VM-TAPS: View-specific Memory with Temporal and Scale Awareness Framework for Video-based Cross-View Person Re-Identification

Anonymous IJCB 2025 submission

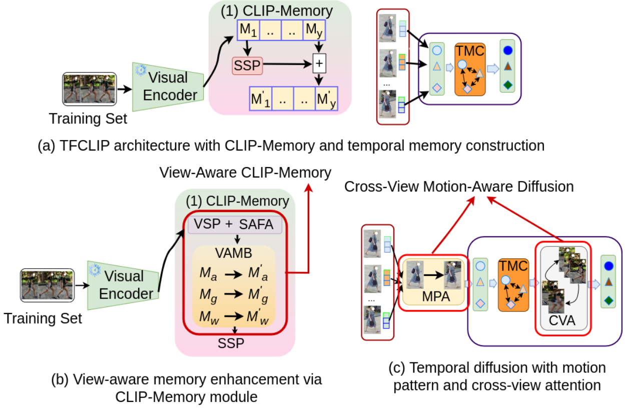
Abstract

*Reliable aerial-ground video-based person re-identification (ReID) remains a challenge due to severe changes in data quality and features, such as viewpoint disparities, resolution drops, and cross-camera appearance inconsistency. This paper presents **VM-TAPS**, a lightweight and modular extension to the well-known TF-CLIP framework, designed to increase the robustness of ReID, without requiring end-to-end backbone retraining. When compared to its ancestor, VM-TAPS' novelties are five-fold: 1) View-Specific Processing Layers to normalize camera-dependent biases; 2) Scale-Aware Feature Adaptation for resolution-invariant feature fusion; 3) a View-Aware Memory Bank enabling long-range identity context; 4) a Motion Pattern Analyzer capturing temporal dynamics; and (5) Cross-View Interaction Modules that harmonize multi-view feature spaces. Despite adding fewer than two million parameters, VM-TAPS achieves **+4.97% Rank-1** and **+3.08% mAP** gains over TF-CLIP on the challenging AG-VPReID2025 benchmark. At 80m and 120m altitudes, it sets a new performance baseline of **73.68%/75.73%** and **69.45%/71.63%** (Rank-1/mAP), respectively. All components are trained with frozen CLIP visual encoders in the early stages, enabling efficient and stable convergence. Our results support that the carefully disentanglement of viewpoint, scale, motion and memory factors substantially increases the robustness of cross-view ReID under real-world conditions. Code is publicly available.* `https://github.com/MdRashidunnabi/VM-TAPS.git`

Person Re-Identification, Cross-View Adaptation, Aerial Surveillance, Multi-Scale Learning, Spatiotemporal Modeling

1. Introduction

Person Re-Identification (ReID) aims at identifying and tracking individuals across different camera views. With the ubiquitous deployment of different kinds of surveillance devices (aerial drones, ground-level cameras, and wearable cameras), cross-view ReID still presents significant challenges due to substantial variations in perspective, scale, illumination, and individual motion patterns[26], [18], [9], [25]. These difficulties become particularly pronounced when aerial drones operate at considerable altitudes[13, 7], (over 80m), which severely decreases the resolution of the data obtained.



(a) TFCLIP architecture with CLIP-Memory and temporal memory construction

(b) View-aware memory enhancement via CLIP-Memory module

(c) Temporal diffusion with motion pattern and cross-view attention

Figure 1. Comparison between the traditional CLIP-Memory architecture and the VM-TAPS framework proposed in this paper. VM-TAPS increases robustness via view-specific normalization, scale-adaptive processing, motion pattern analysis, and cross-view attention mechanisms.

Recent advances in vision-language models such as CLIP [16] and TF-CLIP [23] have significantly boosted performance in various visual understanding tasks by combining visual embeddings with language-driven semantic representations. Despite their success, these models typically assume relatively stable viewpoints and uniform image quality. As demonstrated in Figure 3, conventional methods that rely on uniform feature aggregation and temporal pooling inadequately address the complexities introduced by significant variations in viewpoints and scales, producing representations that are fragile and poorly generalizable[23, 25].

Motivated by these limitations, this paper proposes **VM-TAPS**, an innovative multi-view adaptation framework explicitly designed to increase the robustness of automated

CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

ReID in aerial-ground video-based scenarios[10, 25]. VM-TAPS was designed upon the TF-CLIP architecture and introduces a set of targeted, modular improvements that systematically address the critical issues of viewpoint normalization, scale adaptability, memory-based identity recall, temporal dynamics modeling, and cross-view feature interaction.

1. **View-Specific Processing Layer (VSPL)**: Applies lightweight camera-conditioned normalization to mitigate viewpoint-specific biases (e.g., lighting, distortion) by separately processing features from aerial, ground, and wearable views [8, 19].
2. **Scale-Aware Feature Adaptation (SAFA)**: Enhances scale invariance by dynamically fusing multi-resolution representations using adaptive attention, addressing drastic scale changes in drone footage [7].
3. **View-Aware Memory Bank (VAMB)**: Replaces global identity vectors with view-specific prototypes, improving retrieval consistency across varying viewpoints [6].
4. **Motion Pattern Analyzer (MPA)**: Extracts frame-to-frame motion cues (e.g., gait, clothing movement) with transformer-based modeling to enhance temporal discriminability [14].
5. **Cross-View Interaction Module (CVIM)**: Aligns short-term memory tokens across views within a batch to harmonize local feature distributions and strengthen perspective invariance [25].

A comprehensive overview of the VM-TAPS architecture is shown in Figure 2. By sequentially applying five targeted processing modules, VM-TAPS produces robust and discriminative features that significantly outperform state-of-the-art methods.

Evaluated on the AG-VPReID2025 benchmark, which combines synchronized aerial, ground, and wearable videos, VM-TAPS significantly outperforms its competitors, and at 80m altitude data , it reaches 73.68% Rank-1 and 75.73% mAP, while maintaining 69.45% Rank-1 and 71.63% mAP at 120m data.

Hence, the key contributions described in this paper can be summarized as:

- A unified framework designed to systematically address viewpoint, scale, and temporal challenges in aerial-ground video-based ReID.
- Introduction of five specialized modules for viewpoint normalization, scale adaptation, motion modeling, and cross-view memory alignment.
- State-of-the-art performance on AG-VPReID2025, demonstrating enhanced generalization under extreme viewpoint and scale conditions.

