Neurocomputing 501 (2022) 270–284

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Revisiting instance search: A new benchmark using cycle self-training

Yuqi Zhang [a], Chong Liu [b], Weihua Chen [a], Xianzhe Xu [a], Fan Wang [a], Hao Li [a], Shiyu Hu [c,d], Xin Zhao [c,d,*]

[a] Machine Intelligence Technology Lab, Alibaba Group, China
[b] State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, China
[c] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[d] University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 20 January 2022
Revised 30 March 2022
Accepted 6 June 2022
Available online 9 June 2022
Communicated by Zidong Wang

Keywords:
Instance search
Self training
Instance detection
Instance retrieval

ABSTRACT

Instance search aims at retrieving a particular object instance from a set of scene images. Although studied in previous competitions like TRECVID, there have been limited literature or datasets on this topic. In this paper, to overcome the generalization issue when arbitrary categories are involved in search and to benefit from the large amount of unlabeled data, we propose a cycle self-training framework which trains the instance search pipeline with automatic supervision. Given the two-stage pipeline with a localization and ranking module, the cycle self-training includes a ranker-guided localizer, and a localizer-guided ranker, each carefully designed to handle noisy labels that come with self-supervision. Furthermore, we build and release large-scale groundtruth annotations for instances to facilitate the algorithm evaluation and analysis in this research topic, especially for small objects in complex background. The datasets are publicly available at https://github.com/instance-search/instance-search. Extensive experiments show the effectiveness of the proposed cycle self-training framework and the superior performance compared with other state-of-the-art methods.

1. Introduction

Is it possible to search the same instance in scene images from multiple cameras? Imagine trying to identify the owner of an unattended suitcase in an airport for security reasons, by looking through all surveillance cameras and tracking it, or searching for the accomplices of a suspect by identifying those who helped transfer a briefcase. Unlike person re-identification which mainly focuses on retrieving a particular person, the object of interest here could be the suitcase, the briefcase, or any arbitrary object. Instance search is just designed for these cases.

For clarification, we define instance search as to locate and retrieve object instances of arbitrary category from whole scene images (Fig. 1 (a)). The topic is related to many vision tasks, e.g., person re-identification (ReID), person search, content based image retrieval (CBIR), and single object tracking (SOT) as compared in Fig. 2. Person ReID uses pre-detected person images for retrieval, and the category is limited to the person category. Person search steps further than person ReID because the query person should search from whole scene images. Image retrieval often focuses on salient objects in limited categories, e.g., buildings. Single object tracking handles tiny objects of arbitrary classes. However, the search is executed only in a local region by the spatial–temporal constraint. Single object tracking searches through one video while instance search operates through all candidate scene images. More discussions among these tasks are listed in Section 2.

Instance search is very challenging for the following reasons: (1) **Precise localization and ranking**: The same instance captured in different cameras might appear in various locations, scales, viewpoints, with various deformations, and occlusions in complex background. (2) **No labeled training data**: There is often no training data for the task. Generalization should be guaranteed for objects of any category. There is room for current studies in both methods and data usage.

The first instance search dataset TRECVID [1] aims to search and locate instances from movies. However, the public availability and data scale are limited, which slows down the academic research. In recent years, instance search mainly focuses on salient objects, which makes the problem simpler. However, these methods with classic features [2–4] or deep features [5–7] fail to localize non-salient objects. Motivated by the Siamese networks [8], recent works [9,10] also employ query-guided Siamese networks for precise localization. However, Siamese networks lack discrimination capability and fail to reject distractor instances with similar appearances [11].

* Corresponding author.

https://doi.org/10.1016/j.neucom.2022.06.027
0925-2312/

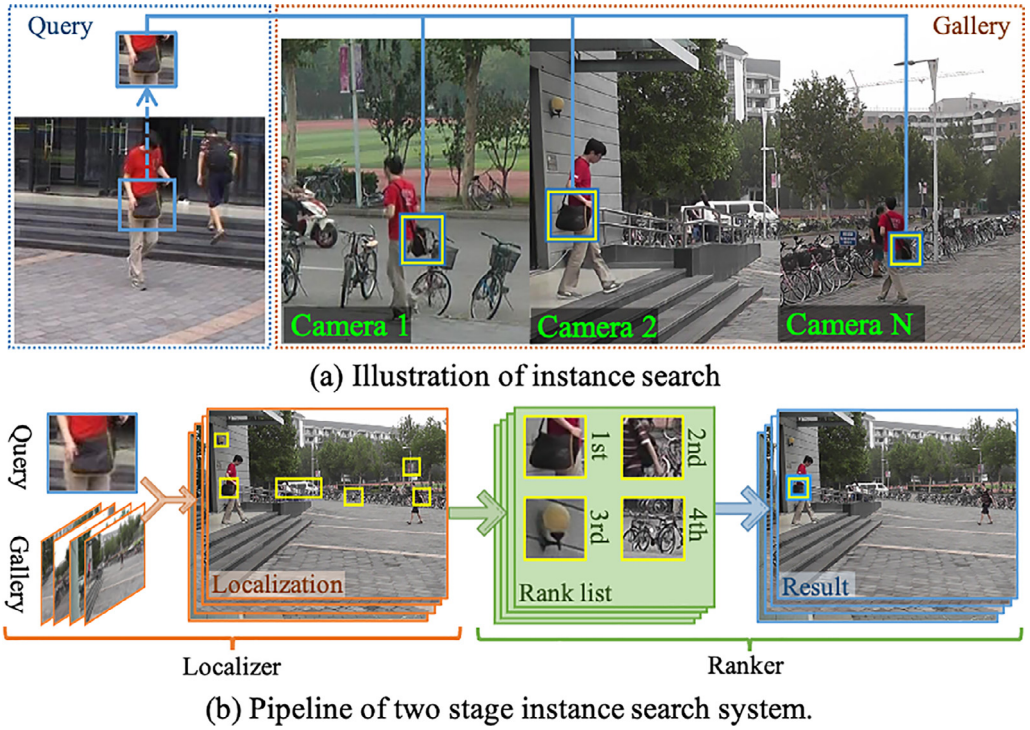
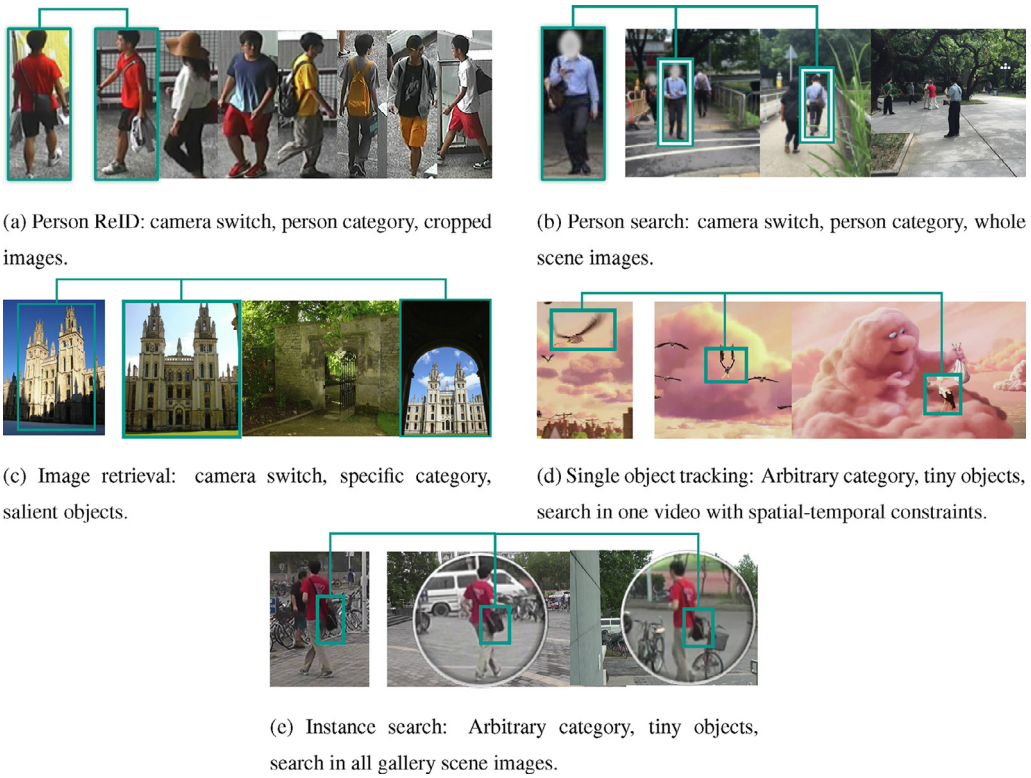


Fig. 1. (a) Illustration of the instance search task. The instance search problem setting is closer to real-world applications and more challenging. **(b)** Pipeline of instance search system. Two modules are included: 1) The localizer to locate candidate bounding boxes in an initial ranking list. 2) The class-agnostic ranker to further refine the results.



