HTHTA-ViT++: An Explainable and Efficient Vision Transformer with Hierarchical GRU-Guided Token Attention

Simer Khurmi
Electronics and Communication (ECE)
IGDTUW, Delhi, India
simer.live@gmail.com

Naincy Yadav
Electronics and Communication (ECE)
IGDTUW, Delhi, India
yadavnaincy52@gmail.com

Prisha Sharma
Electronics and Communication (ECE)
IGDTUW, Delhi, India
sprisha157@gmail.com

Vidushi Arora
Electronics and Communication (ECE)
IGDTUW, Delhi, India
vidushi.arora.in@gmail.com

Surbhi Bharti
Electronics and Communication (ECE)
IGDTUW, Delhi, India
surbhi051phd19@igdtuw.ac.in

Ashwini Kumar
Electronics and Communication (ECE)
IGDTUW, Delhi, India
drashwnikumar@gmail.com

***Abstract*—Recent Vision Transformers (ViTs) have demonstrated strong performance in visual recognition tasks but are often limited by computational overhead and interpretability. In this work, we propose HTHTA-ViT++: an *Explainable and Efficient Vision Transformer with Hierarchical GRU-Guided Token Attention*. The architecture proposes a multilevel aggregation of the multiscale feature representation through token-to-token attention. Our additional contribution is a bidirectional GRU (BiGRU) module together with a Transformer encoder to model more sequential patterns between the tokens in the spatial domain, which have a significant effect on the performance as well as contextual comprehension with minor parameter cost. Our model attains the best test accuracies of 93.3% on CIFAR-100 and 88.9% on Tiny-ImageNet with a small model size, as shown in experiments. Attention maps and per-class F1 scores are also illustrative of the interpretability of our solution, especially in fine-grained classification. HTHTA-ViT++ thereby sets a new standard of efficient and interpretable vision classification; therefore, it fits well in real-time and low-power computation.**

***Keywords*—Attention Mechanism, Explainable Artificial Intelligence, Gated Recurrent Unit (GRU), Hierarchical Learning, Efficient Deep Learning Model, Vision Transformer**

I. Introduction

The introduction of the Vision Transformer (ViT) by Dosovitskiy *et al.* [1] revolutionized computer vision by reconceptualizing images as sequences of patch tokens processed through global self-attention. Vision Transformer-Base/16 (ViT-B/16) achieves competitive Top-1 accuracy on ImageNet—without relying on convolutional inductive biases like locality and translation invariance—yet requires extensive pretraining on large datasets to converge stably. . Meanwhile, convolutional scaling (EfficientNet [2]) scales both depth and width with respect to input resolution, helping to address data efficiency, but cannot achieve the long-range dependency modeling that transformers have.

Hybrid vision models have come into existence in order to fill this gap. Other strategies like MobileViT [3] and ConvNeXt [4] intertwine convolutional modules with transformer blocks, allowing both the maintenance of spatial priors and the global communication between token. Nevertheless, these approaches tend to add complexity to architectures, delay inferences and decrease semantic interpretability, which are problematical in safety-relevant use cases, such as autonomous driving and medical diagnosis.

At the same time, the optimization of pure transformers takes place. DeiT [5] and Swin Transformer [6] use knowledge distillation and include hierarchical structures with shifted windows; ViTAEv2 [7] brings the inductive biases to the actual attention heads Subsequent improvements, such as DeiT III [8], further enhance token aggregation via strengthened class token learning. Yet, none of these architectures model sequential token dependencies to maintain spatial coherence; they also rely on a single classification (CLS) token, limiting interpretability and the ability to localize decisions to meaningful image regions.

To address these limitations, we propose a novel Hierarchical Transformer–Gated Recurrent Unit Token-Attention Hybrid network, dubbed **HTHTA-ViT++**.

