Supplementary Material for Paper "Neural Feature Search for RGB-Infrared Person Re-Identification"

Yehansen Chen[1]*, Lin Wan[1]*, Zhihang Li[2]†, Qianyan Jing[1], Zongyuan Sun[1]
[1]School of Geography and Information Engineering, China University of Geosciences, Wuhan, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{chenyehansen, wanlin, jingqianyan, sunzongyuan}@cug.edu.cn; zhihang.li@nlpr.ia.ac.cn

1. The Architecture of Our Baseline Network

In our implementation, the baseline network adopts a commonly used two-stream structure (Fig. 1) and takes ResNet-50 as the backbone.

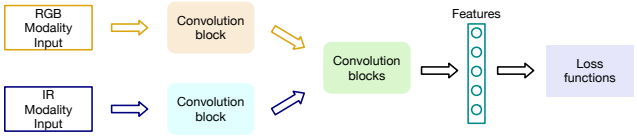

Figure 1. Two-stream structure of the baseline network.

Architecture details of the baseline network are shown in Table 1. All images are resized to 288 × 144 as the network inputs. The stride of the last convolutional block is set to 1 so as to obtain fine-grained feature maps. The other hyper-parameters are following [1] without tuning.

Table 1. Architecture details of our two-stream baseline network.

Layer name	Output size	50-layer	Type
conv1	144×72	7×7, 64, stride 2	modality-specific
conv2	72×36	3×3 max pool, stride 2	shared
		$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	
conv3	36×18	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	shared
conv4	18×9	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	shared
conv5	18×9	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	shared

2. Visualization of Retrieved Examples

As shown in Fig. 2, the top-5 NFS retrieval results of 16 randomly selected query examples on the SYSU-MM01 dataset are plotted. We not only follow the original evaluation protocol, but also evaluate the *Visible-Infrared* setting.

In detail, the first column includes randomly selected samples from the query set, and retrieval results are sorted from left to right in descending order of cosine similarity scores. Due to lack of color information in IR images, some cases are even difficult for human (e.g., query D and K). But the proposed method can retrieve correct results, which demonstrates the effectiveness of NFS in narrowing the large modality gap. According to the retrieval results for query B, C, D, and I, we also observe that our method exhibits certain robustness for high sample noises such as background clutter and partial occlusions. Interestingly, we discover that even if some persons change their clothes (e.g., query E), NFS can still return accurate retrieval results by mining other discriminative cues, perhaps the face or shoes. Another interesting phenomenon is that performance of *Visible-Infrared* setting is usually better than that of *Infrared-Visible* one. The main reason is that visible images often provide richer appearance information than their IR counterparts. Although there are still a few failure cases, most of these images (such as query F, O, and P) only present back views of the person-of-interest with limited identity-related information (e.g. face, texture, or logo of clothes). In conclusion, NFS exhibits promising performance in either query setting.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

*Equally-contributed first authors
†Corresponding author

Figure 2. The top-5 retrieval results for 16 randomly selected query samples (8 samples per query setting) on the SYSU-MM01 dataset with our neural feature search method. Correct retrieved samples are in green boxes and wrong matchings are in red boxes (best viewed in color). Numerical values report cosine similarity scores of image pairs.