Adversarially Occluded Samples for Person Re-identification

Houjing Huang [1,2] Dangwei Li [1,2] Zhang Zhang [1,2] Xiaotang Chen [1,2] Kaiqi Huang [1,2,3]
[1] CRIPAC & NLPR, CASIA [2] University of Chinese Academy of Sciences
[3] CAS Center for Excellence in Brain Science and Intelligence Technology
{houjing.huang, dangwei.li, zzhang, xtchen, kaiqi.huang}@nlpr.ia.ac.cn

Abstract

Person re-identification (ReID) is the task of retrieving particular persons across different cameras. Despite its great progress in recent years, it is still confronted with challenges like pose variation, occlusion, and similar appearance among different persons. The large gap between training and testing performance with existing models implies the insufficiency of generalization. Considering this fact, we propose to augment the variation of training data by introducing Adversarially Occluded Samples. These special samples are both ***a) meaningful*** *in that they resemble real-scene occlusions, and* ***b) effective*** *in that they are tough for the original model and thus provide the momentum to jump out of local optimum. We mine these samples based on a trained ReID model and with the help of network visualization techniques. Extensive experiments show that the proposed samples help the model discover new discriminative clues on the body and generalize much better at test time. Our strategy makes significant improvement over strong baselines on three large-scale ReID datasets, Market1501, CUHK03 and DukeMTMC-reID.*

1. Introduction

Image based person re-identification is a fundamental task in video surveillance that aims to retrieve a particular person from a great number of person images that have been detected under various cameras. It was originally proposed in multi-camera tracking but gradually grows into an independent research task due to its complexity as well as wider application in real world [53].

Though it has been researched for years, the meaningful task is still confronted with many challenges. The main obstacles reside in low resolution of images, similar clothes among different persons, background clutter, various view points and body poses, *etc*. What's more, the test persons are never seen during training. In this respect, it can not be overstated that person re-identification is exactly one kind of zero-shot learning [27].

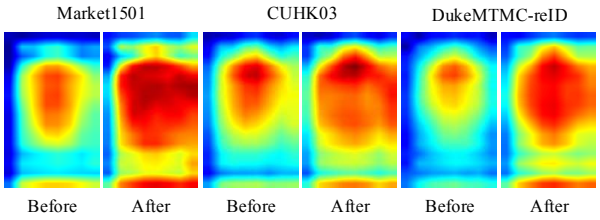

Figure 1: Activation maps (from layer Conv4 of ResNet-50 [15]) averaged over the whole test sets, before and after applying our method. Warmer color denotes higher value. Results on three large-scale datasets Market1501 [52], CUHK03 [20] and DukeMTMC-reID [29] are displayed. The original model mainly focuses on upper regions of the body, our method guides it to also pay attention to other parts *e.g*. pants, shoes and head, for better generalization.

The recent advance of deep neural networks [18, 33, 37, 15] provides a promising way for learning invariance from data and building robust models. Accordingly, there have been some efforts on extracting discriminative features [19, 51, 45, 36], learning a better similarity metric [8, 5], or combining them in an end-to-end model [54]. These works improve the performance of person ReID to some extent but the gap between training and testing performance still remains. One typical phenomenon is that during training the model easily satisfies the training constraint, decreasing the loss function to a relatively low level, and then stays convergent. Whereas the performance on test set is still far from perfect.

Increasing variation in training data is an effective way of improving generalization, especially for deep learning paradigm that benefits from huge data volume [18]. However, unlike other tasks such as object recognition, it's rather expensive to gather an adequately large dataset for ReID due to the cross-camera data collection procedure. Alternatively, data augmentation is an important technique to make full use of the current dataset without extra cost. Common methods [18] in this direction include randomly cropping a

proportion of the training image, mirroring the image, and jittering the pixel values, *etc*.

In this paper, we propose a special type of data augmentation that generates **occluded** samples **adversarial** to the existing ReID models. Specifically, after training a ReID model in the traditional way, we search for discriminative regions[1] on training images with the help of network visualization techniques. These regions are what the model mainly relies on for making decisions. Then we generate new samples by occluding (a proportion of) these regions (maintaining original labels), which we further apply to retrain and improve the ReID model. We argue that this kind of samples are both meaningful and effective. **a) Simulating Occlusion Is Meaningful.** Images of the same person under different cameras tend to have different views, *e.g*. frontal and back, due to which self occlusion naturally exists. In addition, environmental objects like bicycles and other pedestrians also cause some occlusion. These cases are common in real scene and require the model to capture every possible clues for discriminating true positive from distractors. So it's meaningful that we generate occlusion to simulate the real world scenario. **b) Adversary Is Effective to Help the Model Jump out of Local Optimum.** Through visualization, we find that the model is only prone to occlusion in some regions of the images, as shown in Figure 3. From the aspect of training, only by occluding these regions can we provide the model with pressure and thus momentum to evolve and jump out of local optimum. And intuitively, only by occluding those discriminative parts the model previously depends on can we urge it to discover new ones. Thus mining these adversarially occluded samples are necessary.

