360° from a Single Camera: A Few-Shot Approach for LiDAR Segmentation

Laurenz Reichardt[1] Nikolas Ebert[1,2] Oliver Wasenmüller[1]
[1]Mannheim University for Applied Science, Germany
[2]RPTU Kaiserslautern-Landau, Germany
l.reichardt@hs-mannheim.de, n.ebert@hs-mannheim.de, o.wasenmueller@hs-mannheim.de

Abstract

Deep learning applications on LiDAR data suffer from a strong domain gap when applied to different sensors or tasks. In order for these methods to obtain similar accuracy on different data in comparison to values reported on public benchmarks, a large scale annotated dataset is necessary. However, in practical applications labeled data is costly and time consuming to obtain. Such factors have triggered various research in label-efficient methods, but a large gap remains to their fully-supervised counterparts. Thus, we propose ImageTo360, an effective and streamlined few-shot approach to label-efficient LiDAR segmentation. Our method utilizes an image teacher network to generate semantic predictions for LiDAR data within a single camera view. The teacher is used to pretrain the LiDAR segmentation student network, prior to optional fine-tuning on 360° data. Our method is implemented in a modular manner on the point level and as such is generalizable to different architectures. We improve over the current state-of-the-art results for label-efficient methods and even surpass some traditional fully-supervised segmentation networks.

1. Introduction

1.1. Label Efficient LiDAR Segmentation

Recent advancements in the application of deep learning methods for LiDAR perception have yielded impressive results on public benchmarks. It is desirable for these methods to perform consistently across different devices and specifications. However, in practice the heterogeneous characteristics of LiDAR sensors (field of view, number of beams, rotational frequency, etc.) result in substantial variations in data. This leads to a decline in performance when deep learning methods are applied to different sensors or tasks. The severity of this sensor domain problem is unique to 3D pointclouds and has prevented the widespread adoption of pretrained feature extraction backbone networks, even between different tasks of the same dataset.

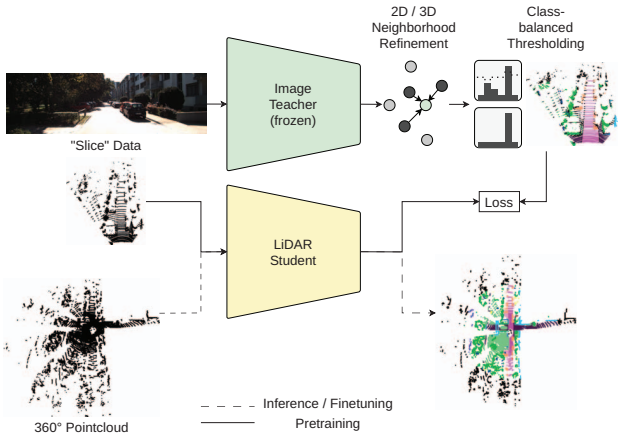

Figure 1. Our few-shot method ImageTo360. A frozen image teacher is used to pretrain the LiDAR student on single camera predictions ("Slice" Data). Later finetuning and inference is performed using the entire 360° of LiDAR data, without image data. During pretraining / finetuning the loss of the respective LiDAR student method is used, following the original implementations.

As a result, the practical application of the methods require large annotated datasets in order for them to perform on par with their public benchmark counterparts [46]. This has led to research in efficient labeling [48, 40] and Label-efficient training as natural fits to counteract the extensive efforts and costs necessary to obtain such datasets. While recent improvements in label-efficient training show promise, this field still requires more research, as most methods under-perform compared to their fully-supervised counterparts.

Building on the readily available camera images in autonomous driving and robotics we propose ImageTo360 for LiDAR semantic segmentation. In this paper we present:

- A streamlined, yet effective and practically viable few-shot learning approach for 360° LiDAR semantic segmentation from a single camera view. Image data is

not required at inference time

- The study of effective methods to improve the quality of 2D information for 3D pretraining
- We show that information from a limited view is sufficient to train a network for 360° data.
- A comprehensive analysis in the field of label-efficient LiDAR semantic segmentation, on the SemanticKITTI [2] benchmark, showing the effectiveness of our approach. We achieve state-of-the-art results with only 1% data, over comparable label-efficient methods.

2. Related Work

The continuous and geometrically complex nature of LiDAR data makes the labeling process especially time-consuming and costly. This is exasperated by drift issues and memory requirements, limiting labeling on reconstructed scenes [2].

