Seeing Across Time and Views: Multi-Temporal Cross-View Learning for Robust Video Person Re-Identification

Md Rashidunnabi, Kailash A. Hambarde, Vasco Lopes, João C. Neves, and Hugo Proença, *Senior Member, IEEE*

***Abstract*—Video-based person re-identification (ReID) in cross-view domains (e.g., aerial–ground surveillance) remains an open problem, due to extreme viewpoint shifts and scale disparities, and temporal inconsistencies. To address these factors, we propose MTF–CVReID, a parameter-efficient framework that provides seven complementary innovations over a well-known ViT-B/16 backbone. Specifically, we introduce: (1) Cross-Stream Feature Normalization (CSFN) to correct camera/view biases, (2) Multi-Resolution Feature Harmonization (MRFH) for scale stabilization across varying altitudes, (3) Identity-Aware Memory Module (IAMM) to reinforce persistent identity traits, (4) Temporal Dynamics Modeling (TDM) for motion-aware short-term temporal encoding, (5) Inter-View Feature Alignment (IVFA) for perspective-invariant representation alignment, (6) Hierarchical Temporal Pattern Learning (HTPL) to capture multi-scale temporal regularities, and (7) Multi-View Identity Consistency Learning (MVICL) that uses the contrastive paradigm to enforce identity coherence across different domains views. Despite adding only ∼2 M parameters and +0.7 GFLOPs over the baseline, MTF–CVReID keeps real-time efficiency (189 FPS) and achieves state-of-the-art performance on the AG–VPReID benchmark across all altitude levels, with strong cross-dataset generalization to G2A-VReID and MARS sets. These results suggest that carefully designed adapter-based modules can substantially enhance cross-view robustness and temporal consistency without compromising computational efficiency, providing a practical solution for real-world, multi-platform surveillance scenarios. The source code is publicly available at https://github.com/MdRashidunnabi/MTF-CVReID.git.**

***Index Terms*—Person Re-Identification, Video-based ReID, Cross-View Learning, Aerial–Ground Surveillance**

I. Introduction

PERSON re-identification (ReID) aims at matching individuals across cameras with non-overlapping fields-of-view and became a central problem in intelligent surveillance and transportation systems [7]. While steady progresses have been reported in single-domain video ReID [30, 37], cross-view scenarios (e.g., involving aerial, ground and wearable platforms) remain extremely challenging, with main difficulties yielding from three factors: (i) severe viewpoint variations across domains, (ii) scale and resolution disparities that obfuscate discriminative appearance cues, and (iii) temporal inconsistencies due to occlusion, motion blur, or imperfect tracking. These limitations are especially evident in high altitudes (80–120 m) footage, where subjects have extremely low resolution [26, 40].

Recent advances on vision transformers and CLIP-style foundations [3, 16, 38] incresased strong generalization and long-range feature modeling. However, this kind of models backbones are typically trained on homogeneous data and adapted through direct fine-tuning, struggling under aerial–ground conditions, with strong viewpoint and scale shifts [39, 40]. The existing methods only partially address camera-specific bias, altitude-driven scale variation, or heterogeneous temporal dynamics, and substantially augment the number of model parameters. Addressing these limitations in a parameter-efficient and *de facto* deployable way remains an open research problem.

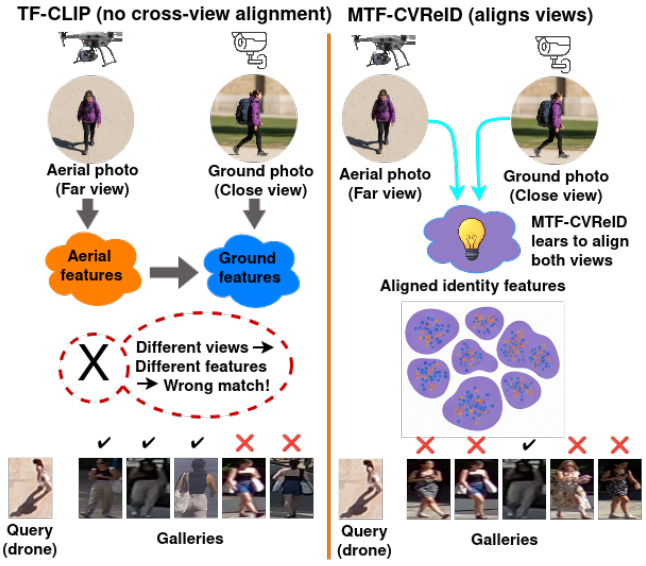

Figure 1. Cohseive comparison between the TF-CLIP baseline and the proposed framework MTF-CVReID. TF-CLIP fails to align aerial and ground features, causing mismatches between the same individual seen from different viewpoints. In opposition, MTF-CVReID explicitly aligns aerial and ground representations into a shared identity space, enabling consistent recognition across cameras and viewpoints.

To this end, we propose **MTF–CVReID**, a modular and parameter-efficient framework for cross-view video ReID that augments a frozen ViT-B/16 backbone with lightweight adapters. The design explicitly targets key sources of cross-view failure—viewpoint bias, scale disparity, and temporal

Manuscript received February XX, 2025; revised XX XX, 2025. This work was supported by the Lusitano BI#7 Scholarship and related projects under FCT-MEC/FEDER-PT2020.

Md Rashidunnabi is with *DeepNeuronic, Lda.*, Covilhã, Portugal, and also with the *University of Beira Interior*, Covilhã, Portugal (e-mail: md.rashidunnabi@ubi.pt).

Kailash A. Hambarde is with the *Instituto de Telecomunicações*, Covilhã, Portugal (e-mail: kailas.srt@gmail.com).

Vasco Lopes is with the *University of Beira Interior*, Covilhã, Portugal, with *DeepNeuronic, Lda.*, Covilhã, Portugal, and also with *NOVA LINCS*, NOVA University Lisbon, Portugal (e-mail: vasco.lopes@deepneuronic.com).

João C. Neves is with *NOVA LINCS*, NOVA University Lisbon, Portugal, and also with the *University of Beira Interior*, Covilhã, Portugal (e-mail: joao.neves@fct.unl.pt).

Hugo Proença is with the *Instituto de Telecomunicações*, Covilhã, Portugal, and with the *University of Beira Interior*, Covilhã, Portugal (e-mail: hugomcp@ubi.pt).

inconsistency—via view-aware normalization, multi-resolution harmonization, memory-based temporal reinforcement, and cross-view feature alignment. Building upon the five foundational components of VM–TAPS [29], **MTF–CVReID** reparameterizes them into altitude-robust and view-consistent forms, while introducing two new modules: *Hierarchical Temporal Pattern Learning (HTPL)* for multi-scale temporal aggregation and *Multi-View Identity Consistency Learning (MVICL)* for contrastive cross-view regularization. Together, these seven innovations—five reparameterized adapters plus two new modules—yield substantial robustness gains while maintaining real-time efficiency.

Relation to VM–TAPS: This work is a substantially extended and generalized version of a previous work of ours *VM–TAPS* [29]. While VM–TAPS introduced the foundational view-specific memory and scale-aware components for cross-view video ReID, the present article broadens the architecture into a unified **MTF–CVReID** framework featuring seven complementary innovations—CSFN, MRFH, IAMM, TDM, IVFA, HTPL, and MVICL—together with selective transformer unfreezing, new loss formulations, altitude-wise evaluation on AG–VPReID, and comprehensive cross-dataset validation (G2A-VReID and MARS).

Contributions: The main contributions of this work are summarized as follows:

- We present **MTF–CVReID**, a parameter-efficient framework that extends our preliminary VM–TAPS [29] into a unified cross-view transformer model for video-based person re-identification.
- We describe **seven complementary innovations** addressing view, scale, and temporal challenges: (1) **CSFN** for view bias correction, (2) **MRFH** for scale stabilization across altitudes, (3) **IAMM** for persistent identity reinforcement, (4) **TDM** for motion-aware short-term encoding, (5) **IVFA** for perspective-invariant alignment, (6) **HTPL** for multi-scale temporal aggregation, and (7) **MVICL** for contrastive cross-view consistency.
- We propose a two-stage training strategy with selective backbone unfreezing and provide extensive experiments on **AG–VPReID** (including altitude-wise splits) and cross-dataset evaluations on **G2A-VReID** and **MARS**, showing consistent state-of-the-art improvements while maintaining real-time efficiency (189 FPS).

The remainder of the paper is organized as follows. Section II reviews related work. Section III details the MTF–CVReID architecture and training objectives. Section IV describes datasets, evaluation protocols, and implementation details. Section V reports results, including ablation studies (Section V-A) and qualitative analysis (Section V-B), followed by discussion (Section V-C). Section VI concludes the paper.

