This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DFormerv2: Geometry Self-Attention for RGBD Semantic Segmentation

Bo-Wen Yin, Jiao-Long Cao, Ming-Ming Cheng, Qibin Hou*
VCIP, CS, Nankai University
bowenyin@mail.nankai.edu.cn, houqb@nankai.edu.cn

Abstract

Recent advances in scene understanding benefit a lot from depth maps because of the 3D geometry information, especially in complex conditions (e.g., *low light and over-exposed). Existing approaches encode depth maps along with RGB images and perform feature fusion between them to enable more robust predictions. Taking into account that depth can be regarded as a geometry supplement for RGB images, a straightforward question arises: Do we really need to explicitly encode depth information with neural networks as done for RGB images? Based on this insight, in this paper, we investigate a new way to learn RGBD feature representations and present DFormerv2, a strong RGBD encoder that explicitly uses depth maps as geometry priors rather than encoding depth information with neural networks. Our goal is to extract the geometry clues from the depth and spatial distances among all the image patch tokens, which will then be used as geometry priors to allocate attention weights in self-attention. Extensive experiments demonstrate that DFormerv2 exhibits exceptional performance in various RGBD semantic segmentation benchmarks. Code is available at: https://github.com/VCIP-RGBD/DFormer.*

1. Introduction

Semantic segmentation, aiming at assigning each pixel in an image to a specific pre-defined category label, has been a fundamental area of research in computer vision due to its broad range of applications, such as in intelligent transportation systems and autonomous driving [30]. However, approaches based solely on RGB data often suffer significant performance degradation in complex scenarios, such as cluttered indoor environments or low-light conditions. In recent years, advancements in 3D modular sensors have made depth data more accessible. Integrating RGB-D data makes scene understanding more robust and accurate and thus becomes pivotal in advancing high-level vision tasks. Furthermore, RGB-D data have demonstrated remarkable potential, surpassing the RGB-based paradigm in various downstream tasks, including autonomous driving [26], SLAM [50], and robotics [36].

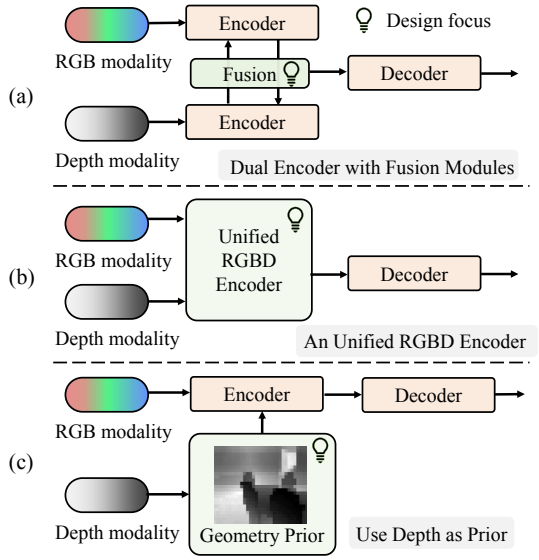


Figure 1. Comparisons among the main RGBD segmentation pipelines and our approach. (a) Use dual encoders to encode RGB and depth respectively and design fusion modules to fusion them [28, 61]; (b) Adopt an unified RGBD encoder to extract and fuse RGBD features [1, 59]; (c) Our DFormerv2 use depth to form a geometry prior of the scene and then enhance the visual features.

Fig. 1(a) presents the architecture of current mainstream RGB-D models. As depicted, it utilizes a dual encoder architecture [61, 62], wherein one encoder extracts features from the RGB modality, while the other processes depth information. Meanwhile, a fusion strategy is performed to achieve interaction between the information of these two modalities during the encoding process. Despite the success, the majority existing RGBD segmentation approaches adopt identical backbone architectures to extract features from both RGB and depth data for fusion, neglecting the inherent differences between the RGB and depth.

A series of studies have sought to identify optimal methods for processing depth maps and integrating them with

*Qibin Hou is the corresponding author.

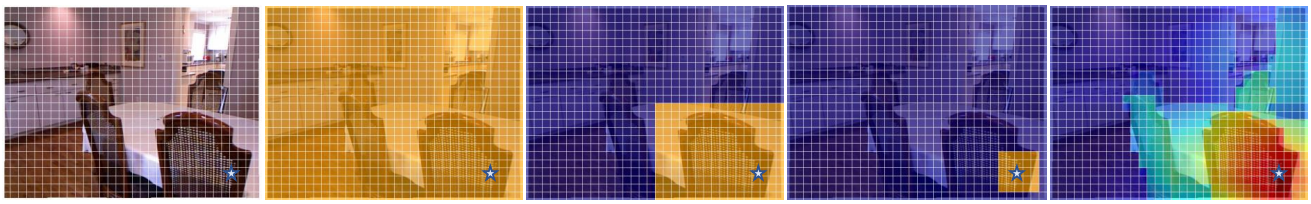


Figure 2. Comparison between geometry self-attention (GSA) and other attention mechanisms, *i.e.,* vanilla attention [10], window attention [9, 35], and local attention [52, 57]. The 'star' sign means the current query's position. In GSA, colors closer to red represent smaller decay rates, while colors farer away red represent larger ones. In other attention mechanisms, the bright color means the receptive field.

RGB data. Asyformer [11] employs a dual-stream asymmetric backbone, *i.e.,* using a more efficient encoder for depth data to reduce redundant parameters during feature extraction. PrimKD [20] proposes a knowledge distillation (KD)-based method to guide multimodal fusion in RGB-D semantic segmentation, with an emphasis on leveraging the primary RGB modality. Furthermore, as shown in Fig. 1(b), DFormer [59] presents an efficient RGB-D model that encodes both RGB and depth data in a unified encoder via representation learning manners [23, 47], yet allocates more computational resources to processing the RGB data. These methods acknowledge that RGB and depth carry distinct information, each contributing differently to semantic segmentation. However, they fail to account for the unique characteristics of the depth modality fully. In conclusion, how to effectively and efficiently utilize depth information remains an open question and warrants further exploration.

