Transformer fusion for indoor RGB-D semantic segmentation

Zongwei Wu, Zhuyun Zhou, Guillaume Allibert, Christophe Stolz, Cédric Demonceaux, Chao Ma

Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Transformer Fusion for Indoor RGB-D Semantic Segmentation

Zongwei Wu[a], Zhuyun Zhou[a], Guillaume Allibert[b], Christophe Stolz[c], Cédric Demonceaux[a], Chao Ma[d,**]

[a]*ICB, UMR CNRS 6303, University of Burgundy, Dijon, France*
[b]*Université Côte d'Azur, CNRS, I3S, Nice, France*
[c]*ImViA, University of Burgundy, Dijon, France*
[d]*MOE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China*

ABSTRACT

Fusing geometric cues with visual appearance is an imperative theme for RGB-D indoor semantic segmentation. Existing methods commonly adopt convolutional modules to aggregate multi-modal features, paying little attention to explicitly leveraging the long-range dependencies in feature fusion. Therefore, it is challenging for existing methods to accurately segment objects with large-scale variations. In this paper, we propose a novel transformer-based fusion scheme, named TransD-Fusion, to better model contextualized awareness. Specifically, TransD-Fusion consists of a self-refinement module, a calibration scheme with cross-interaction, and a depth-guided fusion. The objective is to first improve modality-specific features with self- and cross-attention, and then explore the geometric cues to better segment objects sharing a similar visual appearance. Additionally, our transformer fusion benefits from a semantic-aware position encoding which spatially constrains the attention to neighboring pixels. Extensive experiments on RGB-D benchmarks demonstrate that the proposed method performs well over the state-of-the-art methods by large margins.

1. Introduction

Recent developments in depth sensors provide geometric information at a low cost. Since the depth information along with images can naturally contribute to scene understanding, RGB-D semantic segmentation has drawn increasing attention Wang and Neumann (2018); Wu et al. (2020); Zhou et al. (2022a); Wang et al. (2022a).

When merging the depth cues and images, three typical challenges arise: (1) Multi-modal fusion. RGB input contains rich information on visual changes, while depth images are sensitive to occluded boundaries. How to extract, preserve, and fuse these modality-specific features is as yet an open issue for RGB-D semantic segmentation. (2) Noisy response in each modality. On the one hand, the similar visual appearance between neighboring objects can adversely affect the model's discriminability. On the other hand, the depth quality may be influenced by environmental factors during acquisition, such as object distances, as discussed in previous works Chen et al. (2020); Fan et al. (2021); Ji et al. (2021). (3) Feature alignment. As shown in Fig. 1(3), current fusion approaches assume that the sensor calibration is precise and different modalities are accurately aligned at the pixel level, which is not always the case in practice. Despite the recent advances Wang and Neumann (2018); Hu et al. (2019); Chen et al. (2020, 2021a), we observe that most existing works are still based on pixel-wise fusion, whose limited awareness of contextualized cues causes the main performance bottleneck.

**Corresponding author.
e-mail: chaoma@sjtu.edu.cn (Chao Ma)

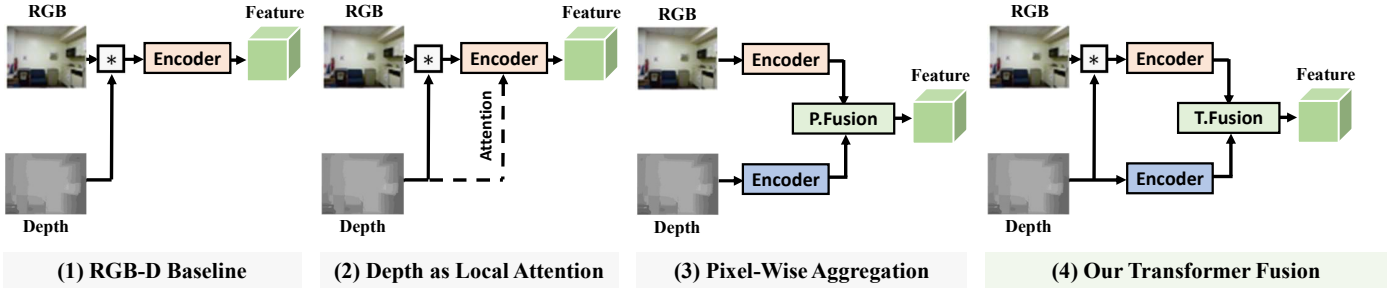

Fig. 1: **Comparison of different RGB-D fusion strategies.** (1) Conventional RGB-D early fusion schemes. (2) Previous attempts to improve the RGB-D learning with local depth awareness Wang and Neumann (2018); Wu et al. (2020). (3) Pipeline of most existing two-stream networks with pixel-wise feature fusion strategies Hu et al. (2019); Chen et al. (2020). **P.** stands for **Pixel-Wise Correlation**. (4) Our transformer fusion which explores contextualized geometric cues to better deal with objects sharing a similar visual appearance. **T.** stands for **Transformer Fusion**.

Recently, transformer has shown its capability in modeling long-range dependencies in various vision tasks Dosovitskiy et al. (2021); Liu et al. (2021b); Chen et al. (2021b); Zhu et al. (2021). Compared to convolution, transformer is built upon global attention with inter key-query correlation. We observe that by extending the inter key-query correlation to cross-modal key-query correlation, transformer attention suggests a natural way to aggregate RGB-D features. Inspired by this observation, we propose to first extract both mixed RGB-D and modality-specific depth features. Then we leverage the depth cues to retrieve geometric information from mixed RGB-D features. As shown in Fig. 1(4), the key idea is to leverage contextualized transformer attention to improve the early fusion with enhanced awareness of depth cues. As such, we can better deal with objects sharing a similar visual appearance but at different camera distances or with occlusion, which is challenging for indoor semantic segmentation.

Specifically, our transformer fusion with geometric cues, termed TransD-Fusion, consists of three parts: a self-enhancement module, a bi-directional cross-calibration module, and a depth-guided query design. The enhancement module is realized through the vanilla transformer self-attention. The bi-directional calibration module aims to refine each modality with complementary information: for the depth image, we expect to suppress unsatisfactory responses due to measurement bias; while for the RGB image, we expect to strengthen the edge awareness on neighboring objects with a similar visual appearance. Finally, the depth-guided query strategy ensures the effective segmentation of objects with strengthened discriminability.

To enable position awareness and leverage locality into our TransD-Fusion, we propose a semantic-aware position encoding generator (S-PE) built upon convolutions. It takes a modality-specific sequence as input and generates a category-aware position encoding. We expect our encoding to spatially constrain the attention around the neighboring area to better segment objects. Moreover, our positional embedding can be learned from hierarchical features, yielding a simple yet efficient encoding for RGB-D fusion. Finally, to tackle the limitations of CNN-based backbones, we implement our TransD-Fusion on Swin-Transformer Liu et al. (2021b) to better model contextualized dependencies. In brief, our contributions are summarized as follows:

- We propose a novel transformer-based multi-modal fusion to replace the existing pixel-wise fusion modules for RGB-D semantic segmentation.
- We design a semantic-aware position encoding (S-PE) scheme to improve our transformer fusion. The S-PE is dynamically generated from a modality-specific sequence of tokens by a convolutional layer, yielding a spatial constraint on neighboring features for accurate segmentation.
- Our proposed network performs favorably over the state-of-the-art methods on large-scale benchmark datasets by large margins.

2. Related Work

2.1. RGB-D Semantic Segmentation

How to deal with complementary depth is a key research topic for RGB-D semantic segmentation. At an early stage, Gupta et al. (2014) proposes to explore the geometric cues by transforming the depth map into an HHA image. Afterward, researchers take RGB-HHA as input and design various fusion strategies. Several preliminary works Gupta et al. (2014); Hazirbas et al. (2016); Wang et al. (2016) fuse the RGB-D images from the input side, treating depth/HHA as additional channels. D-CNN Wang and Neumann (2018) further proposes a depth-aware re-calibration weight to strengthen the discriminatory power during feature modeling. Since then, networks with early-fused RGB-HHA have shown great advances with different forms of weight functions Chen et al. (2019); Xing et al. (2019, 2020). However, the proposed depth-aware operations are sensitive to depth noise, which might be the performance bottleneck while dealing with unsatisfactory geometry. Recent works Shen and Stamos (2021, 2020) have shown the great potential of DHS representations which can serve as an alternative to HHA for 2D/3D object detection and instance segmentation. However, DHS plays the role of a pseudo-3D representation and requires processing with a 3D network, which demands more computational cost compared to HHA.

