

INCORPORATING INTRA-CLASS VARIANCE TO FINE-GRAINED VISUAL RECOGNITION

YAN BAI^{*,1,2}, FENG GAO^{*,2}, YIHANG LOU^{1,2}, SHIQI WANG³, TIEJUN HUANG², LING-YU DUAN²

¹SECE of Shenzhen Graduate School, Peking University, Shenzhen, China

²National Engineering Lab for Video Technology, Peking University, Beijing, China

³Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore

ABSTRACT

Fine-grained visual recognition aims to capture discriminative characteristics amongst visually similar categories. The state-of-the-art research work has significantly improved the fine-grained recognition performance by deep metric learning using triplet network. However, the impact of intra-category variance on the performance of recognition and robust feature representation has not been well studied. In this paper, we propose to leverage intra-class variance in metric learning of triplet network to improve the performance of fine-grained recognition. Through partitioning training images within each category into a few groups, we form the triplet samples across different categories as well as different groups, which is called Group Sensitive TRiplet Sampling (GS-TRS). Accordingly, the triplet loss function is strengthened by incorporating intra-class variance with GS-TRS, which may contribute to the optimization objective of triplet network. Extensive experiments over benchmark datasets CompCar and VehicleID show that the proposed GS-TRS has significantly outperformed state-of-the-art approaches in both classification and retrieval tasks.

Index Terms— Fine-grained visual recognition, Metric learning, Intra-class variance

1. INTRODUCTION

Fine-grained visual recognition aims to reliably differentiate fine details amongst visually similar categories. For example, fine-grained car recognition [1, 2] is to identify a specific car model in an image, such as “Audi A6 2015 model”. Recently, more research efforts in fine-grained visual recognition have been extended to a variety of vertical domains, such as recognizing the breeds of animals [3, 4, 5], the identities of pedestrians [6, 7, 8] and the types of plants [9, 10, 11], etc. The challenges of fine-grained visual recognition basically relate to two aspects: inter-class similarity and intra-class variance. On the one hand, the instances of different fine categories may exhibit highly similar appearance features. On the other hand, the instances within a fine category may produce significantly

variant appearance from different viewpoints, poses, motions and lighting conditions.

To mitigate the negative impact of inter-class similarity and/or intra-class variance on the fine-grained visual recognition, lots of research work has been done [12, 13, 14]. Various part-based approaches [12, 13] have been proposed to capture the subtle “local” structure for distinguishing classes and reducing the intra-class variance of appearance features from the changes of viewpoint or pose, etc. For example, for fine-grained birds recognition in [13], zhang *et al.* proposed to learn the appearance models of parts (*i.e.*, head and body) and enforce geometric constraints between parts. However, part-based methods rely on accurate part localization, which would fail in the presence of large viewpoints variations. In addition, recently, more promising methods [14, 15, 16] based on metric learning, which aims to maximize inter-class similarity distance and meanwhile minimize intra-class similarity distance, have been proposed. In particular, a sort of triplet constraint in [14] is introduced to learn a useful triplet embedding based on similarity triplets of the form “sample *A* is more similar to sample *P* in the same class as sample *A* than to sample *N* in a different class”.

On the other hand, some methods [17, 18] utilize multiple labels, which are meant to denote the intrinsic relationship of properties in images, to learn a variety of similarity distances of relative, sharing or hierarchical attributes. In [17], multiple labels are leveraged to inject hierarchical inter-class relationship of attributes into learning feature representation . Lin *et al.* [18] utilized bipartite-graph labels to model rich inter-class relationships based on multiple sub-categories, which can be elegantly incorporated into convolutional neural network. However, those methods focus on the inter-class similarity distance, whereas the intra-class variance and its related triplet embedding have not been well studied in learning feature representation. When a category exhibits high intra-class appearance variance, intra-class triplet embedding is useful to deal with the complexity of feature space.

In this paper, we propose a novel Group Sensitive TRiplet Sampling (GS-TRS) approach, which attempts to incorporate the modeling of intra-class variance into triplet network. A so-called grouping is to figure out a mid-level representation within each fine-grained category to capture the intra-class

*YAN BAI and FENG GAO are joint first authors.
LING-YU DUAN is the corresponding author.

