Abstract

As 3D perception problems grow in popularity and the need for large-scale labeled datasets for LiDAR semantic segmentation increase, new methods arise that aim to reduce the necessity for dense annotations by employing weakly-supervised training. However these methods continue to show weak boundary estimation and high false negative rates for small objects and distant sparse regions. We argue that such weaknesses can be compensated by using RGB images which provide a denser representation of the scene. We propose an image-guidance network (IGNet) which builds upon the idea of distilling high level feature information from a domain adapted synthetically trained 2D semantic segmentation network. We further utilize a one-way contrastive learning scheme alongside a novel mixing strategy called FOVMix, to combat the horizontal field-of-view mismatch between the two sensors and enhance the effects of image guidance. IGNet achieves state-of-the-art results for weakly-supervised LiDAR semantic segmentation on ScribbleKITTI, boasting up to 98% relative performance to fully supervised training with only 8% labeled points, while introducing no additional annotation burden or computational/memory cost during inference. Furthermore, we show that our contributions also prove effective for semi-supervised training, where IGNet claims state-of-the-art results on both ScribbleKITTI and SemanticKITTI.

1. Introduction

With the ever growing interest in 3D scene understanding for autonomous vehicles, semantic segmentation for LiDAR point clouds has also risen in popularity. To accurately and robustly learn the dense prediction task of generating per point class labels, a high volume of data is not only valuable but required. However manually labeling outdoor LiDAR scenes for semantic segmentation is both time consuming and expensive for large scale datasets.

There are two recently explored paths in the literature for reducing the labeling cost of outdoor LiDAR scenes: (i) by employing weak-supervision, where all frames have incomplete labels (e.g. by using line-scribbles [32]) and (ii) by employing semi-supervision, where a subset of frames are labeled and the rest remain completely unlabeled [15].

Commonly, LiDAR semantic segmentation models suffer from error prone boundary estimation between classes, as well as high false negative rates on both small objects and distant sparse regions. This is caused by the sparsity of LiDAR point clouds which severely reduces the number of points that fall on such regions to form an understandable and well separable geometry. As expected, these errors are further amplified when dealing with incomplete supervision, especially with scribble labels that completely forgo labeling boundaries. It can even be argued that such hard cases potentially need more representation within the dataset for correct and robust learning, something that clearly lacks under data-efficient settings.

These errors are severely reduced when operating on a denser representation of a scene (see Fig. 1 - top). Luckily, LiDAR sensors are commonly paired with cameras that are
not only cheaper but also provide a dense signal in the form of an RGB image that allows better separable boundaries (especially with the aid of RGB color channels), as well as orders of magnitude more pixels than points on small objects and distant regions. It is for this reason that all autonomous vehicles are equipped with a high resolution camera facing the front of the car to provide a denser and more complete understanding of the critical ego-vehicle path.

Our goal in this work is to leverage this high resolution image within our 3D pipeline to target the common weaknesses of LiDAR semantic segmentation models trained under incomplete supervision (weak labels). However we face two major challenges: (i) we need to retain our low annotation budget to have a scalable solution, therefore we cannot use additional annotated datasets or pretrained models in our setup; (ii) we need to tackle the issue of the horizontal field-of-view (FOV) mismatch between a LiDAR sensor and camera, where only a subset of points that fall onto the camera FOV have valid correspondence.

To this extent, we propose the Image-Guidance network (IGNet) that comprises of two core modules: (M1) domain adaptive image feature distillation that allows us to keep our low annotation budget and (M2) one-way contrastive learning that combats the FOV mismatch by leveraging image features to supervise out-of-image points. Throughout this work, we strictly associate the 2D domain with RGB images and 3D with LiDAR point clouds.

M1: Firstly, we train a 2D semantic segmentation model to generate per pixel high level features that better capture shape and context for sparse regions. By training on synthetic data, we avoid introducing any additional annotation requirements. We establish point-to-pixel correspondence between the LiDAR point cloud and the camera image (Fig. 1 - bottom), and distill the information from the generated features onto a 3D network via an auxiliary loss.

However, training on synthetic data yields yet another challenge: There exists a domain gap between synthetic images and real images that hinder performance in 2D. To further improve the quality of our image features, we propose using a domain adaptation (DA) pipeline to align our source domain onto the corresponding image. We further supervise the DA task via weak image labels generated by projecting the LiDAR labels onto the corresponding image.

