TAFD: TEXT-ANCHORED FEATURE DEFORMATION FOR AERIAL–GROUND PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Aerial-ground person re-identification requires matching identities across substantial changes in viewpoint and visual appearance. A central challenge is to establish cross-view correspondence while retaining the view-specific cues needed for identity discrimination. Current methods typically utilize direct feature alignment strategies, which do not explicitly model how visual evidence should be reorganized across viewpoints. Therefore, we propose Text-Anchored Feature Deformation (TAFD), which formulates cross-view matching through language-conditioned directional transformations in feature space. A Semantic View Anchoring (SVA) module learns aerial and ground semantic anchors using shared and view-private text prompts encoded by a frozen text encoder. Conditioned on the difference between these anchors, a Directional Feature Deformation (DFD) module employs two independent operators, one for aerial-to-ground transformation and the other for ground-to-aerial transformation, to predict spatial offsets, resample multi-level patch features, and aggregate target-view evidence. To preserve identity during transformation, a Target-Space Identity Consistency (TIC) objective supervises transformed descriptors using target-view classifiers and batch identity prototypes constructed from stop-gradient target features, while native features retain view-specific classification and metric supervision. Experiments on CARGO and AGReIDv1 demonstrate the effectiveness of TAFD for aerial-ground person retrieval. On CARGO, TAFD achieves 72.59% mAP and 76.60% Rank-1 for aerial-to-ground retrieval. On AGReIDv1, the same architecture achieves average bidirectional performance of 80.79% mAP and 87.51% Rank-1.

1 INTRODUCTION

Person re-identification (ReID) matches a person across cameras with disjoint fields of view and has been studied extensively in ground-camera networks (He et al., 2021; Li et al., 2023). The growing use of unmanned aerial vehicles extends this task to aerial-ground camera networks (Nguyen et al., 2023; Zhang et al., 2024). In this setting, a top-down image and a ground-view image may reveal different parts of the same person. The former emphasizes the head, shoulders, and body outline, whereas the latter often preserves clothing and frontal or lateral appearance. Therefore, cross-view retrieval must bridge these differences, while aerial-aerial and ground-ground retrieval must still retain the details that distinguish identities within each view.

Existing aerial-ground ReID methods have mainly pursued view-invariant or view-aware representations. The View-Decoupled Transformer (Zhang et al., 2024) separates view-related and view-unrelated information, while ViSA aligns local evidence through view-aware semantic queries (Zhang et al., 2026). These methods establish the value of semantic alignment, yet leave a fundamental tension: pressing aerial and ground images into a shared identity space may suppress cues that are useful for aerial-aerial (AA) or ground-ground (GG) retrieval. GSAlign (Li et al., 2025) offers a more direct structural approach. It progressively deforms visual patch features with a learnable thin-plate-spline module and filters them with visibility-aware masks. However, correcting a feature is not the same as specifying where it should go. As shown in Fig. 1(a), GSAlign does not explicitly require a corrected aerial feature to match the identity representation learned for the ground view, or vice versa. Because both the backbone and correction module receive identity gradients, either can absorb the error. The intended direction of correction therefore remains underconstrained, resulting in suboptimal performance. Together, these observations suggest two coupled requirements: preserving

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

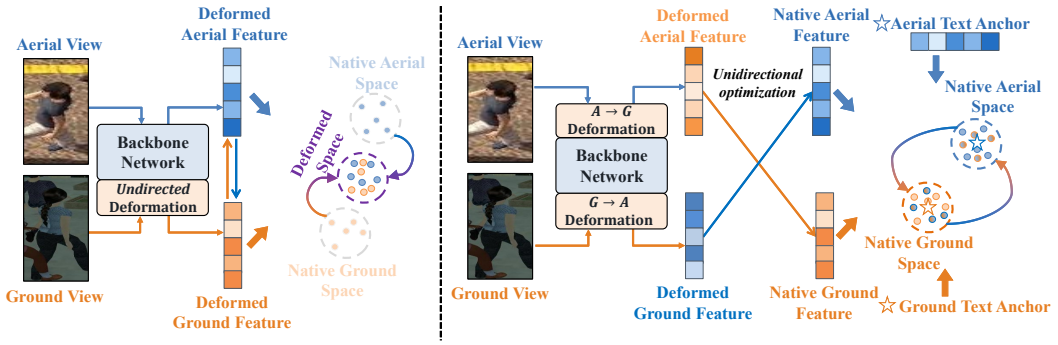

Figure 1: Motivation for text-anchored feature deformation. (a) Target-agnostic correlation provides no explicit destination identity space for deformed features, which may blur the native aerial and ground feature structures. (b) TAFD uses aerial and ground text anchors to guide two separate transformations, A→G and G→A. Each transformed feature is optimized in its target-view identity space, while native features retain view-specific supervision. Blue and orange denote aerial and ground feature spaces, respectively, while stars denote text anchors.

native-view identity evidence and giving each structural conversion an explicit target-view identity objective.

To achieve this goal, as shown in Fig. 1(b), we propose Text-Anchored Feature Deformation (TAFD). Rather than forcing one embedding to serve every retrieval relation, TAFD keeps a native feature for the observed view and constructs a separate converted feature for the opposite view. Native aerial and ground features are trained with their respective identity-classification and triplet objectives. Two independent deformers perform A→G and G→A conversion. Since the visible evidence changes asymmetrically, these transformations need not be inverses of each other. At inference, each image retains both native-view and converted-view information, allowing cross-view alignment without replacing the representation used for within-view retrieval.

