Learning Feature Embedding with Strong Neural Activations for Fine-Grained Retrieval

Chen Shen *
Zhejiang University, Alibaba Group
zjushenchen@gmail.com

Chang Zhou
Alibaba Group
zhouchang.zc@alibaba-inc.com

Zhongming Jin
Alibaba Group
zhongming.jinzm@alibaba-inc.com

Wenqing Chu *
Zhejiang University, Alibaba Group
wqchu16@gmail.com

Rongxin Jiang †
the State Key Laboratory of Industrial Control Technology, Zhejiang University

Yaowu Chen
Zhejiang Provincial Key Laboratory for Network Multimedia Technologies, Zhejiang University

Xian-Sheng Hua †
Alibaba Group
huaxiansheng@gmail.com

ABSTRACT

Fine-grained object retrieval, which aims at finding objects belonging to the same sub-category as the probe object from a large database, is becoming increasingly popular because of its research and application significance. Recently, convolutional neural network (CNN) based deep learning models have achieved promising retrieval performance, as they can learn both feature representations and discriminative distance metrics jointly. Specifically, a generic method is to extract activations of the fully-connected layer as feature descriptors and simultaneously optimize classification constraints (*e.g.*, softmax loss) and similarity constraints (*e.g.*, triplet loss) to improve the representative capability of the features. However, the typical fully-connected layer activations are more focused on representing global attributes of the corresponding image, thus relatively less sensitive to specific local characteristics. Therefore, the features learned through these approaches in general are not sufficiently capable for retrieving fine-grained objects. To attack this issue, we propose an effective feature embedding by simultaneously encoding original global features and discriminative local features, in which the local features are extracted by exploiting strong neural activations on the last convolutional layer. We present that the novel feature embedding can dramatically enlarge the gap between inter-class variance and intra-class variance, which is the key factor to improve retrieval precision. In addition, we show our architecture can also be applied in person re-identification. Experimental results on multiple challenging benchmarks demonstrate that our method outperforms the current state-of-the-art approaches by large margins.

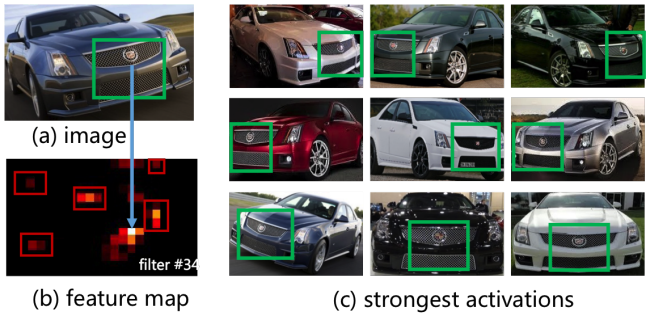

Figure 1: Visualization of the feature map. (a) An image from Stanford cars-196 dataset. (b) The feature map of a random $conv_5$ filter (the basic network architecture is VGGNet [27]). The arrow indicates the strongest response and its corresponding position in the raw image. The red rectangles mark the weaker activation areas of the feature map. (c) Other images that have the strongest responses of the same filter. The cars in these images are all belonging to the same fine-grained category – Cadillac CTS-V 2012. The green rectangles mark the receptive fields of the strongest responses.

KEYWORDS

Fine-Grained Object Retrieval, Strong Neural Activations, Deep Convolutional Neural Networks, Person Re-identification

1 INTRODUCTION

Fine-grained object retrieval, which aims at searching objects belonging to the same subordinate class as the query object from a large database, is becoming more and more important and meaningful in computer vision community. For example, fine-grained car retrieval can be used for numerous purposes in intelligent transportation, surveillance and public security areas, such as tracking a suspicious car over multiple surveillance cameras when the license

ThematicWorkshops'17, October 23–27, 2017, Mountain View, CA, USA.

DOI: https://doi.org/10.1145/3126686.3126708

*This work was done when the authors were visiting Alibaba as research interns.
†Corresponding authors.

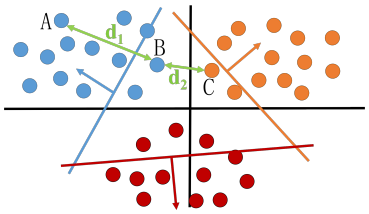

Figure 2: A schematic diagram of three fine-grained category feature distributions. Each color represents a specific category and each point represents the global fully-connected layer features of an image from the corresponding category.

plate is not reliable. Recently, deep convolutional neural networks (CNN) [16, 27] have made great progress in fine-grained categorization on many fine-grained object datasets such as birds [30], dogs [13], cars [15], *etc.* And CNN features extracted from the fully-connected layer can be viewed as a universal discriminative representation to retrieve objects from the same sub-category, as classification task can take full advantages of category annotations and learn an outstanding feature representation to differentiate subordinate classes. In the meanwhile, distance metric learning (*e.g.*, contrastive information [6] or triplets [5]) has been exploited to alleviate the issue of large intra-class variance and small inter-class variance, which is known as the main challenge of fine-grained image understanding. Considering the above two aspects, a generic fine-grained object retrieval method [39] is to extract activations of the CNN fully-connected layer as global feature representations and simultaneously optimize softmax loss and triplet loss to improve the representative capability of the features. That means, the problem is formulated as constructing a multi-task network, which effectively learns fine-grained features by jointly optimizing both classification and similarity constraints.

