This supplementary material contains the following five sections. In Section A we demonstrate that training DVF using ground truth projected 3D labels can generalize to multiple 2D detectors, including detectors that are not fine-tuned on KITTI *train* set. In Section B we conduct a range-based evaluation of DVF and show significant improvements over sparse fusion at far range. In Section C we discuss the effect of the simulated confidence sampling strategy on the performance of DVF models. In Section D we discuss the effect of the drop-out percentage on the performance of DVF models. In Section E we present the standard deviation of DVF models trained on KITTI dataset. In Section F we provide more examples for visualizing DVF. In Sections G we discuss limitations of the proposed method and discuss potential solutions. Finally, in Section H we briefly discuss potential negative impact of our work.

A. Inference using Pretrained 2D Object Detectors

One of the advantages of training DVF models directly with ground truth projected 3D bounding box labels rather than noisy, detector-specific, 2D predictions, is that it enables the generalization to any 2D object detector during inference time. To show this, instead of using CLOCs 2D predictions [18], we train a single DVF +PV-RCNN using ground truth labels and conduct inference using multiple 2D object detectors. Specifically, we use Detectron2 [33] pretrained Cascade-RCNN [1] and Mask-RCNN [8] models with ResNet-50 [9] backbone to extract 2D predictions without fine-tuning on KITTI images. In addition, we fine-tune the same models on KITTI *train* set and show 3D and BEV $AP|_{R_{40}}$ on *val* set compared to models that are not fine-tuned on KITTI *train* set. In Table 8 we observe that the trained model DVF +PV-RCNN gains are consistent compared to the baseline PV-RCNN [22] regardless of the pretrained 2D model. In addition, due to the small size of KITTI *train* set, fine-tuning does not have a significant advantage over using predictions from pretrained 2D object detectors.

B. Range-based Evaluation

DVF fuses at the voxel level which augments LiDAR point cloud information with dense details from image data. This is contrary to Pointpainting [27] where fusion is at the 3D point level, and is therefore sparse, especially at mid-to-long range. We show that dense fusion is especially useful at mid-to-long range, where more objects are occluded and LiDAR returns are sparse. In Table 9 we conduct a range-based evaluation for the 3D predictions of Pointpainting [27] and DVF applied to PV-RCNN [22]. We evaluate 3D AP $|_{R_{40}}$ on three range bins $[0.0, 20.0]$m, $[20.0, 40.0]$m and greater than 40m. Inference is done using 2D predictions from Cascade-RCNN [1]. In addition, we conduct inference using ground truth 2D bounding boxes for both models to determine an upper bound on the potential gains for Pointpainting and DVF. Dense voxel fusion shows significant improvement over Pointpainting [27] of +1.65% and +5.05% at long range (i.e., greater than 40m) using Cascade-RCNN [1] 2D predictions and ground truth labels respectively.

C. Effect of Simulated Confidence

To study the effect of the simulated confidence of the projected 3D bounding boxes on the performance of DVF models, we conduct an experiment using SECOND [37] where we vary the minimum confidence value a from 0.0 to 1.0, while the maximum confidence b is fixed at 1.0. For each experiment, we drop-out 50% of the ground truth samples added to each scene to simulate missed image detections. Looking at Figure 6, we observe that low confidence values result in poor performance since missed detections are already simulated using the drop-out strategy. In addition, very high a results in simulating an overly confident 2D object detector. The optimal values for SECOND [37] are around $[0.5, 0.8]$. The minimum simulated confidence is set at 0.8 for all our experiments. We have also experimented with sampling confidence values proportional to the size of the projected 3D bounding box, but did not achieve better results over the presented uniform sampling strategy.

D. Effect of Drop-out Percentage

During the training of DVF models, a percentage of the objects (i.e., ground truth samples) added to the point cloud, are not projected back to the foreground mask. The goal here is to simulate image missed detections. To study the effect of the drop-out percentage on the performance of DVF models, we conduct an experiment using SECOND [37] where we vary the drop-out percentage of the ground truth samples from 0% to 100%. Here, we set the maximum number of added ground truth samples per scene to 5.0. Similar to the implementation of ground truth sampling in OpenPCDet [26] library, only samples that do not collide with other samples in the current scene are added. Looking at Figure 7 we observe that dropping 100% of the added ground truth samples from the foreground mask results in a model that achieves almost the same performance as the baseline. We reason that the model learns to ignore the image information as it misses many objects that can be easily detected from the LiDAR data. On the other hand, setting the drop-out percentage 0.0% results in a model that relies too much on image information and thus is not robust against image missed detections. Dropping around 50% of the ground truth samples from the foreground mask is a good compromise. We set the drop-out percentage to 50%

