# ProtoTransfer: Cross-Modal Prototype Transfer for Point Cloud Segmentation

Pin Tang[1] Hai-Ming Xu[2] Chao Ma[1*]
[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2] Australian Institute for Machine Learning, University of Adelaide
{pin.tang,chaoma}@sjtu.edu.cn, hai-ming.xu@adelaide.edu.au

## Abstract

*Knowledge transfer from multi-modal, i.e., LiDAR points and images, to a single LiDAR modal can take advantage of complimentary information from modal-fusion but keep a single modal inference speed, showing a promising direction for point cloud semantic segmentation in autonomous driving. Recent advances in point cloud segmentation distill knowledge from strictly aligned point-pixel fusion features while leaving a large number of unmatched image pixels unexplored and unmatched LiDAR points under-benefited. In this paper, we propose a novel approach, named ProtoTransfer, which not only fully exploits image representations but also transfers the learned multi-modal knowledge to all point cloud features. Specifically, based on the basic multi-modal learning framework, we build up a class-wise prototype bank from the strictly-aligned fusion features and encourage all the point cloud features to learn from the prototypes during model training. Moreover, to exploit the massive unmatched point and pixel features, we use a pseudo-labeling scheme and further accumulate these features into the class-wise prototype bank with a carefully designed fusion strategy. Without bells and whistles, our approach demonstrates superior performance over the published state-of-the-arts on two large-scale benchmarks, i.e., nuScenes and SemanticKITTI, and ranks 2nd on the competitive nuScenes Lidarseg challenge leaderboard.*

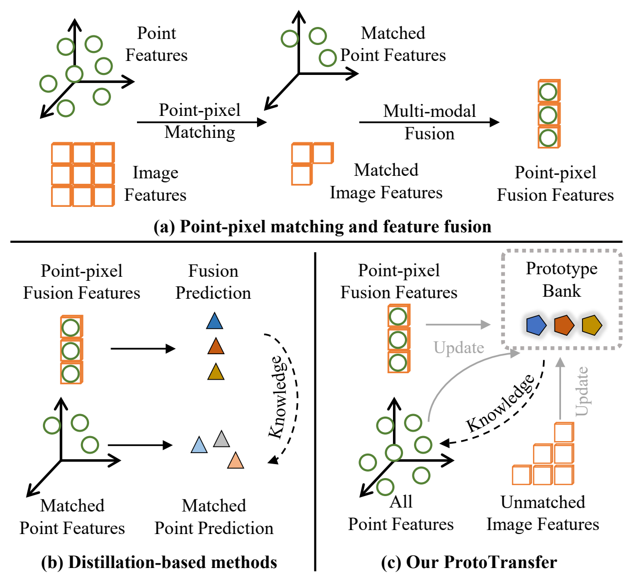



Figure 1. **(a) Point-pixel matching and feature fusion:** only features of strictly matched point-pixel pairs can be fused. **(b) Distillation-based methods** are only performed between fusion prediction and the corresponding matched point prediction while other unmatched points have no chance to mimic the fusion prediction. **(c) Our ProtoTransfer** performs knowledge transfer from prototype bank to all point features directly without being restricted by matching requirements. The prototype bank is updated using fusion, point, and unmatched image features.

## 1. Introduction

Semantic segmentation on point cloud [6, 15, 37, 40] has attracted increasing attention from the computer vision community for its crucial role in scene understanding of 3D space. Although LiDAR point cloud can provide accurate location and depth information of interested scenes, the sparse and textureless shortages inevitably restrict its semantic segmentation performance. On the other hand, 2D images consist of dense pixels with rich color and subtle textures [34]. Therefore, incorporating these two complementary modals altogether can be a more plausible solution.

Recently, there are mainly two stream approaches of utilizing multi-modal data, i.e., fusion-based methods [10, 50] and distillation-based methods [42]. The former methods project point clouds to the camera coordinate to obtain a point-to-pixel mapping, based on which point features are fused with corresponding image features to produce the final point-wise segmentation during both the training and evaluation phases. Although robust and accurate segments are achieved by fusing different sensors, the fusion-based methods may suffer from heavy memory and time consumption for processing these two modal data simul-

[*] Corresponding author.

taneously. In order to benefit from multi-modal fusion while bypassing the computation burden, the latter methods [42] propose to utilize the distillation technique to transfer knowledge learned from cross-modal fusion at the training stage and discard the fusion module for evaluation. The impressive segmentation performance suggests the feasibility of the methods in this stream. Therefore, this paper further explores this research direction.

Unlike common distillation approaches in image classification tasks [14, 44], distilling from multi-modal to the single LiDAR modal may meet quite different pitfalls: (1) Not all LiDAR points can be used in knowledge distillation because LiDAR-to-image is partially matched. Due to the difference in the way sensors collect data, not all the LiDAR points can be aligned to the dense image pixels. For example, we empirically find that only around 16.07% LiDAR points in SemanticKITTI dataset can be matched to corresponding image pixels, and Fig. 1(a) also demonstrates this observation. Thus, classic logit-/probability-style distillation strategies [42] can only be performed on the well-aligned point clouds, leading to a sub-optimal knowledge distillation. (2) Massive image pixels are overlooked. Since only partial image pixels[1] will contribute to the modal fusion based on the above matching constraints, a large number of pixels are not used in knowledge distillation, resulting in a great loss of rich semantic information in dense pixels.

In this paper, we propose a novel approach for point cloud semantic segmentation which successfully avoids the above pitfalls. Specifically, following the existing work [17, 34, 42], we first construct a modal fusion module to combine the feature representations of matched LiDAR points and image pixels. Next, in order to transfer the knowledge from the fusion module to each point feature representation, we create a class-wise prototype bank to accumulate the fusion features learned in the fusion module and encourage the similarity between features of all LiDAR points and corresponding-class prototypes as high as possible. By this means, every LiDAR point has a fair chance to mimic the fusion features as shown in Fig. 1(c), and we dub our method as **ProtoTransfer**. Moreover, to make full use of semantic information inhered in dense image pixels, we propose to incorporate the image features into the class-wise prototype bank through a cleverly designed fusion strategy. Since images in the point cloud semantic segmentation datasets usually lack per-pixel annotations, we further use a pseudo-labeling scheme to generate pseudo-labels for each pixel and thus make the update of pixel-feature into class-wise prototypes become possible. In summary, the main contributions of this paper are three aspects:

- We investigate the pitfalls of point cloud semantic segmentation on distillation-based methods and find that a large number of image features are not well-utilized and point features are not well distillation-benefited.
- We introduce the prototype bank concept into point cloud semantic segmentation and propose a novel approach ProtoTransfer to successfully overcome the above pitfalls.
- We conduct experiments on both SemanticKITTI and nuScenes benchmarks to demonstrate the effectiveness of our approach and also achieve a 2nd place on the competitive nuScenes Lidarseg leaderboard.

