Aligning Hierarchical Vision Transformers as Multi-Scale Dense Feature Extractors

Rebanta Dey, Tribikram Dhar, Samrat Dutta
TCS Research, Bangalore, India

***Abstract*—Dense prediction requires representations that are multi-scale, semantically aligned, and resolution stable, yet existing hierarchical vision transformers are task-specific, optimization-sensitive, and inefficient under scale variation. We propose MAHViT, a self-supervised hierarchical vision transformer designed as a multi-scale dense feature extractor. MAHViT introduces a dual-Layer Normalization data-flow that stabilizes deep hierarchical attention by jointly preserving gradient flow and feature diversity, enabling scalable self-supervised pre-training. To efficiently connect hierarchical encoders with dense decodes, we propose a Feature Diversity Sampling (FDS) neck that suppresses redundant tokens while retaining semantically distinct features. Pre-trained using a BYOL-based objective with resolution-adaptive alignment, MAHViT encoder learns scale-consistent representations. The encoder is evaluated on semantic segmentation task by pairing a Progressive Upsampling decoder with the FDS Neck. Through extensive evaluations, MAHViT achieves state-of-the-art or competitive multi-scale mIoU on COCO-2017, ADE20K, and Cityscapes across multiple model scales, while maintaining favorable FLOPs–accuracy trade-offs suitable for edge deployment.**

***Index Terms*—Hierarchical Vision Transformer, Self-Supervised Pre-Training, Semantic Segmentation**

I. Introduction

Convolutional Neural Networks (CNNs) have long been the standard approach for dense prediction tasks in computer vision due to their strong local feature extraction and weight-sharing properties. Hierarchical CNN architectures are particularly effective as they produce multi-scale features that capture both local details and global context. However, these models rely heavily on strong inductive biases, which can limit their ability to generalize across diverse datasets.

The introduction of Vision Transformers (ViTs) [1] shifted the community toward attention-based architectures [2]–[4] due to their strong generalization capabilities. ViT-based models now serve as the backbone of modern vision foundation models [5], [6]. However, standard ViTs produce fixed-resolution feature maps that are well suited for classification but less effective for dense prediction tasks such as segmentation, detection, and depth estimation. Hierarchical Vision Transformers (HViTs) [3], [4], [7] address this by generating multi-resolution features, but they are often trained in a supervised manner and may suffer from instability across scales. Supervised learning remains widely used in computer vision because labeled data provides clear training signals. However, large-scale annotation is expensive and often encourages models to learn shortcuts that limit their transferability.

Self-supervised learning addresses this limitation by enabling models to learn intrinsic image structures without human labels. Approaches include contrastive learning [8], [9], feature-alignment methods [5], [10], [11], predictive latent approaches [12], and masked image modeling [13], [14]. While most of this literature relies on ViT architectures due to their strong global representation learning, their single-scale outputs remain a limitation for dense prediction tasks.

In this work, we study the impact of self-supervised pre-training for a hierarchical vision transformer and demonstrate its effectiveness on semantic segmentation. We improve traditional HViTs by introducing a stable training pipeline with a dual-normalization transformer block, termed the Feature Efficient Transformer (FeT), to mitigate representation collapse and vanishing gradients. To evaluate the encoder, we adopt a multi-stage PUP decoder for semantic segmentation and introduce the FDS (*Feature Diversity Sampling*) Neck to bridge the encoder and decoder. This module reduces feature redundancy, enabling faster convergence, lower FLOPs, and more efficient inference suitable for edge deployment. Although our experiments focus on semantic segmentation, the proposed encoder and training strategy remain task-agnostic.

