Convolutional Low-Resolution Fine-Grained Classification

Dingding Cai, Ke Chen, Yanlin Qian, Joni-Kristian Kämäräinen

***Abstract*—Successful fine-grained image classification methods learn subtle details between visually similar (sub-)classes, but the problem becomes significantly more challenging if the details are missing due to low resolution. Encouraged by the recent success of Convolutional Neural Network (CNN) architectures in image classification, we propose a novel resolution-aware deep model which combines convolutional image super-resolution and convolutional fine-grained classification into a single model in an end-to-end manner. Extensive experiments on multiple benchmarks demonstrate that the proposed model consistently performs better than conventional convolutional networks on classifying fine-grained object classes in low-resolution images.**

***Index Terms*—Fine-Grained Image classification, Super Resolution Convoluational Neural Networks, Deep Learning**

The problem of image classification is to categorise images according to their semantic content (*e.g.* person, plane). Fine-grained image classification further divides classes to their "sub-categories" such as the models of cars [1], the species of birds [2], the categories of flowers [3] and the breeds of dogs [4]. Fine-grained categorisation is a difficult task due to small inter-class variance between visually similar sub-classes. The problem becomes even more challenging when available images are low-resolution (LR) images where many details are missing as compared to their high-resolution (HR) counterparts.

Since the rise of Convolutional Neural Network (CNN) architectures in image classification [5], the accuracy of fine-grained image classification has dramatically improved and many CNN-based extensions have been proposed [6], [7], [8], [9], [10], [11]. However, these works assume sufficiently good image quality and high resolution, (*e.g.* typically 227×227 for AlexNet [5]) while with low resolution images the CNN performance quickly collapses [12], [13]. The challenge raises from the problem of how to recover necessary texture details from low-resolution images. Our solution is to adopt image super-resolution (SR) techniques [14], [15], [16], [17], [18] to enrich imagery details. In particular, inspired by the recent work on CNN-based image SR by Deng *et al*. [19] we propose a unique end-to-end deep learning framework that combines CNN super-resolution and CNN fine-grained classification – a resolution-aware Convolutional Neural Network (RACNN) for fine-grained object categorisation in low-resolution images. To our best knowledge, our work is the first end-to-end learning model for low-resolution fine-grained object classification.

Our main principle is simple: the higher image resolution, the easier for classification. Our research questions are: Can computational super-resolution recover some of the important details required for fine-grained image classification and can such SR layers be added to an end-to-end deep classification architecture? To this end, our RACNN integrates deep residual learning for image super-resolution [20] into typical convolutional classification networks (*e.g.* AlexNet [5] or VGG-Net [21]). On one hand, the proposed RACNN has deeper network architecture (*i.e.* more network parameters) than the straightforward solution of conventional CNN on up-scaled images (*e.g.* bicubic interpolation [22]). Our RACNN learns to refine and provide more texture details for low-resolution images to boost fine-grained classification performance. We conduct experiments on three benchmarks, Stanford Cars [1], Caltech-UCSD Birds-200-2011 [2] and Oxford 102 Flower Dataset [3]. Our results answer the aforementioned questions: super-resolution improves fine-grained classification and SR-based fine-grained classification can be designed into a supervised end-to-end learning framework, as depicted in Figure 1 illustrating the difference between RACNN and conventional CNN.

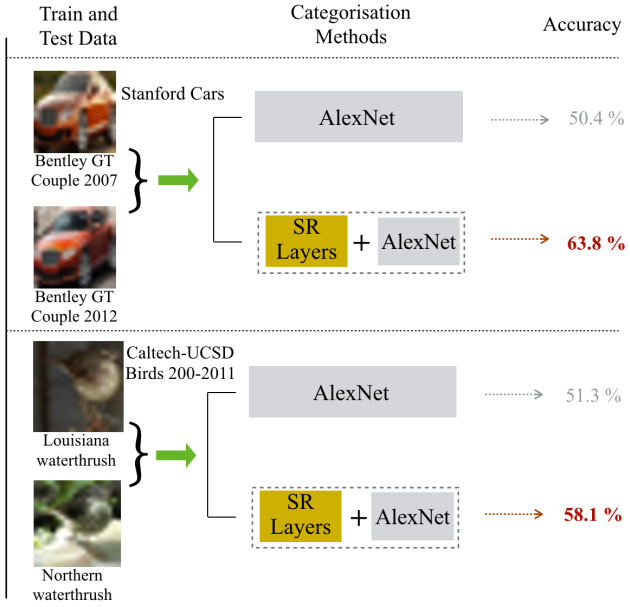


Fig. 1: Owing to the introduction of the convolutional super-resolution (SR) layers, the proposed deep convolutional model (the bottom pipelines) achieves superior performance for low resolution images.