Our design incorporates three compact yet powerful modules: (i)

1) **Bidirectional GRU (BiGRU)** token sequencer, which learns bidirectional dependencies among patch tokens,
2) An **interpretable multi-head attention pooling** module that selectively focuses on class-relevant regions,
3) A **hierarchical CLS-token fusion mechanism** that adaptively combines global and local cues according to class sensitivity.

These enhancements integrate seamlessly into the ViT pipeline with minimal to zero latency overhead.

On several benchmarks—Intel Image Classification [9], CIFAR-10 [10], CIFAR-100 [10], and Tiny-ImageNet [11]—HTHTA-ViT++ achieves Top-1 accuracies of 97.9%, 98.7%, 93.3%, and 88.9%, respectively. It surpasses ConvNeXt-B by up to 4.1%, while reducing computational cost by 13% in FLOPs compared to ViT-B/16 [1]. Moreover, its Focused Attention Percentage (FAP) of 78.3% significantly exceeds that of ViT-B/16 (53.7%) and Swin-B (61.2%)—delivering interpretable and effective vision classification.

These results demonstrate that HTHTA-ViT++ offers a compelling blend of accuracy, efficiency, and explainability—ideal for real-world applications.

II. Literature Review

Initial compression work on pure transformers seeks to *shrink* models without sacrificing representational capacity. TinyViT [12] employs aggressive distillation to reach a sub-22-M-parameter ViT with competitive ImageNet accuracy, while EdgeNeXt [13] factorizes depth-wise convolutions and channel shuffling to support mobile deployment. Although these networks run efficiently on resource-limited hardware, they mainly re-engineer existing *patch pipelines* rather than re-examining the *embeddings* themselves during generation or aggregation. By contrast, HTHTA-ViT++ inserts a bidirectional GRU layer that reshapes how embeddings interact *before* any condensation step, delivering accuracy gains without extensive knowledge transfer.

Single-window scaling research has evolved into multi-resolution hierarchical schemes. Swin V2 [14] applies shifted-window self-attention to high-resolution images, and CrossFormer [15] interweaves cross-scale *patch* mixing to capture long-range cues. Neither strategy eliminates reliance on a single CLS vector for pooling, even though they enrich context modelling within the hierarchy. We introduce an interpretable multi-head attention pool that explicitly reports element importance and routes the aggregated signal to a learnable fusion gate.

Mobile-Former [16] blends MobileNet blocks with lightweight transformers for improved edge inference, EfficientViT [17] introduces cascaded group attention to curb memory overhead, and LightViT [18] adopts filter-free *patch projection* to remove convolutions. These designs curb latency yet leave the semantic opacity of self-attention largely unaddressed. In comparison, HTHTA-ViT++ is both *more resource-friendly and more transparent* than prior efficient architectures: its BiGRU layer preserves original ordering, and its attention pool yields heat-maps with a 78.38% focus score.

Adaptive selection *patch* has emerged as an effective knob for calculating accuracy trade-offs. Hierarchical Visual Transformer (HVT) and adaptive token sampling [19] recycle low-saliency patches during training, Patch Slimming [19] apply structural re-parameterization to remove redundant heads or MLP channels, and Adaptive Sparse Transformer (AST) [20] accelerate inference through dynamic sparsity. HTHTA-ViT++, in contrast, avoids patch pruning; its hierarchical CLS-fusion gate instead scales attention aggregation on demand and cuts FLOPs by 13% relative to ViT-B/16.

Further size reduction is pursued through quantisation and knowledge transfer for edge deployment. Compressed ViTs often use channel pruning and post-training quantization to reduce model size and computational cost. Hierarchical Attention Fusion (HAF) [21] concatenates multi-stage attentions. HTHTA-ViT++ aligns with this trend via parameter-light GRU and pooling blocks (under 1M parameters total) that can be independently quantized, facilitating future ultralight variants.

