HSPFormer: Hierarchical Spatial Perception Transformer for Semantic Segmentation

Siyu Chen, *Student Member, IEEE*, Ting Han, *Graduate Student Member, IEEE*, Changshe Zhang, Jinhe Su, Ruisheng Wang, *Senior Member, IEEE*, Yiping Chen, *Senior Member, IEEE*, Zongyue Wang, and Guorong Cai, *Senior Member, IEEE*

***Abstract*— Semantic perception in driving scenarios plays a crucial role in intelligent transportation systems. However, existing Transformer-based semantic segmentation methods often do not fully exploit their potential in understanding driving scene dynamically. These methods typically lack spatial reasoning, failing to effectively correlate image pixels with their spatial positions, leading to attention drift. To address this issue, we propose a novel architecture, the Hierarchical Spatial Perception Transformer (HSPFormer), which integrates monocular depth estimation and semantic segmentation into a unified framework for the first time. We introduce the Spatial Depth Perception Auxiliary Network (SDPNet), a framework for multiscale feature extraction and multilayer depth map prediction to establish hierarchical spatial coherence. Additionally, we design the Hierarchical Pyramid Transformer Network (HPTNet), which uses depth estimation as learnable position embeddings to form spatially correlated semantic representations and generate global contextual information. Experiments on benchmark datasets such as KITTI-360, Cityscapes, and NYU Depth V2, demonstrate that HSPFormer outperforms several state-of-the-art networks, and achieves promising performance with 66.82% top-1 mIoU on KITTI-360, 83.8% mIoU on Cityscapes, and 57.7% mIoU on NYU Depth V2, respectively. The code will be made publicly available at https://github.com/SY-Ch/HSPFormer.**

***Index Terms*— Semantic segmentation, pyramid transformer, position embedding, multi modilaty.**

Received 1 February 2024; revised 13 August 2024 and 3 November 2024; accepted 29 December 2024. This work was supported in part by the Natural Science Foundation of Xiamen, China under Grant 3502Z202373036; in part by the National Natural Science Foundation of China under Grant 42371457, Grant 41971424, Grant 42301468, Grant 42371343, and Grant 61902330; in part by the Key Project of Natural Science Foundation of Fujian Province, China, under Grant 2022J02045; and in part by the Natural Science Foundation of Fujian Province, China, under Grant 2022J01337, Grant 2022J01819, Grant 2023J01801, Grant 2023J01799, Grant 2022J05157, and Grant 2022J011394. The Associate Editor for this article was D. F. Wolf. *(Siyu Chen and Ting Han are co-first authors.) (Corresponding author: Jinhe Su.)*

Siyu Chen, Jinhe Su, Zongyue Wang, and Guorong Cai are with the School of Computer Engineering, Jimei University, Xiamen 361021, China (e-mail: chensy@ieee.org; sujh@jmu.edu.cn; wangzongyue@jmu.edu.cn; guorongcai.jmu@gmail.com).

Ting Han and Yiping Chen are with the School of Geospatial Engineering and Science, Sun Yat-sen University, Zhuhai 519082, China (e-mail: ting.devin.han@gmail.com; chenyp79@mail.sysu.edu.cn).

Changshe Zhang is with the School of Ocean Information Engineering, Jimei University, Xiamen 361021, China (e-mail: xduzcs@163.com).

Ruisheng Wang is with the Schulich School of Engineering, Department of Geomatics Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada (e-mail: ruiswang@ucalgary.ca).

Digital Object Identifier 10.1109/TITS.2025.3525542

I. Introduction

SEMANTIC segmentation is a crucial step of scene perception and understanding in Advanced Driver Assistance System (ADAS), and it assigns distinct categories to individual road image pixels. This process has led to the development of insightful algorithms in the fields of computer vision and intelligent transportation [1], [2], [3], [4], [5], [6], [7]. Most previous methods rely simply on color and texture information to distinguish different semantic object classes without considering the impact of spatial information on target attention [8], [9], [10], [11], [12]. In complex environments, models may struggle to focus precisely on target objects, a phenomenon we refer to as "attention shift" (Fig. 1(a)). When the model's attention spreads beyond the target area, the lack of spatial cues often leads to over-segmentation or under-segmentation errors (Fig. 1(c)). To address this issue, we propose the DepthEmbed module, which incorporates depth information as a spatial prior into the attention mechanism to enhance the model's spatial alignment capabilities. DepthEmbed first processes the depth map through a Convolutional Neural Network (CNN) to suppress noise and improve feature stability. It then fuses depth features pixel-by-pixel with RGB image features to create a multimodal representation, which serves as input to the attention mechanism, accurately establishing spatial relationships between target and background areas. Additionally, DepthEmbed introduces depth features as a bias in the self-attention mechanism, allowing the model to dynamically adjust attention weights and align precisely with target contours, thus preventing the attention from spreading to non-target areas like the sky or road (Fig. 1(b)). By extracting depth features across multiple scales using a multi-scale convolutional structure, DepthEmbed ensures spatial consistency at various scales, enabling the model to focus more accurately on target regions within complex backgrounds, improving boundary precision (Fig. 1(d)).

Recent years, to effectively distinguish the relationships between objects, Transformer improves the inherent locality of convolutional neural networks (CNNs) and form attention perception from a global perspective [13], [14], [15], [16]. In traditional single-model strategy (Fig. 2(a)), position embedding (relative [13], absolute [17], and learnable position embedding [18] etc.) provides location dependence for pixels, but it does not consider the spatial correlation between image

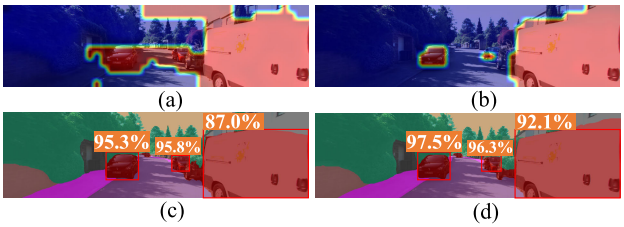


Fig. 1. Incorporating depth position embedding into HSPFormer (b, d) has corrected the issue of attention shift in previous methods (a, c), leading to a significant improvement in semantic segmentation performance, as highlighted in red with corresponding IoU.

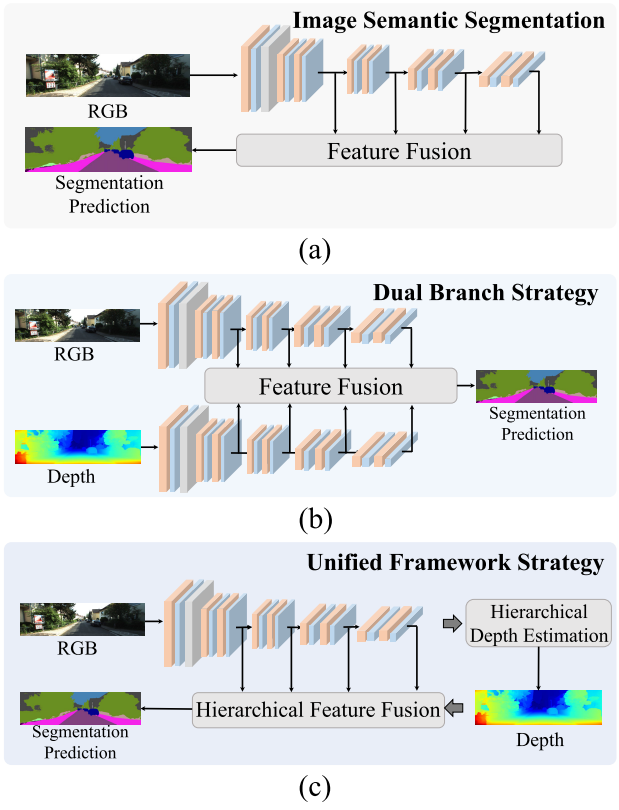


Fig. 2. Illustration of three different architectures for semantic segmentation. (*a*) Image semantic segmentation, (*b*) Multi-modality semantic segmentation in dual branch strategy, and (*c*) Our unified framework strategy to incorporate monocular depth estimation into semantic segmentation.

pixels and the real world. For instance, adjacent pixels in the image may have a large distance difference in the real scene. To incorporate spatial perception into the feature learning and extraction, many studies have focused on the strategy of multi-modality data fusion [19], [20], [21], [22], [23], [24], such as 3D point clouds and depth / disparity images as prior references for RGB images, as shown in Fig. 2(b). However, these methods are not suitable for autonomous driving tasks due to significant computational overheads and the need for real-time 2D-3D alignment. To this end, we first evaluated the trade-off between accuracy and speed in traffic scene perception using RGB images and depth maps. As shown in Fig. 2(b), we designed the dual-branch structure Hierarchical Spatial Perception Transformer network named HSPFormer-DBS, which uses spatial information as explicit position embedding to establish a connection between pixels and the real-world scenario. HSPFormer-DBS improved performance by over 8% while introducing an additional 48M parameters.

