EVALUATION OF SUPER-RESOLUTION ON BIRD DETECTION PERFORMANCE BASED ON DEEP CONVOLUTIONAL NETWORKS

Ce Li[2,3] *Hanwen Hu*[2] *Baochang Zhang*[2,1⋆]

[1] State Key Laboratory of Satellite Navigation System and Equipment Technology, Shijiazhuang, China
[2] Department of Automation, Beihang University, Beijing, China
[3] Department of Computer Science, China University of Mining & Technology, Beijing, China
⋆✉ bczhang@buaa.edu.cn; correspondence

ABSTRACT

Recent advances in image super-resolution and object detection algorithms have offered unprecedented potential for reconstructing low-resolution images and detecting various objects. In this paper, we aim to analyze reliability of bird detection from Low-Resolution (LR) images. We collect a dataset named BIRD-50[1] and a public dataset named CUB-200 of real bird images with different scale low-resolutions, then conduct a study to quantify the performance of several state-of-the-art Super-Resolution (SR) reconstruction algorithms using deep convolutional networks. By analyzing the influence of the resolution reduction on the bird detection, we demonstrate the functionality of SR on the bird detection performance improvement. Further experimental results analysis indicates that the inclusion of SR algorithms results in significant improved detection accuracies.

Index Terms— Low-resolution Image, Bird Detection, Deep Convolutional Networks

1. INTRODUCTION

Driven by the growing availability of aerial vehicles (AVs), bird detection and recognition play an essential role in diverse applications, such as tracking bird, monitoring bird population, bird behavior study [1]. However, bird detection from images obtained in the wild poses many challenges due to image size, resolution and the relatively small size of the objects.

In tacking the challenges of low-resolution and small object size in bird detection, we focus on performance of state-of-the-art bird detection techniques on High-Resolution (HR) birds reconstructed from single Low-Resolution (LR) bird images. Super-Resolution (SR) has been extensively studied in the past few years [2, 3, 4], especially the recentl earning-based methods [5, 6, 7, 8, 9, 10, 11]. Super-Resolution Convolutional Neural Network (SRCNN) [12] predicts the nonlinear LR-HR mapping via a fully convolutional network and outperforms tranditional non-DL(Deep Learning) methods. Then, Very Deep Super-Resolution (VDSR) [5] is proposed stacking 20 convolutional layers and trained with a high learning rate to solve the gradient explosion in SRCNN. On the other hand, Fast Super-Resolution by CNN (FSRCNN) [13] is proposed to accelerate SRCNN in 2016, which achieves a speed up of more than 40 times with superior reconstruction quality. Following the rule of "the deeper the better" in SR, another Deep Recursive Residual Network (DRRN) [11] is constructed with 52 convolutional layers and trained with residual learning and recursive learning, which significantly outperforms the previous DL and non-DL single image super-resolution (SISR) methods in 2017.

Likewise, the Convolutional Neural Networks (CNNs) [14, 15, 16] have also been achieved in object detection. Recent DL-based object detection methods, especially Region-based Convolutional Neural Networks (R-CNNs) [17, 18], usually follow the paradigm of *proposal + classification* that transfers the knowledge learned in recognition tasks to the problem of detection. The proposals are usually generated by bottom-up methods that exploit low-level cues [19, 20] requiring less candidates [17]. Faster R-CNN [21] are best known for a practical object detection in terms of both speed (5fps including all steps on GPU) and accuracy (1st-place winning in ILSVRC and COCO 2015 competitions). Since then, the framework of Faster R-CNN have been adopted and successfully generalized to many other methods, such as 3D object detection [22], part-based detection [23], and image captioning [24]. However, as far as we know, there is no study that investigate the effect of super-resolution on bird detection performance with varying bird resolutions.