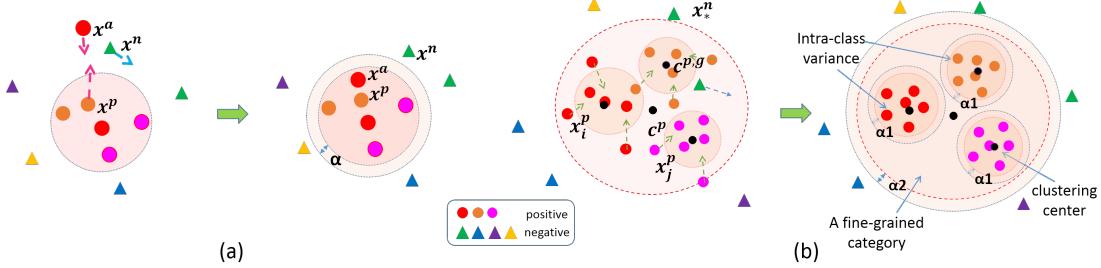


Fig. 1. Illustration of traditional triplet loss (a) and the intra-class variance (ICV) incorporated triplet loss (b). The instances denoted by different colors in (b), which can be sorted out by grouping in terms of some features or attributes. The ICV triplet loss further enforces that the samples within each group should be drawn closer. By contrast, the traditional triplet loss in (a) does not take the intra-class structure into account (Best viewed in color).

variance and intra-class invariance. In practice, clustering can be applied to implement the grouping. Given a fine-grained category, instances are clustered to a set of groups. To formulate the triplet loss function, we need to consider the inter-class triplet embedding and the inter-group triplet embedding. The latter works on intra-class variance. The proposed GS-TRS has been proved to be effective in triplet learning, which can significantly improve the performance of triplet embedding in the presence of considerable intra-class variance.

Our main contributions are twofold. Firstly, we incorporate the modeling of intra-class variance into triplet network learning, which can significantly mitigate the negative impact of inter-class similarity and/or intra-class variance on fine-grained classification. Secondly, by optimizing the joint objective of softmax loss and triplet loss, we can generate effective feature representations (*i.e.*, feature maps in Convolution Neural Network) for fine-grained retrieval. In extensive experiments over benchmark, the proposed method outperforms state-of-the-art fine-grained visual recognition approaches.

The rest of this paper is organized as follows. In Section 2, we formulate the problem of injecting the modeling of intra-class variance into triplet embedding for fine-grained visual recognition. In Section 3, we present the proposed GS-TRS approach. Extensive experiments are discussed in Section 4, and finally we conclude this paper in Section 5.

2. PROBLEM STATEMENT

2.1. Problem Formulation

Let $S_{c,g}$ denote a set of instances of the g_{th} group in fine-grained category c , and S_n are a set of instances not in category c . Assume each category c consists of G groups, where the set of distinct groups may represent intra-class variance, and each individual group may represent intra-class invariance. The objective of preserving intra-class structure in metric learning is to minimize the distances of samples in the same group for each category when the distances of samples from different categories exceed a minimum margin α .

$$\begin{aligned} & \min \sum_{g=1}^G \sum_{x_i, x_j \in S_{c,g}} \|x_i - x_j\|^2 \\ & \text{s.t. } \sum_{x_i \in S_{c,g}, x_k \in S_n} \|x_i - x_k\|^2 \geq \alpha, \end{aligned} \quad (1)$$

where samples x_i and x_j from category c fall in the same group g ; x_k is from the other category; and α is the minimum margin constraint between samples from different categories.

Eq (1) can be optimized by deep metric learning using triplet network. The remaining issue is to model the intra-class variance of each fine-grained category and properly establish triplet units to accommodate the variance structure.

