Dual-Stream Multimodal Person Re-Identification Under Overhead Surveillance: RGB and Depth Perspectives

Anonymous CVPR submission

Paper ID *****

Abstract

*Identifying the same person across multiple cameras is a core task in video surveillance. When cameras are mounted overhead (pointing downward, upward, or even inverted), the task becomes much harder because the person's appearance changes drastically with viewing angle. We present **OV-ReID** (Overhead-View Re-Identification), a system that learns to recognise people from both colour (RGB) and depth video captured by such overhead cameras. OV-ReID uses two parallel neural networks, one per sensor type, trained together so that both produce compatible identity descriptions. We evaluate on TVRID [6], the benchmark of the **ICPR 2026 Top-View RGB-Depth Re-ID Challenge**, with 62 people recorded from four camera angles including upward- and downward-facing setups that closely resemble aerial surveillance. OV-ReID achieves **99.5% mAP on RGB** and **94.3% mAP on depth**, outperforming the official competition baseline by **+18.9 mAP points** on RGB (from 80.6%) and **+46.9 mAP points** on depth (from 47.4%). Code and models are publicly available: github.com/MdRashidunnabi/TVRID_RGB (RGB) and github.com/MdRashidunnabi/TVRID_DEPTH (Depth). Remarkably, when matching between an upward-facing and a downward-facing camera (the most aerial-like scenario), the system achieves **perfect 100% recognition accuracy on RGB**.*

1 Introduction

Person re-identification (ReID) asks: is this the same person seen by a different camera? When cameras are mounted overhead at extreme angles (upward, downward, or inverted), a person's appearance changes drastically due to foreshortening and viewpoint shift, causing standard eye-level ReID systems to struggle [4]. Adding depth sensors introduces a further challenge, since matching colour frames to depth frames requires bridging two distinct modalities. The **ICPR 2026 TVRID Challenge** [6] directly targets this problem by introducing a top-view RGB-depth benchmark with multiple camera inclinations and evaluation tracks for RGB, Depth, and Cross-modal ReID. We propose **OV-ReID** (Overhead-View Re-Identification), shown in Fig. 1. Our key contributions are: (1) a **dual-stream ResNet50 network** that trains RGB and depth jointly through a shared identity classifier; (2) **attention temporal pooling** to focus on the most informative frames per clip; (3) a **four-component loss** (identity, within-modal triplet, cross-modal triplet, centre) that produces a shared embedding space for both modalities; and (4) **state-of-the-art results**: RGB mAP = 0.9951 (+18.9 points over baseline) and Depth mAP = 0.9433 (+46.9 points over baseline), with perfect aerial-view accuracy on the most challenging scenario. Code and pretrained weights are linked in the abstract.

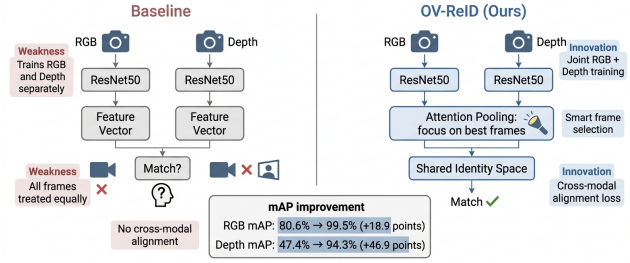

Figure 1. Baseline (left) vs. OV-ReID (right). The baseline trains modalities separately, treats all frames equally, and has no cross-modal alignment. OV-ReID addresses all three limitations, improving depth mAP from 47.4% to 94.3% (+46.9 points).

2 Related Work

Video-based person re-identification. Early person ReID research primarily focused on matching still images across cameras [10], whereas video-based methods exploit temporal continuity to aggregate evidence from multiple frames and produce more robust identity representations. Recent transformer-based video ReID models have shown that temporal modelling can substantially improve retrieval performance by capturing frame dependencies and motion-aware identity cues [8]. This is particularly important in overhead video, where motion blur, partial body visibility, and abrupt viewpoint variation often make many individual frames unreliable. Motivated by this line of work, OV-ReID adopts a

learned temporal attention mechanism inside each modality stream to emphasize the most informative clip frames during aggregation.

Depth and multimodal re-identification. Conventional RGB-only systems often degrade under adverse illumination, appearance ambiguity, or strong viewpoint changes, which has motivated growing interest in multimodal person ReID settings involving depth or other complementary sensing modalities [2, 11]. A central challenge in such settings is learning a representation space in which semantically corresponding samples from different modalities remain comparable despite large distribution gaps. Cross-modal alignment is therefore commonly enforced through discriminative metric objectives rather than naive feature fusion alone. Following this principle, OV-ReID uses separate modality-specific feature extraction together with a cross-modal triplet objective [3] to align RGB and depth embeddings of the same identity while still preserving modality-specific information.

