Self-Supervised Collaborative Distillation: Enhancing Lighting Robustness and 3D Awareness

Wonjun Jo[1] Hyunwoo Ha[1] Kim Ji-Yeon[1] Hawook Jeong[2] Tae-Hyun Oh[3]

[1]POSTECH [2]RideFlux [3]KAIST

Abstract

As deep learning continues to advance, self-supervised learning has made considerable strides. It allows 2D image encoders to extract useful features for various downstream tasks, including those related to vision-based systems. Nevertheless, pre-trained 2D image encoders face two key challenges: nighttime lighting conditions and limited 3D awareness, which are required for robust perception and 3D understanding of reliable vision-based systems. To address these issues, we propose a novel self-supervised approach, ***Collaborative Distillation****, which improves light-invariance and 3D awareness in 2D image encoders while retaining semantic context, integrating the strengths of 2D image and 3D LiDAR data. Our method significantly outperforms competing methods in various downstream tasks across diverse lighting conditions and exhibits strong generalization ability. This advancement highlights our method's practicality and adaptability in real-world scenarios.*

1. Introduction

While deep learning models have shown considerable progress [9, 31, 35, 78], many of them rely on supervised learning, which requires extensive human labeling—a costly process [22, 79]. In contrast, self-supervised learning methods, which are label-efficient, have shown significant progress in the image domain [5, 10, 49, 77]. These self-supervised learning methods allow the image encoders to learn versatile features that are effective for downstream tasks, such as semantic segmentation and depth estimation, which are beneficial in vision-based systems [8, 23].

However, the pre-trained model by these self-supervised learning approaches often faces two key challenges. The first is lighting variation, particularly at night. In real-world scenarios, vision-based systems must operate reliably even under low-light or nighttime conditions. Yet, pre-trained image encoders often struggle in such environments, likely because their pre-training datasets, such as ImageNet-1K [17] and LVD-142M [49], are predominantly composed of daylight images. The second challenge is 3D awareness, a crucial capability for applications like autonomous driving, surveillance, and robot navigation. Although recent pre-trained image encoders exhibit a certain degree of 3D awareness, there is still room for improvement [19].

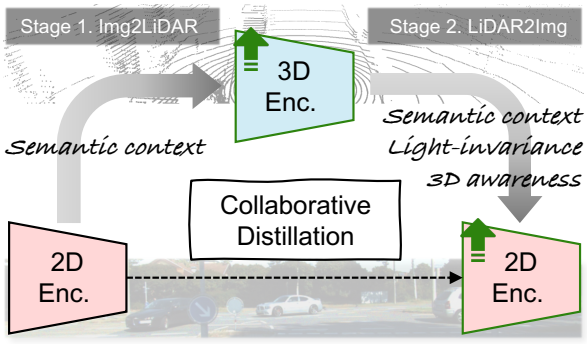


Figure 1. **Collaborative distillation scheme for complementary improvement.** Our proposed collaborative distillation method addresses two challenges in 2D image representation: light invariance and 3D awareness. In stage 1, the 3D LiDAR encoder learns semantic context. In stage 2, the 2D image encoder learns light-invariance and 3D awareness, preserving the original semantic context.

To address these challenges, we leverage 3D LiDAR data. 3D LiDAR offers two key advantages: (1) invariance to lighting conditions [18] and (2) inherent 3D information, both of which 2D images generally lack. The main question is how to enable a 2D image encoder to learn these favorable properties from 3D LiDAR data effectively. Since raw 3D LiDAR data lacks semantic context, directly injecting information with only 3D LiDAR data may degrade the 2D encoder's original semantic context and generalization ability. To prevent this degradation of the 2D encoder, a careful design is needed instead of a straightforward approach.

Therefore, we propose a two-stage self-supervised distillation paradigm, *collaborative distillation*, to preserve the original strengths of 2D encoders while enhancing their light-

invariance and 3D awareness (see Fig. 1). A key rationale for this design is the inclusion of a preliminary step that provides the 2D encoder's semantic context to the 3D encoder before the transfer of the 3D encoder's knowledge (image-to-LiDAR distillation). Then, we distill the 3D encoder's knowledge back to the 2D encoder (LiDAR-to-image distillation). Through this process, the 2D encoder acquires light-invariant and 3D-aware representations while retaining its original semantic context.

We demonstrate the effectiveness of our method on downstream tasks, including depth estimation, semantic, and depth-aware video panoptic segmentation, using both in-domain and out-of-domain datasets. We summarize our main contributions as follows:

- To the best of our knowledge, this study is the first self-supervised representation learning to enhance the robustness of 2D image encoders under challenging lighting conditions while improving their 3D awareness.
- We propose a novel collaborative distillation method that effectively exchanges the complementary properties of images and LiDAR through a two-stage distillation process.
- Our method enhances 2D image encoders across domains, demonstrating the potential of its strong generalization ability and adaptability in various vision-based systems.

2. Related Work

Our work is related to 2D images and 3D LiDAR self-supervised representation learning. We focus on improving 2D representation in day and nighttime conditions and 3D awareness. Therefore, we brief the related work: 2D self-supervised representation learning, image-to-LiDAR distillation, improving 2D encoders in nighttime conditions, and improving 3D awareness of 2D representation.

2D self-supervised representation learning. 2D self-supervised representation learning approaches have led to remarkable advancements in image understanding. These methods aim to learn visual features from the unlabeled data, which can be applied to various downstream tasks [39]. Recently, three pretext tasks are commonly chosen as main strategies: (1) contrastive learning, where a 2D encoder learns to extract features that are invariant to augmentations [10, 11, 24, 26, 46, 48, 61, 67]; (2) masked image modeling, where the encoder is trained to reconstruct masked parts of an image [3, 25]; and (3) self-distillation, where a student model learns by predicting the features of a teacher model [5, 49, 77]. However, most of these methods are trained on daylight images, leaving their effectiveness in nighttime conditions underexplored. Additionally, 2D self-supervised learning methods have insufficient 3D awareness [19]. Our method tackles both challenges in a self-supervised distillation approach.

