CLIMB-ReID: A Hybrid CLIP-Mamba Framework for Person Re-identification

Chenyang Yu[1], Xuehu Liu[2], Jiawen Zhu[1], Yuhao Wang[3], Pingping Zhang[3]*, Huchuan Lu[1,3]

[1] School of Information and Communication Engineering, Dalian University of Technology, Dalian, China
[2] School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China
[3] School of Future Technology, School of Artificial Intelligence, Dalian University of Technology, Dalian, China
{yuchenyang, jiawen, 924973292}@mail.dlut.edu.cn;liuxuehu@whut.edu.cn;{zhpp, lhchuan}@dlut.edu.cn

Abstract

Person Re-IDentification (ReID) aims to identify specific persons from non-overlapping cameras. Recently, some works have suggested using large-scale pre-trained vision-language models like CLIP to boost ReID performance. Unfortunately, existing methods still struggle to address two key issues simultaneously: efficiently transferring the knowledge learned from CLIP and comprehensively extracting the context information from images or videos. To address these issues, we introduce CLIMB-ReID, a pioneering hybrid framework that synergizes the impressive power of CLIP with the remarkable computational efficiency of Mamba. Specifically, we first propose a novel Multi-Memory Collaboration (MMC) strategy to transfer CLIP's knowledge in a parameter-free and prompt-free form. Then, we design a Multi-Temporal Mamba (MTM) to capture multi-granular spatiotemporal information in videos. Finally, with Importance-aware Reorder Mamba (IRM), information from various scales is combined to produce robust sequence features. Extensive experiments show that our method outperforms other state-of-the-art methods on both image and video person ReID benchmarks.

Code — https://github.com/AsuradaYuci/CLIMB-ReID

Introduction

Person Re-IDentification (ReID) (Gao et al. 2023; Zhang et al. 2020; Liu et al. 2021b; Zhang et al. 2024) aims to identify an individual from images or videos captured by non-overlapping camera systems. Various approaches have been developed over the past decade, including methods based on CNNs (Dai et al. 2019; Gao et al. 2024; Chen et al. 2022; Zhu et al. 2024a; Zou et al. 2023; Wang et al. 2022) and Transformers (Zhang et al. 2021a; Chen et al. 2024; Diao et al. 2024; Zhu et al. 2023; Diao et al. 2025). While these methods demonstrate promising performance, they still encounter challenges in achieving robust feature representations. Recently, some works (Li et al. 2024; Li and Gong 2023; Diao et al. 2025) have utilized large-scale pre-trained vision language models like Contrastive Language-Image Pretraining (CLIP) (Radford et al. 2021), to enhance the ReID performance. Unfortunately, current methods still struggle to tackle two main challenges: efficiently transferring knowledge from CLIP and sufficiently extracting contextual information from images or videos.

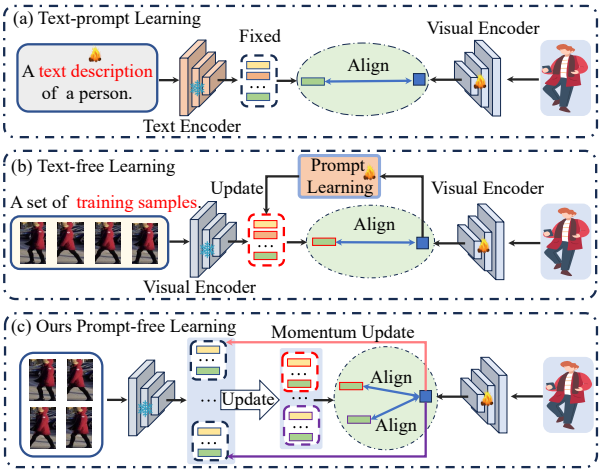

Figure 1: Comparison of typical CLIP-based ReID methods and our prompt-free framework. (a) Text-prompt learning. (b) Text-free learning. (c) Our prompt-free learning.

In fact, when applying CLIP to ReID, the first challenge is that person samples are typically annotated with one-hot labels, which lack semantic information. As shown in Fig. 1 (a), Li *et al.* (Li, Sun, and Li 2023) propose CLIP-ReID to address this issue by generating label-specific text features with prompt learning. Wang *et al.* (Wang et al. 2024a) further propose SVLL-ReID for generating diverse texts to enhance CLIP's performance in ReID. Given that CLIP's text and visual encoders are aligned in the feature space, Yu *et al.* (Yu et al. 2024) propose an one-stage text-free framework. As shown in Fig. 1 (b), it leverages the visual encoder to generate an identity-specific CLIP Memory to replace text descriptions. Although the above methods achieve significant performance improvements, the use of prompt learning inevitably introduces additional training parameters.

On the other hand, leveraging spatiotemporal information is crucial for video-based person ReID. Existing methods (Liu et al. 2024a; Wang et al. 2024d; Wu et al. 2022; Wang et al. 2024c; Yan et al. 2023) rely on the self-attention

*Corresponding author.

mechanism (Vaswani et al. 2017) to capture global spatiotemporal information. Although good performances are achieved, their exponential complexity results in significant computational overhead. Recently, Mamba (Gu and Dao 2023) has been favored for its good performance and low computational overhead in sequence modeling tasks, such as image classification (Wang et al. 2024b) and object segmentation (Yang, Xing, and Zhu 2024). Motivated by the above observation, we propose CLIMB-ReID, a novel framework that synergizes the impressive power of CLIP with the remarkable computational efficiency of Mamba.

Specifically, our CLIMB-ReID framework is composed of two key components: Multi-Memory Collaboration (MMC) and Multi-Temporal Mamba (MTM). Technically, we first design the MMC strategy to efficiently transfer knowledge from CLIP. As shown in Fig. 1 (c), we initialize multiple memories and select some distinct samples from the training batch to update these memories. Subsequently, the updated memories are employed to collaboratively guide the learning process of the visual encoder. By utilizing distinct memories to complementarily represent the same person, we can reduce intra-class variance, resulting in more robust features. We further design a MTM to capture multi-granular spatiotemporal information in videos. Specifically, we begin by slicing the video sequence at different strides along the temporal dimension. Unlike methods (Li et al. 2025; Zhu et al. 2024c) that simply add sequential scanning ways, we propose an Importance-aware Reorder Mamba (IRM) that reorders tokens in a sequence according to their importance. The reorder operation ensures that tokens with similar semantics are fed into IRM in adjacent order, thereby strengthening IRM's local modeling capabilities. Finally, information from various temporal scales is combined to produce robust sequence features. Experiments conducted on three video-based and two image-based person ReID benchmarks clearly demonstrate the effectiveness of our methods.

