Ferdowsi University of Mashhad

Journal of Computer and Knowledge Engineering

https://cke.um.ac.ir

Information and Communication Technology Association of Iran

Discriminative Cross-Modal Attention Approach for RGB-D Semantic Segmentation*

Research Article

Emad Mousavian [1], Danial Qashqai [2], Shahriar B. Shokouhi [3]

10.22067/cke.2025.88682.1117

Abstract *Scene understanding through semantic segmentation is a vital component for autonomous vehicles. Given the importance of safety in autonomous driving, existing methods are constantly striving to improve accuracy and reduce error. RGB-based semantic segmentation models typically underperform due to information loss in challenging situations such as lighting variations and limitations in distinguishing occluded objects of similar appearance. Therefore, recent studies have developed RGB-D semantic segmentation methods by employing attention-based fusion modules. Existing fusion modules typically combine cross-modal features by focusing on each modality independently, which limits their ability to capture the complementary nature of modalities. To address this issue, we propose a simple yet effective module called the Discriminative Cross-modal Attention Fusion (DCMAF) module. Specifically, the proposed module performs cross-modal discrimination using element-wise subtraction in an attention-based approach. By integrating the DCMAF module with efficient channel- and spatial-wise attention modules, we introduce the Discriminative Cross-modal Network (DCMNet), a scale- and appearance-invariant model. Extensive experiments demonstrate significant improvements, particularly in predicting small and fine objects, achieving an mIoU of 77.39% on the CamVid dataset, outperforming state-of-the-art RGB-based methods, and a remarkable mIoU of 82.8% on the Cityscapes dataset. As the CamVid dataset lacks depth information, we employ the DPT monocular depth estimation model to generate depth images.*

Key Words Attention Mechanism, Autonomous Driving, Deep Learning, RGB-D Semantic Segmentation.

1- Introduction

In recent years, semantic segmentation methods have seen a dramatic development due to the advancement of deep learning methods. Semantic segmentation aims to assign a semantic label to each pixel within an image. Concurrently, advancements in depth estimation methods and the development of RGB-D techniques have highlighted the significant role of geometric information in enhancing semantic segmentation performance. Depth information is less sensitive to illumination variations and, by providing 3D information, can be used as complementary data for RGB images. Therefore, RGB-D semantic segmentation models [1, 2, 3] have been developed to enhance the results of RGB-based models. The benefit of integrating both RGB and depth information is shown in Fig. 1.

RGB-D semantic segmentation models use different approaches to fuse RGB and depth features. Early methods [4, 5] utilized element-wise addition for cross-modal feature fusion, whereas more recent approaches [6, 7, 8] have introduced attention-based modules to achieve improved fusion. These fusion modules are designed to apply attention mechanisms either to individual modalities [1, 3] or to their interactions [2, 7]. Specifically, methods [1, 3, 6] incorporate channel-wise attention within their fusion modules, while [2, 7] and [9] utilize channel and spatial attention in sequential and parallel configurations, respectively. Despite this improvement, a potential issue arises when the RGB branch carries less information than the corresponding depth branch, or vice versa. In such cases, attending separately to feature maps of these modalities can degrade performance by failing to take advantage of the complementary information available from the other modality.

To address this issue, we introduce the Discriminative Cross-Modal Attention Fusion (DCMAF) module. Unlike prior attention-based approaches, this module improves cross-modal fusion by evaluating the discriminative power

* Manuscript received 2024 June 26, Revised 2025 February 23, accepted 2025 April 6.

[1] Corresponding author. Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran, **Email**: emad_mousavian@elec.iust.ac.ir

[2] Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran, **Email**: danialqashqai99@gmail.com

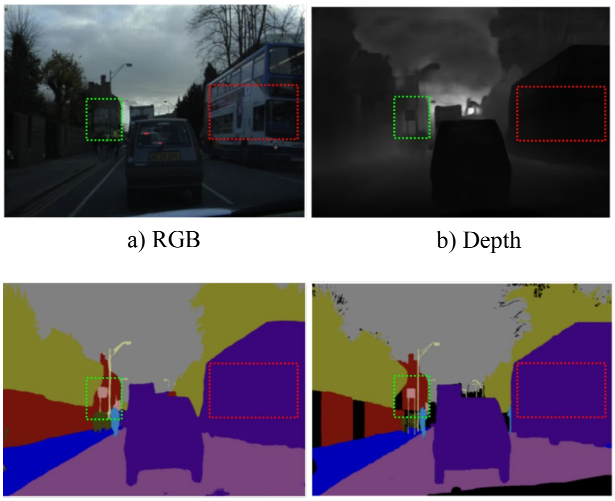
[3] Associate professor, Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran, **Email**: bshokouhi@iust.ac.ir

of each modality through element-wise subtraction. Therefore, the proposed DCMAF module: 1) enhances the fusion of complementary data by performing feature map comparison between modalities; and 2) facilitates gradient flow during backpropagation to update weights that extract the most salient features from the depth and RGB branches. By following the architecture of ACNet [1] and using the proposed DCMAF module, we present the Discriminative Cross-Modal Network (DCMNet). Furthermore, we enhance our model's performance by effectively integrating spatial and channel attention modules.

Experimental results on the CamVid [10] and Cityscapes [11] datasets show that our method enhances semantic segmentation performance, particularly in the accurate prediction of small and fine objects. Given that the CamVid dataset does not include depth information, we use the DPT model [12], a well-established network for monocular depth estimation, to generate synthetic depth images for this dataset. On the CamVid dataset, our proposed model achieves an mIoU of 77.39%, which outperforms the state-of-the-art RGB-based models. Furthermore, on the Cityscapes dataset, our model attains an mIoU of 82.8%, demonstrating a clear improvement over recent, comparable models.

The main contributions of this paper are summarized as follows:

- We introduce a Discriminative Cross-modal Attention Fusion (DCMAF) module to apply attention based on the distinction between RGB and depth modalities.
- We use the DPT monocular depth estimation model to generate synthetic depth images for the CamVid dataset.
- We propose DCMNet for RGB-D semantic segmentation, which is state-of-the-art on challenging outdoor semantic segmentation datasets.



a) RGB b) Depth

c) Prediction d) Ground-Truth

Fig. 1 Despite differing feature distributions, the complementarity of RGB images (a) and depth information (b) enhances semantic segmentation accuracy. For instance, green boxes highlight regions with prominent depth features, while red boxes indicate regions rich in color texture features. Comparison of the predicted segmentation (c) and ground truth (d) demonstrates the effectiveness of leveraging this complementary information.