In this paper, in order to show how these algorithms would perform on bird images acquired in the wild, we collect a dataset named BIRD-50 and a public dataset named CUB-200 with varying scale low-resolutions, and propose a deep CNN based framework to handle the problems of bird detection in low-resolution images as shown in Fig. 1. In this framework, we first conduct a comprehensive study on the

[1]BIRD-50 will be publicly valiable at website: https://github.com/bczhang/bczhang/

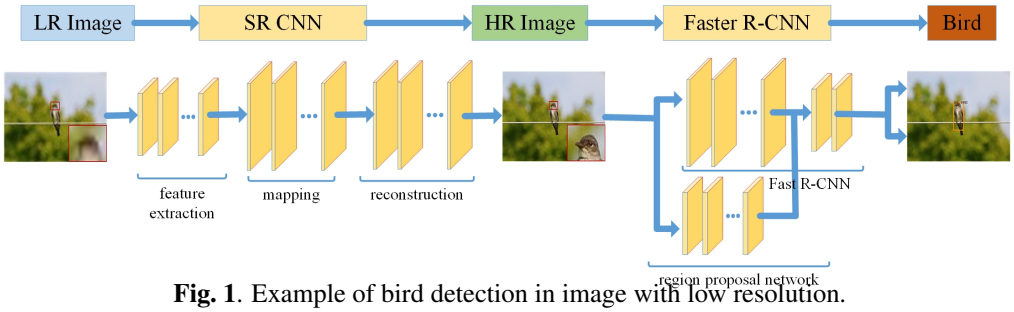


Fig. 1. Example of bird detection in image with low resolution.

reconstruction quality of three DL-based state-of-the-art SR algorithms. Then, we demonstrate the functionality of them by comparing the bird detection performance before and after SR in a deep network, and show the significant improved detection accuracies with inclusion of SR. The main contributions are:

1. We collect a dataset named BIRD-50 and a public dataset named CUB-200 with varying scale low-resolutions to evaluate the performance of small-sized bird detection in LR images.
2. We evaluate extensive single image SR algorithms on bird detection in the wild using Faster R-CNN.

2. FRAMEWORK AND METHODS

We first briefly describe the network structures of three comparative SR methods including VDSR [5], FSRCNN [13], DRRN [11] and bird detection method faster R-CNN [21].

2.1. Super-Resolution

We evaluate three state-of-the-art SR algorithms with publicly available source code from Internet. Fig. 2 simply overviews the network structures of evaluated algorithms, respectively. The strategies used in network structures are illustrated in Table 1. All SR algorithms aim at learning an end-to-end mapping function F between the input LR image Y and the output HR image X.

Before VDSR, the baseline SRCNN consists of 9×9 sized patch extraction filters, 1×1-sized non-linear mapping filters, and 5×5-sized reconstruction filters. As shown in Fig. 2(a), VDSR first introduces residual learning from ResNet [25] and successfully stacks 20 weight layers (3×3 for each layer) in the residual branch. VDSR uses extremely high learning rates that is 10^4 times higher than SRCNN [26], and reconstruction information (receptive field) is much larger (41×41 *vs.*13×13). Denoting a residual image $R^i=Y^i-X^i$ given a training LR X^i and HR Y^i image-pair, the cost function of VDSR is represented in Table 1, $F\left(X^i;\theta\right)$ is the network prediction for X^i with parameter θ.

FSRCNN [13], developed from SRCNN [12], is a real-time (24fps) SR method that achieves tens of (17.36) times

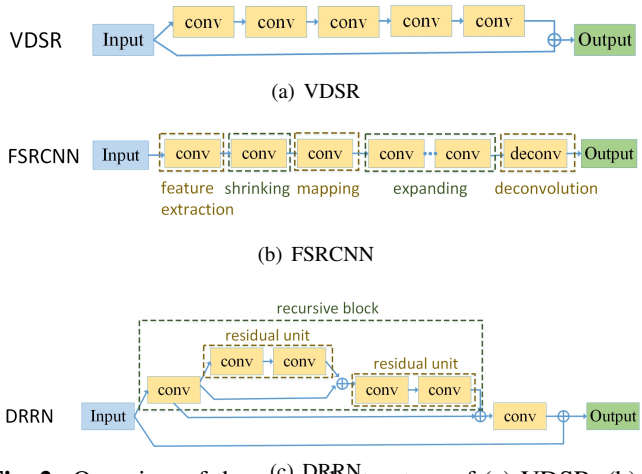


Fig. 2. Overview of the network structure of (a) VDSR, (b) FSRCNN, and (c) DRRN.

