Counterfactual Attention Learning for Fine-Grained Visual Categorization and Re-identification

Yongming Rao[*], Guangyi Chen[*], Jiwen Lu[†], Jie Zhou
Department of Automation, Tsinghua University, China
State Key Lab of Intelligent Technologies and Systems, China
Beijing National Research Center for Information Science and Technology, China
{raoyongming95,guangyichen1994}@gmail.com; {lujiwen,jzhou}@tsinghua.edu.cn

Abstract

Attention mechanism has demonstrated great potential in fine-grained visual recognition tasks. In this paper, we present a counterfactual attention learning method to learn more effective attention based on causal inference. Unlike most existing methods that learn visual attention based on conventional likelihood, we propose to learn the attention with counterfactual causality, which provides a tool to measure the attention quality and a powerful supervisory signal to guide the learning process. Specifically, we analyze the effect of the learned visual attention on network prediction through counterfactual intervention and maximize the effect to encourage the network to learn more useful attention for fine-grained image recognition. Empirically, we evaluate our method on a wide range of fine-grained recognition tasks where attention plays a crucial role, including fine-grained image categorization, person re-identification, and vehicle re-identification. The consistent improvement on all benchmarks demonstrates the effectiveness of our method. [1]

1. Introduction

Attention is one of the most fundamental mechanism of human visual perception. When facing a complex scene, humans are able to select regions of interest, and employ attention to narrow down the search and speed up recognition. Many efforts [57, 65, 50, 18, 45, 14, 1, 6, 36] have been made to model the mechanism of human attention in computer vision systems, which aim to facilitate high-performance recognition by discovering discriminative regions and mitigating the negative effects brought by diverse visual appearance, cluttered backgrounds, occlusions, pose variations, *etc*. Since subtle differences are key to distinguish subordinate visual categories, visual attention mechanism has proven to be especially effective in fine-grained visual recognition tasks and become a core component in many state-of-the-art methods [47, 31, 7, 65, 50, 19, 28, 25, 62, 44].

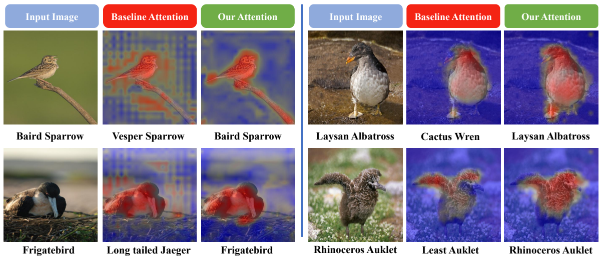


Figure 1: Attention visualization on CUB. We respectively show the original images, baseline attention maps, and attention maps with counterfactual learning. In the left part, we observe that our attention maps can better focus on the object. While comparisons in the right part show that our models prefer to look at the whole object instead of some parts. Best in color.

Despite the widespread use, the problem of how to learn effective attention is still barely studied. Most existing methods learn the visual attention in a weakly-supervised manner, *i.e.*, the attention modules are simply supervised by the final loss function, without a powerful supervisory signal to guide the training process. This likelihood based approach only explicitly supervises the final prediction (*e.g.*, class probabilities for classification task) but ignores the causality between the prediction and attention. Previous methods also did not teach the machine how to distinguish between the main clues and biased clues. For example, if most training samples of one specific class appear with sky as background, then the attention model may be very likely to treat the sky as a discriminative region. Although these biased clues may also be beneficial to the classification on the current datasets, the attention model should only focus on the discriminative patterns, *i.e.* the main clues. Besides, directly learning from data may encourage the model to only focus on some cer-

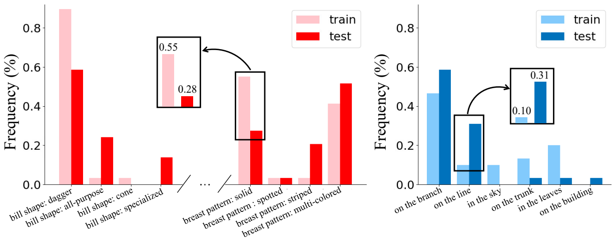
[*] Equal contribution. [†] Corresponding author.
[1] Code is available at https://github.com/raoyongming/CAL

tain attributes of the objects instead of all attributes, which may limit the generalization ability on test set. Therefore, we argue that this attention learning scheme is sub-optimal, where the effectiveness of the learned attentions is not always guaranteed, and the attention may lack discriminative power, clear meaning and robustness. As shown in Figure 1, misleading and scattered attentions can still be observed from a well-trained attention model and potentially lead to the wrong predictions. To better understand this phenomenon, we analyze the statistics of both intrinsic attributes and external environments on the CUB dataset (see Figure 2), where we use the attributes provided by the dataset and manually collect the environment statistics. We see there are biases for both attributes and environment, which indicates either background and single part are not reliable clues for classification. Therefore, it is desired to design new attention learning method beyond conventional likelihood maximization to mitigate the effects of data biases.

Because of the lack of effective tool to evaluate the quality of attentions quantitatively, correcting misleading attentions is a very challenging task. One straightforward solution is to use extra annotations like bounding boxes or segmentation masks to obtain the regions of interest explicitly such as [1]. However, this kind of method requires considerable cost of human labor and is hard to scale up. Considering the critical role that attention plays in fine-grained visual recognition tasks, it is necessary to design a method to measure the quality of attentions without additional human supervision and further optimize the learned visual attentions.

