International Journal of Intelligent Robotics and Applications (2021) 5:104–118
https://doi.org/10.1007/s41315-021-00167-2

REGULAR PAPER

# Learning hierarchical and efficient Person re-identification for robotic navigation

**Jiangning Zhang[1] · Chao Xu[1] · Xiangrui Zhao[1] · Liang Liu[1] · Yong Liu[1] · Jinqiang Yao[2] · Zaisheng Pan[1,3]**

Received: 26 August 2020 / Accepted: 18 March 2021 / Published online: 19 April 2021


## Abstract

Recent works in the person re-identification task mainly focus on the model accuracy while ignoring factors related to efficiency, e.g., model size and latency, which are critical for practical application. In this paper, we propose a novel ***H**ierarchical and **E**fficient **Net**work* (HENet) that learns hierarchical global, partial, and recovery features ensemble under the supervision of multiple loss combinations. To further improve the robustness against the irregular occlusion, we propose a new dataset augmentation approach, dubbed random polygon erasing, to random erase the input image's irregular area imitating the body part missing. We also propose an ***E**fficiency **S**core* (ES) metric to evaluate the model efficiency. Extensive experiments on Market1501, DukeMTMC-ReID, and CUHK03 datasets show the efficiency and superiority of our approach compared with epoch-making methods. We further deploy HENet on a robotic car, and the experimental result demonstrates the effectiveness of our method for robotic navigation.

**Keywords** Person re-identification · Hierarchical features · Robotic navigation · Random polygon erasing · Efficient score

## 1 Introduction

The person re-identification aims at retrieving corresponding images of a given person among the gallery person database, which has vast promising applications such as video surveillance and criminal investigation. Since Yi et al. (2014) first apply the deep neural network to solve the ReID task, innumerous methods (Zhang et al. 2017; Lawen et al. 2020; Sun et al. 2018; Zhu et al. 2017; Fu et al. 2018; Wang et al. 2018) emerge in succession. However, there are still challenges to apply the current person ReID methods to practical application. *In the article, we focus on how to efficiently design and train the ReID model, and the idea can easily help to boost the performance of other methods.*

Network architecture is one of the most concerned problems. Zheng et al. (2019a) first use a network to estimate pose information and then combine it with the ReID model to improve performance. However, this method suffers from an extra-time consumption for extracting pose information, which is unsuitable for practical application. Song et al. (2018) propose a mask-guided model named MGCAM to learn separate feature from body and background, and DSA (Zhang et al. 2019) constructs a set of densely semantically aligned part images to guide the *main full image stream* to learn densely semantically aligned features from the original

✉ Jiangning Zhang
186368@zju.edu.cn

Chao Xu
21832066@zju.edu.cn

Xiangrui Zhao
xiangruizhao@zju.edu.cn

Liang Liu
leonliuz@zju.edu.cn

Yong Liu
yongliu@iipc.zju.edu.cn

Jinqiang Yao
Eluse@qq.com

Zaisheng Pan
panzs@zju.edu.cn

[1] APRIL Lab, College of Control Science and Engineering, Zhejiang University, Hangzhou, Zhejiang, China

[2] Zhejiang Communications Group Inspection Technology, Hangzhou, Zhejiang, China

[3] Zhejiang Chipkong Technology, Hangzhou, Zhejiang, China

image. Nevertheless, such mask-guided methods also suffer from a similar time consumption for requiring an extra module to generate the body mask. The attention-based method (Liu et al. 2017) introduces an attention mechanism to enhance model discrimination, and methods (Zhong et al. 2019; Liu et al. 2018) involve the idea of GAN to enrich the training dataset for reducing the impact of the limited dataset. Zheng et al. (2019b) employ a generative module to generate high-quality cross-id composed images. However, these methods generally require extra structures or operations that increase the complexity of the network in practical application. So many stripe-based methods are proposed, which are easy to follow and have pretty good performance. Sun et al. (2018) first propose a stripe-based PCB that divides the image into six stripes and obtains a good result. Recently proposed HPM (Fu et al. 2018) acquires a better performance than other methods by using a pyramid structure. However, nearly all methods only focus on the accuracy yet ignore the efficiency of the model that is equally important for practical application. Considering the above reasons, we propose a stripe-based HENet that consists of global, partial, and recovery branches, which can efficiently extract discriminative features without extra complicated operations.

Besides the network structure, some researchers focus on loss function designing because a stronger constraint can improve the model performance without increasing the model complexity. Wen et al. (2016) propose a center loss to learn the feature center of each class, and Xiao et al. (2017) design a non-parametric OIM loss to further leverage unlabeled data. Hoffer and Ailon (2015) apply the triplet loss to the ReID task, and this loss improves the network performance in a large margin by punishing intra-class and inter-class distances. We summarize center loss and OIM in one non-parametric category, considering the design intention, while triplet and quadruplet losses in one hard sample mining category. Since different loss functions have different design intentions, we can simultaneously use different kinds of loss functions, expecting mutual complementation between them to improve the model performance.

As we all know, deep neural networks' performance usually degrades in challenging scenarios, such as pose change, illumination intensity, and especially body occlusion. Liu et al. (2018) propose a novel Pose-transfer network to generate images with different poses from only one input image, which greatly extends the training dataset. Zhong et al. (2019) design a network named CamStyle, which serves as a data augmentation approach that smooths the camera style disparities. Besides the GAN-based data augmentation method, Zhong et al. (2017b) make a *Random Erasing* (RE) operation on the training dataset, and the results indicate that RE is very effective. To further reduce the impact of irregular occlusions inside the pedestrian images, we propose a new *Random Polygon Erasing* (RPE) method that will be detailedly illuminated in Sect. 3.3.

To explore and solve the above issues, we propose a novel HENet that learns global, partial, and newly designed recovery features simultaneously. During the training stage, we apply different loss function combinations to different branches, expecting mutual complementation among different kinds of loss functions. Furthermore, a novel RPE data augmentation method is proposed to boost performance, and we propose an ***Efficiency Score*** (ES) metric to evaluate model efficiency. The designed HENet does not need extra structure or operation during the testing stage, so it is more suitable for practical applications, e.g., robotic navigation. As shown in Fig. 1, we compare with different epoch-making stripe-based methods because they have relatively simple structures that are friendly to the actual hardware and require no elaborate training techniques. Besides, our approach has a comparable volume and speed, as well as achieves higher R1 and mAP scores than other methods.

Specifically, we make the following four contributions:

- We propose an efficient HENet that learns hierarchical features ensemble under the supervision of multiple loss combinations, and it is more suitable for practical application than other methods.
- We propose a new data augmentation approach, dubbed RPE, to imitate the irregular occlusion, which significantly boosts the model performance.
- We propose an ES metric to evaluate the model efficiency for practical application.
- Our approach achieves a much better result than other epoch-making methods in three commonly used datasets, as well as acquires the highest efficiency score that shows its superiority for practical industrial application, as shown in Fig. 1.