Our contributions are summarized as follows:

1) We propose Multi-scale Aligned Hierarchical Vision Transformer (MAHViT), a self-supervised hierarchical vision transformer encoder architecture for multi-scale dense feature extraction.
2) We improve upon existing HViTs by introducing dual normalization strategy in the data-flow mechanism of our transformer block to mitigate representation collapse and vanishing gradient.
3) We propose a diversity-aware feature sampling technique a learnable FDS Neck to reduce feature redundancy during dense prediction applications.
4) We showcase the efficacy of our work by training a multi-stage PUP decoder for semantic segmentation, which handles the multi-scale features from the pre-trained HViT.

II. Methodology

Our framework follows an encoder–decoder architecture built on a hierarchical vision transformer backbone. The encoder generates multi-scale features, which the decoder uses for dense prediction. Before describing the task-specific components, we first outline the architectural modifications

Fig. 1: Patch similarity visualization of MAHViT-Advanced, computed using cosine similarity between patches after the resolution adaptation phase (II-B). The input resolution is set to 1536×1536, and the reference patch is marked with $\times$

that enable hierarchical transformers to act as effective dense feature extractors.

A. Encoder Architecture

1) Hierarchical Vision Transformer as a Dense Feature Extractor: Hierarchical Vision Transformers (HViTs) are well suited for dense prediction tasks such as detection and segmentation because they generate multi-scale feature maps through progressive downsampling. This preserves fine-grained spatial information in early stages while capturing broader semantic context in deeper layers, naturally aligning with decoder architectures used in dense prediction.

Our Feature Efficient Transformer (FeT) encoder processes an input image $\mathcal{I}_{H\times W\times 3}$ and generates four progressively downsampled feature maps, through a stack of multi-scale transformer blocks of shape

$$F_i = \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$$

where $i \in \{1, 2, 3, 4\}$. Each stage expands the receptive field and captures increasingly global dependencies, enabling the model to encode both high-frequency local structures and long-range contextual cues.

2) Positional-free spatial encoding: Fixed absolute positional embeddings often degrade under scale changes due to interpolation errors. We mitigate this by removing fixed embeddings entirely and injecting spatial biases through the MixFFN module [7]. Its 3×3 depth-wise convolution introduces local structure while remaining resolution-agnostic, and producing feature representations that transfer reliably to new input sizes.

3) Learnable Sequence Reduction for high-resolution scalability: Transformers scales with quadratic memory and computational cost with respect to sequence length, making high-resolution image processing expensive. To mitigate this, we adopt the learnable sequence-reduction technique from [4], which compresses key–value tokens before attention. This allows each query to attend to a smaller set of keys, reducing the complexity from $\mathcal{O}\left(N^2\right)$ to $\mathcal{O}\left({}^{N^2}/_{R}\right)$, where R denotes the Sequence Reduction (SR) ratio. As a result, hierarchical vision transformers can process more tokens without exceeding memory constraints. Sequence Reduction is particularly effective for vision tasks because neighboring image patches often contain redundant information, such as similar textures and structures. The learnable reduction merges these redundant key–value features while preserving important spatial cues, enabling our Multi-Head Sequence Reduction Attention (MHSRA) to focus on salient structures while maintaining representational quality.

4) Dual-LN Data Flow mechanism: Training deep attention-based models is often hindered by vanishing gradients and representation collapse. Pre-normalization improves training stability by removing mean bias and normalizing

