A multitask joint framework for real-time person search

Ye Li1,2,3 · Kangning Yin2 · Jie Liang2 · Zhuofu Tan2 · Xinzong Wang1 · Guangqiang Yin2 · Zhiguo Wang2

Received: 14 February 2022 / Accepted: 15 July 2022 / Published online: 17 August 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
Person searches generally involve three important parts: person detection, feature extraction and identity comparison. However, a person search integrating detection, extraction and comparison has the two following drawbacks. First, the accuracy of detection will affect the accuracy of comparison. Second, it is difficult to achieve real-time results in real-world applications. To solve these problems, we propose a multitask joint framework for real-time person search (MJF) that optimizes person detection, feature extraction and identity comparison. For the person detection module, we propose the YOLOv5-GS model, which is trained with a person dataset. YOLOv5-GS combines the advantages of the Ghostnet and the squeeze-and-excitation block and improves the speed of person detection. For the feature extraction module, we design a model adaptation architecture, which can select different networks according to the number of people. It can balance the relationship between accuracy and speed. For identity comparison, we propose a 3D pooled table and a matching strategy to improve identification accuracy. On the condition of 1920×1080-resolution video and a 200-ID table, the IR and the FPS achieved by our method reach 82.69% and 25.14, respectively. Therefore, the MJF can achieve real-time person search.

Keywords Person search · Multitask · Joint framework · Real time

1 Introduction
Person search [1] is mainly used to determine whether there is a specific person in an image or a video sequence. Person search has great potential in applications related with video surveillance, such as searching for lost people or suspects. These applications are closely related to public security and safety, therefore person search has been receiving increasing attentions over recent years. In real-world applications, person search can be divided into two tasks: person detection and person re-identification. Detection is the premise of reidentification and neither can exist effectively without the other. However, most scholars only study person re-identification [2–4]. Their achievements are difficult to implemented in real-world applications. In addition, although some scholars propose an end-to-end network [5–7] combined with two tasks, the accuracy is relatively low since the detection accuracy affects the identification accuracy. Furthermore, person detection aims to distinguish a person from the background, while person identification aims to distinguish between different IDs. Therefore, a single network cannot meet real-time requirement of person search.

To address the problems above, we propose a multitask joint framework for real-time person search (MJF) that consists of the person detection, feature extraction and identity comparison, as shown in Fig. 1.

Regarding the task of person detection, we integrate a Ghost [8] and squeeze-and-excitation block (SE) [9] and propose the YOLOv5-GS network to improve the speed and accuracy. In addition, considering that the original YOLOv5 model is mainly used to detect multiple types of targets...
rather than single person, we train it with the CrowdHuman person dataset, which makes the network more suitable for person detection. As a result, our model achieves 58.2 FPS for 1920 \times 1080-resolution video.

Regarding the task of feature extraction, we design a model adaptation architecture (MAA) can select different feature extraction models to balance the relationship between accuracy and speed. The MAA chooses a simple model to extract features when there are many people in a frame, and chooses a complex model to extract features when there are a few people in a frame. Therefore, the module can improve accuracy as well as speed. We choose ResNet-18, ResNet-34 and ResNet-50 networks [10] as the basic models, and embed Non-local block [11]. Considering that the features of a person are different in different orientations, we insert the person orientation classification module into the network, and extract orientation information from the front, back and side. The experiments show that the MAA can extract the features of 605 people in one second even if we use ResNet-50 as the feature extraction model, and mAP and Rank-1 can reach 93.2% and 95.9%, respectively.

Regarding the task of identity comparison, we design the 3D pooled table, which is used to store the features of each ID under three orientations. In addition, we propose a matching strategy that confirms 4 frames in 5 consecutive frames. After confirmation, the feature is matched with the features in the 3D pooled table. If the match is successful, then the feature under the corresponding ID and orientation are updated; otherwise, the new ID and orientation are initialized. We use the cosine distance to evaluate the similarity between different IDs. On the condition of 1920 \times 1080-resolution video and a 200-ID table, the IR and FPS achieved by our method reach 82.69% and 25.14.

The main contributions of this paper are summarized as follows:

- We optimize person detection and propose the YOLOv5-GS model. It integrates Ghostnet and SE block, which reduces the parameters and improves the detection speed.
- We propose a model adaptation architecture. The architecture can flexibly select feature extraction networks according to the number of people and greatly improve the accuracy and speed. We insert the person orientation classification module into the network, and extract orientation information from the front, back and side.
- We design a 3D pooled table and propose an identity matching strategy. The stability and reliability of person identification are improved by identity confirmation in 4 frames within 5 consecutive frames and feature matching in the same direction.