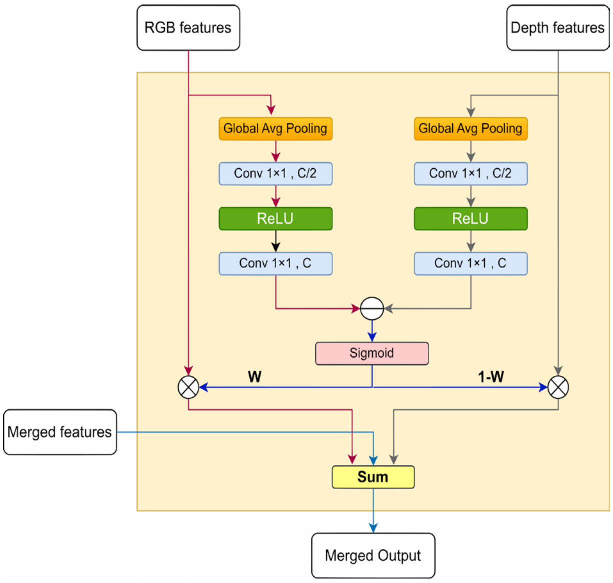


Fig. 2. Structure of Discriminative Cross-Modal Attention Fusion (DCMAF) Module.

2- Related Works

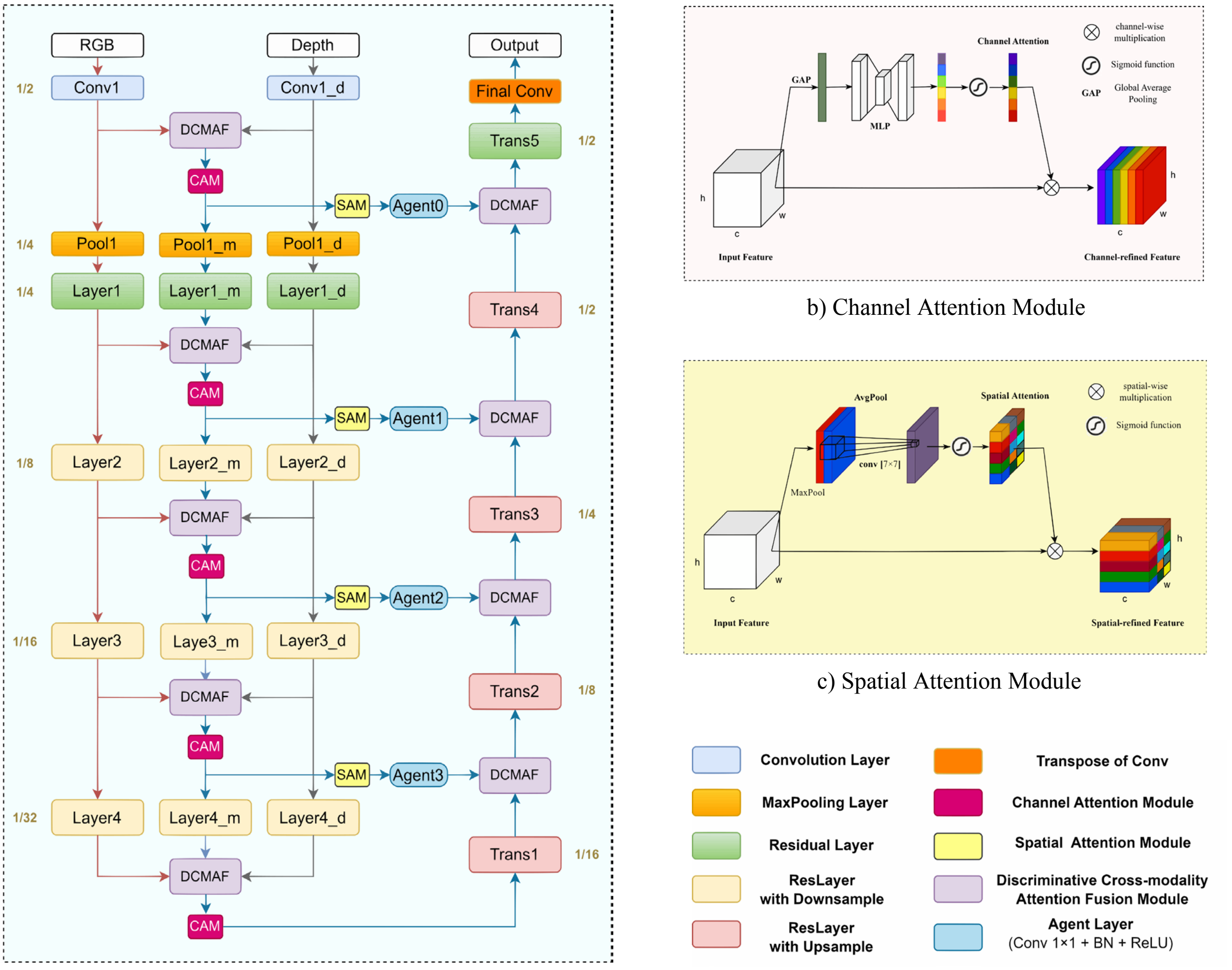
This section provides an overview of prior RGB and RGB-D semantic segmentation methods, due to their significant overlap.

A. Semantic Segmentation

Semantic segmentation is an image analysis technique that assigns a semantic label to each pixel within an image. The FCN [13] network was the first successful network in semantic segmentation, which used fully convolutional layers in its structure and solved the problem of resizing the input image. The FCN approach has had a major influence on subsequent research to enhance segmentation results. For instance, [14, 15] adopted encoder-decoder architectures with skip connections; [16, 17, 18] focused on expanding the receptive field; [19, 20, 21] used attention modules; and [9, 22, 23] incorporated vision transformers into their architectures.

