VAug-CLIP: Pose-Guided Augmentation for CLIP-based Video Person Re-identification

Guquan Jing[1,2], Peng Gao[1,2], Yujian Lee[1,2], Hui Zhang[1]*
[1] Department of Computer Science, Beijing Normal-Hong Kong Baptist University, Guangdong, China
[2] Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
{guquanjing1, gaopeng1225, yujianlee1119}@gmail.com, amyzhang@bnbu.edu.cn

***Abstract*—Video-based person re-identification (Re-ID) aims to retrieve particular pedestrians from video sequences over non-overlapping cameras. While existing methods attempt to improve performance with designed networks or multi-modal fusion, they often struggle with misalignment, occlusion, and cross-modal domain gaps. To address these, we propose VAug-CLIP, a novel CLIP-based network that leverages 2D pose priors to alleviate data challenges. Our method features three key components. Before training stage, Pose-guided Augmentation that contains a Pose-based Alignment Module (PAM) rectifies misaligned frames by clustering pose features and recalculating bounding boxes, and a occlusion masking strategy that mines and applies realistic occlusion patterns from the dataset. Furthermore, we design an Augmentation Memory (AM) with a coarse-to-fine update to enhance CLIP-based sequence embeddings, complemented by a Temporal Interaction (TI) module for fine-grained temporal modeling. Extensive experiments on four benchmarks demonstrate that VAug-CLIP consistently achieves state-of-the-art performance, validating the effectiveness of our method.**

***Index Terms*—Video-based person re-identification, multi-modal learning**

I. Introduction

Video-based person re-identification (Re-ID) is an important task in computer vision that aims to match the same pedestrian from video sequences captured by non-overlapping cameras [1]–[4]. Unlike single-shot images used in image-based Re-ID, video sequences contain richer temporal relationships and additional spatial dependencies, enhancing feature representations for the target pedestrian. However, these richer visual cues often come with inherent data challenges, primarily stemming from misalignment and occlusion.

Recent methods [1], [5]–[7] often adopt a two-stage paradigm to construct comprehensive pedestrian representations from video sequences. This typically involves first extracting frame-level features, followed by spatial-temporal feature aggregation via proposed modules. While achieving improved performance, they remain vulnerable to data challenges. Misaligned and occluded frames frequently persist, posing significant challenges to robust feature learning. Some works [8]–[10] leverage other modality such as 2D pose to generate enhanced fusion representations. Despite their benefits, additional modalities introduce computational costs and the challenge of cross-modal feature inconsistency. As shown in Figure 4 (a), these methods introduce a significant domain gap between RGB and multi-modal data, leading to inconsistent features and often requiring complex fusion models. Furthermore, inherent noise such as occlusion and misalignment often leads to sub-optimal feature correlations during aggregation.

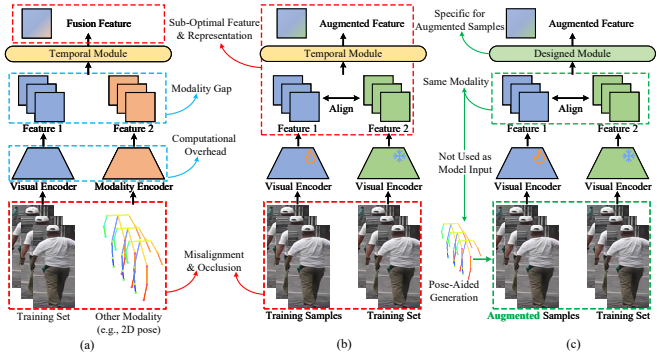

Fig. 1. Illustration of existing architecture in video Re-ID and related challenges. (a) Multi modal methods. (b) TF-CLIP training framework. (c) Our VAug-CLIP framework.

To enhance the quality of representation, recent advancements [2]–[4] leverages Contrastive Language-Image Pre-training (CLIP) [11], showcasing its promising capability in Re-ID. Notably, TF-CLIP [3] introduces a one-stage and text-free learning paradigm, pioneering a novel training manner for video Re-ID. It replace text feature with a dynamically updated CLIP-Memory. This memory is refined via designed module to ensure the framework remains robust to intra-class variations. Furthermore, a temporal module is employed to capture and diffuse temporal dependencies across frame-level tokens. By achieving improved retrieval accuracy, TF-CLIP provides a novel perspective in CLIP-based video Re-ID. Nevertheless, as shown in Figure 4 (b), the data challenges including misalignment and occlusion persist within their framework. These noisy data samples distort the identity-specific cues in CLIP-Memory and degrade temporal dependencies, thereby limiting the model's ability to extract discriminative semantic patterns and causing sub-optimal representation generation.