Our contributions can be summarized as follows:

- In this paper, we propose a novel framework named CLIMB-ReID for person ReID. To the best of our knowledge, this is the first use of Mamba to person ReID.
- We propose a Multi-Memory Collaboration (MMC) strategy to efficiently transfer the knowledge from CLIP.
- We propose a Multi-Temporal Mamba (MTM) to capture multi-granular spatiotemporal information in videos.
- Extensive experiments reveal that our method achieves superior performance on three video-based ReID datasets (MARS, LS-VID, and iLIDS-VID) and two image-based ReID datasets (Market-1501 and MSMT17).

Related Works

Person ReID with CLIP

To address person ReID, various approaches have been developed over the past decade, including methods based on CNNs (Zou et al. 2023; Gong et al. 2024; Shi et al. 2025; Diao et al. 2021) and Transformers (Zhang et al. 2021a; Gao et al. 2024; Zhu et al. 2023). For example, Wu *et al.* (Wu et al. 2022) propose a contextual alignment Transformer to capture temporal information for video-based ReID. Recently, the remarkable CLIP has demonstrated an exceptional ability in learning robust representations. Consequently, some methods have begun to transfer the knowledge from CLIP to the ReID task. To overcome the issue of missing text labels, Li *et al.* (Li, Sun, and Li 2023) design a two-stage training strategy for image-based person ReID. Wang *et al.* (Wang et al. 2024a) further propose to integrate self-supervision and pre-trained CLIP to improve the ReID performance. Yu *et al.* (Yu et al. 2024) propose an one-stage framework that extracts the identity-specific sequence feature to replace the text feature. Although the above methods have achieved great success, they inevitably introduce additional training stages or require additional prompt networks. To facilitate more efficient knowledge transfer from CLIP, this paper proposes a parameter-free Multi-Memory Collaboration (MMC) that is both text-free and prompt-free.

Mamba in Computer Vision

The newly proposed Mamba (Gu and Dao 2023) has garnered significant attention due to its remarkable potential for global perception with linear complexity. Inspired by the success of Mamba, several works have explored its effectiveness in computer vision. For example, Ma *et al.* (Ma, Li, and Wang 2024) propose a hybrid CNN-SSM block for biomedical image segmentation. Zhu *et al.* (Zhu et al. 2024c) propose a bidirectional Mamba framework for image classification. However, these Mamba-based methods still fall short of outperforming Transformer-based methods in ReID. Therefore, we propose a hybrid structure that leverages the strengths of both CLIP and Mamba to achieve improved performance with reduced computational complexity. On the other hand, Mamba is a recurrent model, and the order of the hidden states significantly impacts the performance of long-range dependency modeling. To address this direction-sensitive issue, Liu *et al.* (Liu et al. 2024b) introduce a cross-scan module to traverse the spatial domain. Li *et al.* (Li et al. 2025) further propose the VideoMamba framework to explore different scanning methods for video understanding. Different from the above methods, we improve the scanning method for person ReID by proposing an importance-aware reorder scanning strategy on multi-scale temporal inputs. The reorder operation ensures that tokens with similar semantics are fed into IRM in adjacent order, thereby strengthening the local modeling capabilities.

Our Method

As illustrated in Fig. 2, our proposed framework mainly includes two components: Multi-Memory Collaboration (MMC) and Multi-Temporal Mamba (MTM).

Preliminaries

Given a video sequence $V = \{I_t\}_{t=1}^{T}$ containing T frames, $I_t \in \mathbb{R}^{H\times W\times 3}$ represents the t-th frame, where H and W represent the number of height and width, respectively. CLIP's visual encoder f_θ is used to extract deep features. The visual encoder first divides each frame into N patches, then the frame-level feature $Z_t = \{z_t^{cls}; z_t^1; z_t^2; \cdots, z_t^N\} \in \mathbb{R}^{(1+N)\times D}$ is obtained after L Transformer layers, where

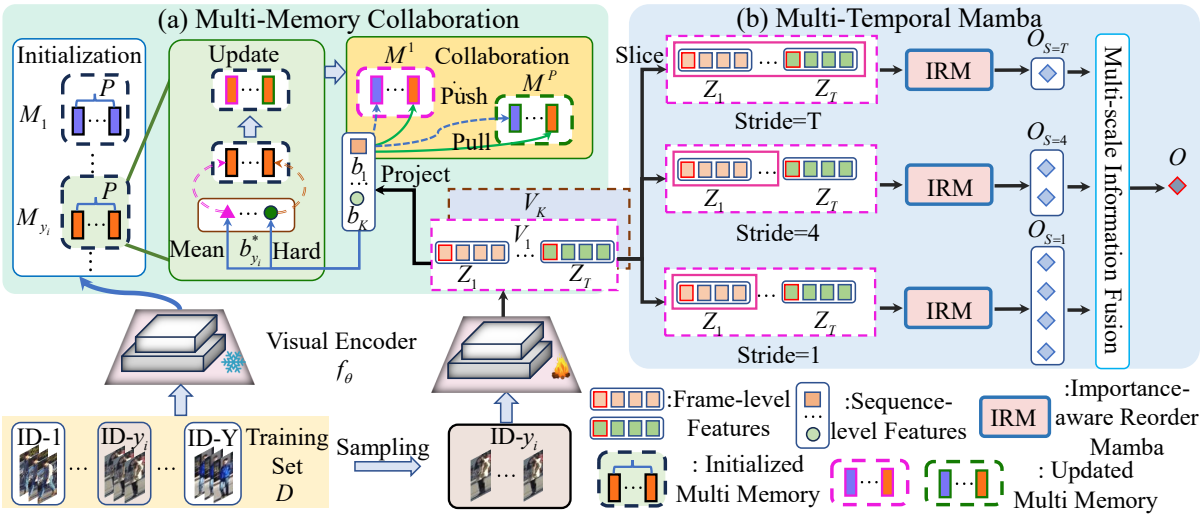

Figure 2: Illustration of the proposed CLIMB-ReID framework. The Multi-Memory Collaboration (MMC) is designed to efficiently transfer the knowledge from CLIP in a parameter-free form. Then, the Multi-Temporal Mamba (MTM) is proposed to capture multi-granular spatiotemporal information in videos. Finally, information from various temporal scales is combined to produce robust sequence features.

z_t^{cls} and z_t^n represent D-dimensional feature vectors of the class token and patch tokens, respectively. After that, a linear layer is employed to project the class token z_t^{cls} into a visual-language unified space denoted as $v_t \in \mathbb{R}^{1\times d}$. Finally, the temporal average pooling is used to derive the sequence-level feature b_{y_i}.

