

Confidence Guided Stereo 3D Object Detection with Split Depth Estimation

Chengyao Li, Jason Ku, and Steven L. Waslander

Abstract— Accurate and reliable 3D object detection is vital to safe autonomous driving. Despite recent developments, the performance gap between stereo-based methods and LiDAR-based methods is still considerable. Accurate depth estimation is crucial to the performance of stereo-based 3D object detection methods, particularly for those pixels associated with objects in the foreground. Moreover, stereo-based methods suffer from high variance in the depth estimation accuracy, which is often not considered in the object detection pipeline. To tackle these two issues, we propose CG-Stereo, a confidence-guided stereo 3D object detection pipeline that uses separate decoders for foreground and background pixels during depth estimation, and leverages the confidence estimation from the depth estimation network as a soft attention mechanism in the 3D object detector. Our approach outperforms all state-of-the-art stereo-based 3D detectors on the KITTI benchmark.

I. INTRODUCTION

3D object detection is vital to applications such as autonomous driving. Many LiDAR-based methods [1], [2], [3] achieve strong performance due to the accurate depth information that LiDAR provides. Compared with LiDAR, a stereo camera setup is less expensive and provides more dense information. In addition, stereo detection could add redundancy to an autonomous driving system and help reduce safety risks in combination with LiDAR methods. Recent stereo-based methods [4], [5], [6], [7] have shown promising performance, although the detection performance gap between stereo and LiDAR configurations is still considerable.

One recent state-of-the-art stereo-based approach is Pseudo-LiDAR [4], [5], which first estimates disparities with a stereo matching network, converts the estimated disparities into a 3D point cloud, and then feeds the estimated point cloud to a LiDAR-based 3D object detector. However, compared with the LiDAR point cloud, the estimated point cloud often has poor depth estimation, particularly as depth increases, where it no longer preserves the shape of the objects. One attribute of this method is that the stereo matching algorithm jointly estimates both foreground and background pixels and does not learn specifically the depth and shape of the foreground objects. [8] shows that the depth distribution and pattern for foreground pixels and background pixels are different, and treating foreground and background pixels equally leads to sub-optimal results in a monocular depth estimation pipeline. In this paper, we propose to use two separate decoders for foreground and background pixels in a

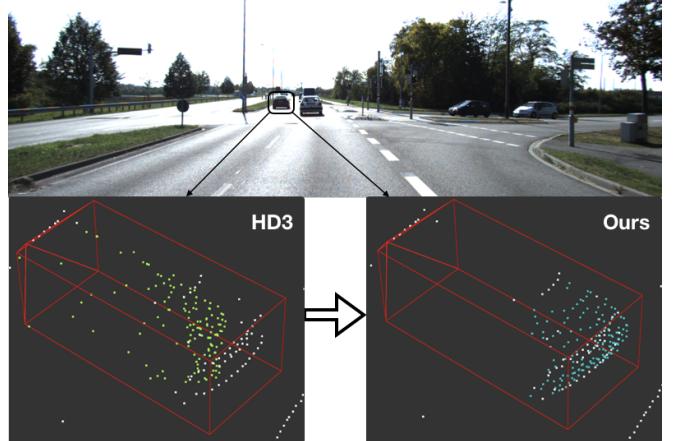


Fig. 1: A comparison between our proposed depth estimation module with our baseline HD³ on KITTI dataset. Using LIDAR measurements (shown in white) as a reference, our proposed method (shown in dark cyan) is able to learn the depth and shape of the car more accurately compared with our baseline (shown in yellow). The 3D bounding box is shown in red.

stereo matching network to provide better estimates of object shape and depth. The foreground and background masks can be obtained from image segmentation. In cooperation with the point cloud loss from [9], we show in Fig. 1 that with our approach the depth and shape of the object can be significantly improved. With further experiments, we show that such improvement also leads to better 3D object detection performance.

In contrast to LiDAR measured point clouds, the accuracy of stereo point clouds greatly varies across a scene. For constant pixel-level disparity error, the depth error increases as depth increases because of the effect of triangulation. In addition, the estimated point clouds also suffer from poor depth estimates at object boundaries, because it can be hard for a stereo matching algorithm to determine whether a pixel belongs to the object or the background [6]. In the original Pseudo-LiDAR pipeline [4], [5], the estimated point cloud is directly fed into a 3D object detector which does not consider any uncertainty information. To address this issue, we propose to encode the uncertainty output from the depth estimation module as an additional layer in the point cloud serving as a soft attention mechanism. This method allows the network to focus on points with high confidence and mitigates the effect of low confidence points such as points

Authors are with the University of Toronto Institute for Aerospace Studies. email: (chengyao.li@mail.utoronto.ca), (kujason.ku@mail.utoronto.ca), (stevenw@utias.utoronto.ca)

