IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

CAR-Stereo: Confidence-aware Adaptive Disparity Refinement for Real-time Stereo Matching

Chanill Park[1], Janghyun Kim[1], Minseong Kweon[2], and Jinsun Park[3]

***Abstract*— In this paper, we propose a novel real-time disparity refinement method that enables precise structure perception. We construct a compact full-resolution cost volume from residuals around the initial disparity and adaptively eliminate redundant information on a per-pixel basis by leveraging the confidence. The core idea of our method comprises residual cost volume construction and an adaptive range masking strategy. The residual cost volume is constructed from refinement candidates around the initial disparity, based on the assumption that the ground-truth disparity is near the initial disparity. Compared to the conventional cost volume constructed over the entire set of disparity candidates, our approach achieves computational efficiency and maintains precise structural information by operating at full-resolution. Moreover, we propose an adaptive range masking strategy that filters refinement candidates for each pixel by leveraging confidence values. This approach effectively eliminates redundant information present in cost volumes that are composed of uniformly sampled refinement candidates. Experimental results on the Scene Flow and KITTI 2012 benchmarks demonstrate that our method achieves real-time performance and sets a new state-of-the-art among real-time stereo matching algorithms.**

I. Introduction

Recently, depth perception [6], [7], [8], [9], [10] has emerged as a key component in autonomous driving, enabling vehicles to perceive 3D structures and estimate the distances to surrounding obstacles. Among these, stereo matching [6], [11], [12], [13], [14] offers a reliable and geometry-driven solution by estimating the disparity from rectified image pairs. Accurate detection of fine-grained structures (*e.g.*, wire fence, poles, etc.) is essential to ensure the safety and reliability of autonomous vehicles [15], [16].

However, contemporary deep learning methods for stereo matching struggle to adequately balance accuracy with efficiency, particularly for fine-grained or thin-structures. Performance-oriented models [17], [18], [19] process high-resolution cost volumes with 2D or 3D convolutions, thereby improving matching accuracy. However, the inference speed is slow due to the massive computational burden. To reduce the computational burden, efficiency-oriented models [20], [4], [21] typically construct the cost volume at a lower resolution and subsequently upsample disparity map to the original resolution. This downsampling process inevitably leads to the loss of precise structural details and ambiguity in object boundaries, consequently imposing significant constraints on the accurate recognition of fine structures. This challenge of information loss is further exacerbated by the inefficiencies inherent in conventional cost volume construction methods. Conventional cost volumes are often built upon a predefined, full-range depth basis applied uniformly across the entire image. This uniformity can introduce redundant information, particularly in texture-rich regions where initial disparity estimates are already reliable and extensive exploration is unnecessary. In addition, high accuracy requires resource-intensive methods that increase memory and latency, making the approach unsuitable for real-time applications.

To address these limitations, we propose CAR-Stereo, a novel disparity refinement framework composed of a residual

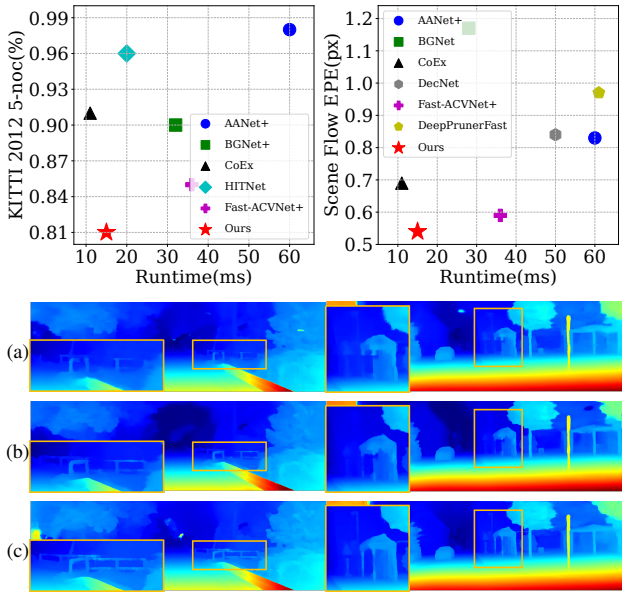


Fig. 1. **Top:** Performance and runtime comparisons with efficiency-oriented methods on the KITTI 2012 [1] and Scene Flow [2] datasets. **Bottom:** Qualitative results on KITTI 2012 and 2015 [3] test dataset. (a) CoEx [4], (b) Fast-ACVNet+ [5], and (c) Ours.

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP-2025-RS-2023-00260098, 50%), in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00358935, 40%), and in part by IITP under the Artificial Intelligence Convergence Innovation Human Resources Development grant funded by the Korea government (MSIT) (IITP-2025-RS-2023-00254177, 10%). *(Corresponding author: Jinsun Park.)*

[1]Chanill Park and Janghyun Kim are with the Department of Information Convergence Engineering (Artificial Intelligence Major), Pusan National University, Busan 46241, Republic of Korea alt990@pusan.ac.kr; jangjoa41@pusan.ac.kr

[2]Minseong Kweon is with the Minnesota Robotics Institute (MnRI), University of Minnesota, Twin Cities, Minneapolis, MN 55455, USA kweon021@umn.edu

