Scale-Aware Vision-Language Adaptation for Extreme Far-Distance Video Person Re-identification

Anonymous CVPR submission

Paper ID 3

Abstract

Extreme far-distance video person re-identification (ReID) is particularly challenging due to scale compression, resolution degradation, motion blur, and aerial-ground viewpoint mismatch. As camera altitude and subject distance increase, models trained on close-range imagery degrade significantly. In this work, we investigate how large-scale vision-language models can be adapted to operate reliably under these conditions. Starting from a CLIP-based baseline, we upgrade the visual backbone from ViT-B/16 to ViT-L/14 and introduce backbone-aware selective fine-tuning to stabilize adaptation of the larger transformer. To address noisy and low-resolution tracklets, we incorporate a lightweight temporal attention pooling mechanism that suppresses degraded frames and emphasizes informative observations. We retain adapter-based and prompt-conditioned cross-view learning to mitigate aerial–ground domain shifts, and further refine retrieval using improved optimization and k-reciprocal re-ranking. Experiments on the DetReIDX stress-test benchmark show that our approach achieves an average mAP of 35.73 across aerial-ground (A2G), ground-aerial (G2A), and aerial-aerial (A2A) protocols, improving substantially over the existing baseline. These results show that large-scale vision-language backbones, when combined with stability-focused adaptation, significantly enhance robustness in extreme far-distance video person ReID.

1. Introduction

Person re-identification (ReID) aims to match individuals across non-overlapping camera views [1] and has progressed rapidly with vision transformers, CLIP-based text encoders, and large-scale pretraining [2]. These advances have significantly improved performance on standard ground-level benchmarks and enabled applications in surveillance, crowd analysis, and public safety. However, extreme far-distance aerial scenarios operate under fundamentally different visual constraints. As the altitude of unmanned aerial vehicles (UAVs) and subject distance increase, pedestrians often occupy only a few pixels and are affected by severe scale compression, resolution degradation, motion blur, and strong viewpoint differences between aerial and ground cameras, as shown in Figure 1. Under such conditions, fine-grained appearance cues become unreliable, and models developed for close-range imagery degrade sharply. The assumptions underlying conventional ReID, sufficient spatial detail, and consistent viewpoints no longer apply.

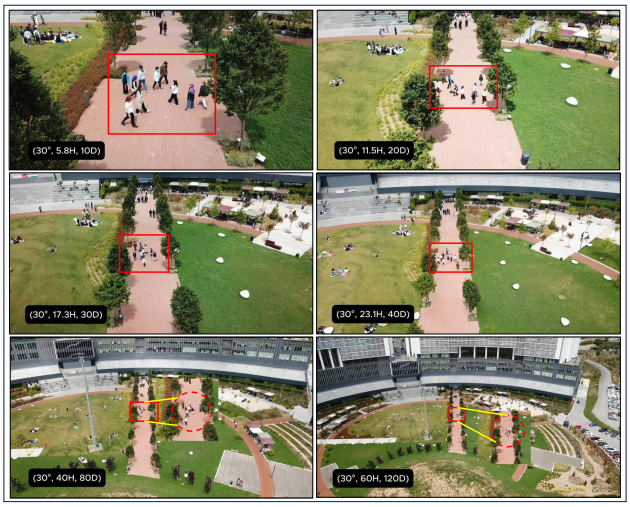

Figure 1. Illustration of extreme far-distance conditions in the DetReIDX dataset. As UAV altitude and horizontal distance increase, pedestrians undergo severe scale compression and resolution degradation. The red bounding boxes highlight the same group of individuals as their pixel footprint shrinks dramatically, showing the visual challenges of aerial-ground video person ReID.

Recent stress-test datasets and benchmarks for UAV-based person analysis highlight this limitation explicitly. In particular, the DetReIDX was introduced to evaluate robustness under long-range viewpoints, aerial-ground cross-domain matching, and session-level appearance changes.

The benchmark reveals that strong detection and ReID baselines can collapse under extreme distance and viewpoint variation. Improving performance, therefore, requires adaptation strategies that explicitly account for scale degradation and noisy video tracklets, rather than direct transfer from conventional settings.