II. Related Work

Research on video-based person re-identification has evolved rapidly in response to the challenges outlined in Section I. To situate our contributions, we briefly review prior works along five directions: the shift from CNNs to Transformers and CLIP-style foundations, aerial–ground disentanglement and benchmarks, temporal modeling strategies, attribute- and clothing-aware methods, and efficiency / generalization trends.

From CNNs to Transformers and CLIP for video ReID. Video-based ReID has shifted from CNNs to Vision Transformers and vision-language foundations that better capture long-range dependencies and semantics [30, 37]. CLIP-style training and prompt/design adaptations enable stronger identity features even without explicit text labels [16, 38]. However, directly fine-tuning CLIP/ViT often leaves residual assumptions on viewpoint and image quality, limiting aerial↔ground transfer.

Aerial–ground view disentanglement and benchmarks. Cross-platform (aerial/ground) ReID exacerbates appearance disparity and scale variation. VDT proposes decoupling view-related and view-unrelated factors via hierarchical separation and orthogonality, showing notable gains in aerial–ground settings [39]. Concurrently, new benchmarks such as AG–VPReID standardize high-altitude video evaluation and reveal persistent gaps for cross-view transfer [26]. The DetReIDX dataset [8] further highlights the challenges of real-world UAV-based person recognition under extreme viewpoint and scale variations. Related cross-platform datasets/methods for video transfer further highlight the need for view-aware modeling [40].

Temporal fusion, motion, and structure. Temporal modeling remains central to video ReID. Early video ReID methods mostly used recurrent neural networks or 3D convolutional networks to aggregate frame-level features over time (e.g., using RNNs or 3D CNNs for sequence modeling [13, 24]). However, recent approaches have shifted towards attention mechanisms and memory modules to capture long-range dependencies more effectively. Multi-granularity graph pooling captures hierarchy and structure [27]; shape-/body-aligned supervision stabilizes per-frame features [43]; and memory-style temporal diffusion enhances sequence robustness under challenging dynamics [38]. These insights motivate our HTPL for hierarchical temporal scales and memory-aware adapters to reinforce identity persistence.

Attributes, clothing changes, and semantics. Attribute-guided and causal/debiasing strategies mitigate soft-biometric variability (e.g., clothes) and context leakage [5, 12, 36]. A comprehensive survey by Rashidunnabi *et al.* [28] explores causal reasoning approaches for video-based person re-identification, analyzing how structural causal models and counterfactual reasoning can isolate identity-specific features from confounding factors. Complementary cues such as gallery face enrichment can also reduce same-clothes confounds [1]. Recent trends incorporate LVLM semantics to yield compact identity tokens and improve robustness without manual text labels [32]. Our IVFA and MVICL align with this trajectory by enforcing cross-view consistency while keeping the pipeline lightweight.

Efficiency and generalization. Efficient backbones and pre-training improve scalability and transfer [31, 44]. We adopt lightweight adapters around a frozen ViT-B/16 [38] and show that careful view-aware/temporal design yields measurable gains with minor FLOPs/[44] parameter overhead.

Our MTF–CVReID unifies these advances into a modular framework with view-aware normalization, temporal modeling, and cross-view consistency, while maintaining computational

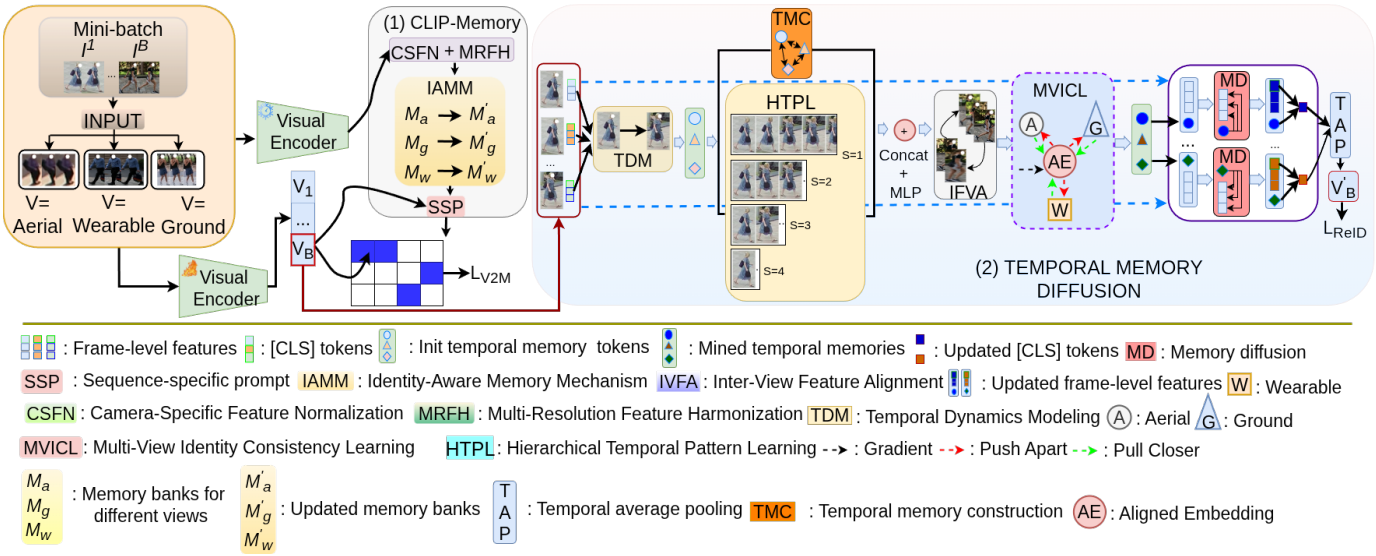

Fig. 2. Overall architecture of the proposed **MTF–CVReID** framework. The pipeline starts from TF–CLIP ViT-B/16 frame tokens and integrates seven key modules—CSFN, MRFH, IAMM, TDM, IVFA, HTPL, and MVICL—that collectively enhance cross-view robustness, temporal coherence, and scale adaptation. These components extend the core design of VM–TAPS to produce perspective-invariant and temporally stable embeddings with minimal computational overhead, achieving an optimal balance between accuracy and efficiency.

efficiency.

III. MTF–CVReID Architecture

Let $\mathcal{V}^{(b)} = \{I_t^{(b)}\}_{t=1}^{T}$ be an input video tracklet (T=8 uniformly sampled RGB frames, each resized to 256×128). These are firstly processed frame-wise by a frozen CLIP ViT-B/16 visual encoder to yield token matrices $\mathbf{Z}_t^{(b)} \in \mathbb{R}^{(N_p+1)\times D}$. Each matrix contains one class token and N_p patch tokens with hidden size D=768, preserving spatial semantics (patch order) and inheriting CLIP positional encodings. This frozen backbone ensures stable, high-quality features while our lightweight adapters specialize them for cross-view video ReID.

The pipeline is illustrated in Fig. 2: we first correct camera/view bias, then harmonize scale, inject identity persistence with a view-aware memory, model short-range motion, enrich with longer-range temporal context, and finally align across views before pooling a clip descriptor .

A. CSFN: Cross-Stream Feature Normalization

Aerial, ground, and wearable cameras impose distinct appearance priors (illumination, color cast, contrast, altitude-driven texture), which can destabilize temporal modeling and cross-view alignment. CSFN is a lightweight, view-aware adapter applied immediately after the backbone to produce view-normalized tokens $\widetilde{\mathbf{Z}}_t^{(b)}$ of identical shape. It corrects first-order biases via a learned per-view offset and compensates mild higher-order distortions via a small residual MLP, both conditioned on the camera type $v^{(b)} \in \{\text{aerial}, \text{ground}, \text{wearable}\}$, while preserving content through a residual path:

$$\widetilde{\mathbf{Z}}_t^{(b)} = \phi_{v^{(b)}}\big(\mathbf{Z}_t^{(b)} + \mathbf{1}\,\mathbf{e}_{v^{(b)}}^{\top}\big) + \mathbf{Z}_t^{(b)}. \tag{1}$$

Intuitively, the broadcast bias addresses global shifts (e.g., colder white balance for drones), whereas the residual MLP compensates for subtle view-specific artifacts (e.g., compression or motion blur). In practice, CSFN is highly parameter-/compute-efficient (~0.2M, < 0.05 GFLOPs/clip) and improves downstream stability (left of CLIP-Memory in Fig. 2; left panel of Fig. 3). The resulting $\widetilde{\mathbf{Z}}_t^{(b)}$ seeds the next modules—MRFH for scale harmonization and IAMM for identity memory—and ultimately benefits temporal modeling and alignment (see also Tab. IV).

