DepthMatch: Semi-Supervised RGB-D Scene Parsing Through Depth-Guided Regularization

Jianxin Huang, *Student Member, IEEE*, Jiahang Li, *Graduate Student Member, IEEE*, Sergey Vityazev, *Senior Member, IEEE*, Alexander Dvorkovich, *Member, IEEE*, and Rui Fan, *Senior Member, IEEE*

***Abstract*—RGB-D scene parsing methods effectively capture both semantic and geometric features of the environment, demonstrating great potential under challenging conditions such as extreme weather and low lighting. However, existing RGB-D scene parsing methods predominantly rely on supervised training strategies, which require a large amount of manually annotated pixel-level labels that are both time-consuming and costly. To overcome these limitations, we introduce DepthMatch, a semi-supervised learning framework that is specifically designed for RGB-D scene parsing. To make full use of unlabeled data, we propose complementary patch mix-up augmentation to explore the latent relationships between texture and spatial features in RGB-D image pairs. We also design a lightweight spatial prior injector to replace traditional complex fusion modules, improving the efficiency of heterogeneous feature fusion. Furthermore, we introduce depth-guided boundary loss to enhance the model's boundary prediction capabilities. Experimental results demonstrate that DepthMatch exhibits high applicability in both indoor and outdoor scenes, achieving state-of-the-art results on the NYUv2 dataset and ranking first on the KITTI Semantics benchmark.**

***Index Terms*—Heterogeneous features, semi-supervised learning, RGB-D scene parsing.**

I. Introduction

SCENE parsing task aims to assign each pixel in an image to a specific category and has broad applications in fields such as autonomous driving and robotics [1], [2], [3], [4]. With the widespread adoption of deep learning techniques, deep learning models trained via supervised learning have shown significant performance improvements in various image segmentation tasks compared to traditional geometry-based methods [5], [6], [7]. However, single-modality networks relying solely on RGB images experience a substantial performance drop under challenging conditions, such as poor lighting or adverse weather [8], [9], [10]. To tackle this problem, FuseNet [11] first explores the fusion of heterogeneous features derived from RGB and depth images, leveraging spatial information from depth data to improve scene parsing performance under extreme conditions. To fully exploit the complementary information in heterogeneous features, RoadFormer series [8], [12] designs a novel attention-based feature fusion strategy, significantly enhancing the capability of feature integration. DPLNet [13] introduces a lightweight multimodal prompt learning module to fuse RGB-D features, reducing training overhead while maintaining superior results.

Although depth information significantly improves scene parsing performance, existing RGB-D scene parsing methods still primarily rely on supervised training strategies to achieve satisfactory results, which require extensive pixel-level annotations. This limitation greatly hinders the deployment of advanced models, especially when sufficient annotations are unavailable [14]. Semi-supervised learning reduces annotation reliance by leveraging inexpensive unlabeled data to enhance scene parsing performance. Recent semi-supervised approaches commonly adopt pseudo-labeling frameworks based on consistency regularization [15], where the model is required to produce consistent predictions under different augmentations or perturbations applied to the unlabeled data, thus improving the model's applicability across diverse conditions. In this context, CowMix [16] demonstrates that combining Cutout [17] and CutMix [18] into unlabeled images is an effective regularization technique. UniMatch [19] inherits the weak-to-strong consistency regularization method from FixMatch [20], constructing a simple yet effective regularization framework. Unfortunately, in the field of RGB-D scene parsing, semi-supervised training strategies remain underexplored.

To overcome the aforementioned challenges, we propose DepthMatch, a semi-supervised learning framework specifically designed for RGB-D scene parsing. Our method is built on consistency regularization, training the model to generate consistent

Received 11 April 2025; accepted 13 May 2025. Date of publication 2 June 2025; date of current version 8 July 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62388101, Grant 62473288, and Grant 62233013; in part by the Fundamental Research Funds for the Central Universities, NIO University Program (NIO UP); and in part by Xiaomi Young Talents Program. The associate editor coordinating the review of this article and approving it for publication was Dr. Aykut Koc. *(Corresponding author: Rui Fan.)*

Jianxin Huang, Jiahang Li, and Rui Fan are with the College of Electronics & Information Engineering, Tongji University, Shanghai 201804, China, also with the Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 201804, China, also with the Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, China, also with the State Key Laboratory of Intelligent Autonomous Systems, Tongji University, Shanghai 201804, China, and also with the Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, China (e-mail: jasonhuang@tongji.edu.cn; lijiahang617@tongji.edu.cn; rui.fan@ieee.org).

