Cognitive Computation
https://doi.org/10.1007/s12559-021-09898-7

3W-AlignNet: a Feature Alignment Framework for Person Search with Three-Way Decision Theory

Yuting Yang[1] · Duoqian Miao[1] · Hongyun Zhang[2]

Received: 18 November 2020 / Accepted: 15 June 2021

Abstract
Person search aims to locate and recognize a specified person from a gallery of uncropped scene images, which combines pedestrian detection and person re-identification (re-ID). Existing methods based on Faster R-CNN have been widely used to tackle the two sub-tasks jointly, but they ignore the feature misalignment problem, i.e., re-ID feature localization is not fully aligned with the detected bounding boxes (BBoxes). Due to the fine-grained property of re-ID, it is crucial to extract accurate appearance features. In addition, the granularity of BBoxes detected from gallery images is quite different, and it is defective to treat gallery boxes with different granularity as equal in estimating their similarities with the query. Three-way decision methods are fields of research on human-inspired computation. Inspired by them, we propose a three-way-based feature alignment framework (3W-AlignNet) to optimize the re-ID feature localization. The framework is implemented by iteratively generating new BBoxes and features from previous BBoxes. The three-way decision theory is applied to avoid the mismatch problem caused by increasing Intersection over Union (IoU). We further propose a Granularity Weighted Similarity (GWS) algorithm to relieve the granularity mismatch problem. Extensive experiments show that our method outperforms all other state-of-the-art end-to-end methods on two widely used person search datasets, CUHK-SYSU and PRW.

Keywords Person search · Person re-identification · Three-way decision · Multi-granularity

Introduction

Person search [1] task aims to identify a specified person in a gallery of scene images. It combines pedestrian detection [2–4] and person re-identification (re-ID) [5–7], i.e., generating bounding boxes (BBoxes) from gallery images and matching the gallery boxes with the query. Person search compensates for the functional limitations of re-ID, and it is more suitable for practical applications, such as video surveillance systems, and people finder systems. Due to the complex changes in human posture, scene lighting, occlusion, and background clutter, etc., it has attracted more and more attention in recent years.

Existing methods divide person search into two sub-tasks and solve them jointly or separately. For two-stage methods [8–10], they train the detector and the re-ID model independently. The detector crops candidate people from gallery images and the re-ID model identifies the query person from candidates. In contrast, one-stage methods [1, 11, 12] apply multi-task frameworks based on Faster R-CNN [13] to solve detection and re-ID jointly. Although the multi-task framework is efficient and easy to train, it suffers from the feature misalignment problem, i.e., feature localization in the re-ID branch is not fully aligned with the detected BBoxes. Specifically, the detection branch predicts the regressor to revise region proposals into BBoxes, but the re-ID branch extracts pedestrian features before the regressor is applied. Therefore, the network could not convey accurate regression information to the re-ID branch. Due to the fine-grained property of re-ID, such misaligned features will lead to significantly unsatisfactory performance, especially when the proposals are far away from the BBoxes.

✉ Duoqian Miao
dqmiao@tongji.edu.cn

Yuting Yang
yytshirley@tongji.edu.cn

[1] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

[2] Key Laboratory of Embedded System and Service Computing Ministry of Education, Tongji University, Shanghai 201804, China

Published online: 06 July 2021

Motivated by the above observations, we propose an Iterative Alignment Strategy (IAS) at the inference phase. Our main purpose is to reduce the regressor between proposals and BBoxes so as to extract aligned features. IAS is implemented by a cascade iteration strategy, i.e., we apply BBoxes generated in the first iteration as region proposals to generate new BBoxes and features. As shown in Fig. 1, when applying the second iteration, new proposals (BBoxes generated in the first iteration) are very close to new BBoxes. Hence the feature misalignment is tiny enough to be negligible.

Moreover, we notice that the performance of our IAS is limited by the decrease of detection quality caused by Intersection over Union (IoU) mismatch (mentioned in [14]), i.e., the network is trained with 0.5 IoU threshold to distinguish negatives/positives, which performs best for samples of IoU close to the threshold during testing. Namely, the network fails to handle samples with much higher IoU than the threshold in the second iteration. To address this issue, we propose a three-way based feature alignment framework (3W-AlignNet). The thinking model of three-way cognitive computations [15] is developed from rough sets theory [16]. Yao [17] extends it to a much broader frontier, which outlines a unified theory of three-way decision [17]. In recent years, the methodology is widely used in many theoretical and practical fields, such as fuzzy sets theory [18], shadowed sets theory [19], face recognition [20], and sentiment classification [21]. Three-way decision explores thinking, problem-solving, and information processing in threes, namely sets of three parts or items. Here the three parts correspond to positive, negative and boundary, respectively. We firstly divide BBoxes into the above three parts according to their detection confidence. We then retain positive BBoxes and remove negative ones. For BBoxes in the boundary domain, we implement similar iterative operations as IAS on them. Afterward, we combine new BBoxes with retained BBoxes to achieve better performance. In general, our 3W-AlignNet gives joint consideration to detection quality and feature localization.

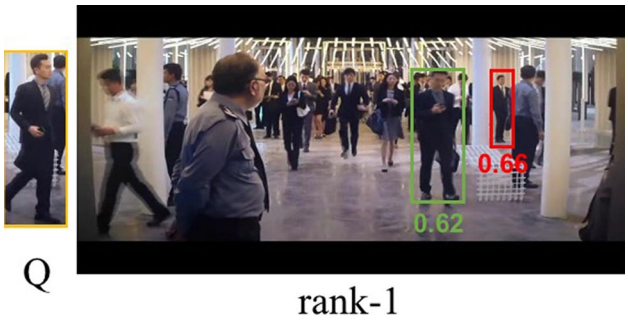

Fig. 2 Example of false match. The notion Q denotes the query person. The man in green box is the correct result to be matched and the man in red box is the false match

Another challenge for person search is the granularity mismatch problem. We refer to the resolution of the BBox as the granularity of it. The higher the resolution, the finer the granularity, and vice versa. Given a query image Q and a gallery G, the granularity of BBoxes detected from G is quite different. Compared with BBoxes with close granularity to Q, those with much coarser granularity than Q tend to be easier to get higher similarity scores with Q. Figure 2 shows an example of a false match. Given the man in the yellow box as the query, the man in the green box is the correct result to be matched, but the network misidentifies the man in the red box as the query. We can see that people in coarse-grained BBoxes tend to be blurred and their appearance features are not discriminative enough to be distinguished from others. To relieve this issue, we propose a multi-granularity BBox reweighting algorithm, named Granularity Weighted