## 2 Related work

### 2.1 Deep Person ReID

Hand-crafted methods had been dominating the person ReID task until learning-based methods arrived. The first work using deep learning to solve ReID issue is Yi et al. (2014), which greatly improves the model performance and promotes the development of ReID. Recently, many stripe-based methods (Sun et al. 2018; Fu et al. 2018; Wang et al. 2018) that focus on learning local features are proposed, which are easy to follow and have high accuracy. Some other works (Zheng et al. 2019a; Song et al. 2018) leverage extra information, e.g., human pose and body mask, to improve the model performance. However, they generally consume extra storage space and time because an extra sub-network is

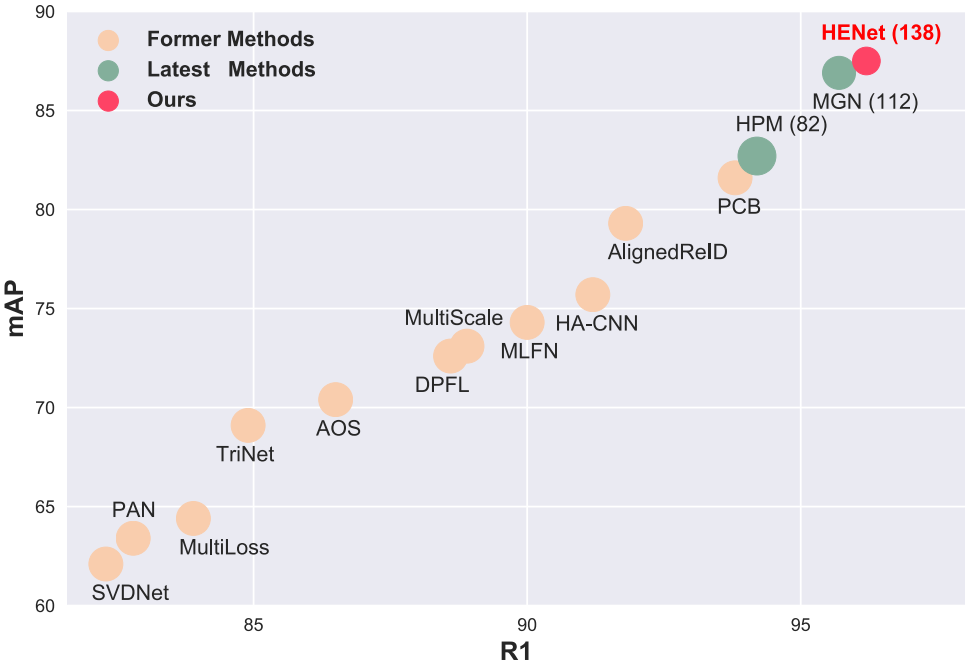


**Fig. 1** Performance of different epoch-making methods on Market1501. For our and latest popular methods, the marker size indicates the model size, and the number in parenthesis indicates the running frame rate. Our proposed HENet outperforms all other methods in R1 and mAP, as well as has an equilibrium model size and a relatively faster-running speed

needed to extract pose or mask information. Attention-based approach (Liu et al. 2017) introduces an attention mechanism to locate the active salient region that contains the person. The local feature-aligned method is extremely popular in recent years, and Zhao et al. (2017) propose a part-aligned representation to deal with the body part misalignment problem. AlignedReID (Zhang et al. 2017) performs an alignment/matching by calculating the shortest path between two sets of local features without requiring extra supervision.

Methods as mentioned above usually require extra information (e.g., human pose) or operation (e.g., body alignment), which increases application costs in practical application. Stripe-based methods (Sun et al. 2018; Fu et al. 2018; Wang et al. 2018), which divide the image into several stripes and extract local features for each stripe, can be viewed as pure methods because they neither contain extra complicated structures or operations nor need extra label information. Thus we design our partial branch by equally cropping deep feature maps into small stripes to learn partial information, which is easy to use and has strong feature extraction ability for practical application.

## 2.2 Loss function for Person ReID

Cross-entropy loss is usually employed as a supervisory signal to train a classification model, and many researchers focus on designing the loss function for the ReID task. Wen et al. (2016) propose a center loss that simultaneously learns the feature center of each class and penalizes different feature centers, while Xiao et al. (2017) propose a non-parametric OIM loss to further leverage unlabeled data. Works Hoffer and Ailon (2015), Chen et al. (2017a) employ a distance comparison idea to improve the network performance by punishing intra-class and inter-class distances. In the paper, we adopt multiple loss combinations in training to extract hierarchical features ensemble, expecting mutual complementation between different kinds of loss functions for improving the model performance.

## 2.3 Data augmentation for Person ReID

Some GAN-based methods commonly contribute to the model performance by enriching the training dataset. For instance, Camera Style (Zhong et al. 2019) use CycleGAN to style-transfer labeled images to other cameras, which reduces the over-fitting and increases data diversity. Liu et al. (2018) propose the pose-transfer network to synthesize images with various poses from only one input image. At the same time, PTGAN (Wei et al. 2018) bridges the domain gap among datasets when transferring a person from one dataset to another. Besides generating more diverse images, some related works aim to solve body occlusion by adding erasing operations. Wei et al. (2017) use an adversarial erasing method to localize and expand object regions progressively. Zhong et al. (2017b) propose a RE dataset augmentation approach, which first selects a random rectangle region in an image and then erases its pixels with fixed values. To reduce the impact of irregular body occlusions, we propose a novel

*random polygon erasing* (RPE) approach, which brings a considerable improvement without adding complex structure or operation to the network.

## 3 Our approach

### 3.1 The structure of HENet

As shown in Fig. 2, the proposed HENet consists of three branches to extract hierarchical features. Specifically, global branch $G_1$ learns the global feature, while partial branch $P_4$ equally splits the whole feature map into four horizontal spatial bins for extracting partial features. The newly designed $R_{16}$ branch learns the recovery feature under the reconstruction loss supervision besides the CE loss. Considering the efficiency and performance for practical application, we employ ResNet50 (He et al. 2016) structure as the backbone of HENet that is in line with other methods.

$GM_1$ is the global feature map belonging to branch $\mathbf{G_1}$ after convolution operations. $GM_4$ and $PM_{4,j}$ are global and partial feature maps belonging to branch $\mathbf{P_4}$, where $PM_{4,j}$ is the $j_{th}$ bin that is split from $GM_4$. Then we use the pooling layer to generate global and partial intermediate features, $GI_1$, $GI_4$, and $PI_{4,j}$.

$$GI_i = Pooling(GM_i), \quad PI_{4,j} = Pooling(PM_{4,j}), \tag{1}$$

where $i = 1, 4$ and $j = 1, 2, 3, 4$. Subsequently, convolution with 1x1 kernel is employed to generate final global feature $\boldsymbol{f}_{G_i}$ with 512 vector dim and partial feature $\boldsymbol{f}^j_{P_4}$ with 256 vector dim.

$$\boldsymbol{f}_{G_i} = Conv(GI_i), \quad \boldsymbol{f}^j_{P_4} = Conv(PI_{4,j}). \tag{2}$$

In the training stage, $\boldsymbol{f}_{G_i}$ are fed into linear layer $FC_{512}$ where Tri and CE losses serve as loss functions, while $\boldsymbol{f}^j_{P_4}$ are fed into linear layer $FC_{256}$ where OIM and CE losses are used.