Although the global fully-connected layer features achieve competitive results on fine-grained classification (*e.g.*, 89% accuracy on Stanford cars-196 dataset [15], which is very close to the state-of-art result of 92.8% [14]), the retrieval precision is far from satisfactory (*e.g.*, only 68% on top-40 retrievals [39]). There are two observations that motivate our work. Firstly, as we can see from Figure 1(b), the strong activations of conv_5 (*i.e.*, the last convolutional layer of VGGNet [27]) filters probably represent some discriminative local semantic regions of the images, *e.g.*, the filter #34 is most activated by a unique grille shape of a specific car model, which can be verified in Figure 1(c). However, for a specific conv_5 filter, some weaker activations (red rectangles on Figure 1(b)), which may just indicate some redundant content, are also encoded in the global fully-connected layer features, *i.e.*, the fully-connected layer activations encode global attributes of the raw images, thus relatively less sensitive to specific local characteristics. Secondly, Figure 1(c) shows that for different sample images, the positions of discriminative local regions are unfixed due to differences in viewpoints or poses. Thus, the corresponding strong activations on the feature maps of the same conv_5 filter appear at different positions.

These two observations show that the activations of the conv_5 layer from the same fine-grained category images have various distributions. Consequently, the global fully-connected layer feature vectors suffer relatively large intra-class variance because the global fully-connected layer activations before the "ReLU" layer are just some linear combinations of the convolutional layer activations. Figure 2 illustrates that we can easily classify different fine-grained categories, but the retrieval precision will decline significantly under the circumstances of large intra-class variance (*e.g.*, the Euclidean distance $d_1 > d_2$, so C will be ranked higher than A when a query point is close to B).

In this paper, we propose an effective feature embedding for fine-grained object retrieval. Inspired by [7, 9], we select and pool the strong neural activation areas on each feature map of the last (*i.e.*, highest-level) convolutional layer, then form a feature representation of specific local characteristics through fully-connected procedure. The new features, which can be called as local highlighted fully-connected features in contrast to global fully-connected features, are focused on describing discriminative local semantic regions of raw images. After that, we concatenate the discriminative local features with the global features and project to a low-dimensional feature space through fully-connected procedure to form the final feature embedding. The final feature representation integrates both advantages, preserving global information and being more sensitive to specific local location-independent characteristics. Besides, the representative capability of local semantic features can also be reinforced through the back propagation pattern [17] during training. We demonstrate that, under some reasonable assumptions, the novel feature embedding can dramatically enlarge the gap between inter-class variance and intra-class variance, which is the key factor to improve retrieval precision.

In addition to the above said, our proposed method has two additional benefits. Firstly, our discriminative local features are extracted in a simple and unsupervised manner (*i.e.*, part annotations are not required). Notice that some previous models of fine-grained object recognition yield excellent performance by directly localizing and describing the critical parts [10, 14, 26, 38]. However, these approaches either require more or less part level annotations, complicated artificial participations, or suffer complex optimization, making them not as simple and applicable as ours. Secondly, our novel feature embedding can be seamlessly incorporated into the existing fine-grained retrieval frameworks. For instance, in this work we follow the learning strategy of [39], which is simultaneously optimizing both softmax loss and triplet loss as aforementioned, while replace the original global fully-connected layer features used by [39] with our novel feature embedding. We conduct extensive experiments on two commonly used fine-grained object datasets: Stanford cars-196 [15] and CUB-200-2011 [30]. Experimental results demonstrate that our introduced feature embedding gains a much smaller intra-class variance and attains a consistent and significant performance gain compared with [39], which is one of the current state-of-the-art fine-grained retrieval methods. Furthermore, because person re-identification (re-ID) can be considered as some kind of fine-grained retrieval problems if we view each person identity as a sub-category, we also evaluate our method on the classic person re-ID dataset: CUHK03 [18]. Similarly, we outperform the current state-of-the-art person re-ID approaches by notable margins.

2 RELATED WORK

In this section, we briefly review the recent works on fine-grained categorization, triplet distance metric learning, image retrieval and person re-ID. We only emphasize on the methods that are most relevant to our approaches.

Fine-Grained Object Categorization. Most of recent methods focus on improving the categorization accuracy by capturing the subtle appearance differences in specific object parts. A common pipeline to address this problem consists of two main steps: (1) learning a set of part detectors; (2) encoding part features into the final representation. Some previous techniques rely heavily upon the use of keypoint or part annotations [4, 33, 37]. Lately, Krause *et al.* [14] propose a method based on generating parts using co-segmentation and alignment, using no part annotations. Further on, Zhang *et al.* [38] explore a unified framework based on two steps of deep filter response picking which is free of any object/part annotation at both training and testing stages. However, these approaches need complicated artificial participations. In this paper, we introduce a simple but effective method to extract one discriminative local semantic feature representation for each high-level convolutional filter. The method in [38] is relevant to ours, but the way to use filter responses is significantly different: Zhang *et al.* [38] pick deep filters which respond to specific patterns to learn part detectors, while we directly leverage strong neural filter activations to represent discriminative local semantic features.

Triplet Distance Metric Learning. Triplet loss deep convolutional network aims at learning a feature representation in Euclidean space such that data points from the same class are clustered together, while those from different classes are pushed apart from each other. It is of vital important for fine-grained object verification and retrieval, as it can effectively measure both intra-class and inter-class similarity. Wang *et al.* [31] propose a deep ranking model to directly learn the similarity metric by sampling triplets from images. Recently, Zhang *et al.* [39] jointly optimize triplet information and traditional classification objective simultaneously and introduce label structures embedded into the framework by generalizing the triplet loss. Liu *et al.* [20] employ a coupled cluster triplet strategy to pick more stable anchor candidates. However, these methods neglect how to select negative triplet candidates. In this paper, we introduce an online training strategy to pick more discriminative negative samples.