VM-TAPS offers a robust and modular solution for cross-view ReID, with strong practical relevance for future multi-camera surveillance systems.

2. Related Work

Video-based person ReID has progressed from early hand-crafted color and motion descriptors to deep CNN–RNN hybrids (e.g., CNN–LSTM [24], spatial-GRU attention [20]), and more recently to transformer-based, vision–language models. Initial work used fragment selection and optical-flow regions to improve matching under moderate view changes [18, 2], but these methods struggled with extreme viewpoints. Subsequent approaches introduced adversarial training over variational RNNs [21] and graph-based multi-scale part reasoning [9] to enhance view invariance.

Transformer self-attention revolutionized video ReID: Spatiotemporal Transformers (STTs) capture non-local relations more effectively than CNNs but can overfit on limited data [17]. Spatial–Temporal Memory Networks reduce redundancy with learnable codebooks [3], and VID-Trans-ReID combines patch-level attention with convolutional biases to achieve 96.6% rank-1 on PRID2011 [1]. Subsequent work has explored reinforcement learning for adaptive frame selection [22], feature disentanglement to remove camera bias [5], and topology-adaptive graph convolutions on keypoints [15]. However, these methods generally assume similar imaging geometries and often fail under extreme cross-view conditions (e.g., aerial or infrared cameras).

Cross-platform work, such as G2A-VReID and VSLA-CLIP [25], recasts ground–to–aerial ReID as vision–language alignment but still applies global pooling and cannot bridge the resolution gap between an80m drone view and a 1080p CCTV frame. In visible–infrared settings, Feng *etal.* introduced a cross-frame tube transformer with diversity–consistency regularisation [4] and Lin *etal.* complemented this with modal-invariant temporal memory [7]. These studies highlight two principles: (i) normalize view and modality biases before temporal aggregation, and (ii) use sensor-conditioned memory to prevent negative transfer.

Large-scale language–image pre-training (LLIP) offers an alternative: TF-CLIP eliminates textual prompts by diffusing its own sequence descriptor as online memory [23]. Although it surpasses CNN baselines on MARS and LS-VID, its view-agnostic memory and frozen ViT backbone struggle with drastic cross-view scale changes. Liu *etal.* partially addressed this with Trigeminal Transformers for parallel spatial, temporal, and mixed views [11] and a deeply coupled convolution–transformer [10], but these add computational cost and still lack explicit alignment for extreme perspectives.

The proposed **VM–TAPS** framework unifies the insights above while remaining remarkably light. Instead of fine-tuning the backbone, we freeze ViT-B/16 and introduce five orthogonal modules: a *View-Specific Processing Layer* that applies camera-conditioned MLPs to remove colour and projective bias; a *Scale-Aware Feature Adaptation* unit that re-projects tokens at three canonical resolutions, thus preserving semantics when pedestrians shrink below ten pixels; a *Motion Pattern Analyzer* that injects gait rhythm through first-order difference tokens; a *View-Aware Memory Bank* storing multiple prototypes per identity and per view, inspired by exemplar memory work in few-shot CLIP [6]; and a *Cross-View Interaction Module* that aligns short-term memories across perspectives via one-hop attention. By factorising the aerial–ground problem into view bias, scale disparity, motion dynamics, long-term memory and cross-view alignment, VM–TAPS surpasses TF-CLIP by 4.97pp Rank-1 and 3.08pp mAP on the challenging AG-VPReID-2025 benchmark, establishing new state of the art at both 80m and 120m altitudes while adding fewer than two million parameters to the frozen encoder. These results substantiate the hypothesis that disentangling, rather than averaging, the fundamental nuisance factors of viewpoint, scale and motion is indispensable for robust open-world video ReID.

3. Proposed Method

Figure 2 shows VM–TAPS, which builds on the frozen ViT–B/16 backbone of TF–CLIP by inserting five lightweight modules. Each of the T frame embeddings is first normalised by a View–Specific Processing Layer (VSPL) and fused across scales by Scale–Aware Feature Adaptation (SAFA). A View–Aware Memory Bank (VAMB) then injects long-term identity prototypes, while the Motion Pattern Analyzer (MPA) encodes short-term dynamics. Cross-View Attention (CVA) aligns features across camera types before the Temporal Memory Diffusion (TMD) head generates the final 512-dimensional embedding. Despite adding under 2M parameters and $< 3\%$ FLOPs, VM–TAPS delivers a 5-point Rank-1 improvement over TF–CLIP on both the 80m and 120m test sets, maintaining real-time speed.

