Camera- and LiDAR-based Person Re-identification

Sebastian Krebs[1,2] and Dariu M. Gavrila[1]

***Abstract*— In this paper, we introduce a novel method for creating appearance embeddings to identify individual persons using an object re-identification (ReID) framework. We present CLFormer (Camera LiDAR Transformer), a transformer-based architecture that incorporates multi-modal data from both camera and LiDAR sensors. We introduce the 3D Cuboid-Inclusive Point Embedding (3D-CIPE), which leverages rich data from LiDAR point clouds and 3D cuboids to add a learnable embedding into the transformer structure. Additionally, through ablation studies, we explore and analyze various strategies for the early and late fusion of multi-modal input data.**

To evaluate our proposed CLFormer, we reinterpret the nuScenes dataset [1] for ReID purposes and use it for our experiments. Our method demonstrates a significant improvement in performance, outperforming the image-only baseline with an increase of 2.3 in mean Average Precision (mAP).

2025 IEEE Intelligent Vehicles Symposium (IV) | 979-8-3315-3803-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/IV64158.2025.11097607

I. Introduction

In the intelligent vehicles domain the knowledge of the vehicles surroundings and other traffic participants is a key building block to allow safe and reliable applications, like driver assistance systems or autonomous driving. These tasks are generally addressed as part of the environment perception, using the input of single or multiple sensors such as cameras, LiDARs, and radars. In this paper, we focus on vulnerable road users (VRUs), such as pedestrians and riders, who are of particular interest due to their vulnerability and dynamic behavior. A typical perception pipeline includes the detection of objects of interest in the raw sensor data, followed by subsequent tracking. Tracking methods are employed to mitigate errors in the detection process and allow the integration of information over time. Tracked objects enable subsequent processing steps, such as intention or motion prediction [2]. Numerous tracking methods exist, among which the tracking-by-detection paradigm stands out as one of the best performing and commonly adopted techniques [3]. As the name suggests, detections are associated over time with the set of existing tracks. For association, a similarity measure between the detection and track is required. For the tracking of persons using a learned appearance embedding as part of the similarity measure has shown to be a good performing choice [4], [5], [6], [7].

Person re-identification (ReID) focuses on generating such appearance embeddings to uniquely identify individuals, primarily in surveillance settings with multiple overlapping static cameras. However, advances in this field can be directly applied to tracking methods.

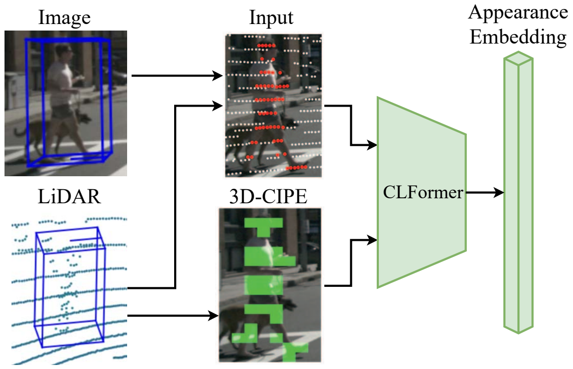


Fig. 1: Schematic depiction of our proposed novel multi-modal person ReID method. We use LiDAR point clouds, camera images, and 3D cuboids. Aligned image and LiDAR crops are used as input to generate an appearance embedding for each person. LiDAR points located within the annotated 3D cuboid are used to generate the 3D Cuboid-Inclusive Point Embedding (3D-CIPE) which is added as embedding to the transformer.

The combination of LiDAR and camera information has recently shown great benefits for detection methods [8]. However, extending person ReID approaches to multi-modal information is a little explored field. LiDAR sensors are available in most intelligent vehicle settings [1], yet there are only a few works that use this information exclusively [9], [10] and do not combine it with image information.

This paper presents a novel person ReID method, which combines image and LiDAR information in an intelligent vehicles setting, as shown in Figure 1. We introduce and investigate different fusion strategies for combining the two modalities to enhance our transformer-based person ReID method. While LiDAR data may not provide detailed appearance information, such as clothing or hair color, it can effectively guide visual attention. Additionally, the patch-based nature of the transformer architecture facilitates an implicit part-based approach, which is beneficial for handling partial occlusions. To better utilize the multi-modal information available to our method, we make use of additional auxiliary tasks. Lastly, we present a novel way to integrate object-centered information based on the LiDAR point clouds and available 3D cuboids into the transformer. We extract whether an image region contains a LiDAR point from within the 3D cuboid to create an abstract object mask and embed this information into the transformer. We coin the resulting method CLFormer, which combines the best-

Corresponding author: sebastian.krebs@mercedes-benz.com
[1] Intelligent Vehicles Group at TU Delft, Netherlands
[2] Perception and Maps Department at Mercedes-Benz AG, Germany

performing fusion strategy - channel concatenation - with our 3D Cuboid-Inclusive Point Embedding (3D-CIPE). To benchmark CLFormer, we repurpose the nuScenes dataset [1] for person ReID to validate our approach and compare it to baseline and other ReID methods. The contributions of our work are summarized as follows:

- To the best of our knowledge, we are the first to present a multi-modal transformer-based person ReID method that uses LiDAR in addition to camera information to learn more robust appearance embeddings.
- We introduce the 3D Cuboid-Inclusive Point Embedding (3D-CIPE) module, which embeds object-centered information from LiDAR point clouds and 3D cuboids into the transformer architecture.
- We propose and investigate various fusion strategies to effectively combine LiDAR and image sensor data within our ReID method.

