Incremental Generative Occlusion Adversarial Suppression Network for Person ReID

Cairong Zhao, Xinbi Lv, Shuguang Dou, Shanshan Zhang, *Member, IEEE*, Jun Wu, *Senior Member, IEEE*, and Liang Wang, *Fellow, IEEE*

***Abstract*— Person re-identification (re-id) suffers from the significant challenge of occlusion, where an image contains occlusions and less discriminative pedestrian information. However, certain work consistently attempts to design complex modules to capture implicit information (including human pose landmarks, mask maps, and spatial information). The network, consequently, focuses on discriminative features learning on human non-occluded body regions and realizes effective matching under spatial misalignment. Few studies have focused on data augmentation, given that existing single-based data augmentation methods bring limited performance improvement. To address the occlusion problem, we propose a novel Incremental Generative Occlusion Adversarial Suppression (IGOAS) network. It consists of 1) an incremental generative occlusion block, generating easy-to-hard occlusion data, that makes the network more robust to occlusion by gradually learning harder occlusion instead of hardest occlusion directly. And 2) a global-adversarial suppression (G&A) framework with a global branch and an adversarial suppression branch. The global branch extracts steady global features of the images. The adversarial suppression branch, embedded with two occlusion suppression module, minimizes the generated occlusion's response and strengthens attentive feature representation on human non-occluded body regions. Finally, we get a more discriminative pedestrian feature descriptor by concatenating two branches' features, which is robust to the occlusion problem. The experiments on the occluded dataset show the competitive performance of IGOAS. On Occluded-DukeMTMC, it achieves 60.1% Rank-1 accuracy and 49.4% mAP.**

***Index Terms*— Batch-based incremental occlusion, occlusion suppression, occluded person re-identification.**

Manuscript received May 25, 2020; revised December 24, 2020 and February 6, 2021; accepted March 27, 2021. Date of publication April 6, 2021; date of current version April 12, 2021. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62076184, Grant 61673299, Grant 61976160, and Grant 61573255; in part by the Key Laboratory of Advanced Theory and Application in Statistics and Data Science, East China Normal University, Ministry of Education, through the Open Research Fund; and in part by the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiantao Zhou. *(Cairong Zhao, Xinbi Lv, and Shuguang Dou contributed equally to this work.) (Corresponding author: Cairong Zhao.)*

Cairong Zhao, Xinbi Lv, Shuguang Dou, and Jun Wu are with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China (e-mail: zhaocairong@tongji.edu.cn).

Shanshan Zhang is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China, and also with the Jiangsu Key Laboratory of Image and Video Understanding for Social Security, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: shanshan.zhang@njust.edu.cn).

Liang Wang is with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wangliang@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TIP.2021.3070182

I. Introduction

PERSON Re-Identification (re-id) aims to match pedestrian images captured from non-overlapping camera views and has significantly progressed in recent years. Existing approaches [16], [18], [24], [29], [30], [11], [36]–[38] focus on the holistic person images that cover a full glance of one person. However, in the realistic scenario, a person may be occluded by walking out of the field of the camera's view or static obstacles (e.g., cars, trees, walls, other persons), as shown in Fig. 1 (b)-(f). Then, the camera only captures the partial pedestrian. In this case, the retrieval performance of those methods degrades significantly. Both occluded re-id and partial re-id have gradually attracted researchers' attention to address the occlusion problem.

Compared with holistic re-id, occluded re-id and partial re-id are more challenging. This is due to the image suffering occlusion and containing less discriminative pedestrian information, which leads to matching the wrong persons. Besides, less pedestrian information usually causes a spatial misalignment of the pedestrian's non-occluded body region between different images, which provokes distracting noises for part-to-part matching. Generally, occluded re-id mainly aims to solve the interference of occlusions, where partial re-id aims to address the problem of spatial misalignment. Recently, some works have been proposed to solve the occluded/partial re-id.

On occluded re-id works, Zhou *et al.* [1] proposed an Attention Framework of Person Body (AFPB). AFPB automatically generates artificial occlusions for holistic images and uses multi-task losses to optimize the network to discriminate a person's identity and whether it is occluded. He *et al.* [2] proposed an alignment-free matching approach called Foreground-aware Pyramid Reconstruction (FPR). The approach calculates the error of reconstruction over spatial pyramid features to measure similarities between two persons and ignore the variation in the feature's size. Miao *et al.* [3] proposed the Pose-Guided Feature Alignment (PGFA) framework. PGFA utilizes the pose landmarks of humans to guide the network to focus on human non-occluded regions, and only matches shared visible regions for retrieval. In [4], Wang *et al.* proposed that learning high-order relation and topology information between occluded image pairs could determine their discriminative features and robust alignment. In [15], Zhuo *et al.* propose a teacher-student learning framework to learn a robust occlusion model from the full-body

Fig. 1. Illustration of person re-id: (a) holistic re-id, (b) - (c): partial re-id, (d) to (f): occluded re-id.

person domain to that of an occluded person. MHSA [48] proposes a Multi-Head Self-Attention branch and an Attention Competition Mechanism to prune unimportant information and capture key local information from occluded person images. HPNet [52] extracts part-level features and predicts visibility of each part, based on human parsing, to reduce background noise and correct alignment relationships between part-to-part.

On partial re-id works, Zheng *et al.* [5] propose a local patch-level matching model called Ambiguity-sensitive Matching Classifier (AMC). They also introduce a global part-based matching model called Sliding Window Matching (SWM). He *et al.* [6] propose an alignment-free method called Deep Spatial Feature Reconstruction (DSR). DSR sparsely reconstructs the spatial probe maps from the spatial maps of gallery images faster than the SWM. He *et al.* [7] further utilize a dictionary learning-based Spatial Feature Reconstruction (SFR) to match different size feature maps. Sun *et al.* [8] propose a Visibility-aware Part Model (VPM) to extract region-level features and compare two images by their shared regions. Based on VPM, Gao *et al.* [41] propose that learn discriminative features with pose-guided attention and self-mines part visibility.

To address the occlusion problem, existing methods continuously attempt to design complex modules to capture implicit information (such as pose landmarks and mask maps). This is done to force the network to focus on discriminative features on non-occluded body regions and further realize matching under spatial misalignment. In general, data augmentation does not require any extra parameter learning. It is an effective method to enhance the robustness of the model to variations of data, including occlusion. However, it brings limited performance improvement due to the random generation strategy for hard training samples. Moreover, traditional single-based data augmentation methods conduct an operation on the single 2D image, such as random cropping [25], random cutout [9], and random erasing [10], sometimes makes the training data more complex and diverse. The training, consequently, becomes harder to converge.

Inspired by the easy-to-hard learning strategy and the idea of antagonism, we propose a novel Incremental Generative Occlusion Adversarial Suppression (IGOAS) network to tackle this challenging problem. Specifically, we first propose an incremental generative occlusion (IGO) block. Unlike the traditional single-based data augmentation methods, it adopts an easy-to-hard approach to generate occluded data instead of random, which makes the network more robust to occlusion by gradually learning harder occlusion instead of hardest occlusion directly. It also controls a uniform size and position of occlusion in each batch image. Secondly, a G&A framework is proposed to make the model ignore the generated occlusion region that producing an adversarial process. It contains a global branch and an adversarial suppression branch. The generated occluded data are further input into this framework for feature extraction and learning. We use ResNet-50 [13] Stage 1, 2, 3 as the backbone for feature extraction. In the global branch, a ResNet-50 baseline is continuously adopted to extract steady global features.

Some attention works [14], [22]–[24], [32]–[35], [39], [40] effectively exploit the attention mechanism to capture and focus on attentive regions. For example, Chen *et al.* [43] propose a spatial-temporal attention-aware learning (STAL) method to learn the quality scores of these spatial-temporal video units, which aims to attend to the salient parts of persons in videos. Zhang *et al.* [44]. propose a multi-scale spatial-temporal attention (MSTA) model to exploiting the importance of local regions of each frame to the whole video representation in both spatial and temporal domains.

