Mixture of Submodules for Domain Adaptive Person Search

Supplementary Material

In this supplementary material, we provide additional experimental results, loss functions for baseline, and qualitative results to complement the main paper.

1. Additional Experiments

In this section, we conduct additional experimental results to investigate the effectiveness of our framework.

Ablation study with the different weights $\lambda_{\text{da\_id}}$. In Table 1, we analyzed our MoS with the ablation evaluations with respect to the different weights $\lambda_{\text{da\_id}}$. In all experiments, we set the number of Top-K and negative samples as $K = 2$ and $N^{-} = 128$, respectively. The results show that our MoS with the different weights $\lambda_{\text{da\_id}}$ effectively learn a domain-invariant and discriminative representation, consistently boosting the performance of mAP and top-1 accuracy on CUHK-SYSU [4] → PRW [5] scenario. When $\lambda_{\text{da\_id}}$ is set as 0, it can be seen as the mixture of submodules with the domain- and task-conditional routing policy. Specifically, as $\lambda_{\text{da\_id}}$ is increased from 0, the performance of our MoS is improved. When the $\lambda_{\text{da\_id}}$ is above 0.1, it converges to high mAP and top-1 accuracy. The results indicate that our counterpart domain sample generation method effectively synthesizes augmented samples with consistent identity across domains, thereby efficiently learning domain invariant feature representations through contrastive domain alignment. Since the result with $\lambda_{\text{da\_id}} = 0.1$ has shown the best performance of mAP and top-1 accuracy on CUHK-SYSU [4] → PRW [5] scenario, we set $\lambda_{\text{da\_id}}$ as 0.1 for the remaining experiments.

Quantitative Evaluation for Detection Task. We present quantitative evaluation for the person detection task in Table 2. The results demonstrate that our methods also enhance the capability to localize the person. Specifically, as the PRW dataset [5] tends to be easier for object detection compared to CUHK-SYSU dataset [4], both our MoS and the baseline [1] achieve high recall and Average Precision (AP) on the CUHK-SYSU [4] → PRW [5] scenario. On the other hand, on PRW [5] → CUHK-SYSU [4] scenario, our MoS achieves a large performance improvement in both recall and AP, improving the recall by 1.4% and AP by 2.6% over the DAP [1]. It also indicates the effectiveness of the domain- and task-specific modeling for joint optimization with contradictory objectives in cross-domain scenarios.

# of Top-K	PRW	
	mAP	top-1
$\lambda_{\text{da\_id}} = 0$	36.0	81.5
$\lambda_{\text{da\_id}} = 0.001$	36.7	81.3
$\lambda_{\text{da\_id}} = 0.01$	36.9	81.5
$\lambda_{\text{da\_id}} = 0.1$	**37.1**	**81.9**
$\lambda_{\text{da\_id}} = 1$	36.7	81.2

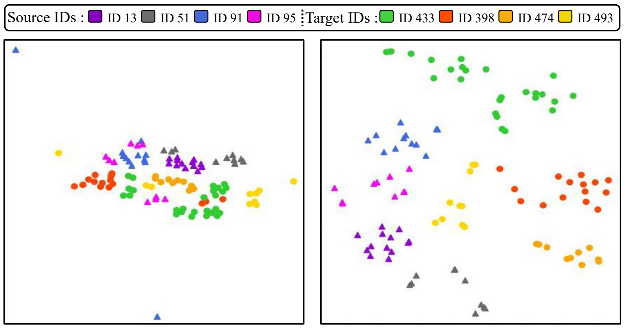
Table 1. **Ablation study for the different importance weights $\lambda_{\text{da\_id}}$ of contrastive domain alignment loss on CUHK-SYSU [4] → PRW [5].**

Methods		PRW		CUHK-SYSU	
		Recall	AP	Recall	AP
DA	DAPS [1]	**97.2**	**90.9**	77.7	69.9
	Ours	96.8	90.6	**79.1**	**72.5**

Table 2. **Quantitative evaluation for detection task on domain adaptive person search scenarios, including CUHK-SYSU [4] → PRW [5] and PRW [5] → CUHK-SYSU [4].**