Image-to-LiDAR distillation. While traditional 3D self-supervised approaches learn only from 3D LiDAR point clouds [12, 28, 47, 57, 62, 68, 72, 75], recent strategies also leverage the semantic information of the image domain by pre-training 3D encoders on image-LiDAR pairs. SLidR [58] pioneered an image-to-LiDAR distillation, where 2D image superpixels and matched 3D point clouds are compared through contrastive learning. Building on SLidR, several methods [40, 44, 50] develop loss functions to push performance further or revisit the data utilization [33]. ScaLR [53] scaled up the dataset and the model size. These methods are based on the pre-trained image encoder. However, directly using this encoder may cause unreliable distillation in challenging scenarios like low lighting or nighttime. We focus on improving the 2D image encoder itself in these scenarios.

Improving 2D encoders in nighttime condition. There are two streams of focusing improvement in low light, including nighttime conditions. Most studies focus on enhancing low-light images to make them more visible [7, 32, 41, 43, 64]. However, these methods primarily improve image quality and do not enhance feature representations for downstream tasks. Another line of research focuses on robust recognition in low light [20, 36, 56, 60, 63]. Even so, these studies are limited to their specific task, *e.g.*, depth estimation, object detection, or pose estimation. Our method is the first self-supervised representation designed to improve 2D image encoders under nighttime lighting. Unlike the above streams, this enables it to generalize across various downstream tasks.

Improving 3D awareness of 2D representation. There have been several studies aimed at improving the 3D awareness of 2D representations. Various methods [2, 29, 30, 66] incorporate 3D priors through multi-view geometry or masked image modeling with RGB-D data. Recently, FiT3D [73] lifts 2D features into 3D Gaussians and applies multi-view consistent rendering. Condense [74] enforces 2D-3D feature consistency via ray marching [45], enabling training of 2D and 3D encoders. Unlike the above studies that focus solely on 3D priors in indoor scenarios, several studies [13, 14, 27, 38, 65, 69] have utilized LiDAR-to-image distillation in outdoor autonomous driving, targeting a specific task, *e.g.*, 3D or BEV-based object detection. In contrast, we aim to enhance 2D encoders that take a single 2D image as input, enabling them to perform various downstream tasks in both in and out-of-domain for broad applicability.

3. Collaborative Distillation

In this section, we introduce a self-supervised collaborative distillation method for 2D image encoders to learn representation robust to lighting conditions and exhibit 3D awareness. This method is divided into stages 1 and 2 (see Fig. 2), which we describe sequentially in Secs. 3.1 and 3.2.

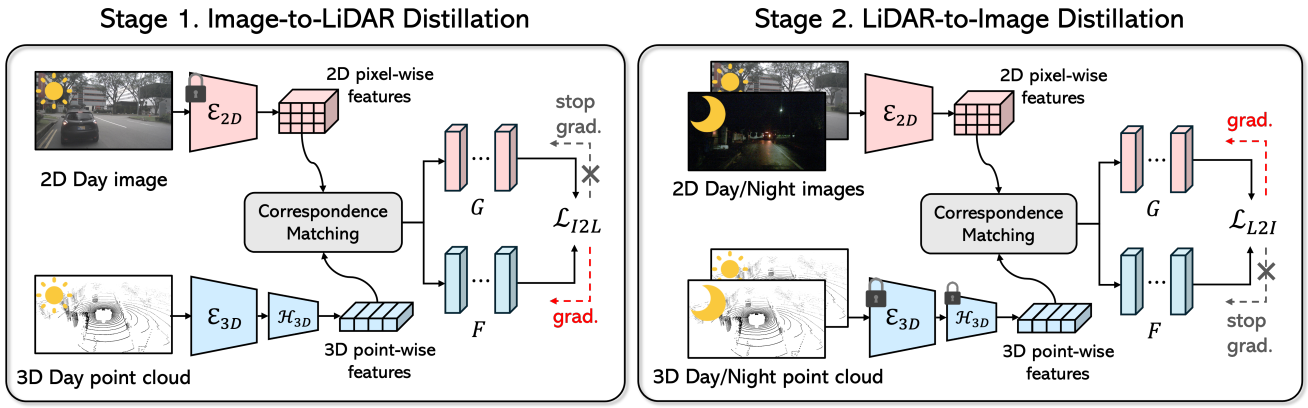


Figure 2. **Overall pipeline**. Our method consists of two stages: (1) Image-to-LiDAR distillation, where the semantically rich features from the 2D encoder $\mathcal{E}_{2D}$ are distilled into the 3D encoder $\mathcal{E}_{3D}$ using day images and LiDAR data, and (2) LiDAR-to-image distillation, where the representations learned by the 3D encoder $\mathcal{E}_{3D}$ are distilled back into the 2D encoder $\mathcal{E}_{2D}$. In this second stage, both day and nighttime images and LiDAR data are used, allowing the 2D encoder to learn robust features invariant to lighting conditions and exhibit 3D awareness.

3.1. Stage 1: Image-to-LiDAR Distillation

The goal of stage 1 is to provide its original semantic context to the 3D LiDAR encoder as a preliminary step. This helps the pre-trained 2D image encoder to retain the semantic context in the next stage.

To perform stage 1, we prepare the pre-trained 2D encoder $\mathcal{E}_{2D}$, followed by a bilinear upsampling layer to restore the reduced feature map to the original image resolution. For the 3D part, we prepare a randomly initialized 3D encoder $\mathcal{E}_{3D}$, followed by a 3D linear head $\mathcal{H}_{3D}$ to align the 3D feature dimension to the 2D one. For simplicity, we omit to note the bilinear upsampling layer and the 3D linear head. Since the 2D encoder $\mathcal{E}_{2D}$ is pre-trained on daylight images, we use only image-LiDAR pairs captured during the daytime to distill reliable features from the 2D encoder at this stage.

