Person Search Challenges and Solutions: A Survey

Xiangtan Lin[1], **Pengzhen Ren**[2], **Yun Xiao**[2], **Xiaojun Chang**[1*], **Alex Hauptmann**[3]

[1]Monash University
[2]Northwest University
[3]Carnegie Mellon University

{xiangtan.lin, cxj273, alex.hauptmann}@gmail.com, pzhren@foxmail.com, yxiao@nwu.edu.cn

Abstract

Person search has drawn increasing attention due to its real-world applications and research significance. Person search aims to find a probe person in a gallery of scene images with a wide range of applications, such as criminals search, multi-camera tracking, missing person search, etc. Early person search works focused on image-based person search, which uses person image as the search query. Text-based person search is another major person search category that uses free-form natural language as the search query. Person search is challenging, and corresponding solutions are diverse and complex. Therefore, systematic surveys on this topic are essential. This paper surveyed the recent works on image-based and text-based person search from the perspective of challenges and solutions. Specifically, we provide a brief analysis of highly influential person search methods considering the three significant challenges: the discriminative person features, the query-person gap, and the detection-identification inconsistency. We summarise and compare evaluation results. Finally, we discuss open issues and some promising future research directions.

1 Introduction

Person search [Xu *et al.*, 2014] aims to find a query person in a gallery of scene images. Historically, person search was an extended form of person re-identification (re-id) problem [Liu *et al.*, 2020b; Liu *et al.*, 2020a; Li *et al.*, 2019b; Cheng *et al.*, 2018; Liu *et al.*, 2018b; Cheng *et al.*, 2017; Liu *et al.*, 2017b]. Therefore, early researches on person search focused on an image-based setting, which uses person image as the search query [Xiao *et al.*, 2017; Liu *et al.*, 2017a; Chang *et al.*, 2018; Gao *et al.*, 2019; Xiao *et al.*, 2019]. Meanwhile, research in text-based person search [Li *et al.*, 2017b; Wang *et al.*, 2020b] has made significant advances in the past few years. Text-based person search is handy when a probe image is unavailable but free-form natural language. The two types of person search are illustrated in Figure 1.

*Corresponding Author

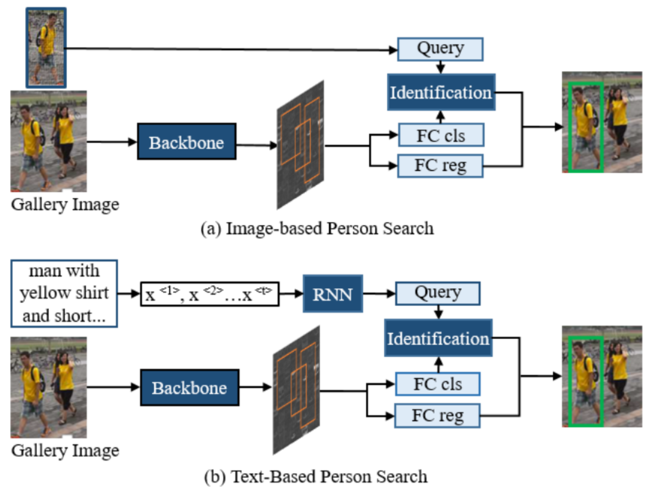

Figure 1: The general frameworks of person search. (a) Image-based person search in which person image is available as the search query against a gallery of images. Image-based person search involves two sub-tasks, person detection and person identification. (b) Text-based person search in which search query is free form natural language. A general text-based person search framework typically learns text feature through an RNN variant network and then align text features with visual elements from the detection network to identify the person in the target images.

Person search faces more challenges than person re-id problem. Unlike the person re-id setting where the cropped person images are provided, and the primary challenge is just to bring the query-person gap. Person search needs to deal with an additional detection challenge so that the detected person can be used for the downstream identification task. The additional detection task poses more challenges due to the influences of poses, occlusion, resolution and background clutter in the scene images. Such detection results may be inconsistent with the identification task (Figure 2). Similarly, text-based person search is also more challenging than the traditional text-image matching problem [Li *et al.*, 2017b] as it needs to learn discriminative features first before the text-person matching.

Person search is fast-evolving, and existing person search methods are diverse and complex. Researchers may leverage the rich knowledge concerning object detection, person re-id, and text-image matching separately. Systematic surveys

Survey	Covering	Analysis
[Islam, 2020]	Image-based	Components
Ours	Image-based, Text-based	(Challenges: **Solutions**) Discriminative person features: **Deep feature representation learning** Query-person gap: **Deep metric learning** Detection-identification inconsistency: **Identity-driven detection**

Table 1: Summary of the main differences between the previous survey and ours. This survey focuses more on challenges and solutions.

concerning person search bring more values to the community. Especially, as far as we know, there is no existing survey covering the text-based person search. [Islam, 2020] surveyed works on image-based person search and neglected the text-based person search. Furthermore, [Islam, 2020] didn't discuss the joint challenge of person detection and identification, especially the detection-identification inconsistency challenge as illustrated in Figure 2. Therefore, we survey works beyond image-based person search and provide a systematic review of the diverse person search solutions. We summarise the main differences between the previous survey [Islam, 2020] and ours in Table 1.

