3D-VITAL: Visibility-aware Identity Training from Artificial Liftings of 3D Reconstruction model for Cross-View Object Re-Identification

William Grolleau[1,2] **Astrid Sabourin**[1] **Guillaume Lapouge**[1] **Catherine Achard**[2]
[1]Université Paris-Saclay, CEA, List, F-91120 Palaiseau, France
[2]Sorbonne University, CNRS, Institute of Intelligent Systems and Robotics (ISIR), Paris, France
{william.grolleau,guillaume.lapouge}@cea.fr

Abstract

In this paper, we address the Cross-View Object Re-Identification (ReID) task, aiming to match identities across drastically different viewpoints. We focus specifically on Single-View Aerial-Ground ReID (SV AG-ReID), where models train on one viewpoint and must generalize to another at inference. To bridge the viewpoint gap of Cross-View ReID, related work employs generative models to synthesize novel views from source images as additional training data. We leverage recent advances in large-scale single-view 3D mesh reconstruction to lift source images into textured meshes and to render them from new viewpoints. However, this pipeline produces unreliable data in two distinct ways, and addressing both is vital for synthetic data to serve as an effective training signal. First, the reconstruction may fail and produce geometrically incorrect meshes, resulting in severe corruption in rendered novel views. Second, even when the geometry is correct, the rendered novel views contain hallucinated textures in regions not visible in the source image, which may degrade training if used naively. We address both shortcomings by exploiting a signal that prior generative methods do not naturally expose. Indeed, the reconstructed mesh carries a per-pixel partition between regions visible from the source view and regions that the reconstruction model had to hallucinate, available as a deterministic byproduct of rendering. We introduce this per-pixel visibility map and exploit it in a two-stage design. Cross-View Consistency Filtering (CVCF) discards geometrically incorrect meshes by ensuring multi-view consistency between source and novel views over visible patches. Visibility-Aware Attention Supervision (VAAS) leverages the visibility map to guide the model's attention toward reliable regions of the rendered view during training. Combined, these two components enable our method, 3D-VITAL, to better leverage the synthetic signal from single-view 3D mesh reconstruction, benefiting from wider viewpoint augmentation. 3D-VITAL sets a new state-of-the-art on three SV AG-ReID benchmarks (AG-ReID, AG-ReID.v2, MOO). Code and data will be released at https://github.com/TurtleSmoke/3D-VITAL.

1 Introduction

Object Re-Identification (ReID) is a pattern recognition problem aiming to match the same object across non-overlapping cameras. This matching is especially challenging in Aerial-Ground ReID (AG-ReID), an emerging setting driven by the deployment of Unmanned Aerial Vehicles (UAVs). Here, the gap between top-down aerial views and frontal ground views causes severe geometric distortions, self-occlusions, and pose variations [1, 2, 3, 4, 5, 6]. However, annotating data covering both viewpoints is costly, which has motivated the Single-View AG-ReID (SV AG-ReID) protocol [7],

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

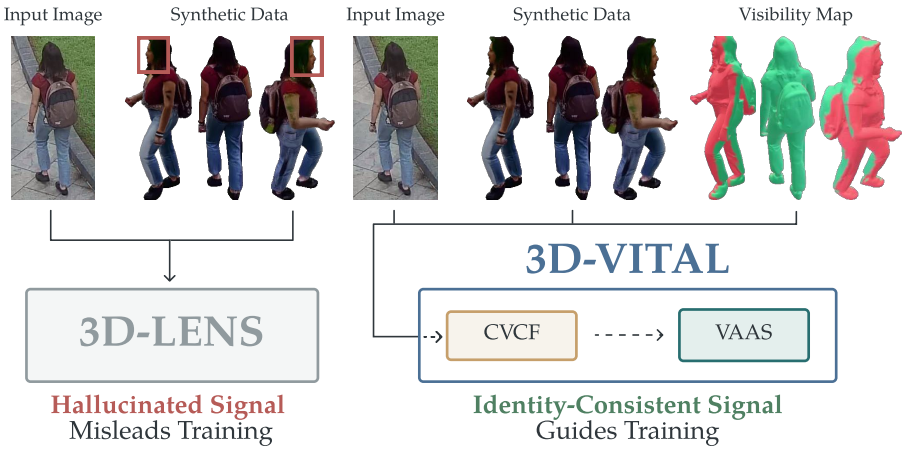


Figure 1: **Motivation of 3D-VITAL. Left:** Prior synthetic-real ReID pipelines such as 3D-LENS treat each synthetic sample as wholly reliable or wholly unreliable, which lets misleading identity signal from hallucinated regions (red rectangle) degrade training. **Right:** 3D-VITAL instead derives a per-pixel visibility map (visible regions as green, hallucinated as red) from the reconstructed mesh, separating observed regions from hallucinated ones. Cross-View Consistency Filtering (CVCF) uses this map to reject geometrically incorrect samples. Visibility-Aware Attention Supervision (VAAS) uses it to guide the model's attention toward observed regions during training.

in which only one viewpoint is available during training and the model must generalize across unseen viewpoints at inference time.

3D-LENS [7] addresses this task by lifting single-view images into textured meshes using recent single-view 3D mesh reconstruction [8]. It then renders novel views to simulate the missing viewpoint, providing cross-view supervision from single-view real data. However, this synthetic data is not always reliable for training for two reasons. When reconstruction fails to recover the correct geometry, the rendered novel views carry no usable identity signal. Even when the geometry is correct, some regions of the reconstructed object are not seen from the source view and are hallucinated by the reconstruction model. These models are trained on generic objects to produce visually plausible images, so hallucinated regions contain texture consistent with the object but without identity cues. Using such regions naively can degrade training, since the rendered novel views are only partially reliable.

Working with unreliable training data is a well-studied problem in computer vision. Prior work either infers reliability from optimization dynamics [9, 10] or derives it from external models [11, 12, 13]. However, neither approach applies in our setting, since 3D reconstruction models are explicitly optimized to produce visually plausible textures even in hallucinated regions, the rendered image alone provides no signal from which reliability could be estimated. Also, the 2D generative paradigms used in ReID, such as GANs and diffusion models, do not expose an explicit geometric signal from which a per-pixel reliability map could be derived.

Closer to our setting, existing synthetic data pipelines for ReID do not exploit such a per-pixel signal. They either trust the synthetic data [14, 15] or filter it at the image level [16, 17]. Both treat reliability as a property of the whole sample, so a novel view with correct geometry but partially hallucinated appearance is discarded entirely, wasting the valid identity signal it contains in its visible regions.

In this paper, we address these limitations by computing a visibility map directly from the reconstructed mesh (Figure 1). This per-pixel labeling separates visible regions from hallucinated regions. Visible regions have direct correspondence to the source image and are therefore more likely to carry its identity signal, whereas hallucinated regions consist of pixels synthesized without source correspondence, potentially injecting identity-inconsistent noise. Building on this signal, we propose 3D-VITAL, which exploits visibility through a two-stage pipeline. The first stage, **Cross-View Consistency Filtering (CVCF)**, rejects source images whose meshes are geometrically incorrect. The second stage, **Visibility-Aware Attention Supervision (VAAS)**, operates on the surviving samples. We validate 3D-VITAL on three SV AG-ReID benchmarks (AG-ReID, AG-ReID.v2, MOO).

In a nutshell, the main contributions of this paper are as follows:

- We identify that 3D reconstruction methods enable the computation of a visibility map, providing a dense reliability signal that prior generative methods do not naturally expose.
- We introduce Cross-View Consistency Filtering (CVCF) and a Visibility-Aware Attention Supervision (VAAS). CVCF scores reconstruction quality only over visible patches, while VAAS uses the visibility map to guide the model's attention toward reliable regions.
- We achieve state-of-the-art results on three SV AG-ReID benchmarks (AG-ReID, AG-ReID.v2, MOO), with ablations isolating the contribution of each component.