2 Related work

2.1 Person detection

Object detection models can be divided into two categories. One is two-stage, which has two steps: object positioning and object recognition. This category includes R-CNN [12], Fast R-CNN [13] and Faster R-CNN [14]. The candidate boxes are generated in the first step, and the candidate boxes are judged in the second step. The generation and judgment of the boxes are two processes. This type of algorithms has high accuracy but slow speed. The other one is one-stage, such as SSD [15] and YOLO [16–19]. This type can achieve high inference speed and meet the real-time requirements. What is more, there are many detection method proposed in recent years [20–26].

YOLO is the first end-to-end network and can predict the locations and categories simultaneously. The YOLO series algorithms are divided into three modules: backbone, neck and head. YOLO redefines object detection as a regression
problem, which leads to high speed [16, 17]. Specifically, the detection speed of YOLO is faster than that of other object detection algorithms due to the simplification of the network [18] and the use of tricks [19], which can achieve real-time detection in the application of object detection [27].

2.2 Feature extraction

Feature extraction can be divided into global feature extraction and local feature extraction in the field of person re-identification. Generally, the input of the global feature extraction network should be the whole image. Many researchers have used ResNet (including ResNet-18, ResNet-34 and ResNet-50), which can solve degradation problems. In addition, a series of networks was developed to achieve better performance, such as ResNeXt [28] which combines ResNet and Inception [29], ResNest [30] that fuses split-attention module. However, the method of using global features has poor performance in real scenes for the problem of occlusion. Thus, local feature extraction becomes extremely important. The global feature consists of multiple local features, which are extracted from a certain region. These methods include image segmentation [31–35] and attitude extraction [36–40]. Zhao et al. [37] proposed a multistage feature decomposition and tree-like structure competitive feature fusion network, which is based on human body region guidance. It extracts people features with the help of 14 key position information item of the human body. To solve the problem of person dislocation, Zhao et al. [41] proposed an effective person alignment network, which included the convolutional network and the region extraction network. It extracted the most discriminative human body areas and spliced them into the final person features.

2.3 Identity comparison

There are two key issues that need to be addressed in identity comparison. One is to confirm that the Tracklets in the scene belong to the same ID. The other is to confirm that the ID is the comparison target. Xiao et al. [42] proposed using the LUT to store the features of each ID between the detection network and the person re-identification network. Xiao trained an end-to-end person re-identification network using the OIM loss function. Li et al. [43] modified the LUT and proposed the pooled table to store the features of each camera. Considering changes in color, lighting and angle of people at different cameras, Li proposed a new loss function Triplet Online Instance Matching Loss.

3 Method

3.1 Person search process

The person search can be divided into four steps (detect person, extract feature, fuse feature and identify person), and we provide the explanation mathematically about how to search the person.

Detect person We input an frame and detect the pedestrians, in which Q is query (database) composed of M images (person), denoted as \( \{ q_i \}_{i=1}^M \).

Extract feature We train the model to extract the feature \( f_j \) of the person \( j \) (see Eq. 1).

\[
f_j = F(q_j),
\]

where \( F(\cdot) \) is the feature extraction model.

Fuse feature To improve the robustness of the feature, we fuse multiple image features of the same person (see Eq. 2).

\[
fuse_j = \text{Fuse}(f_j^k),
\]

where \( \text{Fuse}(\cdot) \) is the feature fusion model and \( f_j^k \) represents the feature of \( k \)th frame of \( j \) person.

Identify person \( G \) is a feature gallery (database) composed of \( N \) images, denoted as \( \{ g_i \}_{i=1}^N \). They belong to \( N \) different identities 1, 2, ..., \( N \). Given a fusion feature \( fuse_j \), its identity is determined by:

\[
i^* = \arg \max_{i \in 1,2,...,N} \text{sim}(fuse_j, g_i),
\]

where \( i^* \) is the identity of probe \( fuse_j \), and \( \text{sim}(\cdot, \cdot) \) is some kind of similarity function.

3.2 YOLOv5-GS

It is difficult to achieve real-time person search using the current methods since the detection is time-consuming. In this paper, we design the YOLOv5-GS model, which uses the idea of YOLOv4 and the architecture of YOLOv5. YOLOv5 adopts the CSP [44] that can effectively reduce network parameters, the focus structure that can decrease information loss, and the SPP pyramid structure that is suitable for multisize input. Inspired by these methods, we insert the Ghostnet and SE block into the YOLOv5 to improve the accuracy and speed. Ghostnet can reduce the parameters of the backbone, while the SE layer can lower the accuracy degeneration caused by Ghostnet. Although inserting the SE layer may increase the calculation, the model can strengthen valuable features and suppress invaluable features by adjusting the channel attention, which leads to the enhancement of the learning ability of the
network. Specifically, the person dataset CrowdHuman is used to train YOLOv5-GS.