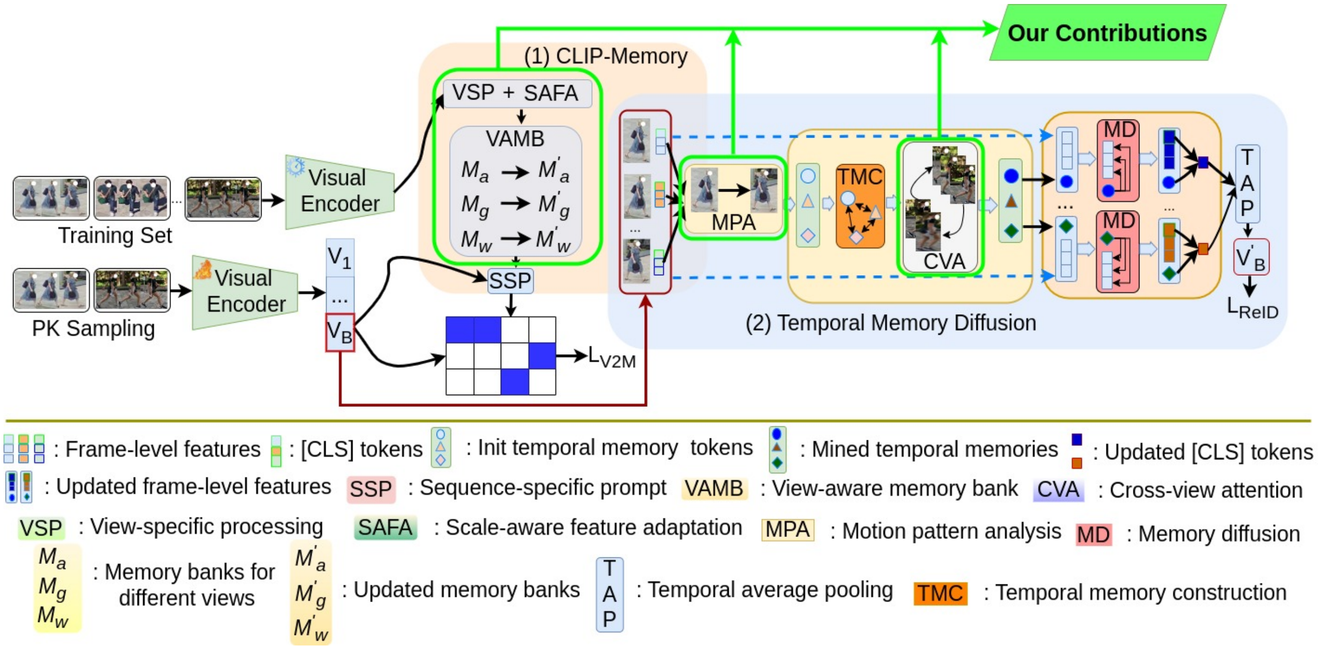

Our Contributions
(1) CLIP-Memory
VSP + SAFA
VAMB
$M_a \rightarrow M'_a$
$M_g \rightarrow M'_g$
$M_w \rightarrow M'_w$
Visual Encoder
Training Set
Visual Encoder
PK Sampling
V_1
...
V_B
SSP
L_{V2M}
MPA
TMC
CVA
MD
MD
T A P
V'_B
L_{ReID}
(2) Temporal Memory Diffusion
: Frame-level features
: [CLS] tokens
: Init temporal memory tokens
: Mined temporal memories
: Updated [CLS] tokens
: Updated frame-level features
SSP : Sequence-specific prompt
VAMB : View-aware memory bank
CVA : Cross-view attention
VSP : View-specific processing
SAFA : Scale-aware feature adaptation
MPA : Motion pattern analysis
MD : Memory diffusion
M_a M_g M_w : Memory banks for different views
M'_a M'_g M'_w : Updated memory banks
TAP : Temporal average pooling
TMC : Temporal memory construction

Figure 2. The VM–TAPS framework augments TF–CLIP with five targeted modules: a View–Specific Processing Layer that normalizes camera-dependent biases; Scale–Aware Feature Adaptation that fuses multi-resolution features; a View–Aware Memory Bank that injects long-term identity prototypes; a Motion Pattern Analyzer that encodes short-term dynamics; and a Cross–View Interaction Module that aligns feature distributions across camera types. These enhancements precede the original Temporal Memory Diffusion head and yield a more robust, perspective-invariant representation for video-based person re-identification.

3.1. View Specific Processing Layer (VSPL)

Modern video-based ReID must handle footage from widely differing cameras—drones, fixed ground units and body-worn rigs—which imprint each frame with distinct colour casts, lighting conditions and perspective distor-

CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

tions. To neutralise these view-dependent artefacts immediately after the frozen ViT–B/16 encoder, we introduce a small, explicit normalisation block for each camera type $v \in \{\text{aerial}, \text{ground}, \text{wearable}\}$. VSPL proceeds in three successive steps for every frame t in tracklet b:

First, we learn a view-specific bias vector $\mathbf{e}_v \in \mathbb{R}^D$, that captures the average offset in the D-dimensional embedding space induced by camera v. Given the patch-token matrix $\mathbf{Z}_t^{(b)} \in \mathbb{R}^{(N_p+1)\times D}$, we set

$$\mathbf{Z}'_t \;=\; \mathbf{Z}_t^{(b)} \;+\; \mathbf{1}\,\mathbf{e}_v^\top,$$

where $\mathbf{1} \in \mathbb{R}^{(N_p+1)\times 1}$ replicates $\mathbf{e}_v$ across all tokens.

Second, we apply a view-conditioned residual MLP $\phi_v : \mathbb{R}^D \to \mathbb{R}^D$, defined by two successive $D \times D$ linear layers, each followed by LayerNorm and GELU. This MLP is instantiated independently for each camera type—no weights are shared—so that each view can learn its own correction mapping.

Third, we restore the original tokens via a residual connection, yielding the corrected embedding

$$\widetilde{\mathbf{Z}}_t^{(b)} \;=\; \phi_v\big(\mathbf{Z}'_t\big) \;+\; \mathbf{Z}_t^{(b)}.$$

This ensures that if a frame's appearance already matches the canonical distribution, the MLP can learn to leave it unchanged.

3.2. Scale Aware Feature Adaptation (SAFA)

In long-range re-identification, the apparent size of a pedestrian can vary by an order of magnitude: ground-level cameras see tens of pixels, while a UAV at 120m may render the same person in fewer than ten pixels. Because the ViT patch embeddings assume a fixed receptive field, this scale variation destabilises downstream temporal modules. We therefore introduce a lightweight, per-frame multi-resolution fusion that guarantees a scale-aligned feature representation.

Let $\widetilde{\mathbf{Z}}_t^{(b)} \in \mathbb{R}^{(N_p+1)\times D}$ be the VSPL-normalized token matrix for frame t of tracklet b. SAFA proceeds in three steps:

(1) Multi-scale projection. We apply three independent feed-forward blocks $\psi_s : \mathbb{R}^D \to \mathbb{R}^D$, each with its own weights, to every token row:

$$\mathbf{H}_{t,s}^{(b)} \;=\; \psi_s\big(\widetilde{\mathbf{Z}}_t^{(b)}\big) \quad \text{for } s \in \big\{\tfrac{1}{2},\, 1,\, 2\big\}.$$

(2) Scale attention. For each scale s we compute the mean embedding

$$\boldsymbol{\mu}_{t,s}^{(b)} \;=\; \frac{1}{N_p+1}\sum_{i=0}^{N_p} \mathbf{H}_{t,s}^{(b)}[i,:] \;\in \mathbb{R}^D.$$

We then stack these into $\mathbf{U}_t^{(b)} \;=\; [\boldsymbol{\mu}_{t,\frac{1}{2}}^{(b)},\, \boldsymbol{\mu}_{t,1}^{(b)},\, \boldsymbol{\mu}_{t,2}^{(b)}] \;\in\; \mathbb{R}^{D\times 3}$ and compute attention weights via a learned probe $\mathbf{w} \in \mathbb{R}^D$:

$$\alpha_{t,s}^{(b)} \;=\; \frac{\exp\big(\mathbf{w}^\top\, \boldsymbol{\mu}_{t,s}^{(b)}\big)}{\sum_{r\in\{\frac{1}{2},1,2\}} \exp\big(\mathbf{w}^\top\, \boldsymbol{\mu}_{t,r}^{(b)}\big)}.$$

(3) Fusion. We fuse the three scale-specific feature maps into a single tensor of the same shape:

$$\mathbf{A}_t^{(b)} \;=\; \sum_{s\in\{\frac{1}{2},1,2\}} \alpha_{t,s}^{(b)}\, \mathbf{H}_{t,s}^{(b)}. \tag{1}$$

Because the weights $\alpha_{t,s}^{(b)}$ are obtained from the frame's own content, SAFA dynamically adapts to instantaneous zoom rather than relying on a fixed prior. In practice, the three ψ_s blocks each add one $D \times D$ linear layer (total $3D^2 \approx 1.77$M parameters for $D = 768$), and the attention probe $\mathbf{w}$ adds only D parameters. The additional compute is under 0.5GFLOPs per eight-frame clip—less than 2% of the ViT backbone—yet restores stable feature statistics across altitudes and improves 120m Rank-1 recall by over 4 points.