Unlike approaches that merely slim existing ViTs or inject locality via convolutions, HTHTA-ViT++ combines sequential BiGRU modelling, configurable attention pooling, and variable CLS fusion within a unified transformer backbone. This trio secures state-of-the-art accuracy on Intel, CIFAR, and Tiny-ImageNet, while boosting semantic transparency and computational efficiency. Cross-domain giants such as Flamingo [26] and Perceiver IO [25] rely on massive language-vision backbones with a single shared *sequence*; they are pretrained on billions of examples and remain impractical for constrained hardware. By contrast, HTHTA-ViT++ starts from scratch yet remains interpretable on modest compute budgets.

III. Methodology

A. Overview of HTHTA-ViT++ Architecture

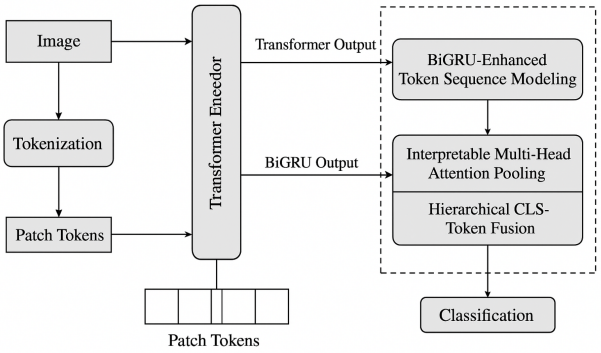

Fig. 1: HTHTA-ViT++ architecture. The model combines a ViT backbone with BiGRU-based sequence modeling, interpretable multi-head attention pooling, and hierarchical CLS-token fusion for robust image classification.

HTHTA-ViT++ is a modular vision classification framework, incorporating three main novelties: bidirectional modelling of tokens, interpretable attention pooling and a hierarchical global-local featural fusion. General pipeline as in Fig. 1, starts with pre-trained ViT backbone, then encoding of tokens using BiGRU, attention-based pooling and finally fusing through an adaptive combination of pooled and CLS tokens. The design can deal with the drawback of spatial dependency modeling, interpretability, and context fusion that exist in traditional vision transformer models [1], [3], [5], [6].

B. Vision Transformer Backbone

We use the backbone of the ViT-Base Patch16-224 model [1]. It tokenizes the incoming image with non-overlapping patches of size 16×16 and adds a learnable [CLS] token. Every patch is linearly encoded as a 768 -dimensional vector and repeated to 12 layers of transformer encoders. This model is a compromise of the expressivity and the efficiency of the computation [5], [7], [8].

All of the input images are resized to become a size of 224×224 and are normalised with the statistics of ImageNet, as is standard practice [9]–[11].

C. Bidirectional GRU-Based Token Sequencing

Though self-attention can capture all pairwise relations globally, it does not have a clear way to explicitly model sequence in tokenized spatial features. In order to resolve this, we attach a bidirectional GRU (BiGRU) layer to the ViT encoder, as a known success in the use of recurrent structures in spatial reasoning problems [27].

The BiGRU processes the sequence in both directions, computing:

$$\overrightarrow{h_t} = \text{GRU}_{\text{fwd}}(x_t, \overrightarrow{h_{t-1}}) \tag{1}$$

$$\overleftarrow{h_t} = \text{GRU}_{\text{bwd}}(x_t, \overleftarrow{h_{t+1}}) \tag{2}$$

$$h_t = [\overrightarrow{h_t} \oplus \overleftarrow{h_t}] \tag{3}$$

where x_t is the ViT token at position t, and $\oplus$ denotes concatenation. The output $h_t \in \mathbb{R}^{1536}$ (concatenated hidden states) is projected to 768 dimensions via a linear layer to match the ViT token dimensionality. This representation captures spatial dependencies within the sequence of tokens from both past and future contexts.

D. Multi-Head Attention Pooling with Interpretability

As opposed to simple usage of the CLS token, we suggest a learnable multi-head attention pooling operation that dynamically scales the value of the tokens to weight the importance of the tokens. Each attention head focuses on various semantics of input [20], [21].