Such factors have lead to research on training in a weakly supervised manner. A common approach is the integration of region proposals, e.g. through clustering, during the training process [41, 27]. ScribbleKITTI [47] directly trains with weak scribble annotations, combining various methods such as self-training, multi-scale voxel class distribution information and the mean teacher framework [43]. Besides its susceptibility to various hyperparameters, ScribbleKITTI requires a large computational budget. Furthermore, generalization to different network architectures requires individual adaptions. ScribbleKITTI scales from weakly annotated data, which requires additional human effort.

LaserMix [24] combines compositional mixing of unlabeled and fully labeled data in a semi-supervised approach. However, the angular partitioning of their composition approach is tuned to the hardware of the KITTI dataset, which may impact performance with different LiDAR sensors. LiDAL [21] integrates selective pseudo-labeling, based on the uncertainty between multiple augmented versions of sequential LiDAR scans. Due to the use of multiple scans, their method has a high memory footprint and requires pose information. HybridCR [26] combines weak supervision with consistency loss and learned augmentation.

Genova *et al*. [14] take an image pseudo-labeling approach at urban mapping for temporally assembled LiDAR pointclouds. Their method utilizes a series of filters which remove up to 99% of points, including those of moving entities. While similar in concept, our solution fundamentally differs in that it can function on single-scan data and moving classes such as cars and pedestrians are considered. Their closed-source method was tested on a non-public dataset, making the necessary implementations for comparison infeasible.

2.1. LiDAR Domain Adaption

Domain adaption (DA) has been a natural candidate tackling the LiDAR cross-sensor gap. While DA is fundamentally different to label-efficient methods, the shared goal is to reduce or entirely eliminate the need for target domain data. This field is mainly split into Real→Real and Synthetic→Real domain adaption.

Some compose intermediate representations of source domain data, e.g. through voxel completion [58] or mesh creation [17], prior to sampling target domain data or using the dense data itself for training. Multi-modal methods such as xMUDA [23] and ADAS [13] bridge LiDAR characteristics with image information. CosMix [37] uses a synthetic trained teacher network with compositional mixing similar to LaserMix, while SynLiDARs [52] PCT module learns to reconstruct synthetic pointclouds in the appearance of the target domain.

A reoccurring issue of DA methods is the evaluation on a limited number of overlapping classes between different public datasets [23, 32, 13, 38, 58, 17, 36]. There is no consensus on mapping overlapping labels, limiting the expressiveness of results when compared to label-efficient or fully-supervised methods. Synthetic methods have the benefit that annotations can be generated to match those of the target dataset for evaluation [37, 52].

2.2. Pointcloud Pretraining

Advances in image self-supervised pretraining [18, 8] have kindled research in adapting these concepts to pointclouds. Variants of masked pretraining use reconstruction [59, 31], occlusion completion [49], or occupancy prediction [30], but are specialized towards specific network architectures or limited to data generated from 3D models. Other methods of pointcloud pretraining include contrastive methods [54, 62]. However, the application of these methods has been limited to indoor pointclouds and to impact fully-supervised training, not in the context of label-efficient training.

Pointcloud pretraining also includes multi-modal strategies. Janda *et al*. [22] add image depth estimation features to their contrastive method. Other multi-modal methods such as the SLiDR series [39, 28] and CLIP2Scene [7] utilize knowledge distillation with pretrained image, image/text backbones for 3D semantic segmentation, followed by fine-tuning with a low amount of annotated data. However, the teacher backbones in these methods lack knowledge specific to the segmenting of street accuracy, and perform on-par with LiDAR only methods.

3. Method

We made three main observations based on the current state-of-the-art. Methods with weak annotations [27, 47] or

semi-supervised training [37, 21] intricately combine a variety of techniques dependent on a large number of hyperparameters. Others include image information in the training process, but constrain 3D networks into learning the 2D features of discrete pixel grids [22] instead of in continuous space, or use a general 2D backbone [7, 39, 28]. Thirdly, most pretraining methods are specific to certain architectures or representations (voxels, range-projections, etc.)

These observations inspire us to follow a "back to the basics" approach for ImageTo360. We assume that camera data together with registered LiDAR data is readily available in autonomous driving and robotics. Our method leverages this image information in a 2D supervised teacher network pretraining of a 3D student network. Since the teacher predictions are only available within the camera field of view, we follow up by optional fine-tuning on 360° data. Our proposed few-shot method is visualized in Fig. 1.

Pretrained image segmentation networks are readily available [10] and can be used as "off-the-shelf" teacher networks. From a practical perspective image segmentation is a logical choice to introduce specific knowledge into label efficient LiDAR segmentation. In order to not bind the 3D student network to 2D features, our method uses high quality pseudo-labels as a general representation of segmentation knowledge.