Image Retrieval. Some recent approaches mainly focus on instance-level image retrieval such as highly variable scenes. MOP-CNN [8] presents a scheme of extracting CNN activations for local patches at multiple scale levels, performing VLAD pooling [11] of the activations at each level separately and concatenating the results. Following this idea, Ng *et al.* [36] introduce an approach for extracting convolutional features from different layers, Babenko *et al.* [2] propose a local features aggregation method based on sum pooling, and Paulin *et al.* [25] illuminate an unsupervised framework to learn patch-level descriptors. These approaches above follow the pipeline of generating abundant local patch descriptors and aggregating them to provide a new global representation. Inspired by their methods, we also explore the local features, however, there are two differences compared with theirs: (1) We extract the local features from feature maps of the last convolutional layer, instead of raw images; (2) For each filter of the last convolutional layer, we only utilize the most activated area to represent a local semantic region under the assumption that fine-grained object's discriminative features can be reflected by strong neural activations, while these approaches above generate a mass of patch windows of raw images towards the goal of representing highly variable scenes, not fine-grained objects.

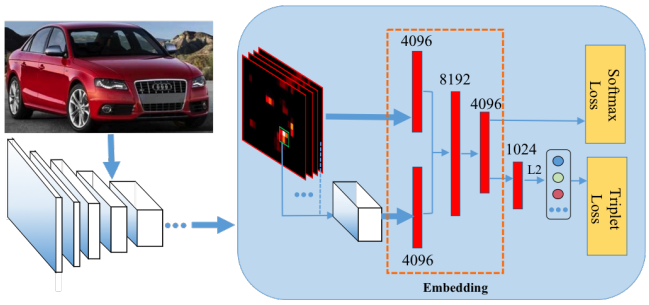

Figure 3: Illustration of feature embedding and multi-task framework.

Person Re-identification. CNN-based deep learning methods can incorporate feature extraction and metric learning into a unified neural network and jointly optimize these two components. Thus, they make remarkable progress to significantly boost the re-ID performance [3, 18, 21, 34, 35]. Generally speaking, two types of CNN models are commonly employed [43]: classification model and Siamese model. For large-scale datasets, such as CUHK03 [18] and Market-1501 [42], the classification model achieves excellent performance without careful training sample selection [32, 41, 43]. We mainly focus on the classification model because they are much similar to our method. However, we replace the global fully-connected layer features with our novel feature embedding. Experimental results show that our method yields state-of-the-art re-ID accuracy on CUHK03 [18] compared with the recent approaches.

3 PROPOSED METHOD

In this section, we firstly present our motivation of exploiting strong neural activations on feature maps of high-level convolutional filters and give the proof that under some reasonable assumptions, the final feature embedding can enlarge the gap between inter-class variance and intra-class variance. Then, we detail the carefully designed local semantic feature extracting methods. Finally, we illustrate the architecture and the loss function of our multi-task end-to-end network, as shown in Figure 3.

3.1 Motivation and Formulation

As mentioned in Section 1, we make a further observation on feature maps of $conv_5$ filters (see Figure 1). For a certain input image, a portion of $conv_5$ filters are activated and the responses on the feature maps of these filters mostly have the following distribution: the strongest activation may occur at any position, surrounded by some relatively strong activations, and other weaker activations are randomly distributed in different places. In the meanwhile, strong activation areas of these filters probably represent some discriminative

local semantic regions of the raw images, which are shared among the same sub-category (*e.g.*, particular shapes of headlight, taillight, fog light, grille or even typical car body type for car images), while other weaker activation areas may just indicate some redundant content. The phenomena above inspire us to directly exploit strong neural activations towards the goal of approximately representing local semantic features, which are helpful to distinguish the correct sub-category images from other subordinate classes. Hence, we propose to compactly select one strong neural activation area for each filter of the last convolutional layer and pool location-independent feature vectors to represent the discriminative local features. We choose the last convolutional layer because deeper convolutional layer generally has more powerful semantic representation capability. Notice that for the corresponding input image, there are also part of conv_5 filters being less activated (*i.e.*, even the strongest response on these feature maps has a extremely small numerical value) and the response distributions on the feature maps of these filters are either messy or sparse. Extracting local features from these filters seems uninterpretable and meaningless, but due to their tiny numerical activation values, the influence of these filters can be almost negligible compared with those activated ones.

Then, we propose an effective feature embedding by concatenating the above discriminative local features with original global features, and projecting to a low-dimensional feature space through fully-connected procedure. The aforementioned motivation is under the assumption that discriminative local features of a specific fine-grained category can be reflected by some typical strong neural activations, *i.e.*, for images from a certain sub-category, their local features are more similar to each other than the images from another fine-grained class. Going a step further, we can naturally assume that, if we utilize the Euclidean distance of the above local features as distance metric, then the distance between an anchor image and a positive image (*i.e.*, from the same sub-category) is smaller than that between the anchor image and a negative image (*i.e.*, from a different sub-category). Under such assumption, we prove that, compared with the original global fully-connected layer features, our proposed feature embedding can enlarge the gap between inter-class variance and intra-class variance, which is the key factor to improve retrieval precision. Technically speaking, we give the theorem and proof as follows.

