Med-T: A pixel-level position information aware transformer for medical visual task

Jianbo Huang[a], Ke Zhao[a], Jing Wang[a], Kexin Zhang*[b]
[a]Shenyang Institute of computing technology, Chinese Academy of Sciences, No.16 Nanping East Road, Shenyang, China; [b]College of Information Engineering of Liaoning University of Traditional Chinese Medicine, No. 79 Chongshan East Road, Shenyang, China
* Corresponding author: Kxzh@sina.com

ABSTRACT

The model of transformer structure occupies a dominant position in the field of multimodal large model. While previous studies have highlighted the potential of Visual Transformer (ViT) models, their reliance on large datasets poses challenges in domains like medicine, where obtaining extensive data can be difficult. In such scenarios, traditional convolutional neural networks (CNNs) often outperform transformer-based models due to their ability to capture pixel-level fine-grained information. In this paper, we proposed soft mask operation and fine-grained information awared visual transformer Med-T, a CNN-Transformer hybrid visual backbone network tailored for visual feature extraction task on limited datasets. Through extensive evaluation across three small datasets, Med-T consistently outperforms alternative approaches, showcasing the efficacy of leveraging the pixel-level position information extraction ability of CNN branch.

Keywords: Medical image processing, fine-grained information capturing, visual transformer, small sample

1. INTRODUCTION

In the modern era of medical informatics, the integration Artificial Intelligence(AI) technology into clinical diagnostics has become ubiquitous. Computer Vision (CV) is a crucial branch of artificial intelligence that enables computers to detect complex objects in clinical diagnosis. At the heart of CV lies the backbone network, a linchpin whose efficacy in extracting features profoundly influences downstream tasks' performance. Two predominant types of backbone models have been employed for images feature extraction. The first type harnesses classical Convolutional Neural Network (CNNs), leveraging convolutional kernels to glean features from neighboring regions surrounding a center pixel. The most representative CNNs are ResNet [1], ResNeXt [2], and UNet [3]. The other method adopts a transformer structure, integrating Natural Language Processing (NLP) techniques into computer vision tasks. It splits images into a series of pixel regions, subsequently transformed into token representations akin to NLP task. For example, Vision Transformer (ViT) [4], a pioneering transformer-based model, evenly segments images into uniform patches, augmenting each with positional encoding to discern spatial relationships.

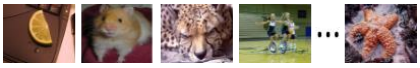
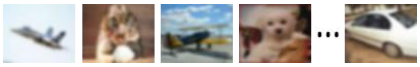
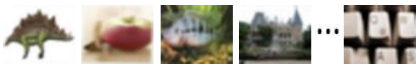
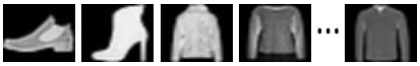
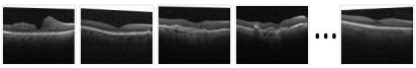
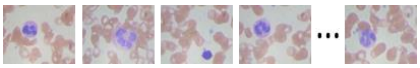
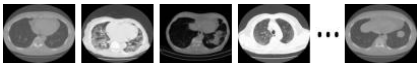
Although the model of transformer structure is outstanding in the field of multimodal large model. Dosovitskiy noted ViT's suboptimal performance on small datasets. Nevertheless, a significant persists, particularly in domains like medicine, where datasets are often severely limited. A poignant concern given the stark contrast between vast datasets like ImageNet [5] and meager medical datasets such as HAM10000 [6], Retinal OCT [7], Blood Cell Images [8], and COVID-CT [9]. As shown in Table 1, while ImageNet boasts millions of samples, medical datasets barely scrape a few thousand, exacerbating issues of class imbalance and hindering model generalization. For instance, despite HAM10000's ostensibly ample sample size, disparities between category counts underscore the challenges posed by imbalanced datasets.

Additionally, Figure 1 showcases the distribution of the A and B channels in the Lab color space for a randomly selected subset of 1000 samples for each dataset. Notably, with the exception of Fashion-MNIST dataset-comprising grayscale images with solely luminance information (thus both A and B channels are 0)-we observed a broad and comparable distribution across the channels in general-purpose datasets. Conversely, medical datasets tend to exhibit a more concentrated distribution. This discrepancy arises from the diverse nature of general-purpose datasets, which encompass

International Conference on Future of Medicine and Biological Information Engineering (MBIE 2024), edited by Yudong Yao, Xiaoou Li, Xia Yu, Proc. of SPIE Vol. 13270, 1327003 · © 2024 SPIE · 0277-786X · doi: 10.1117/12.3039162

images depicting various objects and scenes. In contrast, medical datasets primarily feature images of specific anatomical regions, resulting in more uniform color tones within the dataset.

Table 1. The number of samples in each dataset.

Dataset	Example	Samples
Genral purpose dataset		
ImageNet		14,000,000
ImageNet1K		1,000,000
Cifar10 [10]		60,000
Cifar100 [10]		60,000
FashionMNIST [11]		60,000
Medicine field dataset		
HAM10000		10,015
OCT-C8		24,000
Blood Cell Images		12,500
COVID-CT		742

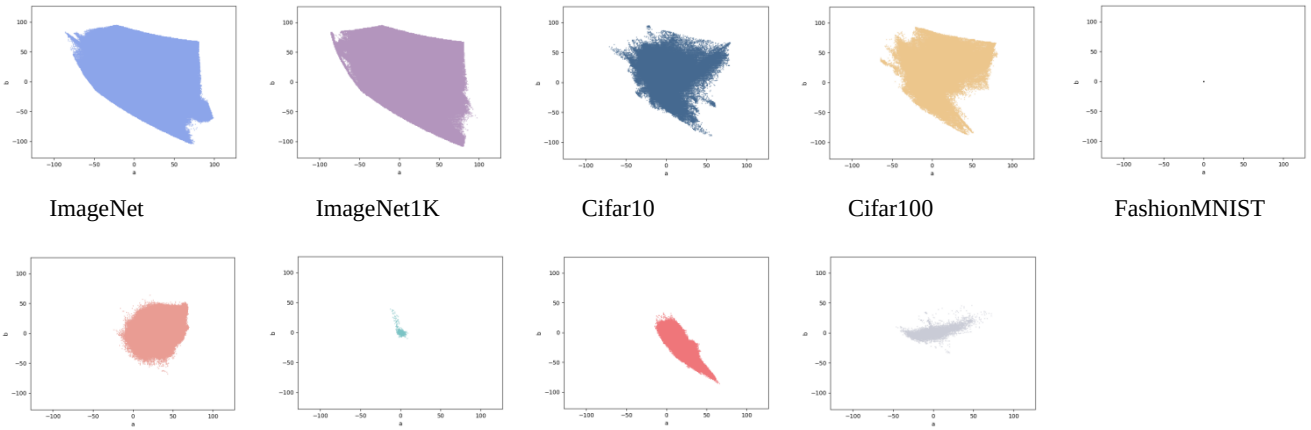


Figure 1. The A and B channels of each dataset sample in Lab space.