To address this issue, several works propose to re-calibrate feature representation with the attention modules. ACNet Hu et al. (2019) adopts a self-enhancement module with channel attention Hu et al. (2018). Sharing the same idea, ShapeConv Cao et al. (2021) directly integrates the channel attention into the convolution function. An alternative is a channel exchanging strategy proposed by Wang et al. (2022b, 2020). SA gate Chen et al. (2020) further leverages spatial attention Woo et al. (2018) to calibrate each modality. Another group of works proposes to enhance feature representation with long-range attention. Li et al. (2016) introduces ConvLSTM models in RGB-D fusion to better model contextualized cues. VCD Xiong et al. (2020) introduces a learned Gaussian convolution kernel to improve spatial-context awareness. Several works Chen et al. (2021a); Wu et al. (2020, 2022b) integrate depth cues with the deformable convolution Dai et al. (2017) to create a more malleable receptive field. Despite the popularity of local-global attention in RGB-D semantic segmentation Li et al. (2016); Chen et al. (2021a); Wu et al. (2020, 2022b); Xiong et al. (2020); Zhang et al. (2021); Su et al. (2021), the capability of modeling long-range dependencies is still limited due to convolution-based feature extraction and fusion. Furthermore, one basic assumption for existing approaches is that the RGB and depth maps are perfectly aligned at the pixel level, which is not always the case in practice due to sensor calibration errors. To tackle these dilemmas, we propose a transformer-based aggregation scheme to explicitly leverage contextualized awareness in multi-modal feature fusion. The concurrent RGB-D models Zhou et al. (2020, 2022a) only leverage single-head non-local attention, while our transformer attention is with multiple heads.

2.2. Transformer Fusion

There are extensive surveys Tay et al. (2020); Han et al. (2022); Khan et al. (2021) of transformers applied in vision tasks. ViT and its successors Dosovitskiy et al. (2021); Radford et al. (2021); Liu et al. (2021b) explore the transformer on feature modeling. DERT and its successors Carion et al. (2020); Gao et al. (2021); Zhu et al. (2021) adopt a transformer on the detection head. In recent works, researchers explore transformer attention to compute the correlation between the different target and source data. For example, Yan et al. (2021); Chen et al. (2021b); Wang et al. (2021) adopt transformer to analyze the similarity between the search image and template image. Yang et al. (2021) shares a similar idea but tackles cross-domain adaption. Yew and Lee (2022) shows that transformer attention can also be used to find the correspondences between two sets of point clouds. In addition to modality-specific tasks, the transformer can also be useful for multi-modal applications. Wang et al. (2022a) realizes the aggregate RGB and depth cues at the token level. Prakash et al. (2021); Song et al. (2021) suggest firstly merging multi-modal features and then leveraging the self-attention to improve the feature representation. An-

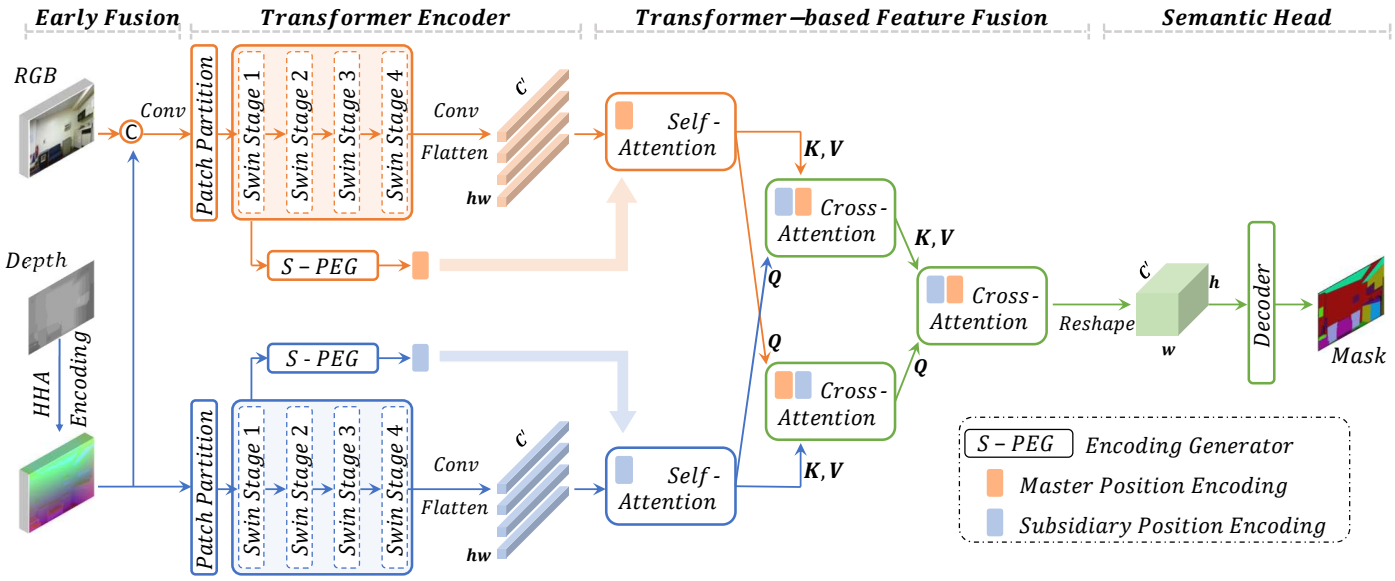

Fig. 2: **Overview.** Details of each module are presented in Section 3.3. Our TransD-Fusion leverages transformer attention to aggregate multi-modal features. The self-attention aims to refine modality-specific features, while the cross-attention makes full use of cross-domain cues to first calibrate and then combine multi-modal information. The transformer fusion benefits from dynamically generated position encodings to constrain the attention around category-aware neighboring pixels.

other work on saliency detection Liu et al. (2021a) adopts the transformer as a dimension regulator to convert the sequence of tokens from the encoder space to the decoder space.

Differently, our model aims to explore multi-modal cues for feature aggregation. We make full use of both self- and cross-attention modules to explicitly preserve, calibrate, and fuse multi-modal information. We show through ablation study that our fusion design performs favorably over other transformer fusion alternatives.

It is also worth noting that attention modules cannot capture order awareness of input tokens. Hence, various research on position encoding (PE) has been conducted to address this issue. In the literature, two main groups of solutions are proposed: absolute PE and relative PE. Absolute PE generates a unique encoding vector for each position, e.g., 2D sinusoidal embeddings Gehring et al. (2017); Vaswani et al. (2017), while relative PE proposes to focus on the relative distance of the elements Shaw et al. (2018); Bello et al. (2019); Yang et al. (2019). In vision tasks, previous studies Dosovitskiy et al. (2021); Liu et al. (2021b); Zhu et al. (2021); Wu et al. (2021); He et al. (2021) have shown that the relative position enables better performance on the image classification task, while the absolute encoding is more suitable for object detection where the pixel position plays a vital role in segmenting and locating objects. CPVT Chu et al. (2021) proposes a conditional PE to leverage the local awareness through a single 2D convolution to improve ViT. However, extending such an idea to RGB-D feature fusion at the semantic level is non-trivial due to the limited feature resolution. In contrast, we propose a modality-dependant and semantic-aware PE to improve our transformer fusion with a better position and category awareness.

3. Our Approach: TransD-Fusion

3.1. Overview

Fig.2 presents the overall framework of our network which is composed of a master-subsidiary two-stream encoder and our proposed transformer feature fusion (TransD-Fusion). The master network is an encoder-decoder pipeline with early-fused RGB-HHA images. The encoder stage takes the transformer backbone to extract features from concatenated RGB-HHA input, while the decoder stage takes the classical convolutional head to output the semantic map. The subsidiary network takes HHA images as input. It processes depth features and aims to enhance the master network with geometric cues via our TransD-Fusion. Details are presented in the following sections.

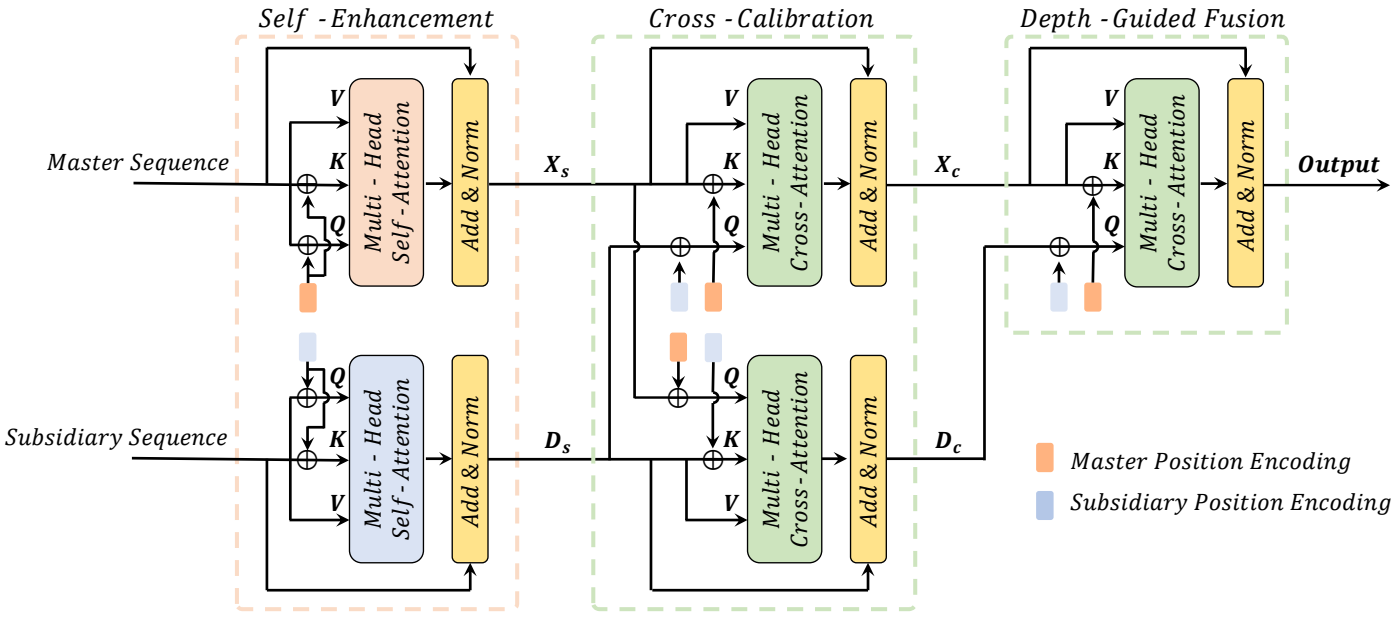

Fig. 3: **Details** of our proposed feature enhancement, calibration, and fusion scheme with transformer attention. Best viewed in color.

