CLLAP: Contrastive Learning-based LiDAR-Augmented Pretraining for Enhanced Radar-Camera Fusion

Supplementary Material

1. Overview

The appendix offers comprehensive explanations of the methodologies introduced in the main text, together with additional experimental results and extended visual analyses. The supplementary material is organized into the following sections:

2. Methodology Supplement

2.1. Sliding Window Feature Matching Mechanism

Cross-modality feature misalignment presents a significant challenge in multi-modal contrastive learning for radar-camera fusion, as naively treating spatially corresponding features as positive pairs often results in suboptimal alignment. To address this limitation, we proposed a method for finding the best positive sample pairs. Specifically, for each anchor $\mathbf{f}_t^j$, We define a local search area with a width of R in the query feature map and use a sliding window of size r (r<R) to slide through the local search area to extract n ($n = R - r + 1$) candidate areas $\mathbf{f}_s^m$. Each candidate area is adaptively aggregated through an attention module $Atten$ to align with the anchor's dimensions, followed by pairwise similarity computation. The highest-scoring candidate $\mathbf{f}_s^{j*}$ is selected as the refined positive pair, effectively replacing coarse positional alignment with a learned, similarity-driven matching strategy. This approach enhances feature correspondence precision while maintaining robustness to spatial discrepancies between pseudo-radar and camera modalities:

$$\mathbf{f}_s^{j*} = \max_{\delta \in \{-R,\dots,R\}} \left(\text{sim}\left(\mathbf{f}_t^j, Atten\left(\left\{ \mathbf{f}_s^m \mid |m - (j+\delta)| \leq \frac{r}{2} \right\} \right) \right) \right) \tag{4}$$

where $\text{sim}(\cdot,\cdot)$ denotes similarity, m is the center of the m^{th} area during the sliding process, and δ is the offset.

2.2. BCSA Module

To further enhance the discriminative power of these refined positive pairs, we introduce a Bidirectional Channel-Spatial Attention module. Within the constructed positive pairs, the channel-wise and spatial distribution information of the feature maps convey distinct semantic meanings. Our module independently applies attention mechanisms along the channel and spatial dimensions:

$$\text{MAT}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{5}$$

where $Q = \mathbf{F}_i \in \mathbf{R}^{B\times D\times C}$ or $\Gamma(\mathbf{F}_i) \in \mathbf{R}^{B\times C\times D}$ is the query, $K,V = \mathbf{F}_{3-i} \in \mathbf{R}^{B\times D\times C}$ or $\Gamma(\mathbf{F}_{3-i}) \in \mathbf{R}^{B\times C\times D}$, and D, C represent the spatial and channel dimensions, respectively. (1) *Channel Attention* learns importance weights for different feature channels, highlighting those most discriminative for the 3D object detection task. (2) *Spatial Attention* learns importance weights for different spatial locations within the feature column, emphasizing the importance of the region of interest in object detection while suppressing interference from background regions, thus enhancing the alignment of pseudo-radar and camera features. Through this bidirectional decoupling and weighting across channel and spatial dimensions, the final refined feature representation is computed as:

$$\begin{aligned} \mathrm{F}_i^{\text{final}} = \mathcal{G}\Big(&\text{LN}\left(\text{MAT}(\mathbf{F}_i, \mathbf{F}_{3-i}, \mathbf{F}_{3-i})\right), \\ &\text{LN}\left(\Gamma\left(\text{MAT}(\Gamma(\mathbf{F}_i), \Gamma(\mathbf{F}_{3-i}), \Gamma(\mathbf{F}_{3-i}))\right)\right)\Big) \end{aligned} \tag{6}$$

where $i \in \{1,2\}$ indexes the feature pair, $\mathcal{G}$ is a gate control mechanism, LN denotes layer normalization, and Γ represents the transposition operation.

The following is the explanation of why columns are selected instead of rows in the local contrastive loss part of the two-stage dual-modal comparative learning strategy in the proposed method paragraph.