Sergey Vityazev is with Ryazan State Radio Engineering University, 390035 Ryazan, Russian Federation (e-mail: vityazev.s.v@ieee.org).

Alexander Dvorkovich is with the Multimedia Technology and Telecom Department, Telecommunications Center, Moscow Institute of Physics and Technology, 141701 Dolgoprudny, Moscow Region, Russia (e-mail: dvork.alex@gmail.com).

Our source code will be publicly available at mias.group/DepthMatch.

Digital Object Identifier 10.1109/LSP.2025.3575640

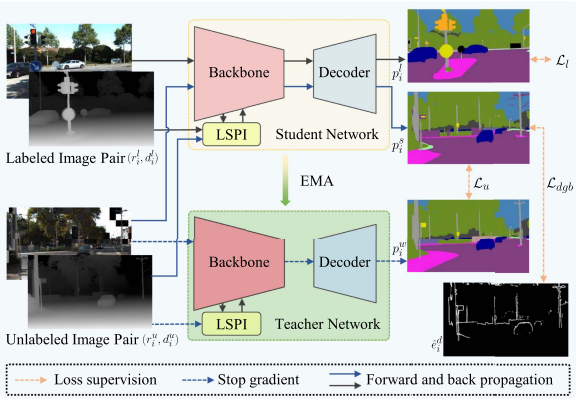

Fig. 1. An overview of our proposed DepthMatch.

predictions on the same data under various strong augmentations to fully leverage unlabeled RGB-D image pairs. Specifically, we design Complementary Patch Mix-up Augmentation strategy (CPMA) to randomly replace the color information with geometric information, encouraging the model to further explore the texture and spatial features in the input data. Considering that RGB images usually contribute more than depth images in RGB-D scene parsing tasks, we propose a Lightweight Spatial Prior Injector (LSPI) to replace traditional complex fusion modules, achieving efficient integration of heterogeneous features. Additionally, we introduce the Depth-Guided Boundary Loss (DGBL) to guide the model's boundary learning on unlabeled data. Experimental results demonstrate that DepthMatch, incorporating all these innovative components and strategies, achieves state-of-the-art (SoTA) performance on the NYUv2 and KITTI datasets, validating its strong advantages and high applicability across both indoor and outdoor scenes.

Our contributions can be summarized as follows:

- We introduce CPMA, a regularization strategy specifically designed for RGB-D scene parsing tasks, which encourages the model to better explore the relationships between texture and spatial features.
- We propose LSPI to achieve more efficient feature fusion. It efficiently injects spatial information from depth images into RGB features during the encoding phase.
- We propose DGBL, which imposes boundary supervision on unlabeled data based on contour information extracted from depth images.

II. Methodology

A. Overall Training Pipeline

The overall architecture of our proposed DepthMatch is illustrated in Fig. 1. Following [15], we utilize the pre-trained DINOv2 [21] model as our encoder, leveraging its self-supervised contrastive learning approach to enhance the quality and generalization of the extracted features. Meanwhile, we employ the Transformer-based DPT [22] as the decoder to generate globally consistent predictions. Unlike traditional supervised RGB-D scene parsing methods [1], [23], [24], DepthMatch simultaneously utilizes a labeled dataset $D_l = \{(r_i^l, d_i^l, m_i^l)\}$ and a larger unlabeled dataset $D_u = \{(r_i^u, d_i^u)\}$ for training, where r_i and d_i represent the i-th RGB-D image pair, and m_i is the corresponding semantic ground truth mask. Following [25], we employ an EMA teacher with the same architecture to generate stable and improved pseudo-labels for unlabeled data. During training, each batch for both the teacher and student networks consists of B_l labeled image pairs and B_u unlabeled image pairs. For labeled data, we use manually annotated labels to supervise model training. The supervised loss $\mathcal{L}_l$ for labeled data is defined as:

$$\mathcal{L}_l = \frac{1}{B_l} \sum_{i=1}^{B_l} \mathcal{L}_{cls}(p_i^l, m_i^l), \tag{1}$$