B. RGB-D Semantic Segmentation

Semantic segmentation of RGB-D images is based on the interaction and fusion of RGB and depth information. Recent research has shown that fusing depth information with RGB images results in enhanced outcomes in comparison to RGB-based models. The feature fusion of RGB-D semantic segmentation models can be categorized into four types: early fusion, mid-term fusion, late fusion, and multi-level interactive fusion. In early fusion, RGB and depth images are concatenated and fed into the model input. This approach does not fundamentally alter the model structure compared to RGB-based models, and the network input is only increased to four channels, similar to the SegNet [14] model. Such a simple combination ignores the complementary nature of RGB images and depth information. Mid-term fusion [24] uses two parallel



b) Channel Attention Module

c) Spatial Attention Module

a) DCMNet model

Fig. 3. Architecture of DCMNet for RGB-D semantic segmentation. (a) Encoder-decoder architecture of DCMNet with RGB and Depth inputs. (b) Channel Attention Module (CAM). (c) Spatial Attention Module (SAM).

encoders to extract each modal separately and only fuses cross-modal features at the end of the encoder network. Late fusion [25] fuses RGB and depth feature maps after being separately encoded and decoded. Due to the gradual loss of information in the two networks, this method is unable to reconstruct the rich information of each branch. The multi-level interactive fusion method is relatively more complex, and most recent models [7, 26, 27] use this method. In this approach, feature maps extracted from depth and RGB images are fused at multiple levels, resulting in superior accuracy compared to alternative fusion methodologies.

3-Proposed Method

In this section, we first introduce the Discriminative Cross-Modal Attention Fusion (DCMAF) module and then explain the architecture of the proposed Discriminative Cross-Modal Network (DCMNet). We also propose the efficient use of channel and spatial attention modules.

A. Discriminative Cross-Modal Attention Fusion Module

Conventional RGB-D methods typically employ a fusion module to attend to the channels of each modality independently before the fusion operation. This approach lacks direct connections between the corresponding channels in the RGB and depth branches, leading to the fusion of cross-modal information without adequate consideration of complementary features. To address this limitation, we propose the Discriminative Cross-Modal Attention Fusion (DCMAF) module. The DCMAF module leverages element-wise subtraction to compute channel-wise differences between modalities. This subtraction of feature maps is a simple and computationally efficient approach that facilitates cross-modal interaction by highlighting conflicting or complementary information. Subsequently, utilizing the output of this subtraction within an attention mechanism causes the attention weights to become dependent on the relative differences between the RGB and depth features. This establishes an implicit dependency, enabling the model to prioritize modalities based on their differences, and during backpropagation, directs gradients towards the most informative modalities or features for more efficient learning of cross-modal dependencies. Thereby, the DCMAF module creates a meaningful and effective fusion method.

As shown in Fig. 2, the DCMAF module first applies a global average pooling layer to the feature maps of each RGB and depth branch. The resulting vectors represent the amount of global information within each channel. Subsequently, two 1×1 convolutional layers with the

ReLU activation are used for each branch to extract the correlation between the channels. The difference between the RGB and depth vectors is then computed via element-wise subtraction, enabling the module to distinguish between the two modalities. Finally, the resulting vector is passed through a sigmoid function to scale the weight vector to the range of 0 to 1. We multiply the weight vector (W) by the RGB features and the vector (1-W) by the depth features, and then we aggregate the weighted feature maps and the feature maps extracted from the middle branch by the addition operation. It is worth noting that in the first DCMAF module within the network, only the weighted feature maps of the RGB and depth branches are summed. The DCMAF module is formulated as follows:

$$\begin{aligned} W_L^c(\mathcal{D}_L^c, \mathcal{I}_L^c) &= \sigma\Bigg(w_{1,L}^c\left(\mathcal{R}e\left(w_{0,L}^{c/2}\left(\mathcal{I}_{avg,L}^c\right)\right)\right) - \\ & w'^{c}_{1,L}\left(\mathcal{R}e\left(w'^{c/2}_{0,L}\left(\mathcal{D}_{avg,L}^c\right)\right)\right)\Bigg), \\ \mathcal{D}_L' &= \mathcal{D}_L^c \otimes \left(1 - W_L^c(\mathcal{D}_L^c, \mathcal{I}_L^c)\right), \\ \mathcal{I}_L' &= \mathcal{I}_L^c \otimes W_L^c(\mathcal{D}_L^c, \mathcal{I}_L^c), \\ M_L &= M_{L-1}' + \mathcal{D}_L' + \mathcal{I}_L', \end{aligned} \tag{1}$$

where $\mathcal{D}_{avg,L}^c$ and $\mathcal{I}_{avg,L}^c$ are the average vectors of the depth ($\mathcal{D}_L^c$) and RGB ($\mathcal{I}_L^c$) feature maps, respectively, calculated across c channels at level L. The parameters $w_{0,L}^{c/2}$, $w_{1,L}^c$, $w'^{c/2}_{0,L}$‘ and $w'^{c}_{1,L}$ denote the independent weights of the 1×1 convolutional layers, and W_L^c represents the weight obtained after applying the sigmoid function (σ). M_L is the output of the DCMAF module, derived from the weighted element-wise summation of the RGB and depth feature maps ($\mathcal{I}_L'$ and $\mathcal{D}_{\mathrm{L}}'$, respectively) with the merged feature maps of the previous level (M_{L-1}').

B. Network Architecture

We introduced the Discriminative Cross-Modal Network (DCMNet) by following the ACNet [1] architecture. DCMNet uses a multi-level interactive fusion strategy with three branches of ResNet50 [28] as its encoder network. These branches are responsible for extracting features from RGB images, depth information, and cross-modal fused features. The proposed model architecture is shown in Fig. 3(a). Initially, the network applies a 7×7 convolution with a stride of 2 on the RGB and depth images, and then the extracted feature maps of both branches are fed to the first DCMAF module to merge information with a discriminative cross-modal attention-based approach. Generally, this fusion process is performed at five levels with output strides of 2, 4, 8, 16, and 32. Since attention in DCMAF is applied between corresponding channels in two branches, the weight assigned to channels with same numerical amount of information will be equal, regardless of the quantity of useful information in both channels. Therefore, a channel attention module (Fig. 3(b)) will be applied after each DCMAF. The middle branch is dedicated to processing cross-modal information, leveraging complementary features from both RGB and depth modalities to specifically extract fused features. Within the encoder network, skip connections are utilized to transfer multi-level features to the decoder. For enhanced efficacy, a spatial attention module (Fig. 3(c)) is applied alongside the Agent block. The Agent block is a 1×1 convolutional layer followed by a batch normalization layer and a ReLU activation, which offers computational efficiency.

The decoder network uses residual layers similar to the encoder network. Simple residual blocks are used instead of bottleneck blocks, which contain transposed convolutional layers, to perform upsampling. The upsampled feature maps in the decoder network are fused with the corresponding skip connection feature maps at different levels by the DCMAF module. Finally, a residual layer without upsampling, followed by a transposed convolution, is used to generate the semantic segmentation prediction.

