Supplementary Material
TF-CLIP: Learning Text-free CLIP for Video-based Person Re-Identification

Anonymous submission

More Experimental Details

Details of the Datasets

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). MARS is one of the largest public datasets for video-based person ReID, which consists of 1,261 pedestrians and 20,715 tracklets captured by 6 cameras, and each individual appears in at least two cameras. Meanwhile, each identity has 13.2 tracklets on average. Whereas many sequences may have a poor quality since the bounding boxes are generated by the DPM detector (Felzenszwalb et al. 2010) and the GMMCP tracker (Dehghan, Modiri Assari, and Shah 2015), the failures of tracking and detections will affect the ReID accuracy. LS-VID is one of the largest video ReID benchmarks which collects 3,772 identities existing in 14,943 video tracklets captured by 15 cameras. And the average sequence length is 200 frames. Faster RCNN (Ren et al. 2015) is utilized for pedestrian detection. iLIDS-VID contains 600 image sequences of 300 peoples from two non overlapping camera views in an airport arrival hall. The frame lengths in each sequence vary from 23 to 192, with an average length of 73.

Besides, we follow common practices and adopt the Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) to measure the performance. For iLIDS-VID, we randomly split the probe/gallery identities following the protocol from (Wang et al. 2014). Persons are randomly split into two subsets with equal size as train and test sets, and the performance is reported as the average results of ten trials. For MARS, we use the original splits provided by (Zheng et al. 2016) which use the predefined 8,298 sequences of 625 peoples for training and the remaining 11,310 sequences of 636 peoples for testing. For LS-VID, we use the original splits provided by (Li et al. 2019) which use the predefined 2,831 sequences of 842 peoples for training and the remaining 11,333 sequences of 2,730 peoples for testing.

Why not evaluate our approach on DukeMTMC-VideoReID. DukeMTMC-VideoReID (Zheng, Zheng, and Yang 2017) is one of the standard benchmarks for video-based person re-identification that was once widely used. However, this dataset collection involved non-consensual video surveillance of students on Duke University campus. It is unlikely that all students even knew they were being recorded, and their relative lack of power with respect to the institution surveilling them also raises concerns about the ability to meaningfully object to the surveillance. Therefore, this dataset has been retracted by the authors, emphasizing that the use of this dataset is no longer authorized. Based on the above facts, we do not evaluate our approach on DukeMTMC-VideoReID.

Methods	Source	Backbone	MARS	
			mAP	Rank-1
STMP (Liu et al. 2019)	AAAI19	C	72.7	84.4
M3D (Li, Zhang, and Huang 2019)	AAAI19	C	74.1	84.4
GLTR (Li et al. 2019)	ICCV19	C	78.5	87.0
TCLNet (Hou et al. 2020)	ECCV20	C	85.1	89.8
MGH (Yan et al. 2020)	CVPR20	C	85.8	90.0
GRL (Liu et al. 2021c)	CVPR21	C	84.8	91.0
BiCnet-TKS (Hou et al. 2021)	CVPR21	C	86.0	90.2
CTL (Liu et al. 2021a)	CVPR21	C	86.7	91.4
STMN (Eom et al. 2021)	ICCV21	C	84.5	90.5
PSTA (Wang et al. 2021)	ICCV21	C	85.8	91.5
SINet (Bai et al. 2022)	CVPR22	C	86.2	91.0
MFA (Gu et al. 2022)	TIP22	C	85.0	90.4
LSTRL (Liu, Zhang, and Lu 2023)	ICIG23	C	86.8	91.6
Baseline		C	83.9	89.0
TF-CLIP(Ours)		C	86.3	90.9
DIL (He et al. 2021b)	ICCV21	T	87.0	90.8
STT (Zhang et al. 2021)	Arxiv21	T	86.3	88.7
TMT (Liu et al. 2021b)	Arxiv21	T	85.8	91.2
CAVIT (Wu et al. 2022)	ECCV22	T	87.2	90.8
DCCT (Liu et al. 2023)	TNNLS23	T	87.5	<u>92.3</u>
Baseline		T	<u>88.1</u>	91.7
TF-CLIP(Ours)		T	**89.4**	**93.0**

Table 1: Comparison with state-of-the-art CNN- (C) and Transformer-based (T) methods on MARS.