2. t-SNE Visualization

To explain the effectiveness of our task-specific mixture of submodules, we provide the distribution of instance features for both detection and ReID in Fig. 1. We train our task-specific mixture of submodules on PRW [5] → CUHK-SYSU [4] scenario. Using test images from both CUHK-SYSU [4] and PRW [5], we extract RoI feature maps from the mixture of submodules for person detection and ReID, respectively. We then employ t-SNE [3] to transform these feature maps into two-dimensional points, allowing us to visualize the distribution of instance features. Each color represents the person identity for the source (*i.e.* CUHK-SYSU [4]) and target (*i.e.* PRW [5]) domain. The results show that, for person detection, the mixture of submodules is able to cluster instance features of different identities into a single group. In contrast, for person ReID, it effectively separates the instance features into distinct groups. It demonstrated that our MoS framework can learn identity-irrelevant representations across domains for person detection, as well as identity-relevant representations for person ReID, effectively dealing with the contradictory objectives in cross-domain scenarios of these two sub-tasks (*i.e.* person detection and ReID) within a unified framework.



(a) MoS (Detection) (b) MoS (ReID)

Figure 1. **Visualization of feature distribution by our MoS for (a) person detection and (b) ReID.** Data projection in 2-D space is attained by t-SNE based on the feature representation. Each color represents a different person identity. Our mixture of submodules for person detection can cluster the instance features with different identities into the same group, while the mixture of submodules for person ReID clusters the instance features into distinct groups.

3. Loss functions

Following the baseline [1], we adopt several loss functions for baseline $\mathcal{L}_{\text{base}}$, including detection loss $\mathcal{L}_{\text{det}}$, ReID loss $\mathcal{L}_{\text{id}}$, image-level domain alignment $\mathcal{L}_{\text{da\_img}}$ and instance-level domain alignment loss $\mathcal{L}_{\text{da\_ins}}$. In this section, we describe these losses in detail.

Detection Loss. For person detection, we adopt cross-entropy loss on the magnitude of these vectors v_{det} (*i.e.* person classification probabilities) such that

$$\mathcal{L}_{\text{det}} = -y \log(v_{\text{det}}) - (1-y)\log(1-v_{\text{det}}), \qquad (1)$$

where y denotes the person class label. We also adopt Smooth-L1 loss of regression vectors between ground-truth and predicted boxes following the Faster R-CNN [2].

ReID Loss. To enhance the discriminative power of the instance vector, we adopt a memory-based loss [1]. This loss is applied to the direction of instance vectors v_{id} for ReID, maximizing similarities among positive samples z^+ within the same identities, while minimizing the similarities among negative samples within different identities in memory $\mathbf{M}$ as follows:

$$\begin{aligned}
\mathcal{L}_{\text{id}} &= -\log \frac{\exp(v_{\text{id}} \cdot z^+/\tau)}{\sum_{z\in\mathbf{M}} \exp(v_{\text{id}} \cdot z/\tau)} \\
&\sum_{z\in\mathbf{M}} \exp(v_{\text{id}} \cdot z/\tau) = \\
&\sum_{k=1}^{N_t^c} \exp(v_{\text{id}} \cdot z_k/\tau) + \sum_{k=1}^{N_t^o} \exp(v_{\text{id}} \cdot z_k/\tau) + \\
&\sum_{k=1}^{N_s^c} \exp(v_{\text{id}} \cdot z_k/\tau) + \sum_{k=1}^{N_t^n} \exp(v_{\text{id}} \cdot z_k/\tau)
\end{aligned}, \qquad (2)$$

where N_t^c, N_t^o, N_s^c, and N_t^n denote the number of target cluster centroids, samples not belonging to any target cluster, source cluster centroids, and target hybrid hard case samples, respectively.

image-level domain alignment Loss. To align the intermediate feature distribution across domains, we adopt image-level domain alignment loss [1], as follows:

$$\mathcal{L}_{\text{da\_img}} = -\lambda_{\text{da\_img}} \sum_{N_I} [o \log D(f_{\text{id}}) + (1-o)\log D(f_{\text{id}})], \qquad (3)$$

where f_{img}, D, o, $\lambda_{\text{da\_img}}$, and $D(\cdot)$ are the intermediate feature, the discriminator, the domain, the weight for image-level domain alignment loss, and the discriminator.

Instance-level domain alignment Loss. We also adopt instance-level domain alignment [1] to learn the commonness of all persons across domains as follows:

$$\mathcal{L}_{\text{da\_ins}} = -\lambda_{\text{da\_ins}} \sum_{N_v} [o \log D(v_{\text{id}}) + (1-o)\log D(v_{\text{id}})], \qquad (4)$$

where v_{id}, N_v and $\lambda_{\text{da\_ins}}$ are the instance vectors, the number of instance vectors, the weight for instance-level domain alignment loss.

