applied sciences

MDPI

Article

Multi-Attention-Guided Cascading Network for End-to-End Person Search

Jianxi Yang and Xiaoyong Wang *

School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China
* Correspondence: wangxiaoyong@cqjtu.edu.cn

Abstract: The key procedure is to accurately identify pedestrians in complex scenes and effectively embed features from multiple vision cues. However, it is still a limitation to coordinate two tasks in the unified framework, thus leading to high computational overhead and unsatisfactory search performance. Furthermore, most methods do not take significant clues and key features of pedestrians into consideration. To remedy these issues, we introduce a novel method named Multi-Attention-Guided Cascading Network (MGCN) in this paper. Specifically, we obtain the trusted bounding box through the detection header as the label information for post-process. Based on the end-to-end network, we demonstrate the advantages of jointly learning to construct the bounding box and attention module by maximizing the complementary information from different attention modules, which can achieve optimized person search performance. Meanwhile, by imposing an aligning module on re-id feature extracted network to locate visual clues with semantic information, which can restrain redundant background information. Extensive experimental results for the two benchmark person search datasets are provided to demonstrate that the proposed MGCN markedly outperforms the state-of-the-art baselines.

Keywords: person search; multiple information; end-to-end network; attention module

Citation: Yang, J.; Wang, X. Multi-Attention-Guided Cascading Network for End-to-End Person Search. *Appl. Sci.* **2023**, *13*, 5576. https://doi.org/10.3390/app13095576

Academic Editor: Chilukuri K. Mohan

Received: 10 March 2023
Revised: 15 April 2023
Accepted: 28 April 2023
Published: 30 April 2023

1. Introduction

Person search is an active research area that attempts to locate and identify query people from real-world scenes, and has gained significant attention in recent years. To date, the person's information that has been collected from multiple monitors or sensors cry out for effective search tools [1–3]. Compared to the person re-identification (re-id) task, this process usually contain a two-stage process, including person detection and person re-id. Notably, as shown in Figure 1, the person search task is primarily focused on locating the target person within the context of the entire image. In particular, in the real scene, the captured images have problems such as changes in scale, resolution, and different views, which brings significant challenges for the person search task.

Considering the complexity of the person search task, it is crucial to guide the discriminant information learning for the re-id task based on the bounding box of the target person generated by the detection network. Thanks to the well-trained deep learning framework, many object-detection methods have been proposed to label the bounding boxes for multiple objects, which provides reliable label information for the re-id task. With the continuous advancements of deep learning technology, the introduction of Regions with Convolutional Neural Network features (R-CNN) [4] to extract more discriminative features for the object detection tasks. Generally, R-CNN generates about 2000 regional proposals for each image by the selective search strategy [5]. After that, numerous CNN-based methods have been proposed for completing object-detection tasks. Girshick et al. [6] introduced Fast R-CNN method, which jointly optimized the classification and the boundary box regression tasks. Ren et al. [7] proposed a Faster R-CNN network to generate region proposals through additional subnets, which has obtained more and more attention recently in the detection

task. Compared with object detection tasks, re-id tends to focus on local information from a person's image [8,9]. Faced with the re-id task, there are various approaches that have been proposed by effectively learning the discriminability information from multiple cropped people [10]. Since deep learning methods have demonstrated their unique ability in many challenging re-id tasks, researchers have proposed some efficient feature representation methods to explore deep features from personal images. Qian et al. [11] presented a multi-scale depth learning framework to capture people's feature representations at different scales. Li et al. [12] introduced a memory module to store the features from target samples, which can execute invariance constraints without increasing computation cost. Luo et al. [13] suggested a spatial transformer network to construct effective feature representation through local affine transformation. Even though the above re-id studies have obtained satisfactory performance, most of them utilize the cropped person images without considering the rich context information from the scene.

Figure 1. Person search: detection a target person from whole scene images.

To tackle this issue, person search is proposed to be exploited for search tasks within a given scene, which skillfully combines detection and re-identification tasks [14,15] in a unified framework. Typically, these methods are required to detect multiple persons of different scales in real-world images and re-identify them based on the detected labeled information. Zheng et al. [16] introduced a novel dataset for person search tasks, which contains videos collected by six cameras. This method proves that the detection task can be beneficial for improving the re-id performance of the target pedestrian. Chen et al. [17] segmented the foreground person from the original image to obtain each person's identification information. The re-identification algorithm is then employed to extract discriminative features from both the original image and the foreground person separately. Although the aforementioned two-step methods have yielded satisfactory results, they increased the calculation cost and time. Therefore, in practical applications, numerous joint frameworks have been suggested to solve the person search problem. Xiao et al. [18] introduced an online instance matching (OIM) loss function that integrates person detection and re-identification tasks into a single convolutional neural network. Furthermore, Munjal et al. [19] extended the OIM algorithm to take context information from queries and gallery images into consideration. It can be found that both one-step and two-step person search methods need reliable embedded feature representation to achieve accurate retrieval. Above all, person detection tends to extract global features in the scene, and then distinguish the different features between the background and instances, while re-id task needs to distinguish the differences between various similar persons. Therefore, an effective multi-attention end-to-end framework should be introduced for completing person search

tasks, which is competent to extract multiple noteworthy features based on the valuable information collected from multiple people.

To address the aforementioned challenges, we propose a novel Multi-Attention-Guided Cascading Network (MGCN) for end-to-end person search task, which integrates multiple attention information from the target person in a unified one-step network to achieve fine-grained person search tasks. Inspired by the Faster R-CNN network, MGCN adaptively labeled the trustworthy bounding box for each person in various scenes. Furthermore, we adopt a multi-attention-guided network to directly obtain feature embedding representations for the re-id task, which takes both channel attention and spatial attention into consideration to obtain fine-grained information from each person. Moreover, MGCN adopt aligning module to focus on region from the labeled person with more discriminative information to constrict interference of background information. Finally, in the test stage, we utilize bipartite graph processing to explore the similarity between various people in the search results to obtain optimal matching results. The entire MGCN process is illustrated in detail in Figure 2.

- "In this paper, we propose a novel method termed Multi-Attention-Guided Cascading Network (MGCN), which can integrate multiple discriminative information with coarse-to-fine features from people to be retrieved for end-to-end person search tasks. By this unified framework, the trusted bounding box and person discriminant information can be exploited effectively".
- "Moreover, we explore the attention maps of different clues to extract the person's semantic regions, which promotes the network attention to different salient information from pedestrians. In addition, in order to reduce the interference of background information in the scene, we design the alignment module in our end-to-end network".
- "Experiments on two challenging datasets confirm that MGCN markedly outperforms other person search methods and the proposed multi-attention module has significant advantages".