M2: Next, we tackle the issue of the horizontal FOV mismatch between the camera and the LiDAR sensor. As our image-guidance module requires valid point-pixel correspondences, the auxiliary supervision remains limited to points that fall onto the image. To extend the supervision to points outside of the image, we propose using a one-way contrastive loss guided by a teacher model, allowing points that fall within the image to guide points that fall outside.

Here we observe that the number of pixel-to-outside-point-pairings remains limited as each LiDAR scan has a fixed associated image. This reduces the effect of the contrastive learning, especially since this single image alone often contains zero to a few object instances of each class. To combat this, we introduce a simple mixing strategy called FOVMix, where we cut and paste an image with its corresponding points from one scene onto another. With FOVMix, we are not only able to generate new pixel-point pairings to aid the contrastive learning but also increase the variability within each mini-batches.

To summarize:

- We propose using a synthetically trained 2D semantic segmentation model to guide the 3D network's feature space in order to improve boundary, distant region and sparse object segmentation.
- We employ weakly-supervised domain adaptation to further align the 2D features with our dataset.
- We extend the supervision from the image-guidance network to points out of the camera field-of-view via a one-way supervised contrastive loss.
- We propose a new mixing strategy called FOVMix to introduce additional variety into the dataset along with additional point-pixel pairings to extract further performance from our contrastive loss.

We achieve state-of-the-art results for weakly-supervised semantic segmentation on ScribbleKITTI [32]. We further show that IGNet can also be utilized for semi-supervised LiDAR segmentation to yield state-of-the-art results on both ScribbleKITTI and SemanticKITTI [2].

It should be noted that our proposed modules are only required during training, thus the performance boost comes without any additional computational or memory burden compared to the baseline 3D model during inference. Finally, as only synthetic data is required, we also do not introduce any additional annotation costs.

2. Related Work

Data Efficient LiDAR Semantic Segmentation: LiDAR semantic segmentation research has heavily focused on understanding how to best process the unordered data structure, with earlier focus on direct point based neural networks [14, 23, 24, 30, 33] having later shifted to sparse convolutional networks [9, 18, 28, 34, 42]. As architectures mature, we observe another developing area of interest: data efficiency within LiDAR semantic segmentation.

As known, the dense prediction task requires a large-scale annotated dataset, which is especially difficult and expensive to obtain for LiDAR point clouds [2]. Recent work therefore investigate two paths that aim to reduce this associated labeling cost: (i) weakly-supervised learning, where every frame is partially labeled, and (ii) semi-supervised learning, where only a subset of frames are labeled and the...
remaining stay completely unlabeled. However such approaches always come at the cost of performance, as reducing the number of labels within a dataset reduces the supervision provided to the model. Current popular literary work that deal with incomplete labels aim to extend the supervision to unlabeled points by (i) self-supervised training [5, 11, 44] where a model is trained on self-generated pseudo-labels or (ii) relying on a guidance network to generate on the fly targets (e.g. mean teacher [29, 31, 32]).


For 3D in particular, ScribbleKITTI [32] provides the first realistic benchmark for weakly supervised LiDAR semantic segmentation by introducing the scribble-annotated dataset. In their work, to reduce the gap to fully supervised training, they propose the SSLSS pipeline where they utilize a mean teacher setup [29] to stretch the supervision to unlabeled points, and extend CBST with a range component to deal with the increased sparsity of LiDAR point clouds. For works on indoor point clouds, PSD [37] utilizes similar consistency checks to align clean and perturbed outputs of unlabeled points. WS3D [19] utilizes region-level boundary awareness and instance discrimination to improve indoor and outdoor 3D semantic segmentation with simulated weak labels. Furthermore for semi-supervised learning, DiAL [31] uses a simple MT setup, GPC [15] proposes using a pseudo-label guided point contrastive loss, SSPC [8] utilizes self-training and LaserMix [16] uses a mixing operation to bring supervision to unlabeled frames. CPS [7] utilizes a Siamese structure to induce cross supervision.

**Multi-Modality with LiDAR and Image:** As mentioned, the additional information available in the corresponding RGB image does provide meaningful advantages that can improve LiDAR perception. Yet the task of incorporating this information within a robust pipeline is not trivial.