where p_i^l represents the model's prediction for the i-th labeled image pair, m_i^l is the corresponding ground truth mask, and $\mathcal{L}_{cls}$ is the pixel-wise cross-entropy loss for class prediction. For unlabeled image pairs, the model first generates pseudo-labels p_i^w on weakly augmented image pairs (r_i^w, d_i^w), which are then used to supervise the learning of their strongly augmented counterparts (r_i^s, d_i^s). During semi-supervised training, weak augmentations include random scaling and random flipping of RGB-D image pairs, while strong augmentations encompass our proposed CPMA as well as color jittering, grayscaling, and Gaussian blurring. By training the model to make invariant predictions under various data augmentations, it can learn and extract latent patterns and features from unlabeled images [20]. The unsupervised loss $\mathcal{L}_u$ thus can be formulated as:

$$\mathcal{L}_u = \frac{1}{B_u} \sum_{i=1}^{B_u} \mathbb{1}(\max(p_i^w) \geq \tau) \mathcal{L}_{cls}(p_i^s, \hat{p}_i^w), \tag{2}$$

where $\hat{p}_i^w$ is the pseudo-label derived from p_i^w using the argmax operation, p_i^s represents the model's prediction for (r_i^s, d_i^s), and $\mathbb{1}(\max(p_i^w) \geq \tau)$ introduces a predefined confidence threshold $\tau = 0.95$ to reduce noise in pseudo-labels by excluding those that do not meet the threshold from training [15]. This ensures that during the early training iterations, the model primarily learns from high-quality manually labeled images, gradually expanding to high-confidence pseudo-labeled images as training progresses [20].

B. Complementary Patch Mix-Up Augmentation

The consistency regularization-based semi-supervised algorithms require the model to make consistent predictions on unlabeled images under different data augmentations [19], [20]. However, simply transplanting strong data augmentation methods designed for RGB images to RGB-D data often leads to suboptimal results, as it overlooks the inherent modality differences between RGB-D image pairs [14]. To address this issue, we design a simple yet effective CPMA based on the modality-specific characteristics of RGB-D data. Previous studies have demonstrated that RGB images contain rich color and texture information, while depth images provide accurate geometric cues that are complementary in terms of detail and structure [26], [27]. By swapping certain patches between the RGB image and its corresponding depth image during strong augmentation, CPMA replaces the original color and texture information of

the patches with geometric information. This encourages the model to better explore the spatial features within the input data, resulting in improved neural representations [28]. Specifically, CPMA can be formulated as follows:

$$\hat{r}_i^u = M \cdot r_i^u + \hat{M} \cdot d_i^u, \tag{3}$$

where M and $\hat{M} = 1 - M$ form a pair of complementary random patch masks. Following [29], we set the patch size of the mask to 32 pixels, and the masking ratio for the RGB image will be further discussed in Section III-D.

C. Lightweight Spatial Prior Injector

Feature fusion modules that enhance and integrate RGB and depth features are crucial in RGB-D scene parsing networks [30]. However, using complex feature fusion modules to symmetrically integrate heterogeneous features often introduces unnecessary overhead parameters [13]. Inspired by [13], we propose an LSPI to progressively inject spatial priors from the depth image into RGB features during the encoding process, achieving effective fusion of heterogeneous features. Specifically, the RGB-D image pair (r_i, d_i) is encoded into $(\boldsymbol{F}_i^r, \boldsymbol{F}_i^d)$ using two patch embedding layers. The heterogeneous embeddings $(\boldsymbol{F}_i^r, \boldsymbol{F}_i^d)$ are then downsampled to inject depth priors into the RGB embeddings, and the fused embeddings are then upsampled to their original dimensions. The LSPI can be formulated as:

$$\hat{\boldsymbol{F}}_i^r = w_f\left(w_r(\boldsymbol{F}_i^r) + w_d(\boldsymbol{F}_i^d)\right) + \boldsymbol{F}_i^r, \tag{4}$$

where the linear layers w_r and w_d reduce the feature dimensions of $\boldsymbol{F}_i^r$ and $\boldsymbol{F}_i^d$ to one-quarter of their original size, and the linear layer w_f fuses these heterogeneous embeddings and restores them to their original dimensions. The enhanced embedding $\hat{\boldsymbol{F}}_i^r$ is then processed through the encoder and decoder to produce the final predictions.