Theorem 3.1. *For images of a typical sub-category, assuming their feature representations $\mathbf{x} \in \mathbb{R}^n$, we define the intra-class variance*

$$V_{intra} = \mathbb{E}\|\mathbf{x} - \mu_{\mathbf{x}}^{+}\|_2^2,$$

and the inter-class variance

$$V_{inter} = \mathbb{E}\|\mathbf{x} - \mu_{\mathbf{x}}^{-}\|_2^2,$$

where $\mu_{\mathbf{x}}^{+}, \mu_{\mathbf{x}}^{-}$ is the expectation of all the feature vectors among the same sub-category and among the other different categories. Suppose $\mathbf{x_g} \in \mathbb{R}^n$ is a traditional global feature, $\mathbf{x_s} \in \mathbb{R}^n$ is a local feature learned from strong activation prior, and $\mathbf{x_c} = [\mathbf{x_g}; \mathbf{x_s}]$ is the concatenating feature. $\mathbf{x_s}^{+}$ has the same class label as $\mathbf{x_s}$, while $\mathbf{x_s}^{-}$ has the different class label as $\mathbf{x_s}$. If we have

$$\mathbb{E}(\|\mathbf{x_s} - \mathbf{x_s}^{-}\|_2^2 - \|\mathbf{x_s} - \mathbf{x_s}^{+}\|_2^2) > 0, \tag{1}$$

then, we can prove that

$$\mathbb{E}\|\mathbf{x_c} - \mu_{\mathbf{x_c}}^{-}\|_2^2 - \mathbb{E}\|\mathbf{x_c} - \mu_{\mathbf{x_c}}^{+}\|_2^2 > \mathbb{E}\|\mathbf{x_g} - \mu_{\mathbf{x_g}}^{-}\|_2^2 - \mathbb{E}\|\mathbf{x_g} - \mu_{\mathbf{x_g}}^{+}\|_2^2. \tag{2}$$

Proof. Because

$$\mathbb{E}(\|\mathbf{x_s} - \mathbf{x_s}^{-}\|_2^2 - \|\mathbf{x_s} - \mathbf{x_s}^{+}\|_2^2) > 0,$$

we can obtain

$$\mathbb{E}(\|\mathbf{x_c} - \mathbf{x_c}^{-}\|_2^2 - \|\mathbf{x_c} - \mathbf{x_c}^{+}\|_2^2) > \mathbb{E}(\|\mathbf{x_g} - \mathbf{x_g}^{-}\|_2^2 - \|\mathbf{x_g} - \mathbf{x_g}^{+}\|_2^2).$$

Without loss of generality, we normalize the feature as $\frac{\mathbf{x}}{\|\mathbf{x}\|_2}$, then we have $\|\mathbf{x}\|_2 = 1$. Thus,

$$\mathbb{E}(\|\mathbf{x} - \mathbf{x}^{-}\|_2^2 - \|\mathbf{x} - \mathbf{x}^{+}\|_2^2)$$

is equivalent to

$$\mathbb{E}(\mathbf{x}^T\mathbf{x}^{+} - \mathbf{x}^T\mathbf{x}^{-}) + const_a,$$

and

$$\mathbb{E}\|\mathbf{x} - \mu_{\mathbf{x}}^{-}\|_2^2 - \mathbb{E}\|\mathbf{x} - \mu_{\mathbf{x}}^{+}\|_2^2$$

can be written as

$$\|\mu_{\mathbf{x}}^{+}\|_2^2 - \|\mu_{\mathbf{x}}^{-}\|_2^2 + const_b.$$

Because

$$\sum_{i=1}^{m}\sum_{j=1}^{m} x_i^{(d)} x_j^{(d)} = \Big(\sum_{i=1}^{m} x_i^{(d)}\Big)^2,$$

where m is the number of data points, and $x_i^{(d)}$ represents the d-th dimension of the i-th point, then we can see that

$$\mathbb{E}(\mathbf{x}^T\mathbf{x}^{+} - \mathbf{x}^T\mathbf{x}^{-})$$

is equivalent to

$$\|\mu_{\mathbf{x}}^{+}\|_2^2 - \|\mu_{\mathbf{x}}^{-}\|_2^2.$$

It means

$$\mathbb{E}(\|\mathbf{x} - \mathbf{x}^{-}\|_2^2 - \|\mathbf{x} - \mathbf{x}^{+}\|_2^2)$$

is equivalent to

$$\mathbb{E}\|\mathbf{x} - \mu_{\mathbf{x}}^{-}\|_2^2 - \mathbb{E}\|\mathbf{x} - \mu_{\mathbf{x}}^{+}\|_2^2 + const_c.$$

$const_a, const_b, const_c$ represents some constants. In this way, we can prove that

$$\mathbb{E}\|\mathbf{x_c} - \mu_{\mathbf{x_c}}^{-}\|_2^2 - \mathbb{E}\|\mathbf{x_c} - \mu_{\mathbf{x_c}}^{+}\|_2^2 > \mathbb{E}\|\mathbf{x_g} - \mu_{\mathbf{x_g}}^{-}\|_2^2 - \mathbb{E}\|\mathbf{x_g} - \mu_{\mathbf{x_g}}^{+}\|_2^2.$$

□

Actually, there are also some approximations in implementation, such as $\mathbf{x_c}$ is followed by fully-connected procedure to reduce dimension and form the final feature embedding. However, as can be seen in Section 5, the experimental performances show the efficiency of our simplified proposal. By utilizing our feature embedding, the intra-class variance is significantly reduced and the inter-class variance has a slight raise, making the gap between inter-class variance and intra-class variance substantially increasing.

3.2 Local Semantic Feature Extraction

How to select strong neural activation areas and pool location-independent feature vectors, is at the core of our approach, thus we present two techniques to solve this issue.