II. Related Work

This section reviews recent advancements in person ReID, with a focus on deep learning applications and the integration of multi-modal data. For foundational methods, readers may refer to the comprehensive survey by Zheng et al. [11]. Current strategies predominantly leverage deep learning to extract unique person features through positive and negative sample training, often integrating triplet loss [12] with cross-entropy loss [13] for metric learning.

CNN-based methods have been extensively explored, including notable approaches such as BoT [14], SBS [15], AGW [16], and MGN [17]. BoT [14] set a new baseline by incorporating a series of enhancements into the training strategy, which is further improved by SBS [15]. AGW [16] extends BoT by integrating attention mechanisms into the CNN architecture, while MGN [17] extends the CNN to a multi-branch network structure to simultaneously learn global and local features. Recent research has increasingly focused on addressing occlusions, a major challenge in the field, which introduce non-object appearances into the learning process. To mitigate this, data augmentation-based approaches [18] train networks to recognize and disregard occluded regions by artificially occluding parts of the image during training. Other approaches segment the person patch into several parts and learn local part-based representations [19], [20], [21], [22]. These parts can be created using semantic networks [22], information from a pose estimation network [20], [21], or predefined static areas such as a fixed set of rows [19]. Part-based approaches generate a feature embedding for each part individually, while also extracting visibility [20], [21], [22] or quality [19] scores. Other techniques address occlusion by synthesizing complete images from sequential data [23], using human shape as additional supervision [24], or by excluding occluded samples during training [25].

Transformer-based methods have recently gained traction in person ReID, with models such as TransReID [26], AAFormer [27], Performer [28], DC-Former [29], LoGoViT [30] adapting the Vision Transformer (ViT) architecture [31] to this task. TransReID [26] presents the first ViT-based ReID framework incorporating enhancements such as overlapping patches and embeddings that capture orientation or camera-specific information. Various extensions have been proposed [26], [27], [29], [30] to improve the models performance and its ability to handle partially occluded persons. TransReID [26] utilizes patch tokens in different permutations in addition to the class token, while LoGoViT [30] processes the patch tokens through an additional transformer layer. DC-Former [29] increases the number of class tokens, whereas AAFormer [27] introduces new part tokens.

Multi-modal and LiDAR-only methods: The integration of LiDAR or depth data into ReID is relatively uncharted. Early techniques employed RGB-D data (image and depth) to derive anthropometric features [32], [33], [34], or utilized skeleton data from the Kinect RGB-D sensor [35], [36]. More recent transformer-based approaches have merged RGB with depth data [37], [38] or other auxiliary inputs like depth or contour plots [39]. In addition to depth data, a body of work has focused on multi-modal ReID using RGB and thermal imagery [40], [41], [42]. These modalities can be processed separately [40], [42] or jointly by a common backbone [41]. Information across modalities is shared using token permutation [40], spatial- and frequency-based token selection and aggregation [41], or by an unsupervised collaborative learning strategy utilizing deep and shallow features [42]. Recently, two methods have been presented which focus on solely using LiDAR points for re-identification [9], [10]. [10] presents a graph-based complementary enhancement encoder to extract features from multiple point clouds followed by a transformer-based temporal fusion to estimate the final ReID features. In [9] a Siamese network tracker is extended to generate ReID features for vehicles and pedestrians using the point cloud input.

Despite advancements in multi-modal ReID, the potential of LiDAR data remains underutilized. While prior work has explored depth, infrared, or camera-only modalities, to the best of our knowledge, no method has yet combined LiDAR and camera data for person ReID. LiDAR provides rich spatial information that can significantly enhance ReID performance, especially in occluded scenarios. Our work addresses this gap by incorporating LiDAR data into our transformer-based approach, leveraging its spatial context to guide attention mechanisms.

III. Methodology

In this section, we explain our proposed appearance embedding network for person re-identification (ReID) (see Figure 2 for an overview). Our method transforms input sensor data crops into a learned latent appearance space. The distance between two appearance embeddings should be minimal for the same person and maximal for different persons.

A. Camera-Only Baseline

Our camera-only ReID baseline uses a transformer-based architecture similar to [26], [29], [30]. Building on the Vision

