





Modality Synergy Complement Learning with Cascaded Aggregation for Visible-Infrared Person Re-Identification (Supplementary Materials)

Yiyuan Zhang¹, Sanyuan Zhao¹, Yuhao Kang¹, and Jianbing Shen^{1,2}

¹ School of Computer Science, Beijing Institute of Technology

² SKL-IOTSC, Department of Computer and Information Science,
University of Macau

{yiyuanzhang.ai, yuhaokangai, shenjianbingcg}@gmail.com
zhaosanyuan@bit.edu.cn

1 Theoretical Analysis

In this paper, we find that the visible image contains a high diversity of information, but it contains relatively more noise information, while the infrared image contains less information, but it contains relatively less noise information. Therefore, we propose MSCLNet to combine the advantages of the two types of images, so that we can extract information about identities with diversity and less noise. At the same time, we propose a new aggregation strategy, which can help us to reduce the intra-class discrepancy and enhance the inter-class discrimination in a fine grained manner.

1.1 Modality Synergy Module

Before two-stream features are fed into Modality Synergy Module, we utilize Instance Normalization to enhance instance discrimination across modalities. Visible and infrared features f_i^v, f_i^r are not in the same representation space. This needs us to construct a common feature space to project the multi-modal features. In particular, we take the Mogrifier LSTM [3] to deal with the two-stream context-irrelevant features. f_i^v, f_i^r are taken as sequential inputs in Mogrifier LSTM, where they share the same hidden state to encode the multi-modal features as shown in Fig. 1.

In conclusion, Modality Synergy Module fine-grained synergizes cross-modal features even though they are almost context-irrelevant. This makes the representation spaced really enhanced. And the detailed formulations of Modality

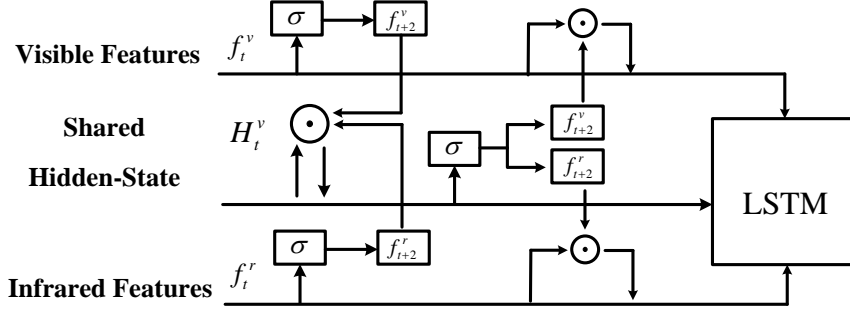


Fig. 1: **Detailed Illustration of Modality Synergy Module.** In specific, we share the Hidden state of Mogrifier LSTM to construct common representation space.

Synergy Module are as follows:

$$\begin{aligned}
f_i^s &\triangleq (\hat{f}_i^v, \hat{f}_i^r, y) = \mathcal{S}(\hat{f}_i^v, \hat{f}_i^r, y_i, \theta_s) \\
f_i^v &= 2\sigma(\mathbf{Q}_i h_{i-1}^{\text{prev}}) \odot f_{i-2}^v, \text{ for odd } i \in [1 \dots N_v] \\
f_i^r &= 2\sigma(\mathbf{Q}_i h_{i-1}^{\text{prev}}) \odot f_{i-2}^r, \text{ for odd } i \in [1 \dots N_r] \\
h_i^{\text{prev}} &= 2\sigma(\mathbf{R}_i f_{i-1}) \odot h_{i-2}^{\text{prev}}, \text{ for even } i \in [1 \dots r] \\
I_t &= \sigma(w_{fI} f_{i,t} + w_{hI} h_{t-1} + b_I) \\
F_t &= \sigma(w_{fF} f_{i,t} + w_{hF} h_{t-1} + b_F) \\
o_t &= \sigma(w_{fo} f_{i,t} + w_{ho} h_{t-1} + b_o) \\
c_t &= F_t \odot c_{t-1} + I_t \odot \tanh(w_{fg} f_{i,t} + w_{hg} h_{t-1} + b_g) \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}$$