Therefore, we proposed a unified framework strategy HSPFormer-UFS to integrate monocular depth estimation and semantic segmentation into a unified framework, as shown in Fig. 2(c). HSPFormer-UFS comprises two main components: a Spatial Depth Perception Auxiliary Network (SDPNet) and a Hierarchical Pyramid Transformer Network (HPTNet). The primary contribution of the proposed method is that it provides depth information with hierarchical consistency for features at different layers. More importantly, semantic features and spatial features have corresponding relationships, and spatial features at different levels also have corresponding relationships. Additionally, using depth features predicted by depth estimation as learnable positional embedding guides the feature learning and semantic segmentation, effectively utilizing spatial observation information in visual perception. By incorporating spatial feature biases into visual features and modeling the correlation between pixels and real scenes, the network effectively correct the generation and focus of global attention. As shown in Fig. 1(b, d), our proposed HSPFormer focuses attention more accurately on target objects, and the segmentation results demonstrate a significant improvement in accuracy due to this focused attention. It significantly outperforms the CMX [21] and CMNeXt [19] models, with similar model efficiency (see Fig. 3). The main contributions of this paper are as follows:

1. The proposed HSPFormer-DBS utilizes depth information as position embedding to ensure accurate position relationship in real-world driving scenarios.
2. We designed HSPFormer-UFS to integrates monocular depth estimation and semantic segmentation, emphasizing positional relationships while contributing to the trade-off between performance and efficiency.
3. HSPFormer excels in real traffic scenarios semantic segmentation, and its robust scene perception capabilities contribute to enhancing the safety of autonomous driving. Our method demonstrates remarkable performance in KITTI-360 with a top mIoU of 67.32%, and outperforms the previous state-of-the-art method a +2.23% improvement.

II. Related Work

A. Semantic Segmentation

Convolutional Neural Networks (CNNs) have made remarkable advancements in the domain of semantic segmentation [25], [26], [27], [28]. CNNs have adopted various strategies [29], [30] to expand their receptive field. Despite these enhancements, the inherent local nature of their receptive field still limits their ability to understand global contextual relationships in high-resolution images. The emergence of Transformer has greatly improved global feature space learning. SETR [15] and Swin Transformer [13] have demonstrated the powerful feature extraction capability of Transformer in semantic segmentation, making outstanding contributions to advancing research. They accomplish this with self-attention that establishes long-range connections among features. Subsequently, there are many methods to introduce the pyramid

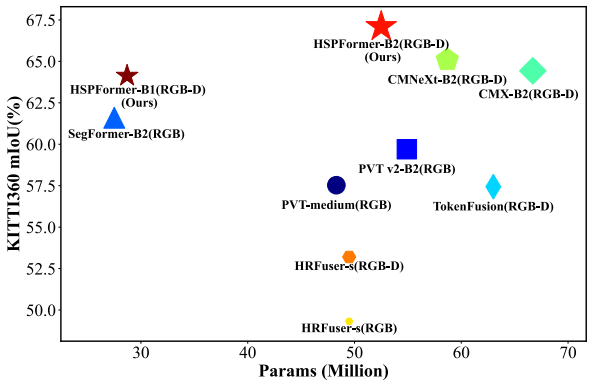


Fig. 3. Performance vs. model efficiency on KITTI-360. HSPFormer-B2, denoted by red star, achieves a new state-of-the-art 66.82% mIoU while maintaining competitive efficiency to other state-of-the-art methods.

structure in CNNs to the design of Transformer backbones [18], [31], [32], [33], [34], [35]. After that, some improvements are made to combine the advantages of CNN and Transformer to obtain stronger feature representations [36], [37], [38], [39], [40], [41]. However, current research predominantly explores the stacking and fusion of Transformer blocks and CNN blocks, without delving into the nuanced fusion strategies between CNN and Transformer within these blocks

B. Multi-Modality Architecture

RGB images tend to interpret color and texture information, but may encounter difficulties in scenarios with similar textures or lack of textural clues. The introduction of depth maps enhances a model's ability to interpret scenes by providing depth and spatial information. Currently, many studies have demonstrated that multi-modality data fusion is able to significantly improve the accuracy of semantic segmentation [42], [43], [44], [45]. Especially with the fusion of RGB-D data, the addition of depth information can markedly enhance the segmentation capabilities of models. Subsequent methods explore more sophisticated architectures and fusion strategies, such as using dual-branch structures [23], [46], [47] or designing various fusion modules [20], [48]. Additionally, some approaches combined monocular depth estimation with semantic segmentation [49], [50] for mutual optimization and constraint, providing new insights for our study. Existing multimodal fusion methods typically treat RGB and depth maps as two separate components, which does not fully take into account the structural characteristics of Transformers. Adding more branches also leads to an increase in the model's parameter count. Therefore, it is necessary to design a completely new fusion approach.

C. Position Embedding

Position embedding is crucial for understanding the token positions in a sequence in vision Transformer. The vanilla approaches (fixed-length absolute position embedding [14], [17], [51]), limited the model's ability to handle diverse data. Subsequent methods incorporating relative position embedding [13], [52], [53] to enhance the performance of Transformer in image understanding. Furthermore, many research introduced learnable implicit position embedding using CNNs to understand local relationships among tokens or pixels [31], [37], [54], [55]. In multimodal tasks such as RGB-D, RGB and depth images are commonly handled as separate components for feature extraction. This approach does not relate pixels to real-world situations or take into account the spatial connections between pixels.

III. Method

Our goal is to accommodate monocular depth estimation to the semantic segmentation simulating multi-modality operation, and then utilize depth information to generate hierarchy-coherent representations to address the concern of attention shift in semantic segmentation. To this end, we develop Hierarchical Spatial Perception Transformer (HSPFormer) network as shown in Fig. 4 (Sec. III-A), consisting of a Spatial Depth Perception Auxiliary network (SDPNet) (Sec. III-B) and a Hierarchical Pyramid Transformer network (HPTNet) (Sec. III-C).

A. HSPFormer: Hierarchical Spatial Perception Transformer

Unlike typical semantic segmentation methods (Fig. 2(a)) that solely extract image features for prediction, in HSPFormer, we aim to enable the image itself to learn the underlying spatial relationships of scene objects and formalize them as depth position embedding at different scales to assist in refining pixel category predictions. Meanwhile, HSPFormer is an efficient segmentation framework without hand-crafted and computationally demanding modules. HSPFormer consists of two main components: (1) a spatial depth perception auxiliary network based on monocular depth estimation to generate multi-scale spatial information; and (2) corresponding hierarchical pyramid Transformer network to fuse images features and spatial information coarse-to-fine to produce multi-level semantic features, as depicted in Fig. 4.

For an image of size $H \times W \times 3$, it serves as input for the SDPNet to obtain multi-level features at resolution of $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$ relative to the raw image. Differing from the dual-branch strategy (Fig. 2(b)) of existing RGB-D semantic segmentation methods, we simultaneously extract features from the image and predict the depth features (Fig. 2). Therefore, the depth estimation head is added within the auxiliary network to predict depth information at the corresponding resolution, realizing spatial perception. During the training phase, the network is supervised on the predicted depth features.

Subsequently, in the HPTNet, we design parallel Transformer encoders to further learn the image features. In the Transformer encoder, the learned depth features are used as position embedding to correct the attention shift caused by the traditional methods being disconnected from the actual scene in global attention. The features generated by the Transformer encoders maintain resolution consistency. Then, we pass multi-level features to the decoder to restore them to the original resolution at $H \times W \times N_{cls}$ for prediction after feature fusion, where N_{cls} is the number of classes.

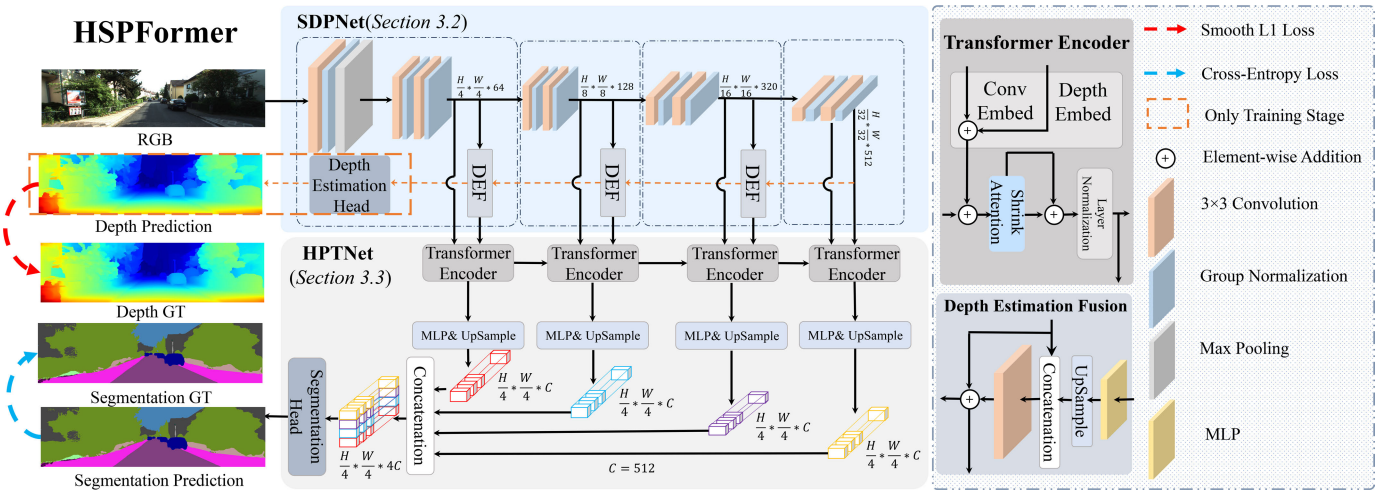


Fig. 4. The proposed HSPFormer framework consists of two main modules: A spatial depth perception auxiliary network (SDPNet) to extract coarse features and estimate multi-level depth maps; and a hierarchical pyramid Transformer network (HPTNet) to fuse images and spatial features coarse-to-fine and predict the segmentation mask. Note that supervision for depth estimation is conducted during training stage.