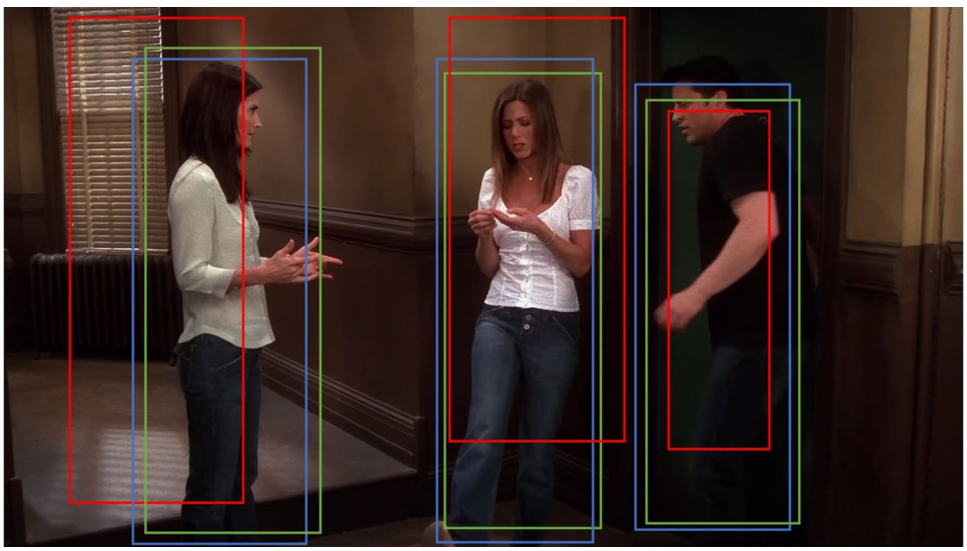

Fig. 1 The illustration of IAS. The red boxes denote region proposals generated by RPN in the first iteration. The green boxes represent BBoxes generated in the first iteration and the blue boxes represent BBoxes generated in the second iteration

Similarity (GWS), which suppresses BBoxes with coarse granularity by incorporating granularity difference into similarity measurement.

The main contributions of our works are as follows:

- We propose a three-way based feature alignment framework (3W-AlignNet) for person search to relieve the feature misalignment problem, which provides a good idea for the combination of three-way decision theory and deep learning methods.
- We propose a multi-granularity BBox reweighting algorithm (GWS) to relieve the granularity mismatch problem, which incorporates granularity difference into similarity measurement.
- Our method is computationally light and outperforms all state-of-the-art end-to-end methods on CUHK-SYSU and PRW datasets.

The remainder of this paper is organized as follows. "Related Work" provides a survey of the literature, "Methodology" describes the proposed method, "Experiments and Results" reports on the experiments and results, and "Conclusions" makes some conclusions.

Related Work

Person Search Person search aims to locate and recognize people from unconstrained images. Existing methods divide this task into two sub-tasks and solve them jointly or separately. For two-stage methods, Chen et al. [8] put forward the contradictory objective problem and suggest to train two detection and re-ID models. Lan et al. [9] propose a Cross-Level Semantic Alignment (CLSA) to tackle the multi-scale matching problem. Wang et al. [10] design an Identity-Guided Query (IDGQ) detector and a Detection Results Adapted (DRA) re-ID model separately to eliminate the inconsistencies between the two sub-tasks.

In contrast to the above two-stage methods, other works train the detection and re-ID model in an end-to-end manner. For example, Xiao et al. [1] prove that pedestrian detection and person re-ID can be solved in an end-to-end framework. They apply Faster R-CNN as the backbone network and employ a parallel re-ID branch with the detection branch. Then an online instance matching (OIM) loss function is proposed for re-ID. Xiao et al. [11] enhance the discrimination of features by applying center loss. Yan et al. [12] build a graph learning framework and employ context information for person search. Liu et al. [22] and Munjal et al. [23] propose query-guided networks to leverage the query image extensively. Dong et al. [24] propose a Siamese network with an additional instance-aware branch to alleviate the negative effects of context information. Chen et al. [25] analyze the contradictory goals of two sub-tasks and solve the problem by the Norm-Aware Embedding (NAE) model, which disentangles the person embedding into norm and angle for detection and re-ID, respectively.

Three-Way Decision In this paper, we continue to use the framework of end-to-end methods. Based on the NAE [25] model, we propose IAS to alleviate the feature misalignment problem. The strategy largely improves the performance of re-ID but is still limited by the IoU mismatch problem. Three-way decision theory [15], which offers new theories, models, and tools for cognitive analytics, provides a smart alternative to solve it. The methodology of three-way decision is widely applied in many fields, including traditional machine learning fields and deep learning fields. Li et al. [20] introduce three-way decision to face recognition and propose a sequential three-way decision and granulation for cost-sensitive face recognition. Zhang et al. [21] propose a three-way enhanced network for sentiment classification, which successfully combines traditional models with deep learning methods. Chen et al. [26] propose a graph-based keyphrase extraction model with three-way decision. The above works enrich the theoretical foundation of three-way decision and indicate that the thinking model of three-way cognitive computations is applicable for many practical decision problems. Our method is inspired and guided by three-way decision theory and relevant applications, and we hope it will illuminate other fields related to cognitive computation.

Methodology

In this section, we first present an overview of the whole structure of our baseline. Then we describe the proposed IAS, which aims to alleviate the feature misalignment problem. Subsequently, we describe our 3W-AlignNet to relieve the IoU mismatch problem. Finally, a useful GWS algorithm is introduced.

End-to-End Framework for Person Search

As aforementioned, we aim to tackle two sub-tasks together in a multi-task end-to-end framework. We take the Norm-Aware Embedding (NAE) [25] network as our baseline. Figure 3 shows the structure of NAE, in which a multi-task head for detection and re-ID is built upon the top of Faster R-CNN.