[3]Jinsun Park is with the School of Computer Science and Engineering, Pusan National University, Busan 46241, South Korea, and also with the Center for Artificial Intelligence Research, Pusan National University, Busan 46241, Republic of Korea jspark@pusan.ac.kr

cost volume and an adaptive range masking strategy. We first construct a residual cost volume by restricting the disparity search to a residual range around the initial disparity estimate at full-resolution. This approach reduces computational complexity while preserving fine structural details by avoiding exhaustive disparity range searches. Building upon this, we propose a pixel-level adaptive masking strategy guided by pixel-wise confidence to construct a more effective and accurate cost volume. Specifically, for high-confidence regions, we suppress most refinement candidate matches, retaining only a narrow set to reduce the influence of noisy or redundant disparities. In contrast, for low-confidence regions, we retain a broader set of refinement candidate disparities to allow flexible matching refinement across a wider range of hypotheses. As shown in Fig. 1, the proposed method demonstrates the best performance and real-time operation on KITTI 2012 [1] and Scene Flow [2], and also shows superior perception results for precise structures through visual comparison. This adaptive strategy effectively balances precision and robustness, improving real-time stereo matching accuracy. Experimental results show that our algorithm has achieved real-time performance and state-of-the-art results on Scene Flow and KITTI 2012 datasets compared to existing efficiency-oriented methods. In summary, our main contributions can be summarized as follows:

- We introduce the full-resolution disparity refinement algorithm to preserve the fine structure of depth information, while enhancing computational efficiency by constraining the refinement to the residuals around the initial disparity.
- We propose an adaptive range masking strategy to eliminate redundant information from the residual cost volume by adaptively filtering each pixel based on disparity confidence.
- Our method achieves real-time performance and sets a new state-of-the-art among efficiency-oriented models on the Scene Flow and KITTI 2012 datasets.

II. Related Work

Deep Learning based Stereo Matching In deep learning-based stereo matching, the cost volume serves as a core component that encodes the matching cost across a range of disparities by stacking features from the left and right images. Disparity estimation is then performed by aggregating this volume using a neural network [22]. Early studies [17], [23], [24], [25], [20], [26] introduced the use of 3D convolutional neural networks (3D CNNs) to regularize the entire cost volume. These 3D CNNs effectively capture contextual information across both spatial and disparity dimensions, substantially improving matching accuracy.

Recent approaches have investigated two complementary strategies: multi-scale feature aggregation and coarse-to-fine refinement. The multi-scale approaches improve robustness by incorporating contextual cues across multiple resolutions. Their main drawback is the increased computational and memory overhead required for multi-scale feature processing. CasStereoNet [18] employs a cascaded approach that progressively constructs cost volumes with increasing resolution by utilizing multi-scale feature maps extracted from a feature pyramid. On the other hand, coarse-to-fine strategies [19], [12] refine the disparity predictions using an initial estimate as a guide. PCW-Net [19] reduces the search space by warping features using the initial disparity and computing costs within a narrow residual range.

However, this method applies a uniform residual range across all pixels. This uniformity leads to a degradation in accuracy, particularly in areas with high-confidence. To overcome these limitations, our work introduces an adaptive range masking strategy. DeepPruner [27] learns to prune the initial disparity search space before cost volume construction. In contrast, our method operates at the refinement stage, adaptively filtering candidates within a constructed residual cost volume based on pixel-wise confidence scores. This allows for more fine-grained, localized adjustments. More recently, foundation models [28], [29] have emerged, prioritizing zero-shot generalization to unseen domains. The contribution of CAR-Stereo complements these works, aiming to achieve state-of-the-art performance while balancing accuracy and efficiency, a trade-off that remains crucial for real-time applications such as autonomous driving.

Efficiency-oriented Stereo Matching Recent lightweight stereo matching networks [25], [21], [30], [27] have adopted a strategy of constructing the 4D cost volume at a reduced spatial resolution and subsequently upsampling the disparity map. However, these strategy inevitably leads to blurred object boundaries. In order to restore the high-frequency information lost due to the low-resolution cost volume, MobileStereoNet [31] enhances upsampling quality by incorporating a refinement network. AANet [20] efficiently restores high-frequency details lost in low-resolution cost volumes through an adaptive aggregation module that flexibly adjusts sampling locations and weights. CoEx [4] restores high-frequency details by dynamically weighting low-resolution cost volume disparities with high-resolution reference features. HITNet [32] achieves real-time performance through hierarchical iterative refinement on image tiles. Fast-ACVNet [5] leverages attention to refine the cost volume and achieve robust depth estimation. While this strategy alleviates the high computational and memory demands of full-resolution 3D cost volume processing, its reliance on low-resolution representations often results in the loss of fine details and degraded accuracy near object boundaries.

To address these issues, our model has been designed to achieve real-time performance without sacrificing high-resolution details. Our framework first generates an initial disparity map using a lightweight network [4]. To refine this estimated disparity, we construct a cost volume using full-resolution features, but restrict the disparity search to a narrow range centered around the initial estimate. This enables refinement at a higher spatial resolution while maintaining computational efficiency. By operating entirely on the full-resolution cost volume, our approach mitigates the loss of fine structures and the ambiguity of object