D. Depth-Guided Boundary Loss

Recent studies [31] have demonstrated that introducing boundary supervision can improve model performance in handling complex object boundaries. This improvement arises because cross-entropy loss independently computes the classification accuracy of each pixel, lacking direct guidance for locating and precisely segmenting boundary pixels [32]. In contrast, boundary supervision encourages the model to focus more on the distribution of boundary pixels [31]. Unfortunately, existing boundary supervision methods [31], [32], [33] rely on boundary information extracted from masks, making them unavailable for unlabeled images in semi-supervised learning. To overcome these limitations, we propose the DGBL to further guide the model in aligning predicted boundaries with ground-truth boundaries on unlabeled images. More concretely, we compute the boundary maps e_i^p and e_i^d from the model's predicted mask p_i^u and the corresponding depth image d_i^u, respectively, and binarize the resulting boundary maps using a threshold of 0.1 to obtain $\hat{e}_i^p$ and $\hat{e}_i^d$. Finally, we use the mean squared error to measure the difference between the predicted and ground truth boundaries. The DGBL can be formulated as follows:

$$\mathcal{L}_{dgb} = \frac{1}{N}\sum_{j=1}^{N} y_{ij}^p \cdot (y_{ij}^p - y_{ij}^d)^2, \tag{5}$$

where y_{ij}^p and y_{ij}^d represent the boundary values in $\hat{e}_i^p$ and $\hat{e}_i^d$, respectively, and N is the total number of pixels in the image. By multiplying the loss for each pixel by y_{ij}^p, we ensure that the loss is only calculated for boundary pixels in the predicted mask map.

E. Overall Loss Function

At the beginning of training, the model often struggles to generate pseudo labels with well-defined boundaries [19]. As training progresses, the quality of these pseudo labels improves. Therefore, we gradually decrease the weight of $\mathcal{L}_{\text{dgb}}$ during training to increase the model's focus on $\mathcal{L}_u$ for unlabeled data. Inspired by previous studies [34], we introduce a weight decay strategy to dynamically adjust the weight of $\mathcal{L}_{\text{dgb}}$. The time-dependent function is defined as $f(t) = 1 - \frac{t-1}{t_{\max}}$, where $t = (1, 2, \ldots, t_{\max})$ represents the current training epoch, and $t_{\max}$ denotes the maximum training epoch. For each batch of training data, we minimize the following loss function:

$$\mathcal{L}_{total} = \mathcal{L}_l + \lambda_u\left(\mathcal{L}_u + f(t)\mathcal{L}_{dgb}\right), \tag{6}$$

where λ_u balances the influence of the unlabeled data, while $f(t)$ controls the weight of $\mathcal{L}_{dgb}$ and progressively decays during training.

III. Experiments

A. Datasets and Evaluation Metrics

We conduct experiments on RGB-D scene parsing datasets for both indoor and outdoor environments. For indoor scenes, we used the popular NYUv2 dataset, which consists of 1,449 labeled RGB-D image pairs. Following the standard partition, the dataset was divided into 795 training samples and 654 testing samples. All images are resized to a resolution of 630 × 476 pixels. On the other hand, we evaluate the performance of DepthMatch in outdoor scenarios using the KITTI Semantics [35] dataset. This dataset contains 200 labeled images across 19 categories and an additional 4,000 unlabeled images. All images are resized to a resolution of 1,274 × 378 pixels, with depth data obtained using ViTAStereo [36]. We evaluate model performance using the commonly adopted metrics: mean Intersection over Union (mIoU) and instance Intersection over Union (iIoU). We also use model parameters (Params) and floating-point operations (FLOPs) to assess the model efficiency. Additionally, the evaluation metrics used for the KITTI Semantics benchmarks are available on the official webpage: www.cvlibs.net/datasets/kitti/eval_semseg.php?benchmark=semantics2015.