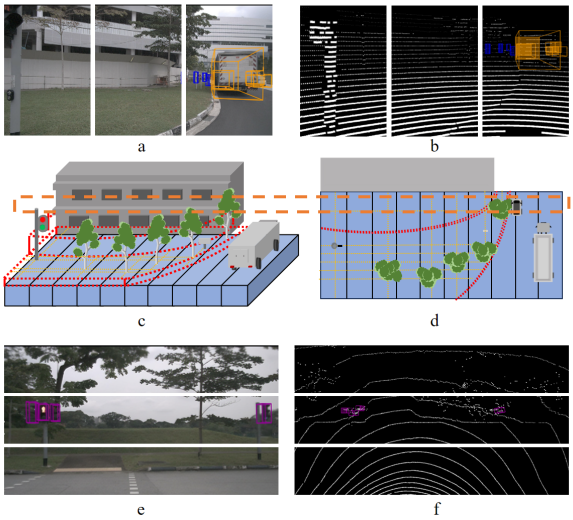

Figure 7. Figures (a) and (b) present the visualizations of the image and point cloud from the front view, respectively. Figures (c) and (d) illustrate simulated representations of the same modality observed from different viewpoints. Figures (e) and (f) depict visualizations of different modalities under varying viewpoints.

2.3. Column Selection in Local Contrastive Loss

As illustrated in Fig. 7, Figures (a) and (b) indicate that both modalities convey consistent content when evaluated within the same column unit. Figures (c) and (d) demonstrate that column-based units maintain consistent content representation across views, whereas row-based units introduce noticeable inconsistencies, as highlighted by the yellow dashed rectangles. Figures (e) and (f) reveal that row-based optimization units exacerbate inconsistencies in content representation across modalities and views.

3. Visual supplementation

3.1. Visualization of Experimental Results

Figure 8 provides a visual comparison between the results produced by our proposed method and those generated by the CRN baseline. The green solid rectangle denotes the ground truth bounding box, the red dotted rectangle represents the prediction from the baseline model, and the blue dotted rectangle corresponds to the prediction obtained using our method. As observed, the predictions produced by our approach exhibit greater alignment with the ground truth, demonstrating improved localization accuracy compared to the baseline. This visual evidence highlights the effectiveness of our method in enhancing detection performance.

3.2. Global Contrastive Loss

As illustrated in Fig. 9, the left panel shows the inconsistency map produced by CLLAP without the Global Contrastive Loss, while the right panel shows the result after adding this loss. In both maps, brighter colors indicate a larger discrepancy between each feature and its local neighborhood. Without the global supervision, high-response regions are scattered across both foreground objects and background areas, and even the interior of the target exhibits many fragmented, high-valued blobs, suggesting unstable and noisy representations. After introducing the Global Contrastive Loss, the responses become much more compact: large values concentrate around object boundaries, whereas the interior of the object and most background regions are strongly suppressed. This indicates that the proposed loss encourages globally coherent, object-level features and clearer separation between foreground and background, which aligns with the quantitative gains in detection

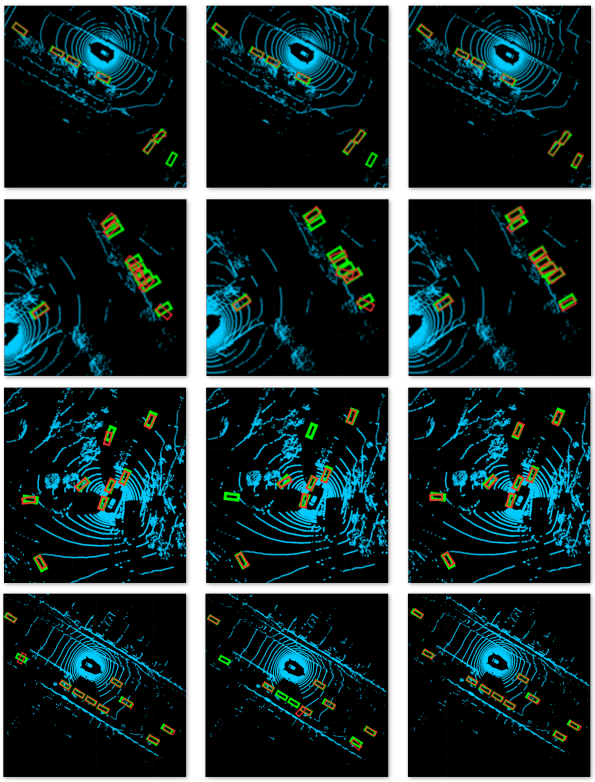

Figure 8. Comparison of our method with baseline visualization results.