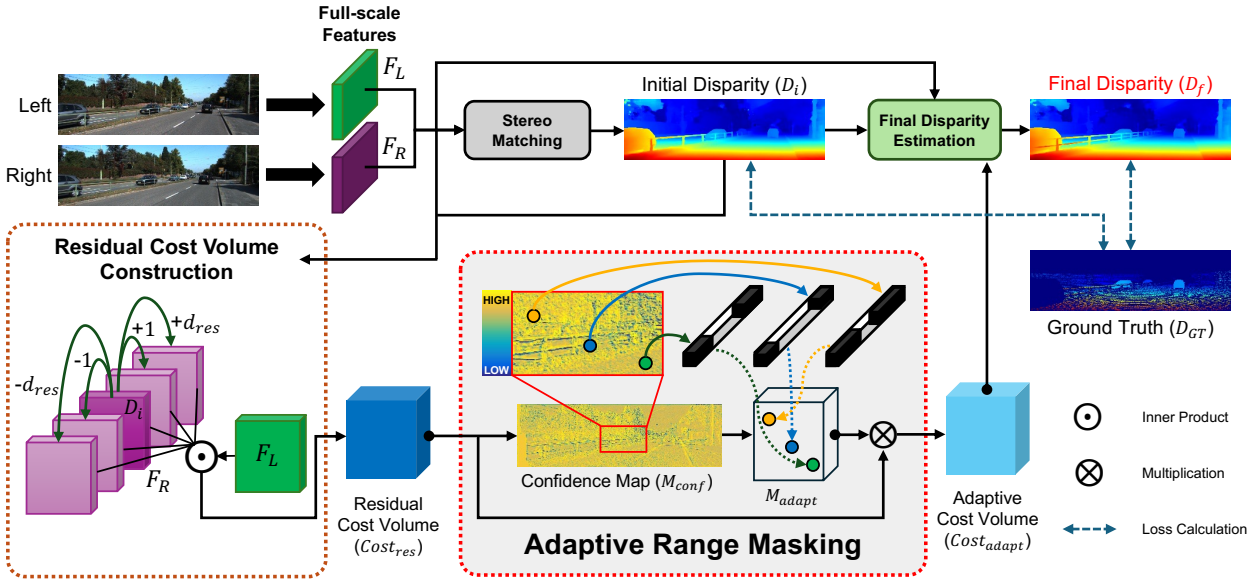


Fig. 2. **Overview of the proposed architecture.** A MobileNetV2-based feature extractor produces an initial disparity map D_i. A residual cost volume $Cost_{res}$ is constructed around D_i using full-resolution features, then filtered by adaptive range masking to create $Cost_{adapt}$. Final disparity estimation refines D_i with $Cost_{adapt}$ to generate D_f. Both D_i and D_f are supervised by ground-truth D_{GT} during training.

boundaries commonly observed in low-resolution processing. Furthermore, by restricting the search to a high-probability region rather than the entire disparity range, it achieves real-time performance without significant loss in accuracy.

III. Method

A. Overall Architecture

Figure 2 shows the overall framework of CAR-Stereo, which employs a coarse-to-fine strategy. The first stage involves the extraction of multi-scale features and initial disparity regression. In this step, we use a MobileNetV2 [33] model pretrained on ImageNet [34] to extract features from each rectified stereo image at five scales $\{1, 1/4, 1/8, 1/16, 1/32\}$ relative to the original image resolution. The feature map at the original resolution is subsequently used to construct a residual cost volume which is used to refine high-resolution disparity, while the remaining resolutions are utilized in the efficiency-oriented model [4] to obtain the initial disparity.

Subsequently, the initial disparity map is utilized to construct the residual cost volume in conjunction with the full-scale left and right features. The residual cost volume is designed to enable efficient computation by allowing the search to be conducted only within a residual range around the initial disparity. It is also constructed at full-scale, thereby preventing the loss of fine-grained regions. The proposed adaptive range masking strategy generates an adaptive cost volume by selectively masking refinement candidates in the residual cost volume, leveraging pixel-wise confidence. Consequently, the final disparity estimation module produces the final disparity map from the adaptive cost volume, left feature, and right feature. The following sections provide a detailed description of each module.

B. Initial Disparity Estimation

First, an initial disparity map is generated from multi-scale features. The initial disparity is predicted by an efficiency-oriented model [4], which constructs the cost volume at a low spatial resolution to achieve high efficiency. However, the limited resolution of this cost volume results in a lack of spatial precision, causing the predicted disparity map to miss fine details and produce inaccurate boundaries [32]. To overcome these limitations, the estimated initial disparity map is leveraged as a geometric prior for a subsequent refinement process performed at full-resolution. By constraining the disparity search to a narrow range centered around the initial estimate, the refinement process maintains computational efficiency while improving accuracy. This enables effective recovery of detailed structural information and sharp object boundaries.

C. Residual Cost Volume Construction

To improve the matching performance for precise structures, we adopt an efficient residual cost volume at the original resolution, inspired by the 3D warping volume [19]. For this purpose, we assume that the initial disparity D_i is in a certain range from the ground-truth D_{GT}, satisfying the following condition:

$$|D_i - D_{GT}| \leq d_{res}, \tag{1}$$

where d_{res} is the residual (*i.e.*, maximum disparity discrepancy) between D_i and D_{GT}. Therefore, we can search for a better disparity estimation D_f that satisfies the following condition:

$$|D_f - D_{GT}| \leq |D_i - D_{GT}| \leq d_{res}. \tag{2}$$

This condition implies that we can narrow our additional disparity search range for the refinement down to

$[D_i - d_{res}, D_i + d_{res}]$. Note that conventional stereo matching algorithms adopt a fixed search range $[0, d_{max}]$ where d_{max} is the maximum disparity. Therefore, if $d_{res} < \frac{d_{max}}{2}$ is satisfied, our disparity refinement process is computationally efficient compared to conventional matching-based ones. Therefore, a cost volume for the refinement can be constructed around D_i with small residuals, resulting in a significant reduction in computational complexity.

In contrast to previous approaches that rely on upsampled features, our method constructs the residual cost volume directly from the full-scale features (*i.e.*, F_L, F_R) extracted from the initial layers of the backbone [33]. This allows our method to retain more fine-grained information. To generate the residual cost volume, the inner product of F_L and F_R is calculated and normalized by the number of channel vector features N_c, and the residual cost volume $Cost_{res}$ is calculated as follows:

$$Cost_{res}(k,u,v) = \frac{1}{N_c}\langle F_L(u,v), F_R(u_k,v)\rangle, \quad (3)$$

$$d_k = -d_{\text{res}} + k, \qquad k \in [0, 2d_{\text{res}}], \quad (4)$$

$$u_k = u - D_i(u,v) - d_k, \quad (5)$$

where u and v denote the pixel coordinates, k denotes an index for the residual displacement search range, d_k denotes the k-th residual offset between $-d_{res}$ and d_{res}, and $\langle\cdot,\cdot\rangle$ denotes the inner product of two feature vectors.

