Vision-Perceptual Transformer Network for Semantic Scene Understanding

Mohamad Alansari[1], Hamad AlRemeithi[1,2], Bilal Hassan[1,3], Sara Alansari[1], Jorge Dias[1,3], Majid Khonji[1,3], Naoufel Werghi[1,3,4] and Sajid Javed[1,3]

[1]*Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, U.A.E.*

[2]*Research and Technology Development Department, Tawauzn Technology & Innovation, Abu Dhabi, U.A.E.*

[3]*Center for Autonomous Robotic Systems, Khalifa University, Abu Dhabi, U.A.E.*

[4]*Center for Cyber-Physical Systems, Khalifa University, Abu Dhabi, U.A.E.*

Keywords: Attention Mechanisms, Computational Resources, Pyramid Vision Transformers, Scene Understanding, Semantic Segmentation.

Abstract: Semantic segmentation, essential in computer vision, involves labeling each image pixel with its semantic class. Transformer-based models, recognized for their exceptional performance, have been pivotal in advancing this field. Our contribution, the Vision-Perceptual Transformer Network (VPTN), ingeniously combines transformer encoders with a feature pyramid-based decoder to deliver precise segmentation maps with minimal computational burden. VPTN's transformative power lies in its integration of the pyramiding technique, enhancing multi-scale variations handling. In direct comparisons with Vision Transformer-based networks and variants, VPTN consistently excels. On average, it achieves 4.2%, 3.41%, and 6.24% higher mean Intersection over Union (mIoU) compared to Dense Prediction (DPT), Data-efficient image Transformer (DeiT), and Swin Transformer networks, while demanding only 15.63%, 3.18%, and 10.05% of their Giga Floating-Point Operations (GFLOPs). Our validation spans five diverse datasets, including Cityscapes, BDD100K, Mapillary Vistas, CamVid, and ADE20K. VPTN secures the position of state-of-the-art (SOTA) on BDD100K and CamVid and consistently outperforms existing deep learning models on other datasets, boasting mIoU scores of 82.6%, 67.29%, 61.2%, 86.3%, and 55.3%, respectively. Impressively, it does so with an average computational complexity just 11.44% of SOTA models. VPTN represents a significant advancement in semantic segmentation, balancing efficiency and performance. It shows promising potential, especially for autonomous driving and natural setting computer vision applications.

1 INTRODUCTION

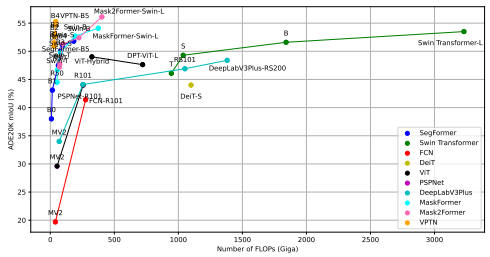
Semantic segmentation, the process of classifying each pixel of an image into distinct semantic categories, is a fundamental task in computer vision with critical applications in autonomous driving, medical imaging, and robotics. Its significance in autonomous driving lies in its capacity for scene analysis, object detection, and behavior prediction (Siam and et al., 2018).

This task, however, faces several hurdles. Complex real-world scenes, with their clutter and partial occlusions, pose a challenge in differentiating objects (Feng and et al., 2020). Variable lighting and noisy images add further complexity. Another issue is the processing of high-resolution images essential for detailed scene understanding, particularly in real-time video streams for autonomous driving (Papadeas and et al., 2021). Moreover, acquiring accurate annotations for model training is labor-intensive and prone to inconsistencies, risking overfitting and limited model generalization (Feng and et al., 2020).

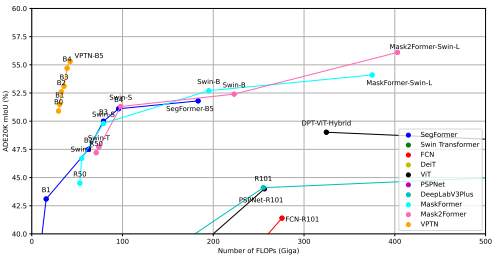
In addressing these challenges, deep learning, especially Convolutional Neural Networks (CNNs), has been widely adopted for its effective feature extraction (Li and et al., 2019). The recent focus has shifted towards transformer-based models, recognized for enhanced performance in complex segmentation scenarios (Chitta and et al., 2022). These models, such as SETR (SEgmentation TRansformers) (Zheng et al., 2021), incorporate Vision Transformers (ViTs) for their capability to handle diverse image scales. However, ViTs typically require high computational resources and may struggle with tasks beyond image