Extensive experiments demonstrate that adversarial occlusion effectively urges the model to discover new clues from the body and improves its performance, as illustrated in Figure 1 and 7. The proposed method has obvious and consistent improvement on three large-scale datasets for person ReID, *i.e*. Market1501 [52], CUHK03 [20] and DukeMTMC-reID [29].

The contribution of this paper is summarized as follows:

- We analyze the generalization problem of person ReID and propose Adversarially Occluded Samples as a way of data augmentation.
- We conduct extensive experiments and provide some insights into better harnessing these samples.
- The proposed method has significant and consistent improvement on three large-scale datasets.

[1]In this paper, *discriminative regions* and *critical regions* are interchangeably used.

2. Related Work

Person ReID. A person ReID model typically consists of two components, a feature extractor and a similarity metric. Deep learning was first adopted in person ReID by [46] and [20], and it is becoming more and more popular for this task. A line of later works focused on improving either the feature extractor, the similarity metric or both of them. For a better feature extractor, some of them concentrate on learning discriminative latent parts of body, extracting and fusing features from these parts for a robust representation [19, 51, 45]. To achieve a large data volume for training convolutional neural network (CNN), Xiao *et al*. [43] also combined several datasets and proposed a domain aware dropout method to learn a single model from all these samples. With access to larger scale dataset like Market1501 [52], generic object recognition models can also be finetuned as feature extractors for ReID [53]. For a better similarity metric, Ding *et al*. [8] introduced triplet loss [30] into person ReID. Chen *et al*. [5] further analyzed the relationship of intra- and inter-class distance and improved the triplet loss. Zhong *et al*. [57] re-ranked the ranklist as a post-processing method which improved the retrieval performance significantly. It can also be viewed as a type of metric learning. Zheng *et al*. [54] combined verification loss [2] with identification loss and jointly learned a metric in the model.

Data Augmentation. As a free way of generating large number of extra data, data augmentation is beneficial for preventing deep models from overfitting and has been successfully applied to tasks like object recognition [18, 33, 37, 15], object detection [11, 28, 10] and semantic segmentation [26, 4, 16]. Common practices include randomly mirroring images, cropping a portion or slightly jittering pixel values of images. These techniques are also applicable to person ReID for the same purpose. Recently, Zheng *et al*. [56] proposed a novel way to generate new person images using a Generative Adversarial Network (GAN) [12] and these additional images helped improve a ReID model.

What we propose is a special type of data augmentation, which is complementary to these existing methods. For example, we can mirror the image, randomly crop a portion and further occlude a discriminative part, *etc*. to achieve richer variation.

Our work is also inspired by A-Fast-RCNN [41], yet with much difference in implementation. A-Fast-RCNN addressed object detection and simulated all kinds of rare occlusions by occluding some cells of a grid on the feature maps using a learned network. It finetuned the model with occlusion masks and learned the mask net simultaneously in an adversarial manner. In terms of using adversarial masks to improve the baseline model, our method is similar to theirs. However, we are different in the type of masks and how to generate them. We use a continuous re-

gion hoping that it can occlude a relatively complete part of the body, instead of some small and scattered patches. Besides, we generate masks in an offline manner through the help of network visualization techniques, which is easier to handle than the online manner.

A parallel work by Zhong *et al*. [58] occludes a rectangular part of images during training, with the position and size of the rectangle randomly selected from a range. This shares much similarity with our Random sampling method in Section 4.7.

Network Visualization. The black-box characteristic of deep neural network necessitates some technique to reveal why the network works or fails and how it can be improved. Works in this direction include Deconvolution [48], Guided Backpropagation [34], Class Activation Map (CAM) [59], Grad-CAM [32], *etc*. They typically visualize the weights of filters, activation maps, images that activate a certain neuron, the regions and the patterns of input images that the model depends on when making decisions, *etc*. We utilize this technique to determine which parts of the input images are critical for the model's decision.

Adversarial Examples (Samples). Related works [38, 13] discover that by intentionally imposing some noises to the input images, the recognition network can be fooled totally. These perturbed samples are called *adversarial examples*. Researchers combined these samples with clean data to train the recognition model and found that it helped the model perform better on the clean data. The samples we generate are also hard (or adversarial) to the network, but with more analogy to real world scenarios. Similarly, our samples also prove to be beneficial for improving the original model.

3. Method

The proposed framework includes three phases. First, we train a ReID model till convergence. Then we utilize some network visualization technique to determine which regions are critical for the model to make decisions. Finally, we combine original and adversarially occluded samples (occluding discriminative regions while maintaining original labels) to train the ReID model as in the first phase. Each phase can have various options, yet we focus on one concrete implementation in this paper, as shown in Figure 2.

