CRF based Deep Saliency Detection and Generative Background Manipulation

Yoonsik Kim [1] **Jae Woong Soh** [1] **Sungyeob Han** [1] **Adayev Marat** [2]

Abstract

We proposed a method to detect a saliency map of object and manipulate the background of the given object. Saliency detection, which is to identify the most visually distinctive objects in an image, is useful as a pre-processing step for many computer vision tasks. Recently, great features from convolutional neural network (CNN) has shown considerable progress in this area. However, CNN-based saliency detection results blurry map near the boundary of salient objects. Applying conditional random field (CRF) in super-pixel domain can refine the saliency map from CNN, and make sharp and fine boundaries. Generative background manipulation is the scheme that neural networks find a different and suitable background for the given object. To manipulate the background for the object, we learn the probability model of the dataset by generative adversarial networks. In our project, we first present the saliency detection using CNN and CRF, then we manipulate the saliency map by transferring saliency map to the image domain using GAN.

1. Introduction

Saliency detection, extracting visually important or distinctive objects, has been an important research area in computer vision. Early works were done by extracting hand-crafted features (Jiang et al., 2013). Results from perceptual research (Einhäuser & König, 2003), visual contrast is the most important feature in visual saliency, therefore had been usually used. The hand-crafted features tend to perform well in standard cases, but they are not sufficiently robust for all complicated cases. Recently, deep neural network has shown great performances in many areas. Especially, convolutional neural network (CNN) has shown extensive power in most computer vision tasks such as object recognition (Simonyan & Zisserman, 2014), object detection (Ren et al., 2015), semantic segmentation (Long et al., 2015), super-resolution (Dong et al., 2014) and etc. In saliency detection, the powerful features from CNNs has been improved the accuracy (Li & Yu, 2015), (Li & Yu, 2016). However, deep features from CNN are typically based on object recognition models, thus they include more high-level features than spatial information. They mostly operate in patch-level instead of pixel-level, which means labeling each pixel from CNN features are a challenging problem. By using fully convolutional networks and multi-scale features from CNN, which is first introduced in FCN (Long et al., 2015), it is possible to get some spatial information, however they are somewhat blurry especially in boundaries of salient objects. Post-processing with Conditional Random Field (CRF) from probabilistic graphical model can compensates above limitation.

Background manipulation is an interesting challenge in computer vision society, because it shows various spectacles and applies many fascinating techniques. In this project, we propose to manipulate image except for salient region part. We first extract a saliency map on target image based on CRF. Using the saliency map, we manipulate the background part, which is generated by iGAN (Zhu et al., 2016). The background images are suitable for saliency-part, then we should explore the background in the neighborhood of the manifold or generate background using conditional relation (Mirza & Osindero, 2014).

In this paper, we discuss the methods:

- We optimize the saliency detection based on Convolutional Neural Networks (CNN) by using Conditional Random Fields.
- We discuss the mathematical backgrounds of manipulating background and explain why generating images with many categories.
- We propose the method to change backgrounds by manipulating on the latent vector space.

[1]Department of Electrical and Computer Engineering, SNU [2]Department of Computer Science and Engineering, SNU. Correspondence to: Sungyeob Han <syhan@cml.snu.ac.kr>.

Final Reports of M1522.001300 Probabilistic Graphical Models (2017 Fall) in Seoul National University,