4- EXPERIMENTS AND RESULTS

In this section, we first introduce the CamVid and Cityscapes datasets. Then, we explain the implementation details and compare the experimental results with the state-of-the-art methods in terms of accuracy and computational complexity. We also study the effectiveness of the proposed methods and finally discuss the pros and cons of the proposed model.

A. Datasets

1) *CamVid [10]:* This dataset contains driving scenes for semantic segmentation. It consists of 701 images at a resolution of 720 × 960, with 11 semantic classes. There are 367, 101, and 233 images for training, validation, and testing, respectively. In this paper, we use DPT-Hybrid [12] as a successful monocular depth estimation network to estimate the depth images of the CamVid dataset. A few examples of RGB and synthetic depth images of the CamVid dataset are shown in Fig. 4.

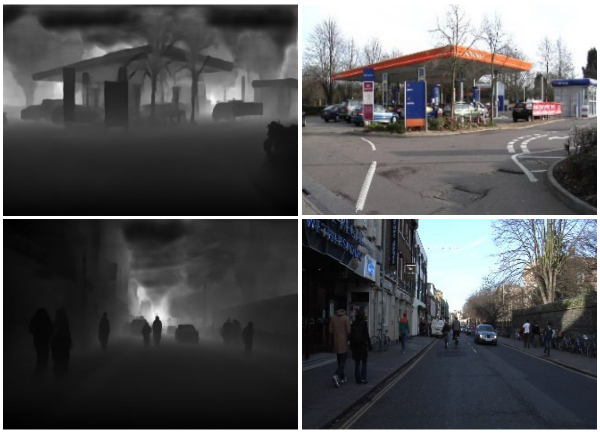


Fig. 4. Some examples of synthetic depth images estimated by the DPT-Hybrid model.

2) *Cityscapes [11]:* This RGB-D dataset includes urban street scenes from 50 different cities. It consists of 5,000 well-annotated images with 19 semantic classes. The images have a high resolution of 2048×1024 and

are split into 2,975 images for training, 500 for validation, and 1,525 for testing.

TABLE I
Comparison of our model with other methods on CamVid dataset. "I" and "D" denote RGB and depth maps, respectively. *: RGB-D methods that are well-trained on RGB-depth CamVid dataset.

Model	**Modality**	**Params (M)**	**mIoU (%)**
SegNet [14]	I	29.5	55.6
EDANet [31]	I	0.68	66.4
ICNet [34]	I	26.5	67.1
LBN-AA [35]	I	6.2	68
BiSeNet1 [33]	I	49	68.7
S2-FPN18 [32]	I	17.8	69.5
S2-FPN34 [32]	I	27.9	71
BiSeNetV2 [36]	I	-	72.4
STDC1-Seg75 [37]	I	8.4	73
BiSeNetV2L [36]	I	-	73.2
STDC2-Seg75 [37]	I	12.5	73.9
S2-FPN34M [32]	I	27.9	74.2
SGACNet-R34-NBt1D* [9]	I & D	35.2	75.83
RedNet-R50* [5]	I & D	81.8	75.94
SA-Gate-R50* [2]	I & D	64.7	76.01
ESANet-R50* [3]	I & D	54.2	76.13
FRNet-R34* [38]	I & D	101.9	76.27
ACNet-R50* [1]	I & D	124.8	76.30
DCMNet (Ours)	I & D	133.5	**77.39**

B. Implementation Details

We train our proposed DCMNet using the CamVid and Cityscapes datasets on two NVIDIA T4 GPUs. To generate the CamVid depth information, we leverage DPT-Hybrid, which is trained on the large-scale MIX 6 dataset and fine-tuned on the KITTI driving scenes dataset [29]. The ResNet50 backbone is pre-trained on the ImageNet [30] dataset. Since the depth data has only one channel, we convert the three channels of the first layer of ResNet50 to a single channel by averaging. For data augmentation, different scales with ratios of {0.75, 0.85, 1, 1.25, 1.5, 1.75} are used, along with random horizontal flipping and random cropping with dimensions of 544×704 for the CamVid and 512×1024 for the Cityscapes datasets. Histogram equalization and ColorJitter were also used for RGB images. The batch size is set to 8, and the number of traning epochs is set to 600 and 800 for the CamVid and Cityscapes datasets, respectively. We used the CrossEntropy loss function and exploited the class-weighting scheme introduced in [31] due to the class imbalance of the CamVid dataset. The momentum is 0.9, and the weight decay is 0.0005 for the SGD optimizer. We employ a cosine annealing schedule with an initial learning rate of 0.05. All reported accuracy results are based on the mean intersection over union (mIoU) metric.

C. Comparison With SOTA Models

The quantitative results of our method and other models on the CamVid dataset are compared in Table I. The results show that our model outperforms other state-of-the-art methods. It is worth noting that some methods used a combination of training and evaluation sets to train their models, while our model was able to reach this result using only the training set and without pretraining on the additional driving scene datasets. The DCMNet model achieves about a 3.19% improvement in mIoU over [32], the best RGB-based method, with an mIoU of 77.39%. To have a fair comparison with other RGB-D methods, we have trained some RGB-D models on the RGB-depth CamVid dataset. As shown in Table II, we also achieve the highest IoU for 8 out of 11 classes compared to other methods. Fig. 5 presents the qualitative results obtained on the CamVid dataset, showing the remarkable semantic segmentation performance of our proposed method.

Table III illustrates the performance of our proposed DCMNet on the Cityscapes. Achieving a significant 82.8% mIoU, DCMNet surpasses the state-of-the-art RGB-D model, CMX. Importantly, this performance is achieved with reduced parameter and computational complexity compared to CMX, which employs the MiT-B4 vision transformer. Furthermore, some of the qualitative results obtained on the CityScapes dataset are presented in Fig. 6, demonstrating the substantial semantic segmentation performance of our proposed method.

TABLE II
DCMNet class IoU comparison on CamVid. *: RGB-D method that are well-trained on RGB-depth CamVid dataset.

