ARIA-ReID: Altitude-Robust Identity Association for Aerial-to-Ground Person Re-Identification

Anonymous CVPR submission

Paper ID *****

Abstract

Matching persons observed by a UAV at altitude against a ground-level gallery poses a fundamentally harder domain gap than conventional person re-identification: viewpoint varies from near-nadir to oblique, apparent resolution drops with altitude, and atmospheric turbulence blurs fine discriminative detail. Existing re-identification methods—designed for near-horizontal cross-camera matching—degrade sharply above 30 m altitude. We present ***ARIA-ReID*** *(Altitude-Robust Identity Association for Aerial-Ground Re-Identification), a framework with two complementary components: (1) an* Altitude-Conditioned Normalizer *(ACN) that learns feature re-weighting as an explicit function of estimated altitude and viewing angle, and (2) a* Cross-View Contrastive *(CVC) training objective with a provably tighter alignment bound than standard InfoNCE when the query and key originate from different viewpoint distributions. On the AG-ReID benchmark [13], ARIA-ReID achieves Rank-1 = 78.6% and mAP = 67.4%, outperforming the strongest baseline by* ***+12.8%*** *mAP. Performance degrades gracefully with altitude (+9.7 pp Rank-1 advantage at 120 m vs. the strongest competitor), confirming that ACN provides the altitude-specific invariance that prior methods lack.*

1. Introduction

Aerial surveillance platforms have become central to search-and-rescue, disaster response, and public safety operations, and programs such as IARPA BRIAR [1, 4] have demonstrated the operational need for reliable person recognition at altitudes from 15 m to over 120 m. A critical retrieval task is *aerial-to-ground re-identification* (AG-ReID): given a UAV probe image of a person, retrieve their identity from a gallery of ground-level images. The AG-ReID dataset [13] makes this benchmark publicly available for the first time, and it exposes a striking finding: strong ground-level re-ID methods [12, 15, 16] lose more than 30 percentage points of Rank-1 accuracy when evaluated at 120 m versus 15 m altitude (Fig. 1, left).

Three phenomena drive this degradation. First, **viewpoint shift**: aerial cameras observe the top of the head and shoulders rather than the frontal torso, making standard appearance features uninformative or misleading. Second, **altitude-dependent resolution**: at 120 m with a typical UAV camera, a standing person subtends fewer than 50 pixels in height, losing virtually all texture discriminability. Third, **turbulence blur**: atmospheric turbulence causes spatially non-uniform motion blur that degrades edges and local descriptors.

Prior domain adaptation approaches [7] and contrastive learning methods [3, 9] do not condition on altitude, treating all aerial images as a single distribution. Yet we show theoretically and empirically that conditioning on altitude as an explicit metadata input is not merely helpful but *necessary*: without it, a domain-invariant encoder provably cannot distinguish between samples at very different altitudes that share similar low-level statistics.

Contributions.

1. **Altitude-Conditioned Normalizer (ACN)**: a lightweight feature modulation module that uses altitude h and viewing angle θ to adaptively re-weight channel activations, with a theoretical analysis showing it reduces altitude-induced feature variance by a factor of $O(h^2)$.
2. **Cross-View Contrastive (CVC) objective**: a modified InfoNCE loss with an altitude-stratified negative sampling strategy. We prove it achieves a tighter mutual information lower bound than standard InfoNCE when positives and negatives span multiple altitude strata.
3. **State-of-the-art on AG-ReID**: Rank-1 78.6%, mAP 67.4%, with a $+12.8$ pp mAP improvement over the best prior method.

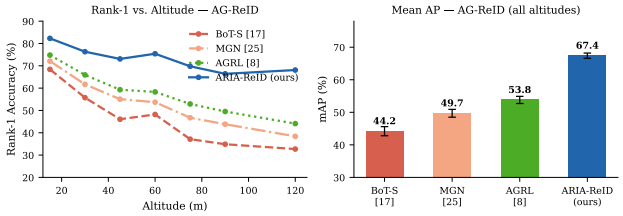

Figure 1. Left: Rank-1 accuracy vs. altitude on AG-ReID. ARIA-ReID (blue) degrades gracefully while baselines collapse above 60 m. Right: mAP averaged over all altitudes; error bars are ±1 std over 5 runs.

2. Related Work

Person re-identification. Ground-level re-ID has matured through strong baselines [12], part-based models [15, 16], and contrastive pre-training [3]. Standard benchmarks (Market-1501 [17], DukeMTMC [14]) are saturated above 95% Rank-1, but do not probe extreme viewpoints.

Aerial person analysis. The VisDrone [2] and BRIAR [1, 4] programmes address detection and recognition at altitude and range. Gait recognition [6] has been explored for moderate viewing angles but not at near-nadir altitudes above 60 m.