I. Related Work

Fine-Grained Image Categorisation – Recent algorithms for discriminating fine-grained classes (such as animal species

or plants [2], [4], [23] and man-made objects [1], [24], [25]) can be divided into two main groups. The first group of methods utilises discriminative visual cues from local parts obtained by detection [6], [26] or segmentation [8], [27], [28]. The second group of methods focuses on discovering inter-class label dependency via pre-defined hierarchical structure of labels [10], [29], [30], [31], [32] or manually-annotated visual attributes [33], [34]. Significant performance improvement is achieved by convolutional neural networks (CNNs), but this requires a massive amount of high quality training images. Fine-grained classification from low-resolution images is yet challenging and unexplored. The method proposed by Peng *et al*. [35] transforms detailed texture information in HR images to LR via fine-tuning to boost the accuracy of recognizing fine-grained objects in LR images. However, in [35], their strong assumption requiring HR images available for training limits its generalisation ability. In addtion, the same assumption also occurs in Wang's work [36]. Chevalier *et al*. [12] design a CNN-based fine-grained object classifier with respect to varying image resolutions, which adopts ordinary convolutional and fully-connected layers but misses considering super-resolution specific layers in convolutional classification networks. On contrary, owing to the introduction of SR-specific layers in RACNN, our method can consistently gain notable performance improvement over conventional CNN for image classification on fine-grained classification datasets.

Convolutional Super-Resolution Layers – Yang *et al*. [37] grouped existing SR algorithms into four groups: prediction models [38], edge-based methods [39], image statistical methods [40] and example-based methods [14], [17], [18], [41], [42], [43], [44], [45]. Recently, Convolutional Neural Networks have been adopted for image super-resolution achieving state-of-the-art performance. The first attempt using convolutional neural networks for image super-resolution was proposed by Dong *et al*. [18]. Their method learns a deep mapping between low- and high- resolution patches and has inspired a number of follow-ups [20], [46], [47]. In [47], an additional deconvolution layer is added based on SRCNN [18] to avoid general up-scaling of input patches for accelerating CNN training and testing. Kim *et al*. [46] adopt a deep recursive layer to avoid adding weighting layers, which does not need to pay any price of increasing network parameters. In [20], a convolutional deep network is proposed to learn the mapping between LR image and its residue between LR and HR image to speed up CNN training for very deep network. Convolutional layers designed for image super-resolution (namely SR-specific convolutional layers) have been verified their effectiveness to improve the quality of images. In this work, we incorporate the state-of-the-art residual CNN layers for image super-resolution [20] into a convolutional categorisation network for classifying fine-grained objects (*i.e*. AlexNet [5], VGG-Net [21] and GoogLeNet [48]). In the experiments, SR-specific convolutional layers are verified to improve classification performance.

Contributions – Our contributions are two-fold:

- Our work is the first attempt to utilise super-resolution specific convolutional layers to improve convolutional fine-grained image classification.
- We experimentally verify that the proposed RACNN achieves superior performance on low-resolution images which make ordinary CNN performance collapse.

II. Resolution-Aware Convolutional Neural Networks

Given a set of N training images and corresponding class labels $\{\boldsymbol{X}_i, y_i\}, i = 1, 2, \cdots, N$, the goal of a conventional CNN model is to learn a mapping function $y = f(\boldsymbol{X})$. The typical cross entropy (ce) loss $L_{\text{ce}}(\cdot)$ on softmax classifer is adopted to measure the performance between class estimates $\hat{y} = f(\boldsymbol{X})$ and ground truth class labels y :

$$L_{\text{ce}}(\hat{y}, y) = -\sum_{j=1}^{l} y_j \log(\hat{y}_j), \tag{1}$$

where j refers to the index of element in vectors, and l denotes the dimension of softmax layer (*i.e*. the number of classes). In this sense, CNN solves the following minimisation problem with gradient descent back propagation:

$$\min \quad \sum_{i=1}^{N} L_{\text{ce}}(f(\boldsymbol{X}_i), y_i). \tag{2}$$

For fine-grained categorisation in low-quality images, we propose a novel resolution-aware convolutional neural network, which is illustrated in Fig. 2. In general, our RACNN consists of two parts: convolutional super-resolution layers (see Sec. II-A) and convolutional categorisation layers (see Sec. II-B). In Sec. II-C, we describe an end-to-end training scheme for the proposed RACNN.

A. Convolutional Super-Resolution Layers

In this section, we present convolutional super-resolution specific layers for the resolution-aware CNN, the goal of which is to recover texture details of low-resolution images to feed into the following convolutional categorisation layers.