3.2. Master-Subsidiary Network

Early fusion has been widely exploited in RGB-D semantic segmentation Wang and Neumann (2018); Chen et al. (2019); Xing et al. (2019, 2020). It promotes the geometric constraint in the visual appearance from the input side. Nevertheless, the inflexibility of further analysis of multi-modal features at the semantic level severely limits the model performance. To address this issue, we design a master network with early-fused input and a subsidiary stream to enable high-level manipulation with transformer fusion.

Given the RGB image $I \in \mathbb{R}^{3\times H\times W}$ and the geometric feature HHA map $D \in \mathbb{R}^{3\times H\times W}$, we can obtain the *master* feature $X \in \mathbb{R}^{3\times H\times W}$:

$$X = Conv_{1\times 1}([I, D]), \tag{1}$$

where [] denotes the concatenation along the channel dimension. In such a way, the master feature contains both photometric and geometric information and fits the input shape of the transformer backbone.

To extract multi-modal features, X and D are firstly fed into the patch partition to form two sequences of tokens separately, and then fed into the Swin-Transformer Liu et al. (2021b) encoders. A Swin-Transformer layer contains window-based multi-head self-attention (W-MSA), shifted window partitioning configurations (SW-MSA), and a point-wise multi-layer perceptron (MLP) with layer norm (LN). For the i^{th} layer, $i \in \{1,...,L\}$, it takes the sequence z_{i-1} as input, and outputs the new sequence z_{i+1}:

$$\begin{aligned}
\hat{z}_i &= W\text{-}MSA(LN(z_{i-1})) + z_{i-1};\\
z_i &= MLP(LN(\hat{z}_i)) + \hat{z}_i;\\
\hat{z}_{i+1} &= SW\text{-}MSA(LN(z_i)) + z_i;\\
z_{i+1} &= MLP(LN(\hat{z}_{i+1})) + \hat{z}_{i+1}.
\end{aligned} \tag{2}$$

Compared to CNN backbones Simonyan and Zisserman (2015); He et al. (2016), transformer encoders Vaswani et al. (2017); Liu et al. (2021b); Zhao et al. (2017) can better model long-range features. Furthermore, we particularly build upon Swin-Transformer Liu et al. (2021b) with window attention which reduces the computational complexity. We refer readers to the original paper Liu et al. (2021b) for more details.

3.3. Transformer feature fusion

Given two sequences of tokens $f_X \in \mathbb{R}^{c\times h\times w}$ and $f_D \in \mathbb{R}^{c\times h\times w}$ from different streams, we first apply convolutions to f_X and f_D, and output two new feature maps. We expect to strengthen local awareness and/or reduce the channel size from c to c'. These two new feature maps are further flattened in

the spatial dimension, obtaining $f_x \in \mathbb{R}^{c' \times hw}$ and $f_d \in \mathbb{R}^{c' \times hw}$. These flattened features are the inputs of our transformer fusion.

As shown in Fig. 3, we propose a three-stage fusion scheme. Firstly, the modality-specific features are enhanced through self-attention. Secondly, a bi-directional calibration is applied with cross-attention. Finally, we initialize a geometry-guided query scheme to accurately segment objects. The attention module is equipped with learnable position encoding to enable both local and semantic awareness. In the following paragraphs, we introduce the details of each component. The benefit of each component can be found in the ablation study Section 5.3 Table 9.

3.3.1. Multi-Head Attention in Transformer.

The attention mechanism is the key component of our TransD-Fusion. Given an input sequence of tokens, it is firstly flattened to a 1D vector and generates three intermediate representations: queries Q, keys K, and values V. The attention is formulated as follows:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{3}$$

where d_k is the scaling factor. Vaswani et al. (2017) shows that multi-head attention with h heads can further contribute to the model performance by paying diverse attention to features from different positions. The multi-head attention is formulated as follows:

$$\begin{aligned} MultiHead(Q,K,V) &= Concat(head_1,...,head_h)W^O \\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{4}$$

where W^O, W_i^Q, W_i^K, W_i^V are the projection matrices.

3.3.2. Self-Enhancement.

While (Q,K,V) are from the same input modality, the attention module becomes multi-head self-attention which can be considered as a self-enhancement. It analyzes long-range dependencies and explores contextual information to further improve the modality-specific features. Taking flattened global feature f_x as an example, the self-enhanced global feature X_s can be formulated as:

$$X_s = f_x + MultiHead(Q_x + P_x, V_x + P_x, K_x), \tag{5}$$

where (Q_x, K_x, V_x) are the associated intermediate representations and P_x is the associated position encoding. Similarly, we can obtain the self-enhanced geometric feature D_s with the associated position encoding P_d.

3.3.3. Cross-Calibration.

The objective of cross-calibration is to reduce the ambiguity in a single modality, e.g., the limited awareness of the geometric cues in visual appearance and measurement bias in geometric features. Different from previous dual attention Woo et al. (2018); Chen et al. (2020), our cross-calibration is based on transformer attention. We take the queries from one input feature, e.g., Q_{D_s}, to compute the correlation with the keys from the other modality, e.g., K_{X_s}. Formally, we have:

$$\begin{aligned} X_c &= X_s + MultiHead(Q_{D_s} + P_d, K_{X_s} + P_x, V_{X_s}), \\ D_c &= D_s + MultiHead(Q_{X_s} + P_d, K_{D_s} + P_x, V_{D_s}), \end{aligned} \tag{6}$$

where (X_s, D_s) are the outputs of the self-enhancement module, $(Q_{X_s}, K_{X_s}, V_{X_s})$ are the associated intermediate representations for master feature X_s, and $(Q_{D_s}, K_{D_s}, V_{D_s})$ for subsidiary feature D_s. We use the same position encodings (P_x, P_d) as in the previous self-enhancement module.

3.3.4. Depth-Guided Fusion.

To combine master and subsidiary streams, similar to cross-calibration, we use the geometry stream to initialize the query strategy. We have:

$$Output = X_c + MultiHead(Q_{D_c} + P_d, K_{X_c} + P_x, V_{X_c}) \tag{7}$$

where (X_c, D_c) are the outputs of the cross-calibration module, in which the same position encodings (P_x, P_d) are used. The depth-guided fusion module contributes to dealing with objects sharing similar appearances.

3.4. Semantic-Aware Position Encoding

We propose a novel position encoding to equip our transformer attention. Specifically, for each modality, we dynamically generate the position encoding from a lower-dimensional

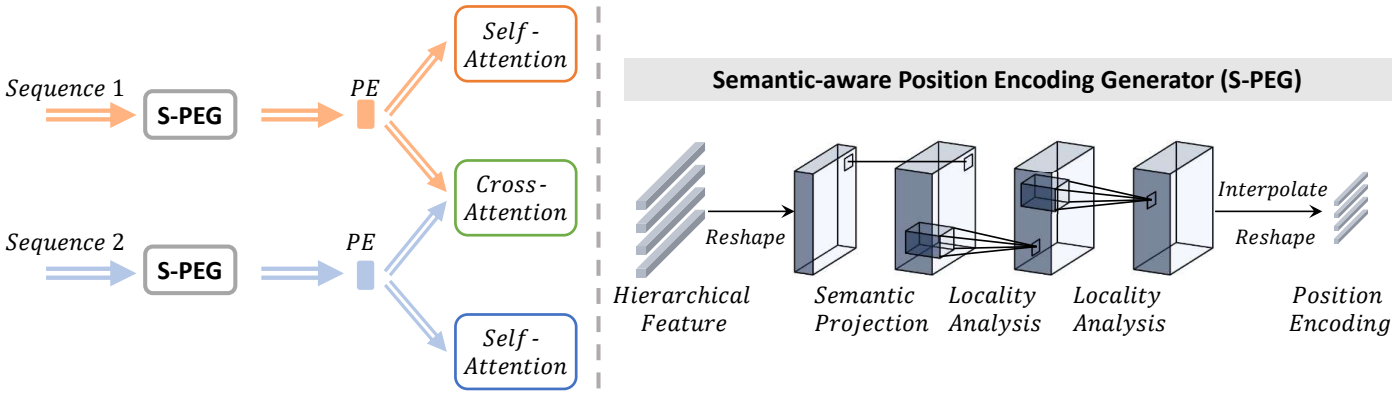

Fig. 4: **Our semantic-aware position encoding (S-PE).** Left: position encoding flows. Right: illustration of encoding generator. Best viewed in color.

feature map with a larger resolution to make full benefits of spatial information, i.e., the output of the first stage of the encoder.