We insert the Ghost Block into the backbone, $N = 1$ in the first Ghost Block and $N = 9$ in the last two. The detailed method is as follows: Fig. 2a shows how the residual block replaces the CSP in the original structure, and Fig. 2b replaces the convolutional layer of the two blocks that make up the Ghost block (as shown in Fig. 2c). We add an SE layer

![Fig. 2](image_url)

**Fig. 2** a Structure of GhostBottleneck [8] when stride = 1. b Structure of GhostBottleneck when stride = 2. c Structure of Ghost Block

---

![Fig. 3](image_url)

**Fig. 3** Architecture of YOLOv5-GS

---

![Fig. 4](image_url)

**Fig. 4** Architecture of MAA

© Springer
after each sampling to enhance the feature learning ability. The architecture of YOLOv5-GS is shown in Fig. 3.

### 3.3 Model adaptation architecture (MAA)

Two problems need to be considered in the feature extraction part. First, the features of the same ID are different under different orientations. If the orientation is ignored, then the accuracy of person identification will decrease after the orientation is changed. Second, the speed of feature extraction decreases with an increase in the number of layers. With a deepening of the network depth and increase in structure complexity, the computing time increases, which affects the real-time performance. To solve the above problems, we design the MAA (as shown in Fig. 4) which consists

| Structure of backbone module | ResNet18 | ResNet-34 | ResNet-50 |
|-----------------------------|----------|-----------|-----------|
| conv<sub>1</sub>             | 7 x 7, 64, stride 2 | x1 |           |
| pool<sub>1</sub>             | 3 x 3, avg, stride 2 | x1 |           |
| layer<sub>1</sub>            | 3 x 3, 64 | x2 | 3 x 3, 64 | x3 |
| nonlocal<sub>1</sub>         | 1 x 1, 64 | x1 | 1 x 0, 64 | x0 |
| layer<sub>2</sub>            | 3 x 3, 128 | x1 | 1 x 1, 256 | x0 |
| nonlocal<sub>2</sub>         | 1 x 1, 128 | x2 | 1 x 1, 128 | x4 |
| layer<sub>3</sub>            | 3 x 3, 256 | x6 | 3 x 3, 256 | x6 |
| pool<sub>2</sub>             | 3 x 3, 512 | x3 | 3 x 3, 512 | x3 |
| fc<sub>1</sub>               | –        | – | (2048, 512) | x1 |
| Bn                          | Bn Feature |           |           |
| fc<sub>2</sub>               | (512, 3) |           |           |

![Fig. 5](image-url) Process of feature storage and comparison. a ID and orientation. b 3D Pooled Table. c Container. Person features of same ID with different orientations is stored in 3D pooled table. Container updates state after processing each frame. After match is confirmed, feature in container replaces feature of corresponding ID and orientation in 3D pooled table.
of three parts: backbone module, network depth adaptation module and orientation and identity module.

**Backbone module** In this section, we study the backbone networks: ResNet-18, ResNet-34 and ResNet-50. The structure of the backbone is shown in Fig. 4. We use non-local block to establish the connection between the pixels that have a certain distance, and the Bn Feature structure to classify orientation and features. The detailed network structure is shown in Table 1.

**Network depth adaptation module** There is a game process between accuracy and speed. The deeper the network is, the higher the recognition accuracy is and the lower the PPS is. By contrast, the shallower the network depth, the lower the recognition accuracy and the higher the PPS. Therefore, the feature extraction module should achieve balance to ensure high accuracy and speed. We design the network depth adaptation module, as shown in Fig. 4. It can select the corresponding network according to the number of targets in each frame and improve the efficiency of feature extraction without decreasing the accuracy. ResNet18 + nonlocal + Bn Feature (RN18), ResNet34 + nonlocal + Bn Feature (RN34) and ResNet50 + nonlocal + Bn Feature (RN50) are selected as the three adaptive branch networks. The strategy is expressed in Eq. (4).

\[
\text{backbone} = \begin{cases} 
\text{RN50} & n \leq \text{th}_1 \\
\text{RN34} & \text{th}_1 < n < \text{th}_2 \\
\text{RN18} & n \geq \text{th}_2, 
\end{cases}
\]  
(4)

where \( n \) is the number of people in one frame, and \( \text{th}_1 \) and \( \text{th}_2 \) represent the thresholds the number of people.

**Orientation and identity module** To reduce the influence of orientations, we proposed the Orientation and identity module, which contains orientation features and ID features, as shown in Fig. 4. We insert the FC2 layer into the Orientation and identity module to extract orientation features. We classify the orientations as front, back and side. The total features have 515 dimensions, among which the person ID feature is represented by a vector of 512 dimensions, and the person orientation feature is represented by a vector of 3 dimensions.