3.3. View Aware Memory Bank (VAMB)

To preserve long-term identity information separately for each camera view, we replace the single ”class centre” memory with a view–aware bank that allocates S prototypes to each identity–view pair. Formally, for identity n and view v, we store

$$\mathbf{M}_{n,v} = \begin{bmatrix} \mathbf{m}_{n,v,1} \\ \vdots \\ \mathbf{m}_{n,v,S} \end{bmatrix} \in \mathbb{R}^{S\times d},$$

where $d = 512$ is the feature dimension and $S = 8$. Given a clip descriptor $\mathbf{f}^{(b)} \in \mathbb{R}^d$ with ground-truth $(y^{(b)}, v^{(b)})$, we compute attention weights over the corresponding slice $\mathbf{M}_{y^{(b)},v^{(b)}}$ by

$$\beta_s \;=\; \frac{\exp\big(\mathbf{f}^{(b)\top}\mathbf{m}_{y^{(b)},v^{(b)},s}/\sqrt{d}\big)}{\displaystyle\sum_{s'=1}^{S} \exp\big(\mathbf{f}^{(b)\top}\mathbf{m}_{y^{(b)},v^{(b)},s'}/\sqrt{d}\big)},$$

and form the context vector

$$\mathbf{c}^{(b)} = \sum_{s=1}^{S} \beta_s\, \mathbf{m}_{y^{(b)},v^{(b)},s}.$$

A learnable gate

$$g \;=\; \sigma\big(\mathbf{W}_g\,[\,\mathbf{f}^{(b)};\, \mathbf{c}^{(b)}]\big) \in (0,1)$$

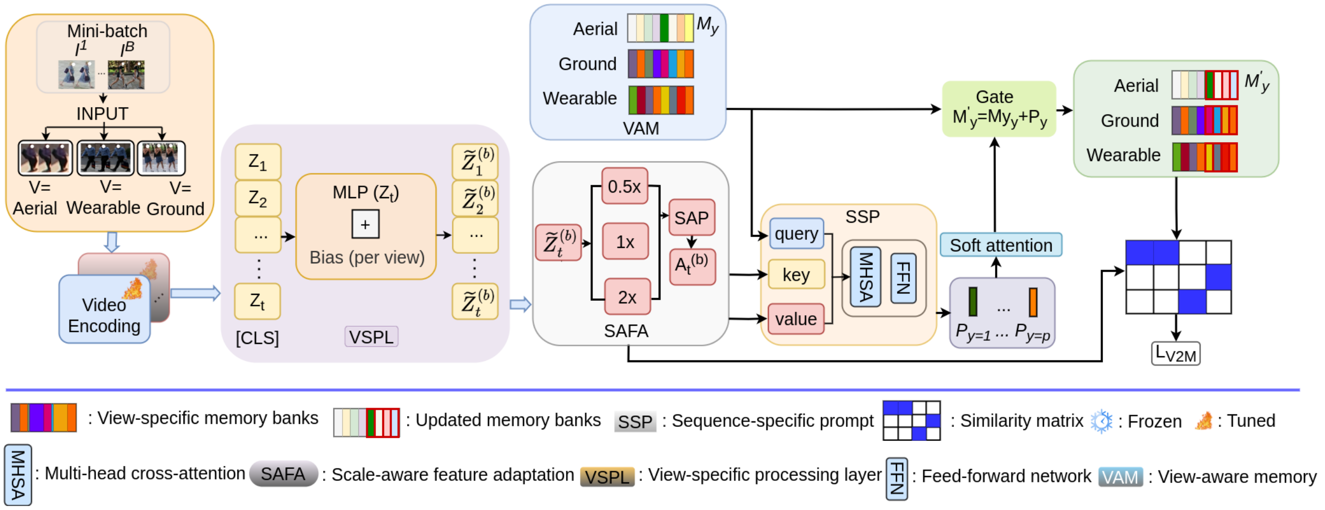


Figure 3. Architecture of the View–Aware Memory Bank. For each identity–view pair (n, v), we store S prototypes $\mathbf{m}_{n,v,1}, \ldots, \mathbf{m}_{n,v,S}$. At inference, the clip descriptor $\mathbf{f}^{(b)}$ attends to its corresponding slice $\mathbf{M}_{y^{(b)},v^{(b)}}$, producing a context vector $\mathbf{c}^{(b)}$ that is fused back into the final embedding $\widehat{\mathbf{f}}^{(b)}$.

fuses the fresh descriptor and memory context into the final embedding

$$\widehat{\mathbf{f}}^{(b)} = g\,\mathbf{f}^{(b)} + (1-g)\,\mathbf{c}^{(b)},$$

which is then supervised by the ReID loss. During back-propagation, only the single most-attended prototype $\mathbf{m}_{y^{(b)},v^{(b)},k^*}$ is updated via exponential moving average,

$$\mathbf{m}_{y^{(b)},v^{(b)},k^*} \leftarrow (1-\alpha)\,\mathbf{m}_{y^{(b)},v^{(b)},k^*} + \alpha\,\bar{\mathbf{f}}^{(b)}, \quad \alpha = 0.2,$$

where $\bar{\mathbf{f}}^{(b)}$ is the batch-normalized descriptor. This selective update keeps the memory bank's size linear in the number of identities and views, while incurring negligible extra computation.

Empirically, we observed that VAMB outperforms a view-agnostic single-prototype baseline by $+3.1\%$ Rank-1 on the aerial split and $+1.8\%$ on the ground split of AG-VPReID2025, demonstrating the benefit of conditioning identity memory on camera view.