In this paper, we present a *counterfactual attention learning* (CAL) method to enhance attention learning based on causal inference. Specifically, we design a tool to analyze the effects of learned visual attention with counterfactual causality. The basic idea is to quantitate the quality of attentions by comparing the effects of facts (*i.e.*, the learned attentions) and the counterfactuals (*i.e.*, uncorrected attentions) on the final prediction (*i.e.*, the classification score). Then, we propose to maximize the difference (*i.e.*, *effect* in causal inference literature [41, 54]) to encourage the network to learn more effective visual attentions and reduce the effects of biased training set.

The proposed method is model-agonistic and thus can serve as a plug-and-play module to improve a wide range of visual attention models. Our method is also computational efficient, which only introduces a little extra computation cost during training and brings no computation during inference while can significantly improve attention models. We evaluate our method on three fine-grained visual recognition tasks including fine-grained image categorization (CUB200-2011 [55], Stanford Cars [23] and FGVC Aircraft [38]), person re-identification (Market1501 [66], DukeMTMC-ReID [46] and MSMT17 [60]) and vehicle re-identification (Veri-776 [30] and VehicleID [29]). By applying our method to a multi-head attention baseline model, we demonstrate our method significantly improves the baseline and achieve state-of-the-art results on all benchmarks.



(a) The attributes

(b) The environments

Figure 2: The biases of both intrinsic attributes and external environments on CUB. We demonstrate the biases on the training and testing sets by the statistics the frequencies in different attributes and environments, taking the *Ringed Kingfisher* as an example.

2. Related Work

Fine-Grained Visual Recognition. Attention mechanism plays an irreplaceable role in fine-grained visual recognition tasks. For example, in fine-grained image categorization task, Sermanet *et al.* [47] pioneer adopting attention mechanism in fine-grained recognition problem and propose a RNN model to learn visual attention. Liu *et al.* [31] extend the idea and employ a reinforcement learning scheme to obtain visual attentions. The subsequent studies such as MA-CNN [65], MAMC [50] and WS-DAN [19] further improve this line of methods and design attention models in a bottom-up manner, which achieve very promising results on fine-grained recognition benchmarks. Attention models have also proven to be effective in person/vehicle re-identification problem to handle the image matching misalignment challenge and improve the discriminative power of CNN features. For instance, Liu *et al.* [28] and Lan *et al.* [25] employ the attention models to locate the discriminative salient regions in images to improve person re-identification. Xu *et al.* [62] and Zhao *et al.* [64] design a body part detector to employ the structure of the human body structure in the attention model. Another group of methods [32, 48, 26, 8] adopts attention mechanism on video-based person re-identification task to discover key parts in videos. Khorramshahi *et al.* [21] propose an adaptive attention model and significantly improve the state-of-the-art of vehicle re-identification task.

Causal Reasoning in Vision. The interest in combining the idea of deep learning and causal reasoning is growing rapidly in recent years. The tool of causality analysis has been successfully used in several areas, including explainable machine learning [40], fairness [24], natural language processing [61], reinforcement learning [20] and adversarial learning [22]. Some efforts also used causality as an effective tool to alleviate the effects of dataset bias in vision tasks, including image classification [33], scene graph gen-

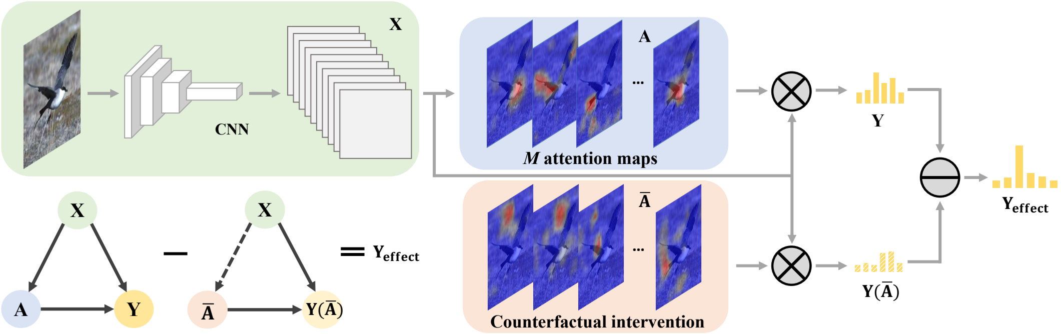


Figure 3: The overall framework of our CAL method. We first apply the counterfactual intervention for original attention by replacing with random attentions. Then, we subtract the counterfactual classification results from original classification to analyze the effects of learned visual attention and maximize them in the training process.

eration [53] and visual commonsense reasoning [59]. In this work, we study causality in the context of visual attention models, which is a new direction that has not been visited.

3. Approach

3.1. Attention Models for Fine-Grained Recognition

We begin by reviewing the attention models for fine-grained visual recognition, on which our method is built. Given an image I and the corresponding CNN feature maps $\mathbf{X} = f(I)$ of size $H \times W \times C$, visual spatial attention model $\mathcal{M}$ aims to discover the discriminative regions of the image and improve CNN feature maps $\mathbf{X}$ by explicitly incorporating structural knowledge of objects. Note that although some of previous methods like [57] propose to equip the backbone network with spatial attention modules, here we follow the mainstreams [47, 31, 65, 50, 19, 28, 28, 25, 62, 64] that learn basic feature maps and attentions separately. Previous studies have demonstrated that this design is more flexible and generic thanks to its model-agnostic nature.