### 3.4 3D Pooled table and matching strategy

In this section, a 3D pooled table is designed to store features of each ID in different orientations, as shown in Fig. 5. Each feature is represented by \( f \in \mathbb{R}^D \), where \( D \) represents the dimension of the feature, the orientation is represented by \( t \in \mathbb{R}^3 \), \( t = 0 \) represents the front, \( t = 1 \) represents the back, and \( t = 2 \) represents the side. Each ID in Table is represented by \( V \in \mathbb{R}^{D \times T} \), where \( T = 3 \) represents the number of types of orientation. In addition, we design a matching strategy to confirm the same ID through container storage. The features and orientations of different IDs are stored in different containers, and the containers also contain the state of feature matching. If a container matches 4 frames in 5 consecutive frames, then the state of the container is changed to confirm, and the features in the container are compared with the person features in the 3D pooled table with the same orientation. The detailed processes are as follows (Algorithm is shown in 1).

**Algorithm 1 Matching Algorithm**

**Input:**
1. Video segment: \( P \)
2. Container: \( C_x, x = 1, 2, 3... \)

**Init:**
1. segment of \( i \)-th person in the first frame: \( p_i^1 \)
2. each person creates a container: \( C_x \) (including \( cou, mis, fea, ori, lab \))

**For each frame** \( p \in P \) do
1. detect the person fragments using YOLOv5-GS: \( p_o \)
2. extract the feature and the orientation using MAA: \( f_o \in \mathbb{R}^D, t_o \in \mathbb{R}^3 \)
3. match \( f_o \) with the existing features \( fea \)
4. if matched then
   5. update the \( fea \) and \( cou \) of the \( C_x \)
   6. else
   7. update the \( mis \) of the \( C_x \)
5. end if
6. if \( mis == 0 \) then
   7. delete \( C_x \)
6. end if
7. if \( cou == 5 \) and \( mis > 0 \) or \( cou == 4 \) and \( mis == 2 \) then
   8. match \( C_x \) with Table
   9. if matched then
   10. update Table
   11. else
   12. new ID
6. end if
8. end if
9. if \( mis > 0 \) or \( cou == 4 \) then
   10. update the \( fea \) and \( cou \) of the \( C_x \)
   11. else
   12. update the \( mis \) of the \( C_x \)
9. end if
10. end if

**Output:**
1. Table
2. Container \( C_x, x = 1, 2, 3... \)

First, the 1st frame inputs the model, and the fragments \( p_i^1 \) of all people in this frame are detected through YOLOv5-GS, where \( i \) represents the \( i \)-th person in the 1st frame. Each detected person creates a container \( C_x, x = 1, 2, 3..., \) where \( x \) represents the \( x \)-th container. In each container, the
corresponding person features will be stored, including the number of matches (cou), number of mismatches (mis), features (fea), orientation of pedestrians (ori), and the person label (lab). Then, the detected fragments of person input MAA for feature and orientation extraction, and the features and orientations to the corresponding container Cx were stored. When the ath frame is input to the model, the fragments p′i of all pedestrians in this frame are first extracted through YOLOv5-GS. The detected fragments of the person are input to network MAA to extract the feature fpi ∈ Rd and the orientation ori ∈ R1. The fpi is matched with the existing feature in the container, and if matched, the corresponding fea and cou in the container are updated.

For containers that do not match, only the mis is updated. If the mis is updated twice in 5 frames, then the container is deleted. If a container matches 4 frames in 5 consecutive frames, then the container is removed and matched the features of the container with the corresponding orientation in the table fpi ∈ Rd, where n represents the nth ID and t represents the orientation. If the match is successful, then the features of the corresponding ID orientation are updated in the table. Otherwise, the new ID is initialized.

4 Dataset and evaluation index

4.1 Person detection comparison

COCO [45] It is a large image dataset designed for object detection, segmentation, person keypoint detection, staff segmentation, and caption generation.

CrowdHuman [46] It is a benchmark dataset to better evaluate detectors in crowd scenarios. The CrowdHuman dataset is large, rich-annotated and contains high diversity. CrowdHuman contains 15,000, 4370 and 5000 images for training, validation, and testing, respectively.

Parameter It refers to the number of parameters in the network and indicates the complexity of the model. It generally includes weight W and bias B.

mAP[0.5] It represents the accuracy of the objection detection and refers to the mean average precision (mAP) under the condition of intersection over union (IoU) was 0.5. IoU is the degree of overlap between the BBox and the Ground Truth. The calculation method of mAP[0.5] can reference the mAP [47].

TS Test speed (TS) is the detection time of GPU for each frame which excludes decoding process.

FPS, FPSmax, FPSavg Frame per second (FPS) is used to evaluate the speed of the image process in videos. The detection speed is affected by the number of pedestrian targets in the image. The more targets, the lower the processing speed. We use the total average method to calculate the FPS, which takes all the frames divided by the time, to avoid the fluctuation of the frame rate as much as possible. The calculation formula is shown in Eq. (5).

\[
FPS = \frac{f_{all}}{t_{all}},
\]

where \(f_{all}\) is the total number of frames in the video, \(t_{all}\) is the time to process the video. FPS\(_{\text{max}}\) and FPS\(_{\text{avg}}\) represents the average frame rate and maximum frame rate, respectively.