Fixed-Size Window Pooling. In this method, we utilize the classic sliding window fashion through the whole feature map of each filter and select the rectangle window in which has the strongest average response. For each filter k of a specific high-level convolutional layer where $k = 1, \ldots, K$ and K is the total number of filters (*e.g.*, 512), let w, h be the fixed width and height of the sliding window, $P^k_{(x,y)}$ the activation value of coordinates(x, y) on the corresponding feature map of filter k and X^k the set of all sliding windows. The window $i^k \in X^k$ is chosen by:

$$\arg\max_{i^k} \sum_{x=x_{i^k}}^{x_{i^k}+w} \sum_{y=y_{i^k}}^{y_{i^k}+h} P^k_{(x,y)}, \tag{3}$$

in which (x_{i^k}, y_{i^k}) represent the left-top coordinates of window i^k. Particularly, we add the constraint that the selected window must contain the pixel corresponding to the strongest activation value. Then, we view each selected window as region of interest (RoI) and apply RoI pooling proposed in [7] to pool fixed size $(w \times h)$ convolutional features for each window. After that, we get a $K \times w \times h$ dimension vector and extract it to be a D-dimension (*e.g.*, $D = 4096$) vector through fully-connected procedure. This local fully-connected vector is proposed to describe local highlighted features of the raw image.

It is noteworthy that when w, h are both equal to 1, the method degenerates to Global Max Pooling [29], which means we just seek a local feature representation by pooling the strongest neural activation.

Variable-Size Window Pooling. This method is almost the same as the above one instead of utilizing an arbitrary size window for each neuron. Let $P^k_{(x,y)}$ be the strongest activation of filter k and (x, y) be the corresponding coordinates. We select an rectangle window around (x, y), in which the activation of the pixel (x_j, y_j) on the boundary satisfies: $P^k_{(x_j,y_j)} >= \eta P^k_{(x,y)}$, where η is a threshold (*e.g.*, $\eta = 0.5$). In other words, we expand pixels towards four orthogonal directions simultaneously until the activation of the specific pixel is less than some threshold.

3.3 Multi-task Framework

Feature Embedding and Network Architecture. As aforementioned, we firstly extract local semantic features through window selecting and RoI pooling procedure. Then we form the novel feature embedding by concatenating the local highlighted features with the global fully-connected layer features and projecting to a low-dimensional feature space, in which the final feature representation preserves global information and is more sensitive to specific local location-independent characteristics. On one hand the final features are directly used for the task of fine-grained categorization. On the other hand, we further project the final feature embedding to a compact Euclidean space and go through L_2 normalization pattern, generating a lower-dimensional vector to accomplish the task of fine-grained retrieval. Learning strategies, which are optimizing softmax loss and triplet loss, are simultaneously employed to learn the feature embedding. Figure 3 illustrates the whole framework intuitively. The multi-task learning architecture is designed to impose knowledge sharing between multiple correlated tasks, boosting the performance of a part or even all of the tasks [40].

Loss Function. Fine-grained object datasets usually have multiple labels, and some labels even have hierarchical structures. For instance, Stanford cars-196 dataset [15] annotates each car image with a ground-truth car type label (such as SUV, Sedan) and a car category label, in the meanwhile, a unique hierarchy is presented for the car category label, which is three levels from top to bottom: make, model, and released year. This structure indicates a direction to make multi-label attributes classification, *i.e.*, simultaneously output the prediction of car fine-grained category, car type and car make, *etc.* Formally, let's assume a training dataset of images x_i, each associated with multi labels y^k_i, where $i = 1, \ldots, N$, $k = 1, \ldots, K$. That is, we have N samples, each with a kind of K labels. The multi-label softmax loss is formulated as:

$$L_{cls} = \sum_k \lambda_k \sum_i -\log\left(\frac{e^{f_{y_i^k}}}{\sum_j e^{f_j^k}}\right), \tag{4}$$

where hyperparameters λ_k $(k = 1, \ldots, K)$ are used to control the balance between the multi-label prediction losses and we use the notation f^k_j to mean the j-th element of the vector of k-th class scores f^k. This enrichment of multi labels forms a mini-multi-task within classification constraints learning, boosting the performance of each classification task. Notice that for the datasets, which just have single category labels such as CUB-200-2011 [30] and CUHK03 [18], we just leverage the conventional softmax loss, which is formulated as:

$$L_{cls} = \sum_i -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right), \tag{5}$$

where the definition of parameters is similar to Eq. 4.

Next, for a standard triplet loss network, the inputs are a batch of triplet units (a_i, p_i, n_i), where a_i is the anchor image from a specific category, p_i an image from the same category, and n_i an image from a different category. As shown in Figure 3, our network projects the final full-connected feature vector of a raw image r_i to a lower-dimensional vector $f(r_i) \in \mathbb{R}^D$ in a compact Euclidean space, after L_2 normalization. The classic triplet loss can be denoted as:

$$L_{sim} = \sum^N \max(\Phi(a_i, p_i) - \Phi(a_i, n_i) + m, 0), \tag{6}$$

where $\Phi(r_i, r_j)$ is the squared Euclidean distance between two vectors $(f(r_i), f(r_j))$ defined above and m is a certain margin satisfied $m > 0$. Our novel triplet units selecting strategy is expounded in Section 4.3.

In summary, the loss function of the whole multi-task framework is defined as:

$$L_{multi-task} = L_{cls} + \mu L_{sim}, \tag{7}$$

where the hyperparameter μ is used to control the balance between the two task losses.

Testing. When a query happens, images in the gallery are ranked according to their similarities with the probe image, in which the

similarity between image features is computed by Euclidean distance. Following the same criterion as training, we project the final feature embedding to a lower-dimensional Euclidean space and go through L_2 normalization pattern. Then, we leverage such features for computing distance.