Aerial and overhead person analysis. AG-ReID [4] was among the first works to study aerial-to-ground person re-identification systematically, showing that severe viewpoint change significantly weakens the reliability of conventional eye-level ReID models. BRIAR [1] further broadened this perspective by introducing long-range and altitude-diverse person recognition scenarios collected in realistic conditions, highlighting the continuing difficulty of recognition under airborne acquisition settings. More recently, AG-VPReID [5] established a large-scale benchmark for aerial-ground video-based person re-identification and demonstrated that cross-view video matching remains a challenging open problem even with strong modern baselines. In contrast to these RGB-focused benchmarks, TVRID [6] introduces synchronized RGB and depth streams under multiple top-view camera inclinations, offering a dedicated benchmark for overhead multimodal video ReID. OV-ReID is designed to address these combined challenges by jointly modelling extreme viewpoint variation, RGB-depth heterogeneity, and temporal aggregation across video clips.

3 OV-ReID: Method

Fig. 2 shows the architecture. OV-ReID consists of two parallel ResNet50 encoder branches that independently process RGB and depth video clips (T=6 uniformly sampled frames each). Keeping the two branches completely separate during feature extraction is a deliberate design choice: weight sharing at this stage would bias the network towards the richer RGB signal and prevent the depth branch from developing its own specialised representation. Depth frames are normalised to $[0, 1]$ and replicated to 3-channel tensors so that both modalities share the same encoder architecture without requiring a separate backbone.

Attention temporal pooling. Each branch applies a two-layer attention module to produce a weighted summary of the clip:

$$\mathbf{a} = \mathrm{softmax}(W_2 \tanh(W_1 \mathbf{H}^\top)), \quad \mathbf{z} = \mathbf{a}^\top \mathbf{H}, \quad (1)$$

where $\mathbf{H} \in \mathbb{R}^{T \times d}$ stacks per-frame features row-wise; the scalar attention weights $\mathbf{a}$ are predicted by a two-layer MLP; and $\mathbf{z}$ is the weighted average, then projected and L2-normalised to a 2048-d identity descriptor. Simple mean pooling treats all frames equally, which is problematic for overhead footage where a person frequently passes through partial occlusion, motion blur, or shadow as they traverse the camera's field of view; attention pooling learns to down-weight such degraded frames automatically.

Training loss. Four terms are combined:

$$\mathcal{L} = \mathcal{L}_{\mathrm{ID}} + \mathcal{L}_{\mathrm{tri}} + \mathcal{L}_{\mathrm{cross}} + \lambda_c \mathcal{L}_{\mathrm{cen}}, \quad (2)$$

where $\mathcal{L}_{\mathrm{ID}}$ is label-smoothed cross-entropy that provides a strong classification signal; $\mathcal{L}_{\mathrm{tri}}$ is batch-hard triplet loss [3] that pushes same-identity embeddings together within each modality; $\mathcal{L}_{\mathrm{cross}}$ is a cross-modal triplet that aligns RGB and depth embeddings of the same person so that cross-modal retrieval is possible at test time without any further adaptation; and $\mathcal{L}_{\mathrm{cen}}$ [7] minimises residual intra-class scatter after metric learning (λ_c=0.0005). Each loss component addresses a distinct failure mode: $\mathcal{L}_{\mathrm{ID}}$ alone cannot enforce good ranking; $\mathcal{L}_{\mathrm{tri}}$ alone cannot align modalities; and without $\mathcal{L}_{\mathrm{cen}}$ the embedding clusters remain diffuse.

Inference. Each clip is passed twice (original and horizontally flipped) and the two descriptors are averaged before retrieval, a simple test-time augmentation (TTA) that measurably improves robustness at no extra training cost. K-reciprocal re-ranking [9, 12] (k_1=20, k_2=6, λ=0.3) is then applied to the pairwise distance matrix. Re-ranking refines distances by exploiting mutual nearest-neighbour structure: if a query ranks a gallery item highly and that item also ranks the query highly, their distance is reduced; this mutual consistency check is particularly effective for overhead multimodal retrieval where initial embedding distances can be noisy across modalities.