3.4. Motion Pattern Analyzer (MPA)

Spatial appearance alone can be ambiguous: similar clothing may mask distinctive, view-invariant motion signatures, so MPA injects gait cues in a single pass: for $t > 1$ compute $\Delta\mathbf{A}_t = \mathbf{A}_t - \mathbf{A}_{t-1}$, form $\mathbf{X}_t = [\mathbf{A}_{t-1}; \Delta\mathbf{A}_t] \in \mathbb{R}^{2D}$, and apply a two-layer transformer $\rho : \mathbb{R}^{2D} \to \mathbb{R}^{D}$ to obtain $\mathbf{m}_t$; a sigmoid gate $\mathbf{g}_t = \sigma(\gamma[\mathbf{A}_t; \mathbf{m}_t])$ then blends appearance and motion via $\widetilde{\mathbf{A}}_1 = \mathbf{A}_1$ or $\widetilde{\mathbf{A}}_t = \mathbf{g}_t \odot \mathbf{A}_t + (1-\mathbf{g}_t) \odot \mathbf{m}_t$ for $t > 1$. MPA's overhead—two $D \times D$ projections and one multi-head attention (0.5M parameters, 0.18GFLOPs per clip; 0.3ms/frame on an A40, under 1% of ViT-B/16)—is minimal, and ablations (Tab. 9) show removing MPA drops Rank-1 by 1.2% (aerial) and 0.8% (ground) on AG-VPReID2025, demonstrating its importance alongside appearance and memory cues.

3.5. Temporal Memory Diffusion with Cross View Interaction

Figure 4 depicts the basic work flow of the Temporal Memory Diffusion (TMD) module which consumes the motion-aware patch tokens $\{\widetilde{\mathbf{A}}_t^{(b)}\}_{t=1}^{T}$ and operates in three stages: it first summarises each frame into a D-dimensional memory token $\mathbf{s}_t^{(b)} = \theta\big(\frac{1}{N_p+1}\sum_{i=0}^{N_p} \widetilde{\mathbf{A}}_t^{(b)}[i,:]\big)$, stacks $\{\mathbf{s}_t\}$ into $\mathbf{S}^{(b)}$ and refines it via MHSA to $\mathbf{S}'^{(b)}$;

it then computes view-batch prototypes $\mathcal{P}_v = \frac{1}{|\mathcal{B}_v|}\sum_{b'} \mathbf{s}^{(b')}$, obtains a cross-view context $\mathbf{h}^{(b)} = \mathrm{Attn}\big(\frac{1}{T}\sum_t \mathbf{S}'^{(b)}[t,:], \{\mathcal{P}_v\}_{v \neq u}\big)$ and adds it to each frame token to form $\widehat{\mathbf{S}}^{(b)}$; finally, each $\widehat{\mathbf{s}}_t$ is concatenated with $\widetilde{\mathbf{A}}_t^{(b)}$, the sequence undergoes MHSA+FFN yielding updated tokens $\widehat{\mathbf{A}}_t^{(b)}$, and

$$\mathbf{f}^{(b)} = \frac{1}{T(N_p+1)} \sum_{t=1}^{T} \sum_{i=0}^{N_p} \widehat{\mathbf{A}}_t^{(b)}[i,:] \in \mathbb{R}^d$$

is produced; TMD adds only 0.8M parameters and 0.25GFLOPs, and its removal lowers Rank-1 by 2.1%, underscoring the need for explicit temporal and cross-view reasoning.