3.1. Baseline ReID Model

We adopt the Identification Model [53, 57] as our baseline. During training, it takes the same structure as generic object recognition models and treats each person as a class. During testing, the classifier, *i.e*. the final FC layer, is stripped off and the remaining layers are used as a feature extractor.

Suppose that we have a training set $\mathcal{I}$ containing C persons, *i.e*. C classes, with N images in total. Each training

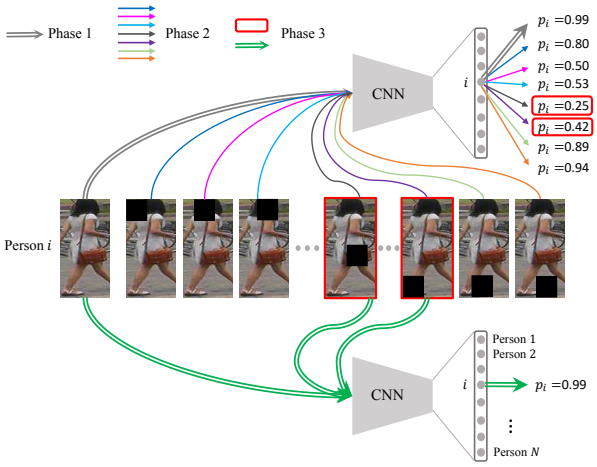

Figure 2: One concrete implementation of our method. It includes three phases. **First**, a classification model for ReID is trained as usual, after which the predicted probabilities for true labels are close to 1. **Second**, we use a sliding window to occlude different regions and let the network predict the class probability. **Finally**, we combine occluded samples with original ones and re-train the network as in the first phase. Those occluded samples with lower predicted probability (harder for the original model) are more likely to be selected in the third phase. The final model is robust to occlusion (predicted probabilities for true labels are close to 1) and also has better test time performance.

sample is denoted by $(I_i, c_i), i \in \{1, 2, ..., N\}$, where c_i is the ground truth class label of image I_i. The network can be viewed as a function that maps an image I_i to a classification score vector $z_i = g(I_i)$, which is further normalized by a softmax function to a probability distribution

$$p(y_{ij}|I_i) = \frac{\exp(z_{ij})}{\sum_{k=1}^{C} \exp(z_{ik})}, \; j = 1, ..., C. \tag{1}$$

The loss of the model on this sample is computed as the Cross Entropy between the true label c_i and the predicted probability for this label y_{ic_i},

$$L_i(\theta) = -\log(p(y_{ic_i}|I_i)), \tag{2}$$

with θ denoting the parameters of the neural network. Thus the loss over the whole training set is as follows

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} L_i(\theta). \tag{3}$$

Minimizing the loss function is equivalent to maximizing the posterior probability of ground truth classes. We use Stochastic Gradient Descent with mini batches to optimize the loss function.

3.2. Search for Critical Regions

After a ReID model is trained, the next step is to find out which regions of training images the model relies on when recognizing persons. Since the deep neural network is trained without any location annotation other than the class labels, we can only figure out what the model depends on through network visualization.

The most straightforward approach is to occlude portions of an input image with a square mask before feeding it to the network, *i.e*. substituting that region with a constant grey intensity, and monitor how the classification probability changes. This method was first introduced by Zeiler and Fergus [48] as one way of measuring the sensitivity of a classification network. They observed that placing the occlusion mask at different places resulted in different degree of decreasing in the classification accuracy, and that when the ground truth object was occluded (or partially occluded), the accuracy dropped significantly, while the vibration caused by occluding other image regions was much smaller.

When applying masks to person images, we find similar phenomenon. As demonstrated in Figure 3, the ReID model degrades dramatically at some occlusion positions, while remaining unaffected elsewhere. We believe that those critical regions indicate what the model depends on when making classification decisions.

3.3. Re-train the Model

In the third phase, we combine original samples and adversarially occluded samples to re-train the model. In this section, we discuss the way of re-training; How occluded samples are selected is introduced in Section 3.4.

The obtained occluded samples are adversarial to the baseline model and intuitively should be used to finetune and improve the model *from that specific training state*. However, we find it tricky to adjust the learning rate for finetuning. On one hand, it requires relatively smaller learning rate compared to training from scratch. On the other, small learning rate may not be sufficient to bring the model out of the local optimum (of overfitting to some body parts). It would be much simpler if we can combine original and occluded samples and train the model just with the same settings as in training the baseline model. The concern that should be addressed is *Would the occluded samples selected based on the baseline model be equally adversarial to this model being trained?* The answer is *Yes*. We discover that models trained in different runs, with the same settings except some randomness, are sensitive to almost the same regions. An example is shown in Figure 3 (a)-(b).

In this way, we can simply follow the common paradigm of data augmentation. Specifically, in each mini batch, each sample is replaced by its occluded counterpart with some probability. The learning rate, iterations and batch size *etc*.

