# LiDAR-based Person Re-identification

Wenxuan Guo[1] Zhiyu Pan[1] Yingping Liang[2] Ziheng Xi[1]
Zhicheng Zhong[1] Jianjiang Feng[1]* Jie Zhou[1]
[1]Department of Automation, BNRist, Tsinghua University [2]Beijing Institute of Technology
{gwx22,pzy20,xizh21,zhongzc18}@mails.tsinghua.edu.cn liangyingping@bit.edu.cn
{jfeng,jzhou}@tsinghua.edu.cn

## Abstract

*Camera-based person re-identification (ReID) systems have been widely applied in the field of public security. However, cameras often lack the perception of 3D morphological information of human and are susceptible to various limitations, such as inadequate illumination, complex background, and personal privacy. In this paper, we propose a LiDAR-based ReID framework, ReID3D, that utilizes pre-training strategy to retrieve features of 3D body shape and introduces Graph-based Complementary Enhancement Encoder for extracting comprehensive features. Due to the lack of LiDAR datasets, we build LReID, the first LiDAR-based person ReID dataset, which is collected in several outdoor scenes with variations in natural conditions. Additionally, we introduce LReID-sync, a simulated pedestrian dataset designed for pre-training encoders with tasks of point cloud completion and shape parameter learning. Extensive experiments on LReID show that ReID3D achieves exceptional performance with a rank-1 accuracy of 94.0, highlighting the significant potential of LiDAR in addressing person ReID tasks. To the best of our knowledge, we are the first to propose a solution for LiDAR-based ReID. The code and dataset are available at https://github.com/GWxuan/ReID3D.*

## 1. Introduction

Person ReID has numerous practical applications, such as video surveillance, intelligent transportation and public security. Most ReID systems use cameras as sensors, aiming to recognize the same individual in images or videos caught by different cameras. With the development of computer vision technology, camera-based ReID has witnessed continuous advancements. However, certain challenges have not yet been effectively addressed. Firstly, cameras introduce limitations in terms of visual ambiguity caused by poor illumination and complex backgrounds [34]. Additionally, current camera-based ReID models primarily learn appearance information [47]. Therefore, variations in human appearance considerably impact the performance of models [19]. Furthermore, camera-based person ReID systems raise personal privacy issues [1, 4] for applications in some areas.

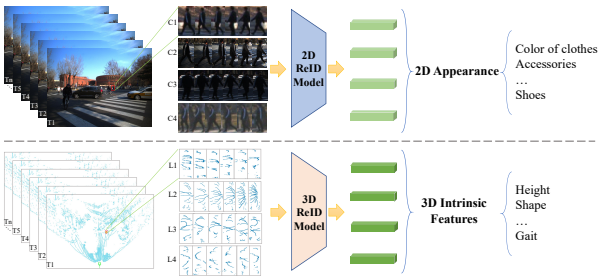



Figure 1. Overview of camera-based and LiDAR-based person ReID. Camera-based methods primarily learn the 2D appearance of human, such as color of clothes, accessories, and shoes. LiDAR-based methods utilize 3D structural information to learn intrinsic features, such as height, body shape, and gait.

In recent years, Kinect-based ReID approaches have been proposed [28, 32] to address these challenges by utilizing depth information. However, Kinect is primarily designed for indoor scenes and has a limited measurement range [17], restricting its applicability in large-scale outdoor scenarios. Recently, researchers have explored the use of radar for gait recognition, identity recognition, and person ReID [7, 26]. While radar is cost-effective, it suffers from poor angular and distance resolution, as well as a limited effective measurement range [7], which poses challenges in discriminating individuals with similar body shapes.

In the past few years, LiDAR has been widely adopted in autonomous driving, driven by its improved accuracy and reduced costs [38]. The successful applications of LiDAR motivate us to explore its potential in person ReID within complex outdoor scenes. LiDAR can offer a broader measurement range and higher resolution, enabling precise perception about individuals in large-scale outdoor settings.

*Corresponding author

LiDAR provides precise 3D structural information without being affected by lighting conditions or complex backgrounds, which enables the extraction of intrinsic features of individuals, such as height, body shape, and gait, irrespective of clothing color. LiDAR has been utilized in the task of gait recognition [34], which focuses on recognizing cooperative subjects walking along a specified route in small-scale scenes. However, ReID primarily focuses on the identification of unscripted pedestrians in large-scale scenes with a small sample size for each identity, which often requires multiple acquisition devices. In our work, we leverage low-cost LiDARs to tackle person ReID challenges, as shown in Fig. 1.

To the best of our knowledge, this paper presents the first study on LiDAR-based person ReID. We build LReID, the first LiDAR-based ReID dataset to facilitate research on utilizing LiDAR point clouds for person ReID. We collect the dataset in several outdoor scenes using multiple collection nodes, each including a Livox Mid-100 LiDAR and an industrial camera. LReID offers several distinctive features: (1) **Real-world scenes**. The dataset is captured in outdoor scenes where pedestrians demonstrate unscripted behavior, leading to occlusions between pedestrians, as well as the presence of dynamic objects like vehicles and bicycles that may impact person ReID. (2) **Data diversity**. LReID encompasses dynamic data and annotations of 320 pedestrian gathered in varying seasons, times of day, and lighting conditions, amounting to a total of 156,000 frames of point clouds and images, thus enabling comprehensive analysis of the impact of different factors on person ReID. (3) **Precision**. The Livox Mid-100 LiDAR has a distance accuracy of 2 cm and an angular accuracy of 0.1°, which provides high-precision 3D structural information for ReID problems.

Additionally, we introduce a simulated dataset, named LReID-sync, including 360,000 frames of point clouds for 600 pedestrians captured by multi-view synchronous LiDARs. LReID-sync comprises annotations for point cloud completion from single view to full views, and Skinned Multi-Person Linear Model (SMPL) parameters [24].