[1]Only 5% image pixels are matched to LiDAR points for a typical 32-beam LiDAR scanner as presented in BEVFusion [22].

## 2. Related Work

**Multi-Modal 3D Semantic Segmentation.** Since different modal can provide complimentary information to each other, multi-modal point cloud semantic segmentation attracts increasing attention [8, 20, 24]. RGBAL [8] casts RGB images to a polar-grid mapping representation and designs an early-mid-level hybrid fusion architecture. Recently, PMF [50] projects LiDAR points to camera coordinates, which is called perspective projection [25]. Then, they use two 2D U-net [27] to extract the image and point features. The multi-scale image and point features in both U-nets are fused to produce better segmentation results. Though satisfactory performance are achieved, these methods need multi-modal inputs during both training and inference phases, which is time-/memory-consuming.

**Prototype Networks.** Prototype-based learning methods has been widely used in machine learning [9, 11, 28]. Recently, a surge of attention is paied to employ prototype networks on various tasks, presenting great potential in few-shot learning [29] and zero-shot learning[18, 36]. Moreover, [48] shows a prototype view of image semantic segmentation network. Another prototype-based semi-supervised method is also proposed [39]. Our work sheds light on the possibility of using prototypes for knowledge transfer from multi-modal to single-modal.

**Cross-Modal Knowledge Transfer.** Knowledge distillation, first proposed by Hinton *et al.* [14], is a common knowledge transfer method, which pushes the student network mimic the soft logits of the teacher network. Very recently, knowledge distillation is introduced into perception tasks in autonomous driving, such as 3D object detection [7, 19, 46] and point cloud segmentation [42]. During training, they use distillation to transfer multi-modal knowledge learned by the multi-modal teacher to a single-modal student. However, these methods suffers from strict point-pixel alignment, leading to massive unmatched image pixels unexplored and points under-benefited. Our work performs knowledge transfer in another way. Contrary to [42], we construct a prototype bank from fusion and unmatched image features and encourage all the point cloud features to learn from the prototypes during model training, thus fully exploiting and transferring multi-modal knowledge.

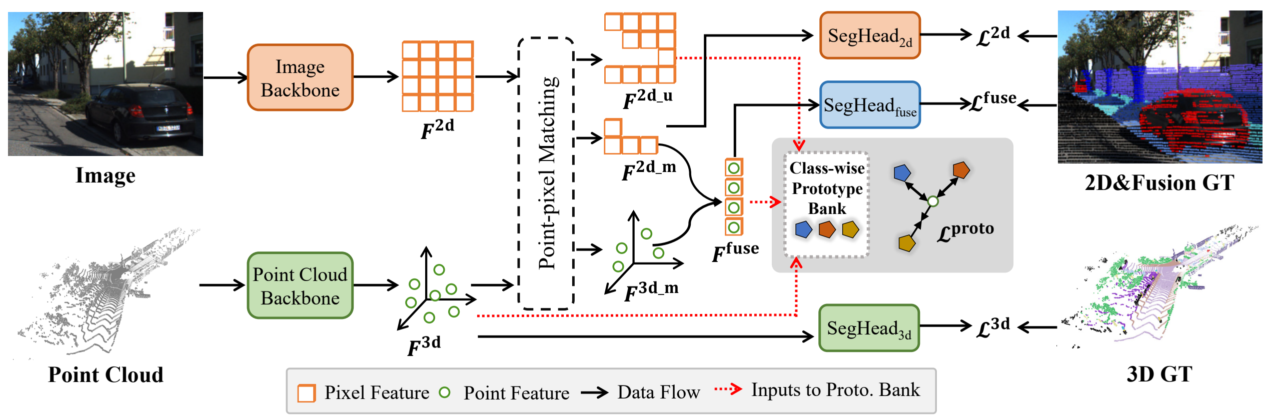



Figure 2. Pipeline overview of our ProtoTransfer. The model first uses modal-specific backbones to generate image features $F^{2d}$ and point features $F^{3d}$ and feeds them into the point-pixel matching module to generate matched point feature $F^{3d\_m}$, matched and unmatched image features $F^{2d\_m}$ and $F^{2d\_u}$, respectively. Then modal-specific segmentation heads are used to produce predictions for loss calculation, i.e., $\mathcal{L}^{3d}, \mathcal{L}^{2d}$ and $\mathcal{L}^{fuse}$, where $\mathcal{L}^{fuse}$ will update the feature $F^{fuse}$ fused from $F^{2d\_m}$ and $F^{3d\_m}$. To take advantage of the fusion features, we introduce a class-wise prototype bank whose input source includes features $F^{3d}$, $F^{fuse}$ and $F^{2d\_u}$ and propose a loss $\mathcal{L}^{proto}$ to transfer fusion knowledge to all point features. Please note 2D&Fusion GT is derived from the 3D GT based on point-pixel matching.

# 3. Methodology

## 3.1. Framework Overview

Given a LiDAR point cloud frame with $N$ unordered points $P = \{p_i \in \mathbb{R}^{d_{\text{in}}}\}_{i=1}^{N}$ where $d_{\text{in}}$ is input feature dimension[2], the goal of point cloud semantic segmentation is to assign a single class label $c \in \{1, 2..., C\}$ to each point. To make up for the sparsity and lack of texture of point clouds collected in outdoor scenes [42], the other modal of the corresponding scenes, raw images $I \in \mathbb{R}^{W \times H \times 3}$, where $W, H$ denote the resolution of a given image, are also provided for model learning to achieve a better segmentation performance.

As the overall structure presented in Fig. 2, the main contributions of our approach are (1) introducing a class-wise prototype bank for knowledge transfer from a fusion modal to all LiDAR points and (2) proposing to make full use of image features to enhance the prototype quality. Specifically, paired point cloud and image are first fed into modal-specific backbones to extract feature representations respectively. Next, based on the given LiDAR-to-image transformation matrix, features are fused for matched LiDAR point and corresponding image pixels, as done in previous modal-fusion-related work [17, 34, 42]. Due to the differing data acquisition mechanisms of different modal sensors, the matching rate between LiDAR points and image pixels is often quite low. Instead of distilling knowledge solely for the limited matched LiDAR points, we propose a novel knowledge transfer module that can enable unmatched points to also benefit from fused feature representations. With this framework, we explore and exploit unmatched image pixels to further enhance the overall point cloud segmentation performance.

During inference, the enhanced point cloud segmentation branch can produce accurate segmentation results without the image backbone and multi-modal fusion, and thus challenges can be tackled within a real-time speed.

