FSLoRA: Harmonizing Detection and Re-Identification via Freq-Spatial Low-Rank Adapter for One-Stage Person Search

Supplementary Material

In this document, we provide more details of our FSLoRA, detailed experiment setup (A) and more experimental results, including quantitative (B) and qualitative results (C).

A. Experiment Setup

In this section, we introduce the datasets, evaluation metrics and implementation details.

A.1. Datasets

Our experiments are implemented on two widely used datasets for person search: CUHK-SYSU [12] and PRW [16]. **(1) CUHK-SYSU** is a large-scale person search dataset containing 18,184 images annotated with 96,143 pedestrian bounding boxes across 8,432 unique identities. The dataset includes two types of images: video frames from movies and street scenes captured by a moving camera, which introduce significant variations in viewpoint, lighting, and occlusion. The dataset is split into a standard training set, consisting of 11,206 images with 5,532 identities, and a test set containing 6,978 gallery images and 2,900 query individuals. Unlike PRW, CUHK-SYSU employs a set of protocols with gallery sizes ranging from 50 to 4,000. In our experiments, we follow the default protocol using a gallery size of 100 unless otherwise specified. **(2) PRW**, captured using six static cameras across a university campus, comprises 11,816 frames derived from surveillance video footage. The dataset includes 43,110 bounding boxes, with 34,304 of these annotated with 932 identities, and the remainder labeled as unknown. The dataset is split into a training set of 5,704 images and 482 identities, and a test set of 2,057 query persons and 6,112 gallery images. In contrast to CUHK-SYSU, the entire gallery set in PRW is used as the search space for each query individual.

In addition to the two standard datasets, a newly introduced dataset, **(3) PoseTrack21** [5], is also applicable to person search. It comprises 42,861 training images featuring 5,474 unique individuals, along with 19,935 gallery images containing 1,313 query identities. Unlike the previously mentioned datasets, PoseTrack21 presents a unique challenge where query images may include multiple individuals due to occlusion.

A.2. Evaluation Metrics

For re-ID subtask, performance is assessed using mean Average Precision (mAP) and Top-k Cumulative Matching Characteristics (CMC). mAP quantifies the accuracy of matching a query to gallery images, averaged over all queries, while Top-k measures the percentage of queries where at least one of the top-k results corresponds to the correct identity. In the detection subtask, we utilize Average Precision (AP) and Recall: AP evaluates the accuracy of bounding box predictions, while Recall indicates the proportion of relevant detections retrieved. In this paper, we adopt the standard evaluation protocols for person search, using the same evaluation metrics as re-ID, namely mAP and Top-k. Top-1 score is utilized unless otherwise specified.

A.3. Implementation Details

All models in our experiments utilize VMamba [10], which is pretrained on ImageNet [4], as the backbone. In particular, module with weights W_0 in Figure 2 corresponds to the linear layers in the feedforward network (FFN) module of VMamba, along with the two linear layers present in the "SS2D" block. For the LoRA configuration, we set r=32 and use 2 expert models, *i.e.*, n=2. Regarding the low-pass and high-pass filters in Eq. 5, $\mathcal{F}_{low}$ represents an ideal low-pass filter with a cutoff frequency of 30, meaning that frequency components below this threshold are preserved while higher frequencies are attenuated. Similarly, $\mathcal{F}_{high}$ denotes an ideal high-pass filter with a cutoff frequency of 40, which removes lower-frequency components while retaining higher-frequency details.

During training, all models are trained on a single A800 NVIDIA GPU for 23 epochs. The initial learning rate is set to 0.0006, which is reduced by a factor of 10 at the 16th and 22nd epoch. The batch size is set to 2. We use the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 0.0005. A warm-up strategy is applied for the first epoch. Consistent with [13, 14], we employ a multi-scale training strategy, where the long sides of the images are randomly resized between 667 and 2000 pixels. For testing, the image resolution is fixed at 1500 $\times$ 900. Our method is implemented using the MMDetection open-source library [3].

