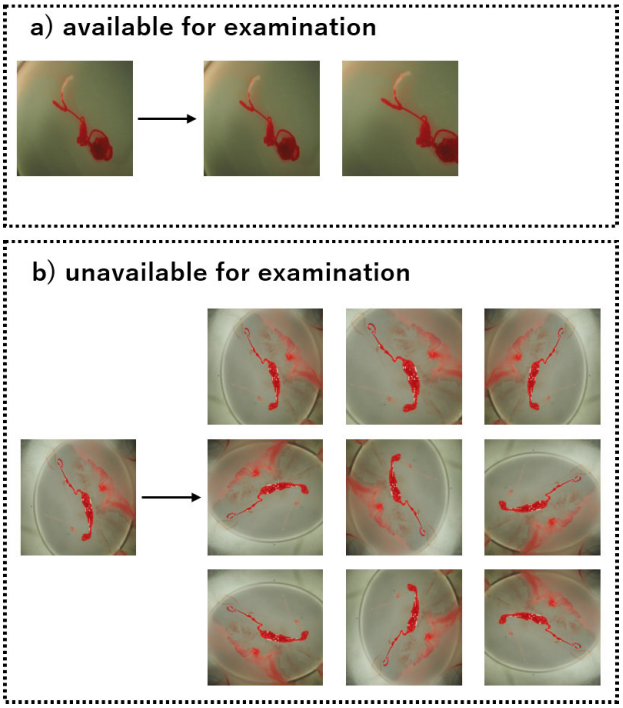


FIGURE 4. (a) The number of available images for examination is increased by a factor of two: the original image and the cropped image. (b) The number of unavailable images for examination is increased by a factor of nine: the original image and the cropped image, in addition, the inverted image and the image rotated by 90 degrees.

Deformable Convolution is that, unlike standard fixed-size convolution kernels, it can adaptively extract features for different parts of the image.

IV. EXPERIMENTS

A. IMPLEMENTATION DETAILS AND EVALUATION METHODS

We use a pancreatic tissue fragment dataset provided by Nagoya University Graduate School of Medicine. The dataset consists of images before staining. We have two classes; 145 images available for examination and 28 images unavailable for examination. The image datasets used in this paper are the images of actual pancreatic disease patients, taken by a specialist physician. Therefore, we believe that the validation can be performed in a situation close to the real environment. Since this is an imbalanced dataset (especially, the number samples in unavailable class is small), the number of images is increased by data augmentation. Data augmentation for training images in two classes is shown in Fig 4. By data augmentation shown in the Figure, the number of available images for examination is increased by a factor of two: the original image and the cropped image. The number of images that are not available for examination is increased by a factor of nine: the original image and the cropped image, in addition, the inverted image and the image rotated by 90 degrees.

We used cross validation because of the small number of images. All images are divided into 28 sets, and 27 sets are used for training and remaining 1 set is used for evaluation. Thus, 28-fold cross-validation is used, in which one test data is shifted and we trained a model 28 times. The reason for using 28 sets is that there are only 28 unavailable images for examination, so one unavailable image for examination is always included in the test, and the remaining 27 images are used for training to make the learning process more efficient. In addition, since the pixel size of the original image is 4608×3456 pixels, the original image is too big to compute with GPU. Thus, the resized images of 1600×1600 pixels are used. This experiment was conducted with a batch size of 5, an epoch of 50, and a learning rate of 0.001. Table 2 summarizes the image details of the data sets presented above.

Because this experiment is a binary classification of available and unavailable for examination, a confusion matrix is used. A confusion matrix is a table that summarizes the classification results and is used as a measure of the performance of binary classification. We use four evaluation measures computed from a confusion matrix; accuracy rate, precision rate, recall rate, and specificity rate.

Accuracy rate is a measure of how well the overall prediction result matches the true value. The higher value is better. Therefore, the accuracy rate indicates the classification accuracy of two classes; available and unavailable for examination. The formula is shown in Equation (5).

$$Accuracy = (TP + TN)/(TP + FP + FN + TN) \quad (5)$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative. In this experiment, positive class means available class and negative class means unavailable class.