## 3.2. Cross-Modal Prototype Transfer

Considering the data structures of point cloud and image pixels are totally different, various network architectures are utilized to extract feature representations for the two modal inputs independently. Specifically, ResNet34 [13] encoder and FCN [23] decoder are used for 2D images to extract dense-grid features $F^{2d} \in \mathbb{R}^{W \times H \times D_{2d}}$, and a sparse convolution [12] based hierarchical point-voxel backbone [42] are designed for 3D point cloud to generate point-wise features $F^{3d} \in \mathbb{R}^{N \times D_{3d}}$. Since LiDAR points and image pixels are not naturally aligned due to the differences in data collection of LiDAR and camera devices, LiDAR-camera transformation matrix is used to find the point-to-pixel correspondence[3] and further obtain the matched pixel feature $F^{2d\_m}$ and matched point features $F^{2d\_m}$. Following the previous work [42], we concatenate the matched point-pixel features and use a multi-layer perceptron (MLP) to obtain the fusion feature

$$F^{\text{fuse}} = \text{MLP}\big(\text{cat}(F^{2d\_m}, F^{3d\_m})\big) \tag{1}$$

where $\odot$ is Hadamard product of two matrices. The MLP is used to reduce feature dimension to $D_{3d}$ and obtain $F^{\text{fuse}}$. With LiDAR per-point ground truth $Y^{3d}$, the fusion branch is got supervised

$$\mathcal{L}^{\text{fuse}} = \mathcal{L}_{\text{CE}}\big(h^{\text{fuse}}(F^{\text{fuse}}), \hat{Y}\big) + \mathcal{L}_{\text{Lovasz}}\big(h^{\text{fuse}}(F^{\text{fuse}}), \hat{Y}\big) \tag{2}$$

[2]Input feature normally contains Cartesian coordinates, intensity of returning laser beam, colors, etc.

[3]Detailed point-to-pixel mapping mechanism is provided in supplementary material.

where $h^{\text{fuse}}$ is the segmentation head of fusion branch. $\hat{Y}$ is the subset of $Y^{3\text{d}}$, representing the ground truth for matched points. $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{Lovasz}}$ denote the cross entropy loss and the IoU-style Lovasz loss [3] respectively.

Attributed to the integration of rich semantic information from 2D images, the segmentation performance of the fusion branch is superior to the LiDAR-sole branch (the bottom branch in Fig. 2). In order to make a single LiDAR point cloud segmentation benefit from the fusion branch, one straightforward way is to introduce a knowledge distillation loss between these two branches, as done in [42]. However, due to the matched point-pixel being limited, only partial point features can benefit from the fusion features and leave massive unmatched points under-benefited.

To bypass the strict matching restriction, we introduce a class-wise prototype bank to accumulate the fusion features during model optimization. Such a prototype bank can be served as a parameter-free segmentation head to regularize the distribution of all point features with LiDAR ground truth

$$\mathcal{L}^{\text{proto}} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{\text{CE}}(p_{c_i}, c_i) + \mathcal{L}_{\text{Lovasz}}(p_{c_i}, c_i), \tag{3}$$

$$\text{where } p_{c_i} := p(y = c_i | F_i^{3\text{d}}) = \frac{\exp\big(\cos(\mathsf{p}_{c_i}, F_i^{3\text{d}})/T\big)}{\sum_{j=1}^{C}\exp\big(\cos(\mathsf{p}_j, F_i^{3\text{d}})/T\big)}$$

where $F_i^{3\text{d}}$ and $c_i$ denote the feature and semantic label of $i$-th point respectively. $\mathsf{p}_{c_i} \in \mathbb{R}^{D_{3\text{d}}}$ represents the prototype in class $c_i$. cos is the cosine function to calculate the similarity of the prototype and point feature. $T$ is the temperature parameter to adjust the scale of similarity measurement and is empirically set to 0.1 in our study.

The proposed cross-modal prototype-based knowledge transfer module encourages the point features to be close to the same-class prototype while staying far away from other-class prototypes. Therefore, through the class-wise prototype bank as a bridge, we successfully transfer the knowledge learned from the fusion feature to the whole point cloud and we term our method as **ProtoTransfer**.

### 3.3. Prototype Bank Initialization and Update

In our work, the prototype bank is designed to be non-parametric and non-learnable and thus the quality of the prototype bank plays a critical role in the success of cross-modal knowledge transfer. Instead of using random vectors to initialize the prototype bank at the beginning of model training, we propose to use the fusion features of the first few iterations to warm up the class-wise prototype bank for a stable optimization procedure.

Meanwhile, since the initial prototype bank can not capture the distribution of fusion features, we also dynamically update the prototype bank throughout the training stage.

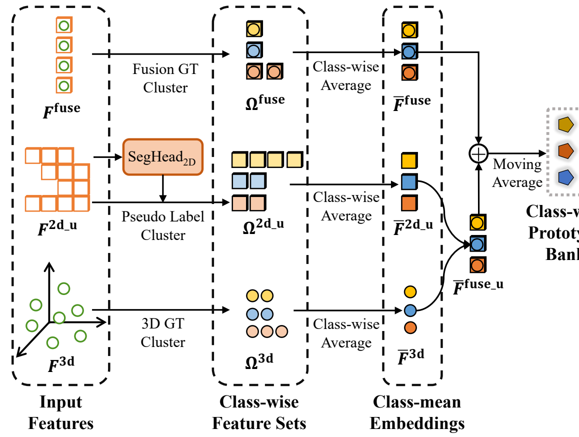


Figure 3. Prototype bank update demonstration. The fusion feature $F^{\text{fuse}}$, unmatched image feature $F^{2\text{d\_u}}$ and point feature $F^{3\text{d}}$ are respectively clustered into class-wise feature sets using corresponding GT or pseudo-label obtained with Eq. (6). Then the feature sets are class-wise averaged to produce class-mean embeddings. $\bar{F}^{2\text{d\_u}}$ and $\bar{F}^{3\text{d}}$ are fused to get $\bar{F}^{\text{fuse\_}u}$. Finally, both $\bar{F}^{\text{fuse}}$ and $\bar{F}^{\text{fuse\_}u}$ are accumulated into the class-wise prototype bank.

Given all fusion features $\{F_i^{\text{fuse}}\}$, we can use their corresponding class labels $\{c_i\}$ to cluster the fusion features into class-wise feature set. As shown in the top branch of Fig. 3, the fusion feature set of class $k$ can be obtained by

$$\Omega_k^{\text{fuse}} = \{F_i^{\text{fuse}} \mid c_i = k\}. \tag{4}$$

We then use the moving average trick to update the class-specific prototype $\mathsf{p}_k$ of class $k$

$$\mathsf{p}_k \leftarrow \alpha \cdot \mathsf{p}_k + (1-\alpha)\cdot \bar{F}_k^{\text{fuse}}, \tag{5}$$
$$\text{where} \quad \bar{F}_k^{\text{fuse}} = \frac{1}{\|\Omega_k^{\text{fuse}}\|}\sum\nolimits_j \Omega_{k,j}^{\text{fuse}}$$

where $\alpha$ is a fixed hyper-parameter to control the prototype update speed and is empirically set as 0.999 in our study. $\bar{F}_k^{\text{fuse}}$ is the class-mean embedding of fusion features in class $k$.