Alansari, M., AlRemeithi, H., Hassan, B., Alansari, S., Dias, J., Khonji, M., Werghi, N. and Javed, S.
Vision-Perceptual Transformer Network for Semantic Scene Understanding.
DOI: 10.5220/0012313800003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 3: VISAPP*, pages 325-332
ISBN: 978-989-758-679-8; ISSN: 2184-4321



(a)



(b)

Figure 1: Performance versus model computational complexity on ADE20K. (a) presents all models, while (b) provides a zoomed-in view where the x-axis is confined to the range of 0-500 Giga Floating-Point Operations (GFLOPs), and the y-axis is focused on the range of 40-60 mean Intersection over Union (mIoU) in (%). In terms of computational complexity, the Vision-Perceptual Transformer Network (VPTN) achieves a new state-of-the-art (SOTA), exhibiting a mean difference of 5.91% mIoU compared to the top-performing Mask2Former, while demanding only 8.67% of its number of GFLOPs.

classification due to their fixed-size input tokenization (Zheng et al., 2021).

To address these limitations, we introduce the Vision-Perceptual Transformer Network (VPTN), a novel approach that combines transformer encoders with a feature pyramid to efficiently produce segmentation maps, as shown in Figure 2, particularly useful in autonomous driving contexts. VPTN navigates multi-scale segmentation challenges effectively, promising improvements in both accuracy and computational efficiency for various computer vision tasks.

2 RELATED WORKS

The field of autonomous driving has witnessed significant advancements, largely driven by the development of deep learning models for complex computer vision tasks (Geiger et al., 2012). The core challenge lies in accurately detecting, recognizing, and segmenting objects, which are critical for navigational decision-making. Deep learning models address-

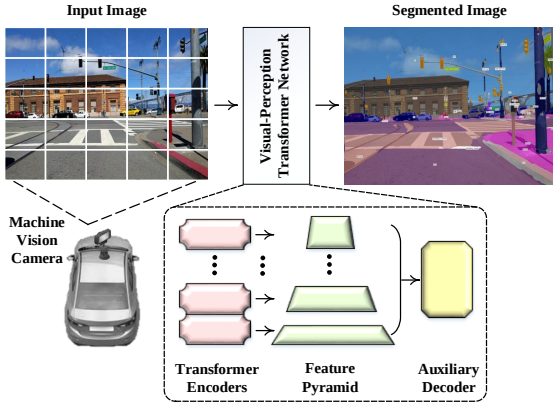


Figure 2: A high-level overview of the proposed VPTN-based semantic scene understanding framework for autonomous driving. The proposed model combines the strengths of transformer encoders and multi-scale feature pyramid to further improve semantic segmentation performance.

ing these challenges can be categorized into Object Detection Models, Semantic Segmentation Models, Transformer-Based Models, and Multi-Task Learning Frameworks, each contributing uniquely to the progress in autonomous driving technology.

2.1 Object Detection Models

Models such as YOLOv4 (Wang et al., 2021) excel in quick object detection and localization. Their rapid detection capabilities, essential for real-time applications, make them suitable for dynamic environments like driving. However, they may struggle with small or obscured objects and may not fully contextualize the scene.

2.2 Semantic Segmentation Models

Models like Mask R-CNN (He and et al., 2020) focus on pixel-wise classification to interpret complex driving scenes. Known for their accuracy in demarcating object boundaries, these models require substantial computational power and extensive training data, especially in diverse environmental conditions.

2.3 Transformer-Based Models

Transformer-based models, exemplified by SETR (Zheng et al., 2021), have improved object detection and semantic segmentation tasks. Their proficiency in global context capture and dependency understanding aids in complex scene analysis. Despite their effectiveness, the high computational demand of these models poses challenges for real-time implementations in autonomous driving.

2.4 Multi-Task Learning Frameworks

Frameworks like UPerNet (Xiao and et al., 2018) demonstrate competence in processing varied image annotations, crucial for detecting and identifying diverse objects in driving contexts. While offering versatility, the complexity of balancing multiple learning tasks requires careful tuning to avoid bias towards a particular task.

2.5 Contribution

Our study introduces the VPTN, an innovative framework tailored for semantic segmentation in autonomous driving. VPTN integrates transformer-based global contextual understanding with multi-task framework precision, excelling in multi-scale variation handling and high-resolution map generation.

