Dual Cross-Entropy Loss for Small-Sample Fine-Grained Vehicle Classification

Xiaoxu Li, Liyun Yu, Dongliang Chang, Zhanyu Ma, and Jie Cao

***Abstract*—Fine-grained vehicle classification is a challenging topic in computer vision due to the high intraclass variance and low interclass variance. Recently, considerable progress has been made in fine-grained vehicle classification due to the huge success of deep neural networks. Most studies of fine-grained vehicle classification based on neural networks, focus on the neural network structure to improve the classification performance. In contrast to existing works on fine-grained vehicle classification, we focus on the loss function of the neural network. We add a regularization term to the cross-entropy loss and propose a new loss function, *Dual Cross-Entropy Loss*. The regularization term places a constraint on the probability that a data point is assigned to a class other than its ground-truth class, which can alleviate the vanishing of the gradient when the value of the cross-entropy loss is close to zero. To demonstrate the effectiveness of our loss function, we perform two sets of experiments. The first set is conducted on a small-sample fine-grained vehicle classification dataset, the Stanford Cars-196 dataset. The second set is conducted on two small-sample datasets, the LabelMe dataset and the UIUC-Sports dataset, as well as on one large-sample dataset, the CIFAR-10 dataset. The experimental results show that the proposed loss function improves the fine-grained vehicle classification performance and has good performance on three other general image classification tasks.**

***Index Terms*—Cross-entropy loss, fine-grained vehicle classification, deep neural networks.**

I. Introduction

With the development of society, the use of vehicles in human life has become increasingly universal and crucial. Research on vehicles has received considerable attention [1]–[3], including applications in the field of computer vision, such as vehicle classification [4]–[7], vehicle detection [8]–[10], vehicle segmentation [11], vehicle re-identification (re-ID) [12], [13], and fine-grained vehicle classification [14], [15]. In this paper, we focus on fine-grained vehicle classification, which refers to the task of identifying the make, model, and year of a vehicle, such as Audi, SUV, 2010. Due to the high intraclass variance and low interclass variance of vehicle images, fine-grained vehicle classification remains a challenging problem [16], [17].

Recently, convolutional neural networks (CNNs) have had a substantial impact on various fields of computer vision, including image classification [18], [19], character recognition [20], [21], object recognition [22], [23], video tracking [24], and image retrieval [25]. Fine-tuning a large CNNs that is trained on a large dataset, such as ImageNet [26], usually produces impressive results for many vision classification tasks [27].

Many fine-grained vehicle classification works based on CNNs have been reported [28]–[30]. Liu *et al.* [31] and Yang *et al.* [4] investigated the use of some of the first deep learning models for fine-grained vehicle classification. Their GoogleNet model outperformed some of the traditional, part-based approaches, reinforcing the belief that the use of deep CNNs for fine-grained classification problems is a viable approach, even if such an approach did not achieve state-of-the-art results at the time.

Valev *et al.* used CNN architectures for the model classification of vehicles and achieved performance competitive with that of state-of-the-art methods [16]. Lin *et al.* proposed a bilinear CNN model consisting of two feature extractors that are combined to obtain an image descriptor for fine-grained image classification [32]. The method performed well on a number of fine-grained datasets and was simple and easy to train. Hu *et al.* proposed a spatially weighted pooling strategy that greatly improved the robustness and effectiveness of the feature representation of most dominant deep CNNs and achieved state-of-the-art performance on the Stanford Cars-196 dataset [33], [34]. Zhao *et al.* proposed a diversified visual attention network that dynamically visits important regions during the training process and performs well on fine-grained object classification [35]. Most of these works improved the structure of the neural network and achieved good performance. In contrast to these studies, this paper focuses on the loss function of the neural network.

Among a multitude of neural network loss functions, cross-entropy loss (CE loss) is one of the most popular. A number of works have attempted to improve CE loss, such as large-margin loss [36], center loss [37], and focal loss [38]. Liu *et al.* considered that CE loss does not explicitly encourage the discriminative learning of features and proposed large-margin loss, which improves the accuracy of classification by increasing

Manuscript received December 24, 2018; revised January 18, 2019; accepted January 19, 2019. Date of publication January 28, 2019; date of current version May 28, 2019. This work was supported in part by the National Natural Science Foundation of China under Grants 61563030, 61402047, and 61763028; in part by the Natural Science Foundation of Gansu Province, China, under Grant 17JR5RA125; and in part by the Hongliu Outstanding Youth Talents Foundation of Lanzhou University of Technology. The review of this paper was coordinated by the Guest Editors of the Special Section on Machine Learning-Based Internet of Vehicles. *(Corresponding author: Zhanyu Ma.)*

X. Li, L. Yu, D. Chang, and J. Cao are with the School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China (e-mail: xiaoxulilut@gmail.com; yuliyunlut@hotmail.com; dlchanglut@hotmail.com; caoj@lut.cn).

Z. Ma is with the Pattern Recognition and Intelligent System Laboratory, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: mazhanyu@bupt.edu.cn).

Digital Object Identifier 10.1109/TVT.2019.2895651

interclass separability and intraclass compactness [36]. Wen *et al.* aimed to enhance the discriminative power of the deeply learned features and proposed center loss, which simultaneously learns a center for the deep features of each class and constrains the distances between the deep features and their corresponding class centers [37]. Li *et al.* proposed focal loss to address this class imbalance by reshaping the standard CE loss such that it down-weights the loss assigned to well-classified examples. Focal loss surpasses the performance of all existing state-of-the-art methods [38]. Liu *et al.* proposed FaceNet and triplet loss for face verification. Triplet loss focuses on enforcing the margin between each pair of faces, from one person's face to all other faces [39]. Liu *et al.* proposed congenerous cosine loss, which improved the accuracy of person recognition by minimizing the cosine distance between samples and their cluster center [13]. All of these works improved different aspects of the original CE loss.

In this paper, we note that CE loss focuses on the probability that a data point is assigned to its ground-truth class and does not place any constraint on the probability that the data point is assigned to a class other than its ground-truth class. During the training process, the probabilities that the data point is assigned to classes other than its ground-truth class can either increase or decrease. Therefore, if we constrain the increase in these probabilities, the optimization speed can be accelerated, and the performance of the network can be improved indirectly.