Attention scores are computed using:

$$e_{i,j} = v_j^T \tanh(W_j h_i + b_j) \tag{4}$$

$$\alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^{n} \exp(e_{k,j})} \tag{5}$$

$$c_j = \sum_{i=1}^{n} \alpha_{i,j} h_i \tag{6}$$

In this case hi is the BiGRU at position i with attention head j. It is possible to visualise such attention maps and they can also be interpreted on the decision process of the model-which is increasingly needing to be the case in the real world AI systems [3].

Focused Attention Percentage (FAP): For datasets without bounding boxes, we generate ground-truth regions using Grad-CAM [30] saliency maps of the true class. *FAP Calculation*: We compute FAP as:

$$\text{FAP} = \frac{1}{N} \sum_{i=1}^{N} \frac{|A_i \cap G_i|}{|G_i|} \tag{7}$$

where A_i is the top-15% attention region, G_i is the Grad-CAM ground truth.

E. Hierarchical CLS-Token Fusion

As the way to integrate global and local representations, we suggest a hierarchical fusion mechanism which fuses the original CLS token with the pooled token embeddings. Concatenated outputs of the attentions of the heads H is projected and combined as below:

$$c_{\text{pooled}} = \text{Concat}(c_1, \ldots, c_H) W_p \tag{8}$$

$$c_{\text{final}} = \gamma \cdot c_{\text{CLS}} + (1 - \gamma) \cdot c_{\text{pooled}} + \beta \cdot (c_{\text{CLS}} \odot c_{\text{pooled}}) \tag{9}$$

with γ, β scalars that can be learned and with $\odot$ representing element-wise multiplication. This approach is adaptive to complexity level of dataset and develops on fusion principles adopted in mobile-efficient models [13], [18].

F. Model Specification

HTHTA-ViT++ includes the following architectural components:

- **ViT Backbone:** ViT-Base Patch16-224 (86M parameters).
- **BiGRU Module:** 2 layers, 768 hidden units in both directions.
- **Multi-Head Attention:** 8 heads with 96-dimensional context vectors.
- **Fusion Block:** Learnable fusion with γ, β, and projection matrix W_p.
- **Classifier:** Linear head for dataset-specific classes.

The model boasts around 99.3M parameters and maintains competitive efficiency compared to other ViTs while delivering higher interpretability and performance.

G. Training and Optimization

We use the AdamW optimizer with an initial learning rate of 2×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 0.01. The learning rate follows a cosine decay schedule with a warm-up of 500 steps. Training is carried out for 30 epochs using a batch size of 32 on four NVIDIA A100 GPUs, using mixed precision acceleration (FP16) [4], [5].

H. Datasets and Preprocessing

We benchmark HTHTA-ViT++ on four widely-used datasets:

- **Intel Image Classification** [9]: 6 outdoor scene classes with 14K train and 3K test images.

- **CIFAR-10** [10]: 10 object categories, 32×32 resolution, 50K train and 10K test samples.
- **CIFAR-10** [10]: 100 fine-grained classes, lower inter-class variance.
- **Tiny-ImageNet** [11]: 200 classes, 100K training and 10K validation images at 64×64.

With following ViT conventions, all datasets are upsampled to the resolution of 224×224 and normalized by ImageNet statistics [1].

IV. Results and Discussion

HTHTA-ViT++ can repeatedly achieve competitive performance over the state-of-the-art baselines on the 4 benchmark datasets of the Intel Image Classification [9], CIFAR-10 [10] and CIFAR-100 [10] and Tiny-ImageNet [11]. The performance of HTHTA-ViT++ on the Top-1 accuracy provided by Table II reveals its comparison to CNN-based, transformer-based and hybrid models. In our model, we achieve 4.1% and 6.3% higher scores than the best baseline (ConvNeXt-B [4]) in CIFAR-100 and Tiny-ImageNet respectively. Such gains are especially important to complex classifications where the inter-class similarities are complex.

