DDAM-PS: Diligent Domain Adaptive Mixer for Person Search

Mohammed Khaleed Almansoori*[1] Mustansar Fiaz*[1,2] Hisham Cholakkal[1]
[1]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE [2]IBM
(mohammed.almansoori, hisham.cholakkal)@mbzuai.ac.ae mustansar.fiaz@ibm.com

Abstract

Person search (PS) is a challenging computer vision problem where the objective is to achieve joint optimization for pedestrian detection and re-identification (ReID). Although previous advancements have shown promising performance in the field under fully and weakly supervised learning fashion, there exists a major gap in investigating the domain adaptation ability of PS models. In this paper, we propose a diligent domain adaptive mixer (DDAM) for person search (DDAP-PS) framework that aims to bridge a gap to improve knowledge transfer from the labeled source domain to the unlabeled target domain. Specifically, we introduce a novel DDAM module that generates moderate mixed-domain representations by combining source and target domain representations. The proposed DDAM module encourages domain mixing to minimize the distance between the two extreme domains, thereby enhancing the ReID task. To achieve this, we introduce two bridge losses and a disparity loss. The objective of the two bridge losses is to guide the moderate mixed-domain representations to maintain an appropriate distance from both the source and target domain representations. The disparity loss aims to prevent the moderate mixed-domain representations from being biased towards either the source or target domains, thereby avoiding overfitting. Furthermore, we address the conflict between the two subtasks, localization and ReID, during domain adaptation. To handle this cross-task conflict, we forcefully decouple the norm-aware embedding, which aids in better learning of the moderate mixed-domain representation. We conduct experiments to validate the effectiveness of our proposed method. Our approach demonstrates favorable performance on the challenging PRW and CUHK-SYSU datasets. Our source code is publicly available at `https://github.com/mustansarfiaz/DDAM-PS`.

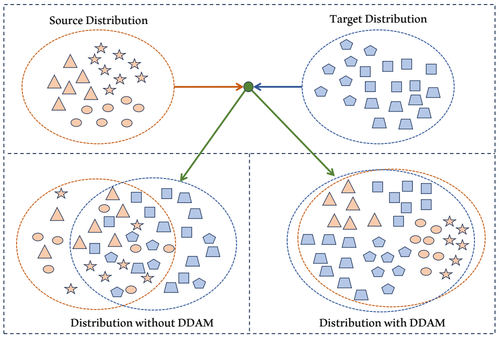

Figure 1: Demonstration of the impact of domain adaptation with and without our proposed diligent domain adaptive mixer (DDAM) module for the person search problem. Suppose, the source and target feature points are localized in hyperspace. In order to better transfer the source knowledge to the target domain, our proposed DDAM finds moderate mixed-domain distribution to bridge the gap between the source and target distributions. Here various shapes and colors donate the different distributions and different person identities correspondingly.

1. Introduction

Person search aims to optimize two conflicting subtasks: detection and re-identification (ReID) [39, 15, 6]. Detection focuses on localizing pedestrians in a given scene, while ReID is responsible for uniquely identifying individuals. This research problem becomes extremely complex due to the utilization of real-world data sources (such as CCTV), which often contain uncropped images with varying specifications, resolutions, lighting conditions, and other variations. While person search has been extensively explored under the fully supervised learning [39, 15, 1, 42] and weakly supervised learning [40, 22] paradigms, adapting it for unsupervised domain adaptation (UDA) generalization remains challenging, as there is a significant disparity between the distributions of the source and target domains.

Unsupervised domain adaptation (UDA) has demonstrated promising results in various domains, including aerial tracking [45], nighttime semantic segmentation [17],

[*] These authors contributed equally to this work.
[‡] Mustansar's work on this paper was done when he was at MBZUAI.

visual recognition [44, 25, 48], and person ReID [49, 32, 7]. Unlike fully supervised learning and weakly supervised learning, UDA focuses on bridging the gap between the ideal training set and real-world scenarios by leveraging labeled source data and transferring learned knowledge to unlabeled target domains. Li et al. [28] are the first to apply UDA to person search and proposed DAPS, a method that employs implicit alignment modules and pseudo-labeling to reduce the discrepancy between source and target domains. However, DAPS suffers from a lack of an explicit bridge to determine which critical information, such as similarity or dissimilarity, should be utilized to mitigate the domain discrepancy. Moreover, the implicit alignment modules employed in challenging real-world scenarios, where the person search (PS) model encounters scene challenges like occlusion and pose variations, as well as environmental challenges such as diverse indoor and outdoor scene distributions, may deteriorate the region of interest. Existing PS [15, 6, 30] methods based on Faster-RCNN [35] strive to jointly optimize the conflicting subtasks of detection and ReID. In an effort to address this issue, Chen et al. [6] introduced norm-aware embedding (NAE) to disentangle the two tasks. However, it still utilizes shared weights for both detection and ReID. Therefore, directly utilizing shared NAE representations for domain adaptation may further increase the complexity of person search.

To address the challenges mentioned above, we propose a diligent domain adaptive bridging mechanism to learn domain-invariant feature representations by introducing a bridge that reduces or minimizes the discrepancy between the two domains. Inspired by [9], we aim to enhance knowledge transfer between the source and target domains by learning mixed-domain representations from both domains. As discussed earlier, a significant domain shift exists between the distributions of the two domains. In Fig. 1, we illustrate the region of interest (RoI) proposals from the source and target distributions in hyperspace. Our bridging mechanism introduces hidden representations, referred to as moderate mixed-domain representations, with the objective of smoothly transferring RoI knowledge from the source domain to the target domain. To achieve this, we enforce two bridge losses on the moderate domain representations, minimizing the distance between the source and target domain representations. Additionally, we employ a disparity loss that regularizes the diversity between the two domains by maximizing the standard deviation. This regularization helps to avoid overfitting to either of the domains and facilitates gradual domain adaptation. Depending on the ambient nature of the mixed-domain representations, the source RoI labels can dominate or the inherent distribution of the target domain can be more exposed. The bridge losses and disparity loss work together to learn mixed-domain representations, allowing the model to effectively transfer source RoI knowledge and enhance discriminability in the target domain for the ReID task. Furthermore, we propose to decouple the norm-aware embeddings to mitigate the conflict between detection and ReID, which in turn simplifies the process of domain adaptation. Through experiments, we demonstrate that our approach surpasses the state-of-the-art method DAPS on the PRW and CUHK-SYSU datasets.

