Cross-Platform Video Person ReID: A New Benchmark Dataset and Adaptation Approach

Anonymous ECCV 2024 Submission

Paper ID #3918

Abstract. In this paper, we construct a large-scale benchmark dataset for Ground-to-Aerial Video-based person Re-Identification, named G2A-VReID, which comprises 185,907 images and 5,576 tracklets, featuring 2,788 distinct identities. To our knowledge, this is the first dataset for video ReID under Ground-to-Aerial scenarios. G2A-VReID dataset has the following characteristics: 1) Drastic view changes; 2) Large number of annotated identities; 3) Rich outdoor scenarios; 4) Huge difference in resolution. Additionally, we propose a new benchmark approach for cross-platform ReID by transforming the cross-platform visual alignment problem into visual-semantic alignment through vision-language model (*i.e.*, CLIP) and applying a parameter-efficient Video Set-Level-Adapter module to adapt image-based foundation model to video ReID tasks, termed VSLA-CLIP. Besides, to further reduce the great discrepancy across the platforms, we also devise the platform-bridge prompts for efficient visual feature alignment. Extensive experiments demonstrate the superiority of the proposed method on all existing video ReID datasets and our proposed G2A-VReID dataset.

Keywords: Dataset · Ground-to-Aerial · Person Re-Identification

1 Introduction

Video-based person Re-Identification (VReID) [2,12,23,24], has been attracting much attention in recent years, as video can provide richer information than single image. Existing research efforts on video-based ReID are mostly based on the data captured from the same platforms, such as ground surveillance cameras. Suppose in this scenario that a suspect has committed a crime in the city where abundant surveillance cameras have been deployed and escaped into the rural areas where there are no deployed ground surveillance cameras in advance. One feasible solution is sending a moving camera with the help of an airbone UAV platform. Thus, the technical crux has been turned into cross-platform video-based person ReID in aerial captured videos, with a given query video tracklet captured by ground cameras.

In this paper, to meet the research need of cross-platform video person ReID, we construct a large-scale benchmark dataset named Ground-to-Aerial Video ReID (G2A-VReID). The G2A-VReID dataset consists of 185,907 images in total, with 5,576 tracklets belonging to 2,788 different person IDs. Each person

ID includes two tracklets captured by the UAV and ground surveillance platforms, respectively. There is an average of 33.3 images for each tracklet. The scale of G2A-VReID dataset is larger than most existing video-based person ReID datasets such as MARS [44], iLIDS [36], PRID-2011 [15], etc.

To capture the videos of the same person by both the ground surveillance camera and the UAV-mounted camera, we simulate the ground-to-aerial platform ReID by fixating a ground surveillance camera at a specific location, while flying a DJI consumer UAV nearby to ensure many people can be captured by both cameras. The ground camera is set at about 2.0 meters above the ground, and the flight altitudes of UAV varies from 20 meters to 60 meters. Additionally, to be more realistic, the flight mode is adjusted randomly among hovering, cruising, and rotating with diverse view angles which greatly enriches the perspectives of the dataset.

Furthermore, the dataset is collected at nine different scenarios, including school campuses, subway station entrances, tourist sites, crossroads, etc. As shown in Fig. 1, the cross-platform video person ReID task is much more challenging than the counterpart in single ground platform, as the tracklets captured in the ground to aerial cross-platform scenarios are featured in drastic variations of view-points, poses, and resolutions. We have evaluated nine existing video-based person ReID algorithms on our newly collected cross-platform dataset. The experimental results showed inferior performances compared with those conventional single-platform datasets. Due to the great challenges of drastic view, pose, and resolution changes, it is not easy to align the visual part features between the cross-platform devices, which is essential in ReID task.

Recently, with the emergence of large-scale pre-trained vision-language models, *e.g.*, CLIP [29], a well-aligned visual-semantic space can be obtained through cross-modality contrastive learning of large web visual data along with high-level language descriptions. Although for the ReID task, there is no language descriptions for each person whose identity is just denoted as an index number, a set of learnable description tokens can also be introduced to roughly describe each ID [25]. In this paper, we propose to transform the cross-platform visual alignment problem into visual-semantic alignment with the help of the foundation model CLIP. To be concrete, a two-stage optimization strategy is utilized, which aims to learn description tokens for each ID in the first stage, and fine-tunes the Image Encoder with aligning visual embeddings to semantic features obtained through the learned description token in the second stage. Our experiments demonstrate that fine-tuning the Image Encoder with the constraint of visual-semantic alignment achieves competitive performance.

However, there are two obvious drawbacks in adapting image-based pre-trained foundation models to video ReID tasks by simply fine-tuning. One is the huge training cost with large-scale trainable parameters, and another is that the image encoder lacks the capability of modeling inter-frame information. Many previous works [1,3,18] deem video as a stack of frames with temporal structure, and are devoted to modeling temporal features with well-designed modules. But these works ignore the complementarity of frames in a video, which proved to be more effective in ReID task [2]. Moreover, from the aerial perspective, temporal information is limited due to severe self-obstruction. As shown in Tab. 2, temporal models [10, 13, 18] show inferior performance on G2A-VReID. In this paper, we present a new perspective that regards a video clip as a disordered set and propose a parameter-efficient Video Set-Level-Adapter (VSLA) module for foundation modal adaptation. Concretely, VSLA consists of a Cross-Frame Attention Adapter (CFAA) and an Intra-Frame Adapter (IFA). CFAA uses cross-frame attention to allow information exchange between frames, enabling our model to collect complementary features in each video set for powerful video-level representations. IFA transfers the visual ability of image-based foundation model to downstream tasks, providing strong intra-frame appearance representation.

Furthermore, we also propose the Platform Bridging Prompt (PBP) module to solve the visual misalignment problem in cross-platform tasks, where the prompts are adopted to provide explicit instruction to the pre-trained models for generating task-specific results. Specifically, the designed PBP is two sets of platform-specific prompts brought in Image Encoder, which aims to guide the model to focus on learning platform-invariant features, thus bridging the semantic gap of visual features between the ground and aerial platforms.