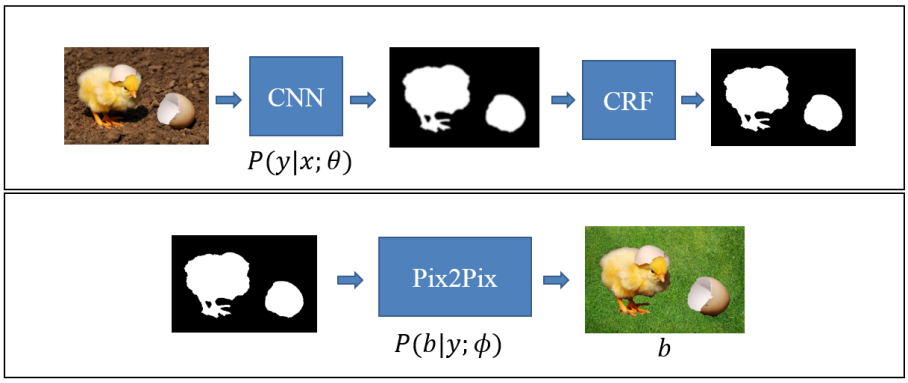

Figure 1. Overall scheme of proposed method

2. Related Works

2.1. Saliency Detection

Saliency detection can be done using handcrafted feature (Jiang et al., 2013) or CNN-based feature (Li & Yu, 2015), (Li & Yu, 2016). In this project, we focus on CNN-based ones. VGG-16 network (Simonyan & Zisserman, 2014) is one of the successful CNN, which has been shown human-level performance for object recognition task. Since, it can represent low-level features to high-level features from hierarchical structure, it is exploited for many computer vision tasks. For saliency detection, features from VGG-16 network are used for better features in discriminating salient objects.

For pixel-wise labeling, fully convolutional networks using multi-scale features (Long et al., 2015), (Li & Yu, 2016) have shown considerable performances. By modifying the VGG network and extracting multi-scale features from different depth of the VGG network, both high-level features and low-level spatial features are unified, therefore can represent considerable results. We adopt multi scale fully convolutional networks.

From recent paper namely DCL (Li & Yu, 2016), multi-scale fully convolutional network and CRF are jointly adopted. CRF in super-pixel domain enhances blurry saliency map from CNN and represents finer map. We adopt CRF as a post-processing step for our saliency detection.

2.2. Background Manipulation

Background generation for saliency detected images is a interesting topics in computer vision and helps people to understand which backgrounds are mathematically matched with the given object. As a result from previous step we have black and white image, where white area - object and black area - background region. In order to solve this image-to-image translation task we used idea of Image-to-Image Translation with Conditional Adversarial Networks (Isola et al., 2016) and release of the pix2pix software associated with this paper. The solution based on conditional GANs, which is conditioned on the input. Then the goal of such GAN is to generate image that is close as much as possible to target given input. For example, in Figure 1, there is shown example of background generation by conditional GAN given white and black input image.

Generative models are the latent variable models that map from latent variables,which are unobserved in training data, to observations in the data space. Traditionally, latent variable models are based on probabilistic graphical model such as Bayesian Networks and Hidden Markov Models. To learn parameters in latent variable models, we use expectation maximization (EM) because latent vectors are not observed. The traditional generative models, such as Boltzmann machine and FVBNS, are costly models which take O(n) time to generate a sample or are intractable to calculate the data distribution (Goodfellow, 2016). In 2014, Ian J. Goodfellow et al. suggested Generative Adversarial Networks (GAN), which improve both time complexity and the quality of images drastically (Goodfellow et al., 2014). Unlike Bayesian Networks, GANs do not calculate the probability between latent vectors and observations directly. Instead, GANs estimate the distribution implicitly based on maximum likelihood.

GANs have many varieties to improve their performances and to use over many applications. Generating backgrounds on the high resolutional image domain is one of the challenging topics in generative models, because GAN is very hard to train. Boundary Equilibrium GAN introduces a equilibrium enforcing method that helps GAN not to be collapsed (Berthelot et al., 2017)

GANs describe the probability distribution of trained datasets by intoducing latent variables. We have to sample a random vector in the latent vector space to generate a sample in the image domain by forward propagation. Therefore, if we want to search a suitable background for the