2.2. Triplet Learning Network

Our proposed GS-TRS approach works on a triplet network model. The main idea of triplet network is to project images into a feature space where those pairs belonging to the same category are closer than those from different ones. Let $\langle x^a, x^p, x^n \rangle$ denote a triplet unit, where x^a and x^p belong to the same category, and x^n belongs to the other category. The constraint can be formulated as:

$$\|f(x^a) - f(x^p)\|^2 + \alpha \leq \|f(x^a) - f(x^n)\|^2, \quad (2)$$

where $f(x)$ is the feature representation of image x , α is the minimum margin between positives and negatives. If the distances between positive and negative pairs violate the constraint in (2), then loss will be back propagated. Thus, the loss function can be defined as:

$$L = \sum^N \frac{1}{2} \max \{\|f(x^a) - f(x^p)\|_2^2 + \alpha - \|f(x^a) - f(x^n)\|_2^2, 0\}. \quad (3)$$

However, there exist two practically important issues in triplet network. First, triplet loss constrains samples of the same class together, while the class-inherent relative distances associated with intra-class variance cannot be well preserved, as illustrated in Fig. 1 (a). Second, triplet loss is sensitive to the selection of anchor x^a , and improper anchors can seriously degrade the performance of triplet network learning.

3. GS-TRS APPROACH

The proposed GS-TRS incorporates intra-class variance into triplet network in which the learning process involves: (1) clustering each category into groups, (2) incorporating intra-class variance into triplet loss, (3) a multiple loss function.

3.1. Intra-class Variance

To characterize intra-class variance, grouping is required. Unlike category labels, intrinsic attributes within a category are



Fig. 2. Exemplar car images from different groups, which are obtained by applying clustering ($K = 5$) to the images of a specific car model in CompCar dataset. Different groups may be interpreted by some attributes (e.g., viewpoints or colors.) latent or difficult to precisely describe (e.g. lighting conditions, backgrounds). Here, we prefer an unsupervised approach to grouping images for each category.

Firstly, we feed image instances in each fine-grained category into the VGG_CNN_M_1024 (*VGGM*) network obtained by pre-training on ImageNet dataset. Then, we extract the last fully-connected layer's output as the feature representation, followed by Principal Component Analysis (PCA) based feature dimension reduction. Finally, K-means is applied to perform clustering:

$$\arg \min \sum_{g=1}^G \sum_{x=1}^{N^{p,g}} \|f(x) - \mu_g\|^2, \quad (4)$$

where G is the number of cluster center μ_g (i.e., group num). $N^{p,g}$ is the number of samples contained in $S_{c,g}$. Each image instance is assigned a group ID after clustering. As illustrated in Fig. 2, grouping often relates to meaningful attributes.

3.2. Mean-valued Triplet Loss

An anchor in triplet units is often randomly selected from positives. To alleviate the negative effects of improper anchor selection, we determine the anchor by computing the mean value of all positives, and formulate a mean-valued triplet loss. Given a positive set $X^p = \{x_1^p, \dots, x_{N^p}^p\}$ containing N^p positive samples and a negative set $X^n = \{x_1^n, \dots, x_{N^n}^n\}$ containing N^n samples from other categories. Thus, the mean-valued anchor can be formulated as:

$$c^p = \frac{1}{N^p} \sum_i^{N^p} f(x_i^p), \quad (5)$$

where $1 \leq i \leq N^p$ and $1 \leq j \leq N^n$. Rather than using randomly selected anchors, the proposed mean-valued triplet loss function is formulated as follows:

$$L(c^p, X^p, X^n) = \sum_i^{N^p} \frac{1}{2} \max\{\|f(x_i^p) - c^p\|_2^2 + \alpha - \|f(x_*^n) - c^p\|_2^2, 0\}, \quad (6)$$

where x_*^n is the negative closest to anchor c^p . It is worthy to note that, although the mean value of positives is considered as an anchor, the backward propagation needs to get all the positives involved. The advantage will be demonstrated in the

subsequent experiments. When the anchor is computed by all of the positives, the triplet $< c^p, x_i^p, x_j^n >$ may not satisfy the constraints $\|f(x_i^p) - c^p\|_2^2 + \alpha \leq \|f(x_j^n) - c^p\|_2^2$. Hence, all the positives involving mean value computing are enforced to perform backward propagation. The partial derivative of positive sample x_i^p is:

$$\frac{\partial L}{\partial f(x_i^p)} = f(x_i^p) - c^p + \frac{1}{N^p} (f(x_*^n) - f(x_i^p)). \quad (7)$$