To this end, we propose *Dual Cross-Entropy Loss*, which is a linear combination of two items: CE loss and a regularization term, and perform two sets of experiments on four datasets. The first set is conducted on a fine-grained vehicle dataset (Stanford Cars-196), and the second set is conducted on three general image classification datasets: the LabelMe dataset [40], the UIUC-Sports dataset [41] and the CIFAR-10 dataset [42]. The experimental results demonstrate that compared with CE loss, the proposed Dual Cross-Entropy Loss has superior performance on the four datasets, accelerates optimization, and improves the performance within limited epochs. The contribution of this paper are briefly summarized as that we proposed a new regularized CE loss, Dual Cross-Entropy Loss, which alleviate the vanishing of the gradient when the value of the CE loss is close to zero, so as to improve generalization performance.

II. Dual Cross-Entropy Loss

Before introducing the proposed loss, we first review multi-class CE loss.

A. Cross-Entropy Loss

Suppose that $D = \{(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_M, y_M)\}$ is a training dataset of M samples, where y_i, a one-hot vector, is the label of the ith sample x_i, and suppose that p_i is a vector in which the j th, $j \in \{1, 2, \ldots, C\}$, element is the probability that sample x_i is assigned to the jth class. Then, the CE loss can be defined as follows (1):

$$L_{CE} = -\frac{1}{M}\sum_{i=1}^{M}(\boldsymbol{y_i}^T \ log(\boldsymbol{p_i})) \tag{1}$$

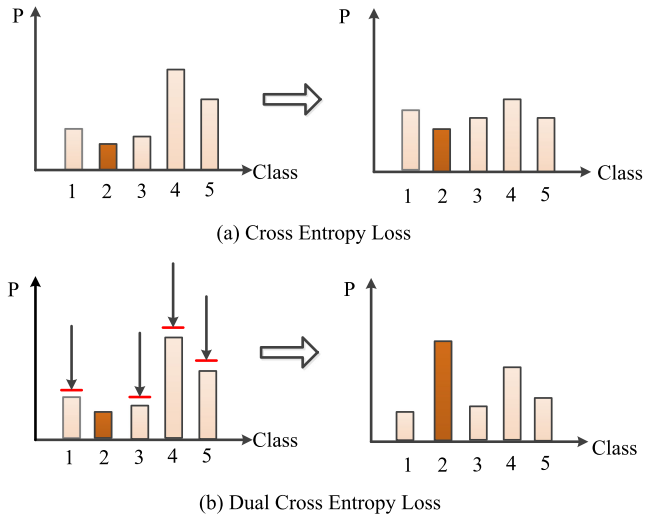


Fig. 1. Motivation of Dual Cross-Entropy Loss. (a) Explanation of CE loss, which focuses on the probability that a data point is assigned to its ground-truth class (labeled in brown) and does not place any constraint on the probability that the data point is assigned to a class other than its ground-truth class. (b) The idea of Dual Cross-Entropy Loss, which not only focuses on the probability that a data point is assigned to its ground-truth class but also adds a constraint on the probability that the data point is assigned to a class other than its ground-truth class.

From the formulation, CE loss focuses only on the probability that a data point is assigned to its ground-truth class and does not place any constraint on the probability that the data point is assigned to a class other than its ground-truth class, as shown in Figure 1 (a). Figure 1 (a) shows, given a data point, the possible changes in the probability distribution over classes due to minimization of the CE loss, where the ground-truth class label of the data point is 2. Except for the second class, the probabilities of some classes increase, and the probabilities of some classes decrease in the process of optimization.

B. Dual Cross-Entropy Loss

In this paper, we propose a new loss that not only focuses on the probability that a data point is assigned to its ground-truth class but also decreases the probabilities that the data point is assigned to a class other than its ground-truth class. We call the proposed loss *Dual Cross-Entropy Loss*. See Equations (2) and (3):

$$L_{DCE} = L_{CE} + \beta L_r \tag{2}$$

$$L_r = \frac{1}{M}\sum_{i=1}^{M}((1 - \boldsymbol{y_i})^T log(\alpha + \boldsymbol{p_i})) \tag{3}$$

where L_{CE} is responsible for increasing the probability that a data point is assigned to its ground-truth class, L_r is responsible for decreasing the probability that the data point is assigned to a class other than its ground-truth class, $\alpha > 0$, $\beta \geq 0$, and the meanings of M, y_i and p_i are the same as those in Equation (1).

When β is larger, the impact of L_r on L_{CE} is greater. When β is equal to 0, our loss L_{DCE} collapses to the CE loss. Our loss aims to increase the probability that a data point is assigned to its ground-truth class while simultaneously decreasing the

Fig. 2. Example images of the Stanford Cars-196 dataset. The top images are the original images in the dataset, and the bottom images are the ground-truth annotations of the bounding boxes.