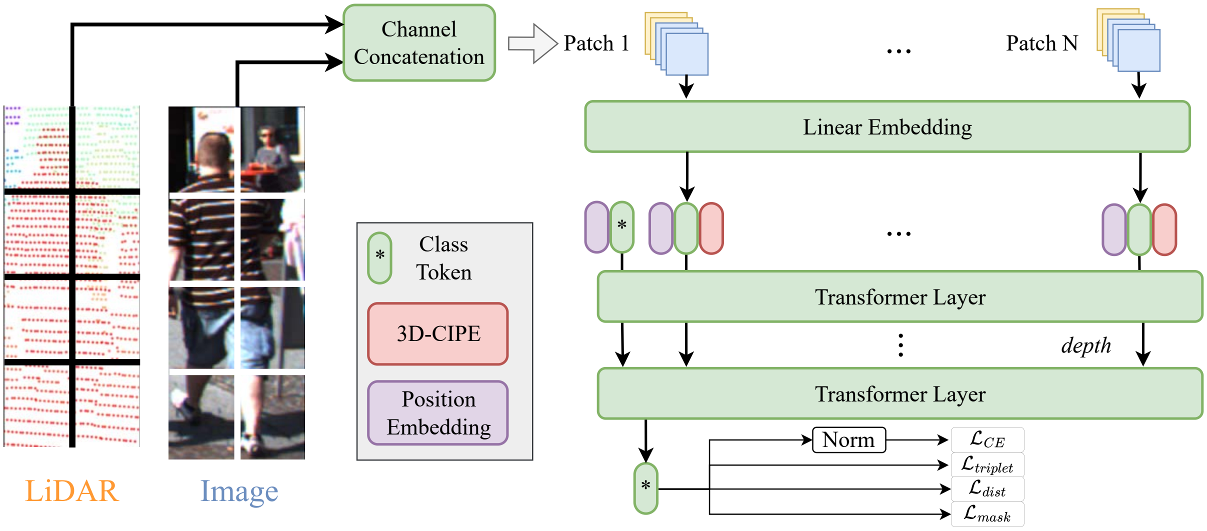


Fig. 2: Our proposed CLFormer method for multi-modal person re-identification. Each input crop is split into N patches, which are embedded using an initial 2D convolution, augmented by a learnable position embedding (purple box). At this stage, we also add our 3D Cuboid-Inclusive Point Embedding (3D-CIPE, see red box). Multi-modal support is added by concatenating the pre-processed LiDAR input to the image crop (see yellow and blue patches at the top). The final class token after the last transformer layer is used as appearance embedding and to compute the losses.

Transformer (ViT) [31], each person crop is resized to a fixed size, resulting in the input $x \in \mathbb{R}^{H \times W \times C}$. The input crop is divided into N patches, each projected into a feature vector of dimension D using a 2D convolution $\mathcal{F}(x)$. To retain spatial information, we augment each patch with a learnable position embedding and add a class token x_{cls} at the beginning of the feature vector for the initial layer:

$$\mathcal{X}_0 = [x_{cls},\ \mathcal{F}(x_1^p),\ \ldots,\ \mathcal{F}(x_N^p)] + \mathcal{P}, \tag{1}$$

with x_i^p being the i-th patch, $i \in [1, N]$, and $\mathcal{P} \in \mathbb{R}^{N+1 \times D}$ (one position embedding for each patch, plus the class embedding). This feature representation is processed by l transformer encoder layers, each consisting of normalization, multi-head self-attention, another normalization, and an MLP layer. The features after the last transformer encoder are normalized and used as input for the ReID tasks.

The ReID task is defined using two losses. First, we define each unique identity in the training set as an individual class and train the network using the cross-entropy loss $\mathcal{L}_{CE}$, encouraging clustering of appearance vectors for identical identities. Second, we use the triplet loss $\mathcal{L}_{triplet}$ as in [12], refining the networks ability to distinguish between crops of the same and different classes. Our final loss is calculated using the weights λ_{CE} and $\lambda_{triplet}$ with:

$$\mathcal{L}_{total} = \lambda_{CE} \cdot \mathcal{L}_{CE} + \lambda_{triplet} \cdot \mathcal{L}_{triplet}. \tag{2}$$

B. 3D Cuboid-Inclusive Point Embedding (3D-CIPE)

To enhance our model with additional 3D information from LiDAR points and 3D cuboids, we introduce the concept of 3D-CIPE. For each LiDAR point, we determine whether it is contained within the 3D cuboid of the object. This binary information is compiled for each patch x_i^p processed by our transformer-based model. Patches that contain at least one LiDAR point identified as being inside the object are labeled as *object* patches. The remaining patches are considered as *non-object* patches. A visual representation of this can be seen in Figure 1, where LiDAR points located inside the objects 3D cuboid are depicted in red, while the resulting patch-based object mask (i.e., the *object* patches) is shown in green.

We extend our architecture, by introducing an object mask embedding. A learnable parameter is added to each patch to differentiate between *object* and *non-object* patches. Specifically, we extend Equation 1 to:

$$\mathcal{X}_0 = [x_{cls},\ \mathcal{F}(x_1^p) + \omega_1,\ \ldots,\ \mathcal{F}(x_N^p) + \omega_N] + \mathcal{P}, \tag{3}$$

with

$$\omega_i = \begin{cases} \omega^O, & \text{if } x_i^p \text{ is } \textit{object} \text{ patch}, \\ \omega^B, & \text{if } x_i^p \text{ is } \textit{non-object} \text{ patch}. \end{cases} \tag{4}$$

Here, ω^O and ω^B are the learnable parameters ($\mathbb{R}^D$) for object and non-object patches, respectively. The augmented feature representation $\mathcal{X}_0$ is then fed into the subsequent transformer layers. Unlike typical positional encodings, which are purely spatial or learned, 3D-CIPE captures whether a patch is physically associated with a 3D object. This information complements visual appearance by providing a binary cue of geometric relevance, enabling the model to better localize and represent objects in cluttered or occluded environments. The resulting enriched token embeddings are passed unchanged into the transformer encoder, allowing the attention layers to leverage this geometric prior implicitly during representation learning.