To mitigate the aforementioned challenges, we propose a novel **V**ideo ReID **Aug**mentation **CILP**-based network (VAug-CLIP) that leverages 2D pose priors. 2D pose offers precise local part positions alongside comprehensive global structural cues. To fully utilize them without causing feature inconsis-

* Corresponding author.

tency, as illustrated in Figure 4 (c), we generates pose-guided augmented samples before training. Specifically, a Pose-based Alignment Module (PAM) and occlusion masking strategy are designed to augment data samples. The PAM identifies misaligned frames by clustering joint features generated from pedestrian poses and rectifies their bounding boxes relative to aligned frames. Furthermore, occlusion pattern mining is performed directly from the source video dataset, applying real-world occlusions onto sampled frames. Our method requires no other modality input during training, but using them to augment image data. This avoids cross-modal inconsistency while effectively alleviating data challenges of misalignment and occlusion. While these augmented samples introduce valuable diversity, existing modules in CLIP-based methods often lack the specialized designs to effectively process such variations. Building upon the text-free paradigm of [3], we introduce tailored refinements specifically designed to integrate and exploit augmented tokens within the existing framework. To effectively incorporate augmented cues, we propose an Augmented Memory (AM) featuring a Coarse-to-Fine Update (C2FU) mechanism. This strategy progressively refines CLIP-based sequence embeddings by integrating augmented tokens from coarse-grained information to fine-grained details, thereby yielding robust and augmented features. Moreover, to perform temporal modeling of these augmented features, we introduce a Temporal Interaction (TI) module consisting of global and semantic branches. This structure captures exhaustive temporal cues from global context and semantic patterns, thereby generating discriminative sequence-level representations. Extensive experiments on four challenging video Re-ID datasets demonstrate that our approach achieves state-of-the-art performance, validating the superiority of the proposed network.

Our contributions are summarized as follows:

- We propose VAug-CLIP, a novel network that leverages 2D pose priors to mitigate data challenges in video Re-ID. It incorporates a Pose-based Alignment Module (PAM) and an occlusion mask generation strategy to mitigate misalignment and occlusions.
- We introduce the Augmentation Memory (AM) with Coarse-to-Fine Update (C2FU) and a Temporal Interaction (TI) module specifically tailored to integrate augmented tokens, which update CLIP-based sequence embeddings and capture comprehensive global-to-semantic temporal cues.
- Extensive experiments on challenging video Re-ID datasets validate the effectiveness of VAug-CLIP, demonstrating the superiority of our proposed method.

II. Related Works

A. Video-based Person Re-identification

Video-based person re-identification (Re-ID) is crucial for retrieving specific pedestrians from video sequences. While early methods model temporal dependencies directly using optical flow, RNN and 3D CNN [12]–[14], they introduced massive parameters and struggled with inherent video data challenges. Recent methods extract spatial-temporal correlations by temporal frame-level feature aggregation [1], [3]–[5]. For instance, DCCT [7] introduces a spatial-temporal complementary learning framework that combines features from CNNs and Transformers. TCViT [1] reduces occlusion through a module that enhances target person features via temporal correlation. TF-CLIP [3] introduces a one-stage text-free paradigm for video Re-ID with CLIP, which avoids the lack of textual descriptions by replacing text features with an CLIP-Memory. Despite these works attempt to boost video ReID performance, these efforts often suffer from misalignment and occlusion. Some methods [8], [15] introduce multi-modal data, such as event data [9], aim to enhance video representations. However, new modalities bring computational burden, complex fusion modules, and a domain gap that compromises feature consistency. Our method requires no other modality input, but using 2D pose to augment image data. This avoids multi-modality drawbacks while effectively alleviating data challenges of misalignment and occlusion.

B. Data Augmentation in Person Re-identification

Data augmentation is widely employed in Re-ID to enhance the model performance. Most methods [16], [17] generate auxiliary data for training. In image-based Re-ID, Zheng et al. [17] propose an end-to-end framework coupling Re-ID learning with data generation. FD-GAN [16] utilizes pose information to create auxiliary images via GANs. However, the quality of these generated images cannot always be guaranteed. For occluded Re-ID, Tan et al. [18] generate photo-realistic occluded samples from real-scene data to enhance model robustness to occlusion. Despite progress across various Re-ID tasks, a comprehensive data augmentation method addressing misalignment and occlusion in video Re-ID remains absent. We propose a pose-guided augmentation method integrated in CLIP-based network to generate distinctive video representations.

III. Methodology

A. Overview

The overall architecture of is illustrated in Figure 2. Before training, for each video tracklet, we first extract 2D pedestrian joints using a pose detector. Each frame yields a set of N joints, represented as $J = \{j^n\}_{n=1}^{N}$. Subsequently, these joints are categorized into five distinct groups for joint featurization. The resulting joint features are then clustered to distinguish between aligned and misaligned frames. For misaligned frames, we employ the proposed Pose-based Alignment Module (PAM) to achieve proper alignment. Following this, real-world occlusion masks are generated and it can randomly applied to the input frames, thereby enhancing model robustness to occlusion. During the training phase, input data is split into two streams following the training strategy [3]: a training set $\mathcal{D}_w$ and a set $\mathcal{D}_s$ sampled via PK sampling [19]. Notably, PAM and occlusion mask are exclusively applied to the $\mathcal{D}_s$. For each input sequence, we