We implement our method on the point level, as LiDAR data is first represented as pointclouds prior to transformation into other representations. As a result our method generalizes across sensors and architectures.

3.1. 2D Supervision

We leverage the Cityscapes [11] dataset due to street scenes similar to those of SemanticKITTI [2]. In order to fully evaluate our method against others, it is necessary to include all SemanticKITTI classes in our analysis, rather than just reporting class-overlap scores. For this reason we fine-tune the network on a subset of the KITTI dataset, however this step is not necessary for the actual implementation of our method.

The 2D predictions are limited to the single camera field of view. When projected into 3D space, using the calibration parameters between camera and LiDAR, these predictions cover a "slice" of the 360° pointcloud (see Fig. 1). This projection inherently has errors, caused by discrepancies in the calibration and synchronization between camera and LiDAR sensors. On top of that, neural network predictions tend to be unfocused at object borders. These discontinuous predictions cause the "Flying Pixels" bleeding effect [44, 34], exemplified in Fig. 2.

While the corresponding pixels of these prediction errors are direct neighbors in 2D grids, this does not necessarily hold true when projected into 3D space. We take inspiration from post-processing methods in range projected LiDAR segmentation [12][29] and propose 3D neighborhood refinement as a method to reintroduce continuous space information into 2D predictions. For our method, we use K-Nearest-Neighbors (KNN), but other neighborhood algorithms are also suitable.

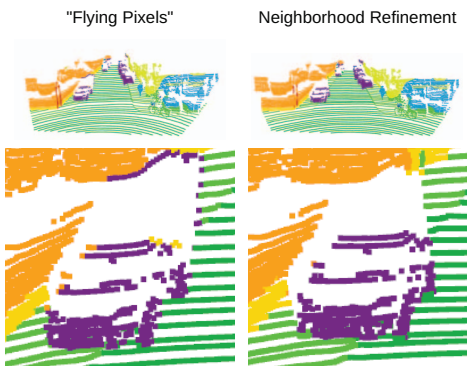

Figure 2. Image segmentation predictions projected into 3D space. Uncertainty at object edges cause "Flying Pixels" [44]; in this example (left image) casting car labels unto the building and vice-versa. Neighborhood refinement counteracts this issue.

First we consider neighborhood majority voting. However KN-Neighborhoods have varying distances and volumes. Simple majority voting would equally weight points regardless of distance. For this reason, we also consider distance-weighted voting ($1 - \text{softmax}(\vec{v}_d)$ with $\vec{v}_d$ as the distance vector). Priority is put on labels closer to the query point. Thirdly, we compare both methods to neighborhood confidence averaging.

3.2. Pretraining from 2D predictions

Pseudo-label confidence thresholding [25] has been an established method in order to remove noisy labels due to incorrect predictions. In a multiclass context, networks tend to bias towards majority classes [5][64], and simple static thresholding can remove minority classes. Class-imbalance inherently affects LiDAR street scenario data (e.g. buildings have more points than a person). We follow established methods [47][38] and integrate class-balanced adaptive thresholding with the goal of further improving pseudo-label quality.

Formally, the class balanced thresholds $\vec{\tau}_c(i)$ of our method can be expressed as

$$\vec{\tau}_c(i) = \frac{\sum[labels = i]}{max\left\{\forall_{i \in c} \sum[labels = i]\right\}} \times (\tau_{max} - \tau_{min}) + \tau_{min} \quad (1)$$

with all dataset classes c, single class i, τ_{min} as the minimum threshold, and τ_{max} as the maximum threshold values. In short, threshold values are scaled within a confidence range (set through τ_{min} and τ_{max}) for each class,

depending on how many times this class occurs in comparison to the majority class. Predictions below the adaptive threshold are not used as pseudo-labels.

For pretraining, the pointclouds are cut into the fields of view of the camera image, significantly reducing training time and memory requirements. Our method can be scaled to include multiple cameras. Optionally, fine-tuning is performed using the annotated 360° LiDAR data, without image data.

4. Evaluation

We evaluate our method ImageTo360 against related works for the spaces of fully-supervised, label-efficient and pretraining methods, as well as against domain adaption. The intention is to show the differences in performance when considering practical application and labeling efforts.