Fusion has been studied for a number of LiDAR based 3D perception tasks in a supervised and weakly-supervised manner [1, 4, 17, 20, 40, 41]. For LiDAR semantic segmentation PMF [43] and LIF-Seg [39] fuse the information from streams that process each modality individually to obtain higher information yielding features. However such approaches not only require image information during inference but also have linearly increasing memory and computation cost. 2DPASS [35] overcomes this by only using a one way information flow during training. Still, training the image stream on only LiDAR projected labels suffer heavily under incomplete annotations where it hinders performance instead of improving it. Sautier et al. [27] proposes a more general approach of self-supervised pretraining through the alignment of pixel- and point regions that still remains susceptible to forgetting (at a reduced scale).

**Mix-Augmentation:** Mixing operations have been very successful in increasing variability in the dataset and producing significant performance boosts for many tasks [6, 10, 12, 21, 25, 36, 38]. CutMix [36] mixes portions of the input and output of one sample image with another. MixMatch [3] applies the same mixing operation to labeled and unlabeled frames in a semi-supervised setting while generating labels via guessing and sharpening for unlabeled parts to provide supervision. Specifically for semi-supervised learning on LiDAR point clouds, LaserMix [16] aims to introduce variability through cylindrical and range-view partitioning and mixing.

3. Data Efficient LiDAR Segmentation

Data efficient LiDAR semantic segmentation aims to reduce the labeling cost associated with the dense prediction task by employing (i) weak supervision, where all frames have incomplete labels (e.g. by using scribble annotations), or (ii) semi supervision, where some frames have labels and others remain unlabeled. In either setting, naively training a model on available labeled points results in a considerable performance drop as only a small subset of points provide supervision. Specifically, we observe an amplified error rate caused by (i) weak boundary estimation between classes and (ii) misclassification of small objects and distant sparse regions, as LiDAR’s increased sparsity by range causes a severe reduction in the number of available points on an object to form an understandable geometry.

3.1. A Baseline Approach: Mean Teacher

As a first step in reducing the performance gap to fully supervised training we employ a generalized approach to utilize all points within the dataset. In specific, to extend the supervision to unlabeled points, following Unal et al. [32], we construct a mean teacher (MT) framework [29], where a student network is trained using a supervised loss \( L \) on an object to form an understandable geometry.

\[
\theta^E_{t} = \alpha \theta^E_{t-1} + (1 - \alpha) \theta_t
\]

The given update rule yields a teacher model that is a better and more robust predictor [22, 29]. To exploit this behaviour, we apply a consistency loss between the teacher and the student to align its outputs to the more accurate predictions, e.g. by minimizing the Kullback-Leibler divergence to the softmax outputs. Formally, for all points \( x \), the loss function can be redefined as:

\[
\mathcal{L} = H(\hat{y}, y) + 1_U(x) \text{KL}(\hat{y} || \hat{y}^E_{\text{EMA}})
\]

with \( \hat{y} \) and \( \hat{y}^E_{\text{EMA}} \) denoting the predictions of the student and teacher models, \( y \) the ground truth labels and \( U \) denoting
Firstly, we apply a forward pass to the camera image using a synthetically-trained semantic segmentation model to extract a high level feature representation ($\theta_{IG} : [0, 255]^3 \rightarrow f_{IG} \in \mathbb{R}^d$). It should be noted that we opt to use synthetic data to avoid introducing any additional annotation burden as the collection of new labeled samples can be easily automated. Using available intrinsic and extrinsic camera matrices $K$ and $[R|t]$ respectively, we project the 3D points cloud in homogeneous coordinates $x_{hom}$ onto the rectified camera coordinates following $x_{rec}^T = K[R|t]x_{hom}^T$ and extract point to pixel mappings $m : x_{rec} \mapsto (k, l)$ with $k = [x_{rec}^{(0)} / x_{rec}^{(2)}]$ and $l = x_{rec}^{(1)} / x_{rec}^{(2)}$. A point to pixel correspondence is considered valid if the pixel $(k, l)$ falls within the image.

While a mean teacher framework does allow us to utilize the entire dataset within our training pipeline, due to the lack of direct supervision, similar to the student, the teacher’s predictions remain uncertain and error prone for points that lie on class boundaries or for sparsely represented classes (e.g. volumetrically small objects or distant regions), especially when trained on weak scribble labels that completely forgo labeling any boundary points.