We first investigate the conventional CNN for the super-resolution task. Given K training pairs of low-resolution and high-resolution images $\{\boldsymbol{X}^{\text{LR}}, \boldsymbol{X}^{\text{HR}}\}^i, i = 1, 2, \cdots, K$, a direct CNN-based mapping function $g(\boldsymbol{X}^{\text{LR}})$ from $\boldsymbol{X}^{\text{LR}}$ (input observation) to $\boldsymbol{X}^{\text{HR}}$ (output target) [18], [47] is learned by minimising the mean square (ms) loss

$$L_{\text{ms}}(\boldsymbol{X}^{\text{LR}}, \boldsymbol{X}^{\text{HR}}) = \frac{1}{2}\sum_{i=1}^{K} \|\boldsymbol{X}^{\text{HR}} - g(\boldsymbol{X}^{\text{LR}})\|^2. \tag{3}$$

Inspired by recent the state-of-the-art residual convolutional network [20] to achieve high efficacy, we design convolutional super-resolution layers as shown on the left hand side of Fig. 2. Similar to [20], our convolutional super-resolution layers learn a mapping function from LR images $\boldsymbol{X}^{\text{LR}}$ to residual images $\boldsymbol{X}^{\text{HR}} - \boldsymbol{X}^{\text{LR}}$. Object function of the proposed convolutional super-resolution layers is as the following:

$$\min \quad \frac{1}{2}\sum_{i=1}^{K} \|\boldsymbol{X}^{\text{HR}} - \boldsymbol{X}^{\text{LR}} - g(\boldsymbol{X}^{\text{LR}})\|^2. \tag{4}$$

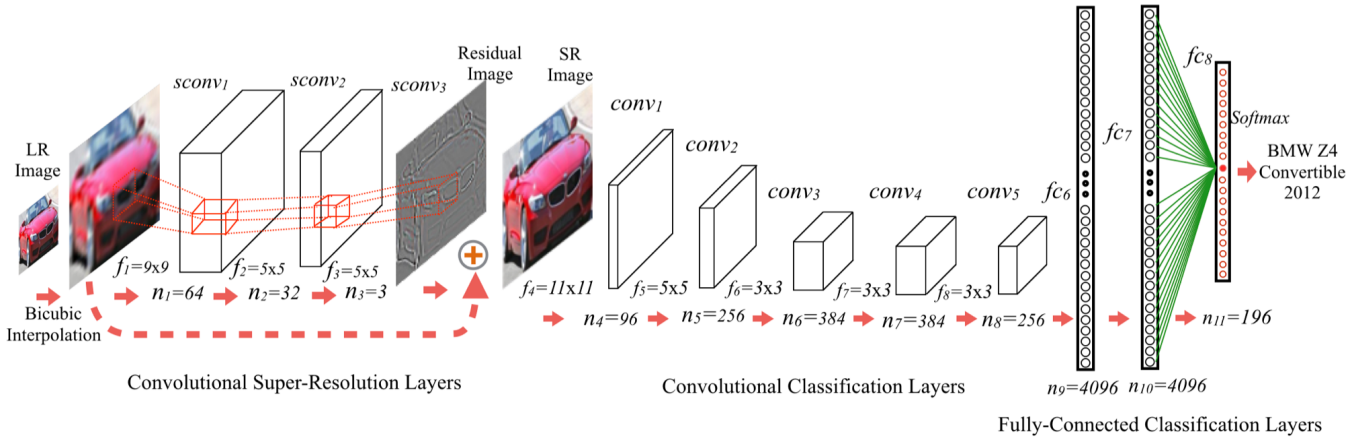


Fig. 2: Pipeline of the proposed resolution-aware convolutional neural network (RACNN) for fine-grained recognition with low-resolution images. Convolutional classification layers from AlexNet are adopted for illustrative purpose, which can be readily replaced by those from other CNNs such as VGG-Net or GoogLeNet

.

The better performance of residual learning yields from the fact that, since the input (LR) and output images (HR) are largely similar, it is more meaningful to learn their residue where similarities are removed. It is obvious that detailed imagery information in the form of residual images is easier for CNNs to learn than direct LR-HR CNN models [18], [47].

We utilise three typical stacked convolutional-ReLU layers with zero-padding filters as convolutional SR layers in RACNN. Following [18], the empirical basic setting of the layers is $f_1 = 9\times9$, $n_1 = 64$, $f_2 = 5\times5$, $n_2 = 32$, $f_3 = 5\times5$ and $n_3 = 3$, which are also illustrated in the left hand side of Fig. 2, where f_m and n_m donate the size and number of the filters of the mth layer respectively. The output of the last convolutional SR layer is summed with the low-resolution input image $\boldsymbol{X}^{\mathrm{LR}}$ to construct the full super-resolution image fed into the remaining convolutional and fully-connected classification layers of RACNN.