In summary, the main contributions are as follows:

- We are the first to collect a large-scale Ground-to-Aerial Video person ReID benchmark dataset for the task of cross-platform video-based person ReID and conducted extensive baseline methods on our dataset.
- We propose to transform the essential cross-platform visual part alignment problem into visual-semantic alignment with the help of CLIP, and propose PBP to further bridge the semantic gap of visual features between the ground and aerial platforms.
- We propose the Video Set-Level-Adapter to efficiently adapt pre-trained image-based visual foundation model to the video ReID tasks. Our methods achieves state-of-the-art performances on three widely used video ReID datasets and our cross-platform benchmark dataset.

2 Related Works

In this section, we provide a concise review of two sets of works closely related to our research.

Video ReID Datasets. Existing works on person ReID can be categorized into image-based ReID and video-based ReID. For video-based ReID, the popular datasets include PRID-2011 [15], iLIDS [36], MARS [44] and LS-VID [22], etc. PRID-2011 comprises multiple person trajectories captured by two static surveillance cameras, encompassing only 400 sequences involving 200 individuals. In contrast, LS-VID is a large-scale benchmark featuring 14,943 sequences of 3,772 persons, with videos captured at various times throughout the day. Many works have achieved superior performances on these datasets. Specifically, FGReID [40] achieved Rank-1 at 96.1% on PRID-2011, SINet [2] got 92.5% of

Rank-1 on iLIDS and DenseIL [14] achieved an mAP of 87.0% on MARS, indicating a saturation trend on these datasets. The existing datasets are all captured with a single platform, i.e. ground surveillance cameras, while we aim to collect a Ground-to-Aerial cross-platform video ReID dataset to support the development of this field.

Video ReID Methods. The object processed in video-based person ReID is a video composed of a sequence of person images. Videos contain richer temporal and spatial information than images. Previous works used 3D CNNs [1, 13, 23], temporal weighting [6, 12, 13, 44, 46], optical flow [8, 11, 26] and many other methods [7, 10, 17, 18] to model the spatiotemporal information of video sequences to alleviate the negative effects of appearance change, occlusion, pose variation, etc. For 3D CNNs, STRF [1] proposed a trainable unit with negligible computational overhead, which is used in conjunction with 3D-CNN to learn discriminate 3D features. For temporal weighting, AP3D [13] assigns attention scores for each spatial region to achieve discriminative parts mining and frame selection. Optical flow refers to the movement of target pixels in an image due to the movement of objects in the image or the movement of the camera in two consecutive frames. STA [12] makes use of color and optical flow information in order to capture appearance and motion information. An essential topic to improve the performance of video-based ReID is the visual part alignment between query and gallery videos. SRL [4] learns the structural relationship between local regions in an image and achieves the alignment of videos. RQEN [32] solves the problem of misalignment caused by the difference in image quality between different regions by compensating between local regions. PiT [42] divides each frame into small patches of different granularity in different directions, allowing the model to align two videos with multi-scale local information. It is relatively easy to align the visual part features between the query and gallery videos for these methods by utilizing a simple stripe partition, as the variations of view, pose, and resolution are limited among the single ground cameras.

To solve the severe misalignment of visual features in cross-platform tasks, we resort to visual-semantic alignment of the CLIP model to align the cross-platform person features.

3 Dataset

In this section, we first introduce how we collect and annotate our G2A-VReID dataset in Sec. 3.1 and Sec. 3.2. Then, we make comparisons with other datasets and highlight the key characteristics of G2A-VReID in Sec. 3.3.

3.1 Dataset Collection

To increase the richness of data and make it closer to the real environment. The videos are captured from 9 different scenarios, including library, crossroads, bus stop, tourist sites, etc. Following the principle of protecting public privacy, we post notes to inform pedestrians of the video capture area in the shooting stage,

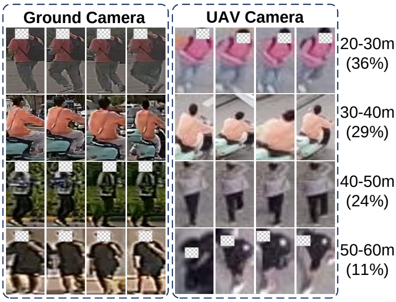

Fig. 1: Visualization of proposed G2A-VReID at different heights.

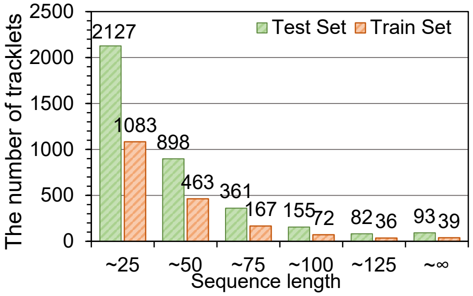

Fig. 2: The distributions of sequence length.

and the dataset will be licensed for non-profit academic research only. Ground surveillance cameras are used to shoot videos from the ground perspective, and a DJI Mavic UAV is adopted to gather videos from the sky perspective. In detail, the surveillance camera is fixed at a height of about two meters above the ground, and the UAV flies at different heights from 20 to 60 meters. The UAV flies in a mode of hovering, cruising, and rotating, making the captured persons contain richer perspectives.

We cropped the captured video at intervals of 0.5 seconds to generate 31,770 frames of scene images. Some of them are shown in Fig. 1. We can see that there are great differences in the viewing perspective and resolution of images taken on different platforms, making it more challenging than existing datasets.

3.2 Annotation

To the best of our knowledge, there is not exist ground-to-aerial cross-platform video-based person Re-ID dataset. One reason is that it takes a lot of effort to annotate a large-scale dataset. We invited 40 experienced annotators and took two weeks to complete the annotation process. During annotation, all persons appeared in the videos are marked with boundary boxes, and each person is cropped from the scene image according to the box. At the same time, we use mosaic to mask the clear face information for privacy protection. Then, the same people in the UAV and surveillance videos are associated and assigned unique IDs. Next, we combine all the images of a person in one camera into one trajectory. Thus, each person has at least two trajectories, one from the surveillance camera and the other from the UAV. Finally, we annotated 185,907 images of 2,788 identities, corresponding to 5,576 tracklets. Fig. 2 shows the distributions of sequence length.

3.3 Characteristics of Our G2A-VReID

Compared with existing VReID datasets [22,36,44] , the characteristics of G2A-VReID are as follows:

Drastic view changes. The tracklets in the query and gallery sets are captured from different types of cameras, specifically ground and aerial views.

Table 1: Comparison of G2A-VReID with other Video-ReID datasets. **CWM** denotes the camera working mode. **AD** is the average duration of each video sequence.

