UCM-VeID V2: A Richer Dataset and A Pre-training Method for UAV Cross-Modality Vehicle Re-Identification

Supplementary Material

7. Details of established dataset

7.1. Data collection

Different from the UCM-VeID, our dataset was collected using three UAV platforms equipped with five cameras to capture images at two adjacent locations. As depicted in Fig. 6, the use of different drones enables the collection of target information from various angles, such as the side, back, and front. The three UAV platforms include the DJI Mini 2, DJI Mini 3 Pro, and DJI M300 RTK. The DJI Mini 2 and Mini 3 Pro utilize their built-in imaging modules to capture high-resolution images at 3840×2160. The DJI M300 RTK is equipped with ZENMUSE H20T which contains an infrared camera with an imaging resolution of 640×512, a wide-angle camera and a zoom camera with an imaging resolution of 1920×1080. To enrich the diversity of the dataset, the UAVs operate in two modes, cruise and spot rotation, capturing vehicles at over eight different locations with complex backgrounds, while flying at altitudes ranging from 30 to 100 meters.

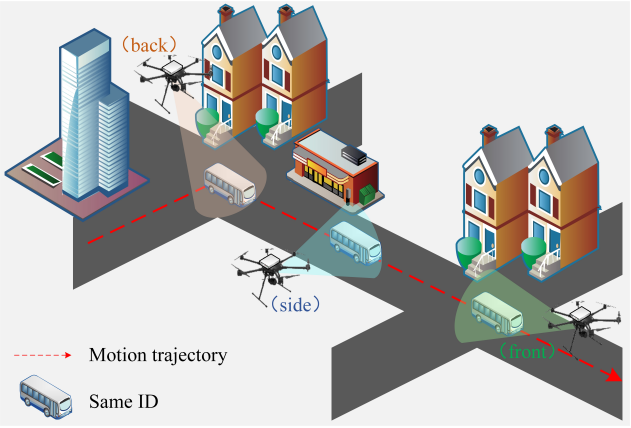


Figure 6. Diagram of data acquisition process.

Data were collected from various locations in Changsha, China, encompassing a wide range of scenarios, weather conditions, and time periods to maximize sample diversity, as shown in Fig. 7. Specifically, the collection included diverse road types such as urban arterial roads, residential areas, viaducts, tunnels, bridges, rural paths, and suburban muddy roads. To ensure data comprehensiveness and representativeness, we captured images under various weather conditions, including sunny, cloudy, foggy, and rainy days. Additionally, data collection was conducted during morning rush hours, noon, evening rush hours, and nighttime to capture traffic dynamics and environmental variations across different time periods. With over 120 man-hours of UAV operation, a total of 535 UAV videos were captured, comprising 338 RGB and 197 IR videos. The total video length amounts to 26 hours, with an average duration of approximately three minutes per video. This diverse and extensive dataset provides a robust foundation for subsequent research and practical applications.

7.2. Annotation Paradigm

Building a vehicle Re-ID dataset is a labor-intensive process, particularly for matching multi-modal images. Given the challenges posed by invisible license plates and significant visual discrepancies between RGB and IR images in UAV videos, expert annotators leveraged auxiliary information such as timestamps, locations, contextual details, and professional judgment to ensure precise vehicle matching and tagging. To streamline this process, we devised a systematic workflow, as shown in Fig. 8.

1. **Frame extraction**. RGB and IR videos were synchronized temporally and spatially, yielding 197 matched video pairs, each covering footage from at least two cameras. Frames were extracted at 3 FPS to create a raw dataset.
2. **Annatation**. Vehicle labels were applied using manual annotation tools or object detection algorithms for assistance.
3. **Batch slicing**. Annotated frames were organized into batches to generate a slice dataset for further processing.
4. **Data cleaning**. The slice dataset was manually reviewed and optimized by additional volunteers to eliminate noise and preserve high-value samples.
5. **Sample labeling**. Each reviewed sample was systematically labeled with identifiers such as ID number, camera number, image sequence number, color, type, and orientation, e.g., 0001_ir_0001_gray_car_7.jpg.

This meticulous labeling process yielded a deterministic dataset ready for diverse applications and analyses. This structured approach ensures that the dataset construction process, despite its complexity, is efficient and accurate, laying a solid foundation for future research and development.