There have been quite a few variants of $\mathcal{M}$, and we can roughly categorize them into two groups. The first type aims to learn "hard" attention maps, where each attention can be represented as a bounding box or segmentation mask that covers a certain region of interest. This group of methods is usually closely related to object detection and semantic segmentation methods. Examples include recurrent visual attention model [39] and fully convolutional attention network [31]. Different from hard-attention models, a wider range of attention models are based on learning "soft" attention maps, which are more easy to optimize. In this paper, we focus on studying this group of methods. Specifically, our baseline model adopts the multi-head attention module used in [65, 50, 19]. The attention model is designed to learn the spatial distributions of object's parts, which can be represented as attention maps $\mathbf{A} \in \mathbb{R}_{+}{}^{H\times W\times M}$, where M is the number of attentions. Using the attention model $\mathcal{M}$, attention maps can be computed by:

$$\mathbf{A}{=}\{\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_M\}{=}\mathcal{M}(\mathbf{X}), \tag{1}$$

where $\mathbf{A}_i \in \mathbb{R}_{+}{}^{H\times W}$ is the attention map covering a certain part, such as the wing of a bird or the cloth of a person. The attention model $\mathcal{M}$ is implemented using a 2D convolutional layer followed by ReLU activation. The attention maps then are used to softly weight the feature maps and aggregate by global average pooling operation φ:

$$\mathbf{h}_i{=}\varphi(\mathbf{X} * \mathbf{A}_i){=}\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{X}^{h,w}\mathbf{A}_i^{h,w}, \tag{2}$$

where $*$ denotes element-wise multiplication for two tensors. Following the practice in [19], we summarize the representation of different parts to form the global representation $\mathbf{h}$:

$$\mathbf{h}{=}\texttt{normalize}([\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_M]), \tag{3}$$

where we concatenate these representations and normalize the summarized representation to control its scale. The final representation $\mathbf{h}$ can be fed into a classifier (*e.g.*, fully connected layer) for image classification task or a distance metric (*e.g.*, Euclidean distance) for image retrieval task. The overall framework of our baseline attention model is illustrated in Figure 3.

3.2. Attention Models in Causal Graph

Before we show our counterfactual method, we first introduce how to reformulate the above model in the language of causal graph. Causal graph is also known as structural causal model, which is a directed acyclic graph $\mathcal{G}{=}\{\mathcal{N}, \mathcal{E}\}$. Each variable in the model has a corresponding node in $\mathcal{N}$ while the causal links $\mathcal{E}$ describe how these variable interact with each other. As presented in Figure 3, we can use nodes in the causal graph to represent variables in the attention model, including the CNN feature maps (or the input image)

X, the learned attention maps A and the final prediction Y. The link $X \rightarrow A$ represents that the attention model takes as input the CNN feature maps and produces corresponding attention maps. $(X, A) \rightarrow Y$ indicates the feature maps and attention maps jointly determine the final prediction. Causal relations between nodes are encoded in the links $\mathcal{E}$, where we call node X is the causal parent of A, and Y is the causal child of X and A. Note that since we do not impose any constraints on the network architecture of backbone models and the implementation details of the attention model, the causal graph can also represent many other attention models. Therefore, our method is model-agonistic and thus can also be extended to a wider range of attention learning problems.

3.3. Counterfactual Attention Learning

Conventional likelihood methods optimize the attention by only supervising the final prediction Y and regard the model as a black box, which ignores how the learned attention maps affect the prediction. On the contrary, causal inference [42] provides a tool to help us think out of the black box by analyzing the causalities between variables. Therefore, we propose to employ the causalities to measure the quality of the learned attention and then improve the model by encourage the network to produce more influential attention maps.

By introducing the causal graph, we can analyze causalities by directly manipulate the values of several variable and see the effect. Formally, the operation is termed *intervention* in causal inference literature, which can be denoted as $do(\cdot)$. When we want to investigate the effect of a variable, the intervention operation is performed by wiping out all the in-coming links of the variable and assigning a certain value to the variable. For example, $do(A{=}\bar{\mathbf{A}})$ in our causal graph means we demand the variable A to take the value $\bar{\mathbf{A}}$ and cut-off the link $X \rightarrow A$ to force the variable to no longer be caused by its causal parent X.

Inspired by causal inference methods [41, 54], we propose to adopt *counterfactual intervention* to investigate the effects of the learned visual attention. The counterfactual intervention is achieved by an imaginary intervention altering the state of the variables assumed to be different [54, 15]. In our case, we conduct counterfactual intervention $do(A{=}\bar{\mathbf{A}})$ by imagining non-existent attention maps $\bar{\mathbf{A}}$ to replace the learned attention maps and keeping the feature maps X unchanged. We can obtain the final prediction Y after the intervention $A{=}\bar{\mathbf{A}}$ according to (2) and (3):

$$Y(do(A{=}\bar{\mathbf{A}}), X{=}\mathbf{X}){=}\mathcal{C}([\varphi(\mathbf{X}{*}\bar{\mathbf{A}}_1), ..., \varphi(\mathbf{X}{*}\bar{\mathbf{A}}_M)]), \tag{4}$$

where $\mathcal{C}$ is the classifier. In practice, we can use random attention, uniform attention or reversed attention as the counterfactuals. Evaluation on these options can be found in Section 4.4.