TAFD uses one shared CLIP visual encoder with view-specific low-rank adapters. Semantic View Anchoring (SVA) combines shared and view-private learnable text prompts with a frozen text encoder to form aerial and ground semantic anchors. Conditioned on the difference between these anchors, Directional Feature Deformation (DFD) uses independent A→G and G→A operators to predict multi-level patch-grid offsets, resample visual features, and aggregate target-view evidence. Crucially, Target-Space Identity Consistency (TIC) supervises each converted feature through its destination-view classifier and batch identity prototypes, with stopped gradients on the target side. Native aerial and ground descriptors continue to receive view-specific classification and metric supervision. Thus, text specifies the destination, TIC makes the converted feature identity-discriminative in that destination, and native-view supervision protects AA and GG retrieval.

Our main contributions are summarized as follows:

- We propose Text-Anchored Feature Deformation (TAFD), which preserves the native identity representation of each view while learning an additional, direction-specific representation toward the opposite view. This design avoids forcing heterogeneous views into a single shared representation and enables one model to accommodate both homogeneous- and heterogeneous-view retrieval.
- We instantiate TAFD with three complementary designs. Semantic View Anchoring (SVA) encodes the destination-view semantics with learnable textual anchors, Directional Feature Deformation (DFD) performs asymmetric A→G and G→A feature transformations through multi-level spatial deformation, and Target-Space Identity Consistency (TIC) explicitly constrains each transformed feature against identity evidence in its target view. Together, they provide both a clear transformation direction and identity-preserving supervision for cross-view feature conversion.
- Extensive experiments on CARGO and AGReIDv1 demonstrate the superior performance of TAFD. On CARGO, TAFD surpasses the previous state of the art by +3.59% mAP and

+5.32% Rank-1 under the challenging A→G protocol, while further improving mAP by **+2.50%** under the ALL protocol.

2 Related Work

Person re-identification (ReID) has been studied under increasingly diverse acquisition conditions. In conventional camera networks, advances in part-based representation learning and transformer backbones have substantially improved identity discrimination (Sun et al., 2019; He et al., 2021). Prior studies have also addressed incomplete observations in occluded ReID (Wang et al., 2020; Li et al., 2021b), language-image correspondence in text-to-image ReID (Shao et al., 2023; Qin et al., 2024), and appearance variations across different sensors or spectra (Wang et al., 2022; Kim et al., 2023; Ren & Zhang, 2024). Vision-language pre-training has further been introduced into image-based ReID to provide semantic supervision beyond visual identity labels (Li et al., 2023). Despite these advances, extending ReID from ground-camera networks to aerial platforms remains challenging because the same person can exhibit substantially different visible body regions and spatial layouts under different viewpoints.

Early aerial ReID benchmarks, including UAV-Human (Li et al., 2021a) and PRAI-1581 (Zhang et al., 2020), mainly focus on retrieval among UAV-captured observations. Aerial-ground ReID further connects UAV imagery with ground-camera networks and requires identity matching under simultaneous viewpoint and visibility changes (Nguyen et al., 2023). Subsequent methods have explored different strategies to reduce this cross-view gap. The View-Decoupled Transformer (VDT) (Zhang et al., 2024) separates view-related and identity-related information and introduces the CARGO benchmark. DTST (Wang & Pishgar, 2024b) selects informative visual tokens to improve cross-view representation, while view-invariant feature learning (Khalid et al., 2025) reduces sensitivity to viewpoint variations. Other approaches incorporate semantic priors into the alignment process. SeCap (Wang et al., 2025a) adapts prompts according to the input view, LATex (Zhang et al., 2025) exploits textual attributes to enhance identity representation, and ViSA (Zhang et al., 2026) aligns view-aware local semantics. SD-ReID (Wang et al., 2025b) follows a generative strategy and employs diffusion models to produce complementary view-specific features. These methods demonstrate the importance of both cross-view correspondence and view-specific identity information for aerial-ground ReID. GSAlign (Li et al., 2025) is particularly relevant to our work because it directly addresses spatial misalignment between aerial and ground observations. Its learnable thin-plate-spline module progressively warps visual patch features toward a canonical geometry, while dynamic visibility masks refine semantic alignment. Although the deformation is regularized and optimized with identity supervision, the corrected feature is not explicitly assigned to a target-view identity space. As a result, the transformation direction remains underconstrained, which may lead to suboptimal feature correction. Our Text-Anchored Feature Deformation (TAFD) instead introduces an explicit target for each cross-view transformation. Semantic View Anchoring (SVA) defines aerial and ground semantic targets through shared and view-private text prompts, while Directional Feature Deformation (DFD) learns separate aerial-to-ground and ground-to-aerial transformations conditioned on these targets. Target-Space Identity Consistency (TIC) further supervises each transformed descriptor using identity evidence from its designated target view, while native descriptors retain view-specific classification and metric supervision. In this way, TAFD learns directed cross-view correspondence without forcing native aerial and ground features to discard their own discriminative structures.

3 Method

3.1 Overview

As illustrated in Fig. 2, TAFD learns two representations for each image, namely a native representation for its observed view and a deformed representation for the opposite view. The framework consists of Semantic View Anchoring (SVA), Directional Feature Deformation (DFD), and Target-Space Identity Consistency (TIC). SVA provides aerial and ground text anchors that specify the destination of feature deformation. DFD learns separate aerial-to-ground and ground-to-aerial transformations conditioned on these anchors. TIC then constrains the transformed feature using identity evidence from its target view, while native features are trained for within-view discrimination.