The purpose of the attention mechanism is to focus on an important region in the image, such as the pedestrian region in the person re-id task. However, without the pixel-level label, the attention module is difficult to distinguish the foreground (pedestrian) from the background. In our network, we artificially generate occlusion masks and regard them as the background. Therefore, the idea of solving the problem is to ignore the region of the occlusion mask and focus more on the foreground. Unlike the above methods, we propose an Occlusion Suppression Module to suppress the occlusion region and learn refined features of non-occluded regions of the pedestrian. Finally, we strengthen the pedestrian feature descriptor by concatenating two branches' features. The experiments on occluded datasets show the competitive performance of the IGOAS. Finally, the contributions of this paper are as follows:

1) We propose an incremental generative occlusion adversarial suppression network. The IGOAS network first generates easy-to-hard occluded data through the IGO block and then suppresses the generated occluded region with the adversarial suppression branch. In this process of adversarial learning, the IGOAS learns a more discriminative and robust feature for the occlusion problem.
2) We propose an incremental generative block to generate easy-to-hard occlusion data. It makes the network more robust to occlusion by gradually learning harder occlusion instead of hardest occlusion directly.
3) We develop an occlusion suppression module in the G&A framework. By suppressing the occlusions, our model can focus less on the background and more on the foreground.
4) The proposed approach achieves superior performance on two occluded datasets—Occluded-DukeMTMC and

Occluded-REID, and competitive performance on two holistic datasets—Market-1501 and DukeMTMC-reID.

The remainder of the paper is organized as follows: In Section II, we review related works about occluded re-id and random data augmentation methods. Section III elaborates on the proposed incremental generative occlusion adversarial suppression network. Section IV presents the experimental results of the comparisons and evaluations. Finally, a conclusion is drawn in Section V.

II. Related Work

The proposed IGOAS network mainly focuses on occluded re-id and the design of the data augmentation method, we briefly review related works in this section.

A. Occluded Re-Id

Recently a few works have attempted to capture implicit information to solve the occlusion problem, including humans pose landmarks, mask maps, and spatial information. The auxiliary information can help the network focus on discriminative features on human non-occluded body regions and realize effective matching under spatial misalignment. However, those models usually contain complex modules. The FPR [2] consists of three components: 1) a fully convolutional network (FCN), 2) a multi-size pyramid pooling, and 3) a complex foreground probability generator. PGFA [3] contains a partial feature branch and a complex pose-guide global feature branch. The pose-guide global feature branch needs a pose estimator to detect human landmarks and guide robust feature representations. The high-order relation framework [4] does not use the pose estimator exclusively. Indeed, it also constructs a high-order relation module and high-order human-topology module for discriminative feature representation and robust alignment. The PVPM [41] needs to train an extra pose-guided visibility prediction model under a self-supervised before feature learning. The co-saliency network [15] builds a teacher network and a student network to learn a robust occlusion model. MHSA [48] needs 8-head self-attention modules and an attention competition mechanism to filter out attention noise and non-key information. HPNet [52] needs an extra human parsing model (trained by COCO [53]) to extract part-level features and predict the visibility of each part.

In short, existing methods attempt to design complex modules to capture implicit information. They also attempt to learn the refined features of human non-occluded body regions and effectively realize matching. Contrarily, we propose a simple global-background framework, embedded with an occlusion suppression module. It can suppress occlusion's response, learn the refined features of non-occluded regions of the human body, and then extract discriminative features. This achieves competitive performances with the above methods.

B. Data Augmentation Methods

Data augmentation is a standard tool for increasing data, avoiding overfitting, and improving the generalization ability of the network. Existing data augmentation methods for re-id generally fall into two categories: single-based data augmentation and batch-based data augmentation.

1) Single-Based Data Augmentation: Common methods of single-based data augmentation include random rotation [46], Gaussian noise [47], color jittering [25], random flipping [42], random cropping [25], random scale [45], random cutout [9], random erasing [10], random patching [11], DropBlock [12], and so on. The object is the single image for each of these methods. The random cutout, random erasing, and random patching are usually adopted to simulate occlusion. Differently, random cutout masks the target region of the image with zero values. Random erasing erases its pixels with random values. And random patching pastes this region with a patch from a patch pool. Dropblock is a special method that works on the 3D feature level in a feature-extracting network rather than a 2D image. It masks out the feature's target region with zero values, this is similar to random cutout. In a way, single-based data augmentation makes the training data more complex. The training process, then, becomes hard to converge. Two recent methods make attempts to solve this problem. We classify them as batch-based data augmentation methods.

2) Batch-Based Data Augmentation: This refers to making a uniform operation on the feature level for each batch input image, which is implemented in the network. Inspired by DropBlock, Batch DropBlock (BDB) [16] randomly drops the same region of all intermediate features in a batch to reinforce the attentive feature learning. Being different from BDB, Slow-DropBlock (SDB) [17] moves the batch-dropping operation from the intermediate feature layer towards the input images in each batch (image-based dropping). However, both BDB and SDB use the fixed-size dropping region to make data augmentation. The diversity of the training data, consequently, weakens.

Therefore, the above methods usually bring limited performance improvement, partly also due to the random generation strategy of hard training samples. To solve this problem, we propose a batch-based incremental occlusion block, with an easy-to-hard sample generation strategy. It generates easy-to-hard occluded data instead of random, which makes the network more robust to occlusion by gradually learning harder occlusion instead of hardest occlusion directly. It also controls a uniform size and position of occlusion in each batch image, which effectively alleviates the issue of training hard to converge. Finally, our method promotes performance improvement more than the above methods.

III. Incremental Generative Occlusion Adversarial Suppression Network

The network framework of the proposed method is shown in Fig. 2. We first elaborate on the G&A framework and the incremental generative occlusion block in Section III- A, B. Following this, we describe the training strategy in Section III- C. Finally, we compare the proposed method with similar methods in Section III- D.

A. The G&A Framework

As shown in Fig. 2, the G&A framework is composed of a backbone network, a global branch, and an adversarial suppression branch.

