Object-Oriented Super-Resolution Imaging

Binqian Zeng(bz866@nyu.edu) **Xinsheng Zhang(xz1757@nyu.edu)**

Abstract

We introduce a pipeline[1] that can quickly provide super-resolution for desired objects in real-time. We create object only image set (OOIS) to train the super-resolution generative adversarial network (SRGAN). The SRGAN trained on our OOIS can capture more details of objects. We also improve the training process of the SRGAN by implementing the modifications proposed by Wasserstein GAN. The improved SRGAN trained on OOIS dataset achieve not only the best evaluation score but also can provide super-resolved image in a shorter time.

1 Introduction

Thinking about the scenario when you zoom in a picture and want to see more explicit of an object, you may be frustrated because all you see are a bunch of pixels without any details. Therefore, you may need an application that can quickly provide super-resolution for desired objects in real-time.

In this project, we propose a fast object-oriented application pipeline for image super-resolution by using deep neural networks. The pipeline contains three modules. The first module is object detection and segmentation in which we detect desired objects and segment object patches by bounding boxes from the background image (Liu et al., 2016). The second module leverages Generative Adversarial Network for image super-resolution(SRGAN) (Ledig et al., 2016) to reconstruct segmented object patches into high-resolution. Instead of using whole images to train SRGAN, we will feed only object images to push model more focus on object details. The third module is for smoothing seams and deblurring the background.

Our object-oriented image super-resolution pipeline is expected to produce high-resolution images and provide better details on desired objects than existing techniques. Moreover, our pipeline is less computational expensive and can be treat as a real-time object super-resolution application.

2 Related Work

Objects Detection

Recently deep learning models have been proved successfully tackling such image classification and object detection problems. One of the successful uses of deep learning for object detection was the OverFeat model proposed by Sermanet et al. (2013). They proposed a multiscale and sliding window approach algorithm using Convolutional Neural Networks(CNNs) for object classification, localization, and detection problems. Followed by Fast-R-CNN, it applies the CNN on the complete input image and then combines the feature vector extracted from pooling layer of Region of Interest (RoI) with a final feed forward network for classification and object detection (Girshick, 2015). Compared with previous approaches, You Only Look Once(YOLO) detection system proposed by Redmon et al. (2015) is a simple convolutional neural network approach which treats object detection as a single regression problem and achieves both high accuracy and speed.

[1] `https://github.com/nyuxz/ds1008_final_project`

Image Super-Resolution

In the aspect of single image super-resolution, SRCNN was proposed by Dong et al. (2014), fits non-linear mapping through three-layer convolutional neural network, which exceeds previous prediction-based methods. Then, DRCN apply recursive neural network on super-resolution problem for the first time (Kim et al., 2016), which achieve similar result as DRCN. LapSRN replaces the pre-defined bicubic interpolation with the learned transposed convolutional layers and optimizes the network with a robust loss function (Lai et al., 2017). The learned transposed convolutional layers enables sharing features between lower levels, which increases the non-linearity at finer convolutional layers to learn complex mappings. The original robust loss function enables the network to learn information from each level. SRDenseNet construct dense blocks,which pass features in each level to all upcoming levels so that all features from each level are concatenated (Tong et al., 2017).

3 Object Only Image Set

The images we used to create Object Only Image Set (OOIS) is obtained from the PASCAL Visual Object Classes Challenge 2012 (PascalVOC) (Everingham et al., 2015). We employ object detection model (Liu et al., 2016) on PascalVOC to predict object bounding boxes for each image. Then we segment object patches by bounding boxes. The collection of all object patches is our new OOIS dataset. To enrich the data samples, we also employ data augmentation on OOIS dataset. For each object patch, we rotate it by a degree between -45 and 45, and also horizontal flip it.