Following [41, 54, 53], the actual effect of the learned attention on the prediction can be represented by the difference between the observed prediction $Y(A = \mathbf{A}, X = \mathbf{X})$ and its counterfactual alternative $Y(do(A = \bar{\mathbf{A}}), X = \mathbf{X})$:

$$Y_{\text{effect}}{=}\mathbb{E}_{\bar{\mathbf{A}}\sim\gamma}[Y(A{=}\mathbf{A}, X{=}\mathbf{X}){-}Y(do(A{=}\bar{\mathbf{A}}), X{=}\mathbf{X})], \tag{5}$$

where we denote the effect on the prediction as Y_{effect} and γ is the distribution of counterfactual attentions. Intuitively, the effectiveness of an attention can be interpreted as how the attention improves the final prediction compared to wrong attentions. Therefore, we can use Y_{effect} to measure the quality of a learned attention.

Furthermore, we can use the metric of attention quality as a supervision signal to explicitly guide the attention learning process. The new objective can be formulated as:

$$\mathcal{L}{=}\mathcal{L}_{ce}(Y_{\text{effect}}, y){+}\mathcal{L}_{\text{others}}, \tag{6}$$

where y is the classification label, $\mathcal{L}_{ce}$ is the cross-entropy loss, and $\mathcal{L}_{\text{others}}$ represents the original objective such as standard classification loss. By optimizing the new objective, what we expect to achieve is two-fold: 1) the attention model should improve the prediction based on wrong attentions as much as possible, which encourages the attention to discover the most discriminative regions and avoid sub-optimal results; 2) we penalize the prediction based on wrong attentions, which forces the classifier to make decision based more on the main clues instead of the biased clues and reduces the influence of biased training set.

Note that in practice, it is not necessary to compute the expectation in Equation (5) and we only sample a counterfactual attention for each observed attention during training, which is also consistent with the idea of stochastic gradient descent. Therefore, the extra computational cost introduced by our method is an additional forward of the attention model and the classifier, which is very lightweight compared with the CNN backbone. Besides, our method introduces no additional computation during inference.

4. Experiments

We assess the effectiveness of our proposed counterfactual attention learning method on several fine-grained visual recognition tasks including fine-grained image categorization, person re-identification and vehicle re-identification. We take the conventional spatial attention as the baseline and compare our counterfactual attention learning method with the baseline method and other state-of-the-art methods. The experimental settings, implementation details and results for different tasks are described below.

4.1. Fine-grained Image Categorization

Fine-grained visual categorization focuses on classifying the subordinate-level classes under a fixed basic-level

Table 1: Comparisons of the top-1 classification accuracy (%) with the SOTA fine-grained image categorization methods on CUB200-2011, Stanford Cars and FGVC Aircraft.

Method	**CUB**	**Cars**	**Aircraft**
RA-CNN [12]	85.3	92.5	-
MA-CNN [65]	86.5	92.8	89.9
MAMC [50]	86.5	93.0	-
NTS-Net [63]	87.5	93.9	91.4
WS-DAN [19]	89.4	94.5	93.0
DCL [9]	87.8	94.5	93.0
Stacked LSTM [13]	90.4	-	-
API-Net [70]	90.0	95.3	93.9
Baseline	89.3	94.0	93.6
Baseline + CAL	**90.6**	**95.5**	**94.2**

category, such as species of bird, types of car and types of aircraft. The objects under the same basic-level category are always high structured and with low inter-class variances. Thus, attention is effective to look for the key difference in detail and discover the discriminative regions.

Datasets and Experimental Settings. We conducted experiments on widely used CUB200-2011 [55], Stanford Cars [23] and Aircraft [38] datasets for fine-grained bird, car and aircraft classification. CUB200-2011 is composed of 5,994 training images and 5,794 testing images from 200 species of birds. Stanford Cars contains 16,185 images of cars from 196 different types and among all collected, 8,144 images are used for training and 8,041 images for testing. FGVC Aircraft consists of 10,000 images of 100 fine-grained aircraft types. Following previous methods, we use 2/3 images for training and 1/3 images for evaluation.

Implementation Details. We adopted the standard ResNet-101 [16] as the backbone network. For attention model, we set the number of attentions to 32 and use the weakly supervised data augmentation method as suggested by [19]. During inference, we use multiple crops and horizontal flipping to boost performance. All experiments are conducted with the same hyper-parameters, including 16 batch size, 448×448 image size, and 1e-5 weight decay. We use 1e-3 initial learning rate and reduce the learning rate by 0.9 times in every 2 epochs.

Results. We compared our method with the baseline attention model and the state-of-the-art methods in Table 1. The proposed counterfactual attention method can improve the strong baseline by 1.3%, 1.5% and 0.6% on CUB200-2011, Stanford Cars and Aircraft, respectively. Our method also outperformed previous state-of-the-art methods. Notably, although a stronger backbone (DenseNet-161) is used in recent API-Net [70] method, our method can still achieve better performance on all three benchmarks. These results clearly demonstrates the effectiveness of our method.

4.2. Person Re-identification

Person re-identification (ReID) is a task to match the query individual from multiple gallery candidates across the non-overlapping camera views. It is a challenging problems because of the intra-class variances due to illumination changes, pose variations, occlusions, and cluttered backgrounds. Attention model has gained great success for person ReID by handling the matching misalignment challenge and enhancing the feature representation [27, 5, 28].