Method	Inference 2D Detector			3D AP			BEV AP		
	Model	Pretrained	Fine-tune	Easy	Mod.	Hard	Easy	Mod.	Hard
PV-RCNN [22]	N/A	N/A	N/A	91.88	84.83	82.55	93.73	90.65	88.55
DVF+PV-RCNN	C-RCNN	COCO	Yes	92.81	85.53	83.01	95.85	91.22	88.94
	C-RCNN	COCO	No	92.55	85.37	82.92	95.32	91.26	88.93
	M-RCNN	Cityscapes	Yes	92.90	85.39	83.00	95.91	91.27	88.97
	M-RCNN	Cityscapes	No	92.95	85.47	82.94	96.02	91.32	88.95

Table 8. 3D and BEV $AP|_{R_{40}}$ results on KITTI [6] *val* set of a single DVF +PV-RCNN trained model using multiple 2D detectors at inference time. Two models are considered; Cascade-RCNN [1] **C-RCNN** and Mask-RCNN [8] **M-RCNN**. **C-RCNN** models are pretrained on COCO dataset [14] and **M-RCNN** are pretrained on Cityscapes dataset [4]. Finally, the same models are fine-tuned on images in KITTI [6] *train* set.

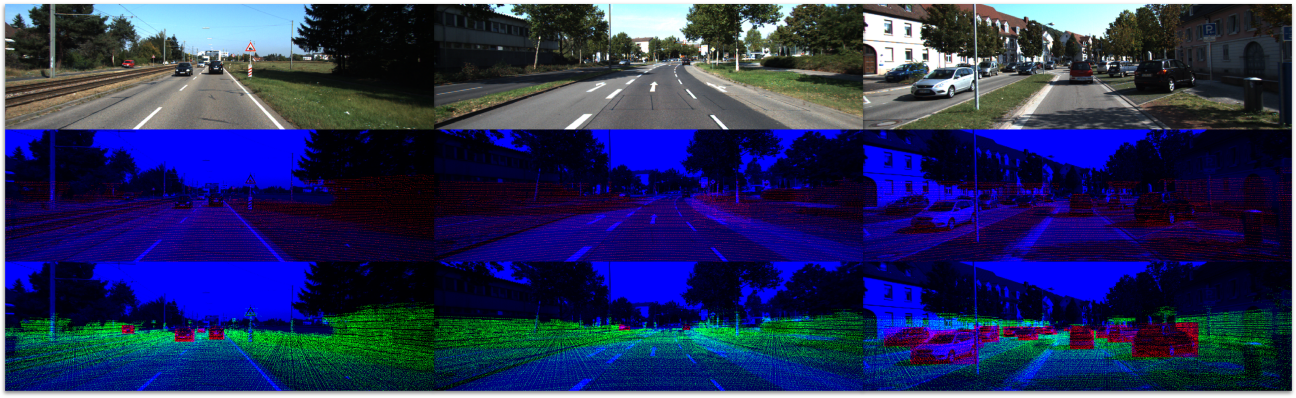

Figure 5. First row shows input images from KITTI [6] dataset, second row shows **red** points represent LiDAR returns projected onto the image plane, and third row shows multi-scale dense voxel centers projected onto the image plane, where **green** and **red** points are associated with background and foreground voxel features respectively. DVF increases the number of correspondences between image and LiDAR features which improves detection at mid-to-long range.

Method	C-RCNN Car - 3D AP			GT Car - 3D AP		
	0-20m	20-40m	40-Inf	0-20m	20-40m	40-Inf
Painted PV-RCNN	88.13	81.39	30.38	93.15	84.23	35.70
DVF+PV-RCNN	**90.62**	**81.48**	**32.03**	**93.60**	**85.29**	**40.75**
Improvement	+2.49	+0.09	+1.65	+0.45	+1.06	+5.05

Table 9. Range-based evaluation of Pointpainting [27] and DVF applied to PV-RCNN [22]. The same training strategy is used for both fusion methods. Inference on the KITTI *val* set is performed using both using Cascade-RCNN [1] 2D predictions **C-RCNN** and ground truth 2D bounding boxes **GT**. Improvements in 3D$AP|_{R_{40}}$ are shown relative to painted models.

for all our experiments.