Precision rate indicates the percentage of true positive in samples predicted as positive. Therefore, the precision rate indicates the accuracy of available images for examination. If there are many false positives, precision rate decreases. The formula is shown in Equation (6).

$$Precision = TP/(TP + FP) \quad (6)$$

Recall rate is the percentage of true positive in positive samples. Therefore, the recall rate also indicates the accuracy of available images for examination. If there are false

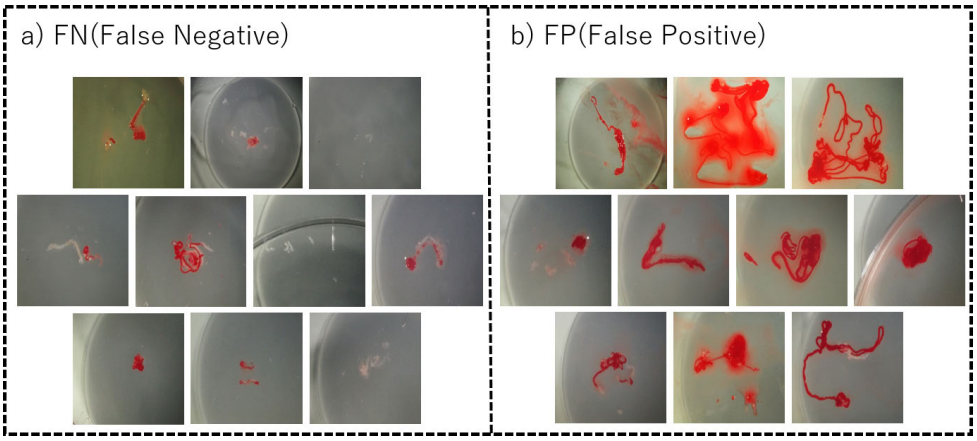


FIGURE 5. Images misclassified by the proposed method. (a) 10 images misclassified by our method in available class. (b) 10 images misclassified by our method in unavailable class.

TABLE 3. Confusion matrices by medical specialties, existing methods, and the proposed method. Available means images that are available for examination, and Unavailable means images that are not available for examination. The number means the number of images classified.

Medical Specialist		Prediction	
		Available	Unavailable
True Label	Available	129	16
	Unavailable	13	15

Resnet34		Prediction	
		Available	Unavailable
True Label	Available	120	25
	Impossible	19	9

Contrastive Learning		Prediction	
		Available	Unavailable
True Label	Available	131	14
	Unavailable	13	15

MetaFormer(Attention)		Prediction	
		Possible	Unavailable
True Label	Available	142	3
	Unavailable	18	10

PoolFormer		Prediction	
		Available	Unavailable
True Label	Available	136	9
	Unavailable	14	14

ConvFormer		Prediction	
		Available	Unavailable
True Label	Available	131	14
	Unavailable	16	12

CAFormer		Prediction	
		Available	Unavailable
True Label	Available	20	125
	Unavailable	2	26

Proposed method		Prediction	
		Available	Unavailable
True Label	Available	135	10
	Unavailable	10	18

negatives, recall rate decreases. The formula is shown in Equation (7).

$$Recall = TP/(TP + FN) \tag{7}$$

Specificity rate is the percentage of true negative in unavailable images for examination. In other words, the specificity rate indicates whether unavailable class for examination is correctly classified. If unavailable images for examination are mis-classified as positive, then pancreatic tissue fragment cannot be examined after staining, and acquisition of tissue fragment must be done again on the other day. Since this is the worse case, our primary goal is to improve the value of specificity rate. The formula is shown in Equation (8).

$$Specificity = TN/(FP + TN) \tag{8}$$

In this experiment, it is difficult to classify whether a pancreatic tissue fragment is available or unavailable for examination when the amount of tissue fragment is small. In particular, it was very difficult even for medical specialists to classify the unavailable samples for examination before staining images.

B. COMPARISON RESULTS

This section shows comparison results with medical specialists and conventional methods. Note that ground truth of each image is the classification result based on post-staining inspection. The accuracy of the specialist is the classification result based on visual inspection of the pre-stained images. Thus, the accuracy of human specialist is different from ground truth. We used two conventional methods; Resnet34 [15] used in [12], the method using contrastive learning [12] that brings the features between pre-stained images and post-stained images with the same class closer. We also evaluated MetaFormer(Attention) [31] and PoolFormer [31] which are related methods to our method. In addition, a comparison was made with ConvFormer [32] and CAFormer [32], the latest methods that utilize the MetaFormer structure. The Loss for each method in this experiment uses Cross-entropy Loss with class balancing weights [2] for imbalanced data. The class Weight of the images available for examination is w0 and the class

TABLE 4. Accuracy at four evaluation measures. Increase Rate is the degree of increase compared to the accuracy of medical specialists. Bold letters indicate the best accuracy in all methods.