4 EXPERIMENTS

4.1 Datasets and Protocols

Fine-Grained Object Retrieval. The empirical evaluation is performed on two fine-grained benchmarks: Stanford cars-196 (Cars) [15] and CUB-200-2011 (Birds) [30]. Stanford cars-196 contains 16, 185 images of 196 car types, while CUB-200-2011 has 11, 788 images of 200 species of birds. Both datasets have a single bounding box annotation in each image. We use the default training / testing split, which gives us around 40 training examples per class for Cars and around 30 training examples per class for Birds. In our experiments, we employ the widely used protocol in object retrieval: mean average precision (mAP), to evaluate all the methods. Particularly for Stanford cars-196, we also report the precision on top-40 retrievals (following the standard of comparison mentioned in [39], using top-40 since each fine-category contains around 40 images).

Person Re-identification. We conduct experiments on the large dataset CUHK03 [18], which consists of more than 14, 000 images of 1, 467 pedestrians. Li *et al.* [18] provide two types of bounding boxes: labeled (human annotated) and detected (automatically generated). In this work, we report results on the labeled set. We follow the protocol used by [32] and randomly pick a set of 100 identities for testing. We draw roughly 20% of all the images in training set for validating the classification task. In our experiments, we employ the commonly used Cumulative Matching Characteristics (CMC) [22] top-1 accuracy to evaluate all the methods and the evaluation results are reported under single-query setting.

4.2 Implement Details

Our algorithm is implemented based on the deep learning framework Caffe [12] and runs on a workstation configured with a NVIDIA M40 GPU card. The basic convolutional network architecture is VGGNet [27]. For all experiments, we extract local features from the conv_5 layer (the last convolutional layer of VGGNet) because deeper convolutional layer has more powerful semantic representation capability. As a large region on the feature map of conv_5 layer corresponds to a huge receptive field of the input image, a relatively small siding window size is chosen for Fixed-Size Window Pooling and we report the experimental results based on 1×1 and 3×3 size (Notice that 1×1 size equally means Global Max Pooling). For Variable-Size Window Pooling, we select $\eta = 0.7$. The local semantic features and the global fully-connected layer features are both 4096-dimension, the concatenating features and the final feature embedding are 8192-dimension and 4096-dimension, and the projected feature vectors for retrieval are 1024-dimension, which can be shown in Figure 3. We fine-tune our network based on the ImageNet pre-trained model and choose by grid search the initial learning rate 0.001 (decreases to its 1/10 every 100 epochs), momentum 0.9, weight decay 0.0005, margin parameter $m = 0.2$ in Eq. 6, regularization parameters $\mu = 0.4$ in Eq. 7. For Stanford cars-196 dataset which has multi labels, we choose $\lambda_{category} = 0.6, \lambda_{type} = 0.3, \lambda_{make} = 0.2$ in Eq. 4.

4.3 Triplet Sampling Strategy

The training process for the experiments and our algorithm of constructing triplet units are detailed as follows. Firstly, we run stochastic gradient descent (SGD) to minimize softmax loss individually (*i.e.*, setting the loss weight of triplet loss to zero) for 20 epochs to reach a relatively stable state (*i.e.*, the classification accuracy maintaining a high level). Then we introduce our novel strategy of generating triplet units and minimize both softmax loss and triplet loss simultaneously for another 280 epochs until convergence. This two-stage procedure aims to make the method of choosing negative samples from the categories similar to positive ones meaningful. After stage one, we normalize the final fully-connected feature vector of each image and compute the mean feature vector of images belonging to the same fine-grained category respectively. Euclidean distances between each mean vector are calculated next and for every fine-grained category, its top-k (k = 14) neighboring categories are marked. To form an online training procedure, we repeat above neighboring categories selecting pattern every 50 epochs. Then, among a mini-batch, in which the batch-size is 64 and the number of positive samples is 8, the anchor is selected as the cluster center of the positive set as mentioned in [20], and the left 56 negative candidates are equally chosen among the closest k (we choose a relatively large number, *e.g.*, $k = 14$, to make sure the samples in a mini-batch are as stochastic as possible) neighboring categories. Within a mini-batch, we form each triplet unit with the anchor, a random positive sample and the hardest negative candidate among all negative samples.

5 RESULTS AND DISCUSSION

5.1 Comparison with State-of-the-art Methods on Stanford Cars-196 Dataset

Extensive experiments are conducted to evaluate our proposed framework on Stanford cars-196 dataset. We compare our method with three state-of-the-art baselines, which are: (1) distance metric learning by triplet loss [31], (2) triplet-based fine-tuning after softmax [24], (3) multi-task learning combined softmax with triplets [39]. The above three methods directly use global fully-connected features and differentiate by learning strategies. As revealed in Table 1, [39] achieves better performance, which proves the superiority of multi-task learning strategy. Therefore, we follow the same learning strategy as [39], while replace the generic fully-connected layer features with our novel feature embedding. The results from Row-4 of Table 1 present a significant performance gain, strongly demonstrating the efficiency of our introduced feature embedding. Then, we generalize the softmax loss to multi-label softmax loss, *i.e.*, simultaneously predict car category, car type and car make, as mentioned in Section 3.3. We just employ multi-label softmax loss for the experiments of Stanford cars-196, because only this dataset provides multi-label annotations. This technique further slightly improve the performance, as can be shown in Row-5 of Table 1. In summary, our method beats the latest state-of-the-art approach [39] by a relative 11.7% increase in mAP and 13.4% increase in top-40 retrieval precision.

Table 1: Performance comparison of state-of-the-art methods on Stanford cars-196.

