A Novel Face Inpainting Approach Based on Guided Deep Learning

Nermin M. Salem
Electrical Engineering Department
Future University in Egypt
Cairo, Egypt
nfawzy@fue.edu.eg

Hani M. K. Mahdi
Computer and Systems Engineering Department
Ain Shams University
Cairo, Egypt
hani.mahdi@eng.asu.edu.eg

Hazem M. Abbas
Computer and Systems Engineering Department
Ain Shams University
Cairo, Egypt
hazem.abbas@eng.asu.edu.eg

Abstract— In the last few years, deep learning has shown significant improvement for many computer vision open problems, especially Image inpainting. Image inpainting is the process of filling missing regions across images. One of the most challenging problems in image inpainting is face inpainting. In this work, a new novel approach for face inpainting is proposed which can capture and preserve the identity of each human face in images while reproducing the missing irregular region in images. A two-stage cascaded model is proposed. It is composed of a shape-predictor of the key-points of the face followed by an inpainting network. The shape-predictor identifies the human face's structure-preserving its local points, i.e. eyes, mouth, nose, and then the inpainting network fills any random-irregular missing regions guided by the obtained knowledge as a priori. The effectiveness of the proposed model was evaluated using the CelebA dataset. The obtained results from the trained model outperform the recently proposed technique with contextual attention.

Keywords— cnn, encoder-decoder, gan, face inpainting, l1 loss, adv. Loss.

I. Introduction

The main reason behind the significant improvement of deep learning methods in image inpainting is the employment of Generative Adversarial Networks (GANs). With the increase of communication through social media, lots of images are exchanged and there are some doubts concerning the reality of the generated images. These doubts have been studied by social media analysts and therefore, many face editing programs are produced. However, they lack the expert-eye of what tools to be used in certain circumstances such as red-eye fixing [1], blemish removal [2].

In recent works, employing deep learning image inpainting methods have been successfully used to reproduce the missing regions in images. Early traditional methods used a center-square mask for the missing region and restored it using context-encoders [3]. Then global and local discriminators were employed for more realistic results [4-5]. The main drawbacks of these methods were the presence of artifacts around the generated region and the tendency to work with low-resolution images. There were several attempts for solving these limitations such as partial convolutions presented in [6] which worked with irregular random masks, Deepfillv2 [7] which worked with a utilized user's sketch as an input, guided inpainting [8] which used another image as a guide for the synthesis process and eye in-painting [9] which focused on eye-inpainting relying on another image of the same person as a reference. The problem of eye in-painting lies in its need for a special type of datasets during training. Although the mentioned deep learning techniques produce good recovered images these methods don't preserve the identity of faces. For example, when inpainting a missing eye, the network will insert an eye corresponding to similar faces obtained from the training set leading to unpleasant results.

To overcome these drawbacks, a new model for face inpainting is proposed here. The proposed model composed of two-cascaded networks:

1. Face-shape predictor: To generate and train the network with the Histogram of Oriented Gradients (HOG). This is the guidance network.
2. Image inpainting network: To generate the missing regions guided with the HOG features in the first stage.

Both networks employ encoder-decoder architecture. This architecture is easier to implement and faster in training. To the best of the authors' knowledge, we are the first to use HOG features as guidance for better inpainting results.

Face-shape predictor network: In this network, HOG [10-12] are generated through the model for the training dataset and the network is trained to hallucinate the missing regions in the structure of images. Image structure recovery is a much simpler task than image inpainting.

The feature descriptor HOG is trained to extract features from the input image. HOG not only captures the localized structure of objects but also provides the structure direction as well, i.e. gradient and orientation. The orientations are computed in a localized patch-wise manner since each image is divided into smaller patches. For each patch, gradients and orientations are computed. HOG is also capable of generating a pixel-wise value of the computed gradients and orientations.

HOG is computed through five stages:

1. Global image normalization: The application of global normalization equalization to decrease the effect of illumination.
2. Computing image gradients: The computation of first-order gradients such as contour and some texture information.
3. Computation of gradient histogram: The production of structure-encoding of image content.
4. Normalization across blocks: The normalization across cells, i.e. the cell is a small spatial region of the image window. This normalized block is referred to as a histogram of oriented gradients.

