

DuoFormer: Leveraging Hierarchical Visual Representations by Local and Global Attention

Xiaoya Tang¹[0009-0002-5638-320X], Bodong Zhang^{1,2}[0000-0001-9815-0303],
Beatrice S. Knudsen³[0000-0002-7589-7591], and
Tolga Tasdizen^{1,2}[0000-0001-6574-0366]

¹ Scientific Computing and Imaging Institute, University of Utah, SLC, UT, USA
xiaoya@sci.utah.edu

² Electrical and Computer Engineering, University of Utah, SLC, UT, USA
bodong.zhang@utah.edu, tolga@sci.utah.edu

³ Department of Pathology, University of Utah, Salt Lake City, UT, USA
beatrice.knudsen@path.utah.edu

Abstract. We here propose a novel hierarchical transformer model that adeptly integrates the feature extraction capabilities of Convolutional Neural Networks (CNNs) with the advanced representational potential of Vision Transformers (ViTs). Addressing the lack of inductive biases and dependence on extensive training datasets in ViTs, our model employs a CNN backbone to generate hierarchical visual representations. These representations are then adapted for transformer input through an innovative patch tokenization. We also introduce a 'scale attention' mechanism that captures cross-scale dependencies, complementing patch attention to enhance spatial understanding and preserve global perception. Our approach significantly outperforms baseline models on small and medium-sized medical datasets, demonstrating its efficiency and generalizability. The components are designed as plug-and-play for different CNN architectures and can be adapted for multiple applications. The code is available at <https://github.com/xiaoyatang/DuoFormer.git>.

Keywords: Vision Transformer · Inductive Bias · Multi-scale features.

1 Introduction

The Vision Transformer (ViT) [4] has significantly advanced the adaptation of transformers from language to vision, demonstrating superior performance over CNNs when pre-trained on large datasets. ViT employs a patch tokenization process that converts images into a sequence of uniform token embeddings. These tokens undergo Multi-Head Self-Attention (MSA), transforming them into queries, keys, and values that capture extensive non-local relationships. Despite their potential, ViTs can underperform similarly-sized ResNets [10] when inadequately trained due to their lack of inductive biases such as translation equivariance and locality [19,14], which are naturally encoded by CNNs. Recent efforts have focused on mitigating ViTs' limitations by integrating convolutions or

adding self-supervised tasks. Prevalent approaches combine CNN feature extractors with transformer encoders [1,4,24,18,28,6,15,5], such as the 'hybrid' ViT [4]. Other methods like knowledge distillation [20] transfer biases from CNNs to ViT. Nonetheless, ViTs' smaller receptive fields compared to CNNs limit their ability to capture detailed spatial relationships [1], which can be partially alleviated by techniques like enriched spatial shifting patches [14].

Histopathology image analysis, critical in medical diagnostics, involves examining whole slide images (WSIs) to detect and interpret complex tissue structures and cellular details. This analysis faces challenges due to the varied scales of visual entities within WSIs, such as the differing sizes of cell nuclei and vascular structures, both of which can contribute to a model's task of distinguishing low- and high-risk kidney cancers. Moreover, vital global features of cancer and its microenvironment, observable only at lower scales, are crucial for various downstream tasks. The neglect of these multiple scales can significantly impair the performance of deep learning models in medical image recognition tasks. CNNs tackle this issue by utilizing a hierarchical structure created by lower and higher stages, which allows them to detect visual patterns from simple low-level edges to complex semantic features. Conversely, ViTs employ fixed-scale patches, thereby overlooking crucial multi-scale information within the same image [19], which can hinder their performance across diverse tasks. By harnessing a hierarchical structure similar to that of CNNs, ViTs can be prevented from overlooking the critical multi-scale features, while also imparting necessary inductive biases. Existing works on directly integrating multi-scale information into ViTs vary primarily in the placement of convolutional operations: during patch tokenization[28,26,8], within[12,8,16] or between self-attention layers, including query/key/value projections[24,28], forward layers [15], or positional encoding [26], etc. Despite the benefits of hierarchical configurations [11], a definitive model for visual tasks has yet to emerge. The challenge remains in effectively producing and utilizing features across various scales. In response, we propose a novel hierarchical Vision Transformer model, outlined as follows:

1. Our proposed multi-scale tokenization involves a single-layer projection, patch indexing, and concatenation, assembling features from different stages of the CNN into multi-scale tokens.
2. We introduce a novel MSA mechanism called scale attention, combined with patch attention. This approach enables the model to recognize connections across scales, expanding ViT's receptive field and bridging the gap between CNN and Transformer architectures.
3. Our proposed scale token, part of the scale attention, is initialized with a fused embedding derived from hierarchical representations. It enriches the transformer's multi-granularity representation and aggregates scale information, serving as the input for the global patch attention.