(a) Person ReID: camera switch, person category, cropped images.

(b) Person search: camera switch, person category, whole scene images.

(c) Image retrieval: camera switch, specific category, salient objects.

(d) Single object tracking: Arbitrary category, tiny objects, search in one video with spatial-temporal constraints.

(e) Instance search: Arbitrary category, tiny objects, search in all gallery scene images.

Fig. 2. Comparison among person re-identification, person search, image retrieval, single object tracking, and instance search.

Therefore, to simultaneously enhance the localization and discrimination ability, we present our method of instance search system as a two-stage pipeline. It includes a query-guided localizer for precise localization and a class-agnostic ranker for robust feature representation, as shown in Fig. 1 (b). Another reason for the two-stage framework is that each stage could be optimized independently with a specially designed strategy. For a query object-of-interest, a localizer locates candidate bounding boxes within

all the gallery scene images. These candidate boxes are further refined by a ranking stage, which represents each box as a feature vector and sorts the results based on their feature similarities to the query object. The pipeline has been evaluated extensively against different localization/ranking algorithms and proved an effective framework on several datasets.

Following the tradition in the instance search community, the dataset usually does not include a training set. The lack of training data urges us to design algorithms with better generalization and adaptation ability. Self-training has become popular in recent years for its capability of training without human annotations. However, self-training for each of the two modules separately [12–14] looks impractical in our setting because we only have full scene images without any detected or cropped objects. Therefore, a cycle self-training framework is proposed, with a ranker-guided localizer and a localizer-guided ranker. Given candidate bounding boxes generated by a localizer, their feature similarities could be refined by the ranker. The refined similarities can provide supervision on training the Siamese-based localizer. Similarly, self-training for the ranker is possible if the localizer helps generate the bounding boxes of objects. As the two stages can benefit each other during training, the self-training of two modules is performed iteratively, named Cycle Self Training (CST).

To our best knowledge, existing instance search datasets either focus on salient objects with limited variations [1,15] or only cover a limited range of object categories [16]. To facilitate the algorithm evaluation in a more challenging scenario, we choose a few popular public datasets and add additional cross-camera instance-level annotations of arbitrary objects to their original images. In particular, we annotate 535 object instances across 6,079 scene images in PRW dataset (named INS-PRW) and 6,972 object instances across 9,648 scene images in CUHK-SYSU dataset (named INS-CUHK-SYSU). Sample annotations can be found in Fig. 5.

Our contributions are summarized as follows:

- A cycle self-training method is proposed to learn the instance searcher from pure scene images without any labeled training data. The cycle consists of a ranker-guided localizer and a localizer-guided ranker. The two modules feed into each other to improve performance through the iterations, as shown in the experiments.
- A new benchmark is built for instance search. We build and release instance-level annotations of several popular datasets for the research community at https://github.com/instance-search/instance-search. Several baselines and comprehensive experiments have been conducted on these datasets. The proposed self-training method outperforms other supervised methods and implies its great potential usage.

2. Related work

2.1. Instance search

Instance search was first introduced as a task in TRECVID. It aims to search for instances of the specific query object, person, or place entity. However, the public availability and data scale are limited. Instance-160 [16] and Instance-335 [15] step further with larger data scale. However, the search region is limited in the single video as the two datasets reuse annotations from single object tracking. Instance search develops from early methods for salient objects to current methods for precise localization. Mohedano et al. [5] produced an assignment map for each local array of activations in a convolutional layer to a visual word. The assignment map was used for fast spatial re-ranking and object localization. Salvador et al. [17] extracted image and region-wise representations pooled from Faster R-CNN. The instance search pipeline is composed of a filtering stage followed by a spatial re-ranking. Zhan et al. [16] used instance-aware semantic segmentation (FCIS) [18] for instance search. Zhang et al. [10] proposed multi-task integration of joint detection and retrieval. The end-to-end framework was improved based on classic Siamese networks by a novel online pairing (OLP) loss and a hard example priority (HEP). Recently, Hong et al. [19] proposed a self-paced learning framework to achieve accurate object localization on the rank list returned by instance search. Some works propose weakly-supervised learning for the lack of labeled data. Guan et al. [20] proposed a novel tag-based weakly-supervised deep hashing framework. Zhao et al. [21,22] proposed a weakly-supervised Deep Multiple Instance Hashing (DMIH) framework for object-based image retrieval and search. The above methods either suffer from precise localization or lack discrimination with only Siamese networks. We are different from these methods as we combine query-guided localizer and ReID network to guarantee precise searching results. More importantly, we do not rely on annotated training data and only learn instance search models from unlabeled data.

2.2. Person search

Person search aims to localize the same query person from a gallery of whole scene images. Ever since the publication of two large-scale datasets, CUHK-SYSU [23] and PRW [24], many methods have been proposed. On the one hand, the popular two-step framework [24] separates the task into person detection and person ReID, which would benefit from the state-of-the-art algorithms in both research communities. Apart from general person detectors, query-guided detectors have also been used for more accurate proposals. Dong et al. [25] proposed a new detection network named Instance Guided Proposal Network (IGPN) to produce query-related proposals. On the other hand, there are also one-stage frameworks. Liu et al. [26] proposed Conv-LSTM based Neural Person Search Machines (NPSM). Chang et al. [27] regard the search process as a conditional decision-making process. Chen et al. [28] proposed Norm-Aware Embedding to disentangle the person embedding into norm and angle for detection and ReID tasks, respectively. Although well studied, the topic only focuses on the person category and relies heavily on person detector and person embedding. The generalization to unseen arbitrary categories could not be preserved.

2.3. Other related topics

We introduce some other topics which may confuse with instance search. The comparison among these datasets can be found in Fig. 2.

2.3.1. Person ReID

Person ReID is the task of associating images of the same person from detected human images. Since person detection is conducted by manual annotation or algorithms, the task does not focus on whole scene images. Current studies focus on global [29,30] or local [31–33] feature learning, loss functions [34,35,30], bag-of-tricks [36,37] and reranking methods [38]. Although well studied, the topic limits the category. Also, the additional person detector is necessary for whole scene images.

2.3.2. Image retrieval

Content Based Image Retrieval (CBIR), a well-studied topic for several years, retrieves images from a large gallery based on their semantic similarities to the query image. The widely used instance-level CBIR datasets include Oxford [39], Paris [40] and Google Landmarks [41], *etc.*. Compact global representations could

be learned from various loss functions [42–44]. Apart from global features, deep-learning-based local descriptors [41,45] have been widely used. Recent methods often combine both global and local features [45,46], with global features to generate the initial rank list and local features to refine the results by further geometric verification.

2.3.3. Single object tracking

Given the initial bounding box provided in the first frame, single object tracking aims to locate the instance in the following frames of the same video. Single object tracking methods also consider arbitrary categories. However, there is a strong assumption that objects should be searched in a local spatial–temporal window rather than the whole image. Huang et al. [47] proposed GlobalTrack which performs searching on the full image without online learning. The method imposes no constraints on the temporal consistency of the target's location or scale to avoid cumulative errors. However, the model alone is indeed a localizer and lacks discrimination capability.

3. Methods

3.1. Brief description

We first introduce the two-stage framework with a localizer and a ranker for instance search in 3.2. We propose self-training for the two modules to learn from unlabeled data arbitrary class in 3.3. The self-training of the two modules can be performed iteratively to achieve better overall performance.

3.2. Two-stage framework

We introduce our two-stage framework with a query-guided localizer [47] and a class-agnostic ranker [48] as shown in Fig. 1. For a given query Q of arbitrary category, the localizer first produces several candidate boxes from all scene images as $C = \{(p_1, d_1), (p_2, d_2), \ldots (p_N, d_N)\}$, where p_i stands for a candidate and d_i for localization score. The localizer alone cannot guarantee discrimination. A ranker is then introduced to extract feature embeddings of the candidates and then use cosine similarities to refine the results. The final ranklist could be written as $l = \{(p_1, s_1), (p_2, s_2), \ldots (p_N, s_N)\}$, where s_i stands for the refined similarity score. With the combination of the two modules, the framework could achieve both precise localization and ranking.

3.2.1. Localizer

It is natural to use Siamese networks for localization tasks. As the query object might appear anywhere in the whole image with any scale, we choose an algorithm like GlobalTrack [47], which does not have any constraints on scale or search range.

GlobalTrack modifies an ordinary two-stage object detector Faster-RCNN by introducing the correlation operation from the query feature map. It converts original RPN and RCNN modules into Query-Guided RPN (QG-RPN) and Query-Guided RCNN (QG-RCNN). We suggest readers GlobalTrack original paper [47] for details.