We adopt ResNet50 [27] as the backbone. Given an image as input, it first passes through the first part (res1-res4) of ResNet50 to extract basic features. Then a Region Proposal Network (RPN) [13] is applied on these feature maps to generate proposals of candidate people. After

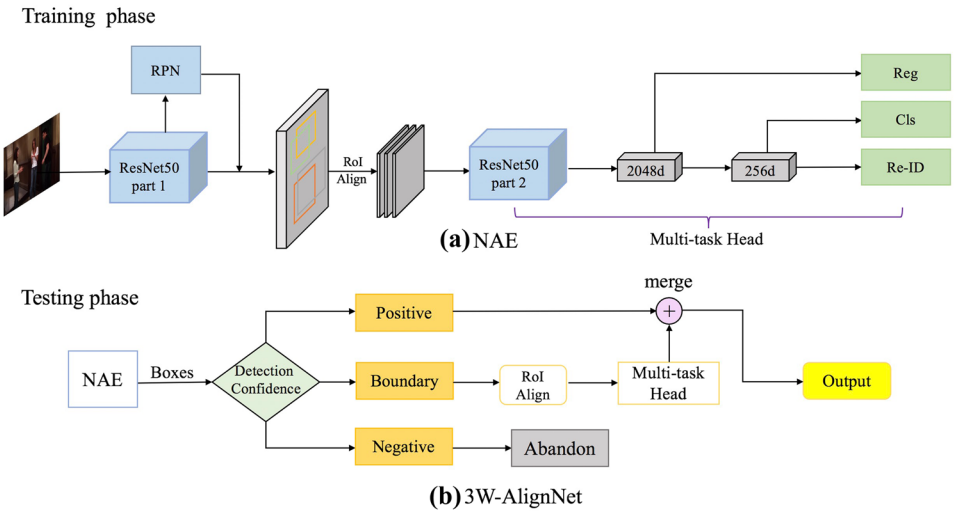

Fig. 3 The illustration of our model. (**a**) The NAE-base. (**b**) The structure of our 3W-AlignNet

non-maximum suppression (NMS) [13], we keep 128 proposals and exploit RoI-Align [28] to pool a $1024 \times 14 \times 14$ region from the stem feature maps for each proposal. These pooled proposals are fed into the second part (res5) of ResNet50 to extract 2048-dimensional 7×7 feature maps. The 2048-dimensional features are used to calculate regressor and then projected to 256-dimensional features, which are decomposed to radial norm r and angle θ in the polar coordinate system for classification and re-ID, respectively. Following the previous works, the Online Instance Matching (OIM) loss [1] is applied to optimize the extracted pedestrian features for re-ID.

During training, the overall feature learning loss function is given as:

$$L = \lambda_1 L_{reg_1} + \lambda_2 L_{cls_1} + \lambda_3 L_{reg_2} + \lambda_4 L_{cls_2} + \lambda_5 L_{oim} \tag{1}$$

where the L_{reg_1} / L_{reg_2} stands for the regression loss used in RPN / R-CNN [29], and the L_{cls_1} / L_{cls_2} stands for the classification loss used in RPN / R-CNN. L_{oim} stands for the OIM loss. λ_3 is set to 10 and others are 1. At inference, BBoxes detected from gallery images are ranked according to the similarities between the query and gallery.

Iterative Alignment Strategy

As Introduction mentioned, re-ID is a fine-grained task, but the baseline suffers from the feature misalignment problem. In this section, we will describe our IAS in detail. A region proposal $p = (p_x, p_y, p_w, p_h)$ contains the four coordinates of an image patch x. In NAE, the task of regression is to regress the region proposal p into a target BBox $b = (b_x, b_y, b_w, b_h)$, using a regressor. The regressor $\Delta = \{\delta x, \delta y, \delta w, \delta h\}$ is defined as follows.

$$\begin{aligned} \delta x = (b_x - p_x)/p_w, \delta y = (b_y - p_y)/p_h \\ \delta w = log(b_w/p_w), \delta h = log(b_h/p_h) \end{aligned} \tag{2}$$

In our baseline, due to the large regressor predicted by the network, the extracted re-ID features are not fully aligned with detected BBoxes. To reduce the regressor, a straightforward way is to make proposals much closer to the ground truth. IAS is described in Algorithm 1. Given an image, we firstly input it into NAE to predict preliminary BBoxes ($boxes_1$) and features ($feats_1$). Then we freeze RPN and take $boxes_1$ as new region proposals ($proposals_2$). RoI-Align is used to pool $proposals_2$ into fixed regions, which are then fed into the multi-task head to generate new BBoxes ($boxes_2$) and features ($feats_2$).

The $boxes_1$ are usually very close to the ground truth. As a result, in the second iteration, the regressor between $proposals_2$ and $boxes_2$ is much smaller than the first iteration. So that the feature misalignment problem would be tiny enough to be negligible.

Algorithm 1 Iterative Alignment Strategy (IAS)

Input:
Gallery image, G
Output:
BBoxes and re-ID features
1: Input G into the NAE network (get $boxes_1$ and $feats_1$)
2: Freeze RPN and take $boxes_1$ as new proposals (get $proposals_2$)
3: For each one of $proposals_2$, exploit RoI-Align to pool it into a fixed region
4: Feed these pooled regions into the multi-task head of NAE again (get $boxes_2$ and $feats_2$)
5: **return** $boxes_2$, $feats_2$

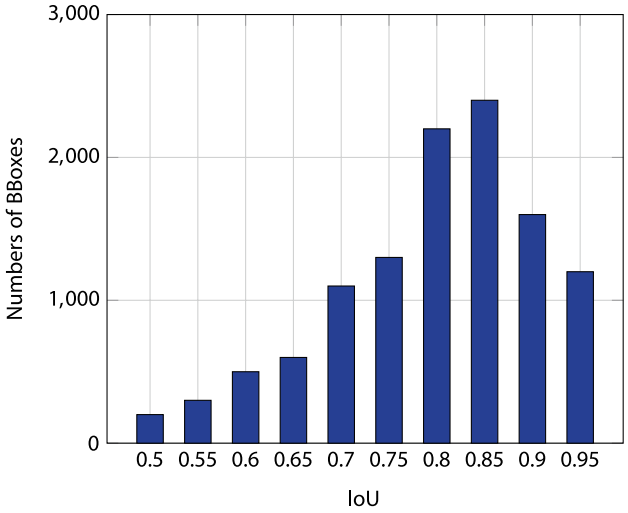

Fig. 4 The IoU statistics of BBoxes of labeled people detected by the NAE network

Three-Way Based Feature Alignment Framework

Although IAS can effectively relieve the feature misalignment problem, the re-ID performance will be inevitably limited by the decrease of detection quality, which is caused by the IoU mismatch. Figure 4 shows the IoU distribution of $boxes_1$ generated by NAE in the first iteration. The NAE network is trained with 0.5 IoU threshold, but there are a lot of BBoxes with IoU > 0.85 after the first iteration. As mentioned in [14] and shown in Fig. 5, the network optimized at a single IoU level is not necessarily optimal at other levels. And it performs best for samples with IoU close to the threshold that the network is trained. In the second iteration, the IoU of proposals is generally higher than the first time. For those samples with much higher IoU (> 0.85) than the training threshold, the detection ability of the network is not powerful enough to handle them.