### 3.4. Unmatched Image Features Exploration and Exploitation

Although prototype-based knowledge transfer has helped all point features benefit from the fusion features and enhanced the point segmentation performance, the fusion features, which are the knowledge source of the prototype bank, are only available on matched point-pixel pairs. Since the point-pixel matching rate is quite low, the rich semantic information lying in a large amount of unmatched image pixels is not explored and exploited.

However, the exploitation of unmatched image pixels is non-trivial and there are at least two challenges: 1) Naturally lacking matching points for modal fusion. Since modal fusion requires well-matched point-pixel features, it is difficult to construct matching points for unmatched pixels.

2) Hardly obtaining semantic labels of unmatched pixels. Since 2D images are provided without annotations and it can be costly to additionally annotate these dense pixels, the semantic labels of unlabeled pixels are thus unknown without the matching correspondence from the LiDAR point cloud ground truth.

Based on our proposed cross-modal prototype transfer framework, we further design a novel strategy to incorporate the unmatched pixel features into the prototype bank to improve the transferred knowledge, and Fig. 3 presents an overview. Specifically, in order to solve the challenge 2), we formulate the image pixel class label generation problem as a weakly-/semi-supervised task [5, 30], i.e., taking the LiDAR-to-image matched pixels as labeled samples and taking other unmatched pixels as unlabeled ones. Considering the segmentation head $h^{\mathrm{2d}}$ for image input has been optimized on features of matched pixels[4], we can generate posterior probability estimation for unmatched pixels

$$p(y|F^{\mathrm{2d\_u}}) = \mathrm{softmax}\big(h^{\mathrm{2d}}(F^{\mathrm{2d\_u}})\big),$$

where $F^{\mathrm{2d\_u}}$ is the features of unmatched pixels. Then the pseudo-labels can be obtained for these unmatched pixels when their maximal posterior probability is greater than a pre-defined confidence threshold $\delta$ (set as 0.8 in our study)

$$\begin{aligned} &y^{\mathrm{2d\_u}} = \mathrm{argmax}_c\, p(y=c|F^{\mathrm{2d\_u}}), \\ &\text{only if} \quad p(y=y^{\mathrm{2d\_u}}|F^{\mathrm{2d\_u}}) \geq \delta. \end{aligned} \tag{6}$$

Given the pseudo-labeled unmatched pixels, it is still non-trivial to fuse the features of unmatched pixels and points due to the challenge 1) mentioned above. One straightforward way is to directly add the features of unmatched pixels to the prototype bank, empirical results in Tab. 4 show that this simple way can not bring a performance gain and we postulate that it may be caused by the distribution gap between image features and prototypes. Alternatively, we propose to first calculate a class-mean embedding $\bar{F}_k^{\mathrm{2d\_u}}$ of unmatched pixel features in the same pseudo-label $k$

$$\begin{aligned} &\bar{F}_k^{\mathrm{2d\_u}} = \frac{1}{\|\Omega_k^{\mathrm{2d\_u}}\|}\sum\nolimits_j \Omega_{k,j}^{\mathrm{2d\_u}}, \\ &\text{where} \quad \Omega_k^{\mathrm{2d\_u}} = \{F_i^{\mathrm{2d\_u}} \mid y_i^{\mathrm{2d\_u}} = k\}, \end{aligned}$$

$\Omega_k^{\mathrm{2d\_u}}$ is class-wise unmatched image feature set and is obtained using pseudo labels but not ground-truth as in Eq. (4). And the class-mean embedding $\bar{F}^{\mathrm{3d}}$ of point features can be obtained from $F^{\mathrm{3d}}$ with LiDAR ground truth. Then, modal fusion can be performed between class-mean embeddings of the same class

$$\bar{F}^{\mathrm{fuse\_u}} = \mathrm{MLP}\big(\mathrm{cat}(\bar{F}^{\mathrm{2d\_u}},\ \bar{F}^{\mathrm{3d}})\big), \tag{7}$$

where the MLP shares parameter with the one in Eq. (1). Finally, the class-wise prototype bank updating in Eq. (5) can be extended as

$$\mathsf{p}_k \leftarrow \alpha\cdot\mathsf{p}_k + (1-\alpha)\cdot(\lambda\cdot\bar{F}_k^{\mathrm{fuse}} + (1-\lambda)\cdot\bar{F}_k^{\mathrm{fuse\_u}}), \tag{8}$$

where $\lambda$ is the balance weight of these two fusion features.

Since the pseudo-labeling scheme for unmatched pixels may not be activated at the beginning of model training, the prototype bank is still initialized with the method mentioned in Sec. 3.3.

### 3.5. Overall Objective Function

Apart from the loss terms in Eq. (2) and Eq. (3), we also have separate losses $\mathcal{L}^{\mathrm{3d}}$ and $\mathcal{L}^{\mathrm{2d}}$ for LiDAR point cloud input and image input respectively. Both of these two losses are composed of a cross-entropy loss and a Lovasz loss as those in Eq. (2). Finally, the overall object function is the weighted sum of these loss terms

$$\mathcal{L} = \omega^{\mathrm{3d}}\mathcal{L}^{\mathrm{3d}} + \omega^{\mathrm{2d}}\mathcal{L}^{\mathrm{2d}} + \omega^{\mathrm{fuse}}\mathcal{L}^{\mathrm{fuse}} + \omega^{\mathrm{proto}}\mathcal{L}^{\mathrm{proto}},$$

where we empirically set $\omega^{\mathrm{3d}} = 2.0$ and $\omega^{\mathrm{2d}} = \omega^{\mathrm{fuse}} = \omega^{\mathrm{proto}} = 1.0$ in our study.

## 4. Experiments

In this section, we first provide details of our experimental setup. Then we evaluate ProtoTransfer on both nuScenes dataset and SemanticKITTI dataset. Finally, extensive ablation studies of our approach are presented.

### 4.1. Experimental Setup

**Datasets. NuScenes [4]** collects 1000 driving scenes from various locations in Boston and Singapore using 1 LiDAR and 6 cameras covering 360° FoV. According to the official setting, it is split into training, validation and test set as 700, 150 and 150 scenes. For point cloud semantic segmentation task, it annotates labels for 16 classes under different traffic and weather conditions. **SemanticKITTI [2]** contains 22 LiDAR sequences numbered from 00 to 21, in which sequence 08 is officially selected as validation set, sequence 00-10 except 08 is training set and 11-21 is test set. Unlike nuScenes, SemanticKITTI has only two front-view cameras.

**Evaluation Metric.** Following previous work [49, 42], we calculate intersection-over-union (IoU) of each class and mean IoU (mIoU) of all classes, which is formulated as $\mathrm{mIoU} = \frac{1}{C}\sum_{c=1}^{C}\frac{\mathrm{TP}_c}{\mathrm{TP}_c+\mathrm{FP}_c+\mathrm{FN}_c}$, where $\mathrm{TP}_c$, $\mathrm{FP}_c$ and $\mathrm{TP}_c$ denotes the number of true positive, false positive and false negative points of class $c$.