In this work, we explore how large-scale vision-language backbones can be effectively adapted for extreme far-distance video ReID. Starting from the official CLIP ViT-B/16 baseline for DetReIDX, we investigate backbone scaling and stability-oriented fine-tuning. We upgrade the visual encoder to ViT-L/14 and introduce backbone-aware selective fine-tuning to preserve pretrained representations while enabling high-level domain adaptation. To improve robustness to frame-level degradation, we incorporate a lightweight temporal attention pooling mechanism that emphasizes informative frames within noisy tracklets. We further retain prompt-based cross-view conditioning and adapter tuning to mitigate aerial-ground domain shifts, and refine retrieval performance through optimized training strategies and k-reciprocal re-ranking [3]. On the DetReIDX benchmark, evaluated across aerial-to-aerial (A2A), aerial-to-ground (A2G), and ground-to-aerial (G2A) protocols [4], our approach improves the average mAP from 28.11 to 35.73, surpassing the recent work results (32.89). These findings demonstrate that large-scale CLIP backbones, when carefully adapted, substantially improve robustness in extreme far-distance video person ReID. We summarize the contribution of this work below:

1. We present a systematic study of CLIP backbone scaling (ViT-B/16 to ViT-L/14) for extreme far-distance aerial-ground video person ReID, demonstrating that increased model capacity improves robustness under severe scale compression.
2. We propose a stability-focused adaptation strategy combining backbone-aware selective fine-tuning, temporal attention pooling, and prompt/adapter tuning to handle noisy, low-resolution video tracklets.
3. We introduce practical optimization refinements, including cosine scheduling and enhanced data augmentation, and enable k-reciprocal re-ranking at inference.
4. Our method achieves 35.73 mAP on the benchmark, improving substantially from the baseline (28.11) and the highest publicly reported score (32.89).

The remainder of this paper is organized as follows. Section 2 reviews related work in ground-level and aerial video person ReID, vision-language adaptation, and retrieval re-ranking. Section 3 describes the proposed scale-aware adaptation framework, including backbone scaling, selective fine-tuning, temporal attention pooling, and optimization refinements. Section 4 presents experimental results on the DetReIDX benchmark, including implementation details and quantitative comparisons. Section 5 discusses the results, limitations and future improvements. Finally, Section 6 concludes the paper.

2. Related Work

Person ReID has been extensively studied in both image-based and video-based settings. With the adoption of transformer backbones and large-scale pretraining, significant progress has been achieved on conventional ground-level benchmarks. However, robustness under extreme far-distance aerial conditions remains largely underexplored. We review related work in ground-level ReID, video-based temporal modeling, aerial and cross-view benchmarks, vision-language adaptation, and retrieval re-ranking.

2.1. Ground-level Person ReID

Early ReID research focused primarily on ground-level camera networks, where pedestrians are captured at moderate distances with sufficient spatial resolution. Benchmarks such as Market-1501 [5] and DukeMTMC4ReID [6, 7] established standardized evaluation protocols and drove rapid development of discriminative feature learning methods. More recently, transformer-based architectures, including TransReID [8] and related global-context modeling frameworks, have demonstrated strong performance by leveraging self-attention and large-scale pretraining.

Despite these advances, most methods are evaluated under relatively stable imaging conditions. When applied to extreme far-distance aerial scenarios, where person instances are severely scale-compressed and contain limited visual detail, their performance degrades substantially. This limitation highlights the challenge of directly transferring ground-level transformer models to aerial settings without explicit adaptation for scale degradation and viewpoint discrepancy.

2.2. Video-based ReID and Temporal Modeling

Video-based ReID extends image-based matching by exploiting temporal cues across tracklets. The MARS benchmark introduced large-scale video evaluation and motivated sequence aggregation strategies beyond simple frame averaging [9]. Early approaches employed recurrent architectures or heuristic pooling, while more recent works adopt attention-based mechanisms to emphasize informative frames and suppress noise. Methods such as spatio-temporal attention (STA) [10] and temporal complementary learning networks (TCLNet) [11] demonstrate the effectiveness of adaptive aggregation for handling occlusion, pose variation, and motion blur. In extreme far-distance aerial footage, tracklets often contain heavily degraded frames due to motion instability and severe resolution loss. Under such conditions, uniform aggregation can amplify noise, making adaptive temporal weighting particularly important.

Table 1. Comparison of aerial–ground benchmark datasets for person detection, ReID, tracking, and action recognition.