To alleviate the IoU mismatch problem, we introduce the framework of three-way decision to Algorithm 1. The structure of our 3W-AlignNet is shown in Fig. 3. Firstly, the NAE network takes an image as input and produces several $bboxes_1$ (see Algorithm 2) as candidates in the first iteration. Then, in order to divide $bboxes_1$ into three parts, it is crucial to choose a proper partition criterion. The detection branch predicts a detection confidence score (*det_score*) for each BBox, which tends to reflect the IoU with the ground truth. Empirically, the larger the *det_score*, the higher the IoU. Therefore, we adopt *det_score* as the partition criteria and predict it by the softmax function, which is used as the last activation function of the NAE baseline to normalize the output of the network to a probability distribution over predicted output classes.

The function takes as input a vector $z = (z_1, ..., z_K)$ of K real numbers, and normalizes it into a probability distribution consisting of K probabilities proportional to the exponentials of the input numbers. The function is defined as follows:

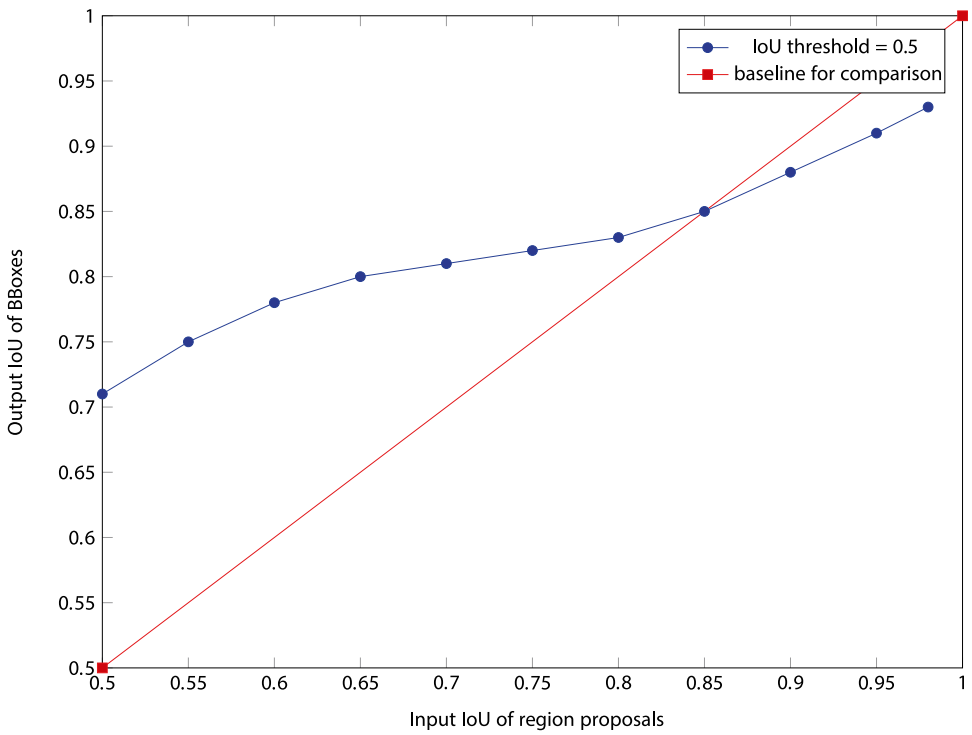

Fig. 5 The output IoU of BBoxes compared with input IoU in the NAE network

$$\sigma(z)_i = \frac{exp(z_i)}{\sum_{j=1}^{K} exp(z_j)}, \ for \ i = 1, ...K \tag{3}$$

where $K = 2$ represents two categories, background and foreground, respectively.

Specifically, we set two thresholds C_1 and C_2 ($0 < C_1 < C_2 < 1$), and divide $bboxes_1$ into the following three parts : positive (foreground BBoxes with higher det_score than C_2), negative (background BBoxes with lower det_score than C_1) and boundary (BBoxes with medium det_score between C_1 and C_2). Then we retain positive $bboxes_1$ and $feats_1$, and remove negative ones. Afterward, similar method as Algorithm 1 is applied to the remaining boundary instances, i.e., we freeze RPN and take $boxes_1$ in boundary domain as new region proposals ($proposals_2$). For each one of $proposals_2$, we exploit RoI-Align to pool it into a fixed region. These pooled regions are then fed into the multi-task head of NAE again to generate $boxes_2$ and $feats_2$. The retained $bboxes_1$ and iteratively generated $bboxes_2$ are finally merged, and their features are applied to the similarity calculation. Our method is described in Algorithm 2. Due to the detection ability of the network (shown in Fig. 5), the regressor between positive $bboxes_1$ and their proposals is relatively small. Retaining samples in the positive domain has few negative effects on feature localization but effectively avoids the IoU mismatch problem. To summarize, our 3W-AlignNet achieves a good balance between detection quality and better features.

Algorithm 2 Three-way Feature Alignment Framework

Input:
Gallery image, G
Two confidence thresholds, C_1 and C_2
Output:
Detected BBoxes and re-ID features
Input G into the NAE network (get $boxes_1$, det_scores_1 and $feats_1$)
Divide $boxes_1$ into three parts according to det_scores_1
if $det\_scores_1 \geqslant C_2$ **then**
  Retain $boxes_1$ and $feats_1$
else if $C_1 < det\_scores_1 < C_2$ **then**
  Freeze RPN and take $boxes_1$ as new proposals (get $proposals_2$)
  Exploit RoI-Align to pool $proposals_2$ into fixed regions
  Feed these pooled regions into the multi-task head of NAE again (get $boxes_2$ and $feats_2$)
else
  Abandan $boxes_1$ and $feats_1$
end if
Merge retained $boxes_1$ and new $boxes_2$ into $boxes$
Merge retained $feats_1$ and new $feats_2$ into $feats$
return $boxes$ and $feats$

We are not the first to apply iterative operations on top of the Faster R-CNN, such as [14, 30]. But our method significantly differs from them in the following aspects.

- The above two methods [14, 30] are proposed in the field of object detection, and they aim at achieving better detection performance. However, our method aims at alleviating the feature misalignment problem in person search and illuminates that person re-ID is very sensitive to the feature localization.
- The above two methods [14, 30] need to process all BBoxes in each iteration, while we just apply BBoxes in the boundary domain to the second iteration, which ensures the efficiency of the network.
- Compared with [14], our second head shares the same parameters with the original network, and no extra training is required.

BBoxes Reweighting Algorithm

In this section, we propose a novel GWS to relieve the granularity mismatch problem. Given a query person, gallery BBoxes are ranked according to their similarities with the query. Previous works treat gallery boxes with different granularity as equal in estimating their similarities with the query, which is defective as mentioned in the Introduction.

