VRCLIP: Multimodal Canonical Correlation Alignment for CLIP-Driven Vision-Radio Person Re-Identification

Supplementary Material

In this supplementary material, we provide further details to accompany our main paper. This document includes:

- A detailed introduction to our newly collected **Visual-Radio ReID (VRR) dataset**. We describe its sensor platform, data acquisition protocols, and unique advantages as the first large-scale benchmark for this task.
- **Supplementary visual results**. We feature classification confusion matrices and t-SNE visualizations of the learned embeddings to further demonstrate our model's discriminative power.
- A comprehensive list of **implementation details and hyperparameters**, outlining the complete experimental setup to ensure full reproducibility.
- An analysis of the **multimodal fusion paradigm**. This comparison against vision-only and radio-only baselines proves that fusing their complementary information is essential for robust performance.

6. (Appendix-A1) Dataset: VRR

The advancement of vision-radio ReID is significantly hampered by the lack of specialized, large-scale public datasets. To bridge this gap, we introduce the Vision-Radio ReID (VRR) dataset, the first and largest collection of its kind designed for robust, multimodal person ReID research. VRR comprises 651,200 temporally synchronized RF-optical frame pairs, totaling over 15 hours of continuous data. It features comprehensive annotations, including fine-grained identity labels and precise 2D positional data, setting a new benchmark for developing and evaluating algorithms in this domain.

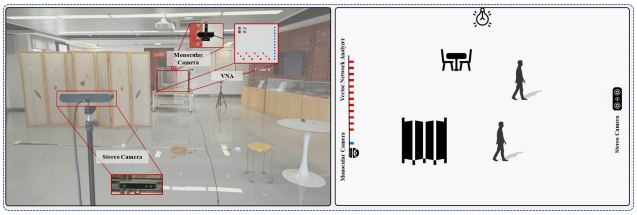

Figure 9. Experiment scenario and multimodal dataset collection system. Rx are the receiving antennas, and Tx are the transmitting antennas.

6.1. Sensor Platform

Our data acquisition platform, shown in Figure 9, integrates three synchronized sensor systems:

UWB MIMO Radar: The core of our system is a custom ultra-wideband (UWB) radar operating from 0.8 to 2.8 GHz. It consists of a vector network analyzer (VNA), a 12-transmit and 12-receive (12T12R) antenna array, and a power divider. The carefully designed antenna layout forms 144 virtual antennas to suppress sidelobes. By generating stepped-frequency continuous-wave (SFCW) signals, the system achieves a high range resolution of approximately 75 mm and excellent obstacle penetration, which is critical for ReID in occluded scenarios.

Optical and Depth Sensors: High-resolution RGB images (1280×720) are captured using a standard monocular camera to provide detailed visual appearance and texture cues. Depth information is obtained from a stereo vision system based on a ZED camera, enabling accurate 2D positional annotations of each subject for ground-truth generation.

All sensors are co-located and synchronized to a local time server using the Network Time Protocol (NTP). To initiate a recording session, the server broadcasts a target timestamp over a TCP connection. Each device begins data acquisition simultaneously upon reaching this timestamp, ensuring millisecond-level synchronization across the different modalities.

6.2. Unique Advantages of VRR

Compared to existing ReID datasets that focus on optical modalities (e.g., RGB-Infrared), VRR offers several unique advantages for pushing the frontiers of person ReID in challenging real-world conditions. *(1) First and Largest Vision-Radio ReID Dataset.* To our knowledge, VRR is the first publicly available large-scale dataset for vision-radio person ReID. Its scale (651k frames, 31 subjects) provides sufficient data to train and validate deep learning models that can effectively learn and fuse features from these two highly distinct modalities. *(2) Unprecedented Diversity in Scenarios and Subjects.* We collected data across numerous indoor and outdoor locations. Indoor scenes feature varying levels of obstruction, including unoccluded, partially occluded (e.g., by furniture), and fully occluded (e.g., by wooden furnitures) conditions. Outdoor data was captured during the morning, afternoon, and night to encompass diverse and challenging illumination profiles. Besides, the dataset features 31 unique participants with diverse demographic characteristics, including variations in age, gender, and body morphology, which helps mitigate model bias and improve generalization. *(3) Rich Modalities with Annotations.* VRR provides high-quality, temporally aligned data

streams. The combination of high-resolution optical images and obstacle-penetrating RF signals facilitates research into novel fusion strategies. Furthermore, every frame is annotated with both the person's identity and their precise 2D position, enabling comprehensive evaluation of both ReID and localization performance.