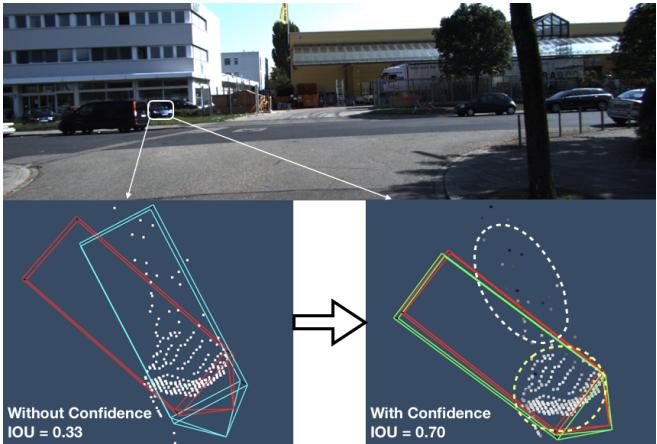


Fig. 2: A comparison between object bounding box detections without confidence (shown in teal) and with confidence (shown in green). Ground truth bounding box is shown in red. In the right image, points with higher confidence are shown in lighter colors. With confidence as an additional layer, the network is able to focus more on the points with high confidence (cycled in yellow) and ignore the points at object boundaries that have low confidence (cycled in white).

on the object boundary, as shown in Fig. 2.

In this paper, we propose CG-Stereo, a confidence guided stereo 3D object detection pipeline. To summarize, our main contributions are as follows:

- We propose the use of separate depth decoders for foreground and background pixels in the stereo matching network, which leads to improved depth estimation accuracy of the foreground pixels and improved object detection performance.
- We propose the use of confidence estimates from the stereo matching algorithm as a soft attention mechanism to guide the object detector network to focus more on the points with higher quality depth information, leading to further improvement in object detection accuracy.
- We demonstrate state-of-the-art performance that exceeds existing stereo-based methods on the challenging KITTI 3D object detection benchmark [10] for all three object classes. Specifically, our approach surpasses the next best-performing method by 1.4%, 6.7%, and 12.7% AP at 0.7 IOU on cars, pedestrians and cyclists, respectively.

II. RELATED WORK

Stereo depth estimation. For stereo vision, depth is often estimated by determining the stereo correspondences between the left and right images. Stereo matching is a well-established field of research [11], with a long history of classical methods [12], [13], [14]. With the recent development of deep learning, end-to-end learning methods have shown significant improvements in this task. A standard learning approach of stereo matching is to construct a 3D cost volume to minimize the matching cost [15], [16], [17].

Chang et al. [15] propose a pyramid pooling module for incorporating global context followed by a stacked hourglass 3D CNN. Yin et al. [17] determine the stereo matching by decomposing the full match density into multiple scales hierarchically first and then compose a global match density. This method not only achieves state-of-the-art results on established benchmarks but also predicts a confidence map that indicates the certainty of the estimation for each pixel. Our stereo 3D detection method takes advantage of the confidence estimation and demonstrates they can be used to improve stereo 3D object detection performances.

LiDAR-based 3D object detection. LiDAR-based object detection methods have shown strong performances and are widely used in autonomous driving since LiDAR provides accurate point clouds in terms of object depth and shape. Recent methods either use voxelization [2], [1], [18], [19], PointNet [20], [3], or a combination of the two [21], [22], [23] to learn features from point cloud data. Taking advantage of the mature pipeline of LiDAR-based detectors, the performance is transferable to stereo-based detectors.

Stereo-based 3D object detection. In recent years, stereo-based 3D object detection methods have shown promising improvements in performance. Stereo R-CNN [7] combines 2D proposals from both left and right images along with the sparse keypoints to generate coarse 3D bounding boxes, and then refines the bounding box using the photometric alignment of left and right regions of interest (ROIs). TLNet [24] projects the predefined anchor box to stereo images to obtain a pair of ROIs, learns to offset these ROIs, and uses triangulation to localize the objects. RT3DStereo [25] proposes to use semantic information together with disparity information to recover 3D bounding boxes. However, they do not take advantage of the semantic information to obtain better depth estimation. Pseudo-LiDAR [4], [5] proposes to mimic the LiDAR signal by converting the depth map to a point cloud, and then feed this point cloud to a LiDAR-based detector. This intuitive method reduces the performance gap between the stereo-based methods and LiDAR-based methods. However, the point cloud from stereo matching preserves streaking artifacts at the object boundaries, leading to inaccurate bounding box estimates. OC-Stereo [6] tries to solve this issue by estimating disparity only on the associated 2D bounding box area. However, this approach requires the 2D detection of the objects to be successful in both left and right images, which is difficult for objects that are truncated on image boundaries or are occluded from one view. It also completely ignores the background pixels which provide context to the 3D scene. Our method estimates the point cloud for both foreground and background pixels and keeps the points belonging to the ground plane since they contain useful contextual information in the 3D detection phase. The most recent state-of-the-art method, DSGN [26], proposes the use of a differentiable 3D volumetric representation of the environment to solve stereo 3D object detection. Their