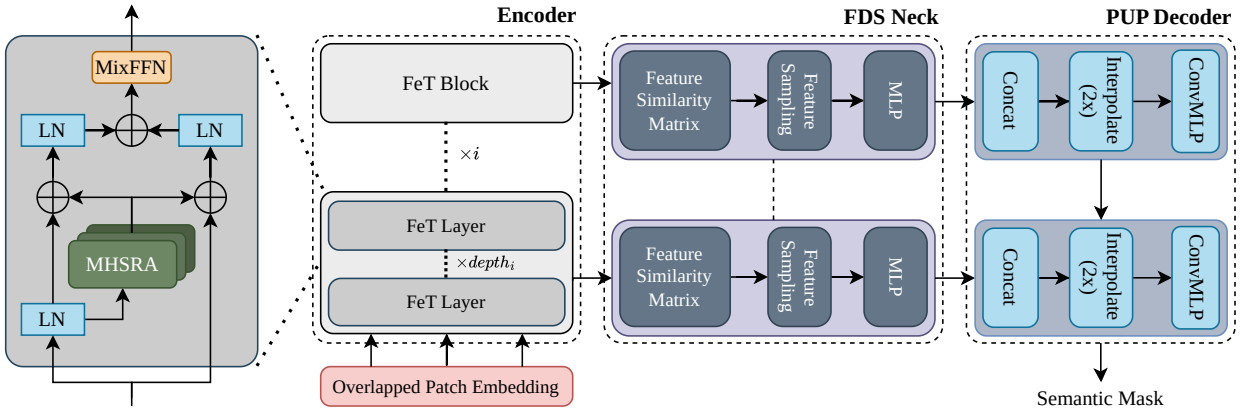


Fig. 2: **MAHViT for Semantic Segmentation Overview:** An input of size $H \times W \times 3$ is converted into overlapping patches using an Overlapped Patch Embedding (OPE) layer and processed by hierarchical Feature Efficient Transformer (FeT) blocks. Outputs from each stage are reshaped into spatial tokens and passed through the Feature Diversity Sampling (FDS) module to reduce redundancy before being sent to the decoder. The decoder then progressively upsamples these features to generate the final segmentation mask.

feature scales before the attention operation, effectively turning the residual path into an identity mapping. However, this normalization constrains features to a uniform-variance manifold, which, across deep and densely stacked blocks, can progressively reduce feature distinctions and increase the risk of representation collapse. In contrast, post-normalization applies normalization after the residual aggregation. While this helps preserve feature diversity, repeated normalization in shallower layers can significantly reduce gradient magnitudes as depth increases, thereby hampering effective optimization. To balance these effects, we adopt a dual Layer Normalization (dual-LN) strategy, inspired by prior work [15], within the data flow of attention blocks. The pre-normalization component stabilizes gradient propagation and alleviates vanishing gradients, while the post-normalization component promotes feature diversity and mitigates representation collapse.

B. Self-Supervised Encoder Pre-Training

Self-supervised learning enables models to capture intrinsic structures from unlabeled data, reducing reliance on annotation-driven shortcuts while improving robustness and generalization. By leveraging large-scale raw images, the encoder learns transferable representations that benefit downstream tasks. We adopt BYOL [10] as our pre-training framework, which aligns feature outputs of an online and a target network through a similarity-based objective without requiring negative pairs. This stable, non-contrastive formulation has proven effective for learning semantic invariances.

We first train the hierarchical transformer encoder with the BYOL objective to build a strong dense feature extractor. This stage encourages the encoder to produce semantically consistent representations suitable for dense prediction tasks. We use standard BYOL augmentations such as center cropping, color jittering, random flipping, and rotation, while removing geometric transformations during the supervised stage to preserve spatial fidelity. Following DinoV3 [6], we also introduce a high-resolution adaptation phase using Gram Anchoring, where high-resolution inputs are provided to the target network to improve scale-consistent feature extraction. The overall pre-training loss is defined as

$$\mathcal{L} = \alpha\mathcal{L} * BYOL(y_o, y_t) + (1-\alpha)\mathcal{L} * GRAM(y_o, y_t)$$

where α balances the BYOL objective and Gram-based alignment. In our experiments, we set $\alpha = 0.5$ during the resolution adaptation stage.

TABLE I: We ablate our model performance with GFLOPs vs FPS on different dataset (left), kindly note that the difference in GFLOPs is due to the addition of dataset specific class head. We also show the efficacy of BYOL pre-training using Linear Evaluation setting on ImageNet-1K (right).

(a) GFLOPs and FPS of different model variant on ADE-20k, Cityscapes and COCO-2017.