Based on point clouds, the identification of pedestrians relies on their static anthropometric features, including height, body shape, and limb structure, as well as their dynamic gait features. Accurately extracting complete shape features of a pedestrian is beneficial for both aspects. To address this, we propose an efficient LiDAR-based framework, termed ReID3D. ReID3D utilizes a pre-training strategy to guide the encoder in learning 3D body features based on LReID-sync. Moreover, in order to extract discriminative static and dynamic features of pedestrians, the ReID network of ReID3D comprises a Graph-based Complementary Enhancement Encoder (GCEE) and a temporal module. Extensive experiments on LReID demonstrate the following: (1) ReID3D outperforms the state-of-the-art camera-based methods, particularly under low light, highlighting the significant potential of LiDAR in addressing person ReID tasks. (2) The use of LReID-sync for pre-training significantly enhances feature encoding capability of the model. (3) Compared to commonly used point cloud encoders, our GCEE exhibits a higher proficiency in extracting comprehensive and discriminative features.

To summarize, our main contributions are as follows:

- To the best of our knowledge, this is the first work on LiDAR-based person ReID, demonstrating the practicality of utilizing LiDAR for person ReID in challenging real-world outdoor scenes.
- We build LReID, the first LiDAR-based person ReID dataset, which is collected in several outdoor scenes with variations in natural conditions. Moreover, we introduce LReID-sync, a new simulated pedestrian dataset designed for pre-training ReID models with tasks of point cloud completion and shape parameter learning.
- We propose a LiDAR-based ReID framework, termed ReID3D, that utilizes pre-training strategy to guide the encoder in learning 3D body features and introduces GCEE for extracting comprehensive and discriminative features. Experimental results on dataset LReID indicate that ReID3D outperforms camera-based methods.

## 2. Related Work

**Person ReID.** Camera-based ReID has been extensively researched in the past decades. Researchers have employed various convolutional neural networks [48] to extract human features from images or videos. Some video-based ReID models adopt recurrent neural networks (RNNs) [25, 45, 57] or transformer blocks [6, 56] to aggregate temporal features. These models primarily focus on extracting appearance features, which serve as efficient but short-lived identifiers. By contrast, intrinsic and behavioral features exhibit minimal change over time. Given the constraints of using cameras at night, infrared sensors have been introduced for RGB-IR cross-modality person ReID [29, 42]. However, the infrared sensor only captures 2D single-channel intensity information, constraining its perceptive ability.

In order to extract intrinsic features of individuals, researches on ReID utilizing Kinect [28, 32] and radar [7] have emerged. Kinect and radar provide 3D structural information to help the model extract physiological and behavioral features, reducing reliance on the appearance of individuals [10, 32]. However, they are constrained by significant hardware limitations [7, 17]. In comparison, LiDAR serves as a more advantageous option, offering superior practicality. Despite this, LiDAR-based person ReID has not been researched.

**Point Cloud Completion.** Point cloud completion is the task to predict missing parts based on the rest of the point

Table 1. Comparison of publicly available 3D datasets for person ReID.

| Dataset | Year | Identity | Sensor | Scene | Unscripted | Natural diversity | Occlusion |
|---|---|---|---|---|---|---|---|
| RGBD-ID [3] | 2012 | 80 | 1 Kinect | Indoor | ✗ | ✗ | ✗ |
| BIWI RGBD-ID [27] | 2014 | 50 | 1 Kinect | Indoor | ✗ | ✗ | ✗ |
| Kinect-REID [30] | 2015 | 71 | 1 Kinect | Indoor | ✗ | ✗ | ✗ |
| **LReID** | 2023 | **320** | **4 LiDARs + 4 Cameras** | **Outdoor** | ✔ | ✔ | ✔ |
| **LReID-sync** | 2023 | **600** | **4 LiDARs** | Simulation | ✔ | ✗ | ✔ |

cloud. To accomplish this, networks need to learn the intrinsic geometric structures and semantic knowledge of the 3D object. Additionally, the learned representations can be transferred to downstream tasks. The entire process does not require human annotations and therefore falls under the category of unsupervised representation learning [44].

Point cloud completion has received increasing attention in the past decade [13, 16, 23, 40]. In terms of completing the missing parts caused by a single viewpoint, Wang *et al*. [36] utilize an encoder-decoder model to recover the occluded points. However, there has been limited research focusing on point cloud completion for pedestrians.

**Person ReID Dataset.** Depending on the used representations, person ReID datasets can be classified into 2D and 3D datasets. Due to the early emergence and widespread use of cameras, 2D datasets are mainly composed of camera-based datasets [12, 21, 22, 37, 54, 55], which have advanced research in person ReID. In addition, some 2D datasets utilize infrared cameras as sensors [29, 42]. Kinect-based datasets [3, 27, 30] serve as typical examples of 3D datasets. However, the current Kinect datasets have limited scales as they are often collected in small indoor scenes with only one cooperative subject. Additionally, Cheng and Liu [7] collected a radar dataset containing 40 identities, but it is not publicly available and also involves only one cooperative subject in each frame. To address these limitations, we introduce a novel LiDAR-based 3D dataset. The comparison of publicly available 3D datasets is shown in Tab. 1.

## 3. Method

ReID3D adopts multi-task pre-training to guide the encoder in learning 3D body features based on LReID-sync, as shown in Fig. 2. The ReID network of ReID3D comprises a Graph-based Complementary Enhancement Encoder (GCEE), which consists of a GCN backbone and a Complementary Feature Extractor (CFE), along with a temporal module, as shown in Fig. 3. The pre-trained GCEE is used to initialize the ReID network.