Datasets	G2A-VReID	LS-VID [22]	Mars [44]	iLIDS [36]	PRID-2011 [15]	3DPeS [3]
identities	2,788	3,772	1,261	300	200	200
tracklets	5,576	14,943	20,715	600	400	1,000
images	185,907	2,982,685	1,067,516	42,460	40,033	200,000
AD (s)	**16.7**	6.7	5.6	2.4	3.3	6.7
camera	2	15	6	2	2	8
view	**ground** & **sky**	ground	ground	ground	ground	ground
CWM	**moving**	fixed	fixed	fixed	fixed	fixed

Consequently, the transitions between the views in the query and gallery tracklets are significantly different, compared to the current video-based person ReID datasets.

Large number of annotated identities. Our G2A-VReID consists of 2,788 person IDs and 185,907 images, corresponding to 5,576 tracklets. The number of identities is significantly higher than all existing datasets except LS-VID [22], as shown in Tab. 1.

Rich outdoor scenarios with large view changes. The G2A-VReID consists of footage from nine diverse scenarios, including libraries, bus stops, subway station entrances, tourist sites, crossroads, and more. This diversity enables G2A-VReID to accurately represent realistic environments for person ReID. In contrast, the videos from Mars [44] are captured in a university campus, while iLIDS [36] only contains videos collected from an airport arrival hall. These datasets exhibit comparatively limited scenarios in terms of environmental diversity.

Huge difference in resolution. As depicted in Fig. 1, the height of the UAV-mounted camera varies significantly, spanning from 20 to 60 meters, leading to varying resolutions of individuals captured in each scene. The width distribution of individuals in images captured by ground cameras primarily ranges from 10 to 70 pixels. Whereas, in UAV-captured images, this range is narrower from 5 to 35 pixels. This discrepancy in resolution distribution between query and gallery images introduce much complexity to the task of ground-to-aerial video-based person ReID.

4 Approach

Fig. 3 illustrates the overall architecture of our proposed method. Our approach focuses on cross-platform video person ReID and aims to parameter-efficiently adapt pre-trained image-based visual foundation models to video person ReID tasks. To bridge visual misalignment in cross-platform tasks, we propose to transform the fundamental visual alignment problem into visual-semantic alignment based on CLIP. Specifically, we design a simple baseline method, named FT-CLIP, through fine-tuning the Image Encoder of CLIP. A two-stage training strategy is employed to optimize our approach. ID-specific description tokens are learned from samples originating from various platforms in the first training stage. Then in the second stage, visual features extracted from different platforms are aligned with the semantic features obtained through the learned

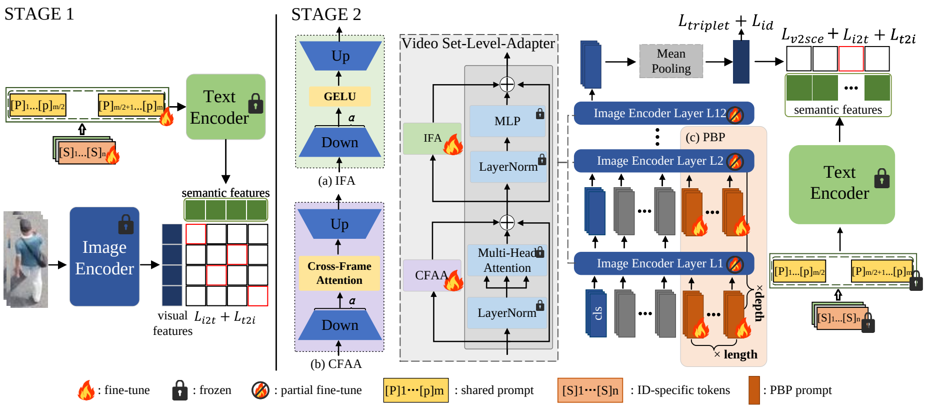

Fig. 3: Overview of our proposed framework. A two-stage training strategy is employed to optimize our approach: ID-specific description and shared text prompt are learned in stage one (left) while freezing Image Encoder and Text Encoder; Video Set-Level-Adapter and PBP are introduced and trained in the second stage (right) while freezing other parameters.

description tokens. Our work shows that FT-CLIP with the constraint of visual-semantic alignment yields competitive performance, but it is not parameter efficient and ignores inter-frame information. Therefore, we propose the Video Set-Level-Adapter for efficient model tuning, termed as VSLA-CLIP, which outperforms FT-CLIP while utilizing fewer parameters. To further bridge the semantic gap in cross-platform tasks, we propose a prompt-based approach called Platform-Bridge Prompt (PBP).

4.1 Revisiting CLIP-ReID

CLIP-ReID [25] is the pioneering approach that employs pre-trained vision-language models for image-based ReID. CLIP [29] relies on text labels to generate text descriptions. However, the labels in ReID tasks are indexes rather than specific text, which lacks the ability to depict detailed information about the corresponding persons. To solve this problem, CLIP-ReID uses a series of ID-specific learnable tokens to learn text descriptions and adapts a two-stage optimization strategy.

In the first training stage, only ID-specific tokens are optimized to learn text descriptions for each ID. Text $\mathcal{TD}$ that feeds into Text-Encoder $\mathbf{E}_t(\cdot)$ is "a photo of $[\mathbf{X}]_1 ... [\mathbf{X}]_{\mathrm{M}}$ person", where $[\mathbf{X}]_i$ is the learnable tokens. Text embedding $\mathbf{T}$ and image embedding $\mathbf{I}$ are obtained by:

$$\mathbf{T} = \mathbf{E}_t(\mathcal{TD}), \quad \mathbf{I} = \mathbf{E}_i(\mathcal{I}), \tag{1}$$

where $\mathbf{E}_i(\cdot)$ is the Image Encoder. The image-to-text contrastive loss $\mathcal{L}_{i2t}$ and text-to-image contrastive loss $\mathcal{L}_{t2i}$ are used to optimize $[\mathbf{X}]_1 ... [\mathbf{X}]_{\mathrm{M}}$. Since there are samples with the same ID in a batch, $\mathcal{L}_{t2i}$ in CLIP-ReID is defined as:

$$\mathcal{L}_{t2i}(y_i) = \frac{-1}{|P(y_i)|} \sum_{p \in P(y_i)} \log \frac{\exp(s(\mathbf{I}_p, \mathbf{T}_{y_i}))}{\sum_{a=1}^{B} \exp(s(\mathbf{I}_a, \mathbf{T}_{y_i}))}, \tag{2}$$

where $\mathbf{T}_{y_i}$ represents the text embedding of ID-y_i, $P(y_i) = \{p \in \{1...B\}, y_p = y_i\}$ is the set of positive samples for $\mathbf{T}_{y_i}$ and B represents the batch size. $\mathcal{L}_{i2t}$ is similar to $\mathcal{L}_{t2i}$. The overall loss function of stage one $\mathcal{L}_{stage1}$ is as follows:

$$\mathcal{L}_{stage1} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i}. \tag{3}$$