4 Methodology

4.1 Pipeline Overview

The first stage of our pipeline is using SSD model to detect multiple objects in an image. Then, we apply SRGAN on object patches, and Bicubic interpolation on background respectively. After achieving objects super-resolution, we embed super-resolved objects into the interpolated background corresponding to their initial locations. The third stage is using smoothing and deblurring techniques to handle splicing seams. The architecture of the pipeline is shown in Figure 1.

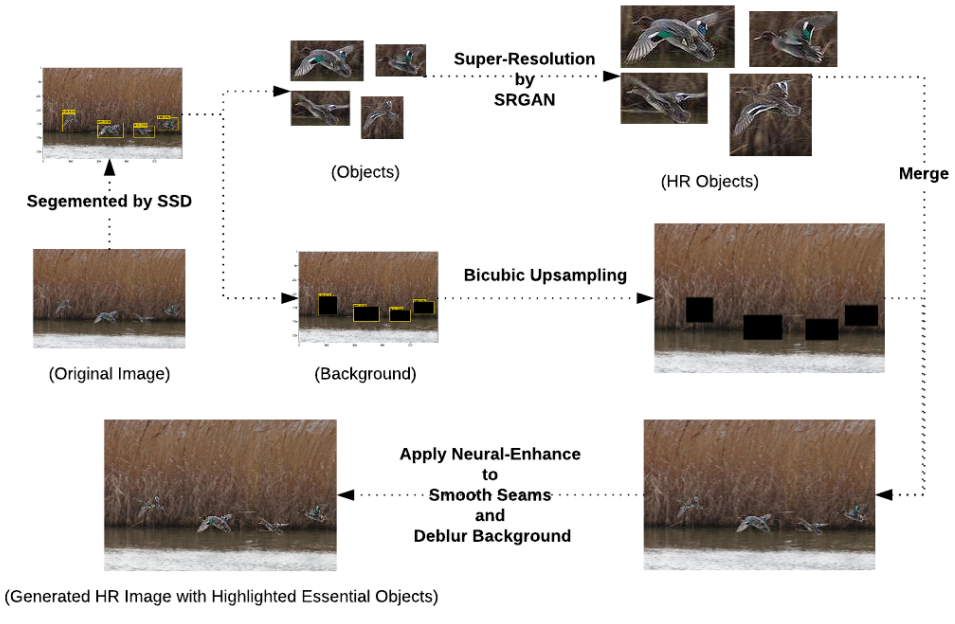

Figure 1: The Object-Oriented Super-Resolution Imaging Pipeline

4.2 Objects Detection and Segmentation

Single Shot MultiBox Detector proposed by Liu et al. (2016) achieves good performance in terms of both precision for object detection and speed. It scores over 74% mean Average Precision (mAP) and can produce 59 frames per second (FPS) on standard datasets such as PascalVOC. Our proposed pipeline is for real-time application, therefore we choose SSD model for object detection task. The architecture of SSD is based on VGG-16 without fully connected layers. Instead of the original VGG fully connected layers, a set of auxiliary convolutional layers (from conv6 onwards) were added, thus enabling to extract features at multiple scales and progressively decrease the size of the input to each subsequent layer (Forson, 2017). We replicate SSD code based on deGroot & Brown (2017), and trained model on PascalVOC2012 benchmark dataset.

4.3 Image Super-Resolution Reconstruction

We employ SRGAN proposed by Ledig et al. (2016) to solve super-resolution task. Comparing with SRResNet or other state-of-art models trained on MSE loss, SRGAN utilizes perceptual loss and adversarial loss to achieve more realistic results. The generator (G) is trained to generate HR output from LR image and try to fool the discriminator (D). The discriminator is trained to distinguish between super-resolved image and real image. As we can see from Figure 2, the main part of G is consisted by five residual blocks. Each residual block contains two sequential convolutional layers and then followed by batch-normalization (BN) and activation layers. Similarly, the main part of D is consisted by seven feed forward blocks, but each block only contains one convolutional layer, then followed by BN and activation layers. After seven blocks, resulting feature maps are fed into two dense layers and sigmoid activation function.