QG-RPN performs operations as below to generate a feature map $\hat{x}$:

$$\hat{x} = g_{qg_rpn}(q, x) = f_{\text{out}}(f_x(x) \otimes f_q(q)) \tag{1}$$

where $q \in \mathrm{R}^{k\times k\times c}$ and $x \in \mathrm{R}^{k\times k\times c}$ denote query ROI feature and gallery feature respectively, k for ROI feature height and width, c for ROI feature channel number. f_x, f_q are projection functions for the gallery and query image, $\otimes$ stands for the convolution operator, f_{out} generates output features $\hat{x}$ and ensures that it retains the size of x. The obtained feature map $\hat{x}$ can be used in the traditional Region Proposal Network(RPN) module during both training and testing. Next, QG-RCNN is used to generate final bounding boxes. After correlating query information, the detection loss function is just the same as Faster-RCNN with the following formula:

$$L_{det} = L_{cls} + \lambda L_{reg} \tag{2}$$

$$L_{reg} = \sum_{i\in\{\mathrm{x,y,w,h}\}} \mathrm{smooth}_{L_1}(t_i - v_i) \tag{3}$$

$$\mathrm{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{4}$$

where L_{cls} stands for the classification loss and L_{reg} for the bounding box regression loss. λ is the parameter to control the two losses. v_i is the ground truth bounding-box regression offset and t_i is the predicted offset. We emphasize that the correlation between query and gallery is essential for the localizer, and other Siamese networks without temporal constraints could also be used.

3.2.2. Ranker

The class-agnostic ranker extracts feature embedding for any candidate objects, and cosine similarities are then computed. The setting is different from custom person re-identification or face recognition because of the unlimited category. The network can be trained with loss functions like

$$L_{reid} = L_{cls} + \alpha L_{trp} \tag{5}$$

$$L_{cls} = \sum_{i=1}^{N} -q_i \log(p_i) \begin{cases} q_i = 0, & y \neq i \\ q_i = 1, & y = i \end{cases} \tag{6}$$

$$L_{trp} = [d_p - d_n + \alpha]_+ \tag{7}$$

where L_{cls} and L_{trp} stand for softmax cross-entropy loss and triplet loss [35], with α balancing their weights. For classification loss, y is truth ID label and p_i as ID prediction logits of class i. For triplet loss, d_p and d_n are feature distances of positive pair and negative pair. α is the margin of triplet loss, and $[z]+$ equals to max(z, 0). For each input object image, features output from the network can be treated as embeddings, and they are ranked based on their cosine similarities to the query feature.

3.3. Cycle self-training

Instance search suffers from the well-known "domain gap" due to different lighting conditions, camera resolutions, backgrounds, demographic, *etc.*. In addition, the algorithm might also suffer from the bias on instance variations presented in training data. As we aim for arbitrary categories of objects, training data may have insufficient sampling over the enormous space, not to mention that annotating data at the instance level is expensive. These challenges hinder the applications of instance search in the real world, so it is desirable to leverage the large amount of unlabeled data to narrow down or eliminate the domain gap with self-training.

As the proposed algorithm architecture includes two modules, our designed self-training framework includes a ranker-guided localizer and a localizer-guided ranker. Most importantly, a cycle self-training framework is designed to improve the two modules iteratively and achieve better overall performance.

An overview of the proposed cycle self-training framework is shown in Fig. 3, and the algorithm details are summarized in Algorithm 1. The framework can be divided into three parts:

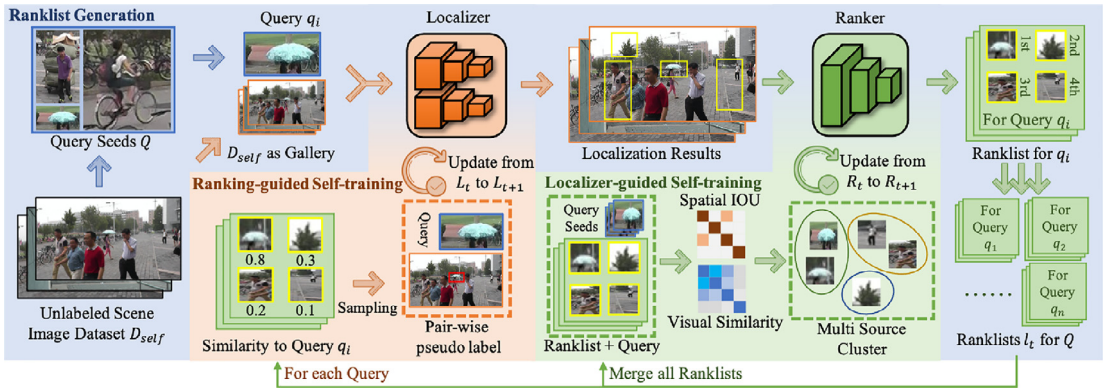


Fig. 3. Illustration of Cycle Self-Training (CST) for instance search. **1)** The blue box is the ranklist generation process, and the ranklist is used for localizer and ranker training. **2)** The orange box is the self-training process of the localizer. The ranklist l_t of t-th iteration is used to update localizer from L_t to L_{t+1}, and L_{t+1} is used for $t+1$-th iteration to generate the new ranklist l_{t+1}. **3)** The green box is the self-training process of ranker. The ranklist l_t of t-th iteration is used to update ranker from R_t to R_{t+1}, and R_{t+1} is used for $t+1$-th iteration to generate the new ranklist l_{t+1}.

- Ranklist generation: At the t-th iteration of the loop, the localizer L_t + ranker R_t together produce a ranklist l_t. This stage is similar to Section 3.2 as no backward propagation is involved.
- Ranker-guided Localizer: l_t is then used for self-training of L_t using the algorithm in Section 3.3.1, producing a better version of localizer L_{t+1}.
- Localizer-guided Ranker: The same ranklist l_t is used to perform self-training of R_{t+1} using algorithm in Section 3.3.2.

Throughout the cyclic style of training, neither the localizer nor the ranker is perfect. However, the designed self-training module can handle noisy pseudo labels robustly, leading to an overall gain and steadily improving pipeline.

Algorithm 1: **C**ycle **S**elf **T**raining (CST)

Input: query set Q, initial localizer L_0, initial ranker R_0, total iterations T
for $t = 0$ **to** $T-1$ **do**
  Generate ranklist l_t by localizer L_t + ranker R_t.
  Localizer self-training for a refined localizer L_{t+1} from ranklist l_t.
  Ranker self-training for a refined ranker R_{t+1} from ranklist l_t.
end for

3.3.1. Ranker-guided Localizer

The self-training of the localizer is depicted in the orange part in Fig. 3. As localizer training needs pairs of images with the same instance ID, the ranker is employed to generate pseudo pair-wise labels.

It is natural to select those highly-ranked predictions to make pseudo pairs as they are more similar to the query. Hard threshold on similarity score or top-K with a fixed K would make the algorithm brittle. Instead, we propose a Similarity-guided Sampling (SGS) strategy. For the ranklist l, we normalize the similarity scores over the whole rank list into a probability format as:

$$prob_i = \frac{s_i}{\sum_{i=1}^{N} s_i} \tag{8}$$

so that $prob_i$ stands for sampling probability for the i-th sample. The pseudo pair-wise labels are sampled according to this normalized probability $prob_i$, so that the samples ranked at the top will contribute more in constructing pseudo pairs and have a larger impact in the localizer training. This strategy can be regarded as a distillation process to transfer knowledge of better similarity scores from the ranker.

3.3.2. Localizer-guided Ranker

SimCLR [49] is a strong framework for image classification self-training and is chosen as our initial setup. A mini-batch of N samples is randomly selected, where each sample is augmented twice to serve as the positive sample and thus results in 2 N data points. For each positive pair denoted as i and j, its contribution to the loss function is written as below:

$$\ell_{i,j} = -\log \frac{\exp\left(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k=1,k\neq i}^{2N} \exp\left(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)} \tag{9}$$

where $sim(u, v)$ denotes the feature similarity between u, v, and τ denotes a temperature parameter.

Compared to the settings of traditional self-supervised learning on image classification or representation learning, there are several additional challenges of performing self-training on images coming from the localization module: (1) distractor images due to imperfect localization, e.g., an object divided into multiple parts. (2) missing data due to imperfect localization, i.e., some object images might not be located.

To handle the data noise and redundancy, we first perform the DBSCAN clustering algorithm as shown in the green part in Fig. 3. Different from the original SimCLR, each sample is assigned a pseudo ID label based on the clustering results, and positive pairs are sampled within each cluster ID. By introducing the additional clustering step, the redundancy in data samples is overcome. It also avoids getting negative pairs for training which actually belong to the same instance.

We propose Multi-Source-Cluster SimCLR (MSC-SimCLR) with both visual and spatial constraints. The pairwise similarity $S_{i,j}$ for clustering is evaluated as the sum of visual feature similarity $V_{i,j}$ and IoU score $IoU_{i,j}$ as $S_{i,j} = V_{i,j} + IoU_{i,j}$, where i,j stands for pair-wise samples. IoU score is introduced because samples generated by the localizer may come from the same scene image, and their overlapping positions can help improve the similarity estimation, especially when the feature similarity is still not well trained yet.