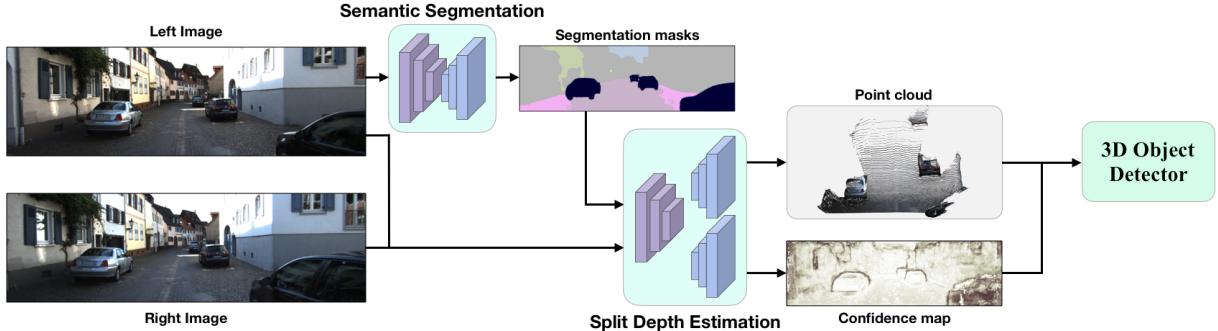


Fig. 3: Overview of our network. A semantic segmentation network first determines the foreground masks and background masks in the left image. The stereo matching network then estimates the disparities of the foreground and background pixels separately using two decoders and outputs the confidence associated with the estimation for each pixel. The disparity map is converted into a 3D point cloud with a confidence score as an additional layer. Points belonging to the background are filtered out except for the ones belonging to the ground plane. The resulting point cloud is then fed into a point cloud-based 3D object detector.

method achieves remarkable results on the KITTI car class, but the performances on pedestrians are not as competitive with the state-of-the-art. In comparison, our decomposed architecture allows us to perform well on pedestrians and cyclists even with the limited training data available for these classes in the KITTI dataset.

III. ARCHITECTURE

The overall pipeline of our method is shown in Fig. 3. First, a semantic segmentation network determines the foreground pixels and background pixels in the left image. We define foreground pixels as the pixels that belong to the objects of interest and background as all other pixels. Then, the stereo matching network estimates the disparities of the foreground and background pixels separately with two separate decoders. It also generates a confidence map associated with each pixel representing the certainty of the network's estimation. The disparity map is converted into a 3D point cloud with confidence as an additional layer. Points belonging to the background are filtered out except for the ones that lie on the ground plane. We use the same method as 3DOP [27] for stereo ground plane generation. The remaining point cloud is finally fed into a LiDAR-based 3D object detector.

A. Semantic Segmentation

For the 3D object detection task, it is common to segment the sensor input depending on the objects of interest. Pseudo-LiDAR [4], [5] converts the stereo image pair to a point cloud and then relies on a LiDAR-based 3D object detector to find objects. However, compared with LiDAR point clouds, the estimated point clouds have lower accuracy and thus are harder to segment. In addition, the texture and color information is lost in this process. We argue that it is possible to leverage image segmentation information in stereo object detection for improved detection accuracy, rather than relying on the LiDAR-based 3D detector exclusively. We also show

that the foreground and background masks from semantic segmentation improve the depth estimation of the foreground pixels in Section III-B, which contributes to higher 3D object detection accuracy.

B. Stereo Split Depth Estimation

Our stereo depth estimation is performed via stereo matching and the proposed formulation is agnostic to any stereo matching algorithm. We build on top of HD³ due to its state-of-the-art performance and ability to run in real-time. In addition, due to its probabilistic framework for match distribution estimation, an uncertainty associated with the estimate at every pixel can be naturally derived [17].

Stereo Matching Architecture. HD³ is designed for learning probabilistic pixel correspondences in both optical flow and stereo matching tasks [17], and we employ their stereo matching implementation as a baseline. The core idea of HD³ is to decompose the full discrete match distributions of pixel-wise correspondences into multiple scales hierarchically, estimate the local matching distributions at each scale, and then compose them from all levels. The resulting distributions at each pixel in the reference image are referred to as match densities. At each image scale level l , a cost volume is constructed to find the correlation between the pixels in both images and a density decoder is trained to estimate the decomposed match density p^l . For more details, readers may refer to the original HD³ paper [17].

Our modified version of the HD³ network is shown in Fig. 4. We only show one level for simplicity. Instead of relying on one density decoder for the entire image, we have two parallel density decoders with the same structure for foreground and background pixels, respectively. At each level l , each decoder takes the feature maps F^l , cost volume, and the density embedding from the previous level E_{fg}^{l-1}

