Advanced Methods for Top-View RGB-D Person Re-ID (TVRID)

Bipul Gyawali
Howard University, bipulgyawali02@gmail.com
Saurav K. Aryal
Howard University, saurav.aryal@howard.edu

This paper presents a robust framework for **Top-View RGB-D Person Re-Identification (TVRID)**, addressing the specific challenges of the ICPR 2026 competition. Our approach integrates three specialized tracks: an **RGB track** utilizing part-based attention and ResNet backbones (e.g., ViT, Swin) to improve robustness to occlusion; a **Depth track** focusing on body-shape cues learned from 1-channel depth with attention and metric losses; and a **Cross-Modal track** employing dual-stream fusion with optional cross-attention and cross-modal metric losses. By combining identity (ArcFace) and metric (batch-hard triplet, center) losses within a **PyTorch Lightning** framework, our method achieves strong discriminability across same-camera and cross-passage scenarios.

CCS Concepts: Computing methodologies → Computer vision; Computing methodologies → Machine learning approaches; Security and privacy → Privacy-preserving protocols.

Keywords: Top-View Re-ID, RGB-D Fusion, Part-based Attention, Cross-Modal Distillation, PyTorch Lightning, ICPR 2026.

Approach Overview

1. RGB Re-ID Track

- **Part-based models:** We use BPBreID-inspired part-level features via horizontal strip-based part attention, which improves recall in occluded top-view settings (optional use_parts in config).
- **Backbones:** ResNet50 by default; optional Transformer-based backbones (ViT, Swin via timm) to capture long-range spatial dependencies.
- **Attention & pooling:** CBAM or ECA attention, GeM pooling, and BNNeck for stable metric learning.
- **Losses:** ArcFace (additive angular margin), batch-hard triplet loss, and center loss to tighten feature clusters and inter-class margins.

2. Depth Re-ID Track

- **Privacy-preserving depth:** Depth maps are processed as 1-channel inputs; the network learns body-shape and anthropometric-style cues from data rather than hand-crafted geometric features.
- **Modeling:** A ResNet backbone adapted for 1-channel depth, with GeM pooling, BNNeck, and channel/spatial attention (CBAM/ECA). Depth-specific normalization is applied in the data pipeline.
- **Losses:** Same identity and metric setup as the RGB track (ArcFace, batch-hard triplet, center loss).

3. Cross-Modal RGB–Depth Track

- **Fusion strategy:** Dual-stream network (separate RGB and depth encoders) with a fusion module that combines embeddings via **concatenation** or **cross-attention**, aligned with related work on Transformer-based fusion (e.g., CAMI-style cross-modal attention).
- **Feature learning:** Identity loss is applied on each modality and on the fused embedding; **cross-modal triplet loss** enforces agreement between RGB and depth for the same identity. Optional **cross-modal distillation** (L2 or KL) can be enabled

to further align the two streams.

- **Shared-specific:** The codebase includes building blocks for shared-specific feature decomposition; the main cross-modal model uses dual-stream fusion and cross-modal metric losses.

Results

The performance of the proposed framework is evaluated based on the Mean Average Precision (mAP) across three evaluation tracks. The system achieves a Mean Overall mAP of 0.4847.

In the individual tracks, the RGB track achieves the highest performance with an overall mAP of 0.7823, followed by the Depth track with an overall mAP of 0.3608, and the Cross-Modal track with an overall mAP of 0.3109.

References

1. Somers, V., De Vleeschouwer, C., & Alahi, A. (2022). Body Part-Based Representation Learning for Occluded Person Re-Identification. *ArXiv*. https://doi.org/10.1109/WACV56688.2023.00166
2. He, S., Luo, H., Wang, P., Wang, F., Li, H., & Jiang, W. (2021). TransReID: Transformer-based Object Re-Identification. *ArXiv*. https://arxiv.org/abs/2102.04378 .
3. Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., & Yu, N. (2020). Cross-modality Person re-identification with Shared-Specific Feature Transfer. *ArXiv*. https://arxiv.org/abs/2002.12489
4. Hafner, F., Bhuiyan, A., Kooij, J. F., & Granger, E. (2018). Cross-Modal Distillation for RGB-Depth Person Re-Identification. *ArXiv*. https://arxiv.org/abs/1810.11641