The key novelties of our approach are:

- A novel efficient hierarchical pyramiding transformer-based architecture adept at semantic segmentation challenges, particularly in multi-scale scenarios
- Comprehensive ablation studies to fine-tune segmentation heads and loss functions, highlighting the pyramiding technique's effectiveness against conventional state-of-the-art (SOTA) transformer models
- SOTA performance on BDD100k and CamVid, with competitive results on Cityscapes, Mapillary Vistas, and ADE20K. Our models distinguish themselves in balancing computational efficiency and high performance, as illustrated in Figure 1

3 PROPOSED METHODOLOGY

In this research, we propose a deep learning architecture called VPTN for semantic scene understanding. The VPTN is a hybrid architecture consisting of two main components: a transformer-based backbone network and a decoder network, as shown in Figure 3.

3.1 VPTN Backbone

The backbone network is designed in a *"progressive shrinking"* fashion to reduce the number of parameters in the network by gradually reducing the feature amps by a factor of 1/4 as the network progresses through its stages (Wang and et al., 2022b). This architecture comprises four cascading stages, each incorporating a stack of transformer blocks (Dosovitskiy and et al., 2020) to maintain long-range dependencies between image regions. The resulting feature pyramid, with a four-level feature pyramid ($F1$, $F2$, $F3$, and $F4$) with a stride of 4, 8, 16, and 32 pixels relative to the input image.

Inputs to transformer blocks are a blend of outputs from previous blocks and lower-level feature maps, facilitating complex pattern recognition and multi-scale integration. Contrasting traditional approaches (Dosovitskiy and et al., 2020), VPTN generates token embeddings from the input image via convolutional layers, enhancing spatial information extraction for images of any size. The network also features a channel attention module, selectively emphasizing or diminishing features across different channels, focusing on relevant features while minimizing noise.

3.2 VPTN Decoder

The VPTN decoder consists of three main components: a pyramid pooling module, a feature fusion module, and a segmentation head.

3.2.1 Pyramid Pooling Module

In VPTN decoder, the pyramid pooling module is used on the last depth of the VPTN backbone network to extract features from multiple scales. The output of the PPM module is a set of four pooled feature maps, each corresponding to a different scale of the input image, as shown in Figure 3. These pooled feature maps are then concatenated and fed into the subsequent feature pyramid depths in the network for further processing.

3.2.2 Feature Fusion Module

The proposed VPTN model employs a pyramid structure with multiple levels of feature maps. Each level captures features at a different scale, allowing the model to capture both local and global information. For instance, low-level features may help to segment fine details such as edges, while high-level features may help to capture the overall shape of objects. In the feature fusion module of the proposed VPTN model, the features from different levels of the pyramid are concatenated to preserve the complementary information and discard any redundant information.

3.2.3 Segmentation Head

The segmentation head, consisting of convolutional, fully connected, and upsampling layers, followed by a softmax function, processes high-resolution feature

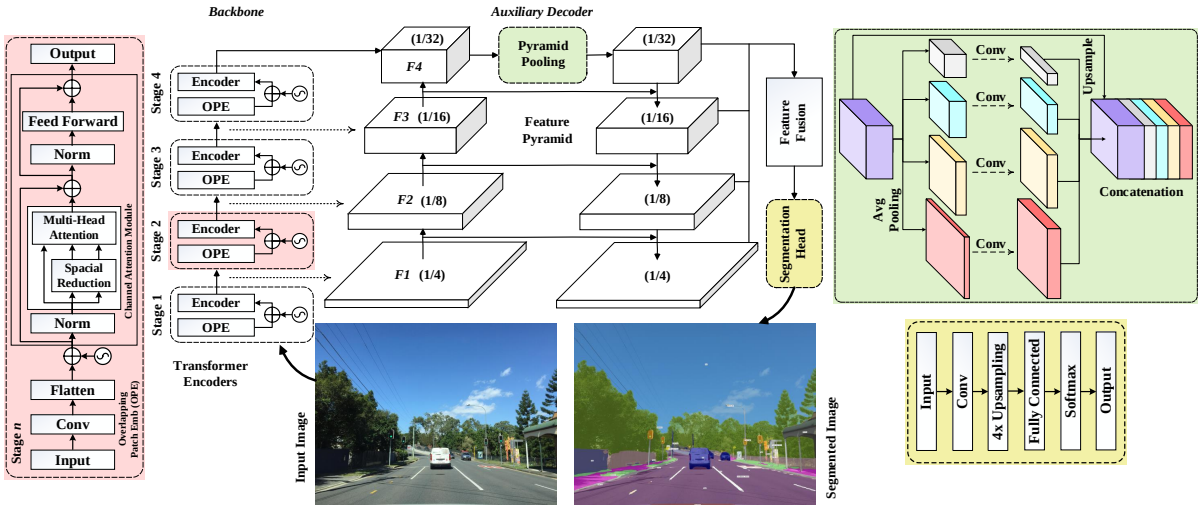


Figure 3: Schematic of the proposed VPTN model, consisting of a pyramidal backbone network (left side) and feature pyramid-based decoder network (right side). The backbone generate hierarchical feature maps at different levels of abstraction, forming a four-level feature pyramid (F1, F2, F3, and F4) with specific stride values. The decoder use the pyramid pooling module on the last depth of the VPTN backbone to extract features from multiple scales and concatenate the pooled feature maps and pass them to subsequent feature pyramid depths for further processing.

maps from the lateral network. It generates a probability distribution across classes, ultimately assigning the highest probability class to each pixel for the final segmentation map.