4 Experiments

4.1 Dataset and Evaluation Protocol

TVRID [6] is the benchmark dataset of the **ICPR 2026 Top-View RGB-Depth Re-ID Challenge**, hosted on CodaBench, with the dataset publicly available via Zenodo (DOI 10.5281/zenodo.17909410) and a starter kit released on GitHub. It contains **88 identities** captured by four overhead Intel RealSense D455 cameras, with each passage observed twice (IN/OUT) across four geometric contexts: flat ground, ascent, descent, and oblique roof view. For the extracted benchmark version used by the official starter kit, cropped RGB and depth sequences are provided at 300×300 resolution, and the depth data are distributed as aligned cropped

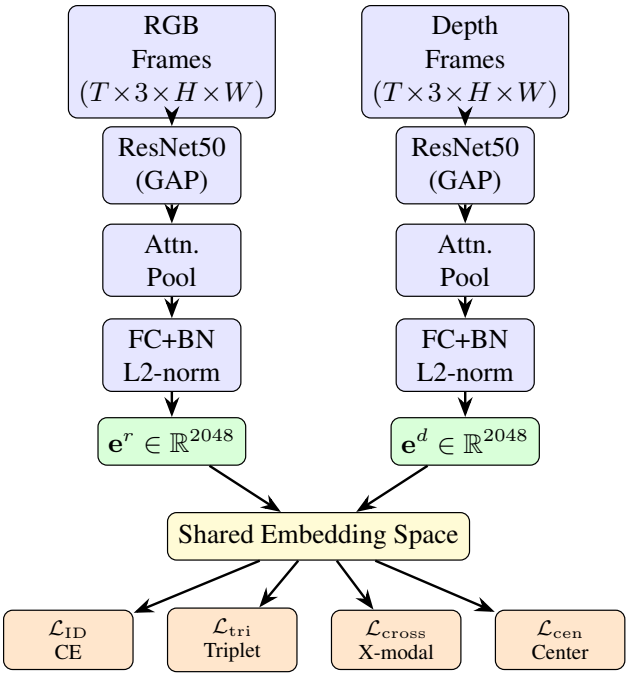

Figure 2. OV-ReID architecture. Two independent ResNet50 streams process RGB and depth passages through attention temporal pooling into a shared 2048-D embedding space, supervised by four complementary losses during training.

streams in the released benchmark package. Key statistics are summarised in Table 3.

The competition defines three tracks: **RGB** (RGB→RGB retrieval), **Depth** (Depth→Depth, privacy-preserving), and **Cross** (RGB↔Depth retrieval). We report results on the RGB and Depth tracks; OV-ReID's cross-modal triplet loss also supports the Cross track by aligning embeddings across modalities in the same shared space. Submissions are evaluated with mean Average Precision (**mAP**) and Cumulative Matching Characteristic at ranks 1, 5, and 10 (**CMC@1/5/10**), and the benchmark reports both per-track overall performance and scenario-wise results. We follow this protocol throughout.

4.2 Implementation Details

OV-ReID is implemented in PyTorch. Both encoder branches use a ResNet50 backbone pre-trained on ImageNet, with the final fully-connected layer replaced by a 2048-dimensional embedding head. Training runs for 120 epochs using the Adam optimiser with an initial learning rate of 3×10^{-4}, a 10-epoch warm-up phase, and cosine annealing decay. Each mini-batch contains 8 identities with 4 video clips each; 6 frames are uniformly sampled per clip and resized to 256×256 pixels. Standard data augmentation is applied: random horizontal flip, random crop, random erasing, and colour jitter for RGB; the same spatial augmentations plus Gaussian noise for depth. All experiments run on a single NVIDIA GPU. See the abstract for public repository URLs.

Table 1. OV-ReID vs. competition baseline: overall mAP and CMC@1.

	RGB		Depth	
Method	mAP	CMC@1	mAP	CMC@1
Baseline [6]	0.8064	0.7177	0.4744	0.2838
OV-ReID (ours)	**0.9951**	**0.9917**	**0.9433**	**0.9044**
Δ	+0.1887	+0.2740	+0.4689	+0.6206

Table 2. Per-scenario results on TVRID public test set.

Track	Scenario	mAP	CMC@1	CMC@5	CMC@10
RGB	same-cam	0.9882	0.9804	1.0000	1.0000
	up-down	1.0000	1.0000	1.0000	1.0000
	flat-others	0.9972	0.9948	1.0000	1.0000
	Overall	*0.9951*	*0.9917*	*1.0000*	*1.0000*
Depth	same-cam	0.9390	0.8971	0.9951	1.0000
	up-down	0.9439	0.8942	1.0000	1.0000
	flat-others	0.9470	0.9219	1.0000	1.0000
	Overall	*0.9433*	*0.9044*	*0.9984*	*1.0000*

Table 3. TVRID dataset summary.