To obtain a text-free model, Yu *et al.* (Yu et al. 2024) employ the pre-trained f_θ for all training sequences to extract identity-specific sequence features denoted as CLIP-Memory. Then, the video-to-memory contrastive learning loss denoted by L_{V2M} is applied for model optimization,

$$L_{V2M}^{(y_i)} = -log\frac{\exp(< b_{y_i}, M_{y_i} >)}{\sum_{j=1}^{N_p}\exp(< b_{y_i}, M_j >)}, \tag{1}$$

where M_{y_i} represents the visual feature centroid of identity y_i. $< \cdot >$ represents the cosine similarity function. N_p represents the number of persons in a training batch.

Multi-Memory Collaboration

From Eq. 1 can be seen that the memory is used as an anchor for similarity comparison. Therefore, it plays an important role in the optimization process. To obtain this memory, CLIP-ReID learns a corresponding text description for each identity. However, the memory obtained by CLIP-ReID is fixed during training. TF-CLIP further utilizes prompt learning to update the memory online. However, this requires the introduction of additional training parameters. To solve the above problems, we propose the Multi-Memory Collaboration (MMC) to migrate CLIP's knowledge more efficiently.

Initialization. Given a video-based person ReID training set $\mathcal{D} = \{(V_i, y_i)\}_{i=1}^{N_s}$ with labels $y_i \in \{1, \cdots, Y\}$, the total number of videos is N_s. As shown in Fig. 2 (a), we traverse the entire training set by the pre-trained f_θ to build multiple memories. For each identity, we maintain P proxies in the memories. Specifically, once all the features belonging to the identity y_i are obtained, the average of them can represent the identity-specific feature $M_{y_i}^p$ to initialize each proxy p in the memories:

$$M_{y_i} = \frac{1}{N_i}\sum_{b\in y_i} b, \tag{2}$$

where N_i is the number of videos belonging to the identity y_i. Therefore, the initialized $M \in \mathbb{R}^{P\times Y\times d}$ has $P \times Y$ entries, in which d represents the dimension of the features.

Update. To obtain diverse memories without introducing additional training parameters, we propose to utilize different samples from the current training batch B to update the memories with the same initialization. Specifically, when the parameters of CLIP's visual encoder are updated, the p-th proxy of label y_i is also updated by:

$$M_{y_i}^x \leftarrow \mu \cdot M_{y_i}^x + (1-\mu)\cdot b_{y_i}^*. \tag{3}$$

where μ is the momentum factor. b^* represents the samples selected from B to update the memories. As shown in Fig. 2 (a), for the p-th proxy of label y_i, b^* can be obtained in the following ways:

$$b_{y_i}^{hard} \leftarrow \arg\min_b b \cdot M_{y_i}^p, b \in B_{y_i}. \tag{4}$$

$$b_{y_i}^{mean} \leftarrow \frac{1}{K}\sum_{b\in B_{y_i}} b. \tag{5}$$

$$b_{y_i}^{rand} \leftarrow b, b \in B_{y_i}. \tag{6}$$

where B_{y_i} is the sample feature set of label y_i in current mini-batch and K represents the corresponding number. $b_{y_i}^{hard}$ is the feature of the hard sample that has the lowest similarity with $M_{y_i}^p$. $b_{y_i}^{mean}$ represents the average feature

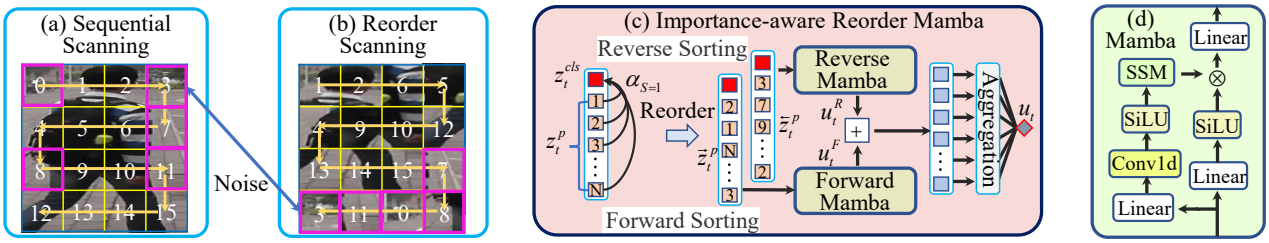

Figure 3: Illustration of our Importance-aware Reorder Mamba (IRM).

of the set. $b_{y_i}^{rand}$ means randomly selecting a sample feature from B_{y_i}. Thanks to the aforementioned selection strategies, we can obtain heterogeneous memories.

Collaboration. We argue that collaborative learning among multiple proxies with the same identity is more effective for optimizing the network. As shown in Fig. 2 (a), we further propose a Multi-Memory Collaboration Loss (MMCL) to achieve the cooperation of multiple memories,

$$L_{MMCL}^{(y_i)} = -\frac{1}{P}\sum_{p=1}^{P} log\frac{\exp(< b_{y_i} \cdot M_{y_i}^{p} >)}{\sum_{j=1}^{N_p}\exp(< b_{y_i} \cdot M_{j}^{p} >)}, \quad (7)$$

where $M_{y_i}^p$ represents the p-th proxy of identity y_i. N_p represents the number of persons in the current mini-batch. By using P different proxies to complementarily represent the same person, we can reduce intra-class variances, leading to more robust features. Similarly, we can construct the corresponding MMC for image-based person ReID.

Multi-Temporal Mamba

Exploiting spatiotemporal information is important for video-based person ReID. In contrast to previous methods that employ attention mechanisms with quadratic computational complexity, we propose a Multi-Temporal Mamba (MTM) that captures multi-scale spatiotemporal information with linear complexity. As shown in Fig. 2 (b), the proposed MTM is composed of Importance-aware Reorder Mamba (IRM) and Multi-scale Information Fusion (MIF).