As illustrated in Fig. 4, given the two sequences with higher resolution, we first project the input sequence into a high-dimensional feature space through semantic projector $\mathcal{P}$. Then, we utilize two convolutional modules $\mathcal{F}$ to strengthen the local awareness of the input sequence. Each module consists of 3×3 convolution, batch normalization, and ReLU activation.

Different from previous works Vaswani et al. (2017); Dosovitskiy et al. (2021); Liu et al. (2021b), PE plays a more important role in our transformer since it should leverage modality-specific cues for RGB-D segmentation. Hence, we propose to learn the PE from each input sequence with convolution, yielding a strengthened locality awareness and becoming category-dependant. The local-aware encoding can also be implemented by the CPVT Chu et al. (2021). However, CPVT uses a simple 2D convolution that takes the sequence $X \in \mathbb{R}^{C \times H \times W}$ as input and generates the position encoding $E \in \mathbb{R}^{C \times H \times W}$ which has the same resolution as input sequence X. Compared to CPVT, one main difference is that our S-PE can be learned from hierarchical features with higher resolutions to fully excavate the spatial cues on the token order. Empirical comparisons can be found in the ablation study Section 5.4 Table 8.

3.5. Architecture

We follow Zhou et al. (2020); Fu et al. (2019) and apply our transformer fusion on the highest-dimensional features where the resolution is minimized. To generate the output semantic map, we adopt the classical DeeplabV3+ Chen et al. (2017a) architecture. The whole training process is supervised by conventional cross-entropy.

In our model, we adopt early fusion together with late fusion. The objective is to fully leverage the depth cues at both the geometric and semantic levels. The idea of using HHA cues to guide RGB-D learning has been widely used in previous RGB-D works, such as DCNN Wang and Neumann (2018), 2.5D Xing et al. (2019), Malleable Xing et al. (2020), DACN Wu et al. (2020), etc. The main difference is that previous works compute local attention (depth weight/offset) from the depth and embed them in convolution, while we explicitly leverage the contextualized awareness to better deal with feature misalignment.

Our fusion strategy substantially differs from the recent fusion works. Specifically, CCFFNet Wu et al. (2022a) adopts spatial and channel attention on features, while our work is fully based on contextualized attention with tokens. Compared to DeepFusion Li et al. (2022), our cross-modal interaction is bi-directional, while DeepFusion is single-directional (Lidar to camera). Finally, compared to CPVT Chu et al. (2021), our positional embedding can better leverage both hierarchical and semantic cues, yielding a simple yet efficient encoding for RGB-D fusion as shown in the ablation study.

4. Experiments

We evaluate our model on three benchmark RGB-D datasets, i.e., NYUv2 Silberman et al. (2012), SUN-RGBD Song et al.

(2015), and Stanford 2D-3D-Semantic Indoor Dataset (SID) Armeni et al. (2017). We analyze the performance with common metrics, i.e., Pixel Accuracy (PixelAcc), Mean Accuracy (mAcc.), Mean Region Intersection Over Union (mIoU), and Frequency Weighted Intersection Over Union (f.w.IoU). Let s_i be the number of pixels with the ground truth class i. n_{ij} denotes the number of pixels with ground truth class i and but predicted as class j. N_c denotes the number of total classes, and $s = \sum_i s_i$ is the number of all pixels. Mathematically, the metrics are defined by:

- Pixel Acc: $PixelAcc = \sum_i \frac{n_{ii}}{s}$
- mean Acc: $mAcc = \frac{1}{N_c} \sum_i \frac{n_{ii}}{s}$
- mean Intersection over Union: $mIoU = \frac{1}{N_c} \sum_i \frac{n_{ii}}{s_i + \sum_j n_{ji} - n_{ii}}$
- Frequency Weighted Intersection over Union: $f.w.IoU = \frac{1}{s} \sum_i s_i \frac{n_{ii}}{s_i + \sum_j n_{ji} - n_{ii}}$

While our TransD-Fusion is based on transformer attention, we do not require any additional training samples with our proposed approach. We follow the conventional training/testing split which has been widely used for both CNN-based Cao et al. (2021); Hu et al. (2019); Zhou et al. (2022a) and transformer-based methods Girdhar et al. (2022); Liu et al. (2022); Zhang et al. (2023). Specifically, on NYUv2 with 40 categories, we follow the widely-used split with 795 images used for training, and the rest 654 images are for testing among the 40 classes. On SUN-RGBD with 37 categories, we follow the widely-used split with 5,285 images for training and the rest 5,050 images for testing. On SID with 13 categories, we train our model on areas 1, 2, 3, 4, and 6, and Area 5 is for testing. During training, we resize the images to a random ratio between 0.5 and 2.0 and explore left-right flipped images. We choose the standard SGD optimizer with momentum to train our model following the "poly" learning rate policy. The initial learning rate is set to 0.007, the momentum is fixed to 0.9, and the weight decay is set to 0.0001. For inference, we evaluate our model with multi-scale testing strategies, i.e., $\{0.5, 0.75, 1.0, 1.25, 1.5, 1.75\}$. Similar to previous works Gupta et al. (2014); Wang and Neumann (2018); Chen et al. (2020); Cao et al. (2021), we take RGB and HHA images as input. The HHA maps are generated according to Gupta et al. (2014) during pre-processing. To make a fair comparison, our transformer backbone is initialized with the weights pre-trained on ImageNet-1K Deng et al. (2009) as CNN backbones.

4.1. Quantitative Comparison

Table 1 illustrates the quantitative comparison on NYUv2. We observe that the models with transformer encoders Girdhar et al. (2022); Wang et al. (2022a) outperform CNN approaches. Our TransD-Fusion even surpasses transformer counterparts on mIoU and sets a new state-of-the-art record, i.e., 55.5% with 1.7 FPS. We also report the performance of the SUN-RGBD dataset and SID dataset. Our TransD-Fusion (Swin-B) outperforms the concurrent ShapeConv Cao et al. (2021) which is also built upon DeepLabV3+ with a large margin: 1.4% $\uparrow$ mIoU on SUN-RGBD and 1.6% $\uparrow$ mIoU on SID. The leading performances on indoor benchmarks validate our effectiveness.

4.2. Qualitative Comparison

Fig. 5 illustrates semantic maps generated by the SOTA CNN model ShapeConv Cao et al. (2021), transformer baseline (with DeeplabV3+ Chen et al. (2017a)), and our TransD-Fusion. Compared to ShapeConv, we observe that transformer models can better generate contextualized features and yield results closer to the ground truth. Compared to the transformer baseline, TransD-Fusion can further explore geometric cues to distinguish objects sharing similar visual appearances, leading to a more accurate semantic segmentation.

We can also observe that with the help of our TransD-Fusion, our network can better deal with large variations, i.e., significant differences in scale and appearance between different objects or regions in a scene, such as a stove and a sofa. In these challenging cases, a conventional transformer-based model fails to accurately capture the context clues. One of the performance bottlenecks is mainly due to the lack of a step-by-step cross-modal fusion design. Differently, we carefully design the proposed TransD-Fusion with both self- and cross-attention blocks, as well as the semantic-aware positional encodings, our

Table 1: **Quantitative comparison** on RGB-D benchmark datasets.

Source	Method	Backbone	PixelAcc	mAcc	mIoU	f.w.IoU
Comparison on NYUv2 datasets						
ECCV'20	*Malleable* Xing et al. (2020)	ResNet-101	76.9	-	50.9	-
ECCV'20	*SAGate* Chen et al. (2020)	ResNet-50	77.9	-	52.4	-
SPL'21	*RTLNet* Yue et al. (2021)	ResNet-50	77.2	-	53.1	-
TIP'21	*SGNet* Chen et al. (2021a)	ResNet-101	76.8	63.1	51.1	-
ICRA'21	*ESAnet* Seichter et al. (2021)	ResNet-34	-	-	51.6	-
CVPR'21	*InverseForm* Borse et al. (2021)	ResNet-101	78.1	-	53.1	-
ICCV'21	*ShapeConv* Cao et al. (2021)	ResNext-101	76.4	63.5	51.3	63.0
PR'22	*CANet* Zhou et al. (2022a)	ResNet-101	77.1	64.6	51.5	-
TMM'22	*PGDENet* Zhou et al. (2022c)	ResNet-34	78.1	66.7	53.7	-
TMM'22	*TET* Zhang et al. (2022)	ResNet-50	77.3	59.7	51.8	-
CVPR'22	*Omnivore* Girdhar et al. (2022)	Swin-B	-	-	54.0	-
CVPR'22	*TokenFusion* Wang et al. (2022a)	SegFormer	**79.0**	66.9	54.2	-
TransD-Fusion (Ours)		Swin-B	78.5	**69.4**	**55.5**	**66.3**
Comparison on SUN-RGBD datasets						
ECCV'18	*DCNN* Wang and Neumann (2018)	VGG-16	-	53.5	42.0	-
ICIP'19	2.5*D* Xing et al. (2019)	ResNet-101	82.4	-	48.2	-
ACCV'20	*CANet* Zhou et al. (2020)	ResNet-101	81.9	-	47.7	-
SPL'21	*RTLNet* Yue et al. (2021)	ResNet-50	81.3	-	45.7	-
ICRA'21	*ESAnet* Seichter et al. (2021)	ResNet-50	-	-	48.3	-
TIP'21	*SGNet* Chen et al. (2021a)	ResNet-101	82.0	60.7	48.6	-
ICCV'21	*ShapeConv* Cao et al. (2021)	ResNet-101	82.2	59.2	48.6	71.3
TETCI'22	*RFNet* Zhou et al. (2022b)	ResNet-34	87.3	59.0	50.7	-
JSTSP'22	*FRNet* Zhou et al. (2022d)	ResNet-34	**87.4**	62.2	51.8	-
TransD-Fusion (Ours)		Swin-B	83.2	**64.1**	**51.9**	**72.8**
Comparison on SID datasets						
TPAMI'17	*Deeplab* Chen et al. (2017a)	VGG-16	64.3	46.7	35.5	48.5
ECCV'18	*DCNN* Wang and Neumann (2018)	VGG-16	65.4	55.5	39.5	49.9
ArXiv'19	*MMAFNet* Fooladgar and Kasaei (2019)	ResNet-152	76.5	62.3	52.9	-
ICCV'21	*ShapeConv* Cao et al. (2021)	ResNet-101	**82.7**	70.0	60.6	71.2
TransD-Fusion (Ours)		Swin-B	**82.7**	**72.0**	**62.2**	**71.5**