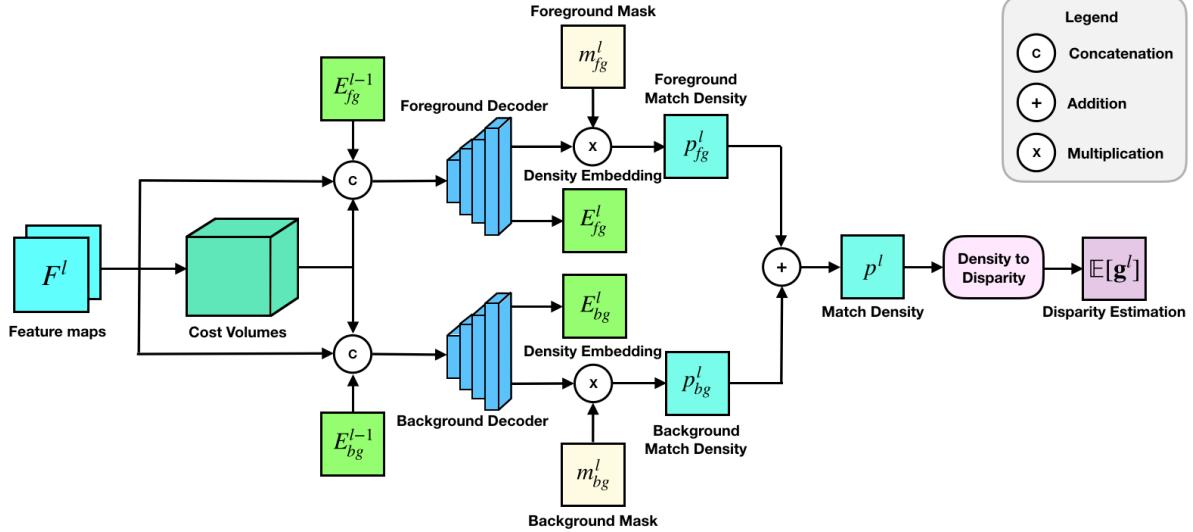


Fig. 4: Modified HD³ network at the l^{th} level. Instead of one density decoder for the entire image, we use separate density decoders for foreground pixels and background pixels, respectively, which allows us to optimize the weights specifically for each task.

or E_{bg}^{l-1} as input, and outputs an estimated match density p_{fg}^l or p_{bg}^l , and the density embedding at the current level E_{fg}^l or E_{bg}^l . Then, we use the foreground and background masks, denoted as m_{fg}^l and m_{bg}^l , to mask out the output from the two decoders, and then fuse them. The match density is then converted into an estimated residual disparity $\mathbb{E}[g^l]$ at the current level. The model-inherent uncertainty can be estimated by applying a softmax operation and a max-pooling operation to the estimated match density at the highest level.

Loss Function. We adapt the foreground-background sensitive loss function from a monocular-based method [8] and add the point cloud loss from [9]. The total loss is defined as

$$L_{total} = \lambda_f L_{fg} + (1 - \lambda_f) L_{bg} + \alpha L_{pc} \quad (1)$$

where λ_f is the weight coefficient representing the degree of preference for foreground pixels. L_{fg} and L_{bg} are the Kullback-Leibler divergence loss for foreground pixels and background pixels, respectively, α is a weight coefficient, and L_{pc} is the point cloud loss, which is set to be a smooth L1 loss of the difference between the foreground point cloud p_c with its ground truth point cloud p_{gt} in camera frame. We include the loss for background pixels because there is interdependence between foreground and background pixels for inferring the depth [8], and background provides context and support for 3D box regression [3]. We include a point cloud loss because it directly penalizes the estimated 3D point cloud and further improves the 3D detection accuracy as demonstrated in Sec. V-B.

C. Point Cloud Generation

The disparity map is converted to 3D points using the camera projection model as shown in Eq. 2.

$$x = \frac{(u - c_u)z}{f_u}, y = \frac{(v - c_v)z}{f_v}, z = \frac{f_u b}{d} \quad (2)$$

where x, y, z is the position of the points, (c_u, c_v) is the camera center, (f_u, f_v) is the focal length, b is the baseline, and d is the estimated disparity for a given pixel.

Image segmentation is a relatively mature field and can provide robust results. As a consequence, rather than feeding the entire point cloud to the 3D object detector, we leverage the segmentation masks and feed only the points that are estimated to be from the foreground pixels. We also keep the points that belong to the ground plane as they provide useful contextual information and supporting information for generating proposals in the detection stage.

D. Confidence Map

In comparison with LiDAR point clouds, the estimated point clouds often suffer from high variance in the accuracy of depth. To consider this difference, we take advantage of the confidence estimation from the stereo matching algorithm and encode this information in the point cloud simply as an additional layer as shown in Eq. 3.

$$p_c = \begin{bmatrix} x \\ y \\ z \\ \sigma \end{bmatrix} \quad (3)$$

where σ represents the confidence for each point. Fig. 5 shows the relationship between the confidence estimation and the estimation error for all the pixels belonging to cars on KITTI validation set. The inverse relationship shown in

Fig. 5 proves that the confidence estimates indicate the accuracy of the estimation and the quality of the estimated point clouds. Adding the confidence estimation as an additional layer is shown to improve object detection performance in Section V-B.