2 Related Work

Estimating reliability. When part of the training signal is unreliable, prior work that aims to increase robustness to input noise can be grouped into two broad categories. A first category infers reliability from the optimization process itself, exploiting the observation that networks fit clean samples before corrupted ones [9, 10]. This property is exploited through peer-network disagreement [18, 19, 20, 21], clustering or Gaussian mixture models [22, 23, 24, 25, 20], or weighted losses based on uncertainty [22, 15, 23, 26]. A second category instead derives reliability from an external model or self-supervised proxy task, providing supervision in some regions of the image. The closest example to our setting is occluded person ReID, where external pose estimators produce heatmaps used as soft attention guidance [11, 12, 13, 27], or where attention weights are learned through self-supervised augmentations [28, 29, 30, 31]. However, in our setting, no such signal is available without explicit supervision since 3D reconstruction models produce plausible textures even in hallucinated regions, leaving no visual cue to distinguish reliable from unreliable patches, and no external model provides this partition for synthetic novel views. In contrast, the underlying 3D mesh exposes which regions are projected from the source view and which were hallucinated by the reconstruction model. Treating these regions differently during training lets us exploit synthetic data without being misled by the noise that hallucination introduces.

Synthetic data for object re-identification. Prior work using synthetic data in ReID treats reliability as a property of the whole sample, either by trusting all generated images or by scoring them at the image level. Early methods integrate GAN-synthesized images [14, 32, 33, 34, 15, 35], diffusion-based generation [36, 37, 38], and more recently single-view 3D mesh reconstruction novel-view rendering [7] into ReID training pipelines. These methods use synthetic data without specific reliability mechanisms, typically only smoothing labels to limit their influence relative to real datasets. However, many argue that curated subsets of synthetic data outperform larger uncurated data [16, 26, 39, 40], with recent ReID methods filtering synthetic data at the image level using ReID-based criteria [17]. Yet, in all of the above, reliability is a property of the full sample, with no mechanism to treat regions of an image independently. Reliability must instead be considered at the pixel level, separating visible regions that carry valid identity signal from hallucinated ones. To our knowledge, no prior synthetic-data pipeline for ReID exposes this distinction, and we are the first to exploit mesh-derived visibility as a per-pixel reliability signal for synthetic training data.

3 Methodology

3D-VITAL (Figure 1) is built upon 3D-LENS and extends it with three components that improve the use of synthetic data from single-view 3D mesh reconstruction. First, we compute a per-pixel visibility map that separates visible regions from hallucinated regions (Section 3.2). Second, Cross-View Consistency Filtering (CVCF) uses this visibility map to reject geometrically incorrect samples (Section 3.3). Third, Visibility-Aware Attention Supervision (VAAS) operates on the surviving samples and uses the visibility map both as a per-patch attention target and as a per-sample weight (Section 3.4). These steps allow 3D-VITAL to exploit valid synthetic signals while mitigating the impact of unreliable regions.

3.1 Preliminaries

Following the baseline [7], each source image in the training set is lifted to a textured mesh and rendered from perturbed camera poses to provide cross-view supervision. Given a source image

$I_i \in \mathbb{R}^{H\times W\times 3}$, a segmenter produces a foreground mask M_i, and a 3D reconstruction model $\mathcal{F}_{3D}$ produces a textured mesh $\mathcal{M}_i$. We calibrate a virtual camera $\Pi_i = [R_i \mid t_i]$ by maximizing the Intersection over Union (IoU) between the foreground mask M_i and the silhouette of $\mathcal{M}_i$ under Π_i. This calibration ensures following viewpoint perturbations are relative to the source viewpoint, and that the visibility map (Section 3.2) is well-defined with respect to the source image. Then, N novel views I_i^{syn} are rendered, composited onto an inpainted background, and aligned in style via style transfer [41], yielding the final synthetic views I_i^{sty}.

The baseline uses a ViT backbone with a CLS token t_{id} supervised by an identity classification loss $\mathcal{L}_{\text{id}}$, and a domain token t_{dom} supervised by a binary cross-entropy loss $\mathcal{L}_{\text{dom}}$ that distinguishes real from synthetic data. We keep both losses as is.

Our additions are the computation of a visibility map that is used with a Cross-View Consistency Filtering (CVCF), which removes incorrect samples before training, and a Visibility-Aware Attention Supervision (VAAS), which introduces a new attention loss term ($\mathcal{L}_{\text{vis}}$) and a weighted variant of the existing triplet term ($\mathcal{L}_{\text{tri}}^{w}$).

3.2 Visibility Map

We aim to compute a per-pixel map separating visible regions in the source image from regions hallucinated by the 3D reconstruction model.

First, we build a visibility mesh $\mathcal{V}_i$ with the same topology as $\mathcal{M}_i$ but with a binary label per triangle. We employ ray casting from the source camera Π_i, marking the first triangle hit by each pixel ray as visible (1) and all other triangles as non-visible (0). Visible triangles are treated as more reliable because their appearance is constrained by the source image, although visibility does not guarantee a faithful reconstruction, while non-visible triangles carry texture hallucinated by $\mathcal{F}_{3D}$, since no observation constrained it.

Second, for each rendered view $I_i^{\text{syn}} = \mathcal{R}(\mathcal{M}_i, \tilde{\Pi}_i)$, we render $\mathcal{V}_i$ under the same camera pose, yielding a binary visibility map $V_i \in \{0,1\}^{H\times W}$. Its pixels stay aligned with the foreground of I_i^{sty}, since background compositing and style transfer do not move the foreground. Spatial augmentations applied to I_i^{sty} during training (horizontal flip, random crop) are applied identically to V_i to preserve pixel correspondence.

Third, we downsample V_i to the $h \times w$ patch grid by average pooling over non-overlapping patches, yielding $\tilde{V}_i \in [0,1]^{h\times w}$. Each entry of $\tilde{V}_i$ is the fraction of visible pixels within its corresponding patch. This patch-resolution map is not fed to the network as input but used downstream as a per-sample reliability score for filtering (Section 3.3) and a supervisory signal for attention (Section 3.4).

3.3 Cross-View Consistency Filtering

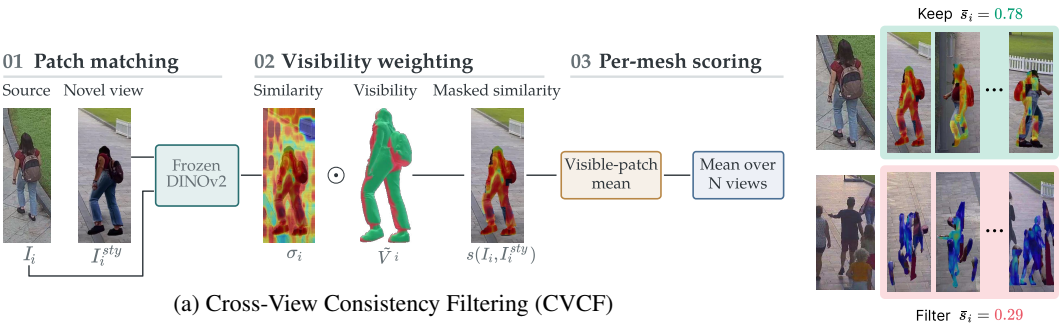
Single-view 3D reconstruction may fail geometrically, producing novel views without identity signal.

To filter these, we compute a consistency score s based on the cosine similarity between DINOv2 features of the source image and its novel views, restricted to visible patches via $\tilde{V}_i$. DINOv2 patch features provide a robust signal for cross-view appearance consistency, while the visibility restriction ensures that only non-hallucinated regions contribute to the score. Unlike conventional image-level filters that completely discard heavily hallucinated views, this approach retains them. As long as the visible regions remain consistent with the source image, these reliable areas still provide a valuable cross-view signal that can be leveraged during training.