Importance-aware Reorder Mamba. As shown in Fig. 2, we take a video sequence $\{I_t\}_{t=1}^{T}$ randomly sampled from the training set as an example. After the CLIP visual encoder $f_\theta(\cdot)$, we can get the representation of each frame Z_t, which consists of class token z_t^{cls} and patch tokens $z_t^p = [z_t^1; z_t^2; \cdots, z_t^N]$. After obtaining the frame-level features, we slice the sequence with different step sizes $\mathcal{S} = [1, 4, \ldots, T]$ and then use the proposed IRM to mine spatiotemporal information of different temporal scales.

As shown in Fig. 3 (a), when scanned sequentially, the alternating appearance of noisy background tokens interferes with the local modeling of semantic tokens. Intuitively, reordering tokens based on semantic similarity enhances Mamba's local modeling ability. As shown in Fig. 3 (b), taking $\mathcal{S} = 1$ as an example, we first propose an importance-aware token reorder strategy for each frame. Specifically, we calculate the cosine similarity α between the class token z_t^{cls} and patch tokens z_t^p as the importance score,

$$\alpha_{\mathcal{S}=1} = \frac{z_t^{cls} \cdot z_t^p}{||z_t^{cls}|| \cdot ||z_t^p||} \quad (8)$$

where $\alpha \in \mathbb{R}^N$ lies between 0 and 1.

As shown in Fig. 3 (c), we then reorder the patch tokens z_t^p bidirectionally, yielding a forward sorting result $\overrightarrow{z}_t^p$ and a reverse sorting result $\overleftarrow{z}_t^p$. Next, we concatenate the class token z_t^{cls} with the reordered patch tokens $\overrightarrow{z}_t^p$ and input them into Mamba for spatiotemporal information mining, which can be expressed as:

$$u_t^F = Mamba([z_t^{cls}, \overrightarrow{z}_t^p]), \quad (9)$$

where $[\cdot, \cdot]$ denotes the concatenation operation. Fig. 3 (d) shows the structure of Mamba (Gu and Dao 2023), consisting of linear layer, convolution, State Space Model (SSM) and SiLU. After modeling the long sequences with distinct orderings, we can obtain two discriminative features u_t^F and u_t^R, and aggregate them to obtain $u_t \in \mathbb{R}^{1\times D}$. The aggregation operation is devised as an element-wise addition of the two features for simplify. Finally, we can obtain the output $O_{\mathcal{S}=1} = \{u_t\}_{t=1}^{T}$ for the input sequence. It is worth noting that when $\mathcal{S} = 1$, no temporal information is involved. Therefore, the proposed IRM block can be directly used for image-based person ReID without any modification.

Multi-scale Information Fusion. For video-based person ReID, it is critical to explore temporal information across frames. Therefore, it is necessary to increase the slice stride. Specifically, before calculating the similarity α, we average the selected t' frames' class tokens as the fragment level class token $z_t'^{cls}$, and then calculate the similarity with the remaining $t' * N$ patch tokens to obtain $\alpha_{\mathcal{S}=t'}$. Then, we bidirectionally sort the patch tokens and concatenate them with the fragment-level class token as the input of IRM. Finally, we obtain the output $O_{\mathcal{S}=t'} = \{u_{t'}\}_{t'=1}^{T//t'}$. As shown in Fig. 2 (b), we aggregate the outputs of all branches to get the final output O, which is defined as:

$$O = Linear(O_{\mathcal{S}=1} + O_{\mathcal{S}=4} + \cdots + O_{\mathcal{S}=T}). \quad (10)$$

Here, we simply use a linear layer to implement the aggregation operation. Surpassing previous methods, our MTM is capable of capturing multi-scale spatiotemporal information, resulting in more robust sequence features.

Training and Inference

During training, we employ three different losses: the Multi Memory Collaboration Loss (MMCL) L_{MMCL}, the triplet

Methods	Source	MARS		LS-VID		iLIDS-VID	
		mAP	Rank-1	mAP	Rank-1	Rank-1	Rank-5
STMP (Liu et al. 2019)	AAAI19	72.7	84.4	39.1	56.8	84.3	96.8
M3D (Li, Zhang, and Huang 2019)	AAAI19	74.1	84.4	40.1	57.7	74.0	94.3
GLTR (Li et al. 2019)	ICCV19	78.5	87.0	44.3	63.1	86.0	98.0
TCLNet (Hou et al. 2020)	ECCV20	85.1	89.8	70.3	81.5	86.6	-
MGH (Yan et al. 2020)	CVPR20	85.8	90.0	61.8	79.6	85.6	97.1
GRL (Liu et al. 2021b)	CVPR21	84.8	91.0	-	-	90.4	98.3
BiCnet-TKS (Hou et al. 2021)	CVPR21	86.0	90.2	75.1	84.6	-	-
CTL (Liu et al. 2021a)	CVPR21	86.7	91.4	-	-	89.7	97.0
STMN (Eom et al. 2021)	ICCV21	84.5	90.5	69.2	82.1	-	-
PSTA (Wang et al. 2021)	ICCV21	85.8	91.5	-	-	91.5	98.1
DIL (He et al. 2021b)	ICCV21	87.0	90.8	-	-	92.0	98.0
STT (Zhang et al. 2021b)	Arxiv21	86.3	88.7	78.0	87.5	87.5	95.0
CAVIT (Wu et al. 2022)	ECCV22	87.2	90.8	79.2	89.2	93.3	98.0
SINet (Bai et al. 2022)	CVPR22	86.2	91.0	79.6	87.4	92.5	-
MFA (Gu et al. 2022)	TIP22	85.0	90.4	78.9	88.2	93.3	98.7
DCCT (Liu et al. 2023)	TNNLS23	87.5	92.3	-	-	91.7	98.6
TMT (Liu et al. 2024a)	TITS24	85.8	91.2	-	-	91.3	98.6
TCVIT (Wu et al. 2024)	AAAI24	87.6	91.7	83.1	90.1	94.3	<u>99.3</u>
TF-CLIP (Yu et al. 2024)	AAAI24	<u>89.4</u>	<u>93.0</u>	<u>83.8</u>	<u>90.4</u>	<u>94.5</u>	99.1
CLIMB-ReID		**89.7**	**93.3**	**85.0**	**91.3**	**96.7**	**99.9**

Table 1: Comparison with state-of-the-art methods on MARS, LS-VID and iLIDS-VID. The **bold** and <u>underline</u> denote the best and second results.