$RM$ is the recovery feature map belonging to branch $\boldsymbol{R}$ after convolution operations, and we also use the pooling layer to generate recovery intermediate feature $RI$.

$$RI = Pooling(RM). \tag{3}$$

The convolution with 1x1 kernel is also employed to generate the final recovery feature $\boldsymbol{f}_{\boldsymbol{R}}$ with 512 vectors dim.

$$\boldsymbol{f}_{\boldsymbol{R}} = Conv(RI). \tag{4}$$

Then a decoder is applied to reconstruct a low-resolution image by inputting $\boldsymbol{f}_{\boldsymbol{R}}$ during the training stage, where the reconstruction and CE losses serve as the cost loss functions.

### 3.2 Loss functions

During the training stage, we employ triplet and CE losses for the global branch, OIM and CE losses for the partial branch, as well as reconstruction and CE losses for the recovery branch.

*CE loss* The person ReID belongs to the multi-classification task, which is generally supervised by the CE loss

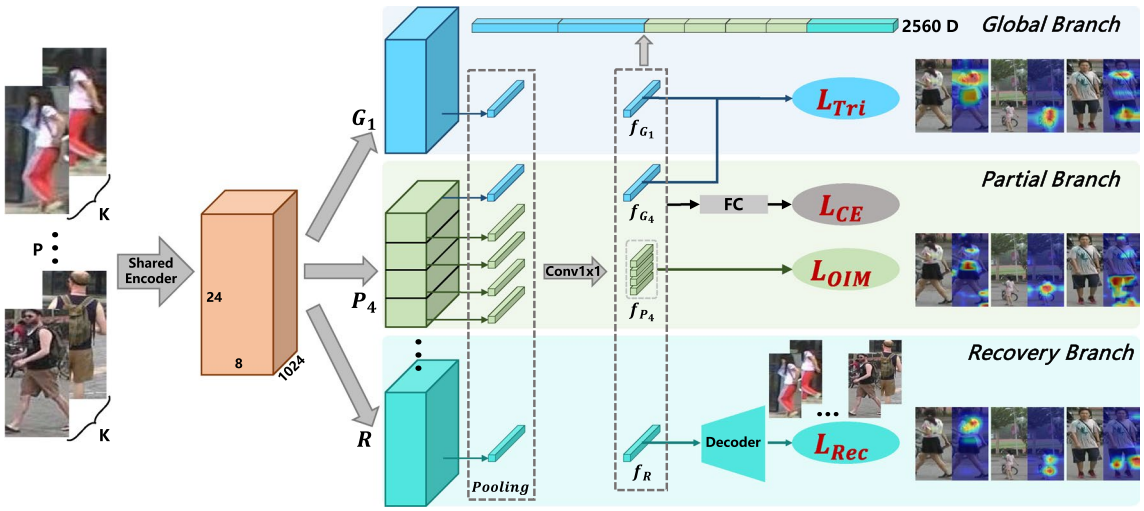


**Fig. 2** Diagram of the proposed HENet. The feature maps gone through a *shared encoder* are split into $P_1$, $P_4$, and $R_{16}$ branches to further extract hierarchical features. During the training stage, global feature $f_G$ uses triplet and CE losses as supervisory signals, partial feature $f_P$ uses OIM and CE losses, and recovery feature $f_R$ uses Rec and CE losses. During the testing stage, hierarchical features are concatenated to form the final representation. The right visual heatmaps are obtained by CAM (Selvaraju et al. 2017) from three branches, where different branches focus on different regions

function. Denote $\boldsymbol{f}_n$ as the $n_{th}$ extracted feature that belongs to $\{\boldsymbol{f}_{G_1}, \boldsymbol{f}_{G_4}, \boldsymbol{f}^j_{P_4}, \boldsymbol{f}_R\}$, we obtain the following equation:

$$\mathcal{L}_{CE} = -\sum_{n=1}^{N} \log \frac{\exp(\boldsymbol{W}_{y_n}\boldsymbol{f}_n)}{\sum_{j=1}^{C} \exp(\boldsymbol{W}_j \boldsymbol{f}_n)}, \tag{5}$$

where $N$ is the batch size, $y_n$ is the real label of the $n_{th}$ extracted feature inside $C$ classes, and $\boldsymbol{W}_j$ is the weight vector for class $j$.

*Triplet loss* As for the global branch, we apply extra triplet loss to improve the model performance by punishing intra-class and inter-class distances in the training stage. Specifically, we first random sample $P$ classes and then random sample $K$ images from each class. For each anchor sample $\boldsymbol{x}^i_a$ that belongs to the class $i$, we select a positive sample $\boldsymbol{x}^i_p$ and a negative sample $\boldsymbol{x}^j_n$ that belong to class $i$ and $j(i \neq j)$ respectively within a batch. The detailed formula is shown below:

$$\mathcal{L}_{Tri}(\theta;\boldsymbol{x}) = \sum_{i=1}^{P}\sum_{a=1}^{K}\left[ m + \max_{p=1\ldots K} D\left(\boldsymbol{x}^i_a, \boldsymbol{x}^i_p\right) - \min_{\substack{j=1\ldots P \\ n=1\ldots K \\ j\neq i}} D\left(\boldsymbol{x}^i_a, \boldsymbol{x}^j_n\right)\right]_+, \tag{6}$$

where $m$ controls the margin between intra-class and inter-class.

*OIM loss* As for the partial branch, we choose OIM loss to further enhance the extracted feature's discrimination. When the category of the training dataset is large and each class only contains few instances, merely using CE loss may result in large variance of gradients in the classifier matrix, thus the model may not learn effectively (Xiao et al. 2017). Non-parametric OIM loss leverages extra unlabeled data in the training stage, which can make this deficiency up to some extent. As a result, we apply both OIM and CE losses to the partial branch, expecting mutual complementation between them. The formula with the input feature $\boldsymbol{x}$ is shown below:

$$p_i = \frac{\exp(\boldsymbol{v}_i\boldsymbol{x}/\tau)}{\sum_{j=1}^{P}\exp(\boldsymbol{v}_j\boldsymbol{x}/\tau) + \sum_{k=1}^{Q}\exp(\boldsymbol{u}_k\boldsymbol{x}/\tau)}, \tag{7}$$

$$\mathcal{L}_{OIM} = \mathrm{E}_x\left[\log p_i\right], \tag{8}$$

where $P$ is total class number, $Q$ denotes the size of circular queue that only contains unlabeled data, $\boldsymbol{v}_j$ and $\boldsymbol{u}_k$ denote weight vectors, and $\tau$ is a temperature parameter that a higher value indicates a softer probability distribution.

*Reconstruction loss* The recovery branch is designed by an adversarial idea. In detail, the CE loss induces the branch $R$ to learn a part of the body feature that is enough for classification, while additional reconstruction loss compels the branch $R$ to learn from the whole image. As the adversarial training goes on, the network learns the feature only from the human body while ignoring the background. Specifically, the recovery feature $\boldsymbol{f}_R$ goes through the *Decoder* to reconstruct a low-resolution image, which is then used to calculate the distance against the real image by the *Mean Square Error* (MSE) loss.

$$\mathcal{L}_{Rec} = \sum(\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i)^2, \tag{9}$$

where $\hat{\boldsymbol{y}}_i$ and $\boldsymbol{y}_i$ denote pixel values of the reconstructed and real images respectively.