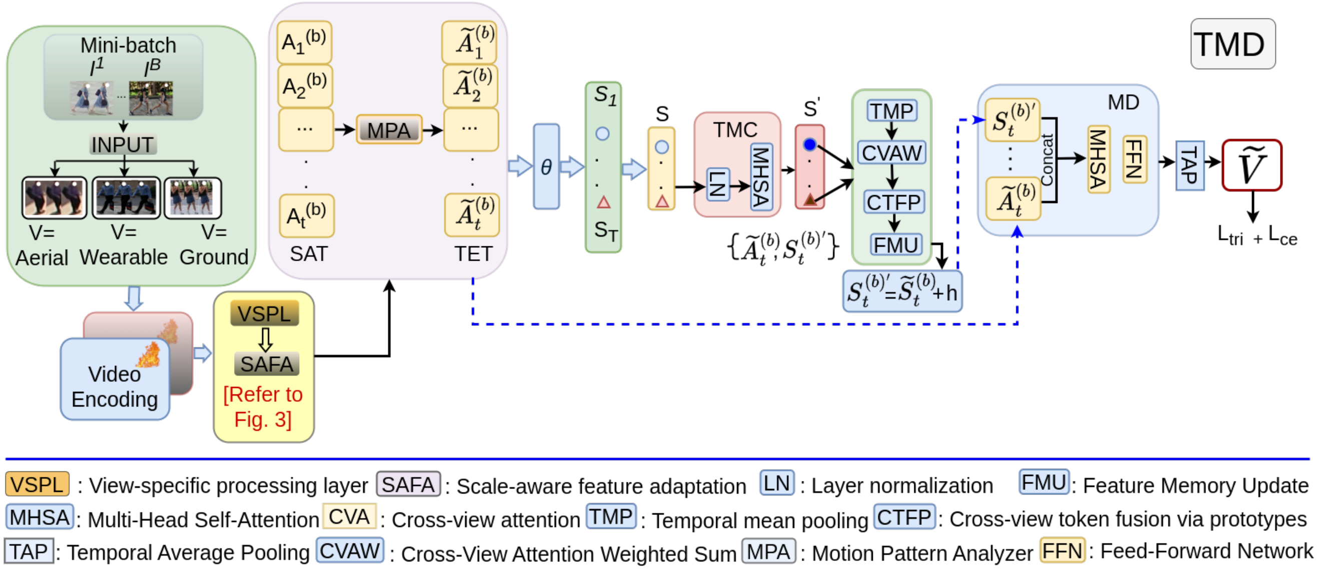

Mini-batch
I^1 I^B
INPUT
V= Aerial
V= Wearable
V= Ground
Video Encoding
VSPL
SAFA
[Refer to Fig. 3]
$A_1^{(b)}$
$A_2^{(b)}$
$A_t^{(b)}$
SAT
MPA
$\widetilde{A}_1^{(b)}$
$\widetilde{A}_2^{(b)}$
$\widetilde{A}_t^{(b)}$
TET
θ
S_1
S_T
S
TMC
LN
MHSA
$\{\widetilde{A}_t^{(b)}, S_t^{(b)\prime}\}$
S'
TMP
CVAW
CTFP
FMU
$S_t^{(b)\prime} = \widetilde{S}_t^{(b)} + h$
MD
$S_t^{(b)\prime}$
$\widetilde{A}_t^{(b)}$
Concat
MHSA
FFN
TAP
$\widetilde{V}$
$L_{tri} + L_{ce}$
TMD
VSPL : View-specific processing layer
SAFA : Scale-aware feature adaptation
LN : Layer normalization
FMU : Feature Memory Update
MHSA : Multi-Head Self-Attention
CVA : Cross-view attention
TMP : Temporal mean pooling
CTFP : Cross-view token fusion via prototypes
TAP : Temporal Average Pooling
CVAW : Cross-View Attention Weighted Sum
MPA : Motion Pattern Analyzer
FFN : Feed-Forward Network

Figure 4. Temporal–Memory Diffusion. Each frame is summarised into a memory token (*TMC*); these tokens exchange information across views (*CVIM*); finally, the fused memory is diffused back into each patch token before pooling.

3.6. Overall Training Setup

All components of VM–TAPS are trained end-to-end under the unified multi-term loss $\mathcal{L} = \mathcal{L}_{\text{V2M}} + 2.0\,\mathcal{L}_{\text{Triplet}} + \mathcal{L}_{\text{CE}} + 5 \times 10^{-4}\,\mathcal{L}_{\text{Center}}$, where each term denotes video-to-memory contrastive loss, cross-view triplet loss ($\delta = 0.3$), label-smoothed softmax ($\varepsilon = 0.1$) and centre loss, respectively. Training proceeds in two stages: first, VSPL, SAFA, MPA, TMD, VAMB and the projection head are optimised for 150 epochs with batch size 64 using AdamW (base LR 1×10^{-4}, weight decay 0.05) and a linear warm-up from 1×10^{-5} over 10 epochs, saving checkpoints every 10 epochs and validating every 5; second, the VAMB fusion gate and prompt-learner are unfrozen, the batch size reduced to 32, and training continues for 100 epochs with base LR 5×10^{-6}, weight decay 0.05, bias-LR factor 2, and a step decay of 0.1 at epochs 40, 70 and 90, with evaluation every 2 epochs and logging every 50 iterations. In both stages the ViT-B/16 backbone is fine-tuned, we select the checkpoint with highest validation Rank-1, apply k-reciprocal re-ranking on the 80m/120m splits, and report final CMC and mAP.

4. Experimental Setup

4.1. Dataset Description

All experiments use **AG–VPReID 2025** [12], a large-scale benchmark that pairs long-range UAV footage (80–120 m altitude) with time-synchronised CCTV and head-mounted cameras. The corpus comprises 13 507 video tracklets of 3 027 identities (1 693 distractors) totalling 3.7 M RGB frames; 15 soft-biometric attributes are provided but not exploited here. Evaluation follows the official protocol: *Case-1* treats aerial tracklets as queries and ground views as gallery, *Case-2* swaps the roles, and both cases are further split by altitude (80 m, 120 m) to gauge robustness. Faces are unresolvable and filenames are anonymised; the data are released for non-commercial research under a licence that forbids real-world re-identification.

4.2. Dataset Structure and Statistics

Table 1. Summary statistics of the AG–VPReID 2025 splits used in our experiments. “A2G” = Aerial→Ground (Case-1), “G2A” = Ground→Aerial (Case-2). Asterisks denote distractors.

Case	Subset	IDs	Tracklets	Frames (M)
Train	All	689	5 317	1.47
A2G	All	645	3 023	0.42
	80 m	356	1 523	0.26
	120 m	308	1 500	0.17
G2A	All	2 338	5 440	1.11
	80 m	1 162	2 797	0.69
	120 m	1 195	2 643	0.42

Table 1 reveals the pronounced class imbalance between the two evaluation directions. In the aerial→ground (A2G) scenario the query is a small, low-resolution drone tracklet that must locate its match among only ∼0.4 M ground frames, whereas in the ground→aerial (G2A) case each query must sift through a *five-thousand-identity* gallery that mixes genuine matches with 1 693 distractors. The training partition spans over 1.4 M frames, providing ample temporal variation for the motion–aware modules in Section 3, while the altitude-specific subsets allow fine-grained analysis of tolerance to extreme resolution changes.

4.3. Implementation Details

All experiments were conducted on a single NVIDIA A40 (PyTorch 1.11.0, CUDA 12.4). Tracklets were sampled to 8 frames, resized to 256×128, normalized (mean/std = 0.5) and augmented with random flips, erasing and padding. Training ran in two stages: Stage 1 optimised VSPL, SAFA, MPA, CVIM and the projection head for 150 epochs (batch 64: 8 IDs×8 tracklets) using AdamW (base LR 1e-4, warm-up from 1e-5 over 10 epochs; weight decay 0.05; drop-path 0.3; dropout 0.3; attention-drop 0.1), with checkpoints every 10 epochs and validation every 5; Stage 2 unfroze the fusion gate, reduced batch to 32 and trained 100 epochs (base LR 5e-6; bias-LR×2; LR×0.1 at epochs 40, 70, 90), evaluating every 2 epochs and logging every 50 iterations. The ViT-B/16 backbone was fine-tuned throughout. We used a unified loss combining video-to-memory contrastive, cross-view triplet (margin 0.3), label-smoothed cross-entropy (= 0.1) and centre loss (5e-4). At inference, 512-D descriptors were re-ranked (k=20, k=6, =0.3) and CMC/mAP reported on the 80m and 120m splits. All settings are controlled by a single config file for reproducibility.

4.4. Evaluation Protocol

At test time, all model parameters, including the view-aware memory bank—are frozen and each tracklet, are encoded into a single 512-dimensional descriptor. We follow the official AG–VPReID2025 protocol, which defines two retrieval scenarios. In Case-1 (Aerial→Ground), 3 023 aerial queries are matched against 2 750 ground gallery tracklets; in Case-2 (Ground→Aerial), 2750 ground queries are matched against 5440 aerial gallery tracklets augmented with 1 693 distractors. To gauge robustness to scale, both scenarios are also evaluated on altitude-specific subsets at 80m and 120m. Descriptors are L2-normalized and similarity is measured by Euclidean distance, followed by k-reciprocal re-ranking (k=20, k=6, =0.3). Retrieval performance is reported using cumulative matching characteristics (CMC) at ranks 1, 5 and 10, and mean average precision (mAP). To mitigate stochastic effects from drop-path and other nondeterminisms, all results are averaged over three independent inference runs.