Let $\phi(\cdot)$ denote patch features extracted by a frozen DINOv2 encoder. For each patch j of a style-aligned novel view I_i^{sty}, we compute its maximum cosine similarity to any patch k in the source image I_i,

$$\sigma_i^{(j)} = \max_k \cos(\phi(I_i^{\text{sty}})^{(j)}, \phi(I_i)^{(k)}). \tag{1}$$

The image consistency score weights each patch similarity by its visibility:



(a) Cross-View Consistency Filtering (CVCF)

(b) Kept and Filtered

Figure 2: **Cross-View Consistency Filtering (CVCF). (a)** Frozen DINOv2 features match each patch of a style-aligned novel view I_i^{sty} to its most similar source-image patch, giving cosine similarities σ_i (Equation (1)). The patch map $\tilde{V}_i$ masks these similarities to source-visible regions (Equation (2)), so hallucinated patches do not contribute. The visible-patch mean, averaged over the N novel views, gives one consistency score $\bar{s}_i$ per source reconstruction (Equation (3)), used to keep the top samples. **(b)** A correct mesh (top) stays consistent across views in the visible regions, even when hallucinated regions are large, while an incorrect mesh (bottom) is inconsistent.

$$s(I_i, I_i^{\text{sty}}) = \frac{\langle \tilde{V}_i, \sigma_i \rangle}{\|\tilde{V}_i\|_1}. \tag{2}$$

The score is one-directional from novel view to source, since a bidirectional match would fail because novel viewpoints naturally occlude some source patches. We aggregate the N novel views per source image by averaging their scores:

$$\bar{s}_i = \frac{1}{N} \sum_{n=1}^{N} s(I_i, I_i^{\text{sty},n}). \tag{3}$$

This appearance-consistency score serves as a proxy for geometric correctness, since it does not enforce one-to-one or geometric correspondence. To filter out geometrically incorrect meshes, we keep the synthetic views of the 50% of source images with the highest $\bar{s}_i$, while every real image remains in training. Sensitivity to this choice is reported in Table 13.

3.4 Visibility-Aware Attention Supervision

Surviving samples contain the source image's identity signal in visible regions and misleading identity appearance in hallucinated ones. The visibility map exposes this partition directly, and we exploit it through two mechanisms. A visibility-aware identity loss ($\mathcal{L}_{\text{vis}}$) supervises the model to focus on visible regions, while a visibility-weighted triplet loss ($\mathcal{L}_{\text{tri}}^{w}$) reduces the impact of heavily hallucinated views on the overall optimization. This attention supervision acts as a regularization signal rather than a strict isolation, since final-layer patch features already mix information from other patches.

Visibility-Aware Identity Loss

We supervise the attention weights of the CLS identity token at the final transformer layer to concentrate on visible patches, using the visibility map as a dense target distribution. We apply this supervision solely at the last transformer layer, allowing earlier layers to develop global context freely, and only to synthetic views. Hallucinated regions are kept in the input, since they still provide object context such as silhouette and pose (Table 6). Ablation across the number of supervised layers is reported in Table 10.

In the final transformer layer, the identity token t_{id} attends to itself, the domain token, and the P image patches. We extract only the attention weights directed at the P image patches, average them across all attention heads, and re-normalize them to sum to one, yielding the attention distribution $\alpha_i \in \mathbb{R}^P$.

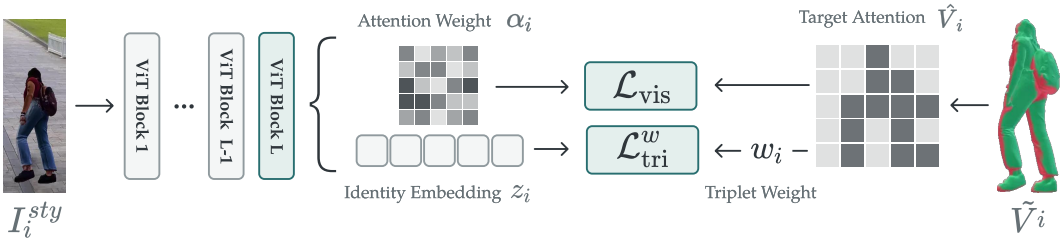


Figure 3: **Visibility-Aware Attention Supervision (VAAS).** A style-aligned novel view I_i^{sty} is processed by a ViT. At the final transformer layer, we extract the attention weights α_i of the identity token t_{id} over the image patches, averaged across heads. The visibility map $\tilde{V}_i$ provides two supervision signals. *Top:* $\mathcal{L}_{\text{vis}}$ is the cross-entropy between α_i and the normalized visibility target $\hat{V}_i$ (Equation (4)), pulling identity-token attention toward visible patches. *Bottom:* the per-sample weight $w_i = \frac{1}{P}\sum_j \tilde{V}_i^{(j)}$ scales the anchor's contribution to $\mathcal{L}^w_{\text{tri}}$ (Equation (5)), reducing the influence of heavily hallucinated samples.

We also normalize the visibility map to a target probability distribution: $\hat{V}_i = \tilde{V}_i/(\epsilon + \sum_{j=1}^{P} \tilde{V}_i^{(j)})$, where ϵ is a small constant for numerical stability. We define the attention loss as the cross-entropy of this extracted attention against the visibility target distribution:

$$\mathcal{L}_{\text{vis}} = -\sum_{j=1}^{P} \hat{V}_i^{(j)} \log \alpha_i^{(j)}. \tag{4}$$

A comparison against alternative attention targets that ignore visible-vs-hallucinated information is reported in Table 9.

Visibility-Weighted Triplet Loss

The visibility map yields a per-sample reliability scalar $w_i = \frac{1}{P}\sum_{j=1}^{P} \tilde{V}_i^{(j)}$, representing the fraction of visible regions in the rendered view. We use w_a, obtained by setting $i = a$, to weight the sample's contribution as a triplet anchor. Given an anchor a, positive p, and negative n drawn from a mini-batch, with embedding distance $D(\cdot,\cdot)$, the standard soft-margin triplet $\log\left(1 + \exp\left(D(a,p) - D(a,n)\right)\right)$ is replaced by

$$\mathcal{L}^w_{\text{tri}} = w_a \cdot \log\left(1 + \exp\left(D(a,p) - D(a,n)\right)\right). \tag{5}$$

The weighted terms are normalized by the sum of the anchor weights in the batch, so anchors with low w_a contribute proportionally less to the triplet loss. For real source images, we set $w_i \equiv 1$ and recover the standard term.

Full Objective

The total loss is:

$$\mathcal{L} = (\lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{dom}}\mathcal{L}_{\text{dom}}) + \lambda_{\text{tri}}\mathcal{L}^w_{\text{tri}} + \lambda_{\text{vis}}\mathcal{L}_{\text{vis}}, \tag{6}$$

with λ_{id} and λ_{dom} set to 1 and 0.5 respectively as in the baseline, and with λ_{tri}, and λ_{vis} being hyperparameters controlling the contribution of each term. Synthetic samples contribute to $\mathcal{L}_{\text{id}}$ and $\mathcal{L}_{\text{dom}}$ as in the baseline regardless of their visibility. Their influence is reduced only through the anchor weight of $\mathcal{L}^w_{\text{tri}}$, and $\mathcal{L}_{\text{vis}}$ is applied to synthetic views only. This joint optimization framework encourages the learning of discriminative representations that are resilient to synthetic hallucinations.