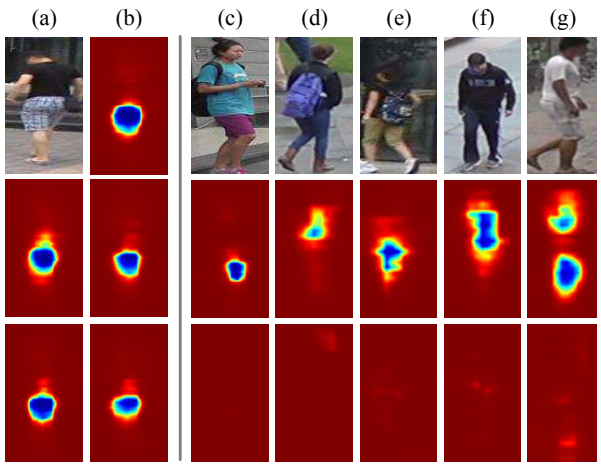

Figure 3: Sensitivity of trained classification model to occlusion, where each heatmap pixel (warmer color with higher value) is the predicted true-class probability when the training image ($h \times w = 256 \times 128$) is occluded by a 64×64 mask centered at that position. Column **(a)-(b)** show sensitivity maps of five baseline models that are trained identically, except for some randomness in training. This shows that independently trained models are equally prone to occlusion at similar positions (blue regions). For column **(c)-(g)**, 1st row is the original training images, while 2nd and 3rd rows are the sensitivity maps of the baseline model and the re-trained model respectively. The original model performs recognition depending on some body parts. After re-trained with the adversarial samples, the model takes more regions into account and grows robust to occlusion.

are the same as in training the baseline model. The superiority over finetuning is that in the early stage of training, the model already has access to those adversarial samples thus able to escape the local optimum with a large learning rate. After dozens of epochs through the dataset, the model is shaped to work well on both original and occluded samples.

3.4. Select Occluded Samples

As introduced in Section 3.2, we search for discriminative regions using a sliding occlusion mask. Thus occluding at each position yields a new sample, which is viewed as a candidate for re-training. Formally, for an image with resolution $H \times W$ and the occlusion mask with size $d \times d$, when we move the mask with horizontal and vertical stride s_w and s_h respectively, the total number of mask positions can be computed as $N_{pos} = \left(\left\lfloor \frac{W-d}{s_w} \right\rfloor + 1\right) \times \left(\left\lfloor \frac{H-d}{s_h} \right\rfloor + 1\right)$. Thus we have a candidate pool of size N_{pos} for each original training image.

During re-training, when a sample in a mini batch is to

be replaced by its occluded version, we select an occluded sample from the candidate pool. One option is to sort the candidates according to how hard they are for the trained ReID model and always select the hardest sample. We call this sampling strategy **Hard-1**.

Beyond this, we notice that 1) placing a mask at neighboring positions has almost equal influence to the original model. 2) When the model is sensitive to a large region, *e.g.* the upper clothes in Figure 3 (f), or more than one region, *e.g.* both the T-shirt and the shorts in Figure 3 (g), multiple mask positions can have comparable effects and should be equally useful in re-training. Therefore, for overcoming the shortcomings of using only the hardest position, we introduce a more flexible strategy. We normalize the N_{pos} influence values for each training image to a distribution and sample from it, so that different occlusion positions can be fairly considered according to the influence they cause. Concretely, for an image, we denote the true class probability predicted by the original model as p. After applying the mask at position $i, i \in \{1, 2, ..., N_{pos}\}$, the probability turns into p_i. Then we compute

$$\hat{p}_i = \begin{cases} p - p_i, & \text{if } p > p_i \\ 0, & \text{otherwise} \end{cases}, \tag{4}$$

which can be normalized to a distribution

$$\bar{p}_i = \frac{\hat{p}_i}{\sum_{j=1}^{N_{pos}} \hat{p}_j}, \ i = 1, 2, ..., N_{pos}. \tag{5}$$

And from this distribution, we sample occluded images. We denote this approach as **Sampling**.

4. Experiments

4.1. Datasets and Evaluation Metrics

We conduct our experiments on three large-scale person ReID datasets, Market1501 [52], CUHK03 [20] and DukeMTMC-reID [29, 56]. Two common evaluation metrics are used, Cumulative Match Characteristic (CMC) [14] for which we report the Rank-1 accuracy, and mean Average Precision (mAP) [52].

Market1501 contains 12,936 training images from 751 persons, 29,171 test images from another 750 persons. To increase the difficulty of retrieval, the gallery set additionally contains 2,798 distractor images with just body parts or background. **CUHK03** contains detected and human cropped images, both with 14,096 images from 1,467 identities. Following [57] for a larger test set, we adopt the new train/test protocol in which 767 identities are used for training and 700 for testing, taking the same evaluation procedure as Market1501. Besides, we experiment on the detected images, which are more close to real world scenario. **DukeMTMC-reID** takes the same format as Market1501, with 16,522 images (702 persons) for training and 19,889 (another 702 persons and 408 distractor persons) for testing.