faster and better super-resolution quality than other contemporary SR methods. As shown in Fig. 2(b), FSRCNN can be decomposed into feature extraction, shrinking, mapping, expanding and deconvolution. We denote feature extraction layer as $Conv(5,d,1)$, where the filter number n_1 can be regarded as the feature dimension d. Then, a smaller filter number s is adopted in shrinking part denoted as $Conv(1,s,d)$. In the non-linear mapping, m mapping layers contain same numbers of filter s as $m\times Conv(3,s,s)$. Similarly, expanding and deconvolution layer are $Conv(1,d,s)$ and $DeConv(9,1,d)$, respectively. The cost function is represented in Table 1, $F\left(Y^i;\theta\right)$ is the network output for X^i with parameter θ.

DRRN [11], generalized version of VDSR [5], consists of several stacked recursive blocks and residual units to multi-path mode local residual learning the reconstruction residual between the LR and HR images. The residual image is then added to the global identity mapping from the input LR image. The whole network structure of DRRN is illustrated in Fig. 2(c). In each recursive block, the recursive block number and the residual unit number are two key parameters. In Table 1, $\mathcal{R}$ is the recursive block, and f_{Rec} is the last convolutional layer to recontruct the residual. VDSR is actually viewed as a special case of DRRN without residual unit.

Table 1. Strategies used in VDSR [5], FSRCNN [13] and DRRN [11]. Y^i is the network output for X^i with parameter θ. f_* is a function for the convolutional or deconvolutional layer, further explained in each network description.

Strategy & Cost Function
VDSR:
$Y^i = f_{Rec}\left(f_{d-1}\left(f_{d-2}\left(\ldots\left(f_1\left(X^i\right)\right)\ldots\right)\right)\right) + X^i$
$\mathcal{L}\left(\theta\right) = \frac{1}{2N}\sum_{i=1}^{N}\left\|Y^i - X^i\right\|^2$
FSRCNN:
$Y^i = f_{DeConv}\left(f_{m+2}\left(\ldots\left(\left(f_2\left(f_1\left(X^i\right)\right)\right)\right)\ldots\right)\right)$
$\mathcal{L}\left(\theta\right) = \frac{1}{N}\sum_{i=1}^{N}\left\|F\left(Y^i;\theta\right) - X^i\right\|^2$
DRRN:
$Y^i = f_{Rec}\left(\mathcal{R}_{\mathcal{B}}\left(\mathcal{R}_{B-1}\left(\ldots\left(\mathcal{R}_1\left(X^i\right)\right)\ldots\right)\right)\right) + X^i$
$\mathcal{L}\left(\theta\right) = \frac{1}{2N}\sum_{i=1}^{N}\left\|Y^i - X^i\right\|^2$

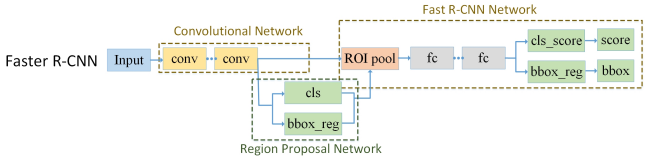


Fig. 3. Overview of the network structure of Faster R-CNN.

2.2. Bird Detection

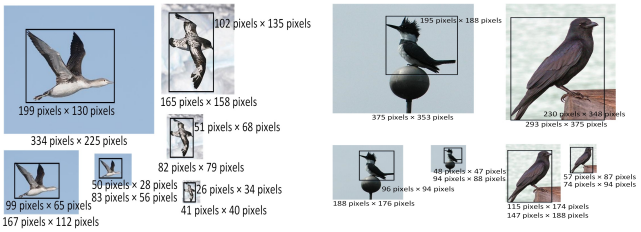
After SR reconstruction, we use the Faster R-CNN [21] to detect bird. Faster R-CNN consists of two modules, Region Proposal Network (RPN) and Fast R-CNN detector, shown in Fig. 3. To generate region proposals, RPN is taken $n \times n$ spatial window of convolutional feature map as input, and implemented with a convolutional layer followed by two sibling1×1 convolutional layers for bouding box regression and classification. Then, for the Fast R-CNN detector, the anchors is key to share features. The bird classification score and bounding box is learned by