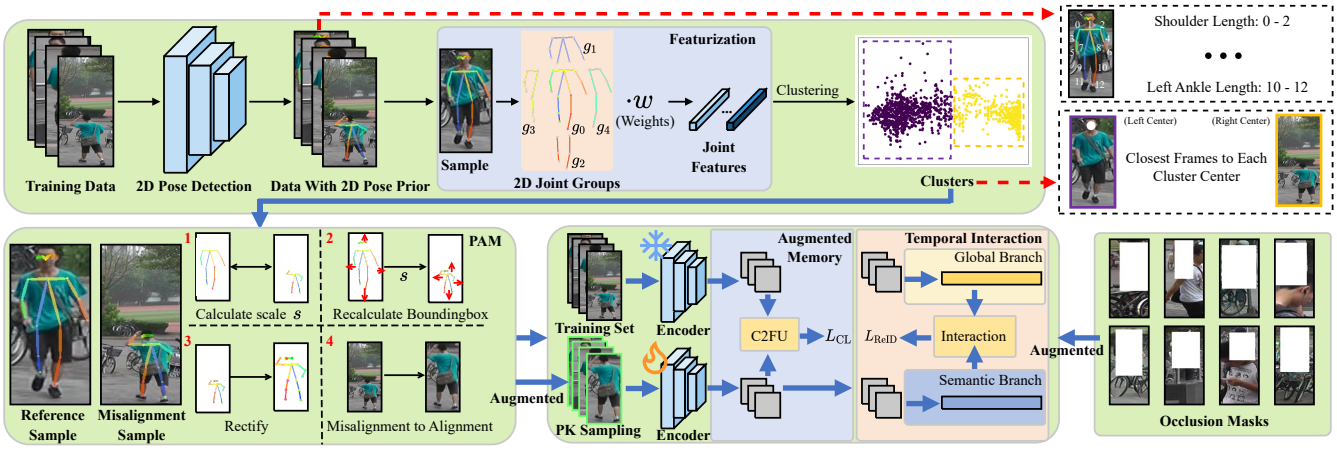

Fig. 2. Overview of VAug-CLIP framework. For the Pose-based Alignment Module (PAM) and occlusion mask generation, we first extract 2D pose and then transform them into joint features. These features are clustered to identify aligned and misaligned frames, followed by a PAM rectification and the application of realistic occlusion masks. This pose-guided augmentation is applied to the PK sampling batch. Subsequently, the Augmented Memory (AM) and Temporal Interaction (TI) modules process data with the augmented samples to yield discriminative representations.

apply the PAM and occlusion mask to only a portion of the frames. This means that PAM and OM are not applied to the entire input sequence. Finally, both sets are fed into the Augmented Memory and Temporal Interaction module to obtain discriminative representations.

B. Joint Featurization

We identify misaligned and aligned frames within a video tracklet by finding that misaligned frames typically exhibit significantly smaller pedestrian skeleton lengths. However, to account for partially occluded frames that could still be well-aligned, we consider different body parts and assign weights to each. In detail, we partition the body joints into five distinct groups for a pose, which are the whole body g_0, upper body g_1, lower body g_2, left body g_3, and right body g_4:

$$\begin{aligned} g_0 &= \{\text{shoulder}, \text{elbow}, \text{wrist}, \text{hip}, \text{knee}, \text{ankle}\} \\ g_1 &= \{\text{shoulder}, \text{elbow}, \text{wrist}, \text{hip}\} \\ g_2 &= \{\text{knee}, \text{ankle}\} \\ g_3 &= \{\text{left\_shoulder}, \text{middle\_shoulder}, \text{left\_elbow}, \\ &\quad \text{left\_wrist}, \text{left\_hip}, \text{left\_knee}, \text{left\_ankle}\} \\ g_4 &= \{\text{right\_shoulder}, \text{middle\_shoulder}, \text{right\_elbow}, \\ &\quad \text{right\_wrist}, \text{right\_hip}, \text{right\_knee}, \text{right\_ankle}\} \end{aligned} \tag{1}$$

For each group $g \in \{g_0, g_1, g_2, g_3, g_4\}$, we extract its joint feature by summing the distances between its constituent joints. This process is further weighted by the visibility confidence of each joint, obtained from the 2D pose detector. The featurization process can be formulated as:

$$f_g = \sum_{(i,k)\in S(g)} v_i v_k \cdot \sqrt{(x_i - x_k)^2 + (y_i - y_k)^2}. \tag{2}$$

Here, $S(g)$ denotes the set of all relevant joint pairs within group g (*e.g.*, shoulder-elbow pairs for upper body). The coordinates for joints i and k are (x_i, y_i) and (x_k, y_k), with v_i and v_k representing their corresponding visibility confidences. The overall joint feature F for a pose is constructed by concatenating the feature f_g from all groups. Afterwards, joint features are clustered to identify aligned and misaligned frame sets. We employ K-Means for this clustering. Figure 2 shows an example of the resulting clusters.

C. Pose-based Alignment Module

To explicitly mitigate misalignment, we propose a Pose-based Alignment Module (PAM) that adjusts misaligned frames based on their overall joint features to achieve proper alignment. As previously discussed, aligned frames typically exhibit larger overall joint feature values, reflecting longer pose skeleton lengths. Thus, the cluster with a higher average feature magnitude (value) is designated as the aligned cluster, while the other is considered the misaligned cluster. From the aligned cluster, we select the frame with overall joint feature closest to the cluster centroid to serve as the reference frame.