4 EXPERIMENTAL SETUP

4.1 Datasets Description

In our rigorous assessment of VPTN's performance, we leverage five publicly available datasets, encompassing urban road and driving environments, as well as generic semantic segmentation scenarios. Each dataset (Cordts and et al., 2016; Yu et al., 2020; Neuhold and et al., 2017; Brostow and et al., 2009; Zhou et al., 2017) includes separate training, validation, and test sets. However, the test set annotations for Cityscapes, BDD100K, and Mapillary Vistas are not publicly available, so we use their validation sets for evaluation. A summary of the datasets and their training-validation splits is detailed in Table 1. This selection ensures a comprehensive assessment of VPTN across varied environments and conditions.

4.1.1 Cityscapes

The Cityscapes (Cordts and et al., 2016) dataset focuses on urban scenes with 5,000 images annotated in 19 categories, reflecting complex city elements and diverse weather and lighting conditions.

4.1.2 BDD100K

The BDD100K (Yu et al., 2020) dataset contains 100,000 driving videos and 10,000 images of urban scenes, aligned with Cityscapes in terms of object classes, but offering broader environmental diversity.

4.1.3 Mapillary Vistas

The Mapillary Vistas (Neuhold and et al., 2017) dataset provides over 25,000 high-resolution images with annotations in 66 classes, sourced from various devices and covering a wide geographic range.

4.1.4 CamVid

CamVid (Brostow and et al., 2009) dataset is a smaller dataset with 701 road scene images, offering detailed annotations in 32 categories, captured through a car-mounted camera in Cambridge, UK.

4.1.5 ADE20K

ADE20K (Zhou et al., 2017) dataset is a broad dataset for generic semantic segmentation, featuring over 20,000 images annotated in 150 fine-grained and 1,000 common categories, covering both indoor and outdoor scenes.

4.2 Evaluation Metric

Our study uses the mean Intersection over Union (mIoU) as the primary metric for assessing VPTN's

Table 1: Dataset Training and Evaluation Sets Statistics.

	Dataset	For Training (#images)	For Evaluation (#images)
Urban road/driving	Cityscapes	Train Set (2,975)	Val Set (500)
	BDD100K	Train Set (7,000)	Val Set (1,000)
	Mapillary Vistas	Train Set (18,000)	Val Set (2000)
	CamVid	Train + Val Sets (469)	Test Set (232)
Generic	ADE20K	Train Set (20,210)	Val Set (2,000)

performance. mIoU is a standard measure in segmentation tasks, accounting for both true positive and false positive predictions. The mIoU score is calculated as:

$$mIoU = \frac{1}{C} \times \sum_{i=1}^{C} (IoU_i), \quad (1)$$

where C is the number of classes, and IoU_i is the IoU value computed for the i^{th} class.

4.3 Training Protocol

All experiments are conducted using Python 3.10.0 (64-bit) and the and Pytorch 1.8.1 on a workstation with an Nvidia GeForce RTX 3080 GPU for VPTN model development. Training was conducted over 100 epochs with a batch size of 8, testing five loss functions to optimize performance. AdamW optimizer with a 0.001 starting learning rate and 0.01 decay, accompanied by a warmup scheduler (power of 0.9, 10 epochs, 0.1 ratio), facilitated the training. Model evaluations were performed every epoch with flip augmentation enabled.

5 EXPERIMENTATION RESULTS

This section explains various ablation studies related to the VPTN model. Following that, we assess the performance of the proposed model using both subjective and objective measures.

5.1 Ablation Study

5.1.1 Optimizing Decoder and Hyperparameters Selection

Extensive ablation studies were conducted to fine-tune the VPTN model, particularly focusing on optimizing the decoder and hyperparameters, presented in Table 2. The Cityscapes dataset was primarily used for these studies due to its manageable size and relevance to urban scenes. Images were resized to 512×512 and the VPTN-B3 network was selected to expedite the training process. Our experiments included exploring various decoder options and loss functions. UPerNet emerged as the most effective decoder, and the weighted cross-entropy loss function was identified as the best fit for our model. We also investigated the use of Trainable Structure Tensors (TST) (Hassan and et al., 2021) as a pre-processing technique but found it less effective compared to using the original dataset.