B. SDPNet: Spatial Depth Perception Auxiliary Network

The goal of this network is, given an input image, to generate both multi-level features and corresponding depth predictions. These features obtained by the feature extractor are enhanced during the process of transitioning resolution from high to low and granularity from coarse to fine, which promotes local understanding for semantic segmentation. More precisely, the four-layer network structure generates features F_i^R at four different scales, with size of $\frac{H}{2^i} \times \frac{W}{2^i} \times C_i$, where $i \in 1, 2, 3, 4$, and $C_i \in 64, 128, 256, 512$. Subsequently, the multi-scale features are sent to parallel Transformer encoders for global feature transformation and encoding.

1) Dual Branch Strategy: Considering that Transformer require position embedding to provide sequence relationships as priors for the image, traditional position embedding often rely on fixed-value sequences that lack relevance to the actual scene, such as sine/cosine and conditional position embedding, etc. This leads to attention shifts, resulting in incomplete object segmentation. Therefore, we aim to adopt a multi-modality strategy by treating the depth image as an additional input, thereby forming a dual branch strategy (DBS) as shown in Fig. 2(b). We employ depth features as learnable position embedding to assist in image feature extraction, establishing a correspondence between sequence relationships and actual scene, ultimately optimizing semantic segmentation. We use shared-weight CNN with the image extractor to generate depth features F_i^D at size of $\frac{H}{2^i} \times \frac{W}{2^i} \times C_i$. Subsequently, we perform element-wise summation with image features. However, this approach increases computational demands and reduces inference speed. Moreover, depth images typically generated from point cloud projection are sparse, and even though depth completion can be performed, it introduces additional noise.

2) Unified Framework Strategy: In contrast to the DBS, we exclusively utilize images as inputs to extract multi-level features while concurrently predicting the corresponding depth information, as shown in Fig.2(c). Specifically, we incorporate the depth estimation module at each layer. The depth estimation module consists of a MLP and a convolution layer with padding. Features x_{i+1} from $i+1$ layer are fused with features x_i from layer i after passing through the MLP. The fused features are then processed through convolution to obtain the depth prediction. The calculation process can be formalized as follows:

$$Depth_{out} = Conv(Cat(x_i, MLP(x_{i+1}))) + x_i \quad (1)$$

where Cat denotes the operation of concatenating features along channels. Finally, We adopt the smooth L1 loss [56] to supervise the predicted depth map. The discrepancy between the predicted $D_{pred}(i, j)$ and GT $D_{true}(i, j)$ at pixel location (i, j) is denoted by $\Delta D(i, j) = D_{\text{pred}}(i, j) - D_{\text{true}}(i, j)$. The smooth L1 loss is expressed as:

$$L_{smooth} = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} Smooth_{L1}(\Delta D(i, j)) \quad (2)$$

$$Smooth_{L1}(\Delta D) = \begin{cases} 0.5(\Delta D)^2 & \text{if } |\Delta D| < 1, \\ |\Delta D| - 0.5 & otherwise. \end{cases} \quad (3)$$

where h and w denote the height and width of the depth map. It is important to note that this supervision is only applied during training stage.

C. HPTNet: Hierarchical Pyramid Transformer Network

In the HPTNet, we design parallel Transformer encoders for global feature extraction. Subsequently, lightweight decoders are appended to process the obtained features to generate the final prediction. In contrast to the conventional approach of adding Transformer encoders at the end of CNNs, we believe that this approach fixes local context within the features and makes it challenging to effectively represent global contextual information. Therefore, we use parallel modules to gradually generate rich feature representations.

1) ConvEmbed: For each Transformer encoder, we employ a consistent structure. To model the local continuity information, unlike the patch embedding process in ViT, we utilize the ResNet [57] to extract pixel-level ConvEmbed generated from

SDPNet to tokenize image features, unifying a $H_i \times W_i \times C_i$ vector into $H_i W_i \times C_i$, where H_i, W_i, and C_i denote the dimension of features in i layer.

2) DepthEmbed: The DepthEmbed module functions to provide spatial prior information within the model, helping it more accurately capture spatial relationships in the image. For the input image and the estimated depth prediction, the DepthEmbed module first receives the depth map and preprocesses it through a CNN, performing noise suppression and smoothing to obtain more reliable depth features. The processed depth features are then pixel-wise added to the RGB features extracted by the ConvEmbed module, embedding the depth into the image feature sequence. This step produces a multimodal feature, which serves as the input F_i^{fuse} to the shrink attention block.

3) Shrink Attention: Subsequently, we transform F_i^{fuse} into Q, K, and V vectors, followed by linear spatial reduction operations on K and V, as described in PVT v2. Thus, the process of attention block is:

$$Attention = Softmax(\frac{Q(SRA(K))^T}{\sqrt{d}})SRA(V) \quad (4)$$

where SRA operation means average pooling spatial reduction. Then, we apply a FFN with an activation function and a normalization layer to process the features. The output vector has the same dimension as the input.

4) MLP Decoder: HPTNet incorporates a lightweight decoder consisting only of MLP layers, avoiding the computationally overheads in other methods. Firstly, multi-level features from the Transformer encoders are compressed to a uniform number C of channels. Then, they are upsampled at the size of $\frac{H}{4} \times \frac{W}{4} \times C$ individually. Thirdly, aggregation is performed by concatenation. Finally, an additional MLP layer is applied to predict the segmentation mask at the size of $\frac{H}{4} \times \frac{W}{4} \times N_{cls}$ from the fused features.

IV. Experiment

A. Datasets and Metric

1) KITTI-360 [58]: The large-scale street scene dataset was recorded in various suburbs of Karlsruhe in Germany, serving as an extension of the original KITTI dataset [60]. It offers rich sensory information and full annotations for dense semantic segmentation. The dataset includes 49,004 training images and 12,276 testing images with the resolution of $1,408 \times 376$. The semantic label definitions are consistent with Cityscapes [61], encompassing 19 classes for evaluation. Furthermore, the depth data is predicted using the binocular estimation algorithm IGEV [62].

2) NYU Depth V2 [59]: The NYU Depth V2 dataset comprises 1,449 RGB-D images of indoor scenes used for semantic segmentation, with 795 images for training and 645 images for testing. The image resolution is 640×480 pixels, and annotations are available for 40 different categories.

3) Cityscapes [61]: Cityscapes is an RGB-D dataset designed for urban street scene analysis during road-driving conditions. It comprises a collection of 5,000 images, categorized into training, validation, and testing (2,975 / 500 / 1,525) subsets. Each image is meticulously annotated with dense labels across 19 categories. The dataset encompasses a diverse range of urban environments, representing 50 unique cities, and with image shape of $2,048 \times 1,024$.

4) Metric: The mean Intersection over Union (mIoU) is used as the evaluation metric to validate the performance of semantic segmentation.

B. Implementation Details

Our model training was conducted on four NVIDIA 4090 GPUs, with a batch size of 4 per GPU. Initially, the network model was pre-trained on ImageNet-1K [64]. During training, we applied various data augmentation techniques to enhance the model's generalization capability, including random flipping, random scaling within the range of $[0.5, 2]$, random color jittering, and random Gaussian blur. We opted for the AdamW optimizer with a weight decay of 0.05. The initial learning rate was set to 6e-5, and a cosine annealing strategy with a warm-up phase was employed for learning rate scheduling. To simplify the training and evaluation process, we used a basic cross-entropy loss function for supervised segmentation prediction and a smooth L1 loss function for supervised depth map prediction. When evaluating the mIoU of PVT and HSPFormer, we did not employ a sliding window approach; instead, we predicted directly from the original images.