4.2. Implementation Details

Model. We use ResNet-50 [15] as the model, changing the output size of the classifier to the number of identities in the training set. Euclidean distance is used for similarity metric. These settings are the same as IDE (R) in [57].

Optimization. We initialize the last layer of ResNet-50 with Gaussian weights and zero bias while reserving the ImageNet pre-trained weights in all other layers. We use Stochastic Gradient Descent (SGD) to optimize the model, with a momentum of 0.9 and weight decay of 5e-4. The learning rate of the last layer is set to 0.02 and other layers to 0.01, all of which are multiplied by 0.1 after every 25 epochs. The training is terminated after 60 epochs. During training, we insert a dropout layer before the classifier to regularize the network. The dropout rates for Market1501, CUHK03 and DukeMTMC-reID are set to 0.6, 0.5, 0.5 respectively.

Preprocessing. It has been demonstrated in [45] (arXiv v1) that for person ReID, a $width : height$ ratio of $1 : 2$ is superior to $1 : 1$, which we also verify in the experiment. So we resize all input images to size $w \times h = 128 \times 256$. Input images are randomly mirrored with a probability of 0.5. The batch size is set to 32 and single-GPU training is used.

Occlusion. Occlusion is implemented by substituting a specific region with all zeros after image normalization. Only square masks are used and the stride of sliding window is empirically set to 10. We analyze area ratio of occlusion mask in Section 4.4 and fix it to 0.1 elsewhere. Input images are replaced by their occluded counterparts with a probability of 0.5 during re-training, except in Section 4.5 where different occlusion probabilities are analyzed.

4.3. Visualizing Sensitivity and Activation Maps

The baseline ReID model is prone to occlusion in two aspects. **First**, for the trained classification network, the predicted probability for true class drops dramatically if input training images are replaced by their adversarial counterparts. The 2nd row of column (c)-(g) in Figure 3 demonstrates this, where the blue regions imply vulnerability. **Second**, the ranking accuracy at test time is also affected severely when some certain part of the query image is occluded. Illustration of this can be found in Figure 4. After re-training the model with adversarial samples, both types of vulnerability decrease largely.

We argue that when occluded images are fed to the network at training time, it strives to discover new discriminative clues from other parts of the body in order to recognize them correctly. This is proved by the changes in the activa-

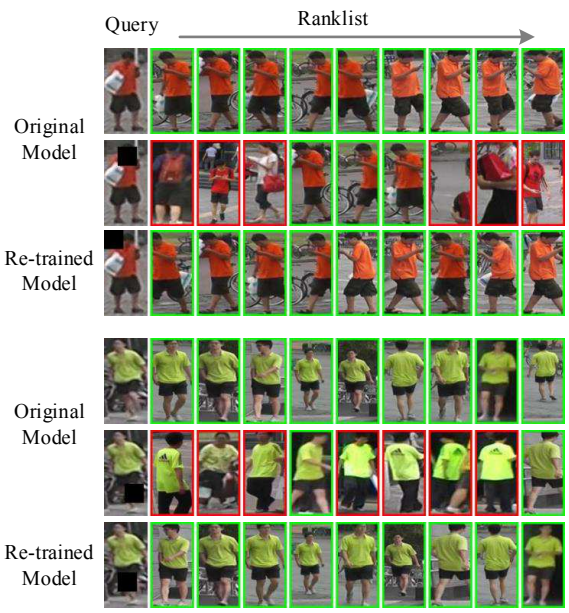

Figure 4: Vulnerability of ReID model to occlusion at test time. The queries are occluded at the position most adversarial to the original or re-trained model, *i.e*. with lowest mAP score. Images with green and red boundary denote true positive and false positive respectively. The original model fails undesirably when occlusion happens at some positions, which is frequent on the test set. After re-training, the model is much more robust to occlusion. *Zoom in to see the black square occlusion masks on four query images.*

tion maps of intermediate layers. In Figure 1, we average the activation maps over the whole test set, and compare the difference before and after re-training. It shows that the original model mainly concentrates on upper part of the body, while the re-trained model also pays attention to other regions like head, shoulder and shoes. Activation maps of the original models imply that the training constraint is fairly easy for the network to satisfy, as a result it just takes the shortest path to goal and overfits to some discriminative parts. The activation maps after re-training indicate that the model has become more cautious when making decisions.