We denote $area(b) = b_w * b_h$ as the area of a BBox $b = (b_x, b_y, b_w, b_h)$. Given a query BBox q and a gallery BBox g, the similarity calculation between q and g is defined as follows:

$$sim(q, g) = \frac{\sum_{i=1}^{n} x_{q_i} x_{g_i}}{|X_q| \cdot |X_g|} \tag{4}$$

where $X_q = \{x_{q_1}, x_{q_2}...x_{q_n}\}$ and $X_g = \{x_{g_1}, x_{g_2}...x_{g_n}\}$ represent the features of q and g respectively.

We define $d(q, g)$ to measure the granularity difference between q and g as follows:

$$d(q, g) = \begin{cases} \frac{area(g)}{area(q)}, & \text{if } area(g) < area(q), \\ 1, & \text{otherwise.} \end{cases} \tag{5}$$

The closer the value of d is to 1, the smaller the granularity difference is.

We design a reweighting function to assign different weights to samples with different granularity. The main idea is to suppress the weights of BBoxes with coarse granularity and increase the weights of those with fine granularity.

The linear decay reweighting function $w(q, g)$ is defined as follows:

$$w(q, g) = \begin{cases} 1 - \frac{(k_1 - d(q,g)) * (1 - k_2)}{k_1}, & \text{if } d(q, g) < k_1, \\ 1, & \text{otherwise.} \end{cases} \tag{6}$$

Two thresholds k_1 and k_2 are defined. As Fig. 6 shows, k_1 indicates the value of the abscissa $d(q, g)$ where $w(q, g)$ begins to decrease. k_2 represents the value of the ordinate

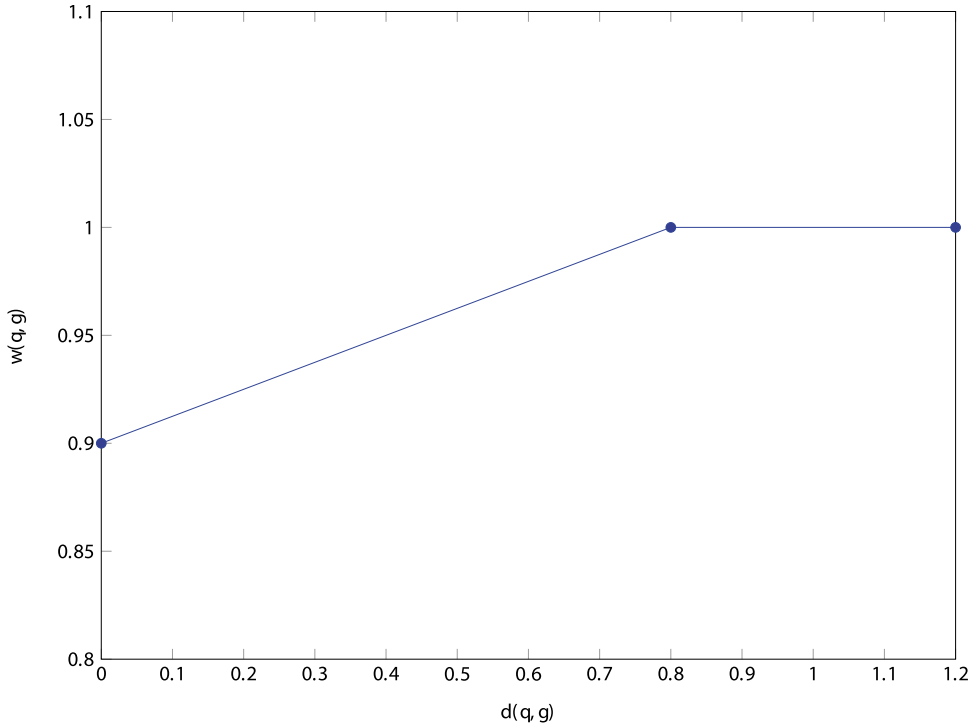

Fig. 6 The illustration of the reweighting function $w(q, g)$

$w(q, g)$ where $d(q, g)$ is equal to 0. If $d(q, g) \geqslant k_1$, $w(q, g)$ is set to 1 and remains unchanged; If $d(q, g) < k_1$, $w(q, g)$ is gradually reduced according to the linear decay factor.

The overall GWS is then defined as:

$$GWS(q, g) = sim(q, g) * w(q, g) \tag{7}$$

GWS effectively reduces the negative influence of the granularity mismatch problem and helps to improve the accuracy of the final results.

Experiments and Results

Datasets

CUHK-SYSU [1] is a large-scale person search dataset captured by a moving camera in the street/urban scene or chosen from the movie snapshots. A total of 18,184 scene images and 96,143 BBoxes with annotations are collected. Each labeled person has a specific Person-ID and appears in at least two different scene images from different angles. Those unlabeled people are marked as unknown identities. The training set contains 11,206 scene images and 5,532 identities, while the testing set contains 6,978 gallery images and 2,900 query people. In the testing set, gallery size is between 50 and 4,000 for each query person. We set gallery size as 100 in all experiments by default.

PRW [31] consists of 11,816 video frames extracted from videos taken at different locations on a university campus. It contains 932 labeled people and 34,304 labeled BBoxes. And annotations are divided into labeled identities and unlabeled identities. The training set includes 5, 704 frames and 482 identities, while the testing set contains 6, 112 gallery images and 2, 057 query images with 450 identities. The size of the gallery is significantly larger than the default setting of CUHK-SYSU.

Evaluation Protocol

We adopt the Cumulative Matching Characteristic (CMC) [25] and the mean Averaged Precision (mAP) [25] as the

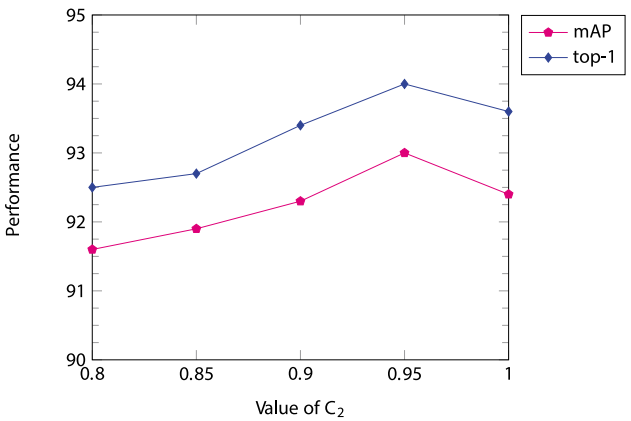

Fig. 7 Influence of different values of C_2 on the performance on CUHK-SYSU

performance metrics, which are widely used in person re-ID and object detection tasks. We calculate an averaged precision (AP) by computing the area under the Precision-Recall curve for each query and then average the APs across all the queries to obtain the final mAP.