	Accuracy rate		Precision rate		Recall rate		Specificity rate	
	Accuracy	Increase rate	Accuracy	Increase rate	Accuracy	Increase rate	Accuracy	Increase rate
medical specialist	83.24%	-	90.85%	-	88.97%	-	53.57%	-
Resnet34	74.57%	-8.67%	86.33%	-4.52%	82.76%	-6.21%	32.14%	-21.43%
Contrastive Learning	84.39%	1.15%	90.97%	0.12%	90.34%	1.37%	53.57%	0.00%
MetaFormer(Attention)	87.86%	4.62%	88.75%	-2.10%	**97.93%**	**8.96%**	35.71%	-17.86%
PoolFormer	86.71%	3.47%	90.67%	-0.18%	93.79%	4.82%	50.00%	-3.57%
ConvFormer	82.66%	-0.58%	89.12%	-1.73%	90.34%	1.37%	42.86%	-10.71%
CAFormer	25.59%	-57.65%	90.91%	0.06%	13.79%	-75.18%	**92.86%**	**39.29%**
Proposed method	**88.44%**	**5.20%**	**93.10%**	**2.25%**	93.10%	4.13%	64.29%	10.72%

TABLE 5. Result of ablation study w/o "Deformable". Accuracy at four evaluation measures. Increase Rate is the degree of increase compared to the accuracy of MetaFormer using simple convolution.

	Accuracy rate		Precision rate		Recall rate		Specificity rate	
	Accuracy	Increase rate	Accuracy	Increase rate	Accuracy	Increase rate	Accuracy	Increase rate
MetaFormer(Conv)	78.61%	-	86.00%	-	88.97%	-	25.00%	-
Proposed method	88.44%	9.83%	93.10%	7.10%	93.10%	4.13%	64.29%	39.29%

TABLE 6. our: DeformableFormer using only pancreatic tissue fragment images without pre-trained models. PoolFormer: A method that uses only pancreatic tissue fragment images for training without prior training with ImageNet. Pre-Train (overall): A method that updates the entire model using PoolFormer pre-trained in ImageNet. Pre-Train (final layer only): A method in which only the output layer is updated using PoolFormer pre-trained in ImageNet and the rest of the layers are fixed.

	Accuracy rate		Precision rate		Recall rate		Specificity rate	
	Accuracy	Increase rate	Accuracy	Increase rate	Accuracy	Increase rate	Accuracy	Increase rate
Proposed method	88.44%	-	93.10%	-	93.10%	-	64.29%	-
PoolFormer	86.71%	-1.73%	90.67%	-2.43%	93.79%	0.69%	50.00%	-14.29%
Pre-Train(overall)	56.07%	-32.37%	87.91%	-5.19%	55.17%	-37.93%	60.71%	-3.58%
Pre-Train(final layer only)	50.87%	-37.57%	85.71%	-7.39%	49.66%	-43.44%	57.14%	-7.15%

Weight of the images unavailable for examination is w1 in the following equation.

$$w_0 = ((n0 + n1)/2)/n0 \tag{9}$$

$$w_1 = ((n0 + n1)/2)/n1 \tag{10}$$

where n0 is the 145 images available for examination and n1 is the 28 images unavailable for examination.

Table 3 shows the confusion matrices of each method. Available in Table indicates available images for examination, and Unavailable means unavailable images for examination. The number in the Table indicates the number of images classified. From Table 3, we see that the conventional methods using contrastive learning [12], MetaFormer using Attention [31], PoolFormer [31] and ConvFormer [32] increased the accuracy of positive class compared to the specialists. Conversely, the number of images that could be correctly classified as unavailable for examination has not increased. It is also clear that the latest method, CAFormer [32], is biased toward unavailable for examination classification and does not classify well. In contrast, the proposed method increased in the number of correctly classified images in both available and unavailable for examination. Therefore, we believe that our method can extract good features from pancreatic tissue fragments of small and complex shape.

Table 4 shows the accuracy at four evaluation measures. Accuracy shows the accuracy of each of four evaluation measures, and Increase Rate shows the rate of increase in the accuracy of each method compared to the accuracy of the medical specialist. In addition, bold letters indicate the best accuracy.