In this survey, we aim to provide a cohesive analysis of the recent person search works so that the rationals behind the ideas can be grasped to inspire new ideas. Specifically, We surveyed recently published and pre-print person search papers from top conference venues and journals. We analyse methods from the perspective of challenges and solutions and summarise evaluation results accordingly. At the end of the paper, we provide insights on promising future research directions. In summary, the main contributions of this survey are:

- In addition to image-based person search, we cover text-based person search which was neglected in the previous person search survey.
- We analyse person search methods from the perspective of challenges and solutions to inspire new ideas.
- We summarise and analyse existing methods' performance and provide insights on promising future research directions.

2 Person Search

Person search is a fast-evolving research topic. In 2014, [Xu *et al.*, 2014] first introduced the person search problem and pointed out the conflicting nature between person detection and person identification sub-tasks. Person detection deals with common human appearance, while the identification task focuses on a person's uniqueness. After [Xiao *et al.*, 2017] introduced the first end-to-end person search framework in 2017, we have seen an increasing number of image-based person search works in the last three years. Meanwhile, in 2017, GNA-RNN [Li *et al.*, 2017b] set the benchmark for text-based person search. We draw a timeline to present the person search works in Figure 3 and show the two divisions: image-based and text-based person search.

Figure 2: The detection-identification inconsistency problem. The detection model learns person proposal based on common person appearance using intersection-over-union (IoU) over certain threshold, which may result in less accurate bounding boxes compare to person for the identification task.

Person search addresses person detection and person identification simultaneously. There are three significant person search challenges to be considered when developing a person search solution. Firstly, a person search model needs to learn discriminative person features from scene images suitable for matching the query identity. Inevitably, the learnt person features differ from the query identity features to some degrees. Therefore the second major challenge is how to bring the gap between the query and the detected person. The third challenge is related to the conflicting nature between person detection and person identification. Person detection deals with common person appearance, while the identification task focuses on a person's uniqueness. The detected person may not be suitable for identity matching. For instance, a partial human body could be considered a person during detection and is inconsistent with the query identity at the identification stage, which may be a full person picture.

In this section, we analyse person search methods regarding above-mentioned three challenges and corresponding solutions from the following three aspects for both image-based and text-based person search:

- *Deep feature representation learning.* Addressing the challenge of learning discriminative person features from gallery images concerning background clutter, occlusion and poses etc.
- *Deep metric learning.* Addressing the challenge of bringing query-person gap by using loss functions to guide feature representation learning.
- *Identity-driven detection.* Addressing the challenge of mitigating the detection-identification inconsistency by incorporating query identities into the detection process.

2.1 Deep Feature Representation Learning

Deep feature representation learning focuses on learning discriminative person features concerning distractors in the gallery images. The majority of the early methods exploited global person features, including context cues, while refining person proposals. Such as RCAA [Chang *et al.*, 2018] utilises the relational spatial and temporal context in a deep reinforcement learning framework to adjust the bounding boxes constantly. However, these methods didn't consider the background clutter in the proposal bounding boxes, resulting in a situation where different persons with similar backgrounds are close in the learnt feature space. SMG [Zheng

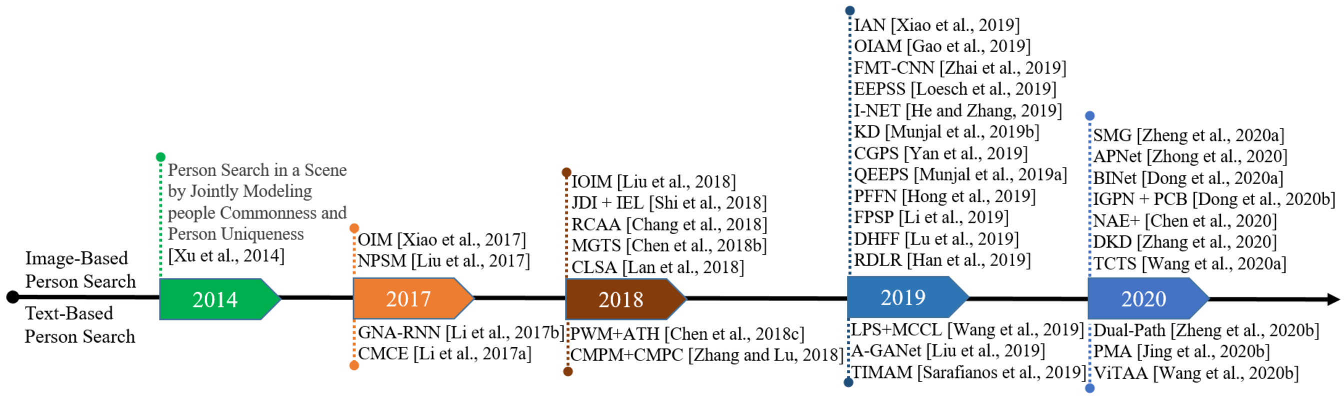

Figure 3: Timeline of person search studies. Above the timeline are image-based person search works. Below the line are text-based person search methods.

et al., 2020a] eliminates background clutter using segmentation masks so that the learnt person features are invariant to the background clutter. NAE [Chen *et al.*, 2020] separates persons and background by norms and discriminates person identities by angles. Person detection and object detection, in general, face the multi-scale matching challenge. To learn scale-invariant features, CLSA [Lan *et al.*, 2018] and DHFF [Lu *et al.*, 2019] utilise multi-level features from the identification network to solve the multi-scale matching problem with different multi-metric losses.

