LOW-RESOLUTION VISUAL RECOGNITION VIA DEEP FEATURE DISTILLATION

Mingjian Zhu[1], *Kai Han*[1,2], *Chao Zhang*[1*], *Jinlong Lin*[3], *Yunhe Wang*[2]

[1]Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
[2]Huawei Noah's Ark Lab
[3]School of Software & Microelectronics, Peking University
zhumingjian@pku.edu.cn,kai.han@huawei.com,
c.zhang@pku.edu.cn,linjl@ss.pku.edu.cn,yunhe.wang@huawei.com

ABSTRACT

Here we study the low-resolution visual recognition problem. Conventional methods are usually trained on images with large ROIs (regions of interest), while the regions and insider images are often small and blur in real-world applications. Therefore, deep neural networks learned on high-resolution images cannot be directly used for recognizing low-resolution objects. To overcome this challenging problem, we propose to use the teacher-student learning paradigm for distilling useful feature information from a pre-trained deep model on high-resolution visual data. In practice, a distillation loss is used to seek the perceptual consistency of low-resolution images and high-resolution images. By simultaneously optimizing the recognition loss and distillation loss, we formulate a novel low-resolution recognition approach. Experiments conducted on benchmarks demonstrate that the proposed method is capable to learn well-performed models for recognizing low-resolution objects, which is superior to the state-of-the-art methods.

Index Terms— Low-Resolution Recognition, Deep Convolutional Networks, Teacher-Student Paradigm;

1. INTRODUCTION

The recent success of deep neural networks has largely boosted the computer vision applications such as visual recognition [1, 2, 3, 4], person re-identification [5], face recognition [6, 7]. Usually, most of visual recognition approaches are learned and conducted on datasets with some certain assumptions, especially, the region of interest (ROI) corresponding to the desired object is often large enough (e.g., the size of ROI of each object in LFW dataset [8] is larger than 16×16). However, in real-world applications the regions of desired objects (e.g., faces of pedestrian may be low-resolution (LR) in surveillance video) may be relatively small and blur, which contain less information and have more noise. Therefore, directly applying models trained on high-resolution images cannot achieve acceptable performance on these low-resolution images. In addition, the cost of replacing existing low-definition cameras with high-definition devices is very expensive. Therefore, effective methods for learning new recognition models suitable for very low-resolution recognition (VLRR) problem [9] is urgently required.

To overcome the aforementioned problems, a number of approaches for recognizing low-resolution images have been proposed recently. In particular, Chu et al. [10] proposed a coupled mappings methods using cluster-based regularized simultaneous discriminant analysis to recognize low-resolution objects. Zhang et al. [11] developed a distance metric learning algorithm for recognizing a low-resolution human face by projecting high-resolution (HR) and low-resolution images into a unified space which utilizes coupled marginal discriminant mappings. Lu et al. [12] proposed the deep coupled ResNet model, which considered not only the discriminability of HR and LR features, but also the similarity between them. Wang et al. [9] took advantage of super resolution, domain adaptation, and robust regression, and formulated a dedicated deep learning method. Although these methods have made tremendous efforts for enhancing the performance of low-resolution image recognition, such methods did not utilize information from models learned on high-resolution images.

On the other side, a possible way to solve the VLRR problem is to utilize the visual super-resolution (SR) technique [13, 14, 15, 16], which receives a low-resolution image and then outputs a high-resolution image through series of transforms for approximating original high-resolution images, which probably can be easier recognized by original models. Dong et al. [13] proposed the SRCNN method, which directly learned an end-to-end mapping between the low-resolution and high-resolution images. Kim et al. [14] presented a highly accurate single-image super-resolution method called VDSR which learns residuals only and uses extremely high learning rates enabled by adjustable gradient clipping. Shi et al. [16] proposed a CNN architecture called ESPCN where the feature maps were extracted in the LR

[*] Corresponding author. This research is partially supported by National Natural Science Foundation (NSF) of China (grant no.61671027) and National Basic Research Program of China (973 Program) (grant no.2015CB352303).