1.2 Algorithm

We summarize the Algorithm of our proposed MSCLNet in Algorithm. 1

2 Experiment Details

The baseline of MSCLNet is AGW*, which means AGW [4] with Random Erasing [5]. And we adopted pre-trained ResNet-50 [1] on ImageNet [2] as the backbone network. What's more, before feeding the images into the model, we pre-processed these images. We changed the size of each image to 288×144 and augmented images through random cropping with zero-padding, random horizontal flipping and random erasing (80% probability, 80% max-area, 20% min-area). During the training stage, we used the SGD optimizer, and the learning rate decayed over each epoch. During the testing stage, we used the model to extract

Algorithm 1 Modality Synergy Complement Learning with Cascaded Aggregation

Input: visible images \mathcal{V} and infrared images \mathcal{R}
Target : Query $x_v^Q \in \mathcal{V}$ get $x_r^G \in \mathcal{R}$, or Query $x_r^Q \in \mathcal{R}$ get $x_v^G \in \mathcal{V}$
 TRAIN PROGRESS($\mathcal{V} \iff \mathcal{R}$)
 select randomly $x_v \in \mathcal{V}, x_r \in \mathcal{R}$
 for $i = 1, 2, \dots, N$
 (1) Get visible, infrared features f_i^v, f_i^r via extractors.
 $\mathcal{L}_{synergy} \leftarrow \mathcal{L}_{div}(\theta_v, \theta_r), \mathcal{L}_t(\theta_v, \theta_r)$.
 (2) Synergize f_i^v, f_i^r to get f_i^s then complement f_i^s with f_i^v, f_i^r .
 $\mathcal{L}_{com} \leftarrow \mathcal{L}_{local}(\theta_s), \mathcal{L}_{global}(\theta_s)$.
 (3) Optimize distribution of feature embeddings \hat{f}_s .
 $\mathcal{L}_{cascade} \leftarrow \mathcal{L}_{sub}(\theta_s, \theta_v, \theta_r), \mathcal{L}_{intra}(\theta_s, \theta_v, \theta_r)$ and
 $\mathcal{L}_{inter}(\theta_s, \theta_v, \theta_r)$
 (4) Backward \mathcal{L}_{total} and update $\theta_v, \theta_r, \theta_s$.
 end for
Output: Optimized model of the proposed method.

the query and the gallery features from single modality by the feature extractors .

3 More Visualization Results

In this section, we add more visualization results. In Fig. 2, it is obvious that features extracted by the model can be discriminative about the pedestrians. The learned features are mostly related to the body and feet of the pedestrians across modalities. Hence, we are convinced that our model successfully extracts the discriminative information about different identities, and the ability to infer retrieval results is enhanced. Meanwhile, as shown in Fig. 3, we calculate the distance matrix according to the features extracted from the model. For each Query image, we select the top 10 images with the highest similarity among gallery images and rank them in the order of similarity. The final results show that the gallery images with the highest similarities are all of the same identities. Retrieval results illustrate the precision of our model.

