Semantic Guided Long Range Stereo Depth Estimation for Safer Autonomous Vehicle Applications

WeiQin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, David Suter, and Alireza Bab-Hadiashar, *Senior Member, IEEE*

***Abstract*—Autonomous vehicles in intelligent transportation systems must be able to perform reliable and safe navigation. This necessitates accurate object detection, which is commonly achieved by high-precision depth perception. Existing stereo vision-based depth estimation systems generally involve computation of pixel correspondences and estimation of disparities between rectified image pairs. The estimated disparity values will be converted into depth values in downstream applications. As most applications often work in the depth domain, the accuracy of depth estimation is often more compelling than disparity estimation. However, at large distances (> 50m), the accuracy of disparity estimation does not directly translate to the accuracy of depth estimation. In the context of learning-based stereo systems, this is mainly due to biases imposed by the choices of the disparity-based loss function and the training data. Consequently, the learning algorithms often produce unreliable depth estimates of under-represented foreground objects, particularly at large distances. To resolve this issue, we first analyze the effect of those biases and then propose a pair of depth-based loss functions for foreground objects and background separately. These loss functions can be tuned and can balance the inherent bias of the stereo learning algorithms. The efficacy of our solution is demonstrated by an extensive set of experiments, which are benchmarked against state of the art. We show on the KITTI 2015 benchmark that our proposed solution yields substantial improvements in disparity and depth estimation, particularly for objects located at distances beyond 50 meters, outperforming the previous state of the art by 10%.**

***Index Terms*—stereo matching, depth estimation, disparity estimation, loss function**

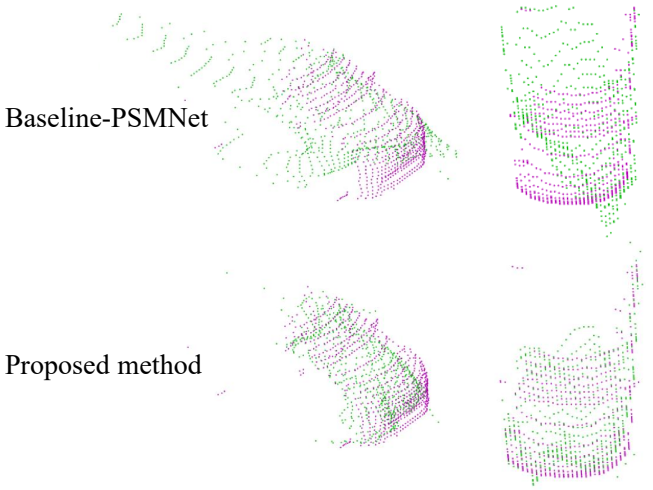


Fig. 1: An example of back-projected 3D points and estimated disparity of an object at 43 meters from the camera (image 000038_10 of KITTI 2015 dataset). Ground truth points are plotted in magenta and predicted points are shown in green. Top and bottom rows show the result of the baseline method [1] and our proposed method. The picture shows that the proposed method significantly improves the depth estimation of a far object. (Best view in colors and zoom in for details.)

I. Introduction

ACCURATE depth perception is critical in intelligent transportation applications such as autonomous driving. The safety of autonomous vehicles strongly correlates with the precision of depth measurements. Precise depth measurements enable reliable cruise control [2], accurate object detection and avoidance [3]–[5], accurate object recognition and classification for efficient lane change and other manoeuvring, and many other applications. Furthermore, precise 3D positions of selected landmarks will result in accurate localization of autonomous vehicles [6]. Therefore, reliable and precise depth information is required to avoid catastrophic road accidents. Laser-based systems such as LiDAR are commonly employed for measuring depth. LiDAR is well known for its accuracy and precision where the captured depth only has errors of an order of centimetres and is currently used in many autonomous vehicles. However, it also has some serious practical limitations, such as a high price tag, reliability issues in different environments and limited resolution [7]. Furthermore, when dealing with long-distance measurements, LiDAR suffers from possible misalignment with other camera sensors due to different coordinate systems and synchronization issues among multiple sensors with varying acquisition periods [8].

A feasible alternative approach to achieve reliable long-range depth estimation is via camera-based depth perception using stereo matching algorithms. However, conventional *discrete* stereo matching algorithms such as semi-global matching (SGM) [9] are prone to pixel-locking (e.g. biased distribution of sub-pixel disparity values towards the integer val-

W. Chuah, R. Hoseinnezhad and A. Bab-Hadiashar are with the School of Engineering, RMIT University, Melbourne, Australia, (e-mail: wei.qin.chuah@student.rmit.edu.au, {rezah,abh}@rmit.edu.au).

R. Tennakoon is with School of Science, RMIT University, Melbourne, Australia, email: ruwan.tennakoon@rmit.edu.au

D. Suter is with School of Science, Edith Cowan University, Joondalup, WA 6027, Australia, email: d.suter@ecu.edu.au

Manuscript received April xx, xxxx.

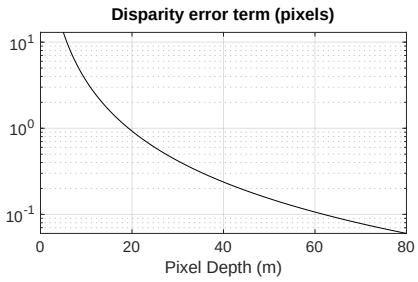


Fig. 2: Variations of the disparity loss term (a pixel disparity error term) with the depth of the pixel, corresponding to 1 meter depth error.

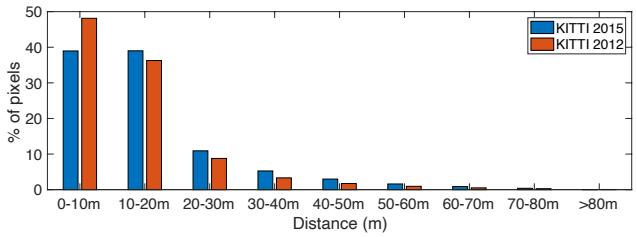


Fig. 3: Distribution of pixels associated with different distances in KITTI 2012 [15] and KITTI 2015 [16] stereo dataset.

ues) [10]. Therefore, these approaches fail to compute accurate disparity at large distances. To alleviate this problem, various methods such as two-stage shifted matching [11], symmetric refinement [12] and disparity smoothing filters [13] were proposed. On the contrary, stereo matching algorithms that are designed to regress continuous disparity values [14] do not suffer from pixel-locking and have been shown to outperform discrete algorithms, especially for long-range depth estimation.

In recent years, there has been significant interest in developing end-to-end learning-based stereo matching models to regress disparity from a rectified pair of stereo images. As these learning-based stereo matching models are designed to estimate continuous disparity values, they are resistant to pixel-locking bias. However, the combination of existing disparity-based loss functions and the commonly used training data (e.g. KITTI [15], [16]) biases the models toward emphasizing more on objects and background areas located at near distances, at the expense of farther objects [17], [18]. This effect can exceptionally be detrimental in safety-relevant driver assistance applications, where distant objects are of interest (e.g. high speed driving on highways [19]). In this work, we aim to improve the long-range depth estimation performance of learning-based stereo matching models, by alleviating the biases caused by the training data and loss function.