SIE and OLP. Following (Li, Sun, and Li 2022; He et al. 2021a), we also use Side Information Embedding (SIE) to make the model aware of the camera information. Overlapping Patches (OLP) usually can further enhance the model by changing the stride in the token embedding. However, this operation will significantly increase the consumption of computing resources, and will cause GPU out of memory when applying OLP to video-based person ReID. Therefore, in this paper we set the stride of patch to 16 instead of 12.

A more detailed schematic of the CLIP-Memory Module. A more detailed schematic of the CLIP memory module is shown in Fig. 1.

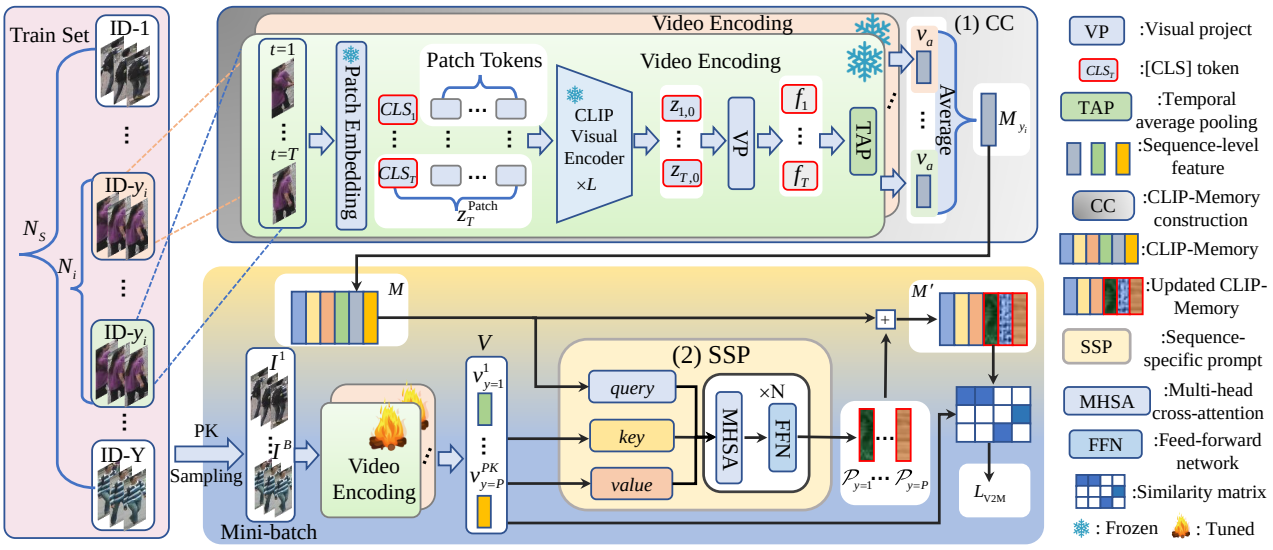

Figure 1: Illustration of the proposed CLIP-Memory.

Methods	Source	Market1501		MSMT17	
		mAP	Rank-1	mAP	Rank-1
ABD-Net (Chen et al. 2019)	ICCV19	88.3	95.6	60.8	82.3
SAN (Jin et al. 2020)	AAAI20	88.0	**96.1**	55.7	79.2
CDNet (Li, Wu, and Zheng 2021)	CVPR21	86.0	95.1	54.7	78.9
DRL-Net (Jia et al. 2022)	TMM22	86.9	94.7	55.3	78.4
AAformer (Zhu et al. 2021)	Arxiv21	87.7	95.4	63.2	83.6
TransReID (He et al. 2021a)	ICCV21	<u>89.5</u>	95.2	69.4	86.2
DCAL (Zhu et al. 2022)	CVPR22	87.5	94.7	64.0	83.1
CLIP-ReID* (Li, Sun, and Li 2022)	AAAI23	**90.4**	95.5	<u>73.2</u>	<u>88.0</u>
+CLIP-M		89.6	95.0	72.0	87.4
+SSP		**90.4**	<u>95.7</u>	**73.9**	**88.5**

Table 2: Comparison with state-of-the-arts on Market1501 and MSMT17. The star $*$ in the superscript indicates the results obtained by re-running on our machine according to the official open source code. The **bold** and <u>underline</u> denote the best two results.