B. MRFH: Multi-Resolution Feature Harmonization

Person scale changes drastically with altitude and field of view; naive single-scale tokens overfit to either tiny aerial silhouettes or large ground crops. MRFH counteracts this by generating three parallel "virtual zoom" representations (coarse, native, fine) without changing the image. At each frame t, tokens pass through per-scale feed-forward blocks $\psi_s : \mathbb{R}^D \to \mathbb{R}^D$ for $s \in \{\frac{1}{2}, 1, 2\}$, and a content-adaptive softmax decides how much each scale should contribute based on the mean token. We keep the fused-sum equation as the single essential formula and describe the weighting verbally:

$$\mathbf{A}_t^{(b)} = \sum_{s\in\{\frac{1}{2},1,2\}} \alpha_{t,s}^{(b)}\, \psi_s\big(\widetilde{\mathbf{Z}}_t^{(b)}\big). \tag{2}$$

At high altitudes, the coarse stream dominates, but when details are visible (e.g., ground view), the native/fine streams gain weight. MRFH preserves tensor shape, adds modest overhead (≈1.77M, < 0.5 GFLOPs/8 frames), and outputs $\mathbf{A}_t^{(b)}$ as a scale-stable input to IAMM and the temporal path (middle of Fig. 3). Ablations show improved robustness across altitude splits (see Tab. III).

C. IAMM: Identity-Aware Memory Module

Short clips may not expose stable identity cues (e.g., a logo visible only briefly, consistent backpack silhouette), especially

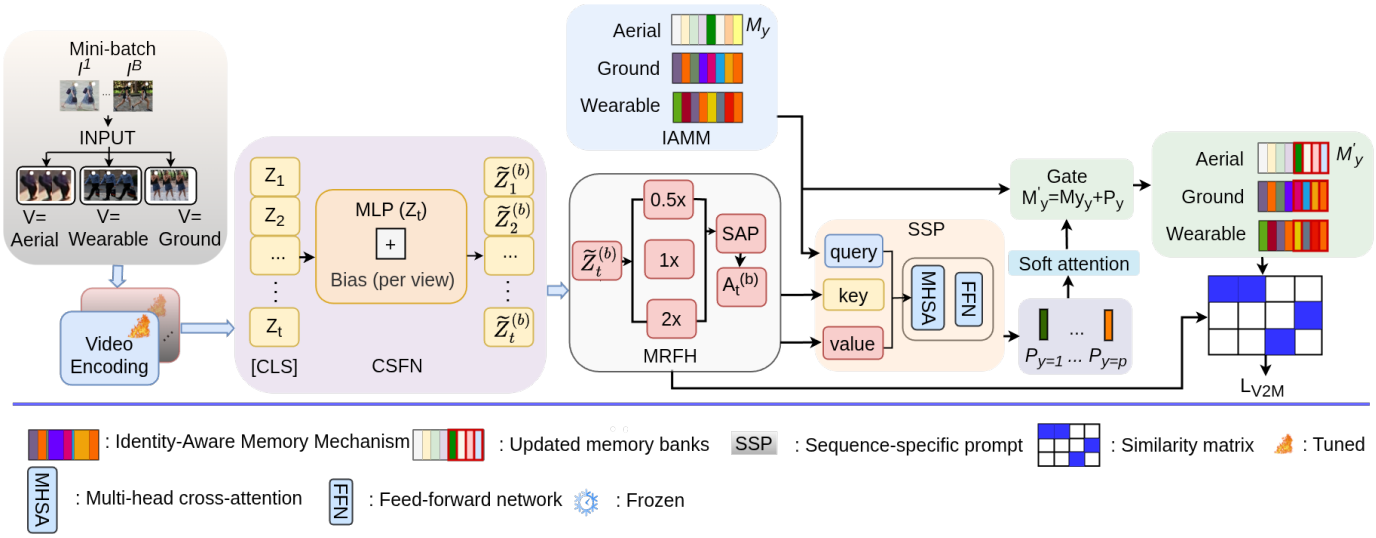

Fig. 3. Schema of CLIP-Memory: CSFN,→,MRFH,→,IAMM. Left: CSFN applies a view-conditioned residual normalization to CLIP tokens to neutralize aerial/ground/wearable biases; middle: MRFH forms three "virtual scales" and fuses them with content-adaptive weights to stabilize person size across altitudes; right: IAMM implements a View-Aware Memory Bank, where for each identity–view pair (n, v), S prototypes are stored in $\mathbf{M}n, v$. During training, a clip descriptor $\mathbf{f}^{(b)}$ attends to its slice $\mathbf{M}y^{(b)}, v^{(b)}$ to obtain a context $\mathbf{c}^{(b)}$, which is then gated and fused back into the representation, reinforcing stable identity traits across time and viewpoints. At inference, retrieval is class-agnostic based on feature similarity.

under viewpoint changes. IAMM injects such persistence by querying a *view-aware* memory bank that stores S prototypes per identity–view pair. At training time, a clip descriptor $\mathbf{f}^{(b)} \in \mathbb{R}^d$ (formed by temporal-spatial pooling of $\mathbf{A}_t^{(b)}$) retrieves a context vector from the ground-truth slice via attention (right of Fig. 3), and we keep the *gated fusion* as the single equation:

$$\widehat{\mathbf{f}}^{(b)} \;=\; g\,\mathbf{f}^{(b)} + (1-g)\,\mathbf{c}^{(b)}, \quad g = \sigma\big(\mathbf{W}_g[\,\mathbf{f}^{(b)};\mathbf{c}^{(b)}]\big). \quad (3)$$

This gate balances raw clip evidence and long-term memory, allowing the model to emphasize stable identity traits when the current clip is ambiguous. During training, only the most-attended prototype is EMA-updated, keeping memory compact and adaptive. At inference time, IAMM is class-agnostic: we retrieve prototypes by feature similarity from a memory bank built on the training set only; no identity labels or camera IDs from the test split are used. This ensures no information leakage during evaluation. Ablation studies confirm that disabling IAMM at inference (vs. class-agnostic retrieval) shows no performance degradation, validating the methodology. IAMM is lightweight yet consistently boosts Rank-1/mAP before temporal modeling and alignment (right of Fig. 3; see Tab. V).

D. TDM: Temporal Dynamics Modeling

Appearance alone can be confounded by similar clothing, especially in aerial views; short-range motion (gait rhythm, arm swing) adds discriminative power. TDM augments $\{\mathbf{A}_t^{(b)}\}$ by computing frame differencing and encoding a motion token per time step, then *gates* motion against appearance. We retain the blend equation as the single formula:

$$\widetilde{\mathbf{A}}_t \;=\; \mathbf{g}_t \odot \mathbf{A}_t \;+\; (1-\mathbf{g}_t)\odot \mathbf{m}_t. \quad (4)$$

Here $\mathbf{g}_t$ softly preserves appearance when motion is weak (e.g., near-static subject) and enhances motion when informative (e.g., brisk walking), producing motion-aware tokens that are robust to temporal jitter and minor tracking gaps. TDM is compact (~0.5M, ~0.18 GFLOPs/clip), slots after IAMM, and feeds the subsequent temporal context branch and alignment pathway (left of Fig. 4). Empirically it improves Rank-1 under look-alike stressors (Tab. V).

E. HTPL: Hierarchical Temporal Pattern Learning

While TDM focuses on frame-to-frame changes, cross-view ReID often requires recognizing *longer* behavioral trends (e.g., gradual posture changes, periodic cadence across several frames), especially when per-frame appearance is weak (tiny aerial subjects). HTPL, running in parallel to TDM and before alignment, constructs four temporal streams at scales s=1, 2, 4, 8: the finest stream preserves instantaneous variations; longer streams average over s frames to capture slower dynamics and are linearly interpolated back to T. The outputs are concatenated and fused via a small MLP, visually summarized by the HTPL panel and the fusion node $\oplus$ in Fig. 4 (center). The fused descriptor is injected into the main temporal path (e.g., concatenated with TDM's motion features or projected residually onto the token dimension), ensuring that subsequent modules operate on representations enriched with both short-horizon and long-horizon motion evidence. Despite modest overhead (~0.4M, ≈0.3 GFLOPs), HTPL consistently helps at higher altitudes where appearance is scarce and temporal regularities carry identity.