Table 4 shows that the method using contrastive learning [12] improved the accuracy in the three evaluation measures in comparison with medical specialist. However, the accuracy of the specificity rate were not improved. This indicates that the method using constrastive learning [12] is not able to classify unavailable images for examination well. In addition, MetaFormer(Attention) [31] and PoolFormer [31] improved the accuracy in two evaluation measures in comparison with medical specialist, but pecision rate and specificity rate decreased. Because the contrastive learning based method [12] and PoolFormer [31] perform feature extraction at a fixed size such as convolution and pooling, we considered that the accuracy is reduced when the amount of pancreatic tissue fragments to be classified is small. Similarly, the MetaFormer(Attention) [31] uses global attention over the entire image, we considered that the accuracy decreased when the amount of pancreatic tissue fragments to be classified is small because the features can not be extracted well. ConvFormer [32] and CAFormer [32], the newest methods in the MetaFormer structure, also showed reduced classification accuracy. ConvFormer [32] uses Depthwise Separable Convolution to extract features with a defined size, similar to Poolformer. Therefore, as with PoolFormer, it can classify with accuracy not much different

from that of specialists in the available for examination class, but it cannot classify well in the unavailable class for examination. CAFormer [32] uses Attention in the latter block. It can be considered that the global feature extraction with this was not suitable for the pancreatic tissue fragment images and therefore could not classify them correctly.

In contrast, the proposed method improved the accuracy at all four evaluation measures in comparison with medical specialists. Therefore, the proposed method shows overall higher performance than other methods. In particular, the best specificity rate by our method demonstrated that it is highly effective for classifying images that are not available for examination, which is the primary goal of the study.

In addition, the images misclassified by the proposed method are shown in Fig 5. Figure 5(a) are the 10 images misclassified by the proposed method in the available classes. The pancreatic tissue fragments were considered too small to contain sufficient pancreatic tissue fragments. Figure 5(b) are the 10 images misclassified by our method in unavailable class. There were many blood portions that were not subject to classification, and we considered the pancreatic tissue fragments to be well contained by our method.

C. ABLATION STUDY

In this section, we conduct experiments to verify whether Deformable Convolution was a factor for improving the accuracy in DeformableFormer. The effectiveness of ''Deformable'' was evaluated by comparing the proposed DeformableFormer with the method using a simple convolution in a TokenMixer of MetaFormer.

The result of ablation study is shown in Table 5. The table 5 shows that Deformable Convolution [7] improved the accuracy in all four evaluation indices, with a accuracy rate of 9.83%, a precision rate of 7.10%, a recall rate of 4.13%, and a specificity rate of 39.29%. Therefore, ''Deformable,'' is considered to be effective for classifying pancreatic tissue fragments of small and complex shape.

D. VISUALIZATION RESULTS

In this section, the visualization results of the offset in Fig 3(b) are shown in Fig 6 in order to confirm that Deformable Convolution [7] is functioning properly. Fig 6(a) shows two images used for visualization. Fig 6(b) is to check which part is used for convolution in the case of the background area. The green point is the center point of the convolution and the black point is the location used for the convolution. Fig 6(c) shows which part of the blood was used for convolution. Fig 6 (d) shows which part of the pancreatic tissue fragment, which is the classification target, is used for convolution. Fig 6 shows that Deformable Convolution is functioning properly because the locations used for feature extraction differ among the background, blood, and pancreatic tissue fragments.