$$\mathcal{L}\left(p_i, t_i\right) = \frac{1}{N_{cls}}\sum_i \mathcal{L}_{cls}\left(p_i, p_i^*\right) + \lambda\frac{1}{N_{reg}}\sum_i p_i^*\mathcal{L}_{reg}\left(t_i, t_i^*\right),$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$ are classification loss and regression loss corresponding with p_i and t_i, respectively. p_i is the predicted probability of anchor i being a bird. p_i^* is a label value 1 or 0 representing positive or negative anchor. Similarly, t_i and t_i^* are predicted bounding box and ground truth, respectively.

3. EXPERIMENTAL RESULTS

In this section, we aim to exploit the effectiveness of single image SR algorithms on bird detection in wild using deep CNN. All SR and faster R-CNN methods are tested using Matlab and Caffe library.



(a) *BIRD-50* (b) *CUB-200*

Fig. 4. Examples of (a) *BIRD-50* and (b) *CUB-200* datasets.

Table 2. Average PSNR (dB) of different SR methods.

Method	BIRD-50		CUB-200	
	Scale-2	Scale-4	Scale-2	Scale-4
VDSR	36.52	32.05	32.65	28.70
FSRCNN	35.63	31.52	32.03	28.34
DRRN	36.09	32.11	32.44	28.76

3.1. Bird Image Datasets

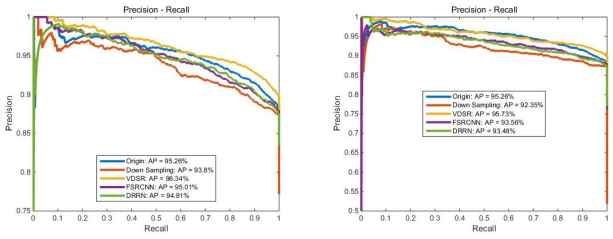
BIRD-50 (BUAA Birds50) is a manually collected dataset in part by us and other team in BUAA, which is annotated with 50 bird species including 2751 images (about 55 images per category) captured under wild conditions. CUB-200 (Caltech-UCSD Birds200 [27]) is a public dataset annotated with 200 bird species including 6033 images (about 30 images per class), mostly North American bird species. Each image is annotated with a bounding box, a rough bird segmentation and a set of attribute labels. We perform downsampling operations using the bicubic interplolation with resolution reduction of scales$\times 2$ and$\times 4$, respectively. In total, three groups of images with Original, Scale-2 and Scale-4 resolutions are used for experiments. In Fig. 4(a) and Fig. 4(b), the sizes of example images are ranging from 375 pixels$\times$353 pixels (large) to 41 pixels$\times$40 pixels (small), and the sizes of bird bounding boxes are from 199 pixels$\times$130 pixels (large) to 26 pixels$\times$34 pixels (small).

3.2. Performance Evaluation

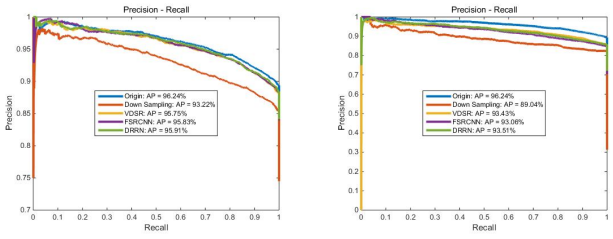
We first compare the reconstruction quality among VDSR [5], FSRCNN [13] and DRRN [11] on two datasets, then investigate the corresponding bird detection performance using Faster R-CNN [21], and finally dicuss the effectiveness of SR algorithms on bird detection in the wild LR images.

Reconstruction Quality. The average signal-to-noise ratio (PSNR, unit:dB) is calculated [28]. In our experiments, a PC with Intel i7 CPU, GTX 1070 GPU, and pretrained VDSR, FSRCNN, DRRN protocol are applied to finetune LR bird images. Similar to [5, 13, 11], we choose $d = 20$ for VDSR, FSRCN and DRRN. One third part of data is used for training and the other data is for testing in the reconstruction task.