Contribution: Our contributions can be summarized as follows: (1) We propose an explicit diligent domain adaptive mixing mechanism to reduce the gap between the source and target domains in the person search domain adaptation problem. Specifically, we learn mixed domain representations that bridge the discrepancy between the two domains and facilitate the swift transfer of source information to the target domain, thereby promoting UDA person search tasks. (2) To enhance domain adaptation ability and generate elegant mixed domain representations, we introduce two bridge losses and a disparity loss. (3) To alleviate the conflict between detection and ReID and further improve domain adaptation, we propose the decoupling of the NAE representation. (4) Experimental results demonstrate the promising performance of our method on two datasets, outperforming state-of-the-art methods. These results highlight the merits of our approach.

2. Related Work

2.1. Person Search

Person search aims to unify the sub-tasks of localization of pedestrians [2, 34, 35] and re-identification of the person of interest [46, 21, 29] in an end-to-end model. The PS problem becomes a popular research topic, and methods start to focus on the challenges of the two contradictory objectives. The challenge comes when pedestrian detection aims to extract common features to improve localization, while ReID pushes to extract unique features of the same individual. PS problem can be classified as two-stage [26, 13, 19, 5] and one-stage [6, 15, 14, 47] methods. In two-stage methods, first detection is performed to locate the pedestrians employing off-the-shelf detectors, and later re-identification task is performed over the cropped pedestrians for identity discrimination. Although two-stage methods provide promising performance, they face immense computational costs.

On the contrary, one-stage methods perform both subtasks simultaneously in an end-to-end manner. These one-stage methods exploit the two-stage detector i.e., Faster RCNN [35], and combine additional ReID loss for pedestrian identity discrimination. For example, OIM [39, 50] utilized Faster RCNN to implement an end-to-end Person search model. NAE [6] disentangle the detection and the ReID into a norm and angle Euclidian representation, allowing to minimize the cross-task conflict. Furthermore, inherited disadvantages of Faster-RCNN affect the gains for PS, thus sequential models [30, 47, 15, 14, 27] mitigate the low-quality proposal of the RPN. Seqnet [30] sequential structure allowed the model to focus on reducing the cross-task conflict by getting a better proposal and for the final stage to focus more on the ReID. The COAT [47] utilized transformer encoders to shuffle patches of individuals with each other in other to generalize better for unseen images. Studies such as PS-ARM [15] introduce the attention-aware relation mixer to exploit the relations between different local regions within RoI of a person.

The works [1, 42] motivate to further disentangle the two-sub tasks, by moving away from Faster R-CNN structure due to limitation and computational resources. To address these issues, AlignPS [42] uses anchor free approach to eliminate the need for using low-quality proposals. In addition, utilizing an aligned feature aggregation module mitigates the issues of scale, region, and task alignment. Cao et al. [1] introduces Deformable Detr [53] for PS that simultaneously predicts the detection and ReID. However, these fully supervised methods (FSL) methods suffer from the issue of domain gap which, degrades the performance of the model. To minimize the domain shift issue, recent studies in weakly supervised person search (WSPS) [40, 22] have access to bounding boxes with ID annotations. Although these issue helps to reduce the domain gap, they still require label data. Another recent study is DAPS [28] which introduces the concept of UDA in PS. DAPS focuses on the domain alignment between the source and the target domain, and also the pseudo-labeling framework for the target domain. In contrast, we propose a novel bridging mechanism that enhances the discriminative learning for ReID by bridging the gap between the source and target domains as well as minimizing the cross-task conflict with localization to ease the domain generalization task.

2.2. Domain Adaptation for Person ReID

Unsupervised domain adaptation (UDA) for person ReID approaches are exposed to labeled source domain and translate the learned knowledge to the underlying target domain in an unsupervised manner. The UDA person ReID approaches are classified into three categories based on their training strategies including GAN transferring [11, 37], joint training [52, 18], and fine-tuning [8, 16]. GAN transferring approaches utilize GAN models to disentangle the style discrepancy and transfer the learned information from the source to the target domain. For joint training, the approaches employ a memory bank that combines the source and target data and jointly trains without building a bridge between the two domains to improve the target domain features. However, for fine-tuning methods, they train the model for source data and fine-tune over target data using pseudo labels. The key component is to mitigate the effect of noisy pseudo labels. Nevertheless, UDA person ReID is based on cropped pedestrians and cannot be directly extended for the person search problem. The DAPS [28] proposed a clustering mechanism to provide high-quality pseudo labels to expedite the target domain training. However, DAPS implicitly utilizes the source and target data while ignoring the explicit bridge mechanism to alleviate the gap between the two domains. Therefore, inspired by [9], we introduced an explicit mechanism to learn what similar/dissimilar information can be employed to improve the target domain features for ReID.