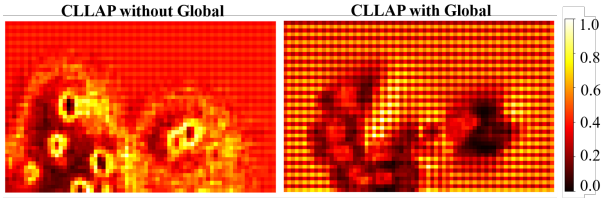

Figure 9. Comparison of feature heatmaps before and after the application of Global Contrastive Loss is presented.

accuracy reported in Sec. 4.4.

4. Supplementary Experiments

Implementation Settings. Our proposed model is implemented using the *PyTorch* framework and trained on *NVIDIA GeForce RTX 4090* and *NVIDIA H800 Tensor Core GPUs*. We adopt the *SGD* optimizer with a learning rate of 2×10^{-4}, momentum of 0.9, and weight decay of 1×10^{-5}. The batch size is set to 6 during pretraining.

(a) Rain Scene **(b) Snow Scene**

Figure 10. Adverse Weather example

4.1. Generalization to adverse weather

To further assess whether the proposed pretraining improves the robustness of the learned representations, we evaluate the models on adverse-weather scenarios without any additional training. Following the corruption protocol of the CVPR 2023 work[6], we synthesize adverse-weather data by injecting noise and weather-related corruptions into the original nuScenes samples; representative examples are shown in Fig. 10. We then directly apply the models pretrained with our method and the corresponding baselines to this corrupted benchmark, using exactly the same weights as on the clean nuScenes validation set. As summarized in Tab. 7, our pretrained models consistently outperform their non-pretrained counterparts under adverse-weather conditions, suggesting that the learned features generalize beyond the clean training distribution and confer improved robustness to realistic corruptions.

4.2. Impact of the weighting parameter in the loss function

To study how the balance between the global and column-wise contrastive terms affects performance, we ablate the weighting factor λ in the loss:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{global}} + \mathcal{L}_{\text{col}}. \tag{7}$$

For efficiency, all models are trained on a randomly sampled 20% subset of the nuScenes training set. As reported in Tab. 8, we evaluate $\lambda \in \{1, 1/6, 1/12\}$. Setting $\lambda = 1/6$ yields the best detection accuracy, indicating that a moderate weight on the global term provides a good trade-off between global regularization and the local column-wise supervision. In contrast, $\lambda = 1$ over-emphasizes the global loss and slightly degrades performance, while $\lambda = 1/12$ under-weights it and limits the benefit of global contrastive learning. Therefore, we adopt $\lambda = 1/6$ as the default in all experiments.

Table 7. Performance under adverse-weather corruptions on the nuScenes validation set. Models are directly evaluated on corrupted data without additional training.

Weather	Baseline		w. CLLAP (ours)	
	mAP (↑)	NDS (↑)	mAP (↑)	NDS (↑)
Rain	32.84	46.57	**34.39**	**47.09**
Snow	13.93	33.40	**15.45**	**34.48**

4.3. Impact of the GMM parameter variations

Additionally, to verify that GMM parameter variations have a minimal impact on our framework, we evaluated GMM with 4, 5, and 6 components by Chamfer distances(a measure of similarity between two sets):

$$CD(P,Q) = \frac{1}{|P|}\sum_{p\in P}\min_{q\in Q}|p-q|^2 + \frac{1}{|Q|}\sum_{q\in Q}\min_{p\in P}|p-q|^2 \tag{8}$$

where P/Q denote the pseudo-radar/real-radar points. The results shown in Tab. 9 reveal that the setting of GMM parameters had little effect on the results.