Local discriminative features are useful when two persons exhibit similar appearance and can't be discriminated against merely by full-body appearance. APNet [Zhong *et al.*, 2020] divides the body into six parts and uses an attention mechanism to weigh the body parts' contribution further. Unlike APNet, which uses arbitrary body parts, CGPS [Yan *et al.*, 2019] proposes a region-based feature learning model for learning contextual information from a person graph. BINet [Dong *et al.*, 2020a] uses the guidance from the cropped person patches to eliminate the context influence outside the bounding boxes.

Deep feature representation learning in text-based person search learns visual representations for the detected person most correspondent to the textual features. Similar to image-based person search, text-based person search methods exploit global and local discriminative features. GNA-RNN [Li *et al.*, 2017b] exploits global features in the first text-based LSTM-CNN person search framework and uses an attention mechanism to learn the most relevant parts. GNA-RNN only attends to visual elements and doesn't address various text structure. To address this problem, CMCE [Li *et al.*, 2017a] employs a latent semantic attention module and is more robust to text syntax variations. To address the background clutter problem, PMA [Jing *et al.*, 2020a] uses pose information to learn the pose-related features from the map of the key points of human. To further distinguish person with similar global appearance, PWM+ATH [Chen *et al.*, 2018b] utilises a word-image patch matching model to capture the local similarities. ViTAA [Wang *et al.*, 2020b] decomposes both image and text into attribute components and conducts a fine-grained matching strategy to enhance the interplay between image and text.

2.2 Deep Metric Learning

Deep metric learning tackles the query-person gap challenge with loss functions to guide the feature representation learning. The general purpose is to bring the detected person features close to the target identity while separating them from other identities. Similarity metrics such as Euclidean distance and cosine similarity are common measures to evaluate the similarity level among those query-person pairs. The identification task is generally formulated as a classification problem where conventional softmax loss trains the classifier. Softmax has a major problem of slow convergence with a large number of classes. OIM (Eq: 2) [Xiao *et al.*, 2017] addresses this issue while exploiting large number of identities and unlabeled identities. OIAM [Gao *et al.*, 2019] and IEL [Shi *et al.*, 2018] further improve the OIM method with additional center losses. Different from OIM variances, I-Net [He and Zhang, 2019] introduces a Siamese structure with an online pairing loss (OPL) and hard example priority Softmax loss (HEP) to bring the query-person gap. RDLR [Han *et al.*, 2019] uses the identification loss instead of regression loss for supervising the bounding boxes.

In the landmark OIM approach, the OIM loss effectively closes the query-person gap utilising labelled and unlabeled identities from training data. The probability of detected person features x being recognised as the identity with class-id i by a Softmax function:

$$p_i = \frac{\exp(v_i^T x/\tau)}{\sum_{j=1}^{L} \exp(v_j^T x/\tau) + \sum_{k=1}^{Q} \exp(u_k^T x/\tau)}. \quad (1)$$

Where v_i^T is the labelled person features for the i_{th} identity in the lookup table (LUT). v_j^T is the j_{th} labelled person features in the LUT. u_k^T is the k_{th} unlabelled person features in the LUT. τ regulates probability distribution. OIM objective is to maximize the expected log-likelihood of the target t.

$$\mathcal{L} = \mathrm{E_x}\left[\log p_t\right]. \quad (2)$$

Metric learning in text-based person search is to close the text-image modality gap. The main challenge in text-based person search is that it requires the model to deal with the complex syntax from the free-form textual description. To tackle this, methods like ViTAA, CMCE, PWM+ATH [Wang *et al.*, 2020b; Li *et al.*, 2017a; Chen *et al.*, 2018b] employ attention mechanism to build relation modules between visual and textual representations. Unlike the above three methods, which are all the CNN-RNN frameworks, Dual Path [Zheng *et al.*, 2020b] employs CNN for textual feature learning and proposes an instance loss for image-text retrieval. CMPM+CMPC [Zhang and Lu, 2018] utilizes a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss to learn discriminative image-text representations. Similar to CMPM+CMPC, MAN [Jing *et al.*, 2020b] proposes cross-modal objective functions for joint embedding learning to tackle the domain adaptive text-based person search.