CNN-based TF-CLIP

Implementation Details

CLIP provides two alternatives (Li, Sun, and Li 2022), namely a transformer and a CNN with a global attention pooling layer. For the CNN, we choose ResNet-50 (He et al. 2016), where the last stride changes from 2 to 1, resulting in a larger feature map to preserve spatial information. The global attention pooling layer after ResNet-50 reduces the dimension of the embedding vectors from 2048 to 1024, matching the dimensions of the text features. And the number of transformer layers in SSP is set to 2. During training, we adopt the Restricted Random Sampling (RRS) (Li et al. 2018) to generate sequential frames. We sample 8 frames from each video sequence and each frame is resized to 256×128. In each mini-batch, we sample 4 identities, each with 4 tracklets. Thus, the number of images in a batch is $4 \times 4 \times 8 = 128$. We also adopt random flipping and random erasing (Zhong et al. 2020) for data augmentation. We train our framework for 120 epochs in total by the Adam optimizer (Kingma and Ba 2014). Following CLIP-ReID (Li, Sun, and Li 2022), we first warm up the model for 10 epochs with a linearly growing learning rate from 3.5 $\times 10^{-6}$ to 3.5 $\times 10^{-4}$. Then, the learning rate is divided by 10 at the 40th and 70th epochs. Our model is implemented on the Pytorch platform and trained with one NVIDIA Tesla A30 GPU (24G memory). The original sequence-level feature $v \in \mathbb{R}^{1\times1024}$ and the aggregated feature $\hat{v} \in \mathbb{R}^{1\times1024}$ are concatenated to obtain the final video representation during testing. And the Euclidean distance is employed as the distance metric for ranking.

Comparison with State-of-the-arts

In this section, we further compare our CNN-based TF-CLIP with other state-of-the-art methods on MARS. The results

Model	Components			MARS			
	TMC	MD	SSP	mAP	Rank-1	Params(M)	FLOPs(G)
A1	×	×	×	88.4	91.6	86.94	11.26
A2	✓	×	×	88.8	91.9	89.93	11.26
A3	✓	✓	×	89.2	92.3	97.02	11.72
A4	×	×	✓	88.8	92.1	94.21	12.11
A5	✓	×	✓	89.0	92.5	97.17	12.11
A6	✓	✓	✓	89.4	93.0	104.26	12.53

Table 3: Comparison of different components in TMD on MARS.

Model	Components			Params(M)	FLOPs(G)	MARS			LS-VID		
	CLIP-M	SSP	TMD			mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
B1	×	×	×	126.78	14.24	88.1	91.7	97.4	80.6	88.8	96.3
B2	✓	×	×	86.94	11.26	88.4	91.6	97.9	80.3	87.7	95.7
B3	✓	✓	×	94.21	12.11	88.8	92.1	97.6	81.3	89.9	96.2
B4	✓	×	✓	97.02	11.72	89.2	92.3	97.7	81.6	88.9	96.4
B5	×	×	✓	136.86	14.70	89.3	92.7	97.8	82.3	90.3	96.7
B6	✓	✓	✓	104.26	12.53	89.4	93.0	97.9	83.8	90.4	97.1

Table 4: Comparison of different components and the computational cost on MARS and LS-VID.

Methods	MARS		Params	FLOPs
	mAP	Rank-1	(M)	(G)
TAP	88.4	91.6	86.94	11.26
Cov1D	88.1	91.8	87.31	11.26
Transf$_{cls}$	87.5	91.5	106.23	11.41
Transf	88.2	91.4	106.22	11.39
TMD	89.2	92.3	97.02	11.72

Table 5: Comparison of different temporal fusion methods.

N	MARS		Params	FLOPs
	mAP	Rank-1	(M)	(G)
1	88.5	91.9	91.06	11.68
2	88.8	92.1	94.21	12.11
3	87.4	92.1	97.36	12.53
4	88.4	92.3	100.51	12.95

Table 6: Effect of different layers in SSP.

Methods	MARS	
	mAP	Rank-1
$ModelB2$	88.4	91.6
-wo CLIP-M	84.9	90.2
$ModelB4$	89.2	92.3
-wo CLIP-M	87.3	91.0

Table 7: Comparison w/wo CLIP-Memory.

are shown in Tab. 1. It is observed that our CNN-based method still outperforms the most state-of-the-art methods on MARS. Specifically, compared with $Baseline$, the proposed CNN-Based TF-CLIP brings 2.4% mAP and 1.9% Rank-1 accuracy gains on MARS, respectively. This experimental results clearly confirm the effectiveness of the proposed method.