### 3.3 Random polygon erasing

To reduce the irregular occlusion's impact inside the person image, we propose a novel data augmentation method named RPE. As depicted in Fig. 3, the left column show the erasing results by RE (Zhong et al. 2017b), while the right two columns are from RPE with different vertex number, i.e., 10 and 40. For an input image $\boldsymbol{I}$ with width $W$, height $H$, and area $S$, it has a probability $p$ to undergo the RPE operation. We first

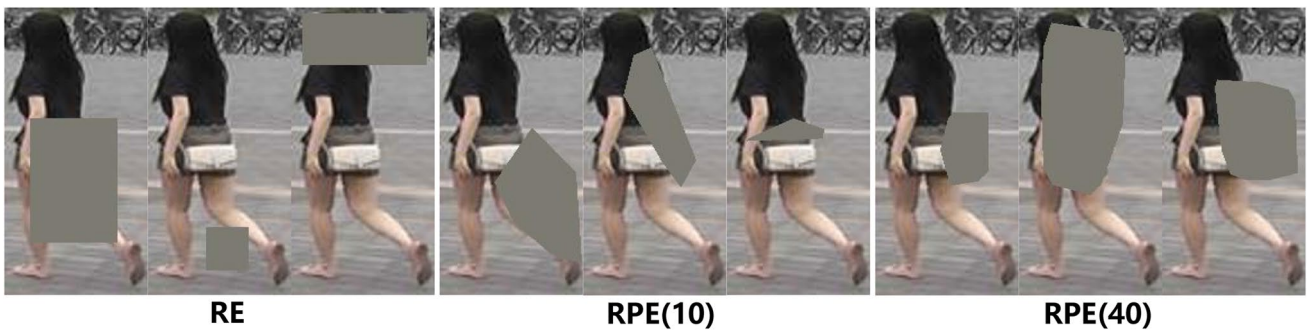


**Fig. 3** Visualization of RE and RPE operations. The gray indicates erased area and the number indicates the vertex number of the selected polygon

random generate the area ratio value $s_e$ between $s_l$ and $s_h$, as well as the aspect ratio $r_e$ between $r$ and $1/r$ ($r < 1$). Then the erasing region with width $W_e = \sqrt{S \times s_e / r_e}(W_e < W)$ and height $H_e = \sqrt{S \times s_e \times r_e}(H_e < H)$ can be generated. Furthermore, a central reference point $P_c$ is randomly selected, whose coordinates $P_{c,x}$ and $P_{c,y}$ are in range $(W_e/2, W - W_e/2)$ and $(H_e/2, H - H_e/2)$ respectively. After that, $n$ points are randomly selected around the central reference point and the minimum external polygon mask can be calculated by generated points. Finally, pixels of $\boldsymbol{I}$ that inside the mask are erased with a random value in the range [0, 255]. In fact, we have experimentally demonstrated that it does not matter what kind of values the erased mask area is filled in, e.g., fixed or random values. So we fixed the values to the mean over the dataset. The detailed procedure of RPE is showed in Algorithm 1, and we set $p = 0.5$, $s_l = 0.02$, $s_h = 0.45$, $r = 0.35$ and $n = 20$ in all experiments by default.

**Algorithm 1:** Random Polygon Erasing Procedure

**Input :** Input image $\boldsymbol{I}$, Image size $W$ and $H$, Erasing probability $p$, Number of point $n$, Erasing area ratio range $s_l$ and $s_h$, Scale ratio $r$.
**Output:** Erased image $\boldsymbol{I}^*$.
**Initialization:** $p_1 \leftarrow$ Rand (0, 1), $cnt \leftarrow 1$.

```
if p_1 ≥ p then
    I* ← I;
    return I*.
else
    while True do
        S_e ← Rand (s_l, s_h) × S;
        r_e ← Rand (r, 1/r);
        W_e = √(S_e/r_e), H_e = √(S_e × r_e);
        if W_e ≤ W and H_e ≤ H then
            P_{c,x} ← Rand (W_e/2, W − W_e/2);
            P_{c,y} ← Rand (H_e/2, H − H_e/2);
            w_min ← Max (P_{c,x} − W_e/2, 0);
            w_max ← Min (P_{c,x} + W_e/2, W);
            h_min ← Max (P_{c,y} − H_e/2, 0);
            h_max ← Min (P_{c,y} + H_e/2, H);
            while cnt ≤ n do
                p_{cnt,x} ← Rand (w_min, w_max);
                p_{cnt,y} ← Rand (h_min, h_max);
                cnt ← cnt + 1;
            end
            mask ← MinPolygon ([(p_{i,x}, p_{i,y})]);
            I(mask) ← Rand (0, 255);
            I* ← I;
            return I*.
        end
    end
end
```

## 4 Experiment

### 4.1 Implementation details

We use pre-trained *ResNet*_50 on ImageNet to initialize the HENet, and the *Decoder* consists of 4 convolution-deconvolution groups. During the training stage, we resize the input image to $384 \times 128$, and form batches by first random sampling four classes and then random sampling four images for each class. Adam is used as the optimizer with parameter settings ($\beta_1 = 0.9$, $\beta_2 = 0.999$) and weight decay $5e^{-4}$. We train the model for 200 epochs, set the base learning rate to $2e^{-4}$, and decay the learning rate to a tenth when arriving at 160 epochs and 200 epochs. HENet is trained in a single TITAN X GPU on PyTorch framework (Paszke et al. 2017).

### 4.2 Dataset and evaluation protocol

*Market1501* includes 32,668 images of 1501 persons detected by the DPM from six camera views. This dataset

is divided into the training set with 12,936 images of 751 persons, the testing set with 3368 query images, and 19,732 gallery images of 750 persons.

*DukeMTMC-ReID* is a subset of the DukeMTMC dataset that contains 36,411 images of 1812 persons from eight camera views. This dataset is divided into the training set with 16,522 images of 702 persons, the testing set with 2228 query images, and 17,661 gallery images of 1110 persons.

*CUHK03* consists of 14,097 images of 1467 persons from six camera views and has two annotation types: manually labeled bounding boxes and DPM-detected bounding boxes (used in the paper). This dataset is divided into the training set with 7365 images of 767 persons, the testing set with 1400 query images, and 5332 gallery images of 700 persons.

*Protocols* we used to evaluate the model performance contain *mean average precision* (mAP), as well as *Cumulative Matching Characteristic* (CMC) at *rank-1* (R1), *rank-5* (R5), and *rank-10* (R10), which can be used in both Market1501 and DukeMTMC-ReID. As for CUHK03, we further adopt the protocol proposed in Zhong et al. (2017a) where the experiments are conducted with 20 random splits for computing averaged performance. Moreover, we propose a new metric named *Efficiency Score* (ES) to evaluate the efficiency of the model for practical application, which considers another three factors besides R1 and mAP, i.e., feature dim (FD), model size (V), and forward speed (S).

### 4.3 Comparison with epoch-making methods

We compare the proposed HENet with some epoch-making methods published in recent 2 years on the datasets mentioned above. As shown in Table 1, HENet-base only contains global and partial branches, while other experiments of the proposed HENet are conducted by adding different components over HENet-base, e.g., RPE, loss, and both components. All experiments of our method only use random horizontal flipping and random erasing data augmentation methods that are in accordance with other epoch-making methods.