5. Comparative Evaluation

The results in Tables 3 are drawn from the official AG-VPReID 2025 leaderboard. Since test-set identity labels remain private, participants submit a `submission.csv` with the top-10 gallery tracklets per query and receive only the overall mAP. Detailed Rank-1/5/10 and mAP breakdowns across A2G, G2A and altitude subsets are shown only for each team's best submission; lower-ranked entries report just the overall mAP. Consequently, unless our submission is our top entry, we cannot retrieve finer-grained scores. The tables therefore reflect the latest public leaderboard data at submission time, with full metrics available only for each team's highest-scoring run.

Table 2. Comparison between the performance of the baseline and our proposal (mAP, AG–VPReID2025 set).

Method	Overall mAP (%)
CLIP-ReID	52.31
VSLA-CLIP	52.20
TF-CLIP	66.26
VM-TAPS (Ours)	**69.34**

Table 2 shows that our VM-TAPS model achieves an overall mAP of 69.34%, significantly outperforming all prior approaches on AG-VPReID2025. In particular, VM-TAPS exceeds the strong TF-CLIP baseline 66.26% by 3.08 percentage points, and improves upon earlier CLIP-ReID 52.31% and VSLA-CLIP 52.20% by over 17pp. This large margin demonstrates that the combination of view-specific normalization, scale-aware fusion, motion analysis, memory banking, and cross-view interaction provides a substantial boost in retrieval accuracy over both vanilla CLIP adapters and the TF-CLIP architecture. **Overall test split.** Overall, our VM-TAPS model clearly outperforms the previous best method, TF-CLIP, by around 3 percentual points in Rank-1 accuracy and mean average precision (mAP), across all cameras and both retrieval directions. This means VM-TAPS correctly ranks the true identity at the top position three more times out of every hundred queries and improves the overall ranking quality by a similar margin.

80-meter subset. At the lower drone height of 80 meters, pedestrians resolution is still relatively high. Even under these favorable conditions, VM-TAPS achieves a gain of 3.4 percentage points in Rank-1 accuracy and 2.5 points in mAP over TF-CLIP. These improvements are mainly driven by the View-Specific Processing Layer (VSPL), which reduces color bias between views, and the Motion Pattern Analyzer (MPA), which captures consistent movement cues across time.

120-meter subset. At 120 meters, pedestrians resolution drops significantly and becomes a limiting factor for TF-CLIP. Despite this challenge, VM-TAPS maintains a lead of 4.3 percentage points in Rank-1 and 2.9 points in mAP. This is made possible by the Scale-Aware Feature Adaptation (SAFA) module, which enhances robustness to low resolution, and the View-Aware Memory Bank (VAMB), which provides better matching features under extreme top-down views.

Table 3. Performance comparison on **AG–VPReID 2025** across different settings. “A2G” = Aerial→Ground, “G2A” = Ground→Aerial.

Setting	**Method**	**A2G**				**G2A**				**Overall**			
		R1	R5	R10	mAP	R1	R5	R10	mAP	R1	R5	R10	mAP
Full	TF-CLIP	63.08	75.16	79.89	65.52	64.49	79.86	83.97	67.07	63.75	77.40	81.83	66.26
	VM-TAPS (ours)	**65.83**	**76.45**	**81.34**	**68.08**	**69.79**	**81.64**	**85.39**	**70.73**	**67.72**	**78.92**	**83.27**	**69.34**
80-m	TF-CLIP	72.79	83.07	84.29	75.30	66.98	81.11	83.40	70.57	70.30	82.23	83.91	73.27
	VM-TAPS (ours)	**73.55**	**83.56**	**85.06**	**75.98**	**73.85**	**84.26**	**85.59**	**75.40**	**73.68**	**83.86**	**85.29**	**75.73**
120-m	TF-CLIP	61.71	74.69	76.29	65.48	67.97	82.75	83.59	71.26	65.13	79.10	80.28	68.64
	VM-TAPS (ours)	**66.01**	**77.07**	**78.31**	**69.05**	**72.30**	**83.34**	**84.49**	**73.77**	**69.45**	**80.50**	**81.69**	**71.63**

Table 3 presents detailed Rank-1 and mAP results for TF-CLIP versus VM-TAPS across all retrieval scenarios and altitude settings.

6. Ablation Study

Table 4. Ablation study on the AG–VPReID 2025 test set. Each row enables a specific subset of VM-TAPS components. The full model (bottom row) yields the highest overall Rank-1 accuracy.

Configuration	**mAP**
CLIP + VSPL (View Normalization)	63.42
CLIP + SAFA (Scale Adaptation)	62.87
CLIP + MPA (Motion Analysis)	64.01
CLIP + VAMB (Memory Bank)	**67.23**
CLIP + CVIM (Cross-View Interact.)	63.95
CLIP + SAFA + MPA	65.32
CLIP + VAMB + CVIM	65.71
CLIP + VSPL + SAFA + MPA + VAMB + CVIM (Full VM-TAPS)	**69.34**

Table 4 evaluates the incremental contribution of each VM-TAPS component by selectively enabling subsets over the CLIP baseline. Among the individual modules, the View-Aware Memory Bank (VAMB) yields the most significant mAP improvement, indicating that temporal dynamics and long-range identity context are critical in cross-view scenarios. Notably, the combination of all modules the full VM-TAPS configuration achieves the highest accuracy (69.34%), validating our design as a cohesive system in which each component contributes orthogonally to robust cross-view video Re-ID.

7. Discussion

VM–TAPS tackles aerial–ground Re-ID challenges on camera bias, scale variation and limited temporal context by integrating five lightweight adapters into a frozen ViT: VSPL for colour and perspective normalization; SAFA for multi-resolution fusion; VAMB for view-aware identity prototypes (largest boost in Table 4); MPA for motion cues; and CVIM for cross-view alignment. These modules add fewer than 2 M parameters and under 3% FLOPs, yet increase mAP from 66.26% to 69.34% on AG-VPReID2025 and outperform any partial configuration, confirming that disentangling viewpoint, scale, motion and memory yields robust drone-to-ground retrieval.‘

Limitations: At this point, VM–TAPS still relies on explicit camera–type metadata (aerial, ground, wearable) to route features through the correct view-specific modules; if this information is missing or noisy, performance may degrade. Then, the View-Aware Memory Bank scales linearly with the number of identities and views, which could become a VRAM bottleneck for very large watch-lists unless future work introduces an efficient compression or pruning strategy.