Dataset	Camera	Format	PIDs	BBoxes	Height (m)	Distance (m)
AG-ReID.v2	UAV+CCTV	Still	1615	100.6K	15–45	–
G2APS-ReID	UAV+CCTV	Still	2788	200.8K	20–60	–
DetReIDX	DSLR+UAV	Video+Still	334	13M	5–120	10–120

2.3. Aerial and Cross-View ReID Benchmarks

UAV-based person analysis introduces challenges that are not captured by conventional ground-only datasets. UAV-Human [12] and P-DESTRE [13] extend ReID to aerial platforms, incorporating detection, tracking, and both short and long-term matching tasks. For explicit aerial-ground cross-view matching, AG-ReID [14] and related datasets reveal the substantial appearance gap between top-down UAV imagery and horizontal ground cameras.

More recently, DetReIDX (see Table 1) was proposed as a stress-test benchmark [4] targeting extreme far-distance conditions. It explicitly models altitude variation, scale compression, cross-view mismatch, and session-level appearance drift. Experimental findings show that strong detection and ReID baselines can collapse under such conditions, underscoring the need for scale-aware and stability-oriented adaptation strategies. These characteristics make DetReIDX an appropriate benchmark for evaluating robustness in extreme far-distance aerial-ground ReID.

2.4. Vision-Language Pretraining and Efficient Adaptation

Vision-language models (VLMs), particularly CLIP, have demonstrated strong transferability through large-scale image-text pretraining [2]. CLIP-based features have been incorporated into ReID frameworks to improve generalization and mitigate the absence of semantic class labels. Approaches such as CLIP-ReID leverage prompt learning to bridge identity supervision and semantic embedding spaces.

Beyond prompt optimization, parameter-efficient adaptation methods, including adapters, low-rank updates, and conditional prompts (e.g., CoOp and CoCoOp) [15], enable stable fine-tuning of large pretrained backbones while reducing overfitting risk. Such strategies are particularly relevant in extreme far-distance aerial-ground ReID, where severe degradation and limited effective visual detail increase the risk of representation drift during adaptation.

2.5. Re-Ranking for Retrieval-based ReID

Post-processing remains an important component of retrieval-based ReID systems. In particular, k-reciprocal re-ranking refines similarity relationships by exploiting reciprocal nearest neighbors in the embedding space, leading to consistent improvements in mAP and ranking accuracy. Careful parameter tuning is often required to maximize gains for a given backbone and feature distribution [3].

In contrast to prior work, which primarily focuses on ground-level settings or moderate aerial conditions, we target the extreme far-distance regime evaluated by DetReIDX. We investigate how large-scale vision-language backbones can be systematically scaled and stably adapted to address severe scale compression and noisy video tracklets. Section 3 describes our proposed scale-aware adaptation framework in detail.

3. Scale-Aware Adaptation Framework

We build upon the official CLIP-based [2] video ReID baseline for the DetReIDX benchmark and introduce a series of scale-aware modifications tailored for extreme far-distance (XFD) aerial-ground scenarios. Our approach focuses on three key aspects: (1) increasing backbone capacity through model scaling, (2) stabilizing adaptation of the larger transformer, and (3) improving robustness to degraded video tracklets via adaptive temporal aggregation and refined optimization. We first describe the baseline framework and then present the proposed modifications.

3.1. Baseline Framework

We adopt the official CLIP-based video ReID baseline released for DetReIDX. The framework performs tracklet-level retrieval under aerial-ground cross-view conditions and includes a vision transformer backbone, prompt-based cross-view conditioning, and parameter-efficient adaptation modules.

Backbone and frame-level feature extraction. The baseline employs a Vision Transformer backbone with patch size 16×16 (ViT-B/16). All frames are resized to 256×128 during both training and testing [16].

Tracklet construction and sampling. Video ReID is performed at the tracklet level. For each training instance, a fixed-length sequence of $L = 16$ frames is sampled and encoded independently by the backbone. A softmax triplet sampling strategy is adopted to support both identity classification and metric learning [17]. The batch size is set to 16 tracklets. During inference, the same sequence length

Table 2. Baseline training configuration.

Parameter	Stage 1	Stage 2
Optimizer	Adam	Adam
Base Learning Rate	3.5e-4	1.0e-4
Max Epochs	120	120
Weight Decay	1e-4	2.5e-4
Weight Decay (Bias)	1e-4	1e-4
Images Per Batch	16	16
ID Loss Weight (λ_{id})	0.25	
Triplet Loss Weight (λ_{tri})	1.0	
I2T / T2I Weights	1.0	

is used, and retrieval is conducted using cosine similarity between ℓ_2-normalized features.