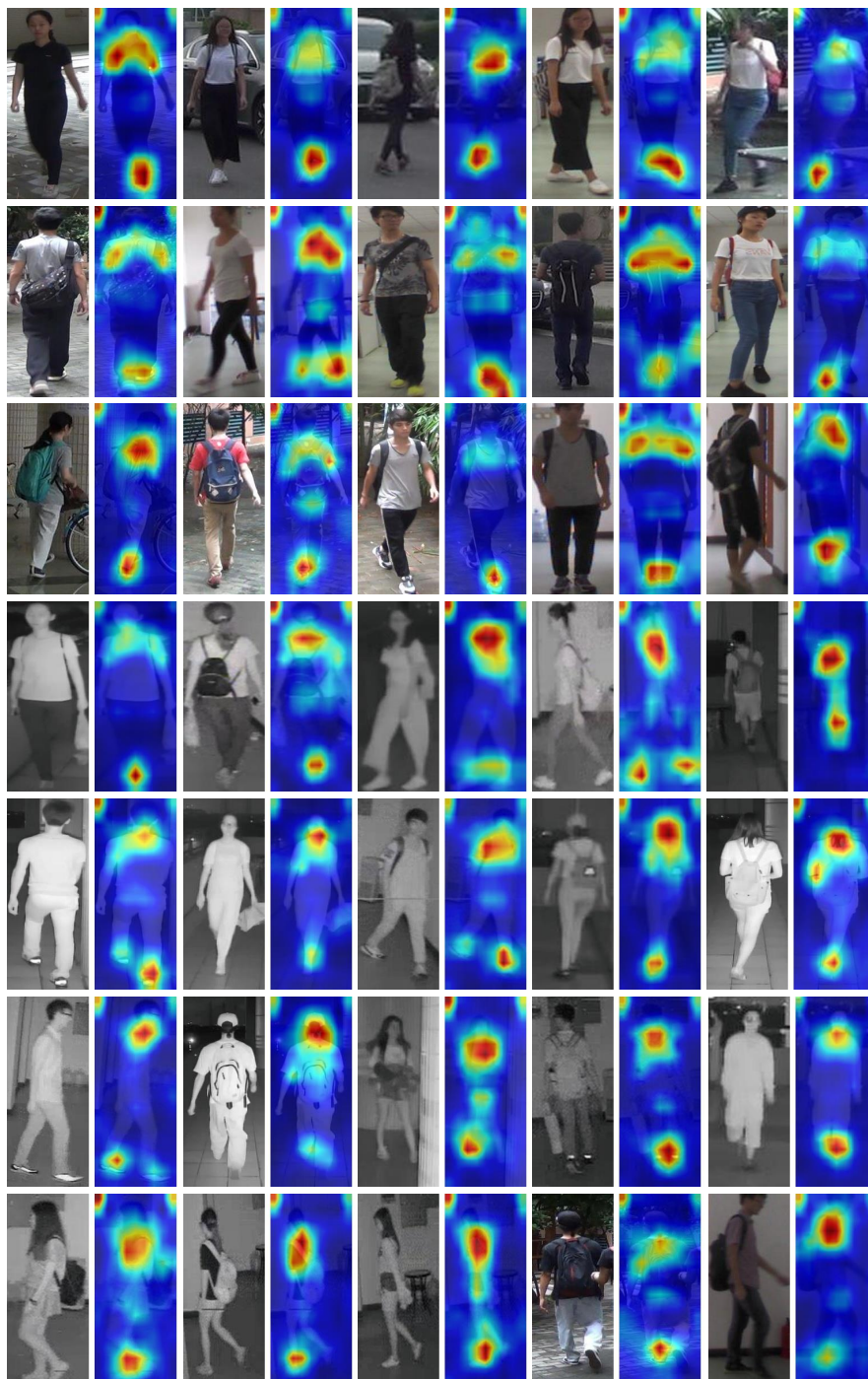


Fig. 2: Visualization: Heatmap.