3.3.3. Query seed

An initial set of query objects is required to start the iterative self-training process. To do so, we employ various methods to generate queries automatically. A general object detection COCO-trained YOLOv4 model detects objects in common categories, e.g., cats, dogs. Region proposal algorithms like selective search and

Fig. 4. Annotation pipeline by reusing person search annotations. Row (a) represents person search annotations, Row (b) represents visual similarities by models, and Row (c) represents Refined similarities from IOU.

edge box are also used to find "objectness" boxes, e.g., a book on the desk or traffic signal on the road. We keep each data source with the same ratio to cover all possible objects.

4. The proposed INS dataset

There are several criteria for a dataset to be suitable for our designated purpose: (1) The raw data should come from a multi-camera setup, i.e., there should be sufficient objects appearing in multiple cameras; (2) The raw scene images need to be available so that the background distractor objects could be taken into account instead of focusing only on cropped images of people; (3) The objects should have varied sizes in complex background. To benefit from the existing well-known datasets, PRW [24] and CUHK-SYSU [23] are chosen based on the criteria mentioned above. In addition to their original annotations for person search, we provide further annotations to facilitate instance search based on their raw data.

4.1. Annotation description

We ask a group of well-trained human annotators for data annotation. The annotators are experienced and professional in annotating objects under different cameras. Researchers also double-check the annotations after the annotation. We first manually draw bounding boxes for candidate objects of interest. After the objects are obtained from each scene image, we associate them based on the appearance cues and their relationship with person ID groundtruth provided by PRW and CUHK-SYSU. As shown in Fig. 4, if the candidate boxes have overlap with person boxes of the same ID, their similarities will be greatly increased because they may correspond to the same object instance. The final similarity between a pair of objects is represented as $score_{ij} = sim_{ij} + \max_k \mathbb{1}(iou_{ik} * iou_{jk})$, where sim_{ij} stands for the visual similarity between objects i,j and iou_{ik} stands for intersection over union between i and an annotated person k, $\mathbb{1}(x)$ is an indicator function taking value of 1 only when $x > 0$. After the association, any bounding boxes which do not find any matching instance across cameras would be discarded.

Fig. 5 presents samples from the annotation. We summarize some important features as follows:

- **Different scenes** in Fig. 5a, 5b. The annotated objects cover both indoor and outdoor scenes which makes the annotation closer to real applications.
- **Tiny objects** in Fig. 5c. Some annotations are very small which requires precise localization.
- **Different instances in the same scene image** in Fig. 5d. Different from single object tracking, our annotation allows multiple different instances in the same scene image.
- **Category variates**. We cover a large number of category variates including backpacks, skirts, albums, books, traffic lights, and so on.
- **Different viewpoints, deformations, and occlusions** in Fig. 5e, 5f. We emphasize that these objects can still be clarified by human annotators. No distractors with similar appearances are included and thus the problem is not ill-conditioned.

4.2. Dataset statistics

In total, 535 object instances are annotated in PRW, covering 6,079 scene images, which is denoted as INS-PRW. 6,972 object instances are annotated in CUHK-SYSU, covering 9,648 scene images, which is denoted as INS-CUHK-SYSU. Table 1 compares these two new datasets with several previous related datasets in various aspects. Compared to the existing instance search datasets, our data covers a much larger number of instances. These make the evaluation more challenging and urge us to develop a more robust algorithm.

We define the normalized area as $Area_{patch}/Area_{whole}$ where $Area_{patch}$ and $Area_{whole}$ represent the pixel-level area of the patch and the whole scene image. Fig. 6 shows the distribution of normalized area and aspect ratio ($width/height$) of the annotated boxes on PRW. For our annotated objects, the boxes occupy a much smaller area in the full image than the person boxes. The aspect ratio distribution of the new annotated objects is more spread out than those of person boxes, meaning that our objects are of wider variations. Fig. 7 further lists the category distribution of the datasets by assigning them into the 20 categories of Pascal VOC (by human annotators). Over 90% of the objects are excluded from VOC classes. These unseen objects can be roughly categorized into backpacks, luggage, furniture, logos, and umbrellas. Since the person category has been well studied in person search, we do

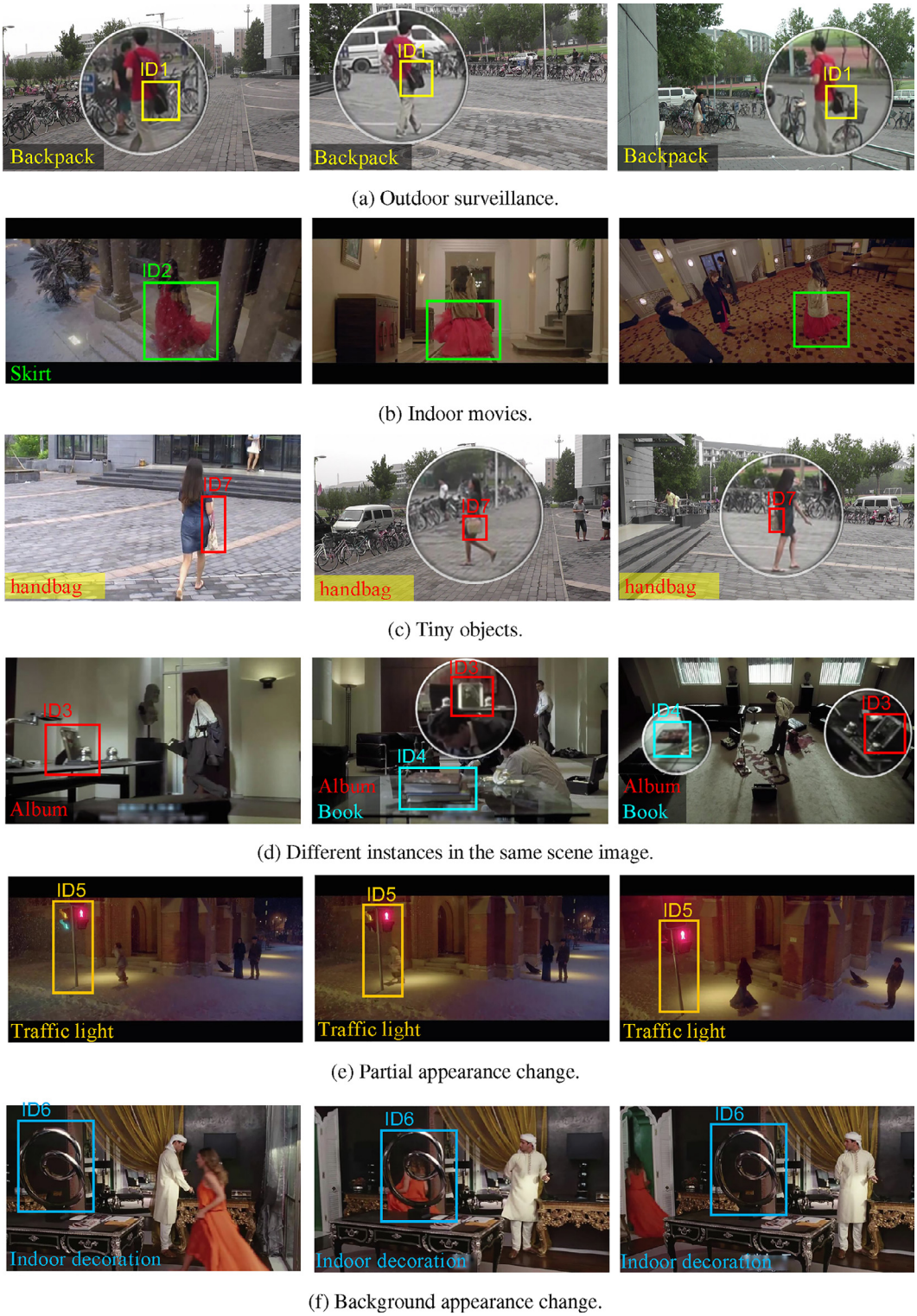


Fig. 5. Sample images of the annotated instances. Target instances vary in different scenes, scales, categories and may or may not be appendages. These instances have different perspectives and appear in different scenes.

not include this category in our annotation. A significant portion of the objects does not belong to person accessories (not carried by a person), which means person detection/ReID methods cannot solve the problem. Overall, the provided annotations make the task more challenging compared to the original person search task and the existing instance search datasets.

5. Experiments

We first demonstrate the effectiveness of the proposed cycle self-training pipeline (Section 6.1), then we report ablation studies where we vary the algorithmic setting of cycle self-training (Section 6.2), and finally, we evaluate the generalization capability on other categories.

Table 1
Comparison among instance search datasets and relevant datasets.