Prompt-based cross-view conditioning (PBP). To reduce aerial-ground domain mismatch, the baseline incorporates prompt-based cross-view conditioning. A prompt of length 1 is inserted across 9 transformer layers. Metadata signals, including altitude, horizontal distance, and viewing angle, are discretized into bins (18 altitude, 18 distance, and 3 angle bins) and encoded as conditioning inputs. This design injects physical context related to viewpoint and scale variation into the representation learning process.

Training strategy. The baseline is trained using a multi-term objective combining identity classification, triplet metric learning, and CLIP-style cross-modal alignment. The overall loss is defined as:

$$\mathcal{L} = \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{tri}}\mathcal{L}_{\text{tri}} + \lambda_{\text{i2t}}\mathcal{L}_{\text{i2t}} + \lambda_{\text{t2i}}\mathcal{L}_{\text{t2i}}, \quad (1)$$

where the loss weights and optimization parameters are summarized in Table 2.

Stage 1 of training focuses on stable representation learning with cosine-style decay, with some warm-up. Stage 2 training performs further fine-tuning using a reduced learning rate and multi-step scheduling to refine retrieval performance and improve ranking stability.

Inference settings. During evaluation, tracklet features are extracted using fixed-length sequences of 16 frames and ℓ_2 normalized before retrieval. Cosine similarity is used for distance computation. Re-ranking is disabled in the default baseline configuration, considering it may degrade the learned representation. The full inference configuration is summarized in Table 3.

Parameter-efficient adaptation via adapters. The baseline further uses lightweight adapter modules [18] as a parameter-efficient mechanism for adapting the pretrained transformer to the target domain. Other optional components provided by the baseline, such as VCAH and QATW, are disabled initially. Temporal Attention pooling is also disabled in the baseline, resulting in the use of a standard mean pooling.

Table 3. Baseline inference configuration.

Parameter	Setting
Batch Size	16
Sequence Length	16
Feature Normalization	ℓ_2 normalization
Distance Metric	Cosine
Re-Ranking	Disabled

3.2. Scale-Aware Adaptation

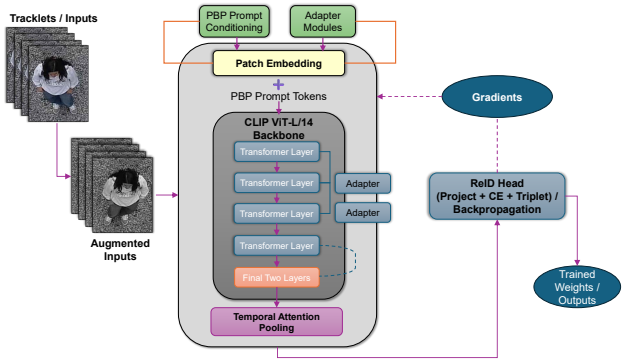

Figure 2. Overview of the proposed CLIP ViT-L/14-based video ReID framework. Augmented tracklets are encoded by a CLIP backbone enhanced with PBP prompt conditioning and adapter modules, with only the final two transformer blocks unfrozen during fine-tuning. Frame-level features are aggregated via temporal attention pooling and optimized using identity and triplet losses.

While the baseline provides a strong foundation for extreme far-distance video ReID, performance remains limited under severe scale compression and heavy frame degradation. We therefore introduce several targeted modifications focused on backbone capacity, stable large-model adaptation, robust temporal aggregation, re-ranking and optimized training strategies. An overview of the proposed pipeline is shown in Figure 2.

3.2.1. Backbone Scaling: ViT-B/16 → ViT-L/14

We modify the baseline configuration by switching the CLIP visual encoder from ViT-B/16 to the larger ViT-L/14 backbone, thereby increasing representational capacity under severe resolution degradation and scale compression. In extreme far-distance settings, pedestrian regions often oc-

cupy only a small number of pixels, and fine-grained appearance cues such as texture and local structure become unreliable. A larger transformer backbone provides improved modeling capacity to extract robust global and mid-level features from such degraded inputs [2, 16].

ViT-B/16 produces 768-dimensional embeddings, whereas ViT-L/14 outputs 1024-dimensional embeddings. Consequently, we adjust the associated projection layers, classification heads, prompt embeddings, and adapter modules to ensure dimensional consistency throughout the architecture. These adjustments are strictly architectural and do not alter the training objective or data sampling strategy. Apart from the backbone upgrade and the necessary dimensional alignment, the overall training pipeline is preserved. This design allows us to isolate the impact of backbone scaling while maintaining comparability with the original baseline.