### 3.2. Image Guidance via Feature Distillation

To target these weaknesses we propose using image feature distillation from a trained 2D semantic segmentation model. But before we dive deep into the details, it is important to establish motivation.

RGB images provide a much denser representation of a scene compared to LiDAR point clouds. This increased density along with the available color channels allow easier distinction of both class boundaries as well as small objects and distant regions. 2D semantic segmentation models can therefore learn better separable and richer features for such pixels. Following this observation, we propose introducing an image guidance (IG) network to exploit the mature features of a trained 2D semantic segmentation model.

Firstly, we apply a forward pass to the camera image using a synthetically-trained semantic segmentation model to extract a high level feature representation ($\theta_{IG} : [0, 255]^3 \rightarrow f_{IG} \in \mathbb{R}^d$). It should be noted that we opt to use synthetic data to avoid introducing any additional annotation burden as the collection of new labeled samples can be easily automated. Using available intrinsic and extrinsic camera matrices $K$ and $[R|t]$ respectively, we project the 3D points cloud in homogeneous coordinates $x_{hom}$ onto the rectified camera coordinates following $x_{rec}^T = K[R|t]x_{hom}^T$ and extract point to pixel mappings $m : x_{rec} \mapsto (k, l)$ with $k = [x_{rec}^{(0)} / x_{rec}^{(2)}]$ and $l = x_{rec}^{(1)} / x_{rec}^{(2)}$. A point to pixel correspondence is considered valid if the pixel $(k, l)$ falls within the image.

We extend our 3D model with an auxiliary head that maps the final layer features to the image feature dimension $d$. During training, we introduce a new consistency term between the student and the IG teacher that is applied to all points that have a valid pixel correspondence. Formally, we restate the loss function to include image-guidance as:

$$\mathcal{L} = H(\hat{y}, y) + 1_U(x) \text{KL}(\hat{y} \| \hat{y}_{EMA}) + \mathcal{L}_{IG}$$

with $\mathcal{L}_{IG} = 1_I(x, m(x)) \text{KL}(sm(f) \| sm(f_{IG}))$  

with $I$ denoting the set of points with valid pixel correspondence, $sm$ denoting the softmax operation, $f, f_{IG} \in \mathbb{R}^{N \times C}$ denoting the feature representations of the 3D auxiliary head and IG decoders respectively.

With the addition of the auxiliary loss, the 3D network aims to mimic the more mature representation of the 2D network for points with pixel correspondences. In other words, we introduce a new teacher model, where boundary points along with small and distant objects more richly defined due to the denser representation, to further and better guide the student on unlabeled points. An illustration of the proposed module can be seen in Eq. 2 - red.

It should be noted that the IG network is only required during training and can be completely removed for inference alongside the auxiliary head, causing no additional memory requirements or time costs to the overall 3D model.
3.3. 2D Weakly-Supervised Domain Adaption

As mentioned before, in order to train $\theta_{IG}$ for semantic segmentation, we resort to synthetic data. It has the desirable property that even dense annotations can be automatically generated so that no additional labeling cost is introduced. However, a model trained on synthetic source data $(I_s, S_s)$, usually experiences a performance drop when applied to real-world target images $I_t$ due to the domain gap.

To tackle this, we propose employing a domain adaptation pipeline to improve the quality of the extracted features and better align with the data from our real-world training set. Following current literature [13], we reestablish a mean teacher framework [29] and use the teacher model to generate pseudo labels $P_t$ for the target domain images by freezing the unlabeled image predictions. We train the 2D network with a linear classification layer $\gamma$ not only on the synthetic image-label pairings $(I_s, S_s)$ but also on the target images with pseudo labels $(I_t, P_t)$. Formally, the loss for the 2D model can be defined as:

$$L = L_S + L_{DA}$$

with $L_S = H(\gamma(\theta_{IG}(I_s)), S_s)$

and $L_{DA} = H(\gamma(\theta_{IG}(I_t)), P_t)$

Formally, we define the one-way supervised contrastive loss with $\lambda_p$:

$$L_{DA} = \lambda_p H(\gamma(\theta_{IG}(I_t)), P_t)$$

(5)

with $\lambda_p = \lambda_p$ for pixels with valid point mapping and 1 otherwise. An illustration of the proposed weakly-supervised domain adaptation pipeline can be seen in Fig. 2 - blue.