Models	**Building**	**Tree**	**Sky**	**Car**	**Sign Symbol**	**Road**	**Pedestrian**	**Fence**	**Pole**	**Sidewalk**	**Bicyclist**	**mIoU**
LBN-AA [35]	83.2	70.5	92.5	81.7	51.6	93	55.6	53.2	36.3	82.1	47.9	68
BiSeNet1 [33]	83	75.8	92	83.7	46.5	94.6	58.8	53.6	31.9	81.4	54	68.7
S2-FPN18 [32]	83	77.2	91.8	88.9	48.2	95.7	56.4	43.4	32.4	84.8	62.5	69.5
S2-FPN34 [32]	85.3	77.4	91.7	91.2	49.6	95.7	59.1	46.8	33.2	85.4	66.5	71
S2-FPN34M [32]	86	78.8	92.6	**92.2**	56.2	96	67.1	47.3	42.1	**86.8**	70.7	74.2
ACNet-R50* [1]	88.3	80.2	93.4	90.1	61.6	**96.1**	70.2	56.6	47.6	**86.8**	68.38	76.30
DCMNet (Ours)	**88.5**	**80.6**	**93.6**	90.9	**64.4**	95.5	**71.4**	**60.5**	**48.9**	85.4	**71.5**	**77.39**

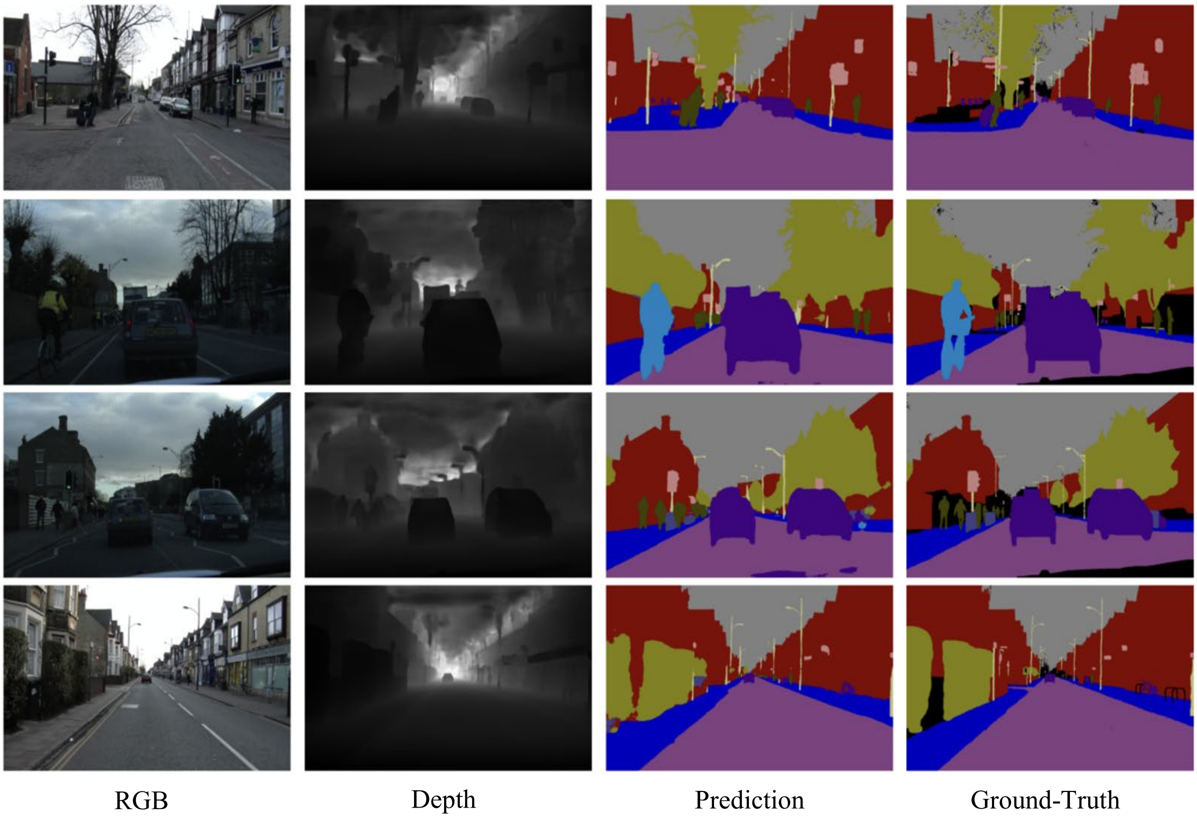


Fig. 5. Qualitative results of our DCMNet on CamVid dataset. Our model excels at segmenting intricate details such as fine and small objects in cluttered backgrounds. From left to right: RGB, Depth, Prediction, and Ground-truth.