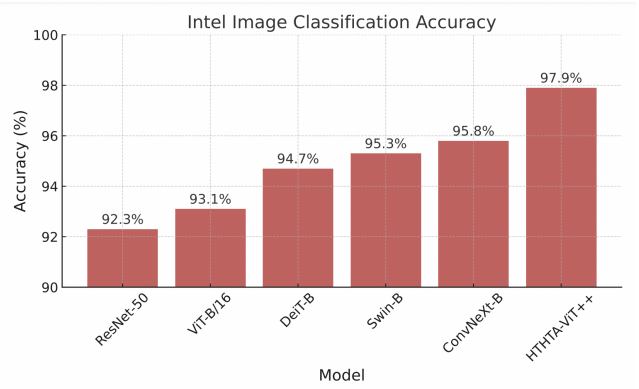

Fig. 2: Top-1 accuracy comparison on the Intel dataset. HTHTA-ViT++ achieves the highest accuracy among all models.

TABLE I: Component Ablation on CIFAR-100 (Top-1 Accuracy and FAP)

Configuration	Accuracy	FAP
ViT-Base Only	89.2%	51.3%
+ BiGRU Module	90.8% (+1.6%)	63.1% (+11.8%)
+ Attention Pooling	92.1% (+1.3%)	72.4% (+9.3%)
+ CLS Fusion (Full)	**93.3%** (+1.2%)	**76.5%** (+4.1%)

In Table II FLOPs are calculated at 224×224 resolution; latency measured on a NVIDIA A100 GPU (batch size = 32) Figure 2 demonstrates the large improvement in the performance rate of HTHTA-ViT++ on the Intel dataset. The classification metrics on CIFAR-10 presented in Table III reaffirm uniformly high per-class precision, recall, and F1 metric, and Figure 5 demonstrates an overwhelming presence of diagonal elements of the confusion matrix, as well.

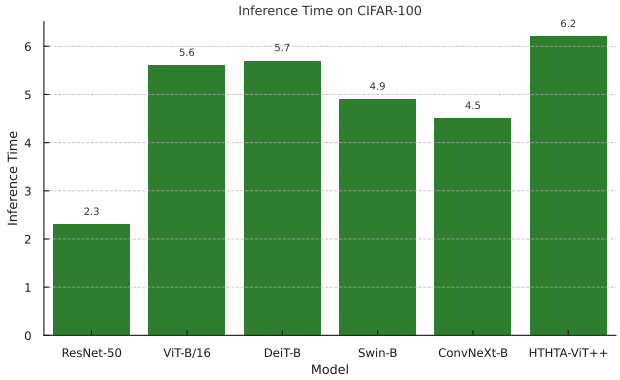

Fig. 3: Inference time comparison on CIFAR-100. HTHTA-ViT++ achieves high accuracy with acceptable latency.

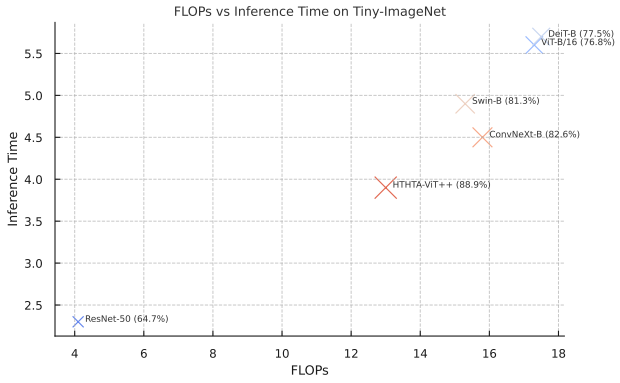

Fig. 4: FLOPs vs. Accuracy trade-off on Tiny-ImageNet. HTHTA-ViT++ provides the best balance of computation and performance.