B. Implementation Details

Our network is trained using the AdamW optimizer [37], with a polynomial decay strategy for the learning rate [38]. The initial

TABLE I
COMPARISON OF DEPTHMATCH'S PERFORMANCE WITH SOTA ALGORITHMS ON THE TEST SET OF THE NYUV2 DATASET

Method	Publication	Backbone	Params (M) ↓	FLOPs (G) ↓	mIoU (%) ↑
ACNet	ICIP'19 [39]	ResNet-50	116.6	126.7	48.1
FRNet	JSTSP'22 [26]	ResNet-34	-	-	53.6
CMX	T-ITS'23 [40]	MiT-B5	181.1	167.8	56.9
Sigma	WACV'25 [41]	VMamba-S	69.8	138.9	57.0
DFormer	ICLR'24 [24]	DFormer-L	39.0	65.7	57.2
RoadFormer+	T-IV'24 [8]	ConvNeXt-B	152.4	281.5	57.6
FCDENet	IOTJ'25 [27]	MiT-B4	128.9	-	57.6
GeminiFusion	ICML'24 [42]	MiT-B5	137.2	300.5	57.7
SEFNet	SPL'24 [23]	ResNet-101	90.7	118.1	58.5
DPLNet	IROS'24 [13]	MiT-B5	88.6	-	59.3
DepthMatch (1/4)	Ours	DINOv2-S	26.9	68.1	56.6
DepthMatch (1/2)	Ours	DINOv2-S	26.9	68.1	58.7
DepthMatch	Ours	DINOv2-S	26.9	68.1	59.9
DepthMatch*	Ours	DINOv2-S	26.9	68.1	**61.4**

* Using SUN-RGBD [43] training set as unlabeled data.

TABLE II
COMPARISON ON THE KITTI SEMANTICS BENCHMARK IN TERMS OF MIOU AND IIOU FOR BOTH CLASS AND CATEGORY

Method	Class		Category		Rank
	mIoU (%)	iIoU (%)	mIoU (%)	iIoU (%)	
VideoProp [44]	72.82	48.68	88.99	**75.26**	5
RoadFormer+ [8]	73.13	45.88	88.75	73.46	4
UJS [45]	75.11	47.71	89.53	75.75	3
WRP [46]	76.44	**50.92**	89.63	73.69	2
DepthMatch	**76.60**	48.48	**90.16**	75.05	**1**

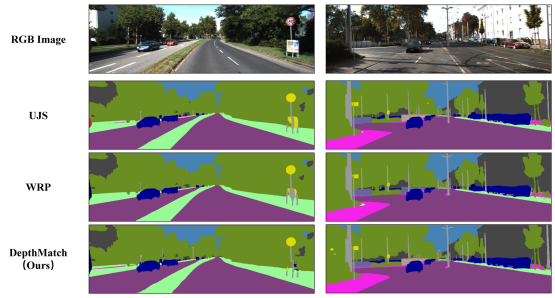

Fig. 2. Qualitative comparison on the KITTI Semantics dataset. The results are produced by the official KITTI online benchmark suite.

TABLE III
ABLATION STUDY OF MASKING RATIO FOR CPMA

Masking Ratio	0	0.05	0.1	0.15	0.2
mIoU (%) ↑	60.2	60.4	**60.9**	60.5	60.5

TABLE IV
ABLATION STUDY OF THE MODULE PART OF DEPTHMATCH

Data Type	LSPI	CPMA	DGBL	Params (M) ↓	mIoU (%) ↑
RGB	×	×	×	24.9	59.6
RGB-D	✓	×	×	26.9	60.2
RGB-D	✓	✓	×	26.9	60.9
RGB-D	✓	×	✓	26.9	60.8
RGB-D	✓	✓	✓	26.9	**61.4**

learning rate is set to 2×10^{-4} with a weight decay of 10^{-2}, and learning rate multipliers of 2.5×10^{-2} are applied to the weight-sharing backbone. It is worth noting that for the KITTI Semantics, we employed a multi-scale inference strategy with scaling factors of $\{1.05, 1.5, 2.0, 2.5\}$.

C. Comparison With SoTA Networks

To validate the effectiveness of our proposed DepthMatch, we compared it with multiple SoTA RGB-D scene parsing methods on two popular datasets, NYUv2 and KITTI Semantics. Quantitative experimental results demonstrate that DepthMatch significantly outperforms all other SoTA methods with fewer model parameters on the NYUv2 dataset. Even when trained with only 1/4 of the labeled images from the NYUv2 training set, DepthMatch maintains exceptional performance. Additionally, by incorporating supplementary unlabeled data, DepthMatch achieves 61.4% mIoU, demonstrating its effective capability to leverage unlabeled data for enhanced model performance. Notably, DepthMatch yields an inference speed of 74 FPS when processing images with a resolution of 630×476 pixels on an NVIDIA RTX 3090 GPU. To further verify the superiority of our method in outdoor driving scene parsing, we submitted the test results produced by DepthMatch to the KITTI Semantics benchmark for performance comparison. As shown in Table II, DepthMatch ranks first on the KITTI Semantics benchmark. Fig. 2 further presents the qualitative comparison on the KITTI Semantics dataset, demonstrating the outstanding performance of our method in both global scene understanding and local boundary detection. These results fully indicate the robustness and high applicability of DepthMatch across indoor and outdoor scenes.