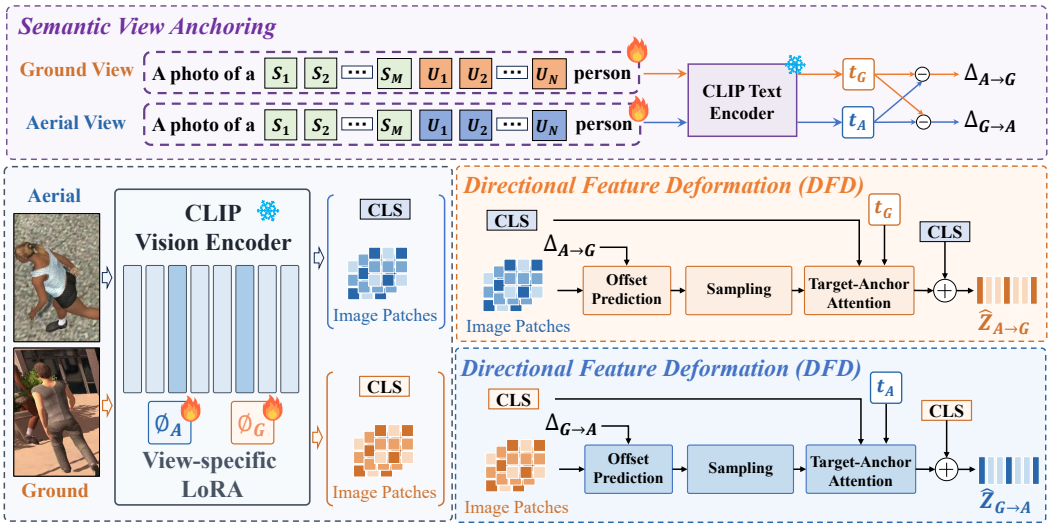

Figure 2: **Overview of Text-Anchored Feature Deformation (TAFD).** Semantic View Anchoring (SVA) encodes shared and view-private prompts into aerial and ground text anchors. A shared CLIP visual encoder with view-specific LoRA extracts native descriptors and multi-level visual tokens. For each transformation direction, an independent Directional Feature Deformation (DFD) operator uses the anchor difference to predict sampling offsets, resamples patch features, and aggregates them with class tokens under the guidance of the target-view anchor. The aggregated response is fused with the native descriptor to obtain a target-view feature.

Let A and G denote the aerial and ground views, respectively. Given an image x from view $v \in \{A, G\}$, a CLIP visual encoder $\mathcal{B}_{\theta,\phi_v}$ extracts a native descriptor and visual tokens. The backbone parameters θ are shared across views, while view-specific low-rank parameters ϕ_v adapt the encoder to aerial or ground inputs. If the encoder has L layers, we may select any nonempty set $\mathcal{S} \subseteq \{1, \ldots, L\}$. We collect the projected class token $\mathbf{c}^\ell$ and patch tokens $\mathbf{P}^\ell$ from each selected layer and use the final projected class token as the native descriptor $\mathbf{z}$:

$$\big(\mathbf{z}, \{\mathbf{c}^\ell, \mathbf{P}^\ell\}_{\ell \in \mathcal{S}}\big) = \mathcal{B}_{\theta,\phi_v}(x), \qquad \mathbf{z}, \mathbf{c}^\ell \in \mathbb{R}^D, \quad \mathbf{P}^\ell \in \mathbb{R}^{H_p W_p \times D}. \tag{1}$$

Here, D is the feature dimension and $H_p \times W_p$ is the patch-grid size. The choice of $\mathcal{S}$ determines the feature levels available to DFD without changing the single-pass visual extraction.

3.2 Semantic View Anchoring

The central question in directional deformation is where a feature should be mapped. To provide an explicit destination, SVA constructs an aerial text anchor and a ground text anchor. Each prompt combines shared learnable context tokens $\mathbf{S}$ and view-private tokens $\mathbf{U}_v$:

$$\begin{aligned} \mathcal{P}_v &= [\mathbf{w}_{\text{pre}}, \mathbf{S}, \mathbf{U}_v, \mathbf{w}_{\text{post}}], \\ \mathbf{t}_v &= E_T(\mathcal{P}_v), \qquad v \in \{A, G\}. \end{aligned} \tag{2}$$

Here, $\mathbf{w}_{\text{pre}}$ and $\mathbf{w}_{\text{post}}$ denote fixed template words, and E_T is the frozen CLIP text encoder. The shared tokens model semantics common to both views, whereas the private tokens allow each anchor to represent view-specific characteristics. Since the prompts are shared by all images, $\mathbf{t}_A$ and $\mathbf{t}_G$ describe views rather than individual identities. For source view s and target view $t \neq s$, we define $\mathbf{\Delta}_{s\to t} = \mathbf{t}_t - \mathbf{t}_s$. The two directions are given by

$$\mathbf{\Delta}_{A\to G} = \mathbf{t}_G - \mathbf{t}_A, \qquad \mathbf{\Delta}_{G\to A} = \mathbf{t}_A - \mathbf{t}_G. \tag{3}$$

The two differences specify opposite destinations in the same visual feature space. Although E_T remains frozen, gradients from the retrieval objectives update the prompt tokens and thereby adapt the view anchors to the task.