Encoder Variant	Parameters (M)		ADE-20k		Cityscapes		COCO-2017	
	Encoder	Decoder	GFLOPs	FPS	GFLOPs	FPS	GFLOPs	FPS
MAHViT-Tiny	5.1	1.5	8.95	36.29	6.48	58.78	7.63	50.95
MAHViT-Lite	20.3	5.4	21.73	33.63	19.26	55.73	20.41	46.41
MAHViT-Base	33.0	8.8	39.14	23.95	34.20	38.11	36.5	32.08
MAHViT-Advanced	44.5	8.8	44.48	17.73	39.54	26.23	41.84	23.58

(b) ImageNet-1K eval.

BYOL	Supervised
75.1	82.1
79.5	84.3
83.1	86.1
86.9	88.9

C. FDS Neck

Hierarchical transformer encoders produce a large number of intermediate tokens, many of which are redundant due to overlapping receptive fields and repeated attention refinement. This redundancy becomes more pronounced in self-supervised backbones, where the absence of label constraints allows correlated or low-variance representations to accumulate. Passing all tokens to the decoder is inefficient and can dilute meaningful structure, as redundant features tend to dominate subsequent attention operations. To address this, we introduce the Feature Diversity Sampling (FDS) neck, a lightweight module that filters encoder outputs based on feature distinctiveness. Given an encoded token set $X \in \mathrm{R}^m$, FDS computes pairwise cosine similarities and assigns each token a redundancy score defined by its maximum similarity with other tokens. Tokens with lower scores are more unique. FDS retains the least redundant 50% of tokens, preserving diverse and informative features while discarding repetitive ones. This reduces decoder computation and improves feature quality without introducing additional losses, supervision, or architectural complexity, providing a simple and task-agnostic way to deliver compact feature sets for dense prediction tasks.

D. Decoder Architecture for Semantic Segmentation and Supervised Training

In the decoder, we employ a progressive upsampling (PUP) strategy to gradually restore dense features to higher spatial resolutions. Each decoder block upsamples its input by a factor of two via bilinear interpolation and fuses the result with shallow features from earlier encoder stages. Convolutional layers then refine the fused representations, leveraging their local inductive bias and weight-sharing properties to capture fine spatial details and maintain boundary precision. This progressive reconstruction enables smooth recovery of semantic structures while remaining computationally efficient. Although the architecture can be applied to various dense prediction tasks, our experiments focus on semantic segmentation.

Building on the encoder's role as a dense feature extractor, we attach the FDS neck and PUP decoder to adapt the network for semantic segmentation. These modules are trained in a supervised manner to generate high-quality pixel-level masks from the pre-trained hierarchical features. To address class imbalance and enhance boundary delineation, training is guided by a combination of multi-class Dice loss and multi-class Focal loss.

III. Result and Discussion

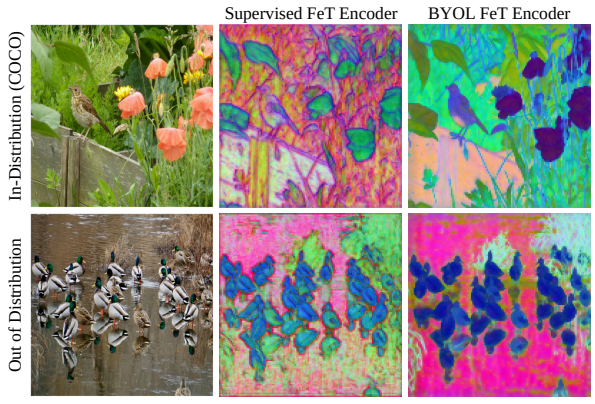
A. Experimental Setup

1) Dataset: We pre-train our encoder architecture on ImageNet-1K [16], a common benchmark for large scale pre-training, using both training and validation set. We used the same augmentation strategy used in BYOL for generating the two views. In the post-training resolution adaptation phase, we removed the local crops and provided only global crops to both the networks.