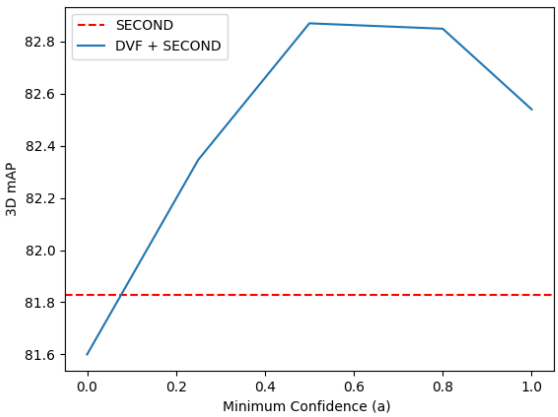

Figure 6. The effect of the minimum confidence on the performance of DVF + SECOND [37]. Here, we also show the performance of the baseline.

E. Standard Deviation of Experiments

Table 10 shows the standard deviation of DVF models. The standard deviation is computed based on running 3 experiments for each model. In addition, we also report the improvement relative to the baseline models.

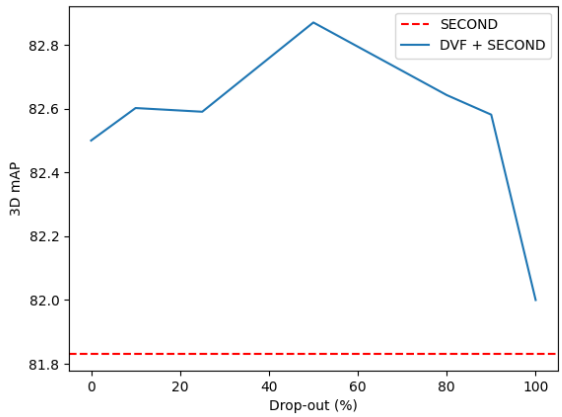

Figure 7. The effect of the drop-out percentage on the performance of DVF + SECOND [37]. Here, we also show the performance of the baseline.

Method	Car - 3D AP			Car - BEV AP		
	Easy	Mod.	Hard	Easy	Mod.	Hard
DVF + SECOND	0.43	0.24	0.19	0.51	0.34	0.11
Improvement	1.74	1.01	0.71	1.83	0.83	0.30
DVF + Voxel-RCNN	0.12	0.09	0.10	0.15	0.11	0.12
Improvement	0.52	0.61	0.31	0.66	0.63	0.54
DVF + PV-RCNN	0.23	0.12	0.08	0.04	0.11	0.07
Improvement	1.19	1.01	0.58	2.48	1.01	0.62

Table 10. The standard deviation of 3D and BEV AP $|_{R_{40}}$ on KITTI *val* set for DVF models on car class. The improvement relative to the baseline is also presented.

F. DVF Visualization

Figure 5 shows how DVF increases the number of correspondences between LiDAR point cloud and image pixels. Here, the second row shows the sparse point cloud projected onto the image plane. In the last row, we project voxel centers associated with occupied voxels onto 2D predicted foreground mask. Voxel centers with a foreground probability of more than 0.9 are colored in red, while the remaining voxels are considered background and are colored in green.

G. Limitations

Training Strategy. Our training strategy consists of training with ground truth 2D bounding boxes while simulating image missed detections throughout the training phase. We have demonstrated that training using this strategy generalizes better than training with erroneous 2D predictions. However, our proposed training strategy does not simulate errors in bounding box dimensions. One possible approach to address this limitation is to train DVF models by adding random noise to the center, width and height of ground truth 2D boxes to simulate small shifts in 2D predictions. Moreover, we could also simulate 2D prediction confidence based on occlusion, size and number of LiDAR returns per 3D bounding box.

Object Boundaries. One of the main advantages of using 2D bounding boxes for fusion compared to segmentation masks is the efficient inference time of 2D object detectors compared to segmentation networks. However, 2D bounding boxes do not capture object boundaries and thus are less accurate than segmentation masks. Hence, background voxels around bounding box boundaries are incorrectly weighted as foreground voxels. One possible approach to address to address this limitation is by modelling 2D bounding box detections as a mixture of Gaussian distributions. The intuition is that our confidence of whether a pixel is foreground or background within a bounding box falls off as we move away from the center of the bounding box. The parameters of each Gaussian distribution corresponding to a 2D prediction can be learned based on the width, height and depth distribution of LiDAR points within the bounding box.

H. Potential Negative Impact

Some potential negative impacts of 3D object detectors include using models for ethically questionable surveillance applications like detecting, counting, and tracking of peaceful demonstrators. Therefore, it is critical to enforce the anonymization of the input sensor data. For example, license plates of vehicles and faces of pedestrians and cyclists should be completely anonimized before conducting inference.