4.4. Influence of Mask Size

Occlusion mask size is the kind of factor that directly determines the degree of adversary (or hardness) for the model. We experiment with Hard-1 under eight mask areas, 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175 and 0.2. For each mask size, we re-run step 2 and 3 in Figure 2. The test scores of re-trained models are plotted in Figure 5. We can see that **1)** Even for small mask size 0.025, *i.e*. 28×28, the improvement over the baselines is obvious. **2)** As mask size increases, the performance goes up at first but afterwards

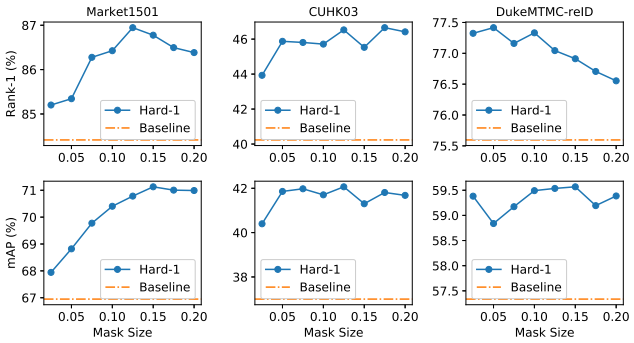

Figure 5: Performance of the ReID models re-trained with method Hard-1, under various mask sizes.

converges or even declines, though different datasets have some difference in the trend. We guess that too large a mask may occlude excessive proportion of the discriminative information and make the images too difficult to recognize. *The subsequent experiments are conducted with mask size 0.1.*

4.5. Influence of Occlusion Probability

During re-training the ReID model, we replace each original sample with its occluded counterpart with some probability. Here we analyze the effect of different occlusion probabilities, for both sampling strategies Hard-1 and Sampling (Section 3.4). The test scores are shown in Figure 6. We observe that **1)** For Hard-1, it performs much better when the occlusion probability is smaller. We speculate that when the proportion of occluded samples is too large, the most discriminative regions are always absent in training. As a result, the data distribution becomes extremely hard for the model and discounts the benefits it brings. **2)** For Sampling, it is much more robust to various occlusion probabilities. **3)** The comparison between Hard-1 and Sampling indicates that under higher occlusion probability, it's better to dynamically choose occlusion regions to achieve both adversary and variation in the training data. **4)** When the probability is merely 0.2, the performance gain of both Hard-1 and Sampling over the baseline is prominent. This demonstrates the effectiveness of our proposed samples. *Occlusion probability is fixed to 0.5 in other experiments.*

4.6. The Necessity of Adversary

The proposed method generates new samples by occluding exactly those discriminative regions of training images. In this section, we aim to verify the necessity of adversary. As introduced in Section 3.4, we can sort the candidate pool for each training image according to recognition probability. To get rid of adversary, we select the occluded image that is easiest for the ReID model. We call this sampling

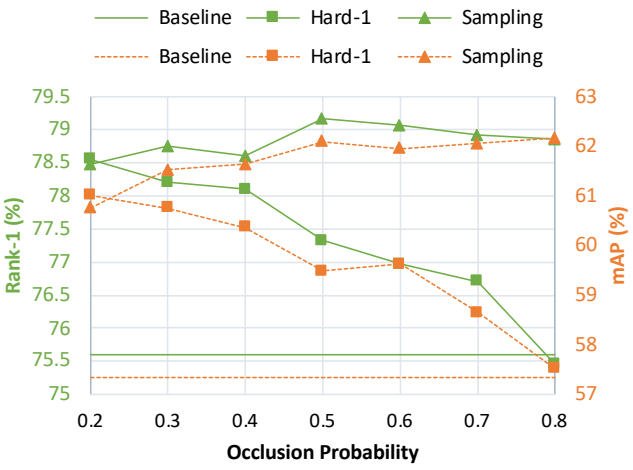

Figure 6: Performance of models re-trained with various occlusion probabilities on DukeMTMC-reID.

Method	Market1501		CUHK03		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Baseline	84.42	66.95	40.23	37.00	75.60	57.34
No Adversary	84.01	66.57	39.41	36.48	76.39	57.40
Random	86.16	69.59	44.56	41.15	78.19	60.94
Hard-1	86.43	70.40	45.71	41.70	77.33	59.49
Sampling	**86.49**	**70.43**	**47.14**	**43.33**	**79.17**	**62.10**

Table 1: Performance of baseline model and models re-trained with different sampling strategies.

strategy **No Adversary**. The comparison of this method with Hard-1 and Sampling is listed in Table 1. We can see that in most cases, the No Adversary method has no improvement over the baseline, even slightly worse. Note that No Adversary introduces no extra hardness to the model, but it introduces variation to the dataset. It can be concluded that increasing variation of training set is not guaranteed to improve the model, if without adversary.

4.7. Different Ways of Selecting Occluded Samples

We compare different sampling strategies in re-training, **Hard-1**, **Sampling**, and **Random**. Here Random occlusion is an easy-to-implement method without the need for a trained model or subsequent model visualization. It just places masks at random positions on images during training a ReID model. We keep its mask size and occlusion probability the same as the other two methods.