Finally, to form the image guidance model $\theta_{IG}$, we copy and freeze the 2D student model (following unsupervised domain adaptation convention [13]) without the linear classifier and use its generated features to guide the 3D student model during training.

3.4. Extending the Supervision Beyond the Image

With image-guidance (Eq. 3) the information distillation from the mature 2D features to the 3D pipeline is limited by the availability of point-pixel correspondences. For many cases, we are limited to a front facing camera, so there exists a big mismatch between the horizontal FOV of the two sensors. Under such a setup, the set of all points with valid pixel correspondence $(I)$ is much smaller than the set of all points without a valid correspondence $(O = I \cap P)$, i.e. $|I| < |O|$. In other words, the lack 360° coverage for the camera means that points with pixel correspondence only make up a small portion of the LiDAR point cloud.

To be able to guide points outside of the image using the 2D domain adapted features, we introduce an extension to the image-guidance loss with a one-way supervised contrastive loss (CL).

Let $I^{(c)} \subseteq I$ and $O^{(c)} \subseteq O$ define two sets of points inside and outside of the image respectively with associated class $c = \arg\max \hat{y}_{EMA}$, given by the teacher’s prediction. Formally, we define the one-way supervised contrastive loss as:

$$L_{CL} = \sum_{c \in O^{(c)}} \sum_{i \in I^{(c)}} -\log \left( \frac{1}{|O^{(c)}|} \sum_{i \in I^{(c)}} \sum_{i' \in I} \exp(f_o \cdot f_{IG,i}/\tau) \right)$$

(6)

with $\tau$ denoting the temperature. The total loss can then be formulated as:

$$L = H(\hat{y}, y) + 1 \nu(x) KL(\hat{y} || \hat{y}_{EMA}) + L_{IG} + \lambda L_{CL}$$

(7)

with $\lambda$ denoting the scale hyperparameter.

As illustrated in Fig. 3, the loss extension aims to apply a pull force to all points towards pixels of the same category while also applying a push to all points away from pixels of a different class. We therefore align the features of points outside of the image with the features of the 2D image-guidance network.
3.5. FOVMix

Finally, we introduce a new mixing operation called FOVMix. Given two data samples \((x_A, y_A, I_A)\) and \((x_B, y_B, I_B)\), the goal of FOVMix is to generate a new training sample \((\tilde{x}, \tilde{y}, \tilde{I})\). Simply put, we take an image from sample A and replace it with the image of sample B. To accompany this, we further take all points that are within the image FOV of sample A, and paste them onto sample B while removing all points of B that were in the same region. An illustration of FOVMix can be seen in Fig. 4.

Formally, we define the mixing operation as:

\[
\begin{align*}
\tilde{x} &= [M_{AA} \odot x_A, (1 - M_{AA}) \odot x_B] \\
\tilde{y} &= [M_{AB} \odot y_A, (1 - M_{AB}) \odot y_B] \\
\tilde{I} &= I_A
\end{align*}
\]

Where \(M_{AB}, M_{AA} \in \{0, 1\}^N\) denote the binary masks that yield the points within the image FOV given the intrinsic projection matrix \(A\) and extrinsic projection matrices \(A\) and \(B\) respectively, \(\odot\) and \([,\] denoting a dot product for masking and concatenation operations. Thus, FOVMix does not depend on a specific sensor/setting, but only relies on the availability of point to pixel correspondences, which is expected for systems with both a LiDAR sensor and camera.

FOVMix is a simple operation that accomplishes two feats: (i) it increases the effectiveness of the one-way contrastive loss by introducing additional pairings of points inside-outside of the image, (ii) it increases the richness of the data within each mini-batch. While FOVMix introduces noise along the boundaries of the image FOV similar to other mixing methods commonly used in dense vision tasks, the increased diversity and richness of each mini-batch is a worthy trade-off against the introduced noise.

4. Experiments

The implementation details alongside the dataset introductions can be found in the supplementary materials.

4.1. Results

Weakly-Supervised LiDAR Segmentation: We report the performance of our image-guidance network (IGNet) trained with scribble-supervision in Tab. 1. As seen, IGNet outperforms previous SOTA, showing improvements across the board for all classes and reaching 96.4\% relative performance when compared to fully supervised training while only using 8\% labeled points. In specific, we observe large gains for small object categories such as bicycle and motorcycle when compared to the previous SOTA SSLSS [32].