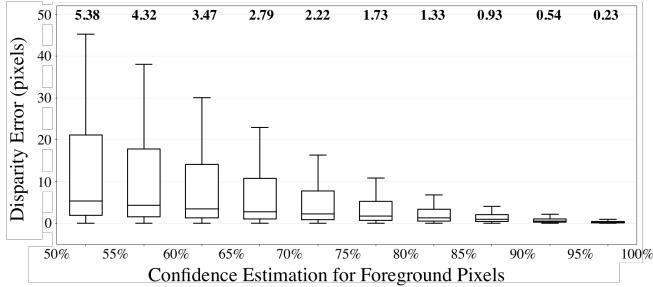


Fig. 5: Disparity Error vs. Confidence Estimation from modified HD³ for all pixels belonging to cars in the *validation* set on KITTI object detection benchmark [10]. Each box represents a 5% range in confidence. The median is shown on top of the box. There is a trend that a higher confidence estimation indicates a higher quality of the estimation output. We do not include plots with confidence lower than 50%, because they contain significantly fewer samples.

E. 3D Box Regression Network.

In the detection phase, we choose the open-source PointRCNN as our 3D object detector for its strong performance, and because it works directly on point clouds without voxelization, allowing us to encode the confidence estimation into the points.

IV. IMPLEMENTATION DETAILS

Semantic Segmentation. We employ VideoProp-LabelRelax [28] as the semantic segmentation network during inference. Since the KITTI semantic segmentation dataset [10] has only 200 labelled images, the network and model was instead trained on Mapillary [29] and Cityscapes [30] before being finetuned on KITTI. There is a class discrepancy between the object detection benchmark and the semantic segmentation benchmark on KITTI. On the object detection benchmark, the cyclist class is a single stand-alone class, but on the semantic segmentation benchmark, the rider and bike classes are separate. To solve this issue, we dilate the rider masks and check if they overlap with a bike mask. If there is an overlap, we keep the union of the original rider mask and the bike mask as a cyclist mask. All other bike masks are discarded.

Stereo Depth Estimation. For our stereo matching algorithm, we use the model pretrained on the FlyingThings3D Dataset [31], and then train the proposed two-decoder network on the training split of the KITTI object detection dataset. The foreground and background decoders have the same pre-trained weights before training on KITTI. To train on KITTI, we use the depth map

generated by depth completion [32] and their corresponding point clouds as ground truth. To obtain the ground truth instance segmentation masks used during training, we follow [9] and project ground truth points within the 3D labels to the image as the foreground masks. Training is performed for 375 epochs, with a batch size of 32 and a learning rate of 5×10^{-4} . The learning rate decays by 0.5 at the 125th, 187th, and 250th epochs. We apply horizontal flipping as data augmentation. Specifically, we increase the number of training samples by switching the left image and right image and horizontally flipping both of them. For this module, We train one model for cars, and another model for pedestrians and cyclists.

3D Object Detection. The region proposal network of PointRCNN is trained for 200 epochs with a batch size of 16 and a learning rate of 0.001, and the 3D box refinement network is trained for 50 epochs with a batch size of 8 and a learning rate of 0.001. For augmentations, we follow the original paper of PointRCNN [3]. Similar to PointRCNN, we subsample 16,384 points for each scene. Specifically, we sample half of the points that have depth larger than 20 m.

V. EXPERIMENTAL RESULTS

We evaluate our proposed method on the widely used KITTI 3D object detection dataset [10]. Specifically, we first compare the 3D object detection results with the state-of-the-art stereo-based detectors, then validate each contribution through ablation studies, and finally show qualitative results. KITTI contains 7,481 stereo image-pairs for training and 7,518 for testing. The benchmark also has annotations for 3 classes, which are cars, pedestrians and cyclists. Each annotation is categorized as easy, moderate, and hard based on the 2D box height, occlusion, and truncation. We follow the same training and validation split as other methods [4], [6], [7]. We also submit our results for all three classes to the online KITTI test server. KITTI recently changed its evaluation metrics on the test server. For the results on *test* set and in the ablation studies, we use the new KITTI metric which is mean average precision with 40 recall positions. For a fair comparison with other approaches, the results on the *validation* set are compared using the original KITTI metric with 11 recall positions.

A. AP Comparison with State-of-the-Art Methods

We compare our method with state-of-the-art stereo-based methods on the KITTI benchmark in Tables I, II and III. For the car class, our approach outperforms all state-of-the-art methods on the KITTI *validation* split in all categories. On the *test* set, our method ranked the first among all stereo-based methods on all three difficulties in both AP_{3D} and AP_{BEV} . Specifically for moderate difficulty, we show a 1.40% increase in AP_{3D} , and a 1.39% increase in AP_{BEV} . For pedestrians and cyclists, our proposed method outperforms all other stereo-based methods by significant margins. Most noticeably, on the test server, we have 6.73% and