In this paper, taking into account the physical meanings of depth maps that reflect the geometrical information of the given scenes, we consider the way of utilizing depth maps from a new perspective. Unlike previous works that use neural networks to simultaneously encode RGB images and depth maps as shown in Fig. 1, we propose directly employing depth maps as geometry priors and using them to guide weight distributions in self-attention, producing a new attention mechanism, called Geometry Self-Attention (GSA). An illustration of how the proposed GSA works and its differences from other self-attention variants can be found in Fig. 2. In each building block, we model the geometric and spatial relationships among all the patch tokens based on the GSA, a more efficient way to fuse RGB and depth information. Our method requires no extra layers to process depth maps and hence needs fewer learnable parameters and computations compared to other types of RGBD segmentation methods. In addition, to reduce the computational burden of vanilla self-attention, we also adopt an axes decomposition operation that decomposes self-attention along both spatial axes of the features.

Based on Geometry Self-Attention, we construct a powerful RGB-D vision backbone, called DFormerv2. We demonstrate the effectiveness of DFormerv2 on popular RGB-D semantic segmentation benchmarks, *e.g.,* NYU DepthV2 [42], SUNRGBD [43], and Deliver [62]. By adding a small decoder head on top of the DFormerv2, our approach sets new state-of-the-art records with less computational cost compared to previous methods. Remarkably, our base scale model, DFormerv2-B, achieves equal performance with the second-best method Gemnifuision (MiT-B5) [28], *i.e.,* 57.7% mIoU on NYU DepthV2, with less than half of the computation costs. Meanwhile, our largest model DFormerv2-L is able to achieve 58.4% mIoU on NYU DepthV2 with 95.5M parameters. Compared with other methods, our DFormerv2 achieves the best trade-off between segmentation performance and computations.

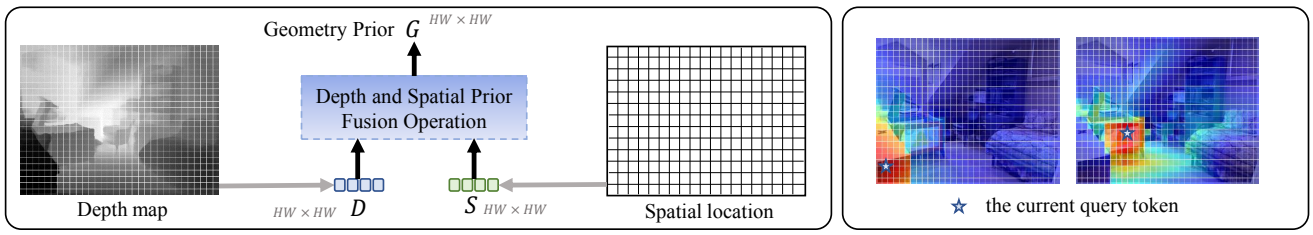
Our main contributions can be summarized as follows:

- To our best knowledge, our work marks the first successful attempt to combine depth information with spatial information as a geometry prior and apply it to the neural network.
- We propose Geometry Self-Attention which introduces geometry prior to self-attention, to construct an efficient RGB-D encoder, termed DFormerv2.
- Our method achieves new state-of-the-art performance with less than half the computational cost of the best current methods on three popular RGB-D semantic segmentation datasets.

2. Related Work

2.1. RGB-D Semantic Segmentation

Semantic segmentation [64], as one of the core pursuits in computer vision, aims to categorize each pixel in an image into a specific category. Recently, significant developments in deep learning technologies [6, 7, 18, 44, 45] have been made in this field. However, some real-world scenes [13, 27, 32, 33, 56] are still challenging to understand using only RGB images, which do not provide sufficient textures, especially in low illumination and fast-moving scenarios. To address this issue, researchers [63, 65] propose to utilize depth, which contains 3D geometry information for the scene, to enhance RGB semantic segmentation, known as RGB-D semantic segmentation. Since then, a series of works have been proposed to achieve the fusion of RGB-D data and leverage the additional information to capture more



(a) Generation of geometry prior

(b) Visualization of geometry prior

Figure 3. Illustration of the geometry prior. (a) Generation process of the geometry prior. (b) Some visualization of the geometry prior, where the 'blue star' means the current query.

details. Here, we delve into the RGB-D fusion schemes and analyze their characteristics.

The current mainstream methods [25, 39] put a lot of effort into designing interaction modules to fuse the RGB and depth features encoded by two parallel pretrained backbones. For instance, methods such as CMX [61], TokenFusion [54], GeminiFusion [28] dynamically fuse the RGB-D representations from RGB and depth encoders and aggregate them in the decoder. These methods significantly push the performance boundaries in the applications of RGB-D semantic segmentation. Nevertheless, they still face two common issues: (1) Treating RGB and depth maps equally with two parallel backbones brings significantly higher computational cost compared to methods based on RGB data; (2) The used backbones are pretrained with RGB images but take an image-depth pair as input during fine-tuning. The inconsistency between the input causes a huge representation distribution shift.

Recently, DFormer [59] proposed an RGB-D representation learning framework and utilizes a unified backbone that pretrained on RGB-D pairs to overcome the two issues. It notices the difference in information density between these two modalities and observes that depth information only requires a small portion of channels to encode. Although it achieves accurate prediction with high efficiency, it overlooks the intrinsic characteristics of depth modality and merely allocates depth with lower computation cost. Differently, in this paper, we propose to generate the geometry prior via depth from the perspective of data characteristics. To the best of our knowledge, this is the first attempt to explicitly use the geometry information of depth without any extra encoding layers.

2.2. Vision Transformer and Prior Knowledge

Vision Transformer (ViT) [10] was the first to introduce transformer architecture to visual tasks, where images are split into small, non-overlapped patch sequences. The biggest difference from CNNs [22, 24, 34] is that transformers [8, 9, 35, 49] use attention as an alternative to convolution layers to enable global context modeling. However, vanilla self-attention incurs a heavy computational burden, as it computes pairwise feature affinities across all patches. Various sparse attention mechanisms [19, 35, 55, 60, 69] have been proposed to alleviate the huge computation cost of self-attention. At the same time, researchers have presented many studies [17, 37, 46, 48, 51] to incorporate prior knowledge into the transformer model to enhance its representation capacity. The original transformers [10] utilize position encoding to provide positional information for each token. For vision tasks, swin-transformer [35] proposes to use relative positional encoding instead of the original absolute position encoding. In contrast, we propose transferring depth into geometry prior knowledge and introducing it into self-attention, termed geometry self-attention. Compared to the position prior, our geometry prior can model the relationships in the 3D domain across the whole image.