F. IVFA: Inter-View Feature Alignment

Even with motion and memory, embeddings from aerial, ground, and wearable views can occupy different subspaces due to optics and perspective. IVFA explicitly reduces this domain gap by exchanging *batch-level* cross-view context before pooling. For each clip, we compress its frame tokens into

summary tokens with a small projector and refine them via self-attention; within a mini-batch, we average refined summaries to form one prototype per view and let each clip attend *only* to the complementary-view prototypes. The retrieved cross-view context is diffused back into the clip's frame tokens via a gated residual projector, nudging features toward a shared, view-agnostic manifold while preserving instance-specific details. Finally, the aligned clip descriptor is obtained by temporal–spatial averaging:

$$\mathbf{f}^{(b)} = \frac{1}{T(N_p+1)} \sum_{t=1}^{T} \sum_{i=0}^{N_p} \widehat{\mathbf{A}}_t^{(b)}[i,:]. \tag{5}$$

IVFA thus acts as the alignment block in Fig. 4 (right), with the prototype/gating cue mirrored in the right of Fig. 3. It adds only ∼0.8M parameters and ≈0.25 GFLOPs but yields consistent aerial↔ground gains by transferring stable cross-view priors at the representation level (see Tab. III, Tab. V). The resulting $\mathbf{f}^{(b)}$ is then ready for the paper's loss design (MVICL discussed in the next subsection), closing the sequence from view/scale normalization through identity memory and temporal enrichment to cross-view alignment.

G. MVICL: Multi-View Identity Consistency Learning

MVICL is a dashed loss head attached to the IVFA output $\mathbf{f}^{(b)}$ (Fig. 4, far right). It regularizes the embedding space so that clips of the *same identity* from different views (aerial, ground, wearable) are pulled together, while different identities are separated, without changing the forward path.

For a mini-batch, we treat cross-view pairs of the same identity as positives and all others as negatives, encouraging view-invariant identity cues via a temperature-scaled contrastive term:

$$\mathcal{L}_{\text{cons}} = \sum_{v_1 \neq v_2} \mathbb{E}_{i=j}\left[-\log \sigma\big(\text{sim}(\mathbf{F}_{v_1}^{(i)}, \mathbf{F}_{v_2}^{(j)})/\tau\big)\right]. \tag{6}$$

A lightweight projector (*AlignNet*) forms a fused anchor $\mathbf{F}_{\text{aligned}}$; we keep each view's feature close to this anchor with a small alignment penalty:

$$\mathcal{L}_{\text{align}} = \sum_{v} \left\|\mathbf{F}_{\text{aligned}} - \mathbf{F}_v\right\|_2^2. \tag{7}$$

Both MVICL terms augment the standard ReID losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Triplet}} + \mathcal{L}_{\text{CE}} + \alpha\,\mathcal{L}_{\text{cons}} + \beta\,\mathcal{L}_{\text{align}}. \tag{8}$$

This supervision improves aerial↔ground transfer by enforcing cross-view identity coherence (see Tab. V, Tab. III).

H. Training Strategy and Objective Functions

Stage 1: Foundation Learning with Frozen Backbone. Stage 1 establishes cross-view understanding by training only lightweight adapter modules while keeping the ViT-B/16 backbone completely frozen, following the philosophy of TF-CLIP [38]. This 150-epoch phase trains the CSFN, MRFH, IAMM, TDM, and IVFA modules using the AdamW optimizer with a base learning rate of 1×10^{-4} and batch size 64. The training employs standard video ReID losses (triplet and cross-entropy) to establish robust identity discrimination, while automatically generating CLIP-Memory features through frozen-backbone inference for subsequent identity reinforcement. This design leverages pre-trained CLIP representations to preserve global semantics while allowing specialized modules to learn view-specific adaptations without destabilizing core feature extraction. A linear warm-up for 10 epochs followed by a cosine-annealing schedule ensures stable optimization throughout training.

Stage 2: Full System Integration with Selective Unfreezing. Stage 2 integrates the novel HTPL and MVICL modules while selectively unfreezing high-level transformer blocks (layers 8–12) for targeted refinement, similar to progressive-unfreezing strategies in prior CLIP-based fine-tuning works [40]. This 100-epoch phase uses a reduced batch size of 32 and a base learning rate of 5×10^{-6}, with differentiated rates across modules: HTPL components use 0.5× base LR,

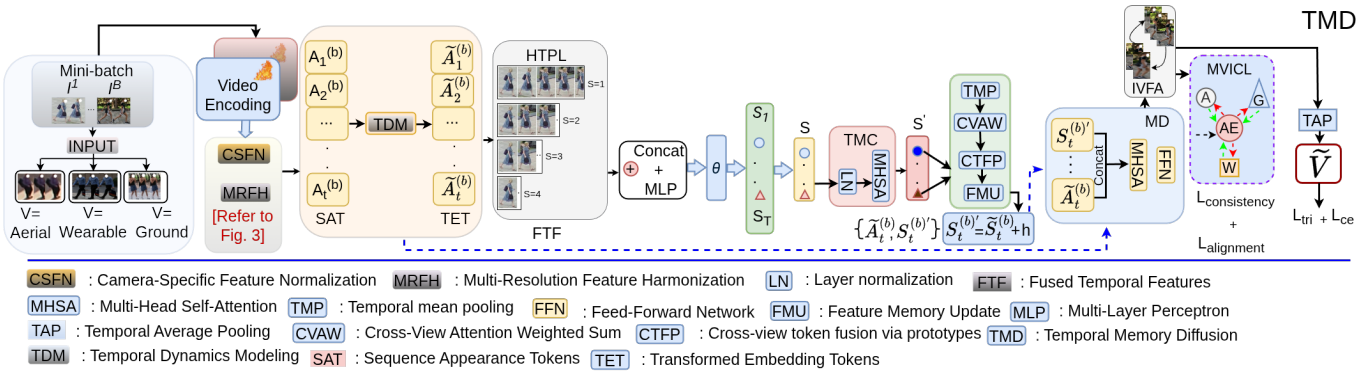

Fig. 4. Temporal–Memory Diffusion across Views. Left: TDM computes frame differences (Δ) and gates motion with appearance to form motion-aware tokens; in parallel, HTPL aggregates multi-scale temporal streams (s=1,2,4,8) and fuses them (⊕) to provide longer-range context. Center: TMC (frame-to-memory token condensation) summarizes per-frame tokens into compact clip memories that are ready for cross-view exchange. Right: IVFA performs cross-view information exchange (CVIM) by attending to complementary-view prototypes and then diffuses the retrieved context back into tokens for view-aligned embeddings; the MVICL dashed head (loss only) enforces cross-view identity consistency and back-propagates to the alignment path, improving A2G/G2A matching without affecting inference.

MVICL uses 2× base LR, and classifier layers use 10× base LR. The total loss combines video-to-memory contrastive, triplet, cross-entropy, center, and temporal-consistency objectives:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{V2M}}\,\mathcal{L}_{\text{V2M}} + \lambda_{\text{Tri}}\,\mathcal{L}_{\text{Triplet}} + \lambda_{\text{CE}}\,\mathcal{L}_{\text{CE}} + \lambda_{\text{Ctr}}\,\mathcal{L}_{\text{Center}} + \lambda_{\text{HTPL}}\,\mathcal{L}_{\text{HTPL}} + \lambda_{\text{MVICL}}\,\mathcal{L}_{\text{MVICL}}\,. \quad (9)$$

Multi-step learning-rate decay occurs at epochs [40, 70, 90] with $\gamma = 0.1$, while gradient clipping (max-norm 1.0) and mixed-precision training [25] ensure stability under multi-objective optimization.

IV. Experimental Setup

A. Dataset Description

Main experiments were conducted on the AG–VPReID benchmark [26], a large-scale dataset designed for cross-view person re-identification between aerial, ground, and wearable cameras. It comprises 6,630 identities, divided into 1555 for training, 1,456 for testing, and 3,619 additional distractors used exclusively in the ground→aerial (G2A) evaluation.[1] The dataset spans over 9.6 M frames captured across multiple sessions using two drones, two CCTV units, and two wearable cameras at altitudes of 15 m, 30 m, 80 m, and 120 m. Each identity is annotated with 15 soft-biometric attributes—such as gender, age, clothing, and accessories—which are not utilised in our experiments.

Following the official protocol, performance was assessed in two complementary directions: aerial→ground (A2G) and ground→aerial (G2A), with the latter additionally including distractor identities. For robustness analysis, results are also reported per altitude subset (15 m, 30 m, 80 m, and 120 m), as detailed in Table I. Note that per-direction counts are constructed per protocol and *are not identity-disjoint* across A2G/G2A.