Methods	Market1501		MSMT17	
	mAP	Rank-1	mAP	Rank-1
ABD-Net (Chen et al. 2019)	88.3	95.6	60.8	82.3
SAN (Jin et al. 2020)	88.0	<u>96.1</u>	55.7	79.2
CDNet (Li, Wu, and Zheng 2021)	86.0	95.1	54.7	78.9
DRL-Net (Jia et al. 2022)	86.9	94.7	55.3	78.4
AAformer (Zhu et al. 2024b)	87.7	95.4	63.2	83.6
TransReID (He et al. 2021a)	89.5	95.2	69.4	86.2
DCAL (Zhu et al. 2022)	87.5	94.7	64.0	83.1
CLIP-ReID (Li, Sun, and Li 2023)	90.4	95.5	73.2	88.0
PCL (Li and Gong 2023)	<u>91.4</u>	95.9	<u>76.1</u>	<u>89.8</u>
TF-CLIP (Yu et al. 2024)	90.4	95.7	73.9	88.5
CLIMB-ReID	**92.6**	**96.8**	**77.8**	**90.5**

Table 2: Comparison with state-of-the-arts on Market1501 and MSMT17.

loss L_{tri} (Hermans, Beyer, and Leibe 2017) and the label smooth cross-entropy loss L_{ce}. Finally, the overall loss L_{total} is defined as:

$$L_{total} = L_{MMCL} + L_{tri} + L_{ce}. \tag{11}$$

During inference, the original sequence-level feature and the multi-scale temporal feature are concatenated to obtain the final representation for video-based person ReID. While for image-based person ReID, the original frame-level feature and the $\mathcal{S}$=1 IRM feature are concatenated for ranking.

Experiments

Datasets and Evaluation Protocols

We evaluate our approach on three video-based person ReID benchmarks, including MARS (Zheng et al. 2016), LS-VID (Li et al. 2019) and iLIDS-VID (Wang et al. 2014). The proposed method is also validated on two image-based person ReID datasets, *i.e.*, Market1501 (Zheng et al. 2015) and MSMT17 (Wei et al. 2018). More details of these datasets can be found in the **Supplementary**. Following common practices, the Cumulative Matching Characteristic (CMC)@$K(K = 1, 5)$ and mean Average Precision (mAP) are adopted to measure the performance.

Experiment Settings

We use the ViT-B/16 from CLIP (Radford et al. 2021) as the feature encoder, which contains 12 Transformer layers with the hidden size of 768. For MMC, we use the "Mean" and "Hard" selection strategies to generate $P = 2$ memories. We set μ to 0.2. During training, we adopt random flipping, random cropping and random erasing (Zhong et al. 2020) for data augmentation. Each frame is resized to 256×128. We train the framework for 60 epochs in total. For video-based person ReID, the mini-batch size is 128, consisting of 4 identities, 4 tracklets for each identity and 8 frames from each tracklet. We utilize the Adam optimizer with the learning rate of 5×10^{-6}. We warm up the model with 10 epochs, linearly increasing the learning rate from 5×10^{-7} to 5×10^{-6}. Afterwards, the learning rate is reduced by a factor of 0.1 at the 30th and 50th epochs. The slice stride in MTM is set to be $\mathcal{S} = [1, 4, 8]$. For image-based person ReID, the mini-batch size is 128, consisting of 16 identities and 8 images for each identity. We utilize the SGD optimizer with the learning rate of 3.5×10^{-4} and the weight decay of 5×10^{-4}. Similarly, we also warm up the model with 10 epochs, linearly increasing the learning rate from 3.5×10^{-5} to 3.5×10^{-4}. The cosine distance is employed as the distance metric for ranking.

Model	Components			LS-VID		MSMT17	
	MMC	IRM	MTM	mAP	Rank-1	mAP	Rank-1
1	×	×	×	80.3	87.7	73.9	88.5
2	✓	×	×	82.0	90.1	76.7	90.1
3	✓	✓	×	83.2	90.7	77.8	90.5
4	✓	×	✓	84.4	91.0	-	-
5	✓	✓	✓	85.0	91.3	-	-

Table 3: Comparison of different components on LS-VID and MSMT17.

Comparison with State-of-the-arts

Results for Video-based Person ReID. Our results for the video-based benchmarks are presented in Tab. 1. It is observed that our method outperforms other state-of-the-arts on MARS, LS-VID and iLIDS-VID. Compared with the attention-based ReID methods (Wu et al. 2022; Zhang et al. 2021b), our method significantly outperforms them. For example, our method achieves 93.3% for Rank-1 and 89.7% for mAP on MARS, which surpasses PSTA (Wang et al. 2021) by 1.8% for Rank-1 and 3.9% for mAP. As a representative method, TMT (Liu et al. 2024a) employs a trigeminal feature extractor to integrate multi-view spatiotemporal cues. Instead, we design a MTM to capture spatiotemporal cues. Our method obtains a higher Rank-1 than TMT on iLIDS-VID, which validates the effectiveness of our method. It is worth noting that TF-CLIP (Yu et al. 2024) also explores transferring CLIP's knowledge to video-based person ReID. Different from TF-CLIP, we propose a parameter-free and prompt-free MMC. As a result, our method achieves 91.3% Rank-1 accuracy on LS-VID, which surpasses TF-CLIP by 1.1%. These comparisons fully demonstrate the effectiveness of our proposed method.

Results for Image-based Person ReID. We compare our method with other state-of-the-art methods on Market1501 and MSMT17. The results are shown in Tab. 2. On the Market1501 and MSMT17 datasets, our method achieves the best results of 96.8%/92.6% and 90.5%/77.8% in Rank-1/mAP accuracy, respectively. Specifically, compared with TransReID (He et al. 2021a), our method brings 8.4% mAP and 4.3% Rank-1 accuracy gains on MSMT17, respectively. It is worth noting that CLIP-ReID (Li, Sun, and Li 2023), TF-CLIP (Yu et al. 2024) and PCL (Li and Gong 2023), also explore transferring CLIP knowledge to image-based person ReID. Unlike them, we introduce a MMC that is both parameter-free and prompt-free. As a result, our method achieves 92.6% Rank-1 accuracy on Market1501, which surpasses PCL by 1.2%. These experimental results validate the superiority of our method.

Ablation Study

We conduct ablation experiments on LS-VID and MSMT17 datasets to assess the impact of different components. The compared results are shown in Tab. 3. $Model1$ means that CLIP-Memory proposed in TF-CLIP is applied as the baseline. $Model3$ means the slice stride is 1. $Model4$ means the slice stride is larger than 1. More ablation studies can be found in the **Supplementary**.