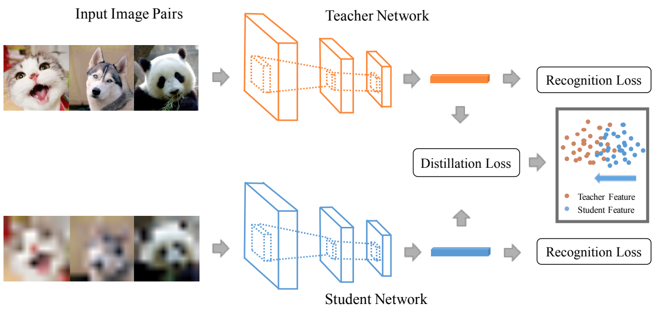

Fig. 1. The schematic of the proposed feature distillation method for recognizing LR images. Best viewed in color.

space. They proposed a sub-pixel convolution layer which is capable of super-resolving LR data into HR space with very little additional computational cost. Ledig et al. [15] presented SRGAN, a generative adversarial network (GAN) for image super-resolution (SR). It is the first framework capable of inferring photo-realistic natural images for $4\times$ upscaling factors. Although high visual quality images can be generated by these methods, the details in them are actually different to that in the original images, which cannot provide enough useful information for the subsequent recognition task.

Although a variety of methods have been proposed for addressing the low-resolution visual recognition problem, the information in high-resolution images has not been fully investigated. To this end, this paper proposes an effective approach for recognizing low-resolution visual objects by extracting useful information from the model trained on high-resilution images to seek the perceptual consistency of these images, as shown in Fig.1. In the first stage, we train the teacher model with HR images using corresponding labels. Then, a student model is established for recognizing low-resolution images by minimizing two objectives, i.e., the Euclidean distance between features of LR model and HR model and the cross entropy loss. Since features of HR images are easier to be correctly classified than those of LR images, making their features similar can significantly increase the recognition accuracy of LR images. Experimental results on benchmark datasets and deep models demonstrate the superiority of the proposed algorithm.

2. METHODS

2.1. Peliminary

In the setting of very low-resolution recognition problem, we have LR images $\{I_i^{lr}\}$, HR images $\{I_i^{hr}\}$, and labels $\{\boldsymbol{y}_i\}$ in the training stage, and make prediction only based on LR test images during testing. In practice, the CNN model $\mathcal{N}_{HR}$ trained on HR images performs much better than models $\mathcal{N}_{LR}$ trained on LR images, that is to say, the HR image features extracted from HR model are more discriminative. Thus, the motivation of this paper is to use the discriminative features extracted from HR model to help the training of LR model $\mathcal{N}_{LR}$.

Recently teacher-student learning paradigm is very popular, i.e., directly learning a student network with the indication of a teacher network which consists of more valuable information [17, 18, 19, 20]. Many techniques have been developed using the teacher-student learning paradigm. [21] proposed representational distance learning, a stochastic gradient descent method that improved classification performance by minimizing the difference between the pairwise distances between representations of teacher and student models at selected layers using auxiliary error functions. [22] proposed to combine the knowledge of multiple teacher networks in the intermediate representations to train a thin deep student network. [23] proposed to minimize the Euclidean distance between features extracted from student and teacher networks.

Considering that the original models learned on high-resolution images cannot directly capture the distinctive features from low-resolution images, which can be definitely used for guiding the training procedure of the network for recognizing LR images. Therefore, we propose to use the teacher-student learning paradigm for enhancing the performance of $\mathcal{N}_{LR}$. Inspired by FITNETS [18], we propose that using a model trained with high-resolution images as teacher model to indicate the training of student model with low-resolution images in order to make the performance of the student model close to that of the teacher model. In the viewpoint of data augmentation, HR images can be considered as auxiliary information for model training. Using HR images in addition to LR images is a kind of data augmentation and helps to model performance. The framework of the proposed method is shown in Fig.1.

2.2. Deep Feature Distillation

One of the most severe problems of low-resolution recognition problem is the lack of labeled low-resolution images. To prepare a LR image, we first down-sample the original image, and then we up-sample it as a LR image to match the size of the model input. The original image is considered as a HR image and it loses information during the period of down-sampling.