The core of PAM is to recalculate the bounding box for each frame in the misaligned cluster based on the reference frame. For the reference frame, we define a pose margin vector $R_r \in \mathbb{R}^4$, where its four components represent the distances from the extremal pose joint (topmost, bottom-most, leftmost, rightmost) including body and head joints to four corresponding boundaries of the bounding box. Our goal is to determine a new margin vector $R_m \in \mathbb{R}^4$ for a misaligned frame. We establish that the ideal margins are proportionally scaled, determined by the relation $R_m = R_r/s$, where s is a scale factor that reflects the difference in pose size between the misaligned frame and the reference frame. To obtain the optimal scale factor, we align the overall joint feature of the misaligned frame $F_m \in \mathbb{R}^5$, with that of the reference frame $F_r \in \mathbb{R}^5$. The optimal value s^* is found by solving the following least-squares minimization problem:

$$s^* = \arg\min_s \|s \cdot F_m - F_r\|_2^2. \tag{3}$$

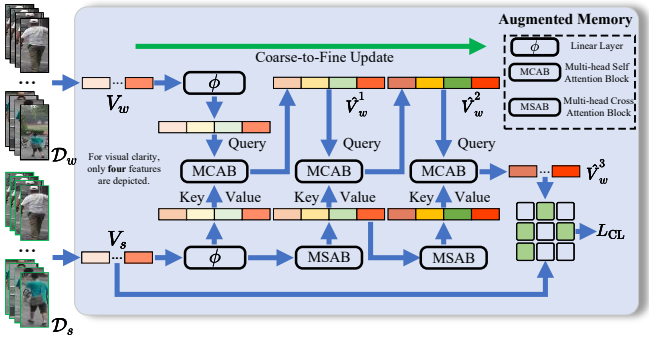

Fig. 3. The illustration of Augmented Memory with Coarse-to-Fine Update.

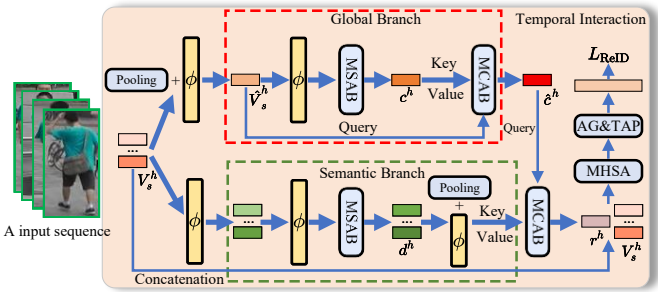

Fig. 4. The illustration of Temporal Interaction module.

This formulation finds the scalar s^* that best maps the pose feature representation of the misaligned frame to that of the well-aligned reference frame.

Once s^* is calculated, we obtain the corrected margin vector R_m. The new, aligned bounding box for the misaligned frame is then constructed by applying these margin values to its own extremal pose joints. Finally, we crop the image patch defined by this new bounding box and resize it to match the aligned frame size for achieving alignment across tracklet. Note that the full set of available pose joints is used when calculating the pose extents for both R_r and R_m to ensure accuracy.

D. Occlusion Mask Generation

To reduce the impact of occlusion, we enhance model robustness to occlusion by applying real-world occlusion masks obtained from existing datasets to the aligned frames. Within the misaligned cluster, frames can be categorized as either unoccluded or occluded. These differ in that occluded frames typically exhibit significantly shorter lengths for partial 2D pose joint pairs (*e.g.*, the length of elbow to wrist) compared to their unoccluded counterparts.

To distinguish these, we calculate the z-score ($z = \frac{L_i - \bar{L}}{\sigma_L}$) for the lengths of relevant joint pairs across different body parts of the pose. Here, L denotes the set of lengths between relevant joint pairs of arbitrary body parts for a specific pedestrian across all input frames. $\bar{L}$ and σ_L represent the mean and standard deviation of L, respectively. By comparing the z-scores of these parts against a predefined threshold, we identify and obtain the occluded frames from the misaligned cluster. Building upon the pedestrian bounding box outputs from the PAM, we extract regions outside the pedestrian in occluded frames as candidate occlusion masks. From these candidates, two masks per tracklet are selected based on the minimal average visibility of the joints in candidate masks, as provided by the pose extractor. To ensure the target pedestrian appears in the results, we additionally filter out occluded frames where poses are located at extreme positions within the frame.

E. Augmented Memory

During the training phase, the PAM and occlusion mask are applied to the $\mathcal{D}_s$ with a probability. A pre-trained CLIP visual encoder with frozen parameters is employed to $\mathcal{D}_w$ to extract representations V_w from input sequences $\{I_t\}_{t=1}^{T}$ in $\mathcal{D}_w$. For a mini-batch $\{I^b\}_{b=1}^{B=P\times K}$ in $\mathcal{D}_s$, comprising P identities and K sequences, we derive representations V_s through a trainable visual encoder. The V_s incorporates augmented tokens from $\mathcal{D}_s$. However, the original CLIP-Memory with SSP in TF-CLIP [3] are designed for standard video tracklets and lack the adaptive capacity to effectively incorporate these augmented tokens. To this end, we design a Coarse-to-Fine Update (C2FU) module, which progressively update cues from augmented tokens through a hierarchical refinement process, ensuring these tokens are integrated comprehensively. In detail, both features are first projected through a linear layer ϕ. We then formulate a cross-attention mechanism, designating V_w as the Query and V_s as the Key and Value to facilitate cross feature interaction. Furthermore, a self-attention block is applied to V_s to augment its spatial dependencies, yielding a more robust global representation. The cross-attention process can be formulated as:

$$\hat{V}_w^j = \mathrm{MCAB}^j(Query, Key, Value), \tag{4}$$

where MCAB denotes the Multi-head Cross Attention (MCA) Block. Each block contains a MLP layer, a MCA and a feed-forward layer, structured with a residual connection. j indexes the index the sequential states within the block. Similarly, the Multi-head Self Attention Block (MSAB) adopts a same structure but employs MSA to refine V_s. We leverage progressive MSABs for augmented tokens and MCABs for CLIP representations to achieve coarse-to-fine update. Eventually, we apply a contrastive learning to the pair (V_s, $\hat{V}_w^3$). Through the AM, augmented tokens are dynamically integrated to enrich the sequence embeddings with detailed augmented cues. This process effectively aligning the representations of the two inputs to enlarge the model's discriminative capacity.