Update Strategy	LS-VID		MARS	
	mAP	Rank-1	mAP	Rank-1
Mean	79.2	88.2	88.4	92.0
Rand	78.7	87.8	88.3	92.3
Hard	79.8	88.8	88.5	91.9
Mean+Rand	81.4	88.8	88.3	**92.4**
Mean+Hard	**82.0**	**90.1**	**88.6**	**92.4**
Rand+Hard	80.6	88.9	88.1	92.1
Mean+Rand+Hard	80.4	88.8	88.2	91.9

Table 4: Comparison with different update strategies on LS-VID and MARS. The best results are highlighted in **bold**.

Scanning Way	LS-VID		MSMT17	
	mAP	Rank-1	mAP	Rank-1
Forward	82.5	90.1	77.1	89.9
Bidirectional	82.8	90.3	77.3	90.1
Four directions	82.7	90.2	77.2	90.1
Reorder(Ours)	83.2	90.7	77.8	90.5

Table 5: Comparison with different scanning ways in IRM on LS-VID and MSMT17.

Effectiveness of MMC. As demonstrated in Tab. 3, incorporating MMC into $Model2$ results in an improvement of 0.5% in mAP and 0.4% in Rank-1 on LS-VID compared to $Model1$. Additionally, a notable improvement is observed on MSMT17. It is evident that our proposed MMC effectively enhances performance. A plausible explanation for this improvement is that the diverse memory collection facilitates the learning of a more discriminative representation.

Different Memory Update Strategies. As shown in Tab. 4, we further verify the impact of different memory update strategies on LS-VID and MARS. Overall, the performance of combining multiple update strategies tends to surpass that of using a single update strategy. The highest performance is achieved when the "Mean+Hard" combination is utilized. Therefore, we configure MMC with the "Mean+Hard" combination in our method.

Effectiveness of IRM. As shown in Tab.3, the proposed IRM leads to a substantial improvement in performance. Compared with $Model2$, $Model3$ using IRM brings 1.1% mAP and 0.4% Rank-1 gains on MSMT17, respectively. Compared with $Model4$, $Model5$ also brings 0.3% Rank-1 gains on LS-VID. We believe that IRM enhances the semantic information within each temporal scale and produces more robust features, thereby further boosting performance.

The Scanning Way of IRM. In Mamba-based approaches, the scanning way is essential for information aggregation. Thus, we further verify the impact of different scanning ways on LS-VID and MSMT17, including "Forward", "Bidirectional", "Four directions" and "Reorder". As shown in Tab.5, scanning methods influence the performance of IRM. The proposed "Reorder" way achieves 77.8% mAP on MSMT17, which surpasses "Four directions" by 0.6%. Compared with "Bidirectional", the proposed "Reorder" also brings 0.4% Rank-1 gains on LS-VID.

Mamba Layer	MARS		Market1501	
	mAP	Rank-1	mAP	Rank-1
1	89.0	92.3	92.6	96.8
2	89.0	92.5	92.5	96.9
3	88.9	92.4	92.6	96.6
4	88.8	92.2	92.4	96.5

Table 6: Comparison with different layers in IRM on MARS and Market1501.

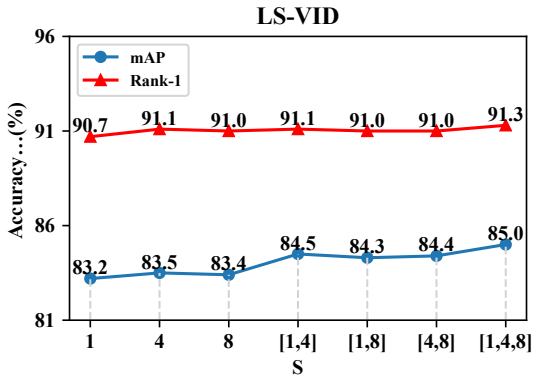

Figure 4: Illustration of the impact of $\mathcal{S}$ on LS-VID.

These results reflect the effectiveness of our method.

Effectiveness of MTM. As shown in Tab.3, the proposed MTM delivers a significant performance boost. Compared with $Model2$, $Model4$ with MTM delivers a 2.4% gain in mAP and a 0.9% improvement in Rank-1 on LS-VID. Moreover, compared with $Model3$, $Model5$ achieves an additional 1.8% gain in mAP and a 0.6% increase in Rank-1 on LS-VID. We believe that MTM can effectively capture multi-scale spatiotemporal information within the sequence, thereby yielding more robust sequence features and contributing to enhanced performance.

The Slice Stride in MTM. The slice stride in MTM dictates the granularity of multi-temporal representations. Therefore, we further examine the effect of varying slice strides on LS-VID. As shown in Fig. 4, slice strides influence the performance of MTM. Specifically, when $\mathcal{S}$=[1,4,8], our method achieves 85.0% mAP on LS-VID, which surpasses single granularity $\mathcal{S}$=8 by 1.6%. The reason is that capturing multi-scale spatiotemporal information is vital for deriving discriminative sequence features. **The Number of Mamba Layers in IRM.** As shown in Tab. 6, we carry out experiments to investigate the effect of the number of Mamba layers in IRM with $Model3$ on MARS and Market1501. It is evident that our method shows robustness to this hyperparameter. To balance the computation and performance, we ultimately select $layer = 1$.

Pure Mamba Structures. As shown in Fig. 5, we explore the performance of some pure Mamba-based methods on MARS, including VMamba (Zhu et al. 2024c) and VideoMamba (Li et al. 2025). From the results in Fig. 5, the performance of these pure Mamba structures is far from ideal. The possible reason is that for person ReID, additional specialized pre-training may be required.

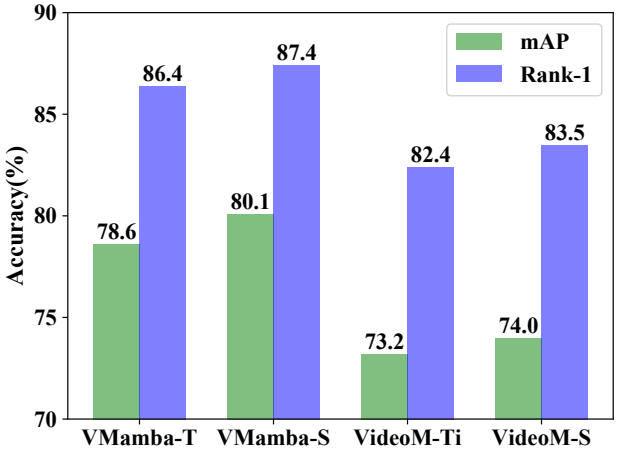

Figure 5: Illustration of the performance of some pure Mamba structures on MARS.

