
Supplementary Material: LVLM-Driven Attribute-Aware Modeling for Visible-Infrared Person Re-Identification

Zhiqi Pang¹ Lingling Zhao¹ Junjie Wang² Chunyu Wang^{1*}

¹ Harbin Institute of Technology, China

² Nanjing Medical University, China

zqpang98@gmail.com, zhaoll@hit.edu.cn, junjiehit@163.com, chunyu@hit.edu.cn

S.I Instruction Set for the LVLM

Each image is input into the LVLM, accompanied by the following instruction:

```
“Analyze the person in the image and answer strictly in this format:”
  “Gender: [male/female],”
  “Glasses: [wearing/without],”
  “Backpack: [carrying/without]”
  “Upper: [clothing type],”
  “Lower: [clothing type],”
  “Requirements:”
  “1. Exclude all color descriptions”
  “2. For glasses: consider all eyewear types”
  “3. For backpack: include shoulder bags and rucksacks”
“Examples:”
  “Good: Gender: female, Glasses: wearing, Backpack: without,
  Upper: blouse, Lower: skirt,”
  “Good: Gender: male, Glasses: without, Backpack: carrying,
  Upper: jacket, Lower: trousers”
```

S.II Details of Attribute-Aware Matching

We group the clusters based on the first three attribute values in the cluster-level attribute arrays. For example, a cluster with the first three attributes being “male,” “without,” and “carrying” is assigned to one group, while a cluster with “male,” “wearing,” and “carrying” is assigned to a different group. Based on the possible combinations of the first three attributes, we obtain a total of eight groups. Typically, each group contains both visible and infrared clusters.

Within each group, following the strategy of PGM [1], we first compute the assignment cost based on the similarity between cluster centers. For example, the similarity between the i -th cluster in the visible modality and the j -th cluster in the infrared modality within the same group is defined as:

$$Sim(c_i^v, c_j^r) = \frac{c_i^v \cdot c_j^{rT}}{\|c_i^v\| \times \|c_j^r\|}, \quad (S.1)$$

where c_i^v and c_j^r denote the centers of the i -th cluster in the visible modality and the j -th cluster in the infrared modality, respectively. Accordingly, the assignment cost between the i -th cluster in the

*Corresponding author: Chunyu Wang

visible modality and the j -th cluster in the infrared modality is defined as:

$$G_{ij} = \frac{1}{\exp(\text{Sim}(c_i^v, c_j^r))}, \quad (\text{S.2})$$

Based on all assignment costs, we construct a cost matrix G . We assume that in the k -th group, the number of visible clusters C_k^v is greater than the number of infrared clusters C_k^r . Subsequently, following the binary linear programming with linear constraints in PGM [1], we obtain a matching matrix H based on the cost matrix G . The matrix H contains C_k^r elements with a value of 1, corresponding to C_k^r positive matches between visible and infrared clusters. We then repeat the above matching process between the remaining $C_k^v - C_k^r$ visible clusters and the same C_k^r infrared clusters, until all visible clusters in the group are matched to infrared clusters. This iterative matching process follows the approach described in PGM [1]. Finally, the matching results from all groups are aggregated to obtain the final matching outcome.

S.III Algorithmic Procedure

Algorithm S.1 LVLM-AAM

Input: Unlabeled data $\{x_i^v\}_{i=1}^{N^v}$ and $\{x_i^r\}_{i=1}^{N^r}$, image encoder parameterized by θ_E , training *epochs* and *iters*.

- 1: Utilize the LVLM to extract an attribute array for each image in $\{x_i^v\}_{i=1}^{N^v}$ and $\{x_i^r\}_{i=1}^{N^r}$
- 2: **for** $i = 1$ to *epochs* **do**
- 3: Extract features of $\{x_i^v\}_{i=1}^{N^v}$ and $\{x_i^r\}_{i=1}^{N^r}$ with θ_E
- 4: Perform clustering within each modality
- 5: Perform attribute-aware refinement within each modality
- 6: Follow Eq.1 to calculate cluster centers
- 7: **for** $j = 1$ to *iters* **do**
- 8: Optimize θ_E to minimize the loss defined in Eq.3
- 9: **end for**
- 10: **end for**
- 11: Save the intra-modality pseudo-labels
- 12: Perform explicit-implicit attribute fusion
- 13: **for** $i = 1$ to *epochs* **do**
- 14: Follow Eq.1 to calculate cluster centers
- 15: Perform attribute-aware matching
- 16: Follow Eq.9 to calculate dynamic text features
- 17: **for** $j = 1$ to *iters* **do**
- 18: Optimize θ_E to minimize the loss defined in Eq.11
- 19: **end for**
- 20: **end for**
- 21: **return** Trained θ_E