978-1-7281-6535-6/20/$31.00 ©2020 IEEE

5. Flattening into a feature vector: The collection of all HOG descriptors from all blocks to produce the final structured image.

The face-shape predictor network is an essential network for an inpainting network guiding inpainting, i.e. provides a road-map of the human face structure which is essential for preserving human face identity while inpainting of the missing regions, as it captures the local structure in human faces.

Image inpainting network: This network is responsible for the inpainting of the missing irregular regions guided with the local structures obtained from the first network. Both network stages employ adversarial model [13].

II. Related Work

Image inpainting [14] usually refers to the art of restoring the missing region in images. Due to its high importance, it motivated a lot of researchers to investigate and search for solutions. Image inpainting techniques can be categorized into three main categories, namely, diffusion-based, patch-based and deep learning-based techniques.

Diffusion-based techniques are the oldest inpainting techniques. These algorithms mainly depend on using the variational method and the Partial Differential Equation (PDE) [15-17]. The pioneer for that field was presented in [16]. The authors produced a model employing non-linear PDEs trained to resemble the technique used by artists specialized in paintings' restoration used in museums. The authors in [17] used an energy minimization technique for computing the inpainted recovered image though a coupled non-linear differential equation. The authors in [18] observed the link between isophote direction and the Navier-Stokes equation. This observation inspired them to propose a solution using a transport equation for the missing domain filling. Diffusion-based techniques were suitable for small missing regions such as scratches.

Patch-based techniques divide the missing region into patches. For each missing patch, the best candidate patch from the surrounding neighborhood patches is selected for filling the missing patch. A patch-based searching technique was presented in [19]. The best-matched patch is selected based on Markov Random Field (MRF). Another technique was presented in [20]. The patch is selected based on annihilation property filter and low rank structured matrix. These techniques are suitable for filling non-complex similar texture images.

Deep learning-based techniques have given a much more accurate and impressive solution for image inpainting problem. Due to its superior results, many kinds of research are inspired to update, use or even implement new approaches. Deep learning approaches' fundamental idea is to train Convolution Neural Networks (CNNs) using a training set of images for inpainting the missing areas/regions in images afterward. This is the most straightforward solution for tackling image inpainting by training a dedicated model for the missing hole shape.

These CNNs always require an input pair composed of the real image, i.e. also known as Ground Truth image (GT), and masked image, i.e. image with the missing area. The training operation begins by training the network to complete that missing region then hopefully the trained model could be used for any other images, i.e. not included in the training set, or even generally for images with different missing areas in terms of shape and size of that missing hole.

Generative Adversarial Networks (GANs) also helped in the impressive inpainting results of deep learning-based. It was first proposed for learning image transforms between two datasets [21-22]. In [21], Pix2Pix was proposed. Pix2Pix employs a training set composed of aligned image pairs used for creating models. These models are capable of converting labels into the real image, converting a sketch to its corresponding real image or converting a black and white image into a colored one. The only limitation for this method lies in its need to have both real and target images paired together in the training dataset to learn such a transformation.

In [22], CycleGAN proposed a modified method that no longer requires pairing images in the training dataset. Without a target image, the authors instead used a virtual result in the target domain while converting the image to its original image domain. If this virtual image is inverted again, it must result in the same original image. This requires two generators for the conversion task. This method showed superior outcomes as it targets the learning of the inverse forward mapping.

III. Network Architecture

The overall network architecture follows an encoder-decoder architecture [23] which showed impressive results for style transfer, super-resolution [24] and image-to-image translation [22]. For the Generators, both generators composed of encoders and decoders. The encoders down samples twice the input using two stride convolutions followed by eight residual blocks [25], i.e. to help to minimize the training degradation problem, employing dilated convolutions [26] with dilation factor equals to two. Both decoders perform the reverse upsampling images to its original resolution using transposed convolutions. Rectified Linear Units (ReLU) is used for all layers of the generator. Instance normalization [27] is used across all network layers leading to impressive improvements to image stylization's quality.