The partial derivative of other positives x_k^p ($k! = i$) is:

$$\frac{\partial L}{\partial f(x_k^p)} = \frac{1}{N^p} (f(x_*^n) - f(x_i^p)). \quad (8)$$

The partial derivative of negative samples is:

$$\frac{\partial L}{\partial f(x_*^n)} = c^p - f(x_*^n). \quad (9)$$

3.3. Incorporating Intra-Class Variance into Mean-valued Triplet Loss

To enforce the preservation of relative distances associated with intra-class variance, we introduce Intra-Class Variance loss (ICV loss) into triplet learning. Let c^p denote a mean center (the mean value of samples) in category c and $c^{p,g}$ denote a group center that is the mean value of samples in group g of category c . For each category c , there are one mean center c^p and G group centers $c^{p,g}$. As illustrated in Fig. 1 (b), each black dot represents the center of a group.

In terms of intra-class variance, x_i^p, x_j^p denote two samples from different groups within c . In terms of inter-class relationship, $x_k^p \in c$ are positives, and $x_*^n \notin c$ are negatives. To incorporate the intra-class variance into triplet embedding, we formulate the constraints as:

$$\begin{aligned} \|c^p - f(x_i^p)\|^2 + \alpha_1 &\leq \|c^p - f(x_*^n)\|^2 \\ \|c^{p,g} - f(x_i^p)\|^2 + \alpha_2 &\leq \|c^{p,g} - f(x_j^p)\|^2, \end{aligned} \quad (10)$$

where α_1 is the minimum margin between those samples from different categories, and α_2 is the minimum margin between those samples from different groups within the same category. Accordingly, we formulate the ICV incorporated mean-valued triplet loss as follows:

$$\begin{aligned} L_{ICV.Triplet} &= L_{inter}(c^p, x_k^p, x_*^n) + \sum_{g=1}^G L_{intra}(c^{p,g}, x_i^p, x_j^p) \\ &= \sum_{k=1}^{N_p} \frac{1}{2} \max\{\|c^p - f(x_k^p)\|^2 + \alpha_1 - \|c^p - f(x_*^n)\|^2, 0\} \\ &\quad + \sum_{g=1}^G \sum_{i=1}^{N^{p,g}} \frac{1}{2} \max\{\|c^{p,g} - f(x_i^p)\|^2 + \alpha_2 - \|c^{p,g} - f(x_j^p)\|^2, 0\}. \end{aligned} \quad (11)$$

3.4. Joint Optimization of Multiple Loss Function

ICV triplet loss alone does not suffice for effective and efficient feature learning in triplet network. Firstly, given a dataset of N images, the number of triplet units is $O(N^3)$, while each iteration in training often selects dozens of triplet units, and only a minority may violate the constraints. So

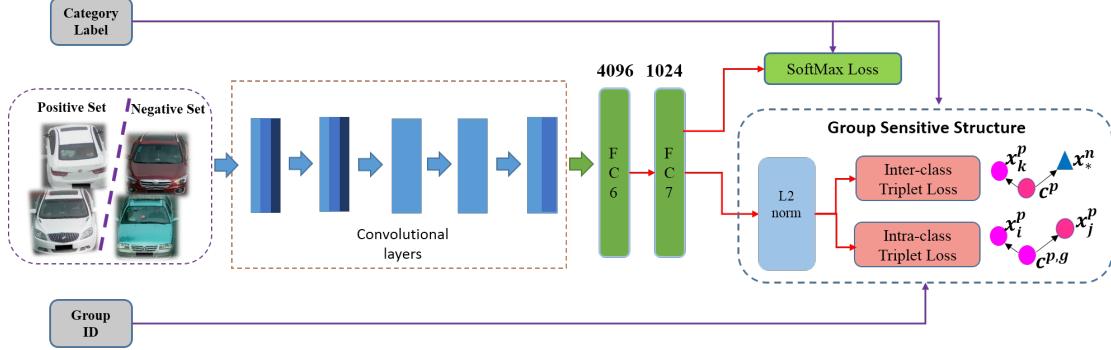


Fig. 3. Illustration of a triplet network by incorporating intra-class variance into triplet embedding, in which the joint learning objective is to minimize the combination of softmax loss and triplet loss (consisting of inter-class and intra-class triplet loss).