In the second stage, the ID-specific tokens and Text-Encoder are frozen. Triplet loss $\mathcal{L}_{tri}$ [31], identity loss $\mathcal{L}_{id}$, and image-to-text cross-entropy loss $\mathcal{L}_{i2tce}$ are used to optimize CLIP Image Encoder. The $\mathcal{L}_{i2tce}$ is defined as follows:

$$\mathcal{L}_{i2tce}(y) = \sum_{k=1}^{N} -q_k \log \frac{\exp(s(\mathbf{I}_y, \mathbf{T}_{y_k}))}{\sum_{y_a=1}^{N} \exp(s(\mathbf{I}_y, \mathbf{T}_{y_a}))}, \tag{4}$$

where q_k denotes smooth label [33] in the target distribution of the k_{th} ID, s represents cosine similarity, and N is the number of identities.

4.2 Visual-Semantic Alignment

We propose to transform the fundamental challenge of cross-platform visual alignment into visual-semantic alignment, and explore the efficacy of fine-tuning to adapt CLIP to video-based ReID tasks with visual-semantic alignment, named the model FT-CLIP. As shown in Fig. 3 (left), learnable ID-specific description tokens $[\mathbf{S}]_\mathrm{i}$ and shared text prompts $[\mathbf{P}]_\mathrm{i}$ are inserted into the Text-Encoder. All the tokens that feed into the Text-Encoder are concatenated as "$[[\mathbf{P}]_1...[\mathbf{P}]_{\mathrm{n}/2} : [\mathbf{S}]_1...[\mathbf{S}]_\mathrm{M} : [\mathbf{P}]_{\mathrm{n}/2+1}...[\mathbf{P}]_\mathrm{n}]$". Semantic features $\mathbf{T}$ can be obtained by:

$$\mathbf{T} = \mathbf{E}_t([[\mathbf{P}]_1...[\mathbf{P}]_{\mathrm{n}/2} : [\mathbf{S}]_1...[\mathbf{S}]_\mathrm{M} : [\mathbf{P}]_{\mathrm{n}/2+1}...[\mathbf{P}]_\mathrm{n}]), \tag{5}$$

where $[\cdot : \cdot]$ represents the concatenating operation, the dimensions of $[\mathbf{P}]_\mathrm{i}$ and $[\mathbf{S}]_\mathrm{i}$ are the same as that of the word embedding.

Inspired by CLIP-ReID [25], we adopt a two-stage optimization strategy. In the first optimization stage, we freeze both the Image Encoder and Text Encoder, using loss function $\mathcal{L}_{stage1}$ in Eq. (3) to optimize ID-specific description tokens and the shared text prompts. In the second optimization stage, Image Encoder is trained to align the video embeddings to semantic features. Given a video sample $\mathcal{V}_i \in \mathbb{R}^{T\times H\times W\times 3}$ with T frames, the CLIP image encoder encodes the T frames independently and mean-pooling is used to fuse the frame embeddings. Visual embeddings $\mathbf{V}_i$ can be obtained by:

$$\mathbf{V}_i = \frac{1}{T}\sum_{j}^{T} \mathbf{E}_i(\mathcal{V}_{ij}), \tag{6}$$

where $\mathcal{V}_{ij}$ represents the j_{th} frame of $\mathcal{V}_i$. The visual to semantic cross-entropy loss $\mathcal{L}_{v2sce}$, $\mathcal{L}_{i2t}$ and $\mathcal{L}_{t2i}$ are adopted to align visual embeddings to semantic features. $\mathcal{L}_{v2sce}$ is similar to $\mathcal{L}_{i2tce}$, defined as:

$$\mathcal{L}_{v2sce}(i) = \sum_{k=1}^{N} -q_k \log \frac{\exp(s(\mathbf{V}_i, \mathbf{T}_{y_k}))}{\sum_{y_j=1}^{N} \exp(s(\mathbf{V}_i, \mathbf{T}_{y_j}))}, \tag{7}$$

where q_k represents the soft label in the target distribution, and N is the number of identities. Meanwhile, triplet loss $\mathcal{L}_{tri}$ with soft-margin and ID loss $\mathcal{L}_{id}$ are also used:

$$\mathcal{L}_{tri} = \max(d_p - d_n + \theta, 0), \tag{8}$$

$$\mathcal{L}_{id} = \sum_{k=1}^{N} -q_k \log(p_k), \tag{9}$$

where θ is the soft-margin of $\mathcal{L}_{tri}$, p_k represents ID prediction logits of class k, d_p and d_n are feature distances of positive pair and negative pair. The overall loss $\mathcal{L}_{stage2}$ is defined as follows:

$$\mathcal{L}_{stage2} = \mathcal{L}_{v2sce} + \beta\mathcal{L}_{tri} + \gamma\mathcal{L}_{id} + \delta\mathcal{L}_{i2t} + \epsilon\mathcal{L}_{t2i}, \tag{10}$$

where β, γ, δ and ϵ balance the importance of the relative losses.

4.3 Video Set-Level-Adapter for Efficient Model Tuning

Video ReID requires the model to learn appearance representation in both intra-frame and inter-frames. We present a novel perspective, where a video sample is regarded as a frame set $\mathcal{S}_i = \{\mathcal{V}_{ij} | j = 1, 2, ..., n\}$ consisting of independent frames, and propose an efficient Video Set-Level-Adapter (VSLA) module. The VSLA consists of two components: an Intra-Frame Adapter (IFA, Fig. 3 (a)) and a Cross-Frame Attention Adapter (CFAA, Fig. 3 (b)). IFA is designed to parameter-efficiently adapt the pre-trained visual foundation model to downstream tasks, it takes raw frames as input and provides image-level appearance representation. CFAA takes a set of frames as input, aggregating the inter-frame complementary information for more powerful video-level representations.

