MFEN: Multi-Frequency Expert Network for Visible-Infrared Person Re-ID

Supplementary Material

Overview. This supplementary material accompanies our main paper entitled "MFEN: Multi-Frequency Expert Network for Visible-Infrared Person Re-ID." It provides additional evidence supporting the main claims. Specifically, it includes the following:

- Complexity analysis, showing that the gain of MFEN does not come with heavy overhead.
- Hyperparameter analysis, showing that the proposed multi-expert design is not overly sensitive to parameter choices.
- Modality-wise frequency statistics, showing that RGB and IR images exhibit distinct band-wise energy distributions.
- Visualization of Random Frequency Augmentation.
- Visualization of retrieval results.
- Visualization of the Multi-Frequency Expert Network, showing that different experts focus on different granularities of identity cues.
- Visualization of Frequency Auxiliary Optimization, showing that frequency-aware supervision improves the feature geometry.

1. Modality-wise Frequency Statistics

To better illustrate our motivation, we quantify the frequency energy distribution of RGB and IR images on SYSU-MM01 under the same four-band partition used by MFEN. As shown in Tab. 1, the two modalities exhibit clearly different band-wise energy distributions. In particular, IR images concentrate much more energy in the lowest-frequency band, while RGB images retain relatively richer energy in the remaining bands. This observation is consistent with our motivation that the modality gap is closely related to illumination-induced frequency differences, and further justifies sample-wise multi-band modeling over uniform whole-spectrum processing.

Table 1. Band-wise frequency energy distribution (%) of RGB and IR images on SYSU-MM01.

Band	RGB	IR	Band	RGB	IR
1 (lowest)	85.4	96.2	3 (mid-high)	4.2	0.8
2 (mid-low)	9.0	2.9	4 (highest)	1.0	0.1

2. Complexity Analysis

As shown in Tab. 2, we analyzed the time overhead and parameter count introduced by each module of our method. The proposed method increases the number of parameters by only 4% over the baseline, with a 6% increase in training time and a 1% increase in inference time. Notably, the RFA and FAO modules incur almost no additional overhead. These results support the claim in the main paper that the gain of MFEN is not obtained by simply increasing model size or introducing expensive test-time computation.

Most of the overhead arises from the MFEN module, which is lightweight and efficient. This is because, for a feature map with an input size of $H \times W \times C$, the time complexities of the Fast Fourier Transform (FFT) and its inverse operator (IFFT) are both $O(CHWlog(HW))$. In contrast, the time complexity of a 1×1 convolution is $O(C^2HW)$. In practice, $log(HW)$ is less than C, which implies that the complexity of the FFT and IFFT is lower than that of a 1×1 convolution. Therefore, the supplementary evidence confirms that MFEN is not only effective, but also practical for VI-ReID systems.

Table 2. Complexity of our method on the SYSU-MM01 dataset.

Method	Params	Train Time	Test Time
baseline	×1.0	×1.0	×1.0
Our method	×1.04	×1.06	×1.01
– RFA	×0.0	×0.0001	×0.0
– MFEN	×0.04	×0.05	×0.01
– FAO	×0.0004	×0.01	×0.0

3. Hyperparameter Analysis

We analyze the margin parameter ρ in the Frequency Euclidean Distance Loss (L_{feu}), as well as the number of experts n.

As shown in Fig. 1 (a), the margin parameter ρ is designed to increase the distance between positive and negative samples in the feature space. Both the model without FAO and our approach (with FAO) achieve the best performance when $\rho = 0.6$. Furthermore, it can be observed that our method, which incorporates a mirrored frequency-domain constraint, significantly enhances performance and demonstrates reduced sensitivity to ρ. This indicates that FAO provides a stable auxiliary signal rather than introducing a fragile extra objective.