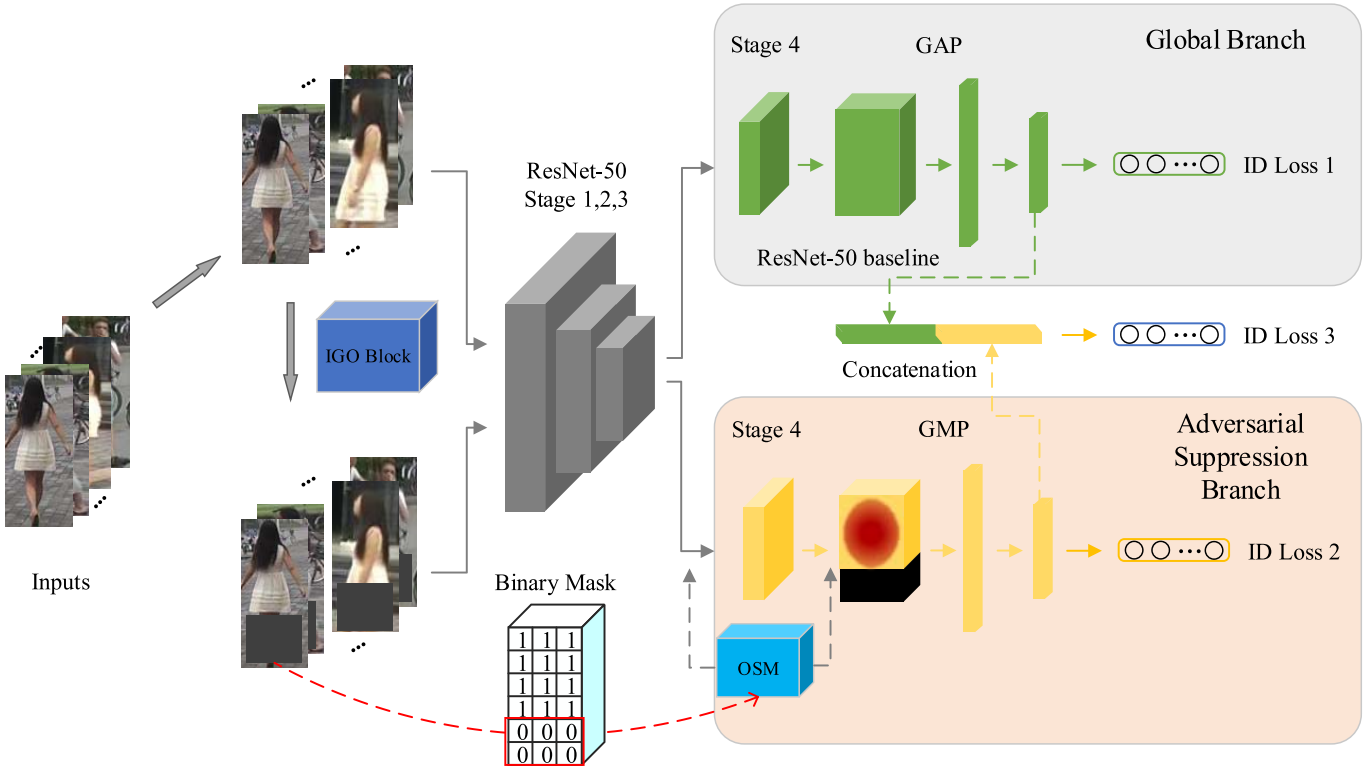


Fig. 2. The flowchart of the proposed IGOAS network. Specifically, in the training phase, the IGO block converts the raw input into occluded data, and then the raw data and the occluded data are entered into the respective branch of the frame for feature extraction. In global branch, we retain the ResNet-50 baseline to extract steady global features of the raw data. In adversarial suppression branch, the OSM and a global max pooling operation are employing to force this branch to suppress the occlusion's response and strengthen discriminative feature representation on non-occluded regions of the pedestrian. Finally, we get a more robust pedestrian feature descriptor by concatenating two branches' features. And in the test phase, the incremental occlusion block won't be performed.

1) Backbone Network: We use the ResNet-50 Stages 1, 2, and 3 as a backbone network. We modify the Stage 4 and do not employ down-sampling operation in the first residual block for a fair comparison with the recent works [3], [8], [18]. In this way, we get a larger feature map of size 2048 × 24 × 8. In our framework, Stages 1, 2, and 3 share weights for fewer parameters learning, and Stage 4 does not. We need Stage 4 to focus on a specific task in each branch.

2) Global Branch: We need the global branch to learn steady global features. Thus, we adopt the ResNet-50 baseline structure as the global branch, considering its competitive performance to re-id. Specifically, following Stage 4, we employ a global average pooling (GAP) to get a 2048-d feature vector. The vector is further reduced to 512-d via a fully connected layer, a batch normalization (BN) layer [26], and a rectified linear unit (ReLU) layer. Finally, a 512-d global feature vector is output for calculating classification loss.

3) Adversarial Suppression Branch: The adversarial suppression branch aims to pay more attention to foreground information by suppressing the response of the generated occlusion region to zero. We develop an occlusion suppression module (OSM) to achieve this goal.

The structure of OSM is shown in Fig. 3. The input feature X is firstly fed in the attention module to obtain the refined feature X'. In this paper, we use CBAM [14] as the attention module of OSM. And then, X_{mask} is obtained by performing an element-wise operation on the binary mask and refined

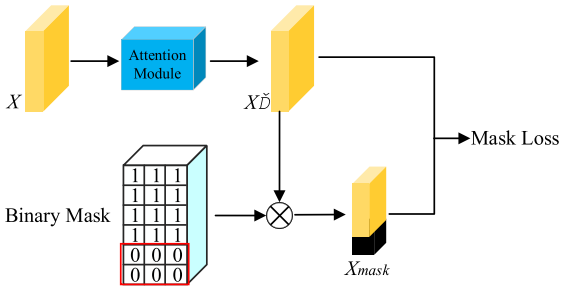


Fig. 3. Structure of OSM. ⊗ denotes element-wise multiplication operation.

features X', where the binary mask is obtained by scaling artificially designed occlusion mask on the image to the same dimension as the input feature X. Finally, X_{mask} supervises X' through the mask loss, so that the model can learn to ignore the background region (black part of X_{mask} in Fig. 3) in the process of back propagation. The mask loss can be formulated as:

$$L_{mask} = MSE(X_{mask}, X') \tag{1}$$

where MSE represents mean square Error. More specifically, the function of mask loss makes the features in the region corresponding to the occlusion mask as zero as possible. Since the position of the occlusion mask is known and random, it can be used as supervision information for the attention module to learn to suppress the generated occlusion's response.

Based on the baseline, we insert the OSM in both Stages 3 and Stage 4 for more discriminative feature learning. Similar to BDB, we employ a global max-pooling (GMP) instead of GAP to get a 2048-d feature vector. The GMP can encourage this branch to identify comparatively weak salient features when the more discriminative body region is occluded. That is, the discriminative body region's feature is easy to select, while the weak-response body region is occluded. However, when the discriminative body region is occluded, the GMP can encourage the branch to strengthen the weak-response body region features. For the GAP, low-response values, except the weak-response features, would still impact the results. This is especially relevant to the feature of the occluded region [16]. In the end, the feature vector is reduced to 512-d for optimization.

Finally, the features from the global branch and adversarial suppression branch are concatenated as the final pedestrian feature descriptor. This strengthens the local non-occluded body region's response based on the steady global representation and suppresses occlusion.

B. Incremental Generative Occlusion Block

To strengthen the network to tackle the occluded re-id, we propose an incremental generative occlusion block based on batch-based data augmentation methods. We randomly generate easy-to-hard occluded data to simulate images that suffer from occlusions, which makes the network more robust to occlusion by gradually learning harder occlusion instead of hardest occlusion directly. In this section, we introduce it in detail.

Given the RGB feature tensor $X \in \mathbb{R}^{B\times C\times H\times W}$ from a batch image, where B, C, H, and W denote each batch's number, the number of channels, the height, and the weight of the feature map. The incremental generative occlusion block randomly generates a uniform occlusion region O in X. Assume that the area of the feature map is $S = W^*H$, we randomly initialize the area of occlusion region O to S_o, where So/S is in the range specified by minimum s_l and maximum s_h. Inspired by easy-to-hard learning strategies, we increase the size of the occlusion mask from small to large. Different from the occlusion generated randomly between the maximum s_l and minimum s_h., the size of the occlusion mask increases with the number of training iterations. The specific formula is shown in step 2 of algorithm 1. The aspect of the occlusion region Ro is randomly initialized between r_l and 1. So the size of O is $h_o = \sqrt{So * Ro}$ and $w_o = \sqrt{So}/Ro$. Then, we randomly initialize a point $P = (x_l, y_l)$ in X, where $0 <= x_l <= H - h_o$ and $0 <= y_l <= W - w_o$. And we set the region, $O = (x_l, y_l, x_l + h_o, y_l + w_o)$, as the target occlusion region. With the selected region in R, G, and B channels, each unit inside is assigned to a random value in [0, 255]. Finally, we get the artificial occluded feature tensor X_o to simulate the images that suffer from occlusions.

The procedure of occlusion simulation is shown in algorithm 1. We visualize the flow of the batch-based occlusion block in Fig. 4. Compared with the single-based methods, the uniform pattern of occlusion in a batch will weaken the diversity of occluded data. Thus, as the number of iterations increases, the block attempts to generate variable-size, variable-position, and easy-to-hard occlusion to enrich the diversity of occlusion patterns. Finally, X_o is entered into the adversarial suppression branch to learn refined features of non-occluded regions.