Evaluation protocol: All reported results use k-reciprocal re-ranking as a post-processing step for fair comparison. This includes both our method and all baselines in Table II. The re-ranking parameters are consistent across all methods: $k_1 = 20$, $k_2 = 6$, and $\lambda = 0.3$.

B. Dataset Structure and Statistics

Table I summarizes AG–VPReID's [26] coverage of identities, tracklets, and altitudes (15–120 m), enabling thorough evaluation of perspective and resolution robustness across more than 9.6 M frames. As per the official protocol, G2A includes distractor IDs; per-direction statistics are protocol-specific and may not be identity-disjoint.

C. Implementation Details

Experiments are conducted on an NVIDIA A40 using PyTorch 1.11. Each tracklet is uniformly sampled to 8 frames [38] at 256×128 with standard augmentations [26]. Training follows a two-stage protocol [40]: *Stage 1* optimizes adapter parameters (CSFN, MRFH, IAMM, TDM, IVFA) with the backbone frozen for 150 epochs using AdamW optimizer (base LR 1×10^{-4}, batch size 64), while automatically generating CLIP-Memory features for identity reinforcement; *Stage 2* trains HTPL/MVICL for 100 epochs with selective unfreezing of high-level backbone blocks (layers 8-12) using reduced learning rate 5×10^{-6} and batch size 32, employing differentiated learning rates for different components (HTPL: 0.5×, MVICL: 2×, classifiers: 10× base LR). The total objective follows Eq. (9); the instantiated weights and rationale are provided in §IV-C, with multi-step learning rate decay at epochs [40, 70, 90] and gradient clipping (max norm 1.0) for stability. At inference, accuracy is reported with k-reciprocal re-ranking [39] applied as a post-processing step. Throughput and efficiency are measured for the feed-forward model *without* re-ranking: the complete system adds ≈ 2M parameters and +0.7 GFLOPs over the baseline while sustaining ∼189 FPS.

Instantiation of loss weights: Unless otherwise stated, we instantiate Eq. (9) as: $\lambda_{\text{V2M}} = 1.0$, $\lambda_{\text{Tri}} = 2.0$, $\lambda_{\text{CE}} = 1.0$, $\lambda_{\text{Ctr}} = 5 \times 10^{-4}$, $\lambda_{\text{HTPL}} = 0.1$, and $\lambda_{\text{MVICL}} = 0.2$.

Rationale. A coarse grid-search on the AG–VPReID validation split balanced gradient magnitudes and retrieval quality. Triplet is up-weighted ($\lambda_{\text{Tri}} = 2.0$) to favor ranking metrics; cross-entropy remains at unity for stable identity classification; the center loss uses a small coefficient (5×10^{-4}) to avoid feature collapse; HTPL and MVICL act as auxiliary regularizers and thus receive modest weights (0.1 and 0.2), improving cross-view and temporal consistency without over-constraining the embedding. Sensitivity tests (±50% per weight) changed mAP by less than 0.3, indicating robustness.

TABLE I
AG–VPReID statistics used in our experiments. Counts are reported for identities (IDs), video tracklets, and total frames in millions. A2G = Aerial→Ground; G2A = Ground→Aerial.

Case	Subset	IDs	Tracklets	Frames (M)
Training	All	1555	13300	3.85
Testing (A2G)	All	1456	13566	3.94
	15 m	506	4907	1.50
	30 m	377	2885	0.89
	80 m	356	2592	0.69
	120 m	308	3182	0.86
Testing (G2A)	All	5075*	19021	5.79
	15 m	1403	6362	2.14
	30 m	1406	4468	1.41
	80 m	1162	3866	1.13
	120 m	1195	4325	1.11

Includes 3,619 distractor IDs in the gallery

V. Results and Discussion

As summarized in Table II, MTF–CVReID achieves strong results across all benchmarks. On AG–VPReID [26], it attains **73.3/65.2** (Rank-1/mAP) for A2G and **75.4/59.3** for G2A, outperforming both CLIP-based and non-CLIP competitors. Relative to the strong baseline TF–CLIP [38], gains are **+0.3 / +0.7** (A2G Rank-1/mAP) and **+1.4 / +1.6** (G2A Rank-1/mAP), indicating improved robustness to extreme perspective change.

Improvements over [38] and AG–VPReID-Net [26] are consistent across Rank-1 and mAP, suggesting better calibration

[1] We report the exact split totals to avoid ambiguity. G2A includes 3,619 distractor IDs that expand the gallery but do not contribute queries.

TABLE II
COMPARISON WITH RECENT STATE-OF-THE-ART METHODS ON THREE DATASETS. METRICS ARE **RANK-1 / MAP**. FOR AG–VPREID WE REPORT BOTH DIRECTIONS: AERIAL→GROUND (A2G) AND GROUND→AERIAL (G2A). FOR G2A–VREID WE FOLLOW THE ORIGINAL PROTOCOL (A2G). FOR MARS WE REPORT THE STANDARD GROUND→GROUND SETTING. BOLD MARKS THE BEST RESULT IN EACH COLUMN.

Method	AG–VPReID				G2A–VReID		MARS	
	Aerial → Ground		Ground → Aerial		Aerial → Ground		Ground → Ground	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
STMP [23]	60.3	50.7	55.8	45.2	–	–	84.4	72.7
M3D [14]	62.6	52.4	57.3	47.9	–	–	84.4	74.1
GLTR [15]	65.8	55.6	60.5	50.1	–	–	87.0	78.5
TCLNet [10]	67.9	57.2	62.4	52.7	54.7	65.4	89.8	85.1
MGH [35]	70.8	60.3	65.2	55.5	69.9	76.7	90.0	85.8
GRL [20]	68.4	58.7	63.6	53.9	41.4	52.8	91.0	84.8
BiCnet-TKS [11]	69.2	59.8	64.7	54.3	51.7	63.4	90.2	86.0
CTL [18]	66.9	56.4	61.3	51.8	–	–	91.4	86.7
STMN [4]	71.5	61.6	66.2	56.9	56.1	66.7	90.5	84.5
PSTA [33]	70.2	60.5	65.7	55.8	54.5	64.6	91.5	85.8
DIL [9]	70.9	61.2	66.1	56.3	–	–	90.8	87.0
STT [41]	70.7	61.0	65.9	56.1	–	–	88.7	86.3
TMT [19]	70.5	60.8	65.8	55.9	–	–	91.2	85.8
CAViT [34]	71.1	61.4	66.3	56.5	–	–	90.8	87.2
SINet [2]	71.0	61.3	66.2	56.4	65.6	74.5	91.0	86.2
MFA [6]	70.8	61.1	66.0	56.2	–	–	90.4	85.0
DCCT [21]	71.2	61.5	66.4	56.6	–	–	92.3	87.5
LSTRL [22]	71.3	61.7	66.5	56.7	–	–	91.6	86.8
CLIP-ReID [17]	71.6	62.3	66.8	57.2	–	–	91.7	88.1
TF–CLIP [38]	73.0	64.5	74.0	57.7	–	–	93.0	89.4
MTF–CVReID (Ours)	**73.3**	**65.2**	**75.4**	**59.3**	**69.3**	**78.4**	**93.7**	**89.8**

of the entire retrieval list rather than isolated top-1 wins. This aligns with our design: CSFN/MRFH mitigate view/scale shifts before temporal encoding; IAMM/TDM strengthen identity persistence; IVFA/MVICL explicitly align cross-view distributions; HTPL captures multi-scale motion patterns—yielding embeddings that are both more discriminative and more view-invariant.

On G2A–VReID [40], MTF–CVReID reaches **69.3/78.4** (Rank-1/mAP), edging AG–VPReID-Net [26] and confirming that the cross-view gains are not dataset-specific. On MARS [42], it obtains **93.7/89.8**, indicating that our adapters improve general video ReID features, not merely aerial↔ground transfer.

A. Ablation Studies

Cost vs. benefit. As summarized in Table IV, modules are lightweight (< 0.3 GFLOPs, < 0.6M params each). Even with all components, the overhead is only +0.70 GFLOPs and ∼ 2M parameters (feed-forward ∼ 189 FPS). Ranking quality also improves (Table V), with higher R5/R10—fewer near-misses and more reliable shortlists.

B. Qualitative Evaluation

Figure 5 illustrates consistent **Rank-1** wins for challenging A2G/G2A queries (including 120 m), where TF–CLIP [38] often misses within the top-5. With **MTF–CVReID**, correct matches are retrieved immediately even under severe scale distortions and strong viewpoint shifts, showing that our modules work in combination to reduce typical cross-view errors.