It should be noted that, in contrast to SSLSS, IGNet does not require self-training. Therefore the training times are considerably reduced (from 5 days to 1 - including the 2D training - using 8 Nvidia RTX2080Ti’s). Still, to further push performance, we can IGNet++. Here, we replace the Cylinder3D backbone of SSLSS with IGNet and therefore employ the same class-range-balanced self-training scheme on top of our image guidance to achieve 63\% mIoU, i.e. 98\% relative performance compared to fully supervised.

Semi-Supervised LiDAR Segmentation: We also show that IGNet can be used for all data-efficient LiDAR semantic segmentation settings. In particular, we report results for (i) semi-supervised training using SemanticKITTI [2] and (ii) semi- and weakly-supervised training on ScribbleKITTI [32], where we carry experiments on a semi-supervised setting while training with a weakly-supervised dataset. We follow Kong et al. [16] and generate a semi-supervised dataset by uniformly sampling frames.

As seen in Tab. 2, IGNet outperforms previous SOTA’s by a considerable margin on almost all cases. Specifically, as expected, we see greater margins of improvement in the ScribbleKITTI semi-supervised benchmark since the image-guidance can be more effectively utilized to learn boundary information despite the lack of any such labels. We also report a direct comparison to the baseline Cylinder3D model where IGNet shows great absolute mIoU improvements of 4.1\% – 9.7\% while introducing no additional memory or computational requirements during inference.

4.2. Ablation Studies

We conduct ablation studies on the ScribbleKITTI [32] dataset, where alongside the mIoU, we also report the relative performance of our model compared to the baseline Cylinder3D [42] trained on densely annotated labels.

Effects of Network Components: We first investigate the effects our proposed components. Starting from a baseline model, we introduce each module one by one, reporting the mIoU and relative performances in Tab. 3. As seen
Table 1. Weakly-supervised 3D semantic segmentation results on ScribbleKITTI. We not only show results from our proposed image-guidance network (IGNet), but also its performance difference compared to the baseline Cylinder3D model and the results of using IGNet within a class-range-balanced self-training pipeline (IGNet++). * indicated methods that do not use Cylinder3D as their backbone.

<table>
<thead>
<tr>
<th>Method</th>
<th>mIoU</th>
<th>car</th>
<th>bicycle</th>
<th>motorcycle</th>
<th>truck</th>
<th>vehicle</th>
<th>person</th>
<th>cyclist</th>
<th>mIoU rel</th>
<th>Δrel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cylinder3D [42]</td>
<td>57.0</td>
<td>88.5</td>
<td>39.9</td>
<td>58.0</td>
<td>58.4</td>
<td>48.1</td>
<td>68.6</td>
<td>77.0</td>
<td>85.4</td>
<td>30.4</td>
</tr>
<tr>
<td>MinkNet* [9]</td>
<td>58.5</td>
<td>91.1</td>
<td>23.8</td>
<td>59.0</td>
<td>66.3</td>
<td>58.6</td>
<td>65.2</td>
<td>75.2</td>
<td>0.0</td>
<td>83.8</td>
</tr>
<tr>
<td>SPVCNN* [38]</td>
<td>56.9</td>
<td>88.6</td>
<td>25.7</td>
<td>55.9</td>
<td>67.4</td>
<td>48.8</td>
<td>65.0</td>
<td>78.2</td>
<td>0.0</td>
<td>82.6</td>
</tr>
<tr>
<td>MT [29]</td>
<td>59.0</td>
<td>91.0</td>
<td>41.1</td>
<td>58.1</td>
<td>85.5</td>
<td>57.1</td>
<td>71.7</td>
<td>80.9</td>
<td>0.0</td>
<td>87.2</td>
</tr>
<tr>
<td>CBST [44]</td>
<td>60.8</td>
<td><strong>92.4</strong></td>
<td>39.1</td>
<td>58.5</td>
<td>78.5</td>
<td>57.0</td>
<td>70.0</td>
<td>77.4</td>
<td>0.0</td>
<td>86.9</td>
</tr>
<tr>
<td>DARS [11]</td>
<td>60.8</td>
<td>91.9</td>
<td>39.3</td>
<td>57.9</td>
<td>78.6</td>
<td>53.3</td>
<td>69.5</td>
<td>77.1</td>
<td>0.0</td>
<td>86.6</td>
</tr>
<tr>
<td>SSLSS [32]</td>
<td>61.3</td>
<td>91.0</td>
<td>41.1</td>
<td>58.1</td>
<td>85.5</td>
<td>57.1</td>
<td>71.7</td>
<td>80.9</td>
<td>0.0</td>
<td>87.2</td>
</tr>
</tbody>
</table>

Table 2. Comparison of state-of-the-art methods for semi-supervised LiDAR semantic segmentation. The uniform frame sampling rate is indicated by [%].