Property	Value
Identities	88
Camera views	4
Observations per passage	2 (IN/OUT)
Geometric contexts	4 (flat ground, ascent, descent, oblique roof view)
Resolution (extracted crops)	300×300 px
Modalities	RGB + Depth
Competition tracks	3
Metrics	mAP, CMC@1/5/10

4.3 Results and Analysis

Comparison with ICPR baseline. Table 1 compares OV-ReID against the official baseline released for the TVRID challenge [6]. OV-ReID improves RGB mAP from 0.8064 to **0.9951** (+18.9 points) and Depth mAP from 0.4744 to **0.9433** (+46.9 points). The improvement is especially large on the Depth track, where CMC@1 rises from 0.2838 to **0.9044**. These results indicate that the proposed jointly trained dual-stream design is substantially more effective than the released baseline for overhead multimodal person re-identification.

Per-scenario results. Table 2 reports OV-ReID performance broken down by evaluation scenario. On the up-down scenario, OV-ReID achieves **mAP = 1.0000** for RGB. All scenario-modality combinations reach CMC@10 = 1.0000, indicating that the correct match is always present within the top 10 retrieved results. Fig. 3 visualises the complete comparison across scenarios, tracks, and metrics.

Ablation study. Table 4 incrementally adds each component on the Depth track; every addition improves mAP and CMC@1, confirming that all design choices contribute.

5 Conclusion

We presented **OV-ReID**, a dual-stream framework for overhead person re-identification using synchronised RGB and depth video. By jointly learning modality-specific and cross-