Description of stage 1. Let $P = [\mathbf{p}_1, ..., \mathbf{p}_N]$ and $I \in \mathbb{R}^{H \times W \times 3}$ denote the paired 3D LiDAR point cloud and 2D image, where $\mathbf{p}_i \in \mathbb{R}^3$ denotes the i-th 3D point, and H, W and N are the height and width of the image, and the number of points, respectively. We extract pixel-wise features from the 2D encoder $\mathcal{E}_{2D}$. Simultaneously, we extract point-wise features from the 3D encoder $\mathcal{E}_{3D}$. We then use pre-computed correspondence indices to match pixel-wise features with point-wise features, creating M pairs of matched pixel-wise features $\mathbf{G} = [\mathbf{g}_1, ..., \mathbf{g}_M]$ and matched point-wise features $\mathbf{F} = [\mathbf{f}_1, ..., \mathbf{f}_M]$, where $\mathbf{g}_i, \mathbf{f}_i \in \mathbb{R}^D$ denote the i-th matched features. For each pair of 3D feature $\mathbf{f}_i$ and 2D feature $\mathbf{g}_i$, we apply an image-to-LiDAR distillation loss to make the 3D feature $\mathbf{f}_i$ similar to 2D feature $\mathbf{g}_i$, defined as follows:

$$\mathcal{L}_{\text{I2L}} = \frac{1}{M} \sum\nolimits_{i \in M} \left\| \text{sg}[\mathbf{g}_i] - \mathbf{f}_i \right\|_2 , \tag{1}$$

where $\mathbf{f}_i$ and $\mathbf{g}_i$ are l_2-normalized and $\text{sg}[\cdot]$ stands for stop-gradient operator.

Through the image-to-LiDAR distillation, we expect the trained 3D encoder $\mathcal{E}_{3D}$ to extract the 3D feature $\mathbf{F}$ containing the 2D encoder $\mathcal{E}_{2D}$'s semantic context.

Correspondence matching. Given a calibrated relative pose between the LiDAR and the camera, the 3D-to-2D projection $\mathcal{T} : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ outputs the projected 2D coordinate on the image I, *i.e.*, $\mathbf{x}_i = \mathcal{T}(\mathbf{p}_i)$. Then, we collect the M pixel indices from all visible $\mathbf{x}_i$ in the image. Since pixel and point indices are paired by $\mathcal{T}$, we use these pairs to retrieve the corresponding pixel-wise and point-wise features. Finally, we obtain the M pixel-point feature pairs $\{\mathbf{g}_i, \mathbf{f}_i\}$.

3.2. Stage 2: LiDAR-to-Image Distillation

The goal of stage 2 is to transfer light-invariant and 3D-aware information from the 3D encoder $\mathcal{E}_{3D}$ to the 2D encoder $\mathcal{E}_{2D}$. Since 3D LiDAR data remains the same regardless of day or night, the extracted features $\mathbf{F}$ are expected to be light-invariant. Additionally, because the data is inherently 3D, the extracted features $\mathbf{F}$ are also 3D-aware. By leveraging the 3D feature, we can improve the 2D encoder $\mathcal{E}_{2D}$.

In stage 2, we continue to use the pre-trained 2D encoder $\mathcal{E}_{2D}$ and the 3D encoder $\mathcal{E}_{3D}$ trained in stage 1. At this stage, we use both day and nighttime images and LiDAR to allow the 2D encoder $\mathcal{E}_{2D}$ to learn to extract the light-invariant and 3D-aware features from day and night images.

Description of stage 2. We implement the LiDAR-to-image distillation by switching the gradient direction, reusing the weights of the 2D encoder $\mathcal{E}_{2D}$ and 3D encoder $\mathcal{E}_{3D}$ from stage 1 without adding any additional layer. The only difference is the LiDAR-to-image distillation loss, which is the same as image-to-LiDAR distillation loss $\mathcal{L}_{\text{I2L}}$, except that the stop-grad is applied to the point feature $\mathbf{f}_i$. The

LiDAR-to-image distillation loss is defined as follows:

$$\mathcal{L}_{\text{L2I}} = \frac{1}{M} \sum\nolimits_{i \in M} \|\text{sg}[\mathbf{f}_i] - \mathbf{g}_i\|_2 \,, \tag{2}$$

where $\mathbf{f}_i$ and $\mathbf{g}_i$ are l_2-normalized.

Through LiDAR-to-image distillation, we expect the further trained 2D encoder $\mathcal{E}_{\text{2D}}$ to extract 2D features $\mathbf{G}$ capturing LiDAR properties *i.e.*, light-invariance and 3D awareness, while preserving its semantic context due to stage 1.

Robustness to the lighting variation after our method. We visualize the features of the 2D encoder to validate the effects of our method. To examine that our 2D encoder $\mathcal{E}_{\text{2D}}$ can extract features robust to lighting conditions, we prepare day and night images from Cityscapes [15].[1] Figure 3 shows that our method generates robust feature maps across both day and nighttime conditions. In contrast, other methods generate feature maps that are heavily influenced by the pixel intensity of the given images. This demonstrates that our method successfully trains the model to extract the expected robust features of lighting conditions.

(a) Input image (b) DINOv2 (c) Ours

Figure 3. **Feature visualization**. Our feature maps (c) are more robust to lighting conditions compared to those from the encoder pre-trained with DINOv2 (b) that shows strong reactions to pixel intensity in both day and night images. The colors of features are obtained using principal component analysis (PCA).

4. Experiments

We first describe the experimental setup (Sec. 4.1), and then present results for in-domain and out-of-domain downstream tasks (Secs. 4.2 and 4.3). Finally, we provide ablation studies on the collaborative distillation method (Sec. 4.4).

4.1. Experimental Setup

Encoders. We use the WaffleIron-768 [52] as a 3D encoder and employ various pre-trained 2D encoders. Our primary models are ViT-S/14 to ViT-G/14, pre-trained by the DINOv2 [49] method. The above pre-trained models are fine-tuned on downstream tasks in our experimental setup. We conduct all ablation studies using the ViT-S/14 model.

Datasets. We pre-train all models on the nuScenes dataset [4], which contains 168K training and 36K validation images. Of these, 20K training and 3.6K validation images are night images. For in-domain experiments, nuScenes and nuImages [4] datasets are leveraged. For out-of-domain experiments, we follow the training and evaluation protocol of DINOv2 [49] with out-of-domain datasets such as KITTI [21], NYUd [59], CityScapes [15], and ADE20k [76]. For the multi-task experiment in night scenarios, we convert half of the training day images and all test day images of the Cityscapes-DVPS [54] into night images using the Stable Diffusion [55]-based image translation method [51].