Task	Dataset	# images	# IDs	# boxes	# queries	category	camera switch	tiny	public
person ReID	Market-1501 [50]	32,668	1,501	32,668	3,368	person	✓	-	✓
	Duke [51]	36,441	1,404	36,441	2,228	person	✓	-	✓
person search	CUHK-SYSU [23]	18,184	8,432	23,435	2,900	person	×	✓	✓
	PRW [24]	11,816	932	34,304	2,057	person	✓	✓	✓
CBIR	Oxford [39]	5,062	11	-	55	building	✓	×	✓
	Paris [40]	6,412	11	-	55	building	✓	×	✓
	INSTRE [52]	28,543	250	28,543	1250	arbitrary	×	✓	✓
	DeepFashion2 [53]	491k	43.8k	801k	24,402	clothes	×	×	✓
SOT	GOT-10k [54]	1.5M	10,000	1.5M	10,000	563 classes	×	✓	✓
	LaSOT [55]	3.87M	1,550	3.87M	1,550	85 classes	×	✓	✓
INS	TRECVID-INS [1]	23,614	50	23,614	50	arbitrary	×	×	×
	Instance-160 [16]	11,885	160	11,885	160	COCO 80	×	✓	×
	Instance-335 [15]	40,914	335	40,914	335	arbitrary	×	×	✓
	INS-CUHK-SYSU	9,648	6,972	16,780	6,972	arbitrary	×	✓	✓
	INS-PRW	6,079	535	7,834	1,537	arbitrary	✓	✓	✓

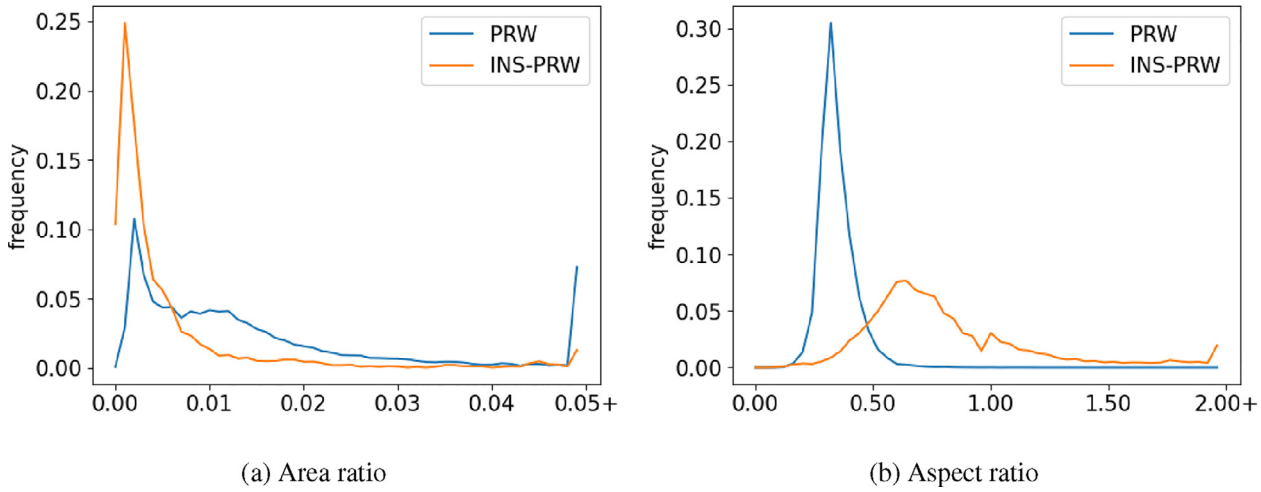


Fig. 6. Statistics of the annotated datasets compared with the original data. Area ratios in (a) indicate smaller annotations than person search. Aspect ratios in (b) shows a wider range than person search.

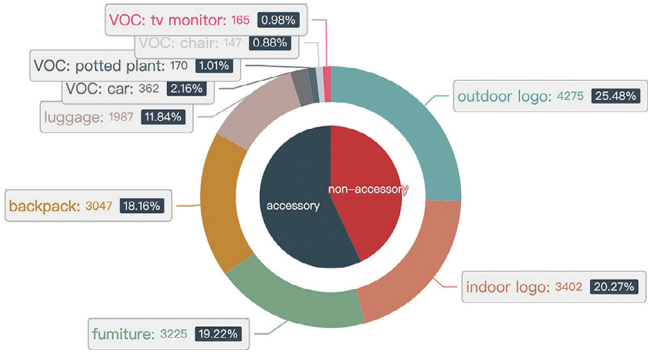


Fig. 7. Category statistics of the INS-CUHK-SYSU. There is only a small portion of VOC predefined categories.

5.1. Datasets

We first propose a two-stage pipeline for INS in the subsequent subsections by collecting objects of varied categories, including GOT-10 k [54], YouTube BB [56] and VID 2015 [57]. These datasets are defined as D_{sup} and the model is defined as 'pretrained' in this paper. Multiple bounding box annotations of the same instance under different views should be provided for supervised learning, which limits the dataset scale. Inspired by the recent progress in self supervised learning, we introduce more datasets including multiple object tracking [58,59], general object detection [60], person detection [61], vehicle tracking [62] and person search [24,23] datasets. We name this more general dataset as D_{self} in our paper.

To evaluate the instance search accuracy for non-salient objects in different views, we use the annotated datasets INS-CUHK-SYSU and INS-PRW as the test set. The overall test set settings are listed as follows:

5.1.1. INS datasets

We choose Instance-160 [16], Instance-335 [15] and our annotated datasets. Other widely used CBIR datasets focus on salient objects and are thus not suitable for our task.

5.1.2. Person search datasets

Two widely used person search datasets CUHK-SYSU [23] and PRW [24] are used to evaluate the generalization of the self-training methods.

5.2. Evaluation protocols

Given a query image with a bounding box, an instance search algorithm should produce multiple prediction boxes (with rank-

ing) on the gallery images. A prediction is regarded as true positive only if it has enough overlap ($iou > 0.5$) with the ground truth bounding box which has the same ID as the query. Rank-1 accuracy and mAP of the rank list are used to evaluate the instance search algorithm. Following the original gallery setting in CUHK-SYSU and PRW, we randomly select 100 local gallery images for a query in INS-CUHK-SYSU and search through the whole gallery set in INS-PRW. No context information near the query bounding box should be allowed because the instance may appear at any location.

5.3. Implementation details

For the localizer, Resnet50_FPN is used as the backbone. The channel number for the backbone feature and ROI feature c is 256. ROI features are extracted by ROI align [63] with the output size as $k = 7$. The size of each input image is normalized so that its longer edge is no larger than 1,333 pixels and shorter edge no larger than 800 pixels. The regression weight λ in Equ. 2 is set to 1.0. We use SGD with a momentum of 0.9 and weight decay of 0.0001. The initial learning rate is 0.001, and it decays with a factor of 0.1 at epoch 8. We train for a total of 12 epochs with a batch size of 4 pairs.

For the ranker, Resnet50 is used as the backbone. We use global average pooling on the last layer of Resnet50 backbone and get a feature of 2,048. We then add a fully connected layer with the dimension of 512 to involve compact features. Then the bottleneck is connected to fully connected layers with the class number. The parameter α in Equ. 5 is set as 1.0. We use SGD with a momentum of 0.9 and weight decay of 0.0005. The initial learning rate is 0.001, and it decays by a cosine annealing strategy with 30 epochs. The input image size is 224×224, and the batch size is 64 per GPU.

6. Baseline experiments

As tasks of instance search do not include a training or fine-tuning step, it is crucial to choose the appropriate pre-trained model which is robust enough to be generalized to the instance search datasets. For the localization module, we take the open-sourced pre-trained model coming from GlobalTrack. Bag-of-Tricks [48] model is employed for the pre-training of the ranker.

We demonstrate the performance of several choices of localization methods and ranking networks in Table 2. The localization methods being tested include: (1) Edge box [64]; (2) YOLO v4 [65] General detection pretrained on COCO 80 classes; (3) SiamRPN [8] (4) GlobalTrack [47]. Two choices of the rankers are listed as well: (1) Bag-of-Tricks model pre-trained on person ReID Duke [51] dataset, denoted as "Ped-ReID"; (2) Bag-of-Tricks model as described in Sec. 3.2, denoted as "Any-ReID".

In addition, the localizer itself can be used to generate a ranked list, which can be treated as the instance search result without an extra ranker. The standalone localizer is the typical setup of a one-stage search framework and is listed for comparison as well in Table 2. Generally speaking, the localizer alone as a one-stage framework achieves a lower performance compared with its two-stage counterpart. By extracting feature vectors in a more dedicated way, the list from localization can be further re-ranked, and the overall accuracy is greatly improved.