C. Backbone Selection

As shown in Tab. I, we assess the performance, parameters, and efficiency of different backbone networks. On one hand, the SegFormer with MiT-B2 as the backbone outperforms PVT v2-B2 due to its sliding window and MLP decoder. However, when removing the sliding window and MLP decoder, the mIoU of SegFormer decrease to 59.18%, which is lower than the 59.70% of PVT v2. On the other hand, PVT v2 has the same parameters and inference speed with SegFormer. Therefore, we select PVT v2 as the backbone network for the subsequent experiments.

D. Experimental Design

In our experiments, we conducted two comparative studies and six ablation studies. Tab. I presents the performance of our method on the KITTI360 and NYU DepthV2 datasets and demonstrates its advantages on the Cityscapes dataset. Tab. II details the experimental results for 16 common categories in the KITTI360 dataset, highlighting the significant improvements achieved by our model. Tab. III analyzes the impact of different modules and frameworks on the baseline, including parameters, mIoU, and accuracy. It is noteworthy that the DepthEmbed module, specifically designed for handling depth maps, was not tested with pure RGB inputs. Tab. IV compares our proposed DepthEmbed position embedding with other mainstream position embedding methods to validate the effectiveness our DepthEmbed. Tab. V shows the effects of integrating DepthEmbed into other Transformer architectures, illustrating its versatility. Tab. VI evaluates the accuracy of our

TABLE I

MAIN RESULTS ON KITTI-360 [58], NYU DEPTH V2 [59] AND CITYSCAPES [61]. THE BEST RESULTS ARE IN BOLD. THE PARAMETERS AND MIOUS ARE REPORTED FROM CMX [21] AND CMNEXT [19]

Method	Input	KITTI-360				NYU Depth V2		Cityscapes	
		Param(M)	Backbone	mIoU(%)	inference time(s)	Backbone	mIoU(%)	Backbone	mIoU(%)
PVT (2021) [18]	RGB	28.2	PVT-small	57.53	0.0182	PVT-large	44.6	PVT-large	78.6
SegFormer (2021) [31]	RGB	25.9	MiT-B2	61.37	0.0192	MiT-B4	52.0	MiT-B2	81.0
PVT v2 (2022) [54]	RGB	29.1	PVT v2-B2	59.70	0.0167	PVT v2-B4	51.6	PVT v2-B2	80.6
HSPFormer-UFS (Ours)	RGB	27.7	PVT v2-B1	63.73	0.0126	-	-	PVT v2-B2	82.1
HSPFormer-UFS (Ours)	RGB	52.5	PVT v2-B2	**66.82**	0.0215	PVT v2-B4	**57.0**	PVT v2-B4	**83.2**
TokenFusion (2022) [23]	RGB-Lidar	26.0	MiT-B2	54.55	0.0696	-	-	-	-
CMX (2023) [21]	RGB-Lidar	66.7	MiT-B2	64.31	0.3073	-	-	-	-
CMNeXt (2023) [19]	RGB-Lidar	58.7	MiT-B2	65.26	0.1106	-	-	-	-
TokenFusion (2022) [23]	RGB-Event	26.0	MiT-B2	55.97	0.0696	-	-	-	-
CMX (2023) [21]	RGB-Event	66.7	MiT-B2	64.03	0.3073	-	-	-	-
CMNeXt (2023) [19]	RGB-Event	58.7	MiT-B2	66.13	0.1106	-	-	-	-
SA-Gate (2020) [47]	RGB-Depth	-	-	-	-	ResNet-101	52.4	ResNet-101	81.7
PGDENet (2022) [63]	RGB-Depth	107.0	ResNet-34	56.34	0.0349	ResNet-101	53.5	-	-
TokenFusion (2022) [23]	RGB-Depth	26.0	MiT-B2	57.44	0.0696	MiT-B2	54.2	-	-
CMX (2023) [21]	RGB-Depth	66.7	MiT-B2	64.43	0.3073	MiT-B5	56.9	MiT-B4	82.6
CMNeXt (2023) [19]	RGB-Depth	58.7	MiT-B2	65.09	0.1106	MiT-B4	56.9	-	-
HSPFormer-DBS (Ours)	RGB-Depth	77.1	PVT v2-B2	**67.32**	0.0350	PVT v2-B4	**57.8**	PVT v2-B4	**83.8**

TABLE II

PERFORMANCE COMPARISON USING DIFFERENT BACKBONE MODELS FOR MULTIPLE CATEGORIES. THE BEST RESULTS ARE IN BOLD. (%)

Method	Input	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic sign	Vegetation	Terrain	Sky	Person	Car	Truck	Bus	Motorcycle	Bicycle
PVT v2 (2022)	RGB	95.2	82.0	87.3	64.0	49.2	38.6	47.4	88.3	74.0	93.0	50.0	92.6	70.9	91.3	51.0	42.4
SegFormer (2021)	RGB	95.3	81.8	87.2	65.5	48.6	36.2	46.9	88.9	75.4	92.9	60.9	92.9	72.8	93.4	54.6	44.0
CMX (2023)	RGB-Depth	95.8	84.1	88.4	66.9	50.8	41.4	46.8	89.4	76.0	93.5	62.7	92.0	66.2	88.8	39.7	41.5
CMNeXt (2023)	RGB-Depth	95.2	82.1	88.2	64.0	49.7	41.5	48.8	89.1	75.1	93.9	63.6	93.6	80.2	95.6	55.1	47.7
HSPFormer-UFS (Ours)	RGB	**96.7**	**84.9**	**88.9**	**67.0**	**53.4**	**43.2**	**50.6**	**89.8**	**77.1**	**94.1**	**64.8**	**94.3**	**80.7**	**96.7**	**55.3**	**50.1**

TABLE III

ABLATION STUDY OF THE PROPOSED MODULES ON KITTI-360 VAL SET. THE BEST RESULTS ARE IN BOLD. WE COMPARED THE PROPOSED METHOD WITH PVT V2 [54], WHICH SERVED AS THE BASELINE

#	Baseline	Input	SDPNet		HPTNet			Param(M)	mIoU(%)	Acc(%)
			DBS	UFS	ConvEmbed	DepthEmbed	MLP Decoder			
1	✓	RGB						29.1	59.7	65.0
2	✓	RGB					✓	25.3	61.4	69.0
3	✓	RGB			✓		✓	51.5	63.5	71.5
4	✓	RGB		✓	✓	✓	✓	52.5	**66.8**	**76.2**
5	✓	RGB-Depth	✓					55.0	60.5	66.7
6	✓	RGB-Depth	✓				✓	28.1	62.3	71.2
7	✓	RGB-Depth	✓		✓		✓	51.8	63.8	72.3
8	✓	RGB-Depth	✓			✓	✓	51.7	66.3	75.9
9	✓	RGB-Depth	✓		✓	✓	✓	77.1	**67.3**	**78.3**

TABLE IV

SEMANTIC SEGMENTATION PERFORMANCE OF DIFFERENT POSITION EMBEDDINGS (PES) ON KITTI-360 VAL SET. THE BEST RESULTS ARE IN BOLD. WE APPLY DIFFERENT PES TO THE PVT V2-B2 [54] TO ENSURE FAIRNESS IN THE COMPARISON

Position Embeddings	Input	mIoU(%)
Sine/Cosine PE [14], [17]	RGB-Depth	57.4
Learnable PE [18], [53]	RGB-Depth	59.1
Implicit PE [31], [54]	RGB-Depth	62.3
DepthEmbed in UFS (Ours)	RGB	64.2
DepthEmbed in DBS (Ours)	RGB-Depth	**66.3**

model in inferring depth maps on the NYU Depth V2 dataset and compares it with specialized depth estimation models. Tab. VII explores the impact of depth map completeness on model performance. Finally, Tab. VIII examines the influence of various CNN scales on ConvEmbed and DepthEmbed, ultimately selecting ResNet34 as the backbone for HSPFormer in the B2 specification, balancing computational load and mIoU.

E. Comprehensive Evaluation

1) Qualitative Results: We conduct qualitative experiments on KITTI-360 and NYU Depth V2 datasets. In this section, we provide a detailed showcase of the visualization results and compare with state-of-the-art methods to demonstrate the effectiveness of out approach.