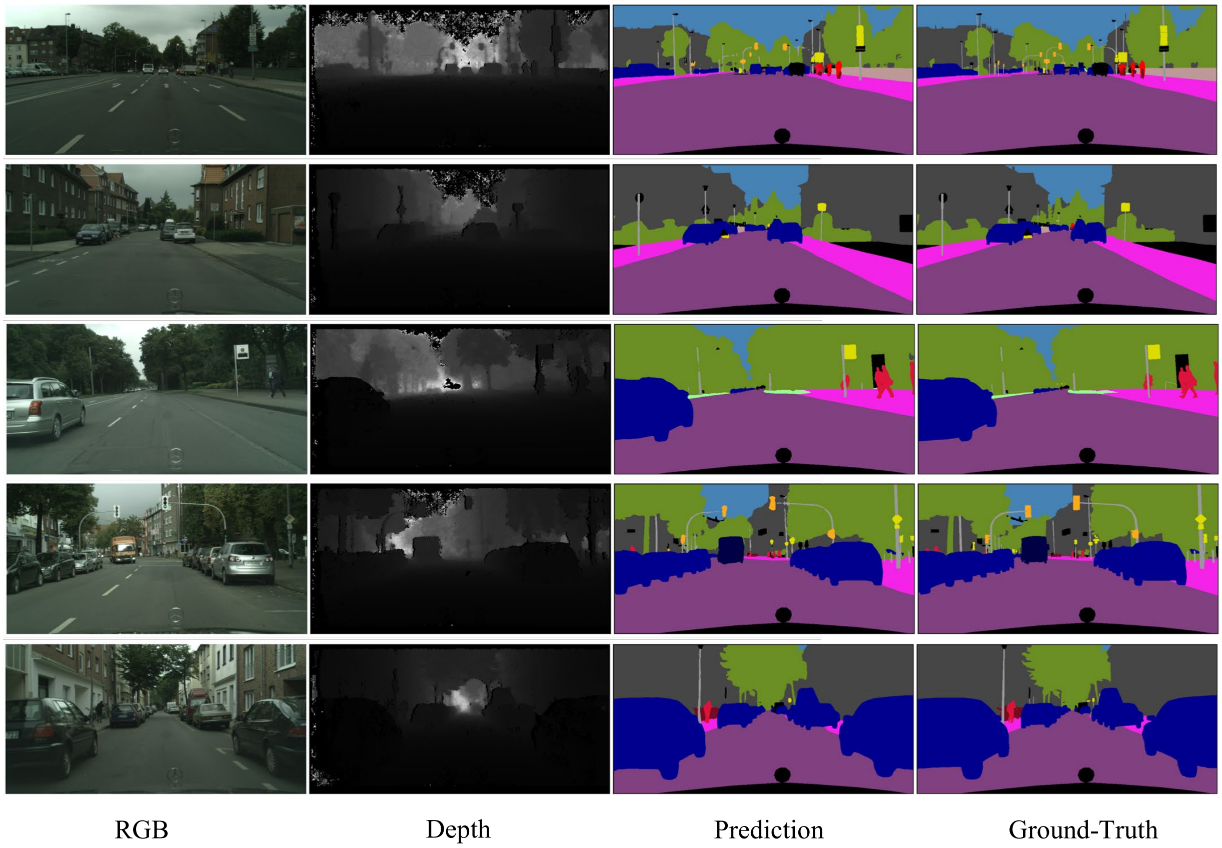


Fig. 6. Qualitative results of our DCMNet on Cityscapes dataset. From left to right: RGB, Depth, Prediction, and Ground-truth.

TABLE III

Comparison of our model with other methods on Cityscapes *val* set. FLOPS are estimated for inputs with a size of 512 × 1024. *: methods that are well-trained on Cityscapes dataset. †: methods that are re-tested on Cityscapes dataset.

Model	Backbone	Params (M)	FLOPs (G)	mIoU (%)
RDFormer [39]	-	-	-	78.3
SA-Gate† [2]	ResNet-50	65.97	406	80.7
SA-Gate† [2]	ResNet-101	110.6	588.7	81.7
ABFNet [8]	MiT-B4	-	-	80.9
ACNet-R50* [1]	ResNet-50	116.6	394.6	81.1
CLGFormer [27]	ResNet-34	-	-	81.4
CMX† [7]	MiT-B2	66.57	228.5	81.6
CMX† [7]	MiT-B4	139.86	394.7	82.6
DCMNet (Ours)	ResNet-50	133.52	394.7	**82.8**

D. Ablation Studies

We conduct ablation studies on the CamVid and Cityscapes datasets. Initially, we evaluate the effectiveness of the main components of our proposed DCMNet. Thereafter, we compare the DCMAF module with several fusion strategies that have publicly available code and are compatible with our model architecture.

1) Effectiveness of Main Components: We verify the contributions of each component within the proposed DCMNet, the results are shown in Table IV. We evaluate the results in four cases. In the first case, we use the ACNet [1] model structure, which includes the ACM module prior to the fusion operation, resulting in the mIoU of 76.30%. In the second case, we remove the ACM modules and employ the DCMAF module along with the channel attention module to fuse the features of the two parallel encoders. This represents an increase of 0.64%. In the third case, we leverage the spatial attention module in the skip connections. This module improves the results by 0.23% by transferring more salient features to the decoder. Finally, in our proposed DCMNet model, the DCMAF module is used in both the encoder and decoder. This resulted in an improvement of 0.22% over the third case. In this case, instead of simple fusion with addition operations, the DCMAF module is used, which has an attention-based fusion approach. This leads to a weighted aggregation of feature maps from different levels of the decoder and transferred features from the encoder, resulting in a better reconstruction of the image information in the decoder.

2) Effectiveness of DCMAF: As shown in Table V, we evaluate the performance of DCMAF against Summation (element-wise addition), the Gated Fusion from LSD-GF [25], the Attention Complementary Module (ACM) from ACNet [1], the Separation-and-Aggregation Gate (SA-Gate) from [2], and the Cross-Modal Feature Rectification Module (CM-FRM) from CMX [7] on the Cityscapes validation set. Notably, DCMAF achieves an mIoU of 82.8%, demonstrating competitive performance against CM-FRM, which has the highest accuracy, with fewer FLOPs and parameters. Specifically, DCMAF maintains the lowest FLOPs among the attention-based methods at 394.7G, comparable to the simple summation approach, while significantly outperforming it in mIoU. These results confirm the DCMAF module's ability to effectively fuse RGB and depth features while maintaining a favorable tradeoff between performance and computational overhead.

TABLE IV

An ablation study for the effectiveness of different modules on CamVid. ACM: Attention-Complementary Module [1], CAM: Channel Attention Module, SAM: Spatial Attention Module, DCMAF-En: Discriminative Cross-Modal Attention Fusion module in the encoder, and DCMAF-De: Discriminative Cross-Modal Attention Fusion module in the decoder

ACM [1]	CAM	SAM	DCMAF-En	DCMAF-De	mIoU(%)
✓					76.30
	✓		✓		76.94
	✓	✓	✓		77.17
	✓	✓	✓	✓	**77.39**

TABLE V

Comparison of DCMAF with various fusion modules on Cityscapes *val* set. FLOPS are estimated for inputs with a size of 512 × 1024.

Fusion Module	Params (M)	FLOPs (G)	mIoU(%)
Summation [4,5]	122.36	394.6	79.5
Gated Fusion [25]	146.72	411.3	80.4
ACM [1]	122.36	394.7	81.5
SA-Gate [2]	137.63	398.9	82.4
CM-FRM [7]	278.49	431.5	82.9
DCMAF (Ours)	133.52	394.7	82.8

E. Discussion

The proposed DCMAF module is well-suited for optimally attending to RGB and depth features. Deploying this module enables the extraction and processing of salient cross-modal features at multiple levels, which significantly improves the segmentation of challenging regions, particularly fine and small objects. As shown in Table II, the proposed model has considerably improved on classes with fine structures (e.g., Fence, Bicyclist, Pole, Sign symbol, and Pedestrian). The DCMAF module can be integrated into various multi-modal deep network architectures without requiring significant modifications. Although the proposed model exhibits improved performance on the CamVid dataset, its performance on the Tree, Road, and Sidewalk classes is less effective. This could be due to depth estimation errors in the model [12]. Additionally, the use of three encoder branches increases computational expense. Fig. 7 shows some qualitative examples of the pros and cons of the proposed model on the CamVid dataset.