4 Experiments

4.1 Implementation

Datasets and Evaluation protocol . Throughout the paper, the notation $X \rightarrow Y$ denotes a query from view X matched against a gallery of view Y, independent of which view the model was trained on. We evaluate 3D-VITAL on two pedestrian AG-ReID benchmarks, AG-ReID [2] and AG-ReID.v2 [3], and on MOO [42], a synthetic cattle dataset that tests generalization beyond pedestrians. AG-ReID contains 21,983 images and 388 identities from one ground camera and one UAV. AG-ReID.v2 contains 100,502 images and 1,615 identities across UAV, wearable, and CCTV platforms. MOO contains 128,000 images and 1,000 identities with elevations from -25° to 90°. We follow the Single-View AG-ReID protocol of [7], training each model on a single viewpoint and querying cross-viewpoint at inference. For AG-ReID and MOO, we train one model per view, yielding two training viewpoints per dataset (Aerial and Ground for AG-ReID, Top and Side for MOO). We evaluate both cross-view directions for each training view, giving four evaluations per dataset. For AG-ReID.v2, we train three models, one per view (Aerial, Wearable, and CCTV). The Wearable view acts as a second ground-like viewpoint and is evaluated against Aerial. We evaluate Aerial $\leftrightarrow$ CCTV and Aerial $\leftrightarrow$ Wearable in both directions for each training view, giving eight inference scenarios for AG-ReID.v2.

Implementation Details . We build on 3D-LENS [7] with a ViT backbone and add our contributions incrementally to isolate their impact. We render synthetic views with a wider azimuth range than the baseline, since 3D-VITAL exploits viewpoint diversity without being harmed by the larger hallucinated regions it introduces (Section 4.3). Images are resized to 256×128 and augmented with random cropping, horizontal flipping, and random erasing. The loss weights λ_{tri} and λ_{vis} from Equation (6) are both set to 1. Sensitivity to these weights is reported in Tables 11 and 12. Training runs for 120 epochs with SGD (momentum 0.9, weight decay 10^{-4}) under cosine scheduling, with a learning rate of 8×10^{-4} on AG-ReID and AG-ReID.v2 and 8×10^{-3} on MOO.

4.2 Comparison with State-of-the-Art

We compare 3D-VITAL against prior methods on AG-ReID (Table 1), AG-ReID.v2 (Table 2), and MOO (Table 1). The baselines span standard ReID architectures (BoT [43], ViT [44], AGW [45], TransReID [46]), viewpoint-agnostic methods (RotTrans [47], VDT [5]), large-scale pretraining (PASS [48]), generative adaptation (DCAC [38]), and the 3D-lifting baseline 3D-VITAL builds on (3D-LENS [7]). All baselines use a ViT-Base backbone pretrained on ImageNet-21k [49], except PASS, which uses its own pretraining.

3D-VITAL achieves the best mAP and Rank-1 across all four AG-ReID configurations (Table 1), reaching 53.2 mAP and 62.6 R-1 on Aerial $G \rightarrow A$, which is $+3.7$ mAP and $+4.9$ R-1 pp over 3D-LENS and far ahead of the best non-3D-lifting baseline (DCAC at 35.2 / 42.7). The gain is even more pronounced on MOO Side training $T \rightarrow S$ (Table 1), where 3D-VITAL reaches 47.8 mAP and 99.6 R-1, a $+15.7$ mAP pp jump over 3D-LENS.

The pattern extends to AG-ReID.v2 (Table 2), where 3D-VITAL improves over 3D-LENS under CCTV training (*e.g.*, $A \rightarrow C$: 46.3 / 54.5 vs. 41.4 / 50.8 mAP / R-1). Gains are smaller under Wearable training, where the higher image resolution (222×96 vs. 187×84 for Aerial and 157×59 for CCTV) makes most reconstructions usable and the fixed 50% filtering ratio overly aggressive. Even so, 3D-VITAL still outperforms every prior method other than 3D-LENS in these columns.

These consistent gains demonstrate the effectiveness of 3D-VITAL's approach to improving the reliability of synthetic data from single-view 3D mesh reconstruction, enabling it to learn better cross-view representations.

4.3 Ablation Study

Component Analysis . To assess the individual contribution of each 3D-VITAL component, we conduct an ablation study on the AG-ReID dataset. Following the same synthetic data generation as the baseline [7], we report the contribution of each 3D-VITAL component in Table 3.

Table 1: **State-of-the-Art Comparison on AG-ReID and MOO.** mAP and Rank-1 (%) on cross-view scenarios across Aerial (A), Ground (G) for AG-ReID, and Top (T), and Side (S) for MOO. Best in **bold**, second-best underlined.

Methods	Year	AG-ReID								MOO							
		Aerial Training				Ground Training				Top Training				Side Training			
		$A \to G$		$G \to A$		$A \to G$		$G \to A$		$T \to S$		$S \to T$		$T \to S$		$S \to T$	
		mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
BoT [43]	2019	41.7	55.7	29.6	34.3	45.3	51.8	49.6	60.2	37.1	95.9	39.1	43.5	19.3	43.1	19.4	94.2
ViT [44]	2020	42.6	58.1	31.9	35.8	46.9	54.6	49.8	60.1	36.5	93.9	37.8	41.6	19.8	43.5	20.0	94.8
AGW [45]	2021	41.4	55.3	32.2	36.7	42.3	51.0	47.3	59.5	36.1	89.2	37.8	44.0	13.5	40.8	13.8	76.6
TransReID [46]	2021	32.7	46.4	23.5	25.2	44.2	50.7	49.7	62.6	37.6	88.2	39.0	40.5	22.8	43.5	21.5	94.7
RotTrans [47]	2022	35.2	50.2	23.2	35.8	44.8	50.3	48.8	59.3	41.4	99.2	41.6	42.4	30.7	43.2	30.7	97.4
PASS [48]	2022	39.1	52.2	26.0	28.6	47.7	56.2	47.8	58.0	36.8	99.3	37.7	40.1	26.7	45.0	23.1	93.0
SBS [50]	2023	16.8	27.9	15.5	15.2	27.1	32.3	27.5	37.5	14.1	63.3	18.9	19.8	0.8	1.1	0.9	0.9
DCT [51]	2023	22.0	35.8	14.0	14.2	36.2	43.8	40.6	52.3	35.6	88.7	36.0	39.1	18.5	40.0	16.0	90.2
VDT [5]	2024	39.0	53.3	31.7	36.1	48.3	56.0	51.4	63.2	37.9	97.9	39.1	43.7	18.1	43.2	17.8	92.9
DCAC [38]	2025	43.4	60.0	35.2	42.7	49.7	61.6	52.9	66.4	4.0	28.6	3.8	38.9	14.3	33.4	13.0	75.3
3D-LENS [7]	2026	49.8	62.9	49.5	57.7	53.3	65.2	55.1	67.4	51.4	**100.0**	52.2	55.6	32.1	57.9	32.9	99.0
3D-VITAL	**2026**	**52.1**	**66.1**	**53.2**	**62.6**	**56.0**	**67.0**	**57.9**	**68.6**	**57.0**	**100.0**	**53.2**	**60.0**	**47.8**	**99.6**	**45.1**	**99.6**

Table 2: **State-of-the-Art Comparison on AG-ReID.v2.** mAP and Rank-1 (%) on cross-view scenarios across Aerial (A), CCTV (C), and Wearable (W). Best in **bold**, second-best underlined.