3. Method

The overall framework of our proposed diligent domain adaptive mixer for person search, DDAM-PS, is illustrated in Fig. 2. It jointly takes input from both the source and target domains. The base network of our framework is DAPS [28], which incorporates an implicit domain alignment module (DAM) to reduce the gap between the two domains. The source and target domain images are fed into a ResNet50 [24] backbone network to extract feature embeddings. These embeddings are then input to the region proposal network (RPN) [35] to generate ROI-Aligned proposal candidates. To enhance the ReID task within the baseline, we introduce a diligent domain adaptive mixing (DDAM) mechanism. This mechanism aims to smooth out the extreme differences between the source and target domains, allowing for better domain adaptation. To achieve this, we fuse the source and target domain proposals to generate new mixed domain proposal representations. For the detection task, we employ a combination of box regression head ($\mathcal{L}_{bbox}$) and person vs background classification head ($\mathcal{L}_{Bg/Person\_cls}$) to compute detection losses. For the ReID task, we impose OIM [39] ReID loss, denoted as $\mathcal{L}_{ReID}$, on the source and target domain features. Pseudo-labels for the target domain are generated using a clustering strategy. In order to generate moderate mixed-domain adaptive representations, we introduce two bridge losses, $\mathcal{L}^{f}_{bridge}$ and $\mathcal{L}^{\varphi}_{bridge}$. The $\mathcal{L}^{f}_{bridge}$ loss is applied by utilizing the NAE embedding of the target-domain, source-domain, or mixed-domain representations to evaluate the distance across the domains. On the other hand, $\mathcal{L}^{\varphi}_{bridge}$ is enforced using a hybrid memory projection module to measure the discrepancy between the mixed domain memory projections and the two domains. Additionally, we employ a disparity loss to regulate the domain mixing mechanism and prevent overfitting to either of the two extreme domains. As mentioned earlier, person search models face challenges in jointly optimizing the two subtasks of object detection and ReID. When adapting these models for UDA, the complexity further increases. To address this issue, we propose to decouple the norm-aware embeddings (NAE). This decoupling not only allevi-

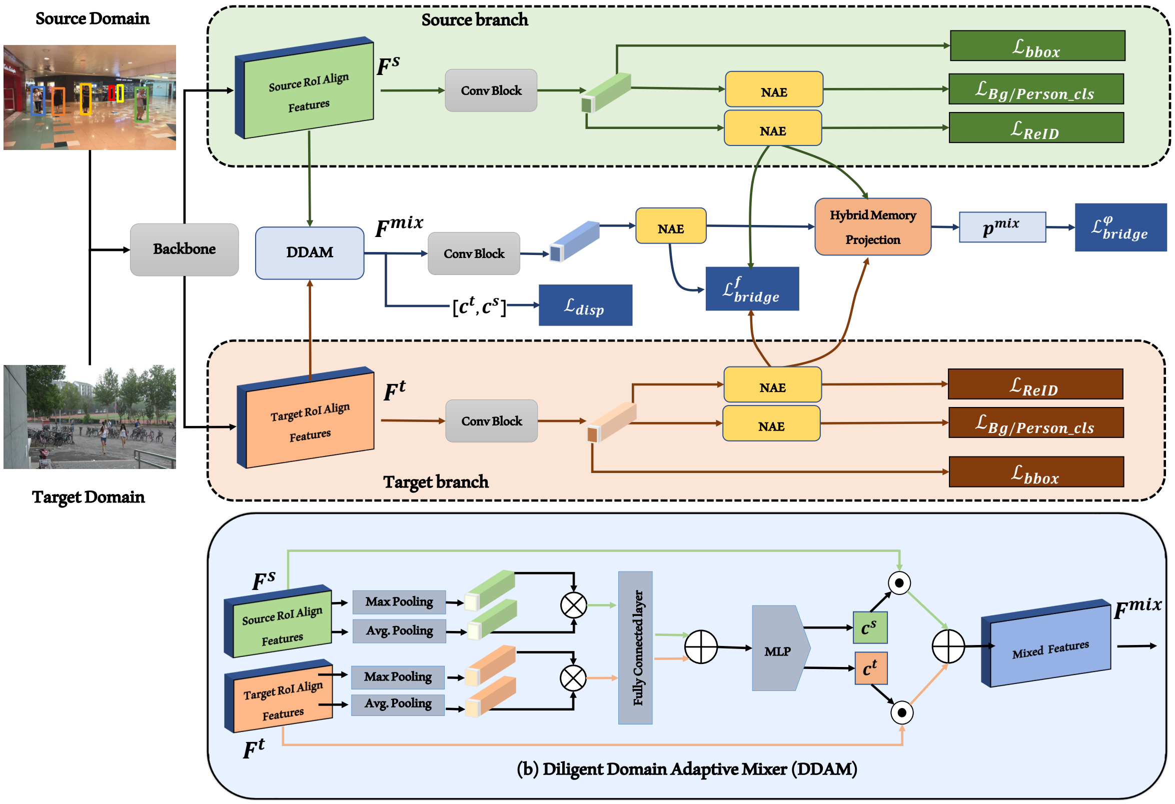

Figure 2: The illustration of our proposed diligent domain adaptive mixer person search (DDAM-PS) framework. The source and target stem features are computed using a backbone and input to the RPN [35] to compute RoI align features. These source and target RoI align features (F^s, F^t) are fed to the diligent domain adaptive mixer (DDAM) module to generate the mixed domain representations (F^{mix}), to reduce the domain gap for unsupervised domain adaptation (UDA), as shown in (b). To generate moderate mixed domain representations, we employ two bridges losses ($\mathcal{L}^{f}_{bridge}$ and $\mathcal{L}^{\varphi}_{bridge}$) and a disparity loss ($\mathcal{L}_{disp}$). The $\mathcal{L}^{f}_{bridge}$ loss is applied using the NAE embedding of the target-domain, source-domain, or mixed-domain representations to evaluate the distance across the domains. While $\mathcal{L}^{\varphi}_{bridge}$ is enforced using a hybrid memory projection module to measure the discrepancy between the mixed domain memory projections and the two domains. The disparity loss is enforced to regulate the mixed domain features, to avoid overfitting using constraint weights (c^s, c^t), obtained from the DDAM module. In addition, we propose to decouple the NAE module and apply separate NAE for both conflicting subtasks i.e., detection and ReID. This decoupling facilitates to adopt it for the UDA ReID task.

ates the conflict between the two subtasks but also improves the PS domain adaptation framework.