In the training stage, we first use HR images to train the teacher model with softmax loss. The teacher model is used to guide the training of student mdoel for LR images. For model simplicity, we choose the same basic architecture for the teacher model and student model, such as AlexNet [1], VGG-16 [24], and ResNet-18 [2]. And we denote the teacher and student deep nested functions up to their last feature layers as $f_t(\cdot)$ and $f_s(\cdot)$ respectively. After getting pre-trained models, we fine-tune the student model following the indication of the teacher model with frozen weights.

We propose to make the output feature $f_s(I_i^{lr})$ of the student model close to the feature $f_t(I_i^{hr})$ of the teacher model, thus we use the Euclidean distance (i.e. square error) between them as the distillation loss:

$$L_{distillation} = \frac{1}{2}\|f_s(I_i^{lr}) - f_t(I_i^{hr})\|_2^2. \tag{1}$$

The loss is used for letting the two features stay in the same domain and get close to each other. Training two models separately makes the feature extracted by them stays in different domains. Adding the distillation loss here can efficiently help the extracted feature of student model turn to the domain of teacher model.

Softmax loss is also used to train the student model:

$$\boldsymbol{p}_i = \text{softmax}(f_s(I_i^{lr})W + \boldsymbol{b}), \tag{2}$$

$$L_{recognition} = -\,\boldsymbol{y}_i \left(\log(\boldsymbol{p}_i)\right)^T, \tag{3}$$

where softmax$(\cdot)$ is the softmax function, $W \in \mathbb{R}^{d\times C}$ and $\boldsymbol{b} \in \mathbb{R}^C$ are the weight matrix and bias vector of the last fully connected layer, d is the feature dimension and C is the number of categories. $\boldsymbol{p}_i$ is the predicted probability vector and $\boldsymbol{y}_i$ is the true label vector of image i.

Combining the recognition loss and distillation loss, we train the student model parameters by minimizing the following loss function:

$$L = L_{recognition} + \lambda L_{distillation}, \tag{4}$$

where λ is the hyper-parameter for balancing the recognition loss and distillation loss. Note that the teacher model is fixed in this step, and the loss will only influence the training of the student model and make no influence to teacher model. The model can be trained via standard back-propagation and stochastic gradient descent. Alg. 1 summarizes the proposed method for training the CNN model to accurately recognize LR images, namely DFD (deep feature distillation).

Algorithm 1 DFD for Low-Resolution Image Recognition

```
Input: Training image pairs $\{I_i^{lr}\}$ and $\{I_i^{hr}\}$, and their corresponding ground-truth labels $\{\boldsymbol{y}_i\}$.
  Module 1: Prepare the teacher model.
    Train the teacher model $\mathcal{N}_{HR}$ with all the HR images $\{I_i^{hr}\}$ and corresponding labels $\{\boldsymbol{y}_i\}$;
  Module 2: Learn the student model.
  repeat
    Rondomly select a batch of image pairs $\{I_i^{hr}\}$ and $\{I_i^{lr}\}$ and coresponding ground-truth labels $\boldsymbol{y}_i$;
    Input them to $\mathcal{N}_{LR}$ and $\mathcal{N}_{HR}$, respectively;
    Update paramteres in $\mathcal{N}_{LR}$ after back-propagation;
  until convergence
Output: The optimized student network $\mathcal{N}_{LR}$
```

3. EXPERIMENTS

In real-world setting, most of low-resolution recognition methods directly recognize subjects from LR images, without any HR images. Specifically, in the training stage we introduce HR images as auxiliary information for model training by assuming that each training image has both LR and HR versions available. In the testing stage, only LR images are fed into the model for prediction.

Table 1. Top-1 accuracies (%) of different methods on CIFAR-10 and SVHN datasets, respectively. Base network without DFD stands for the result of using only LR images to train.

Methods	CIFAR-10		SVHN	
	8×8	16×16	8×8	16×16
AlexNet	64.08	69.10	81.81	87.77
DFD (AlexNet)	**66.23**	**71.60**	**81.97**	**88.35**
VGG-16	78.41	88.02	90.40	95.52
DFD (VGG-16)	**79.97**	**89.35**	**91.11**	**95.90**
ResNet-18	79.86	89.12	90.45	95.55
DFD (ResNet-18)	**81.26**	**90.41**	**92.07**	**96.02**

3.1. Datasets

The **CIFAR-10** dataset [25] consists of 60000 32×32 color images in 10 classes, with 6000 images per class. Since the CIFAR-10 dataset does not provide low-resolution images, We first down-sample the original image by a factor s for LR images of $32/s \times 32/s$. Then we resize them back to 32×32 used as the LR images in the experiment.