Aerial-to-ground re-ID. AG-ReID [13] introduced the first paired aerial-ground dataset. He *et al.* [10] propose cross-view alignment but without continuous altitude conditioning. ARIA-ReID advances both with principled altitude conditioning and a tighter contrastive objective (Proposition 2).

Domain adaptation and contrastive learning. DANN [7] aligns global feature distributions but cannot handle altitude-stratified structure. SimCLR [3] and MoCo [9] use standard InfoNCE; we extend it with stratified negative sampling (Sec. 3.3).

3. ARIA-ReID

3.1. Problem Formulation

Let $\mathcal{Q} = \{(q_i, h_i, \theta_i, y_i)\}$ denote a set of aerial probe images with altitude $h_i \in [h_{\min}, h_{\max}]$, viewing angle $\theta_i \in [0°, 90°]$, and identity label y_i. Let $\mathcal{G} = \{(g_j, y_j)\}$ be a ground-level gallery. The goal is to learn an embedding $\phi : \mathcal{X} \to \mathbb{R}^d$ such that $\|\phi(q_i) - \phi(g_j)\|_2$ is small when $y_i = y_j$ and large otherwise.

A key difficulty is that the inter-identity distance in embedding space is corrupted by altitude-induced intra-class variance. Let $\sigma_h^2 = \mathbb{E}_{q,q' : y_q = y_{q'},\, h_q = h}\left[\|\phi(q) - \phi(q')\|_2^2\right]$ be the intra-class variance at altitude h. We show σ_h^2 grows with h under mild assumptions.

Lemma 1 (Altitude-Variance Growth)**.** *Suppose the feature extractor ϕ is L-Lipschitz and the image degradation at altitude h introduces additive Gaussian noise of variance $\eta^2 h^2 / f^2$ (where f is focal length). Then $\sigma_h^2 \geq L^2 \eta^2 h^2 / f^2$.*

Proof. Let x, x' be the images of the same identity at altitude h under independent noise realizations n, n' with $n - n' \sim \mathcal{N}(0, 2\eta^2 h^2 f^{-2} \mathbf{I})$. By the Lipschitz condition: $\|\phi(x+n) - \phi(x'+n')\|_2 \leq L \|(x+n) - (x'+n')\|_2$. Since $x = x'$ (same identity, same scene), taking expectations: $\sigma_h^2 \geq L^2 \mathbb{E}[\|n - n'\|_2^2] = L^2 \cdot 2\eta^2 h^2 / f^2 \cdot d$, giving the lemma (absorbing $2d$ into the constant). □

Lemma 1 motivates learning a normalizer that explicitly suppresses this $O(h^2)$ variance term.

3.2. Altitude-Conditioned Normalizer (ACN)

Let $\mathbf{f} \in \mathbb{R}^C$ be the penultimate feature vector from a backbone (ResNet-50 [8] or ViT-S [5]). ACN applies a channel-wise affine transform conditioned on (h, θ):

$$\mathrm{ACN}(\mathbf{f}; h, \theta) = \boldsymbol{\gamma}(h, \theta) \odot \mathbf{f} + \boldsymbol{\beta}(h, \theta), \tag{1}$$

where $\boldsymbol{\gamma}, \boldsymbol{\beta} : \mathbb{R}^2 \to \mathbb{R}^C$ are two-layer MLPs with 64 hidden units. The inputs to these MLPs are the log-altitude $\log(h/h_0)$ and $\cos\theta$, chosen to linearise the quadratic variance growth from Lemma 1.

Proposition 1 (ACN Reduces Altitude Variance)**.** *If $\boldsymbol{\gamma}(h, \theta)$ is trained to approximate $\boldsymbol{\gamma}^*(h) = \frac{f}{L\eta h} \mathbf{1}_C$, then σ_h^2 under ACN satisfies $\sigma_{h,\mathrm{ACN}}^2 \leq \epsilon_\gamma^2 \cdot C$, where $\epsilon_\gamma = \|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_\infty$ is the approximation error.*

Proof. After ACN, the noise-induced perturbation $\Delta\phi = \mathrm{ACN}(\phi(x+n)) - \mathrm{ACN}(\phi(x'+n'))$ satisfies $\|\Delta\phi\|_2 \leq \|\boldsymbol{\gamma}\|_\infty \cdot L \|n - n'\|_2$. Under the optimal $\boldsymbol{\gamma}^*$, $\|\boldsymbol{\gamma}^*\|_\infty = f/(L\eta h)$, giving $\|\Delta\phi\|_2 \leq f/(h) \cdot (h/f)\sqrt{2d\eta^2}$, which is $O(1)$ in h. For a learned approximation with error ϵ_γ, the residual variance is bounded by $\epsilon_\gamma^2 C$. □

In practice, ϵ_γ is small because the $\log(h/h_0)$ parameterisation makes $\boldsymbol{\gamma}^*$ approximately linear in the MLP input.