4.2 Feature extraction comparison

Market-1501 [47] The dataset is collected by 6 different cameras. The whole dataset contains 32,668 images, including 12,936 for training and 19,732 for testing.

DukeOrientation Images with different orientations (front, back and side) were selected from the DukeMTMC dataset [48]. The DukeOrientation dataset contains 3587 images, including front 1140, back 873 and side 1574.

mAP It represents the mean average precision of persons ReID. The calculation method can reference the [47].

Rank-1 It represents the probability that a query identity hit the first object of the candidate lists. The calculation method of Rank-1 is same with the CMC top-1 [49].

PPS Person per second (PPS) refers to the number of pedestrians feature extraction within one second. The calculation formula is shown in equation (6).

\[
PPS = \frac{n_{\text{extra}}}{t_{\text{extra}}},
\]

where \(n_{\text{extra}}\) is the pedestrian number of the feature extraction, \(t_{\text{extra}}\) is the time cost.

4.3 Real-time comparison

UESTC-PR A dataset is a very important component in person search. However, most of the existing datasets only contain cropped pedestrian images, which is different from real world scenarios. To simulate the actual application scenario and minimize the production costs of the dataset, we propose a new dataset collection method and collected the UESTC-PR dataset in this way. The UESTC-PR dataset was collected in the small square of the innovation center at the University of Electronic Science and Technology of China (UESTC). Five cameras were adopted and distributed in all directions of the central square. When a pedestrian passes by, the different cameras take pictures of the different poses and backgrounds of the same identity. Thus, it can meet the requirement of person re-identification for diverse poses among the samples in the dataset. In addition, due to the large square and the high position of the camera, the acquired pedestrians are small, which is in line with the actual scene. To control the impact of light on the person recognition accuracy, the
video collection was conducted in three periods at different times of day, namely, 8:30–9:30 am, 11:00–12:00 am and 17:00–18:00 pm. The dataset covers diverse backgrounds, illumination conditions, attitudes, angles, brightnesses, etc. It can meet ReID’s requirements for the dataset. The datasets could be used for training a model for the joint detection and identification task. The pedestrian pictures are cropped from the original surveillance videos, which can be used for the detection task, as well as the joint detection and ReID task. The statistics are clearly shown in Table 2.

**IR** Identification rate (IR) is used to evaluate the search accuracy. The number of correct identifications represents the number of correct identifications in the identity comparison process, and the total number represents the total number of identity comparisons. The calculation formula is shown in Eq. (7).

\[
IR = \frac{n_{\text{cor}}}{n_{\text{all}}},
\]  

where \(n_{\text{cor}}\) is number of correct identification, \(n_{\text{all}}\) is the total number of identification.

5 Experiments

5.1 Pre-processing method

We describe some pre-processing technology in this subsection, including the dataset pre-processing skills and the parameter settings.

**Dataset pre-processing skills** We resize the input image of the person detection task to 640*640, 1088*1088 and 1920*1920. The hue, saturation and lightness of the image are set to 0.015, 0.7 and 0.4, respectively. The image is turned left and right with a probability of 50%.

**Parameter settings** The IoU between ground truth and anchor is set to 0.2 and batchsize is set to 16. We adopt momentum gradient descent and the momentum weight is set to 0.937, and the weight attenuation coefficient is 0.0005. The learning rate is 0.01 and epoch is 300.

5.2 Comparison of person detection

In this section, we first compare the performance of SSD, YOLOv5, YOLOv5-G (YOLOv5 + Ghostnet) and YOLOv5-GS (YOLOv5 + Ghostnet + SE layer) in the field of object detection on the COCO dataset. Then, we select YOLOv5-GS as the person detection model. In addition, we test the detection performance of YOLOv5 and YOLOv5-GS on the UESTC-PR dataset.

**A comparison of different models in object detection** We test the performance of SSD, YOLOv5, YOLOv5-G and YOLOv5-GS on the COCO dataset. We compare the parameter, mAP and TS. The experimental results are shown in Table 3.

Compared with the YOLOv5 network, the parameters of YOLOv5-GS decrease from 7.2M to 4.4M, the GPU inference time shortens by 0.4 ms per frame. Although the mAP of YOLOv5-GS is decreased, the inference speed is increased. At the same time, the results of YOLOv5-G and YOLOv5-GS demonstrate that the SE module can compensate for the decline in mAP caused by Ghostnet. The mAP achieved by YOLOv5-GS can be increased by up to 3.3% compared with YOLOv5-G. It is obvious that YOLOv5-GS finds a balance point between the detection accuracy (mAP[0.5]) and the detection speed (TS). Based on the above experiments, we adopt YOLOv5-GS as the person detection model.