IFA consists of two mapping matrices in a bottleneck structure. It runs in parallel with MLP blocks within each layer of the Image Encoder. As shown in Fig. 3, the Image Encoder in CLIP (ViT-Base-16) consists of alternating layers of Multi-Head Self-Attention (MSA) [34], Multi-Layer Perceptron (MLP) and LayerNorm (LN), which can be formulated as:

$$\mathbf{x}_i^{'} = \mathrm{MSA}(\mathrm{LN}(\mathbf{x}_{i-1})) + \mathbf{x}_{i-1}, \tag{11}$$

$$\mathbf{x}_i = \mathrm{MLP}(\mathrm{LN}(\mathbf{x}_i^{'})) + \mathbf{x}_i^{'}. \tag{12}$$

We denote the input of IFA as $\mathbf{x}_i' \in \mathbb{R}^{T\times(N+1)\times D}$, where $N = HW/P^2$, D represents the dimension and T is the number of frames. The down-projection layer $\mathbf{W}_{down}$ projects $\mathbf{x}_i'$ to $\mathbf{x}_i'' \in \mathbb{R}^{T\times(N+1)\times\alpha}$, where α is a hyper-parameter. Then $\mathbf{x}_i''$ goes through a GELU σ and up-projection layer $\mathbf{W}_{up}$. The process can be formulated as:

$$\mathrm{IFA}(\mathbf{x}_i^{'}) = \sigma(\mathbf{x}_i^{'}\mathbf{W}_{down})\mathbf{W}_{up}, \tag{13}$$

$$\mathbf{x}_i = \mathrm{MLP}(\mathrm{LN}(\mathbf{x}_i^{'})) + \mathbf{x}_i^{'} + \mathrm{IFA}(\mathbf{x}_i^{'}). \tag{14}$$

Unlike LoRA [19], which adds trainable pairs of rank decomposition matrices in parallel to every pre-existing weight matrix, IFA is solely in parallel with MLP. Therefore, adopting IFA results in far fewer parameters, accounting for only 5.5% ($\alpha = 256$) of the whole Image Encoder (ViT-Base-16).

CFAA is also a bottleneck architecture with a cross-frame attention layer in the middle. Our model $\mathbf{M}(\cdot)$ with CFAA is immune to frame ordering [5], which can be formulated as:

$$\mathbf{M}(\{\mathcal{V}_{ij}|j = 1, 2, ..., n\}) = \mathbf{M}(\{\mathcal{V}_{i\pi(j)}|j = 1, 2, ..., n\}), \tag{15}$$

where π is any permutation [41]. We denote the input of CFAA as $\mathbf{x}_{i-1} \in \mathbb{R}^{T\times(N+1)\times D}$, the down-projection layer projects $\mathbf{x}_{i-1}$ to $\mathbf{x}'_{i-1} \in \mathbb{R}^{T\times(N+1)\times\alpha}$. The cross-frame attention layer has the same structure as Multi-Head Self-Attention (MSA) [34]. To aggregate the complementary information among T frames, we reshape the input of cross-frame attention layer $\mathbf{x}'_{i-1}$ to $\mathbf{x}'^{\mathsf{T}}_{i-1} \in \mathbb{R}^{(N+1)\times T\times\alpha}$, and the attention is done in the second dimension of $\mathbf{x}'^{\mathsf{T}}_{i-1}$, thus enabling visual information to exchange across frames. Then, we restore the output of cross-frame attention layer from $\mathbf{x}''^{\mathsf{T}}_{i-1} \in \mathbb{R}^{(N+1)\times T\times\alpha}$ to $\mathbf{x}''_{i-1} \in \mathbb{R}^{T\times(N+1)\times\alpha}$, with $\mathbf{x}''_{i-1}$ passing through up-projection layer. For CFAA, $\mathbf{x}_i^{'}$ in Eq.(11) can be obtained by:

$$\mathbf{x}_i^{'} = \mathrm{MSA}(\mathrm{LN}(\mathbf{x}_{i-1})) + \mathbf{x}_{i-1} + \mathrm{CFAA}(\mathbf{x}_{i-1}). \tag{16}$$

4.4 Platform-Bridge Prompt

Making visual embeddings align with semantic features could effectively alleviate the feature misalignment in cross-platform ReID tasks, but yet to be improved. We additionally introduce Platform-Bridge Prompt (PBP) to bridge platform differences further. PBP is designed to guide model focusing on platform differences. As illustrated in Fig. 3, we add a series of platform-specific learnable prompts in the Image Encoder. Specifically, there are only two sets of prompts, one corresponding to the ground platform and the other to the UAV platform. Applying PBP can be viewed as changing the inputs of each MSA layer in Vision Transformer (ViT [9]). We denote the inputs of the MSA layer as $\mathbf{h} \in \mathbb{R}^{(N+1)\times D}$, where $N = HW/P^2$ and D represents the dimension. The MSA layer with PBP can be formulated as follows,

$$f_k(\mathbf{h}, \mathbf{p}_k) = \begin{cases} MSA_k([\mathbf{h} : \mathbf{p}_k^{ground}]) & \text{if } k < d \text{ and } \mathbf{h} \in Set^{ground} \\ MSA_k([\mathbf{h} : \mathbf{p}_k^{uav}]) & \text{if } k < d \text{ and } \mathbf{h} \in Set^{uav} \\ MSA_k(\mathbf{h}) & \text{if } k \geq d, \end{cases} \tag{17}$$

where $\mathbf{p}_k^{ground} \in \mathbb{R}^{l\times D}, \mathbf{p}_k^{uav} \in \mathbb{R}^{l\times D}$, d and l are the depth and length of PBP, [:] denotes the concatenation operation, MSA_k represents the k_{th} MSA layer in Image Encoder, Set^{uav} and Set^{ground} are two sets containing the samples from the UAV and the samples from the ground platform respectively.

Table 2: Comparison with state-of-the-art methods. † represents the model initialized by the weight of CLIP [29] released by OpenAI, and ‡ represents the model initialized by weight of ViFi-CLIP [30]. We use **bold** to indicate the best results of our methods, and underlines to highlight the best results of other methods. On all datasets, our method outperforms the comparisons significantly.