As a result, constructing the residual cost volume at full-resolution introduces minimal computational overhead, as it involves only local disparity offsets around an initial estimate. This design allows our method to achieve accurate disparity estimation by utilizing full-resolution features, while maintaining efficiency through a limited residual search range.

D. Adaptive Range Masking Strategy

Simply constraining the residual range to construct the cost volume fails to account for the pixel-wise disparity ambiguity. When the reliability of a specific pixel (u, v) in D_i is high, employing a wide search range may degrade the refinement accuracy by introducing unnecessary matching refinement candidates that increase the likelihood of incorrect correspondences. To address this problem, we introduce an adaptive range masking strategy that adaptively masks refinement candidates using confidence values that represent the characteristics of each pixel. We utilize the confidence map to guide this process, directly representing the complexity of matching. The confidence map is predicted based on a residual cost volume that encodes the distribution of matching costs across refinement candidates. The confidence estimator [35] is implemented as a lightweight two-stage CNN. Specifically, we feed the residual cost volume, which contains the distribution of matching costs, into this two-stage CNN-based confidence estimator to predict a confidence value for each pixel.

Based on the predicted confidence map and the maximum per-pixel radius, the per-pixel adaptive masking range $r_{mask}(u,v)$ for each pixel is calculated as follows:

$$r_{mask}(u,v) = 1 + (d_{res} - 1)\cdot(1 - M_{conf}(u,v)), \quad (6)$$

where $M_{conf}(u,v)$ denotes the confidence value at (u,v) from the confidence estimator. This formula adaptively adjusts the pixel-wise masking range from 1 to d_{res} in inverse proportion to the confidence of each pixel, thereby implementing an efficient masking strategy based on matching difficulty. In regions with high matching confidence, the masking range is adaptively constrained to its lower bound of 1, thereby preserving the initially estimated disparity values with minimal perturbation. In contrast, when confidence is low, the search range approaches the maximum search range to allow examination of more refinement candidates. Based on the calculated $r_{mask}(u,v)$, the adaptive masking is defined as follows:

$$M_{adapt}(k,u,v) = \begin{cases} 1, & \text{if } |d_k| \le r_{mask}(u,v), \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

This formula ensures that the residual offset d_k is taken into account during the refinement process only when it lies within the adaptive masking range $r_{mask}(u,v)$. If $|d_k| \le r_{mask}(u,v)$, $M_{adapt}(k,u,v)$ is set to 1, allowing the value of the corresponding refinement candidate to be directly reflected. Otherwise, $M_{adapt}(k,u,v)$ is set to 0 to reduce redundant information from unnecessary refinement candidates. Finally, the masked adaptive cost volume $Cost_{adapt}$ is computed as follows:

$$Cost_{adapt}(k,u,v) = Cost_{res}(k,u,v)\cdot M_{adapt}(k,u,v). \quad (8)$$

In regions where matching is challenging, the masking range was maintained to allow for the consideration of diverse refinement candidates. We reduce the masking range in relatively easy-to-match regions to maximally utilize existing accurate disparity information.

E. Final Disparity Estimation

Our final disparity estimation is based on PCW-Net [19]. Initially, the reconstructed error $F_L(u,v) - F_R(u_k,v)$, F_L, D_i, and the proposed $Cost_{adapt}$ are utilized as inputs. Subsequently, five dilated convolutional layers with progressively increasing dilation rates and three residual blocks are applied to expand the receptive field. This structure captures multi-scale contextual information, enabling a more accurate prediction by considering both broad context and fine details.

F. Loss Functions

We employ the Smooth L1 loss (SL_1) for training on both D_i and D_f. Each loss is computed only within the valid region ($D_{GT} > 0$), and the final loss $\mathcal{L}$ is defined as the weighted sum of the two losses as follows:

$$\mathcal{L} = \lambda_0 \cdot SL_1(D_i - D_{GT}) + \lambda_1 \cdot SL_1(D_f - D_{GT}), \quad (9)$$

where λ_0 and λ_1 represent the loss weights for each stage.

IV. Experiments

A. Datasets

To evaluate our CAR-Stereo, we utilized the following datasets: Scene Flow [2], KITTI 2012 [1] and 2015 [3].

TABLE I

QUANTITATIVE EVALUATION OF STATE-OF-THE-ART METHODS ON THE TEST DATASETS OF KITTI 2012 [1] AND KITTI 2015 [3].

* INDICATING EXECUTION TIME MEASURED IN THE SAME ENVIRONMENT.

Target	Method	KITTI 2012 [1]								KITTI 2015 [3]			Runtime (ms)
		3-noc	3-all	4-noc	4-all	5-noc	5-all	EPE noc	EPE all	D1-bg	D1-fg	D1-all	
Accuracy	PSMNet [23]	1.49	1.89	1.12	1.42	0.90	1.15	0.5	0.6	1.86	4.62	2.32	410
	DeepPruner-Best [27]	-	-	-	-	-	-	-	-	1.87	3.56	2.15	182
	ACVNet [5]	1.13	1.47	0.86	1.12	0.71	0.91	0.4	0.5	1.37	3.07	1.65	200
	PCWNet [19]	1.04	1.37	0.78	1.01	0.63	0.81	0.4	0.5	1.37	3.16	1.67	440
	Defom-Stereo [29]	**0.94**	**1.18**	**0.72**	**0.90**	**0.59**	**0.74**	0.3	0.4	**1.25**	**2.23**	**1.41**	300
Speed	DeepPruner-Fast [27]	-	-	-	-	-	-	-	-	2.32	3.91	2.59	61
	AANet+ [20]	1.55	2.04	1.20	1.58	0.98	1.30	0.4	0.5	1.99	5.39	2.55	60*
	DecNet [36]	-	-	-	-	-	-	-	-	2.07	3.87	2.37	50
	BGNet+ [21]	1.62	2.03	1.16	1.48	0.90	1.16	0.5	0.6	1.81	4.09	2.19	32
	CoEx [4]	1.55	1.93	1.15	1.42	0.91	1.13	0.5	0.5	1.79	3.82	2.13	11*
	HITNet [32]	1.41	1.89	1.14	1.53	0.96	1.29	0.4	0.5	1.74	**3.20**	**1.98**	20
	Fast-ACVNet+ [5]	1.45	1.85	1.06	1.36	0.85	1.09	0.5	0.5	**1.70**	3.53	2.01	36*
	CAR-Stereo (Ours)	**1.38**	**1.80**	**1.02**	**1.34**	**0.81**	**1.06**	0.4	0.5	1.86	3.72	2.17	15*