S.IV Sensitivity Analysis

The parameter η represents the threshold for the difference between the attribute array of a single image and the cluster-level attribute array. Images exceeding this threshold are considered unreliable within the cluster and are therefore excluded. As shown in Figure S.1, setting $\eta = 2$ yields the best performance on both datasets. Setting η either too small or too large leads to performance degradation. Specifically, a too-small η imposes overly strict refinement conditions, resulting in the exclusion of many images and a subsequent reduction in training diversity. Conversely, a too-large η retains the majority of images, thereby diminishing the effectiveness of attribute-aware refinement and reducing the overall reliability of the pseudo-labels.

The parameter λ_{tsc} controls the weight of the text semantic contrastive loss. As shown in Figure S.2, the model achieves higher performance when $\lambda_{tsc} = 0.5$. In contrast, when $\lambda_{tsc} = 0.00$, the model performance significantly drops, demonstrating the effectiveness of the text semantic contrastive loss

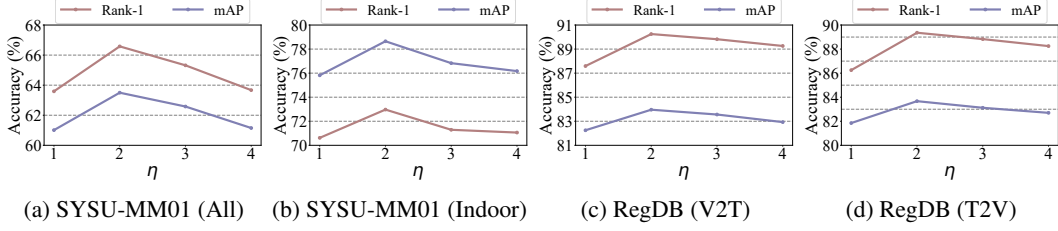


Figure S.1: Impact of hyperparameter η on performance. “All,” “Indoor,” “V2T,” and “T2V” denote SYSU-MM01 (All Search), SYSU-MM01 (Indoor Search), RegDB (Visible to Thermal), and RegDB (Thermal to Visible), respectively.

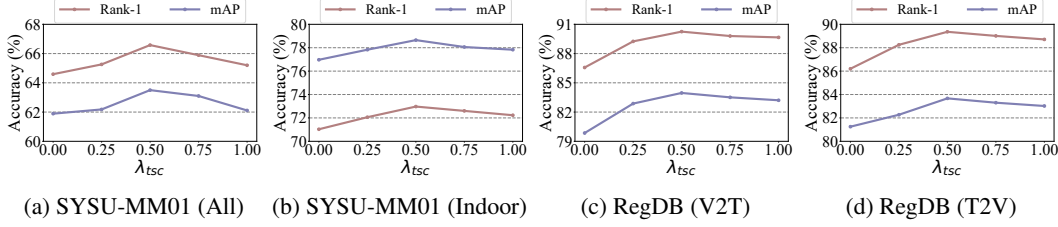


Figure S.2: Impact of hyperparameter λ_{tsc} on performance. “All,” “Indoor,” “V2T,” and “T2V” denote SYSU-MM01 (All Search), SYSU-MM01 (Indoor Search), RegDB (Visible to Thermal), and RegDB (Thermal to Visible), respectively.

in leveraging text semantics to enhance model performance. However, when λ_{tsc} exceeds 0.50, the model performance gradually declines. This may be because, although the text semantic contrastive loss contributes to performance improvement, the text descriptions only capture partial semantic information of the images. Therefore, the optimization process should still be primarily guided by the image-based pseudo-labels.

S.V Broader Impacts

The proposed LVLM-AAM enhances recognition performance by leveraging attribute arrays extracted by the LVLM, thereby facilitating the practical deployment of visible-infrared person re-identification and advancing applications in intelligent surveillance and public safety. Furthermore, the application scenarios of person re-identification highlight the need to strengthen access controls to safeguard public privacy. While the LVLM has the potential to serve as powerful tools driving technological progress, its use should be carefully regulated to prevent misuse.

References

- [1] Zesen Wu and Mang Ye. Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In *CVPR*, pages 9548–9558, 2023.