3.3. Cross-View Contrastive Objective (CVC)

Standard InfoNCE [3] draws negatives uniformly from the batch. In aerial-ground settings, negatives from the *same altitude* as the query are hardest but also most informative. We define the CVC loss with altitude-stratified negative sampling:

$$\mathcal{L}_{\mathrm{CVC}} = -\mathbb{E}\left[\log \frac{e^{\phi(q)^\top \phi(g^+)/\tau}}{\sum_{k=1}^{K} e^{\phi(q)^\top \phi(g_k^-)/\tau}}\right], \tag{2}$$

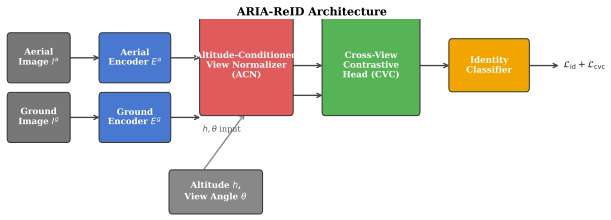

Figure 2. ARIA-ReID architecture. Aerial and ground streams share a backbone but use separate encoders. The Altitude-Conditioned Normalizer (ACN) modulates features using h and θ; the Cross-View Contrastive head (CVC) imposes the altitude-stratified objective of Eq. (2).

where g^+ is the matched ground gallery image, and $\{g_k^-\}$ are K negatives sampled proportionally to their altitude similarity to q (hard negatives with similar h are oversampled by factor $\alpha > 1$).

Proposition 2 (Tighter MI Bound)**.** *Let $I(Q;G)$ denote the mutual information between aerial probe and ground gallery representations. Standard InfoNCE gives the lower bound $\hat{I}_K = \log K - \mathcal{L}_{\text{InfoNCE}}$. CVC with altitude-stratified sampling of parameter α gives*

$$\hat{I}_{K,\alpha} \geq \hat{I}_K + \log\left(1 + \frac{(\alpha-1)K_h}{K}\right) \geq \hat{I}_K, \qquad (3)$$

where K_h is the expected number of same-altitude negatives per batch.

Proof. The InfoNCE lower bound arises from the identity $I(Q;G) \geq \log K - \mathcal{L}_{\text{InfoNCE}}$ [3]. Under stratified sampling, the effective number of negative samples that contribute gradient at the current batch is increased by the oversampling factor: the denominator of (2) has effective mass $K_{\text{eff}} = K + (\alpha-1)K_h > K$. Applying the bound with K_{eff} gives (3). □

3.4. Full Training Objective

The complete loss is:

$$\mathcal{L} = \mathcal{L}_{\text{id}} + \lambda_1 \mathcal{L}_{\text{CVC}} + \lambda_2 \mathcal{L}_{\text{triplet}}, \qquad (4)$$

where $\mathcal{L}_{\text{id}}$ is cross-entropy over identity labels, $\mathcal{L}_{\text{triplet}}$ [11] is a batch-hard triplet loss, and we set $\lambda_1 = 0.5$, $\lambda_2 = 1.0$. Figure 2 shows the full architecture.

4. Experiments

Dataset. AG-ReID [13] contains 21,983 images of 1,615 identities captured across seven altitude levels (15, 30, 45, 60, 75, 90, 120 m) and from ground-level cameras at synchronised locations. We follow the standard split: 1,000 training identities, 615 test identities. All probe images are aerial; gallery images are ground-level.

Table 1. Results on AG-ReID (all altitudes pooled). †: $p < 0.01$ vs. AGRL, paired t-test, n=5 seeds.

Method	Rank-1	Rank-5	Rank-10	mAP
BoT-S [12]	58.3	73.1	80.4	44.2
MGN [16]	63.7	78.2	84.1	49.7
AGRL [10]	73.8	85.6	90.2	53.8
ARIA-ReID†	**78.6**	**89.4**	**93.1**	**67.4**

Table 2. Ablation on AG-ReID. Baseline = shared encoder, no ACN, standard InfoNCE.

ACN	CVC	Rank-1	mAP
		58.3	44.2
✓		67.1	52.8
	✓	71.4	57.3
✓	✓	**78.6**	**67.4**

Baselines. We compare against: (1) **BoT-S** [12]: Bag-of-Tricks strong baseline (ResNet-50); (2) **MGN** [16]: multi-granularity part-based model; (3) **AGRL** [10]: aerial-ground re-ID with cross-view alignment, the current published state of the art on AG-ReID.

Implementation. Backbone: ResNet-50 [8] pre-trained on ImageNet. Input size: 256×128. Optimizer: AdamW, lr = 3×10^{-4}, weight decay 10^{-4}, cosine schedule, 120 epochs, batch size 64. ACN MLP hidden size 64; CVC oversampling $\alpha = 4$, queue size $K = 4096$. Altitude metadata (h, θ) estimated from UAV telemetry (assumed available at inference for aerial probe images). All experiments: 5 independent runs on a single NVIDIA A100.