While transfer learning offers a solution to the challenge of limited data samples in certain domains, significant disparities between the source and target domains, or substantial variations in samples sizes, can impede direct knowledge transfer. In such cases, models may require multiple transfers or adaptations to effectively transition from the source to the target domain. Consequently, exploring the applicability of vision transformers within contexts of constrained sample sizes becomes imperative.

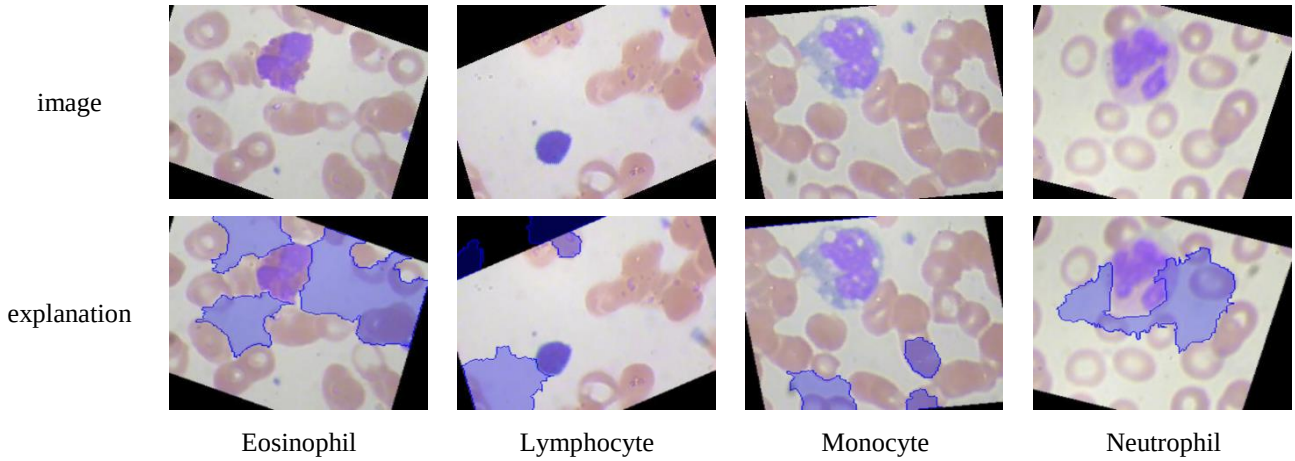


Figure 2. The explanation for prediction of ResNet50 on four types of cell slices.

In addition, due to the ViT model maps each patches to tokens and adds positional encoding to each token, it results in all pixels within a block having the same positional information, making it difficult to capture fine-grained pixel level position information within a patch. While reducing the size of each patch could capture finer details, it comes at the coast of significantly increased computational burden.

Furthermore, during the training stage, images typically encompass both foreground and background elements, with the foreground object often constituting the target of interest for detection. However, complex backgrounds can introduce noise and interfere with model training. Illustrated in Figure 2, where four types of cells share identical backgrounds distinguished solely but variations in the blue region, we employed ResNet50 which trained on Blood Cell dataset for cell image prediction and LIME [12] for resulting interpretation. We found that the distribution of blue area is not concentrated. Although the ResNet50 correctly predicted the category label, but certain unrelated background features were erroneously associated with detected targets, even the target area that needs to be detected is not considered as features. Such outcomes represent undesirable model behaviors that underscore the need for robust background handling mechanisms in computer vision tasks.

In response to these challenges, we proposed the Fine-grained information aware transformer Med-T — a hybrid architecture that combines the strengths of CNN and transformer visual backbone model. This model has exhibited superior feature extraction capability across multiple datasets compared to similar structured models in recent studies.

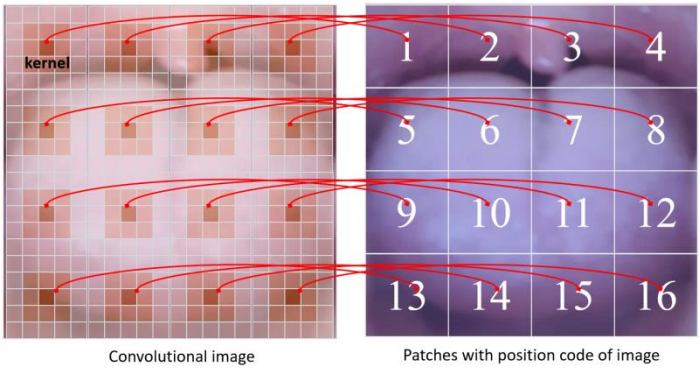


Figure 3. Convolution kernel captures fine-grained position information as a supplementation for patches.

In the Med-T, we employ CNN block with various structure as the encoder to capture pixel level information and generate convolutional images, as depicted in Figure 3. Subsequently, we employed the Scale-Invariant Feature Transform (SIFT) operator to creating a soft mask map and mask image that preserves both features points and their local neighborhoods. During the decoding stage, the mask image serves as the query in the decoder block, while the convolutional image encoded by CNN block acts as the key and value for decoding and mask the attention score map using soft mask map. Our contributions are summarized as follows:

- We proposed Med-T, a hybrid lightweight model architecture that utilizes a CNN block to encode convolutional image for decoder. This configuration enables the model to effectively encode fine-grained spatial relationships, enhancing visual transformer's ability to extract detailed positional information at the pixel level. Finally, the whole image feature can be represented by a low-dimensional vector for downstream task.
- We partition the convolutional feature map and direct its focus towards the mask image, thereby enhancing the model's attention on the feature point region.
- We introduce soft mask operation during the decode stage to extract and dilate feature points area, enabling the model to focus on relevant areas. It also allows the model to smoothly handle tokens at the edge of the mask.

2. RELATED WORK

Most of the previous visual backbone models can be divided into two main categories. The first category of backbone network is based on CNNs. CNNs are difficult to replace in fine-grained feature extraction tasks and have been used as the backbone network in multi-modal fields [13-17]. In CNNs, the receptive field of the model gradually increases as the number of convolution layers increases, which allows the model to extract high-level features. Therefore, in the multi-modal field, most studies mainly use the classic pre-trained CNNs as the backbone network to extract visual features for downstream task, such as ResNet, VGG [18], ResNeXt.