**The effectiveness of YOLOv5-GS in person detection** The model, which is trained with the COCO dataset, is not suitable for person detection. A large number of nonperson targets consume the detection time. We retrain the YOLOv5-GS model with the CrowdHuman dataset. The validation dataset we adopted is UESTC-PR, and we test the relation between FPS and resolution. The experimental results are shown in Table 4.

The results show that the FPS\(_{\text{avg}}\) of YOLOv5-GS reached 58.2 in 1920 \times 1920-resolution video, which meets the

| Dataset          | DukeMTMC | Market-1501 | UESTC-PR |
|------------------|----------|-------------|----------|
| Pedestrians      | 36,411   | 32,668      | 60,437   |
| Identities       | 1812     | 1501        | 499      |
| Training ID      | 702      | 751         | 251      |
| Query ID         | 702      | 750         | 247      |
| Gallery ID       | 1110     | 750         | 248      |
| Size             | Big      | Big         | Small    |
| Shooting angle   | Smooth inspect | Smooth inspect | Overlook |
| Motion state     | Walking  | Walking     | Walking + cycling |
| Detection + identification | No      | No          | Yes      |

The dataset collected by ourself is shown in bold

 Springer
real-time application requirements. In addition, with an increase in the resolution of the input image, the FPS\textsubscript{avg} of YOLOv5-GS fluctuates between 55.7 and 60.6, and the YOLOv5 fluctuates between 35.6 and 37.3. The experiment also proves that YOLOv5-GS can improve the detection speed.

5.3 The comparison of feature extraction

The size of the input images is set to $128 \times 64$. We train the model on the Market1501 dataset and use some tricks. The dimension of the fully connected layer in our paper is set to 512. Compared with the dimension of 256 and 1024, the accuracy and speed reach the balance point when the dimension is 512 since a small dimension leads to a low accuracy and a large dimension leads to a low speed. We test mAP and Rank-1 to evaluate the network performance. The PPS is used to evaluate the speed of the feature extraction. The results are shown in Table 5.

It can be seen from Table 5 that with the addition of the nonlocal and the Bn Feature, the mAP achieved by ResNet-18, ResNet-34 and ResNet-50 can be increased by up to 2.6%, 3.8% and 3.9%, respectively. Compared with the nonlocal, the Bn feature has a greater influence on mAP. The nonlocal increases the depth of the network and reduces the PPS, while the Bn Feature has little influence on the PPS. Based on the above results, ResNet18 + nonlocal + Bn Feature (RN18), ResNet34 + nonlocal + Bn Feature (RN34) and ResNet50 + nonlocal + Bn Feature (RN50) are selected as the three adaptive branch networks. According to the results in Table 5, we calculate the corresponding relationship between the number of people and the selection of the feature extraction network. The selection strategy is shown in Table 6.

According to Table 6, RN50 is selected when the number of pedestrians in the frame is fewer than 24, RN34 is selected when the number of pedestrians in the frame is fewer than 25 and RN18 is selected when the number of pedestrians in the frame is fewer than 28.

5.4 Comparison of identity

In this section, we study the impact of resolution and number of IDs in the 3D pooled table on the IR and FPS. The experimental results are shown in Table 7.

Table 7 shows that an increase in the number of IDs results in a decrease in IR and FPS. This is because the more IDs there are, the more people will need to be compared. In addition, the IR decreases with decreasing resolution, but the FPS is improved to some extent. The main reason is that the poor feature extracted from low resolution pedestrian images leads to a decrease of IR. However, this kind of image improves the speed of detection. In conclusion, under the condition of $1920 \times 1080$-resolution video and a 200-ID table, the IR and FPS achieved by our method reached 82.69% and 25.14, respectively.

5.5 Comparison with the state-of-the-art method

In this section, we compare the method proposed in this paper with other methods. The experiment includes the comparison of the feature extraction model and the comparison of the multitask joint framework.

| Network | Tricks | mAP | Rank-1 | PPS |
|---------|--------|-----|--------|-----|
| ResNet-18 | × | × | 89.2 | 90.4 | 825.979 |
| | ✓ | × | 89.7 | 90.8 | 723.284 |
| | × | ✓ | 91.1 | 92.5 | 795.735 |
| | ✓ | ✓ | 91.8 | 93.1 | 709.321 |
| ResNet-34 | × | × | 89.4 | 90.7 | 717.265 |
| | ✓ | × | 90.7 | 92.2 | 664.488 |
| | × | ✓ | 90.5 | 91.6 | 726.795 |
| | ✓ | ✓ | 93.2 | 94.2 | 637.340 |
| ResNet-50 | × | × | 89.3 | 91.2 | 625.607 |
| | ✓ | × | 90.1 | 91.4 | 616.927 |
| | × | ✓ | 93.2 | 94.2 | 624.981 |
| | ✓ | ✓ | 93.2 | 95.9 | 605.556 |