F. Temporal Interaction

Beyond the [CLS] token, patch tokens have been shown to retain discriminative semantic information in video Re-ID [3], [20]. To effectively extract semantic cues from augmented tokens, we employ a dual-branch architecture comprising global and semantic branches. Given a input sequence embedding $V_s^h \in V_s$, the representation is split into global and semantic branches. In the global branch, we first apply average pooling

to obtain average of all tokens, followed by a linear layer to derive the feature $\hat{V}_s^h$. Afterwards, a MSAB is employed to model temporal correlations c^h. To further refine these dynamics, we utilize an MCAB where $\hat{V}_s^h$ serves as the Query and c^h as the Key and Value, effectively yielding the augmented temporal representation $\hat{c}^h$. This process can be expressed as:

$$\begin{aligned} c^h &= \mathrm{MSAB}(\phi(\frac{1}{N}\sum_{n=1}^{N} V_{s,n}^h)) \\ \hat{c}^h &= \mathrm{MSCB}(\hat{V}_s^h, c^h) \end{aligned} \tag{5}$$

In the semantic branch, we bypass the average pooling operation on V_s^h at the beginning to preserve fine-grained semantic cues. Similar to the global branch, a MSAB is first utilized to extract temporal semantic details d^h. These features are subsequently aggregated via average pooling to yield the compact feature $\hat{d}^h$. To utilize complementary cues, a MCAB is employed to model the interaction between the global cues $\hat{c}^h$ and the semantic details $\hat{d}^h$, where d^h serves as the Query to anchor the distinct temporal details. Inspired by [3], we concatenate the resulting features r^h with the original V_s^h and apply a MSAB to propagate the refined temporal information across all tokens. The final sequence-level representation is derived through a combination of aggregation and Temporal Average Pooling (TAP). Departing from previous temporal designs, our TI module effectively captures the rich semantics embedded in augmented tokens, facilitating the learning of discriminative feature.

G. Loss Function

We use contrastive loss and Re-ID loss. For the contrastive loss, it can be formulated as:

$$L_{\mathrm{CL}}(i) = -\frac{1}{|P_i|}\sum_{k\in P_i}\log\frac{\exp(\mathrm{sim}(\mathbf{v}_s^k, \mathbf{v}_w^3)/\tau)}{\sum_{l=1}^{B}\exp(\mathrm{sim}(\mathbf{v}_s^l, \mathbf{v}_w^3)/\tau)}, \tag{6}$$

where P_i is the set of positive pairs and $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity. τ is temperature hyper-parameter. For the Re-ID loss L_{ReID}, the cross-entropy loss with label smoothing and the triplet loss with batch hard mining are used.

IV. Experiments

A. Datasets and Evaluation Protocols

We conduct comprehensive evaluations of our method on four challenging video-based Re-ID datasets: MARS [27], DukeMTMC-VID [28], iLIDS-VID [29] and PRID-2011 [30]. We leverage the DukeMTMC-VID dataset in order to make reasonable comparisons with previous methods. MARS is a large-scale dataset with 17,503 tracklets (1,261 identities, along with 3,248 distractors) captured by six cameras. DukeMTMC-VID (Duke-Video) is another large-scale dataset comprises 4,832 tracklets across 1,812 identities. iLIDS-VID contains 600 tracklets (300 identities) from two non-overlapping cameras. PRID-2011 features pedestrian tracklets from two disjoint cameras, involving 385 and 749 identities. The first 200 identities are present in both cameras. Consistent with established video Re-ID research, we evaluate our method using mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) metrics.

TABLE I
PERFORMANCE (%) COMPARISON ON MARS AND ILIDS-VID. THE BEST AND SECOND RESULTS ARE MARKED AS **BOLD** AND UNDERLINED.

Methods	MARS		iLIDS-VID	
	Rank-1	mAP	Rank-1	Rank-5
PSTA [5]	91.5	85.8	91.5	98.1
DenseIL [21]	90.8	87.0	92.0	98.0
SGWCNN [22]	90.0	85.7	87.8	96.0
MFA [23]	90.9	86.2	93.3	99.3
SINet [24]	91.0	86.2	92.5	-
CaViT [25]	90.8	87.2	93.3	98.0
PiT [26]	90.2	86.8	92.1	98.9
RGCN [8]	91.1	86.5	90.2	98.5
TMT [6]	91.8	86.5	91.3	98.6
DCCT [7]	92.3	87.5	91.7	98.6
VSLA-CLIP [2]	91.8	88.6	95.3	-
TCViT [1]	91.7	87.6	94.3	99.3
TF-CLIPS [3]	93.0	89.4	94.5	99.1
CLIMB-ReID [4]	<u>93.3</u>	**89.7**	**96.7**	**99.9**
Ours	**93.6**	<u>89.6</u>	**96.7**	<u>99.3</u>

TABLE II
PERFORMANCE (%) COMPARISON ON PRID-2011 AND DUKE-VIDEO. THE BEST AND SECOND RESULTS ARE MARKED AS **BOLD** AND UNDERLINED.