The models in the second category are based on transformer structure [19-26]. The ViT serves as a visual extractor in the CLIP [27] model, which splits images into a sequence of patches and aligns them with word tokens of text, this method converts visual task into natural language processing task. This operation facilitates the alignment of visual features with text features in multi-modal tasks. In the BLIP [28] model, an improved version of ViT is used as a visual feature extractor to encode images, which is also based on the ViT architecture. ViTL [29] is inspired by the ViT model, but it places the main computational burden on the transformer features fusion stage, which makes the model nearly ten times faster than previous VLP models. Ze et al. use a sliding window to capture information between different blocks, and reduces the amount of calculation [30]. In T2T-ViT [31], Yuan, Li et al. continuously merge two adjacent tokens, and the image is ultimately structured into one token. The Conformer [32] is a model that combine CNN with Transformer, interacting the local features of each stage with global features. Muhammad et al. proposed EdgeNeXt [33], this model is also based on CNN-ViT structure, which used a split depth-wise transpose attention (SDTA) encoder to effectively learn both local and global representations. In TransXNet [34], Meng et al. used a token mixer called D-Mixer to aggregate sparse global information and local details, which allows the network to see a wider range of contextual information.

3. METHOD

The primary objective of this study is to enhance the fine-grained position information-capturing capability of transformer-based models on small datasets. To achieve this, we introduced Med-T, a hybrid model amalgamating the CNN model’s local awareness capabilities, particularly effective on small datasets. To benchmark our model against established baselines, we conducted a comprehensive evaluation across various performance metrics.

In this section, we first present the architectural framework of our proposed model (Section 3.1). Subsequently, we defined the task at hand and elucidated the criteria for evaluation (Section 3.2).

3.1 Model architecture

The structure of our model is shown in Figure 4. Initially, we employed the SIFT operator to detect salient feature points within the image. Subsequently, we expanded the pixels corresponding to these feature points through dilation and create a mask to isolate areas outside these points. Following this, we resized all images to a uniform size.

When the mask image is segmented, each patch may contain a portion of the mask pixels. To generate the feature map, we employ a soft masking approach. Initially, we binarized the mask image with 0 pixels representing the masked area and 1 pixel for the unmasked area. Subsequently, we applied average pooling to divide the image into 14×14 feature maps. The value of each patch is inversely proportional to the number of mask pixels it contains; patches with more mask pixels have values closer to 0. Tokens with values lower than the average of all tokens are then truncated to 0. Finally, these tokens are tiled and transposed to produce the soft mask map, this stage is shown in the bottom of Figure 4.

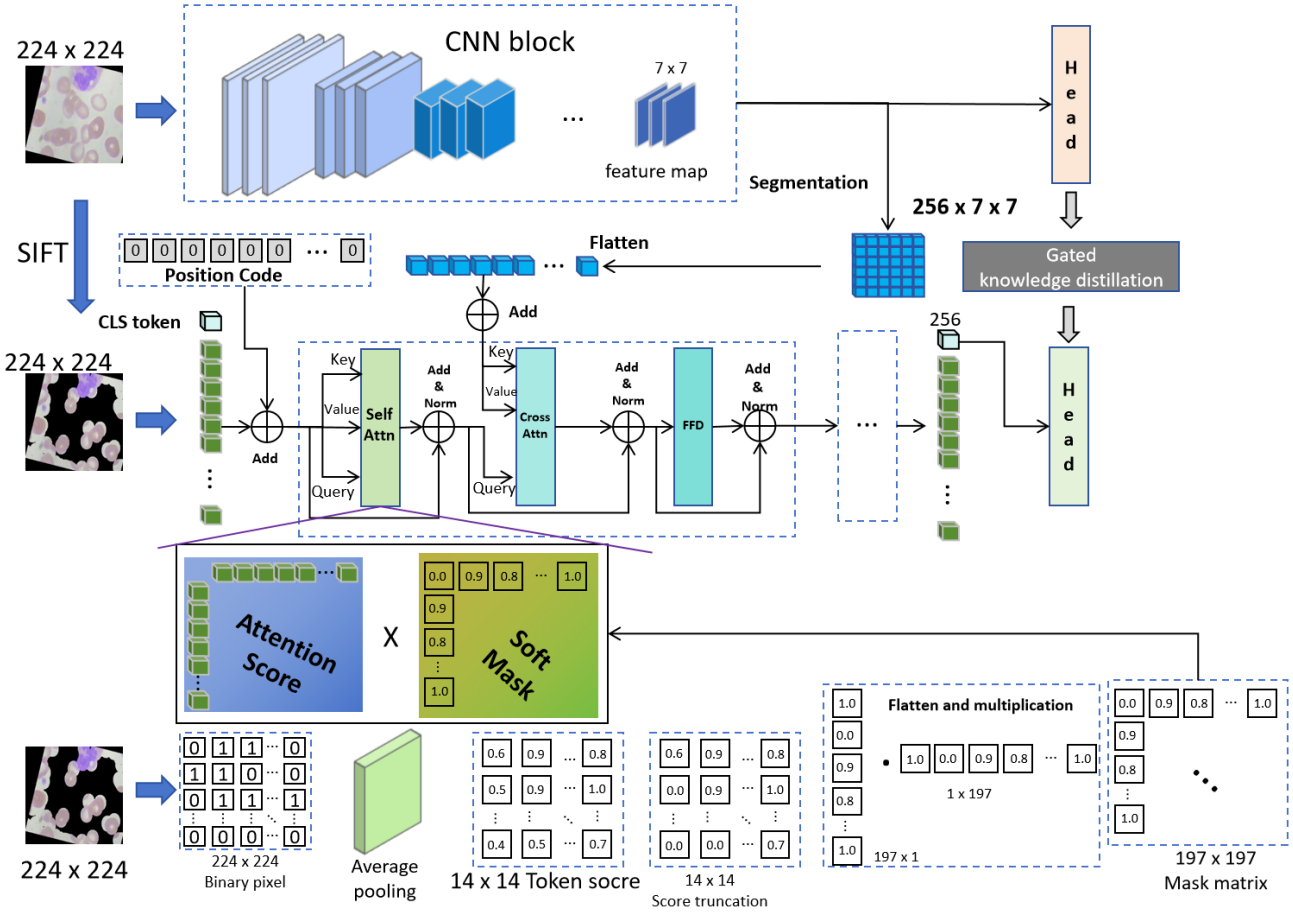


Figure 4. Architecture of Med-T. At the top is the CNN branch, which is used to extract fine-grained information at the pixel level. The middle part is the decoder. At the bottom is the Soft Mask step, which creates a soft mask map.