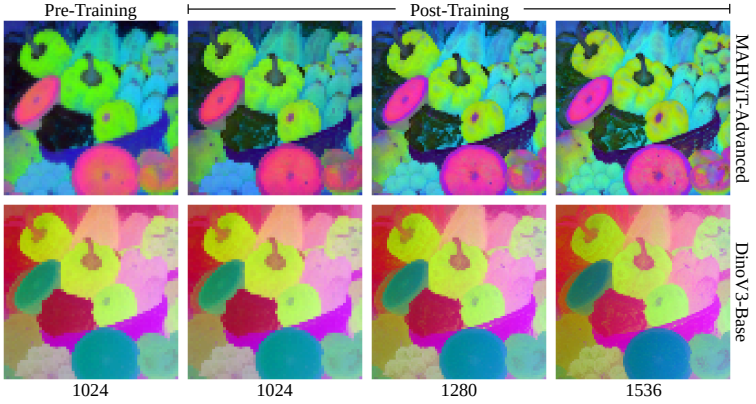
To showcase the efficacy of our pre-trained encoder, we evaluated our encoder on semantic segmentation task on three popular benchmark - COCO-2017 [17], ADE20K [18] and Cityscapes [19].

2) Implementation Detail: All experiments were conducted using 4 Nvidia RTX 6000 ADA. We adopted gradient accumulation technique and mixed precision to maximize efficiency and batch capacity. We set the training resolution to 256 for optimal speed and accuracy during the pre-training stage, and increased the resolution up to 768 during the post-training stage. We set the batch size to 1536, 0.996 as the momentum update in Exponential Moving Average (EMA) update and kept AdamW as the optimizer with weight decay set to 1e-4. We set a warmup schedule of 20 epochs with a base LR of 1e-5 and increased to 1e-4 via cosine annealing strategy. During the decoder training, we kept the same LR of 1e-4 and used ReduceOnPlateau learning rate scheduling.

During the segmentation decoder training, we used random training resolution with scales set between



(a) The first three component of the PCA, mapped to RGB space. We show the output of the third stage of the MAHViT-Advanced encoder. We set the input resolution to 2048×2048.



(b) We show feature space stability under resolution change by showing the PCA map. We compared MAHViT-Advanced with DinoV3-Base [6] and showed similar resolution agnostic behavior.

Fig. 3: PCA visualization of pre-training strategy and input resolution stability on the geometry of encoder feature space.

`[0.5,0.75,1.,1.25]` with base resolution set at 512, and ran the experiments for 50 epochs for COCO, 20 for Cityscapes and 100 for ADE20k. We set the overall batch size to 16 for all variants during the decoder training without gradient accumulation.

B. Ablation Study

1) Model Variants: We scaled up our model to 4 different variants - Tiny, Lite, Base and Advanced. Table Ia shows the the number of Floating Point Operations (FLOPs) and inference FPS on ADE-20k, Cityscapes and COCO-2017 Validation dataset. For the calculation of FPS, we have not used any accelerated implementation like TensorRT, to demonstrate the real-time performance of our models. We set the resolution to 512×512 during and used a batch size of 1.

2) SSL Pre-Training: Self-Supervised pre-training provides a strong foundation for a universal feature extractor. Here we showed the effect of self-supervised pre-training on the classification result of ImageNet-1K and also the patch feature result.

a. Linear Probing: we compare two settings, MAHViT trained from scratch with a classification head, and MAHViT using a BYOL-pretrained encoder with linear probing. The self-supervised trained encoder performs comparable to the fully supervised trained encoder. We show this result in table Ib.

b. Patch Feature Study: We visualize PCA maps for both methods on In-Distribution (ID) and Out-of-Distribution (OOD) images. For ID evaluation, the supervised encoder uses weights trained on the same dataset. As shown in Fig. 3a, the self-supervised encoder produces cleaner, more semantically consistent features, which are better suited for dense prediction tasks like segmentation. In contrast, supervised training often learns shortcuts that improve metrics but may not reflect true object semantics.