2) KITTI-360: Fig. 5 shows the representative visual results on the KITTI-360 dataset. As seen, HSPFormer yields

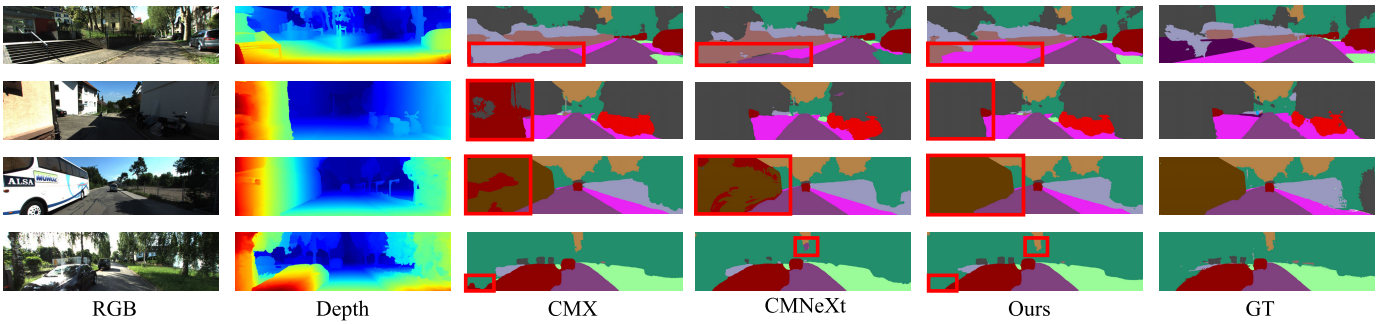


Fig. 5. Qualitative results on KITTI-360 [58]. The depth maps are predicted by SDPNet in our framework. Compared to CMX [21] and CMNeXt [19], Our HSPFormer predicts masks with substantially finer details near boundaries and reduces wide range errors, as highlighted in red boxes.

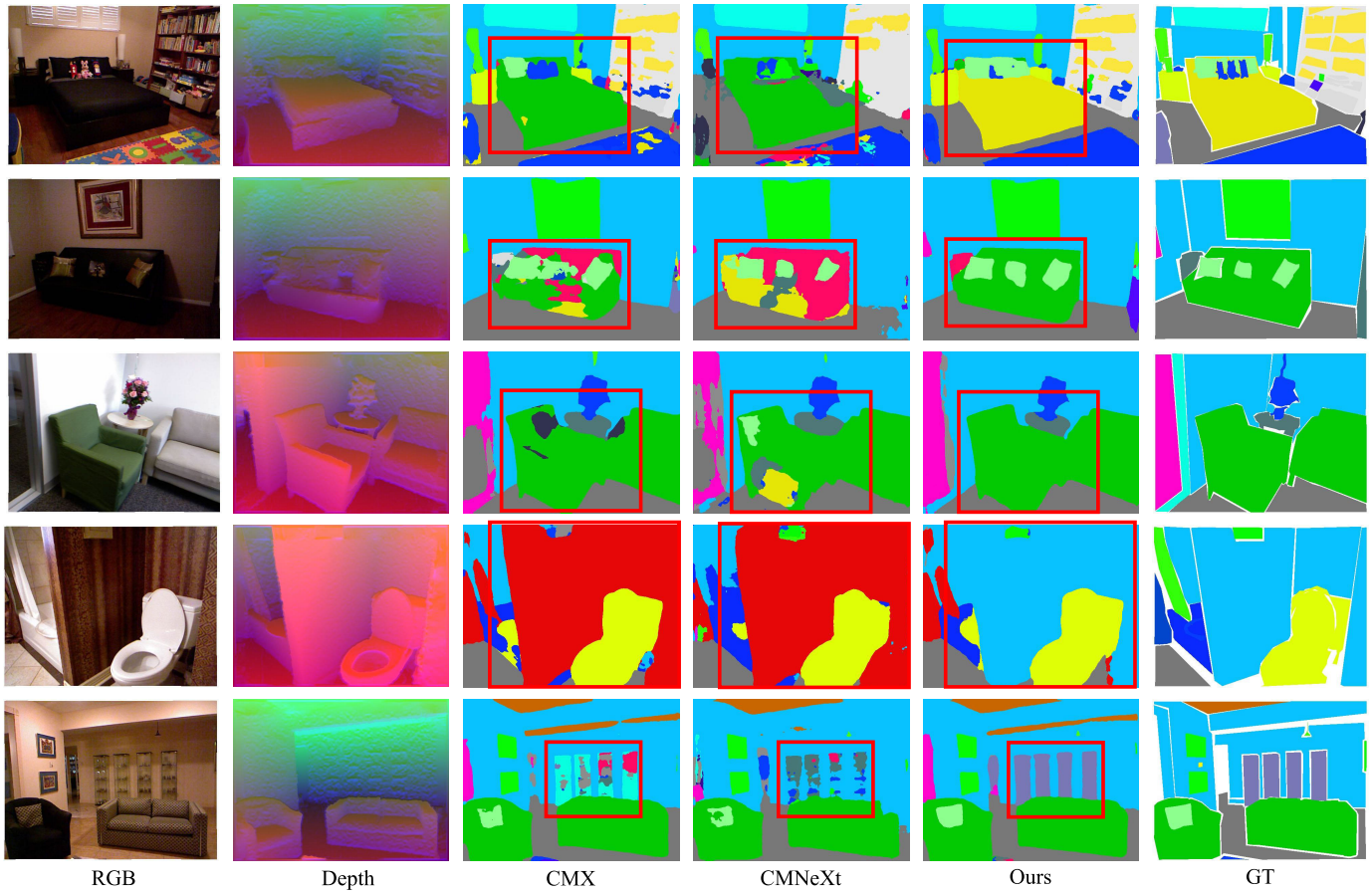


Fig. 6. Qualitative results on NYU Depth V2 [59]. Compared to CMX and CMNeXt, our HSPFormer predicts masks with correct labels and complete targets.

more precise segmentation results in comparison with some top-performing methods and shows strong robustness to various scenarios with occlusions, shadows, textures, densely arranged targets and small objects. Specifically, HSPFormer effectively distinguishes between sidewalks and roads with similar textures, identifies complete buildings and trucks, and recognizes small objects. This is due to the auxiliary spatial observations for semantic segmentation under our unified framework of depth estimation. As shown in the second column of Fig. 5, the depth maps predicted by UFS produce the position and distance information for the pixel positions in the real world. HSPFormer uses depth position embedding to provide powerful visual feature learning and representation ability to generate correct attention, significantly enhancing the completeness and accuracy of the boundaries of objects.

3) NYU Depth V2: Fig. 6 shows the qualitative results on NYU Depth V2, where HSPFormer provides complete targets and correct labels than CMX and CMNeXt. For example, the integrity of objects such as sofas, windows and walls is greatly improved while boundary segmentation is also improved.

4) Quantitative Results: Tab. I presents a comparison of our proposed HSPFormer with several state-of-the-art semantic segmentation models in terms of parameter (M), mIoU (%), and inference time (s) on the KITTI-360, NYU Depth V2, and Cityscapes datasets, respectively.

TABLE V
THE PROPOSED DEPTHEMBED IS INSTALLED FOR PERFORMANCE COMPARISON WITH OTHER TRANSFORMER BACKBONE NETWORKS ON KITTI-360 VAL SET. THE BEST RESULTS ARE IN BOLD. THE SF DENOTES THE SEGFORMER [31], AND ST DENOTES THE SWIN-TRANSFORMER [13]

Method	Backbone	DepthEmbed	Input	mIoU(%)
ST	Swin-S	-	RGB	57.6
ST	Swin-S	in UFS	RGB	58.3
ST	Swin-S	in DBS	RGB-Depth	**61.7**
SF	MiT B2	-	RGB	61.3
SF	MiT B2	in UFS	RGB	**66.0**
SF	MiT B2	in DBS	RGB-Depth	63.4
Ours	PVTv2 B2	-	RGB	59.7
Ours	PVTv2 B2	in UFS	RGB	**66.8**
Ours	PVTv2 B2	in DBS	RGB-Depth	66.3

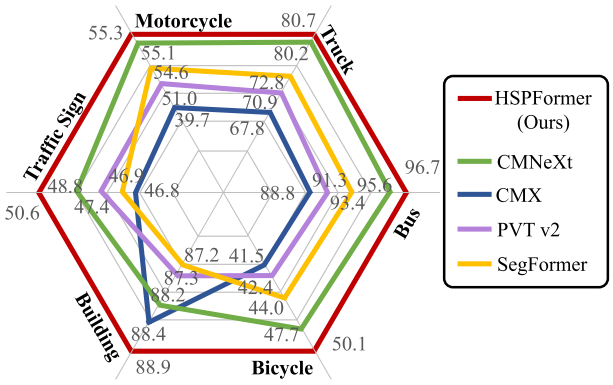


Fig. 7. Radar chart comparing the predictive performance of six prominent categories in the KITTI-360 dataset. HSPFormer-UFS is used for comparison with four top-leading methods.

5) KITTI-360: As shown in Tab. I, the proposed HSPFormer-UFS with single RGB input and PVT v2-B2 backbone achieves a 7.12% improvement in mIoU compared to the baseline, a 9.29% improvement over PVT, and a 5.45% improvement over SegFormer. When we use PVT v2-B1 as the backbone, our method also shows significant improvement over other approaches, with at least a 2% increase in mIoU while maintaining the fastest inference speed.