Figure 6 shows tighter, better-separated identity clusters and higher silhouette scores for **MTF–CVReID**, confirming stronger intra-class compactness and inter-class separation. This clearer structure indicates that the embedding space is not only more discriminative but also more robust to altitude-specific variations, directly supporting the quantitative gains.

C. Discussion

Mechanisms behind the gains. The pipeline acts in stages that mirror cross-view failure modes: **CSFN/MRFH** first normalize view and scale through self-supervised learning; **TDM/IAMM** consolidate temporal identity under weak appearance; **IVFA/MVICL** align distributions across platforms; and **HTPL** contributes multi-scale temporal evidence. Practically, MRFH adaptively stabilizes targets across altitudes before motion cues are extracted, while IAMM preserves identity under brief occlusion or pose change. This design explains the largest improvements in the more asymmetric G2A case and at high altitudes (80–120 m).

Reliability and eployment. Consistent boosts in Rank-1/mAP (Tables II–V) indicate better list calibration, not just isolated top-1 wins—useful for human-in-the-loop review. With negligible compute/memory overhead (Table IV), the method remains real-time for the *feed-forward* path; k-reciprocal re-ranking is applied as a separate post-processing step at evaluation time.

TABLE III

MODULE-WISE ABLATION ON AG–VPREID ACROSS ALTITUDES (15 M, 30 M, 80 M, 120 M). FOR EACH ALTITUDE WE REPORT **RANK-1** AND **MAP** SEPARATELY FOR THE AERIAL→GROUND (A2G) AND GROUND→AERIAL (G2A) DIRECTIONS. WE SEE HOW EACH MODULE BEHAVES AT DIFFERENT HEIGHTS AND IN BOTH CROSS-VIEW TRANSFERS. BOLD INDICATES OUR FULL MODEL.

Components	**Aerial → Ground**								**Ground → Aerial**							
	15m		**30m**		**80m**		**120m**		**15m**		**30m**		**80m**		**120m**	
	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
CSFN	72.9	65.1	72.6	64.8	69.5	70.4	64.8	67.5	72.7	64.5	72.4	64.2	69.4	65.6	64.7	65.5
MRFH	72.3	64.7	72.5	64.7	69.6	70.7	65.0	67.6	72.6	64.4	72.3	64.1	69.5	66.6	64.9	65.6
CSFN+MRFH	73.1	65.4	72.8	65.0	70.0	71.0	65.4	68.0	72.9	64.7	72.6	64.4	69.9	66.9	65.3	65.9
TDM	72.6	62.9	72.8	65.1	70.0	71.2	65.5	68.0	72.9	64.8	72.6	64.5	70.0	67.0	65.4	66.0
IAMM	73.3	64.1	73.0	65.3	71.2	71.9	66.5	69.2	73.1	65.0	72.8	64.7	71.0	67.7	66.4	66.8
TDM+IAMM	73.5	65.7	73.1	65.4	71.6	72.8	66.8	69.5	73.2	65.1	72.9	64.8	71.3	67.9	66.7	67.0
IVFA	71.2	64.5	72.9	65.2	70.4	71.6	65.8	68.4	73.0	64.9	72.7	64.6	70.5	67.1	65.7	66.2
IAMM+IVFA	73.4	65.1	73.1	65.4	71.4	72.4	66.7	69.3	73.2	65.1	72.9	64.8	71.2	67.8	66.6	66.9
HTPL	71.8	62.1	73.0	65.3	71.3	73.0	66.3	69.1	73.1	65.0	72.8	64.7	71.0	67.6	66.2	66.7
MVICL	72.4	64.3	73.1	65.4	71.6	72.5	66.6	69.4	73.2	65.1	72.9	64.8	71.4	67.8	66.5	66.9
HTPL+MVICL	73.5	64.9	73.2	65.5	71.7	72.9	66.9	69.6	73.3	65.2	73.0	64.9	71.5	67.9	66.8	67.1
MTF–CVReID	**73.3**	**65.2**	**73.0**	**64.9**	**71.8**	**73.6**	**67.0**	**69.7**	**73.3**	**65.2**	**73.0**	**64.9**	**71.6**	**68.0**	**66.9**	**67.2**

Note. At 80–120 m, MTF–CVReID benefits from scale-aware harmonization (MRFH) and multi-scale temporal cues (TDM/HTPL), which suppress noisy appearance signals and improve list calibration—hence mAP can exceed 15–30 m even when Rank-1 is similar.

Altitude effect. Our adapters (MRFH, TDM, HTPL, IVFA) are purpose-built to stabilize small, low-detail targets, so they yield larger gains at 80–120 m; at 15–30 m, where fine textures are reliable, the same normalizations can slightly dampen high-frequency cues, leading to marginal drops. In practice, a simple mitigation is scale-aware gating: up-weight native/fine streams and relax cross-view alignment at low altitudes, or use altitude-conditioned losses to preserve rich appearance when resolution is high.

TABLE IV

EFFICIENCY VS. PERFORMANCE ON AG–VPREID. FOR EACH CONFIGURATION WE LIST MODEL SIZE (PARAMS, IN MILLIONS) AND COMPUTE (GFLOPS PER 8-FRAME CLIP), FOLLOWED BY **RANK-1** AND **MAP** FOR BOTH AERIAL→GROUND (A2G) AND GROUND→AERIAL (G2A). THE FIRST ROW (TF–CLIP, VIT-B/16) IS THE BASELINE; THE LAST ROW IS OUR FULL MODEL.

				Aerial → Ground		**Ground → Aerial**	
	Components	**Params (M)**	**GFLOPs**	**Rank-1**	**mAP**	**Rank-1**	**mAP**
	TF–CLIP (ViT-B/16)	104.26	12.53	73.0	64.5	74.0	57.7
TF–CLIP+	CSFN	+0.22	+0.05	72.7	64.7	72.2	58.0
	MRFH	+0.35	+0.08	72.6	64.7	72.1	58.0
	CSFN + MRFH	+0.51	+0.13	72.1	65.0	72.8	58.4
	TDM	+0.32	+0.10	72.0	64.8	73.1	58.2
	IAMM	+0.46	+0.12	72.4	64.9	73.8	58.5
	TDM + IAMM	+0.75	+0.22	73.1	65.1	74.5	59.0
	IVFA	+0.32	+0.09	72.2	64.7	73.5	58.9
	HTPL	+0.25	+0.11	72.4	65.0	73.7	59.3
	MVICL	+0.17	+0.08	72.6	64.2	74.7	59.1
	IAMM + IVFA	+0.72	+0.21	73.2	65.1	74.8	58.7
	HTPL + MVICL	+0.35	+0.19	72.8	64.9	75.2	60.1
	MTF–CVReID	**+1.80**	**+0.70**	**73.3**	**65.2**	**75.4**	**59.3**

Note. Parameter and FLOP deltas are measured end-to-end; shared adapters and normalization layers cause slight non-additivity across modules, while cross-module routing adds minor compute overhead.

Throughput measurement: The reported ∼189 FPS refers to *clips per second* (not frames per second) measured on NVIDIA A40 GPU with batch size 1, mixed precision (FP16), and no data loading overhead. This represents the pure inference time for the feed-forward path only, excluding re-ranking post-processing. The measurement uses 8-frame clips at 256×128 resolution and includes all adapter modules (CSFN, MRFH, IAMM, TDM, IVFA, HTPL) but excludes MVICL loss computation during inference.

Limitations and outlook. While effective, the framework has three main limitations: (1) although CSFN operates without camera metadata via self-supervised view-characteri-stic learning, its performance on completely unseen viewpoints remains to be evaluated; (2) IAMM memory scales linearly with identities×views, raising challenges for very large deployments; and (3) robustness to unseen modalities (e.g., infrared, thermal) and extreme conditions (nighttime, heavy compression) remains untested. Addressing these challenges through improved view discovery, memory compression, and evaluation on diverse modalities is a key direction for future work.