Our ablation studies and evaluation are performed using the SemanticKITTI [2] dataset in a single-scan context. Results are reported on evaluation sequences 08, either on the full 360° or the "slice" of data within the left camera field of view (covering 90°). For image teacher networks we use the OpenMMLab [10] implementation of Segformer [53] and DeepLabV3+ [6] in combination with the left image data of the dataset. For our 3D student networks we use SPVCNN [42] and Cylinder3D [63].

For 3D student fine-tuning, we randomly sample a small percentage of the training split and use the entire 360° pointclouds and corresponding ground-truth annotations. Fine-tuning is performed according to the original methods hyperparameters, except with $0.2 \times learning\ rate$, $0.5 \times epochs$. Additionally we use translation, pointcloud mixing (similar to CutMix [60] and LaserMix [24]), squeezing, and flipping as augmentation for both networks. Using this augmentation schedule, we reproduced the paper results of the original authors.

We do not include test data or validation data in any training steps of our method. Unless otherwise stated we do not use Test Time Augmentation (TTA) or ensembling, in order to provide a fair comparison with research which forgoes such methods. In the case where we use these methods, we utilize Greedy Soup [50] ensembling, and TTA over 12 variants of data (original, three flips, eight rotations).

4.1. Ablation Study

4.1.1 Impact of Image Segmentation Network

Firstly, we study the impact of an image teacher by pretraining the 3D student network on the predictions within the left camera field of view (∼90°). We follow by fine-tuning on a low-label budget using 360° annotated pointclouds. The baseline comparisons train on the same fine-tuning data, without pretraining.

Improvements in image segmentation theoretically help in pretraining the 3D student. To test this assumption we compare Segformer as a modern transformer architecture with the traditional convolutional DeepLabV3+, as teacher networks. Segformer achieves 58.24 mIOU% compared to the 50.88 mIOU% of DeepLab, an improvement of +14.47% on the left camera field of view predictions.

Results are depicted in Table 2. Our pretraining method significantly improves over the baseline for nearly all reduced label scenarios, however the magnitude of improvement reduces with an increasing ratio of labels. The improvements of Segformer over DeepLabV3+ translate to the low label amount of the SPVCNN architecture. With Cylinder3D as the student, results are comparable across both teacher networks. These results reveal noteworthy observations. Without fine-tuning, networks that have only ever seen slices of LiDAR data produce considerable results when predicting on 360° data. While developments in image segmentation can have an impact, results are heavily influenced by the choice of 3D student architecture. Finally, image pretraining improved fully-supervised training with both image teachers for SPVCNN, showing potential for future research.

We continue with Segformer as our teacher network for further ablation and evaluation.

4.1.2 3D Neighborhood Refinement of 2D Predictions

We study the effects of utilizing neighborhood refinement as a technique for reintroducing 3D information into 2D predictions of Segformer. We examine neighborhood majority label voting, distance weighted label voting, and neighborhood confidence averaging in Table 1. All three approaches demonstrate comparable improvements over the baseline. The use of confidence averaging offers the added benefit of avoiding one-hot labels and as such being compatible with pseudo-label thresholding. For this reason we continue with confidence averaging ($K = 19$) for further ablation and evaluation, improving pseudo-label quality by +4.86%.

4.1.3 Class Balanced Pseudo-Label Thresholds

We compare static and class-balanced adaptive thresholding (see Equation 1) after neighborhood confidence averaging. The results are depicted in Table 3. We use $\tau_{max} = 0.95$ as an upper limit to all adaptive threshold ranges. Threshold parameters are chosen to have a comparable percentage of points removed. Class-balanced adaptive thresholding shows a significant improvement over static thresholds, in total improving pseudo-label quality by +20.15% over the neighborhood refined baseline (61.07 mIoU%), while setting 23% of unreliable labels as unlabeled. More information is kept compared to the best static threshold and as a

Table 1. Evaluation of 3D neighborhood refinement of 2D predictions, using the KITTI validation mIOU in the left camera field of view.

Method	**Neighbors**										
	3	*5*	*7*	*9*	*11*	*13*	*15*	*17*	*19*	*21*	*23*
None (Segformer-baseline [53])						58.2					
KNN Majority Voting	59.21	59.79	60.28	60.63	60.89	61.08	61.18	61.23	61.24	61.24	61.21
KNN Distance Normalized Voting	59.21	59.79	60.28	60.64	60.90	61.09	61.21	61.27	**61.30**	61.29	61.26
KNN Confidence Averaging	59.20	59.78	60.23	60.55	60.78	60.94	61.02	61.07	61.07	61.06	61.02

Table 2. Comparison of Segformer [53] and DeepLabV3+ [6] mIOU as teacher networks. Evaluation is performed on the entire 360° annotated pointcloud of the KITTI validation split.