3.3 Directional Feature Deformation

The viewpoint difference between aerial and ground cameras changes the spatial arrangement of visible body regions. We therefore perform deformation on multi-level patch features. DFD contains

two independent operators, $D_{A\to G}$ and $D_{G\to A}$, rather than forcing a single operator to model both directions. The anchor difference in Eq. 3 specifies the transformation direction, while the patch tokens provide image-specific evidence. Let $\mathbf{p}_{\ell,u} \in \mathbb{R}^D$ be the u-th patch token in $\mathbf{P}^\ell$. At location u of the selected layer ℓ, the corresponding operator predicts a two-dimensional offset from this token and $\mathbf{\Delta}_{s\to t}$:

$$\boldsymbol{\delta}_{\ell,u}^{s\to t} = \delta_{\max} \tanh(W_2^{s\to t}\, \sigma(W_1^{s\to t}[\mathrm{LN}(\mathbf{p}_{\ell,u}); \mathrm{LN}(\mathbf{\Delta}_{s\to t})])), \tag{4}$$

where $[\cdot;\cdot]$ denotes the concatenation operation, $W_1^{s\to t}$ and $W_2^{s\to t}$ are learnable affine maps, LN denotes layer normalization, and σ is a nonlinear activation. The parameter $\delta_{\max}$ bounds the displacement in patch-grid units. Although the two anchor differences have opposite signs, the independently parameterized operators do not impose inverse offset fields. Let $\mathbf{g}_u$ be the center of patch u in the same coordinate system. The deformed feature is obtained through bilinear sampling:

$$\widetilde{\mathbf{p}}_{\ell,u}^{s\to t} = \mathrm{Sample}_{\mathrm{bilinear}}\big(\mathbf{P}^\ell, \mathbf{g}_u + \boldsymbol{\delta}_{\ell,u}^{s\to t}\big). \tag{5}$$

The offsets from Eq. 4 are converted into the sampling coordinates of the patch grid. Thus, DFD reorganizes the extracted visual features without warping the input image.

The sampled patches capture local evidence after spatial rearrangement. To integrate them with the global representation, we use the target-view anchor to attend to both the selected class tokens and the deformed patch tokens. We denote the class tokens concatenated in layer order by $\mathbf{C} = \mathrm{Concat}_{\ell\in\mathcal{S}}\, \mathbf{c}^\ell$ and the concatenated deformed patches by $\widetilde{\mathbf{P}}^{s\to t}$. We then apply direction-specific attention $\mathrm{Attn}_{s\to t}(q, K, V)$ with query q, keys K, and values V. The target-view descriptor is computed as

$$\begin{aligned} \mathbf{h}_{\mathrm{cls}}^{s\to t} &= \mathrm{Attn}_{s\to t}(\mathbf{t}_t, \mathbf{C}, \mathbf{C}), \\ \mathbf{h}_{\mathrm{patch}}^{s\to t} &= \mathrm{Attn}_{s\to t}(\mathbf{t}_t, \widetilde{\mathbf{P}}^{s\to t}, \widetilde{\mathbf{P}}^{s\to t}), \\ \hat{\mathbf{z}}^{s\to t} &= \mathrm{LN}_{s\to t}(\mathbf{z} + \mathbf{h}_{\mathrm{cls}}^{s\to t} + \lambda_{\mathrm{loc}}\mathbf{h}_{\mathrm{patch}}^{s\to t}). \end{aligned} \tag{6}$$

Here, $\mathbf{h}_{\mathrm{cls}}^{s\to t}$ and $\mathbf{h}_{\mathrm{patch}}^{s\to t}$ are the global and deformed local responses to the target anchor. The same attention module aggregates both responses within each direction, while the two directions use separate offset predictors, attention parameters, and output layer normalizations $\mathrm{LN}_{s\to t}$. The λ_{loc} controls the contribution of deformed local evidence. Thus, the anchor difference conditions spatial sampling in Eq. 5, whereas the target anchor conditions feature aggregation in Eq. 6.

3.4 Target-Space Identity Consistency and Optimization

Learning a cross-view transformation alone does not determine whether the resulting feature retains the correct identity in the target view. TIC addresses this issue by evaluating each transformed feature against target-view identity references. Meanwhile, native aerial and ground features receive separate identity classification and triplet objectives. Classification separates training identities within each observed view, while the triplet objective maintains their feature-level ordering. We denote the sum of these native-view terms by $\mathcal{L}_{\mathrm{native}}$.

For a training sample i, let y_i and v_i denote its identity label and observed view, respectively, and let $\mathcal{M}_v = \{i : v_i = v\}$ contain the minibatch indices from view v. We first use the target-view identity head C_t, including its neck and classifier, to supervise a transformed source feature. The classification loss is defined as:

$$\mathcal{L}_{\mathrm{cls}}^{s\to t} = \frac{1}{|\mathcal{M}_s|} \sum_{i\in\mathcal{M}_s} \mathrm{CE}(C_t^{\mathrm{sg}}(\hat{\mathbf{z}}_i^{s\to t}), y_i). \tag{7}$$

Here, CE denotes cross-entropy and sg denotes stop-gradient. The superscript on C_t indicates that its parameters and batch-normalization statistics are held fixed for this loss. Thus, an $s \to t$ feature must be recognized as identity y_i by the target-view head, while the head itself continues to be learned from native target-view images.

Classification specifies the target identity but does not directly compare a transformed feature with target-view instances. We therefore use a directional triplet objective. Let d denote the feature-space distance after the fixed target-view neck. For source sample i, d_i^+ is its largest distance to native target-view features with identity y_i, and d_i^- is its smallest distance to native target-view features of

other identities. Let $\mathcal{V}_{s\to t} \subseteq \mathcal{M}_s$ contain samples for which both types of target-view examples are available. The triplet loss is defined as:

$$\mathcal{L}_{\text{tri}}^{s\to t} = \frac{1}{|\mathcal{V}_{s\to t}|} \sum_{i\in\mathcal{V}_{s\to t}} \log(1 + \exp(d_i^+ - d_i^-)), \tag{8}$$

which reduces the distance to the hardest target-view positive relative to the nearest negative. Native target features and the target-view neck are detached in this consistency path, so this ranking signal acts only on the transformed source feature.