C. Extension to Multi-Modal Input

We extend the image-only baseline by incorporating multi-modal sensor data from camera and LiDAR. In the following, we detail the necessary pre-processing and propose three distinct fusion strategies.

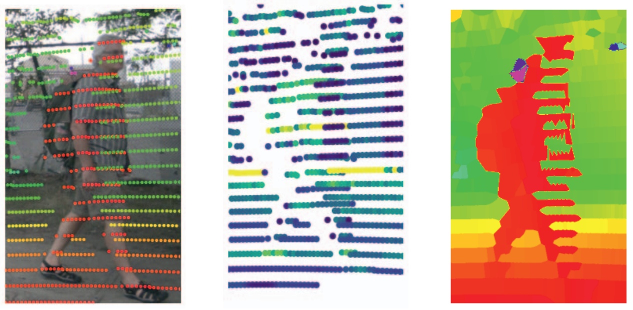
To spatially align the sensors, we project LiDAR points into the cameras image plane. We first select the LiDAR point cloud closest in time to the camera frame. Each point is transformed into the camera coordinate system using extrinsic calibration (rotation and translation), yielding 3D coordinates (X, Y, Z). These are then projected onto the 2D image plane using the cameras intrinsic parameters, resulting in image coordinates (u, v). Each point now has associated (u, v), its 3D camera coordinates (X, Y, Z), and LiDAR intensity I. Points outside the image crop are discarded. We refer to the Z component as the points *depth* in the remainder of this paper. Figure 3 provides a visual representation of this procedure. Given the sparsity of the LiDAR data, we apply nearest-neighbor interpolation, as shown in Figure 3c.

We propose three distinct fusion strategies to integrate this multi-modal data into our network as shown in Figure 4. The first strategy, channel concatenation, involves concatenating the LiDAR data as additional channels to the input image crop (c.f., variable "C" in Sec. III-A), requiring the first convolution layer to process both modalities jointly. The second strategy uses separate embedding layers for each modality, splitting $\mathcal{F}$ into $\mathcal{F}_{im}$ for the image and $\mathcal{F}_L$ for the LiDAR data. The resulting feature vectors can be fused by summation, multiplication, a fully connected layer, or concatenation. For both fusion strategies, the resulting embedded vector $\mathcal{X}_0$ is used as input for the transformer encoders, aligning with the camera-only architecture. The third strategy, cross-modal attention fusion, employs separate transformer backbones to embed the camera and LiDAR data independently. A cross-attention head is used to exchange information between the final camera and LiDAR tokens. Either the LiDAR or camera tokens are used as the Query, while the remaining modality tokens serve as Key and Value. For the Query, we use only the class token, while for the Key-Value modality, we perform experiments using the patch tokens alone and in combination with the class token.

D. Auxiliary Tasks

We also explore auxiliary tasks to leverage the spatial information from the LiDAR modality. First, we add a regression task to estimate the mean depth of each person crop, using the annotated 3D box center as the target depth. This task is trained with the mean squared error loss $\mathcal{L}_{dist}$. Second, we propose estimating a patch-wise depth delta relative to the target depth. For each patch, we calculate the mean LiDAR point depth and determine the delta as the difference between this mean depth and the target depth. The L1 loss $\mathcal{L}_{mask}$ was used to compare the networks per-patch delta estimation with the target values. The final loss function was extended to include these auxiliary tasks using weights λ_{dist} and λ_{mask}:

$$\mathcal{L}_{total} = \lambda_{CE} \cdot \mathcal{L}_{CE} + \lambda_{triplet} \cdot \mathcal{L}_{triplet} + \lambda_{dist} \cdot \mathcal{L}_{dist} + \lambda_{mask} \cdot \mathcal{L}_{mask}. \tag{5}$$



(a) Projected points and camera image. (b) LiDAR intensity for each point. (c) Interpolated LiDAR depth map.

Fig. 3: Example camera and LiDAR crops. We show the distance value of the points with red (close) to green (distant).

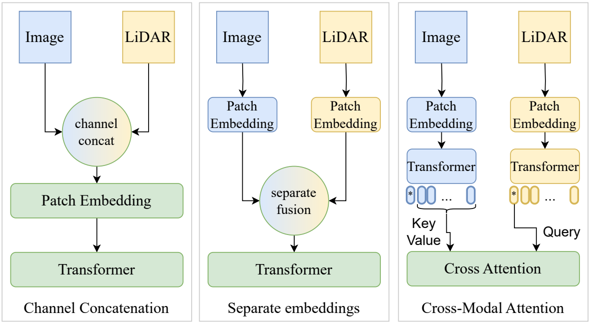


Fig. 4: We propose and investigate three different multi-modal fusion approaches to combine LiDAR and image data.

IV. Experiments

A. Dataset

For our experiments we focus on using the nuScenes driving dataset [1], given its size, quality, and the availability of image and LiDAR data. The dataset offers benchmark protocols for the tasks of object detection, tracking, and prediction among others, but no benchmark setting is provided for person re-identification. To create our ReID dataset, we use all annotated persons in the dataset. For each person (i.e., pedestrians and riders) we extract their corresponding unique id (given by their track id), image and LiDAR crop. We follow the train/val split provided by the nuScenes dataset, to separate train, gallery and query splits. For the latter two, we subdivide the validation split.