Inspired by the recent success of knowledge distillation [Hinton *et al.*, 2015], instead of directly training detection and identification sub-nets, the two modules can be learnt from the pre-trained detection and identification models [Munjal *et al.*, 2019b]. DKD [Zhang *et al.*, 2020] focuses on improving the performance of identification by introducing diverse knowledge distillation in learning the identification model. Specifically, a pre-trained external identification model is used to teach the internal identification model. A simplified knowledge distillation process is illustrated in Figure 4.

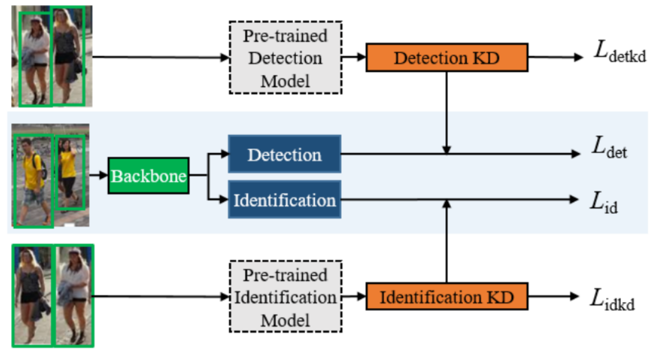

Figure 4: A representative end-to-end person search framework where the detection and identification branches are supervised by pre-trained detection and identification models through knowledge distillation. The detection loss L_{det}, identification loss L_{id} and the knowledge distillation losses L_{detkd} and L_{idkd} can be optimised as a multi-task learning task through back-propagation.

2.3 Identity-driven Detection

The detection-identification inconsistency challenge in image-based person search is tackled by incorporating identities into the detection process. This means during training, ground-truth person identities are used to guide person proposals, or at search time, the query identity information is utilised to refine the bounding boxes. Person search tackles person detection and person identification challenges in one framework. Existing person search methods can be divided into two-stage and end-to-end solutions from the architecture perspective. In two-stage person detection, the detection and identification models are trained separately for optimal performance of both detection and identification models [Zhang *et al.*, 2020; Loesch *et al.*, 2019]. However, due to the detection-identification inconsistent issue, the separately trained models may not yield the best person search result. To address the inconsistency problem between the two branches, TCTS [Wang *et al.*, 2020a] and IGPN+PCB [Dong *et al.*, 2020b] exploit query information at search time to filter out low probable proposals. End-to-end methods share visual features between detection and identification and significantly decrease runtime. However, joint learning contributes to sub-optimal detection performance [Wang *et al.*, 2020a], which subsequently worsen the detection-identification inconsistency problem. To address the problem. NPSM [Liu *et al.*, 2017a] and QEEPS [Munjal *et al.*, 2019a] leverage query information to optimise person proposals in detection process. Differ from the query-guided methods, RDLR [Han *et al.*, 2019] supervises bounding box generation using identification loss. Therefore, proposal bounding boxes are more reliable. In person search settings, the query identity is present in gallery images. Therefore, all methods mentioned above essentially incorporate identities into the detection process.

Text-based person search faces less detection-identification inconsistency challenge since the proposal person is identified by text-image matching without comparing bounding boxes. Therefore, text-based person search mainly focuses on learning visual and language features and improving the matching accuracy. The majority of current text-based person search methods are end-to-end frameworks that consist of a CNN backbone for extracting visual elements and a bi-LSTM for learning language representations. The two modules are jointly trained to build word-image relations from the learnt visual and language feature representations. CMCE [Li *et al.*, 2017a] is the only two-stage framework in which the stage-one CNN-LSTM network learns cross-modal features, and in stage-two, the CNN-LSTM network refines the matching results using an attention mechanism.

3 Datasets and Evaluation

3.1 Datasets

CUHK-SYSU [Xiao *et al.*, 2017] and PRW [Zheng *et al.*, 2017] are the most commonly used datasets for image-based person search. CUHK-SYSU contains 18,184 images, 8,432 person identities, and 99,809 annotated bounding boxes. The training set contains 11,206 images and 5,532 query identities. The test set has 6,978 images and 2,900 query identities. PRW dataset has 11,816 frames and 43,110 person bounding boxes. 34,304 people have identities ranging from 1 to 932, and the rest are assigned identities of -2. The PRW training set has 5,704 images and 482 identities, and the test set has 6112 pictures and 450 identities. LSPS [Zhong *et al.*, 2020] is a new image-based person search dataset, in which a total number of 51,836 pictures are collected. 60,433 bounding boxes and 4,067 identities are annotated. LSPS has a substantially larger number of incomplete query bounding boxes, making it a specialised dataset to evaluate methods exploiting