2 Related Work

Various approaches have explored integrating the hierarchical architecture of CNNs into Vision Transformers (ViTs) across different visual tasks, including video recognition[6], image classification [2,3,5,7,11,12,15,16,18,20,22,23,24,26,27] [28,29,31], object detection[3,7,11,12,16,22,25,26,27,28,31], and segmentation[3,7] [12,16,21,22,26,27,31]. Notable methods emulate the pyramid structure of CNN with stage-wise pooling and convolutional embeddings [11] or integrate pooling within the attention mechanism[6].

Multiple scales have been exploited beyond mere convolution integration. The Swin Transformer [18] utilizes a shifting window strategy, while Dong et al. [3] split multi-heads to perform self-attention in horizontal and vertical stripes. Chen et al. [2] developed a dual-branch architecture that processes varying patch sizes, and Zhang et al. [31] implemented a multi-granularity strategy. PVT [22] reduces feature size progressively using spatial-reduction attention. A recent study [7] employed a spatial decay matrix to enhance self-attention with spatial priors. UNETR [9] constructs a U-shape transformer encoder and decoder for 3D segmentation, while people also replaced UNet skip connections with attention mechanisms[21] for 2D segmentation. Besides, inductive biases can also be integrated through auxiliary tasks such as unsupervised localization to enhance local processing capabilities [17].

3 Methodology

Our model utilizes a CNN as the embedding layer, depicted in Figure 1. Patch tokenization contains two steps represented by the dashed lines in Figure 1: First is extracting hierarchical representations from different stages of the CNN backbone. Second is the projection and patch indexing. After acquiring the multi-scale features, we use our DuoFormer to learn the local dependencies across scales and global dependencies across patches, which are needed for downstream tasks.

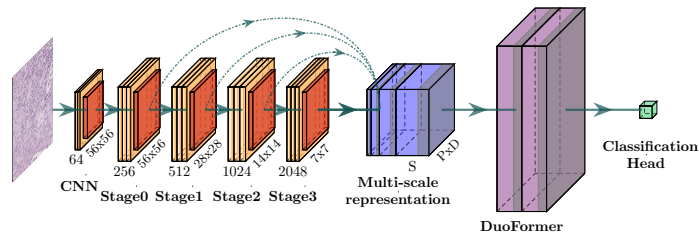


Fig. 1. The pipeline of the proposed DuoFormer. Dimensionalities of the multi-scale representation: S: scale dimension; P: number of patches; D: embedding dimension.

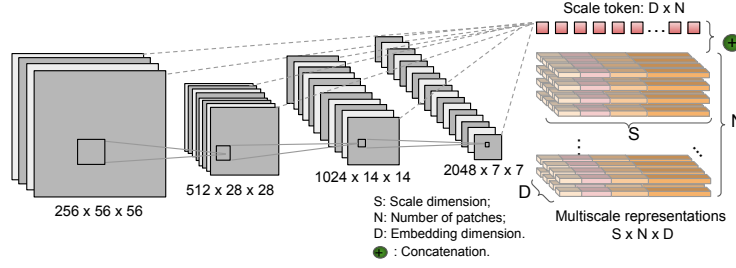


Fig. 2. Visualization of Multiscale Patch Tokenization: This figure depicts the process of converting an image into a sequence of multi-scale patch embeddings, with each color representing a different scale to illustrate the varied dimensions of the patches.

3.1 Multi-scale Patch Tokenization

Given the input to the backbone, $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ with $H = W$, we derive hierarchical outputs from four stages, denoted as $\mathbf{x}_i \in \mathbb{R}^{P_i \times P_i \times C_i}$ for $i \in 0, 1, 2, 3$. Here, $P_i = \frac{H}{4 \cdot 2^i}$ specifies the spatial resolution, and C_i represents the channel dimension. We apply a linear projection to transform these features into embeddings of dimension D . Next, we split the embeddings \mathbf{x}' from each stage into N non-overlapping patches, set as $N = 49$. Each scale yields a sequence of flattened tokens with spatial size $P_i'^2$, where $P_i' = \frac{H}{4 \cdot 2^i \cdot \sqrt{N}}$. We index and concatenate multi-scale embeddings from each patch across all scales to form the multi-scale tokens \mathbf{X}_Σ^t , illustrated in Figure 2. The equation for this process is:

$$\begin{aligned}
 \mathbf{x}'_i &= \text{Projection}(\mathbf{x}_i), \mathbf{x}'_i \in \mathbb{R}^{P_i \times P_i \times D} \\
 \mathbf{x}''_i &\in \mathbb{R}^{P_i'^2 \times N \times D}, P_i'^2 = \frac{HW}{16 \cdot 4^i \cdot N}, i \in 0, 1, 2, 3, \\
 \mathbf{X}_\Sigma^t &= \text{concat}(\mathbf{x}''_i) \in \mathbb{R}^{S \times N \times D}, S = \sum P_i'^2.
 \end{aligned} \tag{1}$$

3.2 Duo Attention Module

Our tokenization directly embeds multiscale spatial information into the scale dimension, inherently enriching the model’s inductive biases. Subsequently, our encoder employs scale and patch attentions to respectively focus on detailed image features and broader contexts, as illustrated in Figure 3(a). Our scale attention adapts the Multi-Head Self-Attention (MSA) framework by incorporating an additional scale dimension. This adaption integrates multi-scale analysis directly into the attention mechanism and alters tensor operations to accommodate multi-dimensional tokens. Details are explained in the equation below and depicted in Figure 3(b).

$$\mathbf{X}'_\Sigma = \mathbf{X}_\Sigma + \text{MSA}(\text{LN}(\mathbf{X}_\Sigma)), \quad \mathbf{Y} = \mathbf{X}'_\Sigma + \text{FFN}(\text{LN}(\mathbf{X}'_\Sigma)) \tag{2}$$

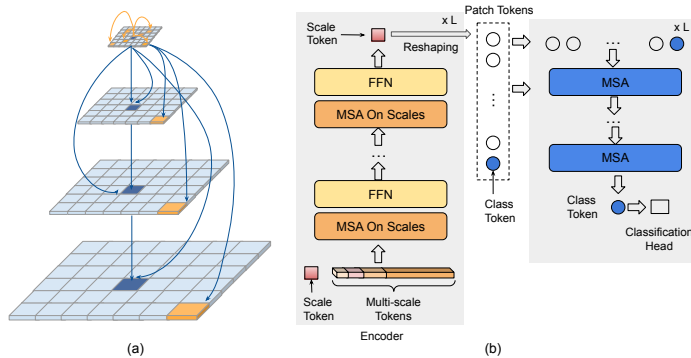


Fig. 3. Illustration of the Duo attentions. Panel (a) shows the local (yellow arrows) and global (blue arrows) dependencies among multi-scale patches, maintaining a consistent grid size of 49; larger patches indicate greater embedding lengths. Panel (b) details the model architecture, including L layers of scale and patch attention blocks in the encoder.

After scale attention, the scale token aggregates key details from all scales for input to patch attention. Patch attention, mirroring standard MSA, omits layer normalization(LN), feed-forward networks (FFN), and residual connections, as shown in Figure 3(b).

3.3 Scale Token

To enhance the hierarchical representations, we use a downsampling strategy involving simple convolutional layers followed by max pooling. This process normalizes the spatial dimensions of all embeddings to N , maintaining consistent channel dimensions. N denotes number of patches, set as 49 in our experiments. These embeddings are then concatenated along the channel dimension and projected into the embedding dimension D using lightweight convolutions. The resulting scale token, concatenated with multi-scale tokens, serves as the input for the scale attention, guiding it effectively, as detailed below.

$$\begin{aligned}
 \tilde{\mathbf{x}}_0 &= \text{MaxPool}(\text{Conv}(\mathbf{x}_0)), & \tilde{\mathbf{x}}_1 &= \text{MaxPool}(\text{Conv}(\mathbf{x}_1)) \\
 \tilde{\mathbf{x}}_2 &= \text{MaxPool}(\mathbf{x}_2), & \tilde{\mathbf{x}}_3 &= \mathbf{x}_3, \quad \text{where } \mathbf{x}_i \in \mathbb{R}^{N \times C_i}, \\
 \tilde{\mathbf{X}}_\Sigma &= \text{concat}(\tilde{\mathbf{x}}_0, \tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \tilde{\mathbf{x}}_3) \in \mathbb{R}^{N \times C}, & C &= \sum C_i, \\
 \text{Scale Token} &= \text{ReLU}(\text{BN}(\text{Conv}(\tilde{\mathbf{X}}_\Sigma))) \in \mathbb{R}^{N \times D}
 \end{aligned} \tag{3}$$