The resulting *nuScenes-ReID* dataset contains the images from all cameras, and was used as the dataset for our experiments reported in the following sections. To improve dataset quality we only consider crops, which are larger than 40 pixel in height, 20 pixel in width, contain more than 25 LiDAR points, and are annotated with an object visibility of 40% or more. We report the final number of identities and crops in Table I for the dataset.

TABLE I: Statistics of our proposed nuScenes-ReID dataset derived from the nuScenes driving benchmark [1].

Dataset	Split	Identities	Crops	Cameras
nuScenes-ReID	train	5,284	53,794	6
	query	1,029	2,421	6
	gallery	1,029	8,220	6

B. Evaluation Metrics

We evaluate our ReID model using the commonly used mean average precision (mAP) and Cumulative Matching Characteristics (CMC) [11]. Evaluation is done using the query and gallery subsets. The appearance embeddings for all samples from the gallery and query are generated. Each query embedding is compared to all gallery embeddings.

The CMC metric specifies the percentage in which the same identity as the query is contained in the closest k selected gallery samples. Commonly, the *rank-1* is used in the ReID setting. To calculate the mAP, the average precision (AP) for each sample is used. The AP provides a measure of how well the network is able to identify the same person, taking the recall and precision into account.

C. Experimental Setup

For all our experiments we use $H = 256$, $W = 128$, $D = 768$, and $l = 12$. If not specified otherwise we use a quadratic patch size of 16 pixel, with a stride of 16. During training we employ random input (both LiDAR and image) augmentations, including horizontal flipping, padding, and cropping. We use a batch size of 256. We use SGD (momentum 0.9, weight decay $1e-4$, lr 0.032) as optimizer. We use a cosine learning rate decay, with a warm-up rate of 40 epochs, and train a total of 120 epochs. Implementation is done using PyTorch. Each experiment is run using a single NVidia RTX A6000 GPU. The weights loaded for the ViT backbone are pre-trained on ImageNet-21K and finetuned on ImageNet-1K. For the multi-modal input, we load the weights for the initial convolution layer ($\mathcal{F}$) for the image channels - if applicable. LiDAR channel weights are initialized randomly.

To ensure the reliability of our experimental results, we conduct each training four times, each with a unique random seed. The seed initializes the state of all random number generators in the training process. This procedure enables us to compute and report the mean and standard deviation for all our evaluation metrics. Early stopping based on the validation mAP value was used to prevent overfitting.

D. Comparison to Baselines and State-of-the-Art

We compare our proposed method CLFormer to various baselines and state-of-the-art methods in Table II. In the first section of the table, we present results for different configurations of our image-based transformer baseline. Experiments were conducted with different backbones (ViT [31], DeiT [43], and CrossViT [44]) and varying patch sizes. Consistent with our expectations and the findings in [26], we observe the best performance with ViT using the smallest patch size and a stride of 12 pixels, increasing the mAP from 77.56 to 78.52 points. This demonstrates the effectiveness of using finer granularity in patch sizes for enhancing the models discriminative power - at the cost of higher computational cost. Additionally, the DeiT model, although slightly behind ViT, shows competitive performance with an mAP of 77.10.

TABLE II: Experimental results of various ReID methods on the NuScenes-ReID dataset. The first section presents different configurations of our image-based transformer baseline. The middle section displays the results of CNN-based approaches, while the final section shows transformer-based approaches. Each metric is reported as the mean and ±*std* from four different training runs.

Model	mAP ↑	Rank-1 ↑
Baseline - ViT	77.56 ±*0.06*	91.14 ±*0.02*
Baseline - DeiT	77.10 ±*0.11*	91.10 ±*0.10*
Baseline - CrossViT	73.91 ±*0.13*	89.46 ±*0.06*
ViT (patchsize 14)	78.26 ±*0.04*	91.22 ±*0.09*
ViT (patchsize 12)	78.52 ±*0.13*	91.16 ±*0.29*
BoT [14]	75.59 ±*0.16*	91.29 ±*0.30*
SBS [15]	76.38 ±*0.02*	91.26 ±*0.14*
AGW [16]	76.48 ±*0.16*	91.70 ±*0.15*
MGN [17]	77.23 ±*0.12*	91.52 ±*0.21*
LoGoViT w/o PM [30]	77.20 ±*0.05*	90.52 ±*0.12*
TransReID [26]	78.80 ±*0.07*	91.43 ±*0.30*
DC-Former [29]	78.89 ±*0.09*	91.59 ±*0.23*
CLFormer (ours)	**79.85** ± *0.10*	**92.76** ±*0.12*

The center of the table shows results for CNN-based methods. For all four methods — BoT [14], SBS [15], AGW [16], and MGN [17] — we use a pretrained ResNet-50 backbone network and incorporate Instance-Batch Normalization (IBN) [45]. As expected, all four methods lag behind the transformer-based baseline (first section), given their less efficient backbone architecture. Among these methods, MGN achieves the best performance with its global and local branches, yielding an mAP of 77.23 and a Rank-1 of 91.52.

In the last section, we present results for more recent transformer-based networks. The LoGoViT network [30] without their Patch Modification (PM) module falls behind our transformer baseline by 0.36 mAP points, achieving an mAP of 77.20 and a Rank-1 accuracy of 90.52. Both TransReID [26] and DC-Former [29] demonstrate improved ReID performance, with increases of 1.24 and 1.33 mAP points, respectively. TransReID achieves an mAP of 78.80 and a Rank-1 accuracy of 91.43, while DC-Former achieves the second-best mAP of 78.89 and a Rank-1 accuracy of 91.59.