Student / Teacher Networks	**SemanticKITTI Labels**							
	0 %	*0.3%*	*1%*	*5%*	*10%*	*20%*	*50%*	*100%*
Cylinder3D [63] (baseline)	-	27.24	34.38	48.90	54.91	59.57	65.44	66.16
Cylinder3D / DeepLabV3+ [6]	46.62	49.76	54.50	58.29	58.62	61.44	64.22	64.88
Cylinder3D / Segformer [53]	50.80	50.74	54.10	59.26	60.00	62.27	64.99	65.82
SPVCNN [42] (baseline)	-	27.25	32.42	49.18	59.10	63.38	65.11	65.81
SPVCNN / DeepLabV3+	45.87	48.53	56.08	**61.96**	**63.07**	**64.38**	65.83	65.94
SPVCNN / Segformer	52.15	**57.30**	**59.53**	61.71	62.41	64.19	**66.10**	**66.44**

Table 3. Ablation of pseudo-label thresholding, applied after neighborhood confidence averaging, using the KITTI validation mIOU in the left camera field of view.

Static threshold				
Threshold τ	0.80	0.85	0.90	0.95
Point reduction %	16.83	20.04	23.63	28.58
mIOU %	70.95	71.69	72.19	72.82
Class-balanced threshold (τ_{max} *= 0.95)*				
Min. threshold τ_{min}	0.5	0.6	0.7	0.8
Point reduction %	15.83	18.31	20.78	23.41
mIOU %	68.75	70.95	72.45	**73.38**

result we continue with adaptive thresholding ($\tau_{min} = 0.8$) for further evaluation.

4.2. Comparison with Label Efficient Methods

We compare our method ImageTo360 with research from the spaces of weakly-supervised, few-shot, and semi-supervised training, as well as self-supervised pretraining. All compared methods in this evaluation rubric share the goal of label-efficient training, fine-tuning on a subset of annotated KITTI data. This rubric excludes efficient annotation methods, which focus on reducing the labeling effort, instead of the data needed. It is common in this field of research to evaluate on the validation set, without ranking on the test benchmark. Also, there is no agreed-upon percentage of labels for evaluation, so we've reported our results (see Table 4) using a variety of percentages.

As the most common reported metric, 1% labels can be considered the most indicative of performance. Comparing amongst the same 3D backbone, ImageTo360 improves +17.58% over LaserMix. We apply TTA and ensembling to our best 1% network to show the full potential of our method, improving over next-best HybridCR by +20.26%. We even surpass ScribbleKITTI with its 8% labeled points. Taken to the extremes by fine-tuning on just 57 annotated samples (0.3%), we outperform all comparable methods using 1% of data, and achieve similar results to those using 5% of data.

Current image backbone pretraining methods usually transfer 2D features to 3D networks. Considering the gap in performance to label-efficient LiDAR-only methods, this multi-modal knowledge is not necessarily beneficial. Our method differs by introducing the specific knowledge of an off-the-shelf segmentation backbone, outperforming all other multi-modal pretraining methods, even without any fine-tuning. From a application standpoint, segmentation backbones are readily available and lead to a considerable performance increase.

4.3. Comparison with domain adaption methods

From a data-perspective, evaluating our ImageTo360 to LiDAR domain adaption (DA) methods is biased. Even with teacher networks trained on a source domain, cross-sensor DA without target domain annotations remains a challenge. Regardless of our method using a modest amount of data, we believe that this comparison is worthwhile when considering practical applications, even more so for safety critical applications where accuracy is of great importance.

Most DA methods evaluate on the class-overlap between public benchmarks. Currently, there is no consensus amongst authors on the class-mapping. This makes the comparison of DA methods to each other challenging. However, the performance margins of each DA method to our method as a reference can give indication on what strategy performs best. We reevaluate our best 1% method, according to the class-mapping of each compared method. The results are depicted in Table 5. The margins of improvement point to the benefits of image information in DA. On AudiA2D2 [15] →KITTI our method performs +41.10% better. The margin of improvement is much greater when considering the best method without image data (SynLiDAR →KITTI with +95.34%). Both results put into view the trade-off between the efforts of annotating a limited amount of data or choosing domain-adaption approaches.