Method	CNN Features	Learning Strategy	mAP	Top-40 Precision
Baseline1 [31]	Global Fully-connected Features	Triplet Loss	57.5	52.5
Baseline2 [24]	Global Fully-connected Features	Triplet Fine-Tune after Softmax	60.7	56.1
Baseline3 [39]	Global Fully-connected Features	**Multi-Task (Softmax and Triplet)**	72.4	67.8
Our Method	**Our Feature Embedding**	**Multi-Task (Softmax and Triplet)**	**79.3**	**75.1**
Our Method	**Our Feature Embedding**	**Multi-Task (Multi-label Softmax and Triplet)**	**80.9**	**76.9**

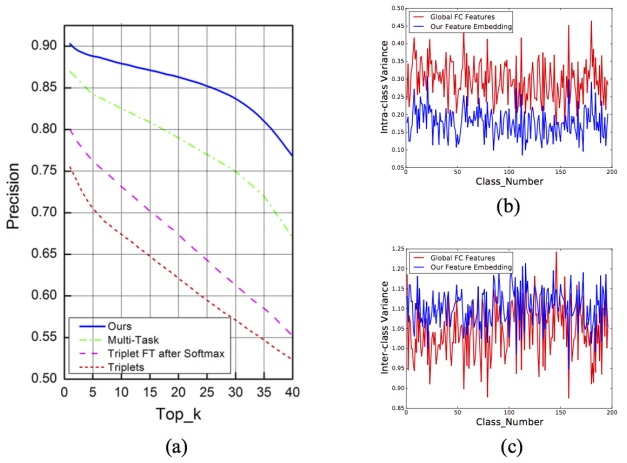

Figure 4: Comparisons on Stanford cars-196 dataset. (a) Top-k retrieval precision. (b) Intra-class variance. (c) Inter-class variance.

Figure 5: Results of top-6 retrievals based on the method utilizing global fully-connected layer features [39] (top) and our novel feature embedding (bottom). The true positive results are in green box, otherwise red.

In addition, Figure 4(a) intuitively illustrates the top-k retrieval precision results. From Figure 4(b) and Figure 4(c), we can clearly see that, by utilizing our feature embedding, the intra-class variance is dramatically decreased and the inter-class variance has a slight raise, compared with the original global fully-connected layer features. These results well verify our theorem in Section 3.1. Figure 5 shows some examples of retrieval results, where we can see that our method is significantly less sensitive to different viewpoints or poses and achieves obvious superiority compared with the existing methods.

We also sufficiently evaluate each of our algorithm components, including using local features alone and different local feature pooling methods. We give the detailed analysis in Section 5.2.

Table 2: mAP on Stanford cars-196 of different algorithm components. "FP" indicates Fixed-size Window Pooling, "GMP" indicates Global Max Pooling, "VP" indicates Variable-size Window Pooling, all mentioned in Section 3.2.

Method	mAP
Global FC Features Alone	72.4
Local Features Alone (FP (3×3))	70.7
Feature Embedding (FP (3×3))	78.3
Feature Embedding (GMP (1×1))	78.9
Feature Embedding (VP)	79.3
Feature Embedding (VP) & Multi-label Strategy	**80.9**

5.2 Effectiveness of Algorithm Components

Taking Stanford cars-196 dataset as an example, we detailed investigate the effects of the proposed algorithm modules, including using the local features alone, and our feature embedding with different local feature pooling techniques. The results are shown in Table 2.

We first evaluate the effect of using local features alone. From Table 2, we can see that the performance of using local features alone is worse than using global fully-connected features, while the final feature embedding is significantly superior to the above two. Such phenomenon is reasonable. As aforementioned, our proposed method is based on the assumption that, for images from a certain sub-category, their local features are more similar to each other than the images from another fine-grained class, *i.e.*, the local features satisfy Eq. 1. It is a weak assumption and we do not suppose local features are better than global fully-connected features. However, under such assumption, we prove that the concatenated features can enlarge the gap between inter-class variance and intra-class variance compared with the original global fully-connected features (see Eq. 2), which is the key factor to improve retrieval precision. As a result, the final feature embedding achieves the best performance.

We can also know that several methods of local feature extraction, including Fixed-size Window Pooling and Variable-size Window Pooling, are all effective designs. Generalizing the softmax loss to multi-label softmax loss can further slightly improve the performance because that it leverages multiple label information of the dataset.

5.3 Comparison with State-of-the-art Methods on CUB-200-2011 Dataset

We also evaluate the performance of our method on another fine-grained object dataset CUB-200-2011 [30], and Table 3 presents the

Table 3: Performance comparison of state-of-the-art methods on CUB-200-2011.

Method	mAP
Global Features via Triplet [31]	42.05
Global Features via Triplet Fine-Tune after Softmax [24]	46.57
Global Features via Multi-Task Learning [39]	55.04
Our Feature Embedding via Multi-Task Learning	**61.98**

Table 4: Performance comparison of several state-of-the-art methods at CMC ranks 1 on CUHK03 labeled dataset.

Method	rank-1
FPNN [18]	20.7
LOMO+XQDA [19]	52.2
Ahmed *et al.* [1]	54.7
Ensembles [23]	62.1
Fused Model [28]	72.4
Xiao *et al.* [32]	76.7
Feature Embedding (Ours)	**82.1**

results. We design this experiment to prove the generalized superiority of our proposed fine-grained feature embedding compared with the baseline of global CNN fully-connected features. Following the same evaluation criterion as Stanford cars-196, we compare our methods with three state-of-the-art baselines, which are all using global fully-connected layer features but different learning strategies. We use the same learning strategy as [39] while replace the fully-connected features with our novel feature embedding. Our method achieves a 61.98% mAP, still outperforming the existing result 55.04% [39] by a significant margin (a relative 12.6% improvement). It strongly proves a consistent performance gain with our feature embedding.