Table 2: Identifying the Optimal Configuration for VPTN-B3 on the Cityscapes Dataset. Values in bold indicate the selected best-performing configuration for our proposed model.

Ablation Experiments		mIoU (%)
Decoder	**UPerNet**	**75.8**
	Lawin	72.7
	FPN	69.4
	FCN	67.6
	FaPN	67.2
	SFNet	65.8
	SegFormer	61.6
Loss Function	**Weighted Cross entropy**	**75.8**
	Cross entropy	73.4
	Ohem Cross entropy	73.3
	Dice	72.4
	Lovasz	55.7
Data Pre-processing	**Baseline**	**75.8**
	Fused Original and TST	68.5
	TST	55.4

5.1.2 Effect of Pyramiding Technique Integration

Our study evaluated the VPTN model's integration of the pyramiding technique against ViT variants on the ADE20K dataset, focusing on mIoU and Giga Floating-Point Operations (GFLOPs) for computational efficiency.

The comparative analysis, detailed in Table 3, highlights VPTN's superior performance. Our VPTN-B4 and VPTN-B5 models outperformed the latest SOTA model, UPerNet-Swin-L, by 2.19% and 3.25% in mIoU, respectively, while requiring significantly less computational power (1.20% and 1.31% of GFLOPs). Across all VPTN models, the average performance was 4.30% lower than UPerNet-Swin-L, but they used only 1.08% of its GFLOPs, showcasing VPTN's efficiency in balancing performance with lower computational demand.

5.2 Comparison with State-of-the-Art

In evaluating the VPTN model, we conducted a thorough comparative analysis with other leading methods across five distinct datasets, focusing on mIoU scores and computational efficiency (GFLOPs). The GFLOPs metrics were computed using specific input scales for each dataset: $\{1024, (512, 1024), 1024, (720, 960), 512\}$ for Cityscapes, BDD100K, Mapillary Vistas, CamVid,

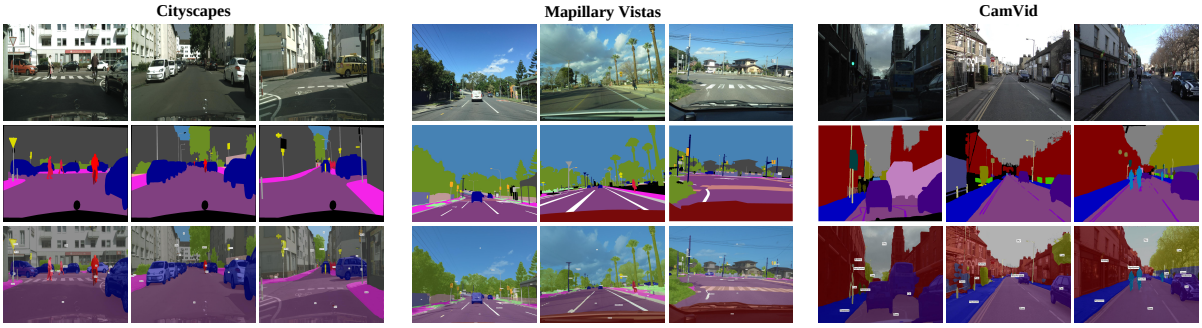


Figure 4: Qualitative segmentation results of the proposed VPTN model using three different datasets (Cordts and et al., 2016; Neuhold and et al., 2017; Brostow and et al., 2009). The top row shows the original frames from each dataset. The middle row shows the ground truth labels, and the last row displays the predicted results.

Table 3: Performance comparison of the proposed VPTN with ViT variants methods using ADE20K validation dataset.

Method	Encoder	Parameters (M)	GFLOPs	mIoU (%)
DPT	ViT-Hybrid	109.16	324.57	49.02
DPT	ViT-L	341.56	721.03	47.63
UPerNet	DeiT-S	52	1099	44.0
UPerNet	Swin-T	60	945	46.1
UPerNet	Swin-S	81	1038	49.3
UPerNet	Swin-B	121	1841	51.6
UPerNet	Swin-L	234	3230	53.5
VPTN (ours)	PVT v2-B0	**6.13**	**29.53**	50.9
VPTN (ours)	PVT v2-B1	16.04	30.82	51.5
VPTN (ours)	PVT v2-B2	27.39	32.51	52.6
VPTN (ours)	PVT v2-B3	47.28	35.55	53.1
VPTN (ours)	PVT v2-B4	64.59	38.90	54.7
VPTN (ours)	PVT v2-B5	83.99	42.25	**55.3**

and ADE20K, respectively.