From the comparison results, our proposed method obtains competitive results on three datasets. Specifically, the proposed HENet obtains 96.0% R-1 for Market1501, 88.9% R-1 for DukeMTMC-ReID, and 67.1% R-1 for CUHK03, which exceeds all other methods. Besides, our approach outperforms others on mAP: 87.2, 78.9, and 66.5% for Market1501, DukeMTMC-ReID, and CUHK03 respectively. To further evaluate each aforementioned component, we conduct experiments on whether to add the component or not. As shown in Table 1 of the last four lines, either the extra RPE data augmentation method or hierarchical losses improves the model performance, and the network can get the best performance when both the components are applied simultaneously.

We also propose a new metric, dubbed *Efficiency Score* (ES), to evaluate the model efficiency. As shown in the last four columns of Table 1, our method has the highest ES, which means that HENet is more suitable for practical application, and the details of ES will be illustrated in Sect. 4.5.

### 4.4 Ablation study

In this section, several ablation experiments are conducted to demonstrate the effectiveness of our approach. Note that all the following experiments are based on Market1501 dataset with the same settings, and the re-ranking (Zhong et al. 2017a) is not used for a fair comparison.

*Structure analysis* Before doing the ablation study of HENet, we conduct a series of experiments with different partial branch quality to illustrate why we use $G_1$ and $P_4$ as the global and partial branches. As shown in Table 2 of the top half, the network has a better performance with the increasing of the partial branch quantity and nearly reaches the peak when quantity equals three. We choose partial branch number as two considering the performance and efficiency in practical application, where the network has a pretty good performance and a low *Feature Dim* (FD). At the bottom of the table, we show the experimental results with two partial branch quality under different configurations and choose P(1,4) as our base model because of its satisfying performance and proper feature dim.

We further conduct an ablation to evaluate the effect of each branch. As shown in Table 3, combining any two branches obtains better performance than each single branch. When using all three branches, the model achieves the highest score: R1=90.62% and mAP=70.97%, which demonstrates that each branch contributes to the model performance.

*Loss settings* We conduct a group of experiments to evaluate each loss term. As shown in Table 4, using combined losses has a better performance than only a single loss. Specifically, we can obtain the best result, i.e., R1=93.67% and mAP=80.60%, when all losses are applied, which illustrates that mixed loss functions help improve the model performance through mutual complementation.

*RPE performance* To evaluate the effectiveness of RPE data augmentation method for the ReID task, we design a set of experiments under different probabilities with different experimental settings: RE, RPE(10), RPE(20), RPE(30), RPE(40), and RE+RPE(20), where the number is the vertex number (N) of the selected polygon. All experiments are based on the P(1, 4) structure without using other data augmentation methods or tricks. As shown in Fig. 4, the network performance gradually increases along with the probability, which can be consistently seen in both *R1* (upper part) and *mAP* (bottom part) metrics. Besides, the effectiveness of RPE increases when N is larger. We can observe that the

**Table 1** Comparisons with epoch-making methods on three commonly used datasets

| Methods | Market1501 | | DukeMTMC-ReID | | CUHK03 | | Efficiency evaluation | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | R1 | mAP | R1 | mAP | R1 | mAP | FD ↓ | V(*MB*) ↓ | S(*FPS*) ↑ | ES ↑ |
| SVDNet (Sun et al. 2017) | 82.3 | 62.1 | 76.7 | 56.8 | 41.5 | 37.3 | - | - | - | - |
| PAN (Zheng et al. 2018) | 82.8 | 63.4 | 71.6 | 51.5 | 36.3 | 34.0 | - | - | - | - |
| MultiScale (Chen et al. 2017b) | 88.9 | 73.1 | 79.2 | 60.6 | 40.7 | 37.0 | - | - | - | - |
| HA-CNN (Li et al. 2018) | 91.2 | 75.7 | 80.5 | 63.8 | 41.7 | 38.6 | - | - | - | - |
| AlignedReID (Zhang et al. 2017) | 91.8 | 79.3 | 71.6 | 51.5 | 36.3 | 34.0 | 2048 | 100 | 207 | 2.55 |
| PCB (Sun et al. 2018) | 93.1 | 81.0 | 82.9 | 68.5 | 63.7 | 57.5 | 1536 | 102 | 192 | <u>3.45</u> |
| HPM (Fu et al. 2018) | 94.2 | 82.7 | 86.6 | 74.3 | 63.1 | 57.5 | 3840 | 356 | 82 | 1.00 |
| MGN (Wang et al. 2018) | 95.7 | 86.9 | 88.7 | 78.4 | 66.8 | 66.0 | 2816 | 263 | 112 | 2.42 |
| HENet(Base) | 92.9 | 79.6 | 84.1 | 69.7 | 60.3 | 57.7 | 2048 | 169 | 160 | 2.76 |
| HENet(+RPE) | 93.6 | 81.2 | 85.6 | 70.6 | 62.6 | 59.1 | 2048 | 169 | 160 | 3.08 |
| HENet(+Losses) | <u>95.6</u> | <u>86.8</u> | <u>88.9</u> | <u>78.7</u> | <u>66.5</u> | <u>65.9</u> | 2560 | 224 | 138 | 3.41 |
| HENet(+Both) | **96.0** | **87.2** | **88.9** | **78.9** | **67.1** | **66.5** | 2560 | 224 | 138 | **3.52** |

The right four columns are efficiency evaluations of different models

The optimal and suboptimal results of each metric are indicated in bold and underline, respectively

**Table 2** Experimental results with different partial branch quantity

| Structure | FD | R1 | R5 | R10 | mAP |
|---|---|---|---|---|---|
| P(1) | 512 | 87.92 | 95.46 | 96.97 | 68.48 |
| P(1,2) | 1536 | 90.15 | 95.52 | 97.08 | 70.03 |
| P(1,2,3) | 2816 | 90.41 | 95.64 | 97.09 | 71.43 |
| P(1,2,3,4) | 4352 | **90.81** | **95.84** | **97.48** | **72.39** |
| P(1,3) | 1792 | 90.08 | 95.64 | 97.12 | 70.12 |
| P(1,4) | 2048 | 90.17 | 95.68 | 97.30 | 70.39 |
| P(1,6) | 2560 | 90.24 | 95.78 | 97.39 | 70.72 |

The optimal result of each metric is indicated in bold

network reaches a relatively good result when N = 20 and a similar performance against RE when N = 40 (The RPE degenerates into RE when N is large). Moreover, the experiment that combines two methods, marked as RE+RPE(20), is further conducted. The results indicate that this approach obtains a much better improvement than any single method, and it reaches a relative peak state when *probability equals 0.5* (R1 = 93.72% and mAP = 80.53%).

To further show the robustness of *Random Polygon Erasing* against occlusion, we add different occlusion levels and different data pre-processing methods, e.g., RE, RPE, and RE+RPE, to the test dataset in Market1501. As shown in Fig. 5, the performance of all models decreases in both metrics with the increasing of the occlusion level, and the accuracy decreases faster when the test dataset is processed by RE+RPE (right sub-graph). The model trained with RPE (blue solid/dotted lines) or RE (green solid/dotted lines) has a stronger tolerance relative to the modified dataset than doing noting (cyan solid/dotted lines). The model trained with RPE and RE (red solid/dotted lines) outperforms others in a considerable margin, especially when the occlusion level is large. In short, results indicate that the RPE can significantly improve the network's robustness, either alone or in conjunction with RE.