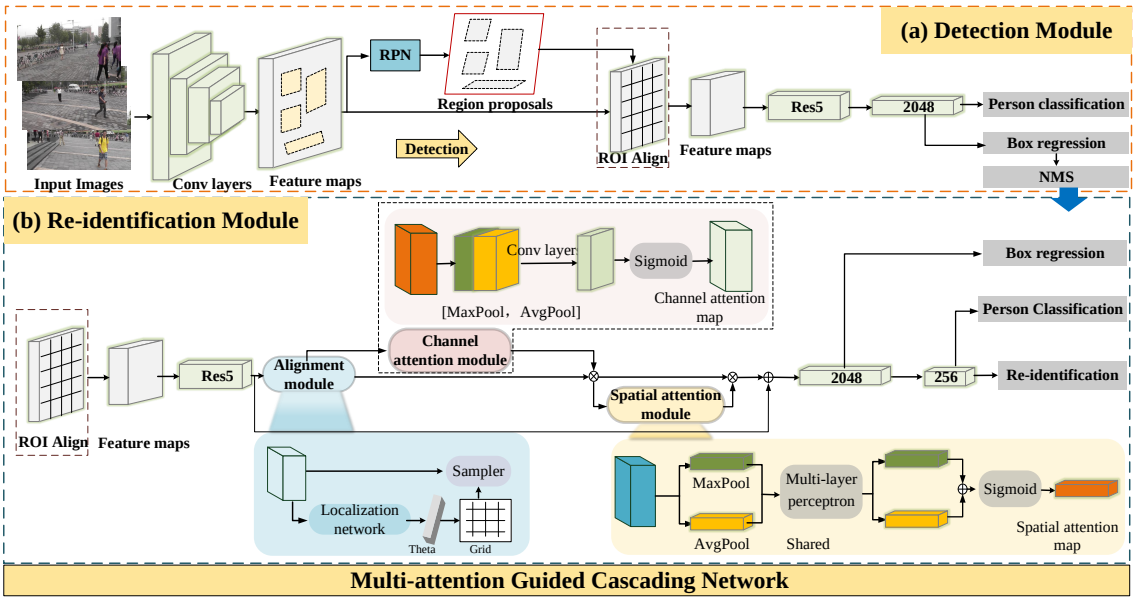

Figure 2. The flow chart of a Multi-Attention-Guided Cascading Network (MGCN), which integrates multi-attention discriminative information from a person in an end-to-end network for person search task. MGCN first mark multiple person by detection network, which provides a high-quality bounding box for each person to obtain label information. Then, MCGN utilize aligning module to locate advanced semantic information from the input image and avoid interference caused by background information. Finally, we introduce a multi-attention module to promote network focus on distinguishing fine-grained information between the multiple person for the re-id task.

We organized this paper as follows. In Section 1, we introduced the development process of person search and some typical algorithms. Meanwhile, Section 2 provided the related works for person re-id and search task. Section 3 introduces MGCN network and corresponding loss function in detail. To demonstrate the superiority of our approach, we conducted person search experiments in Section 4. A detailed conclusion of this paper is presented in Section 5.

2. Related Works

In recent years, numerous methods for person re-identification have been proposed to address various challenges in the field of computer vision. In this section, we summarize the recent algorithms related to person re-id and person search tasks across multiple fields.

2.1. Person Re-Identification

For person re-id tasks, pedestrian images captured by different cameras may have different views, resolutions, and illuminations, which carries significant challenges for identifying the same person under different cameras. Early research mainly employed handcrafted features for re-id tasks. In recent years, deep learning researches have been widely extended in various computer vision tasks with its superior feature representation ability. Wang et al. [20] presented a joint learning framework, which includes a shared sub-network and two specific sub-networks to extract features for matching a given single image and for classifying a given image pair. Luo et al. [21] proposed a neck structure that separates metric and classification loss into two different feature spaces. Ye et al. [22] designed a powerful baseline network, and achieved advanced performance on 12 re-id datasets. Some previous researches have made great progress, but for person re-identification tasks, many previous deep learning-based algorithms mostly extract global person information as input, and the learned features tend to be classified using global appearance.

It can be found that the engagement of person re-id task is to distinguish the differences between various instances, so it is necessary to focus on the local area of people to extract more refined discriminant features. Sun et al. [23] divided the feature representations into several strips of equal length to extract the local features from each person in different region. Furthermore, Wang et al. [24] suggested a multiple granularity framework that extracts global and local features, which partition the target images into several stripes and modifies the number of parts in different local branches to obtain multiple granularity feature representation. Yao et al. [25] counted the classification loss of part features from each detected person, and connected global and local features as the ultimate feature representation. In order to establish the correlation between the different parts, Bai et al. [26] applied long short-term memory to integrate contextual information to explore the complementary relationship between global and local features. Recently, Zhang et al. [27] proposed the part-guided graph convolution network with the aim of simultaneously learning the relationship between and within parts from feature representation and obtain superior performance.

Moreover, some researchers focus on exploring fine-grained semantic information from person images, thus making extracted person feature representations more robust and discriminative. Su et al. [28] introduced a pose-driven deep convolutional network, which adopts human body parts as clues and designs a sub-network to fuse features by pose-driven feature weighting. In order to take the structural information of the human body into consideration, Zhao et al. [29] captured the semantic features from different body parts and combined them with the competitive scheme to preserve the discriminative features. Li et al. [30] integrated attention selection and feature representation to improve person re-id performance in uncontrolled images. In terms of attention-feature fusion, Sun et al. [31] designed a multi-layer feature fusion framework to extract richer feature representation and proposed a novel loss function, which utilize eigenvalue difference orthogonality to reduce the correlation between features.

2.2. Person Search

Person search tasks aim to identify a specific pedestrian according to the given images or videos, which usually contain two steps with person detection and person re-identification task. Compared with person re-id, person search is a more challenging task, which requires detection and distinguishing people in complex environments. In particular, researchers constructed two large-scale datasets, CUHK-SYSU [18] and PRW [16] to verify the performance of the person search algorithms. Therefore, Han et al. [32] presented a framework for refining the localization of persons based on re-id, which can obtain more reliable bounding boxes and provide more discriminative feature embedding for downstream re-id tasks. Wang et al. [33] introduced a task-consistent two-stage framework for person search, which design an identity-guided detector module to produce query-like bounding boxes for the re-id stage. Chen et al. [34] suggested a novel re-id method with remodeling foreground person and original image patches, which can explore more representative features and improve model performance. Although the two-stage method has obtained great results, it is still unsatisfactory in terms of network computing cost and time.