5.4 Comparison with State-of-the-art Methods on CUHK03 Dataset

The CUHK03 dataset is a large-scale challenging person re-ID dataset. Table 4 summarizes the experimental results on the CUHK03 labeled setting. We mainly focus on the comparison with [32] because it is similar to our method. Xiao *et al.* [32] show that for large-scale datasets, such as CUHK03, a carefully designed classification model achieves almost state-of-the-art performance. For fair comparison, we utilize the same basic CNN network and just leverage softmax loss (*i.e.*, set the loss weight of triplet loss to zero) to learn features, following the same settings as [32]. The only difference is that we replace the fully-connected features with our novel feature embedding. Here too, we see the amazing superiority of our model over the existing state-of-the-art approaches, a relative gain of up to 7% (82.1% vs. 76.7%) on CMC rank-1 accuracy.

6 CONCLUSION AND FUTURE WORK

In this paper, we propose a simple but effective feature embedding for fine-grained object retrieval by simultaneously encoding global fully-connected layer features and discriminative local semantic features, in which the local features are extracted by exploiting strong neural activations on the last convolutional layer. We demonstrate that our feature embedding can significantly enlarge the gap between inter-class variance and intra-class variance, which is the key factor to improve retrieval precision. In addition, we show our approach can also be applied in person re-identification. Experimental results on several challenging benchmarks show that our method yields state-of-the-art performances. In the future, we will explore the mechanism of effectively picking specific filters for the corresponding fine-grained category, making the procedure of extracting local semantic features more reasonable and robust.

ACKNOWLEDGMENTS

This paper is partially supported by NSFC (No.31627802) and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. 2015. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3908–3916.
[2] Artem Babenko and Victor Lempitsky. 2015. Aggregating local deep features for image retrieval. In *Proceedings of the IEEE international conference on computer vision*. 1269–1277.
[3] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. 2013. Learning articulated body models for people re-identification. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 557–560.
[4] Yuning Chai, Victor Lempitsky, and Andrew Zisserman. 2013. Symbiotic segmentation and part localization for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision*. 321–328.
[5] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. 2010. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research* 11, Mar (2010), 1109–1135.
[6] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, 539–546.
[7] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 1440–1448.
[8] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. 2014. Multi-scale orderless pooling of deep convolutional activation features. In *European Conference on Computer Vision*. Springer, 392–407.
[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*. Springer, 346–361.
[10] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. 2016. Part-Stacked CNN for Fine-Grained Visual Categorization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[11] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. 2010. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 3304–3311.
[12] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 675–678.
[13] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, Vol. 2.
[14] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. 2015. Fine-grained recognition without part annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5546–5555.
[15] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 554–561.
[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.
[17] Y Le Cun, B Boser, JS Denker, D Henderson, RE Howard, W Hubbard, and LD Jackel. 1989. Handwritten digit recognition with a back-propagation network. In *Proceedings of the 2nd International Conference on Neural Information Processing Systems*. MIT Press, 396–404.

[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 152–159.

[19] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2197–2206.

[20] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang. 2016. Deep Relative Distance Learning: Tell the Difference Between Similar Vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2167–2175.

[21] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. 2016. Multi-Scale Triplet CNN for Person Re-Identification. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 192–196.

[22] Hyeonjoon Moon and P Jonathon Phillips. 2001. Computational and performance aspects of PCA-based face-recognition algorithms. *Perception* 30, 3 (2001), 303–321.

[23] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. 2015. Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1846–1855.

[24] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition.. In *BMVC*, Vol. 1. 6.

[25] Mattis Paulin, Matthijs Douze, Zaid Harchaoui, Julien Mairal, Florent Perronin, and Cordelia Schmid. 2015. Local convolutional features with unsupervised training for image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*. 91–99.

[26] Marcel Simon and Erik Rodner. 2015. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 1143–1151.

[27] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[28] Arulkumar Subramaniam, Moitreya Chatterjee, and Anurag Mittal. 2016. Deep Neural Networks with Inexact Matching for Person Re-Identification. In *Advances in Neural Information Processing Systems*. 2667–2675.

[29] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

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

[31] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1386–1393.

[32] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1249–1258.

[33] Lingxi Xie, Qi Tian, Richang Hong, Shuicheng Yan, and Bo Zhang. 2013. Hierarchical part matching for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1641–1648.

[34] Mang Ye, Chao Liang, Zheng Wang, Qingming Leng, and Jun Chen. 2015. Ranking optimization for person re-identification via similarity and dissimilarity. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 1239–1242.

[35] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 34–39.

[36] Joe Yue-Hei Ng, Fan Yang, and Larry S Davis. 2015. Exploiting local features from deep networks for image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 53–61.

[37] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. 2014. Part-based r-cnns for fine-grained category detection. In *European Conference on Computer Vision*. Springer, 834–849.

[38] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. 2016. Picking deep filter responses for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1134–1142.

[39] Xiaofan Zhang, Feng Zhou, Yuanqing Lin, and Shaoting Zhang. 2016. Embedding label structures for fine-grained feature representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1114–1123.

[40] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*. Springer, 94–108.

[41] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. 2016. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*. Springer, 868–884.

[42] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*. 1116–1124.

[43] Liang Zheng, Yi Yang, and Alexander G Hauptmann. 2016. Person Re-identification: Past, Present and Future. *arXiv preprint arXiv:1610.02984* (2016).