B. Categorisation Layers

The second part in our RACNN is convolutional and fully-connected classification layers with high quality images after super-resolution layers. A number of CNN frameworks [5], [21], [49], [48] have been proposed for image categorisation, and in this paper we consider three popular convolutional neural networks: AlexNet [5], VGG-Net [21] and GoogLeNet [48]. All CNNs typically consist of a number of Convolutional-ReLU-Pool stacks followed by several fully-connected layers. On the right-hand-side of Fig. 2, the typical AlexNet [5] is visualised and employed as convolutional categorisation layers in RACNN. AlexNet [5], the baseline CNN for large-scale image classification over ImageNet [50], consists of 5 convolutional layers (*i.e.* *conv1*, *conv2*, *conv3*, *conv4*, and *conv5*) and 3 fully-connected layers (*i.e.* *fc6*, *fc7*, and *fc8*). VGGNet [21] is made deeper (*i.e.* from 8 layers of Alexnet to 16-19 layers) and more advanced over AlexNet by using very small (*e.g.* 3×3) convolution filters. In our paper, we choose the VGG-Net-16 with 16 layers for our experiments (denoted as VGG-Net in the rest of the paper). GoogLeNet [48] comprises 22 layers but has much less number of parameters than AlexNet and VGG-Net owing to the smaller amount of weights of fully-connected layers. GoogLeNet generally generates three outputs at various depths for each input, but for simplicity only the last output (*i.e.* the deepest output) is considered in our experiments. In our experiments, all three networks are pre-trained on the Imagenet data and fine-tuned with $\{\boldsymbol{X}, y\}$ from fine-grained data as the baseline. For fair comparison, we fine-tune the identical pre-trained CNN models as our convolutional categorisation layers with $\{\boldsymbol{X}, y\}$ by replacing the dimension of final fully-connected layer with the size of object classes l.

C. Network Training

The key difference between the proposed resolution-aware CNN and conventional CNN lies in the introduction of three convolutional super-resolution layers. Evidently, RACNN is deeper than corresponding CNN due to the three convolutional SR layers, which can store more knowledge, *i.e.* network parameters. Before learning RACNN in an end-to-end fashion, we consider two weight initialization strategies for convolutional SR layers in RACNN, *i.e.* standard Gaussian weights and pre-trained weights on the ImageNet data. For fair comparison, we adopt the identical network structure for both initialisation schemes.

For RACNN with Gaussian initial weights, we train the whole network to minimise cross-entropy loss (2) directly. During training, we set learning rates 1 and weight decays 0.1 for the first two SR layers (*sconv1* and *sconv2*) and both learning rate and weight decay are set with 0.1 for the third convolutional SR layer (*sconv3*), while learning rates and weight decays are 0.1 and 0 for all categorisation layers except the last fully-connected layer which uses both learning rate and weight decay 1.

We consider an alternative initialisation strategy for better initial weights for convolutional SR layers. To this end, we

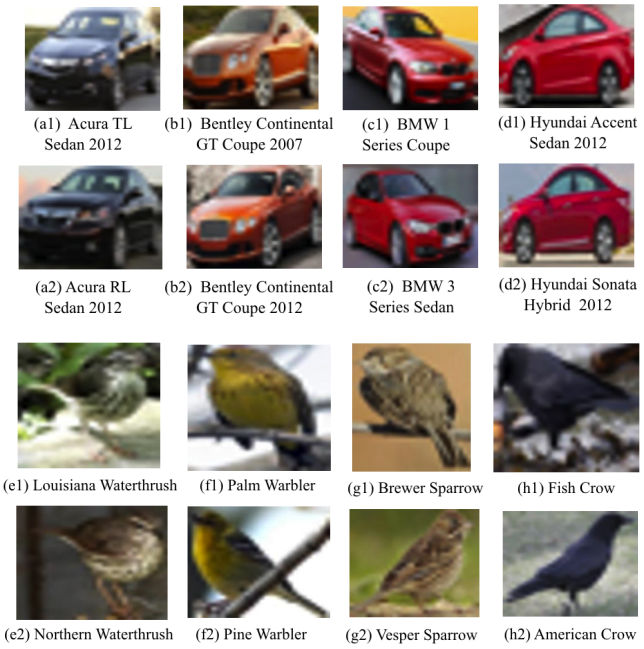


Fig. 3: Low-resolution image samples after removing background from the Stanford Cars and UCSD-Caltech Birds 200-2011 benchmarks.

pre-train the three convolutional SR layers by enforcing the minimal of the mean square loss (4) on ILSVRC 2015 ImageNet object detection testing dataset [51], which consists of 11,142 high-resolution images. Given the pre-trained weights in convolutional SR layers, RACNN is end-to-end trained by minimising the loss function (2) for categorisation. For the goal of direct utilisation of output of convolutional SR layers, we train SR layers in RGB color space with all the channels, instead of only on luminance channel Y in YCbCr color space [18]. Specifically, we generate LR images from HR images (*e.g.* 227×227 pixels) via firstly down-sampling HR images to 50×50 pixels and then up-scaling to the original image size by bicubic interpolation [22]. We then sample image patches using sliding window and thus obtain thousands of pairs of LR and HR image patches. To be consistent with the setting of RACNN using Guassian initial weights, the super-resolution layers are trained with image patches by setting learning rates being 1 and weight decays being 0.1 for the first two SR layers (*sconv1* and *sconv2*) and both learning rate and weight decay being 0.1 for the third SR layer (*sconv3*). Finally, we jointly learn both convolutional SR and classification layers in an end-to-end learning manner with learning rates 0.1 and weight decays 0 for all classification layers except the last fully-connected layer with both learning rate and weight decay set to 1.