the solely ICV triplet loss based learning incurs much slower convergence than classification. Secondly, as the triplet loss works on similarity distance learning rather than hyperplane decision, the discriminative ability of features can be improved by adding the classification loss to the learning objective. Hence, we propose a GS-TRS loss to jointly optimize the ICV triplet combinatin loss and softmax loss in a multi-task learning manner. A simple linear weighting is applied to construct the final loss function as follows:

$$L_{GS-TRS} = \omega L_{softmax} + (1 - \omega)L_{ICV_triplet}, \quad (12)$$

where ω is fusion weight. Fig.3 illustrates the triplet network. Optimizing this multi-loss function helps accomplish promising fine-grained categorization performance as well as discriminative features for fine-grained retrieval. We will investigate the effects of ICV_triplet loss with or without mean-valued anchor on GS-TRS loss in the experiments.

4. EXPERIMENTS

4.1. Experiments Setup

Baselines To evaluate and compare the triplet network based fine-grained visual recognition methods, we setup baseline methods as follows: (1) triplet loss [16], (2) triplet + softmax loss [15], (3) mixed Diff + CCL [19], (4) HDC + Contrastive [20], (5) GS-TRS loss without a mean-valued anchor for each group, *i.e.*, a randomly selected anchor (GS-TRS loss W/O mean), (6) GS-TRS loss with a mean-valued anchor for each group (GS-TRS loss W/ mean). We select the output of L2 Normalization layer as feature representation for retrieval and re-identification (ReID) tasks. For fair comparison, we adopt the base network structure VGG_CNN_M_1024 (VGGM) as in [19]. The networks are initialized with the pre-trained model over ImageNet.

DataSet Comparison experiments are carried out over benchmark datasets VehicleID [19] and CompCar [1]. VehicleID dataset consists of 221,763 images with 26,267 vehicles (about 250 vehicle models) captured by different surveillance cameras in a city. There are 110,178 images available for model training and three gallery test sets. The numbers of gallery images in small, medium and large sets are

800, 1,600 and 2,400 for retrieval and re-identification experiments. CompCar is another large-scale vehicle image dataset, in which car images are mostly collected from Internet. We select the Part-I subset for training that contains 431 car models (16,016 images) and the remaining 14,939 images for test. Note that all the selected images involve more or less backgrounds. We conduct retrieval and ReID experiments on VehicleID dataset, and retrieval and classification experiments on CompCar dataset.

Evaluation Metrics For retrieval performance evaluations, we use mAP and mean precision @ K . For ReID evaluation, we apply the widely used cumulative match curve (CMC). For classification evaluation, we use the mean percentage of those images accurately classified as the groundtruth.



(a) GS-TRS Loss (ICV triplet+ softmax loss)



(b) triplet loss + softmax loss

Fig. 4. Exemplar Top-10 retrieval results on CompCar dataset. The images with a dashed rectangle are wrong results. The GS-TRS loss with grouping yields better results in (a) than the traditional triplet loss without grouping in (b).

4.2. Performance Comparison on VehicleID Dataset

Retrieval Table 1 lists the retrieval performance comparisons. Note that during the training stage, unlike [8, 19] treating each vehicle model as a category, we treat each vehicle ID as a class (*i.e.*, 13,134 vehicles classes). As listed in Table 1, directly combining softmax and triplet loss has outperformed Mixed Diff+CCL [19] with significant mAP gain of 19.5% in the large test set. Furthermore, our proposed GS-TRS loss

without mean-valued anchors can consistently achieve significant improvements across three different scale subsets. In particular, the additional improvement on large test set reaches up to 4.6% mAP. Compared to [19], the improvement on large set has been up to 23.9% mAP. Moreover, GS-TRS loss with mean-valued anchors can further obtain about 2% mAP gains since using mean values of positives from multiple groups within a category yields more reliable anchors, which contributes to better triplet embedding.