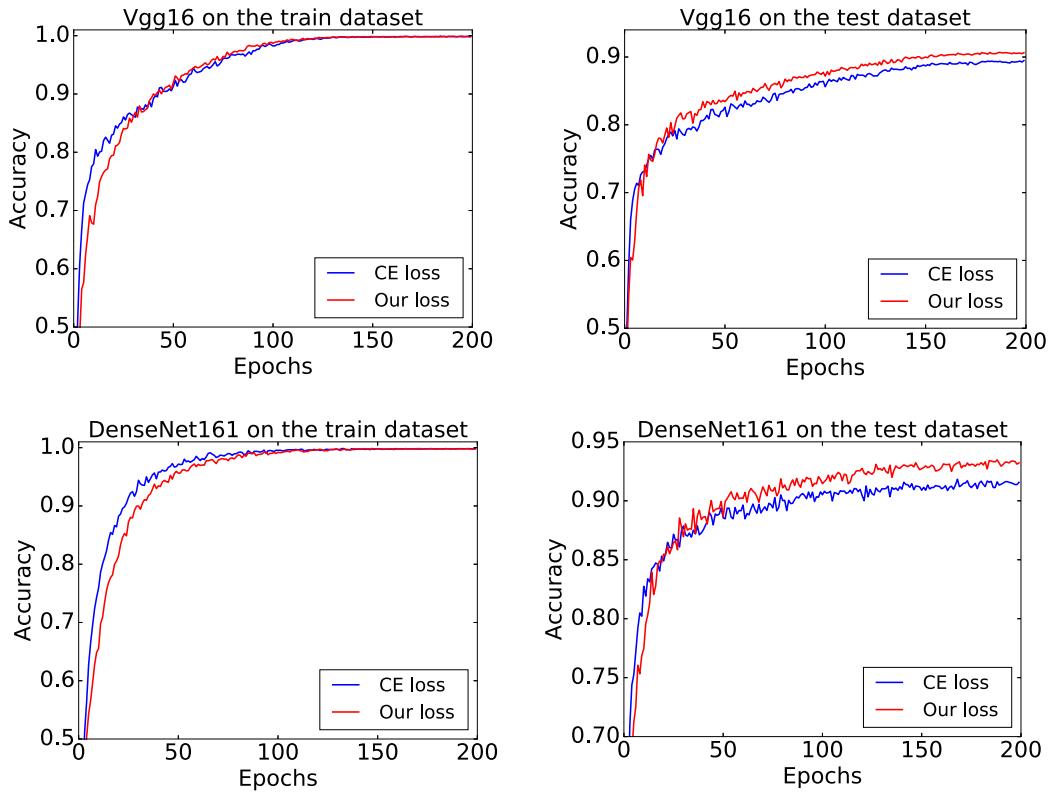


Fig. 3. Curves of the accuracies obtained by the VGG16 and DenseNet161 networks trained by minimizing the CE loss and our loss on the Stanford Cars-196 dataset. Top subfigures: the VGG network with 16 layers. Bottom subfigures: the DenseNet network with 161 layers.

probability that the data point is assigned to a class other than its ground-truth class, as shown in Figure 1 (b).

III. Fine-Grained Vehicle Classification Experiments

Considering our focus on fine-grained vehicle classification, we choose the Stanford Cars-196 dataset.

A. Stanford Cars-196 Dataset

The Stanford Cars-196 dataset contains 16,185 images of 196 classes of cars. The dataset is split into 8,144 training images and 8,041 testing images, in which each class is split roughly 50-50. We employ the information of the bounding boxes provided by the dataset and process images into a specific area of the car, which are used to further train the neural network and evaluate our loss. Several examples of original images and cropped boxes are shown in Figure 2.

Since the Stanford Cars-196 dataset is a small-sample, we use the publicly available deep convolutional neural networks VGG16 and DenseNet161 pretrained on the ImageNet dataset directly and then fine-tune the networks on the Stanford Cars-196 dataset. Specifically, we compare (1) VGG16-CE, (2) VGG16-Ours, (3) DenseNet161-CE and (4) DenseNet161-Ours, where VGG16 and DenseNet161 are two popular convolutional networks and CE and Ours represent CE loss and Dual Cross-Entropy Loss.

To fine-tune VGG16 and DenseNet161, we use the SGD optimization algorithm with a momentum of 0.9 and set the batch size to 32, the number of epochs to 200, and the weight decay parameter to 5×10^{-4}. The entire network does not use the same learning rate. In the convolutional layer, we adopt a fixed learning rate of 0.0001, and in the fully convolutional layer, we decrease the learning rate from 0.01 to 0.0 via the cosine annealing method. We process the images according to the size of each image bounding box, resize the images to a uniform size of 224×224, and run VGG16 and DenseNet161 by minimizing the CE loss and the proposed loss.

For the setting of α and β, we first randomly extract one fifth of the training dataset of the Stanford Cars-196 dataset as the validation dataset and then select the α and β with the best performance on the validation dataset. Finally, we set $\alpha = 1.0$ and $\beta = 4.5$ to train the original training dataset of the Stanford Cars-196 dataset, and the experimental results are shown in Figure 3.

The upper two subfigures of Figure 3 show that the accuracy of VGG16-CE on the training dataset increases much faster than that of VGG16-Ours in the first 50 epochs, whereas after 50 epochs, the rate of increase in the accuracy of VGG16-CE is very similar to that of VGG16-Ours. In terms of the increasing accuracy on the test dataset, VGG16-Ours surpasses VGG16-CE within the first several epochs and remains superior in all following epochs. A similar phenomenon is observed for DenseNet161; see the bottom two subfigures of Figure 3. Therefore, our proposed loss improves upon the performance of CE loss.

According to Figure 3, the proposed Dual Cross-Entropy Loss alleviates the vanishing of the gradient. The reason is that when the accuracy on the training dataset is close to 1.0, the original CE loss will be close to 0, the gradients of all parameters will be quite small, which easily results in vanishing gradients, and the accuracy does not increase in optimization. However, for the proposed Dual Cross-Entropy Loss, the vanishing of the gradient are alleviated since the accuracy still increases at the end of optimization.

B. Effect of Reducing the Number of Epochs on Performance

To further demonstrate the superiority of our loss function, we reduce the number of epochs and compare VGG16-CE, VGG16-Ours, DenseNet161-CE and DenseNet161-Ours. Specifically, we reduce the number of epochs from 200 to 150, 100 and 50 while keeping the other settings unchanged. The classification accuracies are listed in Table I.

Table I shows that on the Stanford Cars-196 dataset, when the number of epochs is 200, the classification accuracy of VGG16-Ours is 0.9017, and the classification accuracy of VGG16-CE is 0.8952. On the DenseNet161 network, the classification accuracy of our loss is 0.9351, and the classification accuracy of CE loss is 0.9217.

As shown in Table I, as the number of epochs decreases, the performances of all four methods worsen. However, our loss always has better performance than CE loss on the Stanford Cars-196 dataset, regardless of whether VGG16 or DenseNet16 is used. Moreover, the performance of VGG16-Ours (100 epochs) is roughly the same as the performance of VGG16-CE (200 epochs), and the performance of DenseNet161-Ours (50 epochs) is roughly the same as the performance of DenseNet-CE (200 epochs). In summary, our loss improves the accuracy of CE loss and accelerates the optimization of the network for fine-grained vehicle classification.

TABLE I
COMPARISON OF THE CLASSIFICATION ACCURACIES OBTAINED BY THE NETWORKS UNDER CROSS-ENTROPY LOSS (CE LOSS) AND THE PROPOSED DUAL CROSS-ENTROPY LOSS ON THE STANFORD CARS-196 DATASET. VGG16-CE: VGG16 NETWORK TRAINED BY MINIMIZING CE LOSS. VGG16-OURS: VGG16 NETWORK TRAINED BY MINIMIZING THE PROPOSED DUAL CROSS-ENTROPY LOSS. DENSE161-CE: DENSENET161 NETWORK TRAINED BY MINIMIZING CE LOSS. DENSE161-OURS: DENSENET161 NETWORK TRAINED BY MINIMIZING THE PROPOSED DUAL CROSS-ENTROPY LOSS

Methods	200Epochs	150Epochs	100Epochs	50Epochs
VGG16-CE	0.8952	0.8818	0.8819	0.8599
VGG16-Ours	**0.9017**	**0.9006**	**0.8967**	**0.8792**
Dense161-CE	0.9217	0.9226	0.9178	0.9152
Dense161-Ours	**0.9351**	**0.9340**	**0.9337**	**0.9304**

IV. OTHER IMAGE CLASSIFICATION EXPERIMENTS

Since fine-grained data show high interclass similarity, to further evaluate the performance of our loss function on other datasets in which some classes show high interclass similarity, in this section, we choose the LabelMe dataset, the UIUC-Sports dataset and the CIFAR-10 dataset. The first two datasets are small-sample datasets, and the last one is a large-sample dataset.