Methods	Year	AG-ReID.v2															
		Aerial Training								CCTV Training				Wearable Training			
		$C \to A$		$A \to C$		$W \to A$		$A \to W$		$C \to A$		$A \to C$		$W \to A$		$A \to W$	
		mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
BoT [43]	2019	14.4	16.8	32.5	42.4	18.8	21.2	28.7	40.5	42.3	52.6	34.5	40.0	31.6	38.0	25.1	28.9
ViT [44]	2020	14.7	16.7	33.2	43.1	19.1	21.5	27.4	38.1	41.8	51.8	35.0	41.0	31.0	37.1	23.9	27.9
AGW [45]	2021	14.6	17.8	33.2	42.6	19.9	23.9	30.2	40.9	38.3	47.4	31.5	38.2	32.4	39.1	25.3	31.2
TransReID [46]	2021	11.5	12.4	25.3	34.0	15.8	18.4	23.8	34.1	36.6	47.4	31.1	37.3	26.0	33.7	20.7	23.8
RotTrans [47]	2022	15.2	16.6	32.6	42.1	20.7	21.8	30.4	42.5	41.6	52.4	34.8	39.7	31.6	40.4	28.7	31.2
PASS [48]	2022	16.6	21.1	32.4	43.6	16.2	19.4	26.3	38.8	40.5	50.5	36.5	42.4	31.0	39.8	25.8	29.8
SBS [50]	2023	2.5	2.7	3.8	5.1	3.2	3.3	2.7	2.8	14.3	21.5	12.6	14.3	3.8	5.0	3.5	3.0
DCT [51]	2023	5.8	6.2	12.6	20.2	10.4	12.1	12.8	22.1	31.6	43.4	26.0	30.6	17.9	26.6	15.8	17.8
VDT [5]	2024	14.6	17.3	32.5	42.5	18.3	21.4	26.4	37.6	41.1	51.7	33.9	39.8	29.6	37.2	23.3	26.4
DCAC [38]	2025	14.4	18.1	34.9	47.9	19.8	23.5	32.6	46.8	45.6	57.9	39.0	48.4	31.9	41.6	30.1	35.8
3D-LENS [7]	2026	37.8	47.1	42.8	55.7	28.6	35.8	34.3	47.1	47.2	58.5	41.4	50.8	**38.3**	**46.2**	**38.9**	**47.9**
3D-VITAL	**2026**	**39.9**	**47.3**	**43.9**	**56.5**	**31.1**	**37.1**	**35.2**	**48.3**	**50.4**	**61.1**	**46.3**	**54.5**	37.9	45.8	**38.9**	46.9

Each component contributes independently: on Aerial $A \to G$, VAAS_{vis} alone yields $+3.2$ mAP pp (Index 1), VAAS_{tri} $+0.9$ pp (Index 2), and CVCF the largest standalone gain at $+5.1$ pp (Index 4) since broken meshes are the most damaging samples. Combined (Index 5), the three components reach $+6.5$ mAP pp under Aerial training and $+6.3$ under Ground training. CVCF removes broken geometry upstream while VAAS handles partial visibility within the surviving samples and together enable the model to learn more effectively from the synthetic data.

Table 3: **Component Ablation on AG-ReID.** mAP and Rank-1 (%) on Aerial $\to$ Ground and Ground $\to$ Aerial, with gains over Index 0 in green. Best in **bold**.

Index	Components			AG-ReID Aerial				AG-ReID Ground			
	VAAS_{vis}	VAAS_{tri}	CVCF	$A \to G$		$G \to A$		$A \to G$		$G \to A$	
				mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
0				45.6	61.2	46.1	56.4	49.7	61.6	51.3	62.9
1	✓			48.8 (+3.2)	63.9 (+2.6)	50.0 (+3.8)	59.6 (+3.1)	52.2 (+2.5)	63.3 (+1.7)	54.6 (+3.3)	65.8 (+2.9)
2		✓		46.5 (+0.9)	62.2 (+0.9)	46.9 (+0.8)	56.9 (+0.4)	50.6 (+0.9)	62.9 (+1.3)	52.9 (+1.6)	63.6 (+0.7)
3	✓	✓		49.2 (+3.6)	64.2 (+3.0)	49.5 (+3.3)	58.6 (+2.2)	53.0 (+3.3)	64.5 (+2.9)	55.5 (+4.3)	67.6 (+4.7)
4			✓	50.7 (+5.1)	65.6 (+4.4)	52.0 (+5.9)	60.2 (+3.8)	54.0 (+4.3)	65.1 (+3.5)	56.3 (+5.0)	67.4 (+4.5)
5	✓	✓	✓	**52.1 (+6.5)**	**66.1 (+4.9)**	**53.2 (+7.1)**	**62.6 (+6.2)**	**56.0 (+6.3)**	**67.0 (+5.4)**	**57.9 (+6.7)**	**68.6 (+5.7)**

Table 4: **Azimuth Range Ablation on AG-ReID.** mAP and Rank-1 (%) on Aerial → Ground and Ground → Aerial. Best in **bold**, second-best underlined.

Setting	Azimuth	Aerial Training				Ground Training			
		$A \to G$		$G \to A$		$A \to G$		$G \to A$	
		mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
Baseline	$\pm 15^\circ$	46.5	61.7	47.2	57.2	50.0	62.3	51.1	63.3
	$\pm 30^\circ$	**46.9**	**62.6**	**47.9**	**57.9**	**50.6**	**62.6**	**52.9**	**64.3**
	$\pm 60^\circ$	45.6	61.2	46.1	56.4	49.7	61.6	51.3	62.9
3D-VITAL	$\pm 15^\circ$	48.7	63.0	50.0	58.1	54.9	65.3	56.1	67.0
	$\pm 30^\circ$	49.8	64.8	51.0	58.3	55.2	66.4	56.5	66.9
	$\pm 60^\circ$	**52.1**	**66.1**	**53.2**	**62.6**	**56.0**	**67.0**	**57.9**	**68.6**

Table 5: **Filtering Strategy Ablation on AG-ReID.** mAP and Rank-1 (%) on Aerial → Ground and Ground → Aerial. Best per column within each sub-block in **bold**, second-best underlined.

Setting	Filter	Aerial Training				Ground Training			
		$A \to G$		$G \to A$		$A \to G$		$G \to A$	
		mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
Without VAAS	No filter	45.6	61.2	46.1	56.4	49.7	61.6	51.3	62.9
	CLIP I-I	46.6	61.0	47.1	57.1	49.0	61.6	51.8	64.1
	CLIP I-T	46.9	61.7	47.7	57.8	51.8	63.5	54.8	65.8
	DINOv2 CLS	48.5	64.2	50.4	59.5	52.0	63.3	54.3	65.8
	DINOv2 w/o Vis	49.8	64.4	51.2	59.2	52.8	63.8	55.3	65.9
	ReID-CTF [17]	50.0	64.8	51.2	59.4	52.7	64.9	55.1	65.8
	DINOv2 Vis (ours)	**50.7**	**65.6**	**52.0**	**60.2**	**54.0**	**65.1**	**56.3**	**67.4**
With VAAS	CLIP I-I	49.8	64.1	50.8	59.3	52.4	63.4	55.6	65.3
	CLIP I-T	49.6	63.9	50.7	59.1	52.9	63.0	55.6	65.3
	DINOv2 CLS	50.5	65.0	51.5	60.3	54.2	65.3	56.6	66.7
	DINOv2 w/o Vis	50.3	64.7	51.3	60.7	54.8	65.4	55.9	66.0
	ReID-CTF [17]	50.7	64.9	51.7	60.4	54.8	65.5	55.9	66.5
	DINOv2 Vis (ours)	**52.1**	**66.1**	**53.2**	**62.6**	**56.0**	**67.0**	**57.9**	**68.6**

Novel View Augmentation Analysis . Table 4 reports the effect of the azimuth augmentation range on AG-ReID, with and without 3D-VITAL. Without 3D-VITAL, performance peaks at the moderate $\pm 30^\circ$ range and degrades at $\pm 60^\circ$, dropping by 1.3 mAP pp on Aerial $A \to G$ relative to $\pm 30^\circ$, since wider perturbations expose the baseline to rendered views with larger hallucinated regions. Once 3D-VITAL components are enabled, the accuracy increases instead, reaching 52.1 mAP on Aerial $A \to G$ at $\pm 60^\circ$, a gain of $+2.3$ mAP pp over the same setting at $\pm 30^\circ$. This shift is consistent with the role of CVCF and VAAS, since geometrically broken meshes are filtered out and partial visibility is handled during training, the model can exploit the larger viewpoint diversity offered by wider azimuth perturbations rather than being harmed by it.