The input of Med-T are images $\mathbf{I} = \{I_1, I_2, \ldots, I_n\}$. This stage can be described by the following formula:

$$\mathbf{I} = \{I_1, I_2, \ldots, I_n\}, n = BatchSize \tag{1}$$

$$\mathbf{I}_{mask} = Dilation(SIFT(\mathbf{I})) \tag{2}$$

$$\mathbf{I}_{mask\_image} = Resize(\mathbf{I}_{mask}) \tag{3}$$

$$\mathbf{I}_{conv\_image\_input} = Resize(\mathbf{I}) \tag{4}$$

$$\mathbf{I}_{soft\_mask} = AveragePooling(Binarize(\mathbf{I}_{mask\_image})) \tag{5}$$

$$\mathrm{I}_{soft\_mask}(\mathrm{i}) = \begin{cases} 0 & , \mathrm{I}_{soft\_mask}(\mathrm{i}) < \mathrm{Avg}(\mathbf{I}_{soft\_mask}) \\ \mathrm{I}_{soft\_mask}(\mathrm{i}), & \mathrm{I}_{soft\_mask}(\mathrm{i}) \geq \mathrm{Avg}(\mathbf{I}_{soft\_mask}) \end{cases} \tag{6}$$

$$\mathrm{Matrix}_{\mathrm{soft\_mask}} = Flatten(\mathbf{I}_{soft\_mask})^{\mathrm{T}} \cdot Flatten(\mathbf{I}_{soft\_mask}) \tag{7}$$

We initially input the original image into the convolutional module and then segmented the output of the convolutional module to generate a token with identical dimensions as that of the transformer. Then we used these tokes to focus on the feature of the decoder. In order to enhance the local information capture capability of the CNN branch, we make the CNN branch perform the auxiliary classification task.

In the decoding stage, we used the masked image as the Query, the token output by the CNN as the Key and Value, and adopted the soft mask mechanism to mask the attention score. Traditional attention mask only uses 0 and 1 for masking, but for each patch, it may only contain part of the mask pixels, so we used soft mask, when a token contains more mask pixels, its weight will be lower, which is a good deal with the attention alignment problem at the mask boundary.

$$\mathbf{I}_{conv\_token} = PatchExtractor(\mathbf{I}_{conv\_image}), \mathbf{I}_{token} = PatchExtractor(\mathbf{I}_{mask\_image}) \tag{8}$$

$$\mathbf{I}_{token\_s} = LayerNorm(\mathbf{I}_{token} + \mathbf{Matrix}_{\text{soft\_mask}} \cdot softmax(\frac{Q(\mathbf{I}_{token}) \cdot K^T(\mathbf{I}_{token})}{\sqrt{\dim^K}}) \cdot V(\mathbf{I}_{token})) \tag{9}$$

$$\mathbf{I}_{token\_c} = LayerNorm(\mathbf{I}_{token\_s} + softmax(\frac{Q(\mathbf{I}_{token\_s}) \cdot K^T(\mathbf{I}_{conv\_token})}{\sqrt{\dim^K}}) \cdot V(\mathbf{I}_{conv\_token})) \tag{10}$$

$$\mathbf{I}_{token} = LayerNorm(\mathbf{I}_{token\_c} + FeedForward(\mathbf{I}_{token\_c})) \tag{11}$$

Finally, we can obtain a special token "CLS" that contains global representation of a sample. Then we used this "CLS" token to perform classification task.

3.2 Task definition and measure standard

Table 2. Dataset information.

Dataset	Categories	Training Set	Validation Set	Test Set	Image Size
HAM10000	7	1667	417	500	$3 \times 640 \times 450$
Blood Cell	4	4000	1000	1250	$3 \times 320 \times 240$
Retinal OCT	8	4000	1000	1250	$3 \times 1000 \times 512$

To assess the efficacy of our model on small datasets, we conducted classification task on a subset of the HAM10000, Retinal OCT-C8 and Blood Cell Image datasets. Except for HAM10000 dataset, wherein we addressed class imbalance by limiting each category to a maximum of 500 samples, we randomly selected 6250 samples from each dataset. From these, 4000 samples were designated for training, 1000 for validation, and 1250 for testing purposes. Details regarding the datasets are outlined in Table 2.

Table 3. The parameter setting of all models.

Model	Depth	Head	Embedding Dim	Learning rate
ViT_{small}	4	8	256	1e-5
ViT_{base}	8	16	512	1e-5
Swin-T	[2,2,6,2]	[3,6,12,24]	96	1e-5
Swin-L	[2,2,18,2]	[3,6,12,24]	192	1e-5
EdgeNeXt	[3,3,9,3]	8	[24,48,88,168]	1e-5
TransNeXt-T	[3,3,9,3]	[1,2,4,8]	[48,96,224,448]	1e-5
TransNeXt-S	[4,4,12,4]	[1,2,5,8]	[64,128,320,512]	1e-5
TransNeXt-B	[4,4,21,4]	[2,4,8,18]	[76,152,336,672]	1e-5
ResNet18	-	-	-	1e-4
ResNet34	-	-	-	1e-4
ResNet50	-	-	-	1e-4
ResNet101	-	-	-	1e-4
ResNeXt50	-	-	-	1e-4
ResNeXt101	-	-	-	1e-4
Med-T_{Res18}	4	8	256	1e-4
Med-T_{Res34}	4	8	256	1e-4
Med-T_{Res50}	4	8	256	1e-4
Med-T_{Res101}	4	8	256	1e-4
Med-T_{ResX50}	4	8	256	1e-4
Med-$T_{ResX101}$	4	8	256	1e-4

Performance evaluation of the models was conducted using various metrics including accuracy, top-5 error, precision, recall and f1 score to measure the performance of models.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{12}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{13}$$

$$F1 = 2\frac{Precision \cdot Recall}{Precision + Recall} \tag{14}$$

We implemented our model using the PyTorch framework, leveraging a NVIDIA GeForce RTX 4090 GPU for both training and testing. Our evaluations encompassed five distinct datasets. We resized all images to a uniform size of 224×224 and set the patch size to 16×16. Across all datasets, we maintained a training batch size of 16 and a learning rate decay rate of 0.1. Detailed parameter configurations for each model are provided in Table 3.