Method	MARS		LS-VID		iLIDS	G2A-VReID	
	mAP	rank-1	mAP	rank-1	rank-1	mAP	rank-1
STA [12]	80.8	86.3	-	-	-	-	-
M3D [23]	74.1	84.4	40.1	57.7	74	-	-
GLTR [24]	78.5	87.0	44.3	63.1	86	-	-
VRSTC [16]	82.3	88.5	-	-	83.4	-	-
AP3D [13]	85.1	90.1	73.2	84.5	88.7	67.7	57.5
STGCN [39]	83.7	90.0	-	-	-	-	-
MGH [38]	85.8	90.0	-	-	85.6	76.7	69.9
MG-RAFA [43]	85.9	88.8	-	-	88.6	-	-
AFA [7]	82.9	90.2	-	-	88.5	-	-
TCLNet [17]	85.1	89.8	70.3	81.5	86.6	65.4	54.7
STRF [1]	86.1	90.3	-	-	89.3	-	-
GRL [28]	84.8	91.0	-	-	90.4	52.8	41.4
DenseIL [14]	87.0	90.8	-	-	92	-	-
BiCnet-TKS [18]	86.0	90.2	75.1	84.6	-	63.4	51.7
PSTA [37]	85.8	91.5	-	-	-	64.6	54.5
STMN [10]	84.5	90.5	69.2	82.1	91.5	66.7	56.1
PiT [42]	-	90.2	-	-	92.1	76.3	67.7
SINet [2]	86.2	91.0	79.6	87.4	92.5	74.5	65.6
LSTRL [27]	86.8	91.6	82.4	89.8	92.2	-	-
FT-CLIP‡	88.00	91.62	84.07	90.77	94.00	78.11	69.32
VSLA-CLIP†	88.22	90.91	84.05	90.54	**95.33**	79.14	71.64
VSLA-CLIP‡	**88.60**	**91.82**	**85.20**	**91.66**	**95.33**	**79.70**	**72.55**

5 Experiments

In this section, we first introduce the evaluation protocols and implementation details. Subsequently, we compare our proposed methods with state-of-the-art algorithms. Finally, ablation studies are conducted to investigate the contribution of each component.

5.1 Datasets and Evaluation Metrics

We conduct experiments on our G2A-VReID and three widely used video person ReID datasets, *i.e.*, iLIDS [36], Mars [44], and LS-VID [22]. For G2A-VReID, we roughly divide 2788 identities into training and test sets at a ratio of $1:2$, similar to that in LS-VID [22]. Therefore, there are 930 identities with 1860 tracklets in training set and 1858 identities with 3716 tracklets in the testing set. During the evaluation, we keep the cross-camera search paradigm in ReID task [15, 22, 36, 44]. Query and gallery are composed of video sequences from the ground and UAV cameras respectively, making G2A-VReID more challenging than other datasets.

We employ two standard metrics to evaluate the performance of our model, *i.e.*, Cumulative Matching Characteristic(CMC) at Rank-1 and mean average precision (mAP).

5.2 Implementation Details

ViT-Base-16 [29] is selected as the Image Encoder. The initial weights are chosen as that of ViFi-CLIP [30], whose Image Encoder and Text Encoder have been fine-tuned on the extensive action recognition dataset Kinetics-400 [20]. Sparse temporal sampling strategy [35] is used to generate a clip containing 8 frames, with each frame resized to 256×128. We randomly disrupt the order of the frames in each clip. Each batch has 32 clips corresponding to 8 identities. Adam [21] optimizer is used in both stages. In the first training stage, we optimize the ID-specific description tokens and shared text prompts with a learning rate of 3.5×10^{-4}, while freezing other parameters. In the second training stage, we adopt the initial learning rate 5×10^{-6} with decaying by 0.1 and 0.01 at the 60_{th} and 90_{th} epoch for FT-CLIP, and the initial learning rate 1×10^{-4} with decaying by 0.1 and 0.01 at the 60_{th} and 90_{th} epoch for VSLA-CLIP. The margin θ of triplet loss in Eq. (8) is set as 0.3, the β, γ, δ and ϵ in Eq. (10) are 1.0, 0.25, 1.0 and 1.0, respectively. Each image is padded with 10 pixels and augmented with random cropping, horizontal flipping, and erasing [45].

5.3 Comparison with State-of-the-Art Methods

On G2A-VReID Dataset. We comprehensively evaluate nine state-of-the-art methods [2, 10, 13, 17, 18, 28, 37, 38, 42] on G2A-VReID, and report the results in Tab. 2. As can be seen that, MGH [38] and PiT [42] showed superior performances on our G2A-VReID dataset, *i.e.* MGH achieves 76.7% on mAP and 69.9% on Rank-1, surpassing other models with a large margin. We attribute this to the careful visual alignment strategy adopted by MGH and PiT, which involves splitting the full image into vertical or horizontal stripes and aiming to align the stripes. This strategy mitigates the challenges of self-occlusion inherent in the UAV perspective. Our method, extracting description tokens for each person and aligning visual embeddings with semantic features, effectively solves the cross-platform visual misalignment problem. Our VSLA-CLIP‡ achieves 79.70% mAP and 72.55% Rank-1 on G2A-VReID, surpassing MGH by 3.0% at mAP and 2.65% at Rank-1.

On All Video ReID Dataset. As shown in Tab. 2, all the variants of our methods with aligning visual embeddings to semantic features, show consistent improvement on all datasets. Especially, our method achieves 85.20% mAP and 91.66% Rank-1 on the challenging LS-VID dataset, which greatly improves the mAP by 2.80% and the Rank-1 by 1.86% compared with the state-of-the-art LSTRL [27]. **2)** Models initialized by weights of ViFi-CLIP (ViFi-weight) are marked as ‡, and it is effective compared with the original model weights released by Open AI (marked as †). VSLA-CLIP initialized with ViFi-Weight improves the performance significantly by 1.15% mAP on LS-VID. **3)** It is worth noting that VSLA-CLIP shows better performance than fine-tuning the whole Image Encoder (FT-CLIP), with far fewer tunable parameters. Specifically, VSLA-CLIP‡ outperforms the FT-CLIP‡ by 1.59% mAP on G2A-VReID with tuning parameters (14.5M vs 88.0M).