The last row shows our CLFormer method, which combines multi-modal input concatenation, our proposed 3D-CIPE, and additional auxiliary losses. Our method achieves the best results for both mAP and Rank-1 metrics, with an mAP of 79.85 and a Rank-1 accuracy of 92.76. This represents an increase of 0.96 in mAP over the second-best method (DC-Former) and 2.29 over the baseline (ViT).

TABLE III: Results for various multi-modal fusion configurations using different LiDAR channels (*D*: Depth, *I*: Intensity), input types (*sparse* or interpolated (*Interp.*)), and fusion strategies. Fusion strategies include input channel concatenation (*Concat*), separate embedding (*S*) with linear, product, summation, or concatenation fusion, and CrossAttention (*CA*) with Camera to LiDAR (*C*→*L*) or LiDAR to Camera (*L*→*C*) attention, with or without class tokens (*cls*). See Section III-C. Each metric is presented as mean ± *std*.

Input	Ch.	Fusion	mAP ↑	Rank-1 ↑
Interp.	D+I	S-Product	65.46 ±*3.68*	86.61 ±*1.37*
Interp.	D+I	S-Linear	63.78 ±*2.51*	85.88 ±*1.42*
Interp.	D+I	S-Sum	77.81 ±*0.07*	**91.14** ±*0.25*
Interp.	D+I	S-Concat	59.68 ±*0.35*	84.24 ±*0.47*
Interp.	D+I	$\mathrm{CA}^{C\to L}$ + cls	30.87 ±*6.12*	57.79 ±*7.00*
Interp.	D+I	$\mathrm{CA}^{L\to C}$ + cls	75.56 ±*0.17*	90.72 ±*0.32*
Interp.	D+I	$\mathrm{CA}^{L\to C}$	75.53 ±*0.07*	90.73 ±*0.15*
Sparse	D+I	Concat	77.59 ±*0.03*	90.88 ±*0.27*
Interp.	I	Concat	77.69 ±*0.09*	90.89 ±*0.15*
Interp.	D	Concat	77.66 ±*0.09*	90.96 ±*0.20*
Interp.	D+I	Concat	**77.90** ±*0.05*	**91.14** ±*0.18*

TABLE IV: Results of the ablation study for additional auxiliary tasks and our 3D-CIPE. We show the performance for the image-only baseline, and our multi-modal model using both LiDAR channels interpolated. The columns *D*, *P*, and *E*, indicate if the distance estimation, patch-delta estimation, and 3D-CIPE was added, respectively. For each metric we show the mean and ±*std* from four different runs.

Model	DPE	mAP ↑	Rank-1 ↑
Image	- - -	77.56 ±*0.06*	91.14 ±*0.02*
Image	✓ - -	77.54 ±*0.08*	90.88 ±*0.24*
Image	- ✓ -	77.45 ±*0.06*	90.95 ±*0.23*
Image	✓✓ -	77.47 ±*0.04*	90.90 ±*0.19*
Multi-Modal	- - -	77.90 ±*0.05*	91.14 ±*0.18*
Multi-Modal	✓ - -	77.87 ±*0.08*	91.32 ±*0.16*
Multi-Modal	- ✓ -	77.66 ±*0.05*	91.15 ±*0.10*
Multi-Modal	✓✓ -	77.73 ±*0.07*	91.12 ±*0.23*
Image	- - ✓	79.54 ±*0.07*	**92.78** ±*0.12*
Image	✓✓✓	79.45 ±*0.05*	92.48 ±*0.30*
Multi-Modal	- - ✓	79.68 ±*0.11*	92.77 ±*0.25*
Multi-Modal	✓✓✓	**79.85** ±*0.10*	92.76 ±*0.12*

E. Impact of different Multi-Modal Fusion Strategies

Our results in Table III indicate that the fusion strategy significantly impacts the performance of the ReID method. Among the separate embedding strategies, summation fusion (*S-Sum*) achieved the highest mAP of 77.81% and a Rank-1 accuracy of 91.14%. We attribute this performance, and the observed drop in the other separate fusion settings, to the misalignment between the hidden dimensions of the transformer and the pre-trained model weights optimized for image-only data. Additionally, we observe a nearly two orders of magnitude higher standard deviation in these settings, indicating potential instabilities during training. For CrossAttention strategies, the configuration with LiDAR to Camera attention and class tokens ($\mathrm{CA}^{L\to C}$ +cls) performed well, achieving an mAP of 75.56% and a Rank-1 accuracy of 90.72%. This aligns with our expectation: using the LiDAR token as query enables extraction of rich features from image tokens, unlike the reverse where LiDAR lacks appearance detail.

The impact of including the class token in the key and value seems statistically not significant. Overall, the CrossAttention fusion strategy does not yield the best results, which we attribute to the image-centric nature of the pretrained features, as well as possible limitations in our chosen architecture and dataset scale. Interestingly, the concatenation fusion strategy (*Concat*) with interpolated input and both Depth and Intensity channels (*D+I*) achieved the highest overall performance, with an mAP of 77.90% and a Rank-1 accuracy of 91.14%. This suggests that simple concatenation of multimodal inputs can be highly effective when combined with interpolation. Consequently, we adopt input concatenation as our preferred fusion strategy for subsequent experiments.