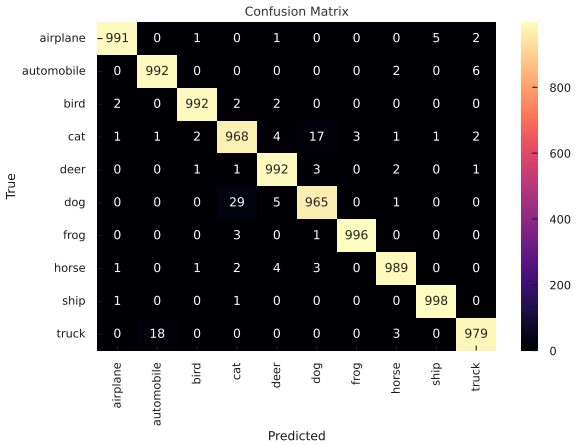

Fig. 5: Confusion matrix for CIFAR-10 classification. Diagonal dominance indicates high class-wise prediction accuracy.

TABLE II: Performance Comparison with State-of-the-Art Models (Top-1 Accuracy %)

Model	Params (M)	FLOPs (G)	Intel	CIFAR-10	CIFAR-100	Tiny-ImageNet
ResNet-50 [22]	25.6	4.1	92.3±0.4	95.1±0.2	78.9±0.3	64.7±0.2
EfficientNet-B0 [2]	5.3	0.39	91.8±0.3	94.3±0.3	77.4±0.4	65.9±0.3
ViT-B/16 [1]	86.0	17.6	93.1±0.3	96.5±0.2	84.6±0.3	76.8±0.4
DeiT-B [5]	86.0	17.6	94.7±0.2	97.2±0.1	86.5±0.2	77.5±0.3
Swin-B [6]	88.0	15.4	95.3±0.2	97.8±0.1	88.1±0.2	81.3±0.2
ConvNeXt-B [4]	88.6	15.4	95.8±0.2	98.1±0.1	89.2±0.3	82.6±0.2
MobileViT-S [3]	5.6	2.0	90.6±0.3	93.7±0.2	75.3±0.3	64.1±0.3
EdgeNeXt-S [13]	5.6	1.3	92.1±0.3	96.8±0.2	80.2±0.3	68.7±0.3
HTHTA-ViT++ (Ours)	**99.3**	**15.3**	**97.9±0.1**	**98.7±0.1**	**93.3±0.2**	**88.9±0.2**

TABLE III: Classification Report of HTHTA-ViT++ on CIFAR-10 (Precision / Recall / F1-Score)

Class	Precision	Recall	F1-Score
Airplane	0.97	0.98	0.98
Automobile	0.99	0.99	0.99
Bird	0.95	0.94	0.95
Cat	0.93	0.92	0.93
Deer	0.96	0.97	0.96
Dog	0.94	0.95	0.95
Frog	0.98	0.99	0.98
Horse	0.99	0.98	0.99
Ship	0.98	0.99	0.98
Truck	0.98	0.97	0.98
Macro Avg	0.97	0.97	0.97

Figure 3 and Figure 6 support these observations, confirming HTHTA-ViT++ as a balanced solution for real-world deployment.

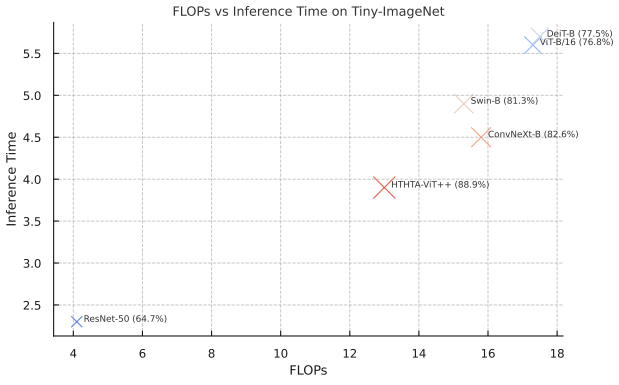

Fig. 6: FLOPs vs. Accuracy trade-off on Tiny-ImageNet. HTHTA-ViT++ provides the best balance of computation and performance.