3.1. Diligent Domain Adaptive Mixer (DDAM)

Inspired by [9], we propose an explicit mixed domain representation learning approach to enhance knowledge transfer between source and target domains for UDA PS.The DDAM module takes n pairs of RoI pooled features from both the source (F^s) and target (F^t) domains and generates domain constraint weights, denoted as c^s and c^t, respectively. The source (F^s) and target (F^t) RoI features are realized with the average and maximum pooling operations. These pooled features are then concatenated for each domain and passed through a shared fully connected (FC) layer. The features from the FC layer are merged via element-wise summation operation and input to a multi-layer perceptron (MLP) followed by a Softmax activation function, yielding the domain constraint weights. The overall procedure to obtain the domain constraint weights is illustrated in Figure 2-(b). The two domain constraint weights c^s and c^t are represented as $[c^s, c^t] = c$, where $c \in \mathbb{R}^2$. Finally, the RoI mixed domain representations are achieved by mixing the source RoI features and target RoI features using the two domain constraint weights as follows:

$$F^{mix} = c^s \cdot F^s + c^t \cdot F^t. \tag{1}$$

Table 1: The quantitative comparison of our's unsupervised domain adaptive (UDA) method with fully supervised state-of-the-art methods on both the CUHK-SYSU and PRW datasets. The performance is evaluated using mAP and top-1 accuracy. Our method scores are in bold.

Method		CUHK-SYSU		PRW	
		mAP	top-1	mAP	top-1
Two-step	CLSA [26]	87.2	88.5	38.7	65.0
	IGPN [13]	90.3	91.4	42.9	70.2
	DPM [19]	-	-	20.5	48.3
	RDLR [23]	93.0	94.2	42.9	70.2
	MGTS [5]	83.0	83.7	32.6	72.1
	TCTS [36]	93.9	95.1	46.8	87.5
End-to-end	OIM [39]	75.5	78.7	21.3	49.9
	RCAA [3]	79.3	81.3	-	-
	NPSM [31]	77.9	81.2	24.2	53.1
	IAN [38]	76.3	80.1	23.0	61.9
	QEEPS [33]	88.9	89.1	37.1	76.7
	CTXGraph [43]	84.1	86.5	33.4	73.6
	HOIM [4]	89.7	90.8	39.8	80.4
	BINet [12]	90.0	90.7	45.3	81.7
	APNet [51]	88.9	89.3	41.2	81.4
	AlignPS [41]	93.1	93.4	45.9	81.9
	AlignPS + [42]	94.0	94.5	46.1	85.8
	NAE [42]	91.5	92.4	43.3	80.9
	SeqNet [30]	93.8	94.6	46.7	83.4
	PSTR [1]	93.5	95.0	49.5	87.8
	OIMNet++ [27]	93.1	93.9	46.8	83.9
UDA	**Ours**	**79.5**	**81.3**	**36.7**	**81.2**

Table 2: Comparison of our method with weakly supervised methods and domain adaptive state-of-the-art PS methods over PRW and CUHK-SYSU datasets. The * indicates the training of R-SiamNet using both CUHK-SYSU and PRW. The best results are in bold.

Methods		PRW		CUHK-SYSU	
		mAP	top-1	mAP	top-1
Weakly-Supervised	CGPS [40]	16.2	68.0	80.0	82.3
	R-SiamNet [22]	21.4	75.2	86.0	87.1
	R-SiamNet$*$ [22]	23.5	76.0	86.2	87.6
UDA	DAPS [28]	34.7	80.6	77.6	79.6
	Ours	**36.7**	**81.2**	**79.5**	**81.3**

3.2. Moderate Domain Mixing

The effectiveness of domain mixing can be hindered by two factors: (1) The RoI pooled feature samples often exhibit diverse backgrounds, and individuals within both the intra-domain and inter-domain distributions may experience appearance variations. This includes challenges related to environmental factors such as indoor and outdoor scenes. (2) From Equation 1, we can generate an infinite number of mixed domain representations by exposing source and target domain's RoI features to the DDAM module. However, only a limited portion of these mixed domain representations is capable of effectively bridging the gap between the two extreme domains. These factors can potentially degrade the quality of the mixed domain representations.

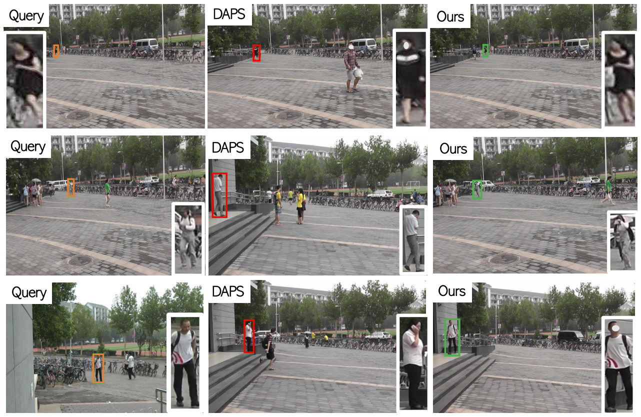

Figure 3: Qualitative comparison between the DAPS [28] and our's approach in three different challenging scenes. Our method predicts correct top-1 matching results. The orange, red, and green colors show the query, failure, and correct, respectively.