### 3.1. Pre-training with Multiple Tasks

Based on our observations, the crucial factors that are likely to impact the performance of ReID models are (1) the variations in information resulted from different viewpoints under cross-view settings, and (2) the incomplete information

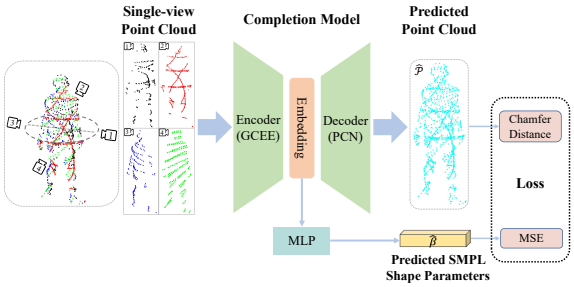



Figure 2. The pre-training approach of ReID3D. Simulated single-view point cloud is taken as input, and the encoder is pre-trained for tasks of point cloud completion and SMPL parameter learning.

acquired from single view. Besides, the collection and annotation of real-world data involve significant costs, while simulated data is low-cost, and comes with rich and accurate annotations. Hence, we leverage simulated data to pre-train the encoder for tasks of point cloud completion and SMPL parameter learning. The overall idea of our proposed pre-training approach is shown in Fig. 2, which enables the encoder to effectively extract anthropometric characteristics and mitigate the influence of viewpoint disparities.

**Multi-task Network.** Throughout we define point cloud $\mathcal{P}$ as a set of points in 3D Euclidean space, $\mathcal{P} = \{p_i | i = 1, 2..., N\}$, where each point $p_i$ can be represented by its coordinates $(x_i, y_i, z_i)$. The network is fed with the single-view point cloud $\tilde{\mathcal{P}}$, then embed it into a latent vector $z$ using encoder $\Phi(\cdot)$. The first branch of the network focuses on the task of point cloud completion, where the decoder $\Psi(\cdot)$ complete the point cloud based on $z$. The process can be formulated as:

$$\hat{\mathcal{P}} = \Psi(\Phi(\tilde{\mathcal{P}})), \tag{1}$$

where $\hat{\mathcal{P}}$ is the predicted complete point cloud.

The second branch aims to learn the characteristics of human body, *i.e*. SMPL shape parameters $\beta$. The latent vector $z$ is expected to encompass shape information. Therefore, a straightforward MLP network is utilized to convert $z$ into predicted SMPL shape parameters $\hat{\beta}$. The process can be formulated as:

$$\hat{\beta} = \text{MLP}(\Phi(\tilde{\mathcal{P}})). \tag{2}$$

**Training.** We adapt the folding-based decoder PCN [49] to complete the single-view point cloud in two steps, out-

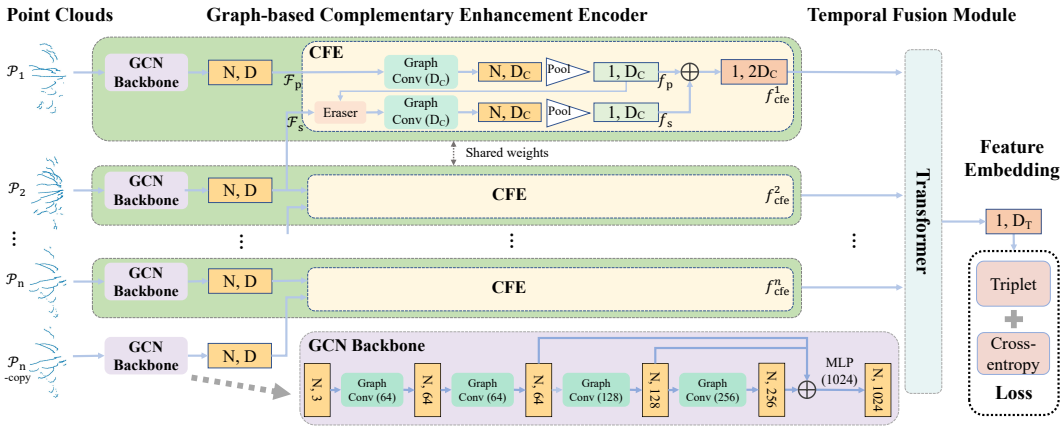



Figure 3. The ReID network of ReID3D. The Graph-based Complementary Enhancement Encoder (GCEE) extracts frame-level features from the pedestrian point clouds, and the transformer module aggregates the sequential features. GCEE consists of a GCN backbone and a Complementary Feature Extractor (CFE).

putting a coarse shape $\hat{\mathcal{P}}_{\text{coarse}}$ and a detailed shape $\hat{\mathcal{P}}_{\text{detail}}$. The pre-training model is trained with a combined loss corresponding to the two branches. In the branch of point clout completion, we use Chamfer Distance (CD) as the difference measure between prediction $\hat{\mathcal{P}}$ and ground-truth $\mathcal{Y}$:

$$\text{CD}(\hat{\mathcal{P}}, \mathcal{Y}) = \frac{1}{|\hat{\mathcal{P}}|}\sum_{\hat{p}\in\hat{\mathcal{P}}}\min_{p\in\mathcal{Y}}\|\hat{p}-p\| + \frac{1}{|\mathcal{Y}|}\sum_{p\in\mathcal{Y}}\min_{\hat{p}\in\hat{\mathcal{P}}}\|p-\hat{p}\|. \tag{3}$$

The loss of the completion branch is a sum of the Chamfer distances on the coarse and detailed shapes weighted by a hyperparameter $\delta$:

$$\mathcal{L}_{\text{com}} = \text{CD}(\hat{\mathcal{P}}_{\text{coarse}}, \mathcal{Y}) + \delta\text{CD}(\hat{\mathcal{P}}_{\text{detail}}, \mathcal{Y}). \tag{4}$$

The branch of shape parameter learning utilizes mean squared error (MSE) as the loss function:

$$\mathcal{L}_{\text{shape}} = \frac{1}{n}\sum_{i=1}^{n}(\beta_i - \hat{\beta}_i)^2, \tag{5}$$

where $n$ is the dimension of $\beta$, *i.e.* $n = 10$.

The pre-training loss is a combination of the two branches with a weighted hyperparameter $\eta$:

$$\mathcal{L} = \mathcal{L}_{\text{com}} + \eta\mathcal{L}_{\text{shape}}. \tag{6}$$

### 3.2. ReID Network

To extract spatio-temporal features from the sequence of point clouds, the ReID Network of ReID3D comprises a GCEE, which consists of a GCN backbone and CFE, along with a temporal module, as illustrated in Fig. 3.