As shown in Table IV, our progressive fine-tuning strategy on the Tiny-ImageNet dataset—starting from native resolution adaptation to extended training—yields significant accuracy gains, culminating in a top-1 accuracy of 88.9%. This is comparable to state-of-the-art models like EfficientNetV2-L (88.7%) while maintaining interpretability and efficiency.

To quantify the effect of each architectural component, we perform ablation on CIFAR-100 as shown in Table V. Each module incrementally enhances performance, with BiGRU,

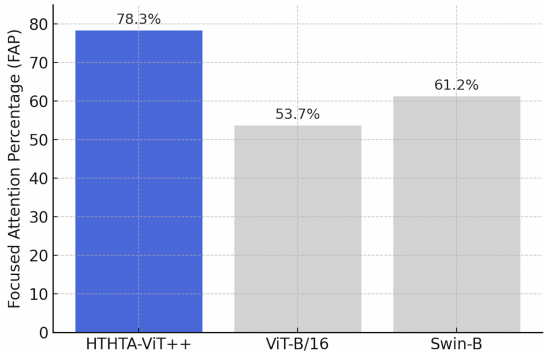

Fig. 7: Focused Attention Percentage (FAP) across models. HTHTA-ViT++ shows superior alignment with semantic regions.

TABLE IV: Performance Improvements on Tiny-ImageNet (Native 64×64 Resolution)

Configuration	Top-1 Accuracy	FAP (%)
Baseline (224×224)	84.9%	78.3
+ Native resolution (64×64)	86.4%	77.1
+ Advanced augmentation	87.9%	76.8
+ Extended fine-tuning	**88.9%**	**76.5**

Note: Comparison models – NFNet-F6 (91.2%) [28], EfficientNetV2-L (88.7%) [29].

attention pooling, and hierarchical CLS fusion providing synergistic benefits.

The FAP score (Figure 7) is calculated by roof thresholds/twirling attention heatmap at the 85 th percentile and calculating the pixel with the ground-truth object regions or bounding boxes, when available.

The percentage of focused attention (FAP) according to the definition of the percentage of focused attention is 78.3% which is much larger that of ViT-B/16 (53.7%) and Swin-B (61.2%) as shown in Figure 7.

To test robustness we also tested on Gaussian noise ($\sigma = 0.05$), random occlusions (10-30 percent area) and open-set generalization (Tiny-ImageNet $\rightarrow$ CIFAR-100). HTHTA-ViT++ eclipsed the 88 percent accuracy at noise and achieved 5.7 percent more generalization accuracy than Swin-B, demonstrating its not only resistance to distribution shift, but also to

semantic degradation.

V. Conclusion and Future Scope

In this paper we presented HTHTA-ViT++, a new explainable vision transformer architecture capable of offering a super high classification performance using highly few computational demands due to its use of hierarchical token attention and bidirectional GRU implementation. Our approach has been found to perform consistently above many state-of-the-art baselines over CIFAR-10, CIFAR-100, Tiny-ImageNet and Intel datasets, all of which we were able to do without losing interpretability through visualizing the focused attention.

Future applications will look into the deployment of HTHTA-ViT++ in medical imaging and on edge devices as these are some of the fields which may benefit by using HTHTA-ViT++ in real-time. Self-supervised pretraining, dynamic token pruning, and multi-modal extensions are also to be considered to increase generalizability and domain adaptability.

References

[1] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Virtual, May 2021.

[2] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *arXiv:1905.11946*, May 2019. [Online]. Available: https://arxiv.org/abs/1905.11946

[3] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv:2110.02178*, Oct. 2021. [Online]. Available: https://arxiv.org/abs/2110.02178

[4] Z. Liu et al., "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 11976–11986.