The results are listed in Table 1. On both Market1501 and CUHK03, Hard-1 is better than random occlusion. However it's the opposite case on DukeMTMC-reID. Though, this may still be reasonable. **For one thing**, using only the hardest sample not only makes the training exceedingly tough, but also misses other effective variation for the training data. **For the other**, the large number of training

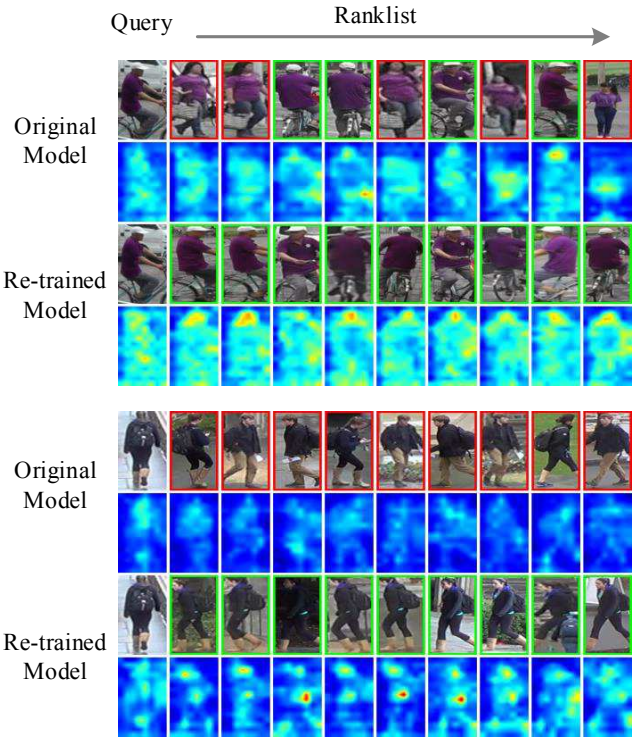

Figure 7: Two example query cases that the re-trained model improves over the original model. Images with green and red boundary denote true positive and false positive respectively. Activation maps for corresponding images are also displayed.

iterations and the large occlusion probability (0.5) give random occlusion the opportunity to generate plentiful effective samples. We conclude that, although less controllable, random occlusion is a promising tradeoff between time efficiency and performance.

The superiority of Sampling proves effective, as demonstrated by the consistent improvement over the Hard-1 and Random approaches. Our method eventually has significant boosting over the original models, increasing the mAP scores by 3.48%, 6.33%, 4.76% and Rank-1 Accuracy by 2.07%, 6.91%, 3.57% on three datasets, respectively.

4.8. Ranklist Analysis

In Figure 7 we compare some query cases that the re-trained model improves over the baseline model. **In the first case**, the original model fails due to the great similarity between the query and distractors, for they have the same clothes colors. Comparing the activation maps, we see that the re-trained model shows advantage in capturing more clues *e.g.* the hat. **In the second case**, the original model mistakenly matches boots with trousers that have the same brown color. The re-trained model successfully cap-

Method	Single Query		Multi Query	
	Rank-1	mAP	Rank-1	mAP
BOW [52]	34.38	14.1	42.64	19.47
WARCA [17]	45.16	-	-	-
PersonNet [42]	37.21	26.35	-	-
DNS [49]	55.43	29.87	71.56	46.03
SCSP [3]	51.9	26.35	-	-
CAN [25]	48.24	24.43	-	-
S-LSTM [40]	-	-	61.6	35.3
Gate-CNN [39]	65.88	39.55	76.04	48.45
CRAFT [7]	71.8	45.5	79.7	54.3
P2S [60]	70.72	44.27	85.78	55.73
CADL [23]	73.84	47.11	80.85	55.58
SpindleNet [50]	76.9	-	-	-
ReRanking (R) [57]	77.11	63.63	-	-
MSCAN [19]	80.31	57.53	86.79	66.7
SSM [1]	82.21	68.8	88.18	76.18
DaF [47]	82.3	72.42	-	-
SVDNet (R) [36]	82.3	62.1	-	-
DeeplyPart [51]	81.00	-	-	-
GAN [56]	83.97	66.07	88.42	76.10
PDM [35]	84.14	63.41	-	-
JLML [21]	85.1	65.5	89.7	74.5
Baseline	84.42	66.95	89.78	75.17
Ours	86.49	70.43	91.32	78.33
Ours + ReRanking	**88.66**	**83.30**	**92.51**	**88.60**

Table 2: Comparison with state-of-the-art methods on Market1501.

tures the features of the hat, boots and some conjunction regions of clothes and avoids the mistake. Besides, in both cases, more regions on the body are activated by the re-trained model than the original one. These changes in the model abound, and more examples can be found in the supplementary material.

4.9. Comparison with State-of-the-art Methods

In this section, we compare our method with state-of-the-art methods. Evaluation results on Market1501, CUHK03 and DukeMTMC-reID are listed in Table 2, 3 and 4 respectively. The baseline we implement (**Baseline**) is competitive to previous works on three datasets. With the proposed re-training method Sampling, our model (**Ours**) achieves state of the art. In addition, we further adopt an effective re-ranking method [57] for post-processing (**Ours + ReRanking**). We eventually achieve impressive performance, with Rank-1 accuracy 88.66%, 54.56% and 84.11% and mAP 83.30%, 56.09%, 78.19% on Market1501, CUHK03 and DukeMTMC-reID respectively.