Algorithm 1 Batch-Based Incremental Generative Occlusion Block

Input: Training data X (Shape: B, C, H, W). Range of occlusion area ratio $[s_l, s_h]$. Range of aspect ratio $[r_l, 1]$. The number of iteration $t \leftarrow 0$.
Output: Artificial occlusion data X_o.
while $t \leq T$ **do**
1. $t \leftarrow t + 1$;
2. Area $S_o = (H \times W) \times (s_l + \frac{(s_h - s_l)t}{t_{max}})$;
3. Randomly generate the aspect ratio R_o = Rand (r_l, 1) ;
4. Calculate occlusion region O's length h_o and width w_o (h_o <H, w_o <W) $h_o = \sqrt{S_o \times R_o}$, $w_o = \sqrt{S_o / R_o}$;
5. Randomly generate the position of the O x_l = Rand (0, H - h_o), y_l = Rand (0, W - w_o), $O = (x_l, y_l, x_l + h_o, y_l + w_o)$;
6. $X(O)$ = Rand (0, 255) ;
7. $X_o = X$;
8. return X_o ;
end while

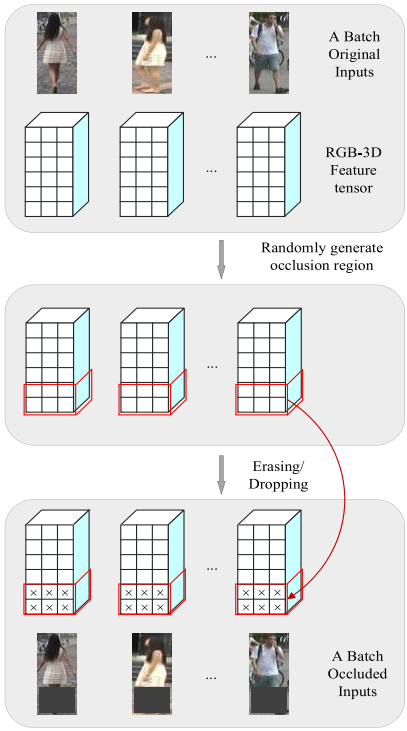


Fig. 4. Simple flow of the batch-based incremental occlusion block in a batch: A cuboid represents the RGB-3D feature tensor of one image. For a batch image, the block randomly generates uniform size and position of occlusion region, such as the red cuboid region. All the units inside will be erasing with random value in [0, 255] to simulate images suffer from occlusions. Notably, the size of occlusion increases with the number of training iterations.

C. The IGOAS Network Training Strategy

In the IGOAS network, the global branch extracts steady global features, whereas the adversarial suppression branch extracts refined features of local non-occluded body regions. Finally, we concatenate the two branches' features to get a

more robust feature descriptor, especially for the occlusion problem.

A multi-classification loss strategy is employed to optimize the network converge faster. During the training phase, we regard re-id as a multi-classification task. Specifically, we utilize softmax loss for the classification and optimization of each branch. Also, we utilize a softmax loss to optimize the fusion feature (concatenated with the two branches' features), considering that it is ultimately used to represent and match. The softmax loss function is presented as follows:

$$L_{softmax} = -\frac{1}{M}\sum_{i=1}^{M} log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T x_i + b_j}}, \tag{2}$$

where the size of each batch and the number of the class is M and C, $x_i \in \mathbb{R}^d$ denotes the ith deep feature in the batch, belonging to the y_ith class, d is the feature dimension. $W \in \mathbb{R}^{d\times n}$ and $b \in \mathbb{R}^n$ are the weights and bias for the last classification layer of the network, which acts as a classifier.

In the IGOAS network, the multi-classification loss L_{cls} is formulated as:

$$L_{cls} = \alpha L_g + \beta L_o + \gamma L_f, \tag{3}$$

where L_g denotes the softmax loss of global branch, L_o denotes the softmax loss of attentive branch, L_f denotes the softmax loss of fusion feature, and α, β, γ denote the weight parameters to balance three losses.

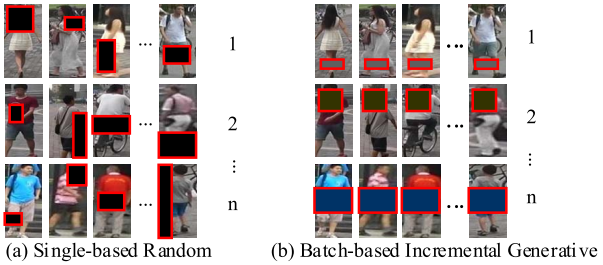
Notably, weight parameters have a great impact on the performance of the network. Empirically setting parameters can improve the efficiency of the experiment. Review the G&B framework, the adversarial suppression branch is more difficult to train and optimize than the global branch due to the diversity of the simulated occlusions. So L_o is typically greater than L_g. Moreover, we have known that the feature extracted from this branch is vital to occluded re-id because we need it to suppress occlusion and focus on the attentive feature of the human non-occluded body region. Therefore, we hope the network gives more attention to the adversarial suppression branch, which means $\beta > \alpha$. For γ, we argue that it is less than either α or β, given that the fusion feature is concatenated by the two branches' features. And the network should give little attention to it. Through the analysis, we finally assign adaptive weights for α, β, γ by using the softmax weight distribution strategy [27] as follows:

$$\begin{cases} \alpha = \dfrac{e^{L_g}}{e^{L_g} + e^{L_o} + e^{L_f}}, \\ \beta = \dfrac{e^{L_o}}{e^{L_g} + e^{L_o} + e^{L_f}}, \\ \gamma = \dfrac{e^{L_f}}{e^{L_g} + e^{L_o} + e^{L_f}}, \end{cases} \tag{4}$$

The results, in Section IV, demonstrate the favorable properties of our adaptive-weight distribution strategy.

Finally, the IGOAS network jointly optimizes the multi-classification loss and mask loss. The total loss is formulated as:

$$L = L_{cls} + L_{mask}, \tag{5}$$



(a) Single-based Random Occlusion Block (b) Batch-based Incremental Generative Occlusion Block

Fig. 5. Comparison of (a) single-based random occlusion block, (b) batch-based incremental occlusion block. In (a), each data in the batch suffers from a variable-size and variable-position occlusion. In (b), all data in the batch suffer from occlusions with a uniform size and position. But as the number of iterations increases, it allows to generate variable-size, variable-position, and easy-to-hard occlusions.

D. Model Comparison

In this section, we compare the IGOAS network with similar methods in network framework and data augmentation.

1) Comparison in Terms of Network Framework: Instead of the complex auxiliary module (such as the pose estimator [3], [4], feature pyramid reconstruction module [2], [6], [7], teacher-student sub-network [15], and high-order relation module [4]), we use a simple ResNet-50 baseline framework and a lightweight adversarial suppression branch (embedded with two suppressing occlusion module) to construct the G&A framework, which requires less parameter learning. The framework can effectively suppress occlusion's response and strengthen the refined features of the non-occluded region, without mining extra implicit information.

2) Comparison in Terms of Data Augmentation: The single-based random occlusion method is visualized in Fig. 5(a). It randomly generates different occlusion for each sample, making the training set more complex and diverse. Sometimes the training process is hard to converge. To address this problem, BDB and SDB do the uniform occlusion operation in the feature level for batch inputs. The overall frameworks of them are shown in Fig. 6(a)-(b). Specifically, BDB randomly drops the same region of all input deep features (deep feature layer) to reinforce the attentive feature learning of local regions. SDB moves the dropping operation from the deep feature layer towards the input image. However, both of them generate fixed-size and patternless occlusions to make augmentation. This limits the performance improvement because the diversity of the occlusion pattern has been weakening.

Unlike BDB or SDB, our batch-based incremental generative occlusion block (as shown in Fig. 6(d)) can generate, variable-position, and easy-to-hard occlusions, as the number of iterations increases, to enrich the diversity of occlusion patterns. And the images in each batch suffer from a uniform size and position of occlusion.

The batch-based random erasing (BRE) block is the basis of the proposed IGO block and shows in Fig. 6(c). BRE block has a completely random size and position, however, the random size is not necessarily the best choice. After a large number of experiments, we found that the aspect ratio, replaced pixel value, and size of the occlusion mask has a significant impact on the performance, especially on the mAP.