**GCN Backbone.** To extract local and global features from point clouds effectively, we employ graph convolutional structure as the backbone. We construct the directed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ by using the k-nearest neighbors (KNN) of each point including self-loop, where $\mathcal{V}$ represents the points and $\mathcal{E}$ represents the set of edges. For the original point cloud, we construct the graph based on the nearest neighbors in its coordinate space. However, we update the graph at each layer of the network in feature space based on the feature similarity among points, rather than fixed spatial positions.

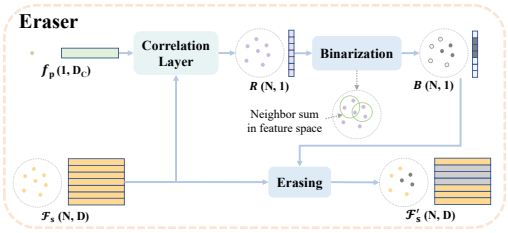



Figure 4. The architecture of the Eraser module in CFE.

Denote that $p_i$ is the central point of once graph convolution operation, and $\mathcal{N}(i) = \{j|(i,j) \in \mathcal{E}\}$ is the set of points in its neighborhood with features $\mathcal{F}(i) = \{f_j|j \in \mathcal{N}(i)\}$. To integrate the global shape structure and local neighborhood feature differences [39], we define $f_{ij} = [f_i, f_j - f_i]$ as the input feature, where $[\cdot,\cdot]$ is the concatenation operation. The convolution process can be formulated as:

$$f_i^{'} = \max_{j\in\mathcal{N}(i)} \sigma(f_{ij} \times k), \tag{7}$$

where $f_i^{'}$ is the output feature of the central point $p_i$, $k$ is the convolution kernel, max is a channel-wise max-pooling function, $\times$ represents matrix multiplication and $\sigma(\cdot)$ is the activation function LeakyReLU.

**Complementary Feature Extractor.** One viable method for extracting frame-level pedestrian features is to apply pooling directly to the features output by the GCN backbone. However, applying the same operation to each frame results in highly redundant extraction, which only emphasizes a part of the most salient features. To this end, inspired by [15], we design CFE to extract complementary features and improving representation learning of the encoder.

For each frame in a sequence, we consider it as the primary frame. The CFE further extracts its features based on the output of the GCN backbone. Then, the next frame is taken as the supplementary frame, from which the Eraser module erases the previously discovered salient features. We duplicate the last frame $\mathcal{P}_n$ as its own supplementary frame. In particular, for a sequence of point clouds $\{\mathcal{P}_i\}_{i=1}^{n} (\mathcal{P}_i \in \mathbb{R}^{N\times 3})$, the feature sequence $\{\mathcal{F}_i\}_{i=1}^{n} (\mathcal{F}_i \in \mathbb{R}^{N\times D})$ is obtained by the GCN backbone, where $N$ is the number of points and $D$ is the dimension of features. For the primary frame, CFE utilizes a graph convolution layer and a global max-pooling layer to extract the most salient features $f_{\mathrm{p}} \in \mathbb{R}^{1\times D_{\mathrm{C}}}$. Then, the Eraser module reconstruct the features of the supplementary frame $\mathcal{F}_{\mathrm{s}}$ guided by the discovered features $f_{\mathrm{p}}$. The reconstructed features $\mathcal{F}_{\mathrm{s}}^{'}$ is fed into similar graph convolution layer and global max-pooling layer to extract auxiliary features $f_{\mathrm{s}}$. $f_{\mathrm{p}}$ and $f_{\mathrm{s}}$ are concatenated to obtain the complementary features $f_{\mathrm{cfe}} \in \mathbb{R}^{1\times 2D_{\mathrm{C}}}$ of the primary frame.

The architecture of the Erasure module is shown in Fig. 4. Firstly, the correlation layer is used to obtain the correlation vector $R \in \mathbb{R}^{N\times 1}$ between $\mathcal{F}_{\mathrm{s}}$ and $f_{\mathrm{p}}$, through computing the semantic relevance between $f_{\mathrm{p}}$ and all the local descriptors of $\mathcal{F}_{\mathrm{s}}$, formulated as:

$$R = \mathcal{F}_{\mathrm{s}} \times (f_{\mathrm{p}} \times \omega)^{T}, \tag{8}$$

where $\omega \in \mathbb{R}^{D_{\mathrm{C}}\times D}$ is a learnable variable projecting $f_{\mathrm{p}}$ to the feature space of $\mathcal{F}_{\mathrm{s}}$. Then, the Binarization module generates the binary mask based on $R$ to identify the points to be erased. Specifically, for each point in $\mathcal{F}_{\mathrm{s}}$, calculate the sum of the correlation values of the point and its $K_{\mathrm{B}}$ nearest neighbors in the feature space. The calculated sum represents the correlation of the region centered on the point. We select the region with the highest correlation value to be erased, through setting the corresponding value in the binary mask $B$ to 0 and others to 1. Finally, $\mathcal{F}_{\mathrm{s}}$ is erased according to $B$ to obtain the reconstructed features $\mathcal{F}_{\mathrm{s}}^{'}$.

**Temporal Fusion Module.** The sequence of pedestrian point clouds includes distinctive dynamic features, such as gait frequency and amplitude of limb swings. We utilize a transformer module [35] with four encoder layers to extract the dynamic features. The transformer module takes the sequence of features $\{f_{\mathrm{cfe}}^{i}\}_{i=1}^{n}$ as input and outputs the final features $f \in \mathbb{R}^{D_{\mathrm{T}}}$.

**Training and Inference.** Following the standard paradigm of camera-based ReID [5, 15, 46, 52], we exploit a combination of cross-entropy and batch-hard triplet terms [14] as the loss function, with the hyperparameter $\gamma$:

$$\mathcal{L} = \mathcal{L}_{\mathrm{ce}} + \gamma \mathcal{L}_{\mathrm{tri}}. \tag{9}$$

During inference, the similarity between query and gallery set is measured using the cosine distance.