Table 3 Comparisons between YOLOv5-GS and others on COCO dataset

| Network | Parameter (M) | mAP0.5 (%) | TS (ms) |
|---------|---------------|------------|---------|
| SSD     | 26.2          | 48.5       | 45.4    |
| YOLOv5  | 7.2           | 56.2       | 3.0     |
| YOLOv5-G| 3.9           | 49.8       | 2.5     |
| YOLOv5-GS | 4.4         | 53.1       | 2.6     |

The results achieved by YOLOv5-GS are shown in bold.

| Network | Resolutions | Resolution (original) | Resolution (resize) | FPS\textsubscript{max} | FPS\textsubscript{avg} |
|---------|-------------|-----------------------|---------------------|------------------------|------------------------|
| YOLOv5  | 1920 × 1080 | 640 × 640             | 37.8                | 36.6                   |
|         |             | 1088 × 1088           | 36.8                | 35.6                   |
|         |             | 1920 × 1920           | 38.4                | 37.3                   |
| YOLOv5-GS | 1920 × 1080 | 640 × 640             | 65.7                | 60.6                   |
|         |             | 1088 × 1088           | 57.5                | 55.7                   |
|         |             | 1920 × 1920           | 59.8                | 58.2                   |
A comparison of different feature extraction models In this experiment, we compare the ResNet-50 that proposed in Table 1 with the other feature extraction models on Market-1501 dataset. The results are shown in Table 8. It can be seen that the mAP and Rank-1 of our method outperform the other methods. This is because nonlocal module and Bn Feature structure can establish the connection between the pixels and classify orientation and features.

A comparison of the multitask joint framework In this experiment, under the condition of 1920 × 1080-resolution video and a 200-ID table, we compare the multitask joint framework with other methods on UESTC-PR dataset. The results are shown in Table 9. It can be seen that the IR and FPS of our method outperform the other methods. The improvement of the IR lies in the orientation guidance and matching strategy. The improvement of FPS lies in the network depth adaptation module that can select the network according to the number of pedestrians in the image.

### 6 Conclusion

We propose a multitask joint framework for real-time person search, including the the YOLOv5-GS, the model adaptation architecture and the 3D pooled table and matching strategy. The YOLOv5-GS integrates Ghostnet and SE block, which reduces the parameters and improves the detection speed. The model adaptation architecture can flexibly select feature extraction networks according to the number of people and greatly improve the accuracy and speed. The stability and reliability of person identification are improved by the 3D pooled table and matching strategy. We conduct extensive experiments to evaluate the proposed method. On the condition of 1920 × 1080-resolution video and a 200-ID table, the IR and FPS achieved by our method could reach 82.69% and 25.14, respectively. This can meet real-time application requirements. We also provide a reference solution for real-time person search.

### Funding
This work was supported in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region (no. 2022D01B05).

### Declarations

Conflict of interest The data that support the findings of this study are available online. These datasets were derived from the following public domain resources: [COCO, CrowdHuman, Market-1501, DukeMTMC].

### References

1. Xu, Y., Ma, B., Huang, R., Lin, L.: Person search in a scene by jointly modeling people commonness and person uniqueness. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 937–940 (2014)
2. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1318–1327 (2017)
3. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3800–3808 (2017)
4. He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M.: Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 558–567 (2019)

5. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: End-to-end deep learning for person search 2 (2). arXiv preprint arXiv:1604.01850

6. Munjal, B., Amin, S., Tombre, F., Galasso, P.: Query-guided end-to-end person search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 811–820 (2019)

7. He, Z., Zhang, L.: End-to-end detection and re-identification integrated net for person search. In: Asian Conference on Computer Vision, pp. 349–364. Springer (2018)

8. Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C.: Ghostnet: More features from cheap operations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1580–1589 (2020)

9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)

11. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)

12. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)

13. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)

14. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)

15. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: Ssd: single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer (2016)

16. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)

17. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)

18. Farhadi, A., Redmon, J.: Yolov3: an incremental improvement. Computer Vision and Pattern Recognition, cite as

19. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.: Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv: 2004.10934

20. D-fes: Deep facial expression recognition system. In: Conference on Information and Communication Technology

21. Singh, H., Dhanak, N., Ansari, H., Kumar, K.: Hdml: habit detection with machine learning. In: International Conference on Computer and Communication Technology (2017)

22. Sharma, S., Kumar, K.: Asl-3dconv: American sign language recognition technique using 3-d convolutional neural networks. Multimed. Tools Appl. 80(17), 26319–26331 (2021)