4. RESULT AND DISCUSSION

Prior research has demonstrated that the ViT model often exhibits suboptimal performance when applied to small datasets, indicating inherent limitations within its architecture, particularly notable in fields such as medical imaging35. In our study, we sought to address this challenge by augmenting the ViT architecture with a convolutional block tailored to extract localized, fine-grained information, serving as the encoder component of our vision transformer. Furthermore, we introduced an approach by utilizing masked images as queries within the transformer decoder and soft mask operation to enhance the performance of the model.

Table 4. The accuracy, top-5 error, precision, recall and f1 score of all models on RetinalOCT and Blood Cell Image datasets.

Model	Accuracy		Top5 Error		Precision		Recall		F1 Score	
	Retinal OCT	Blood Cell	Retinal OCT	Blood Cell	Retinal OCT	Blood Cell	Retinal OCT	Blood Cell	Retinal OCT	Blood Cell
Baseline Model										
ViT_{small}	46.80	22.80	0.72	-	46.73	16.94	46.82	22.88	46.34	15.74
ViT_{base}	46.48	26.08	0.48	-	46.53	19.12	46.90	25.76	46.51	21.59
Swin-T	45.36	30.64	0.88	-	39.69	31.31	45.20	30.66	44.76	30.78
Swin-L	44.32	30.96	0.88	-	45.90	32.34	44.25	30.70	43.91	30.90
EdgeNeXt	44.56	30.56	0.72	-	44.90	30.95	45.05	30.81	44.86	30.65
TransNeXt-T	46.08	41.20	0.40	-	45.85	39.14	45.86	40.94	45.42	39.76
TransNeXt-S	44.08	41.84	0.48	-	45.32	40.16	45.35	41.93	45.04	40.73
TransNeXt-B	38.22	25.36	4.24	-	38.20	25.68	39.49	25.40	37.25	25.13
ResNet18	68.56	75.76	0.08	-	68.25	78.62	68.31	75.82	68.13	75.96
ResNet34	66.40	68.00	8.00	-	67.50	71.80	66.82	68.14	66.68	68.29
ResNet50	49.12	64.16	2.72	-	53.52	64.90	48.95	64.11	48.43	64.37
ResNet101	47.20	49.76	1.92	-	50.64	54.49	47.52	49.73	47.86	51.01
ResNeXt50	51.60	51.36	1.04	-	52.38	53.61	52.47	51.36	52.23	51.78
ResNeXt101	50.16	46.40	1.84	-	52.45	49.66	50.12	46.42	49.71	47.40
Our Model										
Med-T_{Res18}	**69.04**	**77.12**	**0.64**	-	**68.81**	**80.51**	**68.76**	**77.14**	**68.58**	**77.67**
Med-T_{Res34}	57.44	62.24	0.16	-	58.17	68.32	58.49	62.31	55.81	62.77
Med-T_{Res50}	54.48	53.04	3.04	-	55.96	56.65	54.14	53.09	54.37	53.89
Med-T_{Res101}	45.60	34.16	0.72	-	47.14	32.97	45.29	34.35	43.91	31.12
Med-T_{ResX50}	48.96	57.12	1.76	-	55.34	57.11	49.52	57.14	48.17	56.97
Med-$T_{ResX101}$	49.44	51.28	1.84	-	49.98	55.60	49.84	51.19	49.45	51.93

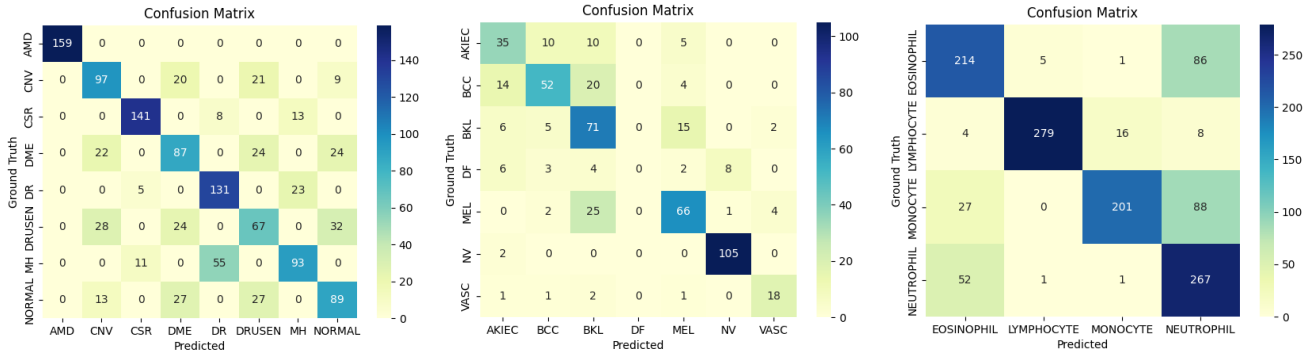
Firstly, we conducted evaluations on subset of Retinal OCT, HAM10000 and Blood Cell Image data sets. Performance metrics including accuracy, top-5 error, precision, recall and f1 score of all model are summarized in Table 4 and Table 5.

Notably, due to the limited number of categories in the Blood Cell Image dataset (only four categories), top-5 error is not considered in the test on this dataset. Our model Med-T_{Res18} exhibited commendable accuracy, achieving a peak performance of 69.04% on the Retinal OCT dataset. Furthermore, compact model variant demonstrated noteworthy accuracies of 70.00% and 77.12% on the HAM10000 and Blood Cell Image datasets respectively. As highlighted by Dosovitskiy, models structured upon the ViT framework often exhibit inferior performance compared to classical CNN models, particularly when trained on datasets with a restricted number of samples. Our utilization of the CNN model as a convolutional encoder, showcasing superior performance over other ViT-based models. This is attributed to the Med-T's adeptness at capturing intricate local features, a capability which proves crucial when discerning subtle distinctions within medical image datasets.

Table 5. The accuracy, top-5 error, precision, recall and f1 score of all models on HAM10000 datasets.