As shown in Fig. 1 (b), our method achieves the best performance when the number of experts is $n = 4$. When $n < 4$, the model's performance improves as the number of experts increases because the model can effectively han-

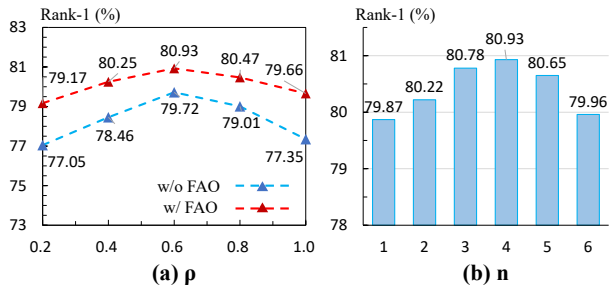

Figure 1. Effect of the hyperparameter ρ and n on SYSU-MM01 [2] dataset.

Figure 2. Visualization of Random Frequency Augmentation. The enhanced image (Aug) is obtained by exchanging the low-frequency amplitudes of the source image with the target image.

dle different frequency bands. However, as n continues to increase, overly granular frequency bands hinder the learning of each expert. This trend is consistent with our main-paper discussion: multiple experts are necessary because different bands capture complementary cues, while an excessively fine partition unnecessarily weakens each expert. Therefore, the multi-expert design is meaningful rather than being a brittle empirical setting.

4. Visualization of Retrieval Results

Fig. 3 visualizes the retrieval ranking lists on RegDB and SYSU-MM01. The green boxes denote correct retrieval results and the red boxes denote incorrect ones. Compared with the baseline, MFEN learns more useful frequency-domain cues and demonstrates better retrieval capability.

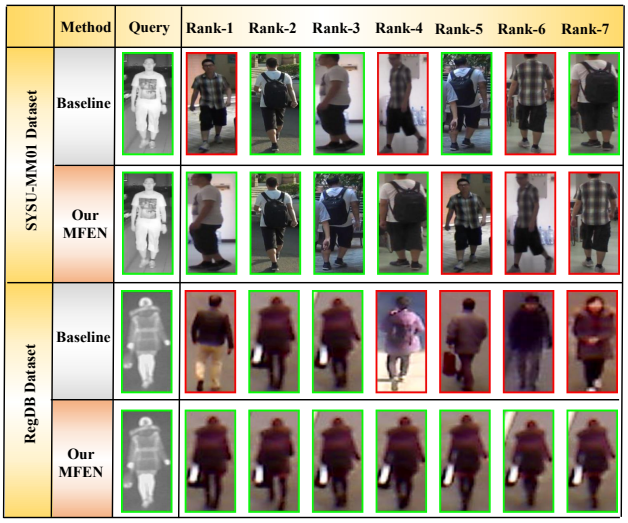

Figure 3. Visualization of retrieval ranking lists on RegDB and SYSU-MM01. Green and red boxes denote correct and incorrect retrieval results.

5. Visualization of Random Frequency Augmentation (RFA)

Fig. 2 presents the visualization results of the proposed RFA. The enhanced image (Aug) is obtained by exchanging the low-frequency amplitudes of the source image with the target image. It is evident that, following the RFA, the illumination conditions of the images become more diverse, effectively narrowing the gap between the two modalities. At the same time, the main body structures of pedestrians remain recognizable, consistent with our design of exchanging only low-frequency amplitude while preserving structural details.

6. Visualization of Multi-Frequency Expert Network (MFEN)

Fig. 4 presents the activation feature maps of the proposed MFEN. In each row, the first column shows the original image and the second column shows the result obtained by combining all experts, while the remaining four columns correspond to the outputs of the four experts, arranged from low to high frequencies. It is evident that different experts, focusing on distinct frequency bands, capture identity cues of varying granularity. Low-frequency experts tend to emphasize overall characteristics, whereas high-frequency experts focus on subtle variations. Finally, when we dynamically combine all experts using a gating mechanism, the model focuses on the most discriminative regions. This visualization directly supports the main-paper claim that the experts are not redundant: they learn complementary cues at different frequency bands, and the fusion module selects

