ProtoTransfer: Cross-Modal Prototype Transfer for Point Cloud Segmentation

Pin Tang1  Hai-Ming Xu2  Chao Ma1*
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.

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 complimentary 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-
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 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 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.

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 paid 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 logsit 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.

3338

---

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

3.1. Framework Overview

Given a LiDAR point cloud frame with $N$ unordered points $P = \{p_i \in \mathbb{R}^{d_{in}}\}_{i=1}^{N}$ where $d_{in}$ 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 $F^{3d,m}$. Following the previous work [42], we concatenate the matched point-pixel features and use a multi-layer perceptron (MLP) to obtain the fusion feature

$$F^{\text{fuse}} = \text{MLP}(\text{cat}(F^{2d,m}, F^{3d,m}))$$

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

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

---

\(^2\)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 \( Y^{2d} \), 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 requirement, 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

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

where \( p_{c_i} := p(y = c_i | F_i^{3d}) = \frac{\exp\left(\cos(p_{c_i}, F_i^{3d}) / T\right)}{\sum_{j=1}^{C} \exp(\cos(p_j, F_i^{3d}) / T)} \)

where \( F_i^{3d} \) and \( c_i \) denote the feature and semantic label of \( i \)-th point respectively. \( p_{c_i} \in \mathbb{R}^{D_{3d}} \) 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.

\[
\bar{F}_{k} = \frac{1}{\Omega_k} \sum_{j} \Omega_{k,j} F_{k,j}^{\text{fuse}}
\]

where \( \alpha \) is a fixed hyper-parameter to control the prototype update speed and is empirically set as 0.999 in our study. \( F_{k}^{\text{fuse}} \) 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 $h^{2d}$ for image input has been optimized on features of matched pixels\footnote{Please note that the segmentation head for 2D images already exists in our training framework.}, we can generate posterior probability estimation for unmatched pixels

$$p(y | F^{2d, u}) = \text{softmax}(h^{2d}(F^{2d, u})), \quad \text{where } F^{2d, u} \text{ is the features of unmatched pixels. Then the pseudo-labels can be obtained for these unmatched pixels when their maximal posterior probability is greater than a pre-defined confidence threshold } \delta \text{ (set as 0.8 in our study)}$$

$$y^{2d, u} = \arg \max_c p(y = c | F^{2d, u}), \quad \text{only if } p(y = y^{2d, u} | F^{2d, u}) \geq \delta.$$  

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 $\bar{F}^{2d, u}_k$ of unmatched pixel features in the same pseudo-label $k$

$$\bar{F}^{2d, u}_k = \frac{1}{|\Omega^{2d, u}_k|} \sum_j \Omega^{2d, u}_{k,j}, \quad \text{where } \Omega^{2d, u}_k = \{F^{2d, u}_i | y^{2d, u}_i = k\},$$

$\Omega^{2d, u}_k$ 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 $\bar{F}^{3d}$ of point features can be obtained from $F^{3d}$ with LiDAR ground truth. Then, modal fusion can be performed between class-mean embeddings of the same class

$$\bar{F}^{fuse, u}_k = \text{MLP}\left(\text{cat}(\bar{F}^{2d, u}_k, \bar{F}^{3d})\right), \quad \text{(7)}$$

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

$$p_k = \alpha \cdot p_k + (1 - \alpha) \cdot (\lambda \cdot \bar{F}^{fuse}_k + (1 - \lambda) \cdot \bar{F}^{fuse, u}_k), \quad \text{(8)}$$

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 $L^{3d}$ and $L^{2d}$ 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

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

where we empirically set $\omega^{3d} = 2.0$ and $\omega^{2d} = \omega^{fuse} = \omega^{proto} = 1.0$ 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

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FP_c + FN_c}, \text{ where } TP_c, FP_c \text{ and } TP_c \text{ denotes the number of true positive, false positive and false negative points of class } c.$$
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 segmentation.
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.

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

<table>
<thead>
<tr>
<th>baseline</th>
<th>$\mathcal{L}^{\text{Proto}}$</th>
<th>update proto. bank</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td></td>
<td>76.21</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>76.50</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>80.51</td>
</tr>
</tbody>
</table>

Table 4. Abalition study of inputs of prototype bank.

<table>
<thead>
<tr>
<th>baseline</th>
<th>$F^{3d}$</th>
<th>$F^{2d}$</th>
<th>$F^{\text{Fuse}}$</th>
<th>$F^{\text{Fuse,2d}}$</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>76.21</td>
</tr>
<tr>
<td>(a)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>76.97</td>
</tr>
<tr>
<td>(b)</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>75.35</td>
</tr>
<tr>
<td>(c)</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td>78.65</td>
</tr>
<tr>
<td>(d)</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>77.11</td>
</tr>
<tr>
<td>(e)</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>75.67</td>
</tr>
<tr>
<td>(f)</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>75.58</td>
</tr>
<tr>
<td>(g)</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>79.83</td>
</tr>
<tr>
<td>(h)</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>80.51</td>
</tr>
</tbody>
</table>

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.

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.
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 $F_{\text{fuse}}$ and $F_{\text{fuse},u}$ 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 $F_{\text{fuse},u}$ is finally accumulated to the prototype bank together with the point-pixel matched fusion features $F_{\text{fuse}}$. 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 $F_{\text{fuse}}$, image feature $F_{2d}$ and point feature $F_{3d}$ 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 $F_{\text{fuse},u}$ plays a more important role. Note that when $\lambda = 0$, i.e., using only $F_{\text{fuse},u}$, 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 $L_{\text{proto}}$ 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


3345