F. Ablation Study: Auxiliary Tasks and 3D-CIPE

The integration of depth-related auxiliary tasks into our network was based on the hypothesis that such tasks would steer the network towards effectively utilizing the LiDAR data in a multi-modal context. To evaluate this, we conduct an ablation study using both the image-only baseline and our multi-modal configuration, which incorporates input concatenation with interpolation. These experiments were performed with and without the auxiliary tasks. Furthermore, we investigated the impact of introducing our 3D-CIPE. The outcomes of these experiments are summarized in Table IV.

The upper and central sections of the table present the incremental activation of the additional distance and patch-delta estimation tasks for the image-only and multi-modal models, respectively. Contrary to our initial expectations, the incorporation of these tasks did not enhance the ReID performance; rather, we observed a marginal decline. This trend was consistent across both the image-only and multi-modal models. Specifically, the inclusion of the distance estimation task did not yield a notable improvement, while the patch delta estimation task seems to contradict the training slightly, leading to decreased performance. In contrast, the implementation of our 3D-CIPE results in a significant improvement in mAP. The image-only baseline experiences an increase in mAP from 77.6 to 79.5 with the integration of 3D-CIPE. A similar improvement is observed for the multi-modal model, with an mAP rise of nearly 2 points.

We further analyzed the computational footprint of the variants in Table IV. The baseline image-only model has 90.6M parameters and processes 39,257 samples/s. Extending to multi-modal input increases parameters slightly to 90.97M and lowers throughput to 37,786 samples/s. Adding the 3D-CIPE mask embedding introduces negligible overhead (90.58M parameters, 37,227 samples/s). The final CLFormer model, combining both extensions, totals 91.1M parameters and processes 35,196 samples/s.

V. Conclusion

This paper presented CLFormer, a novel transformer-based method for multi-modal person re-identification (ReID) using camera and LiDAR data. We adapted the nuScenes dataset for the ReID task, enabling a thorough evaluation of our proposed method. Various fusion strategies were explored, with the early concatenation of additional LiDAR channels found to be the best-performing approach.

Our experiments demonstrated that the integration of LiDAR data significantly improves ReID performance. The proposed 3D-CIPE module, which incorporates object-centered information from LiDAR point clouds and 3D cuboids into our transformer architecture, yielded substantial gains in mAP and Rank-1 accuracy. Conversely, the inclusion of depth-related auxiliary tasks, intended to guide the network to focus more on the LiDAR input, did not result in increased ReID performance. These findings provide valuable insights into the complexities of multi-modal data integration and suggest areas for further optimization.

Overall, our work addresses a gap in the ReID field by effectively leveraging the complementary strengths of LiDAR and camera data. Future work will focus on optimizing the fusion of multi-modal inputs and exploring the integration of multi-modal backbones to overcome current limitations given by image-only pre-trained networks.

References

[1] H. Caesar *et al.*, "nuScenes: A Multimodal Dataset for Autonomous Driving," in *Proc. IEEE CVPR*, 2020, pp. 11 621–11 631.

[2] A. Rudenko *et al.*, "Human motion trajectory prediction: A survey," *IJRR*, vol. 39, no. 8, pp. 895–935, 2020.

[3] W. Luo *et al.*, "Multiple object tracking: A literature review," *AIJ*, vol. 293, p. 103448, 2021.

[4] N. Wojke, A. Bewley, and D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric," in *Proc. ICIP*, 2017, pp. 3645–3649.

[5] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust Associations Multi-Pedestrian Tracking," *arXiv:2206.14651*, 2022.

[6] J. Seidenschwarz *et al.*, "Simple Cues Lead to a Strong Multi-Object Tracker," in *Proc. IEEE CVPR*, 2023, pp. 13 813–13 823.

[7] Y.-H. Wang *et al.*, "SMILEtrack: SiMIlarity LEarning for Occlusion-Aware Multiple Object Tracking," in *Proc. AAAI*, 2024, pp. 5740–5748.

[8] Z. Song *et al.*, "Robustness-Aware 3D Object Detection in Autonomous Driving: A Review and Outlook," *IEEE Trans. ITS*, vol. 25, no. 11, pp. 15 407–15 436, 2024.

[9] B. Thérien *et al.*, "Object Re-Identification from Point Clouds," in *Proc. IEEE WACV*, 2024, pp. 8377–8388.

[10] W. Guo *et al.*, "LiDAR-based Person Re-identification," in *Proc. IEEE CVPR*, 2024, pp. 17 437–17 447.

[11] L. Zheng *et al.*, "Scalable Person Re-identification: A Benchmark," in *Proc. IEEE ICCV*, 2015, pp. 1116–1124.

[12] A. Hermans, L. Beyer, and B. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," *arXiv:1703.07737*, 2017.

[13] Z. Zheng, L. Zheng, and Y. Yang, "A Discriminatively Learned CNN Embedding for Person Reidentification," *ACM TOMM*, pp. 1–20, 2018.

[14] H. Luo *et al.*, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE CVPR Workshops*, 2019.

[15] L. He *et al.*, "Fastreid: A pytorch toolbox for general instance re-identification," in *Proc. of the ACM Conf. on Multimedia*, 2023, pp. 9664–9667.