4.4. Impact of the weighting parameter in sampling

In the paper, the intensity weight w_{int} reflects the relative contribution of each point based on its point intensity. To ensure points with greater reflective intensity are properly emphasized during sampling, greater weights are assigned to them. The intensity weight for each point is calculated as $w_{\text{int}} = I_i^{\frac{1}{2}} / \sum_j I_j^{\frac{1}{2}}$, where I_i is the intensity of point i.

To ensure accurate feature representation, we assign weights based on point sparsity and distance. Sparse regions require greater sampling to capture their structural and semantic information, quantified by the sparsity weight $w_{\text{spa}} = \sum_j D_{ij}^2$, where D_{ij} is the Euclidean distance between point i and its j-th nearest neighbor. Similarly, points

Table 8. The result of different weight settings in the loss factor (using 1/5 of the nuScenes dataset).

weight	mAP(↑)
1	35.13
1/6	**35.74**
1/12	35.29

Table 9. The result of different component settings in GMM.

components	Chamfer Distances(↓)
4	142.0
5	140.7
6	141.2

Table 10. The result of different sampling weight configurations.

Config	α_{int}	α_{dist}	α_{spa}	Chamfer Distances(↓)
1	4	2	4	**114.1**
2	1	1	1	116.6
3	1	2	1	118.6
4	2	1	1	115.3
5	1	1	2	115.5

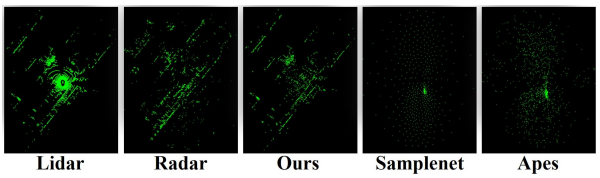

Figure 11. Visualization Comparison

farther from the center of the point cloud, often sparse in radar sensing, are assigned a distance weight $w_{\text{dist}} = 1/D_{iO}^2$, where D_{iO} is the Euclidean distance from point i to the origin. Finally, the overall sampling weight is a linear combination of the three individual weights, with scaling factors α_{int}, α_{spa}, α_{dist} controlling the contributions of each weight: $w_{\text{final}} = \alpha_{\text{int}} w_{\text{int}} + \alpha_{\text{dist}} w_{\text{dist}} + \alpha_{\text{spa}} w_{\text{spa}}$. By normalizing w_{final} for all points, we ensure the sampling probabilities are well-defined. Furthermore, we conducted a sensitivity analysis of the weight using 2000 nuScenes point clouds with 5 settings: 4:2:4, 1:1:1, 1:2:1, 2:1:1, and 1:1:2. The results in Tab. 10 show that the chosen weight of 4:2:4 offers the best performance and the model is relatively robust to weight variations.

4.5. Comparison with Learned Pseudo-Radar Baselines

We further compare the proposed L2R sampling strategy with two learning-based pseudo-radar generation baselines, namely SampleNet[16](*CVPR 2020*) and APES[33] (*CVPR 2023*). As shown in Fig. 11, although these methods achieve lower Chamfer Distance values (SampleNet: `20.9`, APES: `42.0`), their generated point patterns are less consistent with the spatial characteristics of real radar observations. In contrast, the pseudo-radar point clouds produced by our L2R sampling strategy more faithfully preserve radar-like spatial distributions.

To further evaluate their utility for downstream pretraining, we incorporate SampleNet-generated pseudo-radar point clouds into the same pretraining pipeline used in our method. This variant improves the baseline by `+1.59` mAP and `+1.32` NDS, but still remains clearly inferior to our L2R-based pretraining, which achieves `+3.20` mAP and `+1.87` NDS. These results suggest that a lower Chamfer Distance alone does not necessarily translate into more effective pseudo-radar supervision for radar-camera fusion pretraining. More importantly, they demonstrate that the proposed L2R sampling strategy provides pseudo-radar data that are more beneficial for downstream 3D object detection.