Implementation Details

We use PyTorch to implement our model and run experiments on one Tesla V100 GPU. We adopt ResNet50 pretrained on ImageNet [32] as our backbone network. Then a standard RPN is added on top of the first four residual blocks (res1-res4) to generate region proposals. The anchor scales and ratios in RPN are set as (32, 64, 128, 256, 512) and (0.5, 1.0, 2.0) respectively. Afterward, the proposals are reshaped to 14×14 by the RoI-Align layer and passed to the last residual block (res5). During training, we randomly sample five images in each mini-batch and scale them to 900×1500 pixels. Stochastic Gradient Descent (SGD) is used to optimize the model. The momentum of SGD is set to 0.9, and the weight decay is set to 0.0005. The model is trained for 22 epochs, and the initial learning rate is 0.003 (warm-up in the first epoch), dropped to 0.0003 at the 16th epoch.

Ablation Study

In this section, we firstly perform several analytical experiments on CUHK-SYSU and PRW datasets to explore the contribution of each component in our proposed method, which is shown in Table 1. Then we do more experiments on CUHK-SYSU to further explain the superiority of our method.

Aligned Features are Important for Re-ID Table 2 shows the detection and re-ID results of the NAE baseline, Iterative Box [30], and ours. Iterative Box is proposed in object detection, and it is a similar testing phase enhancement method like IAS.

- **IAS vs. NAE.** Our IAS achieves 91.0% and 85.7% w.r.t. Recall and AP, inferior to NAE-base by 1.6 and 1.1 pp. respectively. But it achieves better re-ID performance, 92.4% and 93.6% w.r.t. mAP and top-1, outperforming NAE-base by 0.9 and 1.2 pp. Figure 8 shows the regressor $\Delta = \{\delta x, \delta y, \delta w, \delta h\}$ distribution before and after applying our IAS. We can see that IAS largely reduces the regressor Δ to extract more aligned re-ID features. This is consistent with our viewpoint that aligned features play an important positive role in the re-ID task, even though there is a decline in the detection quality.
- **IAS vs. Iterative Box.** In Table 2, Iterative Box gains a slight increase in re-ID performance, which is brought by better detection quality. Although our IAS achieves inferior detection quality, it achieves better re-ID performance, outperforming Iterative Box by 0.6% and 0.6% w.r.t. mAP and top-1. The experiment result shows the superiority of our method in person search task. Notice that we try to implement more than twice iteration operations on the basis of NAE, but they fail to work better than IAS.

Effectiveness of Three-Way Decision The above detection decline in IAS is caused by IoU mismatch. To alleviate it, we introduce the framework of the three-way decision to IAS. From Table 2, we can see that our 3W-AlignNet achieves 92.4% and 87.3% w.r.t. Recall and AP, superior to IAS by 1.4% and 1.6%, respectively. And it achieves 93.0% and 94.0% w.r.t. mAP and top-1, outperforming the IAS by 0.6 and 0.4 pp. We take C_1 and C_2 as two hyper-parameters and set them according to prior experience and optimal experiment results. Due to the strong classification and representation ability of deep neural networks, most samples are classified by NAE as $det\_score < 0.5$ or $det\_score > 0.8$ in the first iteration. Therefore, C_1 is set as 0.5 to abandon samples in the negative domain, which is the same as the NAE baseline. C_2 is the key value to be explored for better performance. When we set $C_2 < 0.8$ (such as 0.6, 0.7 and others), there are too few samples in the boundary domain and thus the effect is not significant. To divide positive and boundary domain samples, we set C_2 as 0.8, 0.85, 0.9, and 0.95 in our experiment. On CUHK-SYSU, $C_2 = 0.95$ achieves the best performance, while on PRW, $C_2 = 0.9$ is the best choice.

Table 1 Ablation experiments on CUHK-SYSU and PRW

Method	CUHK-SYSU		PRW	
	mAP	top-1	mAP	top-1
NAE-base	91.5	92.4	43.3	80.9
+ GWS	91.8	92.8	43.5	81.2
NAE-base	91.5	92.4	43.3	80.9
IAS	92.4	93.6	44.0	81.6
3W-AlignNet	93.0	94.0	44.3	81.7

Table 2 Influence of detection quality and re-ID feature localization on the performance on CUHK-SYSU dataset. The Iterative Box [30] (mentioned in Methodology) is a similar testing phase enhancement method like IAS

Method	Detection		re-ID	
	Recall	AP	mAP	top-1
NAE-base	92.6	86.8	91.5	92.4
Iterative Box [30]	**93.2**	86.9	91.8	93.0
IAS	91.0	85.7	92.4	93.6
3W-AlignNet	92.4	**87.3**	**93.0**	**94.0**