In contrast, an end-to-end one-step person search framework combines detection and re-id task in a unified model, which focuses on constructing an effective network to obtain search result without extra steps. Chang et al. [35] trained relational context-aware agents to locate the target person from the scene image, which takes into account key information, such as local visual information, and temporal context. Recently, Yan et al. [36] proposed a Feature-Aligned Person Search Network (AlignPS), which designs a feature-aligned network to solve the misalignment issues and follows the person re-id priority principle to extract more discriminative feature embedding. Li et al. [37] took the relationship between detection and re-id into consideration to design a sequential network to progress learning each sub-tasks. In particular, they designed the context bipartite graphs matching algorithm to mine context information from different people. As the transformer network developed, some researchers have proposed the unified framework with a transformer module for person search task. Cao et al. [38] proposed a person search-specialized module with transformers for person detection and re-ID, which integrates detection encoder–decoder and re-id decoder to explore the relationship between different parts from the target person. Yu et al. [39] introduced the end-to-end cascade occluded attention transformer framework to solve the scale, perspective and occlusion problems, respectively. Compared to pedestrian detection, which aims to learn the global features of pedestrians, person re-identification requires more emphasis on the fine-grained details and unique characteristics of each individual. Therefore, we should take multiple fine-grained semantic information for each person into consideration in a one-step framework to mine more discriminative feature representation from complex scenes.

3. Proposed Method

In this paper, we propose a Multi-Attention-Guided Cascading Network (MGCN) for end-to-end person search, which can learn high-quality bounding boxes and fine-grained information from person to achieve discriminative feature embedding. As illustrated in Figure 2, to obtain multiple bounding boxes from the scene image, MGCN first trains the detection network to obtain trustworthy label information and a reliable regression bounding box and gives consideration to provide valuable information for the re-id network. Then, our work can extract high-level semantic information from a person in the re-id network, which can bypass the rigid hard partitioning step and, thus, attain flexible attentive regions in the complex scene. Finally, MGCN finished a person search task by adopting a context bipartite graph matching, which improves the search result effectively.

3.1. Trustworthy Bounding Box Regression

It can be found that the transformer-based feature representation model focuses on extracting the global discriminant representation from the person. However, for some occlusion or complex backgrounds of the image, it is difficult for the re-id model to learn

discriminative feature embedding from person search datasets. It can be found that simultaneously predicting a person's bounding boxes and re-id results usually produce poor results in a single search network. This is because the detection network fails to provide accurate candidate boxes for re-id tasks, which has an impact on the following feature embedding. To this end, we adopt a Faster R-CNN [6] network for the labeled person and generate a trustworthy bounding box in our proposed person search task.

A Faster R-CNN network is introduced in our paper, which sets Resnet-50 as the backbone network, which mainly includes five residual blocks, Region of Interest (ROI) pooling layer, and a region proposal net (RPN). Given the multiple people in complex scenes, we first adopt the first four residual blocks of the network to obtain the feature map of 1024 channels.

In the one-step person search method, the regression box may be inaccurate. Therefore, we added the box generation model regression in the detection part to obtain more reliable bounding boxes. Moreover, MGCN generates a set of pedestrian proposals through the RPN module, which is subsequently fed into an ROI-Align layer to extract discriminative features from candidate proposals. Then the network passes the fifth residual block in the ResNet50 network and is processed to generate 2048 dimensional feature representation. In our paper, the detection head is set after 2048 dimensional feature to obtain trustworthy bounding box regression, which judges whether the current characteristics are people and regression loss to generate a bounding box. To enhance the detection frame quality, we utilize the non-maximum suppression algorithm to decrease the number of candidate frames for repeated detection.

3.2. Multi-Attention-Guided Re-Id Network

The key point of the person search is to obtain an accurate regression box of the target person and extract the fine-grained features from pedestrians through a re-id network. Afterward, we employ a cascaded structure to learn person features from coarse to fine after detecting the frame. In order to obtain high-quality embedded features, MGCN introduces a fine-grained feature extraction module, which obtains the features of salient areas through an alignment module. Furthermore, we introduce a multi-attention module to extract different types of salient features. At each stage of the feature extraction, the features extracted by each module will improve the quality of person detail features from region extraction by the network to a certain extent.

3.2.1. Aligning Module

In this section, MGCN introduces the alignment module to extract the regional features based on the marked person, which can effectively avoid environmental noise and reduce the impact of irrelevant background information. Inspired by [40], the aligning module in our method consists of a localization network, grid generator, and sampler. This enables the module to capture more distinctive visual features from annotated pedestrians. They are employed to learn the positioning parameters, generate a grid of sampling points, and interpolate the feature map to fill in missing pixels caused by the transformation. It can be found that the transformer-based feature representation model focuses on extracting the semantic region feature representation from the labeled person (Figure 3).

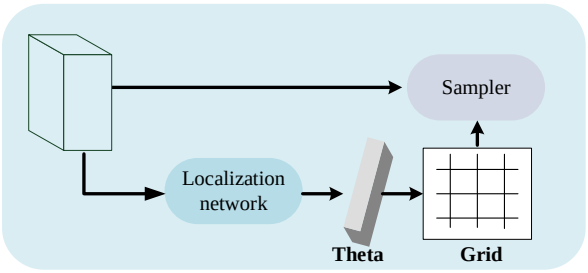

Figure 3. Aligning model.

First, the feature map G is extracted through the ROI-Align layer, where the space converter parameter is θ. To capture the mapping relationship between feature map G and output feature matrix $\bar{G}$, MGCN produces an image grid by transforming the parameters. Finally, the corresponding pixel value is filled with a bilinear sampler and the pixel value outside the original range is assigned to 0. This procedure can be expressed as

$$\bar{G}^c_{x,y} = \sum_{u^s}^{H} \sum_{v^s}^{W} G^c_{(u^s,v^s)} max(0, 1 - |u^t - x|) max(0, 1 - |v^t - y|) \quad (1)$$

where $G^c_{x,y}$ and $\bar{G}^c_{u^s,v^s}$ denote the input features at location (x, y) and output features at location (u^s, v^s) in channel c, respectively. Moreover, when the coordinate point (u^t, v^t) is in close proximity to the coordinate point (x, y) of the input image, aligning module adopts bilinear sampling to increase the pixel value at position (u^s, v^s). According to the output features $\bar{G}$ obtained by the aligning module, MGCN can accurately localize the target person and provide guidance for extracting subsequent salient features. Meanwhile, to emphasize the meaningful details of personnel when training the feature learning model, two attention modules are proposed to obtain identification features.