Method	0.7 IoU			0.5 IoU		
	Easy	Moderate	Hard	Easy	Moderate	Hard
TLNet [24]	18.15 / 29.22	14.26 / 21.88	13.72 / 18.83	59.51 / 62.46	43.71 / 45.99	37.99 / 41.92
Stereo-RCNN [7]	54.11 / 68.50	36.69 / 48.30	31.07 / 41.47	85.84 / 87.13	66.28 / 74.11	57.24 / 58.93
PL:F-PointNet [4]	59.4 / 72.8	39.8 / 51.8	33.5 / 44.0	89.5 / 89.8	75.5 / 77.6	66.3 / 68.2
PL:AVOD [4]	61.9 / 74.9	45.3 / 56.8	39.0 / 49.0	88.5 / 89.0	76.4 / 77.5	61.2 / 68.7
PL++:AVOD [5]	63.2 / 77.0	46.8 / 63.7	39.8 / 56.0	89.0 / 89.4	77.8 / 79.0	69.1 / 70.1
PL++:PIXOR [5]	- / 79.7	- / 61.1	- / 54.5	- / 89.9	- / 78.4	- / 74.7
PL++:P-RCNN [5]	67.9 / 82.0	50.1 / 64.0	45.3 / 57.3	89.7 / 89.8	78.6 / 83.8	75.1 / 77.5
OC-Stereo [6]	64.07 / 77.66	48.34 / 65.95	40.39 / 51.20	89.65 / 90.01	80.03 / 80.63	70.34 / 71.06
DSGN [26]	73.21 / 83.24	54.27 / 63.91	47.71 / 57.83	- / -	- / -	- / -
Ours	76.17 / 87.31	57.82 / 68.69	54.63 / 65.80	90.58 / 97.04	87.01 / 88.58	79.76 / 80.34

TABLE I: **Car Localization and Detection.** AP_{3D} / AP_{BEV} on KITTI validation set. The results are evaluated using the original KITTI metric with 11 recall positions.

Method	AP_{3D}			AP_{BEV}		
	Easy	Moderate	Hard	Easy	Moderate	Hard
RT3DStereo [25]	29.90	23.28	18.96	58.81	46.82	38.38
Stereo-RCNN [7]	47.58	30.23	23.72	61.92	41.31	33.42
PL:AVOD [4]	54.53	34.05	28.25	67.30	45.00	38.40
PL++:P-RCNN [5]	61.11	42.43	36.99	78.31	58.01	51.25
OC-Stereo [6]	55.15	37.60	30.25	68.89	51.47	42.97
DSGN [26]	73.50	52.18	45.14	82.90	65.05	56.60
Ours	74.39	53.58	46.50	85.29	66.44	58.95

TABLE II: **Car Localization and Detection.** AP_{3D} and AP_{BEV} on KITTI test set. The results are evaluated using the new KITTI metric with 40 recall positions. Several methods are not available on the leaderboard.

Method	AP_{3D}			AP_{BEV}		
	Easy	Moderate	Hard	Easy	Moderate	Hard
Pedestrian						
RT3DStereo [25]	3.28	2.45	2.35	4.72	3.65	3.00
OC-Stereo [6]	24.48	17.58	15.60	29.79	20.80	18.62
DSGN [26]	20.53	15.55	14.15	26.61	20.75	18.86
Ours	33.22	24.31	20.95	39.24	29.56	25.87
Cyclist						
RT3DStereo [25]	5.29	3.37	2.57	7.03	4.10	3.88
OC-Stereo [6]	29.40	16.63	14.72	32.47	19.23	17.11
DSGN [26]	27.76	18.17	16.21	31.23	21.04	18.93
Ours	47.40	30.89	27.73	55.33	36.25	32.17

TABLE III: **Pedestrian and Cyclist Localization and Detection.** AP_{3D} and AP_{BEV} on KITTI test set. The results are evaluated using the new KITTI metric with 40 recall positions. Several methods are not available on the leaderboard.

12.72% AP increase in the 3D moderate category at 0.7 IoU for pedestrians and cyclists, respectively. For classes with limited data, our decomposed pipeline allows us to pretrain the sub-modules using additional datasets, which performs significantly better than other methods that lack this ability. The total inference time of our method is 0.57s on average on a GeForce RTX 2080 Ti GPU, which is faster than the current state-of-the-art method DSGN (0.68s) [26] and is comparable with other stereo-based methods on the KITTI leaderboard [10].

B. Ablation Studies

We analyze the effect of each added modules in Table. IV. The baseline is the original HD³ network with background points filtered out and PointRCNN as the 3D detector.

For depth estimation, we follow [8] and use mean absolute relative error (absRel) and scale invariant logarithmic error

(SILog) as the evaluation metrics. To better observe depth estimation accuracy improvements from the modifications made to HD³ for the more challenging but underrepresented pixels at greater depths, we present depth estimation errors for all foreground pixels with greater than 20 m depth.

Effect of Split Depth Estimation. To investigate the effect of split depth estimation, we train the modified HD³ network without the point cloud loss and feed the resulting point cloud directly to the 3D detector without a confidence score layer. The separate decoders allow us to improve the depth estimation for foreground pixels by 29.8% (from 0.047 to 0.033 absRel), and this leads to improvements in the AP_{3D} by 1.81% and AP_{BEV} by 1.31%.