A. The LabelMe Dataset

The LabelMe dataset contains 8 classes of natural scene images: coast, mountain, forest, open country, street, inside city, tall buildings and highways. We randomly select 210 images for each class, of which 100 images, 100 images and 10 images are used for the training dataset, test dataset and validation set, respectively. The total number of images is 1680.

Since the feature quality is vital for image classification performance [43], [44], we resize every image to 256×256 and extract the image features directly using the publicly available deep neural network VGG16 pretrained on the ImageNet dataset. We reserve only the features of the last convolutional layer and simply flatten them, and the feature dimension for each image is 32768. In addition, by selecting one set of values that makes the network achieve the best performance among different values on the validation dataset, α and β are set as 1.0 and 4.5, respectively, on the LabelMe dataset.

On the LabelMe dataset, we select two network structures, a fully connected network of two layers (FC) and a fully connected network of one layer (Softmax). Specifically, we compare (1) FC-CE, (2) FC-Ours, (3) Softmax-CE, and (4) Softmax-Ours, where CE and Ours represent CE loss and the proposed Dual Cross-Entropy Loss, respectively.

In FC, the rectified linear unit function (Relu) and softmax function are the activation functions of the first layer and second layer, respectively. The dropout technique is not applied in FC

TABLE II
COMPARISON OF THE CLASSIFICATION PERFORMANCES ON THE *LabelMe Dataset* (LABELME). THE METHODS INCLUDE *Fully Connected Network Under CE Loss* (FC-CE), *Fully Connected Network Under Dual Cross-Entropy Loss* (FC-OURS), *Softmax Classifier Under CE Loss* (SOFTMAX-CE), AND *Softmax Classifier Under Our Loss* (SOFTMAX-OURS). EACH METHOD IS RUN 60 TIMES. THE MEAN VALUES AND STANDARD DEVIATIONS OF THE CLASSIFICATION ACCURACIES AND THE p-VALUES OF THE PAIRED STUDENT'S T-TEST BETWEEN FC-OURS AND FC-CE AND BETWEEN SOFTMAX-OURS AND SOFTMAX-CE ARE PRESENTED

Methods	Mean	Std.	p-Value
FC-CE	0.86	0.02	0.0021
FC-Ours	**0.87**	**0.019**	
Softmax-CE	0.83	0.079	0.062
Softmax-Ours	**0.86**	**0.031**	

since the dropout technique does not improve the performance of FC. The optimization algorithm [45]–[47] is set as RMSprop with an initial learning rate of 0.001. The batch size and number of epochs are set to 32 and 1500, respectively, and the weight decay is set to 0.01. We monitor the performance of the network on the validation dataset of the LabelMe dataset, and the network weights corresponding to the best performance are used to predict the unseen test data. For Softmax, we remove only the first layer of FC and keep the other settings unchanged.

We run FC-CE, FC-Ours, Softmax-CE and Softmax-Ours on the LabelMe dataset 60 times each. The means and standard deviations of the accuracies are shown in Table II. The means and standard deviations are 0.86 and 0.020 for FC-CE, 0.87 and 0.019 for FC-Ours, 0.83 and 0.079 for Softmax-CE, and 0.86 and 0.031 for Softmax-Ours.

Moreover, to demonstrate that our loss outperforming CE loss on the LabelMe dataset is not due to chance, we conduct paired Student's t-tests between FC-CE and FC-Ours and between Softmax-CE and Softmax-Ours. The p-value of the paired Student's t-test for FC-CE and FC-Ours is 0.0021, which is less than 0.005 (significance level). Therefore, we reject the null hypothesis that FC-CE and FC-Ours have the same mean accuracy. The p-value of the paired Student's t-test for Softmax-CE and Softmax-Ours, however, is 0.062, which is greater than 0.005. Therefore, the null hypothesis that Softmax-CE and Softmax-Ours have the same mean accuracy cannot be rejected. These results show that on the FC network, the CE loss and proposed loss are significantly different.

To further investigate the effect of the proposed loss on the robustness and stability of the network, we present box plots of the accuracies obtained by FC-CE, FC-Ours, Softmax-CE and Softmax-Ours in Figure 4. In Figure 4, the box plot of FC-Ours is more compact than that of FC-CE, and the box plot of Softmax-Ours is more compact than that of Softmax-CE. Meanwhile, FC-Ours and Softmax-Ours have no lower outliers, which indicates that the model using the proposed loss has better stability.

B. The UIUC-Sports Dataset

The UIUC-Sports dataset contains 1578 sports scene images of 8 classes: bocce (137), polo (182), rowing (250), sailing (190), snowboarding (190), rock climbing (194), croquet (236) and badminton (200). Ten randomly sampled images are taken from each class for the validation dataset, and the remaining samples are split equally into a training dataset and a test dataset, resulting in a training dataset of 749 images, a test dataset of 749 images and a validation dataset of 80 images.

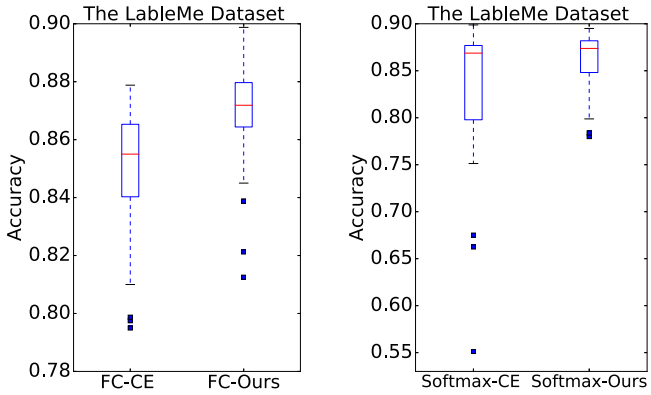


Fig. 4. Box plot comparison of the accuracies obtained by FC-CE, FC-Ours, Softmax-CE, and Softmax-Ours on the LabelMe dataset. The central mark is the median, and the edges of the boxes are the 25th and 75th percentiles. The outliers are marked individually. Each method is run 60 times to produce the box plots.