[4]Please note that the segmentation head for 2D images already exists in our training framework.

| Methods | input | mIoU | barrier | bicycle | bus | car | construction | motorcycle | pedestrian | traffic cone | trailer | truck | driveable | other flat | sidewalk | terrain | manmade | vegetation | Latency (ms) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PolarNet [45] | L | 69.4 | 72.2 | 16.8 | 77.0 | 86.5 | 51.1 | 69.7 | 64.8 | 54.1 | 69.7 | 63.5 | 96.6 | 67.1 | 77.7 | 72.1 | 87.1 | 84.5 | - |
| JS3C-Net [41] | L | 73.6 | 80.1 | 26.2 | 87.8 | 84.5 | 55.2 | 72.6 | 71.3 | 66.3 | 76.8 | 71.2 | 96.8 | 64.5 | 76.9 | 74.1 | 87.5 | 86.1 | - |
| Cylinder3D [47] | L | 77.2 | 82.8 | 29.8 | 84.3 | 89.4 | 63.0 | 79.3 | 77.2 | 73.4 | 84.6 | 69.1 | **97.7** | **70.2** | **80.3** | 75.5 | 90.4 | 87.6 | 63 |
| AMVNet [21] | L | 77.3 | 80.6 | 32.0 | 81.7 | 88.9 | 67.1 | 84.3 | 76.1 | 73.5 | 84.9 | 67.3 | 97.5 | 67.4 | 79.4 | 75.5 | 91.5 | 88.7 | 85 |
| SPVCNN [31] | L | 77.4 | 80.0 | 30.0 | 91.9 | 90.8 | 64.7 | 79.0 | 75.6 | 70.9 | 81.0 | 74.6 | 97.4 | 69.2 | 80.0 | **76.1** | 89.3 | 87.1 | 63 |
| (AF)$^2$-S3Net [6] | L | 78.3 | 78.9 | 52.2 | 89.9 | 84.2 | 77.4 | 74.3 | 77.3 | 72.0 | 83.9 | 73.8 | 97.1 | 66.5 | 77.5 | 74.0 | 87.7 | 86.8 | 270 |
| PMF [50] | L+C | 77.0 | 82.0 | 40.0 | 81.0 | 88.0 | 64.0 | 79.0 | 80.0 | 76.0 | 81.0 | 67.0 | 97.0 | 68.0 | 78.0 | 74.0 | 90.0 | 88.0 | 125* |
| 2D3DNet [10] | L+C | 80.0 | **83.0** | **59.4** | 88.0 | 85.1 | 63.7 | 84.4 | 82.0 | 76.0 | 84.8 | 71.9 | 96.9 | 67.4 | 79.8 | 76.0 | **92.1** | **89.2** | - |
| 2DPASS [42] | L | 80.8 | 81.7 | 55.3 | 92.0 | **91.8** | 73.3 | 86.5 | 78.5 | 72.5 | 84.7 | 75.5 | 97.6 | 69.1 | 79.9 | 75.5 | 90.2 | 88.0 | **44** |
| **ProtoTransfer** [Ours] | **L** | **82.1** | 81.1 | 55.5 | **93.9** | 91.5 | **77.9** | **87.3** | **82.7** | **78.9** | **85.1** | **76.5** | 97.6 | 69.8 | 79.6 | 75.9 | 91.5 | 89.1 | **44** |

Table 1. Semantic segmentation results on nuScenes *test* set. Methods published before the submission deadline (08/03/2023) are listed. L and C respectively denote LiDAR and camera. * The latency of PMF [50] is tested without TensorRT acceleration. The bold numbers indicate the best results, and the blue numbers indicate the second best results.

| Method | **mIoU** | road | sidewalk | parking | other-ground | building | car | truck | bicycle | motorcycle | other-vehicle | vegetation | trunk | terrain | person | bicyclist | motorcyclist | fence | pole | traffic sign | latency (ms) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SqueezeSegV2 [35] | 39.7 | 88.6 | 67.6 | 45.8 | 17.7 | 73.7 | 81.8 | 13.4 | 18.5 | 17.9 | 14.0 | 71.8 | 35.8 | 60.2 | 20.1 | 25.1 | 3.9 | 41.1 | 20.2 | 26.3 | - |
| DarkNet53Seg [2] | 49.9 | 91.8 | 74.6 | 64.8 | 27.9 | 84.1 | 86.4 | 25.5 | 24.5 | 32.7 | 22.6 | 78.3 | 50.1 | 64.0 | 36.2 | 33.6 | 4.7 | 55.0 | 38.9 | 52.2 | - |
| RangeNet53++ [25] | 52.2 | 91.8 | 75.2 | 65.0 | 27.8 | 87.4 | 91.4 | 25.7 | 25.7 | 34.4 | 23.0 | 80.5 | 55.1 | 64.6 | 38.3 | 38.8 | 4.8 | 58.6 | 47.9 | 55.9 | 83.3 |
| 3D-MiniNet [1] | 55.8 | 91.6 | 74.5 | 64.2 | 25.4 | 89.4 | 90.5 | 28.5 | 42.3 | 42.1 | 29.4 | 82.8 | 60.8 | 66.7 | 47.8 | 44.1 | 14.5 | 60.8 | 48.0 | 56.6 | - |
| SqueezeSegV3 [38] | 55.9 | 91.7 | 74.8 | 63.4 | 26.4 | 89.0 | 92.5 | 29.6 | 38.7 | 36.5 | 33.0 | 82.0 | 58.7 | 65.4 | 45.6 | 46.2 | 20.1 | 59.4 | 49.6 | 58.9 | 238 |
| PointNet++ [26] | 20.1 | 72.0 | 41.8 | 18.7 | 5.6 | 62.3 | 53.7 | 0.9 | 1.9 | 0.2 | 0.2 | 46.5 | 13.8 | 30.0 | 0.9 | 1.0 | 0.0 | 16.9 | 6.0 | 8.9 | 5900 |
| TangentConv [32] | 40.9 | 83.9 | 63.9 | 33.4 | 15.4 | 83.4 | 90.8 | 15.2 | 2.7 | 16.5 | 12.1 | 79.5 | 49.3 | 58.1 | 23.0 | 28.4 | 8.1 | 49.0 | 35.8 | 28.5 | 3000 |
| PointASNL [43] | 46.8 | 87.4 | 74.3 | 24.3 | 1.8 | 83.1 | 87.9 | 39.0 | 0.0 | 25.1 | 29.2 | 84.1 | 52.2 | 70.6 | 34.2 | 57.6 | 0.0 | 43.9 | 57.8 | 36.9 | - |
| RandLA-Net [16] | 55.9 | 90.5 | 74.0 | 61.8 | 24.5 | 89.7 | 94.2 | 43.9 | 29.8 | 32.2 | 39.1 | 83.8 | 63.6 | 68.6 | 48.4 | 47.4 | 9.4 | 60.4 | 51.0 | 50.7 | 880 |
| KPConv [33] | 58.8 | 90.3 | 72.7 | 61.3 | 31.5 | 90.5 | 95.0 | 33.4 | 30.2 | 42.5 | 44.3 | 84.8 | 69.2 | 69.1 | 61.5 | 61.6 | 11.8 | 64.2 | 56.4 | 47.4 | - |
| PolarNet [45] | 54.3 | 90.8 | 74.4 | 61.7 | 21.7 | 90.0 | 93.8 | 22.9 | 40.3 | 30.1 | 28.5 | 84.0 | 65.5 | 67.8 | 43.2 | 40.2 | 5.6 | 61.3 | 51.8 | 57.5 | **62** |
| JS3C-Net [41] | 66.0 | 88.9 | 72.1 | 61.9 | 31.9 | 92.5 | 95.8 | 54.3 | 59.3 | 52.9 | 46.0 | 84.5 | 69.8 | 67.9 | 69.5 | 65.4 | 39.9 | 70.8 | 60.7 | 68.7 | 471 |
| SPVNAS [31] | 67.0 | 90.2 | 75.4 | 67.6 | 21.8 | 91.6 | 97.2 | 56.6 | 50.6 | 50.4 | 58.0 | 86.1 | 73.4 | 71.0 | 67.4 | 67.1 | 50.3 | 66.9 | 64.3 | 67.3 | 259 |
| Cylinder3D [47] | 68.9 | 92.2 | 77.0 | 65.0 | 32.3 | 90.7 | 97.1 | 50.8 | 67.6 | 63.8 | 58.5 | 85.6 | 72.5 | 69.8 | 73.7 | 69.2 | 48.0 | 66.5 | 62.4 | 66.2 | 131 |
| RPVNet [40] | 70.3 | **93.4** | **80.7** | 70.3 | 33.3 | **93.5** | **97.6** | 44.2 | **68.4** | 68.7 | 61.1 | 86.5 | **75.1** | 71.7 | 75.9 | 74.4 | 43.4 | 72.1 | 64.8 | 61.4 | 168 |
| (AF)$^2$-S3Net [6] | 70.8 | 92.0 | 76.2 | 66.8 | **45.8** | 92.5 | 94.3 | 40.2 | 63.0 | **81.4** | 40.0 | 78.6 | 68.0 | 63.1 | 76.4 | **81.7** | **77.7** | 69.6 | 64.0 | **73.3** | - |
| 2DPASS [42] | 72.9 | 89.7 | 74.7 | 67.4 | 40.0 | **93.5** | 97.0 | **61.1** | 63.6 | 63.4 | 61.5 | 86.2 | 73.9 | 71.0 | 77.9 | 81.3 | 74.1 | **72.9** | 65.0 | 70.4 | **62** |
| **ProtoTransfer** [Ours] | **73.6** | 91.5 | 77.5 | **71.5** | 42.8 | 92.9 | 97.5 | 57.1 | **68.4** | 76.0 | **68.0** | **86.7** | 74.8 | **72.6** | **80.0** | 76.0 | 61.4 | 70.7 | **65.8** | 66.7 | **62** |