Model	Accuracy	Top5 Error	Precision	Recall	F1 Score
Baseline Model					
ViT_{small}	48.60	5.80	38.81	36.05	33.30
ViT_{base}	47.40	5.40	32.20	35.87	32.81
Swin-T	54.40	4.69	39.69	41.27	38.70
Swin-L	46.40	4.00	30.83	35.49	31.07
EdgeNeXt	52.00	4.40	38.69	39.06	34.93
TransNeXt-T	49.40	7.00	37.76	38.56	35.03
TransNeXt-S	56.80	4.40	47.4	42.96	41.18
TransNeXt-B	45.00	6.60	28.34	33.43	27.58
ResNet18	69.20	1.60	65.52	63.69	63.60
ResNet34	63.00	1.40	59.82	58.95	58.91
ResNet50	59.80	2.20	57.37	54.84	55.65
ResNet101	59.20	1.60	53.24	53.77	53.20
ResNeXt50	59.20	1.00	55.83	54.94	55.19
ResNeXt101	62.60	2.60	59.98	55.43	56.53
Our Model					
Med-T_{Res18}	69.40	**0.80**	59.68	61.52	60.35
Med-T_{Res34}	**70.00**	**0.80**	**61.03**	**63.17**	**61.85**
Med-T_{Res50}	62.60	1.80	52.74	57.52	54.95
Med-T_{Res101}	58.40	3.80	47.70	46.66	44.41
Med-T_{ResX50}	63.20	2.20	55.11	56.80	55.73
Med-$T_{ResX101}$	61.00	1.40	53.13	52.69	52.65

Furthermore, our analysis revealed a marginal decline in model performance as models expanded. This phenomenon can be attributed to the inherent challenges posed by limited data availability, exacerbated by the large parameter space of the model architecture, which may impede effective feature learning from small dataset.



Confusion matrix on Retinal

Confusion matrix on HAM10000 dataset.

Confusion matrix on Blood Cell Image dataset.

Figure 5. Confusion Matrix of Med-T model on test set of three datasets.

The challenge posed by the Blood Cell Image dataset, featuring a mere four categories, lies in the inherent similarity of foreground object features (Each type of cell only differs in local structure). Consequently, distinguishing between these samples poses a considerable challenge of fine-grained information capturing for baseline models. Notably, models structure upon the simplest ViT architecture demonstrated the poorest performance on the Blood Cell Image dataset. Conversely, our model, which amalgamates the strengths of both convolutional and ViT structures, emerges as a standout performer on this dataset. Leveraging the pixel-level feature extraction prowess of convolutional kernels alongside the holistic understanding facilitated by ViT structures, our model attains an accuracy of 77.12% and a precision of 80.51% on the Blood Cell Image dataset.

Figure 5 presents the confusion matrix generated by the Med-T model on the test set of three datasests. Notably, the HAM10000 dataset exhibits imbalanced class distributions, with samples from the “NV” category comprising 66.9% of the total dataset. The model did not predict the class "DF" correctly due to the small number of samples "DF" category. Nevertheless, the prediction of our model is not biased towards the "NV" category. Conversely, when the class distributions are balanced, the confusion matrix displays a more concentrated distribution along the diagonal.

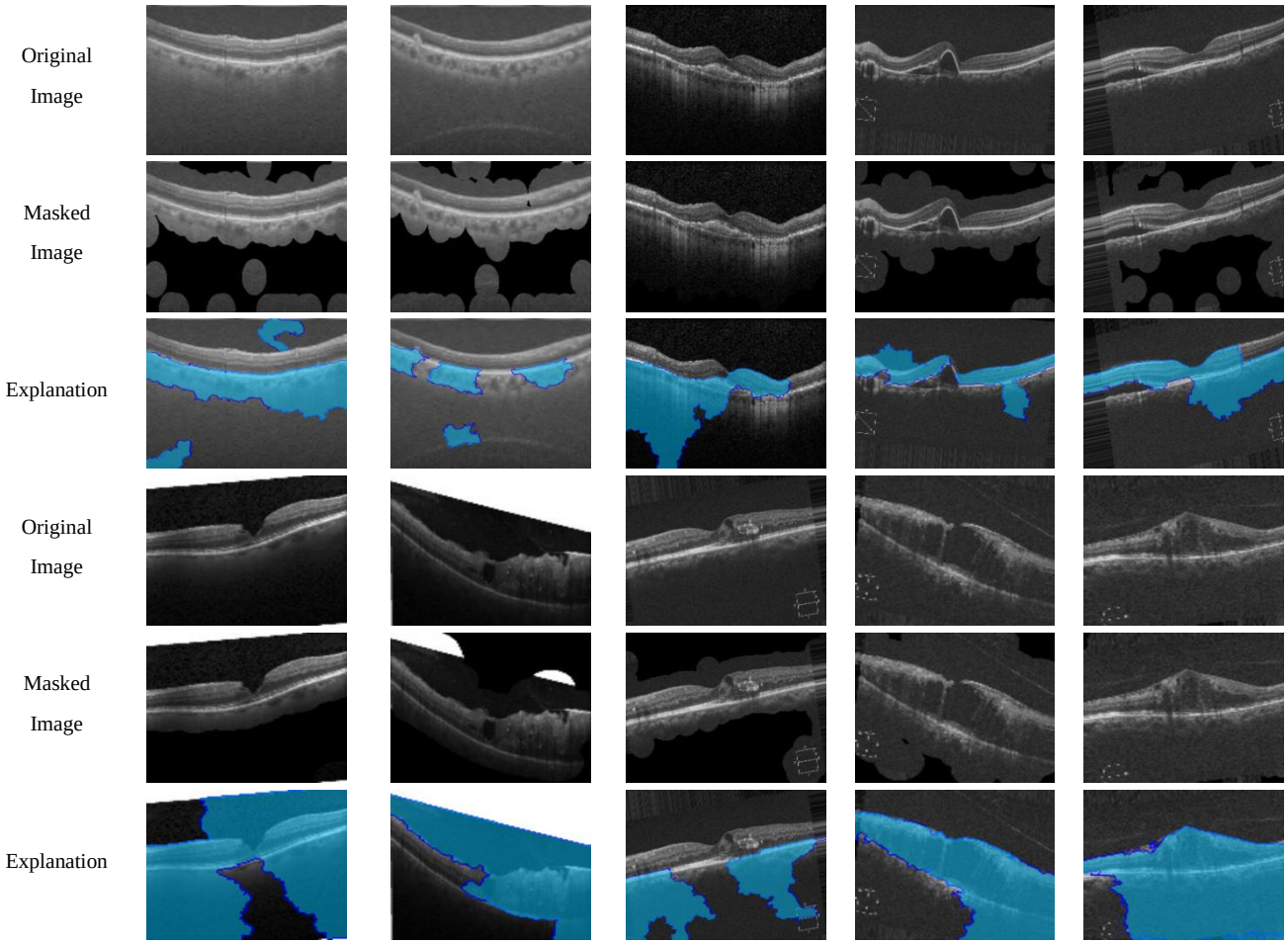


Figure 6. Explanation of model predictions on Retinal OCT C8 dataset using tool LIME. The blue area is the feature that contributes the most to the category predicted by the model.