Data augmentation. For the 2D encoders, we resize images to 224×448 in all stages. For the 3D encoder, we apply random z-axis rotation and xy-axis flipping in stage 1.

[1]For controlled fair comparisons, we use a Stable Diffusion [55]-based image translation [51] to synthesize the night scene corresponding to the day image due to the absence of the day-night pairs.

Hyperparameters. For stage 1, image-to-LiDAR distillation, we pre-train the WaffleIron [52] encoder using the AdamW [42] optimizer, setting weight decay to 3×10^{-2} and a learning rate that starts at 0, increases to 2×10^{-3}, and then gradually decreases to 10^{-5} following a cosine schedule. The batch size is 8, and we train for 49 epochs. For stage 2, LiDAR-to-image distillation, the hyperparameters are nearly identical to those in stage 1, with adjustments to batch size, learning rate, and epochs. The batch size is 32, with learning rates of 2×10^{-5} for ViT and 5×10^{-3} for ResNet50. The training epoch is set to 1. All pre-training is conducted on 4 NVIDIA A100 GPUs. Further details are provided in the supplementary material.

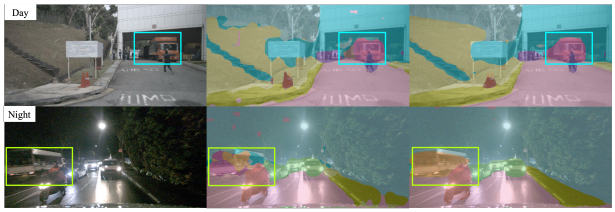
4.2. Transfer to In-Domain Downstream Tasks

We verify whether our method improves robustness to nighttime conditions and enhances 3D awareness on the validation set of the pre-training dataset *i.e.*, in-domain. Therefore, we compare the models pre-trained by DINOv2 with the models after Collaborative Distillation (CD) on in-domain downstream tasks, specifically semantic segmentation on both nuScenes [4] and nuImages [4], and depth estimation on nuScenes. Since the nuScenes dataset lacks 2D semantic segmentation labels, we project the 3D LiDAR semantic segmentation labels onto the images as ground truth. We use the entire training set for downstream tasks, and for the validation set, we split into day and night to check the model's robustness to lighting changes. Each 2D task-specific linear head replaces the pre-training linear head, mapping pixel-wise features to segmentation classes or depth values.

For semantic segmentation, we conduct 1% label fine-tuning to assess effectiveness in label-scarce scenarios and use full-label linear probing to evaluate the effectiveness of the learned representations. We follow the above training and evaluation protocol of [53]. Similarly, for depth estimation, we perform linear probing with all labels to evaluate the learned representations and further fine-tune with all labels to assess how closely the model approaches state-of-the-art robust depth estimation method [20]. We measure mean

Method	Arch.	nuScenes						nuImages	
		full		day		night		full	
		1% FT	100% LP	1% FT	100% LP	1% FT	100% LP	1% FT	100% LP
DINOv2	ViT-S/14	35.2	49.2	36.6	50.6	22.4	27.7	63.2	70.9
+ CD		**35.7 (+0.5)**	**51.9 (+2.7)**	**37.0 (+0.4)**	**52.9 (+2.3)**	**23.9 (+1.5)**	**33.4 (+5.7)**	**64.0 (+0.8)**	**71.7 (+0.8)**
DINOv2	ViT-B/14	39.0	52.3	40.7	53.4	24.5	33.8	70.4	74.9
+ CD		**39.8 (+0.8)**	**55.5 (+3.2)**	**41.4 (+0.7)**	**56.4 (+3.0)**	**26.5 (+2.0)**	**37.4 (+3.6)**	**70.7 (+0.3)**	**76.4 (+1.5)**
DINOv2	ViT-L/14	42.7	53.6	44.0	54.5	26.8	36.5	73.1	75.7
+ CD		**43.7 (+1.0)**	**57.6 (+4.0)**	**44.9 (+0.9)**	**58.4 (+3.9)**	**30.7 (+3.9)**	**42.0 (+5.5)**	**74.4 (+1.3)**	**77.9 (+2.2)**
DINOv2	ViT-G/14	44.4	55.1	45.8	56.1	29.4	37.6	75.0	77.7
+ CD		**47.1 (+2.7)**	**58.8 (+3.7)**	**48.2 (+2.4)**	**59.5 (+3.4)**	**33.3 (+3.9)**	**43.0 (+5.4)**	**75.8 (+0.8)**	**79.4 (+1.7)**

Table 1. **In-domain linear probing and few-shot fine-tuning performance for 2D semantic segmentation.** We compare models pre-trained by DINOv2 with those further improved by our method (CD) on in-domain datasets, nuScenes [4] and nuImages [4]. "Full" includes both day and night sets, while "FT" and "LP" represent fine-tuning and linear probing, respectively. The mIoU results show that our method improves performance across all metrics, including label-scarce and full-label settings for 2D semantic segmentation. The improvement in day images demonstrates that incorporating LiDAR properties helps the 2D representation preserve semantic context and make it more discriminative with 3D depth information. Additionally, the performance gain is especially significant at night compared to day data, demonstrating the effectiveness of our approach in enhancing the model's robustness under nighttime conditions.



(a) Input image (b) DINOv2 (c) Ours

Figure 4. **In-domain semantic segmentation at day and night**. We compare semantic segmentation quality on nuScenes for (a) input images, (b) DINOv2, and (C) our method. We present linear probing results using the ViT-G/14 model. The highlighted regions (cyan and light green boxes) show that our method achieves superior segmentation quality than DINOv2 in day and night.

Intersection over Union (mIoU) for semantic segmentation and Root Mean Squared Error (RMSE) for depth estimation.