[16] M. Ye *et al.*, "Deep learning for person re-identification: A survey and outlook," *IEEE TPAMI*, vol. 44, no. 6, pp. 2872–2893, 2021.

[17] G. Wang *et al.*, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. of the ACM Conf. on Multimedia*, 2018, pp. 274–282.

[18] P. Chen *et al.*, "Occlude Them All: Occlusion-Aware Attention Network for Occluded Person Re-ID," in *Proc. IEEE ICCV*, 2021, pp. 11 833–11 842.

[19] P. Wang *et al.*, "Quality-Aware Part Models for Occluded Person Re-Identification," *IEEE Trans. on Multimedia*, pp. 3154–3165, 2023.

[20] V. Somers, C. De Vleeschouwer, and A. Alahi, "Body Part-Based Representation Learning for Occluded Person Re-Identification," in *Proc. IEEE WACV*, 2023, pp. 1613–1623.

[21] S. Zhou *et al.*, "Depth occlusion perception feature analysis for person re-identification," *PRL*, pp. 617–623, 2020.

[22] Q. Yang *et al.*, "Focus on the Visible Regions: Semantic-Guided Alignment Model for Occluded Person Re-Identification," *Sensors*, p. 4431, 2020.

[23] R. Hou *et al.*, "VRSTC: Occlusion-Free Video Person Re-Identification," in *Proc. IEEE CVPR*, 2019, pp. 7183–7192.

[24] H. Zhu *et al.*, "SEAS: ShapE-Aligned Supervision for Person Re-Identification," in *Proc. IEEE CVPR*, 2024, pp. 164–174.

[25] W. Liu *et al.*, "Multi-object Tracking with Noisy Labels," in *Proc. IEEE PRAI*, 2022, pp. 443–449.

[26] S. He *et al.*, "TransReID: Transformer-Based Object Re-Identification," in *Proc. IEEE ICCV*, 2021, pp. 15 013–15 022.

[27] K. Zhu *et al.*, "AAformer: Auto-Aligned Transformer for Person Re-Identification," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 35, no. 12, pp. 17 307–17 317, 2023.

[28] N. Pervaiz, M. M. Fraz, and M. Shahzad, "Per-former: Rethinking person re-identification using transformer augmented with self-attention and contextual mapping," *The Visual Computer*, pp. 4087–4102, 2023.

[29] W. Li *et al.*, "DC-Former: Diverse and compact transformer for person re-identification," in *Proc. AAAI*, vol. 37, no. 2, 2023, pp. 1415–1423.

[30] N. Phan *et al.*, "LoGoViT: Local-global vision transformer for object re-identification," in *IEEE ICASSP*, 2023, pp. 1–5.

[31] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv:2010.11929*, 2020.

[32] M. Munaro *et al.*, "One-Shot Person Re-identification with a Consumer Depth Camera," in *Person Re-Identification*, ser. ACVPR, 2014, pp. 161–181.

[33] D. Liciotti *et al.*, "Person Re-identification Dataset with RGB-D Camera in a Top-View Configuration," in *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*, 2017, pp. 1–11.

[34] H. Liu, L. Hu, and L. Ma, "Online RGB-D person re-identification based on metric model update," *CAAI Trans. on Intelligence Technology*, vol. 2, no. 1, pp. 48–55, 2017.

[35] A. Wu, W.-S. Zheng, and J. Lai, "Robust Depth-based Person Re-identification," *IEEE Trans. on Image Processing*, vol. 26, no. 6, pp. 2588–2603, 2017.

[36] C. Patruno *et al.*, "People re-identification using skeleton standard posture and color descriptors from RGB-D data," *Pattern Recognition*, vol. 89, pp. 77–90, 2019.

[37] H. Mukhtar and M. U. G. Khan, "CMOT: A cross-modality transformer for RGB-D fusion in person re-identification with online learning capabilities," *Knowledge-Based Systems*, 2024.

[38] A. Chavan *et al.*, "Towards Global Localization using Multi-Modal Object-Instance Re-Identification," *arXiv:2409.12002*, 2024.

[39] W. Wang *et al.*, "DMM: Dual-Modal Model for Person Re-Identification," in *IJCNN*, 2022, pp. 1–8.

[40] Y. Wang *et al.*, "TOP-ReID: Multi-spectral object re-identification with token permutation," *Proc. AAAI*, vol. 38, no. 6, pp. 5758–5766, 2024.

[41] P. Zhang, Y. Wang, Y. Liu, Z. Tu, and H. Lu, "Magic Tokens: Select Diverse Tokens for Multi-modal Object Re-Identification," in *Proc. IEEE CVPR*, 2024, pp. 17 117–17 126.

[42] B. Yang, J. Chen, and M. Ye, "Shallow-Deep Collaborative Learning for Unsupervised Visible-Infrared Person Re-Identification," in *Proc. IEEE CVPR*, 2024, pp. 16 870–16 879.

[43] H. Touvron *et al.*, "Training data-efficient image transformers & distillation through attention," in *Proc. ICML*, 2021, pp. 10 347–10 357.

[44] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE ICCV*, 2021, pp. 357–366.

[45] X. Pan *et al.*, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *Proc. of the ECCV*, 2018, pp. 464–479.