## 4. Person ReID Dataset

In this paper, we build a real-world LiDAR-based ReID dataset LReID and a simulated pedestrian point cloud dataset LReID-sync. LReID is captured by a multimodal imaging system, which consists of 4 Livox Mid-100 LiDAR sensors and 4 industrial cameras [51]. LReID includes 320 identities, 156,000 point cloud frames and synchronized RGB images with corresponding 2D and 3D annotations.

**Data Acquisition.** LReID is collected in two extensive outdoor scenes: a crossroad and a square in front of a building, capturing different time periods and weather conditions, as shown in Fig. 6. We synchronize each pair of LiDAR and camera with $\sim$ 5 ms accuracy and provide an accurate extrinsic and intrinsic calibration. All sequences are recorded with a frame rate of 10 Hz. The images are captured of a resolution of 2048 $\times$ 1536 pixels and the density of LiDAR points is about 30,000 points per frame.

The square scene is located in front of a building with frequent pedestrian traffic, and many pedestrians walk together in groups, which makes it more complex for person ReID tasks. We collected data in the square scene during two different times in summer to analyze the impact of lighting conditions on person ReID. Additionally, to explore the impact of different scenes, weather conditions, and pedestrian attire in different seasons, we also collected data at a crossroads scene in winter. Figure 5 shows the samples of two pedestrians collected from different scenes. Furthermore, we analyze the point distribution in LReID and LReID-sync, as shown in Fig. 6.

**Annotations.** To ensure efficient and accurate annotation, we utilize a point cloud-based 3D detector [18] to locate pedestrians. Additionally, we employ a 3D multi-object tracking method [41] to establish continuous trajectories, and manually rectify them. Furthermore, we adjust the IDs of the pedestrians that appeared repeatedly in the field of view based on the images. After obtaining precise 3D ReID annotations, we project them onto synchronized RGB images and perform manual corrections.

To avoid the influence of absolute coordinate information on the network, we normalize the pedestrian point clouds by subtracting the center coordinates of their respective bounding boxes.

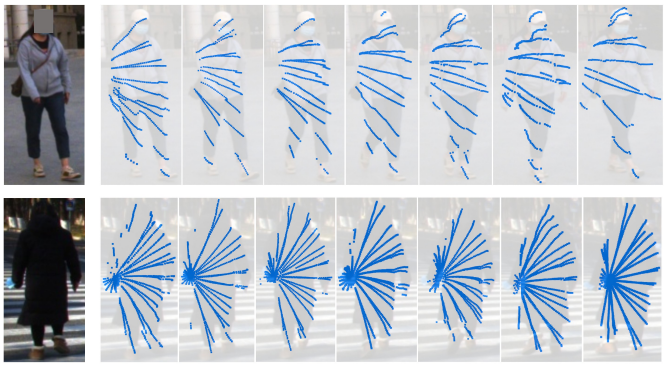


Figure 5. Samples of two pedestrians collected from the square scene in summer (top row), and the crossroads scene in winter (bottom row).

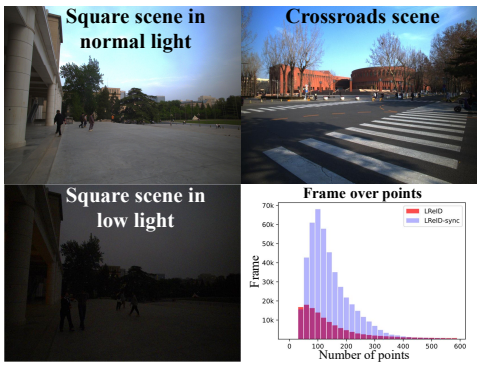



Figure 6. Data acquisition scenes and statistics about points in LReID and LReID-sync (lower right).

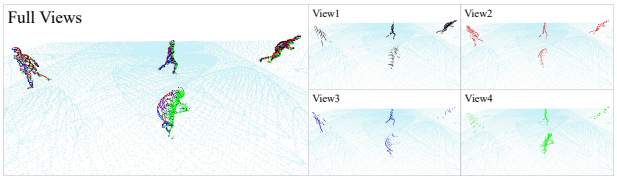



Figure 7. Samples of LReid-sync for full views and each single view.

**Evaluation.** LReID is divided into two splits: a training set with 220 identities and a test set with the remaining 100 identities. Both sets include different scenes, seasons, and lighting conditions. Within the test set, 30 identities are captured under low light, whereas 70 identities are captured under normal light. Following the standard paradigm of video-based datasets [20, 43, 55], the evaluation focuses on recognition across different LiDARs.

We select one sample of each identity in the test set to build the query set and use the other samples as the gallery set. In the test set, we ensure that the data collected by different LiDARs does not overlap in time, which is to prevent the model from recognizing individuals based on human pose at a certain time. To maximize the use of training data, we divide the sequences in training set into several fragments that may partially overlap in time. The evaluation metrics for LReID include Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP), which are consistent with camera-based datasets.

**LReID-sync.** LReID-sync is a novel pedestrian dataset generated using the software Unity3D, which simulates pedestrians in scenes captured by multiple synchronized LiDARs from various views, as shown in Fig. 7. With this setup, LReID-sync accurately simulates the information loss and point cloud sparsity that occur when pedestrians are captured using a single-view LiDAR. Additionally, LReID-sync includes annotations for point cloud completion from single view to full views, as well as SMPL parameters. Therefore, LReID-sync can be utilized for pre-training. LReID-sync comprises 600 pedestrians with various actions, each exhibiting unique body shapes and gaits, ensuring diversity within the dataset.