PatchGAN [28] was employed for both discriminators, i.e. Non-Saturating GAN (NSGAN) version is used for faster convergence. The discriminator architecture is composed of five convolution layers. Each with kernel size 4×4, following

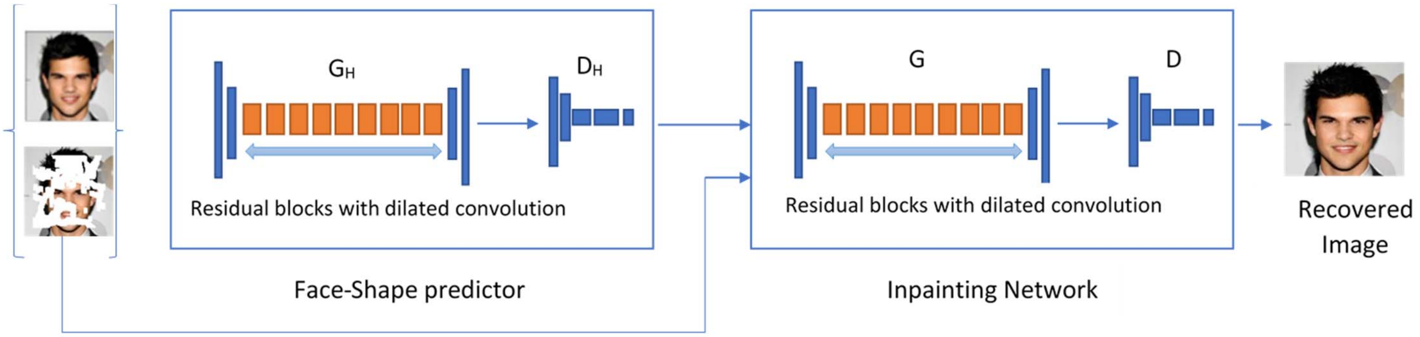
GH
DH
Residual blocks with dilated convolution
Face-Shape predictor
G
D
Residual blocks with dilated convolution
Inpainting Network
Recovered Image

Fig. 1. Proposed Network Architecture. The face-shape network is trained with L1FS and advFS Loss. The inpainting network is trained with L1, Lprec, Lstyle and adv loss.

each convolution layer a Leaky Rectified Linear Unit (Leaky ReLU) with slope α = 0.2. The final layer is responsible for the decision about whether 70×70 overlapping image patches are real or fake. The overall network architecture is shown in Fig. 1.

The input for the Face-Shape predictor is an input-pair composed of the ground truth (GT) image and the masked image. while the input for the Inpainting Network is only the masked image along with the guidance from the first network. The generator and discriminator for the face-shape predictor network are denoted by G_H and D_H while the generator and discriminator for the inpainting network are denoted by G and D respectively.

IV. Loss Functions

For the face-shape predictor network, a joint loss function composed of two-loss measures; $L1_{FS}$ reconstruction loss and an adversarial loss (L_{adv1}). This network takes an input-pair composed of ground truth image I_{gt} and masked image $I_m = I_{gt} \odot (1\text{-}M)$ where M is the binary mask and $\odot$ is the element-wise product operation, i.e. 1 for missing regions and 0 otherwise. The network generator G_H produces the HOG features through training and predicts the missing region $I_{pred}=G_H(I_m, I_{gt})$. Afterward, the discriminator role to determine whether the predicted HOG image from GH is real or fake.

$L1_{FS}$ can be computed from the following Equation:

$$L1_{FS}= \|I_{pred}\text{-} I_{gt}\| \quad (1)$$

And, L_{adv1} can be defined by the following Equation:

$$L_{adv1}= E_{(I_{gt},I_{pred})}[\log D(I_{GT},I_{pred})]+E_{(I_{pred})}[1\text{-}D(I_{pred},I_{gt})] \quad (2)$$

Therefore, the joint loss function can be expressed in the following Equation:

$$L_{FS} = \lambda_1 L1_{FS}+\lambda_{adv1} L_{adv1} \quad (3)$$

Where λ_1 and λ_{adv1} are hyper-parameters with ratio 1:1, respectively.