Table 1. The mAP results of vehicle retrieval task.

Methods	Small	Medium	Large
Triplet Loss [8]	0.444	0.391	0.373
CCL [19]	0.492	0.448	0.386
Mixed Diff+CCL [19]	0.546	0.481	0.455
Softmax Loss	0.625	0.609	0.580
HDC + Contrastive [20]	0.655	0.631	0.575
Triplet+Softmax Loss [15]	0.695	0.674	0.650
GS-TRS loss W/O mean	0.731	0.718	0.696
GS-TRS loss W/ mean	0.746	0.734	0.715

Re-identification Table 2 presents re-identification performance comparisons. Our proposed method GS-TRS loss with mean-valued anchors achieves +30% improvements over Mixed Diff+CCL in the large test set. Such significant improvements can be attributed to two aspects: First, we extend the softmax classification to the granularity level of vehicle ID, rather than the granularity level of vehicle model in [19]. Second, we have improved the similarity distance learning by introducing the intra-class feature space structure and its relevant loss function to triplet embedding. Moreover, from the performance comparisons of combining different triplet loss functions and softmax loss in Top1 and Top5, both the proposed GS-TRS loss without mean-valued anchors and the further improved GS-TRS loss with mean-valued anchors have yielded significant performance gains. More match rate details of different methods from Top 1 to Top 50 on the small test set are given in Fig. 5.

Table 2. The results of match rate in vehicle ReID task.

Method	Small	Medium	Large
Triplet Loss [8]	0.404	0.354	0.319
CCL [19]	0.436	0.370	0.329
Mixed Diff+CCL [19]	0.490	0.428	0.382
Triplet+Softmax Loss [15]	0.683	0.674	0.653
GS-TRS loss W/O mean	0.728	0.720	0.705
GS-TRS loss W/ mean	0.750	0.741	0.732
Triplet Loss [8]	0.617	0.546	0.503
CCL [19]	0.642	0.571	0.533
Mixed Diff+CCL [19]	0.735	0.668	0.616
Triplet+Softmax Loss [15]	0.771	0.765	0.751
GS-TRS loss W/O mean	0.814	0.805	0.789
GS-TRS loss W/ mean	0.830	0.826	0.819

4.3. Performance Comparison on CompCar Dataset

Retrieval Table 3 lists the TopK precision comparisons. The incorporation of intra-class variance into triplet embedding

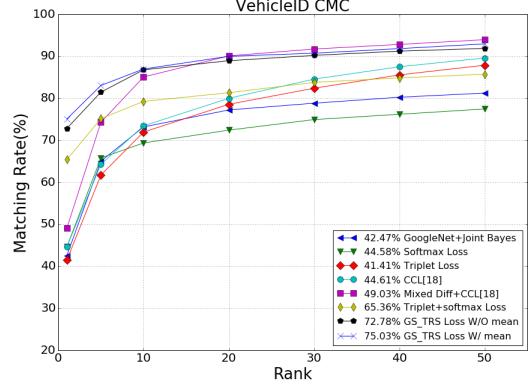


Fig. 5. CMC Results on VehicleID dataset.

can achieve more than 5.6% precision gains at top-500. Overall, the modeling of intra-class variance and its injection into triplet network can significantly improve the discriminative power of feature representation which plays a significant role in fine-grained image retrieval. Fig. 4 gives the retrieval results of an exemplar query over CompCar dataset before and after injecting GS-TRS into triplet embedding.