We reconstruct HR bird images from two groups of LR



(a) Scale-2 on *BIRD-50* (b) Scale-4 on *BIRD-50*



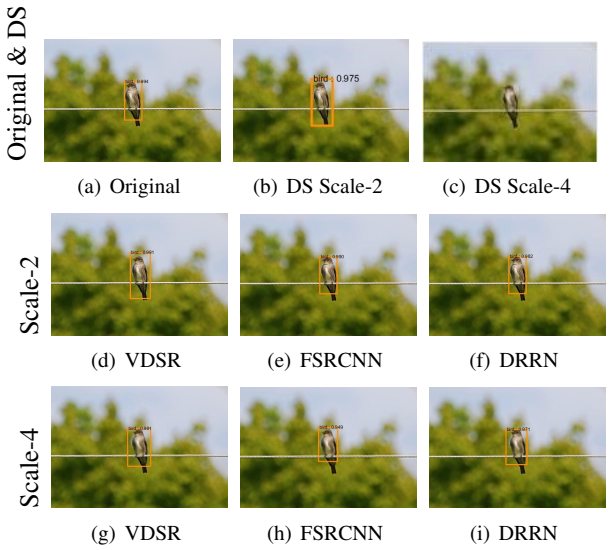
(c) Scale-2 on *CUB-200* (d) Scale-4 on *CUB-200*

Fig. 5. Influence of image resolution on bird detection in *BIRD-50* and *CUB-200* datasets.

images with Scale-2 and Scale-4 resolutions. In Table 2, VDSR and DRRN obviously achieve higher average PSNR than FSRCNN on both two datasets. On Scale-2, VDSR has 0.76 dB PSNR improvement and DRRN has 0.45 dB PSNR improvement than FSRCNN. On Scale-4, VDSR has 0.45 dB PSNR improvement and DRRN has 0.51 dB improvement than FSRCNN. It is clear that DRRN outperforms VDSR in average PSNR on larger datasets with larger scale factors. This is because that reconstruction ability of DRRN network, especially in signal to noise ratio, is relevant to the data size, which enhances SR performance on lower resolution images.

Detection Performance. We further perform bird detection using faster R-CNN [21] set up with the model finetuned on CUB200. In Fig. 5, we plot the detection precision-recall curve results of original images, down sampling images with two scales resolution reduction and reconstructed images via different SR methods. It is noted that the detection curve of VDSR, FSRCNN and DRRN reconstructed images are higher than down sampling images. The SR methods allow for HR images with clear and sharp edges to enhance the detection recall together with precision.

We also summerize the average precision of detection in Fig. 5. Obvisouly, all SR methods enhance the bird detection precisions, and VDSR reconstructed images actually have achieve higher precision than other methods, except for Scale-4 CUB-200 dataset. On Scale-2, VDSR has 2.88% improvement of detection precision, FSRCNN has 2.16% precision enhancement and DRRN has 2.28% enhancement. On Scale-4, VDSR has 4.33% improvement of detection precision, FSRCNN has 2.89% enhancement and DRRN has 3.04% enhancement. It indicates that the inclusion of SR algorithms result in distinct improved detection accuracies.



(a) Original (b) DS Scale-2 (c) DS Scale-4

(d) VDSR (e) FSRCNN (f) DRRN

(g) VDSR (h) FSRCNN (i) DRRN

Fig. 6. Detection results of an *Olive* bird. Larger detection score is better.

Moreover, the detection results of *Olive* bird example are demonstrate in Fig. 6. It is noted that the small-sized *Olive* bird is not detected on DS Scale-4 image, but all of birds in three reconstructed images by SR methods are totally detected and the precision scores are close to that of original image. This is significant since it indicates that applying SR methods prior to detection process provides better results, and VDSR shows detection performance superiority consistent with above reconstruction results.

4. CONCLUSION

This paper carried out three methods to demonstrate the functionality of the SR on bird detection performance before and after SR in a deep framework. Two datasets have been collected and introduced to measure the reconstruction quality and the effect on bird detection. Experiments indicate that the inclusion of SR result in significant improved detection accuracies. In future, we will investigate much deeper DRRN network and more efficient detection network.