Methods	Duke-Video		PRID-2011	
	Rank-1	mAP	Rank-1	Rank-5
GRL [31]	-	-	96.2	99.7
STRF [32]	97.4	96.4	-	-
PSTA [5]	98.3	97.4	95.6	98.9
DenseIL [21]	97.6	97.1	-	-
MSTAT [33]	97.4	96.4	-	-
CaViT [25]	97.6	96.3	95.5	98.9
PiT [26]	-	-	93.2	97.3
FIDN [34]	97.6	97.0	-	-
TMT [6]	97.8	97.5	-	-
HASI [35]	-	-	96.1	98.7
DCCT [7]	<u>98.4</u>	<u>97.6</u>	**96.8**	**99.7**
Ours	**98.7**	**97.7**	**96.8**	<u>98.9</u>

TABLE III
ABLATION STUDY ON THE KEY COMPONENTS.

Components			MARS			Params	FLOPs
PA	AM	TI	mAP	Rank-1	Rank-5		
-	-	-	87.9	91.5	97.4	128.14M	14.31G
-	✓	-	88.5	91.9	97.7	99.51M	12.54G
-	✓	✓	89.0	92.5	97.8	114.63M	13.17G
✓	✓	-	89.2	92.7	97.7	99.51M	12.54G
✓	✓	✓	89.6	93.6	97.8	114.63M	13.17G

B. Implementation Details

We implement our method on the pytorch platform with NVIDIA Tesla V100 GPUs. We leverage ViT-B/16 from CLIP [11] as our visual encoder backbone. 8 frames from the input video sequence are sampled. Each frame is resized to 256

TABLE IV
ABLATION STUDY ON THE JOINT FEATURIZATION.

Method	MARS	
	mAP	Rank-1
Joint Feature w/o w	88.7	91.4
Ours wo w	**89.6**	**93.6**
Only g_0	89.0	92.1
g_0 + g_1 + g_2	89.3	92.7
g_0 + g_3 + g_4	89.2	92.9

TABLE V
ABLATION STUDY ON THE PROBABILITY OF OM.

Proportion	MARS	
	mAP	Rank-1
0	89.3	93.1
0.2	89.5	93.4
0.35	**89.6**	**93.6**
0.6	89.0	92.3

TABLE VI
ABLATION STUDY ON THE PROBABILITY OF PAM.

Proportion	MARS	
	mAP	Rank-1
0	89.2	93.0
0.2	89.3	93.2
0.3	**89.6**	**93.6**
0.5	89.3	92.6

$\times$ 128, and augmented by random cropping and horizontal flipping. Each mini-batch is constructed using a PK sampling strategy, involving 4 identities with 4 tracklets. Training is conducted for 60 epochs via the Adam optimizer. During inference, the final video representation is constructed by concatenating the original sequence-level features with the aggregated features. We employ Euclidean distance as the metric to determine the similarity. We use Alphapose [36] as our 2D pose detector. Probability that the OM is performed is set to 0.35. Probability that the PAM is performed is set to 0.3.

C. Comparison with State-of-the-art Methods

Our comprehensive evaluation in Table I and II demonstrates the effectiveness of our method. We obtain competitive results across various video Re-ID benchmarks. On MARS dataset, our model delivers a superior Rank-1 accuracy of 93.6%, outperforming other CLIP-based methods such as TF-CLIPS [3] and CLIMB-ReID [4]. It also surpasses recent high-performing models like DCCT [7] by 1.3% Rank-1 and 2.1% mAP. We also achieve the second best mAP of 89.6%. For iLIDS-VID, our method achieves a best result of 96.7% Rank-1 and a highly competitive performance with 99.3% Rank-5, significantly exceeding established methods like TF-CLIP. On Duke-Video, we achieve the best performance with 98.7% Rank-1 and 97.7% mAP, slightly surpassing the DCCT. We also reaches a peak Rank-1 of 96.8% on PRID-2011, outperforming previous results like GRL [31] and HASI [35]. These consistent gains across datasets underscore our pose-guided augmentation effectively mitigates misalignment and obstruction. Furthermore, discriminative video representations are captured through the proposed AM and TI modules.

D. Ablation Study

1) Analysis of Main Components: We conduct an ablation study on the MARS dataset to verify the contribution of each proposed component in our framework. As shown in Table III, each component contributes to the final performance. In detail, incorporating the Augmented Memory (AM) into the baseline yields a 0.6% mAP and 0.4% Rank-1 improvement while simultaneously reducing computational complexity, decreasing Params to 99.51M and FLOPs to 12.54G. This implies that AM effectively updates the semantics within augmented tokens, yielding the generation of robust feature through an online update process. The addition of Temporal Interaction (TI) further boosts the results by 0.5% mAP and 0.6% Rank-1, underscoring its efficacy in modeling temporal dependencies. Moreover, applying Pose-guided Augmentation (PA) provides a significant gain, particularly improving Rank-1 accuracy by 1.1% when combined with the full architecture. This demonstrates the robustness afforded by our Pose-guided Augmentation in handling challenging data issues. In particular, the PAM is used to mitigate misalignment, and occlusion masks are applied to improve the model's robustness to occlusion. The full configuration of our method achieves the peak performance of 89.6% mAP and 93.6% Rank-1.