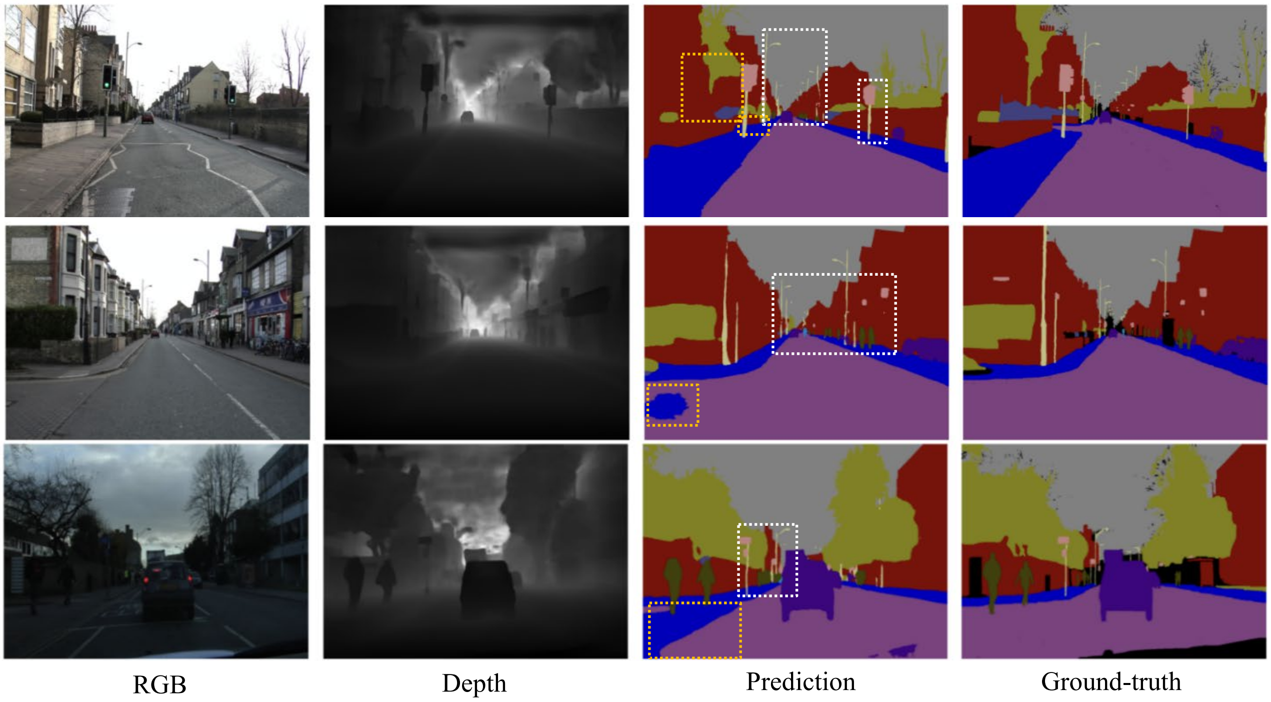


Fig. 7. Visualization of the pros and cons of DCMNet on CamVid. The white and orange boxes indicate pros and cons, respectively.

5- Conclusion

In this paper, we introduce a new module called Discriminative Cross-modal Attention Fusion (DCMAF), which can effectively apply attention-based fusion to discriminate between the RGB and depth modalities. The main idea of the proposed module is to learn a discriminative attention weight from both modalities at each level of the encoder. These weights capture the complementary nature of the modalities and make the model suitable for predicting small and fine objects. We present a new RGB-D semantic segmentation model, Discriminative Cross-Modal Network (DCMNet), by efficiently using the channel- and spatial-wise attention modules, along with the DCMAF module. We evaluate our proposed method on the CamVid and Cityscapes datasets. Since the CamVid dataset lacks depth information, we use the DPT-Hybrid to generate depth images. Experimental results illustrate the effectiveness of our proposed method in enhancing RGB-D semantic segmentation performance. The combination of these advantages makes the DCMNet model ideal for detailed, fine-grained predictions in autonomous driving and robotics applications.

6- References

[1] X. Hu, K. Yang, L. Fei, and K. Wang. (2019, Sep.). ACNet: Attention Based Network to Exploit Complementary Features for RGBD Semantic Segmentation. *IEEE International Conference on Image Processing (ICIP)*. [Online]. Available: https://doi.org/10.1109/ICIP.2019.8803025

[2] X. Chen, K. Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng. (2020, Aug.). Bi-directional Cross-Modality Feature Propagation with Separation-and-Aggregation Gate for RGB-D Semantic Segmentation. European Conference on Computer Vision. [Online]. Available: https://doi.org/10.1007/978-3-030-58621-8_33

[3] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H. M. Gross. (2021, May.). Efficient RGB-D semantic segmentation for indoor scene analysis. In 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 13525–13531. [Online]. Available: https://doi.org/10.1109/ICRA48506.2021.9561675

[4] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. (2016, Nov.). Fusenet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In Asian Conference on Computer Vision, pp. 213–228. [Online]. Available: https://doi.org/10.1007/978-3-319-54181-5_14

[5] J. Jiang, L. Zheng, F. Luo, and Z. Zhang. (2018, Jun.). Rednet: Residual encoder-decoder network for indoor RGB-D semantic segmentation. *arXiv preprint*. [Online]. Available: https://doi.org/10.48550/arXiv.1806.01054

[6] Y. Zhang, Y. Yang, C. Xiong, G. Sun, and Y. Guo. (2022, Jan.). Attention-based dual supervised decoder for RGBD semantic segmentation. *arXiv preprint*. [Online]. Available: https://doi.org/10.48550/arXiv.2201.01427

[7] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen. (2023, Dec.). CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers. *IEEE Transactions on Intelligent Transportation Systems*. [Online]. 24(12), pp. 14679–14694. Available: https://doi.org/10.1109/TITS.2023.3300537

[8] L. Zhong, C. Guo, J. Zhan, and J. Deng. (2024, Dec.). Attention-based fusion network for RGB-D semantic

segmentation. *Neurocomputing*. [Online]. 608, p. 128371. Available: https://doi.org/10.1016/j.neucom.2024.128371