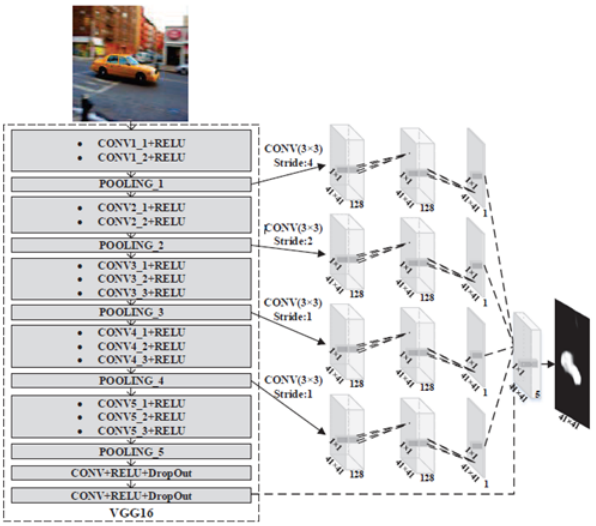

Figure 2. Multi-scale Fully Convolutional Network

given object, we find a relation between latent variables and generated images. Conditional GAN is one of the solution, which introduce condition inputs to make the generator networks sample an image based on the condition inputs (Mirza & Osindero, 2014). The other solution is searching on the latent vector space, which is known as iGAN (Zhu et al., 2016).

3. Proposed methods

Figure 1 shows overall scheme of our proposal. For saliency detection, CNN-based saliency map is first extracted and then CRF algorithm enhances the blurry saliency map. Then as a reverse, transferring saliency map to an image, we generate a reasonable image corresponding to a saliency map. We can think of this whole task in probabilistic point of view. From observation which is an image, which can be referred as x, it is to find latent y which is a saliency map. We can think of this task as a posteriori $P(y|x)$ inference. Then, form latent y which is a saliency map, generating an image b which is most likely to be natural can be thought as likelihood $P(b|y)$ inference.

3.1. Deep Saliency Detection

As shown in Figure 2, the structure of CNN-based saliency detection is mostly from (Li & Yu, 2016). As it is shown, the network is based on VGG-16 but to remain more spatial information, ast two pooling layers were skipped. From each feature map of each pooling layer, separate 1×1 convolutions were done to extract meaningful patches. Multi-scale patches are convolved and bilinear interpolated to obtain coarse saliency map. Then CRF is adopted to get finer saliency map. The model solves binary pixel labeling problem with energy function as follow.

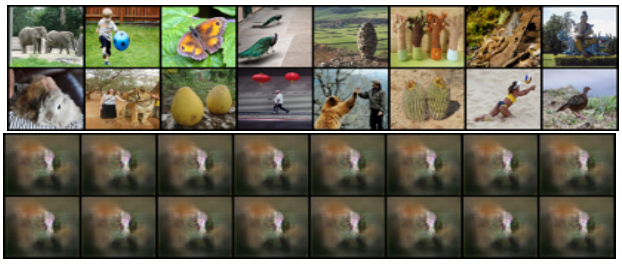

Figure 3. Samples of images in HKU-IS dataset (left), Model Collapsed BEGAN trained on HKU-IS dataset

$$E(L) = -\sum_i \log P(l_i) + \sum_{i,j} \theta_{i,j}(l_i, l_j) \tag{1}$$

where L denotes a binary label assignment for all pixels, $P(l_i)$ is the probability of pixel i having label l_i. First term can be thought as the unary term and the second term is the pair-wise potential. For $theta_{i,j}$,

$$\theta_{i,j} = \mu(l_i, l_j)[w_1 \exp(-\frac{||p_i - p_j||^2}{2\sigma_\alpha^2} - \frac{||I_i - I_j||^2}{2\sigma_\beta^2}) + w_2 \exp(-\frac{||p_i - p_j||^2}{2\sigma_\gamma^2})] \tag{2}$$

where $\mu_(l_i, l_j) = 1$ if $l_i \neq l_j$, and 0 otherwise. The above kernel is widely used in CRF for computer vision tasks, which penalizes the difference in positions and intensities, encourages to have same label within similar colors and near positions. The second part of the kernel is to remove small isolated regions. All the tasks are done in super-pixel wise. Energy minimization is approximately done based on mean field inference to the CRF distribution.