## 4.5 Efficiency evaluation

To further evaluate the efficiency of the model for practical application, we calculate the *Feature Dim* (FD), *Model Volume* (V), and *Forward Speed* (S) attributes for several models that have high scores in *R1* and *mAP*, as shown in Table 1. AlignedReID (Zhang et al. 2017) has the minimum model size and the maximum speed while HPM (Fu et al.

**Table 3** Ablation study of different components combination

| $G_1$ | $P_4$ | $R_{16}$ | R1 | R5 | R10 | mAP | FD |
|---|---|---|---|---|---|---|---|
| ✓ | | | 87.92 | 95.46 | 96.97 | 68.48 | 512 |
| | ✓ | | 89.67 | 95.81 | 96.97 | 69.55 | 1536 |
| | | ✓ | 84.29 | 93.65 | 95.96 | 58.69 | 512 |
| ✓ | ✓ | | 90.17 | 95.68 | 97.30 | 70.39 | 2048 |
| ✓ | | ✓ | 88.98 | 95.28 | 96.91 | 70.42 | 1024 |
| | ✓ | ✓ | 90.15 | 95.89 | 97.16 | 70.34 | 2048 |
| ✓ | ✓ | ✓ | **90.62** | **96.21** | **97.52** | **70.97** | 2560 |

The optimal result of each metric is indicated in bold

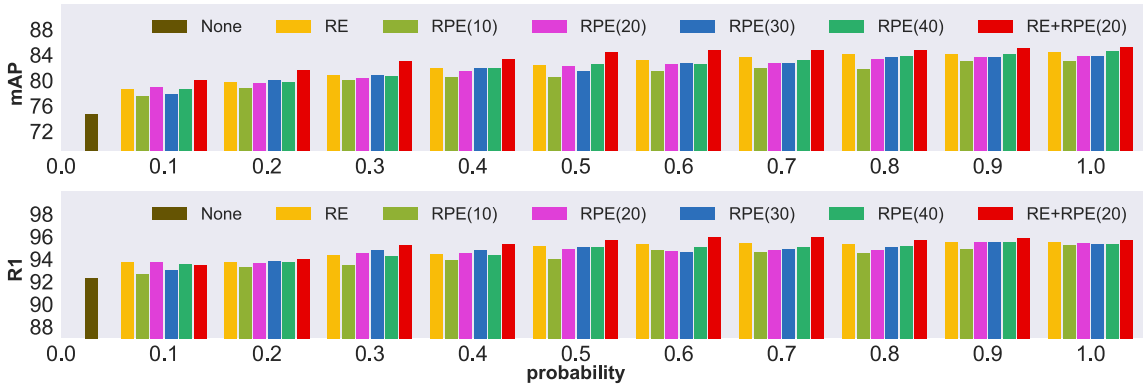


**Fig. 4** Metric evaluations (%) of *R1* and *mAP* with different data augmentations on different levels of occlusion during the training stage in Market1501

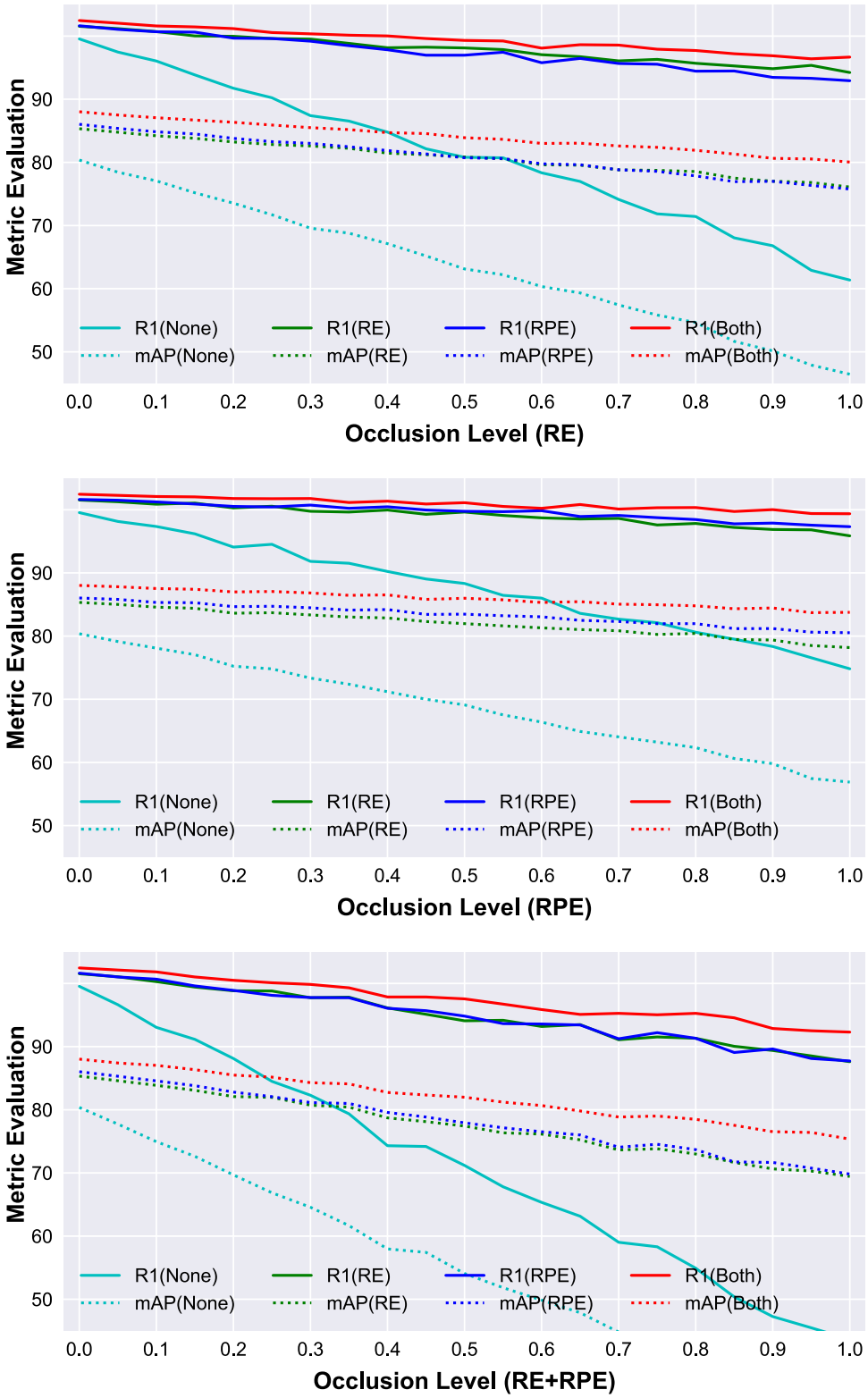


**Fig. 5** Metric evaluations (%) on different levels of occlusion with different methods (from top to bottom are RE, RPE, and RE+RPE) in Market1501. The solid and dotted lines represent the evaluation results on *R1* and *mAP* with different data augmentation methods