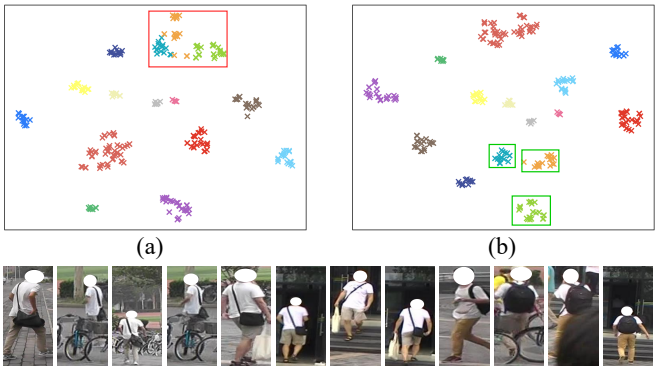

Fig. 5. Visualization of feature distribution results.

2) Analysis of the Joint Featurization: We investigate the impact of applying weights and pose group utilization in equation 1 during joint featurization on MARS in Table IV. Applying visibility weights is crucial for stable joint features. Without them, performance degrade by 0.9% om mAP and 2.2% on Rank-1. This degradation occurs since unobserved or poorly predicted joints are unweighted, corrupting joint feature representation. Furthermore, relying solely on the whole body (g_0) pose group negatively impacts models' performance. Incorporating additional pose groups, such as those focusing on upper/lower body (g_1+g_2) or left/right body joints (g_3+g_4), reliably improves our method effectiveness. This demonstrates that a richer, multi-part feature representation directly enhances the accuracy of our alignment clustering, improving the overall efficacy.

3) Analysis of the Probability of Applying OM: We evaluate the impact of the occlusion mask application probability to model performance on MARS, as detailed in Table V. Performance persistently improves as the probability increased, peaking at 35%. However, when the probability is further raised to 60%, a noticeable performance degradation is ob-

Fig. 6. Visualization of the PAM results.

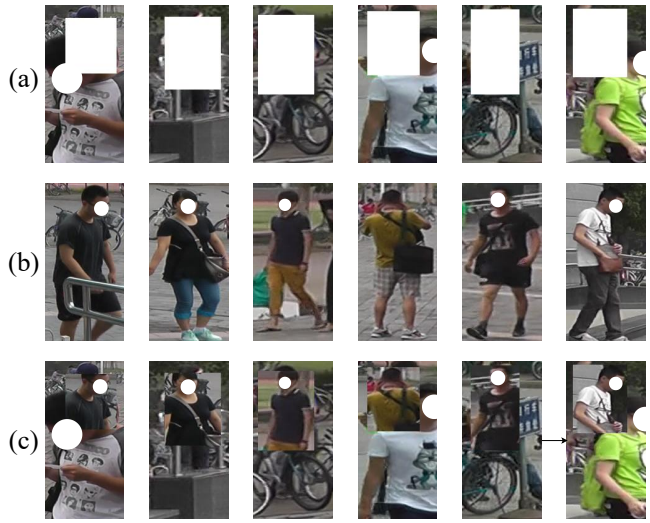

Fig. 7. Visualization of the occlusion mask results.

served. This suggests that while occlusion masks can enhances model robustness to occlusion, applying them excessively could lead to information loss and subsequent performance degradation.

4) Analysis of the Probability of Applying PAM: We investigate the impact of the Pose Alignment Module (PAM) with varying its application probability. As reported in Table VI, our method achieves its optimal performance on MARS at a proportion of 0.3, with slight declines observed at higher or lower values. Deviating from this optimal value leads to a slight performance degradation, suggesting that a balanced alignment frequency is essential for robust feature learning.

E. Visualization

To demonstrate the effectiveness of our method, we provide a series of visualizations. Figure 5 presents the t-SNE visualization of feature distributions from MARS for both the baseline (a) and our VAug-CLIP (b), specifically focusing on challenging pedestrian samples characterized by misalignment and occlusion. We also show these representative pedestrian samples below the t-SNE results. VAug-CLIP yields more compact intra-class feature clusters and larger inter-class margins, indicating the effectiveness of Pose-guided Augmentation. It also highlighted that the learned representations are more discriminative.

As illustrated in Figure 6, the initial video sequences in (a) and (c) often suffer from misalignment, which compromises the overall representation. Processing by the PAM, the frames in the sequence are aligned, shown in (b) and (d). It illustrates how PAM effectively rectifies misalignment, providing the model with aligned inputs for stable feature learning.

Figure 7 demonstrates our occlusion masking. Specifically, (a) presents real-world occlusion masks mined from the dataset, including cyclists, railings, and non-target pedestrians, (b) displays the original video frames, (c) shows the augmented samples through masking. This process enriches the training data with occluded samples, thereby enhancing the model's robustness to occlusion.

V. Conclusion

In this paper, we propose VAug-CLIP, a novel network designed to mitigate data challenges in video Re-ID through pose-guided augmentation. By leveraging 2D pose priors, our Pose-based Alignment Module (PAM) and occlusion masking strategy effectively mitigate misalignment and occlusions. To fully exploit these augmented samples within the CLIP-based architecture, we introduce the Augmentation Memory (AM) with a Coarse-to-Fine Update (C2FU) alongside a Temporal Interaction (TI) module, ensuring integration of augmented tokens and distinctive representation generation. Extensive evaluations across four datasets demonstrate that VAug-CLIP consistently achieves state-of-the-art performance, highlighting the superiority of our method.

Acknowledgment

This work is supported by Guangdong and Hong Kong Universities "1+1+1" Joint Research Collaboration Scheme (2025A0505000003), the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College (2022B1212010006), the Natural Science Foundation of China (62076029), the National Key R&D Program of China (2022YFE0201400), Guangdong Key Lab of AI and Multimodal Data Processing (2020KSYS007), and internal grand of BNBU (R6025A, UICR0300019).