It is worth noting that those methods, GRL (Liu et al. 2021c), TCLNet (Hou et al. 2020) and GLTR (Li et al. 2019), also explore the temporal learning for video-based person Re-ID. Compared with these methods, our proposed method achieves better results on MARS. More specifically, compared with GRL (Liu et al. 2021c), our method improves the performances by 1.5% in terms of mAP accuracy on MARS dataset. Moreover, CTL (Liu et al. 2021a), PSTA (Wang et al. 2021) and LSTRL (Liu, Zhang, and Lu 2023) model spatial-temporal information. Different from them, our model only explore the temporal learning and still achieve competitive results. These experimental results validate the superiority of our method.

Expanding CLIP-Memory to Image-based Person Re-Identification

We further expand our proposed CLIP-Memory module to image-based person ReID. We intensively evaluate our proposed method on two public image-based person ReID benchmarks, including Market1501 (Zheng et al. 2015) and MSMT17 (Wei et al. 2018). The results are shown in Tab. 2. As can be seen from Tab. 2, using the proposed CLIP-Memory module to replace the text branch in CLIP-ReID (Li, Sun, and Li 2022) can obtain promising results. This further demonstrates the potential of our proposed method in extending CLIP to downstream tasks without text labels.

More Ablation Studies

To capture temporal information, we further propose a Temporal Memory Diffusion (TMD) module. And TMD is mainly composed of Temporal Memory Construction (TMC) and Memory Diffusion (MD). To verify the impact of each component in TMD, we conduct several experiments on MARS, and show compared results in Tab. 3. $ModelA1$ means that the proposed one-stage text-free baseline which uses the proposed CLIP-M to replace the text branch.

Effectiveness of TMC. As shown in Tab. 3, compared with $ModelA1$, $ModelA2$ employing the proposed TMC to capture temporal information within the sequence brings 0.4% mAP and 0.3% Rank-1 accuracy gains on MARS, respectively. What's more, compared with $ModelA4$, $ModeA5$ also improves the Rank-1 accuracy by 0.4% on MARS. We can see that our proposed TMC can indeed improve the performance of the network. A reasonable explanation for this improvement is that TMC allows the