B. Quantitative Results

B.1. Results of Multi-View Gallery Setting on PRW Dataset

Since the PRW dataset is collected using six static cameras on a university campus, it provides an ideal benchmark for evaluating multi-view gallery performance, where matches

must be identified across different camera viewpoints. As shown in Table 7, we extend our comparison to additional methods, including HOIM [1], APNet [17], NAE+ [2], PGA [7], SeqNet [9], AGWF [6], COAT [15], and PAD [8]. The results reveal a general decline in performance under the multi-view gallery setting, likely due to challenges such as viewpoint variations, illumination changes, and occlusions across cameras [2, 9]. Despite these difficulties, FSLoRA consistently achieves the highest performance, with 58.6pp and 81.3pp w.r.t. mAP and Top-1 respectively, significantly surpassing existing methods. These findings further validate FSLoRA's effectiveness and robustness in handling cross-view person search in real-world scenarios.

Table 7. Performance comparison on the multi-view gallery of PRW dataset. The best results are bold.

Method	mAP	Top-1
HOIM [1]	36.5	65.0
APNet [17]	38.7	66.7
NAE+ [2]	40.0	67.5
PGA [7]	41.9	68.1
SeqNet [9]	43.6	68.5
AGWF [6]	48.0	73.2
COAT [15]	50.9	75.1
PAD [8]	52.1	77.3
FSLoRA	**58.6**	**81.3**

B.2. Effect of Cutoff Frequency Configuration in FLM

We further investigate how the cutoff frequencies of the low-pass and high-pass filters, denoted as f_{low}^{co} and f_{high}^{co} respectively, influence the performance of FLM on the PRW dataset. As described in Sec. A.3, $\mathcal{F}_{low}$ preserves frequency components below f_{low}^{co}, while $\mathcal{F}_{high}$ retains those above f_{high}^{co}. To analyze their effect, we design two configurations, illustrated as **A** and **B** in Figure 9.

(A) The spectrum is divided into three non-overlapping bands: $f \leq f_{low}^{co}$, $f_{low}^{co} < f < f_{high}^{co}$, and $f \geq f_{high}^{co}$. The mid-band region ($f_{low}^{co} < f < f_{high}^{co}$, *e.g.*, 30–40) is completely filtered out, effectively removing it from the feature spectrum.

(B) Conversely, this configuration defines two overlapping bands: $f \leq f_{low}^{co}$ and $f \geq f_{high}^{co}$, where $f_{high}^{co} < f_{low}^{co}$, leading to an overlap between the two frequency ranges.

As shown in Figure 9, the best performance is achieved when $f_{low}^{co} = 30$ and $f_{high}^{co} = 40$, with configuration **A** consistently outperforming **B**. This suggests that explicitly removing certain mid-frequency components benefits person search, likely because these regions contain task-irrelevant or noisy information. These findings offer valuable insights for future adaptive frequency-band selection in FLM design.

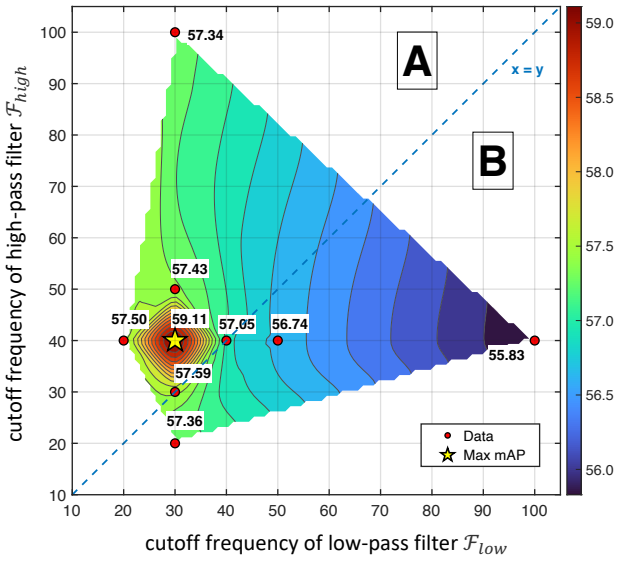


Figure 9. Impact of cutoff frequency configuration in FLM within NAE framework [2]. Region **A** divides the spectrum into three non-overlapping bands, while **B** allows overlapping ranges. Configuration **A** yields higher mAP, suggesting that removing mid-frequency components benefits person search.