4.4. Comparison with fully-supervised methods

For our comparison with fully-supervised semantic segmentation methods we use our best 1% network (Segformer / SPVCNN + TTA, ensembling). We follow standard prac-

Table 4. Evaluation mIOU on the 360° data of the KITTI validation split. Percentages of labels signify the amount of training split data that was used during fine-tuning. The training split data within the camera field of view was used for image teacher pretraining. We compare to the values reported by the authors. Empty "-" values signify that the original papers did not evaluate on that percentage of annotations.

Method	Backbone Networks	Use of images	SemanticKITTI Labels							
			0%	*0.3%*	*1%*	*5%*	*8%*	*10%*	*20%*	*50%*
Label Efficient										
LiDAL [21]	SPVCNN [42]	no	-	-	48.8	59.5	-	-	-	-
ReDAL [51]	SPVCNN	no	-	-	41.8	59.8	-	-	-	-
HybridCR [26]	PSD [61]	no	-	-	52.3	-	-	-	-	-
LaserMix [24]	Cylinder3D [63]	no	-	-	50.6	-	-	60.0	61.9	62.3
ImageTo360 (ours)	Segformer [53] / SPVCNN	yes	52.1	57.3	59.5	**61.7**	**62.1**	**62.4**	**64.2**	**66.1**
	+ TTA, Greedy Soup	yes	-	-	**62.9**	-	-	-	-	-
ImageTo360 (ours)	Segformer / Cylinder3D	yes	50.8	50.7	54.1	59.2	59.6	60.0	62.2	65.0
Weakly Annotated										
ScribbleKITTI [47]	SPVCNN	no	-	-	-	-	61.3	-	-	-
ScribbleKITTI	Cylinder3D	no	-	-	-	-	60.8	-	-	-
Self-supervised Pretraining										
CLIP2Scene [7]	CLIP [33] / SPVCNN	yes	-	-	42.6	-	-	-	-	-
SLidR [39]	Minkowski [9]	yes	-	-	44.6	-	-	-	-	-
ST-SLidR [28]	SwAV [4] / Minkowski	yes	-	-	44.9	-	-	-	-	-
Contrastive Image [22]	ResNet18 [19] / Minkoswki	yes	42.0 (Percentage not stated)							

Table 5. Comparison of our best 1% method to domain adaption. We reevaluate using the label mappings provided by the authors. Evaluation on the KITTI validation sequence.

Method	Backbones	Use of images	mIOU %
A2D2 [15] to KITTI [2] (10 classes)			
xMUDA [23]	SparseVoxel [16] + ResNet34 [19]	yes	49.1
DsCML [32]	SparseVoxel + ResNet34	yes	52.4
ADAS [13] + DsCML	SparseVoxel + ResNet34	yes	54.3
ImageTo360 1% (ours)	SPVCNN [42] / Segformer [53]	yes	**76.62**
Synth4D [38] (synthetic) to KITTI (7 classes)			
GIPSO [38]	MinkUNet [9]	no	40.24
ImageTo360 1% (ours)	SPVCNN + Segformer	yes	**83.25**
NuScenes [3] to KITTI (10 classes)			
Complete & Label [58]	SparseVoxel	no	33.7
ImageTo360 1% (ours)	SPVCNN + Segformer		**79.03**
NuScenes to KITTI (11 classes)			
Fake it, Mix it [17]	Cylinder3D [63]	no	34.3
ImageTo360 1% (ours)	SPVCNN + Segformer	yes	**77.62**
NuScenes to KITTI (11 classes, different mapping)			
GatedAdapters [36]	SalsaNext [12]	no	23.5
ImageTo360 1% (ours)	SPVCNN + Segformer	yes	**72.23**
SynLiDAR [52] (synthetic) - KITTI (all classes)			
SynLiDAR	PCT [52] / Minkowski	no	27.0
CoSMix [37]	Minkowski	no	32.2
ImageTo360 1% (ours)	SPVCNN + Segformer	yes	**62.9**

Table 6. Evaluation of our method compared to fully-supervised methods on the KITTI single-scan semantic segmentation benchmark.

Method	Image training	mIOU %
2DPASS [57]	yes	72.9
Point-Voxel KD [20]		71.2
Cylinder3D [63]		67.8
SPVNAS [42]		66.4
JS3C-Net [56]		66.0
SalsaNext [12]		59.5
KPConv [45]		58.8
ImageTo360 1% (ours)	yes	57.7
SCSSnet [35]		57.6
SqueezeSegV3 [55]		55.9
3D-MiniNet [1]		55.8
RangeNet53++ [29]		52.2

tice and compare to methods using 100% human annotated data on the KITTI test benchmark. Our method is at a natural disadvantage, even more so considering we do not fine-tune with any validation data or use semi-supervised training on the test split.