The **SVHN** dataset [26] is proposed as a large and real-world benchmark. The dataset contains 73257 digits for training, and 26032 digits for testing, including 10 classes, 1 for each digit. We process the data in the procedure similar to the CIFAR-10 dataset.

3.2. Implementation Details

We implement our model using PyTorch package [27]. AlexNet [1], VGG-16 [24], ResNet-18 [2] are used as our backbone networks since these models are classical models for recognition task. SGD optimizer with momentum is used to update the weights. We train the model for 100 epochs with the batch size of 128. We use a momentum of $\mu = 0.9$ and weight decay of $5e-4$. Dropout and batch normalization are also used. We adopt random crop and flip for data augmentation in train set. The hyper-parameter λ is tuned in $\{0.1, 0.2, 0.5, 1, 2, 5, 10\}$. All the experiments are conducted on a NVIDIA Pascal TITAN X GPU.

3.3. Performance on Different Backbone Networks

DFD is general to any CNN architecture, i.e, AlexNet, VGG-16, and ResNet-18. Table 1 shows the performance of methods with different backbone networks under different resolution $(8\times8, 16\times16)$. From the results, DFD improves the performance with different base networks with a significant margin, which verifies the superiority of the proposed method.

3.4. Impact of Hyper-parameter

The hyper-parameter λ in our method is to balance recognition loss and distillation loss. We conduct experiments in the CIFAR-10 dataset and tune λ in $\{0.1, 0.2, 0.5, 1, 2, 5, 10\}$. The resolution of LR images here is 8×8. The performance of different methods are shown in Fig. 2. From the results, DFD always outperforms the corresponding backbone networks which implies that DFD performance is steady while the hyperparameter λ changes. And we can obtain the best performance around $\lambda = 2$, which is a good trade-off point for predicting labels and turning extracted features of student model to that of teacher model.

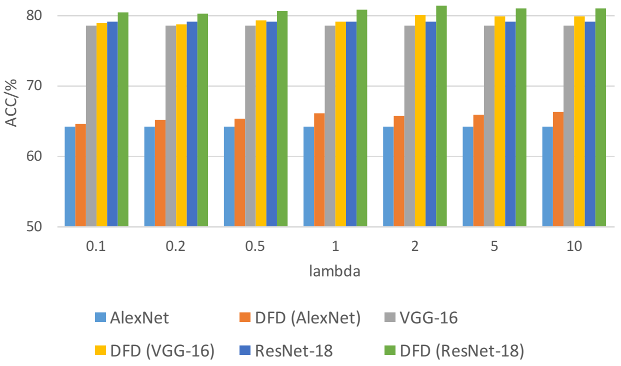

Fig. 2. Influence of hyper-parameter λ: the top-1 accuracies (%) with different base networks w.r.t. λ. Best viewed in color.

3.5. Comparison with Super-Resolution Based Methods

In order to demonstrate the superiority of the proposed method, we further compare our method with the typical super-resolution methods for recognizing low-resolution images in the CIFAR-10 dataset.

We firstly use these methods to transfer the original low-resolution images (with resolution of 8×8) to high-resolution images (*i.e.* 16×16), and then we directly use the ResNet-18 as base model to train and test the new images. Some examples of the generated SR images by exploiting super-resolution techniques are shown in Fig.3. Although these SR images contain much more texture and color details than LR images, we cannot directly evaluate the practical benefits of these generated information.

Thus we apply recognition techniques on the produced images for recognition task. Table 2 shows the performance comparison between our method and super-resolution based methods. HR-baseline stands for the results of using only HR model to train and predict for HR images. LR-baseline stands for the result of using only the LR images to train and predict for LR images. From the results in Table 2 and Fig.3, we find that super-resolution techniques are able to generate pleasing images for human eyes, but the generated images are useless when directly used for improving the performance of image recognition. The SR images produced by the super-resolution techniques are not real images and the generated details in images are irrelevant to classification task, which may explain why directly using super-resolution techniques cannot improve the recognition performance. In contrast, DFD can improve the performance by a large margin, indicating the feature distillation is efficient for VLRR problem.