In addition to instance-level ranking, TIC aligns each transformed feature with the center of its target-view identity. Let $\mathcal{M}_{t,c} = \{j \in \mathcal{M}_t : y_j = c\}$ denote the target-view samples of identity c. Their native descriptors form a minibatch prototype defined as:

$$\boldsymbol{\mu}_c^t = \text{sg}(\frac{1}{|\mathcal{M}_{t,c}|} \sum_{j\in\mathcal{M}_{t,c}} \mathbf{z}_j), \tag{9}$$

where the stop-gradient operation makes the prototype a target-space reference in TIC. Let $\mathcal{I}_t = \{c : |\mathcal{M}_{t,c}| > 0\}$ denote the identities represented in the target view of the minibatch, and let $\cos(\cdot,\cdot)$ denote cosine similarity. Using the target prototypes in Eq. 9, we impose a positive-prototype cosine loss and a prototype contrastive loss as follows:

$$\begin{aligned}
\mathcal{L}_{\cos}^{s\to t} &= \frac{1}{|\mathcal{V}_{s\to t}|} \sum_{i\in\mathcal{V}_{s\to t}} \left[1 - \cos(\hat{\mathbf{z}}_i^{s\to t}, \boldsymbol{\mu}_{y_i}^t)\right], \\
\mathcal{L}_{\text{pNCE}}^{s\to t} &= -\frac{1}{|\mathcal{V}_{s\to t}|} \sum_{i\in\mathcal{V}_{s\to t}} \log \frac{\exp(\cos(\hat{\mathbf{z}}_i^{s\to t}, \boldsymbol{\mu}_{y_i}^t)/\tau)}{\sum_{c\in\mathcal{I}_t} \exp(\cos(\hat{\mathbf{z}}_i^{s\to t}, \boldsymbol{\mu}_c^t)/\tau)}.
\end{aligned} \tag{10}$$

The cosine term pulls the transformed feature toward the native target-view center of the same identity. By itself, this positive-only constraint does not separate the feature from other identities. The prototype contrastive term uses the same center as its positive and the centers of all other target-view identities as negatives. Each identity contributes one prototype to the denominator regardless of the number of its images in the batch, and $\tau > 0$ is the temperature that controls the sharpness of this comparison. Both terms update the transformed feature through detached target prototypes. Training batches contain both views for each sampled identity, and the corresponding loss term is omitted if the required positive or negative target examples are absent. Combining native supervision with the two directional TIC objectives gives as:

$$\mathcal{L} = \mathcal{L}_{\text{native}} + \lambda_{\text{target}} \sum_{s\to t\in\{A\to G, G\to A\}} \left(\mathcal{L}_{\text{cls}}^{s\to t} + \mathcal{L}_{\text{tri}}^{s\to t} + \mathcal{L}_{\cos}^{s\to t} + \lambda_{\text{NCE}}\mathcal{L}_{\text{pNCE}}^{s\to t}\right), \tag{11}$$

where λ_{target} controls the strength of target-space identity consistency relative to native-view supervision, while λ_{NCE} balances prototype contrastive learning within TIC. The CLIP text encoder is frozen, while target-side classifiers and native target features are detached only along the TIC paths.

Inference. During inference, every image is represented by an aerial-space feature and a ground-space feature. We combine the ground- and aerial-space features in Eq. 12 for retrieval. The observed view supplies the native feature, while DFD produces the opposite-view feature:

$$\mathbf{f}(x) = \begin{cases} [\text{Norm}(\mathbf{z}); \text{Norm}(\hat{\mathbf{z}}^{A\to G})], & v = A, \\ [\text{Norm}(\hat{\mathbf{z}}^{G\to A}); \text{Norm}(\mathbf{z})], & v = G. \end{cases} \tag{12}$$

Here, Norm denotes L_2 normalization.

4 Experiments

4.1 Datasets and Evaluation Protocols

We evaluate TAFD on two aerial-ground person ReID benchmarks, CARGO (Zhang et al., 2024) and AGReIDv1 (Nguyen et al., 2023). **CARGO** contains images captured by 13 aerial and ground

Table 1: **Performance comparison with state-of-the-art methods on CARGO under four evaluation protocols.** The best performance is shown in **bold**.

Method	Venue	Protocol 1: ALL		Protocol 2: G→G		Protocol 3: A→A		Protocol 4: A→G	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
SBS	ACM MM'23	43.09	50.32	62.99	72.31	49.73	67.50	29.00	31.25
PCB	TPAMI'21	44.50	51.00	67.60	74.10	44.60	55.00	30.40	34.40
BoT	CVPRW'19	46.49	54.81	66.47	77.68	49.79	65.00	32.56	36.25
MGN	ACM MM'18	49.08	54.81	71.05	83.93	52.96	65.00	33.47	31.87
VV	JAISCR'20	38.84	45.83	62.99	72.31	49.73	67.50	29.00	31.25
AGW	TPAMI'21	53.44	60.26	71.66	81.25	56.48	67.50	40.90	43.57
BAU	NeurIPS'24	38.40	45.20	51.20	61.60	42.60	50.00	36.70	40.40
PAT	ICCV'23	15.30	37.90	24.20	52.70	23.10	50.00	15.50	35.10
DTST	ICME'25	55.73	64.42	72.40	78.57	63.31	80.00	43.49	50.53
ViT	ICLR'21	53.54	61.54	71.34	82.14	64.47	80.00	40.11	43.13
VDT	CVPR'24	55.20	64.10	71.59	82.14	66.83	82.50	42.76	48.12
GSAlign	NeurIPS'25	57.95	65.06	73.86	83.04	65.55	80.00	61.55	64.89
LATex	arXiv'25	67.09	**76.96**	79.90	**90.18**	69.06	80.00	58.88	66.87
ViSA	CVPR'26	65.46	70.51	**83.90**	88.39	67.78	**82.50**	69.00	71.28
TAFD	-	**69.59**	74.68	83.57	89.29	**71.81**	80.00	**72.59**	**76.60**

cameras. Following the standard split, we use 2,500 identities (51,451 images) for training and the remaining 2,500 identities for testing. Following ViSA (Zhang et al., 2026), we evaluate under the overall (ALL), aerial-to-ground (A→G), ground-to-aerial (G→A), aerial-to-aerial (A→A), and ground-to-ground (G→G) protocols. **AGReIDv1** contains 199 training identities with 11,554 images and 189 test identities. Following its standard protocol, we evaluate bidirectional cross-view retrieval under the A→G and G→A settings. For both datasets, performance is evaluated using Cumulative Matching Characteristic (CMC) at Rank-1 and mean Average Precision (mAP).