3.2.2. Backbone-Aware Selective Fine-Tuning

Fine-tuning all parameters of a large transformer can degrade the strong feature representations learned during large-scale pretraining and lead to unstable optimization and overfitting under extreme far-distance conditions. Hence, instead of training the entire architecture on this dataset, the baseline had most backbone parameters frozen, focusing on adapter and prompt training, but we unfreeze the high-level blocks (i.e., blocks 22 and 23) of the architecture for adaptation. The final two blocks were unfrozen since the earlier layers encode general visual features. This strategy preserves the pretrained model's generalization ability while enabling high-level representations to adapt to the target domain [18, 19]. We further use differential learning rates by assigning a smaller learning rate to the unfrozen backbone blocks and a larger learning rate to lightweight modules and heads (e.g., prompts/adapters and the identity classifier). Specifically, the unfrozen backbone blocks are trained using a learning rate scaled to 0.1× of the base learning rate. This improves training stability when adapting the larger ViT-L/14 model and helps avoid catastrophic drift from pretrained weights.

3.2.3. Temporal Attention Pooling

The baseline uses a standard mean pooling because attention pooling is disabled. However, in extreme far-distance aerial settings, individual frames often suffer from severe blur, motion artifacts, occlusion, or extreme scale compression. Standard mean pooling can, therefore, allow degraded frames to negatively influence the final representation. To address this limitation, we enable and introduce a lightweight temporal attention mechanism that adaptively weights frames based on their feature responses [10, 11]. Let $\{f_t\}_{t=1}^{T}$ denote the sequence of frame-level embeddings extracted from a tracklet, where $T = 16$ in our setting and $f_t \in \mathbb{R}^C$.

Table 4. Optimization refinements compared to the baseline configuration.

Parameter	Baseline (ViT-B/16)	Ours (ViT-L/14)
Backbone	ViT-B/16	ViT-L/14
Stride Size	16	14
PBP Prompt Length	1	4
QATW Module	Disabled	Enabled
Instance Norm	Disabled	Enabled
Stage 1 Training		
Optimizer	Adam	Adam
Images Per Batch	16	48
Base Learning Rate	3.5e-4	2.0e-4
Max Epochs	120	50
LR Schedule	Cosine Annealing	Cosine Annealing
Weight Decay	1e-4	1e-4
Stage 2 Fine-Tuning		
Optimizer	Adam	Adam
Images Per Batch	16	24
Base Learning Rate	1.0e-4	1.0e-4
Max Epochs	120	40
LR Schedule	Multi-step (60, 90)	Cosine Annealing

We learn a scalar attention score for each frame using a linear projection:

$$s_t = w^\top f_t, \tag{2}$$

where $w \in \mathbb{R}^C$ is a learnable parameter vector implemented as a fully connected layer. The attention weights are obtained via softmax normalization across the temporal dimension:

$$\alpha_t = \frac{\exp(s_t)}{\sum_{k=1}^{T} \exp(s_k)}. \tag{3}$$

The final tracklet representation is computed as a weighted sum:

$$z = \sum_{t=1}^{T} \alpha_t f_t. \tag{4}$$

This mechanism enables the model to suppress degraded or irrelevant frames while emphasizing temporally consistent and discriminative observations. In our implementation, temporal attention is applied consistently to the intermediate and projected feature representations before classification and retrieval. Compared to the baseline, this adaptive pooling strategy improves robustness under extreme resolution degradation and unstable aerial motion.

3.2.4. Optimization Strategy

In addition to backbone scaling and selective fine-tuning, we refine the optimization strategy to better accommodate the larger ViT-L/14 backbone and improve training stability under extreme far-distance conditions. Compared to the

baseline configuration, we adjust batch size, learning rates, training duration, and scheduling strategy to balance convergence speed and generalization. Table 4 summarizes the key differences between the baseline and our refined optimization setup. Specifically, we increase the Stage 1 batch size from 16 to 48 to expose the larger backbone to more identity diversity per iteration. The base learning rate for Stage 1 is slightly reduced (from 3.5×10^{-4} to 2.0×10^{-4}) to stabilize optimization when training the deeper ViT-L/14 architecture. The total number of training epochs is also reduced in both stages (Stage 1: 120 $\rightarrow$ 50, Stage 2: 120 $\rightarrow$ 40), reflecting faster convergence with the scaled backbone.