3.2.2. Multi-Attention-Guided Module for Person Search

According to the above analysis, person search network combines detection module and the semantic regions feature output by the aligning module, which focuses on the global features from the target person and the visual cues from the local region, respectively. In addition, it is essential to take into account the importance of significant discriminant information from each person to further improve the performance for the re-id stage. For the multi-attention module, MGCN divides the fine-grained feature extraction module into two parts to encourage re-id network focusing on salient features from different types for pedestrians (Figure 4).

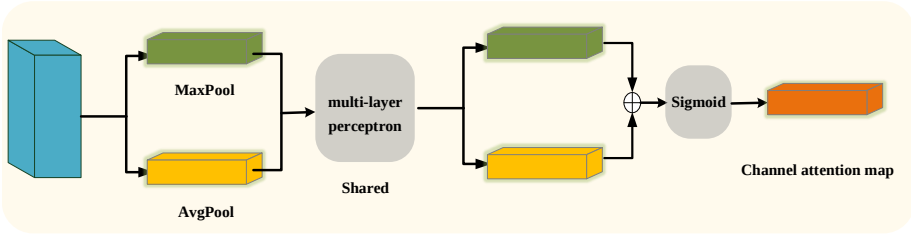

Figure 4. Channel attention model.

In order to generate the pedestrian's channel attention map, MGCN squeezes the spatial dimensions of the feature map by adopting the channel attention mechanism. Inspired by [41], we utilize the maximum pool and the average pool operations simultaneously to collect important clues about pedestrian features. Specifically, MGCN extracts two different spatial context descriptors through the two pooling operations mentioned above, which can be expressed as $\bar{G}^c_{max}$ and $\bar{G}^c_{avg}$. In order to generate a channel attention map $A_c \in \mathbb{R}^{C\times1\times1}$, we use a shared network consisting of a multi-layer perceptron (MLP) with one hidden layer to process the descriptors. Finally, MCGN applies an element-wise summation operation to learn the feature vector, which can be defined as

$$\begin{aligned} A_c(\bar{G}) &= sigmod(MLP(avgpool(\bar{G}))) + (MLP(maxpool(\bar{G}))) \\ &= sigmod(W_{c1}(W_{c2}(\bar{G}))) + (W_{c2}(W_{c1}(\bar{G}))) \end{aligned} \quad (2)$$

where *sigmod* is the activation function. In a shared network, the size of the hidden activation is to as $A_c \in \mathbb{R}^{C/r\times1\times1}$, where r presents the reduction ratio. $W_{c1} \in \mathbb{R}^{C/r\times C}$ and $W_{c2} \in \mathbb{R}^{C\times C/r}$ denote the different weights of MLP. For most attention-based feature extraction networks, one type of attention information is usually extracted in most situations, which has limited guidance for feature extraction. To address this issue, MGCN adopts

multiple attention modules to extract significant pedestrian features through two attention modules complementing each other (Figure 5).

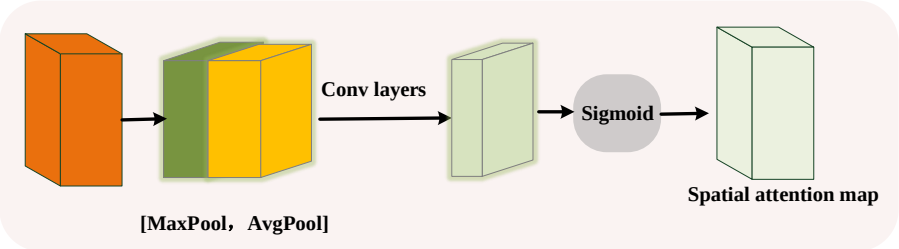

Figure 5. Spatial attention model.

Therefore, we construct a spatial attention map by utilizing the spatial relationship between multiple pedestrian features, which focuses on the location of significant information and complements the channel attention information. First, MGCN applies the maximum pooling and average pooling procedures along the direction of feature channels to generate two attention maps, which are, respectively, represented as $\bar{G}^s_{max}$ and $\bar{G}^s_{avg}$. The resulting features are passed through a standard convolution layer to generate the spatial attention map $A_s \in \mathbb{R}^{H\times W}$.

$$\begin{aligned} A_s(\bar{G}) &= sigmod(conv([(avgpool(\bar{G})); (maxpool(\bar{G}))])) \\ &= sigmod(conv([\bar{G}^s_{avg}; \bar{G}^s_{max}])) \end{aligned} \tag{3}$$

where $\bar{G}^s_{max} \in \mathbb{R}^{1\times W\times W}$ and $\bar{G}^s_{avg} \in \mathbb{R}^{C\times W\times H}$ denote two 2D maps. *conv* is a convolution operation. Above all, we obtain the channel attention representation $A_c \in \mathbb{R}^{C\times 1\times 1}$ and spatial attention representation $A_s \in \mathbb{R}^{1\times H\times W}$, respectively. Furthermore, the result of multi-attention module can be calculated as $\bar{G}_1 = A_c(\bar{G}) \otimes \bar{G}$, $\bar{G}_2 = A_s(\bar{G}) \otimes \bar{G}_1$, where $\otimes$ means element-wise multiplication. $\bar{G}_2$ denotes the attention map, which integrates information of multiple attention to obtain significant feature representation. It is worth noting that pixel-wise fusion operation ensures that the correlation map encodes the information from the key region.

Finally, for a person search network, the feature generates a feature map M through the multi-attention module, which is followed by three headers, including person classification, a box regression, and identity classification. Note that we do not provide a re-id discriminator in the detection part so that the network can detect all person in the scene on a large scale. In the re-id part, we introduce the alignment module and multi-attention module, which can effectively capture the pedestrian's features from coarse to fine, while reducing the impact of background interference.

3.3. Loss Function

In this section, we will describe the loss function used in our model. MGCN consists of two heads, which are utilized for model training in the detection and recognition stages, respectively.