3. Methodology

3.1. Geometry Prior Generation

In the vision transformer, the 2D input image of size $h \times w$ is evenly split into HW small patches, where H and W are the numbers of patches per row and column, respectively. Each patch denoted as P_{ij} is uniquely positioned with a two-dimensional coordinate within the spatial domain, where i and j index the row and column, respectively. When the associated depth map is given, the patch in the depth map at the corresponding position reflects its distance from the camera plane. Based on these two types of priors, we model the geometrical relationships among all the patches and embed them into the self-attention mechanism to form our geometry self-attention.

To be specific, for the depth prior, we perform the average pooling operation for all pixels within the depth patch at position (i, j) to represent its depth location z_{ij} and calculate the distances between each pair of depth patches, which can be defined as:

$$D_{ij,i'j'} = |z_{ij} - z_{i'j'}|, \tag{1}$$

where $D_{ij,i'j'}$ represents the depth distance between the patches at positions (i, j) and (i', j'). D forms a depth relationship matrix of shape $HW \times HW$.

The depth relationship matrix D does not contain the spatial distance information, which is also vital to form the

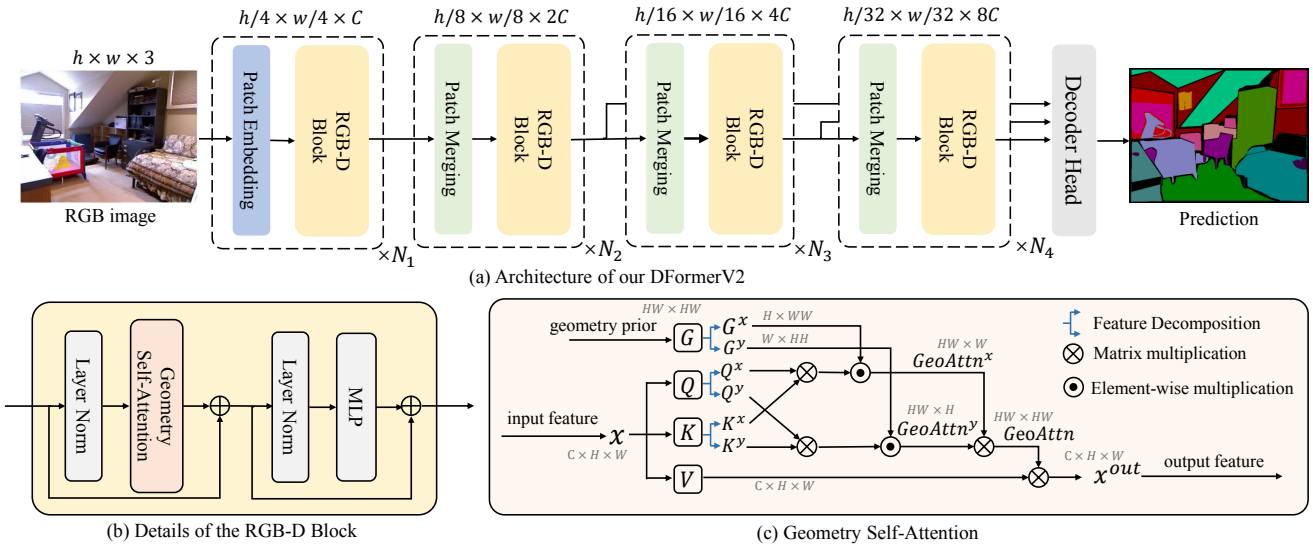


Figure 4. Illustration of our DFormerv2. (a) Overall architecture of our DFormerv2, which contains an encoder with pyramid structure and a decoder head that receives input from the last three stage features. (b) Detailed structure of the basic building block. (c) Detailed illustration of the proposed geometry self-attention mechanism.

geometry clues. Thus, we need to bridge the depth prior with spatial prior as the geometry prior to model comprehensive relationships between image patches. Similar to the processing of depth prior, we calculate the spatial distance among all the image patches with Manhattan distance. This can be defined as:

$$S_{ij,i'j'} = |i - i'| + |j - j'|, \tag{2}$$

where $S_{ij,i'j'}$ represent the spatial Manhattan distance between patches at positions (i, j) and (i', j'). Similar to the depth relationship matrix, we can also produce a spatial relationship matrix S of shape $HW \times HW$.

Given the depth and spatial distance matrices D and S, we perform the fusion operation to build the bridge between them, as shown in Fig. 3. We empirically found that simply using two learnable memories to perform the weighted summation for the depth and spatial priors already works well. It is worth mentioning that more advanced techniques can also be used to generate the geometry prior by fusing the depth and spatial priors. We integrate both types of priors to generate the geometry prior G of shape $HW \times HW$ that stores more comprehensive 3D geometrical relationships for all the image patches. More visualizations of G are shown in Fig. 7.

3.2. Geometry Self-Attention

Given a feature map $x \in \mathbb{R}^{HW \times C}$, self-attention can be simply formulated as follows in each head:

$$\text{SelfAtt}(Q, K, V) = \text{Softmax}(QK^T)V, \tag{3}$$

where Q, K, V are the query, key, and value matrices that can be attained by linear projections. Inspired by [12, 41, 48] that perform positional encoding to provide spatial information for each token, our geometry self-attention can be achieved by introducing the geometry prior G into the self-attention mechanism via a decay manner. This process can be written as:

$$\text{GeoAttn}(Q, K, V, G) = (\text{Softmax}(QK^T) \odot \beta^G)V, \tag{4}$$

where $\odot$ means element-wise multiplication, $\beta \in (0, 1)$ is the decay rate, and β^G means taking each element in G as the power of β to obtain a new matrix. As $\beta \in (0, 1)$ and the elements in G are non-negative numbers, the resulting $\beta^G = [\beta^{g_{ij}}]_{ij} \in (0, 1]^{HW \times HW}$ is a matrix with 1 in diagonal. Small element values mean long geometric distances. β^G embeds the explicit geometry prior into the attention map via multiplication, and GSA obtains the focus in the near regions, as visualized in Fig. 8. Specifically, for a query, the weights of irrelevant key-value pairs are suppressed and the relevant ones are enhanced according to the geometry relationship, benefitting the attention mechanism in modeling intra-object and inter-object relationships. In practical use, Eqn. (4) can also be extended to a multi-head version, and meanwhile, we set different decay rates for different self-attention heads to augment the geometry guidance.