TABLE III
COMPARISON OF THE CLASSIFICATION PERFORMANCES ON THE *UIUC-Sports Dataset* (UIUC). THE METHODS INCLUDE *Fully Connected Network Under CE loss* (FC-CE), *Fully Connected Network Under Dual Cross-Entropy Loss* (FC-OURS), *Softmax Classifier Under CE Loss* (SOFTMAX-CE), AND *Softmax Classifier Under Our Loss* (SOFTMAX-OURS). EACH METHOD IS RUN 60 TIMES. THE MEANS AND STANDARD DEVIATIONS OF THE CLASSIFICATION ACCURACIES AND THE p-VALUES OF PAIRED STUDENT'S T-TESTS BETWEEN FC-OURS AND FC-CE AND BETWEEN SOFTMAX-OURS AND SOFTMAX-CE ARE REPORTED

Methods	Mean	Std.	p-value
FC-CE	0.89	0.01	$3.22e^{-5}$
FC-Ours	**0.90**	**0.007**	
Softmax-CE	0.75	0.15	0.0011
Softmax-Ours	**0.83**	**0.083**	

We compare FC-CE, FC-Ours, Softmax-CE and Softmax-Ours on the UIUC-Sports dataset. The parameter settings of these methods are the same as those used for the LabelMe dataset. In addition, by selecting one set of values that makes the network achieve the best performance among different values on the validation dataset, α and β are set as 1.0 and 7.0, respectively, on the UIUC-Sports dataset. The means and standard deviations of the accuracies are reported in Table III.

The mean and standard deviation of FC-CE are 0.89 and 0.010. The mean of FC-Ours is 0.90, which is higher than that of FC-CE, and the standard deviation of FC-Ours is 0.007, which is smaller than that of FC-CE. Similar results are observed for Softmax-CE and Softmax-Ours. Therefore, for both FC and Softmax, the proposed loss achieves a higher mean accuracy and a lower standard deviation.

We also conduct paired Student's t-tests for FC-CE and FC-Ours and for Softmax-CE and Softmax-Ours on the UIUC-

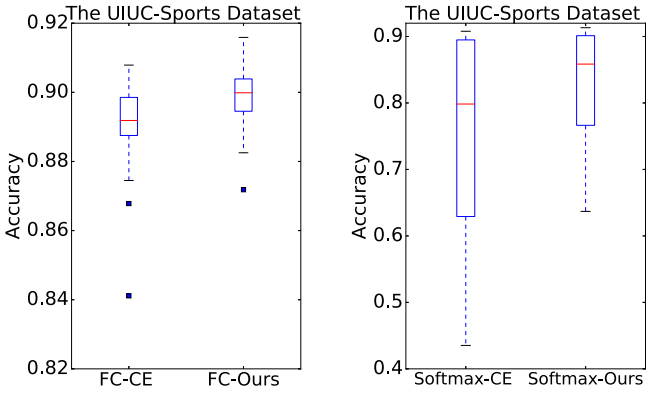


Fig. 5. Box plots of the accuracies obtained by FC-CE, FC-Ours, Softmax-CE, and Softmax-Ours on the UIUC-Sports dataset. The central mark is the median, and the edges of the boxes are the 25th and 75th percentiles. Outliers are marked individually. Each method is run 60 times to produce the box plots.

Sports dataset, and the corresponding p-values are listed in Table III. The p-value of FC-CE and FC-Ours is less than 0.05 (significance level); therefore, the null hypothesis that FC-CE and FC-Ours have the same mean is rejected. Similarly, the null hypothesis that Softmax-CE and Softmax-Ours have the same mean is also rejected. The results indicate that for both FC and Softmax, CE loss and the proposed loss are significantly different on the UIUC-Sports dataset.

Figure 5 shows box plots of the accuracies obtained by FC-CE, FC-Ours, Softmax-CE and Softmax-Ours. From Figure 5, we can see that on the UIUC-Sports dataset, the box plot of FC-Ours is more compact than that of FC-CE, and the box plot of Softmax-Ours is more compact than that of Softmax-CE. Meanwhile, the maximum, median and minimum accuracies of FC-Ours and Softmax-Ours are higher than those of FC-CE and Softmax-CE, respectively. Therefore, compared with the models using CE loss, the models using the proposed loss have better stability.

C. The CIFAR-10 Dataset

The CIFAR-10 dataset contains images of 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. The image size is 32×32. The total number of images is 60,000: the training dataset contains 50,000 images, and the test dataset contains 10,000 images.

We compare two methods, CNNs-CE and CNNs-Ours, on the CIFAR-10 dataset. We construct the CNNs according to the literature [48], use the convolutional network structure based on the ConvPool-CNN-C architecture [48], [49], replace the first dropout layer with a layer that adds Gaussian noise, extend the last hidden layer from 10 units to 192 units, and use 3×3 max-pooling. Meanwhile, we adopt the Adam optimization algorithm and set the batch size and number of epochs to 100 and 200. In the Adam optimization algorithm, the initial learning rate is set to 0.0003, and the momentum is set to 0.9 for the first 100 epochs and then to 0.5. The learning rate is linearly decayed to zero over the second 100 epochs. Regarding the parameters α and β in our loss function, we first randomly extract one fifth of the training dataset of the CIFAR-10 dataset as the validation dataset and then select the α and β with the best performance on the validation dataset. Finally, we set $\alpha = 0.1$ and $\beta = 1.0$ to train the original training dataset. We run CNNs-CE and CNNs-Ours on two sets of CIFAR-10 data, the CIFAR-10 (5:1) dataset and the CIFAR-10 (1:5) dataset. In the first dataset, the ratio of the training data to test data is $5:1$, and in the second dataset, the ratio of the training data to test data is $1:5$. The classification accuracies are reported in Table IV.

TABLE IV
COMPARISON OF THE CLASSIFICATION ACCURACIES OBTAINED BY THE NETWORKS UNDER CROSS-ENTROPY LOSS (CE LOSS) AND THE PROPOSED DUAL CROSS-ENTROPY LOSS ON THE CIFAR10 DATASET. THE RATIOS BETWEEN THE TRAINING AND TEST DATASETS ARE 5:1 AND 1:5

Dataset	CNNs-CE	CNNs-Ours
CIFAR10 (5:1)	0.9080	0.9120
CIFAR10 (1:5)	0.7842	0.7844

As shown in Table IV, on the CIFAR-10 (5:1) dataset, the accuracy of CNNs-Ours is 91.20%, which is 0.4% higher than that of CNNs-CE. On the CIFAR-10 (1:5) dataset, the accuracy of CNNs-Ours is 78.44%, which is approximately 0.2% higher than that of CNNs-CE. The experimental results show that the proposed loss achieves competitive performance on the CIFAR-10 dataset.