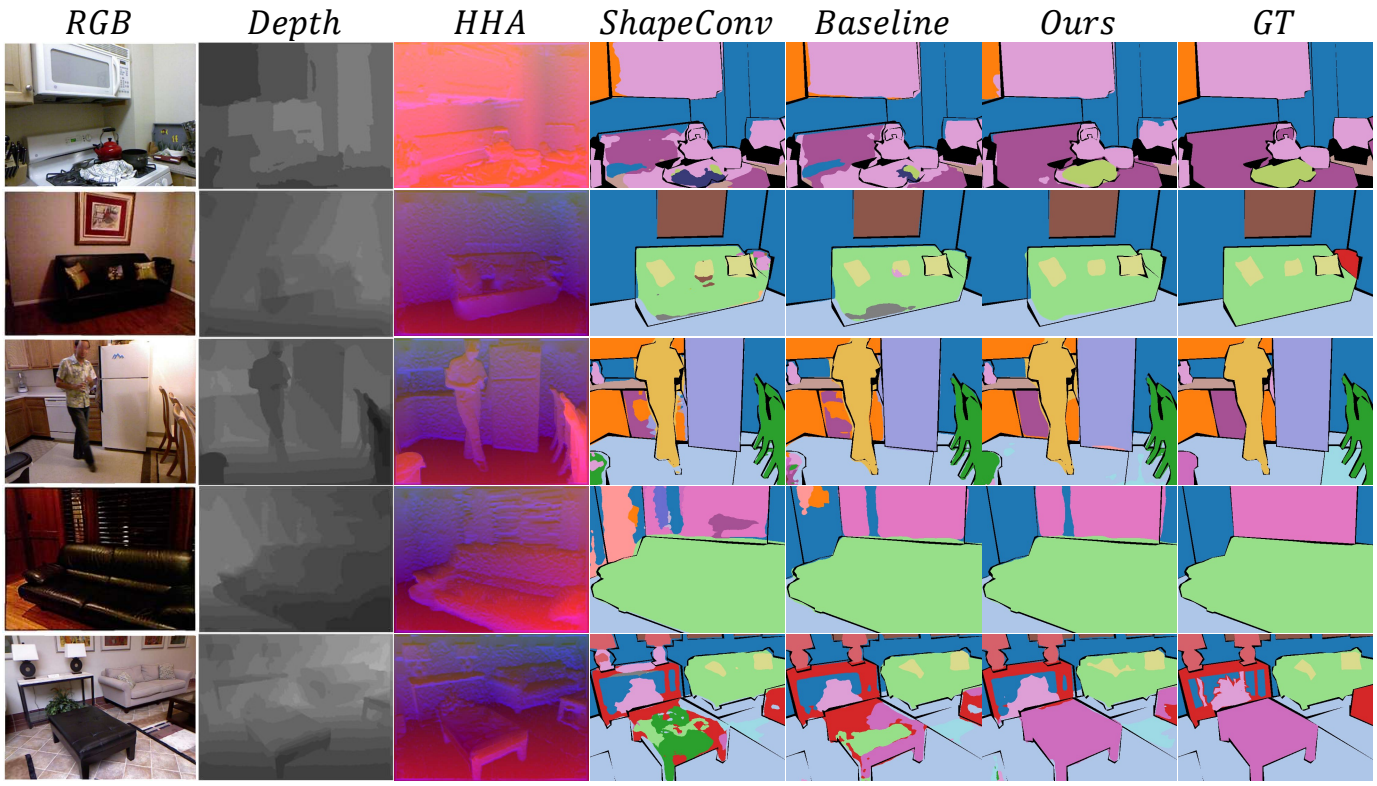

Fig. 5: **Qualitative comparison** with the SOTA CNN model, transformer baseline, and our TransD-Fusion. The black regions in semantic maps indicate the ignored category.

Table 2: **Parameter (M) Comparison**

Model	SAGate	ShapeConv	Omnivore	Ours
Param.	110.9	86.8	95.7	107.2
mIoU	52.4	51.3	54.0	55.5

network can achieve accurate semantic segmentation compared to other counterparts.

4.3. Computational Cost

In Table 2, we present a comparison of parameters. It is evident that our model possesses a comparable model size to state-of-the-art (SOTA) counterparts, yet significantly outperforms them by a wide margin.

5. Ablation Studies

During ablation, without specification, we conduct all the experiments on NYUv2 datasets with the Swin-B backbone under DeepLabV3+ architecture.

5.1. Generalization Capability.

Our TransD-Fusion can be used as a plug-in module. To demonstrate its generalization properties, we conduct experiments with several widely used semantic segmentation architectures, such as Segmenter Strudel et al. (2021), PSPnet Zhao et al. (2017), and DeeplabV3 Chen et al. (2017b) or DeeplabV3+ Chen et al. (2017a). The empirical performance is reported in Table 3. We can observe that our TransD-Fusion can consistently enable progress over the baseline performance in each architecture.

We also verify the generalization capability with a different variant of Swin backbones, i.e., Swin-T, Swin-S, Swin-B, and Swin-L. We refer readers to the original paper Liu et al. (2021b) for more details of the backbone design. To make a fair comparison, all the tests are conducted under the DeepLabV3+ Chen et al. (2017a) architecture. As shown in Table 4, our TransD-Fusion can consistently boost the baseline performance with large gains.

Table 3: **Performance analysis** with different decoders on NYUv2 dataset.

	Segmenter		PSPnet		DeeplabV3		DeeplabV3+	
	Baseline	Ours	Baseline	Ours	Baseline	Ours	Baseline	Ours
mAcc	55.3	56.9 (+**1.6**)	61.4	63.4 (+**2.0**)	63.8	64.4 (+**0.6**)	64.3	69.4 (+**5.1**)
mIoU	42.5	44.3 (+**1.8**)	49.2	50.6 (+**1.4**)	51.8	53.5 (+**2.7**)	52.6	55.5 (+**1.9**)

Table 4: **Generalization evaluation** on the NYUv2 dataset. Our TransD-Fusion can constantly boost performance with different backbones.

	Swin-T		Swin-S		Swin-B		Swin-L	
	Baseline	Ours	Baseline	Ours	Baseline	Ours	Baseline	Ours
mAcc	58.7	60.7 (+**2.0**)	62.6	65.0 (+**2.4**)	64.3	69.4 (+**5.1**)	68.0	69.6 (+**1.6**)
mIoU	47.1	48.4 (+**1.3**)	50.5	51.6 (+**1.1**)	52.6	55.5 (+**2.9**)	55.8	57.1 (+**1.3**)

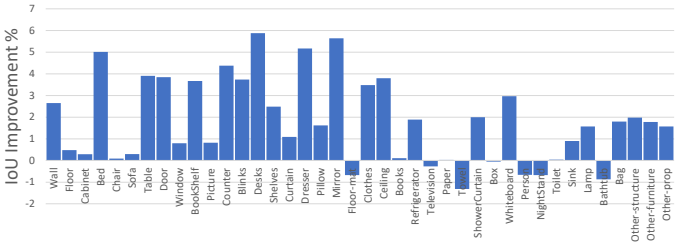

Fig. 6: **Per-class improvement** of our plain TransD-Fusion over the baseline (Swin-B).

5.2. *Network sharing.*

We first conduct experiments by replacing the parallel encoders with a shared-weight Siamese design. As such, we do not add extra learning costs during feature extraction and the improvement compared to the RGB-D single-stream baseline can be purely attributed to our transformer fusion. As shown in Table 5, our fusion strategy can significantly improve the baseline performance with the same encoder parameters. The fusion module only cost an extra 84 Mb learning parameter for the Swin-B baseline and an extra 235 Mb for the Swin-L baseline, while we achieve +2.6 mIoU with Swin-B and +2.2 mIoU with Swin-L. This validates the effectiveness of our fusion strategy. While the weights are not shared, the performance is slightly better compared to the Siamese design, which can be purely attributed to the doubled encoder parameters. We further present in Fig. 6 the per-class improvement with the Swin-B baseline.