As demonstrated in previous works for dense prediction tasks [35], the pyramid structure is often used to encode fine-level features. However, directly using self-attention to encode high-resolution features will introduce high computations and memory costs. Our geometry self-attention also

Model	Backbone	Params	NYUDepthv2			SUN-RGBD		
			Input size	Flops	mIoU	Input size	Flops	mIoU
ESANet$_{21}$ [39]	ResNet-34	31.2M	480 × 640	34.9G	50.3	480 × 640	34.9G	48.2
TokenFusion$_{22}$ [54]	MiT-B2	26.0M	480 × 640	55.2G	53.3	530 × 730	71.1G	50.3
Omnivore$_{22}$ [16]	Swin-Tiny	29.1M	480 × 640	32.7G	49.7	530 × 730	—	—
DFormer$_{24}$ [59]	DFormer-Tiny	6.0M	480 × 640	14.7G	51.8	530 × 730	18.9G	48.8
DFormer$_{24}$ [59]	DFormer-Small	18.7M	480 × 640	28.7G	53.6	530 × 730	37.0G	50.0
DFormer$_{24}$ [59]	DFormer-Base	29.5M	480 × 640	45.3G	55.6	530 × 730	58.4G	51.2
AsymFormer$_{24}$ [11]	MiT-B0+ConvNeXt-Tiny	33.0M	480 × 640	39.4G	55.3	530 × 730	52.6G	49.1
★ DFormerv2-S	DFormerv2-Small	26.7M	480 × 640	33.9G	**56.0**	530 × 730	43.7G	**51.5**
SGNet$_{20}$ [4]	ResNet-101	64.7M	480 × 640	108.5G	51.1	530 × 730	151.5G	48.6
ShapeConv$_{21}$ [3]	ResNext-101	86.8M	480 × 640	124.6G	51.3	530 × 730	161.8G	48.6
FRNet$_{22}$ [68]	ResNet-34	85.5M	480 × 640	115.6G	53.6	530 × 730	150.0G	51.8
EMSANet$_{22}$ [40]	ResNet-34	46.9M	480 × 640	45.4G	51.0	530 × 730	58.6G	48.4
TokenFusion$_{22}$ [54]	MiT-B3	45.9M	480 × 640	94.4G	54.2	530 × 730	122.1G	51.4
Omnivore$_{22}$ [16]	Swin-Small	51.3M	480 × 640	59.8G	52.7	530 × 730	—	—
CMX$_{22}$ [61]	MiT-B2	66.6M	480 × 640	67.6G	54.4	530 × 730	86.3G	49.7
DFormer$_{24}$ [59]	DFormer-Large	39.0M	480 × 640	69.3G	57.2	530 × 730	89.2G	52.5
GeminiFusion$_{24}$ [28]	MiT-B3	75.8M	480 × 640	138.2G	56.8	530 × 730	179.0G	52.7
★ DFormerv2-B	DFormerv2-Base	53.9M	480 × 640	67.2G	**57.7**	530 × 730	86.9G	**52.8**
ACNet$_{19}$ [25]	ResNet-50	116.6M	480 × 640	126.7G	48.3	530 × 730	163.9G	48.1
SA-Gate$_{20}$ [5]	ResNet-101	110.9M	480 × 640	193.7G	52.4	530 × 730	250.1G	49.4
CEN$_{20}$ [53]	ResNet-101	118.2M	480 × 640	618.7G	51.7	530 × 730	790.3G	50.2
CEN$_{20}$ [53]	ResNet-152	133.9M	480 × 640	664.4G	52.5	530 × 730	849.7G	51.1
PGDENet$_{22}$ [67]	ResNet-34	100.7M	480 × 640	178.8G	53.7	530 × 730	229.1G	51.0
MultiMAE$_{22}$ [1]	ViT-Base	95.2M	640 × 640	267.9G	56.0	640 × 640	267.9G	51.1†
Omnivore$_{22}$ [16]	Swin-Base	95.7M	480 × 640	109.3G	54.0	530 × 730	—	—
CMX$_{22}$ [61]	MiT-B4	139.9M	480 × 640	134.3G	56.3	530 × 730	173.8G	52.1
CMX$_{22}$ [61]	MiT-B5	181.1M	480 × 640	167.8G	56.9	530 × 730	217.6G	52.4
CMNext$_{23}$ [62]	MiT-B4	119.6M	480 × 640	131.9G	56.9	530 × 730	170.3G	51.9†
GeminiFusion$_{24}$ [28]	MiT-B5	137.2M	480 × 640	256.1G	57.7	530 × 730	332.4G	**53.3**
★ DFormerv2-L	DFormerv2-Large	95.5M	480 × 640	124.1G	**58.4**	530 × 730	160.5G	**53.3**

Table 1. Results on NYU Depth V2 [42] and SUN-RGBD [43]. Some methods do not report the results or settings on the SUN-RGBD datasets, so we reproduce them with the same training configs. † indicates that we follow the results from [59]. All the backbones are pretrained on ImageNet-1K. We split the models to three sets, *i.e.,* small scale, base scale, and large scale. We can see that our method receives the best results on both datasets.

faces this issue. Thus, inspired by existing sparse attention approaches [9, 12, 21, 57], we use a simple decomposition manner to perform attention along the horizontal and vertical directions separately, as shown in Fig. 4(c). To achieve this, we also need to generate the horizontal and vertical geometry priors. Therefore, we decompose the geometry prior G into G^x and G^y, which reflect the geometry relationship at rows and columns for all the tokens respectively. Specifically, $G^y = [G^y_{ij}]_{i=0,1,..,H-1,j=0,1,...,W-1}$ is a matrix of shape (HW, H), where G^y_{ij} represents the geometry relationship between the patch at (i, j) and all the patches in j-th column. Similarly, we can obtain G^x with shape (HW, W). Then, the calculation of geometry self-attention is formulated as follows:

$$\text{GeoAttn}^y = (\text{Softmax}(Q^y(K^y)^T) \odot \beta^{G^y}), \quad (5)$$

$$\text{GeoAttn}^x = (\text{Softmax}(Q^x(K^x)^T) \odot \beta^{G^x}), \quad (6)$$

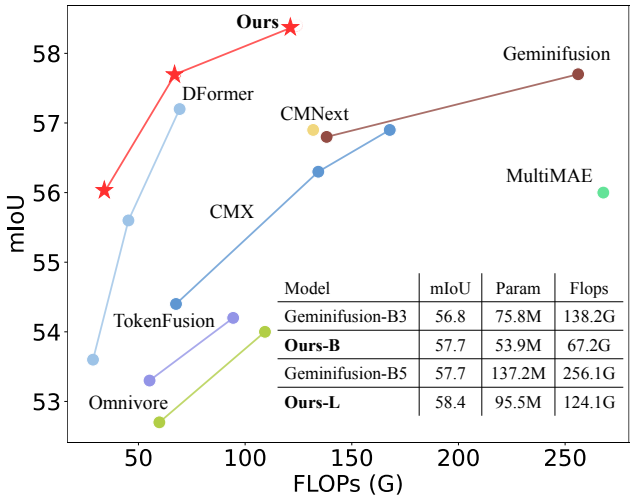
$$\text{GeoAttn} = \text{GeoAttn}^y(\text{GeoAttn}^x V)^T, \quad (7)$$

where $Q^y(K^y)^T$ and $Q^x(K^x)^T$ means perform attention calculation along vertical and horizontal axis.

3.3. DFormerv2 Architecture

Fig. 4 illustrates the overall architecture of DFormerv2, which follows the widely-used encoder-decoder framework. The encoder is composed of four stages, which are utilized to produce multi-scale features. Each stage contains a stack of geometry self-attention blocks. The first three stages perform decomposition on geometry self-attention, while the last one does not. A lightweight decoder head is employed to transform these visual features into RGB-D semantic segmentation results.

Given an RGB image, it is first processed by a stem layer, consisting of two convolutions with kernel size 3 × 3 and stride 2. Then, the RGB features are fed into the hierarchical encoder to encode multi-scale features at $\{1/4, 1/8, 1/16, 1/32\}$ of the original image resolution. Different from existing methods, there is no need for our DFormerv2 to explicitly encode depth maps. We just need to perform the average pooling operation with different pooling kernels and strides on the depth map to the four scales corresponding to the geometry self-attention blocks in the encoder and then utilize them to generate the geome-



Model	mIoU	Param	Flops
Geminifusion-B3	56.8	75.8M	138.2G
Ours-B	57.7	53.9M	67.2G
Geminifusion-B5	57.7	137.2M	256.1G
Ours-L	58.4	95.5M	124.1G

Figure 5. Performance-computation comparisons between our DFormerv2 and other SOTA methods on NYU DepthV2 [42].

try prior for each block. Based on the configurations of the geometry self-attention blocks in each stage, we design a series of encoder variants, termed DFormerv2-S, DFormerv2-B, and DFormerv2-L, respectively, with the same architecture but different model sizes. Detailed configurations of these variants can be found in the supplementary materials.

4. Experiments

4.1. Implementation Details

Pretraining settings. Following DFormer [59] and MultiMAE [1], we perform RGB-D pretraining on ImageNet-1K for our DFormerv2, to endow the encoder with the ability to achieve the interaction between RGB and depth modalities and generate transferable representations with rich semantic and spatial information. The depth maps for ImageNet are generated by depth estimation method [59]. The standard cross-entropy loss is employed as our optimization objective, and the number of training epochs is set to 300, like most pretrained models [34]. Following previous works [59, 61], the AdamW [31] with learning rate 1e-3 and weight decay 5e-2 are adopted, and we set the batch size to 1024. More detailed settings for each variant of DFormerv2 are described in the supplementary materials.

Datasets and settings for finetuning. Following the commonly used experiment settings of RGB-D semantic segmentation works [20, 28, 59], we evaluate our DFormerv2 on two popular datasets, *i.e.,* NYU DepthV2 [42] and SUNRGBD [43]. Additionally, we conduct experiments on the Deliver dataset [62], as done in [28]. In line with DFormer [59], we use a lightweight head [15] as our decoder to build our RGB-D semantic segmentation model. We only adopt two simple data augmentation methods, *i.e.,* random horizontal flipping and random scaling (from

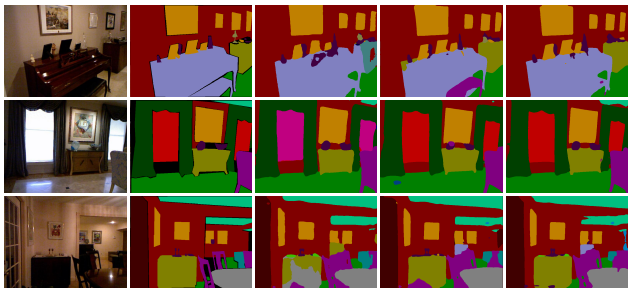


Figure 6. Qualitative comparisons with GeminiFusion-B5 [28] and DFormer-L [59]. 'GT' is the ground truth.

0.5 to 1.75), when finetuning models. We use the cross-entropy loss as the optimization objective and AdamW [31] as our optimizer with an initial learning rate of 6e-5 and the poly decay schedule. For the NYU DepthV2 and SUNRGBD datasets, we crop and resize the images to 480×640 and 480×480 respectively for training. During the evaluation, we adopt the mean Intersection over Union (mIoU), which is averaged across all semantic categories, as the primary evaluation metric to measure the segmentation accuracy. Following recent works [59, 61, 62], we employ multi-scale (MS) flip inference strategies at scales $\{0.5, 0.75, 1, 1.25, 1.5\}$. We adopt the same training and testing strategy as CMNeXt [62] on DeLiVER, where the images are resized to 1024×1024. More details can be found in the supplementary materials.