Before discussing the identified problems in learning-based stereo disparity estimation algorithms, it may be useful to clarify the key terms such as: (1) *near, middle distance, and far* and (2) *foreground / background*. The three terms (*near, middle distance, and far*) have obvious meanings, at least in terms of relative order. In most cases, these terms are associated with some thresholds that are somewhat context, or application, dependent. Thus, it is naive to expect that any fixed setting of these thresholds will be universally useful. In this work, we selected reasonable thresholds that are appropriate for the common datasets that are designed for different autonomous navigation scenarios. While the terms (*background / foreground*) may suggest of far and near distinction, in this paper, we refer foreground as "certain objects of interest" (depending on the application) and the rest as background. For instance, in our analysis, we have purposely chosen objects that are semantically important in autonomous navigation tasks as foreground (e.g. pedestrians and land vehicles) and others such as buildings, vegetation and sky areas as background.

To tackle the aforementioned problems, we first analyze the emphasis and bias of different depth estimation methods. In most of the learning-based stereo disparity estimation algorithms, the implemented loss functions are usually disparity-based [1], [20]–[25]. However, the disparity-based loss function does not uniformly penalize errors at different distances [17]. As illustrated in Figure 2, one metre error at any distance between $0 - 80$ meters translates to disparity-based losses that would penalize nearby objects ($\leq$ 20 m) much higher than far objects. Consequently, the disparity-based loss function biases the training process towards nearby objects and deteriorates the long-range depth estimation performance of the resulting models.

Furthermore, due to the inevitable attribute of the nature that is primarily caused by the perspective-effect (e.g. foreshortening where faraway objects appear to be smaller and vice versa), the front-view of most driving scenes is dominated by (1) foreground that is close to the camera or (2) background (often at a distance). For example, in KITTI 2015 [16], approximately 65% of total pixels in an image are background and have depth value $\leq$ 20 m. More severely, as the foreground often appears to be smaller in size when located at a distance, only 2% of the total pixels are foreground and have depth value $>$ 20 m. The details of pixel data distribution are included in Table I and Figure 3.

Therefore, final solutions of the learning algorithms, utilizing the disparity-based loss functions and the mentioned training data, will result in estimates biased towards the background and are somewhat "blinded" to the foreground (especially those positioned at farther distances). For example, as illustrated in Fig. 1, the depth estimates resulted by PSMNet [1] without addressing the bias issues can be highly unreliable for distant objects (a car located 43 meters from the camera). To address these bias issues, we propose to adjust the bias of the conventional stereo learning algorithms to emphasize on far and foreground objects without losing sight of the close-by objects and background areas.

In short, we propose a simple yet effective solution for regularizing the bias of learning-based stereo disparity estimation algorithms by adjusting the loss function based on the notion of relevant depth ranges and scene contents (e.g. objects positioned at 'mid-range' or 'far' distances in an autonomous navigation scenario). More specifically, we use a depth-based loss function that is divided into foreground and background segments using an off-the-shelf object detector to balance the bias between the two classes. This allows additional and tunable penalization of errors across these classes. The proposed

method can be easily implemented; it is model-agnostic and does not introduce any computational cost overhead during inferencing.

Although shifting the bias or preference of a stereo disparity estimation algorithm from disparity to depth may suggest some deterioration in the performance for the nearby objects and background areas (as depth is reciprocal to disparity), our results demonstrate otherwise. Extensive experiments demonstrate that the proposed loss functions can significantly improve the overall accuracy of disparity and depth estimates at all distances and outperform the baseline stereo disparity estimation algorithm (trained using the disparity loss function). Furthermore, our method also attains state-of-the-art long-range stereo depth estimation performance, outperforming the previous method that utilizes accurate LiDAR (active) measurements for depth refinements (SDN [17]).

The remainder of this paper is organized as follows. Section II describes the related work in the field of learning-based stereo matching networks and depth estimation, and 3D object detection. Section III presents the proposed disparity and depth-based loss functions. Section III also describes the proposed weighted foreground and background loss functions. Experimental results and discussions are presented in Section IV, and Section V concludes the paper.

II. Related Works

A. Long-Range Depth Estimation

In analyzing the long-range stereo matching algorithms, Pinggera *et al.* [10] classified those into two groups of discrete (classification) or continuous (regression) methods, depending on their output. While the discrete methods (e.g. Census [26], semi global matching (SGM) [9] and MC-CNN [27]) are computationally efficient, these methods suffer from the pixel-locking effect due to the fact that the distribution of sub-pixel disparity values is biased towards zero [11]. As depth error at large distances is highly dependent on the accuracy of sub-pixel disparity, discrete methods with pixel-locking effect are unlikely to produce accurate long-range depth estimations [10]. To mitigate the pixel-locking problem, different techniques such as designing new matching costs [28] and interpolation functions [8], [29] have been proposed. Moreover, Nehab *et al.* [12] proposed symmetrical refinement technique that exploits the inherent symmetry of matching cost functions and simultaneously refines the matching coordinates in both stereo images.

In contrast, methods that estimate disparity in the continuous setting are free from the pixel-locking effect and have been shown to outperform the discrete methods [10]. Examples of continuous stereo matching algorithm include local differential matching (LDM) [19], [30], [31], total variation stereo (TV) [14] and learning-based approaches (GC-Net [32], PSMNet [1], GANet [23]). Although the learning-based methods do not suffer from the pixel-locking bias, most of these methods are prone to biases caused by the training data and their loss functions (explained in Section I). In this work, we propose a novel combination of loss functions to standardize the learning attention across all distances and improve the performance of long-range depth estimation in learning-based stereo matching models.

B. Learning-Based Stereo Matching Networks

Recent works [1], [21], [23], [27], [33], [34] have shown that stereo matching, using deep features, illustrate a significant performance boost over traditional hand-crafted features like SIFT [35] and ORB [36] features. Existing end-to-end stereo matching networks utilized CNNs to (1) extract deep representation from input stereo images, (2) perform cost volume aggregations and (3) perform cost volume refinement.

In terms of taxonomy, end-to-end stereo matching networks can be classified into two categories: (1) correlation-based and (2) shifted concatenation-based cost volume construction methods. The correlation-based networks consist of stacked 2D CNNs layers and have significantly lower processing time due to the high efficiency of 2D convolution [20], [25], [34], [37]. The concatenation based networks consist of a combination of 2D CNNs for feature extraction and 3D CNNs for cost volume aggregation and refinement [1], [23], [32]. An interesting exception is the idea of group-wise correlation-based cost volume construction that was proposed to preserve information loss of full correlation [21]. In terms of performance, shifted concatenation-based networks with 3D CNNs layers often outperform correlation-based networks (with 2D CNNs layers) by a large margin, on popular benchmarks (e.g., SceneFlow [20], KITTI [16]).