D. Effect of Reducing the Number of Epochs on Performance

In the previous experiments, all the methods used a fixed number of epochs. To explore the effect of the number of epochs on the performance of the network for the LabelMe and UIUC-Sports datasets, we consider 1500, 800, 200, 100 and 50 epochs for FC-CE and FC-Ours. The other settings remain unchanged. Each method is still run 60 times, and the means and standard deviations of the accuracies are listed in Figure 6. A larger mean and a smaller standard deviation indicate better performance.

As shown in Figure 6, on the LabelMe and UIUC-Sports datasets, when the number of epochs is 1500, FC-Ours has the largest mean and smallest standard deviation. Meanwhile, as the number of epochs decreases, the mean accuracies of both FC-CE and FC-Ours decrease monotonically. However, compared with FC-CE, the mean of FC-Ours does not show an obvious decrease. In addition, on the LabelMe dataset, FC-Ours with 50 epochs has similar mean and standard deviation as FC-CE with 800 epochs. On the UIUC-Sports dataset, FC-Ours with 100 epochs has similar mean and standard deviation as FC-CE with 800 epochs. Therefore, our loss function accelerates the optimization process and reduces the training time.

E. Effect of Varying β on Performance

In the proposed loss, the parameter β is responsible for adjusting the ratio of CE entropy loss and the regularized term. To further show the effect of varying the β of the proposed loss on the performance of the model, we try different values of β and train the FC network by minimizing the proposed loss on the UIUC-Sports dataset. The results for the test data are shown in Figure 7 in box plot format.

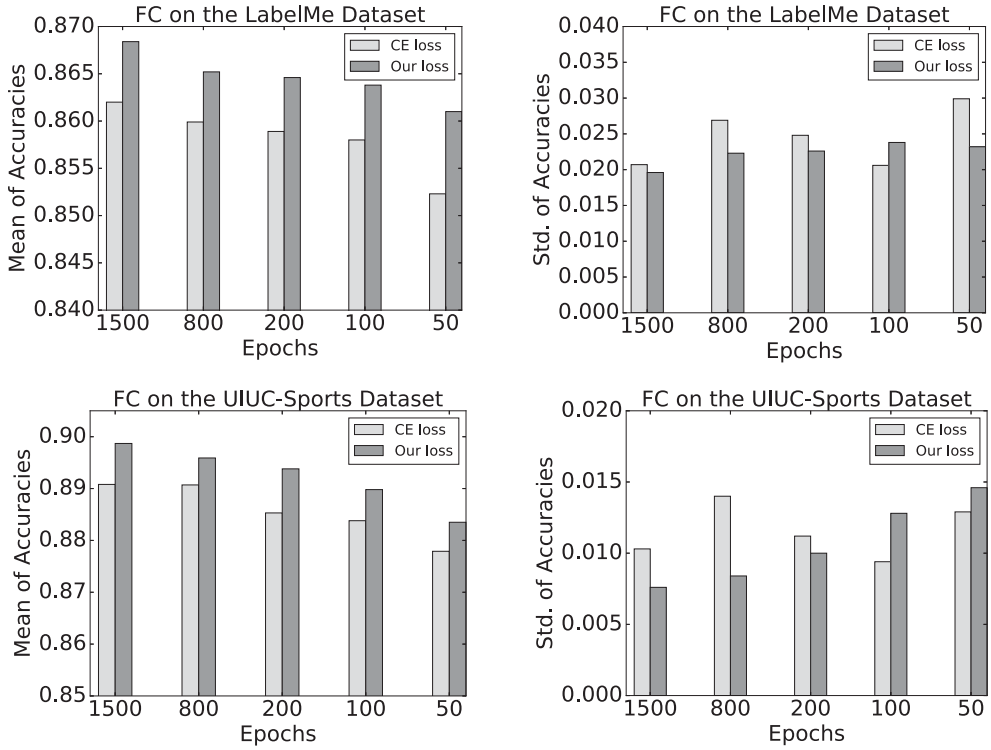


Fig. 6. Comparison of the classification accuracies obtained by the FC network on the UIUC-Sports dataset (UIUC) and the LabelMe dataset (LabelMe) as the number of epochs changes. Each setting for FC is run 60 times, and the mean values (Mean, left subfigures) and standard deviations (Std., right subfigures) of the accuracies are shown.

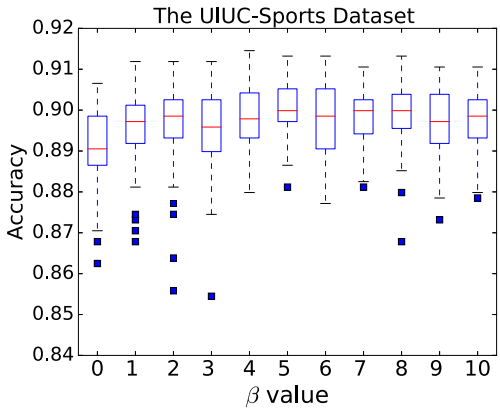


Fig. 7. Box plots of the accuracies obtained by FC-CE on the UIUC-Sport dataset. FC-CE is optimized by varying β and minimizing the proposed loss. FC is run 60 times under each β. The central mark is the median, and the edges of the boxes are the 25th and 75th percentiles.

From Figure 7, as β increases, the overall trend is that the performance first improves gradually and then declines gradually. After declining to the worst value, the performance then improves gradually. The reason for this phenomenon is that when β is small, the regularized term has a positive function for CE loss, but when β is quite large, the regularized term becomes the main part of the loss function. In short, the value of β in the proposed loss can affect the performance of the model. Therefore, in the experiment above, we use the validation dataset to select the parameter β, that is, select the value of β that makes the proposed loss have the best performance on the validation dataset.