In Stage 1, the selectively unfrozen backbone blocks are trained with a reduced learning rate scaled to $0.1\times$ the base learning rate, as described in Section 3.2.2, to prevent destabilization of pretrained representations. Unlike the baseline, which uses multi-step decay during Stage 2, our refined configuration adopts cosine annealing [20] for both training stages. This smooth decay schedule avoids abrupt learning rate drops and provides more stable adaptation of the selectively unfrozen transformer blocks. Additionally, instance normalization at the neck is enabled to further improve robustness under extreme scale degradation. These refinements collectively stabilize large-model adaptation and improve convergence behavior under severe resolution loss.

3.2.5. Data Augmentation

To improve robustness under diverse aerial capture conditions, we extend the baseline augmentation pipeline by incorporating color jitter during training. Specifically, we apply controlled perturbations in brightness, contrast, saturation, and hue to simulate illumination variation and camera-induced color shifts commonly observed in extreme far-distance imagery. In addition to random horizontal flipping and random erasing used in the baseline, these color augmentations encourage the model to rely less on fragile color cues and more on structural and identity-consistent features. During inference, we further apply horizontal flip augmentation and average the features extracted from the original and flipped sequences to improve orientation robustness.

3.2.6. k-Reciprocal Re-Ranking

To further refine retrieval results, we enable k-reciprocal re-ranking as a post-processing step during inference. After computing cosine distances between ℓ_2-normalized query and gallery embeddings, the initial distance matrix is refined using k-reciprocal encoding to improve neighborhood consistency in the embedding space. Specifically, we apply re-ranking with parameters ($k_1 = 28$, $k_2 = 6$, $\lambda = 0.28$). This procedure re-evaluates similarity relationships by considering reciprocal nearest neighbors, reducing the impact of local feature noise, and improving retrieval robustness. Re-ranking is applied only during evaluation and does not affect model training. This inference refinement yields additional improvements in the mean Average Precision (mAP) score.

4. Experiment & Results

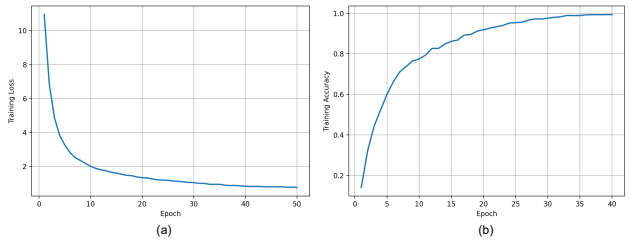

Figure 3. Training across the two-stage optimization process. (a) Stage 1 shows smooth loss convergence under cosine annealing scheduling, while (b) Stage 2 illustrates accuracy improvement during fine-tuning.

The experiments were conducted on a high-performance computing (HPC) cluster using NVIDIA A100 GPUs. Each training run utilized 8 CPU cores and 48 GiB of system memory. This setup was sufficient to train the scaled ViT-L/14 backbone under the proposed adaptation strategy.

4.1. DetReIDX Dataset

We train and evaluate our model on the DetReIDX stress-test benchmark [4], which is designed to evaluate aerial surveillance person ReID under extreme far-distance degradation and aerial–ground cross-view mismatch. The final benchmark score is reported as the average mAP across the three domains (A2G, G2A and A2A). We use mean Average Precision (mAP) as the primary evaluation metric, following standard ReID protocols. The final score is computed as the average mAP across the three evaluation cases.

$$\mathrm{mAP}_{\mathrm{avg}} = \frac{1}{3}\left(\mathrm{mAP}_{\mathrm{A2G}} + \mathrm{mAP}_{\mathrm{G2A}} + \mathrm{mAP}_{\mathrm{A2A}}\right). \quad (5)$$

4.2. Implementation Details

Backbone and Training Setup. The visual encoder is initialized from the CLIP ViT-L/14 checkpoint pretrained on large-scale image–text data. Training is conducted in two stages. In Stage 1, the model is trained for 50 epochs with a batch size of 48 tracklets and a base learning rate of 2×10^{-4}. Stage 2 performs fine-tuning for 40 epochs using a batch size of 24 and a base learning rate of 1×10^{-4}. Selective fine-tuning is applied by unfreezing only the final two transformer blocks (resblocks.22 and resblocks.23), while earlier layers remain frozen. The unfrozen blocks are optimized with a learning rate scaled to $0.1\times$ the base rate to preserve pretrained representations and stabilize adaptation. Adam is used as the optimizer in both training stages. Figure 3 illustrates the smooth convergence behavior of Stage 1 training loss and Stage 2 training accuracy.