To close the performance gaps between the correlation and the shifted concatenation methods, several prior works include context information such as edges [38] and semantics [39] into the network. Moreover, Xu *et al.* [25] proposed AANet: which consists of a new sparse points-based representation for intra-scale aggregation and adaptive multi-scale cross aggregation modules using 2D CNNs layers. As a result, AANet has comparable performance to the shifted concatenated methods, but with real-time inference speed.

Apart from the aforementioned supervised approaches, unsupervised stereo matching networks [40]–[44] have also received substantial amount of interest in recent years. These methods typically rely on reconstruction-based loss function to avoid ground truth supervision. As optimizing the network using the reconstruction-based loss function only is an ill-posed problem, other loss functions such as left-right consistency [42], and occlusion-aware [40], [41] loss functions were proposed to better constraint the overall learning process. In an opposite direction, auxiliary tasks such as optical flow [45], [46], motion segmentation [47] and semantic segmentation [48] were also incorporated to further improve the performance of stereo matching networks.

Despite all the advances in learning-based stereo matching network, the relationship between disparity and depth is rarely discussed. As mentioned in Section I, while current state-of-the-art stereo matching network is capable of performing disparity estimation with high accuracy, the accuracy in disparity estimation does not translate directly to the accuracy in depth estimation, especially for objects that are far away from the camera. In contrast, we turn our attention to this problem

TABLE I: Pixel data distribution of KITTI 2015 dataset.

	$\leq$ 20 m	>20 m	Total
Foreground	14.90%	1.91%	16.81%
Background	64.79%	18.40%	83.19%
Total	79.69%	20.31%	100.00%

and propose a simple yet effective solution, which allows our method to improve the performance of both disparity and depth estimation, particularly for far away foreground objects.

C. Semantic-Guided Stereo Matching Networks

Semantic segmentation involves pixel-wise classification, which provides valuable semantic information for scene understanding. It is commonly incorporated into different low-level tasks such as optical flow [49], depth estimation [50]–[52] and depth completion [53]. Similarly, semantic segmentation is also incorporated in stereo matching task, to leverage its high-level semantic cues. For example, Zhang *et al.* [48] proposed DispSegNet, an unsupervised stereo matching network that concatenates the semantic feature with initial disparity map for refinement. Moreover, Dovesi *et al.* [54] proposed a real-time stereo system named RTS^{2}Net that concatenates semantic class probabilities and disparity volume to calculate disparity residual.

On the other hand, Yang *et al.* [39] proposed SegStereo that incorporates semantic cues in stereo matching networks, by concatenating deep semantic feature embedding with the stereo cost volume. SegStereo consists of a disparity estimation sub-network and a semantic segmentation sub-network that are trained jointly, by using a combination of loss functions consist of a weighted sum of the reconstruction error, a smoothness term, and a segmentation error. In addition, Wu *et al.* [55] introduced SSPCV-Net that fuses multi-scale 4D cost volumes with semantic features obtained using a semantic segmentation sub-network, similar to SegStereo. The resulting spatial pyramid cost volumes are aggregated, refined and up-scaled, using a hourglass (e.g. encoder-decoder) module followed by a 3D feature feature fusion module. The network is optimized using a supervised disparity loss function and a boundary-based smoothness term.

In contrast to the mentioned approaches, our method does not include an additional semantic segmentation sub-network or concatenation of semantic features. Instead, we aim to learn the semantic relationship between foreground objects and background areas to balance the learning bias between these two classes in stereo matching networks.

D. Depth estimation and 3D object detection

Accurate depth information of moving (foreground) objects such as pedestrians, transportation vehicles and cyclist is important in downstream applications such as 3D object detection and autonomous navigation. There are several works that concentrate on stereo depth estimation for 3D object detection. For instance, Pon *et al.* [56] proposed a stereo matching network that focuses on objects of interest while neglecting the background. Qian *et al.* [18] proposed to combine stereo matching and 3D object detection networks into a single pipeline, by designing a novel differentiable module to convert predicted depth maps to pseudo-LiDAR [3]. They used the same stereo matching network proposed in [17].

Although existing methods can achieve impressive results for 3D object detection from RGB images, the performance deteriorates as the distance increases due to the factors discussed in Section I. In this context, You *et al.* [17] proposed SDN, aiming to improve the long range depth estimation by converting a disparity-based stereo matching network [1] into a depth-based stereo matching one. The proposed network converts the disparity cost volume to a depth cost volume thus regressing a depth value for each pixel (instead of disparity). They further proposed a depth propagation algorithm, which fuses extremely sparse (4-beam) LiDAR to rectify the initial depth estimation.

In contrast, we aim to improve the performance of depth estimation of a stereo matching network by adjusting the bias in the common training loss functions and selected datasets. We propose to carefully balance the training signals, preventing any over-emphasis on backgrounds or close objects, as mentioned in Section I. As a result, the proposed method achieves significant improvement in disparity and depth estimation, particularly for distant objects, over the baseline method. More importantly, our results are on par with the variant of the mentioned prior work SDN + GDC [17], *without using any additional information such as sparse LiDAR data points*.

III. Proposed Method

In this section, we will discuss our proposed loss functions and the overall framework. The results of our experimental and ablation studies are presented in Section IV.

A. Loss Function

As the performance of supervised learning neural network largely depends on its loss function, it is crucial to select the appropriate loss function carefully. Also, an optimally designed loss function can mitigate the adverse effect of bias (such as data imbalance, class imbalance) in the training dataset, and therefore improve the overall performance of the trained model [57]. We show that naively employing the disparity-based loss function and the biased training dataset would cause the trained model to overfit[1] to nearby objects and background areas, diminishing the accuracy of long-range depth estimations. To remedy these bias issues, we propose to redesign the loss function by including foreground and background specific depth-based loss functions.

[1]We define *overfit* as: a network is learning to fit accurately at a certain part of the data (dominant data points) with the expenses of lower accuracy for other parts. This is different from the general definition of *overfit* in machine learning: a network is learning to also fit the noise in the training data, and negatively impacts the network capability to generalize to new data.

1) Disparity-based loss function: The disparity-based loss function is implemented to enable the stereo matching network to learn the regression of disparity for each pixel. The commonly employed smooth $L1$ disparity-based loss function is defined as:

$$\mathcal{L}_{disp} = \frac{1}{N}\sum_{i=1}^{N} \text{smooth}_{L_1}(D_i - \hat{D}_i), \tag{1}$$

in which

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}$$

where $\hat{D}_i$ and D_i are the predicted and ground truth disparity values for pixel i, and N is the number of valid pixels.

2) Depth-based loss function: As it was discussed in Section I and shown in Fig. 2, the disparity-based loss function given by Eq. (1) is heavily biased and its minimization would only lead to accurate depth estimates for pixels of nearby objects. To address this issue, we include a depth-based loss function, which is similar to the loss function implemented in the Stereo Depth Network (SDN) [17].