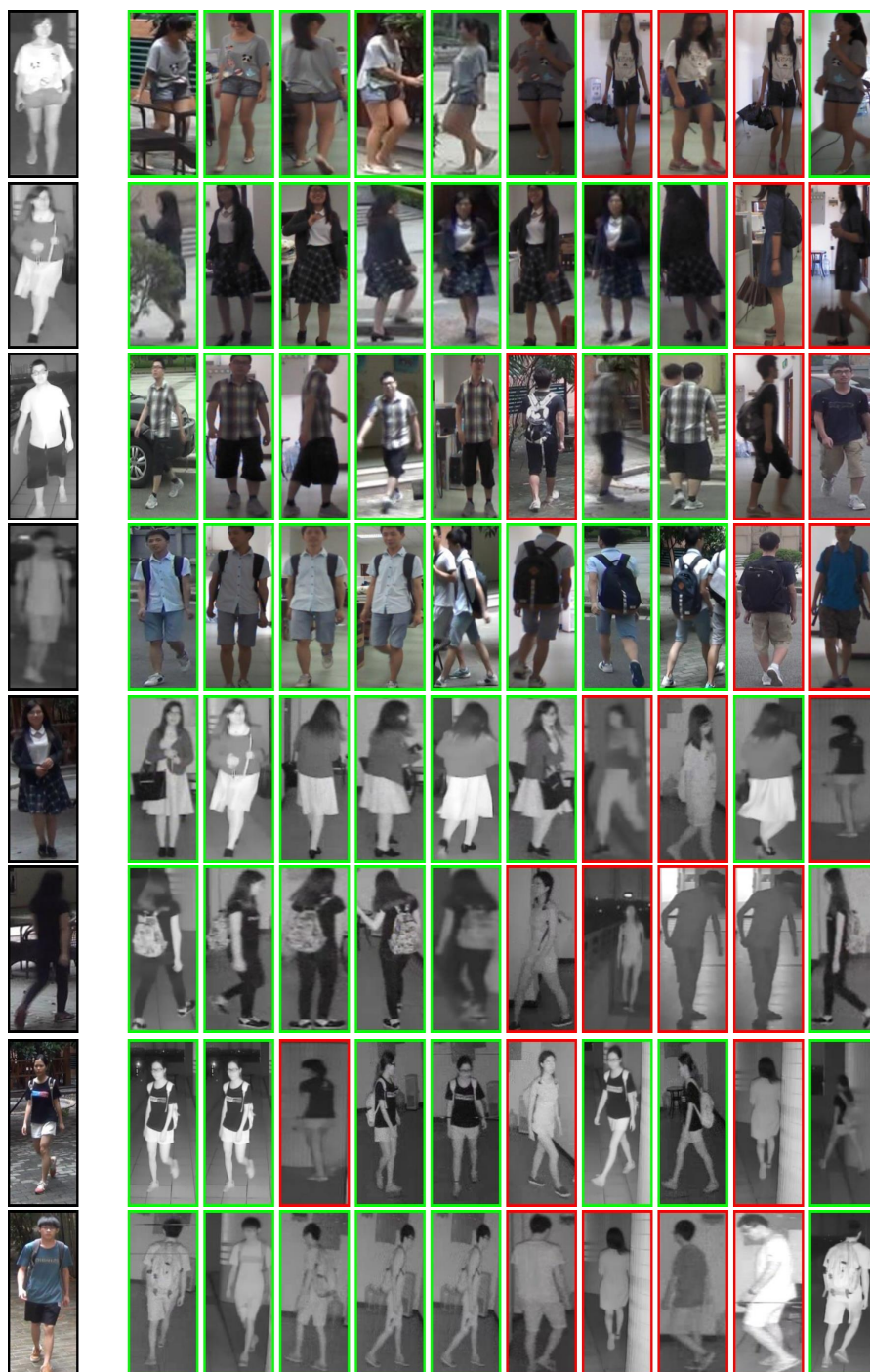


Fig. 3: Visualization: Retrieval Results.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
2. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25** (2012)
3. Melis, G., Kočiskỳ, T., Blunsom, P.: Mogrifier lstm. arXiv preprint arXiv:1909.01792 (2019)
4. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H.: Deep learning for person re-identification: A survey and outlook. arXiv preprint arXiv:2001.04193 (2020)
5. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI. pp. 13001–13008 (2020)