4.2 Implementation Details

We adopt CLIP ViT-B/16 as the visual backbone with a frozen text encoder. The aerial and ground views share the backbone while using separate rank-16 low-rank updates. Input images are resized to 256×128, and features from layers $\mathcal{S} = \{5, 8, 12\}$ are used for deformation. The projected feature dimension is 512. Each view prompt contains 4 shared and 2 view-specific context tokens. For DFD, the displacement range is limited to two patch units, and the final offset layer is zero-initialized. We set $\lambda_{\text{loc}} = 0.25$, $\lambda_{\text{target}} = 1.0$, and $\lambda_{\text{NCE}} = 0.5$, with temperature $\tau = 0.07$. We train all models for 60 epochs using Adam with a batch size of 64. Each batch contains 8 identities, with 4 aerial and 4 ground images sampled for each identity. The learning rates are 3.5×10^{-4} for newly initialized modules, 5×10^{-6} for pretrained visual parameters, and 10^{-4} for low-rank parameters. All learning rates are decayed by a factor of 0.1 at epochs 20 and 40. We employ random horizontal flipping, random cropping, and random erasing for data augmentation.

4.3 Comparison with State-of-the-Art Methods

We compare TAFD with three groups of representative methods. The first group consists of conventional ReID approaches, including SBS (He et al., 2023), PCB (Sun et al., 2019), BoT (Luo et al., 2019), MGN (Wang et al., 2018), OSNet (Zhou et al., 2019), VV (Kuma et al., 2019; Kumar et al., 2020), AGW (Ye et al., 2021), and ViT (Dosovitskiy et al., 2020), which mainly focus on discriminative identity representation learning under conventional camera settings. The second group includes generalizable ReID methods, such as PAT (Ni et al., 2023) and BAU (Cho et al., 2024), which improve robustness to distribution shifts through part-aware representation learning or representation regularization. The third group comprises methods specifically designed for aerial-ground ReID, including Explain (Nguyen et al., 2023), VDT (Zhang et al., 2024), DTST (Wang & Pishgar, 2024a), GSAlign (Li et al., 2025), SeCap (Wang et al., 2025a), LATex (Zhang et al., 2025), and ViSA (Zhang et al., 2026). This diverse set of competitors allows us to examine whether the improvements arise from stronger generic representations or from effectively addressing the unique aerial-ground view discrepancy.

Results on CARGO. As shown in Table 1, conventional ReID models perform reasonably well under within-view retrieval but suffer substantial degradation under the more challenging A→G setting. Generalizable ReID methods do not fully close this gap, suggesting that generic robustness to distribution shifts is insufficient to address the severe geometric and appearance discrepancies between aerial and ground views. In contrast, recent aerial-ground approaches, such as VDT, GSAlign, LATex, and ViSA, achieve clear improvements by explicitly modeling view-specific variations. TAFD further improves cross-view retrieval, achieving 72.59% mAP and 76.60% Rank-1 under A→G, surpassing ViSA by 3.59 and 5.32%, respectively. Under the ALL protocol, TAFD achieves the highest mAP of 69.59%, exceeding LATex by 2.50%. It also obtains the highest A→A mAP of 71.81%, improving over LATex by 2.75% while remaining competitive under G→G. These results are consistent with the motivation of TAFD, which explicitly reduces aerial-ground discrepancies while preserving native-view identity discrimination.

Results on AGReIDv1. Table 2 further evaluates TAFD on AGReIDv1, where its advantage remains consistent across both retrieval directions. Under A→G, TAFD achieves 79.55% mAP and 86.46% Rank-1, improving over LATex by 1.88 and 1.20%, respectively. Under the reverse G→A direction, TAFD achieves the highest mAP of 82.03% together with a competitive Rank-1 accuracy of 88.57%. Averaged over both directions, TAFD reaches 80.79% mAP and 87.51% Rank-1, improving upon the average mAP of LATex from 79.41% to 80.79%. Compared with earlier aerial-ground methods such as VDT, GSAlign, and SeCap, the improvement is also consistent across both directions. These results indicate that the learned deformation is not tied to a particular source-target orientation and transfers effectively to a second aerial-ground benchmark.

Table 2: **Performance comparison with state-of-the-art methods on AGReIDv1 under two retrieval directions.** The best values in each column are shown in **bold**.