Best results in each block are marked in bold

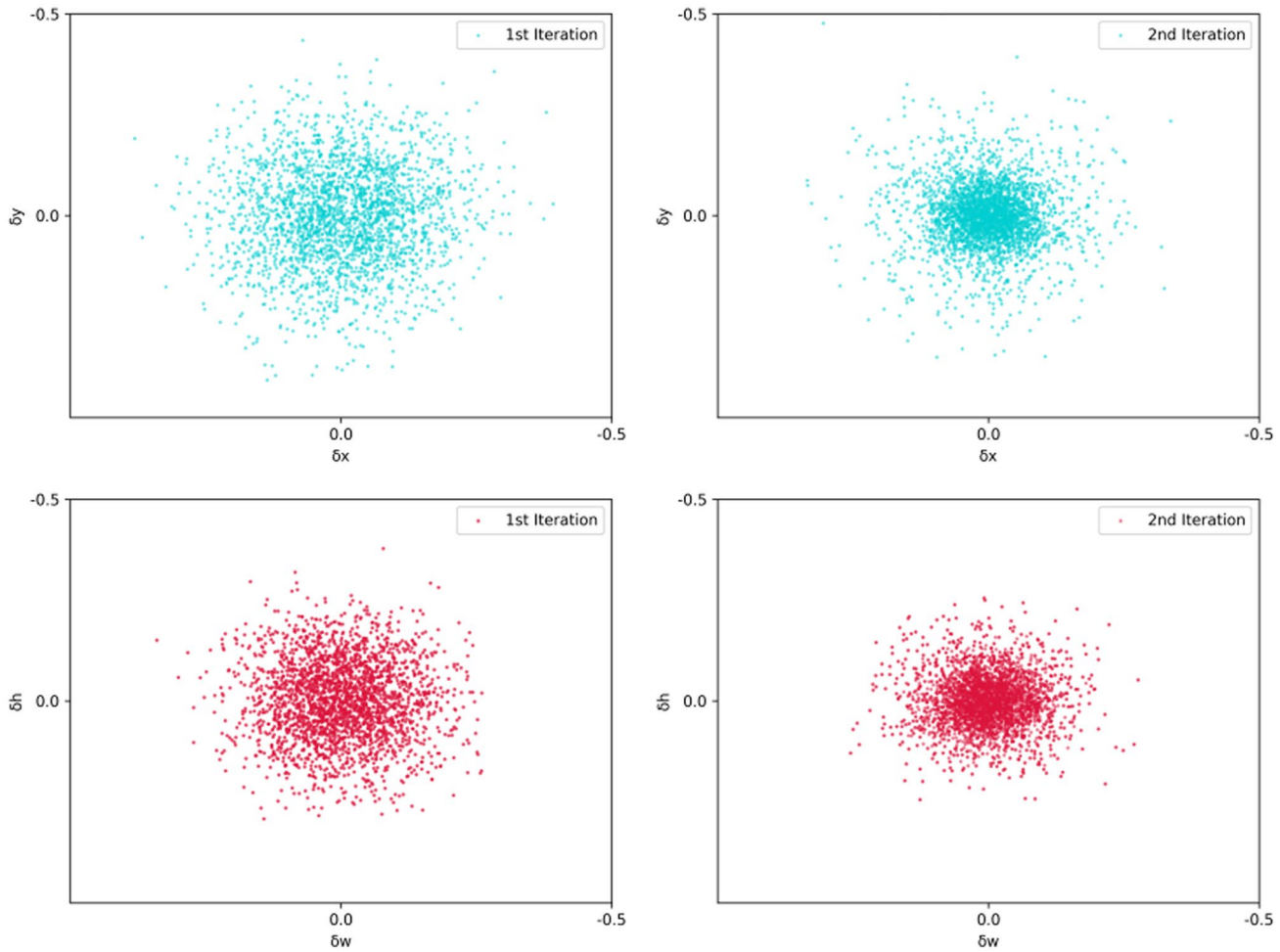

Fig. 8 Regressor Δ distribution before and after applying IAS

The influence of C_2 on the performance on CUHK-SYSU is shown in Fig. 7, which shows the robustness of our method to the value of C_2. In short, 3W-AlignNet is simultaneously influenced by detection quality and re-ID feature localization, and it achieves the best performance when the above two factors are well balanced.

Fig. 9 Top-1 search results for several hard samples. The notion Q denotes the query person, for each we show the top-1 results given by NAE and GWS

Table 3 Comparison of mAP and top-1 accuracy with the state-of-the-art methods on CUHK-SYSU and PRW

	Method	CUHK-SYSU		PRW	
		mAP%	top-1%	mAP%	top-1%
two-stage	DPM+IDE [31]	-	-	20.5	48.3
	CNN+MGTS [8]	83.0	83.7	32.6	72.1
	CNN+CLSA [9]	87.2	88.5	38.7	65.0
	FPN+RDLR [33]	93.0	94.2	42.9	70.2
	TCTS [10]	**93.9**	**95.1**	**46.8**	**87.5**
end-to-end	OIM ([1])	75.5	78.7	21.3	49.9
	IAN [11]	76.3	80.1	23.0	61.9
	NPSM [22]	77.9	81.2	24.2	53.1
	CTXGraph [12]	84.1	86.5	33.4	73.6
	QEEPS [23]	88.9	89.1	37.1	76.7
	BINet [24]	90.0	90.7	45.3	81.7
	NAE [25]	91.5	92.4	43.3	80.9
	NAE+ [25]	92.1	92.9	44.0	81.1
	Ours	**93.0**	**94.0**	**44.3**	**81.7**

Best results in each block are marked in bold

GWS is Helpful Figure 9 shows examples of the visualization results of our GWS, which fixes the incorrect results of NAE. We select $k_1 = 0.8$ and $k_2 = 0.9$, respectively. In Table 1, adding GWS to the NAE baseline on CUHK-SYSU yields a gain of 0.3% and 0.4% for mAP and top-1, respectively. On PRW, GWS yields a gain of 0.2% and 0.3% for mAP and top-1. The overall results demonstrate the efficacy of GWS.

Comparison to the state-of-the-arts

In this section, we will compare our method to state-of-the-art methods in Table 3. All the results are clustered into two categories, i.e., two-stage methods in the upper block and one-stage methods in the lower block. CNN in Table 3 represents a Faster R-CNN detector with ResNet50 backbone.

Comparison on CUHK-SYSU From Table 3, our method achieves 93.0% and 94.0% w.r.t. the mAP and top-1 metrics respectively. The result outperforms all other existing end-to-end methods, including the strong BINet [24] and NAE+ [25]. Our method also achieves comparable results with two-stage ones, including FPN+RDLR [33] and TCTS [10]. Note that, compared with query-guided methods, such as QEEPS [23] and TCTS [10], our method inherits the advantage of NAE, which is computationally light and efficient. The gallery size is set as 100 in Table 3. To evaluate the performance consistency with varying gallery sizes, we evaluate the mAP with a gallery size of [50, 500, 1000, 2000, 4000]. As shown in Fig. 10, the performance of all methods degrades as the gallery size increases, which shows the limitations of current person search algorithms at larger search scales. Nevertheless, our method stands out from all end-to-end methods at all gallery sizes.

Comparison on PRW We further evaluate our method with competitive ones on the PRW dataset. In this experiment, the gallery size of the testing dataset is 6112. Thus PRW is more challenging with less training data and larger gallery size. As shown in the right column of Table 3, our approach outperforms all previous end-to-end methods and surpasses the second-best two-stage method FPN+RDLR [33] by a large margin. The overall results strongly demonstrate the effectiveness of our method again.