For the inpainting network, also a joint loss function was employed through the training phase. It is composed of four loss measures L1 reconstruction loss, adversarial loss (L_{adv}), perceptual loss and finally a style loss. The input for this network composed of an input pair composed of I_{pred} from the first model along with the masked images I_m. The output of the network is a recovered inpainted image I_o with the same resolution as the real image. L1 reconstruction loss is calculated according to Equation (4) which is similar to Equation (1):

$$L1= \|I_o\text{-} I_{gt}\| \quad (4)$$

While L_{adv} is also calculated according to Equation (5) which is similar to Equation (2):

$$L_{adv}= E_{(I_{gt},I_o)}[\log D(I_{GT},I_o)]+E_{(I_{pred})}[1\text{-}D(I_o,I_{gt})] \quad (5)$$

Furthermore, two more losses are included for the generator G, i.e. Perceptual loss [23] and Style loss. These two loss functions were first presented in [6] and have been widely used since then. Perceptual loss [23] also computes L1 reconstruction loss, but with projecting images into feature spaces of VGG-16 [29] model trained on ImageNet. It is defined by the following Equation:

$$L_{prec}= E[\textstyle\sum_k \frac{\|\varphi_k(I_{gt})\text{-}\varphi_k(I_o)\|_1}{N_{\varphi_k(I_{gt})}}] \quad (6)$$

Where φ_k is the activation feature map for the kth layer of the pre-trained model, i.e. VGG-16 [29]. For that purpose, we used layers relu1-1, relu2-1, relu3-1 and relu4-1 similar used in [6]. Style loss also works on the same activation layers of the VGG-16 [29] model. It measures the similarity between the two images. It can be defined by the following Equation:

$$L_{style}= \textstyle\sum_k \frac{1}{C_k C_k} \left\| \frac{G_k(I_o)\text{-}G_k(I_{gt})}{N_k} \right\|_1 \quad (7)$$

Where, $G_k(x)= (\varphi_k(x))^T(\varphi\text{-}k(x))$ is the Gram Matrix. The Gram Matrix has the shape of $C_k \times C_k$ as the feature map has the dimension of $H_k \times W_k \times C_k$ The loss style provides an autocorrelation on each used layer of VGG-16. This loss proved its importance in recent work [30]. It is used as an accurate tool for reducing "checkerboard" artifacts caused by transpose convolution layers [31].

Our overall loss function used is defined by the following Equation:

$$L_{overall}=\eta_l L_1+\eta_{adv} L_{adv}+\eta_{prec} L_{prec}+\eta_{style} L_{style} \quad (8)$$

Where, η_l, η_{adv}, and η_{style} are hyper-parameters. Through training, the values for the hyperparameters were as follows: $\eta_l=1$, $\eta_{adv}=\eta_{prec}=0.1$ and $\eta_{style}=250$.

V. Training data

The model is trained using the CelebA-img-algin dataset [32] with an irregular-mask dataset proposed in [6]. CelebA dataset is used after several pre-processing steps. Initially, all images are resized into 256×256 then 65,534 images are selected for training and 20000 images for each of the validation and testing respectively. The irregular mask dataset has 12000 masks. All mask images are also resized to 256×256. The initial results obtained from the model are shown in the following Fig. 2. During testing, only the masked image is used as an input along with the mask to produce the output recovered image. Further qualitative study is performed with another recent state-of-the-art. Extra Results are found in Appendix A.

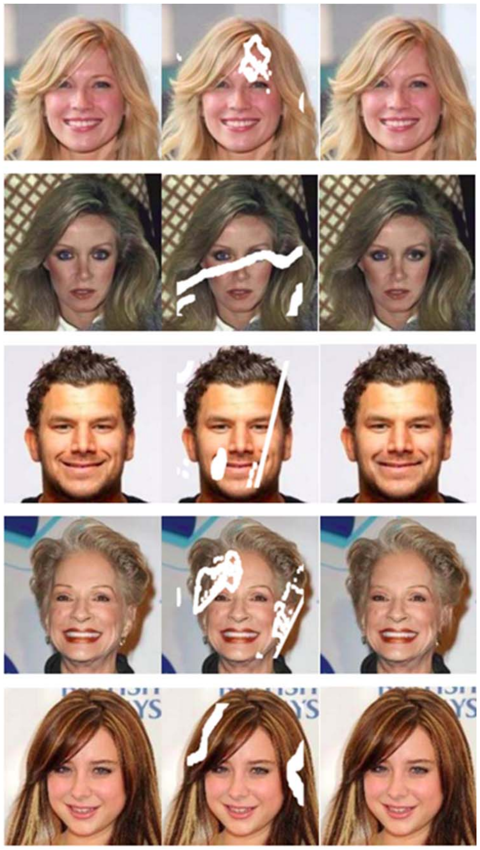

Fig. 2. Results from the trained model. Ground truth image on the left, masked image on the middle and recovered image on the right. During testing, only the masked image is used as an input to produce the recovered inpainted image.