Table 3: Effectiveness of proposed components and comparison of the number of tunable parameters. **baseline** represents training FT-CLIP‡ without Lv2sce in Eq.(7), **VSA** is Visual-Semantic Alignment, IFA represents Intra-Frame Adapter, CFAA is Cross-Frame Attention Adapter and PBP is Platform Bridge Prompt.

Methods	Overall Param(M)	Tunable Param (M)	LS-VID		G2A-VReID	
			mAP	rank-1	mAP	rank-1
AP3D [13]	34.0	24.9	73.2	84.5	67.7	57.5
BiCnet-TKS [18]	33.7	29.3	75.1	84.6	63.4	51.7
STMN [10]	90.9	87.0	69.2	82.1	66.7	56.1
SINet [2]	33.7	27.3	79.6	87.1	74.5	65.6
baseline	86.1	86.1	76.10	84.26	72.80	63.62
baseline+VSA (FT-CLIP‡)	127.4	88.0	84.07	90.77	78.11	69.32
IFA	90.8	**4.7**	77.31	84.86	73.82	65.12
IFA+VSA	132.1	6.6	84.16	90.94	79.01	71.67
IFA+VSA+CFAA (VSLA-CLIP‡)	140.0	14.5	**85.20**	**91.66**	79.70	72.55
IFA+VSA+CFAA+PBP	140.0	14.5	-	-	**81.29**	**74.27**

Our experiments show that adapting pre-trained image-based models to video ReID tasks with the Video Set-Level-Adapter is both effective and efficient, setting a new baseline method for research endeavors in this field.

5.4 Ablation Study

To demonstrate the effectiveness of our proposed components in Sec.4, we conduct ablation studies and compare our method with four other methods.

Effectiveness of Visual-Semantic Alignment. To verify the effectiveness of Visual-Semantic Alignment, we first fine-tune the Image Encoder by directly using two common losses ($\mathcal{L}_{tri}$ and $\mathcal{L}_{id}$ in Eq.(10)) in ReID task, and set this model as our baseline. As shown in Tab. 3, Visual-Semantic Alignment is effective for both finetuning-based methods (FT-CLIP‡ vs. baseline) and adapter-based methods (IFA+VSA vs. IFA). In addition, we conduct ablation experiments to analyze three loss functions for visual-semantic alignment. As shown in Tab. 5, when $\mathcal{L}_{i2t}$, $\mathcal{L}_{t2i}$ and $\mathcal{L}_{v2sce}$ are used jointly, our model achieves the best results on LS-VID.

Effectiveness of Video Set-Level-Adapter. Our goal for proposing the Video Set-Level-Adapter is to efficiently adapt pre-trained image-based visual foundation mode to video-based ReID tasks. Considering that the Video Set-Level-Adapter (VSLA) contains two modules, *i.e.*, an Intra-Frame Adapter (IFA) and a Cross-Frame Attention Adapter (CFAA), we perform ablation experiments separately to verify the effectiveness of each module. As shown in Tab. 3, IFA surpasses the full fine-tuned baseline method (77.31% vs. 76.10% mAP on LS-VID) with significantly less number of tunable parameters (4.7M vs. 86.1M). In addition, CFAA further improves model performance (85.20% vs. 84.16% mAP on LS-VID) while also using a small number of tunable parameters, which indicates that regarding video sequences as a set is effective in Video-based ReID tasks, providing a new solution for adapting Image-based foundation models to video-based tasks.

We also analyze the hyper-parameter α introduced in Sec. 4.3, which determines model's complexity and the number of training parameters. We set α to be

Table 4: Effect of α of Intra-Frame Adapter and Cross-Frame Attention Adapter on LS-VID. TP represents the tunable parameter.

α	TP (M)		LS-VID	
	IFA	CFAA	mAP	rank-1
64	1.2	1.4	79.58	86.71
128	2.4	3.2	83.64	90.00
256	4.7	7.9	**85.20**	**91.66**
384	7.1	14.2	85.09	91.49

Table 5: Ablation experiments for the losses used for Visual-Semantic Alignment on LS-VID.

$\mathcal{L}_{v2sce}$	$\mathcal{L}_{i2t}$	$\mathcal{L}_{t2i}$	mAP
✓			84.73
	✓		84.29
		✓	84.45
	✓	✓	84.71
✓	✓	✓	**85.20**

64, 128, 256, and 384 respectively. As presented in Tab. 4, the performances tend to improve with increasing α, and achieves the best mAP at $\alpha = 256$. Therefore, we fix α to be 256 for other datasets. At this setting, the VSLA module contains only approximately 12.6 million parameters parameters, and VSLA-CLIP achieves 85.20% mAP on LS-VID, surpassing FT-CLIP by 1.13%.

Effectiveness of PBP. The Platform Bridge Prompt (PBP) offers meticulous instructions to enable models to discern differences across platforms. It adeptly steers the model towards obtaining precise and targeted information, thereby bridging the semantic gap in visual features. The depth d and length l are two hyper-parameters in PBP, which are introduced in Sec. 4.4. To analyze the impact of these two parameters on the model, we use grid-search to explore the impact of different value combinations on the model performance. The results for various parameter combinations of the model are presented in Fig. 4, and the optimal performance is achieved when $d = 3$ and $l = 16$.

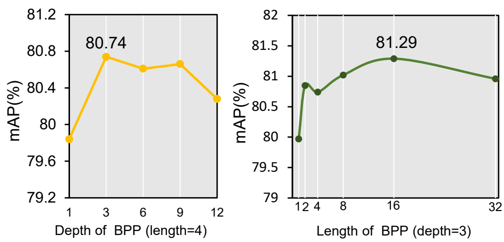

Fig. 4: Analysis on the depth and length of PBP on our G2A-VReID.

6 Conclusion

In this paper, we construct a large-scale benchmark dataset for cross-platform video person ReID, which contains 5,576 tracklets of 2788 IDs and can serve as a potential complement to current ground surveillance system. Besides, we also propose a baseline method solving cross-platform visual misalignment problem by transforming the visual alignment problem into visual-semantic alignment through the vision-language model (*i.e.*, CLIP) and using platform-specific prompts. To efficiently and effectively adapt the pre-trained image-based visual foundation model to Video ReID, We propose a Video Set-Level-Adapter module, which aggregates the inter-frame complementary information for more powerful video-level representations with only 12.6 million trainable parameters. Experimental results demonstrate that our proposed methods achieve state-of-the-art performance and will be a new trend for cross-platform video ReID tasks.