4.2. Comparisons with Other Methods

We compare our DFormerv2 with 17 recent RGB-D semantic segmentation approachs on the NYU DepthV2 [42], SUNRGBD [43], and Deliver [62] datasets. In Tab. 1, we categorize the variants of all methods into three sets based on model scale, *i.e.,* small scale, base scale, and large scale, for a more intuitive and fair comparison. As can be seen, DFormerv2 achieves new SOTA performance across all the model scale settings on the two benchmarks. We also plot the performance computation cost curves of different methods in Fig. 5. DFormerv2 achieves better performance and computation trade-off compared to other methods. Specifically, our largest model, *i.e.,* DFormerv2-L achieves 58.4% mIoU with 95.5M parameters and 124.1G Flops, surpassing the second-best method Gemnifusion [28] by 0.7% with less than its half computations. Similarly, at base and small scales, our DFormerv2 also consistently outperforms other SOTA methods with higher efficiency. In SUNRGBD and Deliver (Tab. 2) datasets, our DFormerv2 also brings significant improvements. Moreover, the visual comparisons between the semantic segmentation results of our DFormerv2 and Gemnifusion [28] are shown in Fig. 6. These improvements demonstrate that our DFormerv2 can more efficiently utilize the geometry prior within the depth maps without ex-

Model	Backbone	Params	**Flops**	**mIoU**
HRFuser [2]	HRFormer-T	30.5M	223.0G	51.9
TokenFusion [54]	MiT-B2	26.0M	55.0G	60.3
★ DFormerv2-S	DFormerv2-S	26.7M	28.9G	**63.7**
CMX [61]	MiT-B2	66.6M	65.7G	62.7
CMNext [62]	MiT-B2	58.7M	62.9G	63.6
★ DFormerv2-B	DFormerv2-B	53.9M	60.8G	**65.2**
CMNext [62]	MiT-B4	116.6M	112.0G	66.3
GeminiFusion$_{24}$ [28]	MiT-B5	137.2M	218.4G	66.9
TokenFusion [54]	MiT-B5	83.3M	144.7G	63.5
★ DFormerv2-L	DFormerv2-L	95.5M	114.5G	**67.1**

Table 2. Results on Deliver [62] dataset. Following [62], the Flops is calculated on the images with shape 512×512.

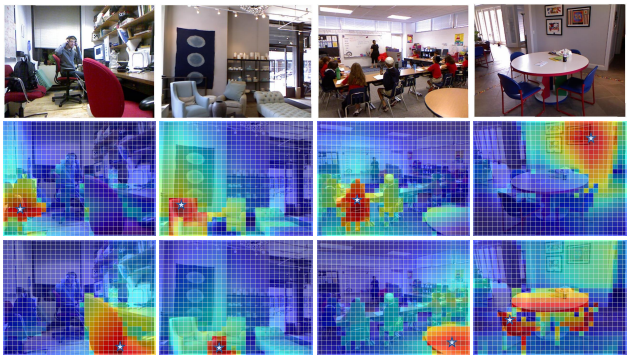


Figure 7. Some visualization samples of the geometry prior. The blue 'star' means the current query token. It is important to note that the visualized prior is only obtained from the depth map.

plicit encoding, and hence yields more accurate predictions with even lower computational cost.

4.3. Model Analysis

Geometry self-attention. The proposed geometry self-attention consists of depth prior, spatial prior, and priors fusion, which are integrated into a unified geometry prior. To evaluate the effectiveness of each component, we present a roadmap from the vanilla self-attention to the geometry self-attention in Tab. 3. First of all, we introduce the depth prior, spatial prior into the vanilla self-attention (steps 1-2), respectively, to observe the impact on performance. These substitutions result in accuracy improvement of 2.6%, and 1.8% on NYUDepthV2 and 1.7%, and 1.3% on SUNRGBD, respectively, compared to the baseline, highlighting the importance of incorporating these priors in self-attention. However, it is also apparent that the simple addition of depth and spatial priors results in only a modest improvement over using just the depth prior, indicating that this method of integration may not be effective. In Step 3, when we introduce fusion operation to bridge the two priors and form the geometry prior, we observe a further improvements on NYUDepthV2 and SUNRGBD, with a negligible increase in computational cost. These results (step 1-3) shows that the geometry prior significantly enhances performance with little increase in complexity. Moreover, in

Step	Attention	Params	Flops	NYUDepthV2	SUNRGBD
0	Vanilla Attn	26.5M	51.4G	51.7	47.8
1	+Only Depth Prior	26.5M	51.4G	54.3 (**+2.6**)	49.9 (**+1.7**)
2	+Only Spatial Prior	26.5M	51.4G	53.5 (**+1.8**)	49.1 (**+1.3**)
3	+Both Priors	26.7M	51.7G	56.2 (**+4.5**)	51.7 (**+3.9**)
4	+decomposition	26.7M	33.9G	56.0 (**+4.3**)	51.5 (**+3.7**)

Table 3. The ablation experiments demonstrate the full roadmap from vanilla Self-Attention to our geometry self-attention on the small scale of DFormerv2. In step 0 and 2, we only input RGB images while we use RGB-D at all the other steps.

Figure 8. Visualization of the focused attention.

Figure 9. Visualization of the features with and without geometry priors. They are randomly picked from the first stage output.

Step 4, we perform a decomposition of the attention mechanism, which further alleviates the computational burden while maintaining almost the same level of performance. Overall, compared to self-attention, integrating geometry priors enables better RGB-D segmentation with minimal computational overhead and a slight increase in parameters.

More insights about the geometry prior. The geometry prior G with a shape $HW \times HW$ is derived from the depth map and represents the geometric relationships between each pair of tokens. To provide more insights into this prior, we randomly select several tokens and visualize their geometry relationships with other tokens in Fig. 7. For each query token, the geometry prior accurately identifies the object to which it belongs and captures the geometric relationship between this object and its nearby counterparts. The perception of the geometry relationship between objects can help our model better distinguish different semantic objects, for example, chairs are often under the table and people often sit on chairs. The focused attention within our GSA is visualized in Fig. 8. Introducing geometry prior to the self-attention mechanism enables the model to better understand the geometric structures of objects and the position relationship between objects in complex scenes resulting in more accurate segmentation results. Additionally, we visu-

Method	Params	Flops	NYUDepthV2	SUNRGBD
Conv	26.9M	34.3G	55.8	51.3
Addition	26.5M	33.5G	54.6	50.4
Hadamard	26.5M	33.6G	54.9	50.9
Memory	26.7M	33.9G	56.2	51.7

Table 4. Different operations to bridge the depth prior and spatial prior on our small scale model.

Settings	NYUDepthV2	SUNRGBD
fixed to 0.25 for all heads	55.7	51.1
fixed to 0.5 for all heads	55.5	51.0
fixed to 0.75 for all heads	55.7	51.2
linearly sampled in [0.5, 1.0)	55.9	51.5
linearly sampled in [0.75, 1.0)	56.0	51.5

Table 5. Effect of different decay strategies in geometry self-attention on DFormerv2-S.

alize the feature maps with and without the use of geometry priors, as shown in Fig. 9. It can be seen that introducing geometry priors can help our model better capture the object's details and improve the segmentation performance.