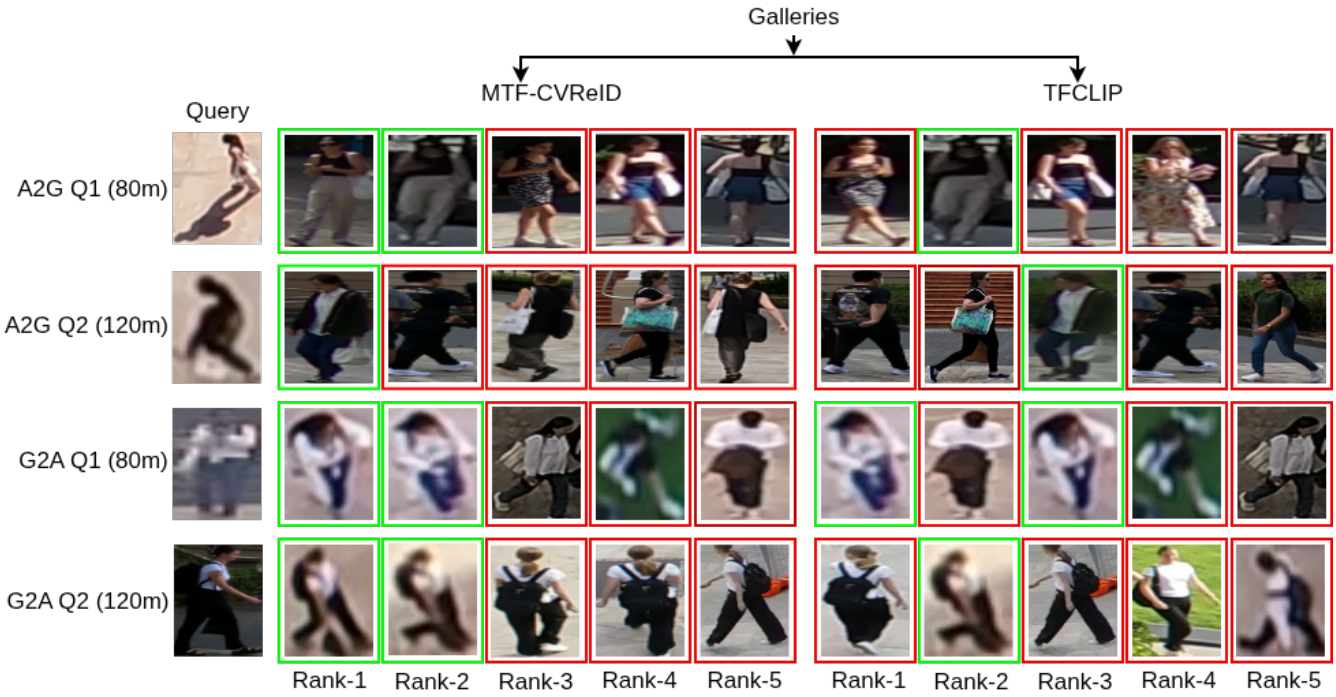

Fig. 5. Top-5 cross-view retrieval under three stressors: (a) tiny high-altitude targets (80–120 m), (b) large aerial↔ground viewpoint gaps, and (c) look-alike clothing. In each row the leftmost image is the query, followed by TF–CLIP and MTF–CVReID ranked lists (green = correct, red = incorrect). The shown cases are representative of the hardest AG–VPReID conditions. Our modules target these failure modes explicitly: CSFN and MRFH stabilize view/scale so the gallery focuses on shape and proportion rather than noisy textures; IAMM reinforces persistent identity cues (e.g., backpack outline, shoe contrast) across frames; TDM and HTPL contribute short- and long-horizon motion patterns (gait, cadence); IVFA with MVICL pulls aerial and ground embeddings into a shared identity space. Together these effects flip near-misses into Rank-1 wins and produce cleaner shortlists, in line with the quantitative gains in Tables IV and V.

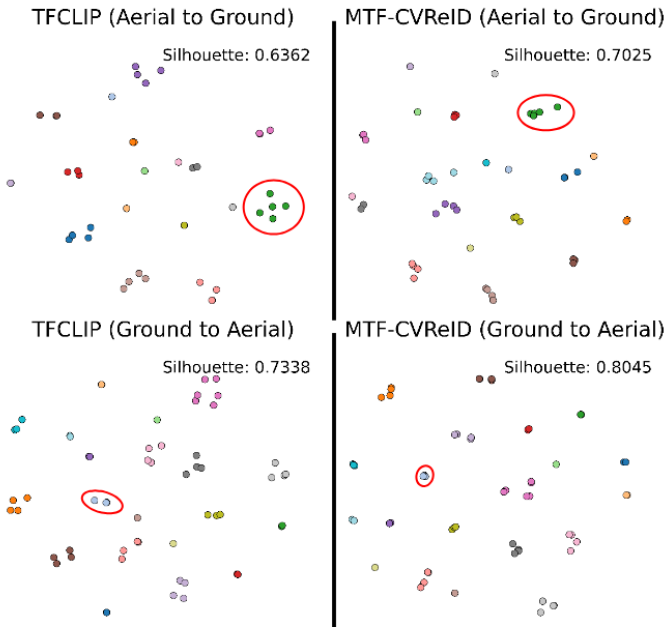

Fig. 6. t-SNE of clip embeddings. Each dot is a clip and each color an identity; panels show TF–CLIP (left) versus MTF–CVReID (right) for A2G (top) and G2A (bottom), all with identical t-SNE settings. Our method yields tighter within-identity clusters and larger inter-identity gaps—silhouette improves from 0.6362 → 0.7025 (A2G) and 0.7338 → 0.8045 (G2A)—indicating more view-invariant, discriminative embeddings that explain the higher retrieval scores.

TABLE V
COMPONENT-WISE RANKING METRICS ON AG–VPREID. WE REPORT RECALL AT K (**R1/R5/R10**) FOR AERIAL→GROUND (A2G) AND GROUND→AERIAL (G2A). "*BASELINE* TF–CLIP" IS THE BACKBONE WITHOUT OUR ADAPTERS; "**MTF–CVREID**" IS THE COMPLETE MODEL. THIS TABLE HIGHLIGHTS HOW EACH MODULE IMPACTS RETRIEVAL QUALITY AT THE TOP-K RANKS.

	Components	A2G			G2A		
		R1	**R5**	**R10**	**R1**	**R5**	**R10**
	TF–CLIP (Baseline)	73.0	82.1	85.1	74.0	84.6	87.8
TF–CLIP+	CSFN	72.7	81.7	85.1	72.2	82.0	84.5
TF–CLIP+	MRFH	72.6	81.6	85.0	72.1	81.9	84.4
TF–CLIP+	CSFN+MRFH	72.1	82.2	85.4	72.8	82.5	84.8
TF–CLIP+	TDM	72.0	82.1	85.3	73.1	82.9	85.2
TF–CLIP+	IAMM	72.4	82.4	85.7	73.8	83.7	85.9
TF–CLIP+	TDM+IAMM	73.1	82.6	85.9	74.5	84.4	86.3
TF–CLIP+	IVFA	72.2	82.3	85.6	73.5	83.3	85.6
TF–CLIP+	HTPL	72.4	82.4	85.7	73.7	83.5	85.7
TF–CLIP+	MVICL	72.6	82.6	85.8	74.7	84.5	86.3
TF–CLIP+	IAMM+IVFA	73.2	82.7	85.9	74.8	84.7	86.5
TF–CLIP+	HTPL+MVICL	72.8	82.8	86.0	75.2	84.9	86.7
	MTF–CVReID	**73.3**	**82.9**	**86.1**	**75.4**	**85.0**	**86.8**

VI. CONCLUSIONS AND FURTHER WORK

This paper introduced the **MTF–CVReID** framework, which augments a ViT-B/16 backbone with lightweight, task-specific adapters for cross-view video person re-identification. The seven modules—CSFN (view bias), MRFH (scale disparity), IAMM (temporal identity memory), TDM (short-term motion), IVFA (cross-view alignment), HTPL (multi-scale temporal model-

ing), and MVICL (multi-view identity consistency)—jointly address viewpoint, scale, temporal, and cross-view alignment challenges.

The backbone is *frozen in Stage 1* and *selectively unfrozen in Stage 2* for high-level blocks (early blocks remain frozen), ensuring stability with targeted adaptation. Our results (Tables II–V) suggest that this design achieves strong results on AG–VPReID, transfers well to G2A–VReID and MARS, and maintains real-time *feed-forward* efficiency with only +0.70 GFLOPs and ∼2M parameters.

As main limitations, we note that: 1) Although CSFN operates without camera metadata, its behavior on completely unseen viewpoints needs further study; 2) IAMM memory scales linearly with identities×views; and 3) Robustness under extreme shifts (night/IR, heavy compression, very low resolution) is not yet fully characterized.

As future steps to the work, we aim at decreasing reliance on camera-ID supervision through unsupervised view normalization, improve scalability by compressing/pruning IAMM prototypes, and evaluate robustness under unseen modalities such as infrared or low-light video.