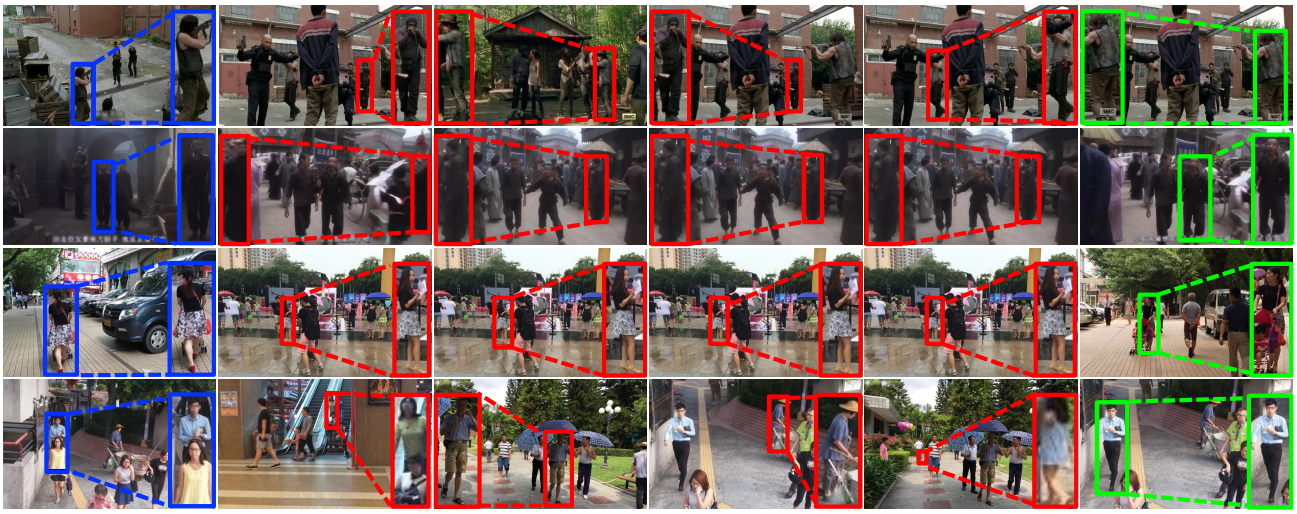
B.3. Effectiveness of SLM and FLM with Equal Parameter Budgets

This experiment aims to verify that the performance improvement of our model stems from the proposed dual-domain design rather than from an increase in parameter count. To this end, we carefully adjust the hyperparameters to ensure that the SLM, FLM, and their combination (SLM+FLM) have approximately equal numbers of parameters when integrated into the NAE [2] framework. The results on the PRW dataset are reported in Table 8.

Under these controlled settings, SLM and FLM individually achieve 58.59pp and 58.40pp w.r.t. mAP, respectively. When both modules are combined, the performance further improves to 59.11pp. These consistent gains under fixed parameter budgets demonstrate that the observed improvements are attributed to the complementary effects

Table 8. Comparison of SLM, FLM, and their combination (SLM+FLM) under equal parameter budgets within the NAE framework [2] on the PRW dataset.

	SLM	FLM	SLM+FLM
mAP	58.59	58.40	59.11
Top-1	88.30	87.95	88.99



(a) Query (b) NAE (c) AlignPS (d) ROI-AlignPS (e) GALW (f) FSLoRA

Figure 10. Visualization of Top-1 retrieval results for several samples. The query is highlighted in blue, with correct matches in green and incorrect matches in red. The detected objects are zoomed in for better view.

of spatial and frequency modulation, rather than increased model capacity. This validates the effectiveness of our dual-domain feature modulation in enhancing person search performance.