For detection head, we employ two objective functions, the regression loss and the classification loss, to optimize the detection module.

$$L_{dec} = L_{c1} + \lambda L_{r1} \tag{4}$$

where λ denotes the balance parameter, which is set to 10. L_{c1} = and L_{r1} represent regression loss and classification loss, respectively.

$$L_{c1} = -\frac{1}{N}\sum_{i=1}^{N} y_i log(p_i) \tag{5}$$

where y_i denotes the ground truth label of the i-th feature and p_i represents the predicted probability that the feature belongs to the i-th class.

$$L_{r1} = -\frac{1}{N_p} \sum_{i=1^{N_p}} L_{l_1}(r_i, \delta_i) \tag{6}$$

where N_p presents the number of positive samples and L_{l_1} denotes the Smooth-l_1 loss. The margin parameter r_i controls the separation between positive and negative features in the embedding space. r_i and δ_i represent the calculated regressor and ground truth regressor, of ith positive sample, respectively. The detection head is trained at 0.5 IoU threshold, which aims to differentiate between positive and negative samples.

For the second head, we adopt three objective functions to optimize the end-to-end person search model, the regression loss, the classification loss, and the identity classification loss.

$$L_{reid} = L_{r2} + L_{nae} \tag{7}$$

where L_{r2} represents classification loss, which is same as L_{r1}. Follow [42], L_{nae} loss contains classification loss and identity classification loss [18]. L_{c1} aims to ensure the features extracted by the re-identification network have both high intra-class consistency and inter-class separability.

Therefore, taking into account the aforementioned factors, the objective loss function L_{total} of MGCN can be expressed as

$$L_{total} = L_{dec} + L_{reid} \tag{8}$$

Moreover, we summarize the algorithm procedure of MGCN in Algorithm 1.

Algorithm 1 Algorithm of MGCN

Require: Person search dataset;
Ensure: Most similar person in each gallery image;
1: **while** not converge **do**
2: **for** $epoch = 1 : 20; epoch++$ **do**
3: By trustworthy bounding box regression generates a set of feature map G;
4: Construct the aligning module as in Equation (1) obtained feature $\bar{G}$;
5: Feed $\bar{G}$ into the channel attention model $A_c(\bar{G})$ according to Equation (2);
6: Feed feature map into the spatial attention model $A_s(\bar{G})$ according to Equation (3);
7: Learn the feature map M of each person by combining above multi-attention-guided module;
8: **end for**
9: **end while**

4. Experiment

In this section, all the experiments are conducted on two publicly available person search datasets CUHK-SYSU [18] and PRW [16]. We conduct a series of comprehensive experiments from various perspectives to validate the effectiveness of MGCN, beyond simply comparing it with state-of-the-art person search methods. In addition, we provide the ablation study to verify the validity of each module.

4.1. Datasets and Evaluation Protocols

In this section, we introduced two standard person search datasets.

CUHK-SYSU [18] is collected from two data sources, including real street snaps and pictures collected in movies. The images collected from real streets include changes in viewpoints, lighting, occlusion, and backgrounds. The dataset consists of 18,184 images featuring 8432 unique identities, and a total of 96,143 bounding boxes with annotations are provided. In our experiment, the training model utilizes 11,206 images with 5523 different

identities, and the testing model adopts 6978 images with 2900 query persons. For the query person, we utilize a gallery with a size of 100 to evaluate the search performance.

PRW [16] is collected from six different cameras located on the campus of the university, which contains 11,816 video frames. The training set in the dataset contains 5134 frames with 482 unique identities, and the testing set contains 6112 frames and 450 unique identities. Each identity in the video has identity information and bounding box information. In the experiment, the query database includes 6112 images with 2057 query persons.

4.1.1. Evaluation Metrics

Following the previous works, we adopt mAP [43] and top-1 accuracy as evaluation metrics to measure the performance of our proposed MGCN, which is commonly used for person search tasks [44,45].

4.1.2. Implementation Details

Our experiment implemented the model on the PyTorch platform [46]. The backbone network of MCGN is ResNet-50 [47], and ImageNet [48] is utilized for the pre-training to acquire the initial weight of the model. In our experiment, the input images are resized as 900×1500. We set the batchsize to 5. Technically, we train the person search network to utilize an SGD optimizer with an initial learning rate of 0.003 and a decrease to 10 in 16 epochs, for 20 epochs in total. In addition, we apply a weight decay of 0.0005 and a momentum of 0.9 in the SGD optimizer. The OIM loss settings are the same as in the previous work [19]. For the detection and re-identification heads, we set the NMS threshold to 0.4 and 0.5, respectively, which are applied to remove redundant detection frames to obtain trusted boxes. In addition, to improve the experimental results, we introduce CBGM [37] method in the test part to re-match the target person.

4.2. Experiment Results on Two Datasets

4.2.1. Evaluation on CUHK-SYSU

We conducted a comparative study between MGCN and various state-of-the-art methods designed to solve the person search problem on the CUHK-SYSU and PRW datasets, which includes DPM [16], MGTS [17], CLSA [49], RDLR [32], IGPN [50], TCTS [33], OIM [18], IAN [51], NPSM [52], RCAA [35], CTXG [53], QEEPS [19], HOIM [54], APNet [55], BINet [56], NAE [42], NAE+ [42] DMRNet [57], PGS [58], AlignPS [36], AlignPS+ [36], and SeqNet [37] methods.

Tables 1 and 2 illustrate the experimental results of our proposed MGCN with a gallery size of 100. We compare MGCN with two-step person search methods and one-step algorithms in Tables 1 and 2, respectively. The experimental results show that our model achieved high performance on CUHK-SYSU dataset with mAP and top-1 accuracy of 94.28% and 94.97%, respectively. Compared with the two-stage person search methods in Table 1, we can observe that, with TCTS, MGCN achieves a significant improvement of 0.48% and 0.3%, respectively in mAP and top-1. Among them, the TCTS model adopts random erasure, label smoothing, and triple loss, but still does not show satisfactory performance compared with MCGN. The starting point of this paper is not just to propose a person search method with the best performance. Compared with TCTS, we combine the detection and re-id tasks in a unified framework to eliminate the difficulty of utilizing two-step progress with different tasks. In addition, MGCN aims to collaborate between multiple spaces in a unified framework to learn a consistency graph that is able to uncover the essential structure of multi-view data. Compared to NAE, MGCN shows significant improvement rates of 2.78% and 2.57% for mAP and top-1, respectively. We can see that MCGN seems more robust than network-based PGS, because it takes the fine-grained features in the re-id process into consideration and considers obtaining more discriminative feature embedding. Compared with the end-to-end method, our method is higher than AlignPS by 2.69%, where AlignPS integrates multi-level information and our method focuses more on salient features during the re-identification stage. In addition, MGCN aims to collaborate multiple tasks in a

unified framework to learn the discriminant pedestrian feature representation that is able to uncover the essential semantic detail features of person search dataset. The results indicate the effectiveness of our multi-attention-guided cascade network.