In terms of the module selection in a two-stage framework, the general object detection performs poorly since the pre-defined 80 classes in COCO could not cover test cases. The series of Siamese networks perform better. However, SiamRPN fails to handle objects with large scale variations or significant location differences. On the other hand, GlobalTrack performs well and is a better choice for the localization module. For the choice of rankers, it is obvious that class-specific feature extraction is inappropriate, and a class-agnostic module is preferable.

Therefore, the following experiments on instance search will be conducted based on the best setting consisting of GlobalTrack and Any-ReID.

6.1. Instance search

6.1.1. Comparison on current INS datasets

We compare our pretrained models and self-trained methods on current instance search datasets. As shown in Table 3, our methods achieve state-of-the-art performance. Previous methods use general object detection or classification models while we use deep features strongly related to the query object. The self-training with more data further improves the performance.

6.1.2. Comparison on proposed INS datasets

Table 4 lists results comparison on the annotated datasets. As previous methods are proposed for simple background with salient objects, their performance can hardly be evaluated on the proposed dataset. State-of-the-art image retrieval method DELG [45] fails to produce precise localization. Siamese networks lack discrimination, and the performance is not satisfying. The proposed two-stage pretrained models together with self-training achieve satisfying performance.

Fig. 8 compares Precision-Recall curve on the proposed INS datasets. Our models achieve the highest recall of about 50% on INS-CUHK-SYSU while about 15% on INS-PRW. The lower recall than person search implies the difficulties of instance search. After self-training, the models improve in both precision and recall.

6.2. Ablation study

6.2.1. Number of query seeds

As the principle of cycle self-training is to leverage the large amount of unlabeled data, Table 5 demonstrates the performance of cycle self-training with different numbers of initial query seeds. More query seeds provide more variations of objects, thus making the model more robust after training. 1,000 query seeds produce moderate performance, while 30,000 query seeds achieve the best performance. Query seeds seem more critical for INS-CUHK-SYSU as the objects are larger and are more likely to be covered by initial query seeds. We use 30,000 query seeds in our experiments unless specified otherwise.

6.2.2. Self-training of localizer

If each query object is treated as a category on its own, the returned list of the same instance across multiple cameras can be regarded as a group of object detection results, which can be evaluated by Recall@k [24] and average precision. Note that these metrics only judge the localization capability by fixing the groundtruth gallery image. Since no other distractor gallery scene images are included, the performance is a lot better than the overall system.

The performance of different localizer self-training methods can be compared with the same rank list from the ranker. As discussed in Section 3.3.1, naive pseudo-pair generation methods cut off the rank list by a fixed length (top-K). From Table 6, it can be observed that $K = 2$ would make a low recall of pseudo labels, and $K = 20$ would make the pseudo labels less accurate. These naive settings both yield poor performance. Instead, the proposed probabilistic approach SGS makes a better tradeoff between precision and recall. The sampling strategy achieves the best performance without the need for specific parameters to generate pseudo labels robustly.

Table 2
The performance comparison of different choices of localizer and ranker. GlobalTrack + Any-ReID is the best setting.

Setting		INS-CUHK-SYSU		INS-PRW	
Localizer	ranker	rank-1	mAP	rank-1	mAP
Edge box	-	0.01	0.09	0.0	0.0
General Det.	-	0.03	0.2	0.0	0.01
SiamRPN	-	16.0	14.2	0.0	0.0
GlobalTrack	-	28.4	27.8	0.2	0.2
Edge box	Ped-ReID	7.9	7.0	1.6	0.4
General Det.	Ped-ReID	27.4	27.3	5.4	1.7
SiamRPN	Ped-ReID	8.0	5.3	1.0	0.1
GlobalTrack	Ped-ReID	35.0	32.8	5.0	2.8
Edge box	Any-ReID	9.5	8.3	2.5	0.5
General Det.	Any-ReID	38.4	37.4	5.7	1.6
SiamRPN	Any-ReID	23.0	19.6	3.2	0.5
GlobalTrack	Any-ReID	**43.1**	**42.1**	**18.7**	**8.5**

Table 3
Performance comparison on current instance search datasets. The pretrained two-stage framework and self-training outperform other methods.

Method	Instance-160	Instance-335
R-MAC [66]	35.8	37.5
CroW [67]	33.8	32.1
CAM [7]	35.8	34.7
BLCF [6]	65.3	48.3
BLCF-SalGAN [6]	65.6	46.9
CIS + XD [16]	72.4	59.3
DASR [15]	77.1	72.4
pretrained	79.6	75.3
self-training	**83.9**	**79.2**

6.2.3. Self-training of ranker

If all labeled instances are cropped into patches, they could be treated as a gallery set with multiple IDs. Thus retrieving cropped gallery images is exactly consistent with a typical ReID task. Therefore, rank-1 and mAP following the ReID convention can be used to evaluate the performance of class-agnostic rankers. Since localization is provided by groundtruth, the performance is a lot better than the overall system.

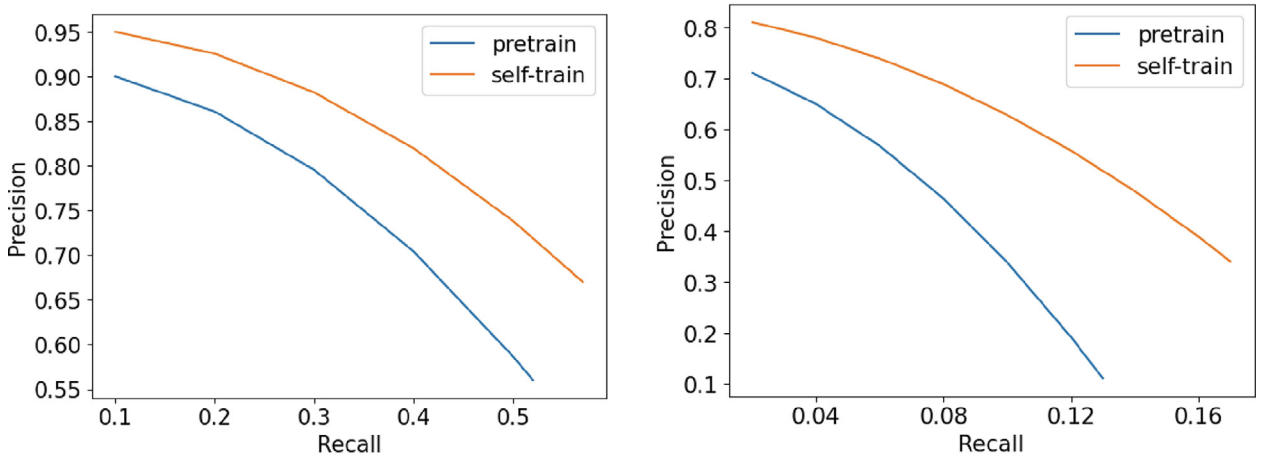
When fixing the localizer, the ranker can be self-trained following the algorithm described in Section 3.3.2. Table 7 lists the comparison between clustering-based methods [12–14] and our proposed solution. On the one hand, clustering-based self-training is very effective, but it is sensitive to the choice of *Ks*; on the other hand, the proposed MSC-SimCLR is not sensitive to noisy data with a large value of *K* and achieves the best performance.

6.2.4. Number of iterations of cycle self-training

Table 8 lists the results achieved by cycle self-training on INS-PRW without any supervised training. As can be seen, both the

Table 4
Performance comparison on annotated instance search datasets. The pretrained two-stage framework and self-training outperform other methods.

Method	INS-CUHK-SYSU		INS-PRW	
	rank-1	mAP	rank-1	mAP
DELG [45]	2.0	1.2	0.0	0.0
SiamRPN [8]	16.0	14.2	0.0	0.0
GlobalTrack [47]	28.4	27.8	0.2	0.2
pretrained	43.1	42.1	18.7	8.5
self-training	**49.4**	**47.4**	**24.2**	**13.4**



(a) INS-CUHK-SYSU PR curve (b) INS-PRW PR curve

Fig. 8. PR curve on the proposed datasets.

Table 5
Performance for different number of query seeds for self-training. More query seeds are beneficial for self-training.

Total number	Number per image	INS-CUHK-SYSU		INS-PRW	
		rank-1	mAP	rank-1	mAP
pretrained	pretrained	43.1	42.1	18.7	8.5
1000	0.004	45.8	43.9	19.1	8.9
10000	0.04	47.4	45.9	20.2	9.8
30000	0.12	49.4	47.4	24.2	13.4
50000	0.2	49.6	47.5	24.3	13.4

Table 6
The performance of self-training for localization. The proposed SGS outperforms naive sampling.

Setting	INS-CUHK-SYSU		INS-PRW	
	Recall@20	mAP	Recall@20	mAP
pretrained	91.4	74.7	76.6	45.5
TOPK = 2	91.7	75.6	74.6	43.7
TOPK = 5	91.8	74.5	78.5	46.5
TOPK = 20	90.1	68.3	71.6	30.4
SGS	92.3	76.7	79.3	47.7

Table 7
The performance of ranker self-training. The proposed MSC-SimCLR outperforms naive thresholding.