Datasets and Experimental Settings. We conducted the experiments on three public person re-identification datasets including Market1501 [66], DukeMTMC-reID [46] and MSMT17 [60]. Market1501 consists of 32,668 images of 1,501 identities detected by 6 cameras. The whole dataset is divided into a training set with 12,968 images of 751 identities and a test set containing 3,368 query images and 19,732 gallery images of 750 identities. DukeMTMC-reID dataset is composed of 1,404 persons captured by 8 cameras. Its training set includes 16,522 images of 702 persons, while the test set contains the remaining 702 persons with 2,228 query images and 17,661 gallery images. MSMT17 is one of the largest ReID datasets which contains 4,101 identities and 126,411 images. The training set is composed of 30,248 images of 1,041 identities, while the remaining 3,060 identities are used for testing.

We followed the settings of [27] and [5] for Market1501 and DukeMTMC-reID datasets, and chose the single-query manner to validate our method. For MSMT17, we followed the settings in the test settings in [60] and [67]. We employed the cumulative matching characteristic (CMC) curve and mean average precision (mAP) as the evaluation metrics.

Implementation Details. We adopted two basic network structures including a standard ResNet50 backbone and the backbone network in SCAL [5] which adds 4 channel attention blocks. We applied the modification backbone due to its competitive performance and relatively concise structure, since recent state-of-the-art methods usually use stronger or more sophisticated backbone such as part model [52, 58] to improve performance. We used † to indicate the models using the modified backbone. The baseline attention block is same as the one in fine-grained categorization task and we set M to 8 for re-identification models. All experiments are conducted with the same hyper-parameters including 80 batch size, 384 × 192 image size, and 2e-4 learning rate. The data augmentation methods includes random cropping, erasing and horizontal flipping. We trained the network for 160 epochs with triplet loss and softmax loss with learning rate reducing by 10 times in every 40 epochs.

Results. As shown in Table 2, we observed that our CAL methods achieve consistent improvement for different baselines on all benchmarks. Specifically, compared with the

Table 2: Comparisons with the state-of-the-art person ReID methods on the Market1501, DukeMTMC-ReID and MSMT17.

Method	Market1501			DukeMTMC-ReID			MSMT17		
	R1	R5	mAP	R1	R5	mAP	R1	R5	mAP
HA-CNN [27]	91.2	-	75.7	80.5	-	63.8	-	-	-
Part-aligned [49]	91.7	96.9	79.6	84.4	92.2	69.3	-	-	-
Mancs [56]	93.1	-	82.3	84.9	-	71.8	-	-	-
PCB+RPP [52]	93.8	97.5	81.6	83.3	-	69.2	68.2	-	40.4
IANet [17]	94.4	-	83.1	87.1	-	73.4	75.5	85.5	46.8
JDGL [67]	94.8	-	86.0	86.6	-	74.8	77.2	-	52.3
SCAL [5]	95.8	**98.7**	89.3	88.9	95.2	79.1	-	-	-
MHN [4]	95.1	98.1	85.0	89.1	94.6	77.2	-	-	-
SFT [37]	93.4	-	82.7	86.9	-	73.2	73.6	-	47.6
OSNet [68]	94.8	-	84.9	88.6	-	73.5	78.7	-	52.9
BAT-Net [11]	95.1	98.2	87.4	87.7	94.7	77.3	79.5	89.1	56.8
Auto-ReID [43]	94.5	-	85.1	-	-	-	78.2	88.2	52.5
MGN+circleloss [51]	**96.1**	-	87.4	-	-	-	76.9	-	52.1
Baseline	94.0	97.7	85.9	85.7	93.6	74.0	75.3	86.4	50.5
Baseline + CAL	94.5	97.9	87.0	87.2	94.1	76.4	79.5	89.0	56.2
Baseline†	94.9	98.3	89.0	88.7	94.7	78.2	81.4	90.3	59.3
Baseline† + CAL	95.5	98.5	**89.5**	**90.0**	**96.1**	**80.5**	**84.2**	**92.0**	**64.0**

strong baseline we obtained 0.6%/0.5% Rank-1/mAP improvement on the Market1501 dataset, 1.6%/2.3% on the DukeMTMC-ReID dataset, and 2.8%/4.7% on the MSMT17 dataset. The improvement on the MSMT17 dataset is larger than other two datasets, since images in MSMT17 have larger intra-class variances. Besides, with our strong attention model, we can achieve the SOTA performance on DukeMTMC-ReID and MSMT17 datasets.

4.3. Vehicle Re-identification

Vehicle Re-Identification (ReID) aims to retrieve all images of a given query vehicle from a large image database, without the license plate clues. The vehicles with different identities can be of the same make, model and color, while the vehicle appearances of the same identity always vary significantly across different viewpoints. Attention model can be applied for matching the key similarity of vehicle images across different viewpoints.

Datasets and Experimental Settings. We conducted the experiments on two widely used vehicle datasets including Veri-776 [30] and VehicleID [29]. Veri-776 dataset contains over 50,000 images from 776 vehicle IDs, where 37,778 images from 576 IDs are split for training and the rest 200 IDs are used for testing. VehicleID is composed of 110,178 images of 13,134 vehicles for training and 111,585 images of 13,133 IDs for testing. Following the experimental settings in [30] and [35], we report the testing results for three subsets including small size subset with 800 vehicles, medium size subset with 1,600 vehicles and large subset with 2,400 vehicles. We employed the CMC curve and mAP as the evaluation metrics for vehicle ReID task.

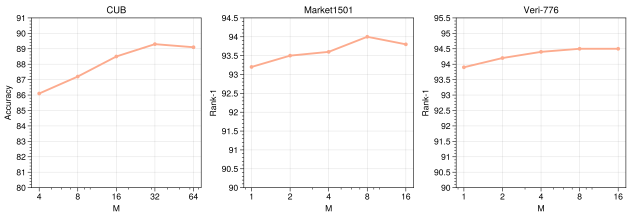


Figure 4: Effects of the number of attentions. We investigate the effects of the numbers of attention on CUB, Market1501 and Veri-776 for fine-grained image categorization, person re-identification and vehicle re-identification respectively and directly use the best hyper-parameters on other datasets.