VI. Qualitative Comparison

To prove the effectiveness of the trained model, a qualitative comparison is performed against the recent work of Deepfillv1 [5], i.e. Deepfillv1 test system had been published. Our guided trained model shows a superior recovered image especially in preserving the face identity as well as looking realistic as shown in Fig. 3.

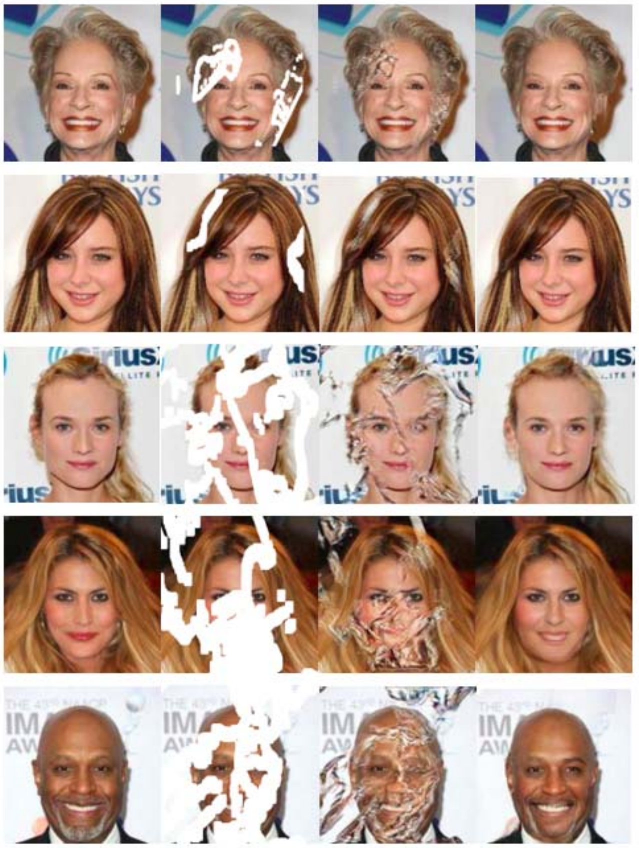

Fig. 3. Qualitative comparisons with Deepfillv1 on the Celeba dataset. it is organized form the left: GT image, masked image, Deepfillv1 results, and our results.

VII. Conclusions and Future Work

A novel face inpainting system is presented on this paper based on a two-step training procedure. The network is composed of two main networks, the first network is responsible for guiding through generating the HOG features for training images. the second inpainting network considers the features extracted and employ these features for better inpainting results. Both networks follow an adversarial model for generating more realistic images. The trained model not only shows a coherent-consistent realistic inpainted image but also preserved and maintained the identity of the human face in images. the model is trained using the CelebA dataset. We plan to investigate other features extractor algorithms and how to extend the network architecture to deal with high-resolution images.

REFERENCES

[1] Z. Zhu, P. Luo, X. Wang and X. Tang, "Multi-View Perceptron: a Deep Model for Learning Face Identity and View," in Advances in Neural Information Processing Systems, 2014.

[2] M. Brand and P. Pletscher, "A conditional Random Field for Automatic Photo Editing," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[3] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell and A. A. Efros, "Context Encoders: Feature Learning by Inpainting," in arXiv preprint arXiv:1604.07379v2, 2016.

[4] S. Iizuka, E. Simo-Serra and H. Ishikawa, "Globally and Locally Consistent Image Completion," in ACM Transactions on Graphics (Proc. SIGGRAPH), Volume 36, Issue 4, 2017.

[5] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu and T. S. Huang, "Generative Image Inpainting with Contextual Attention," in arXiv preprint arXiv:1801.07892v2, 2018.