5.2.1 Cityscapes

On the Cityscapes validation dataset, our VPTN models displayed impressive mIoU scores while requiring significantly fewer GFLOPs compared to the current SOTA model, OneFormer. Specifically, VPTN models had an average of 5.9 percentage points difference in mIoU while using only an average of 17.86% of OneFormer's computational power. Visual segmentations from Cityscapes are shown in Figure 4.

Table 4: Performance comparison of the proposed VPTN with other SOTA methods using the Cityscapes validation dataset.

Method	Encoder	GFLOPs	mIoU (%)
SegFormer (Xie and et al., 2021)	MiT-B0	125.5	78.1
DeepLabV3Plus (Chen et al., 2018)	D-Xception-71	1444.6	79.6
SegFormer (Xie and et al., 2021)	MiT-B1	243.7	80.0
SegFormer (Xie and et al., 2021)	MiT-B2	717.1	81.0
HRNetV2 (Wang and et al., 2021)	HRNetV2-W48	1206.3	81.6
SETR (Zheng et al., 2021)	ViT-L	-	82.2
CMX (SegFormer-B4) (Liu and et al., 2022)	MiT-B4	-	82.6
Panoptic-DeepLab (Cheng and et al., 2020)	SWideRNet	1095.0	83.1
SegFormer (Xie and et al., 2021)	MiT-B5	1460.4	84.0
Mask2Former (Cheng and et al., 2022)	Swin-L	868	84.3
OneFormer (Jain and et al., 2022)	ConvNeXt-XL	775	84.6
VPTN (ours)	PVT v2-B0	118.10	77.4
VPTN (ours)	PVT v2-B1	123.19	78.3
VPTN (ours)	PVT v2-B2	129.89	79.6
VPTN (ours)	PVT v2-B3	141.96	80.1
VPTN (ours)	PVT v2-B4	153.08	81.5
VPTN (ours)	PVT v2-B5	164.20	82.6

5.2.2 BDD100K

Results on the BDD100K dataset's validation set detailed in Table 5 show that VPTN outperformed the previous SOTA, ConvNeXt-B, by 0.04% in mIoU, while utilizing just 4.58% of its GFLOPs. The performance metrics for different models were sourced from the official BDD100K GitHub repository[1] and related research papers.

Table 5: Performance comparison of the proposed VPTN with other SOTA methods using the BDD100K validation dataset.

Method	Encoder	GFLOPs	mIoU (%)
DeiT	DeiT-S	-	61.52
Semantic FPN	ResNet50-FPN	-	61.53
PointRend	ResNet50-FPN	-	61.80
PSANet	ResNet50-Dilate8	205.98	61.99
ViT	ViT-B	-	62.11
DeepLabV3	ResNet101-Dilate8	-	63.23
DPT	ViT-B	-	63.53
HRNet	HRNet48	-	63.93
DeepLabV3Plus	ResNet101-Dilate8	2032.3	64.49
Swin Transformer	Swin-B	1188	65.98
ConvNeXt	ConvNeXt-B	1828	67.26
VPTN (ours)	PVT v2-B0	59.05	60.89
VPTN (ours)	PVT v2-B1	61.61	61.72
VPTN (ours)	PVT v2-B2	64.97	63.13
VPTN (ours)	PVT v2-B3	71.73	64.46
VPTN (ours)	PVT v2-B4	77.69	66.32
VPTN (ours)	PVT v2-B5	83.65	67.29

5.2.3 Mapillary Vistas

The VPTN achieved a maximum mIoU of 61.2% on Mapillary Vistas, compared to 82.6% on Cityscapes. Despite the increased complexity, the VPTN network yields competitive results and surpasses some purely CNN-based architectures. When compared with the current SOTA model, Mask2Former, we found that our models exhibited an average mIoU difference of 10.98 percentage-points, all while utilizing only an average of 15.95% of the computational power (mea-

[1]https://github.com/SysCV/bdd100k-models/tree/main/sem_seg

sured in GFLOPs). Sample segmentations are illustrated in Figure 4.

Table 6: Performance comparison of the proposed VPTN with other SOTA methods using the Mapillary Vistas validation dataset.