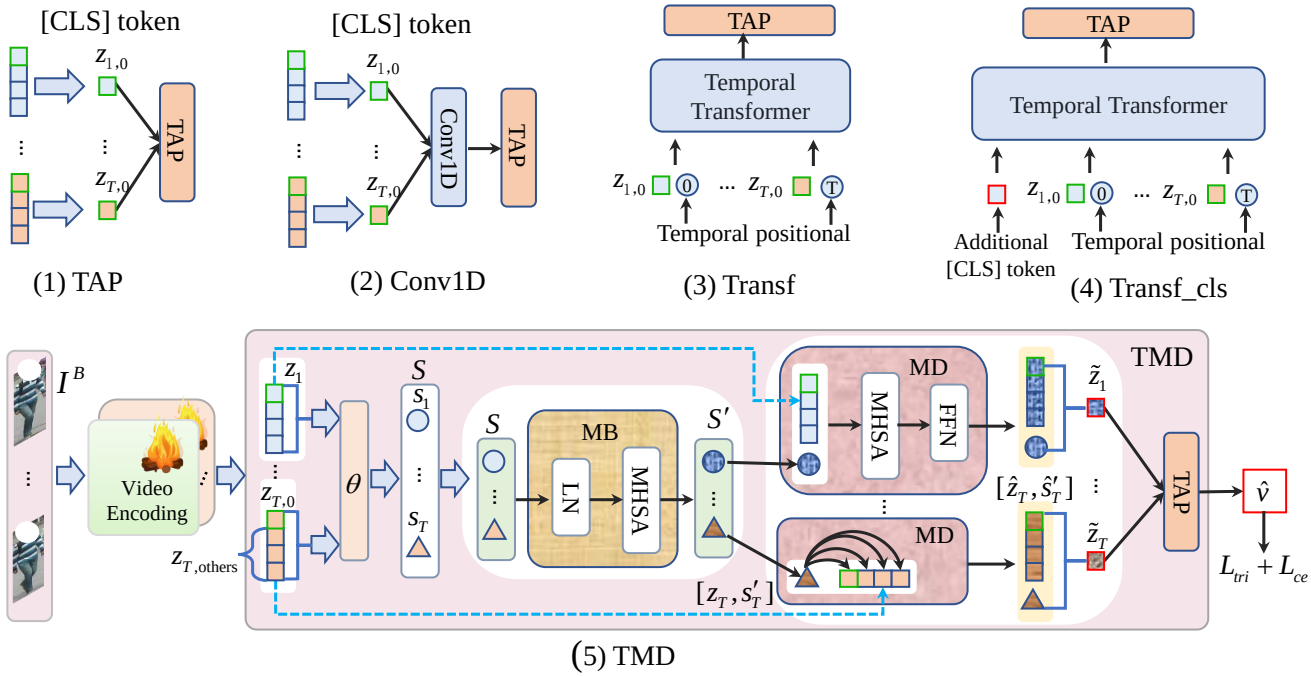

Figure 2: Illustration of (1) TAP, (2) Conv1D, (3) Transf, (4) Transf$_{cls}$ and (5) the proposed TMD module.

frame-level memories in a clip to communicate with each other and can mine temporal information based on the relations obtained from neighbors.

Effectiveness of MD. As shown in Tab .3, the proposed MD method improves the performance remarkably. Compared with $ModelA2$, $ModelA3$ using MD brings 0.4% mAP and 0.4% Rank-1 accuracy gains on MARS, respectively. What's more, compared with $ModelA5$, $ModelA6$ also brings 0.5% Rank-1 accuracy gains on MARS. We believe that MD can diffuse the temporal memories to each token in the original features and obtain more robust sequence features, thus further improving the performance.

Why is there no ablation experiment on V2M loss. Similar to the T2I and I2T loss in CLIP, the video-to-memory contrastive loss denoted by L_{V2M} is proposed to train the SSP module and the visual encoder in the CLIP-Memory. Therefore, V2M loss and CLIP-M are associated. In other words, V2M must be used for supervision when using the CLIP-M module. Otherwise, CLIP-M cannot update the parameters. As a result, as shown in Tab. 7, the ablation study of V2M loss can be equivalent to the corresponding experiments without CLIP-M.

Analysis of Computational Cost. The ablation analysis of computational complexity and the number of parameters are also reported in Tab. 3, 4, 5 and 6. Specifically, as shown in Tab. 4, comparing with the $ModelB2$, $ModelB6$ introduces additional $17.32M$ parameters and $1.27G$ computational complexity (FLOPs). However, compared to $ModelB1$ using text branch, $ModelB6$ reduces parameters and computational complexity by $22.52M$ and $1.71G$, respectively. It is worth noting that $ModelB6$ obtains a higher Rank-1 and mAP (by 1.3% and 1.3%) than $ModelB1$ on MARS. This further confirms the superiority and effectiveness of the proposed method.

As shown in Tab. 6, the proposed method is not sensitive to the hyperparameter of the number of layers in SSP. Compared with $N = 2$, $N = 4$ obtains a higher Rank-1 accuracy by 0.2% on MARS. However, $N = 4$ introduces additional $6.30M$ parameters and $0.84G$ computational complexity. From the perspective of balancing computation and performance, we finally choose $N = 2$, which achieves the best mAP accuracy of 88.8%.

Clarification about the novelty of the TMD module. (1) As shown in Tab. 5, we further investigate the impact of different temporal fusion methods (*e.g.*, TAP, Conv1D, Tranf$_{cls}$ and Transf) with $ModelB2$ on MARS. Moreover, we further visually show the corresponding network structure in Fig. 2. (2) In the TMD module, we first constructs a temporary memory to store time information, and then passes the memory stored information to the frame feature through the MD module. As shown in Fig. 2 (3) and (4), transferring time information can also be achieved by directly use [CLS] tokens of images corresponding to different frames as patch tokens and inputting them into the transformer layer for time dimensional interaction. (3) However, the experimental results in Tab. 5 show that our method outperforms the above methods. Specifically, compared with $Transf$, using TMD brings 1.0% mAP and 0.9% Rank-1 accuracy gains on MARS, respectively. It is worth noting that using TMD reduces parameters by $9.20M$. These experimental results validate the novelty of the TMD module.

Clarification about the novelty of the CLIP-M. (1) To our best knowledge, we are the first to extract identity-specific sequence features denoted as CLIP-M to replace the text features of CLIP. (2) As shown in Tab. 4, compared with $ModleB1$, using CLIP-M brings 0.3% mAP accuracy gains

on MARS. Meanwhile, it is worth noting that using CLIP-M reduces parameters by $39.84M$. These experimental results validate the novelty of the CLIP-M.