5.2.1. *Robustness against Alignment Bias.*

We analyze the robustness of different fusion approaches against sensor misalignment, i.e., RGB and Depth maps are not accurately aligned at the pixel level. Specifically, we simulate a calibration error on NYUv2 by additionally cropping 20 pixels from the RGB input and obtaining a misaligned dataset. We retrain our TransD-Fusion (Swin-B) and the SOTA CNN model ShapeConv with early-fused input. To make a fair comparison, we additionally build two late-fusion baseline networks with the Swin-B backbone. The features are combined with attention modules such as SA gate Chen et al. (2020) (denoted as Swin + **SA**), or with simple pixel-wise concatenation and convolution (denoted as Swin + **Conv**).

The performances under the inferior condition are presented in Fig. 7 and in Tab. 6. Since **SA** and **Conv** are built upon the pixel-wise correlation between different modalities at the semantic level, their performances significantly drop when the features are no more accurately aligned. We observe 1.8% mIoU degradation on Swin + **SA** and 3.5% mIoU degradation on Swin + **Conv**, respectively. In contrast, our TransD-Fusion only drops 0.2% on mIoU. The stable performance against misalignment can be attributed to our fusion design which is built upon the contextualized correlation, yielding a more soft and robust fusion scheme for RGB-D semantic segmentation.

Table 5: **Performance analysis** with different architectures on NYUv2 dataset.

Architecture	Swin-B			Swin-L		
	Baseline	Siamese	Not Sharing	Baseline	Siamese	Not Sharing
mAcc	64.3	67.3 (+3.0)	**69.4** (+2.1)	67.5	69.0 (+1.5)	**69.5** (+0.5)
mIoU	52.6	55.2 (+2.6)	**55.5** (+0.3)	54.3	56.5 (+2.2)	**57.1** (+0.6)
Model Size (MB)	776	860 (+84)	1246 (+386)	1675	1910 (+235)	2595 (+685)

Table 6: **Robustness analysis** on the simulated misaligned NYUv2 dataset. Our TransD-Fusion leads to a more stable and superior performance.

Method	Crop (pixel)	PixelAcc	mAcc	mIoU	f.w.IoU
ShapeConv	40	74.7	62.5	49.2	61.1
Swin + CC	40	76.1	64.1	50.5	62.8
Swin + SA	40	75.7	63.1	50.7	62.2
TransD-Fusion (*MS*)	40	**78.1**	**69.1**	**55.1**	**65.7**
ShapeConv	60	74.6	60.7	48.2	60.8
Swin + CC	60	74.8	63.1	48.8	61.4
Swin + SA	60	75.3	63.7	49.7	61.9
TransD-Fusion (*MS*)	60	**77.9**	**68.8**	**54.8**	**65.5**

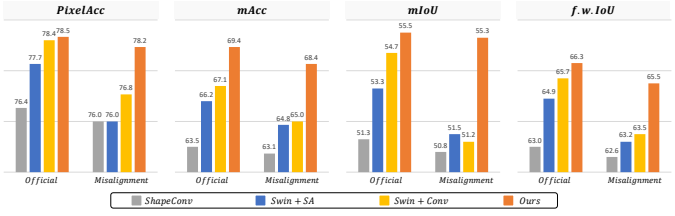

Fig. 7: **Robustness analysis** on the simulated misaligned NYUv2 dataset. Our TransD-Fusion leads to a more stable performance compared to the SOTA fusion approaches.

5.3. Comparison with Fusion Alternatives.

To verify the superior design of our proposed approach, we extensively compare our transformer fusion module with other alternatives. We report the quantitative result in Table 7. Specifically, we test with early fusion ("$F1$"), pixel-wise addition ("$F2$"), concatenation-convolution ("$F3$"), SA gate late Chen et al. (2020) ("$F4$"), SA gate middle Chen et al. (2020) ("$F5$"), Transfuser Prakash et al. (2021) ("$F6$"), and Medusa Song et al. (2021) ("$F7$").

Note that pixel-wise aggregation such as addition and concatenation-convolution are the most widely used naive RGB-D fusion strategies. SA gate Chen et al. (2020) further leverages the conventional cross-modal channel and spatial attention Woo et al. (2018) to improve the feature modeling before the fusion. Compared to the SA gate, our work is based on transformer attention which can better model long-range dependencies during feature fusion. Transfuser Prakash et al. (2021) and Medusa Song et al. (2021) are overall based on self-attention. The main difference is that Transfuser merges the self-enhanced features with the input, while Medusa directly processes with the self-enhanced output. Different from these transformer counterparts, our TransD-Fusion further adopts the cross-attention to explicitly model the cross-modal interaction and realize the feature fusion. Empirically, we conduct experiments by replacing our fusion module with these alternatives. To make a fair comparison, we applied the fusion strategy at the semantic level as ours, i.e., late fusion. SA Gate was initially applied to merge features at each stage. Therefore, we also conduct experiments with SA Gate with both middle (SA-M) and late (SA-L) fusion. As shown in Table 7, while replacing our fusion module with other pixel-wise fusion alternatives,

Table 7: **Empirical comparison** with fusion alternatives on NYUv2 dataset.

# Descrip.	F1 (Early)	F2 (Add)	F3 (Conv)	F4 (SA-L)	F5 (SA-M)	F6 (Transfuser)	F7 (Medusa)	**Ours**
mAcc	64.3	64.6	63.9	64.8	66.2	68.2	67.8	**69.4**
mIoU	52.6	52.8	52.2	53.0	53.3	54.3	53.8	**55.5**

Table 8: **Comparison** with positional encoding alternatives on NYUv2 dataset.

# Descrip.	P1 (w/o)	P2 (Abs)	P3 (Relative)	P4 (L4)	P5 (L3)	P6 (L2)	P7 (CPVT)	**Ours**
mAcc	68.2	67.9	67.5	66.6	67.4	68.8	68.3	**69.4**
mIoU	53.9	54.2	54.9	54.2	54.3	54.9	54.8	**55.5**

the performance significantly drops by around 3% on mIoU. Compared to other fusion methods built upon self-attention, our fusion design yields significantly better performance. The superior empirical results validate the effectiveness of our module with both self- and cross-attention.

While our TransD-Fusion reassembles previous works that all methods are based on transformer attention, our method mainly differs from the step-by-step fusion and our semantic-aware positional encodings. First, our self-calibration module is specifically designed to address the issue of feature calibration in multi-modal fusion. While previous works Chen et al. (2020); Cao et al. (2021) directly merge RGB-D features together, we argue that there is a non-negligible need to first preserve and improve the modality-specific features. This has not been thoroughly explored in prior works, especially coupled with transformer attention. Our module enables more attention to RGB and depth features separately, which is the basis of our further integration of the RGB-D features.

Second, our cross-interaction mechanisms and position encoding schemes are tailored to the specific characteristics of indoor scenes, where there are often large-scale variations and complex spatial relationships between objects. While previous methods with self-attention only and/or with conventional positional encodings Song et al. (2021); Prakash et al. (2021), our model can better capture these relationships and leverage them for improved segmentation accuracy.

Finally, our depth-guided fusion mechanism represents a significant departure from prior approaches, which typically use fixed weighting schemes to combine RGB and depth features. Our approach dynamically weights the two modalities based on the local context and spatial relationships in the scene, resulting in a more adaptive and effective fusion process compared to pixel-wised concatenation Chen et al. (2020); Broedermann et al. (2022).

5.4. Comparison with other position encodings (PEs).

Prior works adopt different PEs that focus on order awareness to improve feature extraction. The PE in our TransD-Fusion plays a more vital role since it should be locality-aware for better segmentation and be category-dependent for multi-modal fusion. To validate the superiority of our proposed PE, we conduct experiments by removing or replacing our encoding with other approaches and report the performance in Table 8. We have: "*P*1" without PE; "*P*2" with absolute PE; "*P*3" with relative PE. Since our PE can be learned from a hierarchical feature with higher resolution to fully excavate the spatial cues, we also conduct experiments to analyze the influence of feature resolution. We denote: "*P*4" for PE learned from the output of Layer 4; "*P*5" learned from Layer 3 output; "*P*6" learned from Layer 2 output. We replace our PE with the concurrent CPVT Chu et al. (2021) by re-implementing it in our TransD-Fusion, denoted as "*P*7". Under consideration of a fair comparison, we

Table 9: **Key components analysis** on NYUv2 dataset.