Tab. I reveals that multimodal inputs consistently outperform single RGB inputs. Notably, RGB-Depth performs better than both RGB-Event and RGB-LiDAR. We attribute this difference to the fact that RGB-LiDAR inputs do not directly use point cloud data but rather project it into a 2D space, generating a pseudo-depth map. These pseudo-depth maps have discrepancies compared to the original RGB images, and any misalignment (semantic or spatial information) can adversely affect the prediction accuracy. Additionally, Event images lose a significant amount of static scene information during capture, further affecting their performance. Our proposed model, HSPFormer-DBS, demonstrates the best performance with RGB-Depth input by integrating depth positional information encoded by DepthEmbed with RGB image features processed by ConvEmbed. This highlights the critical role of depth information in enhancing scene understanding accuracy. Although HSPFormer-UFS sacrifices some precision, it significantly reduces the number of parameters while maintaining excellent performance with pure RGB inputs. This trade-off between precision and parameter efficiency underscores the versatility of model in practical applications. Notably, the B1 version of HSPFormer-UFS achieves a 4.03% improvement in mIoU with fewer parameters than PVT v2-B2. Furthermore, the B2 version of HSPFormer-UFS surpasses CMX and CMNeXt in terms of efficiency while achieving higher mIoU accuracy. These results indicate that our HSPFormer-UFS successfully inherits and refines the design principles of HSPFormer-DBS, especially in the efficiency and performance trade-off.

Tab. II presents the comparison results of HSPFormer-UFS with four state-of-the-art models in terms of IoU across sixteen common categories in driving scenarios. The HSPFormer-UFS achieves the best performance across all categories. Notably, the IoU of RGB-Depth models surpasses that of PVT v2 and SegFormer in most categories, highlighting the significant benefits of depth information. Despite the HSPFormer only utilizing RGB images as input, it outperforms both single RGB input models (PVTv2 and SegFormer) and multi-modal RGB-Depth input models (CMX and CMNeXt) when guided by depth information. Particularly, the predicted spatial information is crucial for accurate segmentation of tiny targets such fences, poles, and traffic signs. This underscores the crucial role and effectiveness of depth information as a positional embedding. Fig. 7 illustrates the IoU comparison of the four algorithms in six more frequently encountered categories in driving scenes. The proposed method exhibits more robustness in traffic environment.

This demonstrates that (1) spatial observations can mimic human visual characteristics by distinguishing different objects based on their spatial positions. Depth information is able to effectively differentiate pixels with distance variations in space, even if these pixels are adjacent in the image. (2) Performing depth estimation from a single RGB image is able to provide spatial information bias for image semantic segmentation without relying on additional input information. Moreover, the information predicted by depth estimation maintains hierarchical consistency with the original image, simultaneously representing absolute and relative structural relationships.

6) NYU Depth V2: Our method maintains robustness across different scenarios. HSPFormer-UFS/DBS outperforms the top-leading CMX and CMNeXt by 0.9% / 0.1%, respectively. With RGB as the input, our method obtains (+5.2%) improvements over PVT v2 when using the same backbone. Compared with the CNN-based methods, our proposed framework demonstrates a significant improvement.

7) Cityscapes: Our methodology demonstrates superior performance across a variety of outdoor settings. Specifically, when utilizing RGB-D input, the HSPFormer-DBS B4 model markedly outperforms the CMX B4 model, registering a significant enhancement of +1.2%. This trend persists even when solely employing single RGB input. Additionally, the HSPFormer-UFS B2 model not only attains results comparable to the CMX B4 model but also exhibits a noteworthy advancement, with a substantial improvement of +1.5% over the

TABLE VI

COMPARATIVE QUANTITATIVE ASSESSMENT OF DEPTH MAPS: EVALUATING HSPFORMER-UFS AGAINST ESTABLISHED DEPTH ESTIMATION ALGORITHMS ON THE UYN DEPTHV2 DATASET. ↑DENOTES HIGHER THE BETTER AND ↓ DENOTES LOWER THE BETTER

Method	RMSE ↓	Abs Rel ↓	log10 ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
SC-DepthV2 [65]	0.532	0.138	0.059	0.820	0.956	-
MobileXNet [66]	0.533	0.418	-	0.797	0.951	0.987
LapDepth [67]	0.393	0.110	0.047	0.885	0.979	0.995
Depthformer [68]	**0.339**	**0.096**	**0.041**	**0.921**	**0.989**	**0.998**
Ours	0.415	0.165	0.076	0.863	0.962	0.984

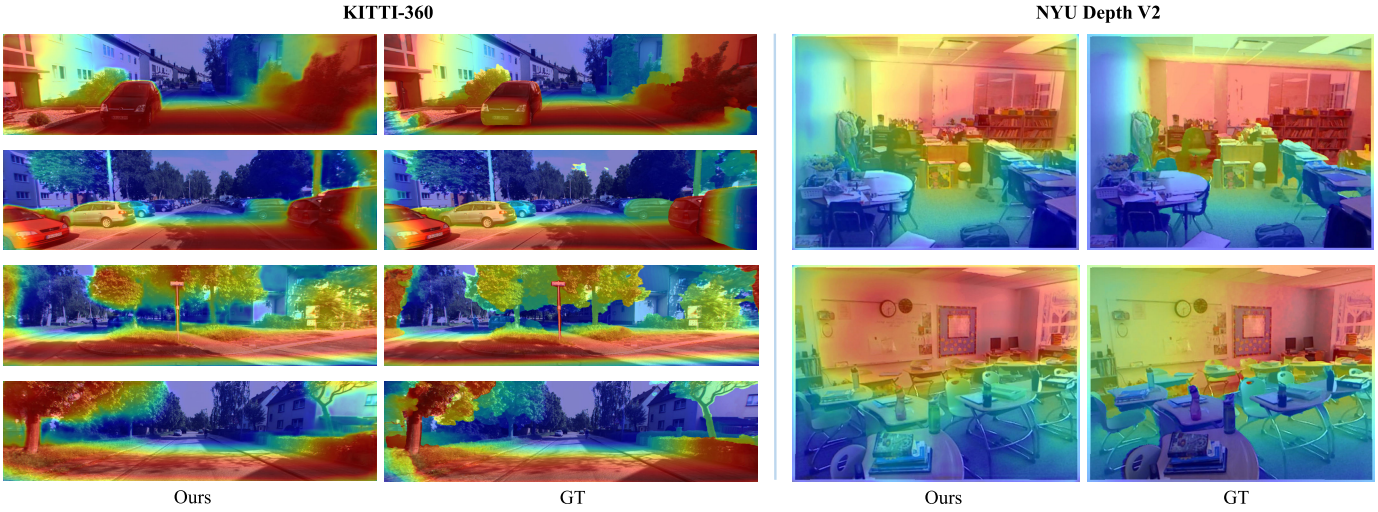


Fig. 8. Visualization results of depth estimation on the KITTI and NYU Depth V2 datasets.

PVTv2 B2 model, highlighting its effectiveness and robustness in complex visual processing tasks.

8) Discussion: HSPFormer establishes the new state-of-the-art both in the KITTI-360, NYU Depth V2 and Cityscapes. These results clearly demonstrate the efficacy of our semantic segmentation framework. While our UFS model falls slightly behind the DBS model (66.82% vs. 67.32%, 57.0% vs. 57.8%, 83.2% vs. 83.8%), the strategy of optimizing position embedding for depth estimation based on RGB input maintains performance and significantly increases model efficiency (52.5M vs. 77.1M).

TABLE VII

ANALYSIS OF THE PROPOSED MODEL WITH DEPTH MAPS OF VARYING QUALITY. THE BEST RESULTS ARE IN BOLD. WE CONDUCT EXPERIMENTS ON NYU DEPTH V2 DATASET

Method	Depth Image	mIoU(%)
HSPFormer-DBS	raw data	56.3
HSPFormer-DBS	repair data	**57.8**
HSPFormer-UFS	raw data	53.1
HSPFormer-UFS	repair data	**57.0**

F. Ablation Studies

To validate the necessity and effectiveness of the proposed modules, we conducted comprehensive ablation experiments in the KITTI-360 dataset to demonstrate their roles and performance within the network.

1) Model Analysis: Ablation experiments of our modules are reported in Tab. III. We see that all designs are able to improve the model in terms of performance and parameter number. Comparing #2 and #3 or #5 and #7, the model with ConvEmbed obtains better mIoU (61.4% vs. 63.5% or 62.3% vs. 63.8%) and better Acc (69.0% vs. 71.5% or 66.7% vs. 72.3%) than the baseline. Specifically, comparing #6 and # 8 or #7 and #9, the model with DepthEmbed shows improvements of approximately 4.0% or 3.5% in mIoU and 4.7% or 6.0% in Acc, respectively. DepthEmbed is effective because it is able to model the association between images and real-world scenes while ensuring the continuity of pixel relationships. As shown in Tab. III #4, our Unified Framework Strategy effectively reduces the parameters and computational overheads introduced by the network while maintaining performance stability to a certain extent (-0.5% mIoU and -2.1% Acc with -24.6M parameters).