However, in contrast to the SDN, the disparity cost volume will not be converted to depth cost volume within the network. We instead propose to convert the predicted disparity map to depth map to ensure our proposed network does not regress depth values. As our aim is to build a passive system, unlike SDN, we do not include laser measurements (to refine our results). The predicted dense disparity map, $\hat{D}$, as well as its corresponding ground truth, D, are converted to depth map, $\hat{Z}$ and Z, via

$$Z = \frac{f \times b}{D} \tag{2}$$

where f and b represent the focal length and baseline of the stereo camera setup. Using those, the depth loss function is defined as follows:

$$\mathcal{L}_{depth} = \frac{1}{N}\sum_{i=1}^{N} \text{smooth}_{L_1}(Z_i - \hat{Z}_i). \tag{3}$$

However, replacing disparity-based with the depth-based loss function comes at the price of having less accuracy for close by objects ($\leq$ 10m) [17]. Therefore, we propose to include both disparity-based and depth-based loss functions. The depth-based loss function regularizes the learning to avoid solution that is over-fitted to the close distance pixels and vice versa. By combining the merits of both loss functions, the overall framework is able to predict accurate disparity / depth for objects at a wide range of distances.

3) Weighted foreground and background loss functions: To balance the bias between foreground and background, due to class imbalance in training datasets, we propose to divide the included depth-based loss function into two terms (foreground and background specific). These terms will be weighted accordingly and the weighting policy will be explained later in this section. We employ Mask R-CNN [58] pre-trained on CityScapes dataset [59] to extract the foreground from the background by performing foreground object segmentation. An example of the segmented foreground object masks is shown in Fig. 4.

Fig. 4: An example of object mask generated using a pre-trained Mask-RCNN [58] on left image sequence 000123_10 of KITTI 2015 [16] stereo dataset.

For simplicity, we only considered transportation vehicles including cars, trucks, vans, buses, bicycles and motorcycles as foreground objects. However, this idea can easily be extended to include other object types. We then combine the object masks and the depth loss function, expressed in Eq. (3), to obtain two new loss functions that are defined as:

$$\mathcal{L}_{depth}^{fg} = \frac{\sum_{i=1}^{N}\left(\text{smooth}_{L_1}(Z_i - \hat{Z}_i)\cdot \mathcal{B}_i\right)}{\sum_{i=1}^{N}\mathcal{B}_i} \tag{4}$$

$$\mathcal{L}_{depth}^{bg} = \frac{\sum_{i=1}^{N}\left(\text{smooth}_{L_1}(Z_i - \hat{Z}_i)\cdot (1-\mathcal{B}_i)\right)}{\sum_{i=1}^{N}(1-\mathcal{B}_i)} \tag{5}$$

where $\mathcal{B}$ is the object masks, $\mathcal{L}_{depth}^{fg}$ is the foreground depth loss and $\mathcal{L}_{depth}^{bg}$ is the background depth loss. Lastly, the overall loss function is proposed as:

$$\mathcal{L} = \mathcal{L}_{disp} + \lambda \cdot \mathcal{L}_{depth}^{fg} + (1-\lambda)\cdot \mathcal{L}_{depth}^{bg} \tag{6}$$

where hyperparameter λ is included to balance the effect on foreground and background learning.

Although the ratio of foreground and background shown in Table I suggests weighting of 0.8 for foreground and 0.2 for background ($\lambda = 0.8$), our experimental results demonstrate otherwise. To investigate this phenomenon, extensive experiments were conducted to study the properties of depth-based loss function and effect of hyperparameter λ on the overall performance of depth estimation, which will be discussed in the next section.

B. Datasets

1) KITTI 2015: This dataset contains images of natural scenes (city and rural areas and highways) collected in Karlsruhe, Germany. It contains 200 training stereo image pairs with sparse ground truth disparities, collected using LiDAR sensor; and 200 testing image pairs without ground truth disparities. KITTI allows performance evaluation by submitting final results to their evaluation server. Following [1], we perform hold-out validation by splitting the 200 training images into 160 for training and 40 for validation. All the results presented in Section IV are computed using the same validation set, unless stated otherwise.

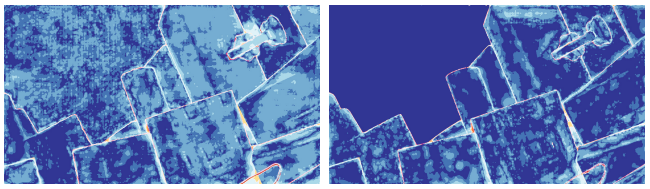
2) DrivingStereo: This dataset is a large scale stereo dataset, covering a diverse set of driving scenarios and different weather conditions; containing over 174,437 stereo pairs for training and 7751 pairs for testing [60]. Sparse ground truth disparities are provided for the training sets only. We use the DrivingStereo dataset to pre-train the stereo matching model, before fine-tuning on smaller KITTI dataset. Similarly,

TABLE II: Ablation study of disparity-based and depth-based loss function using the KITTI 2015 validation set. *The stereo matching networks are pre-trained on Scene Flow and fine-tuned on KITTI 2015 training set.* All pixels are divided into bins according to their true depth values. The performance of depth estimation at different depth intervals is evaluated using the EPE metric. The results illustrate that including depth-based loss function can effectively improve the accuracy of depth estimation, especially at far distances ($\geq$ 20m).

Loss Function		Range (m)							
$\mathcal{L}_{disp}$	$\mathcal{L}_{depth}$	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
		PSMNet [1]							
✓		**0.12**	0.38	0.89	1.50	2.42	3.69	5.39	6.47
	✓	**0.12**	**0.35**	0.87	1.51	2.48	3.68	5.12	5.89
✓	✓	0.13	0.40	**0.86**	**1.46**	**2.33**	**3.53**	**4.61**	**5.66**
		GwcNet [21]							
✓		**0.10**	0.29	0.76	1.48	2.14	3.21	4.28	4.80
	✓	**0.10**	**0.28**	0.78	1.44	2.06	3.22	3.99	4.24
✓	✓	**0.12**	0.31	**0.74**	**1.39**	**1.98**	**2.85**	**3.70**	**3.92**

TABLE III: Ablation study of disparity-based and depth-based loss function. The performance of long-range disparity estimation is evaluated by computing the sub-pixel disparity accuracy, using D1 error metric with 1 pixel and 0.5 pixel thresholds. Only objects and background that are located $50-80$ meters away from the camera are considered in this evaluation.