Semantic segmentation. Table 1 shows that our method outperforms all mIoU metrics, with linear probing on all labels and fine-tuning with 1%. Interestingly, our method is effective in both label-scarce and full-label evaluations while improving performance across all metrics on day images. We assume that by incorporating LiDAR properties, the 2D representation preserves its semantics while becoming more discriminative, as it integrates 3D depth information to differentiate objects with similar appearances but different spatial depths [31, 37, 70]. Additionally, the improvement is larger at night than during the day, indicating that our method reduces the domain gap from lighting differences, enhancing the 2D encoder's robustness to low-light conditions. Our method's strength is further highlighted by the qualitative results in Fig. 4. The results show that our method produces more discriminative segments than DINOv2 in day and night. Our method enhances the 2D encoder's semantic context while remaining robust to lighting changes.

Depth estimation. Table 2 shows the fine-tuning and linear probing RMSE results using all labels on nuScenes. Our method demonstrates consistent performance improvements in day and night settings, suggesting that our approach enhances the 3D awareness of the 2D representation. Although we use only a simple linear layer and MSE loss for depth estimation training, our method achieves comparable performance to state-of-the-art approaches [20] focused on robust depth estimation relying on advanced methods [1].

4.3. Transfer to Out-of-Domain Downstream Tasks

In this section, we demonstrate that our method generalizes well to out-of-domain downstream tasks, including indoor environments, even though it is pre-trained on a nuScenes dataset constructed for outdoor autonomous driving. Therefore, we compare the models pre-trained by DINOv2, two 3D prior injection methods (FiT3D [73] and Condense [74]), and our CD method on out-of-domain downstream tasks. All training and evaluation protocols in this experiment follow the DINOv2 setup. Depth estimation is trained and evaluated on the outdoor KITTI [21] and indoor NYUd [59] datasets, while semantic segmentation is conducted on outdoor Cityscapes [15] and ADE20k [76], which consists of indoor and outdoor data. Note that FiT3D combines the model pre-trained by DINOv2 and the model improved by their method together for downstream tasks (*i.e.*, assembling). Since the FiT3D and Condense model weights are available, we reproduce the results in our experimental setup. In addition, we demonstrate the effectiveness of our 2D encoder in a multi-task learning setup, which jointly handles video

Method	Arch.	full (-)		day-clear (4.81‡)		day-rain (5.90‡)		night (6.37‡)	
		100% FT	100% LP	100% FT	100% LP	100% FT	100% LP	100% FT	100% LP
DINOv2	ViT-S/14	5.72	8.37	5.46	8.14	6.00	8.79	7.00	9.40
+ CD		**5.70**	**7.64**	**5.43**	**7.47**	**5.97**	**7.66**	**6.98**	**8.73**
DINOv2	ViT-B/14	5.46	8.01	5.20	7.86	5.62	8.26	6.87	8.76
+ CD		**5.42**	**7.18**	**5.17**	**7.01**	**5.58**	**7.19**	**6.76**	**8.28**
DINOv2	ViT-L/14	5.29	7.97	5.00	7.81	5.38	8.32	6.57	8.69
+ CD		**5.22**	**6.96**	**4.98**	**6.81**	**5.31**	**7.08**	**6.55**	**7.82**
DINOv2	ViT-G/14	5.18	7.66	4.93	7.49	5.27	8.09	6.55	8.33
+ CD		**5.14**	**6.63**	**4.91**	**6.47**	**5.22**	**6.89**	**6.49**	**7.40**

Table 2. **In-domain linear probing and fine-tuning performance for 2D monocular depth estimation.** We compare models pre-trained by DINOv2 with those further improved by our method (CD) on in-domain datasets, nuScenes [4]. The state-of-the-art performances from md4all [20] are denoted at the top of the Table with ‡. The RMSE results show that our method improves performance in all metrics for 2D monocular depth estimation, demonstrating that our approach enhances 3D awareness of 2D representation.

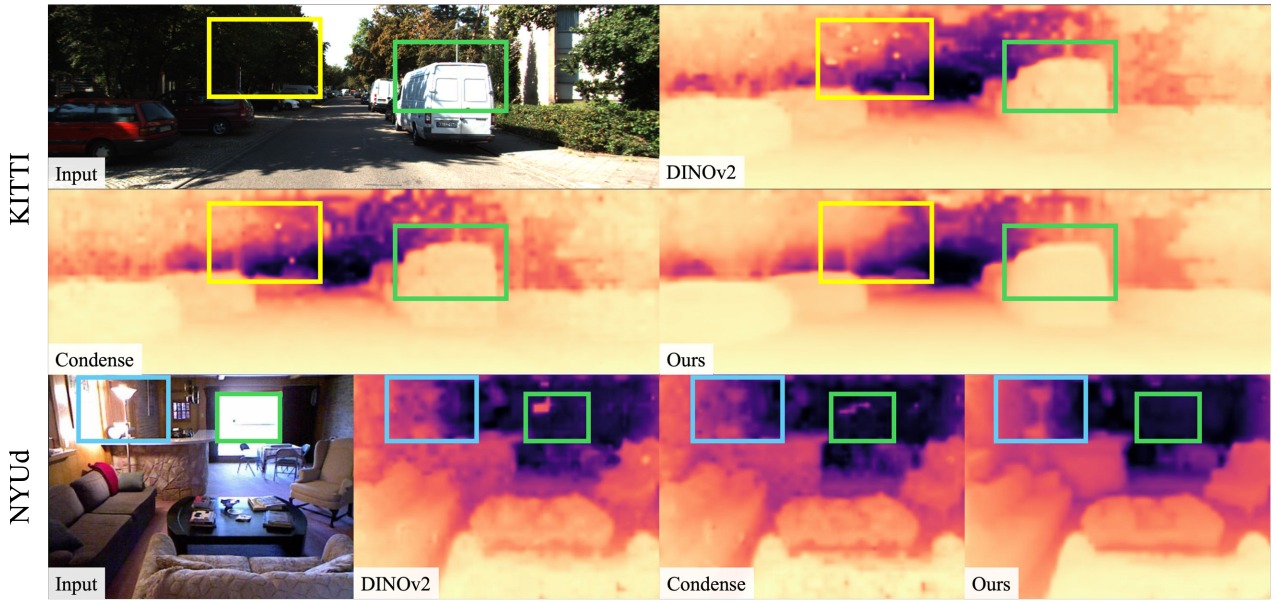
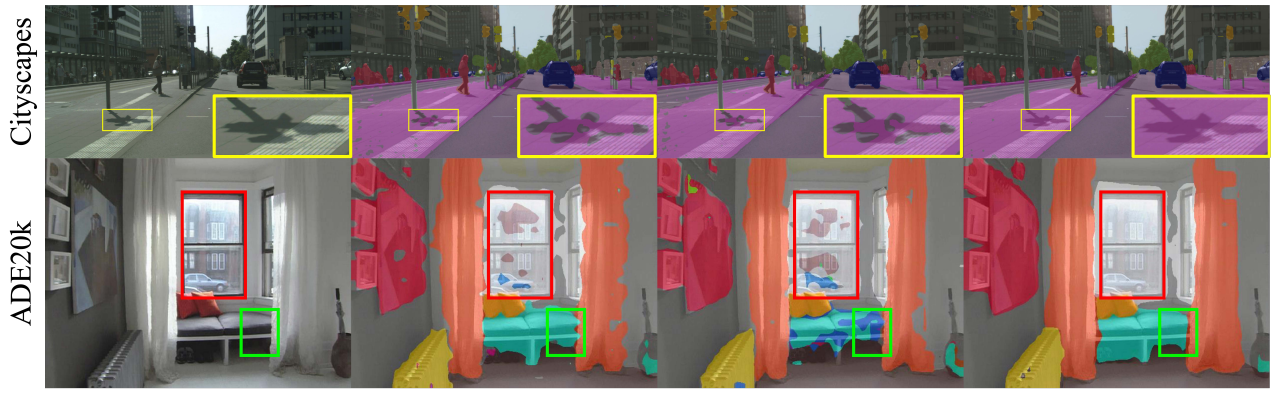