As shown in Table III, configuration #4 demonstrates that our UFS framework effectively reduces network parameters and computational overhead while largely maintaining performance stability, with only a minimal drop of 0.5% in mIoU and 2.1% in accuracy alongside a parameter reduction of 24.6M. Compared to configurations #7 and #8, which have similar parameter counts, the UFS model increases parameters by less than 1M but achieves accuracy gains of 3% and 0.5%, respectively. These results highlight that integrating ConvEmbed and DepthEmbed within the UFS framework provides an advantageous balance between model efficiency and performance.

TABLE VIII
PERFORMANCE COMPARISON USING DIFFERENT BACKBONE MODELS TO OBTAIN CONVEMBED AND DEPTHEMBED. THE BEST RESULTS ARE IN BOLD

Module	CNNs	Param (M)	mIoU(%)
ConvEmbed	ResNet-18	39.0	62.3
	ResNet-34	51.5	63.5
	ResNet-50	55.5	63.9
	ResNet-101	74.4	**64.3**
DepthEmbed	ResNet-18	39.0	64.9
	ResNet-34	51.5	66.3
	ResNet-50	55.5	66.5
	ResNet-101	74.4	**67.6**

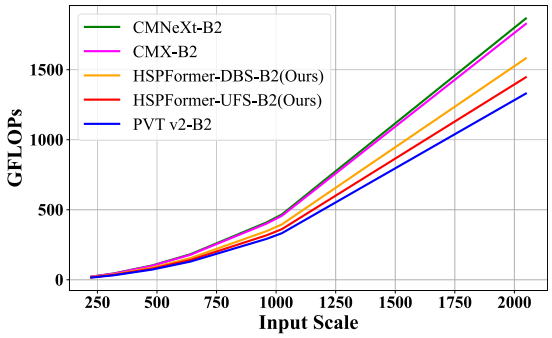


Fig. 9. Performance comparison of different input scales for our method and four top-leading methods.

2) DepthEmbed Importance: Compared to the previous position embedding strategies, DepthEmbed is a learnable position embedding that contains the relationship between pixels and the real world and the consistency of multilevel features. As shown in Tab. IV, our DepthEmbed achieves the best performance of 66.3% mIoU against previous position embeddings. Even when using UFS to predict the depth map as DepthEmbed, we achieve (+6.8% / +4.5% / +1.9%) mIoU improvement over previous methods. This is attributed to the predicted depth map maintaining pixel continuity, locality, and global correlation.

To validate the effectiveness of depth information, we present visual examples of depth estimation using our HSPFormer in Fig. 8. On the KITTI-360 dataset, our predicted depth map effectively distinguishes objects in the streetscape, such as vehicles and vegetation. More importantly, in the third and fourth rows of the Fig. 8, HSPFormer is able to clearly provide distance information for utility poles and trees. This offers powerful spatial prior information for subsequent semantic segmentation based on depth information. Furthermore, to verity the robustness, we provide depth estimation results for indoor scenes. Obviously, the predicted depth map provides spatial distance variations between adjacent pixels in the image, such as separating furniture with hierarchical relationships. Therefore, indoor spatial priors can enhance sensitivity to details and the integrity of target segmentation.

Tab. VI illustrates that the primary design objective of HSPFormer-UFS was to enhance transformers by embedding implicit positional information through the prediction of depth maps. While the core focus of this design is to refine positional embedding rather than depth map production, the quality of the resultant depth maps is commendable. Notably, HSPFormer-UFS demonstrates formidable performance, holding its ground against algorithms dedicated to monocular depth estimation. This outcome underscores a significant insight: the proficiency of HSPFormer-UFS in depth map prediction, though a secondary objective, is substantial and merits recognition, underscoring the model's comprehensive capability and adaptability.

As shown in the Tab. V, after incorporating the depth embedding, the performance of SegFormer improved from 61.3% to 66.0%, Swin Transformer improved from 57.6% to 61.7%, and our method improved from 57.6% to 66.8%, respectively. These results indicate that the multi-scale Transformer structures experience a significant boost when using the UFS framework, while the DBS framework shows a less pronounced improvement. This difference can be attributed to the fact that UFS integrates and optimizes both DepthEmbed and ConvEmbed modules, whereas DBS only incorporates DepthEmbed without ConvEmbed. Since Swin Transformer is a single-scale model, it does not benefit from the advantages provided by ConvEmbed, resulting in less impressive performance gains with the UFS framework. Since the acquired depth map usually carries a large number of missing values. Hence, we have considered the impact of supervising the network with different standards of depth images on the NYU Depth V2 dataset, and the results are presented in Tab. VII. The repair data, as opposed to the sparse raw data, maintain pixel continuity, facilitating more realistic position embedding when matched with RGB image features.

3) The CNNs of Both DepthEmbed and ConvEmbed Contribute to a Better Model: The multi-scale features extracted by the CNN provide multilevel ConvEmbed and DepthEmbed to the pyramid transformer embeddings. Therefore, the performance of CNNs is crucial to the quality of the features. From Tab. VIII, it can be observed that deeper CNNs contribute to better feature extraction performance of ConvEmbed and DepthEmbed.

4) Computational Overhead Analysis: As shown in Fig. 9, with increasing input scale, the growth rate of GFLOPs is as follows: CMNext-B2 > CMX-B2 > HSPFormer-DBS-B2 > HSPFormer-UFS-B2 > PVT v2-B2. These results prove that our proposed methods are able to achieve high-performance semantic segmentation while maintaining computational efficiency to a certain extent.

V. CONCLUSION

In this paper, we introduce HSPFormer, a novel unified framework that integrates monocular depth estimation as learnable position embeddings with Transformers for semantic segmentation. HSPFormer leverages the relationship between pixels and the real world to identify spatial differences among pixels. Through hierarchical feature extraction and depth estimation, consistent spatial features are established for structured scene analysis at corresponding resolutions. By exploiting spatial disparities between pixels, spatial observations can be used to generate more effective

pixel representations, thereby clearly distinguishing objects at different locations.

Comprehensive experiments demonstrate that HSPFormer outperforms many existing semantic segmentation models on three well-known datasets. On the KITTI-360 and Cityscapes traffic environment datasets, our method achieved top mIoU of 67.32% and 83.8%, respectively, exhibiting 2.23% and 1.2% improvements over the previous state-of-the-art approaches. On the NYU Depth V2, HSPFormer achieved a top mIoU of 57.8%. Furthermore, HSPFormer shows significant advantages in real traffic scenarios and the accuracy of various targets segmentation. Thus, the proposed is able to improve the safety and robustness of autonomous systems in ADAS.

Limitations: The model relies on depth data for supervision during training to accurately model spatial relationships between objects; however, this dependency limits its adaptability in scenarios lacking real depth maps. To overcome these limitations, future research could introduce unsupervised depth generation algorithms (e.g., Depth Anything) to provide training supervision, thereby reducing reliance on real depth maps and enhancing the model's generalization and applicability.

REFERENCES

[1] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[2] X. Wang et al., "SegGPT: Towards segmenting everything in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2023, pp. 1130–1140.

[3] M. Yuan et al., "Devil is in the queries: Advancing mask transformers for real-world medical image segmentation and out-of-distribution localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23879–23889.

[4] J. Chen, J. Lu, X. Zhu, and L. Zhang, "Generative semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 7111–7120.

[5] T.-D. Truong, N. Le, B. Raj, J. Cothren, and K. Luu, "FREDOM: Fairness domain adaptation approach to semantic scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19988–19997.

[6] Q. Hu et al., "Label-free liver tumor segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7422–7432.

[7] T. Han et al., "Epurate-net: Efficient progressive uncertainty refinement analysis for traffic environment urban road detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 6617–6632, Jul. 2024.

[8] J. Liu et al., "PolyFormer: Referring image segmentation as sequential polygon generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18653–18663.

[9] H. Zhang et al., "MP-former: Mask-piloted transformer for image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18074–18083.

[10] K. Yan et al., "Two-shot video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 2257–2267.

[11] L. Qi et al., "High quality entity segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4047–4056.

[12] J. Zhu, Y. Luo, X. Zheng, H. Wang, and L. Wang, "A good Student is cooperative and reliable: CNN-transformer collaborative learning for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 11720–11730.

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

[14] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[15] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6881–6890.

[16] K. Wu, H. Peng, M. Chen, J. Fu, and H. Chao, "Rethinking and improving relative position encoding for vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10033–10041.

[17] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2020, pp. 1–11.

[18] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[19] J. Zhang et al., "Delivering arbitrary-modal semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1136–1147.

[20] R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra, "Omnivore: A single model for many visual modalities," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16102–16112.

[21] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14679–14694, Dec. 2023.

[22] R. Bachmann, "Multimae: Multi-modal multi-task masked autoencoders," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Oct. 2022, pp. 348–367.

[23] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 12186–12195.