Table 1. Experimental results on CUHK-SYSU and PRW. The mAP (%) and top-1 are listed.

Method	CUHK-SYSU		PRW	
	mAP	**top-1**	**mAP**	**top-1**
DPM [16]	-	-	20.50	48.30
MGTS [17]	83.00	83.70	32.60	72.10
CLSA [49]	87.20	88.5	38.7	65.00
RDLR [32]	93.00	94.20	42.90	70.20
IGPN [50]	90.3	91.40	47.20	87.00
TCTS [33]	93.90	95.10	46.80	87.50
MGCN	**94.28**	**94.97**	**47.11**	**86.39**

Table 2. Experimental results on CUHK-SYSU and PRW. The mAP (%) and top-1 are listed.

Method	CUHK-SYSU		PRW	
	mAP	**top-1**	**mAP**	**top-1**
OIM [18]	75.50	78.70	21.30	49.90
IAN [51]	76.30	80.10	23.00	61.90
NPSM [52]	77.90	81.20	24.20	53.10
RCAA [35]	79.30	81.30	-	
CTXG [53]	84.10	86.50	33.40	73.60
QEEPS [19]	88.90	89.10	37.10	76.70
HOIM [54]	89.70	90.80	39.80	80.40
APNet [55]	88.90	89.30	41.90	81.40
BINet [56]	90.00	90.70	45.30	81.70
NAE [42]	91.50	92.40	43.30	80.90
NAE+ [42]	92.10	92.90	44.00	81.10
DMRNet [57]	93.20	94.20	46.90	83.30
PGS [58]	92.30	94.70	44.20	85.20
AlignPS [36]	93.10	93.40	45.90	81.90
AlignPS+ [36]	94.00	94.50	46.10	82.10
SeqNet [37]	93.80	94.60	46.70	83.40
MGCN	**94.28**	**94.97**	**47.11**	**86.39**

4.2.2. Evaluation on PRW

The comparison of results on the PRW dataset is presented in Tables 1 and 2. Notably, the PRW dataset has a smaller training set compared to other datasets, while the size of the gallery is larger. MGCN achieves better performance than other methods in most cases. The results show that the proposed model achieves high mAP and top-1 accuracy of 47.11% and 86.39%, respectively, on the PRW dataset. Since PRW dataset has fewer training samples, most of the deep learning methods are prone to over-fitting problems in the model training process. Compared with end-to-end method, two-stage algorithms relies more on previous step label information to train the network model, so processing person data in complex scenes is still not robust. This also demonstrates that joint deal with person detection and re-id task in the unified framework can improve the search performance,

which can reduce the error caused by sub-optimal solution. For example, compared to AlignPS+, MGCN shows significant improvements of 1.01% and 4.29% in mAP and top-1, respectively. Furthermore, compared with baseline NAE, our method is higher than NAE by 5.49% in terms of top-1. We attribute this to our multi-attention structure which generates more fine-grained re-id features, especially when there is a complex setting with background information and personnel details are not obvious. MGCN can obtain a reliable regression bounding box in the detection stage and mine more discriminative pedestrian characteristics from the original scene. Moreover, MGCN integrates multiple attention information in the re-id feature extraction module, which can focus on simultaneously channel and spatial significant information to improve search performance.

4.3. Ablation Study

In this section, we investigate the effectiveness of various components incorporated in MGCN. The experimental results on CUHK-SYSU and PRW datasets are shown in Table 3. We tested the impact of the alignment module and multiple attention modules on our algorithm. Furthermore, we have check-marked the corresponding modules in Table 3. "Aligning Module" denotes that the aligning module is incorporated into the re-id module of the network, enhancing the network's ability to capture the pedestrian details while avoiding background interference. "Channel Attention" denotes the re-id feature extracted network with channel module, which acquires the significant channel information in the local region of the person. Moreover, the "Spatial Attention" refers to a branch in the network that utilizes a spatial attention module to capture fine-grained part information from the local regions of the person images. "Aligning Module and Channel Attention" represents that the spatial attention and aligning module are added on the basis of the re-id network, which integrates the semantic regions-based features and channel features to obtain better results for person search task. "Channel Attention and Spatial Attention" means the our proposed model is trained by double attention modules. "Aligning, Channel Attention and Spatial Attention Module" means person search network, which includes fine-grained kinds of features extracted by three modules.

Table 3. Ablation experimental results on CUHK-SYSU and PRW. The mAP (%) and top-1 are listed.

Aligning Module	Channel Attention	Spatial Attention	CUHK-SYSU		PRW	
			mAP	top-1	mAP	top-1
✓			94.24	95.03	46.83	87.07
✓	✓		93.97	94.38	45.51	82.84
✓		✓	94.23	95.00	44.06	86.49
	✓	✓	94.01	94.62	44.12	81.67
✓	✓	✓	94.28	94.97	47.11	86.39

It is clear from Table 3 that the performance of the baseline model is inferior to that of all other models on all datasets. Specifically, the baseline model achieves the lowest mAP on the PRW dataset, while our proposed MGCN model leads to significant improvement in mAP from 46.83% to 47.11%. In addition, we observed similar improvements in the CUHK-SYSU dataset. Compared with the attention module, adding an alignment module in the network can significantly improve the experimental results. This demonstrates that for people with complex backgrounds, the module can effectively suppress redundant information to obtain significant semantic information about individuals.These findings confirm the effectiveness of MGCN through evaluating the performance of its different components. However, when combining the "Channel attention" and "Spatial attention" modules, the network does not exhibit significant improvement, indicating that solely capturing the details of the person without considering the current scenario information is limited.

Visualization

We visualize the search results in challenging datasets to verify the effectiveness of MGCN. On the left side are the query images, and on the right side are the gallery images showing pedestrians with correct matching. Each green bounding box is marked with the probability of query, while the highest is the correct pedestrian. In Figure 6, our method shows the ability to correctly identify the same person among multiple interfering factors, which demonstrates the robustness and high discrimination of the features extracted by MGCN.