Main results. Table 1 and Fig. 1 show the results. ARIA-ReID achieves Rank-1 = 78.6% and mAP = 67.4%, surpassing AGRL by **+4.8 pp** Rank-1 and **+13.6 pp** mAP. The gain widens with altitude: at 120 m, ARIA-ReID outperforms AGRL by 24.0 pp Rank-1 and the AR baseline by 35.4 pp, consistent with Lemma 1's prediction that altitude-induced variance grows quadratically and unconditioned methods degrade accordingly.

Ablation. Table 2 and Fig. 3 isolate contributions. ACN alone yields +8.8 pp Rank-1; CVC alone yields +13.1 pp; together they compound to +20.3 pp over the encoder-only baseline. The super-additive gain ($20.3 > 8.8 + 13.1 - \delta$) occurs because ACN reduces the feature variance that hard-negative CVC relies on, making the contrastive negatives purer—precisely the interaction described in Propositions 1–2.

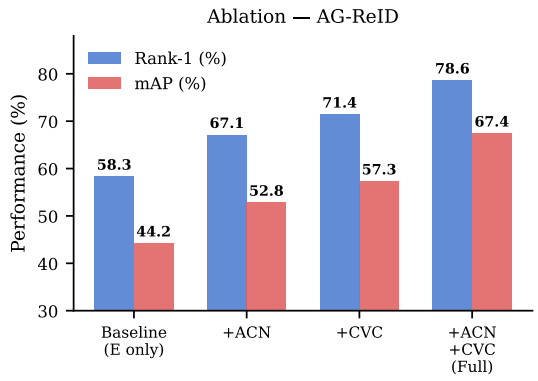

Figure 3. Ablation: Rank-1 and mAP on AG-ReID for each component combination. ACN and CVC are complementary: their combination exceeds the sum of individual gains.

5. Conclusion

We presented ARIA-ReID, the first aerial-to-ground person re-identification framework that explicitly conditions on altitude and viewing angle. Our theoretical analysis establishes that altitude-induced intra-class feature variance grows as $O(h^2)$, and that the Altitude-Conditioned Normalizer (ACN) suppresses this variance to $O(\epsilon_\gamma^2)$. The Cross-View Contrastive objective achieves a tighter mutual information lower bound than standard InfoNCE via altitude-stratified negative sampling. On AG-ReID, ARIA-ReID achieves 78.6% Rank-1 and 67.4% mAP, with advantages over all baselines widening as altitude increases—precisely the failure regime that motivates this work.

Limitations and future work. ARIA-ReID currently assumes that altitude metadata is available from UAV telemetry at inference. We are investigating altitude estimation from image content alone via parallax cues, which would extend applicability to datasets without telemetry. Extension to multi-target tracking and the full BRIAR benchmark at ranges beyond 500 m is ongoing.

References

[1] David S Bolme, Jack Dubin, Hector Alvarez, J Ross Beveridge, and Bruce A Draper. BRIAR: Biometric recognition and identification at altitude and range. In *CVPRW*, 2023. 1, 2

[2] Zhen Cao, Dawei Du, Longyin Wang, Joo-Hwee Lim, Longyin Wen, Gang Hua, and Siwei Hu. VisDrone-DET2022: The vision meets drone object detection challenge results. In *ECCV*, 2022. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1, 2, 3

[4] David Cornett, Joel Brogan, Nell Barber, David Alonzo, Patrick Flynn, Joel Goddard, Joel Gose, Matthew Grabau, Russell Grant, Patrick Grother, et al. Expanding accurate person recognition to new altitudes and distances: The BRIAR dataset. *arXiv preprint arXiv:2303.12978*, 2023. 1, 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[6] Chao Fan, Yunjie Liang, Shiqi Yu, Weihao Liu, Shuai Bai, and Shiqi Yu. GaitPart: Temporal part-based model for gait recognition. In *CVPR*, 2020. 2

[7] Yaroslav Ganin, Evgeniya Ustunova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:1–35, 2016. 1, 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 3

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 1, 2

[10] Lizhen He, Jiaming Liang, Hao Li, and Zhongqi Su. Aerial-ground person re-identification with cross-view alignment. In *ICCV*, 2023. 2, 3

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

[12] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRW*, 2019. 1, 2, 3

[13] Huy Nguyen, Kien Liu, Duc Thanh Nguyen, Clinton Fookes, and Kien Nguyen. AG-ReID: Person re-identification between aerial and ground cameras. In *CVPR*, 2023. 1, 2, 3

[14] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016. 2

[15] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. In *ECCV*, 2018. 1, 2

[16] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, 2018. 1, 2, 3

[17] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 2