c. Resolution Perturbation Study: We analyze the effect of input resolution changes by visualizing the first three principal components mapped to RGB space in Fig. 3b. The results show that our pre-trained encoder effectively handles the increased covariance in feature space, consistently grouping semantically similar features. We compare MAHViT-Advanced with DinoV3-Base and observe comparable stability under resolution changes. Note that the feature space of DinoV3 differs from MAHViT, which leads to different color distributions in the RGB visualization.

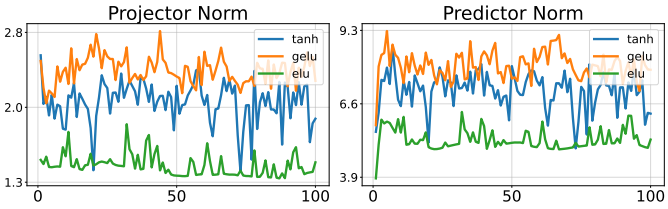


Fig. 4: Projector and predictor norms for different activation functions during BYOL pre-training, measured across the full training schedule (epochs) of the online network.

d. Activation Function Study: We conducted an activation function study during encoder pre-training to examine how different choices affect feature alignment between the online and target networks. Figure 4 compares three activa-

TABLE II: Performance comparison: the left table reports semantic segmentation results on the three benchmarks described in III-A1, while the right table presents ImageNet-1K classification results of our encoder.

(a) We compared our proposed model with recent approaches on the three benchmarks and report the multi-scale mIoU score.

Model	Backbone	Params (M)	COCO	ADE-20k	Cityscapes
SegFormer-B0 [7]	MiT	3.7	35.6	38.0	78.1
EDAFormer-T [20]	EFT	4.9	40.3	42.3	78.7
SegMAN-T [21]	SegMAN	6.4	41.3	43.0	80.3
MAHViT-Tiny (ours)	**FeT**	**6.65**	**38.8**	**38.6**	**81.3**
SegNeXt-B [22]	MSCAN-B	27.6		49.9	83.8
SegMAN-S [21]	SegMAN	29.4	47.5	51.3	83.2
EDAFormer-B [20]	EFT	29.4	45.8	48.9	81.6
MAHViT-Lite (ours)	**FeT**	**25.6**	**45.6**	**43.3**	**83.4**
CGRSeg-B [23]	EfficientFormerV2			45.5	
MaskFormer-T [24]	Swin	42.0		48.8	
SegNeXt-L [22]	MSCAN	48.9		52.1	83.9
MAHViT-Base (ours)	**FeT**	**41.8**	**56.6**	**50.9**	**83.9**
CGRSeg-L [23]	EfficientFormerV2			48.3	
SegMAN-B [21]	SegMAN	51.8	52.6	52.6	48.4
VWFormer-B4 [25]	MiT	64.0	47.6	51.6	84.0
SegFormer-B4 [7]	MiT	64.1	46.5	51.1	83.4
SegFormer-B5 [7]	MiT	84.7	46.7	51.8	83.5
VWFormer-B5 [25]	MiT	84.6	48.0	52.7	84.3
DtDsPR [26]	ViT-B	86.5	47.0		
SegMAN-L [21]	SegMAN	92.4	48.8	53.2	84.2
MAHViT-Advanced (ours)	**FeT**	**53.3**	**61.3**	**56.4**	**84.4**

(b) Linear evaluation on ImageNet-1K. We report the Top-1 accuracy on the ImageNet-val set.