7. (Appendix-A2) Supplementary Visual Results

This appendix provides additional visualizations to further validate our model's fine-grained discrimination and feature learning capabilities, as mentioned in the main paper.

7.1. Classification Confusion Matrix

To evaluate the model's fine-grained discrimination capabilities across different identities, we present the classification confusion matrix in Figure 10. The matrix clearly shows a strong diagonal concentration of values, indicating that the model can distinguish between the vast majority of identity classes with high accuracy. The low off-diagonal values suggest minimal confusion between different individuals, validating the model's excellent classification performance and generalization ability.

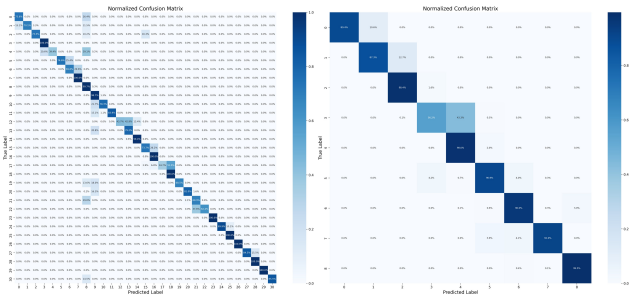

Figure 10. Confusion matrices for person ReID on the VRR (left) and THP (right) datasets.

7.2. t-SNE Visualization of Feature Embeddings

Figure 11 shows a t-SNE visualization of the learned feature embeddings. In this figure, each color denotes a specific lighting quality. The plot reveals that the model learns to distinguish between different illumination conditions in the embedding space, causing features from the same condition to form tight clusters that are well-separated from one another. Crucially, despite this general separability, the feature centroid of the RF signal is shown to be closely aligned with that of the normal lighting condition (RGB 1.0), validating our model's cross-modal calibration capabilities.

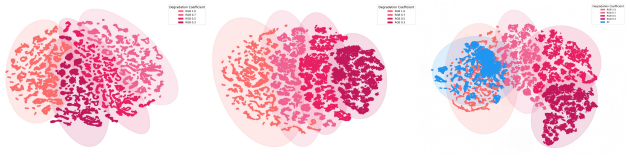

Figure 11. t-SNE visualizations of feature embeddings: (left) initial CLIP visual encoder features, (center) fine-tuned CLIP visual encoder features, and (right) aligned RF and normal lighting condition feature centroids.

8. (Appendix-A3) Detailed Implementation and Hyperparameters

All methods are implemented using the PyTorch framework and trained on a single NVIDIA RTX 4090 GPU. We employ the AdamW optimizer with an initial learning rate of 3×10^{-4} and a weight decay of 1×10^{-4}. A dynamic learning rate schedule is adopted, consisting of a 10% linear warm-up phase followed by a cosine annealing decay down to a minimum learning rate of 1×10^{-6}. This schedule is implemented via a custom LambdaLR scheduler for precise control over the learning rate trajectory. The models are trained for a total of 50 epochs with a batch size of 50. For the data split, a random 10-second segment (approximately 360 sample pairs) was selected for each identity class as training data, while the remaining samples were utilized for testing. To ensure the reproducibility of our results, all experiments are conducted with a fixed random seed of 42.

9. (Appendix-A4) Effect of Multimodal Fusion

To validate the fundamental necessity of our multimodal paradigm, we analyze the performance of unimodal baselines. As presented in Table 3 of the main paper, relying solely on the vision modality (Method A) or the radio modality (Method B) results in a drastic performance collapse, with mAP scores plummeting by 16.1% and 17.7%, respectively. This substantial decline underscores the inherent limitations of a single sensor; the vision modality is susceptible to challenging environmental factors like poor illumination, while the radio modality, though robust, may lack fine-grained appearance details. The remarkable performance leap of our full model demonstrates a powerful synergistic effect, where the fusion of complementary information from both modalities effectively overcomes their individual shortcomings to achieve a far more robust and accurate ReID capability.