[6] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao and B. Catanzaro, "Image Inpainting for Irregular Holes Using Partial Convolutions," in arXiv preprint arXiv:1804.07723v2, 2018.

[7] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu and T. S. Huang, "Free-form image inpainting with gated convolution," in arXiv preprint arXiv:1806.03589, 2018.

[8] Y. Zhao, B. Price, S. Cohen and D. Gurari, "Guided image inpainting: Replacing an image region by pulling content from another image," in arXiv preprint arXiv:1803.08435, 2018.

[9] B. Dolhansky and C. C. Ferrer, "Eye In-Painting with Exemplar Generative Adversarial Networks," in arXiv preprint arXiv:1712.03999, 2017.

[10] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," in International Journal of Computer Vision (2004) 60: 91, 2004.

[12] N. Dalal, "Finding People in Images and Videos," in Human-Computer Interaction [cs.HC], Institut National Polytechnique de Grenoble - INPG, 2006.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D.Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, pages 2672–2680, 2014.

[14] O. Elharrouss, N. Almaadeed, S. Al-Maadeed and Y. Akbari, "Image inpainting: A review," in arXiv preprint arXiv:1909.06399, 2019.

[15] C. Guillemot and O. L. Meur, "Image Inpainting: Overview and Recent Advances," in IEEE Signal Processing Magazine, Volume 31, Issue 1, 2014.

[16] M. Bertalmio, G. Sapiro, V. Caselles and C. Ballester, "Image Inpainting," in ACM Transactions on Graphics (Proceedings of SIGGRAPH), 2000.

[17] C. Ballester, M. Bertalmío, V. Caselles, G. Sapiro and J. Verder, "Filling-in by Joint Interpolation of Vector Fields and Gray Levels," in IEEE Transactions on Image Processing, Volume 10, Issue 8, 2001.

[18] M. Bertalmio and G. S. A. Bertozzi, "Navier-stokes, Fluid Dynamics, and image and Video Inpainting," in Proceeding of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001.

[19] T. Ružić and A. Pižurica, "Context-Aware Patch-Based Image Inpainting Using Markov Random Field Modeling," in IEEE Transactions on Image Processing, Volume 24, Issue 1, 2015.

[20] K. H. Jin and J. C. Ye, "Annihilating Filter-Based Low-Rank Hankel Matrix Approach for Image Inpainting," in IEEE Transactions on Image Processing, Volume 24, Issue 11, 2015.

[21] P. Isola, J.-Y. Zhu, T. Zho and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in CVPR, 2017.

[22] J.-Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in IEEE International Conference Computer Vision (ICCV), 2017.

[23] J. Johnson, A. Alahi and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," in arXiv preprint arXiv:1603.08155v1, 2016.

[24] M. W. Gondal, B. Scholkopf and M. Hirsch, "The unreasonable effectiveness of texture transfer for single image super-resolution," in Workshop and Challenge on Perceptual Image Restoration and Manipulation (PIRM) at the 15th European Conference on Computer Vision (ECCV), 2018.

[25] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

[26] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," in arXiv preprint arXiv:1511.07122, 2016.

[27] D. Ulyanov, A. Vedaldi and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feedforward stylization and texture synthesis," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[28] T.Miyato, T. Kataoka, M. Koyama and Y. Yoshida, "Spectral normalization for generative adversarial networks," in International Conference on Learning Representations, 2018.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein and e. al, "Imagenet Large Scale Visual Recognition Challenge," in International Journal of Computer Vision, 115(3):211–252, 2015.

[30] M. S. M. Sajjadi, B. Scholkopf and M. Hirsch, "Enhancenet: Single Image Super-Resolution Through Automated Texture Synthesis," in The IEEE International Conference on Computer Vision (ICCV), pages 4501–4510, 2017.

[31] A. Odena, V. Dumoulin and C. Olah, "Deconvolution and Checkerboard Artifacts," in Distill, 1(10):e3, 2015.

[32] Z. Liu, P. Luo, X. Wang and X. Tang, "Deep learning face attributes in the wild," in Proceedings of International Conference on Computer Vision (ICCV), 2015.

APPENDIX A