4. More Results

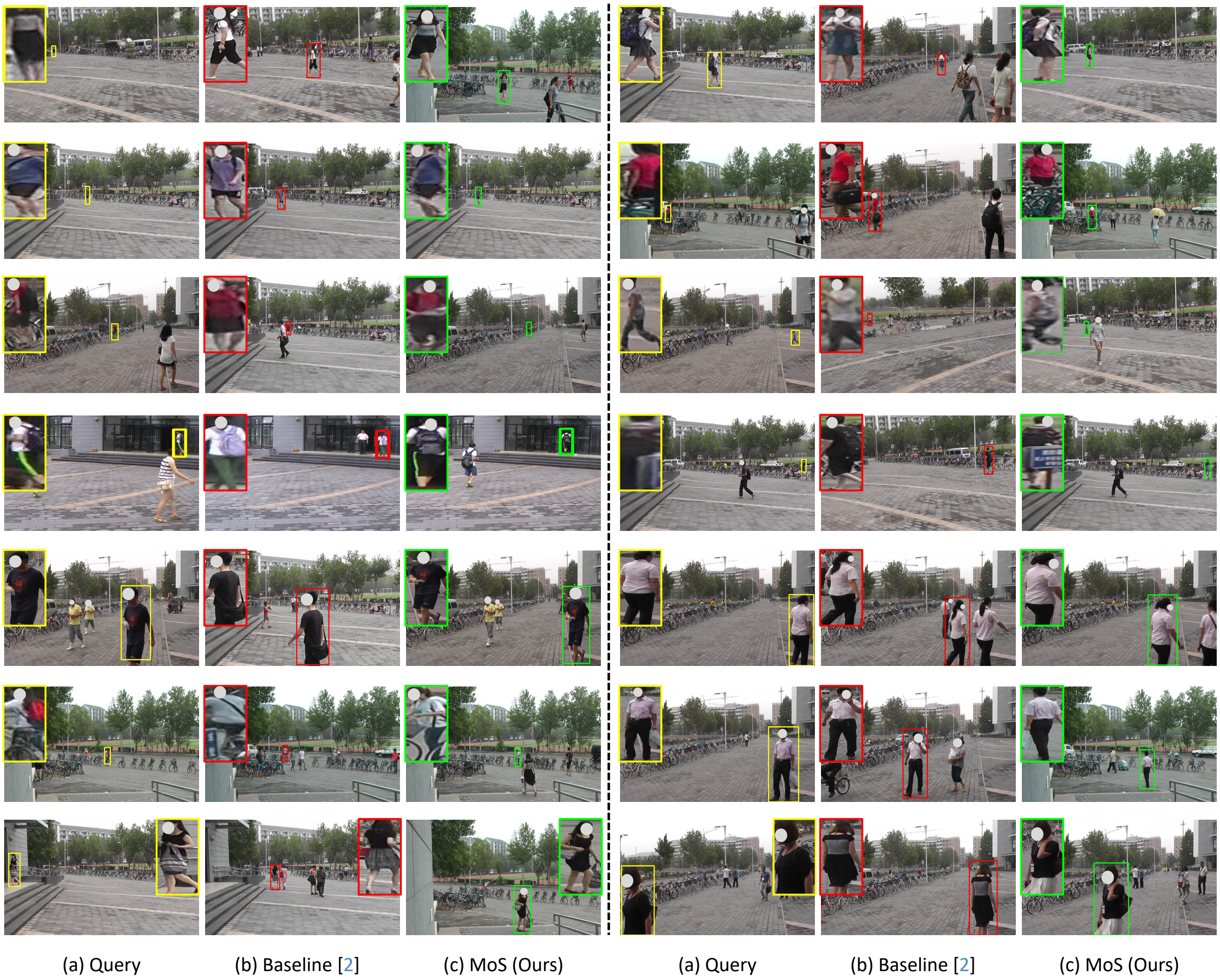
In this section, we provide additional qualitative results on domain adaptive person search scenarios, including CUHK-SYSU [4] → PRW [5] in Fig. 2 and PRW [5] → CUHK-SYSU [4] in Fig. 3.