D. Ablation Studies

All ablation experiments are conducted on the NYUv2 dataset, using the SUN-RGBD training set as unlabeled data. We first investigate the optimal masking ratio for the CPMA. Based on the experimental results in Table III, we set the masking ratio to 0.1. To further validate the effectiveness of our proposed DepthMatch, we perform ablation studies for each contribution, and all results are shown in Table IV. Specifically, we select UniMatchV2 [15] with a DINOv2-S [21] backbone as our baseline model and incrementally add the proposed LSPI, CPMA, and DGBL during training. The ablation results demonstrate that the LSPI module efficiently exploits spatial information from depth images, while the CPMA and DGBL enhance the training on unlabeled data, significantly boosting the model. These results fully validate the effectiveness of each part of the proposed DepthMatch.

IV. CONCLUSION

In this letter, we proposed DepthMatch, a lightweight yet powerful semi-supervised training framework specifically designed for RGB-D scene parsing, aimed at reducing the reliance on manually annotated data. Specifically, we employed the LSPI module to efficiently inject spatial priors from depth images into RGB features. Additionally, we designed CPMA and DGBL to provide more effective regularization strategies and boundary supervision for unlabeled image pairs. Extensive experiments demonstrated that DepthMatch effectively utilized inexpensive unlabeled data to improve model performance, surpassing existing SoTA methods. Our future work will focus on extending our framework to other scene parsing tasks.

REFERENCES

[1] Y. Zhang, W. Zhou, X. Ran, and M. Fang, "Lightweight dual stream network with knowledge distillation for RGB-D scene parsing," *IEEE Signal Process. Lett.*, vol. 31, pp. 855–859, 2024.

[2] R. Fan, H. Wang, P. Cai, and M. Liu, "SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 340–356.

[3] Z. Wu, Y. Feng, C.-W. Liu, F. Yu, Q. Chen, and R. Fan, "S^3 M-Net: Joint learning of semantic segmentation and stereo matching for autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 9, no. 2, pp. 3940–3951, Feb. 2024.

[4] Y. Feng et al., "SNE-RoadSegV2: Advancing heterogeneous feature fusion and fallibility awareness for freespace detection," *IEEE Trans. Instrum. Meas.*, vol. 74, 2025, Art. no. 2512109.

[5] R. Fan and M. Liu, "Road damage detection based on unsupervised disparity map segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4906–4911, Nov. 2020.

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

[7] R. Fan, U. Ozgunalp, B. Hosking, M. Liu, and I. Pitas, "Pothole detection based on disparity transformation and road surface modeling," *IEEE Trans. Image Process.*, vol. 29, pp. 897–908, 2020.

[8] J. Huang et al., "RoadFormer+: delivering RGB-X Scene parsing through scale-aware information decoupling and advanced heterogeneous feature fusion," *IEEE Trans. Intell. Veh.*, early access, Aug. 22, 2024, doi: 10.1109/TIV.2024.3448251.

[9] W. Zhou, X. Lin, J. Lei, L. Yu, and J.-N. Hwang, "MFFENet: Multiscale feature fusion and enhancement network for RGB–thermal urban road scene parsing," *IEEE Trans. Multimedia*, vol. 24, pp. 2526–2538, 2022.

[10] W. Zhou, H. Wu, and Q. Jiang, "MDNet: Mamba-effective diffusion-distillation network for RGB-thermal urban dense prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 4, pp. 3222–3233, Apr. 2025.

[11] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. 13th Asian Conf. Comput. Vis.*, 2017, pp. 213–228.

[12] J. Li, Y. Zhang, P. Yun, G. Zhou, Q. Chen, and R. Fan, "RoadFormer: Duplex transformer for RGB-normal semantic road scene parsing," *IEEE Trans. Intell. Veh.*, vol. 9, no. 7, pp. 5163–5172, Jul. 2024, doi: 10.1109/TIV.2024.3388726.