$\mathcal{L}_{disp}$	$\mathcal{L}_{depth}$	PSMNet [1]		GwcNet [21]	
		<1 px	<0.5 px	<1 px	<0.5 px
✓		17.48	39.84	13.18	34.56
	✓	12.04	28.74	8.52	27.16
✓	✓	9.71	23.91	8.50	26.69



(a) Baseline (b) Ours

Fig. 5: Qualitative results of Scene Flow dataset comparing the performance of (a) disparity-based loss function and (b) the combination of disparity-based and depth-based loss function. By including depth-based loss function, superior results are obtained especially for the pixels located at very far distances.

the dataset is split into training and validation set. Four subsets were randomly selected as the validation set while the remaining are used for training.

3) Scene Flow: This dataset is a large collection of synthetic stereo images with dense disparity ground truth. It comprises three subsets with different settings: FlyingThings3D, Driving and Monkaa. It consists of 35,454 training and 4,370 testing images. The size of each image is 960×540. This dataset is used to train the stereo matching network, to analyze the effect of the disparity-based ($\mathcal{L}_{disp}$) and the depth-based ($\mathcal{L}_{depth}$) loss functions. The results demonstrate that, by combining the two loss functions, can effectively improve the performance of depth estimation, at all distances, and in different scenarios (real-world and synthetic scenes).

C. Metrics

We evaluated the performance of **disparity** estimation using the official D1 metrics, which compute the percentage of outliers (endpoint-error (EPE) of < 3 pixels or $< 5\%$ based on the ground truth) for foreground only, background only and all pixels, respectively (D1-fg, D1-bg, D1-all). In addition, we employed the D1 metric with different pixel thresholds (e.g. 1 pixel and 0.5 pixel) for points located between 50 and 80 meters away from the camera, to analyze the performance of long-range disparity estimation.

We also evaluated the performance of **depth** estimation on the KITTI 2015 dataset. We divided all pixels to the corresponding depth intervals (ranging between $0-80$ meters) based on the ground truth depth value. Then, EPE metric was employed to evaluate the performance of **depth** estimation within each distance interval. This metric provides valuable insights on the performance of depth estimation at different depth ranges.

D. Implementation details

The proposed loss functions are implemented in conjunction with network architecture proposed in PSMNet [1] and GwcNet [21]. The PSMNet is an effective 3D stereo matching network that is commonly used as backbone for disparity estimation [3], [17], [24], [61], [62]. The GwcNet is a recently proposed state-of-the-art stereo matching networks. The networks are implemented using PyTorch framework and is trained end-to-end with Adam ($\beta_1 = 0.9, \beta_2 = 0.999$) optimizer. For pre-training, both networks are trained from scratch using the DrivingStereo [60] dataset. All training setups including data pre-processing, learning rate scheduling,

TABLE IV: Ablation study of the proposed loss function combination using the KITTI 2015 [16] driving dataset. *The stereo matching networks are pre-trained on DrivingStereo datasets and fine-tuned on KITTI 2015 training set.* The performances of disparity estimation under different loss function settings are evaluated using D1 error metric (%), for foreground (Fg), background (Bg) and all pixels respectively. * denotes the weighted $\mathcal{L}_{depth}^{fg}$ and $\mathcal{L}_{depth}^{bg}$ ($\lambda = 0.6$).

Loss Function			PSMNet [1]			GwcNet [21]		
$\mathcal{L}_{disp}$	$\mathcal{L}_{depth}^{fg}$	$\mathcal{L}_{depth}^{bg}$	D1-Fg	D1-Bg	D1-All	D1-Fg	D1-Bg	D1-All
✓			1.89	1.83	1.84	1.66	1.46	1.48
✓	✓		1.59	2.03	1.98	1.62	1.59	1.59
✓		✓	2.06	1.87	1.89	2.13	1.47	1.55
✓	✓	✓	1.53	1.95	1.90	1.51	1.59	1.58
	✓	✓	1.66	1.85	1.83	1.56	1.52	1.52
	✓*	✓*	1.59	1.83	1.80	1.54	1.51	1.51
✓	✓*	✓*	**1.31**	**1.80**	**1.74**	**1.53**	**1.45**	**1.46**

and number of epochs are exactly identical to the original implementation of PSMNet and GwcNet, respectively.

In our implementation, we used the pre-trained models and fine-tuned on KITTI 2015 [16] training set for 300 epochs. The learning rate for fine-tuning process starts at 0.001 and is decreased to 0.0001 after 200 epochs. Following [1], the fine-tuning process of the PSMNet is prolonged to 1000 epochs with learning rate begins at 0.001 and decreased to 0.0001 after $\frac{2}{3}$ of total epochs before submission to KITTI evaluation server. All results presented in Section IV are generated using PSMNet and GwcNet that are pre-trained on DrivingStereo and fine-tuned on KITTI 2015, unless stated otherwise. The batch size for both networks is set to 12 for training on 2 NVIDIA RTX 6000 Quadro GPUs.

In addition, both models were also pre-trained on the Scene Flow dataset and fine-tuned on KITTI 2015, to study the effect of disparity-based ($\mathcal{L}_{disp}$) and depth-based ($\mathcal{L}_{depth}$) loss function (refer to Table II). Similarly, the training procedures were set to be identical to the original implementations.

IV. Experimental Results and Discussion

To validate the effectiveness of each component proposed in this work, several experiments *with different loss function combinations* are conducted using KITTI 2015 [16] validation set and Scene Flow [20] testing sets.

A. Ablation study for disparity and depth loss functions

In this section, we investigate the regularization property of the depth-based loss function, and how it impacts the performance of our trained stereo matching networks at different depth ranges. Depth-based loss function can mitigate the overfitting caused by the disparity-based loss function, by allowing greater training signals for pixels with greater depth value. As shown in Table II and Table III, training using *only* the depth-based loss function achieved better accuracy for objects located at greater distances ($\geq 50m$) compared to training using disparity-based loss function.

Also, by combining the two loss functions, the network achieves even better accuracy for objects located beyond 20m. Although the performance of the close distance pixels ($0 - 20$m) have deteriorated slightly (around 0.02m), this is a relatively small price to pay for significant improvement in the accuracy of long range measurements.

TABLE V: Relationship between balancing term λ and the performance of stereo disparity estimation in PSMNet and GwcNet. The performance is evaluated using the D1 metric for foreground (D1-Fg), background (D1-Bg) and all (D1-All) pixels. The results illustrate that $\lambda = 0.6$ yields the best performance in both PSMNet and GwcNet.

λ	PSMNet [1]			GwcNet [21]		
	D1-Fg	D1-Bg	D1-All	D1-Fg	D1-Bg	D1-All
0.0	2.06	1.87	1.89	2.13	1.47	1.55
0.2	1.68	1.87	1.84	2.04	1.43	1.50
0.4	1.46	1.81	1.77	1.66	1.49	1.51
0.5	1.43	1.83	1.78	1.67	1.51	1.52
0.6	**1.31**	**1.80**	**1.74**	**1.53**	**1.45**	**1.46**
0.8	1.69	2.00	1.96	1.63	1.52	1.53
1.0	1.59	2.03	1.98	1.62	1.59	1.59

Furthermore, it is interesting to note that depth-based loss function also improves the accuracy of measurements for foreground objects irrespective of their depth. Table IV shows that by including the depth-based loss function ($\mathcal{L}_{disp} + \mathcal{L}_{depth}$), the overall accuracy for foreground objects improved by 0.36% in PSMNet and 0.15% in GwcNet. We further demonstrate the effectiveness of depth-based loss function using the Scene Flow dataset. As illustrated in Fig. 5, by including the depth-based loss function, we have consistently achieved incredibly low errors for distant foreground and background. The background of the image included in Fig. 5 has true disparity values ranging between $[1.1, 1.6]$ pixels.