Method	Encoder	GFLOPs	mIoU (%)
SegBlocksRN50 (Verelst and et al., 2023)	EfficientNetLite1	254.4	41.7
NiSeNet (Nag and et al., 2019)	ResNet101	-	48.32
DeepLabV3Plus (Chen et al., 2018)	ResNet50	51.4	49.4
HMSANet (Hua and et al., 2022)	ResNet50	-	52.2
MaskFormer (Li and et al., 2022)	ResNet50	181	55.4
Mask2Former (Cheng and et al., 2022)	ResNet50	226	59.0
HMSANet (Hua and et al., 2022)	HRNet	61.1	
Mask2Former (Cheng and et al., 2022)	Swin-L	868	64.7
VPTN (ours)	PVT v2-B0	118.10	55.9
VPTN (ours)	PVT v2-B1	123.19	56.3
VPTN (ours)	PVT v2-B2	129.89	57.6
VPTN (ours)	PVT v2-B3	141.96	58.8
VPTN (ours)	PVT v2-B4	153.08	60.4
VPTN (ours)	PVT v2-B5	164.20	61.2

5.2.4 CamVid

On the CamVid testing dataset, VPTN-B4 and VPTN-B5 models achieved SOTA results with mIoU scores of 84.1% and 86.3%, respectively, as exhibited in Table 7. These results show a marked improvement over the previous SOTA model, SIW, with an mIoU score of 83.7%. To provide a more tangible perspective of the VPTN's capabilities, Figure 4 visually depicts its segmentation prowess through a selection of randomly chosen images from the CamVid dataset. These images further substantiate the VPTN's capacity to excel in complex real-world scenarios.

Table 7: Performance comparison of the proposed VPTN with other SOTA methods using the CamVid testing dataset.

Method	Encoder	GFLOPs	mIoU (%)
VideoGCRF (Chandra and et al., 2018)	ResNet101	-	75.2
ETC-Mobile - MobileNetV2+ALL (Liu and et al., 2020)	MobileNetV2	-	78.2
DeepLabV3Plus + SDCNetAug	WideResNet38	-	81.7
RTFormer-Base (Wang and et al., 2022a)	-	537.0	82.5
SIW (Yin and et al., 2022)	SegFormer-B5	-	83.7
VPTN (ours)	PVT v2-B0	77.92	80.8
VPTN (ours)	PVT v2-B1	81.84	81.3
VPTN (ours)	PVT v2-B2	85.59	81.7
VPTN (ours)	PVT v2-B3	94.35	83.4
VPTN (ours)	PVT v2-B4	102.06	84.1
VPTN (ours)	PVT v2-B5	109.77	86.3

5.2.5 ADE20K

Finally, we present a comparative analysis of generic semantic segmentation performance on the ADE20K validation dataset, as detailed in Table 8. VPTN's consistently exhibited competitive performance in terms of mIoU scores while demonstrating a notable reduction in computational resource requirements compared to SOTA models, specifically, compared to the SOTA, OneFormer, we observe an average mIoU difference of 9.87 percentage-points lower, coupled with an average computational power (GFLOPs) usage of only 2.55%.

Table 8: Performance comparison of the proposed VPTN with other SOTA methods using the ADE20K validation dataset. The models utilized varying crop sizes to report mIoU and GFLOPs, with annotations [1] representing Crop Size = 640, [2] indicating Crop Size = 896, and [3] denoting Crop Size = 1280.

Method	Encoder	GFLOPs	mIoU (%)
SegFormer (Xie and et al., 2021)	MiT-B0	8.4	38.0
SegFormer (Xie and et al., 2021)	MiT-B1	15.9	43.1
SegFormer (Xie and et al., 2021)	MiT-B2	62.4	47.5
Mask2Former (Cheng and et al., 2022)	ResNet-50	71	49.2
Mask2Former (Cheng and et al., 2022)	Swin-T	74	49.6
SegFormer (Xie and et al., 2021)	MiT-B3	79.0	50.0
Mask2Former (Cheng and et al., 2022)	ResNet-101	90	50.1
SegFormer (Xie and et al., 2021)	MiT-B4	95.7	51.1
SegFormer (Xie and et al., 2021)	MiT-B5	183.3	51.8
Mask2Former (Cheng and et al., 2022)	Swin-B[1]	223	55.1
Mask2Former (Cheng and et al., 2022)	Swin-L[1]	403	57.3
OneFormer (Jain and et al., 2022)	DiNAT-L	678[2]	58.1
OneFormer (Jain and et al., 2022)	DiNAT-L	1369[3]	58.2
VPTN (ours)	PVT v2-B0	29.53	50.9
VPTN (ours)	PVT v2-B1	30.82	51.5
VPTN (ours)	PVT v2-B2	32.51	52.6
VPTN (ours)	PVT v2-B3	35.55	53.1
VPTN (ours)	PVT v2-B4	38.90	54.7
VPTN (ours)	PVT v2-B5	42.25	55.3

6 CONCLUSIONS

Our study introduces the VPTN, a novel architecture for semantic segmentation that optimally balances computational efficiency with advanced performance. The model's superiority is demonstrated through comprehensive evaluations across various datasets and rigorous ablation studies.