4 Experiments

4.1 Experimental Setup

Our evaluation utilized two datasets, Utah ccRCC and TCGA ccRCC [30], with varying ResNet backbones for a thorough analysis. The Utah ccRCC dataset

comprises 49 WSIs from 49 patients, split into training (32 WSIs), validation (10 WSIs), and testing (7 WSIs). Tiles were extracted from marked polygons at 400x400 pixel resolution at 10X magnification with a 200-pixel stride and center-cropped to 224x224 pixels for model compatibility. The training set included 28,497 Normal/Benign, 2,044 Low Risk, 2,522 High Risk, and 4,115 Necrosis tiles, with validation and test sets proportionately distributed. The TCGA ccRCC dataset features 150 labeled WSIs divided into 30 for training, 60 for validation, and 60 for testing, using similar cropping methods but adjusted strides to gather more training patches. It contains 84,578 Normal/Benign, 180,471 Cancer, and 7,932 Necrosis tiles in the training set, with similar distributions in validation and test sets. For model details, please refer to Appendix.

4.2 Result Analysis

In this study, we utilized ResNet18 and ResNet50 backbones [10] pre-trained on extensive datasets, assessing our model under two paradigms: with ImageNet supervised pre-trained and pathology(TCGA and TULIP) self-supervised pre-trained [13] backbones. Results, shown in Table 1, indicate our model surpasses ResNet baselines by over 2% across all settings and outperforms various Hybrid-ViTs in all scenarios. The results underscore our model’s capacity to harness multi-scale features and integrate crucial inductive biases without necessitating additional tasks or additional pre-training of the transformer encoder.

In the supervised pre-training scenario, particularly with TCGA using a ResNet 50 backbone, deeper encoders sometimes hindered performance, highlighting the need for careful design when integrating CNN architectures, especially considering domain shifts. Our DuoFormer improved performance by 3.83%, demonstrating its effectiveness in leveraging multi-scale representations and likely guiding the feature extractor to adapt better to domain shifts when trained together. In the self-supervised pre-trained experiments, our model significantly outperformed the baseline by 9.88% and clearly surpassed the Hybrid-ViTs, showing the superiority of our model in leveraging multi-scale features. These findings suggest that with the proposed designs, the model can effectively capture essential local features while preserving global attention capabilities, thereby addressing the typical inductive bias limitations found in transformers.

4.3 Ablation Studies

Ablation on Scale Attention For ablation studies, we utilized our best models in both settings, employing ResNet18 pretrained on ImageNet for UTAH and ResNet50 pretrained on histopathology images for TCGA. We evaluated the individual contributions of scale and patch attention mechanisms using configurations of 6 layers for UTAH and 8 layers for TCGA. The different depths were chosen to adapt to the larger size of the TCGA dataset and the smaller size of the UTAH dataset. Results in Table 2 indicate that scale attention alone outperforms setups using only patch attention, suggesting the robustness of our

Table 1. Feature extractors were unfrozen unless specified otherwise. Results here are reported as mean values from several independent experiments.

ImageNet Supervised Pretrained		
Model	Params.	Accuracy (%)
TCGA		
ResNet50	23.50M	72.74
ResNet50-ViT Base	112.5M	75.89
ResNet50-ViT Large	197.6M	73.34
ResNet50-DuoFormer (Ours)	186.0M	76.57
UTAH		
ResNet18	11.20M	88.87
ResNet18-ViT Base	99.03M	82.35
ResNet18-ViT Large	184.1M	86.39
ResNet18-DuoFormer (Ours)	89M	91.22
Pathology Self Supervised Pretrained		
TCGA		
Model	Params.	Accuracy (%)
ResNet50-SwaV (Freeze)	0.008M	77.98
ResNet50-SwaV (Freeze)-ViT Base	89.03M	74.00
ResNet50-SwaV (Freeze)-ViT Large	174.1M	85.81
ResNet50-SwaV (Freeze)-DuoFormer (Ours)	124.7M	86.45

Table 2. Ablation study comparing scale and patch attention individually and in combination. Configurations with only scale attention use a single fully-connected (FC) layer to adapt the scale token for the classification head, trained alongside the entire network.