Retrieval Results

As shown in Fig. 3, we visualize the retrieval results of some persons. We can see that the top 5 results of the benchmarks are all disturbed to varying degrees by the samples of other identities with similar appearances. We also find that the top 1 results of the method augmented with SSP are matching in the Fig. 3. What's more, the top 5 results of the proposed TF-CLIP are all matching in the Fig. 3 (1) (3) and (4). Thus, the retrieval results prove our proposed TF-CLIP method can help the network to learn a discriminative embedding space.

Limitations and Ethics

In our work, the main limitation may be the huge parameters and computational complexities, since we fine-tuned the pre-trained CLIP visual encoder VIT-B/16 on the ReID benchmarks. In the future, more techniques will be explored to address this problem, such as prompt tuning and Adapter. In addition, video-based person ReID is a human-centric task, while it does not involve the extraction of personal private information. We want to note that for the community to move in the right direction, the studies on ReID should be aware of potential misuses which violates personal privacy. We strictly abide by the data usage specifications and only use publicly available ReID data for academic research. Besides, there are no ethical problems in our methods. Finally, we confirm that the proposed methods can not be used for any militaries.

Pseudocode

As shown in the Fig. 4, 5 and 6, we further show the Pytorch-like pseudocode of the core modules in TF-CLIP, including CLIP-Memory Construction (CC), Sequence-Specific Prompt (SSP) and Temporal Memory Diffusion (TMD).

References

Bai, S.; Ma, B.; Chang, H.; Huang, R.; and Chen, X. 2022. Salient-to-broad transition for video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 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 *Proceedings of the IEEE International Conference on Computer Vision*, 8351–8361.

Dehghan, A.; Modiri Assari, S.; and Shah, M. 2015. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4091–4099.

Eom, C.; Lee, G.; Lee, J.; and Ham, B. 2021. Video-based person re-identification with spatial and temporal memory networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 12036–12045.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; and Ramanan, D. 2010. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9): 1627–1645.

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021a. TransReID: Transformer-based object re-identification. *Proceedings of the IEEE International Conference on Computer Vision*.

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 *Proceedings of the IEEE International Conference on Computer Vision*, 1490–1501.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014–2023.

Hou, R.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2020. Temporal complementary learning for video person re-identification. In *Proceedings of the European Conference on Computer Vision*, 388–405.

Jia, M.; Cheng, X.; Lu, S.; and Zhang, J. 2022. Learning disentangled representation implicitly via transformer for occluded person re-identification. *IEEE Transactions on Multimedia*.

Jin, X.; Lan, C.; Zeng, W.; Wei, G.; and Chen, Z. 2020. Semantics-aligned representation learning for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11173–11180.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Li, H.; Wu, G.; and Zheng, W.-S. 2021. Combined depth space based architecture search for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6729–6738.

Li, J.; Wang, J.; Tian, Q.; Gao, W.; and Zhang, S. 2019. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 3958–3967.

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

Li, S.; Bak, S.; Carr, P.; and Wang, X. 2018. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 369–378.

Li, S.; Sun, L.; and Li, Q. 2022. CLIP-ReID: Exploiting vision-language model for image re-identification without concrete text labels. *arXiv preprint arXiv:2211.13977*.

Figure 3: (1), (2), (3) and (4) are the top 5 retrieval results of different methods on MARS. The query and gallery both are image sequences. The green box represents the correct matching, while the red box is the opposite.

Algorithm 1: CLIP-Memory Construction

Input:
The training set which consists of Y pedestrians and Ns tracklets.
Parameters:
A pre-trained CLIP visual encoder (Frozen). d=512.
Output:
An identity-specific CLIP-Memory

```
Sequence_f_list = [ ]
Label_list = [ ]
With torch.nograd():
    for n_iter, (imgs, labels) in enumerate (train_dataloader_1):
        ## 1.Video encoding Corresponds to Eq.(2)(3)(4) and (5)
        F_sequence = CLIP_visual_encoder(imgs) ## [1, d]
        Sequence_f_list.append(F_sequence )
        Label_list.append(labels)

## 2. Generate an identity-specific CLIP-Memory corresponds to Eq.(6)
##  Sequence_f_list :  [Ns, d]
CLIP_Memory = Generate_Memory(Label_list,
                              Sequence_f_list) ##  [Y, d]
```

Figure 4: Pytorch-like pseudocode for the core of an implementation of CLIP-Memory Construction (CC).