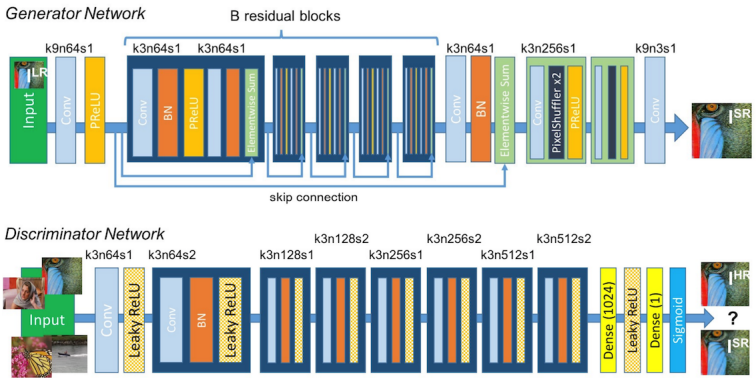

Figure 2: Architecture of Generator and Discriminator Network (Ledig et al., 2016).

Instead of training SRGAN on existing benchmark dataset such as PascalVOC and COCO, we will train SRGAN on our OOIS dataset. We conjecture that SRGAN trained on our dataset will capture more details of objects and reconstruct more realistic object image patches.

The training process for SRGAN is to solve an adversarial min-max problem:

$$\min_{\theta_G} \max_{\theta_D} E_{I_{HR} \sim p_{train}(I_{HR})}[\log D_{\theta_D}(I_{HR})] + E_{I_{LR} \sim p_G(I_{LR})}[\log(1 - D_{\theta_D}(G_{\theta_G}(I_{LR}))] \tag{1}$$

With the inspiration of WGAN (Arjovsky et al., 2017; Gulrajani et al., 2017), we modify the loss function of the original SRGAN to achieve better training stability. We modify the objective function of the generator G as following,

$$\min_{\theta_G}(l\_mse + l\_vgg + l\_gen) \tag{2}$$

where l_mse is pixel-wise MSE loss. l_vgg is based on ReLU activation layers of the pre-trained 16 layer VGG network, and it is defined as euclidean distance between reconstructed image and original

image. Then l_gen is defined as

$$-D_{\theta_D}(G_{\theta_G}(I_{LR})) \tag{3}$$

We train Discriminator D based on following objective function,

$$\min_{\theta_D}(D_{\theta_D}(G_{\theta_G}(I_{LR})) - D_{\theta_D}(I_{HR}) + GP) \tag{4}$$

GP is gradient penalty term. The gradient penalty term is defined as following,

$$GP = E_{\hat{x}\sim p_{\hat{x}}}[(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2] \tag{5}$$

where $p_{\hat{x}}$ is defined as interpolation between I_{LR} and I_{HR}.

4.4 Seams Smoothing & Background Deblurring

Since information in the background is not as important as that provided by objects, we will conduct experiments on existing super-resolution techniques to reconstruct the background and smooth seams. Potential techniques should have acceptable accuracy, visual quality, and high speed.

After merging HR objects back to the up-sampled background, seams exist between objects and the background. We apply pre-trained Neural- Enhance module (alexjc & et al., 2016) to smooth seams and deblur the whole image. The Neural-Enhance module is a product which combines results from Johnson et al. (2016), Shi et al. (2016), Kim et al. (2016), and Ledig et al. (2016). Pre-trained model for image deblurring and repairing are released on alexjc & et al. (2017). The Neural-Enhance module is able to process per 1080p image in about 2 seconds with a single GPU. Sample results from Neural-Enhance are shown in Figure 3.

Figure 3: Sample Results from Neural-Enhance (alexjc & et al., 2016).