Bold: The best, Underline: The second-best

The Scene Flow dataset is a large-scale collection of synthetic stereo images. We used the end-point error (EPE) metric, restricting both training and evaluation to pixels with disparities below 192.

The KITTI 2012 (194 train, 195 test pairs) and KITTI 2015 (200 pairs for both train and test) datasets consist of real-world driving scenes. Evaluation for KITTI 2012 uses outlier percentage and average EPE for non-occluded (noc) and all regions. For KITTI 2015, the D1 outlier rate is used, measured for foreground (fg), background (bg), and all regions.

B. Implementation details

All experiments were performed using a setup with four NVIDIA RTX A6000 GPUs. The weights λ_0 and λ_1 were empirically set to 0.3 and 1.0, respectively.

For the Scene Flow [2] dataset, training is conducted for a total of 30 epochs with a batch size set to 8. The learning rate is set to 1×10^{-3} for the first 7 epochs and then adjusted to 1×10^{-4} for the remaining epochs. For the KITTI datasets, the pre-trained Scene Flow model is fine-tuned for 800 epochs using the mixed training sets from KITTI 2012 [1] and KITTI 2015 [3]. The initial learning rate is set to 1×10^{-3} and reduced by a factor of 0.5 at the 30, 50, and 300 epochs.

C. Performance of CAR-Stereo

Table I presents the results of comparing our CAR-Stereo with state-of-the-art methods on the KITTI 2012 [1] and KITTI 2015 [3] datasets. We include both performance-oriented and efficiency-oriented models to highlight that our method achieves competitive accuracy while maintaining real-time efficiency. Our CAR-Stereo outperforms state-of-the-art models on KITTI 2012 in error rate metrics of 3 pixel, 4 pixel and 5 pixel including non-occluded and all regions. Moreover, our method achieves 15 ms inference time, faster than previous work except for CoEx [4]. This runtime represents a 2.1–4.0× speedup over comparable models such as AANet+ [20] (60 ms), BGNet+ [21] (32 ms), and DecNet [36] (50 ms). With a minimal computational overhead of 4 ms, our method demonstrates a significant 11% accuracy gain on the 3-noc metric compared to the baseline CoEx model. Furthermore, compared with Fast-ACVNet+ [5] on the 3-noc metric, our CAR-Stereo achieved 4.83% accuracy improvement while reducing execution time from 36 ms to 15 ms, representing more than 2× speedup. This indicates that the proposed method is not merely superior in a single metric but has successfully overcome the trade-off between efficiency and accuracy.

On the KITTI 2015 benchmark, our model demonstrates superior performance in D1-all metric compared to existing methods such as AANet+, DecNet, and BGNet+. Compared with Fast-ACVNet+ and HITNet [32], respectively, the proposed model operates about 2× and 1.3× faster while delivering competitive performance. The proposed method is designed to enhance performance on object boundaries and fine structural details, but these areas represent only a small portion of the image. In addition, the D1 metric calculates the proportion of pixels with large disparity errors (*e.g.*, >3 pixels and >5% compared to the ground-truth) across the entire image, making it less sensitive to improvements in such localized areas.

To more accurately assess the effectiveness of our method in capturing all, non-thin, and thin regions, we additionally evaluate the EPE on the KITTI 2015 [38] and Cityscapes [37] validation dataset. To evaluate performance in thin regions, we define thin regions as *Fence*, *Guard rail*, *Pole*, *Traffic light*, and *Traffic sign* classes using the predefined semantic ground-truth labels of KITTI 2015 and Cityscapes. These classes commonly possess thin and elongated geometric characteristics. For this purpose, we evaluated the zero-shot generalization performance on the KITTI 2015 and Cityscapes validation sets with a model trained on the Scene Flow dataset. As shown in Tab. II, CAR-Stereo (D_f) demonstrates the best generalization performance on both datasets in thin-structure regions (EPE-Thin) and across all regions (EPE-All). On the KITTI 2015 dataset, CAR-Stereo (D_f) achieved EPE-Thin improvements of 29% and 11% compared to Fast-ACVNet+ and CoEx, respectively. Similarly, on the Cityscapes dataset, it achieved a 16% improvement over Fast-ACVNet+ and 6% over CoEx. These results demonstrate that CAR-Stereo

TABLE II

ZERO-SHOT PERFORMANCE ON THE KITTI 2015 [3] AND CITYSCAPES [37] DATASETS FOR THIN-STRUCTURE CLASSES. ALL MODELS ARE TRAINED EXCLUSIVELY ON THE SCENE FLOW [2] DATASET.