In order to better learn the mixed domain distribution (P_{mix}), the source distribution (P_s) and target distribution (P_t) should be located on the shortest path [20] (see Fig. 1). Although the baseline learns the domain-invariant representations using the DAM module, this approach does not take into account the extreme classes for each domain which is more likely to affect the class distributions in each domain. Therefore, considering the shortest distance definition, the mixed domain representations should follow the two desired characteristics which are ensured by enforcing the dedicated losses. To bridge the extreme domains in the hyperspace, the distance $d(.)$ should be proportional where $c^s + c^t = 1$ (using softmax function) and $c^s, c^t \in [0, 1]$. Thus, the moderate mixed domain representation can be obtained utilizing domain constraint weights by identifying the closest points to both P^s and P^t as well as localized along the shortest path. The problem can be framed as loss minimization as follows:

$$\mathcal{L}_{bridge} = c^s \cdot d(P_s, P_{mix}^{(c)}) + c^t \cdot d(P_t, P_{mix}^{(c)}). \quad (2)$$

The enforced loss (Eq. 2) controls the gap between two domains by minimizing the shift between the two domains. The $\mathcal{L}_{bridge}$ loss will punish more $d(P_t, P_{mix}^{(c)})$ if $c^t > c^s$, else it will push more $d(P_s, P_{mix}^{(c)})$. The domain constraint weights (c^t, c^s) in DDAM ensure a steady domain adaptive procedure to balance the minimization of the domain shifts from the source to the target domains.

We impose bridge losses on the mixed domain feature representations and feed them to NAE for the ReID task.

Since online instance matching (OIM) [39] utilizes memory to keep the features for the labeled and unknown identities using a lookup table (LUT) and a circular queue (CQ). The LUT is defined as $V \in \mathrm{R}^{D \times L}$ where D and L are the feature dimensions and IDs respectively, and CQ is represented as $U \in \mathrm{R}^{D \times Q}$ where Q is the queue size. It is impractical to directly utilize the OIM for the UDA PS task. Therefore, we extended the OIM for the UDA ReID and introduced a hybrid memory projection module to keep LUT for the known source IDs and pseudo-labeled target IDs. However, we keep a single CQ for both source and target unknown identities. Using the hybrid memory for the extended OIM, we calculate the similarity projection p_k for the input feature sample w.r.t. the LUT IDs as follows:

$$p_k = V_i^T f_k, \tag{3}$$

where k indicates the source domain, target domain, or mixed domain, and f_k denotes the feature embeddings from the kth domain. To quantify the discrepancy in the domain distribution between the mixed domain memory projections and the other two extreme domains, we make employ cross-entropy loss as in Eq. 4. This ensures that the dynamic properties of the hybrid memory projection module are compatible with the bridging method and allows it to work in the person search domain. We employ the $L2-Norm$ loss for the feature space to evaluate the distance across the domains (in Eq. 5) to maintain the shortest path for the mixed feature w.r.t. the source domain and target domain. The proposed two bridge losses are as follows:

$$\mathcal{L}_{bridge}^{\varphi} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k \in [s,t]} c_i^k \cdot [y_k^i log(p_{mix}^i))], \tag{4}$$

$$\mathcal{L}_{bridge}^{f} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k \in [s,t]} c_i^k \cdot ||f_i^k - f_i^{mix}||_2, \tag{5}$$

where k represents the domain (i.e., source or target) and i indicates the index in the minibatch. The y_k^i shows the source label or target pseudo label, the f_i^k denotes the kth domain's representation, p_i^{mix} reflects the mixed domain similarity projection, and f_i^{mix} means the mixed domain features (using the f_i^s and f_i^t), from the proposed DDAM module, respectively.

Another important property is to ensure that the mixed domain is diverse enough so that the source or the target domain does not dominate each other. To maximize the diversity of the domain constraint weights, we utilize the disparity loss. Where within the mini-batch the standard deviation $\sigma(\cdot)$ is used as follow:

$$\mathcal{L}_{disp} = -[\sigma(\{c_i^s\}_{i=1}^n) + \sigma(\{c_i^t\}_{i=1}^n)], \tag{6}$$

where σ denotes the computation of standard deviations in a mini-batch. The imposed disparity loss guarantees that the mixed domain representations are as much diverse as possible to maintain the shortest geodesic path property, which can better bridge the domain gap between the source and target domains.

3.3. Decoupled Norm-aware Embedding

As previously discussed, there is a fundamental conflict between the two subtasks, namely detection, and ReID, within the Faster RCNN-based [35] person search frameworks. These subtasks are exposed to the same backbone network, where detection focuses on capturing common features of pedestrians, while ReID aims to discriminate the uniqueness of individuals. In fully supervised person search settings, norm-aware embeddings (NAE) take the feature vector, pass it through a shared projection layer, and decouple it into two components: norm and angle in the polar coordinate system. However, the introduction of domain adaptation adds an additional layer of complexity to the process. Therefore, we intentionally decouple the NAE for both the detection and ReID tasks. This decoupling not only mitigates the cross-task conflict but also facilitates a more efficient handling of the ReID task in the context of UDA person search problems.

4. Experiments

4.1. Implementation Details

We implemented our method in the PyTorch framework and all experiments are performed over NVIDIA RTX A6000 GPU. Our backbone is ResNet50 [24] pre-trained over ImageNet-1K [10]. We resize the input to 1500×900, adopt a random horizontal flip as augmentation, and trained our model using the Stochastic Gradient Descent (SGD) method. We train the model for 20 epochs when the target dataset is set to PRW witg batchsize of 6 and train for 10 epochs when the target dataset is set to CUHK-SYSU dataset with batch size of 4. The weight decay and momentum are set to 5×10^{-4} and 0.9, respectively. Following [28], we set the learning rate 0.0024, which is reduced at epoch 16 by a factor of 0.1, and warms up at the first epoch. The annotations for the source domains are available during the training, whereas neither the bounding boxes of the pedestrians nor their identity information is accessible for the target domain during the training and test time. Following DAPS [28], we adopt an asynchronized training strategy and employ pseudo-bounding boxes after the α epochs (α=12 for target PRW and α=1 for CUHK-SYSU) on the target branch to supervise the box regression and classification heads. This releases the complexity of the unlabeled target domain images for both detection and ReID training. We utilized DDAM,to generate domain invariant representations, at the training and relinquish it during the inference time.