2018) is in contrast, and our model has an equilibrium size and a relatively faster speed. To evaluate model efficiency, we propose a new metric called *Efficiency Score* (ES), which takes all aforementioned attributes into complete account and can be used as a practical guideline. Specifically, we first choose the reference models $M^1$ and $M^2$ that have the

**Table 4** Ablation study of different loss function combinations

| OIM | Tri | MSE | R1 | R5 | R10 | mAP |
|---|---|---|---|---|---|---|
| Market1501 | | | | | | |
| | | | 88.72 | 95.43 | 97.12 | 69.69 |
| ✓ | | | 91.33 | 96.02 | 97.74 | 73.42 |
| | ✓ | | 92.62 | 96.91 | 98.15 | 78.02 |
| | | ✓ | 90.43 | 95.90 | 97.18 | 71.65 |
| ✓ | ✓ | | 92.90 | 97.30 | 98.25 | 80.04 |
| ✓ | | ✓ | 91.65 | 96.56 | 97.83 | 74.37 |
| | ✓ | ✓ | 93.20 | 97.68 | 98.40 | 79.90 |
| ✓ | ✓ | ✓ | **93.67** | **97.84** | **98.63** | **80.60** |

The optimal result of each metric is indicated in bold

maximum size and minimum R1 score among comparison models. Then we calculate R1 and mAP scores for the comparison model $M^c$ in the following formula:

$$Score^{T}_{M^c} = \frac{M^c_V \times {M^c_S}^2 \times (M^c_M - M^2_M + thr)^3 / M^c_{FD}}{M^1_V \times {M^1_S}^2 \times (M^1_M - M^2_M + thr)^3 / M^1_{FD}}, \quad (10)$$

where $T$ denotes the category of metrics, $FD$ indicates feature dim, $V$ indicates model size, $S$ indicates forward speed, and $thr$ is the metric threshold that equals 30 in the paper. The formula fully considers multiple factors, e.g. $R1$, $mAP$, $FD$, $V$, and $S$, which are important and must be considered on practical application. In general, $V \times S$ is close for different models, and the extra $S$ term is multiplied to enhance the weight of the forward speed. Besides, considering the metric performance and the feature dim, we also add the latter two terms. Final ES score of the comparison model $M^c$ can be obtained as follows:

$$ES_{M^c} = (Score^{R1}_{M^c} + \lambda Score^{mAP}_{M^c})/(1 + \lambda), \quad (11)$$

where $\lambda$ denotes weight, and we set $\lambda$=1 in the paper. As shown in Table 1, in contrast with other methods, our modified network has less feature dim [e.g., 2560 vs. 3840 (Fu et al. 2018)/2816 (Wang et al. 2018)], comparable volume and speed, as well as high R1 and mAP scores than other methods. We also calculate ES of AlignedReID, PCB, HPM, MGN, and our approach in the Market1501 dataset, which are 2.55, 3.45, 1.00, 2.42, and 3.52, respectively. The results indicate that our approach has a higher $ES$ than other epoch-making methods and is superior for practical application.

### 4.6 Robotic navigation

We deploy HENet on the robot, named NanoCar, to evaluate the navigation and tracking effects of the algorithm. As shown in Fig. 6, the NanoCar contains a camera besides typically required sensors for the Laser-SLAM (Simultaneous Localization and Mapping), which means visual information can be used to navigate the robotic car.

In the experiment, wireless image transmission is used to transfer the real-time image from car to sever. Then Yolo-V3