REFERENCES

[1] P. Wu, L. Wang, S. Zhou, G. Hua, and C. Sun, "Temporal correlation vision transformer for video person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 6083–6091.

[2] S. Zhang, W. Luo, D. Cheng, Q. Yang, L. Ran, Y. Xing, and Y. Zhang, "Cross-platform video person reid: A new benchmark dataset and adaptation approach," in *European Conference on Computer Vision*. Springer, 2024, pp. 270–287.

[3] C. Yu, X. Liu, Y. Wang, P. Zhang, and H. Lu, "Tf-clip: Learning text-free clip for video-based person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 6764–6772.

[4] C. Yu, X. Liu, J. Zhu, Y. Wang, P. Zhang, and H. Lu, "Climb-reid: A hybrid clip-mamba framework for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 9, 2025, pp. 9589–9597.

[5] Y. Wang, P. Zhang, S. Gao, X. Geng, H. Lu, and D. Wang, "Pyramid spatial-temporal aggregation for video-based person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 026–12 035.

[6] X. Liu, P. Zhang, C. Yu, X. Qian, X. Yang, and H. Lu, "A video is worth three views: Trigeminal transformers for video-based person re-identification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 9, pp. 12 818–12 828, 2024.

[7] X. Liu, C. Yu, P. Zhang, and H. Lu, "Deeply coupled convolution–transformer with spatial–temporal complementary learning for video-based person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[8] H. Pan, Q. Liu, Y. Chen, Y. He, Y. Zheng, F. Zheng, and Z. He, "Pose-aided video-based person re-identification via recurrent graph convolutional network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7183–7196, 2023.

[9] C. Cao, X. Fu, H. Liu, Y. Huang, K. Wang, J. Luo, and Z.-J. Zha, "Event-guided person re-identification via sparse-dense complementary learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 990–17 999.

[10] R. Lin, A. J. Wang, J. Mo, and M. Li, "Skeletons speak louder than text: A motion-aware pretraining paradigm for video-based person re-identification," *arXiv preprint arXiv:2511.13150*, 2025.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[12] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "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*, 2017, pp. 4747–4756.

[13] G. Chen, Y. Rao, J. Lu, and J. Zhou, "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*. Springer, 2020, pp. 660–676.

[14] X. Gu, H. Chang, B. Ma, H. Zhang, and X. Chen, "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*. Springer, 2020, pp. 228–243.

[15] K. Han, Y. Huang, S. Gong, L. Wang, and T. Tan, "3d shape temporal aggregation for video-based clothing-change person re-identification," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 2371–2387.

[16] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang *et al.*, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," *Advances in neural information processing systems*, vol. 31, 2018.

[17] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2138–2147.

[18] L. Tan, J. Xia, W. Liu, P. Dai, Y. Wu, and L. Cao, "Occluded person re-identification via saliency-guided patch transfer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 5, 2024, pp. 5070–5078.

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

[20] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, and C. Wang, "Transfg: A transformer architecture for fine-grained recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 1, 2022, pp. 852–860.

[21] T. He, X. Jin, X. Shen, J. Huang, Z. Chen, and X.-S. Hua, "Dense interaction learning for video-based person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1490–1501.

[22] Y. Yao, X. Jiang, H. Fujita, and Z. Fang, "A sparse graph wavelet convolution neural network for video-based person re-identification," *Pattern Recognition*, vol. 129, p. 108708, 2022.

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

[24] S. Bai, B. Ma, H. Chang, R. Huang, and X. Chen, "Salient-to-broad transition for video person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7339–7348.

[25] J. Wu, L. He, W. Liu, Y. Yang, Z. Lei, T. Mei, and S. Z. Li, "Cavit: Contextual alignment vision transformer for video object re-identification," in *European Conference on Computer Vision*. Springer, 2022, pp. 549–566.

[26] X. Zang, G. Li, and W. Gao, "Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 12, pp. 8776–8785, 2022.

[27] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer, 2016, pp. 868–884.

[28] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5177–5186.

[29] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*. Springer, 2014, pp. 688–703.

[30] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis: 17th Scandinavian Conference, SCIA 2011, Ystad, Sweden, May 2011. Proceedings 17*. Springer, 2011, pp. 91–102.

[31] X. Liu, P. Zhang, C. Yu, H. Lu, and X. Yang, "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*, 2021, pp. 13 334–13 343.

[32] A. Aich, M. Zheng, S. Karanam, T. Chen, A. K. Roy-Chowdhury, and Z. Wu, "Spatio-temporal representation factorization for video-based person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 152–162.

[33] Z. Tang, R. Zhang, Z. Peng, J. Chen, and L. Lin, "Multi-stage spatio-temporal aggregation transformer for video person re-identification," *IEEE Transactions on Multimedia*, vol. 25, pp. 7917–7929, 2022.

[34] L. Liu, X. Yang, N. Wang, and X. Gao, "Frequency information disentanglement network for video-based person re-identification," *IEEE Transactions on Image Processing*, 2023.

[35] S. Chen, H. Da, D.-H. Wang, X.-Y. Zhang, Y. Yan, and S. Zhu, "Hasi: Hierarchical attention-aware spatio–temporal interaction for video-based person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 6, pp. 4973–4988, 2024.

[36] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 6, pp. 7157–7173, 2022.