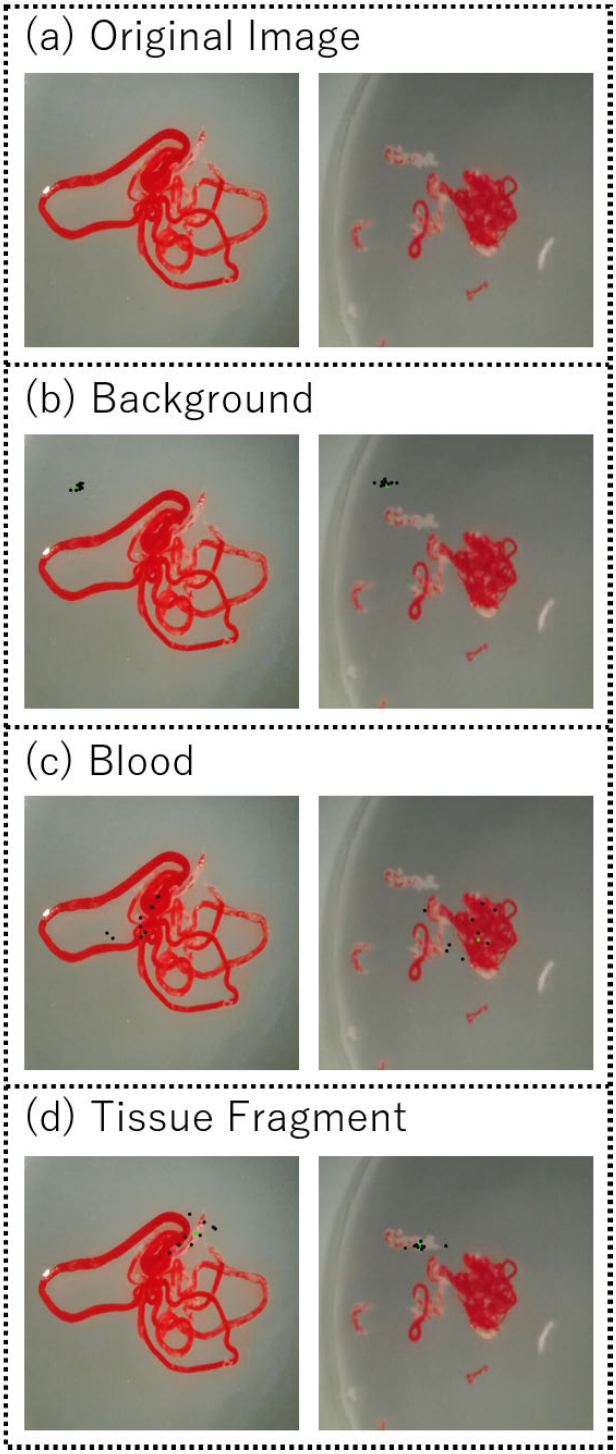


FIGURE 6. Visualized results of offset in Deformable convolution. (a) Image used for visualization. (b) Visualization result for the background area. (c) Visualization result for blood area. (b) Visualization result for pancreatic tissue fragment.

E. COMPARISON WITH PRE-TRAINED MODELS

The pancreatic tissue fragment images used in this study are actual images of people with pancreatic disease. Therefore, the number of images for study is very small. Therefore, we used our proposed method, DeformableFormer, to extract

suitable features and to perform learning efficiently even with a small number of images.

As one of the validations, we used a PoolFormer pre-trained on ImageNet and compared the results. When we use the pre-trained PoolFormer, two methods were used to update the parameters: the first method updates the entire model, while the second method updates only the output layer, leaving the rest fixed. Therefore, we compared the above two methods using trained models against DeformableFormer and Poolformer. The following table 6 is a comparison table of the two methods using the proposed DeformableFormer, Poolformer and PoolFormer pre-trained models. Table 6 shows that the use of pre-trained models significantly decreases the accuracy. This may be due to the fact that the images of the pancreatic tissue fragments used in this study and ImageNet, a common image dataset used in the pre-trained model, are too different. This indicates that it is important to learn efficiently with the number of pancreatic tissue fragment images that are available now, as in the proposed method, DeformableFormer.

V. CONCLUSION

The proposed method improved the accuracy of both classes compared to medical specialists and conventional methods. In particular, when we pay attention to the specificity rate which is the most important metric, Resnet and MetaFormer(Attention), PoolFormer decreased the accuracy by -21.43% and -17.86%, -3.57% in comparison with the medical specialist. Conventional method using contrastive learning did not improve the specificity rate. However, the proposed DeformableFormer improved the specificity rate by 10.72% compared to medical specialists. An increase in the specificity rate means an improvement in the classification of images that are not available for examination. This is because the DeformableFormer extracts features according to the shape of pancreatic tissue fragments. Thus, our method improved the classification accuracy even for images that are not available for examination is scarce.

Although specificity rate of our method outperformed medical specialist, the accuracy is not so high. This may be due to the fact that the number of training images that are unavailable for examination is extremely small. Therefore, we believe that countermeasures against imbalanced image data sets are needed. Therefore, we believe that the accuracy of specificity rate could be further improved by improving loss functions such as Class-Balanced Loss [6], Focal Loss [17], and LDAM Loss [3]. Therefore, the future challenge is to devise loss functions for imbalanced image data sets.

REFERENCES

[1] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[3] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Springer, Aug. 2020, pp. 213–229.

[5] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "A^2-Nets: Double attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[6] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9260–9269.