#	*Master*	*Sub*	*Add*	*Conv*	*SA − M*	*SE* 42 Mb	*CC* 37 Mb	*DGF* 5Mb	Metric mAcc	Metric mIoU
1	✓								66.1	53.8
2	✓					✓			67.0	53.9
3	✓	✓	✓						61.4	51.2
4	✓	✓		✓					67.1	54.6
5	✓	✓						✓	68.5	55.1
6	✓	✓				✓		✓	68.6	55.2
7	✓	✓			✓			✓	66.5	54.3
Ours	✓	✓				✓	✓	✓	**69.4**	**55.5**

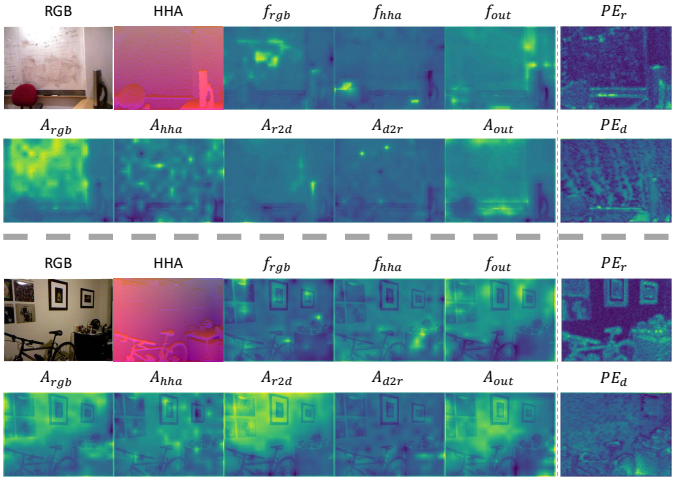

Fig. 8: **Visualization** of our TransD-Fusion. RGB and HHA and the input images. (f_{rgb}, f_{hha}) denote the input feature map for our transformer fusion, while f_{out} denotes the output of our transformer fusion. We also provide the visualization of attention maps. We have A_{rgb}: self-attention on RGB; A_{hha}: self-attention on HHA, A_{r2d}: RGB-guided attention; A_{d2r}: depth-guided attention; A_{out}: fused attention. On the right hand, we present the semantic-aware positional encoding (PE_r, PE_d) for RGB and depth tokens, respectively. It can be seen that our transformer fusion can effectively model long-range dependencies by deeply leveraging the cross-modal properties.

apply CPVT to learn features from Layer 1 output as our S-PE.

Empirical results in Table 8 show that there exists significant degradation on mIoU after removing or replacing our S-PE with conventional PEs. This validates the effectiveness of our S-PE which can better constrain the transformer attention for multi-modal fusion. We also observe that the spatial dimension plays an imperial role in our S-PE. When the spatial resolution decreases, i.e., from Layer 1 output to Layer 4 output, the performances with our S-PE drop as well. Compared to the concurrent CPVT, our superior performance demonstrates that we can better leverage locality awareness.

Note that since our positional encodings are built upon convolutions, we can better constrain the attention into local regions, as shown in Figure 8. Therefore, our network can converge quickly without additional requirements on the training epochs.

5.5. Key Components Analysis of TransD-Fusion

In this section, we conduct studies to verify the importance of the key components of TransD-Fusion: Master stream (master), Subsidiary stream (sub), Self-Enhancement (SE), Cross-Calibration (CC), and Depth-Guided Fusion (DGF). All the experiments are built upon the Swin-B backbone and we report the associated model size for each module. We remove partially or entirely the key components. To make a fair comparison, we additionally conduct experiments with conventional fusion strategies such as element-wise addition (**Add**), concatenation-convolution (**Conv**), and the concurrent SA module Chen et al. (2020) under the same architecture. Note that the SA module is initially applied for middle fusion. Under the consideration of a fair comparison, we adopt the same middle fusion design to merge RGB-D features at each scale. This is denoted as SA-M in Table 9.

We observe from Table 9 that after removing the cross-calibration module, the performance drops since the modality-specific features can no more benefit from complementary cues. Without self-enhancement, the performance further degrades. While further replacing the depth-guided fusion strategy with a pixel-wise fusion module, we can observe a significant drop, i.e., 3.9% ↓ on mIoU with **Add** and 0.5% ↓ on mIoU with **Conv**. These results validate the necessity of leveraging the long-range dependencies for feature fusion. Finally, by comparing lines #5-#6, we observe that the SE plays a minimal role. Therefore we try to replace our SE with the SA module Chen et al. (2020). However, the performance significantly drops, which shows the importance of our self-attention that fully leverages and preserves modality-specific features with contextualized cues.

5.6. Visualization

To better understand our proposed TransD-Fusion, we provide the feature visualization in Fig. 8, where the activation map is generated through the average across channels. We can observe that the extracted feature (f_{rgb}, f_{hha}) are modal-specific, i.e., f_{rgb} contain more texture clues, while f_{hha} are more sensitive to geometric changes. By comparing our fused output f_{out} with the input feature maps, we can observe that our TransD-Fusion can effectively leverage cross-modal information to generate the output enhanced with contextualized awareness.

We also provide the visualization of the attention maps. It can be seen that the self-attention (A_{rgb}, A_{hha}) performs in a similar manner as the encoded features, i.e., focusing on texture and geometric knowledge, respectively. This observation supports our design on the self-enhancement block that aims to improve the modality-specific feature modeling. Then, we apply the cross-attention to enable the bi-directional interaction. From the attention (A_{r2b}, A_{d2r}), we can observe that our cross-modal blocks can efficiently leverage one modality to calibrate and improve the other. Finally, we use the depth features as the query and design a depth-guided fusion block. We can see from the output attention map A_{out} that our module can guide the network by focusing on the global structure and the boundary.

To further understand the transformer attention, we provide in Figure 9 the attention map with the pixel from the sofa as the query. It can be seen that A_{rgb} has activation on objects sharing similar textures, while the A_{hha} has more activations on objects sharing similar depth. Our cross attention A_{r2b}, A_{d2r} further improves the affinity matrix with more contextualized and cross-modal clues, leading to a more global attention output A_{out}.

Finally, we show in Figure 10 the visualization of the Class Activation Map (CAM). It can be seen that our model contains attention to different local regions with respect to the target semantics. Moreover, it can be seen our fusion combines the clues from both RGB texture and HHA geometries for a more refined and accurate output.

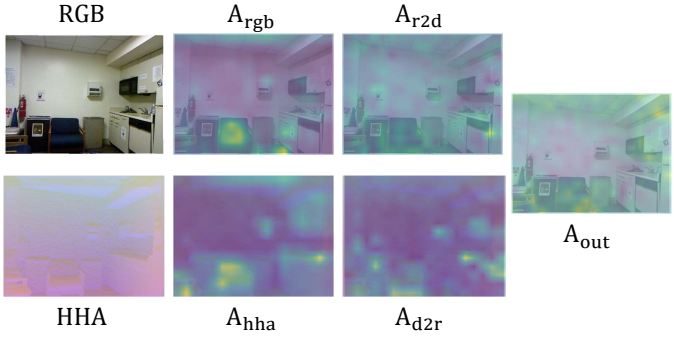

Fig. 9: **Attention Visualization**. We use the pixel from Sofa as a query.

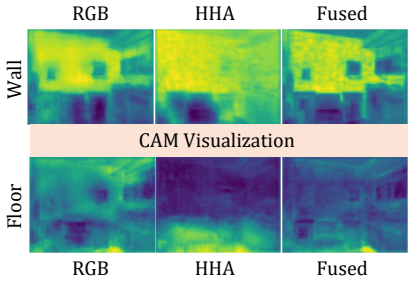

Fig. 10: **Visualization of Class Activation Map (CAM)**. We show the activation map for different target semantics.

6. Conclusion

In this paper, we propose a novel RGB-D fusion scheme for semantic segmentation. Different from previous fusion designs built upon pixel-wise correlation, our network fully explores the transformer attention to aggregate multi-modal features with contextualized cues. Additionally, we design a novel position encoding generator to better leverage the locality awareness into our transformer fusion. Extensive ablation studies verify the generalization property and robustness against misalignment of our TransD-Fusion. The comparison with previous works on fusion design and position encoding further validates the effectiveness of our proposed approach. Experiments on challenging RGB-D benchmarks demonstrate that our TransD-Fusion performs well over the state-of-the-art methods by large margins.

Acknowledgements

This research is supported by the French National Research Agency through ANR CLARA (ANR-18-CE33-0004) and fi-

nanced by the French Conseil Régional de Bourgogne-Franche-Comté.

References

Armeni, I., Sax, S., Zamir, A.R., Savarese, S., 2017. Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 .

Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, Q.V., 2019. Attention augmented convolutional networks, in: Proceedings of the IEEE/CVF international conference on computer vision (ICCV).

Borse, S., Wang, Y., Zhang, Y., Porikli, F., 2021. Inverseform: A loss function for structured boundary-aware segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Broedermann, T., Sakaridis, C., Dai, D., Van Gool, L., 2022. Hrfuser: A multi-resolution sensor fusion architecture for 2d object detection. arXiv preprint arXiv:2206.15157 .

Cao, J., Leng, H., Lischinski, D., Cohen-Or, D., Tu, C., Li, Y., 2021. Shapeconv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers, in: Proceedings of the European Conference on Computer Vision (ECCV).

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence (TPAMI) 40, 834–848.

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017b. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 .

Chen, L.Z., Lin, Z., Wang, Z., Yang, Y.L., Cheng, M.M., 2021a. Spatial information guided convolution for real-time rgbd semantic segmentation. IEEE Transactions on Image Processing (TIP) 30, 2313–2324.

Chen, X., Lin, K.Y., Wang, J., Wu, W., Qian, C., Li, H., Zeng, G., 2020. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV).

Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H., 2021b. Transformer tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Chen, Y., Mensink, T., Gavves, E., 2019. 3D neighborhood convolution: Learning depth-aware features for RGB-D and RGB semantic segmentation, in: International Conference on 3D Vision (3DV).