Methods	Models	Params (M)	Top-1%
Supervised	DeIT-S/16 [2]	22.1	79.8
	PVT-S [4]	24.5	79.8
Self-Supervised	BYOL-ResNet-50 [10]	25.6	74.3
	DINO-ViT-S/8 [11]	21.0	79.2
	MoCo-v3-ViT-S/16 [9]	21.0	73.2
	MoBY-ViT-S/16 [27]	21.0	72.8
	MoBY-Swin-T [27]	29.0	75.0
	DIINO+SelfPatch-ViT-S/16 [28]	21.0	75.6
	MAHViT-Lite (ours)	20.3	79.5
Supervised	PVT-L [4]	61.4	81.7
	Swin-B [3]	87.8	83.5
	ViT-B [1]	86.0	83.2
	CPVT-B [29]	88.0	82.3
Self-Supervised	HiViT-B [30]	66.4	83.8
	MixMAE-Swin-B/W14 [31]	88.0	85.1
	SimMIM-ViT-B [14]	86.0	83.8
	DINO-ViT-B/16 [11]	86.0	80.1
	DINOv2-ViT-B/8 [5]	86.0	84.6
	MAHViT-Advanced (ours)	44.0	86.9

tions by analyzing the predictor and projector norms of the online network. Larger norms generally indicate stronger feature magnitudes and clearer separation in representation space, suggesting greater confidence in distinguishing visual patterns. As shown in the figure, GELU consistently produces the highest norms for both components, and is therefore selected as the activation function in the final encoder design.

C. Patch Similarity Analysis

We use the patch similarity analysis shown in fig 1, to evaluate the quality of learned visual representations by examining whether semantically similar local regions are mapped close to each other in feature space. By retrieving nearest-neighbor patches using cosine similarity, we demonstrate the robustness, invariance and discriminative ability of the feature extractor. This analysis provides intuitive and label-free insight into what visual cues the encoder captures and how reliably it generalizes across local variations.

D. Qualitative Analysis

We present qualitative results of our method in Fig. 5 across all three benchmarks. For mask generation, the confidence threshold is set to 0.5, and random RGB colors are assigned to each class ID. The results show reliable mask predictions on complex structures and textures, highlighting the robustness of the model in challenging scenarios.

We further evaluate segmentation performance on Out-of-Distribution (OOD) images in Fig. 6, comparing MAHViT-Base with MaskFormer-Swin-Tiny trained on the COCO dataset, which lies in a similar parameter range. While MaskFormer-Swin-Tiny is trained end-to-end on COCO, our model demonstrates the benefits of the proposed pre-training and architectural design. Despite using a smaller parameter budget, our model achieves comparable performance, making it more suitable for edge devices.

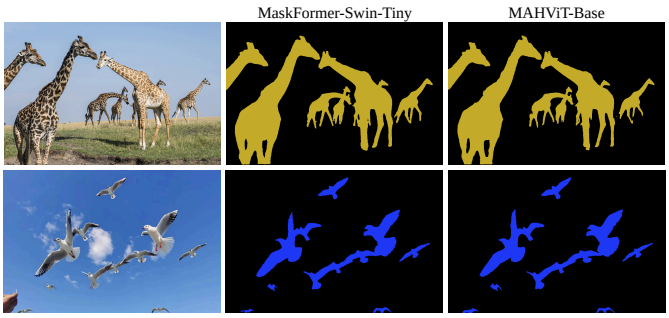


Fig. 6: Comparison with MaskFormer-Swin-Tiny [24] on OOD images using COCO pre-trained weights. Evaluated at 1024×1024 resolution, MAHViT-Base achieves comparable qualitative performance.

E. Comparison to SOTA Approaches

Table IIa shows the performance of our model with the recent approaches on three dataset - COCO-2017, ADE20k and Cityscapes validation dataset. We show that each of our variant of its respective parameter space class achieves either state-of-the art or near state-of-art result on all three benchmarks. We mentioned the multi-scale (MS) mIoU metric wherever available, to make a fair comparison. We kept the same resolution as the training resolution during our evaluation phase and kept mask confidence score to 0.5. We could not mention the FPS result for all the tabulated methods due to the lack of literature on a common dataset and hardware setup, making comparison unfair.