B. Ablation study for foreground and background depth loss functions

We tackle the imbalance between foreground and background by designing a pair of depth-based loss functions, namely foreground specific $\mathcal{L}_{depth}^{fg}$ and background specific

TABLE VI: Mean depth error (m) of KITTI 2015 validation set over various depth range. Our approach significantly improved the accuracy for very far-away pixels without sacrificing the accuracy of closer objects.

Methods	Range (m)							
	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
SDN [17]	0.21	0.35	0.87	1.80	2.67	4.27	5.82	-
SDN+GDC [17]	0.21	0.35	0.84	1.74	2.59	4.14	5.72	-
LR-PSMNet	0.11	0.35	0.84	1.44	2.33	3.29	4.67	5.53
LR-GwcNet	0.10	0.29	0.71	1.35	1.93	2.70	3.69	3.81

TABLE VII: The online benchmark results on KITTI 2015 test sets. All included results are obtained from the official KITTI 2015 benchmark.

Methods	D1-All (%)			D1-Noc(%)		
	Bg	Fg	All	Bg	Fg	All
AANet [25]	1.99	5.39	2.55	1.80	4.93	2.32
PSMNet [1]	1.86	4.62	2.32	1.71	4.31	2.14
SegStereo [39]	1.88	4.07	2.25	1.76	3.70	2.08
DeepPruner [63]	1.87	3.56	2.15	1.71	3.18	1.95
GwcNet [21]	1.74	3.93	2.11	1.61	3.49	1.92
GANet-15 [23]	1.55	3.82	1.93	1.40	3.37	1.73
CFNet [64]	1.54	3.56	1.88	1.43	3.25	1.73
Semantic-Guided Stereo Matching						
DispSegNet [48]	4.20	16.97	6.33	3.86	15.89	5.85
RTS^{2}Net [54]	3.09	5.91	3.56	-	-	-
SSPCVNet [55]	1.75	3.89	2.11	1.61	3.40	1.91
LR-GwcNet	1.67	4.19	2.09	1.53	3.77	1.90
LR-PSMNet	1.65	4.13	2.06	1.52	3.98	1.92
LR-CFNet	**1.64**	**3.10**	**1.89**	**1.49**	**2.72**	**1.70**

$\mathcal{L}_{depth}^{bg}$, that are appropriately weighted. The ratio between the foreground and background data, listed in Table I, suggests a weighting of 0.8 for foreground and 0.2 for background ($\lambda = 0.8$). However, from our experiments, we have found that depth-based loss function has better performance for foreground than background pixels. Therefore, by giving less weights to the foreground pixels and more to background ones, we may achieve a better balance. This is explained in the next subsection.

To support our argument, we have conducted two additional experiments using the disparity-based loss function with either foreground specific ($\lambda = 1$) or background specific ($\lambda = 0$) components. The results are tabulated in Table IV. Within expectation, the results demonstrated that $\mathcal{L}_{depth}^{fg}$ is advantageous for foreground prediction. However, solely including $\mathcal{L}_{depth}^{bg}$ worsens the accuracy for background as well as the overall accuracy. Regardless, $\mathcal{L}_{depth}^{bg}$ is still required to improve the accuracy of background located at far distances. As such, both depth-based loss components are needed to improve the overall accuracy at all distances.

C. Analysis of balancing term λ

Hyperparameter λ balances the contributions of foreground specific $\mathcal{L}_{depth}^{fg}$ and background specific $\mathcal{L}_{depth}^{bg}$ components to the total loss. We study the effect of λ using grid-search between 0 and 1 with interval of 0.2. As it was mentioned earlier, the ratio between foreground and background data in the KITTI 2015 dataset implies the λ to be 0.8 for optimal performance. However, we observed that the optimal results for the overall, foreground and background errors are obtained by setting $\lambda = 0.6$ in both PSMNet [1] and GwcNet [21] (refer to Table V).

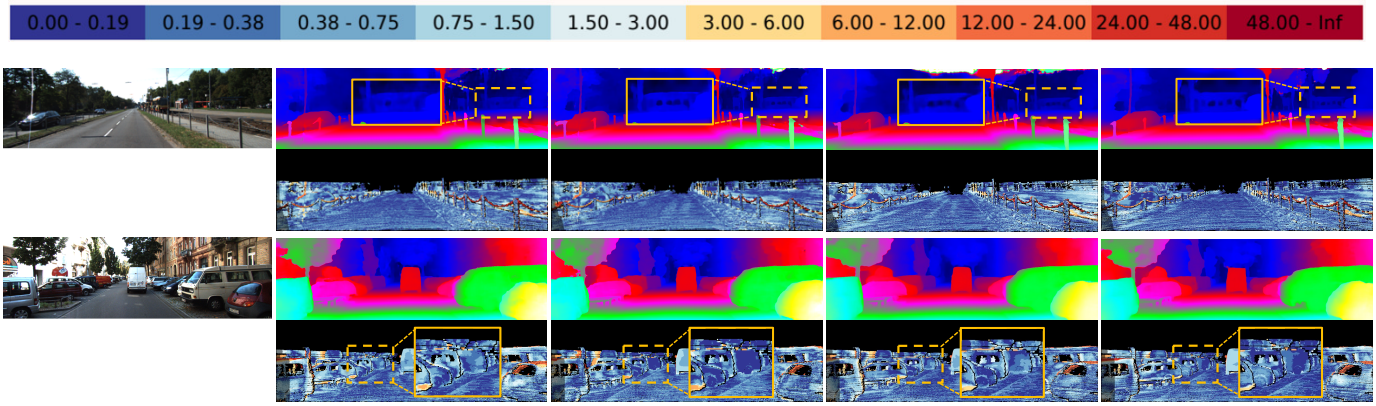
In addition, Table V also shows that including the $\mathcal{L}_{depth}^{fg}$ in loss calculations (by setting $\lambda > 0$) lowers the D1 error for foreground objects. However, the effect of including $\mathcal{L}_{depth}^{bg}$ is less pronounced. A possible explanation for this observation is that $\mathcal{L}_{depth}^{fg}$ and $\mathcal{L}_{depth}^{bg}$ somehow compliment each other and by including a properly weighted combination of both terms, the network produces better accuracy for background, even though the depth-based loss function by itself does not perform well for background depth measurement. However, when the importance of $\mathcal{L}_{depth}^{bg}$ is reduced (say for $\lambda \geq 0.8$), the performance deteriorates quickly.