			CUHK-SYSU		PRW		LSPM	
Method	**Feature**	**Loss**	mAP	top-1	mAP	top-1	mAP	top-1
Non-identity-driven detection								
OIM [Xiao *et al.*, 2017]	global	OIM	75.50	78.70	21.30	49.90	14.40	47.70
IAN [Xiao *et al.*, 2019]	global	Softmax, Center loss	76.30	80.10	23.00	61.90		
OIAM [Gao *et al.*, 2019]	global	OIM, Center loss	76.98	77.86	**51.02**	69.85		
FMT-CNN [Zhai *et al.*, 2019]	global	OIM, Softmax	77.20	79.80				
ELF16 [Yang *et al.*, 2017]	global & local	OIM	77.80	80.60				
IOIM [Liu *et al.*, 2018a]	global	IOIM, Center loss	79.78	79.90	21.00	63.10		
EEPSS [Loesch *et al.*, 2019]	global	Triplet loss	79.40	80.50	25.20	47.00		
JDI + IEL [Shi *et al.*, 2018]	global	IEL	79.43	79.66	24.26	69.47		
RCAA [Chang *et al.*, 2018]	global & context	RL reward		81.30				
I-NET [He and Zhang, 2019]	global	OLP, HEP	79.50	81.50				
MGTS [Chen *et al.*, 2018a]	global & mask	OIM	83.00	83.70	32.60	72.10		
KD-OIM [Munjal *et al.*, 2019b]	global	OIM	83.80	84.20				
CGPS [Yan *et al.*, 2019]	global & context	OIM	84.10	86.50	33.40	73.60		
PFFN [Hong *et al.*, 2019]	global & multi scale	Triplet loss	84.50	89.80	34.30	73.90		
SMG [Zheng *et al.*, 2020a]	global & mask	Binary Cross Entropy	86.30	86.50				
FPSP [Li *et al.*, 2019a]	global	Cross entropy	86.99	89.87	44.45	70.58		
CLSA [Lan *et al.*, 2018]	global & multi-scale	Cross entropy	87.20	88.50	38.70	65.00		
APNet [Zhong *et al.*, 2020]	local	OIM	88.90	89.30	41.90	81.40	18.80	55.70
DHFF [Lu *et al.*, 2019]	global & multi-scale	Multi-Metric loss	90.20	91.70	41.10	70.10		
BINet [Dong *et al.*, 2020a]	global & local	OIM	90.80	91.60	47.20	83.40		
NAE+ [Chen *et al.*, 2020]	global	OIM	92.10	92.90	44.00	81.10		
DKD [Zhang *et al.*, 2020]	global & local		93.60	94.72	**54.16**	**87.89**		
Identity-driven detection								
NPSM [Liu *et al.*, 2017a]	global	Softmax	77.90	81.20	24.20	53.10		
QEEPS [Munjal *et al.*, 2019a]	global	OIM	84.40	84.40	37.10	76.70		
KD-QEEPS [Munjal *et al.*, 2019b]	global	OIM	85.00	85.50				
IGPN + PCB [Dong *et al.*, 2020b]	global		90.30	91.40	47.20	87.00		
RDLR [Han *et al.*, 2019]	global	Proxy Triplet Loss	93.00	94.20	42.90	70.20		
TCTS [Wang *et al.*, 2020a]	global	IDGQ loss	**93.90**	**95.10**	46.80	87.50		

Table 2: Performance of image-based person search methods on CUHK-SYSU, PRW and LSPM datasets.

	Image-based			**Text-based**
Dataset	CUHK-SYSU	PRW	LSPS	CUHK-PEDES
#frames	18,184	11,816	51,836	40,206
#identities	8,432	932	4,067	13,003
#boxes	96,143	34,304	60,433	
#parts(%)	6	0	60	
#cameras		6	17	
#description				80,440

Table 3: Person search datasets statistics.

local discriminative features. CUHK-PEDES is currently the only dataset for text-based person search. It contains 40,206 images of 13,003 identities and 80,440 textual descriptions. Each picture has 2 textual descriptions. The dataset is divided into three parts, 11,003 training individuals with 34,054 images and 68,126 captions, 1,000 validation persons with 3,078 images and 6,158 sentences, and 1,000 test identities with 3,074 pictures 6,156 captions. Dataset statistics are summarised in Table 3.

3.2 Evaluation Metrics

Cumulative matching characteristics (CMC top-K) and mean averaged precision (mAP) are the primary evaluation metrics for person search. In CMC, the top-K predicted bounding boxes are ranked according to the intersection-over-union (IoU) overlap with the ground-truths equal to or greater than 0.5. The mAP is a popular evaluation metric in object detection, in which an averaged precision (AP) is calculated for each query person, and then the final mAP is calculated as an average of all APs.

3.3 Performance Analysis

In this section, we summarise and analyse the evaluation results considering the three significant challenges in person search discussed earlier. We aim to present the influencing factors that contribute to the person search performance. We don't discuss CNN backbones as modern CNN backbones such as ResNet50 and VGG are similar in performance and are mostly interchangeable in different methods.

We summarise the evaluation results of image-based person search methods in Table 2. We annotate feature types and loss functions used for metric learning along with the methods. Image-based person search faces the steep detection-identification inconsistency challenge. Therefore, we divide image-based person search methods into identity-driven detection and non-identity-driven detection methods to analyse the identity-driven detection solution's effectiveness.