Implementation Details. We applied the ResNet50 backbone and the same attention block (M=8) as the baseline. The hyper-parameters are also fixed for the baseline and our method. The loss functions and data augmentation methods are same with person ReID task. We selected 256 samples in a batch with 256 × 256 image size. The initial learning rate is 2e-4 and reducing 10 times in 8000th and 18000th iterations. We trained the network for total 28000 iterations.

Results. We compared the performance of CAL with the baseline attention learning method and other SOTA methods. As shown in Table 3, we obtained 0.9%/2.3% Rank-1/mAP improvement on the Veri-776 dataset and 5.8%/3.3%/4.1% Rank-1 improvement on small/medium/large test settings of the VehicleID dataset. Note that we did not use any extra labels in the training precess, yet achieved the comparable performance with VAML [10] which manually annotates the viewpoints of images to train the view-predictor.

Table 3: Comparisons with the state-of-the-art vehicle ReID methods on the VeRi-776 and VehicleID datasets.

Method	Veri-776			VehicleID								
	Test 11587			Test 800			Test 1600			Test 2400		
	R1	R5	mAP	R1	R5	mAP	R1	R5	mAP	R1	R5	mAP
GSTE [3] -	-	59.4	87.1	-	-	82.1	-	-	79.8	-	-	
AAMI [69]	85.9	91.8	61.3	63.1	83.3	-	52.9	75.1	-	47.3	70.3	-
FDA-NeT [34]	84.3	92.4	55.5	-	-	-	59.8	77.1	65.3	55.5	74.7	61.8
VAML* [10]	89.8	96.0	66.3	88.1	97.3	-	83.2	95.1	-	80.4	93.0	-
AAVER [21]	88.7	94.1	58.5	72.5	93.2	-	66.9	89.4	-	60.2	84.9	-
EALN [35]	84.4	94.1	57.4	75.1	88.1	77.5	71.8	83.9	74.2	69.3	81.4	71.0
DFLNet [2]	93.2	97.6	73.3	78.8	**95.1**	82.8	-	-	-	69.8	**90.6**	75.4
ResNet50	94.5	97.2	72.0	76.7	93.5	84.1	74.9	89.5	81.4	71.0	84.9	78.0
ResNet50 + CAL	**95.4**	**97.9**	**74.3**	**82.5**	94.7	**87.8**	**78.2**	**91.0**	**83.8**	**75.1**	88.5	**80.9**

Table 4: Analysis of different counterfactual attentions. We implement different strategies to generate counterfactual attentions including random attention, uniform attention, reversed attention and shuffle attention. We report the top-1 classification accuracy of classification task.

	CUB	Cars	Aircraft
Baseline	89.3	94.0	93.6
Random Attention	**90.6**	**95.5**	94.2
Uniform Attention	90.2	95.3	94.2
Reversed Attention	89.2	94.1	93.2
Shuffle Attention	90.4	94.3	**94.5**

Table 5: Quantitative analysis of attention. We compare the classification accuracy (%) our method with other three kinds of attention regularization strategies including attention drop, entropy regularization and attention normalization, and evaluate the quality of the learned attention maps using mIoU (%) with ground-truth bounding boxes on CUB.

	Accuracy	mIoU
Baseline	89.3	54.2
+ Attention Dropout	88.9	50.3
+ Entropy Regularization	88.7	51.1
+ Attention Normalization	85.8	46.2
+ CAL	**90.6**	**67.4**

Table 6: Results of single-head attention models. We report the top-1 classification accuracy (%) on three fine-grained categorization datasets.

	CUB	Cars	Aircraft
Baseline (M=1)	85.9	92.1	91.5
+ CAL	**88.2**	**94.2**	**92.9**

4.4. Analysis

We analyzed the influences and sensitivity of some major parameters. We conducted the parameters analysis experiments on three fine-grained visual recognition tasks.

Effects of the type of counterfactual attention. We investigated three different strategies to generate the counterfactual attention maps, namely random attention, uniform attention, reversed attention and shuffle attention (see Supplementary Material for details). The results are presented in Table 4. We see random attention, uniform attention and shuffle attention achieve similar performance while reversed attention fails to improve the baseline on CUB. We think it is because learning attention that is better than reversed attention is relatively easy and cannot provide an effective signal to supervise the attention.

Effects of the number of attentions. The number of heads in attention model is an important hyper-parameter in our baseline model. Therefore, we search the best numbers of attention on CUB [55], Market1501 [66] and Veri-776 [30] for fine-grained image categorization, person re-identification and vehicle re-identification tasks respectively and directly use the searched hyper-parameters on other datasets. For a fair comparison, we use the same hyper-parameters in our models and the baseline models, and did not search the best hyper-parameters for our models separately. The results are presented in Figure 4. Based on these results, we set M to 32 and 8 for fine-grained categorization and re-identification tasks, respectively.

Quantitative analysis of attention. To better verify the effectiveness of CAL, we compare our method with other three kinds of attention regularization strategies including attention drop, entropy regularization and attention normalization (see Supplementary Material for details) and evaluated the quality of the learned attention maps by computing the mean IoU between the rectangular region that covers the high score attentions and the ground-truth object bounding boxes on CUB. The results can be found in Table 5. We see only CAL is effective to simultaneously improve classification accuracy and attention quantitative. Both attention Dropout and Entropy regularization will slightly degrade the final performance under the both metrics. Attention normalization will significantly hurt the performance.