B.4. Number of Extra Parameters

FSLoRA is designed as a lightweight feature decoupling module, ensuring minimal computational and parameter overhead. For instance, when integrated into the NAE [2] framework with a VMamba backbone, it introduces approximately 1.28M additional parameters, which accounts for less than 2% of the total parameters in the baseline model. This demonstrates that FSLoRA facilitates effective task-aware feature decoupling and extraction, while keeping the model complexity almost unchanged.

C. Qualitative Results

We present qualitative comparisons between FSLoRA and some one-stage methods, including NAE [2], AlignPS [13], ROI-AlignPS [14] and GALW [11]. As shown in Figure 10, FSLoRA demonstrates superior localization and identification capabilities, accurately retrieving the target person even in challenging environments with occlusions (rows 3 and 4), low-light conditions (rows 1 and 2) and similar clothes (rows 1, 2 and 3).

In addition, we also provide several failure cases of our method as shown in Figure 11. Our methods occasionally yield incorrect results in situations where the appearances are extremely similar. These qualitative comparisons highlight the robustness of our proposed methods while underscoring the remaining challenges in person search tasks.

Figure 11. Visualization of failure cases. The query and our incorrect Top-1 result are highlighted in blue and red respectively. The detected objects are zoomed in for better view.

References

[1] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Bernt Schiele. Hierarchical online instance matching for per-

son search. *AAAI*, 34(7):10518–10525, 2020. 2

[2] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. In *CVPR*, pages 12615–12624, 2020. 2, 3

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[5] Andreas Doering, Di Chen, Shanshan Zhang, Bernt Schiele, and Juergen Gall. PoseTrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *CVPR*, pages 20963–20972, 2022. 1

[6] Byeong-Ju Han, Kuhyeun Ko, and Jae-Young Sim. End-to-end trainable trident person search network using adaptive gradient propagation. In *ICCV*, pages 925–933, 2021. 2

[7] Hanjae Kim, Sunghun Joung, Ig-Jae Kim, and Kwanghoon Sohn. Prototype-guided saliency feature learning for person search. In *CVPR*, pages 4865–4874, 2021. 2

[8] Hanjae Kim, Jiyoung Lee, and Kwanghoon Sohn. Prototype-guided attention distillation for discriminative person search. *IEEE TPAMI*, 47(1):99–115, 2024. 2

[9] Zhengjia Li and Duoqian Miao. Sequential end-to-end network for efficient person search. *AAAI*, 35(3):2011–2019, 2021. 2

[10] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *NeurIPS*, 37: 103031–103063, 2024. 1

[11] Yanling Tian, Di Chen, Yunan Liu, Shanshan Zhang, and Jian Yang. Grouped adaptive loss weighting for person search. In *ACM MM*, pages 6774–6782, 2022. 3

[12] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, pages 3415–3424, 2017. 1

[13] Yichao Yan, Jinpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, and Ling Shao. Anchor-free person search. In *CVPR*, pages 7690–7699, 2021. 1, 3

[14] Yichao Yan, Jinpeng Li, Jie Qin, Shengcai Liao, and Xiaokang Yang. Efficient person search: An anchor-free approach. *IJCV*, 131(7):1642–1661, 2023. 1, 3

[15] Rui Yu, Dawei Du, Rodney LaLonde, Daniel Davila, Christopher Funk, Anthony Hoogs, and Brian Clipp. Cascade transformers for end-to-end person search. In *CVPR*, pages 7267–7276, 2022. 2

[16] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *CVPR*, pages 1367–1376, 2017. 1

[17] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. Robust partial matching for person search in the wild. In *CVPR*, pages 6827–6835, 2020. 2