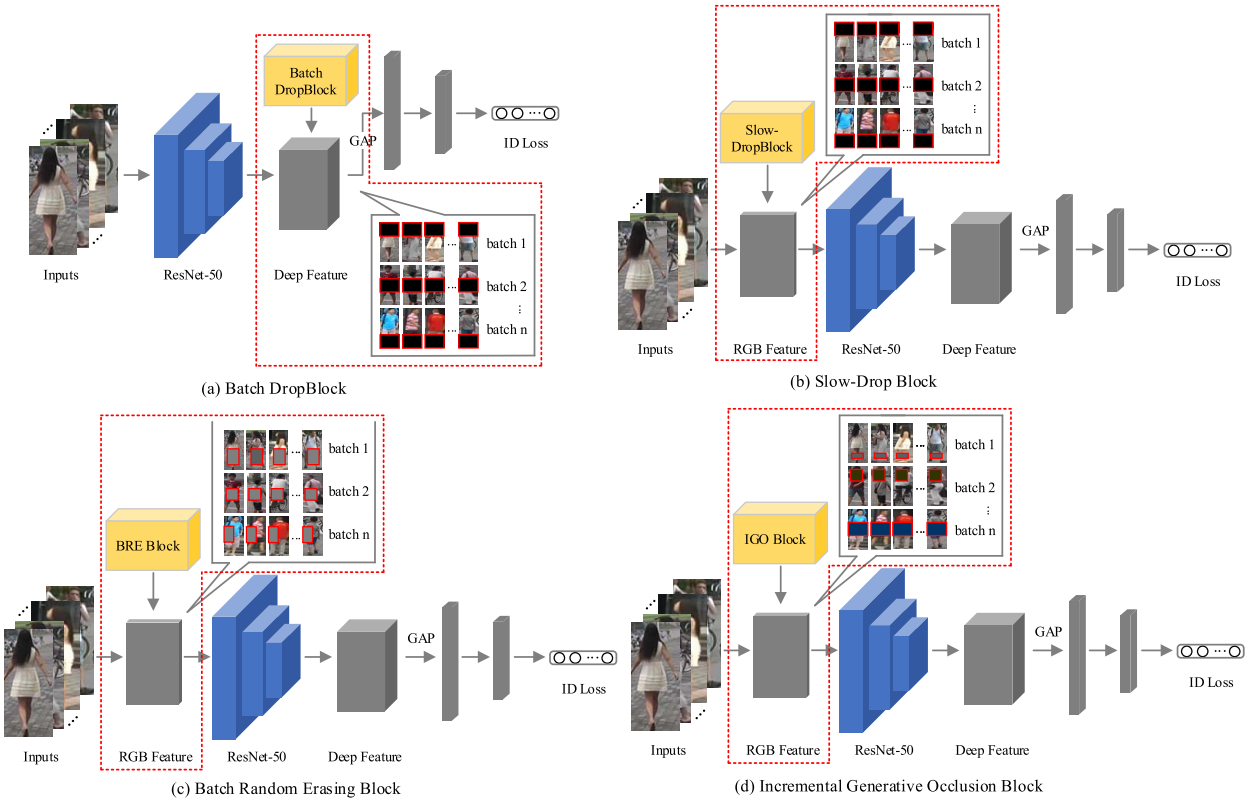


Fig. 6. Comparison of (a) Batch DropBlock, (b) Slow-Drop Block, (c) Batch-based Random Erasing Block and (d) Incremental Generative Occlusion Block. In (a), batch dropblock randomly drops the same region (fixed-size) of all deep features to reinforce the attentive feature learning on local regions. In (b), slow-dropblock moves the dropping operation from the deep feature layer towards the input images ensure inputs diversity for feature learning. In (c), batch-based random erasing block generates the occlusion mask of random size and random position and replaces the original image with mean of ImageNet. In (d), our incremental occlusion block generates variable-position and easy-to-hard occlusions to enrich the diversity of occlusions, more various images under occlusion can be generated. The easy-to-hard learning strategy also make the network more robust to occlusion by gradually learning harder occlusion instead of hardest occlusion directly.

Compared with the BRE block, IGO has made improvements in three aspects. The first is the aspect ratio. Random aspect ratios need to be limited. Specifically, the width is larger than the length for the proposed framework in person re-id task has a better effect. One possible reason is that because our model does not use the complex model to obtain the pose information, the rectangle occlusion mask may block more human information than the wide one, thus affecting the performance of our model. The second aspect is the replaced pixel value. The original version of random erasing uses the mean value of ImageNet. In the three channels of the image, IGO uses different random values of 0-255 as the replacement pixel values. Finally, and most importantly, we design a function to control the size of the occlusion mask increases with the iteration of training. From simple occlusion data to difficult occlusion data, the model can adapt better than random occlusion or large occlusion.

Finally, the blocks with an easy-to-hard learning strategy make the network more robust to occlusion by gradually learning harder occlusion instead of hardest occlusion or random occlusion directly. In Section IV, we compare with related works and show the superiority of our method.

IV. Experiments

In this section, we report on experiments to evaluate our method and compare it with the state-of-the-art methods. To prove the effectiveness of our proposed method, we propose a basic version of the IGOAS network. On the one hand, we directly use a batch-based random erasure block to generate an occlusion mask. On the other hand, we use CBAM replacement in the position of OSM in the network. We call this combined method BRE+CBAM and compare the proposed method with BRE+CBAM on the four datasets in the following.

A. Datasets and Evaluation Measures

We conduct experiments on four representative datasets Occluded-DukeMTMC, Occluded-REID, Market-1501, and DukeMTMC-reID. The details are shown in Table I.

1) Occluded-DukeMTMC: [3] is a large-scale occluded person dataset derived from DukeMTMC-reID that contains 15,618 training images, 17,661 gallery images, and 2,210 query images. Both of which are occluded person images. Fig. 7(a) shows some example images from this dataset. Experimental results on the dataset demonstrate the superiority of the IGOAS for the occluded re-id task.

2) Occluded-REID: [1] is captured by the mobile camera on campus, consist of 2,000 annotated images belonging to 200 identities. Among the dataset, each identity has 5 full-body person images and 5 occluded person images with different types of occlusions. Example images from this dataset are shown in Fig. 7(b). The experiment on the dataset is repeated 10 times to obtain the average results.

TABLE I

DATASET DETAILS. WE EXTENSIVELY EVALUATE THE PROPOSED METHOD ON 4 DATASETS, INCLUDING 2 OCCLUDED AND 2 HOLISTIC

Dataset	Train Nums (ID/Image)	Test Nums (ID/Image)	
		Gallery	Query
Occluded-DukeMTMC	702/15,618	1,110/17,661	519/2,210
Occluded-REID	100/1000	100/500	100/500
Market-1501	751/12,936	750/19,732	750/3,368
DukeMTMC-reID	702/16,522	1,110/17,661	702/2,228

Fig. 7. Example images from: (a) Occluded-DukeMTMC, (b) Occluded-REID, (c) Market-1501, and (d) DukeMTMC-reID.

3) Market-1501: [19] is one of the largest benchmark holistic person datasets, which contains 32,668 images of 1,501 identities from six camera views. A maximum of six cameras captures each identity. There are 751 identities in the training set and 750 identities in the testing set. Some examples are shown in Fig. 7(c).

4) DukeMTMC-reID: [20], [21] contains 36,411 images with 1,812 identities captured from eight different viewpoints. Specifically, there are 16,522 images with 702 identities for training, 17,661 images with 1,110 identities in the gallery, and other 2,228 images with 702 identities for the query. Example images from this dataset are shown in Fig. 7(d).

5) Evaluation Metrics: We use Rank-1 and mean average precision (mAP) to evaluate the performance of the proposed method. All the experiments are performed in a single query setting.

B. Implementation Details

We conduct experiments based on Torchreid [31], a popular framework for deep-learning re-id in Pytorch. In the experiments, the input images are resized to 384 × 128 and augmented by random flipping. We set the batch size to 64 and the training epoch number to 90. The standard Adam [28] algorithm is adopted for fast and robust backpropagation and loss convergence. The learning rate of the network is initialized at 3e-4 and decaying to its 0.1 at 20 and 40 epochs. In the first five epochs, we keep the base layers Stage 1, 2, 3, 4 are frozen, which means only the following layers participated in the training. After the five epochs, all of the layers in the network are open for training.

We simply insert the batch-based incremental generative occlusion block at the beginning of the adversarial suppression branch. Specifically, we clone a copy of the original inputs that do not share the space and memory with the original inputs. An occlusion simulation operation is implemented on the cloned inputs. Finally, the original inputs are entered into the global branch, and the cloned inputs are entered into the adversarial suppression branch for training, respectively.