III. Experiments

A. Datasets and Settings

We evaluate RACNN on three commonly-used datasets: the Stanford Cars [1], the Caltech-UCSD Birds-200-2011 [2] and the Oxford 102 Category Flower [3] datasets. The first one was released by Krause *et al*. for fine-grained categorisation and contains 16,185 images from 196 classes of cars and each class is typically at the level of Brand, Model and Year. By following the standard evaluation protocol [1], we split the data into 8,144 images for training and 8,041 for testing. Caltech-UCSD Birds-200-2011 is another challenging fine-grained image dataset aimed at subordinate category classification by providing a comprehensive set of benchmarks and annotation types for the domain of birds. The dataset contains 11,788 images of 200 bird species, among which there are 5,994 images for training and 5,794 for testing [2]. Oxford 102 Category Flower Dataset consists of 8,189 images which commonly appear in the United Kingdom. These images belong to 102 categories and each category contains between 40 to 258 images. In the standard evaluation protocol [3], the whole dataset is divided into 1,020 images for training, 1,020 for validation and 6,149 for testing. In our experiments the training and validation data are merged together to train the networks.

Images from these datasets are first cropped with provided bounding boxes to remove the background. Cropped images are down-sampled to LR images of the size 50×50 pixels and then up-scaled to 227×227 pixels by bicubic interpolation [22] to fit the conventional CNN, which follows the settings in [35]. Sample LR images from the both benchmarks are illustrated in Fig. 3, which verify our motivation to mitigate the suffering from low visual discrimination due to low-resolution. We compare our RACNN with multiple state-of-the-art methods, the corresponding CNN model for classification (*i.e.* AlexNet [5], VGG-Net [21] and GoogLeNet [48]) and Staged-Training CNN proposed by [35]. The proposed RACNN is implemented on Caffe [52]. We adopt *the average per-class accuracy* [13], [35] for the both datasets (the higher value denotes the better performance).

In our experiments, we used a Lenovo Y900 desktop with one Intel i7-6700K CPU and one Nvidia GTX-980 GPU. The proposed RACNN has deeper structure than the competing networks (*i.e.* AlexNet, VGGNet, GoogLeNet) which requires longer training times as indicated in Table I.

TABLE I: Training times of RACNNs and competing CNNs (seconds / epoch)

.

Methods	Cars [1]	Birds [2]	Flowers [3]
AlexNet	11	8	3
$\text{RACNN}_{\text{AlexNet}}$	111	80	25
VGGNet	133	82	34
$\text{RACNN}_{\text{VGGNet}}$	356	215	90
GoogLeNet	20	25	8
$\text{RACNN}_{\text{GoogLeNet}}$	136	120	59

B. Comparative Evaluation

In Fig. 4, we compare our results with AlexNet [5] and Staged-Training AlexNet [35] for fine-grained classification in low-resolution images. It is evident that our $\text{RACNN}_{\text{AlexNet}}$ consistently achieves the best performance on both benchmarks. Precisely, AlexNet achieves 50.4% and 51.3% ac-

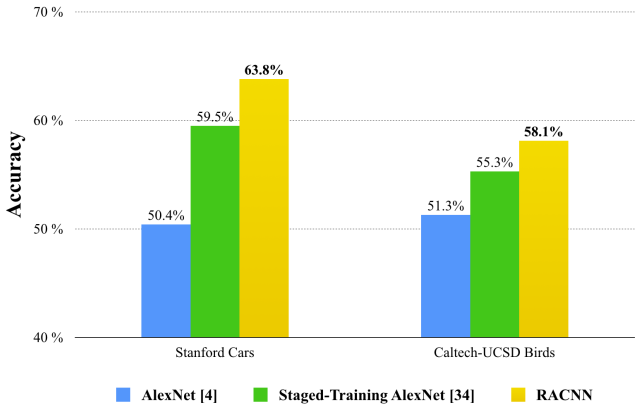


Fig. 4: Comparison to two state-of-the-art methods for classification (average per-class accuracies).

TABLE II: Evaluation on effect of convolutional SR layers to recover high resolution details. We fix all convolutional and fully-connected layers except the last fully-connected layer (*i.e.* extracted features correspond to those with high resolution images). g-RACNN and p-RACNN denote the proposed RACNN with weights initialized with Gaussian (g-) and pre-trained weights (p-) for the convolutional SR layers.