Chu, X., Tian, Z., Zhang, B., Wang, X., Wei, X., Xia, H., Shen, C., 2021. Conditional positional encodings for vision transformers. arXiv preprint arXiv:2102.10882 .

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks, in: Proceedings of the IEEE international conference on computer vision (ICCV).

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database, in: IEEE conference on computer vision and pattern recognition (CVPR).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, (ICLR).

Fan, D.P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.M., 2021. Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks. IEEE Transactions on neural networks and learning systems (TNNLS) 32, 2075–2089.

Fooladgar, F., Kasaei, S., 2019. Multi-modal attention-based fusion model for semantic segmentation of RGB-Depth images. arXiv preprint arXiv:1912.11691 .

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR).

Gao, P., Zheng, M., Wang, X., Dai, J., Li, H., 2021. Fast convergence of DETR with spatially modulated co-attention. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) .

Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N., 2017. Convolutional sequence to sequence learning, in: International Conference on Machine Learning (ICML).

Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., Misra, I., 2022. Omnivore: A single model for many visual modalities, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Gupta, S., Girshick, R., Arbeláez, P., Malik, J., 2014. Learning rich features from RGB-D images for object detection and segmentation, in: Proceedings of European conference on computer vision (ECCV).

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al., 2022. A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) .

Hazirbas, C., Ma, L., Domokos, C., Cremers, D., 2016. Fusenet: Incorporating depth into semantic segmentation via fusion-based CNN architecture, in: Proceedings of Asian conference on computer vision (ACCV).

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR).

He, P., Liu, X., Gao, J., Chen, W., 2021. Deberta: Decoding-enhanced BERT with disentangled attention, in: International Conference on Learning Representations (ICLR).

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR).

Hu, X., Yang, K., Fei, L., Wang, K., 2019. ACNet: Attention based network to exploit complementary features for RGB-D semantic segmentation, in: IEEE International Conference on Image Processing (ICIP).

Ji, W., Li, J., Yu, S., Zhang, M., Piao, Y., Yao, S., Bi, Q., Ma, K., Zheng, Y., Lu, H., et al., 2021. Calibrated RGB-D salient object detection, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR).

Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M., 2021. Transformers in vision: A survey. ACM Computing Surveys (CSUR) .

Li, Y., Yu, A.W., Meng, T., Caine, B., Ngiam, J., Peng, D., Shen, J., Lu, Y., Zhou, D., Le, Q.V., Yuille, A., Tan, M., 2022. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Li, Z., Gan, Y., Liang, X., Yu, Y., Cheng, H., Lin, L., 2016. LSTM-CF: Unifying context modeling and fusion with LSTMs for RGB-D scene labeling, in: Proceedings of the European Conference on Computer Vision (ECCV).

Liu, H., Zhang, J., Yang, K., Hu, X., Stiefelhagen, R., 2022. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. arXiv preprint arXiv:2203.04838 .

Liu, N., Zhang, N., Wan, K., Shao, L., Han, J., 2021a. Visual saliency transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) .

Prakash, A., Chitta, K., Geiger, A., 2021. Multi-modal fusion transformer for end-to-end autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, (ICML).

Seichter, D., Köhler, M., Lewandowski, B., Wengefeld, T., Gross, H.M., 2021. Efficient RGB-D semantic segmentation for indoor scene analysis, in: IEEE International Conference on Robotics and Automation (ICRA).

Shaw, P., Uszkoreit, J., Vaswani, A., 2018. Self-attention with relative position representations. Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) .

Shen, X., Stamos, I., 2020. Frustum voxnet for 3d object detection from rgb-d or depth images, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).

Shen, X., Stamos, I., 2021. 3d object detection and instance segmentation from 3d range and 2d color images. Sensors 21, 1213.

Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from RGBD images, in: Proceedings of the European Conference on Computer Vision (ECCV).

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations (ICLR).

Song, H., Kim, E., Jampan, V., Sun, D., Lee, J.G., Yang, M.H., 2021. Exploiting scene depth for object detection with multimodal transformers, in: 32nd British Machine Vision Conference (BMVC).

Song, S., Lichtenberg, S.P., Xiao, J., 2015. SUN RGB-D: A RGB-D scene understanding benchmark suite, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR).

Strudel, R., Garcia, R., Laptev, I., Schmid, C., 2021. Segmenter: Transformer for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Su, Y., Yuan, Y., Jiang, Z., 2021. Deep feature selection-and-fusion for RGB-D semantic segmentation, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE.

Tay, Y., Dehghani, M., Bahri, D., Metzler, D., 2020. Efficient transformers: A survey. arXiv preprint arXiv:2009.06732 .

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in Neural Information Processing Systems (NeurIPS).

Wang, J., Wang, Z., Tao, D., See, S., Wang, G., 2016. Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks, in: Proceedings of the European Conference on Computer Vision (ECCV).

Wang, N., Zhou, W., Wang, J., Li, H., 2021. Transformer meets tracker: Exploiting temporal context for robust visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Wang, W., Neumann, U., 2018. Depth-aware CNN for RGB-D segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV).

Wang, Y., Chen, X., Cao, L., Huang, W., Sun, F., Wang, Y., 2022a. Multimodal token fusion for vision transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., Huang, J., 2020. Deep multimodal fusion by channel exchanging, in: Advances in Neural Information Processing Systems (NeurIPS).

Wang, Y., Sun, F., Huang, W., He, F., Tao, D., 2022b. Channel exchanging networks for multimodal and multitask dense image prediction. IEEE Transaction on Pattern Analysis and Machine Intelligence (TPAMI) .

Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV).

Wu, K., Peng, H., Chen, M., Fu, J., Chao, H., 2021. Rethinking and improving relative position encoding for vision transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Wu, W., Chu, T., Liu, Q., 2022a. Complementarity-aware cross-modal feature fusion network for rgb-t semantic segmentation. Pattern Recognition 131, 108881.

Wu, Z., Allibert, G., Stolz, C., Demonceaux, C., 2020. Depth-adapted CNN for RGB-D cameras, in: Proceedings of the Asian Conference on Computer Vision (ACCV).

Wu, Z., Allibert, G., Stolz, C., Ma, C., Demonceaux, C., 2022b. Depth-adapted CNNs for RGB-D semantic segmentation. arXiv preprint arXiv:2206.03939 .

Xing, Y., Wang, J., Chen, X., Zeng, G., 2019. 2.5 D convolution for RGB-D semantic segmentation, in: IEEE International Conference on Image Processing (ICIP).

Xing, Y., Wang, J., Zeng, G., 2020. Malleable 2.5 D convolution: Learning receptive fields along the depth-axis for RGB-D scene parsing, in: Proceedings of the European Conference on Computer Vision (ECCV).

Xiong, Z., Yuan, Y., Guo, N., Wang, Q., 2020. Variational context-deformable convnets for indoor scene parsing, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR).

Yan, B., Peng, H., Fu, J., Wang, D., Lu, H., 2021. Learning spatio-temporal transformer for visual tracking, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Yang, J., An, W., Yan, C., Zhao, P., Huang, J., 2021. Context-aware domain adaptation in semantic segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V., 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems (NIPS) .

Yew, Z.J., Lee, G.H., 2022. Regtr: End-to-end point cloud correspondences with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Yue, Y., Zhou, W., Lei, J., Yu, L., 2021. Rtlnet: Recursive triple-path learning network for scene parsing of rgb-d images. IEEE Signal Processing Letters 29, 429–433.

Zhang, G., Xue, J.H., Xie, P., Yang, S., Wang, G., 2021. Non-local aggregation for RGB-D semantic segmentation. IEEE Signal Processing Letters (SPL) 28, 658–662.

Zhang, J., Liu, R., Shi, H., Yang, K., Reiß, S., Peng, K., Fu, H., Wang, K., Stiefelhagen, R., 2023. Delivering arbitrary-modal semantic segmentation, in: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR).

Zhang, X., Zhang, S., Cui, Z., Li, Z., Xie, J., Yang, J., 2022. Tube-embedded transformer for pixel prediction. IEEE Transactions on Multimedia (TMM) .

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR).

Zhou, H., Qi, L., Huang, H., Yang, X., Wan, Z., Wen, X., 2022a. CANet: Co-attention network for RGB-D semantic segmentation. Pattern Recognition (PR) 124, 108468.

Zhou, H., Qi, L., Wan, Z., Huang, H., Yang, X., 2020. RGB-D co-attention network for semantic segmentation, in: Proceedings of the Asian Conference on Computer Vision (ACCV).

Zhou, W., Lv, S., Lei, J., Luo, T., Yu, L., 2022b. Rfnet: Reverse fusion network with attention mechanism for rgb-d indoor scene understanding. IEEE Transactions on Emerging Topics in Computational Intelligence .

Zhou, W., Yang, E., Lei, J., Wan, J., Yu, L., 2022c. Pgdenet: Progressive guided fusion and depth enhancement network for rgb-d indoor scene parsing. IEEE Transactions on Multimedia (TMM) .

Zhou, W., Yang, E., Lei, J., Yu, L., 2022d. Frnet: Feature reconstruction network for rgb-d indoor scene parsing. IEEE Journal of Selected Topics in Signal Processing .

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2021. Deformable DETR: deformable transformers for end-to-end object detection, in: International Conference on Learning Representations (ICLR).