Fig. 5: Qualitative results of MAHViT-Advanced on the ADE20K, COCO-2017, and Cityscapes validation sets. The mask threshold is set to 0.5 with the input resolution of 1024.

IV. Conclusion

We presented MAHViT, a self-supervised hierarchical vision transformer aimed at learning general-purpose multi-scale dense representations. MAHViT mitigates key challenges in hierarchical transformers, including scale-wise misalignment, optimization instability, and feature redundancy, through a dual Layer Normalization data-flow that improves training stability while maintaining feature diversity. To support efficient reuse of hierarchical features in dense prediction pipelines, we introduced a Feature Diversity Sampling (FDS) neck that selectively filters redundant tokens and provides compact, informative representations to downstream decoders. Pre-trained using a BYOL-based objective with resolution-adaptive alignment, MAHViT learns representations that remain consistent across input resolutions. Experimental results on COCO-2017, ADE20K, and Cityscapes demonstrate that MAHViT serves as a strong and reusable dense backbone for semantic segmentation, achieving competitive accuracy–efficiency trade-offs. These findings suggest that self-supervised hierarchical transformers are a promising direction toward more broadly applicable dense feature extractors.

References

[1] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[2] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[4] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[5] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[6] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa *et al.*, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[7] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.

[9] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[10] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[11] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[12] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 619–15 629.

[13] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[14] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9653–9663.

[15] S. Xie, H. Zhang, J. Guo, X. Tan, J. Bian, H. H. Awadalla, A. Menezes, T. Qin, and R. Yan, "Residual: Transformer with dual residual connections," *arXiv preprint arXiv:2304.14802*, 2023.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[18] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[19] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[20] H. Yu, Y. Cho, B. Kang, S. Moon, K. Kong, and S.-J. Kang, "Embedding-free transformer with inference spatial reduction for efficient semantic segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 92–110.

[21] Y. Fu, M. Lou, and Y. Yu, "Segman: Omni-scale context modeling with state space models and local attention for semantic segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 19 077–19 087.

[22] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "Segnext: Rethinking convolutional attention design for semantic segmentation," *Advances in neural information processing systems*, vol. 35, pp. 1140–1156, 2022.

[23] Z. Ni, X. Chen, Y. Zhai, Y. Tang, and Y. Wang, "Context-guided spatial feature reconstruction for efficient semantic segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 239–255.

[24] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in neural information processing systems*, vol. 34, pp. 17 864–17 875, 2021.

[25] H. Yan, M. Wu, and C. Zhang, "Multi-scale representations by varying window attention for semantic segmentation," *arXiv preprint arXiv:2404.16573*, 2024.

[26] S. Chen and J. Lu, "Dual teacher dual student with pixel refinement for weakly supervised semantic segmentation," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–6.

[27] Z. Xie, Y. Lin, Z. Yao, Z. Zhang, Q. Dai, Y. Cao, and H. Hu, "Self-supervised learning with swin transformers," *arXiv preprint arXiv:2105.04553*, 2021.

[28] S. Yun, H. Lee, J. Kim, and J. Shin, "Patch-level representation learning for self-supervised vision transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8354–8363.

[29] X. Chu, Z. Tian, B. Zhang, X. Wang, and C. Shen, "Conditional positional encodings for vision transformers," *arXiv preprint arXiv:2102.10882*, 2021.

[30] X. Zhang, Y. Tian, W. Huang, Q. Ye, Q. Dai, L. Xie, and Q. Tian, "Hivit: Hierarchical vision transformer meets masked image modeling," *arXiv preprint arXiv:2205.14949*, 2022.

[31] J. Liu, X. Huang, J. Zheng, Y. Liu, and H. Li, "Mixmae: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6252–6261.