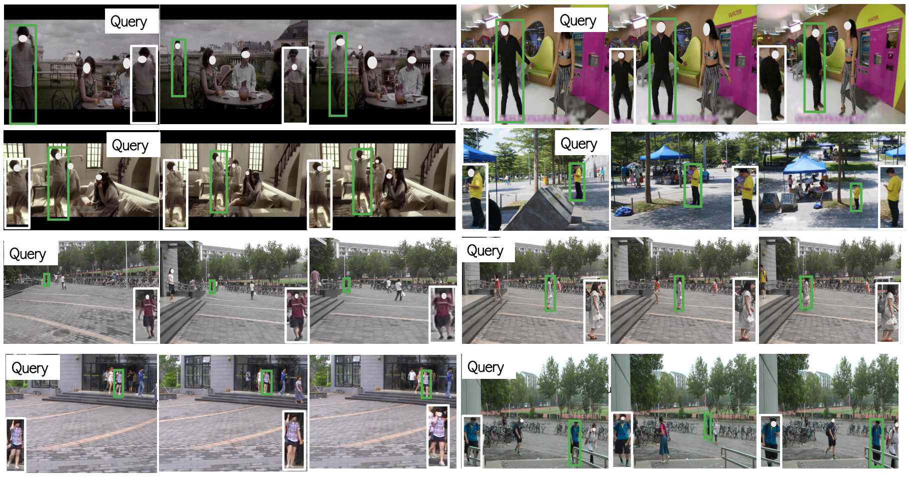

Figure 4: Qualitative analysis on CUHK-SYSU [39] (top 2 rows) and PRW [50] (bottom 2 rows) datasets. We illustrate the top two matching results for different query persons. Our method can effectively bridge the gap using adaptive domain mixing which correctly detects and identifies.

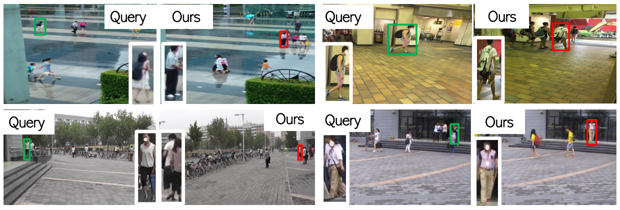

Figure 5: Failure cases on CUHK-SYSU [39] (first row) and PRW [50] (second row) datasets. We demonstrate that our approach incorrectly identifies the query person due to heavy domain conflicts between the domains.

4.2. Datasets and Metrics

Dataset We evaluate our method over the following two datasets, CUHK-SYSU [39] and PRW [50]. **CUHK-SYSU**[39] is a large dataset for a person search with 8,432 ID individuals across 18,184 images accounting for 96,143 bounding boxes. In training, only 5,532 IDs are accessible through 11,206 images while the remaining 2900 IDs and corresponding 6,978 images are used for evaluation. The CUHK-SYSU contains two distinct data sources; 1) street view images that contain a series of variations focusing on viewpoints, lighting, resolutions, and occlusions. 2) Movies and drama serial videos that contain a variety of unique indoor and outdoor challenges. This allows the dataset to add more diversity to the scenes. For evaluation, the images are split into 2900 query persons and the 6978 images are utilized as the gallery set. **PRW**[50] is another dataset consisting of 932 IDs having 11,816 images with 43,110 bounding boxes. The dataset is sampled from videos that were captured from six CCTV university cameras. For training, only 482 IDs are available in 5702 images while the test set has 2057 query persons with a gallery size of 6112 images.

Evaluation Protocols: For the domain adaptation setting, we evaluate our method on the test set of the target domain. In order to quantify the localization/detection task, we use standard object detection protocols such as recall score and average precision. We adopt widely used metrics cumulative matching characteristics (CMC) curves and mean average precision (mAP) to measure the performance of the ReID task. Since ReID reflects the identity of the query person, it is the most challenging metric for the PS task.

4.3. Comparison with State-of-the-art Methods

We compared our method with fully supervised, weakly supervised, and unsupervised domain adaption methods. First, we present a comparison of our UDA method with the fully supervised methods classified as two-stage and one-stage methods in Table 1. Surprisingly, our method outperforms several two-stage and one-stage fully supervised methods including DPM [19], MGTS [5], OIM [39], NPSM [31], IAN [38], and CTXGraph [43]. Second, we also compare our method with weakly supervised and UDA methods in Table 2. Compared to the top-performing weakly supervised method, our approach obtains an absolute gain of 13.2% over the PRW dataset. Compared to DAPS, our method demonstrates outstanding performance depicting the merits of our method and archives 2.0% and 1.9% gain in terms of mAP over both PRW and CUHK-SYSU datasets, respectively.

We present the qualitative comparison of our method with DAPS in Fig. 3 which depicts that our method is able to correctly identify the query person in complex scenes.