Filtering Strategy Analysis . Table 5 compares filtering strategies on AG-ReID with and without VAAS. CLIP-based filters underperform because image-to-image scoring does not capture cross-view identity consistency and image-to-text only verifies subject presence. DINOv2 variants improve on this, but both let hallucinated regions corrupt the score. Diff-Person [17], with its ReID-CTF module, is the strongest non-visibility baseline but introduces a paradox. ReID models penalize the extreme-view samples that are precisely the reason for generating novel views in the first place. Restricting the score to visible patches (ours) avoids both issues and leads to the best performance across all settings.

The interaction with VAAS further highlights the importance of visibility-aware filtering. Without VAAS, our filter leads by a modest margin (52.0 vs 51.2 mAP for ReID-CTF on Aerial $G \to A$). With VAAS, the gap widens to between 1.2 and 1.5 mAP and is consistent across all eight settings. Image-level filters cannot exploit this interaction because they discard the partially-hallucinated samples that VAAS is designed to learn from.

5 Conclusion and Limitations

Conclusion. We presented 3D-VITAL, which turns the visibility map exposed by single-view 3D mesh reconstruction into a training signal for cross-view ReID. Cross-View Consistency Filtering (CVCF) rejects samples with incorrect geometry but retains those with correct geometry and partial hallucination. Our Visibility-Aware Attention Supervision (VAAS) then supervises the CLS-attention and reweights triplet anchors, so the model learns from the visible regions of the surviving samples without being misled by hallucinated ones. Together, the two stages let 3D-VITAL train on a wider viewpoint range than the baseline and set a new state of the art on AG-ReID, AG-ReID.v2, and MOO.

Limitations. 3D-VITAL inherits the failure modes of its upstream components. Low resolution, severe occlusion, or multiple individuals per image may degrade reconstruction quality, causing CVCF to discard samples and reduce the augmented training set. While effective for isolated noisy samples, this filtering creates a bottleneck when poor reconstructions are systematic. Additionally, the fixed filtering ratio introduces a related issue. On the high-resolution Wearable split of AG-ReID.v2, discarding the bottom 50% of samples is overly aggressive because the average reconstruction quality is high. Finally, objects with atypical geometries or fine details that were not seen during the 3D reconstruction model's training distribution often yield degenerate meshes and further limit the applicability of 3D-VITAL to such cases. Finally, as with any ReID method, improved aerial-ground matching could be misused for surveillance of individuals without consent, and deployments should comply with applicable privacy regulations.

Acknowledgments and Disclosure of Funding

This publication was made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-De-France Regional Council. The authors declare no competing interests.

References

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

[2] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Aerial-ground person re-id. In *ICME*, pages 2585–2590, 2023.

[3] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Ag-reid. v2: Bridging aerial and ground views for person re-identification. *IEEE TIFS*, 19:2896–2908, 2024.

[4] Linzhi Shang, Chen Min, Juan Wang, Liang Xiao, Dawei Zhao, and Yiming Nie. Aerial-ground cross-view vehicle re-identification: A benchmark dataset and baseline. *Remote Sens.*, 17(15):2653, 2025.

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

[6] 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 *CVPR*, pages 22119–22128, 2025.

[7] William Grolleau, Astrid Sabourin, Guillaume Lapouge, and Catherine Achard. 3d-lens: A 3d lifting-based elevated novel-view synthesis method for single-view aerial-ground re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 205–221, 2026.

[8] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, Huiwen Shi, Sicong Liu, Junta Wu, Yihang Lian, Fan Yang, Ruining Tang, Zebin He, Xinzhou Wang, Jian Liu, Xuhui Zuo, Zhuo Chen, Biwen Lei, Haohan Weng, Jing Xu, Yiling Zhu, Xinhai Liu, Lixin Xu, Changrong Hu, Shaoxiong Yang, Song Zhang, Yang Liu, Tianyu Huang, Lifu Wang, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Yulin Cai, Jiaao Yu, Yixuan Tang, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu, Chao Zhang, Yonghao Tan, Jie Xiao, Yangyu Tao, Jianchen Zhu, Jinbao Xue, Kai Liu, Chongqing Zhao, Xinming Wu, Zhichao Hu, Lei Qin, Jianbing Peng, Zhan Li, Minghui Chen, Xipeng Zhang, Lin Niu, Paige Wang, Yingkai Wang, Haozhao Kuang, Zhongyi Fan, Xu Zheng, Weihao Zhuang, YingPing He, Tian Liu, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, Jingwei

Huang, and Chunchao Guo. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation, 2025.

[9] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

[10] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

[11] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[12] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[13] 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 (CVPR)*, June 2020.

[14] 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)*, Oct 2017.

[15] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[16] Suncheng Xiang, Dahong Qian, Mengyuan Guan, Binjie Yan, Ting Liu, Yuzhuo Fu, and Guanjie You. Less is more: Learning from synthetic data with fine-grained attributes for person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 19(5s):173:1–173:20, 2023.

[17] Ke Niu, Haiyang Yu, Xuelin Qian, Teng Fu, Bin Li, and Xiangyang Xue. Synthesizing efficient data with diffusion models for person re-identification pre-training. *Machine Learning*, 114(3):66, 2025.

[18] Fang Zhao, Shengcai Liao, Guo-Sen Xie, Jian Zhao, Kaihao Zhang, and Ling Shao. Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In *European conference on computer vision*, pages 526–544. Springer, 2020.

[19] Fengxiang Yang, Ke Li, Zhun Zhong, Zhiming Luo, Xing Sun, Hao Cheng, Xiaowei Guo, Feiyue Huang, Rongrong Ji, and Shaozi Li. Asymmetric co-teaching for unsupervised cross-domain person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12597–12604, 2020.

[20] Mouxing Yang, Zhenyu Huang, and Xi Peng. Robust object re-identification with coupled noisy labels. *International Journal of Computer Vision*, 132(7):2511–2529, 2024.

[21] Min Liu, Fei Wang, Xueping Wang, Yaonan Wang, and Amit K Roy-Chowdhury. A two-stage noise-tolerant paradigm for label corrupted person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4944–4956, 2024.

[22] Haiming Sun and Shiwei Ma. Pro-reid: Producing reliable pseudo labels for unsupervised person re-identification. *Image and Vision Computing*, 150:105244, 2024.

[23] Xian Zhong, Xiyu Han, Xuemei Jia, Wenxin Huang, Wenxuan Liu, Shuaipeng Su, Xiaohan Yu, and Mang Ye. Iclr: Instance credibility-based label refinement for label noisy person re-identification. *Pattern Recognition*, 148:110168, 2024.

[24] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8222–8231, 2019.

[25] 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*, pages 27197–27206, 2024.

[26] Yiqi Jiang, Weihua Chen, Xiuyu Sun, Xiaoyu Shi, Fan Wang, and Hao Li. Exploring the quality of gan generated images for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4146–4155, 2021.

[27] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body part-based representation learning for occluded person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1613–1623, January 2023.

[28] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[29] Zhikang Wang, Feng Zhu, Shixiang Tang, Rui Zhao, Lihuo He, and Jiangning Song. Feature erasing and diffusion network for occluded person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4744–4753, 2022.

[30] Pratyay Dutta and Bir Bhanu. Latent diffusion-guided feature inpainting for occluded person re-identification with hybrid re-ranking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

[31] Peixian Chen, Wenfeng Liu, Pingyang Dai, Jianzhuang Liu, Qixiang Ye, Mingliang Xu, Qi'an Chen, and Rongrong Ji. Occlude them all: Occlusion-aware attention network for occluded person re-id. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11833–11842, 2021.

[32] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[34] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1230–1241, 2018.

[35] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *CVPR*, pages 2004–2013, 2021.

[36] Inès Hyeonsu Kim, JoungBin Lee, Woojeong Jin, Soowon Son, Kyusun Cho, Junyoung Seo, Min-Seop Kwak, Seokju Cho, JeongYeol Baek, Byeongwon Lee, and Seungryong Kim. Pose-dive: Pose-diversified augmentation for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 8640–8650, 2026.