References

1. Aich, A., Zheng, M., Karanam, S., Chen, T., Roy-Chowdhury, A.K., Wu, Z.: Spatio-temporal representation factorization for video-based person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 152–162 (2021) 2, 4, 11
2. Bai, S., Ma, B., Chang, H., Huang, R., Chen, X.: Salient-to-broad transition for video person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7339–7348 (2022) 1, 3, 11, 12, 13
3. Baltieri, D., Vezzani, R., Cucchiara, R.: 3dpes: 3d people dataset for surveillance and forensics. In: Joint Acm Workshop on Human Gesture Behavior Understanding (2011) 2, 6
4. Bao, L., Ma, B., Chang, H., Chen, X.: Preserving structural relationships for person re-identification. In: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 120–125. IEEE (2019) 4
5. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8126–8133 (2019) 10
6. Chen, D., Li, H., Xiao, T., Yi, S., Wang, X.: Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1169–1178 (2018) 4
7. Chen, G., Rao, Y., Lu, J., Zhou, J.: Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. pp. 660–676. Springer (2020) 4, 11
8. Chung, D., Tahboub, K., Delp, E.J.: A two stream siamese convolutional neural network for person re-identification. In: Proceedings of the IEEE international conference on computer vision. pp. 1983–1991 (2017) 4
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 10
10. Eom, C., Lee, G., Lee, J., Ham, B.: Video-based person re-identification with spatial and temporal memory networks. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12016–12025 (2021). https://doi.org/10.1109/ICCV48922.2021.01182 3, 4, 11, 12, 13
11. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1933–1941 (2016) 4
12. Fu, Y., Wang, X., Wei, Y., Huang, T.: Sta: Spatial-temporal attention for large-scale video-based person re-identification. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8287–8294 (2019) 1, 4, 11
13. Gu, X., Chang, H., Ma, B., Zhang, H., Chen, X.: Appearance-preserving 3d convolution for video-based person re-identification. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 228–243. Springer (2020) 3, 4, 11, 12, 13
14. He, T., Jin, X., Shen, X., Huang, J., Chen, Z., Hua, X.S.: Dense interaction learning for video-based person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1490–1501 (2021) 4, 11

15. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Image Analysis: 17th Scandinavian Conference, SCIA 2011, Ystad, Sweden, May 2011. Proceedings 17. pp. 91–102. Springer (2011) 2, 3, 6, 11
16. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Vrstc: Occlusion-free video person re-identification. In: CVPR. pp. 7183–7192 (2019) 11
17. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Temporal complementary learning for video person re-identification. In: ECCV. pp. 388–405 (2020) 4, 11, 12
18. Hou, R., Chang, H., Ma, B., Huang, R., Shan, S.: Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2014–2023 (June 2021) 2, 3, 4, 11, 12, 13
19. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9 10
20. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 12
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 12
22. Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S.: Global-local temporal representations for video person re-identification. In: ICCV. pp. 3958–3967 (2019) 3, 5, 6, 11
23. Li, J., Zhang, S., Huang, T.: Multiscale 3d convolution network for video based person reidentification. In: AAAI. pp. 8618–8625 (2019) 1, 4, 11
24. Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S.: Global-local temporal representations for video person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3958–3967 (2019) 1, 11
25. Li, S., Sun, L., Li, Q.: Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels. arXiv preprint arXiv:2211.13977 (2022) 2, 7, 8
26. Liu, H., Jie, Z., Jayashree, K., Qi, M., Jiang, J., Yan, S., Feng, J.: Video-based person re-identification with accumulative motion context. IEEE transactions on circuits and systems for video technology **28**(10), 2788–2802 (2017) 4
27. Liu, X., Zhang, P., Lu, H.: Video-based person re-identification with long short-term representation learning. In: International Conference on Image and Graphics. pp. 55–67. Springer (2023) 11, 12
28. Liu, X., Zhang, P., Yu, C., Lu, H., Yang, X.: Watching you: Global-guided reciprocal learning for video-based person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13334–13343 (2021) 11, 12
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 2, 7, 11, 12
30. Rasheed, H., khattak, M.U., Maaz, M., Khan, S., Khan, F.S.: Finetuned clip models are efficient video learners. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 11, 12

31. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015) 8
32. Song, G., Leng, B., Liu, Y., Hetang, C., Cai, S.: Region-based quality estimation network for large-scale person re-identification. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018) 4
33. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016) 8
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 9, 10
35. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016) 12
36. Wang, X., Zhao, R.: Person re-identification: System design and evaluation overview. In: Person Re-Identification, pp. 351–370. Springer (2014) 2, 3, 5, 6, 11
37. Wang, Y., Zhang, P., Gao, S., Geng, X., Lu, H., Wang, D.: Pyramid spatial-temporal aggregation for video-based person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12026–12035 (2021) 11, 12
38. Yan, Y., Qin, J., Chen, J., Liu, L., Zhu, F., Tai, Y., Shao, L.: Learning multi-granular hypergraphs for video-based person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2899–2908 (2020) 11, 12
39. Yang, J., Zheng, W.S., Yang, Q., Chen, Y.C., Tian, Q.: Spatial-temporal graph convolutional network for video-based person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3289–3299 (2020) 11
40. Yin, J., Wu, A., Zheng, W.S.: Fine-grained person re-identification. International Journal of Computer Vision **128**(6), 1654–1672 (Jun 2020). https://doi.org/10.1007/s11263-019-01259-0, https://doi.org/10.1007/s11263-019-01259-0 3
41. Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J.: Deep sets. Advances in neural information processing systems **30** (2017) 10
42. Zang, X., Li, G., Gao, W.: Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. IEEE Transactions on Industrial Informatics **18**(12), 8776–8785 (2022). https://doi.org/10.1109/TII.2022.3151766 4, 11, 12
43. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10407–10416 (2020) 11
44. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: ECCV. pp. 868–884 (2016) 2, 3, 4, 5, 6, 11
45. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 13001–13008 (2020) 12

46. Zhou, Z., Huang, Y., Wang, W., Wang, L., Tan, T.: See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4747–4756 (2017) 4