Dataset	Method	EPE							
		Fence	Guard rail	Pole	Traffic light	Traffic sign	Thin	Non-Thin	All
KITTI 2015 [3]	CoEx [4]	**1.748**	0.712	2.775	1.417	2.908	2.060	**2.943**	2.922
	Fast-ACVNet+ [5]	4.039	1.528	2.630	3.285	2.998	2.597	3.369	3.352
	CAR-Stereo (D_i)	1.882	0.661	2.735	1.451	2.594	1.962	3.324	3.287
	CAR-Stereo (D_f)	1.807	**0.587**	**2.507**	**1.256**	**2.525**	**1.822**	2.944	**2.915**
Cityscapes [37]	CoEx [4]	1.716	0.010	3.422	1.354	**2.091**	1.836	4.326	4.251
	Fast-ACVNet+ [5]	4.082	0.174	3.414	**1.274**	2.526	2.040	3.943	3.886
	CAR-Stereo (D_i)	1.688	0.011	3.007	1.537	2.258	1.813	3.569	3.516
	CAR-Stereo (D_f)	**1.585**	**0.009**	**2.774**	1.406	2.168	**1.713**	**3.270**	**3.223**

Bold: The best, Underline: The second-best

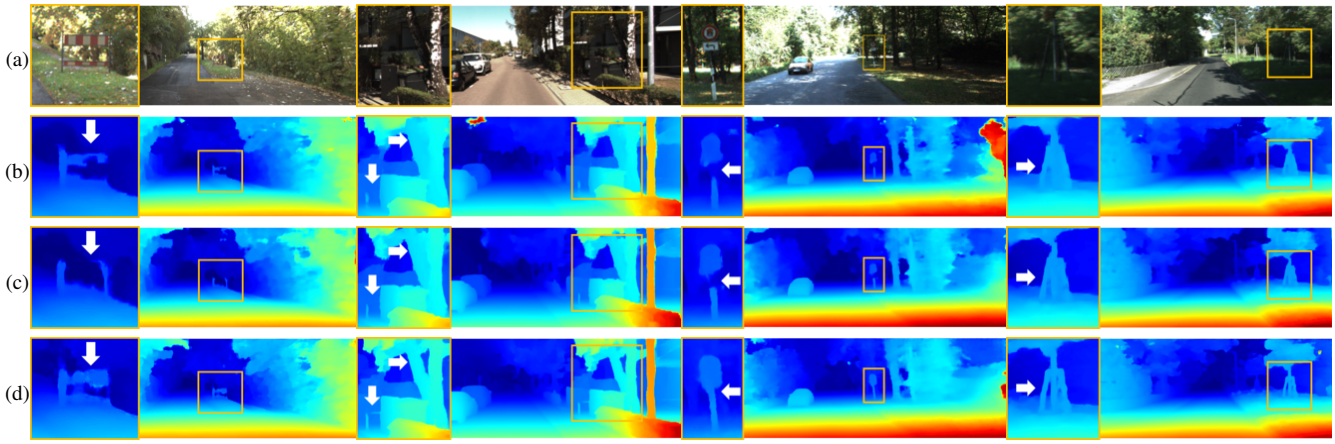


Fig. 3. **Qualitative results on KITTI 2012 [1] and 2015 [3] test dataset.** (a) Left images, (b) CoEx [4], (c) Fast-ACVNet+ [5], and (d) Ours. The highlighted areas in the yellow boxes show that our CAR-Stereo provides more detailed and accurate disparity maps for fine-grained structures (*e.g.*, vehicle barriers, thin poles, and complex tree boundaries). Compared to other methods, our CAR-Stereo reduces blurring and preserves sharp object edges.

TABLE III

EVALUATION OF EFFICIENCY-ORIENTED METHODS ON THE SCENE FLOW [2] DATASET

Method	EPE (px)
DeepPrunerFast [27]	0.97
BGNet [21]	1.17
DecNet [36]	0.84
AANet+ [20]	0.83
CoEx [4]	0.69
Fast-ACVNet+ [5]	0.59
CAR-Stereo (Ours)	**0.54**

Bold: The best, Underline: The second-best

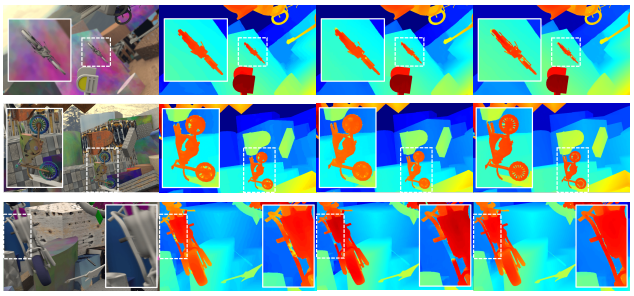


Fig. 4. **Qualitative results on Scene Flow [2] test dataset.** From left to right: Left images, CoEx [4], Fast-ACVNet+ [5], and Ours. The highlighted areas represent regions with fine structures.

consistently provides superior generalization performance for thin-structures across various domains. On both datasets, the refinement from CAR-Stereo (D_i) to CAR-Stereo (D_f) consistently improved performance in the EPE-Thin, EPE-Non-Thin, and EPE-All, indicating the effectiveness of the proposed confidence-aware adaptive disparity refinement.

Our advantage in thin regions is also evident in Fig. 3. CAR-Stereo yields more accurate results than conventional methods, especially for fine-grained structures. For a more detailed results, please see our supplementary material.

On the Scene Flow [2] dataset, the proposed method achieves state-of-the-art results, as shown in Tab. III. In particular, our method achieved a notable 21.7% improvement in EPE, reducing it from 0.69 px to 0.54 px compared to the baseline CoEx model. Compared with the state-of-the-art Fast-ACVNet+ model, the proposed model reduced EPE from 0.59 px to 0.54 px, achieving an 8.5% improvement. As shown in Fig. 4, CAR-Stereo more accurately captures the motorcycle's fine-grained structures (*e.g.*, handlebars, wheel spokes, supporting struts, etc.) than CoEx and Fast-ACVNet+.

In conclusion, the proposed model demonstrated its effectiveness across all three datasets, clearly showcasing the strengths of the methodology through distinct performance improvements, particularly in thin regions.