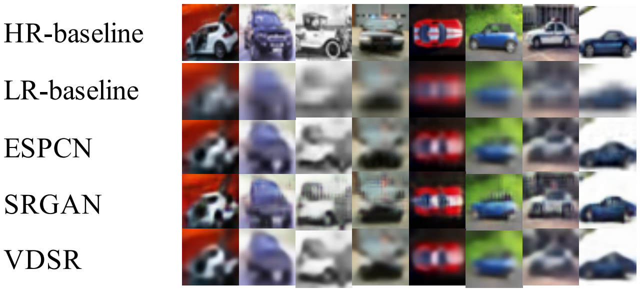

Fig. 3. Image examples of the super-resolution method conducted on CIFAR-10. Best viewed in color.

Table 2. The top-1 accuracies (%) of our method and super-resolution based methods on the CIFAR-10 dataset.

Model	Accuracy
HR-baseline	93.38
LR-baseline	79.86
VDSR [14] + ResNet-18	80.00
ESPCN [16] + ResNet-18	79.41
SRGAN [15] + ResNet-18	79.05
DFD (ResNet-18)	**81.26**

4. CONCLUSION

In this paper, we study the low-resolution problem and present a novel deep feature distillation framework for recognizing very low-resolution images by exploiting the teacher-student learning paradigm. In practice, the student model for LR images can inherit useful information from the teacher model learned on their original HR images. Thus, our deep feature distillation method can improve the performance of student model. The effectiveness of the proposed methods is validated on two benchmark datasets over the state-of-the-art methods. Moreover, our method can be applied to other computer vision tasks, such as denoising, deblurring, *etc*.

5. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] Yunhe Wang, Chang Xu, Chunjing Xu, Chao Xu, and Dacheng Tao, "Learning versatile filters for efficient convolutional neural networks," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., pp. 1615–1625. Curran Associates, Inc., 2018.

[4] Yunhe Wang, Chang Xu, Chao Xu, and Dacheng Tao, "Packing convolutional neural networks in the frequency domain," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[5] Kai Han, Jianyuan Guo, Chao Zhang, and Mingjian Zhu, "Attribute-aware attention model for fine-grained representation learning," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 2040–2048.

[6] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[7] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4873–4882.

[8] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[9] Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S Huang, "Studying very low resolution recognition using deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4792–4800.

[10] Yongjie Chu, Touqeer Ahmad, George Bebis, and Lindu Zhao, "Low-resolution face recognition with single sample per person," *Signal Processing*, vol. 141, pp. 144–157, 2017.

[11] Peng Zhang, Xianye Ben, Wei Jiang, Rui Yan, and Yiming Zhang, "Coupled marginal discriminant mappings for low-resolution face recognition," *Optik-International Journal for Light and Electron Optics*, vol. 126, no. 23, pp. 4352–4357, 2015.

[12] Ze Lu, Xudong Jiang, and Alex ChiChung Kot, "Deep coupled resnet for low-resolution face recognition," *IEEE Signal Processing Letters*, 2018.

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*. Springer, 2014, pp. 184–199.

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[15] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint arXiv:1609.04802*, 2016.

[16] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[18] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[19] Shiming Ge, Shengwei Zhao, Chenyu Li, and Jia Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 2051–2062, 2019.

[20] Yunhe Wang, Chang Xu, Chao Xu, and Dacheng Tao, "Adversarial learning of portable student networks," 2018.

[21] Patrick McClure and Nikolaus Kriegeskorte, "Representational distance learning for deep neural networks," *Frontiers in computational neuroscience*, vol. 10, pp. 131, 2016.

[22] Shan You, Chang Xu, Chao Xu, and Dacheng Tao, "Learning from multiple teacher networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 1285–1294.

[23] Jimmy Ba and Rich Caruana, "Do deep nets really need to be deep?," in *Advances in neural information processing systems*, 2014, pp. 2654–2662.

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

[25] Alex Krizhevsky and Geoffrey Hinton, "Learning multiple layers of features from tiny images," Tech. Rep., Citeseer, 2009.

[26] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS workshop on deep learning and unsupervised feature learning*, 2011, vol. 2011, p. 5.

[27] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," 2017.