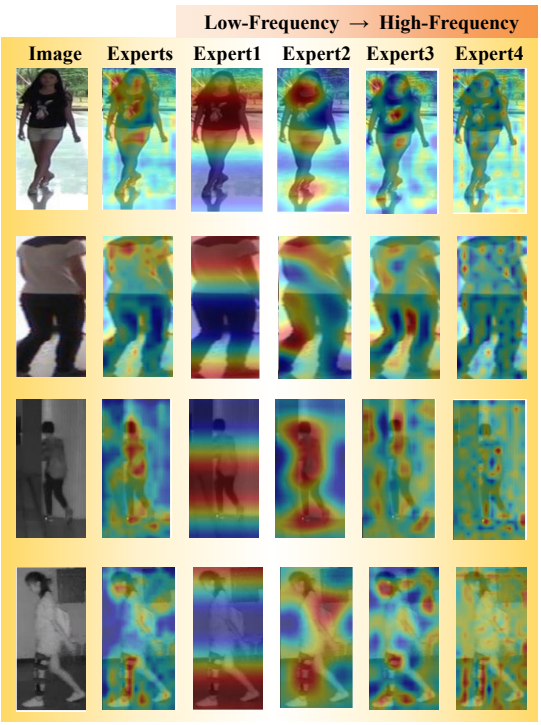

Figure 4. Visualization of the activation feature maps on the SYSU-MM01 dataset. In each row, the first column shows the original image, the second column shows the result obtained by combining all experts, while the remaining four columns correspond to the outputs of the four experts, arranged from low to high frequencies.

a more informative combination for each sample.

7. Visualization of Frequency Auxiliary Optimization (FAO)

Fig. 5 presents the t-SNE [1] visualization of the distributions of image features from the spatial domain and the frequency domain. In the absence of the proposed FAO, the spatial-domain features exhibit a certain degree of misclustering, while the frequency-domain features remain entirely indistinguishable. However, when FAO is introduced, the additional frequency-domain constraints not only render the frequency distribution distinguishable but also collaboratively optimize the spatial feature distribution, resulting in more robust overall representations. This observation is consistent with our formulation of FAO in the main paper: frequency-domain statistics act as an auxiliary constraint that improves the geometry of the final embedding, rather than serving as an isolated branch.

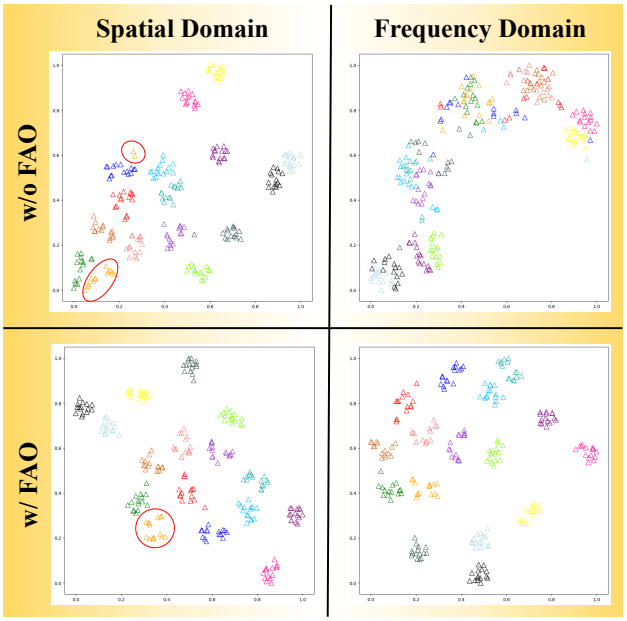

Figure 5. t-SNE [1] visualization of the distributions of image features from the spatial domain and the frequency domain. Different colors represent different identities.

References

[1] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 3

[2] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017. 2