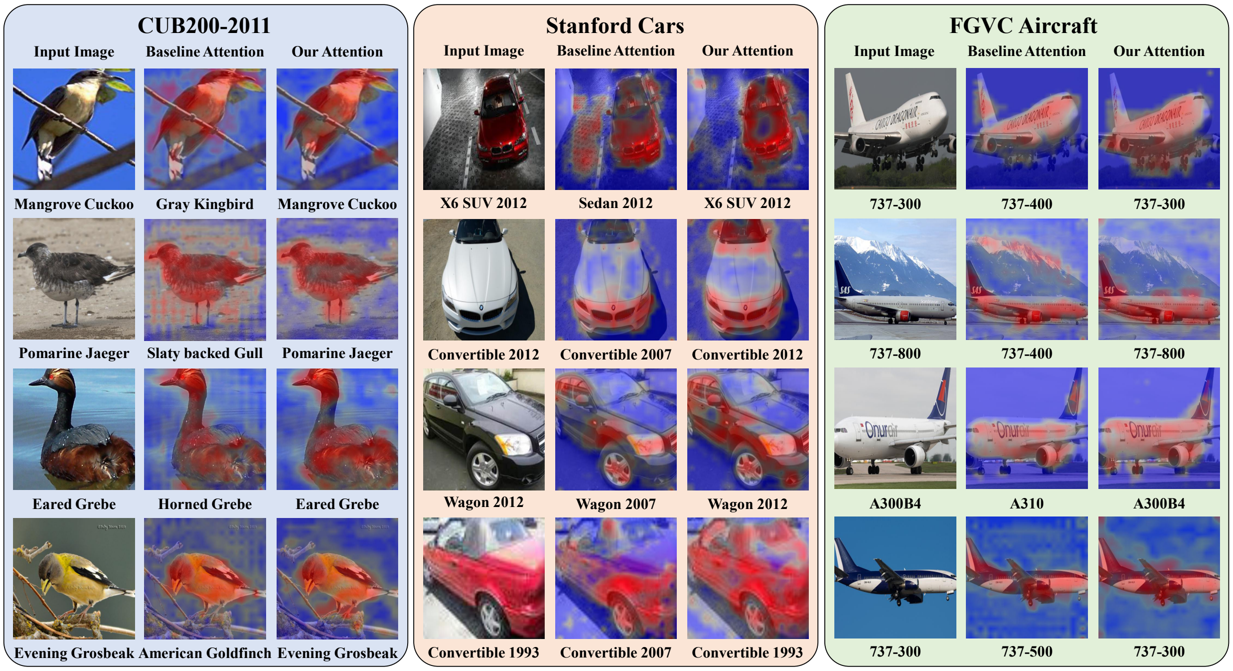


Figure 5: Visualization of the attention maps of our models and the baseline models. We see our method helps the attention models make correct predictions by 1) reducing the misleading and scatter attentions and 2) encouraging the model to focus on the main clues for classification and explore more discriminative regions. Best in color.

Results of single-head attention models. To show the generality for different attention models, we also test CAL on single-head attention models (*i.e.*, baseline models with $M = 1$). The results are presented in Table 6. We see our method can consistently and more significantly improve the relative weak baseline models, which clearly shows our method is suitable for various attention models.

4.5. Visualization

To have an intuitive understanding of our counterfactual attention learning method, we compare the attention maps of our models and the baselines models on CUB200-2011 [55], Stanford Cars [23] and FGVC-Aircraft datasets [38]. The visual results are displayed in Figure 5. We see our method helps the attention models make correct predictions by reducing the misleading and scatter attentions. For example, in the first example of the Stanford Cars dataset, the attention with our CAL method avoids the reflection on the ground. Besides, CAL encourages the model to focus on the main clues for classification and explore more discriminative regions. Taking the Eared Grebe in the CUB200-2011 dataset as example, our attention focuses on the discriminative buttocks region to recognize it. While for the second example of cars and the first one of aircrafts, our attention models tend to explore more discriminative regions such as the rearview mirror and wheel respectively.

5. Conclusion

In this paper, we have presented a counterfactual attention learning method to learn more effective attention based on causal inference. We designed a framework to quantitate the quality of attentions by comparing the effects of facts and the counterfactuals on the final prediction. We also proposed to maximize the difference to encourage the network to learn more effective visual attentions. Our method only brings negligible extra cost during training and introduce no cost during inference. CAL is a model-agnostic framework to enhance attention learning and mitigate the effects of dataset bias, which can be applied to various fine-grained visual recognition tasks. We conducted extensive experiments on three fine-grained visual recognition tasks and demonstrated state-of-the-art performance on all benchmarks.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102803, in part by the National Natural Science Foundation of China under Grant 61822603, Grant U1813218, and Grant U1713214, in part by a grant from the Beijing Academy of Artificial Intelligence (BAAI), and in part by a grant from the Institute for Guo Qiang, Tsinghua University.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.

[2] Yan Bai, Yihang Lou, Yongxing Dai, Jun Liu, Ziqian Chen, and Ling-Yu Duan. Disentangled feature learning network for vehicle re-identification. In *IJCAI*, 2020.