Table 3: Ablation study on the PRW and CUHK-SYSU datasets. Here, we show the merits of our contributions introduced to the baseline (DAPS [28]). The $\mathcal{L}^{f}_{bridge}$ & $\mathcal{L}^{\phi}_{bridge}$ represent the bridge losses, $\mathcal{L}_{disp}$ denotes the disparity loss, and DC-NAE indicates the decoupled NAE. We note that the integration of our noval bridge losses (row 5) and disparity loss (row 6) leads to consistent gain in terms of mAP for both datasets. Similarly, introduced losses (row 7) and decoupled NAE (row 8) obtained better results compared to the baseline. Our final approach (row 9) achieves significant performance gain compared to the baseline and its results are in bold.

Experiments					Target:PRW				Target:CUHK-SYSU			
Exp. No.	$\mathcal{L}^{f}_{bridge}$	$\mathcal{L}^{\phi}_{bridge}$	$\mathcal{L}_{disp}$	DC-NAE	mAP	Top-1	Recall	AP	mAP	Top-1	Recall	AP
1 (baseline)	✗	✗	✗	✗	34.7	80.6	97.2	90.9	77.6	79.6	77.7	69.9
2 (baseline reproduced)	✗	✗	✗	✗	34.4	78.4	92.1	87.5	77.1	78.2	72.8	67.9
3	✓	✗	✗	✗	34.9	78.9	92.5	87.6	77.9	79.2	73.4	68.1
4	✗	✓	✗	✗	34.7	78.6	92.4	87.8	77.4	79.3	73.9	68.5
5	✓	✓	✗	✗	35.1	79.4	92.9	88.1	78.1	79.7	74.1	68.3
6	✗	✗	✓	✗	35.7	79.0	92.6	88.0	78.3	79.8	74.9	68.1
7	✓	✓	✓	✗	35.9	79.5	92.5	88.4	78.5	80.7	75.4	68.7
8	✗	✗	✗	✓	35.5	79.4	93.1	88.2	78.6	80.3	74.7	68.2
9 (Ours)	✓	✓	✓	✓	**36.7**	**81.2**	**93.3**	**88.6**	**79.5**	**81.3**	**76.5**	**68.8**

Table 4: A study on how efficiently the proposed methods can adapt to a reduced-size target dataset.

	Target:PRW				Target:CUHK-SYSU			
Method / Sample Percentage	mAP	Top-1	Recall	AP	mAP	Top-1	Recall	AP
Baseline / 100%	34.4	78.4	92.1	87.5	77.1	78.2	72.8	67.9
Ours / 50%	32.7	77.6	91.4	87.0	76.5	77.1	72.5	65.9
Ours / 75%	34.8	78.9	92.4	87.3	77.2	79.0	74.5	67.8
Ours / 100%	36.7	81.2	93.3	88.6	79.5	81.3	76.5	68.8

More examples from CUHK-SYSU and PRW datasets are shown in Fig. 5. This shows that our DDAM module facilitates correctly localizing and identifying the query person in challenging scenarios. In Fig. 5, we also present failure cases where there exist heavy domain differences.

4.4. Ablation Study

We conduct an ablation study to validate the merits of our method in Table 3. As mentioned earlier, we adopt DAPS [28] as our baseline. For a fair comparison, we reproduce the baseline numbers and report in Table 3 (row 2). We integrated DDAM into the baseline and trained the model using introduced bridge losses (rows 3, 4, and 5) and disparity loss (low 6). We notice that combined bridge and disparity losses have more gain compared to individual bridge losses in terms of mAP for both datasets. When integrating our proposed DDAM (trained with three introduced losses) into the baseline (row 7), the mAP score is significant improved to 35.9% and 78.5% in terms of mAP over PRW and CUHK-SYSU datasets, respectively. This is attributed to the nature of DDAM since the objective is not to improve the quality of feature extraction but to minimize the disparity between the two domains without a label ID class as well as try to maintain diversity for both domains. Similarly, the decoupling of the NAE into the baseline (row 8) leads to improving the mAP scores over both PRW and CUHK-SYSU datasets. This is due to mitigating the issue of conflicting objectives of commoners, uniqueness, and adaption. Separating the NAE for detection and ReID eases the PS process. Finally, combining both contributions (row 9) leads to a significant improvement in performance and obtains mAP scores of 36.7% and 79.5% for both PRW and CUHK-SYSU datasets, respectively.

To further verify the impact of our new module, we studied how much reducing the number of training samples might affect the performance of the model. In Table 4, we see that the model obtains comparable results with the baseline model scores even when one-fourth of the target training set is removed (row 3).

5. Conclusion

We present a novel UDA person search framework that leverages a bridging mechanism to generate domain-invariant representations. Our approach introduces the DDAM module, which produces moderate mixed domain representations that effectively adapt the extremes of the two domains through an adaptive mixing mechanism, facilitating improved knowledge transfer from the source domain. To enhance the discriminability of the model on the target domain, we employ bridge and disparity losses. Additionally, we incorporate an NAE-decoupled module to mitigate the cross-task conflict, resulting in enhanced ReID quality and improved domain adaptation for the person search task. Our proposed contributions significantly enhance the model's domain adaptation abilities for person search. Our experimental studies validate the effectiveness of our proposed method.

Acknowledgement

This work is partially supported by the MBZUAI-WIS Joint Program for AI Research (Project grant number- WIS P008)

References

[1] Jiale Cao, Yanwei Pang, Rao Muhammad Anwer, Hisham Cholakkal, Jin Xie, Mubarak Shah, and Fahad Shahbaz Khan. Pstr: End-to-end one-step person search with transformers. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[2] Jiale Cao, Yanwei Pang, Jin Xie, Fahad Shahbaz Khan, and Ling Shao. From handcrafted to deep features for pedestrian detection: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[3] Xiaojun Chang, Po-Yao Huang, Yi-Dong Shen, Xiaodan Liang, Yi Yang, and Alexander G. Hauptmann. Rcaa: Relational context-aware agents for person search. *Proc. European Conference on Computer Vision*, 2018.