Algorithm 2: Sequence-Specific Prompt

Input:
An identity-specific CLIP-Memory, a training mini-batch which contains P different classes and K video clips for each class. And each clips has T=8 frames.
Parameters:
A pre-trained CLIP visual encoder (Tuned), a sequence-specific prompt(SSP).
d = 512.
Output:
An updated CLIP-Memory.

```
## CLIP_Memory ∈ R^{Y×d}    imgs: [B=PK, T, C, H, W]
for n_iter, (imgs, labels) in enumerate (train_dataloader_2)
    ## 1.Video encoding  corresponds to Eq.(2)(3)(4)(5)
    CLIP_Memory = CLIP_Memory.expand(B, Y, d) ## [B, Y, d]
    F_sequence = CLIP_visual_encoder(imgs) ## [B, d]
    ## 2. Generating SSP corresponds to Eq.(7)
    Prompts = SSP(query=CLIP_Memory,
                  key=F_sequence,
                  value=F_sequence) ## [B, Y, d]
    ## 3.Updating CLIP_Memory corresponds to Eq.(8)
    CLIP_Memory = CLIP_Memory + Prompts   ## [B, Y, d]
    ## 4.Optimize the visual encoder and SSP by Eq.(14)
    Loss_v2m = V2M(CLIP_Memory, f_sequence)
```

Figure 5: Pytorch-like pseudocode for the core of an implementation of Sequence-Specific Prompt (SSP).

Algorithm 3: Temporal Memory Diffusion

Input:
A video clip which has T=8 frames.
Parameters:
A pre-trained CLIP visual encoder (Tuned), a Temporal Memory Construction (TMC) and a memory diffusion (MD). Np=129. D=768.
Output:
A sequence-level feature.

```
imgs: [B=1, T, C, H, W]
for n_iter, (imgs, labels) in enumerate (train_dataloader_2)
        ## 1.Video encoding  corresponds to Eq.(2) and (3)
        F_frame = CLIP_visual_encoder(imgs) ## [T, Np, D]
        ## 2. Constructing a memory token for each frame corresponds to Eq.(9) and (10)
        Init_m_token = TMC(F_frame) ## [T, 1, D]
        F_frame_update = torch.cat([F_frame, Init_m_token], dim=1) ## [T, Np+1, D]
        ## 3. Diffusing the memory token to each frame corresponds to Eq.(11)
        F_frame_update = MD(F_frame_update ) ## [T, Np+1, D]
        ## 4. Aggregating frame-level feature into sequence-level feature corresponds to
Eq.(12) and (13)
        F_sequence = Aggregate(F_frame_update) ## [1, D]
        ## 5.Optimize the visual encoder and TMD by the triplet loss and cross-entropy loss
        Loss_tri = Tri(F_sequence, labels)
        Loss_ce = CE(FC(F_sequence), labels)
```

Figure 6: Pytorch-like pseudocode for the core of an implementation of Temporal Memory Diffusion (TMD).

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 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. *arXiv preprint arXiv:2304.14122*.

Liu, X.; Zhang, P.; and Lu, H. 2023. Video-based Person Re-identification with Long Short-Term Representation Learning. *arXiv preprint arXiv:2308.03703*.

Liu, X.; Zhang, P.; Yu, C.; Lu, H.; Qian, X.; and Yang, X. 2021b. A video is worth three views: Trigeminal transformers for video-based person re-identification. *arXiv:2104.01745*.

Liu, X.; Zhang, P.; Yu, C.; Lu, H.; and Yang, X. 2021c. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 13334–13343.

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

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2014. Person Re-identification by video ranking. In *Proceedings of the European Conference on Computer Vision*, 688–703.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12026–12035.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 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 *Proceedings of the European Conference on Computer Vision*, 549–566. Springer.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2899–2908.

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

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 *Proceedings of the European Conference on Computer Vision*, 868–884.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, 1116–1124.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, 3754–3762.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4692–4702.

Zhu, K.; Guo, H.; Zhang, S.; Wang, Y.; Huang, G.; Qiao, H.; Liu, J.; Wang, J.; and Tang, M. 2021. Aaformer: Auto-aligned transformer for person re-identification. *arXiv preprint arXiv:2104.00921*.