Method	Venue	A→G		G→A		Avg.	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
SBS	ACM MM'23	59.77	73.54	62.27	73.70	61.02	73.62
BoT	CVPRW'19	55.47	70.01	58.83	71.20	57.15	70.61
OSNet	ICCV'19	58.32	72.59	60.99	74.22	59.66	73.41
VV	JAISCR'20	67.23	77.22	69.83	79.73	68.53	78.47
Explain	ICME'23	72.38	81.28	73.35	82.64	72.86	81.96
ViT	ICLR'21	72.61	81.47	73.39	82.85	73.00	82.16
VDT	CVPR'24	74.44	82.91	78.57	86.59	76.50	84.75
GSAlign	NeurIPS'25	75.01	83.75	77.73	84.10	76.37	83.92
SeCap	CVPR'25	76.16	84.03	78.34	87.01	77.25	85.52
LATex	arXiv'25	77.67	85.26	81.15	**89.40**	79.41	87.33
ViSA	CVPR'26	75.56	84.91	77.94	84.10	76.75	84.50
TAFD	-	**79.55**	**86.46**	**82.03**	88.57	**80.79**	**87.51**

4.4 Ablation Study

We conduct a component ablation on AGReIDv1 to examine the contribution of each proposed module, as shown in Table 3. The baseline is a CLIP ViT-B/16 model trained under the same unified re-identification framework.

We first introduce Directional Feature Deformation (DFD), which explicitly transforms aerial and ground representations toward their opposite-view feature spaces. DFD provides the majority of the improvement, increasing the average mAP from 73.00% to 80.08% and the average Rank-1 from 82.16% to 86.78%. This confirms that explicitly modeling directional feature transformation effectively reduces the cross-view discrepancy between aerial and ground observations. We then incorporate Target-Space Identity Consistency (TIC) to constrain transformed features using identity evidence from the destination view. Compared with DFD alone, TIC improves the average Rank-1 from 86.78% to 87.46%, while the average mAP remains nearly unchanged at 80.06%, indicating that target-space supervision helps preserve identity information after deformation. Finally, Semantic View Anchoring (SVA) provides explicit target-view semantic guidance through aerial and ground text anchors. The full model reaches 80.79% average mAP and 87.51% average Rank-1, with a 0.73-point mAP gain over DFD+TIC. These results show that DFD establishes the main cross-view transformation capability, TIC strengthens target-space identity consistency, and SVA further guides the deformation toward the intended view.

Table 3: **Component ablation on AGReIDv1.** The selected full model is highlighted in gray.

Setting	A→G		G→A		Avg.	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Baseline	72.61	81.47	73.39	82.85	73.00	82.16
+ DFD	79.47	85.62	80.69	87.94	80.08	86.78
+ DFD + TIC	79.14	86.46	80.97	88.46	80.06	87.46
+ DFD + TIC + SVA	**79.55**	**86.46**	**82.03**	**88.57**	**80.79**	**87.51**

Table 4: **Further analysis of TAFD on AGReIDv1.** mAP and Rank-1 are reported for A→G and G→A, together with their averages. The best available result is highlighted in **bold**.

(a) Direction of feature deformation.

Setting	A→G		G→A		Avg.	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
A→G only	79.02	85.53	81.48	88.57	80.25	87.05
G→A only	79.16	86.09	81.23	87.84	80.20	86.96
Bidirectional	**79.55**	**86.46**	**82.03**	**88.57**	**80.79**	**87.51**

(b) Sensitivity to λ_{target}.

λ_{target}	A→G		G→A		Avg.	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
0.5	78.69	85.06	80.37	87.53	79.53	86.29
1.0	**79.55**	**86.46**	**82.03**	**88.57**	**80.79**	**87.51**
2.0	77.95	84.50	79.99	87.11	78.97	85.81

(c) Sensitivity to λ_{NCE}.

λ_{NCE}	A→G		G→A		Avg.	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
0.125	79.43	86.55	80.86	87.73	80.14	87.14
0.25	78.88	85.81	80.95	87.63	79.92	86.72
0.5	**79.55**	**86.46**	**82.03**	**88.57**	**80.79**	**87.51**
1.0	78.41	84.22	80.12	87.21	79.27	85.72

(d) Number of private prompt tokens.

Private	A→G		G→A		Avg.	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
0	78.10	85.15	80.97	88.25	79.54	86.70
2	**79.55**	**86.46**	**82.03**	**88.57**	**80.79**	**87.51**
4	79.15	86.18	81.42	88.36	80.79	87.27
8	78.04	85.25	80.61	**89.40**	79.32	87.32
16	77.88	84.50	80.36	89.09	79.12	86.79

4.5 Discussion

Unidirectional vs. Bidirectional Deformation. We first investigate whether deforming features in only one direction is sufficient for aerial-ground matching. As shown in Table 4a, we compare A→G-only deformation, G→A-only deformation, and the proposed bidirectional formulation. The two unidirectional variants optimize cross-view alignment from only one side, whereas the full formulation jointly learns both transformations. The bidirectional design achieves the best overall balance across the two retrieval directions, indicating that the two deformation directions provide complementary supervision rather than reducing cross-view alignment to a single asymmetric mapping.

Target-Alignment Weight. Table 4b studies the sensitivity to λ_{target}. The setting of 1.0 achieves the best overall performance, with 80.79% average mAP and 87.51% average Rank-1. Reducing the weight to 0.5 lowers the average performance, while increasing it to 2.0 also leads to degradation. This suggests that target-view alignment is most effective when applied with a moderate weight. We therefore set $\lambda_{\text{target}} = 1.0$ in the final configuration.

Prototype Contrastive Weight. Table 4c reports the sensitivity to λ_{NCE}. The default value of 0.5 yields the best average performance, reaching 80.79% average mAP and 87.51% average Rank-1. Weaker supervision at 0.125 or 0.25 gives slightly lower results, while increasing the weight to 1.0 reduces the average mAP to 79.27%. This indicates that excessively strong prototype-level constraints may interfere with identity discrimination. We therefore use $\lambda_{\text{NCE}} = 0.5$ by default.