Is Domain Adaptation Necessary? We further investigate the necessity of domain adaptation for our image-guidance network. Starting from a mean teacher framework, we compare the performance of our 3D model when guided by the DAFormer model [13] trained on (i) weak labels that we generate by projecting 3D scribbles onto the image, and (ii) the synthetically generated GTA-V dataset [26], as well as the complete DAFormer pipeline (model + DA) with (iii) GTA-V → ScribbleKITTI, and (iv) GTA-V → ScribbleKITTI with additional projected weak supervision. The results are shown in Tab. 4 which emphasize the importance of DA and the usefulness of the weak supervision.

Where do the Improvements Come From? Our goal when using image features to guide our 3D model is to exploit the better representation capabilities of 2D semantic segmentation models trained on denser representations for (i) border points, where color channels can provide finer separation compared to noisy LiDAR measurements, (ii) small object and sparsely represented regions, where the pixel count remains considerably higher compared to the LiDAR point count. Finally, we conduct an ablation study to investigate if this behaviour can be observed in the model accuracy after introducing the 2D image-guidance module.

In Tab. 5, we isolate the effects of our image guidance module by directly comparing to the mean teacher. Firstly, we show that the introduction of image-guidance does boost the border accuracy significantly (+3.5%). Here, we classify points to be on a border if any of its closes \( N = 16 \) neighbors in 3D space do not share the same class. Second, we observe that IGNet obtains a considerably better performance (+6.6%) on small objects (pedestrians and two-wheelers) compared to the gain in larger objects (+1.5% for four-wheelers). Lastly, when comparing accuracy changes by range, sparsely represented distant regions beyond 25m of range show an improvement of +2.0% when compared to the MT baseline, while close regions only see marginal gains.

<table>
<thead>
<tr>
<th>Method</th>
<th>mIoU</th>
<th>car</th>
<th>bicycle</th>
<th>motorcycle</th>
<th>truck</th>
<th>vehicle</th>
<th>person</th>
<th>cyclist</th>
<th>mIoU rel</th>
<th>Δrel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cylinder3D [42]</td>
<td>57.0</td>
<td>88.5</td>
<td>39.9</td>
<td>58.0</td>
<td>58.4</td>
<td>48.1</td>
<td>68.6</td>
<td>77.0</td>
<td>85.4</td>
<td>30.4</td>
</tr>
<tr>
<td>MinkNet* [9]</td>
<td>58.5</td>
<td>91.1</td>
<td>23.8</td>
<td>59.0</td>
<td>66.3</td>
<td>58.6</td>
<td>65.2</td>
<td>75.2</td>
<td>0.0</td>
<td>83.8</td>
</tr>
<tr>
<td>SPVCNN* [38]</td>
<td>56.9</td>
<td>88.6</td>
<td>25.7</td>
<td>55.9</td>
<td>67.4</td>
<td>48.8</td>
<td>65.0</td>
<td>78.2</td>
<td>0.0</td>
<td>82.6</td>
</tr>
<tr>
<td>MT [29]</td>
<td>59.0</td>
<td>91.0</td>
<td>41.1</td>
<td>58.1</td>
<td>85.5</td>
<td>57.1</td>
<td>71.7</td>
<td>80.9</td>
<td>0.0</td>
<td>87.2</td>
</tr>
<tr>
<td>CBST [44]</td>
<td>60.8</td>
<td><strong>92.4</strong></td>
<td>39.1</td>
<td>58.5</td>
<td>78.5</td>
<td>57.0</td>
<td>70.0</td>
<td>77.4</td>
<td>0.0</td>
<td>86.9</td>
</tr>
<tr>
<td>DARS [11]</td>
<td>60.8</td>
<td>91.9</td>
<td>39.3</td>
<td>57.9</td>
<td>78.6</td>
<td>53.3</td>
<td>69.5</td>
<td>77.1</td>
<td>0.0</td>
<td>86.6</td>
</tr>
<tr>
<td>SSLSS [32]</td>
<td>61.3</td>
<td>91.0</td>
<td>41.1</td>
<td>58.1</td>
<td>85.5</td>
<td>57.1</td>
<td>71.7</td>
<td>80.9</td>
<td>0.0</td>
<td>87.2</td>
</tr>
</tbody>
</table>