Table 2. Semantic segmentation results on SemanticKITTI *test* set. Methods published before the submission deadline (08/03/2023) are listed. The bold numbers indicate the best results, and the blue numbers indicate the second best results.

In order to have a fair comparison, most of the experimental setup and implementations are identical to the recent work [42]. During the model evaluation, the same test-time augmentation as in [42] is also used in our study.

### 4.2. Results on Benchmarks

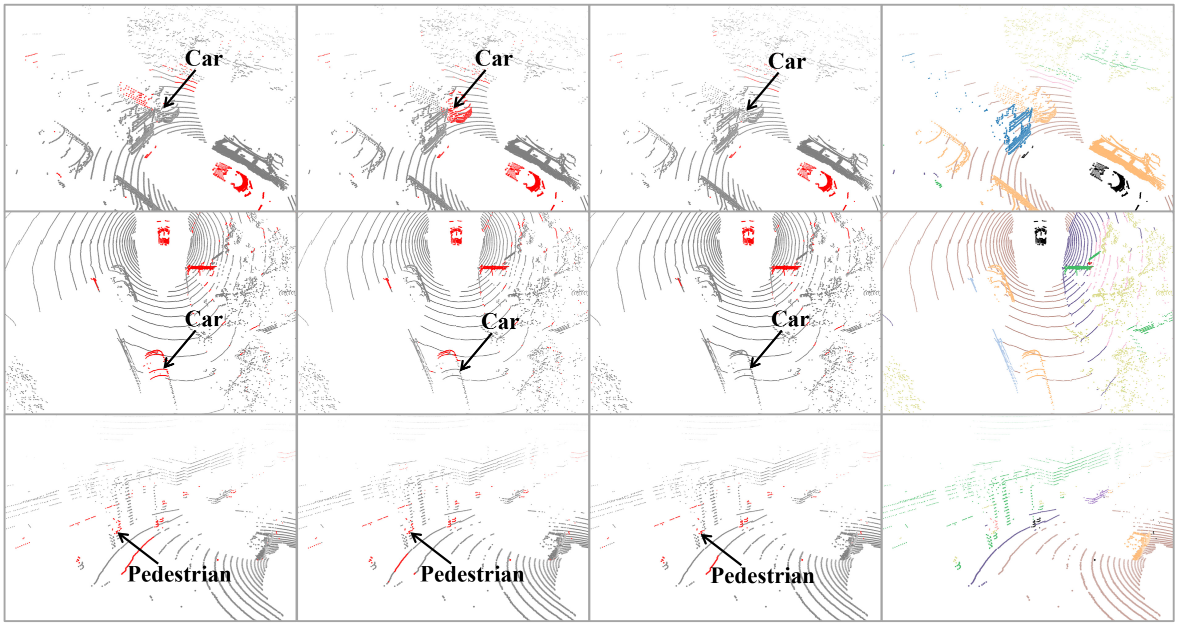
We compare the results of our ProtoTransfer with the published state-of-the-art methods on two large-scale benchmarks, i.e., nuScenes [4] and SemanticKITTI [2].

**NuScenes**. As shown in Tab. 1, ProtoTransfer successfully outperforms all existing methods in terms of both mIoU and latency, demonstrating its efficacy and efficiency. Moreover, our ProtoTransfer not only outperforms single-modal approaches, but also surprisingly surpasses fusion-based methods which require both of LiDAR point clouds and images covering the whole FoV as input for the inference stage. In contrast, our method ProtoTransfer only takes point clouds as input and produces superior segmentation results. Compared to the recently proposed distillation-based method 2DPASS [42], our ProtoTransfer achieves 1.3% performance gain, showing the success of exploring prototypes for knowledge transfer from multi-modal to single-modal. Besides, according to Tab. 1, we can find that our ProtoTransfer performs best on classes of small size and with sparse points, such as traffic-cone, motorcycle and construction vehicle, showing great potential in real-world practice.

