Visible-Infrared Person Re-Identification via Semantic Alignment and Affinity Inference (Supplementary Materials)

This supplementary material presents additional details and results not included in the main paper due to space limitations. We organize the contents as follows:

- Section A provides analysis of the hyper-parameters.
- Section B compares AIM with SIM.
- Section C shows details of the classification loss.
- Section D presents visualizations of retrieval results.

A. Analysis of the Hyper-parameters

Evaluation of parameters $\rho_1$ and $\rho_2$ in Eq. (5). We evaluate the effects of margin $\rho_1$ and margin $\rho_2$ in Eq. (5). We first set $\rho_2$ to 0.7 and experiment with different values for $\rho_1$. As shown in Figure 1, the performance improves with the increase of $\rho_1$ until it reaches 0.01. Increasing $\rho_1$ enhances the diversity of features, thereby reducing the risk of overfitting in the network. However, if $\rho_1$ is set too high, it may cause features of the same pedestrian to be dispersed, leading to reduced discriminability and making it difficult to identify pedestrians accurately. Hence, we set $\rho_1$ to 0.01 to balance diversity and discriminability.

Evaluation of parameters $k_1$ and $k_2$ in AIM. We set $\rho_1$ to 0.01, and vary $\rho_2$ from 0.1 to 0.9 with an interval of 0.2. As shown in Figure 2, we observe an initial improvement in performance with the increase of $\rho_2$, which reaches its maximum value at 0.7. Hence, we set $\rho_2$ to 0.7.

Evaluation of parameters $k_1$ and $k_2$ in AIM. We start from $k_1 = 50$ and $k_2 = 5$. As shown in Figure 3, the performance improves with the increase of $k_1$ until it reaches 20. Increasing $k_1$ allows the network to learn more comprehensive features, thereby improving the performance. However, if $k_1$ is set too high, it may cause the features to be too dispersed, leading to reduced discriminability and making it difficult to identify pedestrians accurately. Hence, we set $k_1$ to 20 to balance diversity and discriminability.

Evaluation of parameters $k_2$ in AIM. We set $k_1$ to 20, and vary $k_2$ from 2 to 10 with an interval of 2. As shown in Figure 4, the performance improves with the increase of $k_2$ until it reaches 6. Increasing $k_2$ allows the network to learn more comprehensive features, thereby improving the performance. However, if $k_2$ is set too high, it may cause the features to be too dispersed, leading to reduced discriminability and making it difficult to identify pedestrians accurately. Hence, we set $k_2$ to 6 to balance diversity and discriminability.
Figure 5. The visualization of retrieval results on SYSU-MM01 in all-search and multi-shot mode. We use different distance measurement methods (e.g., cosine distance, SIM [1], and AIM) to match pedestrian images. The green border represents the correct example, and the red border denotes the wrong example.
to 6, and change $k_1$ from 5 to 30 with an interval of 5. As shown in Figure 3, the performance is improved with the increase of $k_1$ at first, and achieve the best performance when $k_1 = 20$. Therefore, we set the value of $k_1$ to 20.

We set $k_1$ to 20, and change $k_2$ from 2 to 10 with the interval of 2. As shown in Figure 4, the performance is improved with the increase of $\rho_2$ at first, and achieve the best results when $k_2 = 6$. Hence, we set $k_2$ to 6.

B. Compare AIM with SIM

We add AIM and SIM [1] to several methods (e.g., AlignGAN [2], AGW [4], and MPANet [3]) for comparison. As shown in Table 1, AIM achieves better results than SIM. This result further proves the effectiveness of AIM.

Table 1. Evaluation of AIM and SIM on other methods on SYSU-MM01 under multi-shot setting. We retrain the models.

<table>
<thead>
<tr>
<th>Method</th>
<th>all-search</th>
<th>indoor-search</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rank-1 mAP</td>
<td>Rank-1 mAP</td>
</tr>
<tr>
<td>AlignGAN*+SIM</td>
<td>50.85</td>
<td>57.90</td>
</tr>
<tr>
<td>AlignGAN*+AIM</td>
<td>51.63</td>
<td>61.94</td>
</tr>
<tr>
<td>AGW*+SIM</td>
<td>54.17</td>
<td>62.95</td>
</tr>
<tr>
<td>AGW*+AIM</td>
<td>55.01</td>
<td>63.84</td>
</tr>
<tr>
<td>MPANet*+SIM</td>
<td>78.40</td>
<td>84.92</td>
</tr>
<tr>
<td>MPANet*+AIM</td>
<td>78.52</td>
<td>85.96</td>
</tr>
</tbody>
</table>

C. Details of the classification loss

For both baseline and our model, we utilize the classification loss $L_{id}$ to guide the model to focus on identical information. The classification loss can be defined as:

$$L_{id} = L_{sh, id} + L_{sp, id} + L_{cm, id}.$$  \hspace{1cm} (1)

$L_{sh, id}$ uses a shared classifier for both modalities as:

$$L_{sh, id} = -\log P(y_v|C(\tilde{F}_v)) - \log P(y_r|C(\tilde{F}_r)),$$  \hspace{1cm} (2)

where $P(\cdot)$ is the probability of correct prediction, $y_v$ and $y_r$ are labels, and $C(\cdot)$ is a shared classifier. $L_{sp, id}$ replace $C(\cdot)$ with modality-special classifiers $C_v(\cdot)$ and $C_r(\cdot)$. Following [3], we use mean classifiers $\overline{C}_v(\cdot)$ and $\overline{C}_r(\cdot)$ as:

$$L_{cm, id} = \overline{C}_v(\tilde{F}_v) \log \frac{\overline{C}_v(\tilde{F}_v)}{C_v(\tilde{F}_v)} + \overline{C}_r(\tilde{F}_r) \log \frac{\overline{C}_r(\tilde{F}_r)}{C_r(\tilde{F}_r)}.$$  \hspace{1cm} (3)

D. Visualization of retrieval Results

For each query image, we retrieve the top 10 gallery images with the highest similarity and rank them in descending order of similarity. As shown in Figure 5, AIM can achieve more stable matching results than other methods.

References