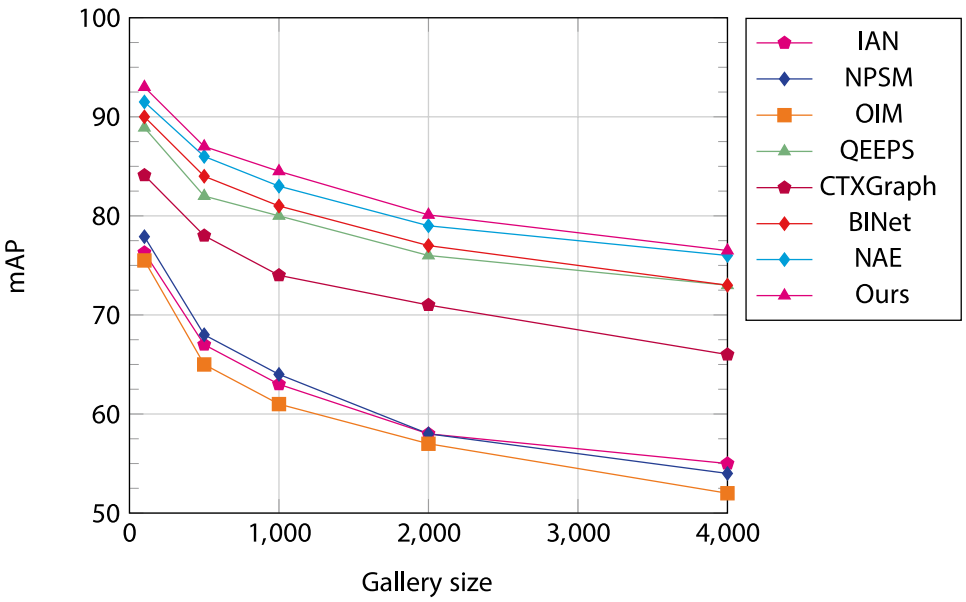

Fig. 10 The mAP of end-to-end methods under different gallery size on CUHK-SYSU

Conclusions

In this paper, we propose a three-way based feature alignment framework for person search, which is inspired by the thinking model of three-way cognitive computations. Our method aims at alleviating the feature misalignment problem in end-to-end methods and illuminates that the re-ID sub-task is very sensitive to the feature localization. We further propose a novel gallery box reweighting algorithm to deal with the granularity mismatch problem between the query and gallery boxes. Extensive experiments have been conducted, and the results validate the superiority of our method. We hope our work will inspire more and more researchers to pay attention to the feature localization mechanism of person search models.

Acknowledgements The research is supported in part by the National Nature Science Foundation of China (Grant Nos. 61976158, 61673301, and 62076182).

Declarations

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent Informed consent is obtained from all individual participants included in the study.

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Xiao T, Li S, Wang B, Lin L, Wang X. Joint detection and identification feature learning for person search. In Proc IEEE Conf Comput Vis Pattern Recognit. 2017;3415–24.
2. Dollár P, Appel R, Belongie S, Perona P. Fast feature pyramids for object detection. IEEE Trans Pattern Anal Mach Intell. 2014;36(8):1532–45.
3. Zhang S, Bauckhage C, Cremers AB. Informed haar-like features improve pedestrian detection. In Proc IEEE Conf Comput Vis Pattern Recognit. 2014;947–54.
4. Yang B, Yan J, Lei Z, Li SZ. Convolutional channel features. In Proceedings of the IEEE International Conference on Computer Vision. 2015;82–90.
5. Liao S, Hu Y, Zhu X, Li SZ. Person re-identification by local maximal occurrence representation and metric learning. In Proc IEEE Conf Comput Vis Pattern Recognit. 2015;2197–206.
6. Cheng D, Gong Y, Zhou S, Wang J, Zheng N. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In Proc IEEE Conf Comput Vis Pattern Recognit. 2016;1335–44.
7. Xiao T, Li H, Ouyang W, Wang X. Learning deep feature representations with domain guided dropout for person re-identification. In Proc IEEE Conf Comput Vis Pattern Recognit. 2016;1249–58.
8. Chen D, Zhang S, Ouyang W, Yang J, Tai Y. Person search by separated modeling and a mask-guided two-stream CNN model. In Proceedings of the European Conference on Computer Vision. 2020;29:4669–82.
9. Lan X, Zhu X, Gong S. Person search by multi-scale matching. In Proceedings of the European Conference on Computer Vision. 2018;536–52.
10. Wang C, Ma B, Chang H, Shan S, Chen X. Tcts: A task-consistent two-stage framework for person search. In Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2020;11952–61.
11. Xiao J, Xie Y, Tillo T, Huang K, Wei Y, Feng J. IAN: the individual aggregation network for person search. Pattern Recogn. 2019;87:332–40.
12. Yan Y, Zhang Q, Ni B, Zhang W, Xu M, Yang X. Learning context graph for person search. In Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2019;2158–67.
13. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2015;91–9.
14. Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection. In Proc. IEEE Conf Comput Vis Pattern Recognit. 2018;6154–62.
15. Yao Y. Three-way decisions with probabilistic rough sets. Inf Sci. 2010;180(3):341–53.
16. Wen P, Li Y, Polkowski L, Yao Y, Tsumoto S, Wang G. Three-way decision: An interpretation of rules in rough set theory. In International Conference on Rough Sets and Knowledge Technology. 2009;642–9.
17. Yao Y. An outline of a theory of three-way decisions. In International Conference on Rough Sets and Current Trends in Computing. 2012;1–17.
18. Yao Y, Wang S, Deng X. Constructing shadowed sets and three-way approximations of fuzzy sets. Inf Sci. 2017;132–53.
19. Yao Y, Wang S, Deng X. Constructing shadowed sets and three-way approximations of fuzzy sets. Inf Sci. 2017;412:132–53.
20. Li H, Zhang L, Huang B, Zhou X. Sequential three-way decision and granulation for cost-sensitive face recognition. Knowl-Based Syst. 2016;91(C):241–51.
21. Zhang Y, Zhang Z, Miao D, Wang J. Three-way enhanced convolutional neural networks for sentence-level sentiment classification. Inf Sci. 2019;477:55–64.
22. Liu H, Feng J, Jie Z, Jayashree K, Zhao B, Qi M, Jiang J, Yan S. Neural person search machines. In Proceedings of the IEEE International Conference on Computer Vision. 2017;493–501.
23. Munjal B, Amin S, Tombari F, Galasso F. Query-guided end-to-end person search. In Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2019;811–20.
24. Dong W, Zhang Z, Song C, Tan T. Bi-directional interaction network for person search. In Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2020;2839–48.
25. Chen D, Zhang S, Yang J, Schiele B. Norm-aware embedding for efficient person search. In Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2020;12615–24.
26. Chen T, Miao D, Zhang Y. A graph-based keyphrase extraction model with three-way decision. In International Joint Conference on Rough Sets. 2020;111–21.
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proc IEEE Conf Comput Vis Pattern Recognit. 2016;770–8.
28. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision. 2017;2961–9.
29. Girshick R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision. 2015;1440–8.
30. Gidaris S, Komodakis N. Attend refine repeat: Active box proposal generation via in-out localization. Proceedings of the British Machine Vision Conference. 2016;90:1–13.
31. Zheng L, Zhang H, Sun S, Chandraker M, Yang Y, Tian Q. Person re-identification in the wild. In Proc IEEE Conf Comput Vis Pattern Recognit. 2017;1367–76.
32. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conf Comput Vis Pattern Recognit. 2009;248–55.
33. Han C, Ye J, Zhong Y, Tan X, Zhang C, Gao C, Sang N. Re-id driven localization refinement for person search. In Proc IEEE/CVF International Conference on Computer Vision. 2019;9814–23.