F. Confusion Matrix

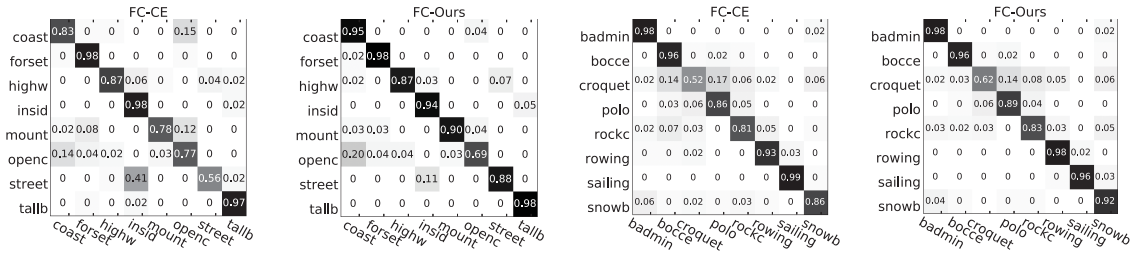
In this section, we present the confusion matrices of FC-Ours and FC-CE for the LabelMe and UIUC-Sports test data, as shown in Figure 8 The results are from one of the 60 runs of FC-Ours and FC-CE. We select a run in which the accuracies of FC-Ours and FC-CE are close to their means. The first two subfigures are for FC-CE and FC-Ours on the LabelMe dataset, and the next two subfigures are for FC-CE and FC-Ours on the UIUC-Sports dataset.

From the first two subfigures in Figure 8, we can see that on the classes of "coast", "mountain", and "street", the accuracies of FC-CE are 0.83, 0.78 and 0.56, respectively, and the accuracies of FC-Ours are 0.95, 0.90 and 0.88, respectively. In particular, FC-CE has considerable confusion on the classes of "mountain" and "street", and FC-Ours produces greatly improved results. Similarly, from the next two subfigures in Figure 8, we can see that on the classes of "croquet", "rock climbing" and "snowboarding", the accuracies of FC-CE are 0.52, 0.81 and 0.86, respectively, and the accuracies of FC-Ours are 0.62, 0.83 and 0.92, respectively. FC-CE has substantial confusion on the classes of "croquet" and "snowboarding", and the results of FC-Ours are greatly improved.

In summary, the confusion matrices in Figure 8 show that our loss achieves improved accuracy for those classes that are easily confused.

G. Discussion

From the experimental results on the LabelMe, UIUC-Sports and CIFAR-10 datasets, compared with CE loss, the proposed loss has superior performance on the first two datasets and



(a) The LabelMe Dataset (b) The LabelMe Dataset (c) The UIUC-Sports Dataset (d) The UIUC-Sports Dataset

Fig. 8. Confusion matrices of the FC networks for the UIUC-Sports dataset and the LabelMe dataset. (a) and (b) FC-CE and FC-Ours on the LabelMe dataset, respectively. (c) and (d) FC-CE and FC-Ours on the UIUC dataset, respectively.

competitive performance on the CIFAR-10 dataset. First, our loss works better on small-sample datasets and performs well on large-sample datasets. Second, on the LabelMe and UIUC-Sports datasets, a comparison of the standard deviations and box plots indicates that our loss can ensure that the network or model has more stable performance compared to CE loss. Third, the proposed loss can help improve the classification accuracy of the model or network for those classes that are easily confused.

The proposed loss possesses these advantages for the following reasons: In the proposed loss, the CE loss is responsible for increasing the probability that a data point is assigned to its ground-truth class, and the regularized term is responsible for decreasing the probability that the data point is assigned to a non-ground-truth class. When the CE loss is close to zero, since the deviation of the regularized term will not be zero, that is, the total loss can continue to decrease, the proposed loss can alleviate the vanishing of the gradient. In addition, in the process of optimization, increases in the probability that the data point is assigned to any non-ground-truth class are constrained, and thus, the probability that the data point is assigned to its ground-truth class increases. Consequently, the proposed loss accelerates the optimization of the model or network.

V. Conclusion

In this paper, we propose a *Dual Cross-Entropy Loss* built on CE loss. The proposed loss can be viewed as a linear combination of CE loss and a regularized loss. The regularization term places a constraint on the probability that the data point is assigned to a class other than its ground-truth class, which can alleviate the vanishing of the gradient when the CE loss is close to zero. The results of two sets of experiments show that compared with CE loss, the proposed loss (1) can accelerate the optimization of the neural network; (2) has better performance on a small-sample fine-grained vehicle classification dataset (Stanford Cars-196), a small-sample scene classification dataset (LabelMe), and a small-sample event classification dataset (UIUC-Sports); and (3) has competitive performance on a large-sample dataset (CIFAR-10).

References

[1] N. Kato *et al.*, "The deep learning vision for heterogeneous network traffic control: Proposal, challenges, and future perspective," *IEEE Wireless Commun.*, vol. 24, no. 3, pp. 146–153, Jun. 2017.

[2] Z. Fadlullah *et al.*, "State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems," *IEEE Commun. Surv. Tut.*, vol. 19, no. 4, pp. 2432–2455, Oct.–Dec. 2017.

[3] Q. Cui, N. Wang, and M. Haenggi, "Vehicle distributions in large and small cities: Spatial models and applications," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10176–10189, Nov. 2018.

[4] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3973–3981.

[5] S. Meta and M. G. Cinsdikici, "Vehicle-classification algorithm based on component analysis for single-loop inductive detector," *IEEE Trans. Veh. Technol.*, vol. 59, no. 6, pp. 2795–2805, Jul. 2010.

[6] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2167–2175.

[7] Y.-K. Ki and D.-K. Baik, "Vehicle-classification algorithm for single-loop detectors using neural networks," *IEEE Trans. Veh. Technol.*, vol. 55, no. 6, pp. 1704–1711, Nov. 2006.

[8] M. Vargas, J. M. Milla, S. L. Toral, and F. Barrero, "An enhanced background estimation algorithm for vehicle detection in urban traffic scenes," *IEEE Trans. Veh. Technol.*, vol. 59, no. 8, pp. 3694–3709, Oct. 2010.

[9] Y. Zhou, H. Nejati, T.-T. Do, N.-M. Cheung, and L. Cheah, "Image-based vehicle analysis using deep neural network: A systematic study," in *Proc. IEEE Int. Conf. Digit. Signal Process.*, 2016, pp. 276–280.

[10] V. D. Nguyen, T. T. Nguyen, D. D. Nguyen, S. J. Lee, and J. W. Jeon, "A fast evolutionary algorithm for real-time vehicle detection," *IEEE Trans. Veh. Technol.*, vol. 62, no. 6, pp. 2453–2468, Jul. 2013.

[11] S. Mahendran and R. Vidal, "Car segmentation and pose estimation using 3D object models," 2015, arXiv:1512.06790.