Methods specifically addressing the detection and identification inconsistency challenge, such as IGPN, RDLR and TCTS, outperform methods addressing the detection and identification separately. Methods exploiting fine-grained

Method	**Feature**	**Loss**	top-1	top-5	top-10
GNA-RNN [Li *et al.*, 2017b]	global	Cross entropy	19.05		53.63
CMCE [Li *et al.*, 2017a]	global	CMCE loss	25.94		60.48
PWM+ATH [Chen *et al.*, 2018b]	global	Cross entropy	27.14	49.45	61.02
Dual-Path [Zheng *et al.*, 2020b]	global	Ranking loss, Instance loss	44.40	66.26	75.07
CMPM+CMPC [Zhang and Lu, 2018]	global	CMPM, CMPC	49.37		79.27
LPS+MCCL [Liu *et al.*, 2019]	global	MCCL	50.58		79.06
A-GANet [Liu *et al.*, 2019]	global	Binary Cross Entropy	53.14	74.03	81.95
PMA [Jing *et al.*, 2020a]	global & pose		53.81	73.54	81.23
TIMAM [Sarafianos *et al.*, 2019]	global	Cross Entropy, GAN Loss	55.41	**77.56**	**84.78**
ViTAA [Wang *et al.*, 2020b]	global & attribute	Alignment loss	**55.97**	75.84	83.52

Table 4: Performance of text-based person search methods on the CUHK-PEDES dataset.

discriminative features without considering the detection-identification inconsistency challenge don't have a clear edge over methods using global features. Our interpretation is that the query identity presents in the gallery images. Therefore, the detected person needs to be consistent with the query identity for better query-person matching. For example, if the detected person features are free from noises, the query should be free of noises. Loss functions play critical roles in guiding feature representation learning, such as using a center loss on top of the OIM loss to bring the same identities closer and separate different identities. Knowledge distillation is a notably effective strategy in training the detection and identification models. KD-OIM, KD-QEEPS and DKD beat the corresponding baseline methods without knowledge distillation.

The performance of the text-based person search methods on CUHK-PEDES is summarised in Table 4. We include feature types and loss functions along with the methods. Text-based person search is essentially a text-image matching problem, and fine-grained discriminative features play a critical role in cross-modal matching. Recent methods exploiting fine-grained discriminative features with novel loss functions outperform methods using global features and vanilla Cross-Entropy loss. Specifically, ViTAA [Wang *et al.*, 2020b] exploiting local discriminative features via attribute-feature alignment achieves the best search results.

4 Discussion and Future Directions

In this survey, we review the recent person search advances covering both image-based and text-based person search. It remains an open question on addressing the three significant person search challenges, namely the discriminative features, the query-person gap and the detection-identification inconsistency. Next, we discuss a few future research directions.

Multi-modal person search. Exiting works focus on search by either image or text. None of them attempted a multi-modal search approach, in which query image and query text complement each other. Multi-modal person search is handy when a partial person image is available such as a passport-sized image. At the same time, the free text provides the rest of the body appearance. Specifically, the CUHK-PEDES dataset can be extended with annotated bounding boxes. Thus CUHK-PEDES has both annotated bounding boxes and textual descriptions, making it a suitable candidate dataset for multi-modal person search.

Attribute-based person search. It is a big challenge for a machine to learn complex sentence syntax. Attribute-based person search method AIHM [Dong *et al.*, 2019] outperforms the text-based method GNA-RNN [Li *et al.*, 2017b] evaluated on cropped person images with attribute annotations. The state-of-the-art text-based person search method ViTAA [Wang *et al.*, 2020b] decomposes textual description to attributes to learn fine-grained discriminative features. Attribute annotated datasets may ease this process and subsequently improve text-based person search performance.

Zero-shot person search. Text-based person search is essentially a zero-shot learning problem, in which the query person is unseen in training. [Dong *et al.*, 2019] formulates the attribute-based person search as a Zero-Shot Learning (ZSL) problem. In zero-shot learning, zero training image is available at training time, and only semantic representations such as textual descriptions are available to infer unseen classes. Text-based person search can leverage the knowledge of zero-shot learning, such as using adversarially generated person features to augment training data.

5 Conclusion

In this survey, we provide a systematic review of the recent works on person search. For the first time, we surveyed papers on text-based person search which is less investigated than image-based person search. We briefly discuss highly regarded methods from the perspective of challenges and solutions. We summarise and compare person search methods' performance and provide insights that a person search method needs to address the joint challenge of discriminative features, query-person gap, and detection-identification inconsistency. We finally discuss some future research directions which may be of interest to incumbent and new researchers in the field.

Acknowledgments

This paper is partially supported by Australian Research Council (ARC) Discovery Early Career Award (DECRA), and National Natural Science Foundation of China under No. 61972315.

References

[Chang *et al.*, 2018] Xiaojun Chang, Po-Yao Huang, Yi-Dong Shen, Xiaodan Liang, Yi Yang, and Alexander G. Hauptmann. RCAA: Relational Context-Aware Agents for Person Search. pages 84–100, 2018.