[3] Yan Bai, Yihang Lou, Feng Gao, Shiqi Wang, Yuwei Wu, and Ling-Yu Duan. Group-sensitive triplet embedding for vehicle reidentification. *TMM*, 20(9):2385–2399, 2018.

[4] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *ICCV*, 2019.

[5] Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. Self-critical attention learning for person re-identification. In *ICCV*, pages 9637–9646, 2019.

[6] Guangyi Chen, Jiwen Lu, Ming Yang, and Jie Zhou. Spatial-temporal attention-aware learning for video-based person re-identification. *TIP*, 28(9):4192–4205, 2019.

[7] Guangyi Chen, Jiwen Lu, Ming Yang, and Jie Zhou. Learning recurrent 3d attention for video-based person re-identification. *TIP*, 29:6963–6976, 2020.

[8] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In *ECCV*, pages 660–676, 2020.

[9] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *CVPR*, pages 5157–5166, 2019.

[10] Ruihang Chu, Yifan Sun, Yadong Li, Zheng Liu, Chi Zhang, and Yichen Wei. Vehicle re-identification with viewpoint-aware metric learning. In *ICCV*, 2019.

[11] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Bilinear attention networks for person retrieval. In *ICCV*, 2019.

[12] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, 2017.

[13] Weifeng Ge, Xiangru Lin, and Yizhou Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *CVPR*, 2019.

[14] Jiayuan Gu, Han Hu, Liwei Wang, Yichen Wei, and Jifeng Dai. Learning region features for object detection. In *ECCV*, pages 381–395, 2018.

[15] York Hagmayer, Steven A Sloman, David A Lagnado, and Michael R Waldmann. Causal reasoning through intervention. *Causal learning: Psychology, philosophy, and computation*, pages 86–100, 2007.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[17] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *CVPR*, 2019.

[18] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, 2018.

[19] Tao Hu, Honggang Qi, Qingming Huang, and Yan Lu. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *arXiv preprint arXiv:1901.09891*, 2019.

[20] Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *NeurIPS*, pages 9269–9279, 2018.

[21] Pirazh Khorramshahi, Amit Kumar, Neehar Peri, Sai Saketh Rambhatla, Jun-Cheng Chen, and Rama Chellappa. A dual-path model with adaptive attention for vehicle re-identification. In *ICCV*, 2019.

[22] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017.

[23] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, pages 554–561, 2013.

[24] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NeurIPS*, 2017.

[25] Xu Lan, Hanxiao Wang, Shaogang Gong, and Xiatian Zhu. Deep reinforcement learning attention selection for person re-identification. *BMVC*, pages 4–7, 2017.

[26] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, pages 369–378, 2018.

[27] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.

[28] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *TIP*, 2017.

[29] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *CVPR*, pages 2167–2175, 2016.

[30] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *ECCV*, pages 869–884. Springer, 2016.

[31] Xiao Liu, Tian Xia, Jiang Wang, Yi Yang, Feng Zhou, and Yuanqing Lin. Fully convolutional attention networks for fine-grained recognition. *arXiv preprint arXiv:1603.06765*, 2016.

[32] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017.

[33] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *CVPR*, pages 6979–6987, 2017.

[34] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *CVPR*, 2019.

[35] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Ling-Yu Duan. Embedding adversarial learning for vehicle re-identification. *TIP*, 28(8):3794–3807, 2019.

[36] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, pages 289–297, 2016.

[37] Chuanchen Luo, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Spectral feature transformation for person reidentification. In *ICCV*, 2019.

[38] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[39] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *NeurIPS*, 2014.

[40] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[41] Judea Pearl. Direct and indirect effects. *arXiv preprint arXiv:1301.2300*, 2013.

[42] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[43] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. In *ICCV*, 2019.

[44] Yongming Rao, Jiwen Lu, and Jie Zhou. Attention-aware deep reinforcement learning for video face recognition. In *ICCV*, 2017.

[45] Yongming Rao, Jiwen Lu, and Jie Zhou. Learning discriminative aggregation network for video-based face recognition and person re-identification. *IJCV*, 127(6):701–718, 2019.

[46] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016.

[47] Pierre Sermanet, Andrea Frome, and Esteban Real. Attention for fine-grained categorization. *arXiv preprint arXiv:1412.7054*, 2014.

[48] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, pages 5363–5372, 2018.

[49] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person reidentification. In *ECCV*, 2018.

[50] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multiattention multi-class constraint for fine-grained image recognition. In *ECCV*, pages 805–821, 2018.

[51] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, 2020.

[52] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. In *ECCV*, pages 480–496, 2018.

[53] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. *CVPR*, 2020.

[54] Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

[55] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[56] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV*, 2018.

[57] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, pages 3156–3164, 2017.

[58] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, pages 274–282, 2018.

[59] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. *CVPR*, 2020.

[60] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018.

[61] Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. Challenges of using text classifiers for causal inference. In *EMNLP*, volume 2018, page 4586. NIH Public Access, 2018.

[62] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, pages 2119–2128, 2018.

[63] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *ECCV*, pages 420–435, 2018.

[64] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person reidentification. In *ICCV*, pages 3219–3228, 2017.

[65] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for finegrained image recognition. In *ICCV*, pages 5209–5217, 2017.

[66] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[67] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019.

[68] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, 2019.

[69] Yi Zhou and Ling Shao. Aware attentive multi-view inference for vehicle re-identification. In *CVPR*, 2018.

[70] Peiqin Zhuang, Yali Wang, and Yu Qiao. Learning attentive pairwise interaction for fine-grained classification. *arXiv preprint arXiv:2002.10191*, 2020.