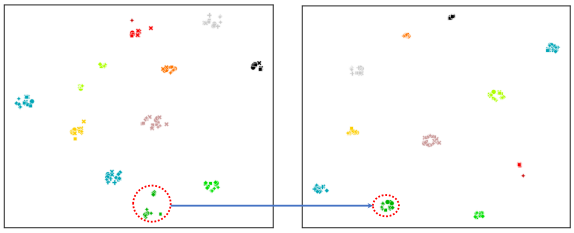

Figure 6: t-SNE visualization of the baseline model and our CLIMB-ReID on the LS-VID test set. Different colored dots represent different identities. Best viewed in color.

Visualization

Some persons from LS-VID are randomly chosen to conduct t-SNE (Van der Maaten and Hinton 2008) visualization experiments. In Fig. 6, we visualize the discriminative feature distributions of different methods. From Fig. 6 (a) to (b), we observe that the features for each identity are more tightly clustered, while the separation between different identities has widened. These visualizations confirm the effectiveness of our method in enhancing feature discrimination.

Conclusion

In this paper, we propose a novel framework named CLIMB-ReID for person ReID. We first propose a Multi-Memory Collaboration (MMC) strategy to efficiently transfer CLIP's knowledge in a parameter-free and prompt-free form. Meanwhile, we further design a Multi-Temporal Mamba (MTM) to capture multi-granular spatiotemporal information in videos. Finally, information from various scales is combined to produce robust sequence features. It is worth noting that the proposed method can be used for both video and image ReID. Extensive experiments show that our proposed method outperforms other state-of-the-art methods on both image-based and video-based ReID benchmarks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No.62101092) and Fundamental Research Funds for the Central Universities (No.DUT23BK050).

References

Bai, S.; Ma, B.; Chang, H.; Huang, R.; and Chen, X. 2022. Salient-to-broad transition for video person re-identification. In *CVPR*, 7339–7348.

Chen, T.; Ding, S.; Xie, J.; Yuan, Y.; Chen, W.; Yang, Y.; Ren, Z.; and Wang, Z. 2019. Abd-net: Attentive but diverse person re-identification. In *ICCV*, 8351–8361.

Chen, Z.; Hu, P.; Zhang, L.; Lu, H.; He, Y.; Wang, S.; Zhang, X.; Hu, M.; and Li, T. 2022. Video object segmentation via structural feature reconfiguration. In *ACCV*, 2729–2745.

Chen, Z.; Zhang, L.; Hu, P.; Lu, H.; and He, Y. 2024. MaskTrack: Auto-labeling and stable tracking for video object segmentation. *TNNLS*, 1: 1–14.

Dai, J.; Zhang, P.; Wang, D.; Lu, H.; and Wang, H. 2019. Video person re-identification by temporal residual learning. *TIP*, 28: 1366–1377.

Diao, H.; Wan, B.; Jia, X.; Zhuge, Y.; Zhang, Y.; Lu, H.; and Chen, L. 2025. Sherl: Synthesizing high accuracy and efficient memory for resource-limited transfer learning. In *ECCV*, 75–95.

Diao, H.; Wan, B.; Zhang, Y.; Jia, X.; Lu, H.; and Chen, L. 2024. Unipt: Universal parallel tuning for transfer learning with efficient parameter and memory. In *CVPR*, 28729–28740.

Diao, H.; Zhang, Y.; Ma, L.; and Lu, H. 2021. Similarity reasoning and filtration for image-text matching. In *AAAI*, 1218–1226.

Eom, C.; Lee, G.; Lee, J.; and Ham, B. 2021. Video-based person re-identification with spatial and temporal memory networks. In *ICCV*, 12036–12045.

Gao, S.; Yu, C.; Zhang, P.; and Lu, H. 2023. Ped-Mix: Mix pedestrians for occluded person re-identification. In *PRCV*, 265–277.

Gao, S.; Yu, C.; Zhang, P.; and Lu, H. 2024. Part representation learning with teacher-student decoder for occluded person re-identification. In *ICASSP*, 2660–2664.

Gong, Y.; Zhong, Z.; Qu, Y.; Luo, Z.; Ji, R.; and Jiang, M. 2024. Cross-modality perturbation synergy attack for person re-identification. *arXiv preprint arXiv:2401.10090*.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, X.; Chang, H.; Ma, B.; and Shan, S. 2022. Motion feature aggregation for video-based person re-identification. *TIP*, 31: 3908–3919.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021a. Transreid: Transformer-based object re-identification. In *ICCV*, 15013–15022.

He, T.; Jin, X.; Shen, X.; Huang, J.; Chen, Z.; and Hua, X.-S. 2021b. Dense interaction learning for video-based person re-identification. In *ICCV*, 1490–1501.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*.

Hou, R.; Chang, H.; Ma, B.; Huang, R.; and Shan, S. 2021. BiCnet-TKS: Learning efficient spatial-temporal representation for video person re-identification. In *CVPR*, 2014–2023.

Hou, R.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2020. Temporal complementary learning for video person re-identification. In *ECCV*, 388–405.

Jia, M.; Cheng, X.; Lu, S.; and Zhang, J. 2022. Learning disentangled representation implicitly via transformer for occluded person re-identification. *TMM*, 25: 1294–1305.

Jin, X.; Lan, C.; Zeng, W.; Wei, G.; and Chen, Z. 2020. Semantics-aligned representation learning for person re-identification. In *AAAI*, 11173–11180.

Li, H.; Wu, G.; and Zheng, W. 2021. Combined depth space based architecture search for person re-identification. In *CVPR*, 6729–6738.

Li, J.; and Gong, X. 2023. Prototypical contrastive learning-based CLIP fine-tuning for object re-identification. *arXiv preprint arXiv:2310.17218*.

Li, J.; Wang, J.; Tian, Q.; Gao, W.; and Zhang, S. 2019. Global-local temporal representations for video person re-identification. In *ICCV*, 3958–3967.

Li, J.; Zhang, S.; and Huang, T. 2019. Multi-scale 3d convolution network for video based person re-identification. In *AAAI*, 8618–8625.

Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2025. Videomamba: State space model for efficient video understanding. In *ECCV*, 237–255.

Li, S.; Leng, J.; Li, G.; Gan, J.; Gao, X.; et al. 2024. CLIP-driven cloth-agnostic feature learning for cloth-changing person re-identification. *arXiv preprint arXiv:2406.09198*.

Li, S.; Sun, L.; and Li, Q. 2023. CLIP-ReID: Exploiting vision-language model for image re-identification without concrete text labels. In *AAAI*, 1405–1413.

Liu, J.; Zha, Z.-J.; Wu, W.; Zheng, K.; and Sun, Q. 2021a. Spatial-temporal correlation and topology learning for person re-identification in videos. In *CVPR*, 4370–4379.

Liu, X.; Yu, C.; Zhang, P.; and Lu, H. 2023. Deeply coupled convolution-transformer with spatial–temporal complementary learning for video-based person re-identification. *TNNLS*, 35: 13753–13763.

Liu, X.; Zhang, P.; Yu, C.; Lu, H.; and Yang, X. 2021b. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *CVPR*, 13334–13343.

Liu, X.; Zhang, P.; Yu, C.; Qian, X.; Yang, X.; and Lu, H. 2024a. A video is worth three views: Trigeminal transformers for video-based person re-identification. *TITS*, 25: 12818–12828.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024b. VMamba: Visual state space model. *arXiv preprint arXiv:2401.10166*.

Liu, Y.; Yuan, Z.; Zhou, W.; and Li, H. 2019. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, 8786–8793.

Ma, J.; Li, F.; and Wang, B. 2024. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Shi, J.; Yin, X.; Chen, Y.; Zhang, Y.; Zhang, Z.; Xie, Y.; and Qu, Y. 2025. Multi-memory matching for unsupervised visible-infrared person re-identification. In *ECCV*, 456–474.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *JMLR*, 9: 2579–2605.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*, 5998–6008.

Wang, B.; Liang, Y.; Cai, L.; Huang, H.; and Zeng, H. 2024a. Image re-identification: Where self-supervision meets vision-language learning. *arXiv preprint arXiv:2407.20647*.

Wang, M.; Li, J.; Lai, B.; Gong, X.; and Hua, X.-S. 2022. Offline-online associated camera-aware proxies for unsupervised person re-identification. *TIP*, 31: 6548–6561.

Wang, Q.; Wang, C.; Lai, Z.; and Zhou, Y. 2024b. Insect-mamba: Insect pest classification with state space model. *arXiv preprint arXiv:2404.03611*.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2014. Person re-identification by video ranking. In *ECCV*, 688–703.

Wang, Y.; Liu, X.; Zhang, P.; Lu, H.; Tu, Z.; and Lu, H. 2024c. TOP-ReID: Multi-spectral object re-identification with token permutation. In *AAAI*, 5758–5766.

Wang, Y.; Zhang, P.; Gao, S.; Geng, X.; Lu, H.; and Wang, D. 2021. Pyramid spatial-temporal aggregation for video-based person re-identification. In *CVPR*, 12026–12035.

Wang, Y.; Zhang, P.; Wang, D.; and Lu, H. 2024d. Other tokens matter: Exploring global and local features of vision transformers for object re-identification. *CVIU*, 244: 104–112.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 79–88.

Wu, J.; He, L.; Liu, W.; Yang, Y.; Lei, Z.; Mei, T.; and Li, S. Z. 2022. CAViT: Contextual alignment vision transformer for video object re-identification. In *ECCV*, 549–566.

Wu, P.; Wang, L.; Zhou, S.; Hua, G.; and Sun, C. 2024. Temporal correlation vision transformer for video person re-identification. In *AAAI*, 6083–6091.

Yan, P.; Liu, X.; Zhang, P.; and Lu, H. 2023. Learning convolutional multi-level transformers for image-based person re-identification. *VI*, 1: 1–12.

Yan, Y.; Qin, J.; Chen, J.; Liu, L.; Zhu, F.; Tai, Y.; and Shao, L. 2020. Learning multi-granular hypergraphs for video-based person re-identification. In *CVPR*, 2899–2908.

Yang, Y.; Xing, Z.; and Zhu, L. 2024. Vivim: a video vision mamba for medical video object segmentation. *arXiv preprint arXiv:2401.14168*.

Yu, C.; Liu, X.; Wang, Y.; Zhang, P.; and Lu, H. 2024. TF-CLIP: Learning text-free CLIP for video-based person re-identification. In *AAAI*, 6764–6772.

Zhang, G.; Zhang, P.; Qi, J.; and Lu, H. 2021a. Hat: Hierarchical aggregation transformers for person re-identification. In *ACM MM*, 516–525.

Zhang, P.; Wang, Y.; Liu, Y.; Tu, Z.; and Lu, H. 2024. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *CVPR*, 17117–17126.

Zhang, T.; Wei, L.; Xie, L.; Zhuang, Z.; Zhang, Y.; Li, B.; and Tian, Q. 2021b. Spatiotemporal transformer for video-based person re-identification. *arXiv:2103.16469*.

Zhang, Z.; Lan, C.; Zeng, W.; and Chen, Z. 2020. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *CVPR*, 10407–10416.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 868–884.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *ICCV*, 1116–1124.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *AAAI*, 13001–13008.

Zhu, H.; Ke, W.; Li, D.; Liu, J.; Tian, L.; and Shan, Y. 2022. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *CVPR*, 4692–4702.

Zhu, J.; Chen, X.; Zhang, P.; Wang, X.; Wang, D.; Zhao, W.; and Lu, H. 2024a. SRRT: Exploring Search Region Regulation for Visual Object Tracking. *TCSVT*, 34: 10551–10563.

Zhu, J.; Lai, S.; Chen, X.; Wang, D.; and Lu, H. 2023. Visual prompt multi-modal tracking. In *CVPR*, 9516–9526.

Zhu, K.; Guo, H.; Zhang, S.; Wang, Y.; Huang, G.; Qiao, H.; Liu, J.; Wang, J.; and Tang, M. 2024b. Aaformer: Auto-aligned transformer for person re-identification. *TNNLS*, 35: 17307–17317.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024c. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.

Zou, C.; Chen, Z.; Cui, Z.; Liu, Y.; and Zhang, C. 2023. Discrepant and multi-instance proxies for unsupervised person re-identification. In *ICCV*, 11058–11068.