[Chen *et al.*, 2018a] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person Search via A Mask-guided Two-stream CNN Model. pages 734–750, 2018.

[Chen *et al.*, 2018b] Tianlang Chen, Chenliang Xu, and Jiebo Luo. Improving Text-Based Person Search by Spatial Matching and Adaptive Threshold. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1879–1887, March 2018.

[Chen *et al.*, 2020] Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-Aware Embedding for Efficient Person Search. pages 12615–12624, 2020.

[Cheng *et al.*, 2017] De Cheng, Xiaojun Chang, Li Liu, Alexander G. Hauptmann, Yihong Gong, and Nanning Zheng. Discriminative dictionary learning with ranking metric embedded for person re-identification. In *IJCAI*, 2017.

[Cheng *et al.*, 2018] De Cheng, Yihong Gong, Xiaojun Chang, Weiwei Shi, Alexander G. Hauptmann, and Nanning Zheng. Deep feature learning via structured graph laplacian embedding for person re-identification. *Pattern Recognit.*, 82:94–104, 2018.

[Dong *et al.*, 2019] Qi Dong, Shaogang Gong, and Xiatian Zhu. Person Search by Text Attribute Query As Zero-Shot Learning. pages 3652–3661, 2019.

[Dong *et al.*, 2020a] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Bi-Directional Interaction Network for Person Search. pages 2839–2848, 2020.

[Dong *et al.*, 2020b] Wenkai Dong, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Instance Guided Proposal Network for Person Search. pages 2585–2594, 2020.

[Gao *et al.*, 2019] Cunyuan Gao, Rui Yao, Jiaqi Zhao, Yong Zhou, Fuyuan Hu, and Leida Li. Structure-aware person search with self-attention and online instance aggregation matching. *Neurocomputing*, 369:29–38, December 2019.

[Han *et al.*, 2019] Chuchu Han, Jiacheng Ye, Yunshan Zhong, Xin Tan, Chi Zhang, Changxin Gao, and Nong Sang. Re-ID Driven Localization Refinement for Person Search. pages 9814–9823, 2019.

[He and Zhang, 2019] Zhenwei He and Lei Zhang. End-to-End Detection and Re-identification Integrated Net for Person Search. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, Lecture Notes in Computer Science, pages 349–364, Cham, 2019. Springer International Publishing.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv:1503.02531 [cs, stat]*, March 2015. arXiv: 1503.02531.

[Hong *et al.*, 2019] Zheran Hong, Bin Liu, Yan Lu, Guojun Yin, and Nenghai Yu. Scale Voting With Pyramidal Feature Fusion Network for Person Search. *IEEE Access*, 7:139692–139702, 2019. Conference Name: IEEE Access.

[Islam, 2020] Khawar Islam. Person search: New paradigm of person re-identification: A survey and outlook of recent works. *Image and Vision Computing*, 101:103970, September 2020.

[Jing *et al.*, 2020a] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. Pose-Guided Multi-Granularity Attention Network for Text-Based Person Search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11189–11196, April 2020. Number: 07.

[Jing *et al.*, 2020b] Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Cross-Modal Cross-Domain Moment Alignment Network for Person Search. pages 10678–10686, 2020.

[Lan *et al.*, 2018] Xu Lan, Xiatian Zhu, and Shaogang Gong. Person Search by Multi-Scale Matching. pages 536–552, 2018.

[Li *et al.*, 2017a] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-Aware Textual-Visual Matching With Latent Co-Attention. pages 1890–1899, 2017.

[Li *et al.*, 2017b] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person Search With Natural Language Description. pages 1970–1979, 2017.

[Li *et al.*, 2019a] Jianheng Li, Fuhang Liang, Yuanxun Li, and Wei-Shi Zheng. Fast Person Search Pipeline. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1114–1119, July 2019. ISSN: 1945-788X.

[Li *et al.*, 2019b] Zhihui Li, Wenhe Liu, Xiaojun Chang, Lina Yao, Mahesh Prakash, and Huaxiang Zhang. Domain-aware unsupervised cross-dataset person re-identification. In *ADMA*, 2019.

[Liu *et al.*, 2017a] Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. Neural Person Search Machines. pages 493–501, 2017.

[Liu *et al.*, 2017b] Wenhe Liu, Xiaojun Chang, Ling Chen, and Yi Yang. Early active learning with pairwise constraint for person re-identification. In *ECML PKDD*, 2017.

[Liu *et al.*, 2018a] Hong Liu, Wei Shi, Weipeng Huang, and Qiao Guan. A Discriminatively Learned Feature Embedding Based on Multi-Loss Fusion For Person Search. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1668–1672, April 2018. ISSN: 2379-190X.

[Liu *et al.*, 2018b] Wenhe Liu, Xiaojun Chang, Ling Chen, and Yi Yang. Semi-supervised bayesian attribute learning for person re-identification. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *AAAI*, 2018.

[Liu *et al.*, 2019] Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang. Deep Adversarial Graph Attention Convolution Network for Text-Based Person Search. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 665–673, New York, NY, USA, October 2019. Association for Computing Machinery.