Dataset	Scale Attn	Patch Attn	Scale Attn & Patch Attn (Ours)
UTAH	90.31	82.35	91.22
TCGA	79.90	74.00	86.45

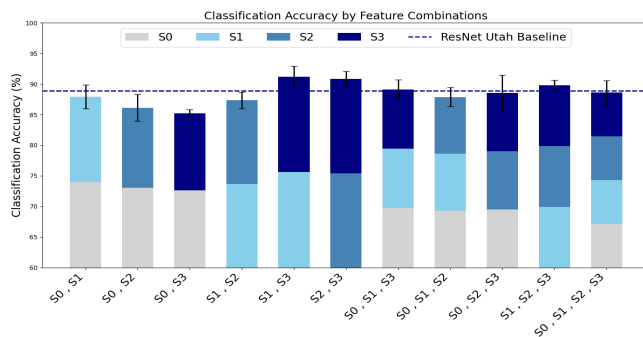
scale token and attention mechanism in harnessing multi-scale features. The results also confirm that the best performance on both datasets is achieved when both attention modules are employed together, emphasizing the necessity of integrating both local and global information for effective visual processing.

Ablation on Scale Token To evaluate the role of the scale token, we conducted experiments comparing configurations with and without a scale token, and against a learnable scale token, as shown in Table 3. Our scale token effectively enhanced local information capture, outperforming the learnable version. Without a scale token, using the first token from scale attention yielded better results than averaging all tokens, likely due to the first token’s representation of the final CNN stage’s output, which provides crucial, concise information. This suggests that averaging introduces noise.

Table 3. Ablation study on the impact of different scale token configurations, including a learnable scale token implemented as `nn.Parameters()`.

Dataset	The first token	Avg. of tokens	Learnable	Ours
	w/o Scale-Token	w/o Scale-Token	w/i Scale-Token	w/i Scale-Token
UTAH	90.61	89.62	88.80	91.22
TCGA	83.22	82.62	83.13	86.45

Ablation on Multi-Scale Representations We explored the impact of different combinations of stages on the UTAH dataset. S_0 represents the shallowest stage (56×56), and S_3 is the deepest (7×7). According to the results, incorporating all stages slightly harmed performance, likely due to overfitting given the small UTAH dataset. Including S_3 generally improved performance, highlighting the final stage’s importance for classification accuracy. Including S_0 often decreased performance, possibly due to its larger spatial embeddings and higher overfitting risk. Conversely, on larger datasets, as shown in Table 1, including all stages proved beneficial. The highest three configurations here used S_1 , S_2 , and S_3 , demonstrating the benefits of multi-scale integration while managing computational complexity.

**Fig. 4.** Ablation study on combinations of hierarchical stages. Stages are represented by colors from light to dark. Bar heights and black error bars show mean accuracies and standard deviations, and the blue dashed line marks the ResNet baseline.

5 Conclusion

In this study, we introduced a novel hierarchical transformer with dual attention mechanisms that enhance visual data interpretation across scales, improving medical image classification. Ablation studies confirm performance optimization, demonstrating the model’s robustness and adaptability across various CNN backbones and tasks, paving the way for broader applications in medical imaging and vision-related fields.

References

1. Araujo, A., Norris, W., Sim, J.: Computing receptive fields of convolutional neural networks. *Distill* **4**(11), e21 (2019)
2. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 357–366 (2021)
3. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 12124–12134 (2022)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
5. d’Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G., Sagun, L.: Convit: Improving vision transformers with soft convolutional inductive biases. In: *International conference on machine learning*. pp. 2286–2296. PMLR (2021)
6. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 6824–6835 (2021)
7. Fan, Q., Huang, H., Chen, M., Liu, H., He, R.: Rmt: Retentive networks meet vision transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5641–5651 (2024)
8. Guo, J., Han, K., Wu, H., Xu, C., Tang, Y., Xu, C., Wang, Y.: Cmt: Convolutional neural networks meet vision transformers. *2022 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 12165–12175 (2022)
9. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. pp. 574–584 (2022)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
11. Heo, B., Yun, S., Han, D., Chun, S., Choe, J., Oh, S.J.: Rethinking spatial dimensions of vision transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 11936–11945 (2021)
12. Hou, Q., Lu, C.Z., Cheng, M.M., Feng, J.: Conv2former: A simple transformer-style convnet for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024)
13. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3344–3354 (2023)
14. Lee, S.H., Lee, S., Song, B.C.: Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492* (2021)
15. Li, Y., Zhang, K., Cao, J., Timofte, R., Van Gool, L.: Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707* (2021)
16. Lin, W., Wu, Z., Chen, J., Huang, J., Jin, L.: Scale-aware modulation meet transformer. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 6015–6026 (2023)