The results on the KITTI benchmark are depicted in Table 6 and can be interpreted two-fold. While ImageTo360 outperforms comparable methods, few-shot training is still an open field of research. A gap remains to current state-of-the-art fully supervised networks, with 2DPASS [57] performing +26.34% better than our method. But with considerable less human effort and annotation costs, our few-shot method outperforms traditional fully-supervised methods such as SqueezeSegV3 [55]. We encourage future label-efficient research to also upload their results on the public benchmark.

5. Conclusion

In this paper, we present ImageTo360, a streamlined approach to few-shot LiDAR semantic segmentation using an image teacher network for pretraining. Our method is designed in a modular manner and at the point level, meaning it can be applied across different 3D network architectures and that components can be exchanged with further developments in deep learning. We evaluated against other label-efficient methods, producing state-of-the-art results. With practical applications in mind, we expanded our evaluation against domain adaption and fully-supervised methods, showing the large performance gaps between the different

fields. The results show that image data of a comparatively low cost camera is sufficient for 3D networks to generate remarkable results on 360° LiDAR data, even outperforming traditional fully-supervised methods.

Acknowledgement

This work was funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) under the grant AuReSi (KK5335501JR1).

References

[1] Iñigo Alonso, Luis Riazuelo, Luis Montesano, and Ana C Murillo. 3d-mininet: Learning a 2d representation from point clouds for fast and efficient 3d lidar semantic segmentation. In *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020. 6

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 4, 6

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. Nuscenes: A multimodal dataset for autonomous driving. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020. 6

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems (NeurIPS)*, 2020. 6

[5] Baixu Chen, Junguang Jiang, Ximei Wang, Pengfei Wan, Jianmin Wang, and Mingsheng Long. Debiased self-training for semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, 2018. 4, 5

[7] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. *arXiv preprint arXiv:2301.04926*, 2023. 2, 3, 6

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[9] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[10] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 3, 4

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[12] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In *International Symposium on Visual Computing (ISVC)*. Springer, 2020. 3, 6

[13] Ben Fei, Siyuan Huang, Jiakang Yuan, Botian Shi, Bo Zhang, Tao Chen, Min Dou, and Yu Qiao. Adas: A simple active-and-adaptive baseline for cross-domain 3d semantic segmentation. *arXiv preprint arXiv:2212.10390*, 2022. 2, 6

[14] Kyle Genova, Xiaoqi Yin, Abhijit Kundu, Caroline Pantofaru, Forrester Cole, Avneesh Sud, Brian Brewington, Brian Shucker, and Thomas Funkhouser. Learning 3d semantic segmentation with only 2d image supervision. In *International Conference on 3D Vision (3DV)*. IEEE, 2021. 2

[15] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020. 5, 6

[16] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[17] Frederik Hasecke, Pascal Colling, and Anton Kummert. Fake it, mix it, segment it: Bridging the domain gap between lidar sensors. *arXiv preprint arXiv:2212.09517*, 2022. 2, 6

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[20] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6

[21] Zeyu Hu, Xuyang Bai, Runze Zhang, Xin Wang, Guangyuan Sun, Hongbo Fu, and Chiew-Lan Tai. Lidal: Inter-frame uncertainty based active learning for 3d lidar semantic segmentation. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2, 3, 6

[22] Andrej Janda, Brandon Wagstaff, Edwin G Ng, and Jonathan Kelly. Self-supervised pre-training of 3d point cloud networks with image data. *arXiv preprint arXiv:2211.11801*, 2022. 2, 3, 6

[23] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 6

[24] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised lidar semantic segmentation. *arXiv preprint arXiv:2207.00026*, 2022. 2, 4, 6

[25] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *International Conference on Machine Learning (ICML)*, 2013. 3

[26] Mengtian Li, Yuan Xie, Yunhang Shen, Bo Ke, Ruizhi Qiao, Bo Ren, Shaohui Lin, and Lizhuang Ma. Hybridcr: Weakly-supervised 3d point cloud semantic segmentation via hybrid contrastive regularization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6

[27] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[28] Anas Mahmoud, Jordan SK Hu, Tianshu Kuai, Ali Harakeh, Liam Paull, and Steven L Waslander. Self-supervised image-to-point distillation via semantically tolerant contrastive loss. *arXiv preprint arXiv:2301.05709*, 2023. 2, 3, 6

[29] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 4213–4220. IEEE, 2019. 3, 6