D. Performance analysis of long range depth estimation

In this section, we compare the performance of long range depth estimation between the proposed method and the current state-of-the-art technique (SDN [17]). We named our method LR-PSMNet and LR-GwcNet (LR: Long Range). As listed in Table VI, our method significantly improves the performance of both models at different distances ($0-80$ meters), especially for estimates with depth values greater than 50m. Compared to SDN, LR-PSMNet and LR-GwcNet do not suffer from performance deterioration for close-by objects as our loss function combination is able to achieve good generalization at all depth ranges. More importantly, our results outperform the SDN + GDC method that uses sparse but accurate depth information, measured by 4-beam LiDAR, to refine the depth estimates [17].

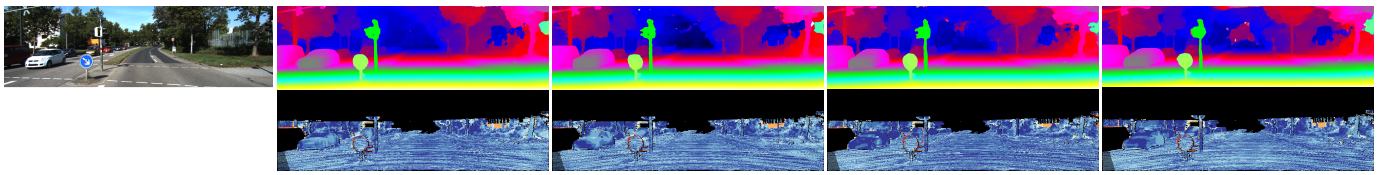
E. KITTI 2015 leaderboard

Although our work focuses on long range depth estimation, we also subjected the proposed method to KITTI performance evaluation exercise. The overall results of LR-PSMNet and LR-GwcNet was 2.06% and 2.09%, as listed in Table VII. By carefully redesigning the loss function, we have achieved



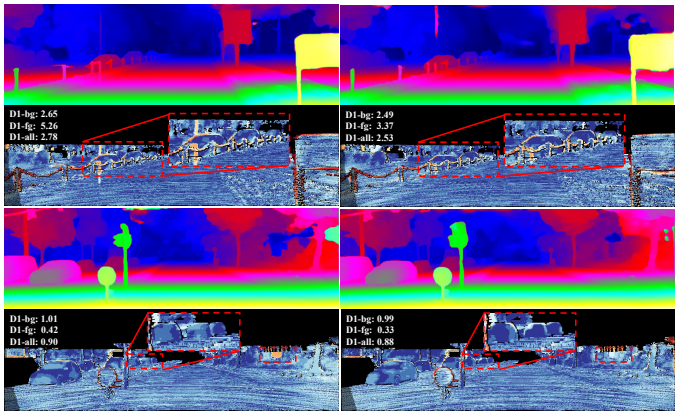
(a) Left Image (b) PSMNet [1] (c) LR-PSMNet (d) GwcNet [21] (e) LR-GwcNet

Fig. 6: Visualization of improvements over the baseline methods on KITTI 2015 dataset. For each example, predicted disparity map is illustrated on top row and error map on the bottom row. Improved areas are highlighted with yellow box. The numerical scale for color mapped on the error maps is provided on top of the figure. (Best viewed in color and zoom in for details.)



(a) Left Image (b) AcfNet (c) GANet (d) LR-PSMNet (e) LR-GwcNet

Fig. 7: Visualization results on KITTI 2015 dataset comparing our results with AcfNet [24] and GANet [23]. (Best viewed in color and zoom in for details.)



(a) CFNet (b) LR-CFNet

Fig. 8: Visualization of improvements of LR-CFNet over its baseline CFNet [64] on KITTI 2015 dataset. Improved areas are highlighted with red dashed box and the corresponding D1 errors for foreground (D1-fg), background (D1-bg) and all (D1-all) pixels are included in the top right corner of each error map. (Best viewed in color and zoom in for details.)

remarkable improvement in the overall performance of stereo disparity estimation in PSMNet and GwcNet. In addition, the rank of the PSMNet [1] method is improved from rank 124 to 72 (recorded on 16^{th} of March 2021). The qualitative comparisons between PSMNet and GwcNet and their long range variant (LR-PSMNet and LR-GwcNet) are included in Fig. 6. The proposed approach also achieves comparable disparity estimation accuracy for foreground and background as compared to other high performing methods (AcfNet [24], GANet [23]) - see Fig. 7 for details. Furthermore, the proposed loss functions also improve the disparity estimation performance of the recently proposed CFNet [64], for foreground objects, with marginal trade-off for background areas. The qualitative comparisons between CFNet and LR-CFNet are also included in 8. Moreover, when evaluated on non-occluded pixels, LR-CFNet outperforms the baseline CFNet (see Table VII).

V. Conclusion

This paper shows that one can effectively improve the performance of a depth estimation network for certain tasks by adjusting the bias (modifying the loss function) of the learning algorithm. We focused on improving the performance of stereo depth estimation for objects positioned at mid-range to far distances, which are arguably of interest in practical applications of autonomous driving. The existing disparity-based loss functions, and the commonly used training data, biases the models towards emphasising more on near-by objects and background areas at the expense of farther objects. To this end, we advocate adjustment of the biases to align the learning with a particular object-distance focus (by including the proposed foreground and background specific depth-based loss functions). We showed that including the depth-based loss

function effectively improved the performance of depth estimation for far-away objects, and by dividing it into foreground and background terms, one can balance the bias between the two classes. Our experimental results demonstrated that the proposed method effectively shifted the emphasis of the learning algorithm and achieved substantial improvement in long range depth estimation while also improving the overall disparity accuracy at all distances.

REFERENCES

[1] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410–5418.

[2] Y. Fang, I. Masaki, and B. Horn, "Depth-based target segmentation for intelligent vehicles: Fusion of radar and binocular stereo," *IEEE transactions on intelligent transportation systems*, vol. 3, no. 3, pp. 196–202, 2002.

[3] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8445–8453.

[4] S. Nedevschi, S. Bota, and C. Tomiuc, "Stereo-based pedestrian detection for collision-avoidance applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 3, pp. 380–391, 2009.

[5] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3d object detection methods for autonomous driving applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3782–3795, 2019.

[6] A. De la Escalera, J. M. Armingol, and M. Mata, "Traffic sign recognition and analysis for intelligent vehicles," *Image and vision computing*, vol. 21, no. 3, pp. 247–258, 2003.

[7] W. Chuah, R. Tennakoon, R. Hoseinnezhad, and A. Bab-Hadiashar, "Deep learning-based incorporation of planar constraints for robust stereo depth estimation in autonomous vehicle applications," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[8] V.-C. Miclea and S. Nedevschi, "A unified method for improving long-range accuracy of stereo and monocular depth estimation algorithms," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020, pp. 1234–1241.