## 5. Experiments

### 5.1. Implementation Details

**ReID Network.** The ReID network is trained for 700 epochs using the AdamW optimizer with a weight decay of 5e-5. The learning rate, initially set to 5e-5, is updated using a Cosine Annealing Learning Rate (CosineAnnealingLR) scheduler with a cycle of 200 epochs. Each frame of pedestrian point clouds in LReID is upsampled or downsampled to 256 points as input to the network. To train our model, we randomly choose 6 identities, and sample 6 sequences for each identity with a sequence length of 30 frames. The loss weight $\gamma$ is set to 1. The neighborhood size is set to 10 and 8 for each graph convolution layer and Binarization module, respectively. And the feature dimensions $D$, $D_{\mathrm{C}}$ and $D_{\mathrm{T}}$ are set to 512, 512 and 1024, respectively.

**Pre-training.** The pre-training model follows a similar training strategy as described above, except that the initial learning rate is set to 1e-4. The loss weight $\eta$ is set to 1. The coefficient $\delta$ in Eq. 4 is set to 0.01 for the first 100 epochs, then increased to 0.1, 0.5 and 1.0 after 100, 200 and 400 epochs. The coarse shape $\hat{\mathcal{P}}_{\mathrm{coarse}}$ contains 128 points, while the detailed shape $\hat{\mathcal{P}}_{\mathrm{detail}}$ contains 512 points. The weights of the pre-trained GCEE are used as initialization for the ReID network.

### 5.2. Comparison with Camera-based Methods

We compare ReID3D with the state-of-the-art video-based methods [2, 9, 15, 50]. The input images are resized in the

Table 2. Comparison with state-of-the-art camera-based methods on LReID. Numbers in bold indicate the best performance and underscored ones are the second best. To ensure fairness, ReID3D without pre-training is evaluated.

| Method | Venue | Modality | Normal light | | | Low light | | | Overall | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Rank-1 | Rank-3 | mAP | Rank-1 | Rank-3 | mAP | Rank-1 | Rank-3 | mAP |
| TCLNet [15] | ECCV 2020 | Camera | **98.6** | 98.6 | **94.88** | 60.0 | 73.3 | 46.18 | 87.0 | 91.0 | 80.27 |
| STMN [9] | ICCV 2021 | | 94.3 | 98.6 | 93.81 | 30.0 | 50.0 | 31.17 | 75.0 | 84.0 | 75.02 |
| SINet [2] | CVPR 2022 | | 97.1 | **100.0** | 94.63 | 43.3 | 60.0 | 43.97 | 81.0 | 88.0 | 79.43 |
| PiT [50] | TII 2022 | | 94.3 | 95.7 | 86.21 | 33.3 | 60.0 | 35.74 | 76.0 | 85.0 | 71.07 |
| B-ReID3D | Ours | LiDAR | 90.0 | 97.1 | 82.04 | 90.0 | 93.3 | 81.64 | 90.0 | 96.0 | 81.92 |
| ReID3D | Ours | | 94.3 | 98.6 | 83.65 | **93.3** | **96.7** | **82.43** | **94.0** | **98.0** | **83.28** |

B-ReID3D: ReID3D without pre-training.

Table 3. Comparison for different pre-training method.

| Pre-training Method | Rank-1 | Rank-3 | mAP |
|---|---|---|---|
| Baseline(w/o pre-training) | 90.0 | 96.0 | 81.92 |
| ReID | 91.0 | 96.0 | 82.13 |
| Completion | 93.0 | 97.0 | 82.73 |
| Completion + SMPL | **94.0** | **98.0** | **83.28** |

resolution of $128 \times 64$. To ensure fairness, ReID3D without pre-training (B-ReID3D) is also evaluated. All methods are evaluated with the same dataset settings and metrics.

The comparative results are shown in Tab. 2, from which the following observations can be obtained: (1) ReID3D and B-ReID3D demonstrate their superiority to the video-based methods, primarily benefiting from the utilization of point clouds, which is unaffected by lighting conditions and complex background. (2) ReID3D achieves state-of-the-art results in overall and low light conditions, but it falls behind video-based methods in normal light conditions. This is because video-based methods make full use of appearance information under normal light. (3) Video-based methods perform poorly under low light, while ReID3D and B-ReID3D demonstrate comparable reliability under both low light and normal light.

### 5.3. Pre-training

To demonstrate the effectiveness of pre-training with simulated dataset LReID-sync, we evaluate the performance of different pre-training method. The following four method are evaluated: (1) ReID3D without pre-training. (2) Pre-training with similar ReID task, in which the pre-training model and loss are consistent with the ReID network. (3) Pre-training with only the branch of point cloud completion. (4) Pre-training with multiple tasks, as described in Section 3.1. The experimental results are shown in Tab. 3. It can be observed that pre-training with multiple tasks makes an improvement to the accuracy of ReID3D, with the rank-1 accuracy increasing from 90.0 to 94.0. This improvement can be attributed to two factors: (1) The simulated data comprises a diverse collection of pedestrian point clouds with various body shapes, which supplements the real-world data. (2) Multi-task pre-training effectively leverages the performance of the encoder. Besides, pre-training with task of ReID or point cloud completion also leads to an improvement in performance.

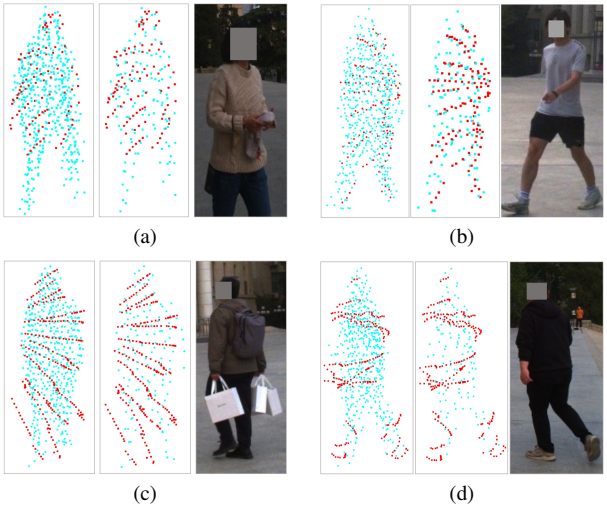



Figure 8. Completion results for real-world pedestrian point clouds obtained by the pre-trained model. In each sample, the detailed shape, coarse shape, and image are shown from left to right. The red points represent the input point cloud, and the blue points represent the predicted complete point cloud.