Methods	Cars [1]	Birds [2]	Flowers [3]
AlexNet [5]	43.75%	44.99%	70.03%
g-RACNN$_{\text{AlexNet}}$	45.77%	47.17%	71.91%
p-RACNN$_{\text{AlexNet}}$	**47.90%**	**51.23%**	**74.24%**
VGG-Net [21]	41.49%	43.46%	67.82%
g-RACNN$_{\text{VGG-Net}}$	42.86%	44.72%	68.03%
p-RACNN$_{\text{VGG-Net}}$	**44.65%**	**49.33%**	**69.17%**
GoogLeNet [48]	46.85%	48.52%	69.28%
g-RACNN$_{\text{GoogLeNet}}$	50.37%	55.16%	69.77%
p-RACNN$_{\text{GoogLeNet}}$	**50.76%**	**57.30%**	**73.51%**

curacies (collected from [35]) for the Stanford Cars and Caltech-UCSD Birds datasets, respectively. Knowledge transfer between varying resolution images (*i.e.* Staged-Training AlexNet [35]) can improve classification accuracy, that is 59.5% for the Stanford Cars and 55.3% for the Caltech-UCSD Birds. However, the staged-training AlexNet [35] relies on the strong assumption that high-resolution images are available for training, which limits to its usage to other tasks. Note that our method is more generic and transforms knowledge of super resolution across datasets, which indicates that our method can be readily applied to other low-resolution image classification tasks. The proposed RACNN$_{\text{AlexNet}}$ significantly beats its direct competitor AlexNet, *i.e.* 63.8% vs. 50.4% on the Stanford Cars dataset and 58.1% vs. 51.3% on the Caltech-UCSD Birds dataset. With the same settings and training samples, the performance gap can only be explained by the novel network structure of RACNN.

C. Evaluation of Convolutional SR Layers

In this experiment, we employ all layers in the AlexNet, VGG-Net and GoogLeNet [48] as categorisation layers in RACNN. Note that, different from the previous experiments, we freeze all categorisation layers by setting learning rates and weights decays to 0 besides the last fully-connected layers of the baseline CNNs, and our RACNN is then fine-tuned with low-resolution data. Such setting treats categorisation layers in RACNN as an identical classifier for evaluating the effect of adding convolutional SR layers. RACNN with initial Gaussian and pre-trained weights are called as g-RACNN and p-RACNN respectively. Comparative results are shown in Table II and Fig. 5. Both g-RACNN and p-RACNN consistently outperform the baseline CNNs in all experiments.

With the same experimental setting except different initial weights for convolutional super-resolution layers, the results of g-RACNN and p-RACNN are reported. Test set accuracies in Table II and Fig. 5 show that p-RACNN is superior to g-RACNN. p-RACNN and g-RACNN share the same network structure but differ only in network weights initialisation of convolutional SR layers. In this sense, better performance of p-RACNN is credited to the knowledge about refining low-resolution images (*i.e.* pre-trained weights), which verifies our motivation to boost low-resolution image classification via image super-resolution. It is noteworthy that since the feature extraction layers are frozen, the networks are not fine-tuned to low-resolution specific features, but all performance boost are owing to recovered high-resolution details important for classifcation by the super-resolution layers.

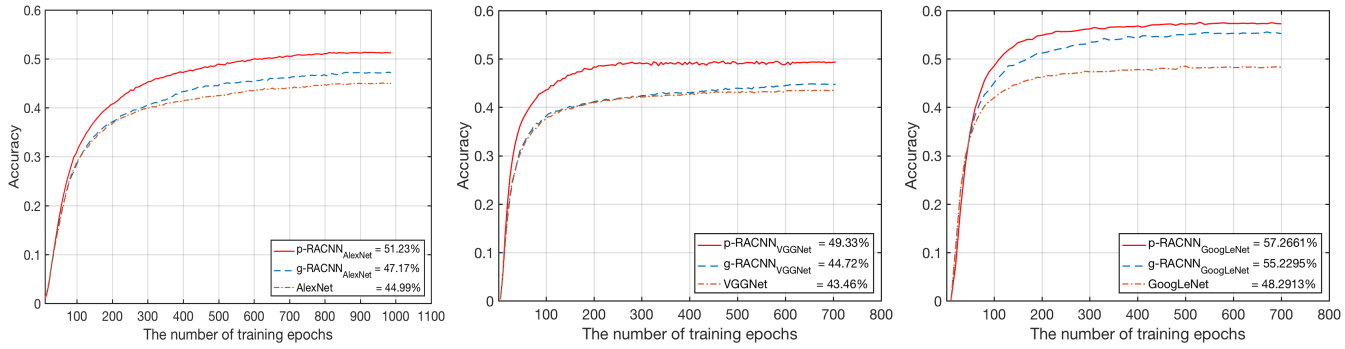
D. Evaluation on Varying Resolution

TABLE III: Comparison with varying resolution level (Res. Level) on the Caltech-UCSD Birds 200-2011 Dataset.