Fusion operation. To build the bridge between depth prior and spatial prior and form the geometry prior, we leverage memory weights to form the geometry clues from the depth and spatial distances among all the image patch tokens. To validate the effectiveness of the fusion operation, we also use some other operations to replace it, including addition, Hadamard product, and convolutional layers. As shown in Tab. 4, we can see that the memory weights leads to better results than other operations.

Decay Rate. When introducing geometry prior into the attention mechanism as formulated in Eq. (4), we use a decay rate β to control the extent of the prior's influence on the features. Here, we investigate how the model performance would change when different decay rate strategies are adopted. The results can be found in Tab. 5. It indicates that assigning distinct decay rates β to different heads in our geometry self-attention introduces multi-scale enhancement and more diversity, which further benefit performance. Thus, we sample the value of β in $[0.75, 1.0]$ by default.

Inference Latency. Real-time inference speed is crucial for the practical deployment of RGB-D models across a wide range of downstream applications [5]. Therefore, we evaluate the inference latency of DFormerv2 alongside other methods to assess their real-time potential. To ensure a fair comparison, all tests are performed on the same hardware setup with a single 3090 RTX GPU, and the same image resolution of 480×640. As shown in Tab. 6, DFormerv2 shows a good trade-off between speed and accuracy.

Discussion on the effect of RGB and depth. Semantic segmentation, assigning each pixel with a category label, can be seen as the combination of classification and segmentation of the objects. Here, we explore how the two modalities

Model	Params	FLOPs	Latency↓	NYU DepthV2
Omnivore [16]	29.1M	32.7G	40.1ms	49.7
DFormer-B [59]	29.5M	45.3G	42.8ms	55.6
DFormerv2-S	26.7M	33.9G	43.9ms	56.0
CMX-B2 [61]	66.6M	67.6G	71.5ms	54.4
DFormer-L [59]	39.0M	69.3G	44.5ms	57.2
GeminiFusion-B3 [28]	75.8M	138.2G	68.2ms	56.8
DFormerv2-B	53.9M	67.2G	50.7ms	57.7
CMX-B5 [61]	181.1M	167.8G	114.9ms	56.9
CMNext-B4 [62]	119.6M	131.9G	98.5ms	56.9
MultiMAE [1]	95.2M	267.9G	76.9ms	56.0
GeminiFusion-B5 [28]	137.2M	256.1G	108.7ms	57.7
DFormerv2-L	95.5M	124.1G	79.9ms	58.4

Table 6. Comparison of the inference latency between our method and recent SOTA models. '↓': the lower the better.

Modality	**Params**	Classification	Segmentation	
		Top-1 Acc↑	wF↑	MAE ↓
RGB	26.5M	83.1	0.818	0.054
Depth	26.5M	43.8	0.715	0.061
RGB+Depth	26.7M	83.4	0.868	0.048

Table 7. Effect of different input modalities on capturing semantic categories and object shapes. Weighted F-measure (wF) and mean absolute error (MAE) are two common metrics for the foreground segmentation tasks [29, 58, 66].

contribute to capturing both semantic categories and object shapes, providing deeper insights into the design of our proposed geometry self-attention mechanism. To do so, we perform experiments on the LUSS [14] dataset, which provides segmentation annotations for 50K images from ImageNet [38]. We divide the data into training, validation, and test sets, and train the model for both classification and foreground segmentation tasks. As shown in Tab. 7, the 3D geometry information within depth mainly helps the model segment the objects and slightly helps capture semantics.

5. Conclusions

We propose DFormerv2, an RGBD vision backbone that incorporates an explicit geometry prior. DFormerv2 leverages depth to model the geometry relationship between image patches and then uses this prior to allocate the weights of attention within the self-attention mechanism, called geometry self-attention. Thanks to this tailored attention mechanism, our method achieves a more effective utilization of the depth modality. Experiments show that DFormerv2 produces better results than recent methods in RGB-D semantic segmentation with far less computational cost.

Acknowledgment. This work was funded by NSFC (No. 62225604, 62176130), the Science and Technology Support Program of Tianjin, China (No. 23JCZDJC01050), and the Shenzhen Science and Technology Program (JCYJ20240813114237048). The Supercomputing Center of Nankai University partially supported computations.

References

[1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. In *ECCV*, 2022. 1, 5, 6, 8

[2] Tim Broedermann, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Hrfuser: A multi-resolution sensor fusion architecture for 2d object detection. In *IEEE ITSC*, 2023. 7

[3] Jinming Cao, Hanchao Leng, Dani Lischinski, Danny Cohen-Or, Changhe Tu, and Yangyan Li. ShapeConv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation. In *ICCV*, 2021. 5

[4] Lin-Zhuo Chen, Zheng Lin, Ziqin Wang, Yong-Liang Yang, and Ming-Ming Cheng. Spatial information guided convolution for real-time RGBD semantic segmentation. *TIP*, 30: 2313–2324, 2021. 5

[5] Xiaokang Chen et al. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In *ECCV*, 2020. 5, 8

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 2

[7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2

[8] Senlin Cheng and Haopeng Sun. Spt: Sequence prompt transformer for interactive image segmentation. *arXiv preprint arXiv:2412.10224*, 2024. 3

[9] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *CVPR*, 2022. 2, 3, 5

[10] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 3

[11] Siqi Du, Weixi Wang, Renzhong Guo, Ruisheng Wang, and Shengjun Tang. Asymformer: Asymmetrical cross-modal representation learning for mobile platform real-time rgb-d semantic segmentation. In *CVPRW*, 2024. 2, 5

[12] Qihang Fan, Huaibo Huang, Mingrui Chen, Hongmin Liu, and Ran He. Rmt: Retentive networks meet vision transformers. In *CVPR*, 2024. 4, 5

[13] Huan Gao, Jichang Guo, Guoli Wang, and Qian Zhang. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In *CVPR*, 2022. 2

[14] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE TPAMI*, 45(6):7457–7476, 2022. 8