References

[1] Junjie Li, Yichao Yan, Guanshuo Wang, Fufu Yu, Qiong Jia, and Shouhong Ding. Domain adaptive person search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 302–318. Springer, 2022. 1, 2, 3, 4

[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

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

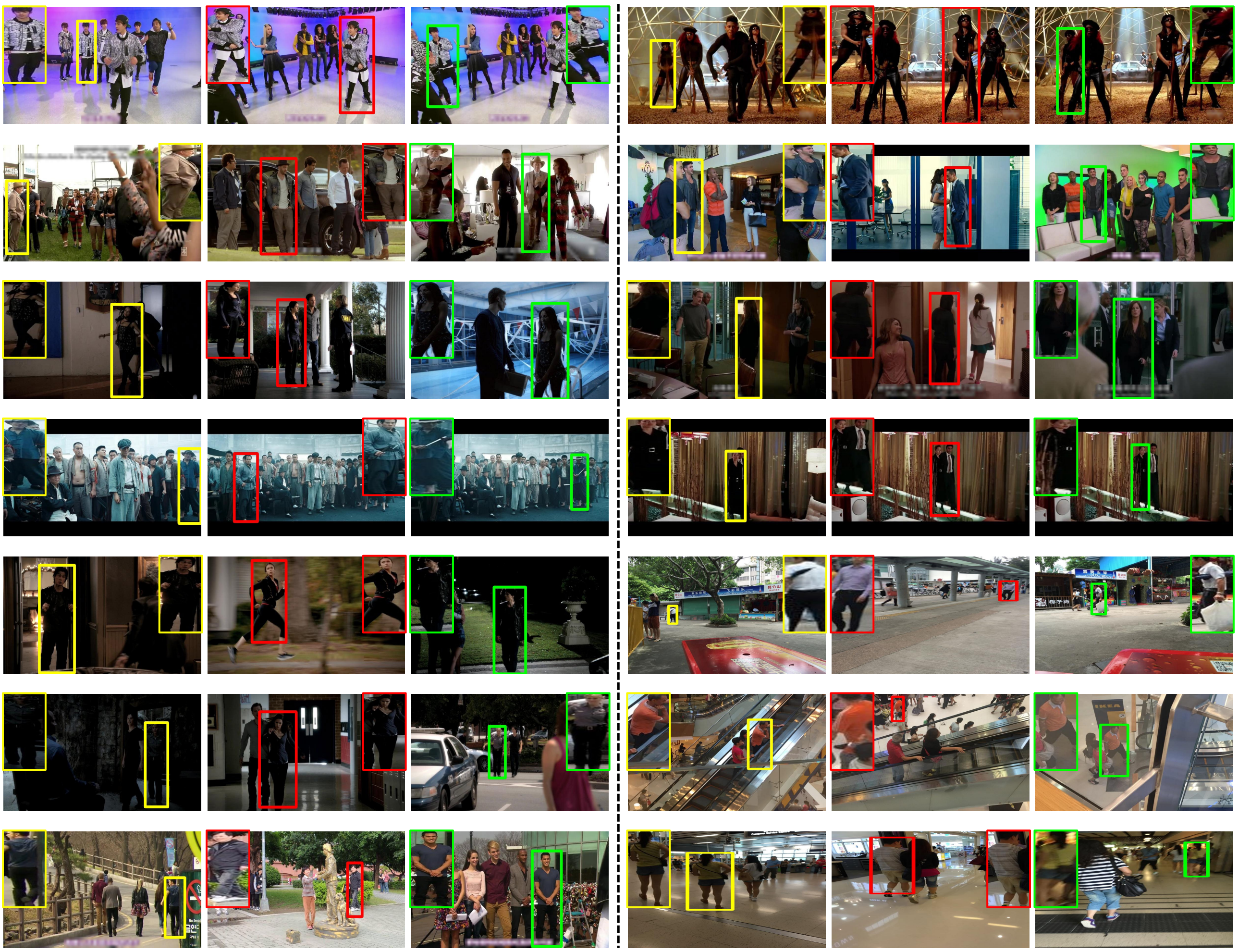


(a) Query (b) Baseline [2] (c) MoS (Ours) (a) Query (b) Baseline [2] (c) MoS (Ours)

Figure 2. **Qualitative comparison between (a) baseline [1] and (b) MoS (Ours) on CUHK-SYSU [4] → PRW [5] scenario.**

[4] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017. 1, 2, 3, 4

[5] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017. 1, 2, 3, 4



(a) Query (b) Baseline [2] (c) MoS (Ours) (a) Query (b) Baseline [2] (c) MoS (Ours)

Figure 3. **Qualitative comparison between (a) baseline [1] and (b) MoS (Ours) on PRW [5] → CUHK-SYSU [4] scenario.**