5 Experiment and Result

5.1 Evaluation Metrics

We use peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) to evaluate and compare super-resolution algorithms. We define PSNR score as following,

$$\textbf{pixel-wise MSE} = \frac{1}{mn}\sum_{0}^{m-1}\sum_{0}^{n-1}||f(i,j) - g(i,j)||^2 \tag{6}$$

$$\textbf{PSNR} = 10\log_{10}(\frac{1}{\textbf{pixel-wise MSE}}) \tag{7}$$

SSIM score is calculated on the y-channel of center-cropped image (Ledig et al., 2016). This score can precisely measure the similarity between the original image and reconstructed image.

5.2 Quantitative Evaluation

We train SRGAN model and modified SRGAN model (improvements are based on WGAN) on three different datasets: PascalVOC, OOIS, and OOIS dataset with augmentations (OOIS_AUG) respectively. We test the model performance on the **Set 5** benchmark dataset (Bevilacqua et al., 2012). The evaluation results are shown in Table 1. As we can see, the modified SRGAN based on WGAN loss function trained on our OOIS with augmentation dataset achieves the best performance on both PSNR and SSIM.

Evaluation on Set 5 Benchmark Dataset		
Model	PSNR	SSIM
SRGAN+PascalVOC	27.671	0.826
SRGAN+OOIS	28.444	0.830
SRGAN+OOIS_AUG	28.776	0.841
Modified SRGAN+PascalVOC	28.279	0.827
Modified SRGAN+OOIS	28.827	0.839
Modified SRGAN+OOIS_AUG	29.323	0.860

Table 1: Evaluation Results. For example, SRGAN+OOIS refers to the original SRGAN model trained on our OOIS dateset.

5.3 Time Analysis

We test 100 image samples on our pipeline with a single GPU. The average time to reconstruct a super-resolved object-oriented image is 3.04s. If we apply SRGAN on the whole image, it will cost 6.18s. Therefore, our object-oriented super-resolution pipeline can provide more realistic object details and in the mean time save computational cost.

5.4 Sample Results

The Figure 4 shows the comparison between objects in original LR images and objects in generated HR images. The super-resolved images provide more details comparing with low resolution images.

Figure 4: Object Super-Resolution Results (Original (Left) vs. Generated (Right))

6 Conclusion and Future Work

Our super-resolution pipeline can fast produce HR images with enriched object details, and it costs around 1s per image. It shows the SRGAN trained on our OOIS dataset achieves better evaluation scores. According to WGAN, the improved SRGAN by changing loss function improves super-resolution result as well. For the future work, the Neural-Enhance module can be replaced by other faster techniques. Based on our pipeline, we will develop an application on devices with limited computational resources.

References

alexjc and et al. Neural-enhance github for super-resolution imaging, 2016. URL `https://github.com/alexjc/neural-enhance/releases`.

alexjc and et al. Pre-trained neural-enhance models for image deblurring and repairing, 2017. URL `https://github.com/alexjc/neural-enhance/releases`.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/arjovsky17a.html.

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, pp. 1–10, 2012. doi: 10.5244/C.26.135. URL https://doi.org/10.5244/C.26.135.

Max deGroot and Ellis Brown. Ssd: Single shot multibox object detector, in pytorch. 2017. URL https://github.com/amdegroot/ssd.pytorch.

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*, pp. 184–199. Springer, 2014.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1): 98–136, 2015.

Eddie Forson. Understanding ssd multibox realtime object detection in deep learning @ONLINE, November 2017. URL https://towardsdatascience.com/understanding-ssd-multibox-real-time-object-detection-in-deep-learning-495ef744fab.

Ross B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015. URL http://arxiv.org/abs/1504.08083.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017. URL http://arxiv.org/abs/1704.00028.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pp. 694–711. Springer, 2016.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1637–1645, 2016.

Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 624–632, 2017.

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*, 2016.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pp. 21–37. Springer, 2016.

Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015. URL http://arxiv.org/abs/1506.02640.

Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013. URL http://arxiv.org/abs/1312.6229.

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*, pp. 1874–1883, 2016.

Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4809–4817. IEEE, 2017.