Notably, in the incremental generative occlusion block, we set the *sl*, *sh*, *r1* to 0.1, 0.33, and 0.5 respectively. And in the inference phase, the images of the test set are directly input into the network framework to obtain the final feature descriptors without occlusion simulation operations.

TABLE II

COMPARISON WITH STATE-OF-THE-ART METHODS ON OCCLUDED-DUKEMTMC

Method	Rank-1	mAP
PCB [18]	42.6	33.7
DSR [6]	40.8	30.4
SFR [7]	42.3	32.0
Ad-Occ [51]	44.5	32.2
Teacher-S [15]	-	-
PGFA [3]	51.4	37.3
HONet [4]	55.1	43.8
MHSA [50]	59.7	44.8
BRE+CBAM	**60.6**	48.3
IGOAS	60.1	**49.4**

C. Experimental Results

We conduct experiments with state-of-the-art methods on four representative datasets.

1) Results on Occluded-DukeMTMC: Table II shows the result of our method and existing state-of-the-art works. Our IGOAS achieves 60.1% Rank-1 accuracy and 49.4% mAP, which is superior to most of the state-of-the-art methods. Compare with the baseline, the Rank-1 is similar and the mAP is increased by 1.1%. Compared to the strongest competing method MHSA (ArXiv 2020), IGOAS outperforms it by +0.4% Rank-1 and +4.6% mAP. Compared to HONet (CVPR 2020), our IGOAS outperforms it by +5.0% Rank-1 and +5.6% mAP.

2) Results on Occluded-REID: Following [1], [15], we take occluded person images as the probes, full-body person images as the galleries, and randomly select half of the identities for training and the rest for a test. The experiments are repeated 10 times to obtain the mean results. As shown in Table III, the IGOAS achieves 81.1% Rank-1 and 91.6% Rank-5. Compared with baseline, the Rank-1 is increased by 5.9% and the mAP increased by 2.8%. This result is the highest Rank-1 we know of and its mAP is second only to the Teacher-S. The results on two occluded datasets are without utilizing extra implicit information, such as pose landmark and mask map. It illustrates the superiority of the proposed batch-based incremental occlusion block and global-background framework. It needs to be mentioned that HONet and PGFA use the Market-1501 as the training set and all the Occluded-REID as testing set.

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON OCCLUDED-REID

Method	Rank-1	Rank-5
SVDNet [52]	63.1	85.1
REDA [10]	65.8	87.9
MLFN [53]	64.7	87.7
PCB [18]	66.6	89.2
AFPB [1]	68.1	88.3
Teacher-S [15]	73.7	**92.9**
PGFA [3]	-	-
HONet [4]	-	-
BRE+CBAM	75.2	88.8
IGOAS	**81.1**	91.6

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS ON TWO HOLISTIC DATASETS

Method	MARKET-1501		DUKEMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
PCB [18]	92.3	77.4	81.8	66.1
DSR [6]	91.3	75.6	82.4	68.7
SFR [7]	93.0	81.0	84.8	71.2
VPM [8]	93.0	80.8	83.6	72.6
FPR [2]	**95.4**	**86.6**	**88.6**	**78.4**
PGFA [3]	91.2	76.8	82.6	65.5
HONet [4]	94.2	84.9	86.9	75.6
MHSA [50]	94.6	84.0	87.3	73.1
BRE+CBAM	93.6	83.5	83.8	70.9
IGOAS	93.4	84.1	86.9	75.1

This is different from the training scheme we used, so the results of the two methods in Table III are shown in "−".

3) Results on Market-1501 and DukeMTMC-reID: We also evaluate the IGOAS network on two holistic datasets, as shown in Table IV. Our method aims to improve the robustness of the model when the images suffer from occlusions. In the two holistic datasets, the image always covers the full glance of one person. IGOAS does not work as well as expected. But it still achieves comparable performances with related methods, which shows the strong generality for re-id tasks. Compared with the baseline, our method has a significant improvement in DukeMTMC-reID. Compared with HONet, the proposed method achieves similar performance. However, HONet uses pose information and therefore introduces more model parameters. Our method does not use extra information, and only solves the occlusion problem by generating occlusion and then suppressing the generated occlusion.

D. Ablation Study

We conduct extensive ablation studies to analyze each component of the IGOAS on two occluded datasets.

TABLE V
PERFORMANCE WITH EACH BRANCH OF IGOAS

Method	Occluded-DukeMTMC		Occluded-REID	
	Rank-1	mAP	Rank-1	Rank-5
G-F	53.3	38.8	68.8	84.4
A-F	57.0	47.5	73.0	86.0
F-F	**60.1**	**49.4**	**81.1**	**91.6**

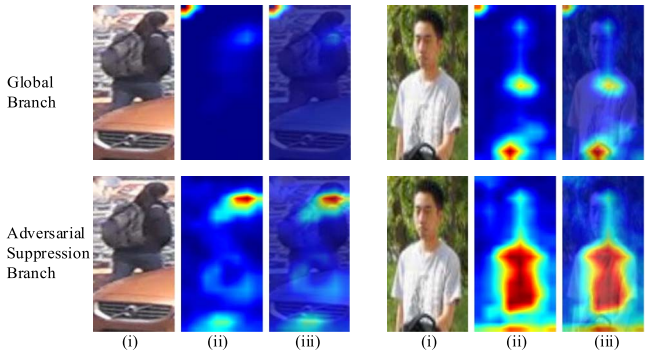


Fig. 8. Visualization of the Stage 4 feature maps learned by different branch. The first and second rows show the results for the global branch and the attentive branch. From left to right, (i) Original images, (ii) Activation map, and (iii) Overlapped image. In the heat map, the response increases from blue to red.

1) G&A Framework: To evaluate the impact of the framework, we conduct a test with each branch's feature representation, including the global branch's feature (G-F), the adversarial suppression branch's feature (A-F), and the fusion feature (F-F). Table V and Fig. 8 show comparisons of the results. A-F are more robust to the occlusion problem, with higher performance by +3.7% Rank-1 on Occluded-DukeMTMC and +4.2% Rank-1 on Occluded-REID. The visualization result in Fig. 8 has validated our intention that the adversarial suppression branch can suppress the occlusion's response and strengthen attentive feature representation, a higher response in the non-occluded body region. Furthermore, the results indicate that the fusion strategy of two branches' features increases the performance improvement. We argue that it strengthens the local non-occluded body region's response based on steady global representation. Finally, we can get a more robust feature representation for the images under occlusions.

We also conduct ablation experiments to evaluate the impact of each component in the adversarial suppression branch, including pooling operation and occlusion suppression module. The results are shown in Table VI. The G&A(+*GAP*) stands for the adversarial suppression branch applies a global average pooling following Stage 4, without any attention module. (+*GMP*) stands for the IGOAS employ a global max-pooling instead of GAP. As is expected, the discriminative body region's feature is easy to select by adopting GMP, which get +1.5% Rank-1 improvement than (+*GAP*). On this basis, we further verify the impact of the attention module. (+*GMP*+*CBAM*) stands for employing with CBAM module. As we can see, the addition of an

TABLE VI
COMPARISON WITH EACH COMPONENT OF ATTENTIVE BRANCH

Method	Occluded-DukeMTMC		Occluded-REID	
	Rank-1	mAP	Rank-1	Rank-5
G&A(+*GAP*)	55.8	44.5	74.6	87.8
G&A(+*GMP*)	57.3	47.2	77.6	89.6
G&A(+*GMP*+*CBAM*)	59.6	48.5	78.2	90.2
G&A(+*GMP*+*OSM*)	**60.1**	**49.4**	**81.1**	**91.6**

TABLE VII
COMPARISON WITH DIFFERENT OCCLUSION OPERATION

Method	Occluded-DukeMTMC		Occluded-REID	
	Rank-1	mAP	Rank-1	Rank-5
G&A	57.6	44.6	74.8	86.8
BDB + G&A	56.1	44.8	77.4	87.8
SDB+ G&A	59.7	48.5	75.0	87.8
BRE+ G&A	58.1	48.1	73.0	86.4
IGO(+*RE*) + G&A	59.1	47.3	73.6	85.2
IGO(+*RD*) + G&A	58.6	47.6	74.2	86.6
IGO(+*BIE*) + G&A	**60.1**	**49.4**	**81.1**	**91.6**
IGO(+*BID*) + G&A	59.2	48.7	78.8	90.2

attention module can improve the performance of the model. (+*GMP*+*CBAM*) gets higher performance by +2.3% Rank-1 on Occluded-DukeMTMC and +0.6% Rank-1 on Occluded-REID. Finally, (+GMP+OSM) gets the highest Rank-1 and mAP, which demonstrates the effectiveness of the OSM in suppressing occlusion and strengthen discriminative pedestrian information.