Setting	INS-CUHK-SYSU		INS-PRW	
	rank-1	mAP	rank-1	mAP
pretrained	68.5	71.5	37.4	24.0
TOPK = 2	69.9	72.6	40.2	27.7
TOPK = 5	70.3	73.0	39.9	28.2
TOPK = 20	69.9	72.8	39.6	27.5
MSC-SimCLR	71.5	74.6	41.2	28.8

Table 8
Performance evolution of cycle self-training for different iterations on INS-PRW. More iterations are beneficial for self-training.

Setting	Localization		Ranking		INS	
	Recall@20	mAP	rank-1	mAP	rank-1	mAP
pretrained	76.6	45.5	37.4	24.0	18.7	8.5
round 1	79.3	47.7	41.2	28.8	22.1	11.8
round 2	79.8	47.9	41.5	29.0	22.5	12.2
round 3	80.2	48.3	42.3	29.6	23.8	12.6
round 4	80.5	48.5	42.4	29.7	24.2	13.4

Table 9
Performance comparison on person search. * represents results on combined datasets (person + non-person). The proposed method generalize well to the person category.

Setting	Method	CUHK-SYSU		PRW	
		rank-1	mAP	rank-1	mAP
sup	OIM	78.7	75.5	49.9	21.3
sup	QEEPS	89.1	88.9	76.7	37.1
sup	BINet	92.4	91.5	81.7	45.3
sup	NAE+	92.9	92.1	81.1	44.0
sup	DC-I-Net	86.5	86.2	55.1	31.8
sup	our baseline	69.8	67.1	68.0	13.4
unsup	IUA	40.9	41.2	36.0	21.7
unsup	our self-train	**79.4**	**76.6**	**71.2**	**26.6**
unsup	our self-train*	**24.4**	**34.6**	**20.2**	**18.4**

Fig. 9. Qualitative evaluation for typical hard challenges. The proposed method outperforms others for different scenes, views, and occlusion due to precise localization and ranking. Best viewed in color.

Table 10
The performance of person-related and person-independent objects n INS-CUHK-SYSU.

Setting	rank-1	mAP
person-related	56.5	51.3
person-independent	44.1	45.6
overall	49.4	47.4

localizer and the ranker keep being improved after several rounds of updates. With the combination of the two modules, the overall performance improves gradually with the updates.

6.3. Model generalization

We then compare performance on person search, which is a special case of instance search. Table 9 compares different SOTA

methods [68,25,28,23] on two popular person search datasets, PRW and CUHK-SYSU. Our supervised baseline trained on general objects achieves promising performance. It even beats IUA [69] which was unsupervised trained on person search datasets. When self-trained on the much larger dataset D_{self}, the performance improves by a large margin compared with the baseline model. The self-training methods even beat supervised learning methods OIM [23] trained on person search datasets. Note that the self-training does not rely on any labels, which indicates its practical application.

Since the annotated data comes from person search datasets, we combine the original person data with our INS data. Thus the combined dataset contains not only the person category but also non-person categories. Current SOTA person search methods fail for arbitrary categories while our method still performs well. Besides, the proposed self-training method has no restriction for categories. It still performs moderately well on the combined dataset, which indicates good generalization capability.

6.4. Qualitative evaluation

To qualitatively analyze different methods and provide guidance for future research, we show the qualitative evaluation results on the proposed instance search datasets in Fig. 9. SiamRPN fails when the gallery ground truth is far from the query. DELG relies heavily on the initial global rank list and thus fails for all cases. General detection and edge box provide candidate proposals with no discriminative ability and thus fail for all cases. GlobalTrack searches throughout the whole scene images but fails for some tiny objects with similar distractors. On the other hand, the proposed method handles tiny objects successfully because of the discriminative ranker.

7. Discussions

7.1. person independent objects

Since some objects are related to the person, person context information may help the instance search of the objects. We thus separate INS-CUHK-SYSU into two subsets by the categories: person-related objects such as luggage and backpack, which counts about 30% of total data. The other data could be served as person-independent objects. Table 10 lists the performance of person-related and person-independent objects in INS-CUHK-SYSU. Person-related objects achieve higher accuracy. The reason could be the informative context clothing of humans. Nevertheless, person-independent objects achieve relatively high results, which again proves the effectiveness of the proposed method.

8. Conclusion

We focus on instance search for small objects in complex background, which is a very challenging problem. A two-stage strong baseline is introduced to precisely localize the searched objects. A cycle self-training framework is proposed to keep the generalization ability to deal with arbitrary classes, with self-training for a ranker-guided localizer and a localizer-guided ranker. The two modules in the pipeline can benefit each other during training without any human-labeled data. The instance search datasets provided in this paper are the first to search unconstrained objects in the wild across multiple cameras. They can be regarded as a valuable extension to tasks like face recognition, person ReID or person search in surveillance scenarios. We believe this framework and the dataset have the potential of being applied in various smart city applications.

CRediT authorship contribution statement

Yuqi Zhang: Conceptualization, Methodology, Software. **Chong Liu:** Formal analysis, Visualization. **Weihua Chen:** Methodology, Formal analysis, Visualization, Writing – original draft. **Xianzhe Xu:** Data curation, Formal analysis. **Fan Wang:** Conceptualization, Methodology, Writing – original draft, Supervision. **Hao Li:** Formal analysis, Resources. **Shiyu Hu:** Formal analysis, Visualization, Writing – original draft. **Xin Zhao:** Formal analysis, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is supported in part by the Youth Innovation Promotion Association CAS.

References

[1] G. Awad, W. Kraaij, P. Over, S. Satoh, Instance search retrospective with focus on trecvid, Int. J. Multimedia Inf. Retrieval 6 (1) (2017) 1–29.
[2] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[3] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), Comput. Vis. Image Understand. 110 (3) (2008) 346–359.
[4] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.
[5] E. Mohedano, K. McGuinness, N.E. O'Connor, A. Salvador, F. Marques, X. Giro-i Nieto, Bags of local convolutional features for scalable instance search, in: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, 2016, pp. 327–331..
[6] E. Mohedano, K. McGuinness, X. Giró-i Nieto, N.E. O'Connor, Saliency weighted convolutional features for instance search, in: 2018 international conference on content-based multimedia indexing (CBMI), IEEE, 2018, pp. 1–6..
[7] A. Jimenez, J.M. Alvarez, X. Giro-i Nieto, Class-weighted convolutional features for visual instance search, arXiv preprint arXiv:1707.02581..
[8] B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, High performance visual tracking with siamese region proposal network, CVPR (2018) 8971–8980.
[9] R. Tao, E. Gavves, A.W. Smeulders, Siamese instance search for tracking, CVPR (2016) 1420–1429.
[10] L. Zhang, Z. He, Y. Yang, L. Wang, X.-B. Gao, Tasks integrated networks: Joint detection and retrieval for image search, IEEE Trans. Pattern Anal. Machine Intell..
[11] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, W. Hu, Distractor-aware siamese networks for visual object tracking, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 101–117.
[12] X. Yan, I. Misra, A. Gupta, D. Ghadiyaram, D. Mahajan, Clusterfit: Improving generalization of visual representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6509–6518.
[13] Z. Zhong, L. Zheng, Z. Luo, S. Li, Y. Yang, Invariance matters: Exemplar memory for domain adaptive person re-identification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 598–607.
[14] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, X. Wang, Unsupervised domain adaptive re-identification: Theory and practice, Pattern Recogn. 102 (2020) 107173.
[15] H.-C. Xiao, W.-L. Zhao, J. Lin, C.-W. Ngo, Deeply activated salient region for instance search, arXiv preprint arXiv:2002.00185..
[16] Y. Zhan, W.-L. Zhao, Instance search via instance level segmentation and feature representation, arXiv preprint arXiv:1806.03576..
[17] A. Salvador, X. Giró-i Nieto, F. Marqués, S. Satoh, Faster r-cnn features for instance search, CVPR workshops (2016) 9–16.
[18] Y. Li, H. Qi, J. Dai, X. Ji, Y. Wei, Fully convolutional instance-aware semantic segmentation, CVPR (2017) 2359–2367.
[19] Y.-G. Hong, H.-C. Xiao, W.-L. Zhao, Towards accurate localization by instance search, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 3807–3815.
[20] Z. Guan, F. Xie, W. Zhao, X. Wang, L. Chen, W. Zhao, J. Peng, Tag-based weakly-supervised hashing for image retrieval, IJCAI (2018) 3776–3782.
[21] W. Zhao, Z. Guan, H. Luo, J. Peng, J. Fan, Deep multiple instance hashing for object-based image retrieval., IJCAI (2017) 3504–3510.
[22] W. Zhao, Z. Guan, H. Luo, J. Peng, J. Fan, Deep multiple instance hashing for fast multi-object image search, IEEE Trans. Image Process. 30 (2021) 7995–8007.