Figure 5. **Qualitative results of out-of-domain depth estimation**. To compare the depth estimation quality of the other methods, we visualize the results of the model pre-trained with DINOv2, improved by Condense [74], and by our method using the ViT-G/14 model. Our method produces clear depth with improved 3D awareness. Additionally, its robustness to lighting results in less noisy depth in the boxed areas. Although our method does not use indoor data in pre-training, it produces the clearest depth map among the models on the indoor NYUd dataset, demonstrating its general transferability.

panoptic segmentation and depth estimation. We adopt the recent framework [31], replacing its image encoder with ours. We train and validate the model with Cityscapes-DVPS [54] datasets. To evaluate the model on nighttime scenarios, we synthesize the night images using the Stable Diffusion [55]-based image translation method [51]. Details and additional experiments are in the supplementary material.

Depth estimation. Table 3 shows the linear-probing RMSE results on the KITTI [21] and NYUd [59] datasets. Our method achieves superior results across all metrics on both indoor and outdoor datasets (KITTI and NYUd). Considering our method is pre-trained solely on the outdoor dataset, nuScenes, this generalization ability is noteworthy. Note that FiT3D and Condense are pre-trained on indoor datasets [6, 16, 71], and FiT3D employs an assembling technique that integrates both the original DINOv2 and its own



(a) Input image (b) DINOv2 (c) Condense (d) Ours

Figure 6. **Qualitative results of out-of-domain semantic segmentation**. To compare the depth estimation quality of the other methods, we visualize the results of the model pre-trained with DINOv2, improved by Condense [74], and by our method using the ViT-G/14 model. Our method produces less noisy segmentation in light changes (see yellow box). Additionally, the results exhibit clear segmentation, thanks to improved semantic context (see red box and green box). Although our method does not use any indoor data in pre-training, it produces the clearest segmentation map among the models on the indoor data in the ADE20k dataset, demonstrating its general transferability.

Method	Arch.	Depth (RMSE ↓)		Seg. (mIoU ↑)	
		KITTI	NYUd	CityScapes	ADE20k
DINOv2	ViT-B/14	2.86	0.397	73.7	47.2
+ FiT3D† [73]		2.79	0.380	-	**49.5**
+ CD		**2.63**	**0.376**	**74.0**	47.9
DINOv2	ViT-G/14	2.60	0.339	73.6	47.9
+ Condense [74]		2.67	0.320	74.1	48.7
+ CD		**2.46**	**0.319**	**75.6**	**51.0**

Table 3. **Out-of-domain linear probing for 2D monocular depth estimation and semantic segmentation.** We compare models pre-trained with DINOv2 with models improved with FiT3D [73], Condense [74], or our method (CD) on out-of-domain datasets. The † means using the assembling technique, which is a strong competitor. For depth estimation RMSE results, our method outperforms other methods in the outdoor KITTI dataset [21] and indoor NYUd [59] dataset. For semantic segmentation mIoU results, our method outperforms other methods at the CityScapes [15] and ADE20k [76] except FiT3D† in ADE20k. These results demonstrate that our method improves 3D awareness of learned 2D representations and retains transferable semantic representations on out-of-domain.

model. Moreover, Fig. 5 shows that our estimated depths are clearer and remain robust under various lighting conditions. In contrast, other methods produce unintended noise in extremely dark or bright pixels.

Semantic segmentation. Table 3 shows the linear probing results on the CityScapes [15] and ADE20k [76] datasets. Our method achieves better mIoU than DINOv2 and Condense [74], demonstrating strong transferability to out-of-domain segmentation tasks. Furthermore, as seen in the in-domain semantic segmentation results in Sec 4.2, our method improves performance not only on day images but also in indoor scenarios. This demonstrates that it preserves the 2D encoder's semantic context while making it more discriminative with 3D depth information. The strength of our method is further supported by the qualitative samples in Fig. 6. In the top row, a dark shadow is present on the road in the input image. Our method discerns the light-invariant semantics (*i.e.*, the road), while other methods fail to capture the underlying semantics, being overly reliant on pixel intensity differences. The bottom row further highlights its ability to capture more discriminative semantics.

Method	Arch.	Video panoptic seg.	Depth
		VPQ ↑	RMSE ↓
DINO [5]	ResNet-50	33.7	5.23
+ CD		**34.2**	**4.42**

Table 4. **Linear probing results on multi-task learning.** We show the linear probing results on the night test images of the Cityscapes-DVPS [54]. We compute Video Panoptic Quality (VPQ) for video panoptic segmentation and Root Mean Square Error (RMSE) for depth estimation. The model [31] improved by our method (CD) outperforms DINO [5], confirming our method's multi-task adaptability and robustness in low-light conditions.