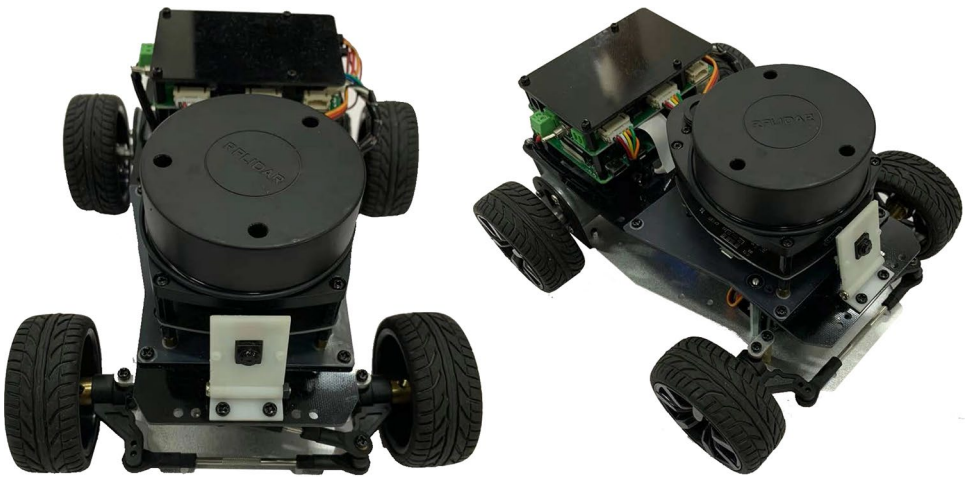

**Fig. 6** A demonstration of the robot named NanoCar for the navigation experiment

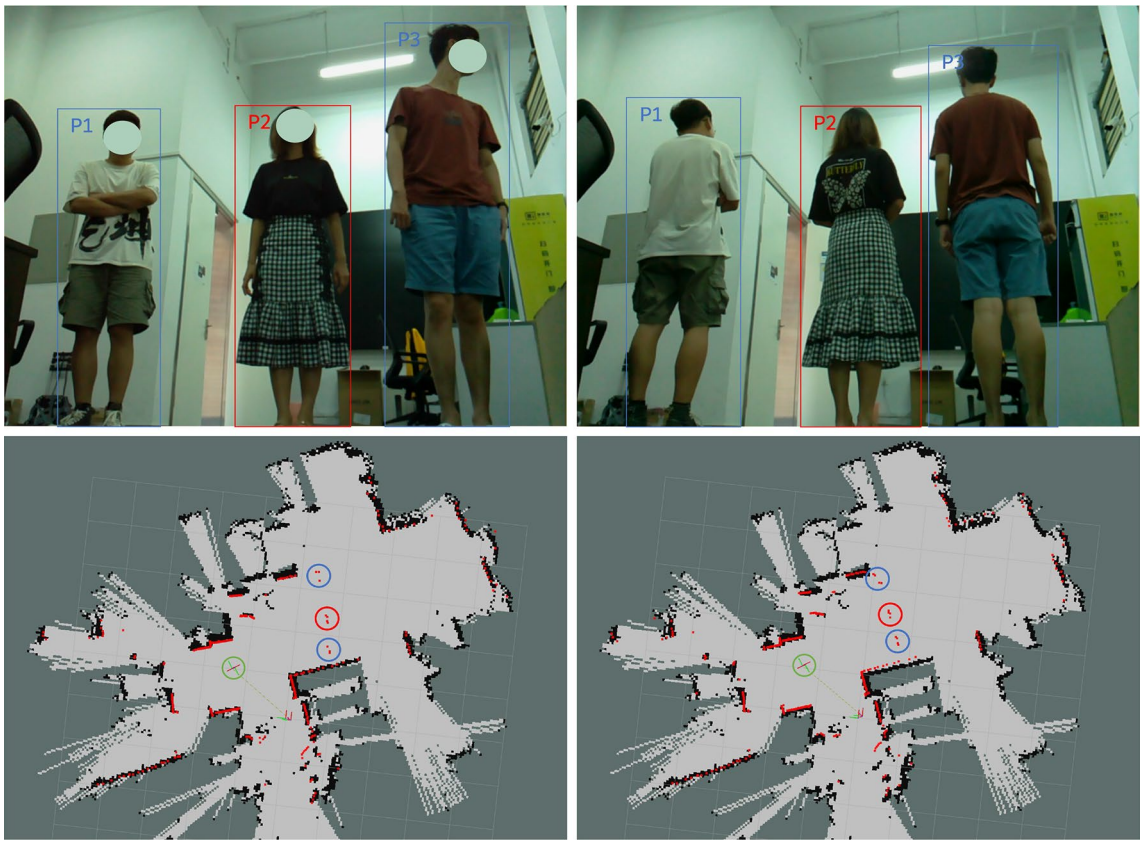


**Fig. 7** Experiment for identifying person in different directions. The figure visualizes RGB results as well as corresponding maps simultaneously. Red, blue, and green circles represent target person, other person, and the position of the car. Please zoom in for more details (colour figure online)

**Fig. 8** Experiment for identifying and navigating the target person. Red, blue, and green circles represent the target person being followed, other people, and the position of the car itself. Please zoom in for more details (colour figure online)

(Redmon and Farhadi 2018) is applied to locate persons, and our proposed HENet is used to re-identify the target person that appeared in the reference frame. As shown in Fig. 7, our proposed method can correctly identify the target person (marked as P2 in red color) even though the person has an opposite direction relative to the reference frame. Thirdly, wireless data transmission is used to transfer location information of the target person to the robotic car. As shown in Fig. 8, we further conduct a robotic navigation experiment in a real complex scene, where factors such as

multiple people, occlusion, and long-distance movement are taken into account. Results indicate that our method can track the target person well over a long distance in the complex multi-person scene while dealing with the occlusion problem. Specifically, we control the movement of the car through the central coordinates of the target person, so as to make the target person in the visual center as much as possible. Accurate real-time location information helps the robotic car navigate and track the target person, which means that the visual information supplied by HENet is significant for practical application, e.g., robotic navigation.

## 5 Conclusion

This paper proposes a novel end-to-end HENet for ReID task, which learns hierarchical global, partial, and recovery features ensemble. Different loss combinations are applied to different branches during the training stage for obtaining more discriminative features without increasing the model complexity. We further propose a new RPE data augmentation method to reduce the impact of irregular occlusions, which improves the network's performance and robustness. Extensive experiments demonstrate the effectiveness and efficiency of our approach, which is more suitable for practical application. We further deploy HENet on a robotic car, in which the experiment demonstrates the signification of our method for practical application such as robotic navigation.

In the future, we will combine our approach with an attention-based method to learn more discriminative features, as well as explore lightweight models so that they can be deployed and run directly on mobile robots.

**Acknowledgements** This work is partially supported by the National Natural Science Foundation of China (NSFC) under Grant no. 61836015 and the Fundamental Research Funds for the Central Universities (2020XZA205).

## References

Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 403–412 (2017)

Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale repre-sentations. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 2590–2600 (2017)

Fu, Y., Wei, Y., Zhou, Y., Shi, H., Huang, G., Wang, X., Yao, Z., Huang, T.: Horizontal pyramid matching for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8295–8302 (2019)

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)

Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International workshop on similarity-based pattern recognition, pp. 84–92. Springer (2015)

Lawen, H., Ben-Cohen, A., Protter, M., Friedman, I., Zelnik-Manor, L.: Compact network training for person reid. In: Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 164–171 (2020)

Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2285–2294 (2018)

Liu, H., Feng, J., Qi, M., Jiang, J., Yan, S.: End-to-end comparative attention networks for person re-identification. IEEE Trans. Image Process. **26**(7), 3492–3506 (2017)

Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., Hu, J.: Pose transferrable person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4099–4108 (2018)

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)

Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp. 618–626 (2017)

Song, C., Huang, Y., Ouyang, W., Wang, L.: Mask-guided contrastive attention model for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1179–1188 (2018)

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)

Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European conference on computer vision (ECCV), pp. 480–496 (2018)

Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM international conference on Multimedia, pp. 274–282 (2018)

Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 79–88 (2018)

Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1568–1576 (2017)

Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European conference on computer vision, pp. 499–515. Springer (2016)

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)

Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: 2014 22nd International Conference on Pattern Recognition, pp. 34–39. IEEE (2014)

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 (2017)

Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Densely semantically aligned person re-identification. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition, pp. 667–676 (2019)
Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp. 3219–3228 (2017)
Zheng, L., Huan g, Y., Lu, H., Yang, Y.: Pose-invariant embedding for deep person re-identification. IEEE Trans. Image Process. **28**(9), 4500–4509 (2019)
Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J.: Joint discriminative and generative learning for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2138–2147 (2019)
Zheng, Z., Zheng, L., Yang, Y.: Pedestrian alignment network for large-scale person re-identification. IEEE Transactions on Circuits and Systems for Video Technology 29(10), 3037–3045 (2018)
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)
Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13001–13008 (2020)
Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camstyle: A novel data augmentation method for person re-identification. IEEE Trans. Image Process. **28**(3), 1176–1190 (2018)
Zhu, F., Kong, X., Zheng, L., Fu, H., Tian, Q.: Part-based deep hashing for large-scale person re-identification. IEEE Trans. Image Process. **26**(10), 4806–4817 (2017)



**Jiangning Zhang** received the B.S. degree in electronic information from Wuhan University, Wuhan, China, in 2017. He is currently working toward the Ph.D. degree in electronic information with the Institute of Cyber Systems and Control, Department of Control Science and Engineering, Zhejiang University, Hangzhou, China. His major research interests include generative adversarial network (GAN), audio signal processing, and neural architecture search (NAS).

**Chao Xu** is pursuing his M.S. degree in College of Control Science and Engineering, Zhejiang University, Hangzhou, China. His major research interests include video generation and video understanding.

**Xiangrui Zhao** is pursuing Ph.D. degree in College of Control Science and Engineering, Zhejiang University after getting the B.S. degree in automation from Huazhong University of Science and Technology in 2018. His major research interests include machine learning in sensor fusion and SLAM.

**Liang Liu** is a fourth-year Ph.D. student in APRIL Lab at Zhejiang University. His main research interest centers on deep learning and computer vision tasks. Currently, he is focusing on self-supervised learning in visual geometry estimation.

**Yong Liu,** Ph.D., Professor of Institute of Cyber-Systems and Control, Zhejiang University. His main research interests include intelligent robot systems, robot perception and vision, deep learning, big data analysis, and multi-sensor fusion.

**Jinqiang Yao,** Master`s Degree, engineer. Deputy Chief engineer of Zhejiang Communications Group Inspection Technology Co.,Ltd., member of China Computer Society. His research directions include Traffic Internet of things, intelligent control, and traffic informatization.

**Zaisheng Pan** graduated from the bachelor's degree of industrial automation of Zhejiang University and the master's degree of industrial automation of Zhejiang University. He works in the College of Control Science and Engineering of Zhejiang University and engages in research and development.