[9] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 807–814.

[10] P. Pinggera, D. Pfeiffer, U. Franke, and R. Mester, "Know your limits: Accuracy of long range stereoscopic object measurements in practice," in *European conference on computer vision*. Springer, 2014, pp. 96–111.

[11] M. Shimizu and M. Okutomi, "Precise subpixel estimation on area-based matching," *Systems and Computers in Japan*, vol. 33, no. 7, pp. 1–10, 2002.

[12] D. Nehab, S. Rusinkiewiez, and J. Davis, "Improved sub-pixel stereo correspondences through symmetric refinement," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1. IEEE, 2005, pp. 557–563.

[13] S. K. Gehrig, H. Badino, and U. Franke, "Improving sub-pixel accuracy for long range stereo," *Computer Vision and Image Understanding*, vol. 116, no. 1, pp. 16–24, 2012.

[14] R. Ranftl, S. Gehrig, T. Pock, and H. Bischof, "Pushing the limits of stereo using variational stereo estimation," in *2012 IEEE Intelligent Vehicles Symposium*. IEEE, 2012, pp. 401–407.

[15] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[16] M. Menze, C. Heipke, and A. Geiger, "Object scene flow," *ISPRS Journal of Photogrammetry and Remote Sensing (JPRS)*, 2018.

[17] Y. You, Y. Wang, W. Chao, D. Garg, G. Pleiss, B. Hariharan, M. E. Campbell, and K. Q. Weinberger, "Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[18] R. Qian, D. Garg, Y. Wang, Y. You, S. Belongie, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. Chao, "End-to-end pseudo-lidar for image-based 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5881–5890.

[19] P. Pinggera, U. Franke, and R. Mester, "Highly accurate depth estimation for objects at large distances," in *German Conference on Pattern Recognition*. Springer, 2013, pp. 21–30.

[20] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4040–4048.

[21] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3273–3282.

[22] X. Cheng, P. Wang, and R. Yang, "Learning depth with convolutional spatial propagation network," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[23] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 185–194.

[24] Y. Zhang, Y. Chen, X. Bai, S. Yu, K. Yu, Z. Li, and K. Yang, "Adaptive unimodal cost volume filtering for deep stereo matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 926–12 934.

[25] H. Xu and J. Zhang, "Aanet: Adaptive aggregation network for efficient stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1959–1968.

[26] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European conference on computer vision*. Springer, 1994, pp. 151–158.

[27] J. Žbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *The journal of machine learning research*, vol. 17, no. 1, pp. 2287–2318, 2016.

[28] R. Szeliski and D. Scharstein, "Sampling the disparity space image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, pp. 419–425, 2004.

[29] V.-C. Miclea, C.-C. Vancea, and S. Nedevschi, "New sub-pixel interpolation functions for accurate real-time stereo-matching algorithms," in *2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2015, pp. 173–178.

[30] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence (IJCAI)*. Vancouver, British Columbia, 1981.

[31] D. Robinson and P. Milanfar, "Fundamental performance limits in image registration," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1185–1199, 2004.

[32] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 66–75.

[33] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5695–5703.

[34] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, and J. Zhang, "Learning for disparity estimation through feature constancy," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2811–2820.

[35] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[36] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[37] J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 887–895.

[38] X. Song, X. Zhao, H. Hu, and L. Fang, "Edgestereo: A context integrated residual pyramid network for stereo matching," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 20–35.

[39] G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia, "Segstereo: Exploiting semantic information for disparity estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 636–651.

[40] A. Li and Z. Yuan, "Occlusion aware stereo matching via cooperative unsupervised learning," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 197–213.

[41] A. Li, Z. Yuan, Y. Ling, W. Chi, S. Zhang, and C. Zhang, "Unsupervised occlusion-aware stereo matching with directed disparity smoothing," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[42] C. Zhou, H. Zhang, X. Shen, and J. Jia, "Unsupervised learning of stereo matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1567–1575.

[43] Y. Zhong, Y. Dai, and H. Li, "Self-supervised learning for stereo matching with self-improving ability," *arXiv preprint arXiv:1709.00930*, 2017.

[44] S. Joung, S. Kim, K. Park, and K. Sohn, "Unsupervised stereo matching using confidential correspondence consistency," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 2190–2203, 2019.

[45] H.-Y. Lai, Y.-H. Tsai, and W.-C. Chiu, "Bridging stereo matching and optical flow via spatiotemporal correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1890–1899.

[46] Y. Wang, P. Wang, Z. Yang, C. Luo, Y. Yang, and W. Xu, "Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8071–8081.

[47] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black, "Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 240–12 249.

[48] J. Zhang, K. A. Skinner, R. Vasudevan, and M. Johnson-Roberson, "Dispsegnet: Leveraging semantics for end-to-end learning of disparity estimation from stereo imagery," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1162–1169, 2019.

[49] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "Segflow: Joint learning for video object segmentation and optical flow," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 686–695.

[50] P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y.-C. F. Wang, "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2624–2632.

[51] J. Jiao, Y. Cao, Y. Song, and R. Lau, "Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 53–69.

[52] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang, "Joint task-recursive learning for semantic segmentation and depth estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 235–251.

[53] N. Zou, Z. Xiang, Y. Chen, S. Chen, and C. Qiao, "Simultaneous semantic segmentation and depth completion with constraint of boundary," *Sensors*, vol. 20, no. 3, p. 635, 2020.

[54] P. L. Dovesi, M. Poggi, L. Andraghetti, M. Martí, H. Kjellström, A. Pieropan, and S. Mattoccia, "Real-time semantic stereo matching," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 780–10 787.

[55] Z. Wu, X. Wu, X. Zhang, S. Wang, and L. Ju, "Semantic stereo matching with pyramid cost volumes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7484–7493.

[56] A. D. Pon, J. Ku, C. Li, and S. L. Waslander, "Object-centric stereo matching for 3d object detection," *arXiv preprint arXiv:1909.07566*, 2019.

[57] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[58] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[59] 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 the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[60] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou, "Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 899–908.

[61] X. Song, G. Yang, X. Zhu, H. Zhou, Z. Wang, and J. Shi, "Adastereo: A simple and efficient approach for adaptive stereo matching," *arXiv preprint arXiv:2004.04627*, 2020.

[62] C. Yao, Y. Jia, H. Di, Y. Wu, and L. Yu, "Content-aware inter-scale cost aggregation for stereo matching," *arXiv preprint arXiv:2006.03209*, 2020.

[63] S. Duggal, S. Wang, W.-C. Ma, R. Hu, and R. Urtasun, "Deeppruner: Learning efficient stereo matching via differentiable patchmatch," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4384–4393.

[64] Z. Shen, Y. Dai, and Z. Rao, "Cfnet: Cascade and fused cost volume for robust stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 13 906–13 915.