[30] Chen Min, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Voxel-mae: Masked autoencoders for pre-training large-scale point clouds. *arXiv preprint arXiv:2206.09900*, 2022. 2

[31] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. *arXiv preprint arXiv:2203.06604*, 2022. 2

[32] Duo Peng, Yinjie Lei, Wen Li, Pingping Zhang, and Yulan Guo. Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 6

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*. PMLR, 2021. 6

[34] Laurenz Reichardt, Patrick Mangat, and Oliver Wasenmüller. Dvmn: Dense validity mask network for depth completion. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2653–2659. IEEE, 2021. 3

[35] Christoph B Rist, David Schmidt, Markus Enzweiler, and Dariu M Gavrila. Scssnet: Learning spatially-conditioned scene segmentation on lidar point clouds. In *Intelligent Vehicles Symposium (IV)*. IEEE, 2020. 6

[36] Mrigank Rochan, Shubhra Aich, Eduardo R Corral-Soto, Amir Nabatchian, and Bingbing Liu. Unsupervised domain adaptation in lidar semantic segmentation with self-supervision and gated adapters. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2022. 2, 6

[37] Cristiano Saltori, Fabio Galasso, Giuseppe Fiameni, Nicu Sebe, Elisa Ricci, and Fabio Poiesi. Cosmix: Compositional semantic mix for domain adaptation in 3d lidar segmentation. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2, 3, 6

[38] Cristiano Saltori, Evgeny Krivosheev, Stéphane Lathuiliére, Nicu Sebe, Fabio Galasso, Giuseppe Fiameni, Elisa Ricci, and Fabio Poiesi. Gipso: Geometrically informed propagation for online adaptation in 3d lidar segmentation. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2, 3, 6

[39] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 6

[40] Dennis Stumpf, Stephan Krauß, Gerd Reis, Oliver Wasenmüller, and Didier Stricker. Salt: A semi-automatic labeling tool for rgb-d video sequences. *arXiv preprint arXiv:2102.10820*, 2021. 1

[41] Tianfang Sun, Zhizhong Zhang, Xin Tan, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Image understands point cloud: Weakly supervised 3d semantic segmentation via association learning. *arXiv preprint arXiv:2209.07774*, 2022. 2

[42] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 4, 5, 6

[43] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems (NeurIPS)*, 2017. 2

[44] Dennis Teutscher, Patrick Mangat, and Oliver Wasenmüller. Pdc: piecewise depth completion utilizing superpixels. In *IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021. 3

[45] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *International Conference on Computer Vision (ICCV)*, 2019. 6

[46] Larissa T Triess, Mariella Dreissig, Christoph B Rist, and J Marius Zöllner. A survey on deep domain adaptation for lidar perception. In *IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*. IEEE, 2021. 1

[47] Ozan Unal, Dengxin Dai, and Luc Van Gool. Scribble-supervised lidar semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 6

[48] Bernie Wang, Virginia Wu, Bichen Wu, and Kurt Keutzer. Latte: accelerating lidar point cloud annotation via sensor fu-

sion, one-click annotation, and tracking. In *IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019. 1

[49] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. Unsupervised point cloud pre-training via occlusion completion. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[50] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*. PMLR, 2022. 4

[51] Tsung-Han Wu, Yueh-Cheng Liu, Yu-Kai Huang, Hsin-Ying Lee, Hung-Ting Su, Ping-Chia Huang, and Winston H Hsu. Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation. In *International Conference on Computer Vision (ICCV)*, 2021. 6

[52] Aoran Xiao, Jiaxing Huang, Dayan Guan, Fangneng Zhan, and Shijian Lu. Transfer learning from synthetic to real lidar point cloud for semantic segmentation. In *AAAI Conference on Artificial Intelligence*, 2022. 2, 6

[53] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 4, 5, 6

[54] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 2

[55] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 6

[56] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *AAAI Conference on Artificial Intelligence*, 2021. 6

[57] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 6

[58] Li Yi, Boqing Gong, and Thomas Funkhouser. Complete & label: A domain adaptation approach to semantic segmentation of lidar point clouds. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 6

[59] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[60] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019. 4

[61] Yachao Zhang, Yanyun Qu, Yuan Xie, Zonghao Li, Shanshan Zheng, and Cuihua Li. Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation. In *International Conference on Computer Vision (ICCV)*, 2021. 6

[62] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[63] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. *arXiv preprint arXiv:2011.10033*, 2020. 4, 5, 6

[64] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *European Conference on Computer Vision (ECCV)*, pages 289–305, 2018. 3