[15] Zhengyang Geng, Meng-Hao Guo, Hongxu Chen, Xia Li, Ke Wei, and Zhouchen Lin. Is attention better than matrix decomposition? *NeurIPS*, 2021. 6

[16] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *CVPR*, 2022. 5, 8

[17] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In *CVPR*, 2022. 3

[18] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *NeurIPS*, 2022. 2

[19] Dongchen Han, Tianzhu Ye, Yizeng Han, Zhuofan Xia, Siyuan Pan, Pengfei Wan, Shiji Song, and Gao Huang. Agent attention: On the integration of softmax and linear attention. In *ECCV*, 2025. 3

[20] Zhiwei Hao, Zhongyu Xiao, Yong Luo, Jianyuan Guo, Jing Wang, Li Shen, and Han Hu. Primkd: Primary modality guided multimodal fusion for rgb-d semantic segmentation. In *ACM MM*, 2024. 2, 6

[21] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *CVPR*, 2023. 5

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[24] Qibin Hou, Cheng-Ze Lu, Ming-Ming Cheng, and Jiashi Feng. Conv2former: A simple transformer-style convnet for visual recognition. *IEEE TPAMI*, 2024. 3

[25] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. ACNet: Attention based network to exploit complementary features for RGBD semantic segmentation. In *ICIP*, 2019. 3, 5

[26] Keli Huang, Botian Shi, Xiang Li, Xin Li, Siyuan Huang, and Yikang Li. Multi-modal sensor fusion for auto driving perception: A survey. *arXiv preprint arXiv:2202.02703*, 2022. 1

[27] Wei Ji, Jingjing Li, Qi Bi, Tingwei Liu, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *MIR*, 2024. 2

[28] Ding Jia, Jianyuan Guo, Kai Han, Han Wu, Chao Zhang, Chang Xu, and Xinghao Chen. Geminifusion: Efficient pixel-wise multimodal fusion for vision transformer. In *ICML*, 2024. 1, 2, 3, 5, 6, 7, 8

[29] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013. 8

[30] Peng-Tao Jiang, Yuqi Yang, Yang Cao, Qibin Hou, Ming-Ming Cheng, and Chunhua Shen. Traffic scene parsing through the tsp6k dataset. In *CVPR*, 2024. 1

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[32] Zheng Lin, Zhao Zhang, Zi-Yue Zhu, Deng-Ping Fan, and Xia-Lei Liu. Sequential interactive image segmentation. *CVM*, pages 753–765, 2023. 2

[33] Chang Liu, Xudong Jiang, and Henghui Ding. Primitivenet: decomposing the global constraints for referring segmentation. *Visual Intelligence*, 2024. 2

[34] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 3, 6

[35] Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2, 3, 4

[36] Nicolas Marchal, Charlotte Moraldo, Hermann Blum, Roland Siegwart, Cesar Cadena, and Abel Gawel. Learning densities in feature space for reliable segmentation of indoor scenes. *RA-L*, 5(2):1032–1038, 2020. 1

[37] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *ICLR*, 2022. 3

[38] Olga Russakovsky et al. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 8

[39] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *ICRA*, 2021. 3, 5

[40] Daniel Seichter, Söhnke Benedikt Fischedick, Mona Köhler, and Horst-Michael Groß. Efficient multi-task rgb-d scene analysis for indoor environments. In *IJCNN*, 2022. 5

[41] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *NAACL*, 2018. 4

[42] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. 2, 5, 6

[43] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, 2015. 2, 5, 6

[44] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 2

[45] Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *CVPR*, 2024. 2

[46] Haopeng Sun. Ultra-high resolution segmentation via boundary-enhanced patch-merging transformer. *AAAI*, 2025. 3

[47] Haopeng Sun, Lumin Xu, Sheng Jin, Ping Luo, Chen Qian, and Wentao Liu. Program: Prototype graph model based pseudo-label learning for test-time adaptation. In *ICLR*, 2024. 2

[48] Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023. 3, 4

[49] Lichun Tang, Zhaoxia Yin, Hang Su, Wanli Lyu, and Bin Luo. Wfss: weighted fusion of spectral transformer and spatial self-attention for robust hyperspectral image classification against adversarial attacks. *Visual Intelligence*, 2024. 3

[50] Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam. In *CVPR*, 2023. 1

[51] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *CVM*, 8(3):415–424, 2022. 3

[52] W Wang, L Yao, L Chen, B Lin, D Cai, X He, and W Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. In *ICLR*, 2022. 2

[53] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *NeurIPS*, 2020. 5

[54] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *CVPR*, 2022. 3, 5, 7

[55] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *CVPR*, 2022. 3

[56] Qi Xu, Yinan Ma, Jing Wu, Chengnian Long, and Xiaolin Huang. Cdada: A curriculum domain adaptation for nighttime semantic segmentation. In *ICCV*, 2021. 2

[57] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021. 2, 5

[58] Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. Camoformer: Masked separable attention for camouflaged object detection. *IEEE TPAMI*, 2024. 8

[59] Bowen Yin, Xuying Zhang, Zhongyu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. *ICLR*, 2024. 1, 2, 3, 5, 6, 8

[60] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *CVPR*, 2022. 3

[61] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *T-ITS*, 2023. 1, 3, 5, 6, 7, 8

[62] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *CVPR*, 2023. 1, 2, 5, 6, 7, 8

[63] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, 2019. 2

[64] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 2

[65] Hao Zhou, Lu Qi, Hai Huang, Xu Yang, Zhaoliang Wan, and Xianglong Wen. Canet: Co-attention network for rgb-d semantic segmentation. *PR*, 124:108468, 2022. 2

[66] Tao Zhou, Deng-Ping Fan, Geng Chen, Yi Zhou, and Huazhu Fu. Specificity-preserving rgb-d saliency detection. *CVM*, pages 297–317, 2023. 8

[67] Wujie Zhou, Enquan Yang, Jingsheng Lei, Jian Wan, and Lu Yu. Pgdenet: Progressive guided fusion and depth enhancement network for rgb-d indoor scene parsing. *IEEE TMM*, 2022. 5

[68] Wujie Zhou, Enquan Yang, Jingsheng Lei, and Lu Yu. Frnet: Feature reconstruction network for rgb-d indoor scene parsing. *JSTSP*, 16(4):677–687, 2022. 5

[69] Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson WH Lau. Biformer: Vision transformer with bi-level routing attention. In *CVPR*, 2023. 3