3.2. Generative Background Manipulation

In this paper, background manipulation is not just a method that changes background images with existing image files. We want to manipulate the background on the image dataset space domain. In other words, we generate new backgrounds for the given object, which is suitable for the object.

3.2.1. GAN OBJECTIVES

We consider the background generating problem as the problem of filling the missing parts. We can consider the probability of the suitable background is the maximum of the probability of background given the segmented object $\mathbf{y}$. Now to find a optimal background $\mathbf{x}_B$,

We

$$\mathbf{x}_B^* = \arg\max p(\mathbf{x}_B|\mathbf{y}) \tag{3}$$

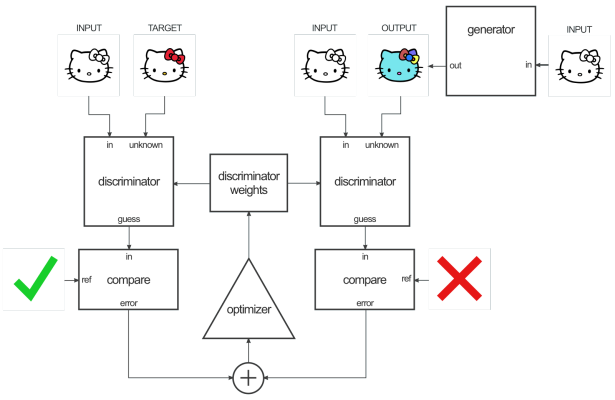

Figure 4. Pix2Pix Network Architecture

We consider Equation 3 as the Bayes Network, then

$$p(\mathbf{x}_B|\mathbf{y}) = \frac{p(\mathbf{x}_B, \mathbf{y})}{p(\mathbf{y}} \tag{4}$$

$$= \left(\prod_{i \in O} \delta(y_i - x_i) \right) P(x_B, x_O) \tag{5}$$

$$= P_{likelihood}(\mathbf{y}|\mathbf{x}) P_{prior}(x) \tag{6}$$

,where O is the observed (given) image, and δ is the Kronecker-delta function. Therefore, we should get a prior probability model of dataset in order to get the optimal background image. We tried to build the background probability model function by 3 methods. First, we used a tradition generative method, Markov Random Fields by optimizing based on KL divergence, but it was intractable. Next, we make the Boundary Equilibrium GAN model to train the prior probability of the dataset.

The HKU-IS datset, however, has too diverse categories and high resolutional property, so the model was collasped even if we used a equilibrium enforcing method, as 3. We decided not to train the marginal prior probability. Therefore we use the conditional probability property to generate a background when the target latent vector is given. We assume that the saliency map is the latent vector. The following equations are derivations for conditional GAN(pix2pix) objectives to find a optimal background.

$$\mathcal{L}_{cGAN} = E_{x,y}\left[\log D(x,y)\right] + E_{x,z}\left[\log(1 - D(x, G(x,z)))\right] \tag{7}$$

$$\mathcal{L}_{L1}(G) = E_{x,y,z}[\|y - G(x,z)_1\|_1] \tag{8}$$

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G) \tag{9}$$

3.2.2. Image to Image Translation

In architecture of the model is shown on 4. The network is includes two main parts, the Generator and the Discriminator. The Generator applies some transform to the input image to map it to the output image. The Discriminator checks whether the input image is similar to an either ground truth image from the dataset or an output of the generator. Then discriminator tries to guess if this was produced by the generator. In order to train the discriminator, initially the generator generates an output image. The same discriminator checks input/target pair and input/output(from generator) pair, then discriminator outputs how real they are. As a result the weights of discriminator are adjusted based on the classification error of the input pairs. The generators weights are adjusted based on two arguments: L1 loss between input/target and discriminator output.