**SemanticKITTI.** We compare our ProtoTransfer with several previous state-of-the-arts works. From Tab. 2, we can see that our proposed ProtoTransfer still achieves the best performance among these methods, leaving a margin of 0.7% compared with the distillation-based 2DPASS [42]. Similarly, ProtoTransfer achieves the best results on classes of small objects such as person and bicycle.

### 4.3. Qualitative Evaluation

We visualize the segmentation results on nuScenes validation set in Fig. 4. As can be observed, our ProtoTransfer achieves the most minor error prediction compared with the two comparison methods. In the first row of Fig. 4, the car and bus are close to each other, neither the baseline method nor the 2DPASS approach can produce an accurate segmen-




**(a) Error by baseline** **(b) Error by 2DPASS** **(c) Error by ProtoTransfer** **(d) Ground-truth**

Figure 4. Qualitative results on nuScenes Lidarseg validation set. Red points are in red. Compared to baseline and the recently proposed distillation-based method 2DPASS [42], our ProtoTransfer achieves better segmentation on region boundaries (the first row), far objects (the second row) and small objects (the third row), thanks to the fully exploited and transferred multi-modal knowledge.

| baseline | $\mathcal{L}^{\text{proto}}$ | update proto. bank | mIoU |
|---|---|---|---|
| ✓ | | | 76.21 |
| ✓ | ✓ | | 76.50 |
| ✓ | ✓ | ✓ | **80.51** |

Table 3. Ablation study of each component in our ProtoTransfer.

tation for the points in the class of car, while our ProtoTransfer precisely segments all points belonging to the car. The second row displays a car far away from the ego-car, both of the two comparing methods mispredict the point clouds in the head area of the car and ProtoTransfer segments the car points perfectly again. The last row presents a pedestrian with only several LiDAR points. Thanks to the fully exploited and transferred multi-modal knowledge, our ProtoTransfer obtains the most accurate segmentation.

## 4.4. Ablation Studies

In this section, we present comprehensive ablation studies to examine the efficacy of each component in ProtoTransfer, ways of prototype bank construction, and distribution of feature representations. All the ablation studies are conducted on the nuScenes validation set.

**Effects of each component.** As shown in Tab. 3, introducing the proposed prototype-based loss term $\mathcal{L}^{\text{proto}}$ into the naive baseline method without updating the prototype bank can only bring minor performance gain. When the prototype bank is dynamically updated with the strategy presented in Sec. 3.4, the segmentation performance is significantly boosted which obviously demonstrates that the quality of the prototype bank is critical.

| | $F^{3d}$ | $F^{2d}$ | $\bar{F}^{\text{fuse}}$ | $\bar{F}^{\text{fuse\_u}}$ | mIoU |
|---|---|---|---|---|---|
| baseline | | | | | 76.21 |
| (a) | ✓ | | | | 76.97 |
| (b) | | ✓ | | | 75.35 |
| (c) | | | ✓ | | 78.65 |
| (d) | ✓ | | ✓ | | 77.11 |
| (e) | | ✓ | ✓ | | 75.67 |
| (f) | ✓ | ✓ | ✓ | | 75.58 |
| (g) | | | | ✓ | 79.83 |
| **(h)** | | | ✓ | ✓ | **80.51** |

Table 4. Abaltion study of inputs of prototype bank.

**Input Source of Prototype Bank.** As presented in Sec. 3, the inputs of the prototype bank include point-pixel fusion features, image features and point features.Tab. 4 shows an ablation study on the effect of different input combinations on final segmentation performance. As can be observed, building up the prototype bank using only point cloud features (a) can generate a 0.76% performance gain over the baseline. We think it is because prototypes constructed from only point features can serve as cluster centers and using the proposed prototype-based loss $\mathcal{L}^{\text{proto}}$ can help point features become more discriminative as they are pushed to be closer to their cluster centers. However, only accumulating image features[5] into the prototype bank leads to a 0.86% performance drop, we postulate it may be because there is

[5]Please note the pseudo-labeling scheme presented in Sec. 3.4 can also be used here to exploit unmatched image features.

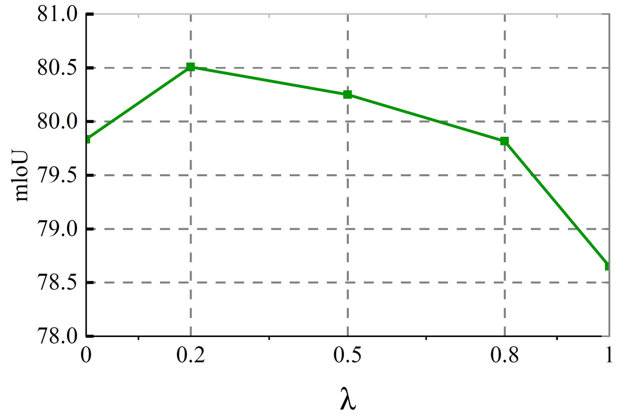


Figure 5. Ablation study of $\lambda$ in Eq. (8). $\lambda = 0.2$ is set as default in ProtoTransfer.

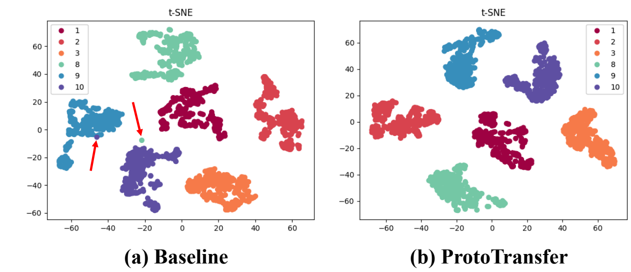


Figure 6. Feature representation visualization of (a) the baseline and (b) our ProtoTransfer. Outlier points have been pointed out by the red arrow. The displayed category ID and their corresponding semantic category is {1: "barrier", 2: "bicycle", 3: "bus", 8: "traffic cone", 9: "trailer", 10: "truck"}.

a distribution gap between image features and point cloud features and thus it is unwise to force point cloud features to imitate image feature cluster centers. On the contrary, when fusing the point-pixel matched features and only using the fusion features for the prototype bank (c) excels 2.44% over the baseline, demonstrating the effectiveness of knowledge transfer from multi-modal to the single LiDAR modal. As shown by Tab. 4 (h), when the fusion features $\bar{F}^{\text{fuse}}$ and $\bar{F}^{\text{fuse\_u}}$ are fully explored and accumulated into the prototype bank, our ProtoTransfer reaches the best mIoU of 80.51%. This ablation study clearly shows that the input source plays a key role in the quality of the prototype bank.