[4] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Bernt Schiele. Hierarchical online instance matching for person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10518–10525, 2020.

[5] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model, 2018.

[6] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. In *CVPR*, 2020.

[7] Deqiang Cheng, Jiahan Li, Qiqi Kou, Kai Zhao, and Ruihang Liu. H-net: Unsupervised domain adaptation person re-identification network based on hierarchy. *Image and Vision Computing*, 124:104493, 2022.

[8] Yongxing Dai, Jun Liu, Yan Bai, Zekun Tong, and Ling-Yu Duan. Dual-refinement: Joint label and feature refinement for unsupervised domain adaptive person re-identification. *IEEE Transactions on Image Processing*, 30:7815–7829, 2021.

[9] Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, and Ling-Yu Duan. Idm: An intermediate domain module for domain adaptive person re-id, 2021.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[11] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 994–1003, 2018.

[12] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Bi-directional interaction network for person search. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2836–2845, 2020.

[13] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Instance guided proposal network for person search. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[14] Mustansar Fiaz, Hisham Cholakkal, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Sat: Scale-augmented transformer for person search. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4820–4829, 2023.

[15] Mustansar Fiaz, Hisham Cholakkal, Sanath Narayan, Rao Muhammad Anwar, and Fahad Shahbaz Khan. Ps-arm: An end-to-end attention-aware relation mixer network for person search. In *Proceedings of the ACCV Asian Conference on Computer Vision*, 2022.

[16] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 6112–6121, 2019.

[17] Huan Gao, Jichang Guo, Guoli Wang, and Qian Zhang. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9913–9923, 2022.

[18] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems*, 33:11309–11321, 2020.

[19] Ross Girshick, Forrest Iandola, Trevor Darrell, and Jitendra Malik. Deformable part models are convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 437–446, 2015.

[20] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Unsupervised adaptation across domain shifts by generating intermediate data representations. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2288–2302, 2013.

[21] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1060–1069, 2022.

[22] Chuchu Han, Kai Su, Dongdong Yu, Zehuan Yuan, Changxin Gao, Nong Sang, Yi Yang, and Changhu Wang. Weakly supervised person search with region siamese networks, 2021.

[23] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-id driven localization refinement for person search, 2019.

[24] 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.

[25] Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu, and Ling Shao. Category contrast for unsupervised domain adaptation in visual tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1203–1214, 2022.

[26] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person search by multi-scale matching. *Proc. European Conference on Computer Vision*, 2018.

[27] Sanghoon Lee, Youngmin Oh, Donghyeon Baek, Junghyup Lee, and Bumsub Ham. Oimnet++: Prototypical normalization and localization-aware learning for person search. In

European Conference on Computer Vision, pages 621–637. Springer, 2022.

[28] Junjie Li, Yichao Yan, Guanshuo Wang, Fufu Yu, Qiong Jia, and Shouhong Ding. Domain adaptive person search, 2022.

[29] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2907, 2021.

[30] Zhengjia Li and Duoqian Miao. Sequential end-to-end network for efficient person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2011–2019, 2021.

[31] Hao Liu, Jiashi Feng, Zequn Jie, Jayashree Karlekar, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural person search machines. *Proc. IEEE International Conference on Computer Vision*, 2017.

[32] Anwesh Mohanty, Biplab Banerjee, and Rajbabu Velmurugan. Ssmtreid-net: Multi-target unsupervised domain adaptation for person re-identification. *Pattern Recognition Letters*, 163:40–46, 2022.

[33] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[34] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-guided attention network for occluded pedestrian detection. *Proc. IEEE International Conference on Computer Vision*, 2019.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[36] Cheng Wang, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Tcts: A task-consistent two-stage framework for person search. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[37] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person reidentification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018.

[38] Jimin Xiao, Yanchun Xie, Tammam Tillo, Kaizhu Huang, Yunchao Wei, and Jiashi Feng. Ian: The individual aggregation network for person search. *Pattern Recognition*, 2019.

[39] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search, 2017.

[40] Yichao Yan, Jinpeng Li, Shengcai Liao, Jie Qin, Bingbing Ni, Xiaokang Yang, and Ling Shao. Exploring visual context for weakly supervised person search, 2021.

[41] Yichao Yan, Jingpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, and Ling Shao. Anchor-free person search. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[42] Yichao Yan, Jinpeng Li, Jie Qin, Shengcai Liao, and Xiaokang Yang. Efficient person search: An anchor-free approach. *CoRR*, abs/2109.00211, 2021.

[43] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[44] Jinyu Yang, Jingjing Liu, Ning Xu, and Junzhou Huang. Tvt: Transferable vision transformer for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 520–530, 2023.

[45] Junjie Ye, Changhong Fu, Guangze Zheng, Danda Pani Paudel, and Guang Chen. Unsupervised domain adaptation for nighttime aerial tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8896–8905, 2022.

[46] Mang Ye, Jianbing Shen, Senior Member, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[47] Rui Yu, Dawei Du, Rodney LaLonde, Daniel Davila, Christopher Funk, Anthony Hoogs, and Brian Clipp. Cascade transformers for end-to-end person search. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[48] Jingyi Zhang, Jiaxing Huang, Zichen Tian, and Shijian Lu. Spectral unsupervised domain adaptation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9829–9840, 2022.

[49] Dingyuan Zheng, Jimin Xiao, Yunchao Wei, Qiufeng Wang, Kaizhu Huang, and Yao Zhao. Unsupervised domain adaptation in homogeneous distance space for person reidentification. *Pattern Recognition*, 132:108941, 2022.

[50] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild, 2017.

[51] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. Robust partial matching for person search in the wild. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[52] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2723–2738, 2020.

[53] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection, 2021.