Table 3. Ablation study where starting from the baseline Cylinder3D, we one-by-one introduce the mean teacher (MT), as well as our proposed image guidance (IG), contrastive loss (CL) and FOVMix modules. Alongside the mIoU, we also report the relative mIoU (rel) compared to the fully supervised baseline.
Table 4. Ablation study showing the effects of domain adaptation for the image-guidance network. (U) indicated unsupervised and (W) indicates weakly-supervised training.

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>mIoU rel</th>
<th>Δ mIoU</th>
<th>Δ rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>SKITTI (W)</td>
<td>-</td>
<td>60.3 93.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GTA-V</td>
<td>-</td>
<td>60.2 93.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GTA-V SKITTI (U)</td>
<td></td>
<td>61.1 95.0</td>
<td>+0.9</td>
<td>+1.4</td>
</tr>
<tr>
<td>GTA-V SKITTI (W)</td>
<td></td>
<td><strong>61.3 95.3</strong></td>
<td><strong>+1.1</strong></td>
<td><strong>+1.7</strong></td>
</tr>
</tbody>
</table>

Table 5. Ablation study on ScribbleKITTI showing where the accuracy improves with our proposed image-guidance module.

<table>
<thead>
<tr>
<th>Method</th>
<th>Border True</th>
<th>False</th>
<th>Object Small</th>
<th>Large</th>
<th>Distance 0-25m</th>
<th>25m+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cylinder3D</td>
<td>62.5 91.8</td>
<td>73.0 94.0</td>
<td>87.7 84.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MT</td>
<td>62.0 92.7</td>
<td>76.9 95.0</td>
<td>88.4 85.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IGNet</td>
<td><strong>65.5 92.6</strong></td>
<td><strong>83.5 96.5</strong></td>
<td><strong>88.8 87.2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Δ MT</td>
<td>+3.5</td>
<td>-0.1</td>
<td><strong>+6.6</strong></td>
<td><strong>+1.5</strong></td>
<td><strong>+0.4</strong></td>
<td><strong>+2.0</strong></td>
</tr>
</tbody>
</table>

Figure 5. Qualitative results comparing state-of-the-art scribble-supervised LiDAR semantic segmentation methods. As seen, utilizing 2D image features as guidance during the training pipeline, IGNet does improve (top) boundary estimation between classes, (middle) small object segmentation, (bottom) distant, sparse object recognition. We change the color of ‘bicyclist’ in (middle) for better visibility.

gains of +0.4%. Here we conclude that image-guidance can indeed compensate for the common weaknesses seen in LiDAR segmentation, especially under weak supervision.

Apart from quantitative results, we also showcase examples from the valid-set illustrating this effect in Fig. 5. Here we show that IGNet can (top) finely determine object boundaries, (middle) better segment small objects (Cylinder3D and SSLSS misidentify some bicyclist points), and (bottom) improve recognition for sparsely represented regions (IGNet correctly segments all three sparse objects).

5. Conclusion

In this work we tackle common weaknesses of data efficient LiDAR semantic segmentation by distilling high level feature information from a synthetically trained 2D semantic segmentation network. We reduce the domain gap between synthetic and real data by employing weakly supervised DA. We extend the supervision from image pixels to out-of-FOV points via a one way contrastive loss and construct new pairings via FOVMix. With our proposed IGNet, we achieve better boundary estimation, increase performance at distant, sparse regions and heavily improve small class segmentation. We achieve SOTA results in both weakly- and semi-supervised 3D semantic segmentation.

Limitations: Compared to the baseline Cylinder3D, IGNet requires roughly twice the training time due to its two stage approach. Furthermore, the feature distillation module requires paired RGB images with LiDAR scans. While all current LiDAR equipped autonomous systems have an accompanying camera setup, our method still relies on the fact that the sensors need to be calibrated for valid pairings.

Acknowledgements: This work was funded by Toyota Motor Europe via the research project TRACE Zurich.
References