[24] L.-Z. Chen, Z. Lin, Z. Wang, Y.-L. Yang, and M.-M. Cheng, "Spatial information guided convolution for real-time RGBD semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 2313–2324, 2021.

[25] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 11976–11986.

[26] X. Chen, X. Chen, Y. Zhang, X. Fu, and Z.-J. Zha, "Laplacian pyramid neural network for dense continuous-value regression for complex scenes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 5034–5046, Nov. 2021.

[27] Q.-D. Pham et al., "Segtransvae: Hybrid CNN–transformer with regularization for medical image segmentation," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–5.

[28] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[29] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[31] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Sys. (NIPS)*, vol. 34, Dec. 2021, pp. 12077–12090.

[32] H. Fan et al., "Multiscale vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6824–6835.

[33] W. Yu et al., "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10819–10829.

[34] S. Chen et al., "MSP-former: Multi-scale projection transformer for single image desnowing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[35] A. Shabani, A. H. Abdi, L. Meng, and T. Sylvain, "Scaleformer: Iterative multi-scale refining transformers for time series forecasting," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2022, pp. 1–17.

[36] Y.-H. Wu, Y. Liu, X. Zhan, and M.-M. Cheng, "P2T: Pyramid pooling transformer for scene understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12760–12771, Nov. 2023.

[37] X. Chu et al., "Conditional positional encodings for vision transformers," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2021, pp. 1–20.

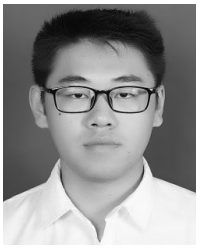
[38] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16519–16529.

[39] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2021, pp. 1–9.

[40] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2021, pp. 22–31.
[41] Y. Yuan et al., "HRFormer: High-resolution vision transformer for dense predict," in *Proc. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 7281–7293.
[42] Z. Wang, X. Huo, Z. Chen, J. Zhang, L. Sheng, and D. Xu, "Improving RGB-D point cloud registration by learning multi-scale local linear transformation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2022, pp. 175–191.
[43] T. Ghosh Mondal, M. R. Jahanshahi, and Z. Y. Wu, "Deep learning-based RGB-D fusion for multimodal condition assessment of civil infrastructure," *J. Comput. Civil Eng.*, vol. 37, no. 4, Jul. 2023, Art. no. 04023017.
[44] B. Zhou, P. Wang, J. Wan, Y. Liang, and F. Wang, "A unified multimodal *De-* and re-coupling framework for RGB-D motion recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 11428–11442, Oct. 2023.
[45] Q. Zhao, Y. Wan, J. Xu, and L. Fang, "Cross-modal attention fusion network for RGB-D semantic segmentation," *Neurocomputing*, vol. 548, Sep. 2023, Art. no. 126389.
[46] X. Hu, K. Yang, L. Fei, and K. Wang, "ACNET: Attention based network to exploit complementary features for RGBD semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1440–1444.
[47] X. Chen et al., "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Aug. 2020, pp. 561–577.
[48] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, "ShapeConv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7088–7097.
[49] Z. Liu, R. Xiong, and T. Jiang, "CI-Net: Clinical-inspired network for automated skin lesion recognition," *IEEE Trans. Med. Imag.*, vol. 42, no. 3, pp. 619–632, Mar. 2023.
[50] J. Liu and Y. Zhang, "High quality monocular depth estimation with parallel decoder," *Sci. Rep.*, vol. 12, no. 1, p. 16616, Oct. 2022.
[51] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2021, pp. 10347–10357.
[52] H. Bao, D. Li, and F. Wei, "BEiT: BERT pre-training of image transformers," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2021, pp. 1–8.
[53] L. Yuan et al., "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 558–567.
[54] W. Wang et al., "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, Sep. 2022.
[55] X. Chu et al., "Twins: Revisiting the design of spatial attention in vision transformers," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9355–9366.
[56] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3273–3282.
[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[58] Y. Liao, J. Xie, and A. Geiger, "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3292–3310, Mar. 2023.
[59] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 7576. Cham, Switzerland: Springer, 2012, pp. 746–760.
[60] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.
[61] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[62] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21919–21928.
[63] W. Zhou, E. Yang, J. Lei, J. Wan, and L. Yu, "PGDENet: Progressive guided fusion and depth enhancement network for RGB-D indoor scene parsing," *IEEE Trans. Multimedia*, vol. 25, pp. 3483–3494, 2022.
[64] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[65] J.-W. Bian, H. Zhan, N. Wang, T.-J. Chin, C. Shen, and I. Reid, "Auto-rectify network for unsupervised indoor depth estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9802–9813, Dec. 2022.
[66] X. Dong, M. A. Garratt, S. G. Anavatti, and H. A. Abbass, "MobileXNet: An efficient convolutional neural network for monocular depth estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 20134–20147, Nov. 2022.
[67] M. Song, S. Lim, and W. Kim, "Monocular depth estimation using Laplacian pyramid-based depth residuals," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4381–4393, Nov. 2021.
[68] A. Agarwal and C. Arora, "Depthformer: Multiscale vision transformer for monocular depth estimation with global local information fusion," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 3873–3877.

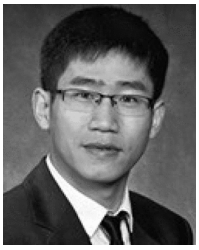
Siyu Chen (Student Member, IEEE) received the bachelor's degree in computer engineering from Jimei University, Xiamen, Fujian, China. He has published several articles in top-tier journals, including IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS and *Alimentary Pharmacology and Therapeutics*. His research spans multiple areas, with a focus on deep learning, 2-D and 3-D object detection, semantic segmentation, 3-D reconstruction, and building extraction.

Ting Han (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree with the School of Geospatial Engineering and Science, Sun Yat-sen University, Zhuhai, Guangdong, China. He has published several articles in top journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, *ISPRS Journal of Photogrammetry and Remote Sensing*, and *International Journal of Applied Earth Observation and Geoinformation*. His research interests include image and point cloud processing, semantic segmentation, 3-D reconstruction, deep learning theory in 3-D vision, and its applications in remote sensing.



Changshe Zhang received the bachelor's degree in microelectronics and engineering from the School of Ocean Information Engineering, Jimei University, Xiamen, China, in 2020. He is currently pursuing the master's degree with the School of Artificial Intelligence, Xidian University, Xi'an, China. His research interests include object detection and medical image segmentation.

Jinhe Su received the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China. He is currently a Teacher with the School of Computer Engineering, Jimei University, Xiamen, Fujian, China. His main research interests are machine learning, computer vision, remote sensing data processing, image processing, and virtual reality.



Ruisheng Wang (Senior Member, IEEE) received the B.Eng. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, the M.Sc.E. degree in geomatics engineering from the University of New Brunswick, Fredericton, NB, Canada, and the Ph.D. degree in electrical and computer engineering from McGill University, Montreal, QC, Canada.

In 2012, he joined the University of Calgary, where he is currently a Professor with the Department of Geomatics Engineering. Prior to that, he was an Industrial Researcher with HERE (formerly NAVTEQ), Chicago, IL, USA, in 2008. His research interests include mobile LiDAR data processing for next-generation map making and navigation.

Yiping Chen (Senior Member, IEEE) received the Ph.D. degree in information communication engineering from the National University of Defense Technology, Changsha, China, in 2011. From 2007 to 2011, she was an Assistant Researcher with The Chinese University of Hong Kong, Hong Kong. She is currently an Associate Professor with the School of Geospatial Engineering and Science, Sun Yat-sen University, Zhuhai, China. She has co-authored more than 80 publications, including in top remote sensing journal, such as IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, *International Society for Photogrammetry and Remote Sensing* (ISPRS), *Journal of Photogrammetry and Remote Sensing* (JPRS), and *International Journal of Applied Earth Observation and Geoinformation* (JAG), and in flagship Computer Vision (CV), Artificial Intelligence (AI) conferences, such as IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Association for the Advancement of Artificial Intelligence (AAAI), and International Joint Conference on Artificial Intelligence (IJCAI). Her main research interests include image processing, mobile LiDAR data analysis, and GeoAI.

Dr. Chen is the Co-Chair of the ISPRS WG I/4 on LiDAR, Laser Altimetry and Sensor Integration from 2022 to 2026. She is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Zongyue Wang received the Ph.D. degree from Wuhan University, Wuhan, Hubei, China. He is currently a Professor with the School of Computer Engineering, Jimei University, Xiamen, Fujian, China. His main research interests are image processing, machine learning, computer vision, and remote sensing data processing.

Guorong Cai (Senior Member, IEEE) received the Ph.D. degree from Xiamen University, Xiamen, Fujian, China. He is currently a Full Professor with the Computer Engineering College, Jimei University, Xiamen. His main research directions are computer vision and point cloud processing, including: 3-D reconstruction, image and point cloud based object detection and recognition, image registration, deep learning theory, and its applications.