**Fusion Strategy of Features for Prototype Bank.** As illustrated in Fig. 3 and Sec. 3.4, we calculate class-mean embeddings for the unmatched pixel features and point features, respectively. Then we concatenate these two embeddings and fuse them by reusing the MLP in Eq. (1). The fused embedding $\bar{F}^{\text{fuse\_u}}$ is finally accumulated to the prototype bank together with the point-pixel matched fusion features $\bar{F}^{\text{fuse}}$. To demonstrate the improvement not only comes from the additional image and point features but also comes from the cleverly designed fusion strategy, we use a naive way to have an investigation, i.e., any combinations among the point-pixel naturally matched fusion feature $\bar{F}^{\text{fuse}}$, image feature $F^{2\text{d}}$ and point feature $F^{3\text{d}}$ in the way of simple summation and use moving average to accumulate them into the prototype bank as presented in Tab. 4 (d)∼(f). We can find that this naive summation of all three features (f) reduces the mIoU to 75.58%. We owe this to the large distribution gap between different kinds of features which cannot be handled in the naive sum-up way. In contrast, our ProtoTransfer reuses the MLP to map the unmatched image features to the fusion feature space, thus successfully bypassing this problem.

**Effect of $\lambda$ Selection.** We have experimented with different values of $\lambda$ in Eq. (8) and the results are presented in Fig. 5. As can be observed, our ProtoTransfer achieves the best performance when $\lambda = 0.2$ which reveals that the unmatched-pixel-feature-based fusion embedding $\bar{F}^{\text{fuse\_u}}$ plays a more important role. Note that when $\lambda = 0$, i.e., using only $\bar{F}^{\text{fuse\_u}}$, is able to achieve a satisfactory segmentation accuracy, demonstrating the benefits brought by the unmatched image features.

**Distribution of Feature Representation.** The essence of introducing the prototype-based loss term $\mathcal{L}^{\text{proto}}$ in our method is to encourage the point features to be close to the same-class multi-modal prototype while staying far away from the other-class prototypes. Hence, it is important to study the impact of prototypes on feature distribution. As can be observed in Fig. 6 (a), outlier points appear in the feature distribution of the baseline method, while our ProtoTransfer is able to produce more compact feature distributions for all semantic classes, demonstrating the effectiveness of our method.

## 5. Conclusion

This work presents a cross-modal knowledge transfer method dubbed ProtoTransfer for point cloud semantic segmentation. ProtoTransfer achieves remarkable segmentation performance but keep a single LiDAR inference speed. By accumulating the fusion features into a prototype bank, all LiDAR points can learn from their class-specific prototypes, thus being well benefited. The unmatched image features are further explored and exploited via a pseudo-labeling scheme and a novel prototype bank update strategy. Through extensive experimental results on nuScenes and SemanticKITTI dataset, the efficacy of our method has been successfully demonstrated.

**Acknowledgements.** This work was supported by Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and the Fundamental Research Funds for the Central Universities.

## 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. *arXiv preprint arXiv:2002.10893*, 2020.

[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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 9297–9307, 2019.

[3] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018.

[4] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[5] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2617–2626, 2022.

[6] Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, and Bingbing Liu. Af2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12547–12556, 2021.

[7] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, and Wanli Ouyang. Monodistill: Learning spatial features for monocular 3d object detection. In *International Conference on Learning Representations*, 2022.

[8] Khaled El Madawi, Hazem Rashed, Ahmad El Sallab, Omar Nasr, Hanan Kamel, and Senthil Yogamani. Rgb and lidar fusion based 3d semantic segmentation for autonomous driving. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 7–12. IEEE, 2019.

[9] Salvador Garcia, Joaquin Derrac, Jose Cano, and Francisco Herrera. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE transactions on pattern analysis and machine intelligence*, 34(3):417–435, 2012.

[10] 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 *2021 International Conference on 3D Vision (3DV)*, pages 361–372. IEEE, 2021.

[11] Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. Neighbourhood components analysis. *Advances in neural information processing systems*, 17, 2004.

[12] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[15] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8479–8488, 2022.

[16] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[17] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12605–12614, 2020.

[18] Saumya Jetley, Bernardino Romera-Paredes, Sadeep Jayasumana, and Philip Torr. Prototypical priors: From improving classification to zero-shot learning. *arXiv preprint arXiv:1512.01192*, 2015.

[19] Bo Ju, Zhikang Zou, Xiaoqing Ye, Minyue Jiang, Xiao Tan, Errui Ding, and Jingdong Wang. Paint and distill: Boosting 3d object detection with semantic passing network. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5639–5648, 2022.

[20] Georg Krispel, Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. Fuseseg: Lidar point cloud segmentation fusing multi-modal data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1874–1883, 2020.

[21] Venice Erin Liong, Thi Ngoc Tho Nguyen, Sergi Widjaja, Dhananjai Sharma, and Zhuang Jie Chong. Amvnet: Assertion-based multi-view fusion network for lidar semantic segmentation. *arXiv preprint arXiv:2012.04934*, 2020.

[22] Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multitask multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[24] Gregory P Meyer, Jake Charland, Darshan Hegde, Ankit Laddha, and Carlos Vallespi-Gonzalez. Sensor fusion for joint 3d object detection and semantic segmentation. In *Pro-*

*ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

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

[26] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[28] Ruslan Salakhutdinov and Geoff Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. In *Artificial Intelligence and Statistics*, pages 412–419. PMLR, 2007.

[29] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[30] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[31] 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*, pages 685–702. Springer, 2020.

[32] Maxim Tatarchenko, Jaesik Park, Vladlen Koltun, and Qian-Yi Zhou. Tangent convolutions for dense prediction in 3d. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3887–3896, 2018.

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

[34] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11794–11803, 2021.

[35] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4376–4382. IEEE, 2019.

[36] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[37] Aoran Xiao, Jiaxing Huang, Dayan Guan, Kaiwen Cui, Shijian Lu, and Ling Shao. Polarmix: A general data augmentation technique for lidar point clouds. In *Advances in Neural Information Processing Systems*.

[38] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. *arXiv preprint arXiv:2004.01803*, 2020.

[39] Haiming Xu, Lingqiao Liu, Qiuchen Bian, and Zhen Yang. Semi-supervised semantic segmentation with prototype-based consistency regularization. In *Advances in Neural Information Processing Systems*.

[40] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16024–16033, 2021.

[41] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3101–3109, 2021.

[42] 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*, pages 677–695. Springer, 2022.

[43] Xu Yan, Chaoda Zheng, Zhen Li, Sheng Wang, and Shuguang Cui. Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5598, 2020.

[44] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[45] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9601–9610, 2020.

[46] Wu Zheng, Mingxuan Hong, Li Jiang, and Chi-Wing Fu. Boosting 3d object detection by simulating multimodality on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13638–13647, 2022.

[47] Hui Zhou, Xinge Zhu, Xiao Song, Yuexin Ma, Zhe Wang, Hongsheng Li, and Dahua Lin. Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation. *arXiv preprint arXiv:2008.01550*, 2020.

[48] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2582–2593, 2022.

[49] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9939–9948, 2021.

[50] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16280–16290, 2021.