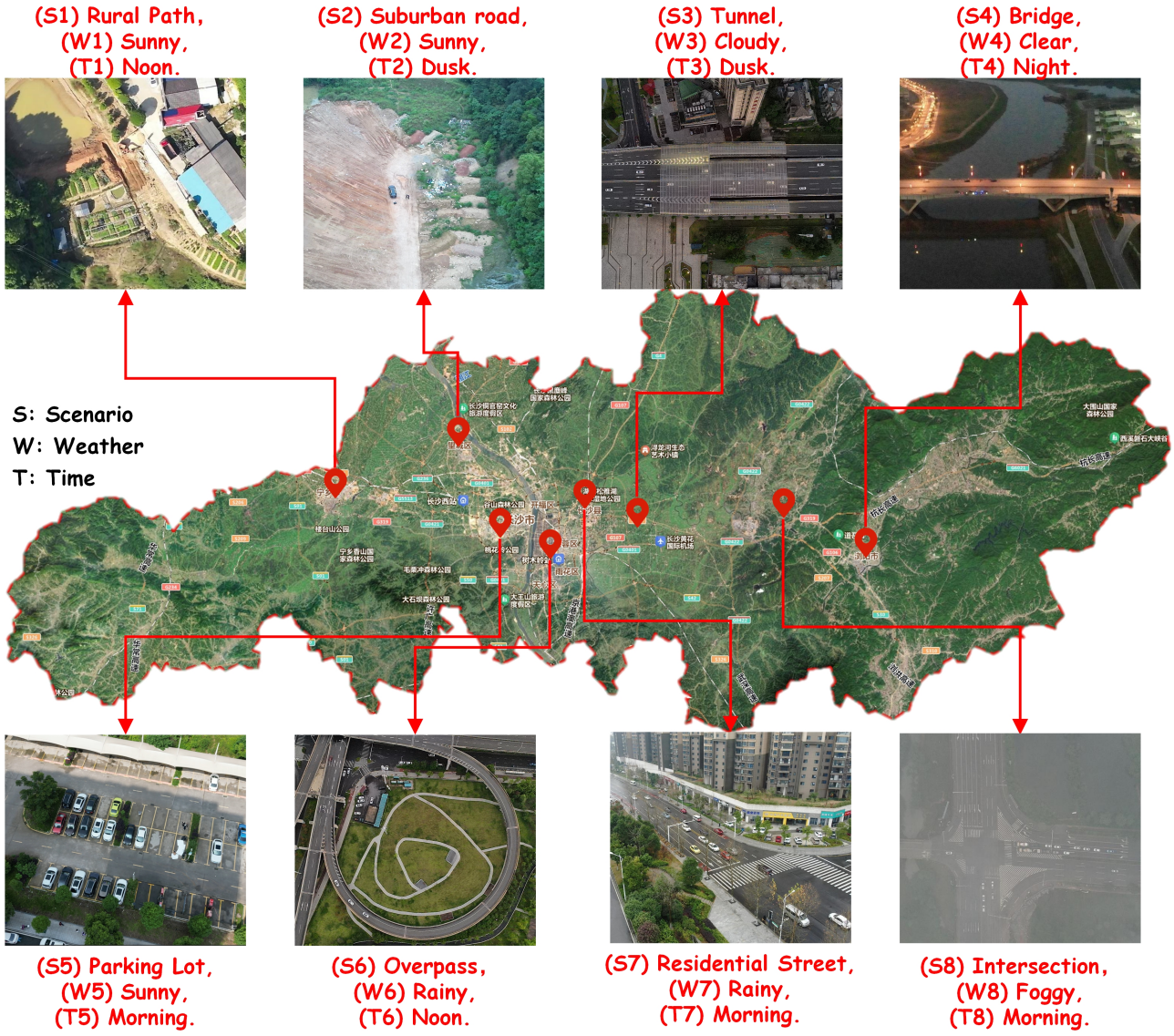


Figure 7. Illustration of data collection location, weather, and time period.