Finally, we employed the model interpreter LIME to elucidate the results predicted by the Med-T_{Res18} model, as depicted in Figure 6 and Figure 7. LIME leverages a trained local surrogate model to explain individual samples, highlighting features contributing to the predicted outcomes. The results indicate that despite the model also taking some irrelevant backgrounds into consideration, it can almost completely cover the targets to be detected. And the blue area is mainly concentrated near the detected target, showing a continuous distribution. The mask operation conveniently covers areas unassociated with the detection task, facilitating precise predictions by enabling the model to focus on the relevant regions during inference.

5. CONCLUSION

Previous research has elucidated the challenges facing ViT models when applied to small datasets, a point reaffirmed by the findings of this paper. However, within the medical domain, high-quality samples are scarce commodities, each bearing immense value. Through rigorous experimentation, we have demonstrated that augmenting transformer architectures with convolutional block to extract pixel-level position information can notably enhance model performance. This integration effectively addresses the limitations of vision transformer models when confronted with limited sample sizes, offering a promising avenue for advancement in medical large model. In this study, we combined convolutional block with transformers and leveraged soft mask operation to bolster the performance of vision transformer models on small datasets. Our findings signify the potential of vision transformer models in medical field. However, it is pertinent to acknowledge the extra computational demands inherent to our approach. Moving forward, we envision a convergence of diverse encoder and feature extraction methodologies with vision transformer models, aiming to further enhance their applicability across medical field.

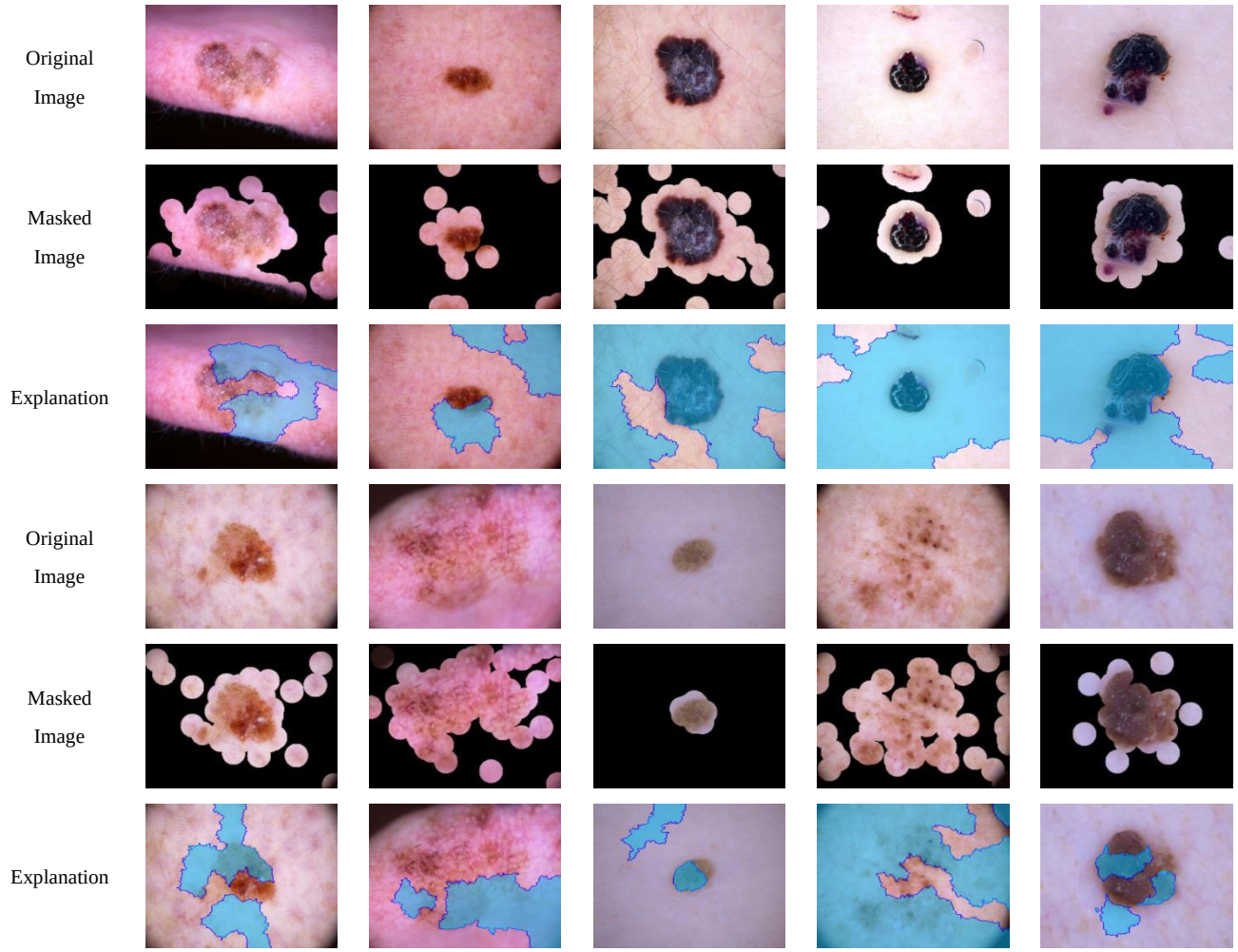


Figure 7. Explanation of model predictions on HAM10000 dataset using tool LIME. The blue area is the feature that contributes the most to the category predicted by the model.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Natural Science Foundation of China (No.82274580). This work is also supported by the Research Project of Liaoning Education Department (serial number: LJKZ0894,2021-549) and the Natural Science Research Project of Liaoning University of Traditional Chinese Medicine (2021-18) .

Data Availability The datasets used for the development and evaluation of methods in this paper are publicly available:

1. Blood Cell Image: https://www.kaggle.com/datasets/paultimothymooney/blood-cells
2. Retinal OCT C8: https://www.kaggle.com/datasets/obulisainaren/retinal-oct-c8
3. HAM10000: https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000/data