Table 3. mean precision @ K on CompCars retrieval task.

mean precision @ K	1	50	500	All (mAP)
Triplet Loss [8]	0.502	0.371	0.198	0.122
Softmax Loss	0.456	0.282	0.167	0.091
Triplet+Softmax Loss [15]	0.719	0.586	0.419	0.349
GS-TRS loss W/O mean	0.734	0.603	0.475	0.376
GS-TRS loss W/ mean	0.756	0.620	0.497	0.393

Classification We train a VGGM network with single softmax loss and set initial learning rate = 0.002 and total iteration = 80K, and then yield 78.24% classification accuracy. Further fine-tuning with triplet+softmax loss can bring about 0.7% classification accuracy improvement, while GS-TRS loss with mean-valued anchors can yield more accuracy improvement of 1.6% (*i.e.*, the classification accuracy is 79.85%). Such improvements demonstrate that preserving intra-class variance is beneficial for fine-grained categorization as well.

5. CONCLUSION

We have proposed a novel approach GS-TRS to improve triplet network learning through incorporating the intra-class variance structure into triplet embedding. The multi-task learning of both GS-TRS triplet loss and softmax loss has significantly contributed to fine-grained image retrieval and classification. How to further optimize the grouping strategy as well as the selection of anchors with respect to meaningful and effective groups is included in our future work.

Acknowledgments: This work was supported by grants from National Natural Science Foundation of China (U1611461, 61661146005, 61390515) and National Hightech R&D Program of China (2015AA016302). This research is partially supported by the PKU-NTU Joint Research Institute,

that is sponsored by a donation from the Ng Teng Fong Charitable Foundation.

6. REFERENCES

- [1] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang, “A large-scale car dataset for fine-grained categorization and verification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3973–3981.
- [2] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei, “3d object representations for fine-grained categorization,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 554–561.
- [3] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar, “Cats and dogs,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3498–3505.
- [4] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li, “Novel dataset for fine-grained image categorization: Stanford dogs,” in *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, 2011, vol. 2.
- [5] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur, “Birdsnap: Large-scale fine-grained visual categorization of birds,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 2019–2026.
- [6] Ejaz Ahmed, Michael Jones, and Tim K Marks, “An improved deep learning architecture for person re-identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3908–3916.
- [7] Dong Yi, Zhen Lei, and Stan Z Li, “Deep metric learning for practical person re-identification,” *arXiv preprint arXiv:1407.4979*, 2014.
- [8] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao, “Deep feature learning with relative distance comparison for person re-identification,” *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.
- [9] Qi Qian, Rong Jin, Shenghuo Zhu, and Yuanqing Lin, “Fine-grained visual categorization via multi-stage metric learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3716–3724.
- [10] Yin Cui, Feng Zhou, Yuanqing Lin, and Serge Belongie, “Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop,” *arXiv preprint arXiv:1512.05227*, 2015.
- [11] Asma Rejeb Sfar, Nozha Boujemaa, and Donald Geman, “Vantage feature frames for fine-grained categorization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 835–842.
- [12] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang, “The application of two-level attention models in deep convolutional neural network for fine-grained image classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 842–850.
- [13] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell, “Part-based r-cnns for fine-grained category detection,” in *European Conference on Computer Vision*. Springer, 2014, pp. 834–849.
- [14] Kilian Q Weinberger and Lawrence K Saul, “Distance metric learning for large margin nearest neighbor classification,” *Journal of Machine Learning Research*, vol. 10, no. Feb, pp. 207–244, 2009.
- [15] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu, “Learning fine-grained image similarity with deep ranking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1386–1393.
- [16] Florian Schroff, Dmitry Kalenichenko, and James Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.
- [17] Xiaofan Zhang, Feng Zhou, Yuanqing Lin, and Shaoting Zhang, “Embedding label structures for fine-grained feature representation,” *arXiv preprint arXiv:1512.02895*, 2015.
- [18] Feng Zhou and Yuanqing Lin, “Fine-grained image classification by exploring bipartite-graph labels,” *arXiv preprint arXiv:1512.02665*, 2015.
- [19] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang, “Deep relative distance learning: Tell the difference between similar vehicles,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2167–2175.
- [20] Yuhui Yuan, Kuiyuan Yang, and Chao Zhang, “Hardware deeply cascaded embedding,” *arXiv preprint arXiv:1611.05720*, 2016.