23. A novel superpixel based color spatial feature for salient object detection. In: Conference on Information and Communication Technology

24. Ansari, H., Vijayvergia, A., Kumar, A.: Dcr-hmm: depression detection based on content rating using hidden markov model. In: Proceedings of IEEE 2nd Conference on Information and Communication Technology, (CICT 2018) (2019)

25. Negi, A., Chauhan, P., Kumar, K., Rajput, R.S.: Face mask detection classifier and model pruning with keras-surgeon. In: 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE) (2020)

26. Vijayvergia, A., Kumar, K.: Selective shallow models strength integration for emotion detection using GloVe and LSTM. Multim. Tools Appl. 80, 28349–28363 (2021)

27. Hu, J., Gao, X., Wu, H., Gao, S.: Detection of workers without the helmets in videos based on yolo v3. In: 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 1–4. IEEE (2019)

28. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)

29. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)

30. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Mueller, J., Manmatha, R., et al.: Resnext: Split-attention networks. arXiv preprint arXiv:2004.08955

31. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepleid: deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159 (2014)

32. Varior, RR., Shuai, B., Lu, J., Xu, D., Wang, G.: A siamese long short-term memory architecture for human re-identification. In: European Conference on Computer Vision, pp. 135–153. Springer (2016)

33. Sharma, S., Kumar, P., Kumar, K.: Lexer: Lexicon based emotion analyzer

34. Kumar, S., Kumar, K.: Lsrc: Lexicon star rating system over cloud, pp. 1–6 (2018)

35. Kumar, S., Kumar, K.: Irscc: Integrated automated review mining system using virtual machines in cloud environment. In: 2018 Conference on Information and Communication Technology (CICT) (2018)

36. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3960–3969 (2017)

37. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, Y., Yi, S., Wang, X., Tang, X.: Spindle net: person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1077–1085 (2017)

38. Vijayvergia, A., Kumar, K.: Star: rating of reviews by exploiting variation in emotions using transfer learning framework. In: 2018 Conference on Information and Communication Technology (CICT) (2018)

39. Darbari, A., Kumar, K., Darbari, S., Patil, P.L.: Requirement of artificial intelligence technology awareness for thoracic surgeons. Cardiothorac. Surg. 29(1), 13 (2021)

40. Kumar, K., Kurhekhar, M.: Sentimentalizer: Docker container utility over cloud. In: International Conference on Advances in Pattern Recognition

41. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aware features from cheap operations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3219–3228 (2020)

42. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3415–3424 (2017)
43. Li, Y., Yin, G., Liu, C., Yang, X., Wang, Z.: Triplet online instance matching loss for person re-identification. Neurocomputing 433, 10–18 (2021)
44. Wang, C-Y., Mark Liao, H-Y., Wu, Y-H., Chen, P-Y., Hsieh, J-W., Yeh, I-H.: Cspnet: A new backbone that can enhance learning capability of cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 390–391 (2020)
45. Lin, T-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, CL.: Microsoft coco: common objects in context. In: European Conference on Computer Vision, pp. 740–755 (2014)
46. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: a benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123
47. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124 (2015)
48. Gou, M., Karanam, S., Liu, W., Camps, O., Radke, R.J.: Dukemtmc4reid: A large-scale multi-camera person re-identification dataset. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1425–1434 (2017). https://doi.org/10.1109/CVPRW.2017.185
49. Bolle, RM., Connell, JH., Pankanti, S., Ratha, NK., Senior, AW.: The relation between the roc curve and the cmc. In: IEEE Workshop on Automatic Identification Advanced Technologies (2005)
50. Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., Ji, R.: Pyramidal person re-identification via multi-loss dynamic training. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8506–8514 (2019). https://doi.org/10.1109/CVPR.2019.00871
51. Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., Sun, J.: Alignedreid: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184
52. Martinel, N., Foresti, G.L., Micheloni, C.: Aggregating deep pyramidal representations for person re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1544–1554 (2019). https://doi.org/10.1109/CVPRW.2019.00196
53. Herzog, F., Ji, X., Teepe, T., Hörmann, S., Gilg, J., Rigoll, G.: Lightweight multi-branch network for person re-identification. arXiv preprint arXiv:2101.10774
54. Wang, H., Fan, Y., Wang, Z., Jiao, L., Schiele, B.: Parameter-free spatial attention network for person re-identification. arXiv preprint arXiv:1811.12150
55. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: a new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. 43(2), 652–662 (2021). https://doi.org/10.1109/TPAMI.2020.2938758
56. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3701–3711 (2019). https://doi.org/10.1109/ICCV.2019.00380
57. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. arXiv preprint arXiv:1611.05431
58. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Muller, J., Manmatha, R., Li, M., Smola, A.: Resnest: Split-attention networks. arXiv preprint arXiv:2004.08955

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.