[37] Xiao-Wen Zhang, Delong Zhang, Yi-Xing Peng, Zhi Ouyang, Jingke Meng, and Wei-Shi Zheng. Viperson: Flexibly generating virtual identity for person re-identification. In *ICCV*, pages 23374–23384, 2025.

[38] Jiachen Li and Xiaojin Gong. Unleashing the potential of pre-trained diffusion models for generalizable person re-identification. *Sensors*, 25(2):552, 2025.

[39] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. In *European Conference on computer vision*, pages 775–791. Springer, 2020.

[40] Shuting He, Hao Luo, Weihua Chen, Miao Zhang, Yuqi Zhang, Fan Wang, Hao Li, and Wei Jiang. Multi-domain learning and identity mining for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2485–2493, 2020.

[41] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8795–8805, 2024.

[42] William Grolleau, Achraf Chaouch, Astrid Sabourin, Guillaume Lapouge, and Catherine Achard. Moo: A multi-view oriented observations dataset for viewpoint analysis in cattle re-identification, 2026.

[43] 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 (CVPRW)*, pages 1487–1495, 2019.

[44] 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 *ICLR*, 2021.

[45] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022.

[46] 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 (ICCV)*, pages 14993–15002, 2021.

[47] Shuoyi Chen, Mang Ye, and Bo Du. Rotation invariant transformer for recognizing object in uavs. In *ACM MM*, pages 2565–2574, 2022.

[48] Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. Pass: Part-aware self-supervised pre-training for person re-identification. In *ECCV*, pages 198–214, 2022.

[49] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses, 2021.

[50] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. In *ACM MM*, pages 9664–9667, 2023.

[51] Wen Li, Cheng Zou, Meng Wang, Furong Xu, Jianan Zhao, Ruobing Zheng, Yuan Cheng, and Wei Chu. Dc-former: Diverse and compact transformer for person re-identification. In *AAAI*, volume 37, pages 1415–1423, 2023.

Table 6: **Hallucinated Pixel Masking on AG-ReID.** Hallucinated pixels of the rendered views are masked out of the input. mAP and Rank-1 (%) on Aerial → Ground and Ground → Aerial. Best in **bold**.

Variant	Aerial Training				Ground Training			
	$A \to G$		$G \to A$		$A \to G$		$G \to A$	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
Visibility map (ours)	**52.1**	**66.1**	**53.2**	**62.6**	**56.0**	**67.0**	**57.9**	**68.6**
Hallucinated pixels masked	50.3	63.1	46.5	52.4	54.3	63.0	55.3	66.0

Table 7: **Hallucinated Fraction Ablation on AG-ReID.** Only novel views with less than the given fraction of hallucinated pixels are retained. mAP and Rank-1 (%) on Aerial → Ground and Ground → Aerial. Best in **bold**, second-best underlined.

Variant	Aerial Training				Ground Training			
	$A \to G$		$G \to A$		$A \to G$		$G \to A$	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
$< 10\%$ hallucinated	47.6	62.9	48.4	57.6	51.3	62.3	53.5	64.3
$< 30\%$ hallucinated	49.1	63.6	51.2	59.6	52.8	64.6	55.3	66.3
$< 50\%$ hallucinated	50.9	64.7	51.7	60.9	54.1	<u>66.4</u>	56.9	67.2
$< 70\%$ hallucinated	<u>51.5</u>	<u>65.8</u>	<u>52.5</u>	<u>61.3</u>	<u>55.3</u>	65.7	<u>57.6</u>	<u>67.7</u>
All views (ours)	**52.1**	**66.1**	**53.2**	**62.6**	**56.0**	**67.0**	**57.9**	**68.6**

A Supplementary Material

A.1 Visibility Map Analysis

Table 9 compares our mesh-derived visibility map against alternatives that ignore which regions were actually visible. The One maps apply uniform attention supervision, while the Foreground map restricts attention to the rendered subject without distinguishing visible from hallucinated regions. Both reach roughly the same accuracy as each other on AG-ReID Aerial and are well below our visibility map with a gap of $+1.5$ to $+1.7$ mAP pp on Aerial training. This shows that the gain from VAAS does not come from subject masking but from the per-patch reliability information that only the mesh can supply. Conversely, directing attention to the hallucinated regions falls below the model trained without attention supervision (Index 4 of Table 3) in every column, confirming that visible regions carry more identity signal than hallucinated ones.

Table 6 shows that hallucinated regions should nevertheless not be removed from the input. Hard-masking hallucinated pixels in the rendered views lowers accuracy in every column, since these regions still provide object context such as silhouette and pose. VAAS therefore guides the identity token toward visible patches without excluding hallucinated content from the representation.

A.2 Low-Visibility Samples

Table 7 retains only the novel views whose hallucinated ratio is below a threshold. Accuracy increases as more hallucinated views are retained, and keeping all views is best in every column. Low-visibility views come from the largest viewpoint shifts and are relevant for large viewpoint changes despite their lower share of visible pixels. For this reason, CVCF targets geometrically incorrect meshes rather than low-visibility views.

A.3 Calibration Noise

The visibility map depends on the calibrated source pose Π_i, for which no ground truth is available. We therefore measure the downstream effect of calibration errors by adding noise of up to $\pm n^\circ$ in azimuth and elevation to the fitted pose (Table 8). A $\pm 5^\circ$ perturbation leaves accuracy almost unchanged, while larger perturbations degrade it progressively. For scale, the fitted rotations average $21.6^\circ/27.0^\circ/0.5^\circ$ (azimuth / elevation / roll) on AG-ReID, $19.3^\circ/20.7^\circ/0.3^\circ$ on AG-ReID.v2, and $23.0^\circ/36.9^\circ/0.8^\circ$ on MOO. The visibility map is therefore robust to small calibration errors, and CVCF removes the meshes whose calibration fails.

A.4 Number of supervised layers

Table 10 tries different values of L, the number of transformer layers on which the CLS-attention loss is applied. Supervising only the last layer ($L = 1$) is best across all eight columns, with 52.1 mAP and 66.1 R-1 on Aerial

Table 8: **Calibration Noise Ablation on AG-ReID.** Noise of up to $\pm n^\circ$ in azimuth and elevation is added to the fitted source pose. mAP and Rank-1 (%) on Aerial $\rightarrow$ Ground and Ground $\rightarrow$ Aerial. Best in **bold**, second-best underlined.

Variant	Aerial Training				Ground Training			
	$A \rightarrow G$		$G \rightarrow A$		$A \rightarrow G$		$G \rightarrow A$	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
No noise (ours)	**52.1**	**66.1**	**53.2**	**62.6**	**56.0**	**67.0**	**57.9**	**68.6**
$\pm 5^\circ$	52.0	**66.1**	52.9	62.4	55.8	66.9	57.8	68.4
$\pm 10^\circ$	51.2	65.8	52.6	61.9	53.4	63.9	56.8	67.7
$\pm 20^\circ$	49.7	63.9	50.6	60.1	52.1	62.0	54.2	66.8
$\pm 30^\circ$	48.1	60.4	47.4	57.9	50.4	59.2	52.2	62.1

Table 9: **Visibility Mask Source Ablation on AG-ReID.** mAP and Rank-1 (%) on Aerial $\rightarrow$ Ground and Ground $\rightarrow$ Aerial. Best in **bold**, second-best underlined.

Variant	Aerial Training				Ground Training			
	$A \rightarrow G$		$G \rightarrow A$		$A \rightarrow G$		$G \rightarrow A$	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
One map	50.5	64.0	51.5	59.4	55.0	66.2	54.4	65.3
Foreground map	50.6	65.1	51.6	59.2	54.8	65.3	53.5	64.9
Hallucinated map	48.2	62.5	50.3	57.8	52.8	63.3	55.3	65.3
Visibility map (ours)	**52.1**	**66.1**	**53.2**	**62.6**	**56.0**	**67.0**	**57.9**	**68.6**

$A \rightarrow G$. Extending the supervision to more layers degrades performance until it stabilizes at $L = 4$ and beyond. The visibility map specifies where the final identity representation should attend, but constraining earlier layers removes their ability to build the global context that the last-layer attention then exploits.