Multi-task learning. In Table 4, we report the multi-task learning performance in nighttime scenes using Video Panoptic Quality (VPQ) [34] for video panoptic segmentation and Root Mean Squared Error (RMSE) for depth estimation. The model with our 2D encoder outperforms the model with DINO [5] encoder in both metrics. As the light-invariant and 3D-aware characteristics are essential for depth-aware video panoptic segmentation task, our superior results imply that

Method	Seg. (mIoU ↑)	Depth (RMSE ↓)		
	nuScenes	nuScenes	KITTI	NYUd
DINOv2	49.2	8.37	2.99	0.443
DINOv2*	49.4	8.21	3.03	0.457
CD	**51.9**	**7.64**	**2.80**	**0.431**

Table 5. **Comparison of futher trained DINOv2 with our method.** We report the linear probing performance of our method and further trained DINOv2, denoted by *. The results show that further trained DINOv2 still significantly underperforms compared to our method, demonstrating that the superior performance of our method stems from specifically being crafted to improve the 2D encoder rather than just additional training on the nuScenes dataset.

Method	Stage 1	Stage 2	Seg. (mIoU ↑)		Depth. (RMSE ↓)
			nuScenes	nuImages	nuScenes
DINOv2	-	-	49.2	70.9	8.37
+ CD	✗	✓	32.2	39.2	9.53
+ CD	✓	✓	**51.9**	**71.7**	**7.64**

Table 6. **Performance with and without stage 1.** We report the linear probing performance of our method with and without stage 1. In the version without stage 1, stage 2 is applied with a 3D LiDAR encoder pre-trained through supervised learning on 3D semantic segmentation. The results show that our method with stage 1 significantly outperforms the one without it, demonstrating that stage 1 is essential for preserving semantic context in the collaborative distillation method.

such knowledge is well-distilled into the 2D encoder, and this 2D encoder can be adapted to joint multi-task learning.

4.4. Ablation Studies

Pre-training by CD vs. DINOv2 protocol. One may question whether the performance improvements reported in Section 4.2 stem primarily from additional training on the nuScenes dataset rather than from the merits of our method. To answer that question, we further train the pre-trained DINOv2 on the nuScenes image data, using the same training protocol of DINOv2. Table 5 shows that the further trained DINOv2 still significantly underperforms compared to our method. This result demonstrates the effectiveness of our method, which is specifically crafted to distill light-invariant and 3D-aware semantics from the 3D encoder. Details and further results are provided in the supplementary material.

Effectiveness of stage 1. The objective of our two-stage design is to ensure the 2D encoder preserves its original semantic context while learning LiDAR properties. To verify the effectiveness of stage 1, we present a baseline that directly injects the LiDAR information into the 2D image encoder without stage 1 (image-to-LiDAR distillation) and only with stage 2 (LiDAR-to-image distillation). For this, we train the 3D encoder from scratch on 3D semantic segmentation and then apply stage 2 without stage 1. Then, we compare this baseline with our method on in-domain downstream tasks. Table 6 shows that removing stage 1 results in a significant performance drop. This demonstrates that stage 1 is crucial for aligning 3D and 2D features to preserve the original semantic context in the proposed collaborative distillation. Without first aligning 3D features with 2D features, we fail to improve the 2D features effectively in stage 2, leading to a substantial performance drop.

Encoder	Stage 1 data	Eval. data	Seg. (mIoU ↑)
3D Encoder	nuScenes-Full	nuScenes-Night	52.7
	nuScenes-Day		**57.8**
2D Encoder	nuScenes-Full	nuScenes-Night	31.0
	nuScenes-Day		**33.4**

Table 7. **Performance of 2D and 3D encoder with and without night data in stage 1.** We report the linear probing performance of the 2D and 3D semantic segmentation tasks for the 2D and 3D encoders on the nuScenes dataset, respectively. The results show that using day-only data in stage 1 is important for distilling reliable image features into the 3D encoder, which also improves the 2D encoder in stage 2.

Impact of night data in stage 1. To investigate the effect of including night images in stage 1, we compare the semantic segmentation performance of the 3D encoder (after stage 1) and the 2D encoder (after stage 2) under two conditions: day-only vs. day+night training in stage 1 on the nuScenes dataset (see Table 7). The results indicate that constructing stage 1 with day images yields superior nighttime performance for both 3D and 2D encoders. This supports our hypothesis that the 2D encoder, pre-trained on day images, could provide unreliable semantic context for night images—hindering the training of the 3D encoder when night data are included at stage 1.

5. Conclusion

In this paper, we present a novel self-supervised collaborative distillation method to improve 2D image encoders under various lighting conditions and their 3D awareness. Our approach demonstrates overall improvements across day/night, in-domain/out-of-domain, and outdoor/indoor scenarios. These improvements are achieved through our carefully designed two-stage paradigm, which fully harnesses image–LiDAR pairs from a single outdoor driving dataset. The 2D image encoder pre-trained on a single dataset successfully transferring across diverse domains demonstrates its potential for generalization ability and adaptability in various vision-based autonomous systems. This success suggests future research directions, such as exploring its generalization ability to new environments and integrating it with different sensor modalities.

Acknowledgment

This project was supported by RideFlux, the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2025-25443318, Physically-grounded Intelligence: A Dual Competency Approach to Embodied AGI through Constructing and Reasoning in the Real World; No. RS-2024-00397663, Real-time XR Interface Technology Development for Environmental Adaptation, 33%), and was also supported by 'Ministry of Science and ICT' and NIPA ("HPC Support" Project).