Res. Level	AlexNet [5]	g-RACNN$_{\text{AlexNet}}$	p-RACNN$_{\text{AlexNet}}$
25×25	31.58%	43.68%	**45.06%**
50×50	44.99%	47.17%	**51.23%**
100×100	51.01%	51.24%	**52.88%**

We further evaluate our proposed RACNN method with respect to varying resolutions on the Caltech-UCSD Birds 200-2011 Dataset. All low-resolution images are first up-scaled to the input image size, *i.e.* 227×227, before training models. The better performance of RACNN$_{\text{AlexNet}}$ over conventional AlexNet is achieved for cross-resolution fine-grained image classification, which is shown in Table III.

We observe that our method performs much better for lower resolution images (*e.g.* 25×25) than relatively high resolution images (*e.g.* 100×100). In details, p-RACNN$_{\text{AlexNet}}$ increases the accuracy by above 13% for 25×25 pixel images but less than 2% improvement on 100×100 resolution images. The reason is that the SR layers of RACNN play a significant role in introducing texture details especially when missing more visual cues of object classification in lower quality images, which further demonstrates our observation and motivation.



(a) AlexNet (b) VGG-Net (c) GoogLeNet

Fig. 5: Training process of AlexNet, VGGNet and GoogLeNet on the Caltech-UCSD Birds Dataset.

In p-RACNN$_{\text{AlexNet}}$, the weights for convolutional SR layers are pre-trained only with 50×50 resolution-level ImageNet images, but our RACNN is applied to varying resolution levels (*i.e.* 25×25 and 100×100). Further improvement on classification performance shows the generalisation of pre-trained weights for varying resolution levels, which demonstrates the generalisation ability of RACNN with pre-trained SR weights.

IV. Conclusion

We propose and verify a simple yet effective resolution-aware convolutional neural network (RACNN) for fine-grained image classification of low-resolution images. The results from extensive experiments indicate that the introduction of convolutional super-resolution layers to conventional CNNs can indeed recover fine details for low-resolution images and clearly boost performance in low-resolution fine-grained classification. This result can be explained by the fact that the super-resolution layers learn to recover high resolution details that are important for classification when trained end-to-end manner together with the classification layers. The concept of our paper is generic and the existing convolutional super-resolution and classification networks can be readily combined to cope with low-resolution image classification.

References

[1] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3d object representations for fine-grained categorization, in: International Conference on Computer Vision Workshops, 2013, pp. 554–561.

[2] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset.

[3] M.-E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: Indian Conference on Computer Vision, Graphics & Image Processing, IEEE, 2008, pp. 722–729.

[4] A. Khosla, N. Jayadevaprakash, B. Yao, F.-F. Li, Novel dataset for fine-grained image categorization: Stanford dogs.

[5] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advance in Neural Information Processing Systems, 2012.

[6] N. Zhang, J. Donahue, R. Girshick, T. Darrell, Part-based R-CNNs for fine-grained category detection, in: European Conference on Computer Vision, 2014.

[7] T.-Y. Lin, A. RoyChowdhury, S. Maji, Bilinear cnn models for fine-grained visual recognition, in: IEEE International Conference on Computer Vision, 2015, pp. 1449–1457.

[8] J. Krause, H. Jin, J. Yang, L. Fei-Fei, Fine-grained recognition without part annotations, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5546–5555.

[9] K. Chen, Z. Zhang, Learning to classify fine-grained categories with privileged visual-semantic misalignment, IEEE Transactions on Big Data.

[10] Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele, Evaluation of output embeddings for fine-grained image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[11] S. Branson, G. V. Horn, S. Belongie, P. Perona, Bird species categorization using pose normalized deep convolutional nets, in: British Machine Vision Conference, 2014.

[12] M. Chevalier, N. Thome, M. Cord, J. Fournier, G. Henaff, E. Dusch, LR-CNN for fine-grained classification with varying resolution, in: IEEE International Conference of Image Processing, 2015, pp. 3101–3105.

[13] Y. Liu, Y. Qian, K. Chen, J.-K. Kämäräinen, H. Huttunen, L. Fan, J. Saarinen, Incremental convolutional neural network training, in: International Conference of Pattern Recognition Workshop on Deep Learning for Pattern Recognition, 2016.

[14] R. Zeyde, M. Elad, M. Protter, On single image scale-up using sparse-representations, in: International Conference on Curves and Surfaces, 2010, pp. 711–730.

[15] J. Yang, J. Wright, T. S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE transactions on image processing 19 (11) (2010) 2861–2873.