TABLE IV

ABLATION STUDY ON SEARCH REFINEMENT CANDIDATES AND ADAPTIVE MASKING IN SCENE FLOW [2].

d_{res}	Adaptive Masking	EPE (px)	Runtime (ms)
2		0.59	13.69
	✓	0.57	14.53
4		0.56	14.24
	✓	<u>0.55</u>	14.60
6		<u>0.55</u>	14.73
	✓	**0.54**	15.07
12		0.56	15.33
	✓	<u>0.55</u>	16.01
24		0.58	16.12
	✓	0.56	17.43

Bold: The best, <u>Underline</u>: The second-best

TABLE V

ABLATION STUDY ON FULL-SCALE FEATURES AND ADAPTIVE MASKING IN SCENE FLOW [2].

Full-scale Features	Adaptive Masking	EPE (px)	Runtime (ms)
		0.60	14.78
	✓	0.59	15.19
✓		<u>0.55</u>	14.73
✓	✓	**0.54**	15.07

Bold: The best, <u>Underline</u>: The second-best

D. Ablation Studies

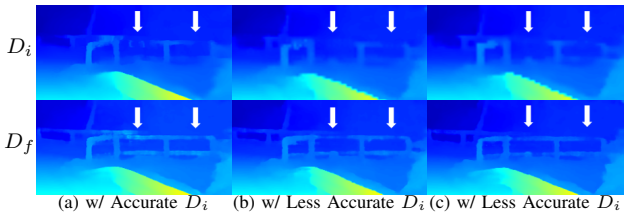
Maximum Search Range Table IV presents the results of analyzing the impact of d_{res}, the maximum residual search range for the residual cost volume construction, and adaptive masking on performance. Here, adaptive masking represents the process of adaptively filtering the residual cost volume based on the confidence of each pixel. When the d_{res} value was too low (4 or less) or too high (12 or more), errors increased due to insufficient matching information and redundant information, respectively. At d_{res} = 6, which optimally balances the speed-accuracy trade-off, our CAR-Stereo outperforms Fast-ACVNet+ [5] by 6.8% in EPE, even without applying adaptive masking. This indicates that our proposed refinement approach with the residual cost volume is sufficiently effective to outperform state-of-the-art models. In addition, adaptive masking achieved further performance improvement from 0.55 to 0.54 by removing redundant information from the residual cost volume with only a small computational overhead of 0.34 ms. Furthermore, adaptive masking consistently reduced EPE by 0.01 to 0.02 px across all d_{res} settings, demonstrating its effectiveness.

Full-Scale Features Table V presents the results of performance variations when full-scale features are employed and depending on whether adaptive masking is applied. Here, full-scale features denote the direct use of features extracted at the original resolution, rather than the 4×-upsampled features obtained from the quarter-resolution representation adopted in the baseline PCW-Net [19]. Full-resolution features lowered average EPE by 0.05 px, and adaptive masking cut it by an additional 0.01 px, producing a final EPE of 0.54 px. Omitting full-scale features and applying adaptive masking introduce minor latencies of 0.05–0.12 ms and 0.4–0.6 ms, respectively, due to the requisite upsampling and masking operations.

TABLE VI

ABLATION STUDY ON ADAPTIVE PRUNING: RUNTIME (ms) AND EPE IN SCENE FLOW [2].

Method	Feature Extraction	Cost Aggregation	Adaptive Pruning	Refine	Total	EPE
Confidence Range Prediction [27]	7.08	4.39	5.98	1.97	19.42	0.61
Adaptive Masking (Ours)	7.08	4.39	**1.63**	1.97	**15.07**	**0.54**



(a) w/ Accurate D_i (d_{res}=6) (b) w/ Less Accurate D_i (d_{res}=6) (c) w/ Less Accurate D_i (d_{res}=15)

Fig. 5. **Qualitative comparisons with large initial disparity errors on KITTI 2015 [3] test dataset.**

Adaptive Pruning Table VI presents ablation study results on the adaptive pruning method. The proposed Adaptive Masking demonstrates superiority over the conventional DeepPruner's Confidence Range Prediction [27] in terms of both time and accuracy. DeepPruner utilizes a network to process sparse disparity estimations, the left image, and the warped right image before constructing the cost volume. In contrast, Adaptive Masking directly estimates confidence from the existing residual cost volume via a network and subsequently applies the mask. Our masking strategy enables the accurate elimination of candidates by leveraging the rich information available after cost aggregation. This structural difference also affects the accuracy.

V. DISCUSSION AND LIMITATIONS

The refinement network restricts the search to a residual range centered around the initial disparity. Thus, ground-truth disparities outside the range cannot be considered if the initial disparity error is large. As shown in Fig. 5(b), replacing the superpixel refinement module of CoEx [4] with bilinear interpolation to produce D_i leads to degraded performance in thin regions. This performance degradation also affects the results shown in Fig. 5(b) D_f. However, extending the refinement range d_{res} to 15 enables more precise predictions in the thin regions, as illustrated in Fig. 5(c) D_f, demonstrating that initial estimation failures in thin structures can be compensated for by adjusting d_{res}. In future work, one can estimate the matching difficulty of each pixel and adaptively adjust the maximum refinement candidate range. This approach will retrieve correct values even for pixels with large initial disparity errors.

VI. CONCLUSION

In this paper, we propose CAR-Stereo, a novel disparity refinement method for real-time stereo matching. This method

constructs a compact, full-resolution cost volume from residuals around the initial disparity and leverages confidence to adaptively eliminate redundant information on a per-pixel basis, thereby achieving computational efficiency while maintaining precise structure perception. Experimental results on the Scene Flow and KITTI 2012 datasets demonstrate that CAR-Stereo provides real-time performance and achieves state-of-the-art accuracy, outperforming existing efficiency-oriented stereo matching algorithms.

REFERENCES

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3354–3361.

[2] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3061–3070.