4. Dataset

We used HKU-IS dataset provided by (Li & Yu, 2016). These dataset targets for detecting the saliency map on the high-resolutional images. Therefore, the dataset covers very diverse categories of objects. When we trained on the background manipulation, we resized image to small square images to help GAN model to be converged.

5. Experiments

5.1. Saliency Detection Experimental Results

5.1.1. Experimental Setup

We evaluate the performance on HKU-IS dataset which is one of most challenging and largest dataset for salient object detection. It is composed of 4447 images including ground truth, most of which have either low contrast or multiple salient objects. From HKU-IS (Li & Yu, 2015) dataset, we used 1447 images for test which is officially divided from provider.

We evaluate the quantitative performance graphically precision-recall (PR) curves, ROC curves (false positive-false negative curve). We also evaluate the quantitative performance using tables when the threshold is determined maximizing the F-measure which which contains the results of average precision, recall, and F-measure. The F-measure is defined as

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{10}$$

5.1.2. Comparison the methods

We explore the saliency detection for our projects and can approach some algorithms. We compare the three saliency detection model YS, MDF, and DCL. YS (Hwang et al., 2017; Kim et al., 2017) where the author is same as the report author is originally designed for skin detection method, but it is adapted to saliency detection method with saliency dataset. It is trained with HKU-IS training images. MDF (Li

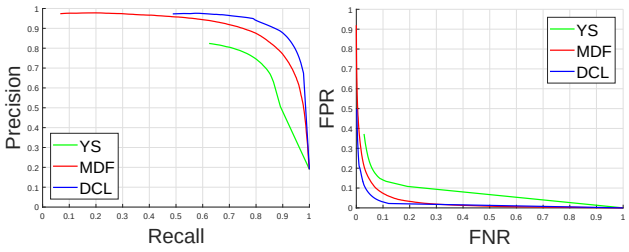

Figure 5. Comparison of PR and ROC curves on HKU-IS datasets (from left to right).

Table 1. Comparison of quantitative results including precision, recall, F-measure at peak F-measure and process time

METHODS	PRECISION	RECALL	F-MEASURE	OPT(S)
YS	0.7817	0.7891	0.7854	**0.2**
MDF	0.8435	0.8375	0.8405	-
DCL	**0.8838**	**0.8933**	**0.8886**	2.5

& Yu, 2015) is initially proposing the multi-scale CNN for saliency detection. DCL (Li & Yu, 2016) is the reference paper of our project report. The quantitative evaluation is presented in Figure 5 and Table 1. We can find that DCL have best performance for most of measures with big margin of other methods. Though, the process time is higher than YS, our algorithm's goal is not reducing the computational time. We also visualize qualitative comparison in Figure 6. Since YS and MDF is super-pixel based method, it has some false positive (near the elephant leg) and false negative (near the chick) region. Furthermore, DCL can refine pixel wisely using CRF and has more clear image with high confidence where the intensity means that confidence of algorithm classifying the salient object.

5.2. Generative Background Manipulation

5.2.1. EXPERIMENTAL SETUP

We trained the pix2pix adversarial network on HKU-IS dataset (aff). The original images in HKU-IS is rectangular, so we resized images to 256×256 for convenience. Generator's architecture is "U-net": convolutional network (Ronneberger et al., 2015), which takes 256x256 image as input, and discriminator is "PatchGAN" (Isola et al., 2016).

5.2.2. BACKGROUND MANIPULATION RESULTS

As training dataset we used same HKU-IS dataset from previous step of project. Change of loss over training time is shown on 7. To train network we used 4134 images. The result of testing step is shown on 8. We observe that the structure of the target object is preserved in the generated backgrounds. We can find some relations between