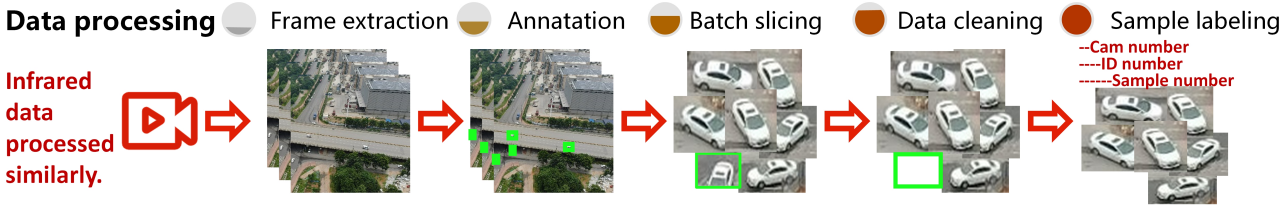


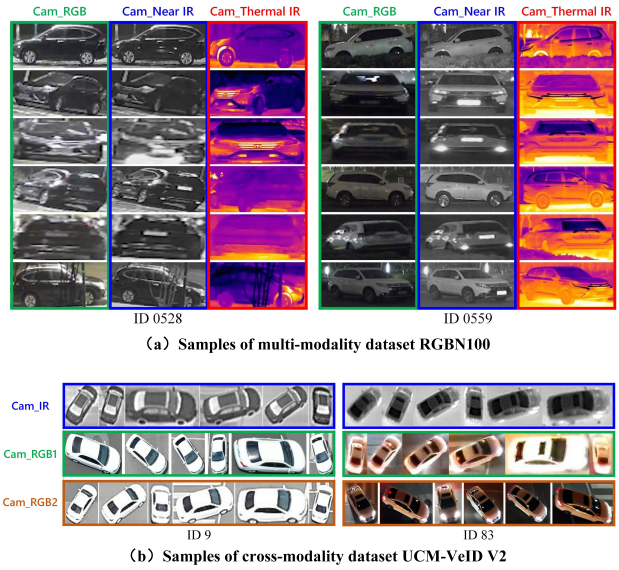
Figure 8. Data processing flow

8. Difference between UCM-VeID V2 with multi-spectral Re-ID dataset RGBN100

Some samples from both datasets are shown in Fig. 9. The main differences between RGBN100 and UCM-VeID V2 lie in the following aspects:

1. **Data Collection Perspective**. RGBN100 utilizes cameras with a human-eye-level perspective for data collection, whereas UCM-VeID V2 employs drones with a

CVPR
#84



（a）Samples of multi-modality dataset RGBN100

（b）Samples of cross-modality dataset UCM-VeID V2

Figure 9. Samples of different dataset

bird's-eye view to collect datasets.

2. **Image Registration**. RGBN100 requires strict alignment of images to ensure feature consistency across different modalities. In contrast, UCM-VeID V2 does not necessitate strict image registration.
3. **Methodology**. Multi-spectral Re-ID tasks aim to capture complementary features of different modality images using paired data, while cross-modality Re-ID tasks focus on learning the common features of images across different modalities.

From the above analysis, it can be concluded:

1. From the data collection perspective, RGBN100 is more demanding, whereas cross-modal vehicle Re-ID datasets are more aligned with real-world applications, offering higher practical value.
2. From the image registration perspective, RGBN100 requires significant time and human resources to align and annotate data, making the construction of large-scale datasets more challenging.
3. From a methodological perspective, cross-modality Re-ID tasks are more flexible, functioning effectively with a single modality, while multi-spectral Re-ID tasks rely on the support of multiple modalities.

According to the above analyses, the cross-modality Re-ID task seems to be more challenging and significant in real-world applications.

9. Limitations

There remain several limitations in this work that needs to be further investigated.

From the dataset aspect:

- **Fine-grained annotation limitations**: Due to the characteristics of UAV-captured data, such as small targets and multi-scale variations, it is challenging to annotate fine-grained features like interior decorations, front bumpers, and headlights.
- **Limited data collection platforms**: The dataset currently relies on a few specific platforms, lacking comprehensive data collection from diverse angles and devices, which may impact the dataset's diversity and representativeness.

From the method aspect:

- **Pre-training Data Dependency**: Current pre-trained weights still depend on large-scale datasets. Compared to pre-training on smaller datasets, the proposed method requires longer training time but deliver better performance. However, compared to other large-scale pre-trained weights, the proposed method achieve faster convergence at the cost of slightly lower performance.