2) Batch-Based Incremental Occlusion Block: We further conduct experiments with different occlusion methods, contains BDB, SDB, BRE, and IGO, to analyze the impact of the batch-based incremental occlusion block (IGO). The results are shown in Table VII. The G&A stands for the network framework without any occlusion simulation. IGO(+*RE*) stands for the IGO employing single-based random erasing in the occlusion block. (+*RD*) stands for employing with single-based random dropping. (+*BID*) stands for employing with the batch-based incremental drop. And (+*BIE*) stands for employing batch-based incremental erasing. Most single-based and batch-based occlusion simulator methods have improved the performance of the model to the occlusion problem. The validity of the data augmentation method is, therefore, proved. Finally, our incremental occlusion block (IGO(+BIE)) achieves the highest performance among all listed methods on two datasets. In Table VIII, we also compare the performance of the baseline network only global features are used by adopting different data augmentation methods. Incrementally adding the difficulty of batch erasing improves the performance of the baseline network. It outperforms BRE by +8.7% Rank-1 and +7.2% mAP on Occluded-DukeMTMC,

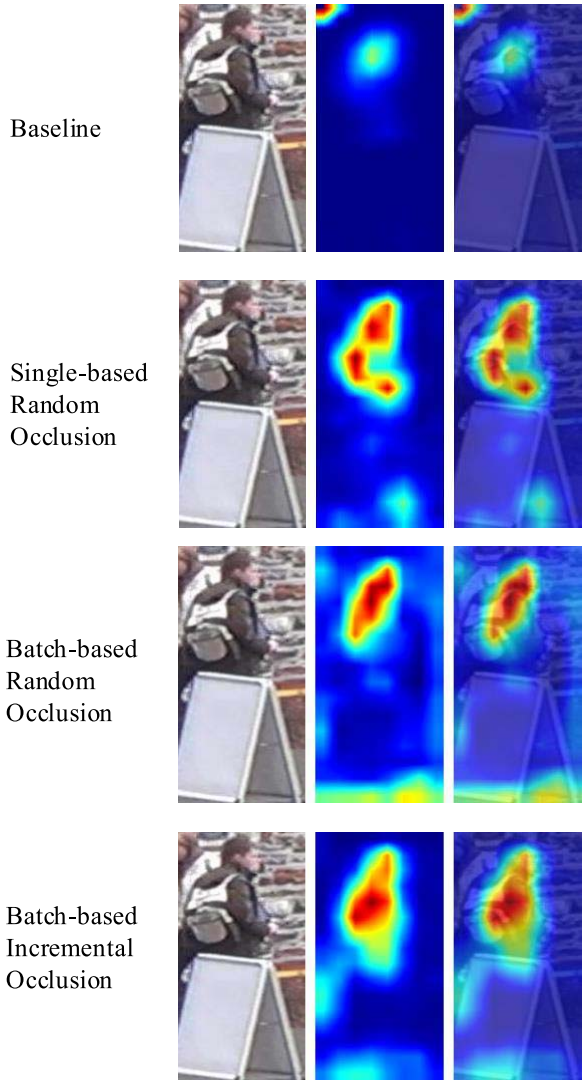


Fig. 9. Visualization of the Stage 4 feature maps learned by different occlusion methods. The first, second, third, and fourth rows show the results for the baseline (ResNet-50) without occlusion simulation, the IGOAS joint with single-based random occlusion, the IGOAS joint with batch-based random occlusion, and the IGOAS joint with batch-based incremental occlusion.

and + 19.4% Rank-1 and +16.4% Rank-5 on Occluded-REID. They prove the superiority and generality of the easy-to-hard learning strategy, which makes the network more robust to occlusion by gradually learning harder occlusion instead of hardest occlusion directly. The visualization comparison of the Stage 4 feature maps in Fig. 9 further shows that our block strengthens the network to suppress occlusion and pay more attention to human non-occluded body region information and less attention to the background.

Easy-to-hard Learning Strategy. We combine two different data augmentation with CBAM for comparison and analyze the easy-to-hard learning strategy joint with attention. As shown in Fig. 10, the single-based method is lower than the batch-based method except for the shock in the early training period. Before the 30th epoch, IGOAS and batch-based combined method are almost the same. This is because, in the early epochs of this training, the IGO block generated only simple and small occlusion masks. After the

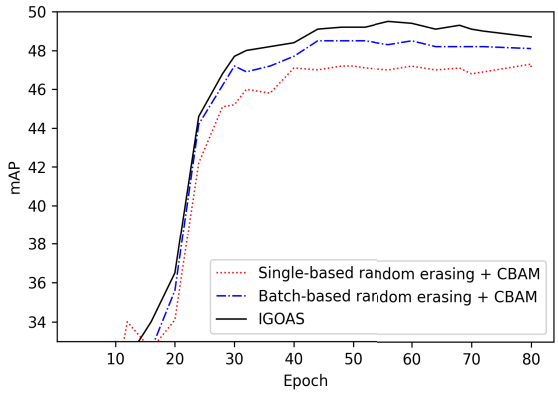


Fig. 10. Comparison of two combined methods and IGOAS on Occluded-DukeMTMC.

TABLE VIII
COMPARISON WITH DIFFERENT OCCLUSION OPERATION ON BACKBONE

Method	Occluded-DukeMTMC		Occluded-REID	
	Rank-1	mAP	Rank-1	Rank-5
BRE + G-F	42.3	33.5	53.0	71.0
BIE + G-F	**51.0**	**40.7**	**72.4**	**87.4**

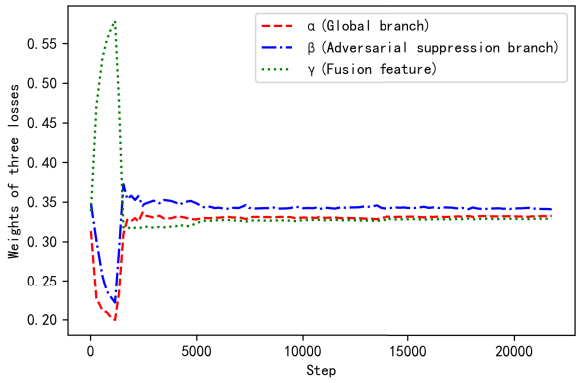


Fig. 11. Weights of the adaptive-weight strategy in the training process.

TABLE IX
COMPARISON WITH DIFFERENT WEIGHTS STRATEGY

Method	Occluded-DukeMTMC		Occluded-REID	
	Rank-1	mAP	Rank-1	Rank-5
IGOAS(+*mw*)	**60.2**	48.6	80.2	90.4
IGOAS(+*aw*)	60.1	**49.4**	**81.1**	**91.6**

30th epoch, the generation of the occlusion mask becomes more difficult, and the suppression effect of OSM is better. At this time, the performance of IGOAS is better than the combined method.

Adaptive-Weight Strategy. We also evaluated the effect of the adaptive-weights strategy in *eq*. (3). Fig. 11 shows the weights of the adaptive-weight strategy during the training process. For the first 5 epochs, the changes in weights are not referenced because some of the network layers are frozen. After the five epochs, all of the layers in the network are open for training. The comparison of weights is $\beta > \alpha > \gamma$. It effectively confirmed the adaptive-weight strategy can optimize the network to give more attention to the adversarial suppression branch. The results in Table IX also confirmed the adaptive-weight strategy achieved higher performance than the mean-weight one. (+*mw*) means optimized by mean-weight strategy, whereas (+*aw*) means optimized by adaptive-weight strategy.