[Liu *et al.*, 2020a] Chong Liu, Xiaojun Chang, and Yi-Dong Shen. Unity style transfer for person re-identification. In *CVPR*, 2020.

[Liu *et al.*, 2020b] Wenhe Liu, Xiaojun Chang, Ling Chen, Dinh Phung, Xiaoqin Zhang, Yi Yang, and Alexander G. Hauptmann. Pair-based uncertainty and diversity promoting early active learning for person re-identification. *ACM Trans. Intell. Syst. Technol.*, 11(2):21:1–21:15, 2020.

[Loesch *et al.*, 2019] Angelique Loesch, Jaonary Rabarisoa, and Romaric Audigier. End-To-End Person Search Sequentially Trained On Aggregated Dataset. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4574–4578, September 2019. ISSN: 2381-8549.

[Lu *et al.*, 2019] Yan Lu, Zheran Hong, Bin Liu, Weihai Li, and Nenghai Yu. Dhff: Robust Multi-Scale Person Search by Dynamic Hierarchical Feature Fusion. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3935–3939, September 2019. ISSN: 2381-8549.

[Munjal *et al.*, 2019a] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-Guided End-To-End Person Search. pages 811–820, 2019.

[Munjal *et al.*, 2019b] Bharti Munjal, Fabio Galasso, and Sikandar Amin. Knowledge Distillation for End-to-End Person Search. *arXiv:1909.01058 [cs]*, September 2019. arXiv: 1909.01058.

[Sarafianos *et al.*, 2019] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. Adversarial Representation Learning for Text-to-Image Matching. pages 5814–5824, 2019.

[Shi *et al.*, 2018] Wei Shi, Hong Liu, Fanyang Meng, and Weipeng Huang. Instance Enhancing Loss: Deep Identity-Sensitive Feature Embedding for Person Search. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4108–4112, October 2018. ISSN: 2381-8549.

[Wang *et al.*, 2020a] Cheng Wang, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. TCTS: A Task-Consistent Two-Stage Framework for Person Search. pages 11952–11961, 2020.

[Wang *et al.*, 2020b] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. ViTAA: Visual-Textual Attributes Alignment in Person Search by Natural Language. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 402–420, Cham, 2020. Springer International Publishing.

[Xiao *et al.*, 2017] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint Detection and Identification Feature Learning for Person Search. pages 3415–3424, 2017.

[Xiao *et al.*, 2019] Jimin Xiao, Yanchun Xie, Tammam Tillo, Kaizhu Huang, Yunchao Wei, and Jiashi Feng. IAN: The Individual Aggregation Network for Person Search. *Pattern Recognition*, 87:332–340, March 2019.

[Xu *et al.*, 2014] Yuanlu Xu, Bingpeng Ma, Rui Huang, and Liang Lin. Person Search in a Scene by Jointly Modeling People Commonness and Person Uniqueness. In *Proceedings of the 22nd ACM international conference on Multimedia*, MM '14, pages 937–940, New York, NY, USA, November 2014. Association for Computing Machinery.

[Yan *et al.*, 2019] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning Context Graph for Person Search. pages 2158–2167, 2019.

[Yang *et al.*, 2017] Jinfu Yang, Meijie Wang, Mingai Li, and Jingling Zhang. Enhanced Deep Feature Representation for Person Search. In Jinfeng Yang, Qinghua Hu, Ming-Ming Cheng, Liang Wang, Qingshan Liu, Xiang Bai, and Deyu Meng, editors, *Computer Vision*, Communications in Computer and Information Science, pages 315–327, Singapore, 2017. Springer.

[Zhai *et al.*, 2019] Sulan Zhai, Shunqiang Liu, Xiao Wang, and Jin Tang. FMT: fusing multi-task convolutional neural network for person search. *Multimedia Tools and Applications*, 78(22):31605–31616, November 2019.

[Zhang and Lu, 2018] Ying Zhang and Huchuan Lu. Deep Cross-Modal Projection Learning for Image-Text Matching. pages 686–701, 2018.

[Zhang *et al.*, 2020] Xinyu Zhang, Xinlong Wang, Jia-Wang Bian, Chunhua Shen, and Mingyu You. Diverse Knowledge Distillation for End-to-End Person Search. *arXiv:2012.11187 [cs]*, December 2020. arXiv: 2012.11187.

[Zheng *et al.*, 2017] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person Re-Identification in the Wild. pages 1367–1376, 2017.

[Zheng *et al.*, 2020a] Dingyuan Zheng, Jimin Xiao, Kaizhu Huang, and Yao Zhao. Segmentation mask guided end-to-end person search. *Signal Processing: Image Communication*, 86:115876, August 2020.

[Zheng *et al.*, 2020b] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path Convolutional Image-Text Embeddings with Instance Loss. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(2):51:1–51:23, May 2020.

[Zhong *et al.*, 2020] Yingji Zhong, Xiaoyu Wang, and Shiliang Zhang. Robust Partial Matching for Person Search in the Wild. pages 6827–6835, 2020.