Number of View-Specific Prompt Tokens. We further study the number of view-specific prompt tokens used in Semantic View Anchoring, as reported in Table 4d. Removing private tokens reduces the average mAP from 80.79% to 79.54%, indicating that view-specific textual cues are beneficial for guiding feature deformation. Using two private tokens achieves the best overall performance, with 80.79% average mAP and 87.51% average Rank-1. Increasing the number of private tokens to four, eight, or sixteen does not yield further improvements and generally decreases the average mAP. This suggests that a small number of private prompt tokens is sufficient to encode view-specific semantics, whereas additional tokens may introduce redundant guidance. We therefore use two private prompt tokens as the default setting.

5 Conclusion

We presented Text-Anchored Feature Deformation (TAFD) for aerial-ground person re-identification. TAFD retains native representations for view-specific identity learning while transforming multi-level features toward the opposite view. Semantic View Anchoring provides aerial and ground text anchors, Directional Feature Deformation learns separate transformations for the two directions, and Target-Space Identity Consistency supervises transformed features in their destination view. Together, these components give cross-view transformation an explicit target without replacing native-view discrimination. TAFD achieves 72.59% mAP and 76.60% Rank-1 for aerial-to-ground retrieval on CARGO and an average bidirectional mAP of 80.79% on AGReIDv1. These results demonstrate competitive retrieval performance across both benchmarks.

AI USE STATEMENT

In this work, we used generative AI tools primarily to assist with academic writing, including language polishing, sentence restructuring, and improving the clarity and presentation of the manuscript. We also used generative AI tools to assist with the implementation of selected portions of the experimental code. The research questions, methodological design, experimental setup, evaluation protocols, and interpretation of experimental results were conceived and conducted by the authors. All AI-assisted text was manually reviewed and revised by the authors.

REFERENCES

Yoonki Cho, Jaeyoon Kim, Woo J Kim, Junsik Jung, and Sung-Eui Yoon. Generalizable person re-identification via balancing alignment and uniformity. *Advances in Neural Information Processing Systems*, 37:47069–47093, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9664–9667, 2023.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15013–15022, 2021.

Wajahat Khalid, Bin Liu, Xulin Li, Muhammad Waqas, and Muhammad Sher Afgan. Bridging the sky and ground: Towards view-invariant feature learning for aerial-ground person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9749–9758, 2025.

Minsu Kim, Seungryong Kim, Jungin Park, Seongheon Park, and Kwanghoon Sohn. Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18621–18632, 2023.

Ratnesh Kuma, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram. Vehicle re-identification: an efficient baseline using triplet embedding. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9. IEEE, 2019.

Ratnesh Kumar, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram. A strong and efficient baseline for vehicle re-identification using deep triplet embedding. *Journal of Artificial Intelligence and Soft Computing Research*, 10(1):27–45, 2020.

Qiao Li, Jie Li, Yukang Zhang, Lei Tan, Jing Chen, and Jiayi Ji. Gsalign: Geometric and semantic alignment network for aerial-ground person re-identification. In *Advances in Neural Information Processing Systems*, 2025.

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*, volume 37, pp. 1405–1413, 2023.

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*, pp. 16266–16275, 2021a.

Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2898–2907, 2021b.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Aerial-ground person re-id. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 2585–2590. IEEE, 2023.

Hao Ni, Yuke Li, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Part-aware transformer for generalizable person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11280–11289, 2023.

Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27197–27206, 2024.

Kaijie Ren and Lei Zhang. Implicit discriminative knowledge learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 393–402, 2024.

Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11174–11184, 2023.

Yifan Sun, Liang Zheng, Yali Li, Yi Yang, Qi Tian, and Shengjin Wang. Learning part-based convolutional features for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):902–917, 2019.

Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6449–6458, 2020.

Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 274–282, 2018.

Shining Wang, Yunlong Wang, Ruiqi Wu, Bingliang Jiao, Wenxuan Wang, and Peng Wang. Secap: Self-calibrating and adaptive prompts for cross-view person re-identification in aerial-ground networks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22119–22128, 2025a. doi: 10.1109/CVPR52734.2025.02060.

Yuhai Wang and Maryam Pishgar. Dynamic token selection for aerial-ground person re-identification. *arXiv preprint arXiv:2412.00433*, 2024a.

Yuhai Wang and Maryam Pishgar. Dynamic Token Selection for Aerial-Ground Person Re-Identification, 2024b. URL https://arxiv.org/abs/2412.00433.

Yuhao Wang, Xiang Hu, Lixin Wang, Pingping Zhang, and Huchuan Lu. Sd-reid: View-aware stable diffusion for aerial-ground person re-identification. *arXiv preprint arXiv:2504.09549*, 2025b.

Zi Wang, Chenglong Li, Aihua Zheng, Ran He, and Jin Tang. Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2633–2641, 2022.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

Pingping Zhang, Xiang Hu, Yuhao Wang, and Huchuan Lu. Latex: Leveraging attribute-based text knowledge for aerial-ground person re-identification. *arXiv preprint arXiv:2503.23722*, 2025.

Quan Zhang, Lei Wang, Vishal M Patel, Xiaohua Xie, and Jianhaung Lai. View-decoupled transformer for person re-identification under aerial-ground camera network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22000–22009, 2024.

Quan Zhang, Zeqiang Cai, Peiming Zhao, Jingze Wu, Cailun Wu, Hongbo Chen, and Jianhuang Lai. View-aware semantic alignment for aerial-ground person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4383–4392, June 2026.

Shizhou Zhang, Qi Zhang, Yifei Yang, Xing Wei, Peng Wang, Bingliang Jiao, and Yanning Zhang. Person re-identification in aerial imagery. *IEEE Transactions on Multimedia*, 23:281–291, 2020.

Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3701–3711. IEEE, 2019.