17. Liu, Y., Sangineto, E., Bi, W., Sebe, N., Lepri, B., Nadai, M.: Efficient training of visual transformers with small datasets. *Advances in Neural Information Processing Systems* **34**, 23818–23830 (2021)
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 10012–10022 (2021)
19. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems* **34**, 12116–12128 (2021)
20. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: *International conference on machine learning*. pp. 10347–10357. PMLR (2021)
21. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 36, pp. 2441–2449 (2022)
22. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 568–578 (2021)
23. Wang, W., Chen, W., Qiu, Q., Chen, L., Wu, B., Lin, B., He, X., Liu, W.: Crossformer++: A versatile vision transformer hinging on cross-scale attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023)
24. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 22–31 (2021)
25. Xia, C., Wang, X., Lv, F., Hao, X., Shi, Y.: Vit-comer: Vision transformer with convolutional multi-scale feature interaction for dense predictions. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5493–5502 (2024)
26. Xu, W., Xu, Y., Chang, T., Tu, Z.: Co-scale conv-attentional image transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 9981–9990 (2021)
27. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10819–10829 (2022)
28. Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., Wu, W.: Incorporating convolution designs into visual transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 579–588 (2021)
29. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 558–567 (2021)
30. Zhang, B., Manoochehri, H., Ho, M.M., Fooladgar, F., Chong, Y., Knudsen, B.S., Sirohi, D., Tasdizen, T.: Class-m: Adaptive stain separation-based contrastive learning with pseudo-labeling for histopathological image classification. *arXiv preprint arXiv:2312.06978* (2023)
31. Zhang, Y., Liu, Y., Miao, D., Zhang, Q., Shi, Y., Hu, L.: Mg-vit: A multi-granularity method for compact and efficient vision transformers. *Advances in Neural Information Processing Systems* **36** (2024)

6 Appendix

Model Training Details All models, including the ResNet baselines, were trained using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, without applying weight decay. For the DuoFormer model, batch sizes were set to 32 for the Utah dataset and 6 for the TCGA dataset. We employed a OneCycle learning rate scheduler that starts from a minimal learning rate, progressively increasing to a set rate of 1×10^{-4} . Each model underwent training for 50 epochs on Utah and 100 epochs on TCGA, utilizing early stopping with patience of 20 and 50 epochs, respectively. We saved the best-performing model from the validation data for inference. Model performance was evaluated using balanced accuracy for both datasets. All computations were performed on an NVIDIA RTX A6000 with 48 GB of memory.

Ablation on Numbers of Heads and Layers We assessed our model’s sensitivity to two hyperparameters: the number of heads and the number of layers in two attention modules. Initially, we fixed the number of heads at 12 and varied the number of layers from 4 to 12 to identify optimal configurations for each dataset. Subsequently, we tested heads from 4 to 12, excluding 10 due to incompatibility with the feature dimension $D = 768$, using the optimal number of layers. We observed that performance generally increases and then decreases with attention depth. Specifically, performance peaks at 6 layers for the Utah dataset and at 8 layers for the TCGA dataset, likely due to the varying sizes of the datasets. Additionally, we noted a similar pattern of initial increase followed by a decrease in performance for the number of heads across both datasets, peaking at 8 heads. Notably, our models with all tested numbers of heads and layers performed better than baseline ResNets, except in one case where performance was slightly worse, demonstrating the effectiveness of our proposed model.

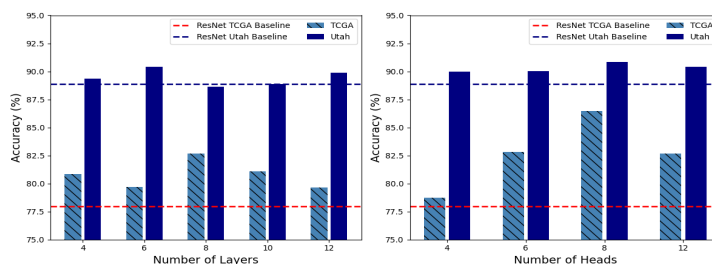


Fig. 5. Ablation studies comparing the number of layers and heads in the dual attention modules for both the TCGA (solid bars) and Utah (striped bars) datasets. The dashed lines represent ResNet baselines for each dataset. Each configuration synchronizes the layers between scale and patch attention.