[16] H. Chang, D.-Y. Yeung, Y. Xiong, Super-resolution through neighbor embedding, in: IEEE Conference on Computer Vision and Pattern Recognition, 2004.

[17] D. Glasner, S. Bagon, M. Irani, Super-resolution from a single image, in: IEEE International Conference on Computer Vision, 2009, pp. 349–356.

[18] C. Dong, C. C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE transactions on Pattern Analysis and Machine Intelligence 38 (2) (2016) 295–307.

[19] D. Dai, Y. Wang, Y. Chen, L. Van Gool, Is image super-resolution helpful for other vision tasks?, in: IEEE Winter Conference on Applications of Computer Vision, 2016, pp. 1–9.

[20] J. Kim, J. Kwon Lee, K. Mu Lee, Accurate image super-resolution using very deep convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1646–1654.

[21] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition.

[22] R. Keys, Cubic convolution interpolation for digital image processing, IEEE transactions on acoustics, speech, and signal processing 29 (6) (1981) 1153–1160.

[23] A. Angelova, S. Zhu, Efficient object detection and segmentation for fine-grained recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 811–818.

[24] M. Stark, J. Krause, B. Pepik, D. Meger, J. J. Little, B. Schiele, D. Koller, Fine-grained categorization for 3d scene understanding, International Journal of Robotics Research (2011) 1543–1552.

[25] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, A. Vedaldi, Fine-grained visual classification of aircraft, arXiv preprint arXiv:1306.5151.

[26] N. Zhang, R. Farrell, F. Iandola, T. Darrell, Deformable part descriptors for fine-grained recognition and attribute prediction, in: International Conference on Computer Vision, 2013.

[27] Y. Chai, V. Lempitsky, A. Zisserman, Symbiotic segmentation and part localization for fine-grained categorization, in: International Conference on Computer Vision, 2013.

[28] E. Gavves, B. Fernando, C. G. Snoek, A. W. Smeulders, T. Tuytelaars, Local alignments for fine-grained categorization, International Journal of Computer Vision 111 (2) (2015) 191–212.

[29] J. Shotton, M. Johnson, R. Cipolla, Semantic texton forests for image categorization and segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[30] S. J. Hwang, K. Grauman, F. Sha, Semantic kernel forests from multiple taxonomies, in: Advances in Neural Information Processing Systems, 2012.

[31] A. Mittal, M. B. Blaschko, A. Zisserman, P. H. Torr, Taxonomic multi-class prediction and person layout using efficient structured ranking, in: European Conference on Computer Vision, 2012.

[32] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, H. Adam, Large-scale object classification using label relation graphs, in: European Conference on Computer Vision, 2014.

[33] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, L. Bourdev, Panda: Pose aligned networks for deep attribute modeling, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[34] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, S. Gong, Transductive multi-view embedding for zero-shot recognition and annotation, in: European Conference on Computer Vision, 2014.

[35] X. Peng, J. Hoffman, X. Y. Stella, K. Saenko, Fine-to-coarse knowledge transfer for low-res image classification, in: IEEE International Conference of Image Processing, 2016, pp. 3683–3687.

[36] Z. Wang, S. Chang, Y. Yang, D. Liu, T. 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.

[37] C.-Y. Yang, C. Ma, M.-H. Yang, Single-image super-resolution: a benchmark, in: European Conference on Computer Vision, 2014, pp. 372–386.

[38] M. Irani, S. Peleg, Improving resolution by image registration, Computer Vision, Graphics, and Image Processing: Graphical models and image processing (1991) 231–239.

[39] R. Fattal, Image upsampling via imposed edge statistics, in: ACM Transactions on Graphics, Vol. 26, 2007, p. 95.

[40] J. Huang, D. Mumford, Statistics of natural images and models, in: IEEE Conference on Computer Vision and Pattern Recognition, 1999, pp. 541–547.

[41] J.-B. Huang, A. Singh, N. Ahuja, Single image super-resolution from transformed self-exemplars, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5197–5206.

[42] J. Yang, Z. Lin, S. Cohen, Fast image super-resolution based on in-place example regression, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1059–1066.

[43] G. Freedman, R. Fattal, Image and video upscaling from local self-examples, ACM Transactions on Graphics 30 (2) (2011) 12.

[44] D. Dai, R. Timofte, L. Van Gool, Jointly optimized regressors for image super-resolution, in: Computer Graphics Forum, Vol. 34, 2015, pp. 95–104.

[45] S. Schulter, C. Leistner, H. Bischof, Fast and accurate image upscaling with super-resolution forests, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3791–3799.

[46] J. Kim, J. Kwon Lee, K. Mu Lee, Deeply-recursive convolutional network for image super-resolution, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1637–1645.

[47] C. Dong, C. C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, in: European Conference on Computer Vision, 2016, pp. 391–407.

[48] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[51] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.

[52] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: ACM international conference on Multimedia, 2014, pp. 675–678.