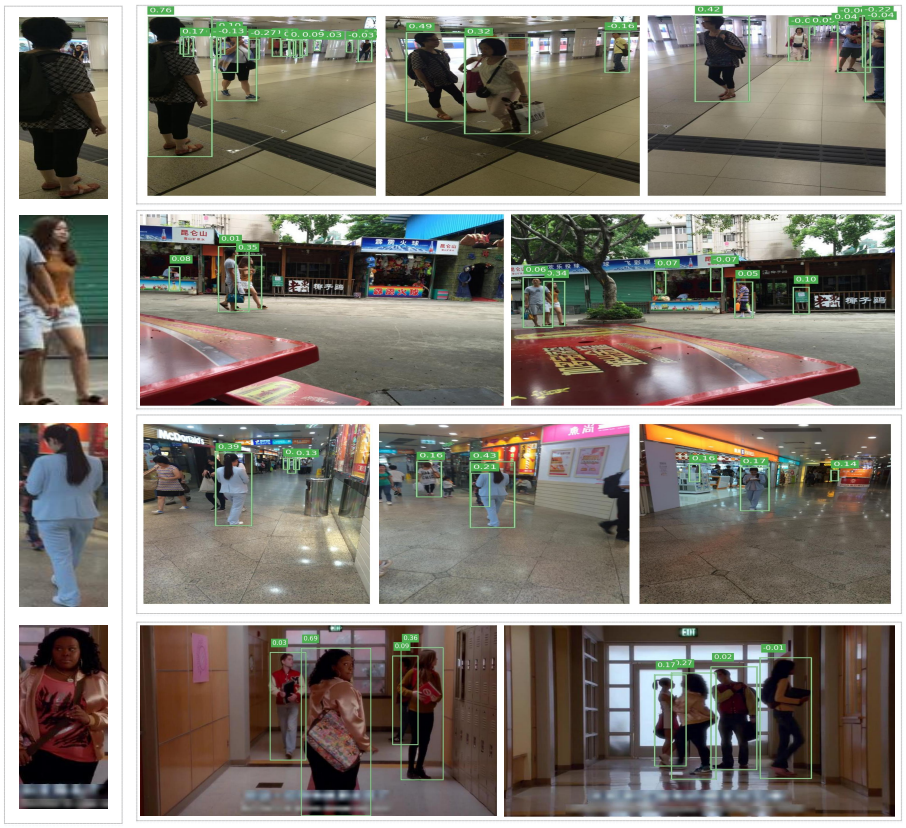

Figure 6. Visualization results of person search. The left side represents pedestrians to be detected and the green bounding boxes represent search results with different probabilities.

5. Conclusions

In this paper, we proposed a novel Multi-Attention-Guided Cascading Network (MGCN), which integrates multiple attention information to achieve end-to-end person search. Our main research contribution is that MCGC takes detection and re-id tasks into consideration to capture people in complex scenes from coarse-to-fine in a framework. Moreover, we introduce multi-attention-guided module to learn the discriminatory feature that is able to uncover the significant region of the person. In order to accurately locate person semantic information, the alignment module is introduced to avoid complex background information and redundant noise to provide accurate visual clues. Facing the optimization of the person search, the method of bipartite graph is adopted to obtain better sorting results. Various experiments show that take the reliability of the detection frame and the fine-grained features extracted from the re-id module into consideration are extraordinarily necessary for person search.

Notably, instead of utilizing the two-step person search scheme, the proposal combine two tasks to extract multiple noteworthy features in a unified framework, such that the error caused by two-step processing can be significantly reduced. In addition, our method still has limitations, as attention mechanisms typically capture local features from objective

images and ignore learning discriminative features from a global perspective. In the future research, we would work on utilizing transformer learning to capture the global relevant information from images to further improve the search performance.

Author Contributions: Conceptualization, J.Y. and X.W.; methodology, X.W.; software, X.W.; validation, J.Y. and X.W.; formal analysis, X.W.; investigation, X.W.; resources, J.Y.; data curation, X.W.;writing—original draft preparation, X.W.; writing—review and editing, J.Y. and X.W.; visualization, J.Y.; supervision, J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Science and Technology Research Program of Chongqing Municipal Education Commission of China (KJZD-M202000702).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets are publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, H.; Wang, Y.; Zhang, Z.; Fu, X.; Zhuo, L.; Xu, M.; Wang, M. Kernelized multiview subspace analysis by self-weighted learning. *IEEE Trans. Multimed.* **2020**, *23*, 3828–3840. [CrossRef]
2. Wang, H.; Yao, M.; Jiang, G.; Mi, Z.; Fu, X. Graph-Collaborated Auto-Encoder Hashing for Multiview Binary Clustering. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 1–13. [CrossRef] [PubMed]
3. Qian, B.; Wang, Y.; Yin, H.; Hong, R.; Wang, M. Switchable Online Knowledge Distillation. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
4. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
5. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
6. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
7. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015.
8. Wang, H.; Peng, J.; Chen, D.; Jiang, G.; Zhao, T.; Fu, X. Attribute-guided feature learning network for vehicle reidentification. *IEEE Multimed.* **2020**, *27*, 112–121. [CrossRef]
9. Wang, H.; Jiang, G.; Peng, J.; Deng, R.; Fu, X. Towards Adaptive Consensus Graph: Multi-view Clustering via Graph Collaboration. *IEEE Trans. Multimed.* **2022**, 1–13. [CrossRef]
10. Wang, H.; Peng, J.; Zhao, Y.; Fu, X. Multi-path deep cnns for fine-grained car recognition. *IEEE Trans. Veh. Technol.* **2020**, *69*, 10484–10493. [CrossRef]
11. Qian, X.; Fu, Y.; Jiang, Y.G.; Xiang, T.; Xue, X. Multi-scale deep learning architectures for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5399–5408.
12. Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; Yang, Y. Invariance matters: Exemplar memory for domain adaptive person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 598–607.
13. Luo, H.; Jiang, W.; Fan, X.; Zhang, C. Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. *IEEE Trans. Multimed.* **2020**, *22*, 2905–2913. [CrossRef]
14. Wang, H.; Peng, J.; Jiang, G.; Xu, F.; Fu, X. Discriminative feature and dictionary learning with part-aware model for vehicle re-identification. *Neurocomputing* **2021**, *438*, 55–62. [CrossRef]
15. Peng, J.; Jiang, G.; Wang, H. Adaptive Memorization with Group Labels for Unsupervised Person Re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *early access*. [CrossRef]
16. Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; Tian, Q. Person re-identification in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1367–1376.
17. Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; Tai, Y. Person search via a mask-guided two-stream cnn model. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
18. Xiao, T.; Li, S.; Wang, B.; Lin, L.; Wang, X. Joint detection and identification feature learning for person search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3415–3424.