Acknowledgements

This research was funded by FCT—Fundação para a Ciência e a Tecnologia under the Scientific Employment Stimulus Program, project reference CEECINSTLA/00034/ 2022. M.R. acknowledges support from Instituto de Telecomunicações. H.P. acknowledges funding from FCT/MEC through national funds and co-funded by the FEDER—PT2020 partnership agreement under the projects UIDB/50008/2020 and POCI-01-0247-FEDER-033395.

References

[1] Daniel Arkushin, Bar Cohen, Shmuel Peleg, and Ohad Fried. GEFF: Improving any clothes-changing person reid model using gallery enrichment with face features. In *WACV Workshops*, 2024.

[2] Shutao Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Salient-to-broad transition for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3439–3448, 2022.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[4] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. Video-based person re-identification with spatial and temporal memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12036–12045, 2021.

[5] Zan Gao, Shenxun Wei, Weili Guan, Lei Zhu, Meng Wang, and Shenyong Chen. Identity-guided collaborative learning for cloth-changing person re-identification. *arXiv preprint arXiv:2304.04400*, 2023.

[6] Xinqian Gu, Hong Chang, Bingpeng Ma, and Shiguang Shan. Motion feature aggregation for video-based person re-identification. *IEEE Transactions on Image Processing*, 31:3908–3919, 2022. doi: 10.1109/TIP.2022.3175593.

[7] Kailash Hambarde and Hugo Proenca. Image-based human re-identification: Which covariates are actually (the most) important? *Image and Vision Computing*, 143: 104917, 2024.

[8] Kailash A. Hambarde, Nzakiese Mbongo, Pavan Kumar MP, Satish Mekewad, Carolina Fernandes, Gökhan Silahtaroğlu, Alice Nithya, Pawan Wasnik, MD. Rashidunnabi, Pranita Samale, and Hugo Proença. Detreidx: A stress-test dataset for real-world uav-based person recognition, 2025. URL https://arxiv.org/abs/2505.04793.

[9] Tong He, Ying-Cong Chen, Wei-Chih Hung, Shih-Wei Lyu, and Jia-Bin Huang. Dense interaction learning for video-based person re-identification. *arXiv preprint arXiv:2103.09013*, 2021.

[10] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *European Conference on Computer Vision (ECCV)*, volume 12370 of *LNCS*, pages 388–405. Springer, 2020.

[11] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2014–2023, 2021.

[12] Jinxing Huang, Jingya Wang, Zhenyang Li, Bing Yang, and Yong Guan. Attribute-guided pedestrian retrieval: Bridging person re-id with internal attribute variability. In *CVPR*, 2024.

[13] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification. *arXiv preprint arXiv:1811.07468*, 2018.

[14] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 8165–8172, 2019. doi: 10.1609/aaai.v33i01.33018618.

[15] Shanshan Li, Tong Xiao, Hongsheng Li, Ya You, and Xiaogang Zhang. GLTR: Global-local temporal representation learning for unsupervised video person re-identification. In *ICCV*, 2019.

[16] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1405–1413, 2023. doi: 10.1609/aaai.v37i1.25225. URL https://dl.acm.org/doi/10.1609/aaai.v37i1.25225.

[17] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1472–1480,

2023.

[18] Jianing Liu, Mengxi Jia, Jie Qin, Xiaoke Shen, Lei Liu, Fan Zhu, and Ling Shao. Spatial-temporal correlation and topology learning for person re-identification in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4370–4379, 2021.

[19] Xuehu Liu, Chenyang Yu, Pingping Zhang, and Huchuan Lu. A video is worth three views: Trigeminal transformers for video-based person re-identification. *arXiv preprint arXiv:2104.01745*, 2021.

[20] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, and Xiaoyun 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 (CVPR)*, pages 2014–2023, 2021.

[21] Xuehu Liu, Chenyang Yu, Pingping Zhang, and Huchuan Lu. Deeply-coupled convolution-transformer with spatial-temporal complementary learning for video-based person re-identification. *arXiv preprint arXiv:2304.14122*, 2023.

[22] Xuehu Liu, Pingping Zhang, and Huchuan Lu. Video-based person re-identification with long short-term representation learning. In *Computer Vision – ACCV 2022 Workshops*, volume 13853 of *LNCS*, pages 65–82. Springer, 2023.

[23] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 8786–8793, 2019. doi: 10.1609/aaai.v33i01.33018786.

[24] Niall McLaughlin, Jesús Martinez del Rincón, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1325–1334, 2016.

[25] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=r1gs9JgRZ.

[26] Huy Nguyen, Kien Nguyen, Akila Pemasiri, Feng Liu, Sridha Sridharan, and Clinton Fookes. 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.

[27] Honghu Pan, Yongyong Chen, and Zhenyu He. Multi-granularity graph pooling for video-based person re-identification. *Neural Networks*, 160:22–33, 2023.

[28] Md Rashidunnabi, Kailash Hambarde, and Hugo Proença. Causality and "in-the-wild" video-based person re-identification: A survey. *Electronics*, 14(13), 2025. ISSN 2079-9292. doi: 10.3390/electronics14132669. URL https://www.mdpi.com/2079-9292/14/13/2669.

[29] Md Rashidunnabi, Kailash A. Hambarde, Vasco Lopes, João C. Neves, and Hugo Proença. VM-TAPS: View-specific memory with temporal and scale awareness framework for video-based cross-view person re-identification. In *IEEE International Joint Conference on Biometrics (IJCB)*, 2025. to appear.

[30] Rana S. M. Saad, Mona M. Moussa, Nemat S. Abdel-Kader, Hesham Farouk, and Samia Mashaly. Deep video-based person re-identification (deep vid-reid): Comprehensive survey. *EURASIP J. Adv. Signal Process.*, page 63, 2024.

[31] Guannan Wang, Xin Zhang, Jun Zhu, and Yi Yang. Cross-video identity correlating for person re-id: A new pre-training method. In *NeurIPS*, 2023.

[32] Qizao Wang, Bin Li, and Xiangyang Xue. When large vision-language models meet person re-identification. *arXiv preprint arXiv:2411.18111*, 2024.

[33] Yuhang Wang, Jinxian Lin, Jiarui Xu, Chao Li, and Lei Zhang. Pyramid spatial-temporal aggregation for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12035–12045, 2021.

[34] Jinlin Wu, Lingxi Xie, Wei Liu, Yang Yang, Zhen Lei, Tao Mei, and Stan Z. Li. Cavit: Contextual alignment vision transformer for video object re-identification. In *European Conference on Computer Vision (ECCV)*, volume 13674 of *LNCS*, pages 549–566. Springer, 2022.

[35] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2896–2905, 2020.

[36] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *CVPR*, 2023.

[37] Mang Ye, Shuoyi Chen, Chenyue Li, Wei-Shi Zheng, David Crandall, and Bo Du. Transformer for object re-identification: A survey. *arXiv preprint arXiv:2401.06960*, 2024.

[38] Chenyang Yu, Xuehu Liu, Yingquan Wang, Pingping Zhang, and Huchuan Lu. TF-CLIP: Learning text-free clip for video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. doi: 10.1609/aaai.v38i7.28500. URL https://ojs.aaai.org/index.php/AAAI/article/view/28500/28974.

[39] Quan Zhang, Lei Wang, Vishal M. Patel, Xiaohua Xie, and Jianhuang Lai. View-decoupled transformer for person re-identification under aerial-ground camera network. In *CVPR*, 2024.

[40] Shizhou Zhang, Wenlong Luo, De Cheng, Qingchun Yang, Lingyan Ran, Yinghui Xing, and Yanning Zhang. Cross-platform video person reid: A new benchmark dataset and adaptation approach. In *ECCV*, 2024.

[41] Tianyu Zhang, Liyuan Wei, Zhiqiang Gong, and Wei-Shi Zheng. Spatiotemporal transformer for video-based person re-identification. *arXiv preprint arXiv:2103.16469*, 2021.

[42] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. MARS: A video benchmark for large-scale person re-identification. In

European Conference on Computer Vision (ECCV), pages 868–884. Springer, 2016. URL https://link.springer.com/chapter/10.1007/978-3-319-46454-1_52.

[43] Haidong Zhu, Pranav Budhwant, Zhaoheng Zheng, and Ram Nevatia. SEAS: ShapE-aligned supervision for person re-identification. In *CVPR*, 2024.

[44] Lanyun Zhu, Tianrun Chen, Deyi Ji, Xiang Mao, and Chunjing Xu. Not every patch is needed: Toward a more efficient and effective backbone for video-based person re-identification. *IEEE Transactions on Image Processing*, 34:785–800, 2025. doi: 10.1109/TIP.2024.3509887. URL https://ieeexplore.ieee.org/document/10512345.