Effect of Point Cloud loss. We analyze the effect of point cloud loss by feeding the point cloud to the 3D detector

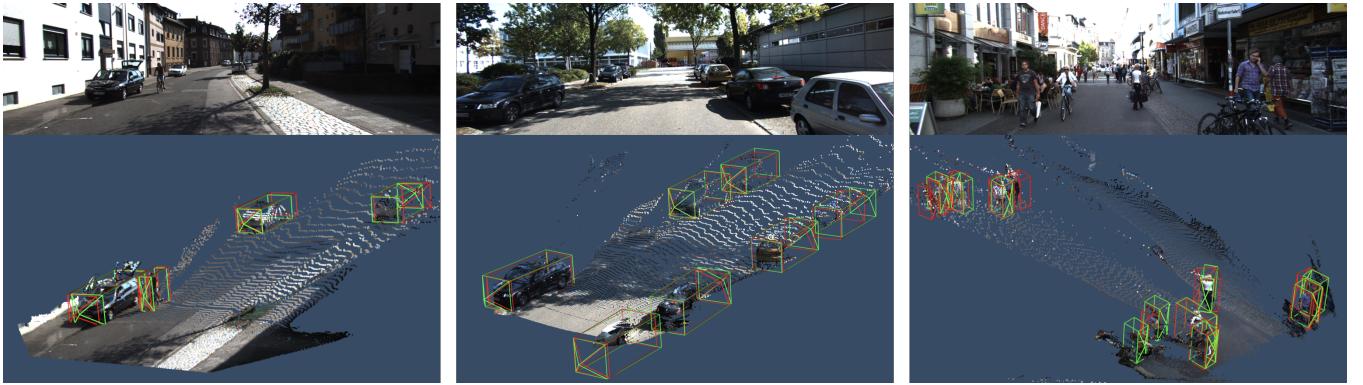


Fig. 6: Qualitative results of our method on several samples in the KITTI *validation* split. The ground truth labels and detections are shown in red and green, respectively.

without confidence estimation. Point cloud loss further improves the foreground depth estimation from 0.033 to 0.027 absRel. Including a point cloud loss in the pipeline also helps to obtain better 3D object detection accuracy, improving the AP_{3D} by 2.27% and AP_{BEV} by 0.29%.

Effect of Confidence Estimation. Finally, adding confidence estimation as an additional layer in the 3D detector boosts the 3D detection performance by another 1.02% and BEV performance by another 1.73%. This shows that the 3D detection network benefits from the confidence estimation generated by the depth estimation module.

Split Depth	L_{pc}	Confidence Feature	Foreground absRel	AP_{3D}	AP_{BEV}
-	-	-	0.047	52.48	67.84
✓	-	-	0.033	54.29	69.15
✓	✓	-	0.027	56.56	69.44
✓	✓	✓	-	57.58	71.17

TABLE IV: **Ablation Studies.** Comparison of depth estimation for foreground pixels, and comparisons of AP_{3D} and AP_{BEV} at 0.7 IoU for moderate difficulty for the car class. Both are evaluated on KITTI *validation* set. L_{pc} denotes the point cloud loss. *absRel* denotes the mean absolute relative error and *SILog* denotes the scale invariant logarithmic error.

Sensitivity to Semantic Segmentation. The performance of our method depends on the quality of the semantic segmentation. On the KITTI benchmark, there are only 200 images with semantic ground truth for training, which limits the performance of the semantic segmentation network. To investigate the performance upper bound of our method, we perform experiments using the labels that are generated from [33] as the segmentation masks on the KITTI *validation* split. Table V shows that with the labelled masks, there is a 5.36% and a 3.84% improvement in AP_{3D} and AP_{BEV} . This experiment suggests that our proposed method has the potential to obtain even better performance on other datasets with more accurate semantic segmentation.

Masks	AP_{3D}	AP_{BEV}
VideoProp-LabelRelax [28]	57.57	71.16
Labels [33]	62.93	75.00

TABLE V: Comparisons of AP_{3D} and AP_{BEV} for moderate difficulty for cars at 0.7 IoU using estimated segmentation masks [28] with using labelled masks [33] on KITTI *validation* set.

C. Qualitative Results

Fig. 6 shows the estimated point cloud, the ground truth bounding box, and the final detections of our proposed method on the KITTI *validation* split. The right image suggests that even with the significant improvements compared with the previous state-of-the-art methods, pedestrian and cyclist classes remain challenging for our stereo-based detector because of the limited training samples and the high intra-class variation.

VI. CONCLUSIONS

In this paper, we present CG-Stereo, a confidence-guided stereo 3D object detection pipeline with split depth estimation. Taking advantage of the mature development of image segmentation, the stereo matching network can learn the depth for foreground and background pixels separately, and achieve better foreground depth estimation and 3D object detection performance as a result. We also show that encoding the confidence estimation from the stereo matching network into the point cloud as a soft attention mechanism guides the 3D object detector to focus more on the accurate points and further boosts 3D object detection accuracy. Our proposed method outperforms all state-of-the-art stereo-based methods on the KITTI 3D object detection benchmark. Future work includes evaluating our proposed method on other autonomous driving datasets [34], [35], [36].