A.5 Triplet loss weight

Table 11 sweeps the weight λ_{tri} of the visibility-weighted triplet loss. Setting $\lambda_{\text{tri}} = 0.25$ underuses the term and yields the weakest results (51.1 mAP and 59.6 R-1 on Aerial $A \rightarrow G$), while $\lambda_{\text{tri}} \in \{0.5, 0.75, 1.0\}$ all reach comparable accuracy. This shows that the visibility-weighted triplet loss is a key component of our method, but the exact weight is not critical as long as it is sufficiently high to influence training.

A.6 CLS-attention loss weight

Table 12 sweeps the weight λ_{vis} of the CLS-attention loss with $\lambda_{\text{tri}} = 1$ fixed. Again, setting $\lambda_{\text{vis}} = 0.25$ does not provide enough signal (51.3 mAP on Aerial $A \rightarrow G$) and higher values yield comparable performance. $\lambda_{\text{vis}} = 1$ is best overall, and we adopt it in our experiments.

A.7 Filtering ratio

Table 13 reports the effect of the filtering percentage, which controls the quantity of filtered novel meshes discarded. At low values (≤ 0.3) filtering is too permissive and low-quality synthesized samples with severe artifacts or poor geometric fidelity leak into training and drag performance down across both protocols, with 0.1 falling to 49.2 mAP on $A \rightarrow G$. At high values (≥ 0.7) filtering becomes overly aggressive and prunes informative views along with the noisy ones, eroding the diversity needed to bridge the aerial-to-ground viewpoint gap. We

Table 10: **Number of Supervised Attention Layers Ablation on AG-ReID.** mAP and Rank-1 (%) on Aerial $\rightarrow$ Ground and Ground $\rightarrow$ Aerial. Best in **bold**, second-best underlined.

Variant	Aerial Training				Ground Training			
	$A \rightarrow G$		$G \rightarrow A$		$A \rightarrow G$		$G \rightarrow A$	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
$L = 2$	51.2	64.4	52.3	61.4	55.2	65.4	56.4	66.1
$L = 4$	50.1	63.8	51.2	59.2	54.0	64.0	55.7	66.0
$L = 8$	50.0	63.8	51.2	59.2	53.8	64.8	55.6	66.4
$L = 12$	50.2	63.7	51.0	59.5	53.6	64.7	55.7	65.7
$L = 1$ (ours)	**52.1**	**66.1**	**53.2**	**62.6**	**56.0**	**67.0**	**57.9**	**68.6**

Table 11: **Triplet Loss Weight λ_{tri} Ablation on AG-ReID.** mAP and Rank-1 (%) on Aerial → Ground and Ground → Aerial. Best in **bold**, second-best underlined.

Variant	Aerial Training				Ground Training			
	$A \to G$		$G \to A$		$A \to G$		$G \to A$	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
λ_{tri}=0.25	51.1	59.6	49.7	59.4	54.4	65.7	56.6	66.4
λ_{tri}=0.50	51.5	65.8	52.9	61.3	55.9	66.9	55.3	66.5
λ_{tri}=0.75	51.9	64.8	52.9	61.3	55.8	66.9	57.8	**68.9**
λ_{tri}=1.00 (ours)	**52.1**	**66.1**	**53.2**	**62.6**	**56.0**	**67.0**	**57.9**	68.6

Table 12: **CLS-Attention Loss Weight λ_{vis} Ablation on AG-ReID.** mAP and Rank-1 (%) on Aerial → Ground and Ground → Aerial. Best in **bold**, second-best underlined.

Variant	Aerial Training				Ground Training			
	$A \to G$		$G \to A$		$A \to G$		$G \to A$	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
λ_{vis}=0.25	51.3	64.8	51.9	59.5	54.2	65.2	56.5	65.9
λ_{vis}=0.50	51.8	65.7	52.0	60.6	55.9	66.8	57.5	68.5
λ_{vis}=0.75	51.9	65.5	52.7	61.1	55.6	66.1	57.4	68.5
λ_{vis}=1.00 (ours)	**52.1**	**66.1**	**53.2**	**62.6**	**56.0**	**67.0**	**57.9**	**68.6**

chose to filter the bottom 50% of samples, which is a good compromise between quality and diversity. However, the ideal threshold depends on the complexity of the dataset and on the robustness of the 3D reconstruction method, which may require tuning for different settings.

A.8 Computational Cost

The initial 3D lifting of source images is the most computationally intensive task of our pipeline, but it amortizes the overall cost by enabling rendering of multiple novel views per mesh. Crucially, this cost is incurred only at training time. At inference, 3D-VITAL uses the same ViT backbone as the baseline, since our components only modify the training data and loss function.

Specifically, for AG-ReID, the 3D reconstruction of 10,000 source images (all successfully segmented images of the training set) requires 24 hours on 8 A100 GPUs. Rendering 100,000 novel views and their associated visibility maps (10 views for each of the 10,000 meshes) takes only 8 hours on one P5000 GPU. Finally, the StyleID step requires 8 hours on 8 A100 GPUs. The final ReID training with 3D-VITAL is equivalent to the ViT baseline, as our components are not computationally intensive and do not require additional forward passes at inference time. Including the remaining generation stages, the pipeline requires 272 A100 and 8 P5000 GPU-hours for the entire AG-ReID training set. This cost is a one-time investment that is independent of downstream ReID training.

For rough context, diffusion-based augmentation methods such as Diff-Person [17] require per-identity fine-tuning of a diffusion model and report three days on 8 RTX 4090 GPUs (576 GPU-hours) for 5,000 identities, which must be repeated whenever the identity set changes. They do not report the cost of filtering their generated

Table 13: **Filtering Ratio Ablation on AG-ReID.** Percentage of filtered novel meshes. mAP and Rank-1 (%) on Aerial → Ground and Ground → Aerial. Best in **bold**, second-best underlined.

Variant	Aerial Training				Ground Training			
	$A \to G$		$G \to A$		$A \to G$		$G \to A$	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
0.10	49.2	62.1	45.9	53.3	53.1	62.5	56.2	68.1
0.20	50.6	62.3	51.5	59.5	55.3	66.0	57.1	**68.7**
0.30	51.7	63.9	52.9	61.8	55.3	65.3	57.3	67.2
0.40	51.8	66.0	53.0	62.4	55.3	65.9	57.6	67.8
0.50 (ours)	**52.1**	**66.1**	**53.2**	**62.6**	**56.0**	**67.0**	**57.9**	68.6
0.60	51.1	65.4	53.0	62.0	55.3	66.2	57.8	68.5
0.70	51.2	66.0	53.0	61.1	55.5	66.9	57.7	68.1
0.80	50.8	65.5	52.8	60.9	54.6	65.5	57.3	68.3
0.90	50.1	64.2	52.3	61.0	53.2	64.2	55.5	65.7

samples within the paper. A100 and RTX 4090 GPUs are not directly comparable in compute throughput, so these numbers should be read as orders of magnitude rather than a precise ratio.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction of this paper accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations of this work in Section 5.

   Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We have introduced all the details in our theoretical result.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: This paper has provided detailed information for reproducing the experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but

reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code will be available at `url`

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: This paper has specified all the training and test details, such as dataset and hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: This paper does not provide error bars due to the limited resources. Also, previous methods do not provide error bars either

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This paper has provided sufficient information on the computer resources, see Appendix A.8

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper meets the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper proposes a training method evaluated on existing public benchmarks and introduces no new data.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks since it does not release data or models that have risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This paper has properly cited the original paper that produced the code package or dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: This paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: This paper only uses LLMs for writing and editing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.