[9] Y. Zhang, C. Xiong, J. Liu, X. Ye, and G. Sun. (2023, Aug.). Spatial information-guided adaptive context-aware network for efficient RGB-D semantic segmentation. *IEEE Sensors Journal*. [Online]. 23(19), pp. 23512–23521. Available: https://doi.org/10.1109/JSEN.2023.3304637

[10] G. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. (2008). Segmentation and recognition using structure from motion point clouds. In Proceedings of the 10th European Conference on Computer Vision (ECCV), pp. 44–57. [Online]. Available: https://doi.org/10.1007/978-3-540-88682-2_5

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. (2016). The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213–3223. [Online]. Available: http://openaccess.thecvf.com

[12] R. Ranftl, A. Bochkovskiy, and V. Koltun. (2021). Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12179–12188. [Online]. Available: http://openaccess.thecvf.com

[13] J. Long, E. Shelhamer, and T. Darrell. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440. [Online]. Available: http://openaccess.thecvf.com

[14] V. Badrinarayanan, A. Kendall, and R. Cipolla. (2017, Jan. 2). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. [Online]. 39(12), pp. 2481–2495. Available: https://doi.org/10.1109/TPAMI.2016.2644615

[15] O. Ronneberger, P. Fischer, and T. Brox. (2015, Oct.). U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, Proceedings, Part III, vol. 18, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28

[16] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. (2017). Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881–2890. [Online]. Available: http://openaccess.thecvf.com

[17] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. (2017, Apr.). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. [Online]. 40(4), pp. 834–848. Available: https://doi.org/10.1109/TPAMI.2017.2699184

[18] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *In Proceedings of the European Conference on Computer Vision (ECCV)*. [Online]. pp. 801–818. Available: http://openaccess.thecvf.com

[19] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. (2019). Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [Online]. pp. 3146–3154. Available: http://openaccess.thecvf.com

[20] Z. Zhong, Z. Q. Lin, R. Bidart, X. Hu, I. B. Daya, Z. Li, W. S. Zheng, J. Li, and A. Wong. (2020). Squeeze-and-attention networks for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13065–13074. [Online]. Available: http://openaccess.thecvf.com

[21] H. Li, P. Xiong, J. An, and L. Wang. (2018, May.). Pyramid attention network for semantic segmentation. *arXiv preprint*. [Online]. Available: https://doi.org/10.48550/arXiv.1805.10180

[22] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. (2021, Dec. 6). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*. [Online]. 34, pp. 12077–12090. Available: https://proceedings.neurips.cc

[23] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, and L. Zhang. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6881–6890. [Online]. Available: http://openaccess.thecvf.com

[24] J. Wang, Z. Wang, D. Tao, S. See, and G. Wang. (2016, Oct.). Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks. In Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, Proceedings, Part V, vol. 14, pp. 664–679. [Online]. Available: https://doi.org/10.1007/978-3-319-46454-1_40

[25] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang. (2017). Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3029–3037. [Online]. Available: http://openaccess.thecvf.com

[26] D. Qashqai, E. Mousavian, S. B. Shokouhi, and S. Mirzakuchaki. (2024, Jul.). CSFNet: A Cosine Similarity Fusion Network for Real-Time RGB-X Semantic Segmentation of Driving Scenes. *arXiv preprint*. [Online]. Available: https://doi.org/10.48550/arXiv.2407.01328

[27] T. Li, Q. Zhou, D. Wu, M. Sun, and T. Hu. (2024, May.). CLGFormer: Cross-Level-Guided Transformer for RGB-D Semantic Segmentation. *Multimedia Tools and Applications*. [Online]. pp. 1–23. Available: https://doi.org/10.1007/s11042-024-19051-9

[28] K. He, X. Zhang, S. Ren, and J. Sun. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. [Online]. Available: http://openaccess.thecvf.com

[29] A. Geiger, P. Lenz, and R. Urtasun. (2012, Jun.). Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3354–3361. [Online]. Available: https://doi.org/10.1109/CVPR.2012.6248074

[30] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. (2009, Jun.). ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255. [Online]. Available: https://doi.org/10.1109/CVPR.2009.5206848

[31] S. Y. Lo, H. M. Hang, S. W. Chan, and J. J. Lin. (2019, Dec.). Efficient dense modules of asymmetric convolution for real-time semantic segmentation. In Proceedings of the 1st ACM International Conference on Multimedia in Asia, pp. 1–6. [Online]. Available: https://doi.org/10.1145/3338533.3366558

[32] M. A. Elhassan, C. Yang, C. Huang, T. L. Munea, X. Hong, A. Adam, and A. Benabid. (2022, Jun.). S^2-FPN: Scale-aware Strip Attention Guided Feature Pyramid Network for Real-time Semantic Segmentation. *arXiv preprint*. [Online]. Available: https://doi.org/10.48550/arXiv.2206.07298

[33] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. (2018). BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 325–341. [Online]. Available: http://openaccess.thecvf.com

[34] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. (2018). ICNet for real-time semantic segmentation on high-resolution images. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 405–420. [Online]. Available: http://openaccess.thecvf.com

[35] G. Dong, Y. Yan, C. Shen, and H. Wang. (2020, Mar.). Real-time high-performance semantic image segmentation of urban street scenes. *IEEE Transactions on Intelligent Transportation Systems*. [Online]. 22(6), pp. 3258–3274. Available: https://doi.org/10.1109/TITS.2020.2980426

[36] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang. (2021, Nov.). BiSeNet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*. [Online]. 129, pp. 3051–3068. Available: https://doi.org/10.1007/s11263-021-01515-2

[37] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei. (2021). Rethinking BiSeNet for real-time semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9716–9725. [Online]. Available: http://openaccess.thecvf.com

[38] W. Zhou, E. Yang, J. Lei, and L. Yu. (2022, May.). FRNet: Feature reconstruction network for RGB-D indoor scene parsing. *IEEE Journal of Selected Topics in Signal Processing*. [Online]. 16(4), pp. 677–687. Available: https://doi.org/10.1109/JSTSP.2022.3174338

[39] Z. Peng, Y. Zheng, Y. Cheng, and Y. Qiao. (2024, Oct.). RDFormer: Efficient RGB-D Semantic Segmentation in Complex Outdoor Scenes. *In* Proceedings of the 2024 5th International Conference on Machine Learning and Computer Application (ICMLCA), pp. 170–175. [Online]. Available: https://doi.org/10.1109/ICMLCA63499.2024.10754213