[4] A. Bangunharcana, J. W. Cho, S. Lee, I. S. Kweon, K.-S. Kim, and S. Kim, "Correlate-and-excite: Real-time stereo matching via guided cost volume excitation," in *IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 3542–3548.

[5] G. Xu, Y. Wang, J. Cheng, J. Tang, and X. Yang, "Accurate and efficient stereo matching via attention concatenation volume," *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, vol. 46, no. 4, pp. 2461–2474, 2023.

[6] Z. Chen, W. Long, H. Yao, Y. Zhang, B. Wang, Y. Qin, and J. Wu, "Mocha-stereo: Motif channel attention network for stereo matching," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 27 768–27 777.

[7] J. Kim, J. Noh, M. Jeong, W. Lee, Y. Park, and J. Park, "Adnet: Non-local affinity distillation network for lightweight depth completion with guidance from missing lidar points," *IEEE Robotics and Automation Letters (RAL)*, vol. 9, no. 9, pp. 7533–7540, 2024.

[8] Y. Liang, Y. Hu, W. Shao, and Y. Fu, "Distilling monocular foundation model for fine-grained depth completion," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 22 254–22 265.

[9] Z. Yan, Y. Lin, K. Wang, Y. Zheng, Y. Wang, Z. Zhang, J. Li, and J. Yang, "Tri-perspective view decomposition for geometry-aware depth completion," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 4874–4884.

[10] J. Kim, U. Shin, S. Heo, and J. Park, "Exploiting cross-modal cost volume for multi-sensor depth estimation," in *Proc. of Asian Conf. on Computer Vision (ACCV)*, 2024, pp. 1420–1436.

[11] J. Park, Y. Jeong, K. Joo, D. Cho, and I. S. Kweon, "Adaptive cost volume fusion network for multi-modal depth estimation in changing environments," *IEEE Robotics and Automation Letters (RAL)*, vol. 7, no. 2, pp. 5095–5102, 2022.

[12] J. Zeng, C. Yao, Y. Wu, and Y. Jia, "Temporally consistent stereo matching," in *Proc. of European Conf. on Computer Vision (ECCV)*. Springer, 2024, pp. 341–359.

[13] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 5249–5260.

[14] L. Bartolomei, F. Tosi, M. Poggi, and S. Mattoccia, "Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 1013–1027.

[15] C. Zhou, J. Yang, C. Zhao, and G. Hua, "Fast, accurate thin-structure obstacle detection for autonomous mobile robots," in *Proc. of Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1–10.

[16] S. Ha, Y. Kim, and J. Park, "Interdimensional knowledge transfer for semantic segmentation on lidar point clouds," *IEEE Robotics and Automation Letters (RAL)*, vol. 9, no. 9, pp. 7501–7508, 2024.

[17] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2017, pp. 66–75.

[18] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2495–2504.

[19] Z. Shen, Y. Dai, X. Song, Z. Rao, D. Zhou, and L. Zhang, "Pcw-net: Pyramid combination and warping cost volume for stereo matching," in *Proc. of European Conf. on Computer Vision (ECCV)*. Springer, 2022, pp. 280–297.

[20] H. Xu and J. Zhang, "Aanet: Adaptive aggregation network for efficient stereo matching," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1959–1968.

[21] B. Xu, Y. Xu, X. Yang, W. Jia, and Y. Guo, "Bilateral grid learning for stereo matching networks," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 497–12 506.

[22] F. Tosi, L. Bartolomei, and M. Poggi, "A survey on deep stereo matching in the twenties," *Int'l Journal of Computer Vision (IJCV)*, pp. 1–32, 2025.

[23] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5410–5418.

[24] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drummond, and Z. Ge, "Hierarchical neural architecture search for deep stereo matching," *Proc. of Advances in Neural Information Processing Systems*, vol. 33, pp. 22 158–22 169, 2020.

[25] S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, and S. Izadi, "Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction," in *Proc. of European Conf. on Computer Vision (ECCV)*, 2018, pp. 573–590.

[26] Q. Wang, S. Shi, S. Zheng, K. Zhao, and X. Chu, "Fadnet: A fast and accurate network for disparity estimation," in *IEEE Int'l Conf. on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 101–107.

[27] S. Duggal, S. Wang, W.-C. Ma, R. Hu, and R. Urtasun, "Deeppruner: Learning efficient stereo matching via differentiable patchmatch," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4384–4393.

[28] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 5249–5260.

[29] H. Jiang, Z. Lou, L. Ding, R. Xu, M. Tan, W. Jiang, and R. Huang, "Defom-stereo: Depth foundation model based stereo matching," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 21 857–21 867.

[30] C. Yao, Y. Jia, H. Di, P. Li, and Y. Wu, "A decomposition model for stereo matching," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 6091–6100.

[31] F. Shamsafar, S. Woerz, R. Rahim, and A. Zell, "Mobilestereonet: Towards lightweight deep networks for stereo matching," in *Proc. of Winter Conf. on Applications of Computer Vision (WACV)*, 2022, pp. 2417–2426.

[32] V. Tankovich, C. Hane, Y. Zhang, A. Kowdle, S. Fanello, and S. Bouaziz, "Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 362–14 372.

[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*. Ieee, 2009, pp. 248–255.

[35] J. Park, K. Joo, Z. Hu, C.-K. Liu, and I. So Kweon, "Non-local spatial propagation network for depth completion," in *Proc. of European Conf. on Computer Vision (ECCV)*. Springer, 2020, pp. 120–136.

[36] C. Yao, Y. Jia, H. Di, P. Li, and Y. Wu, "A decomposition model for stereo matching," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6091–6100.

[37] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.

[38] H. Abu Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented reality meets computer vision: Efficient data generation for urban driving scenes," *Int'l Journal of Computer Vision (IJCV)*, vol. 126, pp. 961–972, 2018.