Table 5. Main results on DetReIDX using the official averaged mAP across A2G/G2A/A2A (Eq. 5).

Method	mAP
CLIP ViT-B/16 (Official baseline)	28.11
DetReIDXv1	32.89
Scale-aware adaptation w/ CLIP ViT-L/14 (Ours)	**35.73**

Prompt and Conditioning Configuration. Prompt-based cross-view conditioning is enabled in the final model with a prompt length of 4 and deep prompt insertion across 9 transformer layers. Camera and metadata conditioning-based on discretized altitude, distance, and viewing angle bins, are activated as described in Section 3.1. Adapter modules, the QATW module, and instance normalization are also enabled in the final configuration.

Learning Rate Scheduling and Regularization. Cosine annealing is used in both training stages to provide smooth learning rate decay and stable convergence. Weight decay is set to 1×10^{-4} in Stage 1 and 2.5×10^{-4} in Stage 2.

Data Processing and Augmentation. All frames are resized to 256×128 and normalized. Each tracklet consists of 16 sampled frames. During training, we apply random horizontal flipping ($p = 0.5$), random erasing ($p = 0.5$), color jitter with brightness, contrast, saturation, and hue parameters of $(0.1, 0.1, 0.1, 0.05)$, and padding of 10 pixels. These augmentations improve robustness to noise and enhance generalization under aerial degradation conditions.

Temporal Aggregation and Inference. Temporal attention pooling aggregates frame-level embeddings into tracklet-level representations. A lightweight linear attention module computes adaptive frame weights through softmax normalization along the temporal dimension, enabling the model to emphasize informative frames while suppressing degraded ones. During inference, tracklet embeddings are ℓ_2-normalized, and cosine similarity is used for retrieval.

K-reciprocal re-ranking. To further improve retrieval consistency, we apply k-reciprocal re-ranking as a post-processing step during inference. After computing cosine distances between ℓ_2-normalized query and gallery embeddings, the distance matrix is refined using k-reciprocal encoding. Unless otherwise specified, we set $k_1 = 28$, $k_2 = 6$, and $\lambda = 0.28$. Re-ranking is applied only during evaluation and does not affect model training.

Table 6. Ablation study on DetReIDX using averaged mAP across A2G/G2A/A2A (Eq. 5).

Variant	Backbone	mAP (%)	Δ
CLIP ViT-B/16	ViT-B/16	28.11	+0.00
+ Optimization		28.18	+0.07
+ k-reciprocal re-ranking		29.40	+1.29
+ Re-ranking tuned		29.47	+1.36
+ Temporal attention pooling		29.71	+1.60
+ Re-ranking tuned (final)		29.84	+1.73
+ Backbone scaling + Adaptation	ViT-L/14	35.73	+7.62

4.3. Results

We report results on DetReIDX using the official evaluation metric defined in Eq. (5), namely the averaged mAP across the three protocols (A2G, G2A, and A2A). Since the benchmark ranking is determined by this aggregated score, we focus on mAP_{avg} for all comparisons. Our full model achieves an averaged mAP of 35.73, representing a +7.62 improvement over the official CLIP ViT-B/16 baseline (28.11). Compared to the strongest publicly reported result (32.89), our method yields a further +2.84 gain. These results indicate that backbone scaling combined with stability-focused adaptation substantially enhances robustness under extreme far-distance degradation and aerial–ground cross-view mismatch.

4.4. Ablation Study

We perform ablation experiments to evaluate the contribution of each component in the proposed scale-aware adaptation framework. All results are reported using the official DetReIDX metric, mAP_{avg}, averaged across A2G, G2A, and A2A (Eq. 5). The ablation results indicate that k-reciprocal re-ranking consistently improves performance in the ViT-B/16 setting. Incorporating temporal attention pooling yields additional gains by adaptively down-weighting degraded frames within noisy tracklets. The most substantial improvement comes from scaling the backbone to ViT-L/14, demonstrating that increased representational capacity, when paired with stability-oriented selective fine-tuning, plays a critical role in handling extreme far-distance scale compression.