REFERENCES

[1] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," 2016 IEEE Con- ference on Computer Vision and Pattern Recognition (CVPR), 770–778 (2016).
[2] Xie, S., Girshick, R., Doll´ar, P., Tu, Z., and He, K., "Aggregated residual transformations for deep neu- ral networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5987–5995 (2017).
[3] Olaf, R., Philipp, F., and Thomas, B., "U-net: Convolutional networks for biomedical image segmentation,"(2015). Preprint at https://api.semanticscholar.org/CorpusID:3719281.
[4] Alexey, D., Lucas, B., Alexander, K., Dirk, W., Xiaohua, Z., Thomas, U., Mostafa, D., Matthias, M.,Georg, H., Sylvain, G., Jakob, U., and Neil, H., "An image is worth 16x16 words: Transformers for image recognition at scale," (2020). Preprint at https://api.semanticscholar.org/CorpusID:225039882.
[5] Olga, R., Jia, D., and Hao, S., "Imagenet large scale visual recognition challenge," International Journal of Computer Vision 115, 211 – 252 (2014).
[6] Tschandl, P., Rosendahl, C., and Kittler, H., "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data 5 (2018).
[7] Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., and Baxter, S. L., "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell 172 (2018).
[8] Alam, M. M. and Islam, M. T., "Machine learning approach of automatic identification and counting of blood cells," Healthcare Technology Letters 6(4), 103–108 (2019).
[9] Loey, M., Manogaran, G., and Khalifa, N. E. M., "A deep transfer learning model with classical data augmentation and cgan to detect covid-19 from chest ct radiography digital images," Neural Computing and Applications.
[10] Krizhevsky, A. and Hinton, G., "Learning multiple layers of features from tiny images," Handbook of Sys-temic Autoimmune Diseases 1 (2009).
[11] Han, X., Kashif, R., and Roland, V., "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," (2017). https://arxiv.org/abs/1708.07747.
[12] Ribeiro, M. T., Singh, S., and Guestrin, C., "why should i trust you?": Explaining the predictions of any classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144 (2016).
[13] Liao, W., Zeng, B., Liu, J., Wei, P., and Fang, J., "Image-text interaction graph neural network for image- text sentiment analysis," Applied Intelligence 52 (2022).
[14] Ruan, S., Zhang, K., Wu, L., Xu, T., Liu, Q., and Chen, E., "Color enhanced cross correlation net for image sentiment analysis," IEEE Transactions on Multimedia, 1–1 (2021).
[15] Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H., and Qian, J., "Multimodal sentiment analysis with image-text interaction network," IEEE Transactions on Multimedia 25, 3375–3385 (2023).
[16] Yang, X., Feng, S., Wang, D., and Zhang, Y., "Image-text multimodal emotion classification via multi-view attentional network," IEEE Transactions on Multimedia 23, 4014–4026 (2021).
[17] Liang, Y., Maeda, K., Ogawa, T., and Haseyama, M., "Deep metric network via heterogeneous semantics for image sentiment analysis," (2021). Paper presented at IEEE International Conference on Image Processing (ICIP).
[18] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," Computer Science (2014).

[19] Xin-Ying, X., Yuanyuan, P., Zhengpeng, Z., Jinjing, G., and Dan, X., "Bit: Improving image-text sentiment analysis via learning bidirectional image-text interaction," (2023). Paper presented at International Joint Conference on Neural Networks (IJCNN).
[20] Renyu, Z., Chengcheng, H., Qian, Y., Sun, Q., Li, X. L., Gao, M., Cao, X., and Xian, Y., "Exchanging-based multimodal fusion with transformer," (2023). Preprint at https://api.semanticscholar.org/CorpusID: 261557178.
[21] Danae S'anchez, V., Daniel, P.-P., and Nikolaos, A., "Improving multimodal classification of social media posts by leveraging image-text auxiliary tasks," (2023). Preprint at https://api.semanticscholar.org/ CorpusID:261824432.
[22] Junyu, C., Jie, A., Hanjia, L., and Jiebo, L., "Improving visual-textual sentiment analysis by fusing expert features," (2022). Preprint at https://api.semanticscholar.org/CorpusID:253801570.
[23] Shyamgopal, K., Karsten, R., Massimiliano, M., and Zeynep, A., "Vision-by-language for training-free compositional image retrieval," (2023). Preprint at https://api.semanticscholar.org/CorpusID:264128238.
[24] Haoxing, C., Yaohui, L., Yan, H., Zizheng, H., Zhuoer, X., Zhangxuan, G., Jun, L., Huijia, Z., and Weiqiang, W., "Boosting audio-visual zero-shot learning with large language models," (2023). Preprint at https://api.semanticscholar.org/CorpusID:265309197.
[25] Xiaoyu, Y., Lijian, X., Hongsheng, L., and Shaoting, Z., "Vilam: A vision-language model with enhanced visual grounding and generalization capability," (2023). Preprint at https://api.semanticscholar.org/
[26] Feng, L., Qing, J., Hao, Z., Tianhe, R., Shilong, L., Xueyan, Z., Hu-Sheng, X., Hongyang, L., Chun-yue, L., Jianwei, Y., Lei, Z., and Jianfeng, G., "Visual in-context prompting," (2023). Preprint at https://api.semanticscholar.org/CorpusID:265351501.
[27] Alec, R., Jong Wook, K., Chris, H., Aditya, R., Gabriel, G., Sandhini, A., Girish, S., Amanda, A., Pamela, M., Jack, C., Gretchen, K., and Ilya, S., "Learning transferable visual models from natural language super-vision," (2021). Paper presented at International Conference on Machine Learning (ICML).
[28] Junnan, L., Dongxu, L., Caiming, X., and Steven, C. H. H., "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," (2022). Paper presented at International Conference on Machine Learning (ICML).
[29] Wonjae, K., Bokyung, S., and Ildoo, K., "Vilt: Vision-and-language transformer without convolution or region supervision," (2021). Paper presented at International Conference on Machine Learning (ICML).
[30] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B., "Swin transformer: Hierarchical vision transformer using shifted windows," (2021). Paper presented at IEEE/CVF International Conference on Computer Vision (ICCV).
[31] Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F. E. H., Feng, J., and Yan, S., "Tokens-to-token vit: Training vision transformers from scratch on imagenet," (2021). Paper presented at IEEE/CVF International Conference on Computer Vision (ICCV).
[32] Peng, Z., Guo, Z., Huang, W., Wang, Y., Xie, L., Jiao, J., Tian, Q., and Ye, Q., "Conformer: Local features coupling global representations for recognition and detection," IEEE Transactions on Pattern Analysis and Machine Intelligence 45, 9454–9468 (2023).
[33] Muhammad, M., Abdelrahman, M. S., Hisham, C., Salman, K., Syed Waqas, Z., Rao Muhammad, A., and Fahad Shahbaz, K., "Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications," (2022). Preprint at https://api.semanticscholar.org/CorpusID:249890419.
[34] Dai, S., "Transnext: Robust foveal visual perception for vision transformers," (2023). Preprint at https://api.semanticscholar.org/CorpusID:265498825.
[35] Dong-Ming, Z., Yi-Mou, L., Hao, C., Zhuotao, T., Xin, Y., Jinhui, T., and Kwang-Ting, C., "Understanding the tricks of deep learning in medical image segmentation: Challenges and future directions," (2022). Preprint at https://api.semanticscholar.org/CorpusID:258556728.