[23] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, in: CVPR, 2017, pp. 3415–3424..
[24] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person re-identification in the wild, CVPR (2017) 1367–1376.
[25] W. Dong, Z. Zhang, C. Song, T. Tan, Instance guided proposal network for person search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2585–2594.
[26] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, S. Yan, Neural person search machines, in: ICCV, 2017, pp. 493–501..
[27] X. Chang, P.-Y. Huang, Y.-D. Shen, X. Liang, Y. Yang, A.G. Hauptmann, Rcaa: Relational context-aware agents for person search, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 84–100..
[28] D. Chen, S. Zhang, J. Yang, B. Schiele, Norm-aware embedding for efficient person search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12615–12624.
[29] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned cnn embedding for person reidentification, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14 (1) (2017) 1–20.
[30] W. Chen, X. Chen, J. Zhang, K. Huang, A multi-task deep network for person re-identification, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 3988–3994.
[31] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: ECCV, 2018, pp. 480–496..
[32] Y. Suh, J. Wang, S. Tang, T. Mei, K. Mu Lee, Part-aligned bilinear representations for person re-identification, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 402–419..
[33] Y. Zhang, Y. Huang, S. Yu, L. Wang, Cross-view gait recognition by discriminative feature learning, IEEE Trans. Image Process. 29 (2019) 1001–1015.
[34] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, IEEE, 2006, pp. 1735–1742.
[35] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: CVPR, 2015, pp. 815–823..
[36] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, J. Gu, A strong baseline and batch normalization neck for deep person re-identification, IEEE Trans. Multimedia..
[37] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, AAAI (2020) 13001–13008.
[38] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, CVPR (2017) 1318–1327.
[39] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: 2007 IEEE conference on computer vision and pattern recognition, IEEE, 2007, pp. 1–8..
[40] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Ziserman, Lost in quantization: Improving particular object retrieval in large scale image databases, in: 2008 IEEE conference on computer vision and pattern recognition, IEEE, 2008, pp. 1–8..
[41] H. Noh, A. Araujo, J. Sim, T. Weyand, B. Han, Large-scale image retrieval with attentive deep local features, in: ICCV, 2017, pp. 3456–3465..
[42] H. Oh Song, Y. Xiang, S. Jegelka, S. Savarese, Deep metric learning via lifted structured feature embedding, CVPR (2016) 4004–4012.
[43] F. Cakir, K. He, X. Xia, B. Kulis, S. Sclaroff, Deep metric learning to rank, CVPR (2019) 1861–1870.
[44] A. Brown, W. Xie, V. Kalogeiton, A. Zisserman, Smooth-ap: Smoothing the path towards large-scale image retrieval, in: European Conference on Computer Vision, Springer, 2020, pp. 677–694..
[45] B. Cao, A. Araujo, J. Sim, Unifying deep local and global features for image search, European Conference on Computer Vision, Springer (2020) 726–743.
[46] P.-E. Sarlin, C. Cadena, R. Siegwart, M. Dymczyk, From coarse to fine: Robust hierarchical localization at large scale, CVPR (2019) 12716–12725.
[47] L. Huang, X. Zhao, K. Huang, Globaltrack: A simple and strong baseline for long-term tracking, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 11037–11044..
[48] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, CVPR Workshops (2019).
[49] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.
[50] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: ICCV, 2015, pp. 1116–1124..
[51] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, European Conference on Computer Vision, Springer (2016) 17–35.
[52] S. Wang, S. Jiang, Instre: a new benchmark for instance-level object retrieval and recognition, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 11 (3) (2015) 1–21.
[53] Y. Ge, R. Zhang, X. Wang, X. Tang, P. Luo, Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5337–5345.
[54] L. Huang, X. Zhao, K. Huang, Got-10k: A large high-diversity benchmark for generic object tracking in the wild, IEEE Trans. Pattern Anal. Mach. Intell. (2019), 1–1.
[55] H. Fan, H. Bai, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, M. Huang, J. Liu, Y. Xu, et al., Lasot: A high-quality large-scale single object tracking benchmark, Int. J. Comput. Vision 129 (2) (2021) 439–461.
[56] E. Real, J. Shlens, S. Mazzocchi, X. Pan, V. Vanhoucke, Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video, CVPR (2017) 5296–5305.
[57] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV) 115 (3) (2015) 211–252.
[58] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, K. Schindler, Motchallenge 2015: Towards a benchmark for multi-target tracking, arXiv preprint arXiv:1504.01942..
[59] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, K. Schindler, Mot16: A benchmark for multi-object tracking, arXiv preprint arXiv:1603.00831..
[60] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755..
[61] S. Zhang, Y. Xie, J. Wan, H. Xia, S.Z. Li, G. Guo, Widerperson: A diverse dataset for dense pedestrian detection in the wild, IEEE Trans. Multimedia 22 (2) (2019) 380–393.
[62] L. Wen, D. Du, Z. Cai, Z. Lei, M. Chang, H. Qi, J. Lim, M. Yang, S. Lyu, Detrac: A new benchmark and protocol for multi-object tracking, arXiv preprint arXiv:1511.04136 2 (4) (2015) 7..
[63] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969..
[64] C.L. Zitnick, P. Dollár, Edge boxes: Locating object proposals from edges, in: ECCV, Springer, 2014, pp. 391–405..
[65] A. Bochkovskiy, C.Y. Wang, H.-Y.M. Liao, Yolov4: Optimal speed and accuracy of object detection, arXiv preprint arXiv:2004.10934..
[66] G. Tolias, R. Sicre, H. Jégou, Particular object retrieval with integral max-pooling of cnn activations, arXiv preprint arXiv:1511.05879..
[67] Y. Kalantidis, C. Mellina, S. Osindero, Cross-dimensional weighting for aggregated deep convolutional features, in: European conference on computer vision, Springer, 2016, pp. 685–701.
[68] B. Munjal, S. Amin, F. Tombari, F. Galasso, Query-guided end-to-end person search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 811–820.
[69] S. Cao, Y. Liu, An iterative unsupervised person search algorithm on natural scene images, in: 2019 Chinese Automation Congress (CAC), IEEE, 2019, pp. 3779–3783..

Yuqi Zhang received his BSc degree in Harbin Institute of Technology (HIT) in 2014 and PhD degree at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China in June 2019. His research interests include person re-identification, image retrieval and gait recognition.

Chong Liu is currently pursuing the Ph.D. degree in computer science with the State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China, under the supervision of Prof. Yi-Dong Shen. His current research interests include machine learning, few-shot learning, cross-modal retrieval, and person reidentification.

Weihua Chen received the Ph.D. degree from the Institute of automation, Chinese Academy of Sciences (CASIA) in 2018. His research areas include computer vision and deep learning, particularly object tracking and person re-identification. He has authored/co-authored over 10 papers, including CVPR, ICCV, AAAI and etc. He received the championships of many challenges, such as Visual Domain Adaptation Challenge in ECCV2020, AI City Challenge in CVPR2021 and Multi-camera Multiple People Tracking in ICCV2021.

Xianzhe Xu received master degree at school of electrical and information engineering, Tianjin University in July 2021. His research interests include person re-identification, image retrieval and image segmentation.

Fan Wang received the B.S. and M.S. degree from Department of Automation, Tsinghua University, Beijing, China, and the Ph.D. degree from Department of Electrical Engineering, Stanford University, California, United States. She is currently with Alibaba Group as a Senior Staff Algorithm Engineer. Her research interests include object tracking and recognition, 3D vision and multi-sensor fusion.

Hao Li holds a PhD from Chinese Academy of Sciences and is in charge of real-scene visual understanding technologies. His research fields include smart interpretation of remote sensing images, X-ray object identification, facial recognition-based clocking in system, new retail, and smart campuses. Related technologies include deep learning model compression, facial recognition, person re-identification, and image search. He Published more than 20 papers and owns more than 20 licensed patents.

Shiyu Hu received her BSc from Beijing Institute of Technology (BIT). and MSc from the University of Hong Kong (HKU). In September 2019, she joined the Institute of Automation of the Chinese Academy of Sciences (CASIA), where she is currently studying for his doctorate. Her current research interests include pattern recognition, computer vision and machine learning.

Xin Zhao received the Ph.D. degree from the University of Science and Technology of China (USTC). He is currently an Associate Professor in the Institute of Automation, Chinese Academy of Sciences (CASIA). His current research interests include pattern recognition, computer vision, and machine learning. He has published the international journal and conference papers, such as the IEEE TPAMI, IEEE TIP, IEEE TCSVT, CVPR, ICCV, AAAI, IJCAI. He received the International Association of Pattern Recognition Best Student Paper Award at ACPR 2011. He received the 2nd place entry of COCO Panoptic Challenge at ECCV 2018.