References

[1] Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4009–4018, 2021. 5

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022. 2

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020. 4, 5, 6

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 7

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 6

[7] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018. 2

[8] Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7546–7554, 2021. 1

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020. 1, 2

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[12] Ye Chen, Jinxian Liu, Bingbing Ni, Hang Wang, Jiancheng Yang, Ning Liu, Teng Li, and Qi Tian. Shape self-correction for unsupervised point cloud understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8382–8391, 2021. 2

[13] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. *arXiv preprint arXiv:2211.09386*, 2022. 2

[14] 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. 2

[15] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 4, 5, 7

[16] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[18] Hang Dong and Timothy D Barfoot. Lighting-invariant visual odometry using lidar intensity imagery and pose interpolation. In *Field and Service Robotics: Results of the 8th International Conference*, pages 327–342. Springer, 2013. 1

[19] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024. 1, 2

[20] Stefano Gasperini, Nils Morbitzer, HyunJun Jung, Nassir Navab, and Federico Tombari. Robust monocular depth estimation under challenging conditions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8177–8186, 2023. 2, 4, 5, 6

[21] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361. IEEE, 2012. 4, 5, 6, 7

[22] 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. 1

[23] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. *arXiv preprint arXiv:2002.12319*, 2020. 1

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[26] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020. 2

[27] Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022. 2

[28] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15587–15597, 2021. 2

[29] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5693–5702, 2021. 2

[30] Ji Hou, Xiaoliang Dai, Zijian He, Angela Dai, and Matthias Nießner. Mask3d: Pre-training 2d vision transformers by learning masked 3d priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13510–13519, 2023. 2

[31] Kim Ji-Yeon, Oh Hyun-Bin, Kwon Byung-Ki, Dahun Kim, Yongjin Kwon, and Tae-Hyun Oh. Uni-dvps: Unified model for depth-aware video panoptic segmentation. *IEEE Robotics and Automation Letters*, 2024. 1, 5, 6, 7

[32] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30: 2340–2349, 2021. 2

[33] Wonjun Jo, Kwon Byung-Ki, Kim Ji-Yeon, Hawook Jeong, Kyungdon Joo, and Tae-Hyun Oh. The devil is in the details: Simple remedies for image-to-lidar representation learning. In *Proceedings of the Asian Conference on Computer Vision*, pages 3172–3188, 2024. 2

[34] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 7

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[36] Sohyun Lee, Jaesung Rim, Boseung Jeong, Geonu Kim, Byungju Woo, Haechan Lee, Sunghyun Cho, and Suha Kwak. Human pose estimation in extremely low-light conditions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 704–714, 2023. 2

[37] Xirui Li, Charles Herrmann, Kelvin CK Chan, Yinxiao Li, Deqing Sun, Chao Ma, and Ming-Hsuan Yang. A simple approach to unifying diffusion-based conditional generation. *arXiv preprint arXiv:2410.11439*, 2024. 5

[38] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *Advances in Neural Information Processing Systems*, 35:18442–18455, 2022. 2

[39] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021. 2

[40] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. *arXiv preprint arXiv:2306.09347*, 2023. 2

[41] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017. 2

[42] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[43] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *Bmvc*, page 4. Northumbria University, 2018. 2

[44] 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. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7102–7110, 2023. 2

[45] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[46] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020. 2

[47] Lucas Nunes, Rodrigo Marcuzzi, Xieyuanli Chen, Jens Behley, and Cyrill Stachniss. Segcontrast: 3d point cloud feature representation learning through self-supervised segment discrimination. *IEEE Robotics and Automation Letters*, 7(2): 2116–2123, 2022. 2

[48] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez,

Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 1, 2, 4

[50] Bo Pang, Hongchi Xia, and Cewu Lu. Unsupervised 3d point cloud representation learning by triangle constrained contrast for autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5229–5239, 2023. 2

[51] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024. 4, 6

[52] Gilles Puy, Alexandre Boulch, and Renaud Marlet. Using a waffle iron for automotive point cloud semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3379–3389, 2023. 4

[53] Gilles Puy, Spyros Gidaris, Alexandre Boulch, Oriane Siméoni, Corentin Sautier, Patrick Pérez, Andrei Bursuc, and Renaud Marlet. Three pillars improving vision foundation model distillation for lidar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21519–21529, 2024. 2, 4

[54] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4, 6, 7

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4, 6

[56] Yukihiro Sasagawa and Hajime Nagahara. Yolo in the dark-domain adaptation method for merging multiple models. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 345–359. Springer, 2020. 2

[57] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[58] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9891–9901, 2022. 2

[59] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 4, 5, 6, 7

[60] Jaime Spencer, Richard Bowden, and Simon Hadfield. Defeat-net: General monocular depth via simultaneous unsupervised representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14402–14413, 2020. 2

[61] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. 2

[62] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9782–9792, 2021. 2

[63] Kun Wang, Zhenyu Zhang, Zhiqiang Yan, Xiang Li, Baobei Xu, Jun Li, and Jian Yang. Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16055–16064, 2021. 2

[64] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2604–2612, 2022. 2

[65] Zeyu Wang, Dingwen Li, Chenxu Luo, Cihang Xie, and Xiaodong Yang. Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8637–8646, 2023. 2

[66] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022. 2

[67] 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. 2

[68] 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)*, pages 574–591. Springer, 2020. 2

[69] Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, Haoyi Zhu, Tong He, Shixiang Tang, Hengshuang Zhao, Qibo Qiu, Binbin Lin, et al. Unipad: A universal pre-training paradigm for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15238–15250, 2024. 2

[70] Hanrong Ye and Dan Xu. Taskprompter: Spatial-channel multi-task prompting for dense scene understanding. In *ICLR*, 2023. 5

[71] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 6

[72] Junbo Yin, Dingfu Zhou, Liangjun Zhang, Jin Fang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Proposalcontrast: Unsupervised pre-training for lidar-based 3d object de-

tection. In *European Conference on Computer Vision*, pages 17–33. Springer, 2022. 2

[73] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2D Feature Representations by 3D-Aware Fine-Tuning. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 5, 7

[74] Xiaoshuai Zhang, Zhicheng Wang, Howard Zhou, Soham Ghosh, Danushen Gnanapragasam, Varun Jampani, Hao Su, and Leonidas Guibas. Condense: Consistent 2d/3d pre-training for dense and sparse features from multi-view images. In *European Conference on Computer Vision*. Springer, 2024. 2, 5, 6, 7

[75] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *IEEE International Conference on Computer Vision (ICCV)*, pages 10252–10263, 2021. 2

[76] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 4, 5, 7

[77] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022. 1, 2

[78] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[79] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020. 1