[12] Z. Wang *et al.*, "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 379–387.

[13] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "Learning deep neural networks for vehicle re-ID with visual-spatio-temporal path proposals," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1918–1927.

[14] M. Biglari, A. Soleimani, and H. Hassanpour, "A cascaded part-based system for fine-grained vehicle classification," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 273–283, Jan. 2018.

[15] J. Liang, X. Chen, M.-L. He, L. Chen, T. Cai, and N. Zhu, "Car detection and classification using cascade model," *IET Intell. Transport Syst.*, vol. 12, no. 10, pp. 1201–1209, 2018.

[16] K. Valev, A. Schumann, L. Sommer, and J. Beyerer, "A systematic evaluation of recent deep learning architectures for fine-grained vehicle classification," *Proc. SPIE*, vol. 10649, 2018, Art. no. 1064902.

[17] Z. Ma *et al.*, "The role of data analysis in the development of intelligent energy networks," *IEEE Netw.*, vol. 31, no. 5, pp. 88–95, 2017.

[18] Y. Jeon and J. Kim, "Active convolution: Learning the shape of convolution for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1846–1854.

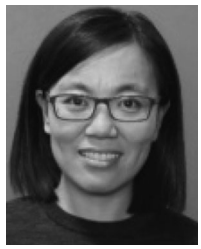
[19] T. Durand, T. Mordan, N. Thome, and M. Cord, "WILDCAT: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, vol. 2, pp. 5957–5966.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

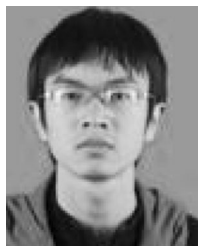
[21] N. Kato, M. Suzuki, S. Omachi, H. Aso, and Y. Nemoto, "A handwritten character recognition system using directional element feature and asymmetric Mahalanobis distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 3, pp. 258–262, Mar. 1999.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, arXiv:1409.1556.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[24] L. Wang, T. Liu, G. Wang, K. L. Chan, and Q. Yang, "Video tracking using learned hierarchical features," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1424–1435, Apr. 2015.
[25] P. Xu, K. Li, Z. Ma, Y.-Z. Song, L. Wang, and J. Guo, "Cross-modal subspace learning for sketch-based image retrieval: A comparative study," in *Proc. IEEE Int. Conf. Netw. Infrastructure Digit. Content*, 2016, pp. 500–504.
[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[27] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.
[28] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teulière, and T. Chateau, "Deep MANTA: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2040–2049.
[29] J. Sochor, A. Herout, and J. Havel, "Boxcars: 3D boxes as CNN input for improved fine-grained vehicle recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3006–3015.
[30] H. Zhang, K. Wang, Y. Tian, C. Gou, and F.-Y. Wang, ''MFR-CNN: Incorporating multi-scale features and global information for traffic object detection," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8019–8030, Sep. 2018.
[31] D. Liu and Y. Wang, "Monza: Image classification of vehicle make and model using convolutional neural networks and transfer learning," 2017, [Online]. Available: http://cs231n.stanford.edu/reports/2015/pdfs/lediurfinal.pdf.
[32] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1449–1457.
[33] Q. Hu, H. Wang, T. Li, and C. Shen, "Deep CNNS with spatially weighted pooling for fine-grained car recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3147–3156, Nov. 2017.
[34] Y. Tian, W. Zhang, Q. Zhang, G. Lu, and X. Wu, "Selective multi-convolutional region feature extraction based iterative discrimination CNN for fine-grained vehicle model recognition," in *Proc. 24th Int. Conf. Pattern Recognit.*, 2018, pp. 3279–3284.
[35] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1245–1256, Jun. 2017.
[36] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 507–516.
[37] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.
[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.
[39] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.
[40] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[41] L.-J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[42] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. TR-2009, 2009.
[43] Z. Ma, J.-H. Xue, A. Leijon, Z.-H. Tan, Z. Yang, and J. Guo, "Decorrelation of neutral vector variables: Theory and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 129–143, Jan. 2018.
[44] Z. Ma, Y. Lai, W. B. Kleijn, Y.-Z. Song, L. Wang, and J. Guo, "Variational Bayesian learning for Dirichlet process mixture of inverted Dirichlet distributions in non-Gaussian image feature modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 449–463, Feb. 2019.
[45] Z. Ma, A. E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, and J. Guo, "Variational Bayesian matrix factorization for bounded support data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 876–889, Apr. 2015.
[46] Z. Ma and A. Leijon, "Bayesian estimation of beta mixture models with variational inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2160–2173, Nov. 2011.
[47] Z. Ma, P. K. Rana, J. Taghia, M. Flierl, and A. Leijon, "Bayesian estimation of Dirichlet mixture model with variational inference," *Pattern Recognit.*, vol. 47, no. 9, pp. 3143–3157, 2014.
[48] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 901–909.
[49] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *ICLR (workshop track)*, 2015. [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a



Xiaoxu Li received the Ph.D. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2012. She is currently an Associate Professor with the School of Computer and Communication, Lanzhou University of Technology, Lanzhou, China. She is a member of the China Computer Federation. Her current research focuses on the machine learning fundamentals with a focus on applications in image and video understanding.



Liyun Yu received the B.E. degree in computer science and technology, in 2016, from the Lanzhou University of Technology, Lanzhou, China, where he is currently working toward the graduate degree. His research interests include machine learning and small-sample image understanding.

Dongliang Chang received the B.E. degree in network engineering from Zhoukou Normal University, Zhoukou, China, in 2016. He is currently working toward the graduate degree with the Lanzhou University of Technology, Lanzhou, China. His research interests include machine learning and computer vision.

Zhanyu Ma received the Ph.D. degree in electrical engineering from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2011. Since 2014, he has been an Associate Professor with the Beijing University of Posts and Telecommunications, Beijing, China. Since 2015, he has also been an Adjunct Associate Professor with Aalborg University, Aalborg, Denmark. From 2012 to 2013, he was a Postdoctoral Research Fellow with the School of Electrical Engineering, KTH Royal Institute of Technology. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in multimedia signal processing, data mining, biomedical signal processing, and bioinformatics.

Jie Cao received the M.E. degree from Xi'an Jiaotong University, Xi'an, China, in 1994. She is currently a Professor and Vice President of the Lanzhou University of Technology, Lanzhou, China. Her research interests include machine learning, pattern recognition, speech and speaker recognition, information fusion, and computer vision.