To showcase the robust feature encoding capabilities attained by the encoder through pre-training, we visualize the completion results for several real-world pedestrian point clouds with different characteristics, as shown in Fig. 8. We can observe that: (1) Intuitively, the detailed and coarse shape closely resemble the actual human shape, which indicates that the encoder has successfully captured the complete features of the human body. (2) The detailed shape is an extension based on the coarse shape, with higher resolution and more information. (3) The pre-trained encoder has the ability to estimate the features of the missing part in the point cloud. As shown in Fig. 8a, the unobserved parts of the legs are accurately completed. (4) Due to motion blur or object carrying, the point cloud may contain some noise points. For example, in Fig. 8b and Fig. 8c, there are red noise points located between the legs and from the carrying bags, respectively. However, the model can effectively filter out the noise points that do not conform to the body shape.

Table 4. Performance of B-ReID3D with different encoders.

| Encoder | Rank-1 | Rank-3 | mAP |
|---|---|---|---|
| PointNet [31] | 74.0 | 88.0 | 60.31 |
| Point Transformer [53] | 83.0 | 93.0 | 74.86 |
| 3DCNN [8] | 34.0 | 51.0 | 28.27 |
| GCEE | **90.0** | **96.0** | **81.92** |

Table 5. Ablation study on the CFE module and Eraser module.

| Encoder Type | Rank-1 | Rank-3 | mAP |
|---|---|---|---|
| B-ReID3D-w/o CFE | 83.0 | 94.0 | 74.15 |
| B-ReID3D-w/o Eraser | 85.0 | 95.0 | 80.47 |
| B-ReID3D | **90.0** | **96.0** | **81.92** |

### 5.4. Ablation Study

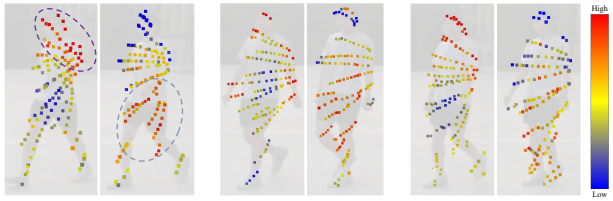
**Effect of Encoders.** To evaluate the effectiveness of our GCEE, we also implement several commonly used encoders in point cloud tasks to replace GCEE in the ReID network for comparison, including PointNet [31], Point Transformer [53], and 3DCNN [8]. All methods use consistent training strategies and employ a same transformer module as the temporal fusion network. We report the results on LReID in Tab. 4, obtaining the following observations: (1) GCEE outperforms other encoders, which is mainly beneficial by its flexible graph structure in the feature space and the strategy of complementary enhancement. (2) Point-based encoders are superior to voxel-based encoder 3DCNN, which indicates that point-based encoders can provide a more detailed understanding of humans in tasks like ReID.

**Effect of CFE.** CFE is a crucial component in our proposed encoder, used to extract complementary features. We compared our full model with two ablated versions: one without CFE and another without the Eraser module, as shown in Tab. 5. Removing CFE results in a lower rank-1 accuracy, with a score of 83.0 compared to our 90.0, which highlights the effectiveness of the complementary enhancement in constructing distinctive and comprehensive features. While removing the Eraser module retains the features of the complementary frame, it might lead to the extraction of similar features from the primary frame, thus restricting the improvement in accuracy, with a score of 85.0 compared to 83.0 of the version without CFE.

**Effect of the Temporal Module.** We evaluate the effectiveness of different temporal modules in combination with GCEE, as shown in Tab. 6. We can observe that the attention-based transformer structure demonstrates significant superiority over the LSTM structure. The rank-1 accuracy of transformer exceeds that of LSTM by 8 points, and that of BiLSTM by 10 points.

Table 6. Results of B-ReID3D with different temporal modules.

| Temporal Module | Rank-1 | Rank-3 | mAP |
|---|---|---|---|
| LSTM [11] | 82.0 | 89.0 | 72.34 |
| Bi-LSTM [33] | 80.0 | 90.0 | 70.92 |
| Transformer [35] | **90.0** | **96.0** | **81.92** |




(a) (b) (c)

Figure 9. Feature visualization of CFE. The features of the primary frame and the supplementary frame are presented on left and right, respectively. The color bar on the right indicates the feature value.

### 5.5. Visualization Analysis

For qualitative analysis, we visualize the feature values output by the two graph convolutional layers in the CFE, as shown in Fig. 9. Three samples are taken from different actions of a person in the test set of LReID, with each sample comprising two consecutive frames. In each sample, the features of the primary frame and the supplementary frame are presented. We can observe that primary ones pay more attention to the head and arms, potentially linked to the discriminative height parameter, while supplementary ones predominantly emphasize the torso and legs. The combination of them creates a complementary and comprehensive feature representation.

## 6. Conclusions

This paper presents the first research on person ReID using precise 3D structural information provided by LiDAR. Firstly, We propose a LiDAR-based ReID framework, named ReID3D, that utilizes pre-training to guide Graph-based Complementary Enhancement Encoder (GCEE) for extracting comprehensive 3D intrinsic features. Moreover, we build the first LiDAR-based person ReID dataset, termed LReID, which contains 320 pedestrians in various outdoor scenes and lighting conditions. Additionally, we introduce LReID-sync, a new simulated pedestrian dataset designed for pre-training encoders with tasks of point cloud completion and shape parameter learning. Our proposed ReID3D demonstrates exceptional performance on LReID, highlighting the significant potential of LiDAR in addressing person ReID tasks.

## Acknowledgement

This work is supported by the National Key Research and Development Program of China under Grant (2018AAA0102803).