[13] S. Dong et al., "Efficient multimodal semantic segmentation via dual-prompt learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 14196–14203, 2023, *arXiv:2312.00360*.

[14] J. Wang, G. Li, G. Qiu, G. Ma, J. Xi, and N. Yu, "Depth-assisted semi-supervised RGB-D rail surface defect inspection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 8042–8052, Jul. 2024.

[15] L. Yang et al., "UniMatch V2: Pushing the limit of semi-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 4, pp. 3031–3048, 2024.

[16] G. French, T. Aila, S. Laine, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, high-dimensional perturbations," 2019, *arXiv:1906.01916*.

[17] T. DeVries, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

[18] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6023–6032.

[19] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7236–7246.

[20] K. Sohn et al., "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 596–608.

[21] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, 2023.

[22] R. Ranftl A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12179–12188.

[23] P. Xiang, B. Yao, Z. Jiang, and C. Peng, "Self-enhanced feature fusion for RGB-D semantic segmentation," *IEEE Signal Process. Lett.*, vol. 31, pp. 3015–3019, 2024.

[24] Y. Bowen et al., "DFormer: Rethinking RGBD representation learning for semantic segmentation," in *Proc. Int. Conf. Learn. Representations*, 2024.

[25] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[26] W. Zhou, E. Yang, J. Lei, and L. Yu, "FRNet: Feature reconstruction network for RGB-D indoor scene parsing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 4, pp. 677–687, Jun. 2022.

[27] W. Zhou, B. Jian, and Y. Liu, "Feature contrast difference and enhanced network for RGB-D indoor scene classification in Internet of Things," *IEEE Internet Things J.*, vol. 12, no. 11, pp. 17610–17621, Jun. 2025.

[28] R. Gong, Q. Wang, M. Danelljan, D. Dai, and L. V. Gool, "Continuous pseudo-label rectified domain adaptive semantic segmentation with implicit neural representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7225–7235.

[29] U. Shin, K. Lee, I. S. Kweon, and J. Oh, "Complementary random masking for RGB-thermal semantic segmentation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2024, pp. 11110–11117.

[30] A. Valada, R. Mohan, and W. Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1239–1285, 2020.

[31] C. Wang et al., "Active boundary loss for semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, pp. 2397–2405.

[32] D. Wu, Z. Guo, A. Li, C. Yu, C. Gao, and N. Sang, "Conditional boundary loss for semantic segmentation," *IEEE Trans. Image Process.*, vol. 32, pp. 3717–3731, 2023.

[33] S Qiu, J Chen, H Zhang, R Wan, X Xue, and J. Pu, "Guided contrastive boundary learning for semantic segmentation," *Pattern Recognit.*, vol. 155, 2024, Art. no. 110723.

[34] Z. Yang et al., "Effective whole-body pose estimation with two-stages distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4210–4220.

[35] H. A. Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *Int. J. Comput. Vis.*, vol. 126, pp. 961–972, 2018.

[36] C.-W. Liu, Q. Chen, and R. Fan, "Playing to vision foundation model's strengths in stereo matching," *IEEE Trans. Intell. Veh.*, early access, Sep. 25, 2024, doi: 10.1109/TIV.2024.3467287.

[37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, pp. 1–10, 2019.

[38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[39] X. Hu et al., "ACNET: Attention based network to exploit complementary features for RGBD semantic segmentation," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1440–1444.

[40] J. Zhang et al., "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14679–14694, Dec. 2023.

[41] Z. Wan et al., "Sigma: Siamese Mamba network for multi-modal semantic segmentation," in *Proc. IEEE/CVF Winter Conf. Appl.Comput. Vis.*, pp. 1734–1744, 2024.

[42] D. Jia et al., "Geminifusion: Efficient pixel-wise multimodal fusion for vision transformer," in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 21753–21767.

[43] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 567–576.

[44] Y. Zhu et al., "Improving semantic segmentation via video propagation and label relaxation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8856–8865.

[45] Y. Cai, L. Dai, H. Wang, and Z. Li, "Multi-target pan-class intrinsic relevance driven model for improving semantic segmentation in autonomous driving," *IEEE Trans. Image Process.*, vol. 30, pp. 9069–9084, 2021.

[46] A. Ganeshan et al., "Warp-refine propagation: Semi-supervised auto-labeling via cycle-consistency," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15499–15509.