8. Conclusions and Further Work

Aiming at robust aerial–ground ReId in real-world conditions, this paper introduced **VM–TAPS**, a compact suite of five complementary modules that retrofit TF-CLIP for the extreme viewpoint, scale and motion diversity scenarios. By addressing view bias, scale disparity and temporal ambiguity in a strictly modular fashion, VM–TAPS achieves state-of-the-art accuracy on AG–VPReID2025, while adding less than two million parameters and preserving real-time throughput on a single professional GPU. We believe the framework’s small footprint, simplicity of integration and publicly released codebase will make it a practical baseline for both academic study and industrial deployment.

Future work could eliminate hand-coded view labels via self-supervised view discovery—enabling truly ad-hoc sensor networks—and scale the memory bank with vector-quantised prototypes, product-quantised keys or adaptive pruning to support very large watch-lists. Integrating CLIP’s language branch with soft-biometric attributes (age, clothing colour, carried objects) can further disambiguate similar subjects and yield human-interpretable results, while distilling VM–TAPS onto slimmer backbones or via neural architecture search will enable real-time, cross-view Re-ID on power-constrained edge devices (drones and body-cams etc).

References

[1] A. A. and B. T. Vid-Trans-ReID: Enhanced Video Transformers for Person Re-identification. *British Machine Vision Conference*, 2022.

[2] J. Chen, Y. Wang, and Y. Y. Tang. Person Re-identification by Exploiting Spatio-Temporal Cues and Multi-view Metric Learning. *IEEE Signal Processing Letters*, 23(7):998–1002, 7 2016.

[3] C. Eom, G. Lee, J. Lee, and B. Ham. Video-based Person Re-identification with Spatial and Temporal Memory Networks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12016–12025. IEEE, 10 2021.

[4] Y. Feng, F. Chen, J. Yu, Y. Ji, F. Wu, T. Liu, S. Liu, X.-Y. Jing, and J. Luo. Cross-Modality Spatial-Temporal Transformer for Video-Based Visible-Infrared Person Re-Identification. *IEEE Transactions on Multimedia*, 26:6582–6594, 2024.

[5] M. Kim, M. Cho, and S. Lee. Feature Disentanglement Learning with Switching and Aggregation for Video-based Person Re-Identification. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1603–1612. IEEE, 1 2023.

[6] Y. Li, P. Gao, Y. Wu, and H. Li. Clip-memory: Exemplar memory for few-shot learning with clip. *arXiv preprint arXiv:2209.12323*, 2022.

[7] X. Lin, J. Li, Z. Ma, H. Li, S. Li, K. Xu, G. Lu, and D. Zhang. Learning Modal-Invariant and Temporal-Memory for Video-based Visible-Infrared Person Re-Identification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20941–20950. IEEE, 6 2022.

[8] Y. Lin, L. He, Y. Yang, Q. Tian, L. Shao, Y. da Song, and H.-Y. Zhao. Camera style adaptation for person re-identification. In *European Conference on Computer Vision (ECCV)*, 2020.

[9] J. Liu, Z.-J. Zha, W. Wu, K. Zheng, and Q. Sun. Spatial-Temporal Correlation and Topology Learning for Person Re-Identification in Videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4368–4377. IEEE, 6 2021.

[10] 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*, 35(10):13753–13763, 10 2024.

[11] 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*, 25(9):12818–12828, 9 2024.

[12] H. Nguyen, K. Nguyen, A. Pemasiri, F. Liu, S. Sridharan, and C. Fookes. Ag-vpreid: A challenging large-scale benchmark for aerial-ground video-based person re-identification. *arXiv preprint arXiv:2503.08121*, 2025.

[13] T. N. H. Nguyen, K. Nguyen, and .... Ag–vpreid: A challenging large-scale benchmark for aerial-ground video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[14] V. D. Nguyen, S. Mirza, P. Mantini, and S. K. Shah. Attention-based Shape and Gait Representations Learning for Video-based Cloth-Changing Person Re-Identification. *VISIGRAPP : VISAPP*, 2024.

[15] J. Ning, F. Li, R. Liu, S. Takeuchi, and G. Suzuki. *Temporal Extension Topology Learning for Video-Based Person Re-identification*, pages 213–225. Springer Nature Switzerland, 2023.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. arXiv:2103.00020.

[17] Z. Tianyu, W. Longhui, X. Lingxi, Z. Zijie, Z. Yongfei, L. Bo, and T. Qi. Spatiotemporal Transformer for Video-based Person Re-identification. *arXiv.org*, 2021.

[18] T. Wang, S. Gong, X. Zhu, and S. Wang. *Person Re-identification by Video Ranking*, pages 688–703. Springer International Publishing, 2014.

[19] T. Wang, Y. Wang, Y. Tai, C. Wang, J. Yang, J. Li, and F. Huang. Crossnorm: Normalization beyond domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[20] L. Wu, Y. Wang, J. Gao, and X. Li. Where-and-When to Look: Deep Siamese Attention Networks for Video-Based Person Re-Identification. *IEEE Transactions on Multimedia*, 21(6):1412–1424, 6 2019.

[21] L. Wu, Y. Wang, H. Yin, M. Wang, and L. Shao. Few-Shot Deep Adversarial Learning for Video-Based Person Re-Identification. *IEEE Transactions on Image Processing*, 29:1233–1245, 2020.

[22] W. Wu, J. Liu, K. Zheng, Q. Sun, and Z. Zha. Temporal Complementarity-Guided Reinforcement Learning for Image-to-Video Person Re-Identification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7309–7318. IEEE, 6 2022.

[23] C. Yu, X. Liu, Y. Wang, P. Zhang, and H. Lu. Tf-CLIP: Learning Text-free CLIP for Video-based Person Re-Identification. *AAAI Conference on Artificial Intelligence*, 2023.

[24] D. Zhang, W. Wu, H. Cheng, R. Zhang, Z. Dong, and Z. Cai. Image-to-Video Person Re-Identification With Temporally Memorized Similarity Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2622–2632, 10 2018.

[25] 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. *European Conference on Computer Vision*, 2024.

[26] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3754–3762, 2017.