REFERENCES

- [1] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, “Joint 3d proposal generation and object detection from view aggregation,” in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.
- [2] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, “Multi-view 3d object detection network for autonomous driving,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1907–1915.
- [3] S. Shi, X. Wang, and H. Li, “Pointrcnn: 3d object proposal generation and detection from point cloud,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 770–779.
- [4] 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 (CVPR)*, 2019, pp. 8445–8453.
- [5] Y. You, Y. Wang, W.-L. Chao, D. Garg, G. Pleiss, B. Hariharan, M. Campbell, and K. Q. Weinberger, “Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving,” in *International Conference on Learning Representations (ICLR)*, 2020.
- [6] A. D. Pon, J. Ku, C. Li, and S. L. Waslander, “Object-centric stereo matching for 3d object detection,” *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020.
- [7] P. Li, X. Chen, and S. Shen, “Stereo r-cnn based 3d object detection for autonomous driving,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7644–7652.
- [8] X. Wang, W. Yin, T. Kong, Y. Jiang, L. Li, and C. Shen, “Task-aware monocular depth estimation for 3d object detection,” *arXiv preprint arXiv:1909.07701*, 2019.
- [9] J. Ku, A. D. Pon, and S. L. Waslander, “Monocular 3d object detection leveraging accurate proposals and shape reconstruction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 867–11 876.
- [10] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3354–3361.
- [11] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *International journal of computer vision*, vol. 47, no. 1-3, pp. 7–42, 2002.
- [12] E. Tola, V. Lepetit, and P. Fua, “A fast local descriptor for dense matching,” in *IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2008, pp. 1–8.
- [13] Y. Ohta and T. Kanade, “Stereo by intra-and inter-scanline search using dynamic programming,” *IEEE Transactions on pattern analysis and machine intelligence*, no. 2, pp. 139–154, 1985.
- [14] V. Kolmogorov and R. Zabih, “Computing visual correspondence with occlusions using graph cuts,” in *Proceedings Eighth IEEE International Conference on Computer Vision (ICCV)*, vol. 2. IEEE, 2001, pp. 508–515.
- [15] J.-R. Chang and Y.-S. Chen, “Pyramid stereo matching network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5410–5418.
- [16] 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 (CVPR)*, 2019, pp. 185–194.
- [17] Z. Yin, T. Darrell, and F. Yu, “Hierarchical discrete distribution decomposition for match density estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6044–6053.
- [18] Y. Zhou and O. Tuzel, “Voxelnet: End-to-end learning for point cloud based 3d object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4490–4499.
- [19] Y. Yan, Y. Mao, and B. Li, “Second: Sparsely embedded convolutional detection,” *Sensors*, vol. 18, no. 10, p. 3337, 2018.
- [20] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, “Frustum pointnets for 3d object detection from rgb-d data,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 918–927.
- [21] Y. Chen, S. Liu, X. Shen, and J. Jia, “Fast point r-cnn,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 9775–9784.
- [22] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, “Pv-rcnn: Point-voxel feature set abstraction for 3d object detection,” *arXiv preprint arXiv:1912.13192*, 2019.
- [23] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, “Pointpillars: Fast encoders for object detection from point clouds,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 697–12 705.
- [24] Z. Qin, J. Wang, and Y. Lu, “Triangulation learning network: from monocular to stereo 3d object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 7607–7615.
- [25] H. Königshof, N. O. Salscheider, and C. Stiller, “Realtime 3d object detection for automated driving using stereo vision and semantic information,” in *Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1405–1410.
- [26] Y. Chen, S. Liu, X. Shen, and J. Jia, “Dsgn: Deep stereo geometry network for 3d object detection,” *arXiv preprint arXiv:2001.03398*, 2020.
- [27] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, “3d object proposals for accurate object class detection,” in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 424–432.
- [28] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro, “Improving semantic segmentation via video propagation and label relaxation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8856–8865.
- [29] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, “The mapillary vistas dataset for semantic understanding of street scenes,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4990–4999.
- [30] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset,” in *CVPR Workshop on the Future of Datasets in Vision*, vol. 2, 2015.
- [31] 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 (CVPR)*, 2016, pp. 4040–4048.
- [32] J. Ku, A. Harakeh, and S. L. Waslander, “In defense of classical image processing: Fast depth completion on the cpu,” in *Proceedings of the 15th Conference on Computer and Robot Vision (CRV)*. IEEE, 2018, pp. 16–22.
- [33] L.-C. Chen, S. Fidler, A. L. Yuille, and R. Urtasun, “Beat the mTurkers: Automatic image labeling from weak 3d supervision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2014, pp. 3198–3205.
- [34] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, *et al.*, “Scalability in perception for autonomous driving: Waymo open dataset,” *arXiv*, pp. arXiv–1912, 2019.
- [35] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuscenes: A multimodal dataset for autonomous driving,” *arXiv preprint arXiv:1903.11027*, 2019.
- [36] M. Pitropov, D. Garcia, J. Rebello, M. Smart, C. Wang, K. Czarnecki, and S. Waslander, “Canadian adverse driving conditions dataset,” *arXiv preprint arXiv:2001.10117*, 2020.