V. CONCLUSION

In this paper, we propose an incremental generative occlusion adversarial suppression network for occluded person ReID. Firstly, a novel batch-based incremental generative occlusion block is employed to generate easy-to-hard occlusion data. It not only makes the network more robust to occlusion by gradually learning harder occlusion instead of hardest occlusion directly or random occlusion size. Then, we proposed a simple global-adversarial suppression (G&A) framework with an occlusion suppression mechanism. In this framework, an adversarial suppression branch, embedded with an occlusion suppression module, effectively suppresses occlusion's response and strengthens attentive feature representation on non-occluded body regions. Finally, a more discriminative pedestrian feature descriptor can be obtained, the experiment results demonstrate that it is robust and effective for occlusion problem.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their critical and constructive comments and suggestions.

REFERENCES

[1] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.
[2] H. Lingxiao, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8450–8459.
[3] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 542–551.
[4] G. Wang *et al.*, "High-order information matters: Learning relation and topology for occluded person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6449–6458.
[5] W. Zheng *et al.*, "Partial person reidentification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4678–4686.
[6] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7073–7082.
[7] L. He, Z. Sun, Y. Zhu, and Y. Wang, "Recognizing partial biometric patterns," 2018, *arXiv:1810.07399*. [Online]. Available: http://arxiv.org/abs/1810.07399
[8] Y. Sun *et al.*, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 393–402.
[9] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*. [Online]. Available: http://arxiv.org/abs/1708.04552
[10] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017, *arXiv:1708.04896*. [Online]. Available: http://arxiv.org/abs/1708.04896

[11] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3702–3712.
[12] G. Ghiasi, T. Y. Lin, and Q. V. Le, "Dropblock: A regularization method for convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2018, pp. 10727–10737.
[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[14] S. Woo *et al.*, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[15] J. Zhuo, J. Lai, and P. Chen, "A novel teacher-student learning framework for occluded person re-identification," 2019, *arXiv:1907.03253*. [Online]. Available: http://arxiv.org/abs/1907.03253
[16] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch DropBlock network for person re-identification and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3691–3701.
[17] X. Wu, B. Xie, S. Zhao, S. Zhang, Y. Xiao, and M. Li, "Diversity-achieving slow-DropBlock network for person re-identification," 2020, *arXiv:2002.04414*. [Online]. Available: http://arxiv.org/abs/2002.04414
[18] Y. Sun *et al.*, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 480–496.
[19] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
[20] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3754–3762.
[21] E. Ristani *et al.*, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 17–35.
[22] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2285–2294.
[23] T. Chen *et al.*, "ABD-net: Attentive but diverse person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8351–8361.
[24] S. Zhou, J. Wang, D. Meng, Y. Liang, Y. Gong, and N. Zheng, "Discriminative feature learning with foreground attention for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4671–4684, Sep. 2019.
[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.
[27] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6036–6046.
[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
[29] C. Zhao, X. Lv, Z. Zhang, W. Zuo, J. Wu, and D. Miao, "Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3180–3195, Dec. 2020.
[30] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global–local-alignment descriptor for scalable person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 986–999, Apr. 2019.
[31] K. Zhou and T. Xiang, "Torchreid: A library for deep learning person re-identification in PyTorch," 2019, *arXiv:1910.10093*. [Online]. Available: http://arxiv.org/abs/1910.10093
[32] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6450–6458.
[33] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.
[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[35] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep Siamese attention networks for video-based person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1412–1424, Jun. 2019.
[36] G. Ding, S. Zhang, S. Khan, Z. Tang, J. Zhang, and F. Porikli, "Feature affinity-based pseudo labeling for semi-supervised person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2891–2902, Nov. 2019.
[37] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 371–381.
[38] J. Guo, Y. Yuan, L. Huang, C. Zhang, J.-G. Yao, and K. Han, "Beyond human parts: Dual part-aligned representations for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3642–3651.
[39] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2119–2128.
[40] R. Zhang *et al.*, "SCAN: Self-and-collaborative attention network for video person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4870–4882, Oct. 2019.
[41] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person ReID," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11744–11752.
[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2014, pp. 1–14.
[43] G. Chen, J. Lu, M. Yang, and J. Zhou, "Spatial-temporal attention-aware learning for video-based person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4192–4205, Sep. 2019.
[44] W. Zhang, X. He, X. Yu, W. Lu, Z. Zha, and Q. Tian, "A multi-scale spatial-temporal attention model for person re-identification in videos," *IEEE Trans. Image Process.*, vol. 29, pp. 3365–3373, 2020.
[45] Y. Liu, Y. Zhong, F. Fei, Q. Zhu, and Q. Qin, "Scene classification based on a deep random-scale stretched convolutional neural network," *Remote Sens.*, vol. 10, no. 3, p. 444, Mar. 2018.
[46] R. Maree *et al.*, "Random subwindows for robust image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 34–40.
[47] D.-H. Shin, R.-H. Park, S. Yang, and J.-H. Jung, "Block-based noise estimation using adaptive Gaussian filtering," *IEEE Trans. Consum. Electron.*, vol. 51, no. 1, pp. 218–226, Feb. 2005.
[48] H. Tan, X. Liu, S. Tian, B. Yin, and X. Li, "MHSA-Net: Multi-head self-attention network for occluded person re-identification," 2020, *arXiv:2008.04015*. [Online]. Available: http://arxiv.org/abs/2008.04015
[49] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5098–5107.
[50] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3800–3808.
[51] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2109–2118.
[52] H. Huang, X. Chen, and K. Huang, "Human parsing based alignment with multi-task learning for occluded person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.
[53] R. A. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7297–7306.

Cairong Zhao received the B.Sc. degree from Jilin University, Changchun, China, in 2003, the M.Sc. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Beijing, China, in 2006, and the Ph.D. degree from the Nanjing University of Science and Technology, Nanjing, China, in 2011. He is currently a Professor with Tongji University, Shanghai, China. He is the author of more than 30 scientific articles in pattern recognition, computer vision, and related areas. His research interests include computer vision, pattern recognition, and visual surveillance.

Xinbi Lv is currently pursuing the master's degree with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include computer vision, deep learning, and person search, in particular, focusing on person re-identification and person search for visual surveillance.

Shuguang Dou is currently pursuing the Ph.D. degree with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include computer vision, deep learning, X-ray, and person re-identification.

Shanshan Zhang (Member, IEEE) received the master's degree in signal and information processing from Tongji University in 2011 and the Ph.D. degree in computer science from the University of Bonn in 2015. From January 2015 to December 2016, she was a Postdoctoral Researcher with the Department of Computer Vision and Multimodal Computing, Max Planck Institute for Informatics, Saarbrucken, Germany. Since December 2016, she has been a Professor with the Nanjing University of Science and Technology, China. Her main research interests include computer vision, pattern recognition, and particularly applications for driverless vehicles.

Jun Wu (Senior Member, IEEE) received the B.Sc. degree in information engineering and the M.Sc. degree in communication and electronic systems from Xidian University, Xi'an, China, in 1993 and 1996, respectively, and the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications, Beijing, China, in 1999. He is currently a Professor with the Department of Computer Science and Technology, Tongji University, Shanghai, China. He joined Tongji University as a Professor in 2010. Before he joined Tongji, he was the Principal Scientist of Huawei and Broadcom. His research interests include wireless communication, information theory, machine learning, and signal processing.

Liang Wang (Fellow, IEEE) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CAS), China, in 2004. From 2004 to 2010, he has been working as the Research Assistant of the Imperial College London, U.K., and Monash University, Australia; a Research Fellow with The University of Melbourne, Australia; and a Lecturer with the University of Bath, U.K. He is currently a Full Professor of the Hundred Talents Program with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, CAS, and the Deputy Director of NLPR. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published at highly ranked international journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and IEEE TRANSACTIONS ON IMAGE PROCESSING; and leading international conferences, such as CVPR, ICCV, and ICDM. He is currently an IAPR Fellow as well as a member of BMVA. He has obtained several honors and awards such as the Special Prize of the Presidential Scholarship of the Chinese Academy of Science.