5. ACKNOWLEDGEMENTS

The work was supported in part by the Natural Science Foundation of China under Contract 61601466, 61672079, 61473086. The work of B. Zhang was supported in part by the Program for New Century Excellent Talents University within the Ministry of Education, China, and in part by the Beijing Municipal Science and Technology Commission under Grant Z161100001616005. Baochang Zhang is the correspondence.

6. REFERENCES

[1] D. W. Stowell, M. Stylianou, Y. Glotin, and Hervé, "Bird detection in audio: A survey and a challenge," *Machine Learning for Signal Processing (MLSP)*, vol. 30, 2016.

[2] C. Fookes, F. Lin, V. Chandran, and S. Sridharan, "Evaluation of image resolution and super-resolution on face recognition performance," *Journal of Visual Communication and Image Representation*, vol. 23, no. 1, 2012.

[3] R. Timofte, V. Smet, and L. Gool, "Anchored neighborhood regression for fast example-based super-resolution," *ICCV*, pp. 1920–1927, 2013.

[4] P. Rasti, T. Uiboupin, S. Escalera, and G. Anbarjafari, "Convolutional neural network super resolution for face recognition in surveillance monitoring," *Springer International Publishing*, pp. 175–184, 2016.

[5] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," *TPAMI*, vol. 38, no. 2, 2015.

[6] N. Rusk, "Accelerating the super-resolution convolutional neural network," *ECCV*, vol. 9905, no. 1, 2016.

[7] A. ElSayed, A. Mahmood, and T. Sobh, "Effect of super resolution on high dimensional features for unsupervised face recognition in the wild," *arXiv*, 2017.

[8] Y. F. Wang, L. J. Wang, H. Y. Wang, and P. H. Li, "End-to-end image super-resolution via deep and shallow convolutional networks," *arXiv*, vol. 1607.07680, 2016.

[9] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," *arXiv*, 2015.

[10] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *TPAMI*, vol. 38, no. 2, 2016.

[11] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," *CVPR*, 2017.

[12] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolu- tional network for image super-resolution," *ECCV*, p. 184?99, 2014).

[13] C. Dong, C. L. Chen, and X. Tang, "Accelearting the super-resolution convolutional neural networks," *ECCV*, 2016.

[14] J. Carreira and C. Sminchisescu, "Cpmc: Automatic ob- ject segmentation using constrained parametric mincuts," *TPAMI*, 2012.

[15] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeul ders, "Selective search for object recognition," *IJCV*, 2013.

[16] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," *CVPR*, 2014.

[17] R. Girshick, "Fast r-cnn," *ICCV*, 2015.

[18] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun, "Object detection networks on convolutional feature maps," *arXiv*, vol. 1504.06066, 2015.

[19] K. E. V. Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders, "Segmentation as selective search for object recognition," *ICCV*, pp. 1879–1886, 2011.

[20] C. L. Zitnick and P. Dollar, "Edge boxes: locating object proposals from edges," *ECCV*, pp. 391–405, 2014.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *TPAMI*, 2015.

[22] S. Song and J. Xiao, "Deep sliding shapes for amodal 3d object detection in rgb-d images," *arXiv*, vol. 1511.02300, 2015.

[23] J. Zhu, X. Chen, and A. L. Yuille, "Deepm: A deep part-based model for object detection and semantic part localization," *arXiv*, vol. 1511.07131, 2015.

[24] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," *arXiv*, vol. 1511.07571, 2015.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, pp. 770–778, 2015.

[26] C. Dong, C. L. Chen, and K. He, "Image super-resolution using deep convolutional networks," *TPAMI*, vol. 38, no. 2, pp. 295–307, 2014.

[27] P. Welinder, S. Branson, and T. Mita, "Caltech-ucsd birds 200," *California Institute of Technology*, 2010.

[28] Z. Wang, A. C. Bovik, and H. R. Sheikh, "Image quality assessment: from error visuality to structural similarity," *TIP*, vol. 13, no. 4, 2004.