[7] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[9] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," in *Shape, Contour and Grouping in Computer Vision*, 1999, pp. 319–345.

[10] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[12] T. Ishikawa, M. Hayakawa, H. Suzuki, E. Ohno, Y. Mizutani, T. Iida, M. Fujishiro, H. Kawashima, and K. Hotta, "Development of a novel evaluation method for endoscopic ultrasound-guided fine-needle biopsy in pancreatic diseases using artificial intelligence," *Diagnostics*, vol. 12, no. 2, p. 434, Feb. 2022.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[14] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18661–18673.

[15] B. Koonce and B. Koonce, "ResNet 34," in *Convolutional Neural Networks With Swift for Tensorflow: Image Recognition and Dataset Categorization*, 2021, pp. 51–61.

[16] T. Kurami, T. Ishikawa, and K. Hotta, "DeformableFormer for classifying endoscopic ultrasound-guided fine-needle biopsy in pancreatic diseases," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2023, pp. 113–114.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands. Springer, Oct. 2016, pp. 21–37.

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[20] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

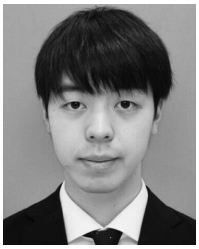
[21] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone selfattention in vision models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[23] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

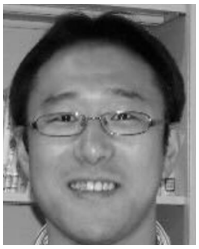
[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[25] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "MLP-mixer: An all-MLP architecture for vision," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24261–24272.
[26] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
[27] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12889–12899.
[28] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6450–6458.
[29] X. Cao, Y. Lian, K. Wang, C. Ma, and X. Xu, "Unsupervised hybrid network of transformer and CNN for blind hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024.
[30] X. Cao, Y. Lian, J. Li, K. Wang, and C. Ma, "Unsupervised multi-level spatio-spectral fusion transformer for hyperspectral image super-resolution," *Opt. Laser Technol.*, vol. 176, Sep. 2024, Art. no. 111032.
[31] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10809–10819.
[32] W. Yu, C. Si, P. Zhou, M. Luo, Y. Zhou, J. Feng, S. Yan, and X. Wang, "MetaFormer baselines for vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 896–912, Feb. 2024.
[33] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 538–547.



TAIJI KURAMI received the B.E. degree in electrical and electronic engineering from Meijo University, Aichi, Japan, in 2023. Since 2023, he has been a Master Course Student with Meijo University. His current research interest includes image recognition.

TAKUYA ISHIKAWA received the M.D. and Ph.D. degrees from Nagoya University School of Medicine, in 2001 and 2011, respectively. During his academic journey, he held various positions and gained valuable experience in gastroenterology. From July 2014 to June 2016, he was a Therapeutic Endoscopy Fellow with the Division of Gastroenterology, University of Calgary, Faculty of Medicine, Calgary, Canada. Prior to that, he was an Assistant Director with the Department of Gastroenterology, Japanese Red Cross Nagoya Daiichi Hospital, Aichi, Japan, from January 2011 to June 2014. He has been a Lecturer with the Department of Gastroenterology and Hepatology, Nagoya University Hospital, Nagoya, Japan, since January 2023. His research interests include endoscopic ultrasound (EUS) and EUS related procedures, especially in the pancreas and biliary diseases. Throughout his career, he has been actively involved in various professional organizations and has received several honors and awards, including the prize of Japan Biliary Association, in 2012, and the ENDO 2020/JGES Best Abstract Award, in 2020.



KAZUHIRO HOTTA received the B.S., M.S., and Ph.D. degrees in computer science from Saitama University, Saitama, Japan, in 1997, 1999, and 2002, respectively. From 1999 to 2002, he was a Research Fellow of Japan Society for the Promotion of Science (JSPS). From 2002 to 2010, he was an Assistant Professor with the University of Electro-Communications (UEC), Tokyo, Japan, and since 2002, he has been an Assistant Professor with UEC. From 2010 to 2018, he was an Associate Professor with Meijo University, Nagoya, Japan, where he has been a Professor, since 2018. He was a Visiting Scholar with the University of Maryland, in 2012. His research interests include pattern recognition, computer vision, and machine learning. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), the Information Processing Society of Japan (IPSJ), The Japanese Society for Artificial Intelligence (JSAI), and the IEEE Computer Society.