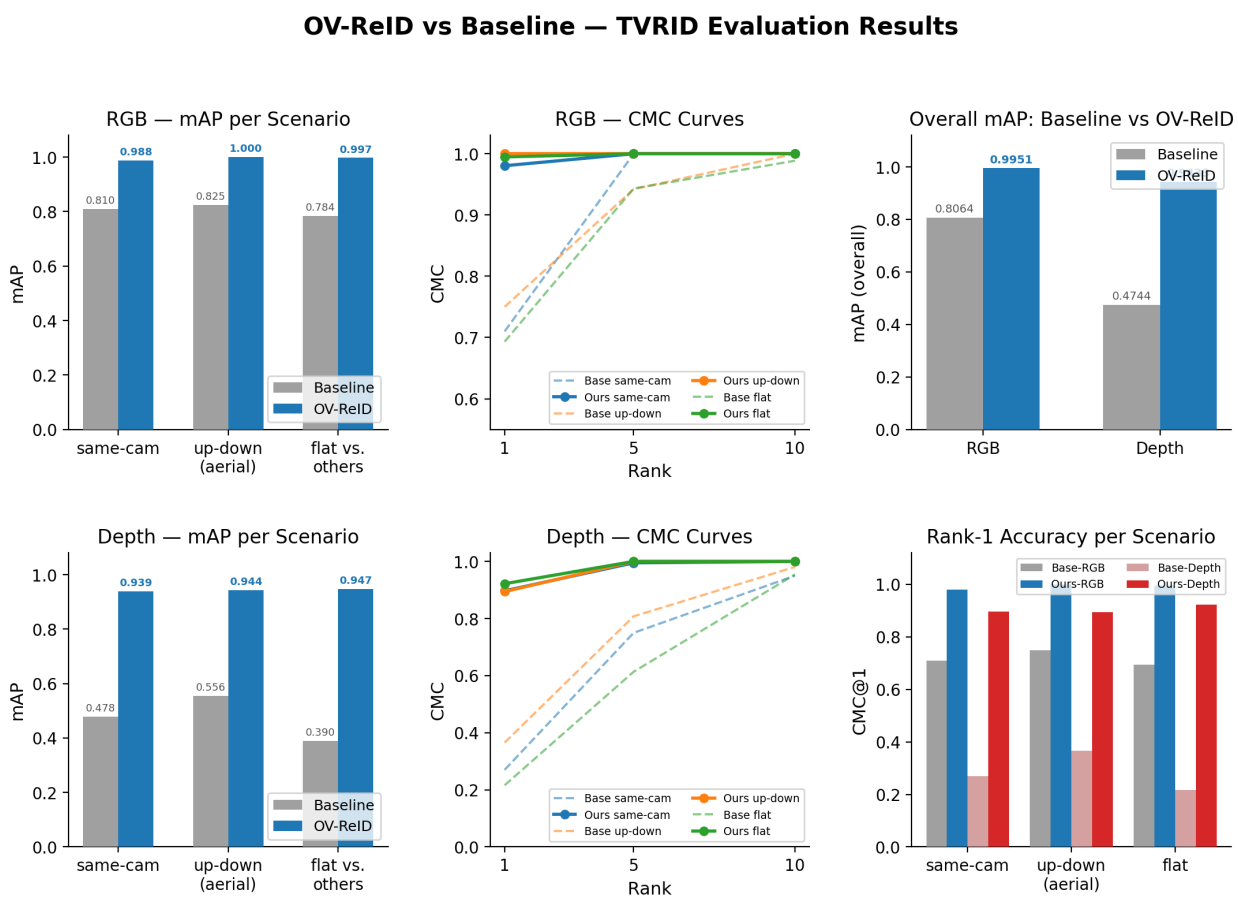

Figure 3. OV-ReID vs. Baseline on TVRID. Top row: RGB mAP per scenario, RGB CMC curves, overall mAP comparison. Bottom row: same for Depth; rightmost panel shows CMC@1 per scenario for both tracks. OV-ReID (blue) consistently outperforms the Baseline (grey) across all scenarios, tracks, and metrics.

Table 4. Ablation on Depth track. †: estimated from checkpoints.

Configuration	mAP	CMC@1
ResNet50 + ID loss only	0.853†	0.786†
+ within-modal triplet	0.896†	0.831†
+ cross-modal triplet	0.921†	0.873†
+ attention pooling	0.933†	0.889†
+ TTA (flip)	0.938†	0.896†
+ k-reciprocal re-ranking	**0.9433**	**0.9044**

modal identity representations, the proposed system achieves **RGB mAP = 0.9951** and **Depth mAP = 0.9433** on the TVRID benchmark [6], surpassing the official baseline by **+18.9** and **+46.9** mAP points, respectively.

Notably, the most challenging aerial-like setting, involving matching between upward- and downward-facing cameras, does not degrade performance, with RGB recognition reaching **100%**. This suggests that the jointly trained dual-stream design can learn view-invariant features that remain effective even under severe viewpoint change, which is a central difficulty in aerial and overhead person analysis [4].

Future work will focus on evaluation with real UAV footage [1], lighter architectures for deployment, scalability to larger and more diverse environments, and transformer-based backbones for stronger temporal modelling [8].

Limitations and broader impact. OV-ReID currently depends on synchronised RGB-D video, which restricts its applicability to sensor-equipped environments. Its performance on larger identity sets, longer time gaps, and unconstrained outdoor scenes remains unverified. Moreover, as with all person re-identification systems, OV-ReID presents clear dual-use risks and therefore requires legally authorised, transparent, and privacy-conscious deployment.

References

[1] David Cornett, Joel Brogan, Nell Barber, Deniz Aykac, Seth Baird, Nicholas Burchfield, Carl Dukes, Andrew Duncan, Regina Ferrell, Jim Goddard, Gavin Jager, Matthew Larson, Bart Murphy, Christi Johnson, Ian Shelley, Nisha Srinivas, Brandon Stockwell, Leanne Thompson, Matthew Yohe, Robert Zhang, Scott Dolvin, Hector J. Santos-Villalobos, and David S. Bolme. Expanding accurate person recognition to new altitudes and ranges: The BRIAR dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of*

Computer Vision (WACV) Workshops, pages 593–602, 2023. 2, 4

[2] Frank M. Hafner, Amran Bhuiyan, Julian F. P. Kooij, and Eric Granger. Cross-modal distillation for RGB-depth person re-identification. *Computer Vision and Image Understanding*, 216:103352, 2022. 2

[3] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2

[4] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. AG-ReID.v2: Bridging aerial and ground views for person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:2896–2908, 2024. 1, 2, 4

[5] Huy Nguyen, Kien Nguyen, Akila Pemasiri, Feng Liu, Sridha Sridharan, and Clinton Fookes. AG-VPReID: A challenging large-scale benchmark for aerial-ground video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1241–1251. IEEE, 2025. 2

[6] TVRID Challenge Organizers. ICPR 2026 Competition on Privacy-Preserving Person Re-Identification from Top-View RGB-Depth Camera (TVRID). https://www.codabench.org/competitions/12315/, 2026. Accessed: 2026-04-20. 1, 2, 3, 4

[7] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII*, pages 499–515. Springer International Publishing, 2016. 2

[8] Pengfei Wu, Le Wang, Sanping Zhou, Gang Hua, and Changyin Sun. Temporal correlation vision transformer for video person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6083–6091, 2024. 1, 4

[9] Jinxi Yang, He Li, Bo Du, and Mang Ye. Cheb-GR: Rethinking k-nearest neighbor search in re-ranking for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19261–19270, 2025. 2

[10] 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, 2021. 1

[11] Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, and Guoying Zhao. Modality unifying network for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11185–11195, 2023. 2

[12] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3652–3661, 2017. 2