5. Discussion

Our experiments suggest that extreme far-distance (XFD) video person ReID benefits primarily from increased representational capacity, provided that adaptation remains stable. Scaling the CLIP visual encoder from ViT-B/16 to ViT-L/14 yields the largest improvement in the official DetReIDX averaged score, indicating that additional transformer depth and width help recover discriminative cues even when pedestrian regions are severely scale-compressed and fine-grained appearance details are unreli-

able. Importantly, these gains are realized only when backbone scaling is coupled with stability-oriented fine-tuning. Selectively unfreezing the final transformer blocks and applying a reduced learning rate to the backbone preserves the pretrained feature space while enabling high-level domain adaptation.

Beyond backbone scaling, the ablation study highlights the complementary role of temporal attention and re-ranking. Temporal attention pooling improves robustness by down-weighting degraded or blurred frames within tracklets, mitigating the impact of motion instability and resolution loss during feature aggregation. In contrast, k-reciprocal re-ranking enhances neighborhood consistency in the embedding space at inference time, improving retrieval reliability. Although each component contributes a smaller absolute gain than backbone scaling, together they establish a stronger and more stable foundation for performance improvements under extreme far-distance conditions.

Limitations. This study is evaluated exclusively on the DetReIDX stress-test benchmark using the official averaged metric across the A2G, G2A, and A2A protocols. While this aggregated score enables standardized comparison, it limits detailed analysis of protocol-specific behavior and makes it difficult to determine which domain transfer setting benefits most from individual components. Moreover, the final system employs a larger ViT-L/14 backbone and optional re-ranking during inference, which increases computational cost and latency compared to the baseline configuration.

Future work. Several direction‘s may further advance extreme far-distance ReID. First, multi-granularity temporal modeling such as hierarchical aggregation across short- and long-term temporal windows, could better handle long tracklets with intermittent visibility. Second, more expressive scale-aware prompting or metadata-conditioned adapters may further improve cross-view alignment under substantial altitude and distance variation. Third, evaluating cross-dataset generalization across additional aerial benchmarks would provide stronger evidence of robustness beyond DetReIDX. Finally, given the privacy implications of aerial surveillance, future research should consider responsible deployment practices and privacy-aware evaluation protocols when developing aerial person recognition systems.

6. Conclusion

We presented a scale-aware adaptation framework for CLIP-based video person re-identification in extreme far-distance aerial-ground scenarios. By scaling the backbone from ViT-B/16 to ViT-L/14 and introducing stability-oriented selective fine-tuning, temporal attention pooling, and optimized training strategies, we significantly improve robustness under severe scale compression and frame-level degradation. On the DetReIDX stress-test benchmark, our approach achieves an averaged mAP of 35.73 across the three official protocols, substantially outperforming the baseline. These findings demonstrate that large-scale vision-language backbones, when paired with stable adaptation and adaptive temporal aggregation, provide an effective foundation for extreme far-distance video person ReID.

References

[1] Lin Wu, Yang Wang, Jianhuang Gao, and Xuelong Li. Deep learning for person re-identification: A survey and outlook. *arXiv preprint arXiv:2001.04193*, 2020. 1

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 3, 5

[3] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3

[4] Khushal Hambarde, Emmanuel Mbongo, Nirav Menghani, Ajay Ramesh, Rogerio Schmidt Feris, and Hugo Proença. DetReIDX: A stress-test dataset for real-world UAV-based person recognition. *arXiv preprint arXiv:2505.04793*, 2025. 2, 3, 6

[5] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[6] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision Workshops (ECCV Workshops)*, 2016. 2

[7] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[8] Shuting He, Haoxi Wu, Peng Wang, Mengdan Zhang, Zhongzhan Huang, and Yonghong Tian. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[9] 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)*, 2016. 2

[10] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Spatial-temporal attention model for video-based person re-identification. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 2, 5

[11] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. TCLNet: Temporal complementary learning for video person re-identification. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 5

[12] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. UAV-Human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[13] S. V. Aruna Kumar, Ehsan Yaghoubi, Abhijit Das, B. S. Harish, and Hugo Proença. The P-DESTRE: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. *IEEE Transactions on Information Forensics and Security (TIFS)*, 16:1696–1708, 2021. 3

[14] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Aerial-ground person re-identification. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 2585–2590, 2023. 3

[15] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. *arXiv preprint arXiv:2203.05557*, 2022. 3

[16] 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. 3, 5

[17] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3

[18] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mohammad Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning (ICML)*, 2019. 4, 5

[19] Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 5

[20] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017. 6