## References


[1] Sk Miraj Ahmed, Aske R Lejbolle, Rameswar Panda, and Amit K Roy-Chowdhury. Camera on-boarding for person re-identification using hypothesis transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12144–12153, 2020. 1

[2] Shutao Bai, Bingpeng Ma, Hong Chang, Rui Huang, and Xilin Chen. Salient-to-broad transition for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7339–7348, 2022. 6, 7

[3] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. Re-identification with rgb-d sensors. In *European Conference on Computer Vision*, pages 433–442. Springer, 2012. 3

[4] Karla Brkić, Tomislav Hrkać, Zoran Kalafatić, and Ivan Sikirić. Face, hairstyle and clothing colour de-identification in video sequences. *IET Signal Processing*, 11(9):1062–1068, 2017. 1

[5] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In *European Conference on Computer Vision*, pages 660–676. Springer, 2020. 5

[6] Ying Chen, Shixiong Xia, Jiaqi Zhao, Yong Zhou, Qiang Niu, Rui Yao, Dongjun Zhu, and Dongjingdian Liu. Restreid: Transformer block-based residual learning for person re-identification. *Pattern Recognition Letters*, 157:90–96, 2022. 2

[7] Yuwei Cheng and Yimin Liu. Person reidentification based on automotive radar point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021. 1, 2, 3

[8] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 8

[9] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. Video-based person re-identification with spatial and temporal memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12036–12045, 2021. 6, 7

[10] Lijie Fan, Tianhong Li, Rongyao Fang, Rumen Hristov, Yuan Yuan, and Dina Katabi. Learning longterm representations for person re-identification using radio signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10699–10709, 2020. 2

[11] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471, 2000. 8

[12] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, pages 1–7, 2007. 3

[13] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3D surface generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 216–224, 2018. 3

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5

[15] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *European Conference on Computer Vision*, pages 388–405. Springer, 2020. 5, 6, 7

[16] Zitian Huang, Yikuan Yu, Jiawen Xu, Feng Ni, and Xinyi Le. Pf-net: Point fractal network for 3D point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7662–7670, 2020. 3

[17] Elise Lachat, Helene Macher, M-A Mittet, Tiemo Landes, and Pierre Grussenmeyer. First experiences with kinect v2 sensor for close range 3D modelling. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40:93–100, 2015. 1, 2

[18] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019. 5

[19] Qingming Leng, Mang Ye, and Qi Tian. A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1092–1108, 2019. 1

[20] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3958–3967, 2019. 6

[21] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *Asian Conference on Computer Vision*, pages 31–44. Springer, 2013. 3

[22] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 3

[23] Minghua Liu, Lu Sheng, Sheng Yang, Jing Shao, and Shi-Min Hu. Morphing and sampling network for dense point cloud completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11596–11603, 2020. 3

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2

[25] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2016. 2

[26] Zhen Meng, Song Fu, Jie Yan, Hongyuan Liang, Anfu Zhou, Shilin Zhu, Huadong Ma, Jianhua Liu, and Ning Yang. Gait recognition for co-existing multiple people using millimeter wave sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 849–856, 2020. 1

[27] Matteo Munaro, Andrea Fossati, Alberto Basso, Emanuele Menegatti, and Luc Van Gool. One-shot person re-identification with a consumer depth camera. *Person Re-Identification*, pages 161–181, 2014. 3

[28] Brent C Munsell, Andrew Temlyakov, Chengzheng Qu, and Song Wang. Person identification using full-body motion and anthropometric biometrics from kinect videos. In *European Conference on Computer Vision*, pages 91–100. Springer, 2012. 1, 2

[29] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017. 2, 3

[30] Federico Pala, Riccardo Satta, Giorgio Fumera, and Fabio Roli. Multimodal person reidentification using rgb-d cameras. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(4):788–799, 2015. 3

[31] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 8

[32] Liangliang Ren, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Multi-modal uniform deep learning for RGB-D person re-identification. *Pattern Recognition*, 72:446–457, 2017. 1, 2

[33] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. 8

[34] Chuanfu Shen, Chao Fan, Wei Wu, Rui Wang, George Q Huang, and Shiqi Yu. Lidargait: Benchmarking 3D gait recognition with point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1054–1063, 2023. 1, 2

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 5, 8

[36] 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. 3

[37] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703. Springer, 2014. 3

[38] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 1

[39] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5):1–12, 2019. 4

[40] Xin Wen, Tianyang Li, Zhizhong Han, and Yu-Shen Liu. Point cloud completion by skip-attention network with hierarchical folding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1939–1948, 2020. 3

[41] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. Ab3dmot: A baseline for 3D multi-object tracking and new evaluation metrics. *arXiv preprint arXiv:2008.08063*, 2020. 5

[42] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. RGB-infrared cross-modality person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5380–5389, 2017. 2, 3

[43] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2018. 6

[44] Aoran Xiao, Jiaxing Huang, Dayan Guan, Xiaoqin Zhang, Shijian Lu, and Ling Shao. Unsupervised point cloud representation learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[45] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, pages 701–716. Springer, 2016. 2

[46] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2899–2908, 2020. 5

[47] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, 28(6):2860–2871, 2019. 1

[48] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2021. 2

[49] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In *2018 International Conference on 3D Vision*, pages 728–737. IEEE, 2018. 3

[50] Xianghao Zang, Ge Li, and Wei Gao. Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Transactions on Industrial Informatics*, 18 (12):8776–8785, 2022. 6, 7

[51] Meng Zhang, Wenxuan Guo, Bohao Fan, Yifan Chen, Jianjiang Feng, and Jie Zhou. A flexible multi-view multi-modal imaging system for large-scale outdoor scenes. In *3DV*, 2022. 5

[52] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-granularity reference-aided attentive feature ag-

gregation for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10407–10416, 2020. 5

[53] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021. 8

[54] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1116–1124, 2015. 3

[55] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016. 3, 6

[56] Yanwei Zheng, Zengrui Zhao, Xiaowei Yu, Dongxiao Yu, et al. Template-aware transformer for person reidentification. *Computational Intelligence and Neuroscience*, 2022, 2022. 2

[57] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person reidentification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4747–4756, 2017. 2