19. Munjal, B.; Amin, S.; Tombari, F.; Galasso, F. Query-guided end-to-end person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 811–820.
20. Wang, F.; Zuo, W.; Lin, L.; Zhang, D.; Zhang, L. Joint learning of single-image and cross-image representations for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26–30 June 2016; pp. 1288–1296.
21. Luo, H.; Jiang, W.; Gu, Y.; Liu, F.; Liao, X.; Lai, S.; Gu, J. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Trans. Multimed.* **2019**, *22*, 2597–2609. [CrossRef]
22. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 2872–2893. [CrossRef]
23. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 480–496.
24. Wang, G.; Yuan, Y.; Chen, X.; Li, J.; Zhou, X. Learning discriminative features with multiple granularities for person re-identification. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 274–282.
25. Yao, H.; Zhang, S.; Hong, R.; Zhang, Y.; Xu, C.; Tian, Q. Deep representation learning with part loss for person re-identification. *IEEE Trans. Image Process.* **2019**, *28*, 2860–2871. [CrossRef] [PubMed]
26. Bai, X.; Yang, M.; Huang, T.; Dou, Z.; Yu, R.; Xu, Y. Deep-person: Learning discriminative deep features for person re-identification. *Pattern Recognit.* **2020**, *98*, 107036. [CrossRef]
27. Zhang, Z.; Zhang, H.; Liu, S.; Xie, Y.; Durrani, T.S. Part-guided graph convolution networks for person re-identification. *Pattern Recognit.* **2021**, *120*, 108155. [CrossRef]
28. Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; Tian, Q. Pose-driven deep convolutional model for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3960–3969.
29. Zhao, H.; Tian, M.; Sun, S.; Shao, J.; Yan, J.; Yi, S.; Wang, X.; Tang, X. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1077–1085.
30. Li, W.; Zhu, X.; Gong, S. Harmonious attention network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2285–2294.
31. Sun, J.; Li, Y.; Chen, H.; Zhang, B.; Zhu, J. Memf: Multi-level-attention embedding and multi-layer-feature fusion model for person re-identification. *Pattern Recognit.* **2021**, *116*, 107937. [CrossRef]
32. Han, C.; Ye, J.; Zhong, Y.; Tan, X.; Zhang, C.; Gao, C.; Sang, N. Re-id driven localization refinement for person search. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9814–9823.
33. Wang, C.; Ma, B.; Chang, H.; Shan, S.; Chen, X. Tcts: A task-consistent two-stage framework for person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11952–11961.
34. Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; Tai, Y. Person search by separated modeling and a mask-guided two-stream cnn model. *IEEE Trans. Image Process.* **2020**, *29*, 4669–4682. [CrossRef]
35. Chang, X.; Huang, P.Y.; Shen, Y.D.; Liang, X.; Yang, Y.; Hauptmann, A.G. Rcaa: Relational context-aware agents for person search. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 84–100.
36. Yan, Y.; Li, J.; Qin, J.; Bai, S.; Liao, S.; Liu, L.; Zhu, F.; Shao, L. Anchor-free person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7690–7699.
37. Li, Z.; Miao, D. Sequential end-to-end network for efficient person search. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 2011–2019.
38. Cao, J.; Pang, Y.; Anwer, R.M.; Cholakkal, H.; Xie, J.; Shah, M.; Khan, F.S. PSTR: End-to-End One-Step Person Search With Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9458–9467.
39. Yu, R.; Du, D.; LaLonde, R.; Davila, D.; Funk, C.; Hoogs, A.; Clipp, B. Cascade Transformers for End-to-End Person Search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7267–7276.
40. Yang, Q.; Yu, H.X.; Wu, A.; Zheng, W.S. Patch-based discriminative feature learning for unsupervised person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 3633–3642.
41. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
42. Chen, D.; Zhang, S.; Yang, J.; Schiele, B. Norm-Aware Embedding for Efficient Person Search and Tracking. *Int. J. Comput. Vis.* **2021**, *129*, 3154–3168. [CrossRef]
43. Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; Yang, Y. Improving person re-identification by attribute and identity learning. *Pattern Recognit.* **2019**, *95*, 151–161. [CrossRef]

44. Wang, Y. Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion. *ACM Trans. Multimed. Comput. Commun. Appl.* **2021**, *17*, 1–25.
45. Wang, Y.; Peng, J.; Wang, H.; Wang, M. Progressive learning with multi-scale attention network for cross-domain vehicle re-identification. *Sci. China Inf. Sci.* **2022**, *65*, 160103. [CrossRef]
46. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in Pytorch. 2017. Available online: https://pytorch.org/tutorials/beginner/basics/autogradqs_tutorial.html (accessed on 9 March 2023).
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
48. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
49. Lan, X.; Zhu, X.; Gong, S. Person search by multi-scale matching. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 536–552.
50. Dong, W.; Zhang, Z.; Song, C.; Tan, T. Instance guided proposal network for person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2585–2594.
51. Xiao, J.; Xie, Y.; Tillo, T.; Huang, K.; Wei, Y.; Feng, J. IAN: The individual aggregation network for person search. *Pattern Recognit.* **2019**, *87*, 332–340. [CrossRef]
52. Liu, H.; Feng, J.; Jie, Z.; Jayashree, K.; Zhao, B.; Qi, M.; Jiang, J.; Yan, S. Neural person search machines. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 27–29 October 2017; pp. 493–501.
53. Yan, Y.; Zhang, Q.; Ni, B.; Zhang, W.; Xu, M.; Yang, X. Learning context graph for person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2158–2167.
54. Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; Schiele, B. Hierarchical online instance matching for person search. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10518–10525.
55. Zhong, Y.; Wang, X.; Zhang, S. Robust partial matching for person search in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6827–6835.
56. Dong, W.; Zhang, Z.; Song, C.; Tan, T. Bi-directional interaction network for person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2839–2848.
57. Han, C.; Zheng, Z.; Gao, C.; Sang, N.; Yang, Y. Decoupled and memory-reinforced networks: Towards effective feature learning for one-step person search. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 1505–1512.
58. Kim, H.; Joung, S.; Kim, I.J.; Sohn, K. Prototype-guided saliency feature learning for person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4865–4874.