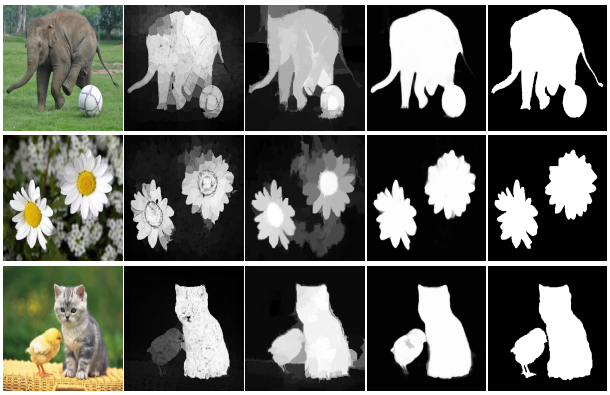

Figure 6. Visual comparison on the HKU-IS dataset: (from left top to right bottom) input, YS, MDF, DCL and ground truth images.

the segmented object and generated backgrounds. we concluded that our method is based on the maximum likelihood method, so the network generates can train the relative information.

6. Conclusions

CRF based saliency detection shows the great performance to detect the boundaries of the target object. We think that the combination of deep neural networks and probabilistic graphical model methods can cover the shortage of each other. We observed that training the generative models on the high-quailty image dataset is very hard, and Giving generator networks conditional variables can be one solution. We will continue to research the method to generate diverse and high-resolutional background images to improve our models. We concluded that learning the relation between objects and backgrounds is a key to understanding the probabilisitc models of images.

Figure 7. Background Manipulation Result on the HKU-IS dataset: (from left to right) saliency map, generated background, and ground truth

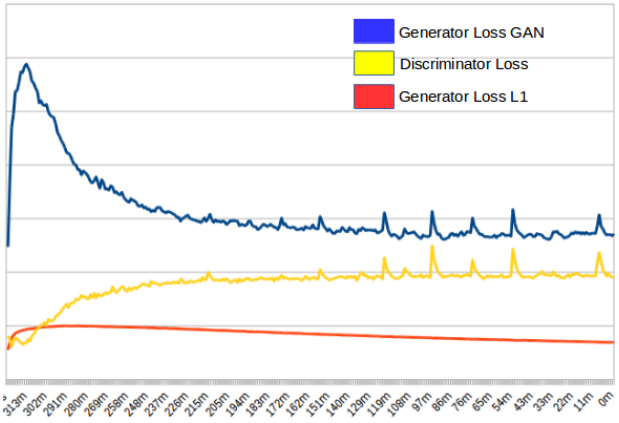

Figure 8. The Training loss of generator network and discriminator network

References

Image-to-image translation in tensorflow. https://affinelayer.com/pix2pix/.

Berthelot, David, Schumm, Tom, and Metz, Luke. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

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

Einhäuser, Wolfgang and König, Peter. Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience*, 17(5): 1089–1097, 2003.

Goodfellow, Ian. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Hwang, Insung, Kim, Yoonsik, and Cho, Nam Ik. Skin detection based on multi-seed propagation in a multi-layer graph for regional and color consistency. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 1273–1277. IEEE, 2017.

Isola, Phillip, Zhu, Jun-Yan, Zhou, Tinghui, and Efros, Alexei A. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016. URL http://arxiv.org/abs/1611.07004.

Jiang, Peng, Ling, Haibin, Yu, Jingyi, and Peng, Jingliang. Salient region detection by ufo: Uniqueness, focusness and objectness. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1976–1983, 2013.

Kim, Yoonsik, Hwang, Insung, and Cho, Nam Ik. Convolutional neural networks and training strategies for skin. In *Image Processing (ICIP), 2017 24th IEEE International Conference on*, pp. 1273–1277. IEEE, 2017.

Li, Guanbin and Yu, Yizhou. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5455–5463, 2015.

Li, Guanbin and Yu, Yizhou. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 478–487, 2016.

Long, Jonathan, Shelhamer, Evan, and Darrell, Trevor. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

Mirza, Mehdi and Osindero, Simon. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Ren, Shaoqing, He, Kaiming, Girshick, Ross, and Sun, Jian. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

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

Zhu, Jun-Yan, Krähenbühl, Philipp, Shechtman, Eli, and Efros, Alexei A. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pp. 597–613. Springer, 2016.