Performance benchmarks against SOTA transformer-based models reveal VPTN's proficiency. It sets a new SOTA on the BDD100K dataset with a 0.04% mIoU increase, using only 4.58% of the previous SOTA's computational resources. On CamVid, VPTN improves mIoU by 3.11%. In Cityscapes and Mapillary Vistas, it closely trails the current SOTA by merely 1.36 percentage-points, requiring just 16.9% of their computational power. Even in the ADE20K dataset, VPTN competes closely with the current SOTA, OneFormer, with a 9.87% mIoU difference while using a mere 2.55% of its GFLOPs. VPTN's ability to strike a compelling balance between high accuracy and low computational demand makes it a prime solution for diverse semantic segmentation challenges, particularly in urban and generic scene understanding.

ACKNOWLEDGEMENTS

This publication acknowledges the support provided by the Khalifa University of Science and Technology under Faculty Start-Up grants FSU-2022-003 Award No. 8474000401.

REFERENCES

Brostow, G. J. and et al. (2009). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30.

Chandra, S. and et al. (2018). Deep spatio-temporal random fields for efficient video segmentation. In *Proceedings of the IEEE Conference on CVPR*, pages 8915–8924.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818.

Cheng, B. and et al. (2020). Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on CVPR*, pages 12475–12485.

Cheng, B. and et al. (2022). Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 1290–1299.

Chitta, K. and et al. (2022). Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Transactions on PAMI*.

Cordts, M. and et al. (2016). The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE Computer Society Conference (CSC) on Computer Vision and Pattern Recognition (CVPR)*, 2016-December.

Dosovitskiy, A. and et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feng, D. and et al. (2020). Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE T-ITS*, 22(3):1341–1360.

Geiger, A. et al. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. *Proceedings of the IEEE CSC on CVPR*.

Hassan, T. and et al. (2021). Trainable structure tensors for autonomous baggage threat detection under extreme occlusion. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12627 LNCS.

He, K. and et al. (2020). Mask r-cnn. *IEEE Transactions on PAMI*, 42.

Hua, Z. and et al. (2022). Dual attention based multi-scale feature fusion network for indoor rgbd semantic segmentation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3639–3644. IEEE.

Jain, J. and et al. (2022). OneFormer: One Transformer to Rule Universal Image Segmentation. *arXiv*.

Li, P. and et al. (2019). Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 7644–7652.

Li, Z. and et al. (2022). Maskformer with improved encoder-decoder module for semantic segmentation of fine-resolution remote sensing images. In *2022 IEEE ICIP*, pages 1971–1975. IEEE.

Liu, H. and et al. (2022). Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems (T-ITS)*.

Liu, Y. and et al. (2020). Efficient semantic video segmentation with per-frame inference. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 352–368. Springer.

Nag, S. and et al. (2019). What's there in the dark. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2996–3000. IEEE.

Neuhold, G. and et al. (2017). The mapillary vistas dataset for semantic understanding of street scenes. *Proceedings of the IEEE ICCV*, 2017-October.

Papadeas, I. and et al. (2021). Real-time semantic image segmentation with deep learning for autonomous driving: A survey. *Applied Sciences*, 11(19):8802.

Siam, M. and et al. (2018). A comparative study of real-time semantic segmentation for autonomous driving. In *Proceedings of the IEEE conference on CVPR workshops*, pages 587–597.

Verelst, T. and et al. (2023). Segblocks: Block-based dynamic resolution networks for real-time segmentation. *IEEE Transactions on PAMI*, 45(2):2400–2411.

Wang, C. Y. et al. (2021). Scaled-yolov4: Scaling cross stage partial network. *Proceedings of the IEEE CSC on CVPR*.

Wang, J. and et al. (2021). Deep high-resolution representation learning for visual recognition. *IEEE Transactions on PAMI*, 43(10):3349–3364.

Wang, J. and et al. (2022a). Rtformer: Efficient design for real-time semantic segmentation with transformer. *arXiv preprint arXiv:2210.07124*.

Wang, W. and et al. (2022b). Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8.

Xiao, T. and et al. (2018). Unified perceptual parsing for scene understanding. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11209 LNCS.

Xie, E. and et al. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090.

Yin, W. and et al. (2022). The devil is in the labels: Semantic segmentation from sentences. *Conference on CVPR*.

Yu, F. et al. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on CVPR*, pages 2636–2645.

Zheng, S. et al. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on CVPR*, pages 6881–6890.

Zhou, B. et al. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on CVPR*.