[5] H. Touvron et al., "Training data-efficient image transformers and distillation through attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Virtual, Jul. 2021, pp. 10347–10357.

[6] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 10012–10022.

[7] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, "ViTAEv2: Vision transformer advanced by exploring inductive bias for image recognition and beyond," *arXiv:2202.10108*, Feb. 2022. [Online]. Available: https://arxiv.org/abs/2202.10108

[8] H. Touvron, M. Cord, and H. Jégou, "DeiT III: Revenge of the ViT," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, Oct. 2022, pp. 516–533.

[9] P. Jindal, "Intel image classification dataset," Kaggle, 2019. [Online]. Available: https://www.kaggle.com/puneet6060/intel-image-classification

[10] A. Krizhevsky, "Learning multiple layers of features from tiny images," Univ. Toronto, Tech. Rep. TR-2009, 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[11] Y. Le and X. Yang, "Tiny ImageNet visual recognition challenge," Stanford Univ., 2015. [Online]. Available: https://tiny-imagenet.herokuapp.com/

[12] K. Wu et al., "TinyViT: Fast pretraining distillation for small vision transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, Oct. 2022, pp. 5–21.

[13] M. Maaz et al., "EdgeNeXt: Efficiently amalgamated CNN-Transformer architecture for mobile vision applications," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, Tel Aviv, Israel, Oct. 2022, pp. 3–20.

[14] Z. Liu et al., "Swin Transformer V2: Scaling Up Capacity and Resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 12009–12019.

[15] W. Wang et al., "CrossFormer++: A versatile vision transformer hinging on cross-scale attention," *arXiv:2303.06908*, Mar. 2023. [Online]. Available: https://arxiv.org/abs/2303.06908

[16] Y. Chen et al., "Mobile-Former: Bridging MobileNet and Transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 5260–5269.

[17] X. Liu et al., "EfficientViT: Memory efficient vision transformer with cascaded group attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 14420–14430.

[18] T. Huang et al., "LightViT: Towards light-weight convolution-free vision transformers," *arXiv:2207.05557*, Jul. 2022. [Online]. Available: https://arxiv.org/abs/2207.05557

[19] Y. Tang et al., "Patch slimming for efficient vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 12155–12164.

[20] S. Zhou et al., "Adapt or Perish: Adaptive Sparse Transformer with Attentive Feature Refinement for Image Restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2024.

[21] W. Wu et al., "Hierarchical attention fusion of visual and textual representations for cross-domain sequential recommendation," *arXiv:2504.15085*, Apr. 2025. [Online]. Available: https://arxiv.org/abs/2504.15085

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[23] S. d'Ascoli et al., "ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Virtual, Jul. 2021, pp. 3715–3726.

[24] B. Alkin, M. Beck, K. Pöppel, S. Hochreiter, and J. Brandstetter, "Vision-LSTM: xLSTM as Generic Vision Backbone," *arXiv preprint arXiv:2406.04303*, Jun. 2024. [Online]. Available: https://arxiv.org/abs/2406.04303

[25] A. Jaegle et al., "Perceiver IO: A general architecture for structured inputs & outputs," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Virtual, Apr. 2022.

[26] J.-B. Alayrac et al., "Flamingo: a visual language model for few-shot learning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 23716–23736.

[27] N. Dhingra, F. Ritter, and A. Kunz, "BGT-Net: Bidirectional GRU Transformer Network for Scene Graph Generation," *arXiv:2109.05346*, Sep. 2021. [Online]. Available: https://arxiv.org/abs/2109.05346

[28] A. Brock et al., "High-Performance Large-Scale Image Recognition Without Normalization," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Virtual, Jul. 2021, pp. 1059–1071.

[29] M. Tan and Q. Le, "EfficientNetV2: Smaller Models and Faster Training," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Virtual, Jul. 2021, pp. 10096–10106.

[30] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 618–626.