5. Conclusion

In this paper, we address the problem of generalization in person ReID, for which we propose to generate Adversarially Occluded Samples based on a trained ReID model and use these extra samples to improve the model. We demonstrate that these samples urge the model to discover new clues on the body and improves its capability. We also analyze the hardness and effects of different mask sizes, occlusion probabilities and sampling methods, which gives some insights into harnessing occlusion based adversarial samples. The significant improvement on three large-scale datasets demonstrates the efficacy of our approach. To explore more potential in our strategy, we will consider more efficient way of finding discriminative regions, such as Grad-CAM [32].

Method	Rank-1	mAP
BOW + XQDA [52]	6.36	6.39
PUL [9]	9.1	9.2
LOMO + XQDA [22]	12.8	11.5
IDE + DaF [47]	26.4	30
ReRanking (R) [57]	34.7	37.4
PAN [55]	36.3	34
DPFL [6]	40.7	37
SVDNet (R) [36]	41.5	37.3
Baseline	40.23	37.00
Ours	47.14	43.33
Ours + ReRanking	**54.56**	**56.09**

Table 3: Comparison with state-of-the-art methods on CUHK03 (*detected* subset), under the new evaluation protocol [57].

Method	Rank-1	mAP
BOW + KISSME [52]	25.13	12.17
LOMO + XQDA [22]	30.75	17.04
GAN [56]	67.68	47.13
OIM [44]	68.1	-
Verif. + Identif. [54]	68.9	49.3
APR [24]	70.69	51.88
ACRN [31]	72.58	51.96
PAN [55]	71.59	51.51
SVDNet (R) [36]	76.7	56.8
Baseline	75.60	57.34
Ours	79.17	62.10
Ours + ReRanking	**84.11**	**78.19**

Table 4: Comparison with state-of-the-art methods on DukeMTMC-reID.

6. Acknowledgement

This work is jointly supported by the National Key Research and Development Program of China (2016YFB1001005), the National Natural Science Foundation of China (Grant No. 61473290, Grant No. 61673375), the Projects of Chinese Academy of Science (Grant No. QYZDB-SSW-JSC006, Grant No. 173211KYSB20160008), and Huawei Technologies Co., Ltd (Contract No.:YBN2017030069).

References

[1] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, 2017.

[2] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a "siamese" time delay neural network. In *NIPS*, 1994.

[3] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, 2016.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv*, 2016.

[5] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *CVPR*, 2017.

[6] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *CVPR Workshop*, 2017.

[7] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai. Person re-identification by camera correlation aware feature augmentation. *TPAMI*, 2017.

[8] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *PR*, 2015.

[9] H. Fan, L. Zheng, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. *arXiv*, 2017.

[10] R. Girshick. Fast r-cnn. In *ICCV*, 2015.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv*, 2014.

[14] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS Workshop*, 2007.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[16] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv*, 2014.

[17] C. Jose and F. Fleuret. Scalable metric learning via weighted approximate rank component analysis. In *ECCV*, 2016.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[19] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017.

[20] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[21] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. In *IJCAI*, 2017.

[22] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[23] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou. Consistent-aware deep learning for person re-identification in a camera network. In *CVPR*, 2017.

[24] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving person re-identification by attribute and identity learning. *arXiv*, 2017.

[25] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *TIP*, 2017.

[26] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[27] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

[28] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[29] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016.

[30] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[31] A. Schumann and R. Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *CVPR Workshop*, 2017.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014.

[34] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv*, 2014.

[35] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.

[36] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[38] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv*, 2013.

[39] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016.

[40] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016.

[41] X. Wang, A. Shrivastava, and A. Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *CVPR*, 2017.

[42] L. Wu, C. Shen, and A. v. d. Hengel. Personnet: Person re-identification with deep convolutional neural networks. *arXiv*, 2016.

[43] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016.

[44] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017.

[45] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian. Deep representation learning with part loss for person re-identification. *arXiv*, 2017.

[46] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *ICPR*, 2014.

[47] R. Yu, Z. Zhou, S. Bai, and X. Bai. Divide and fuse: A re-ranking approach for person re-identification. In *BMVC*, 2017.

[48] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

[49] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016.

[50] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017.

[51] L. Zhao, X. Li, J. Wang, and Y. Zhuang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017.

[52] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[53] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv*, 2016.

[54] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person re-identification. *TOMM*, 2017.

[55] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. *arXiv*, 2017.

[56] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[57] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017.

[58] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *arXiv*, 2017.

[59] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

[60] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng. Point to set similarity based deep feature learning for person re-identification. In *CVPR*, 2017.