ISSN 1213-6972 Journal of WSCG http://www.wscg.eu Vol.28, No.1-2, 2020

Hallucinating Very Low Resolution Face Images to 16x magnification with Age based Attributes

Jihwan Kim
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
okpo65@gmail.com

Sunghee Choi
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
sunghee@kaist.edu

ABSTRACT

Face hallucination is a type of super resolution that restores very low resolution (8×8 pixel) to high resolution (128×128 pixel) face images. Since unique facial features caused by age, e.g.wrinkles, are ignored during restoration, restored face images can be somewhat dissimilar to the original faces, particularly for older people. To solve this problem, we construct a pipeline network to restore face images more realistically by including age attribute, predicted from the low resolution image. Predicted age attribute is divided into young and old groups, where the aging network is the last pipeline stage and only applied when the original face image includes old age attributes. Thus, older people tend to be restored with wrinkles and features similar to their original appearance. Restored images are compared qualitatively and quantitatively with images created by existing methods. We show that the proposed method maintains and restores age related personality features, such as wrinkles, producing higher structural similarity index than other methods.

Keywords

Face hallucination, Pipeline network, Age, Personality, Deep learning

1 INTRODUCTION

Face hallucination (FH) restores a low resolution face image (LR) to high resolution (HR), and is particularly important for face recognition and restoration systems, such as surveillance or CCTV [HHST17]. Generally, FH supports 16 scale restoration for 8×8 pixel and 8 scale restoration for 16×16 pixels. Larger images (16×16 pixel) have considerably more information compared to 8×8 pixel images, which allows employing prior information, such as heatmaps, during their restoration. In contrast, 8×8 pixel images have limited information, which makes restoration difficult. Figure 1 shows that it is difficult to distinguish face shape at all for 8×8 pixel images, whereas 16×16 pixel images can be divided into eye, nose, mouth, and face regions by eye. This paper considers the more challenging 8×8 pixel image restoration.

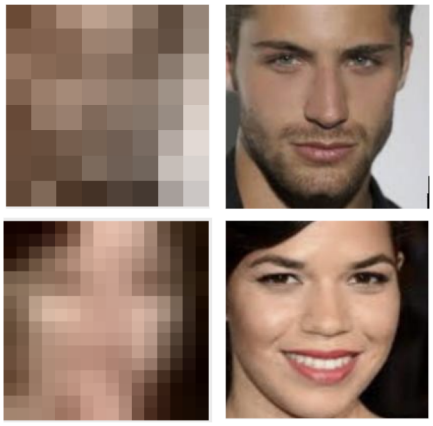

Figure 1: Typical 8×8 and 16×16 pixel images (left) and their corresponding ground truth images (right)

Several methods have been proposed previously to implement FH. For example, Dahl et al. [DNS17] used pixelCNN, Yu et al. [yu2018face] used auto-encoder, Chen et al. [CTL+18] used GAN, and Huang et al. [HHST17] used the wavelet transform. However, these methods have ill-posed problems, producing many HR solutions for a single LR face image. Lower image quality increases restoration difficulty because more information regarding face feature and personality details are lost. Consequently, we cannot solve the FH one-

Figure 2: Restoration from 8×8 to 128×128 pixel images using a wavelet based convolutional network approach [HHST17]

to-many problem between LR and possibly many HRs, even if we use large training datasets [YFHP18]. Therefore, this paper proposes a pipeline based network that includes facial age attributes to mitigate the ill-posed FH problem. Very low resolution images (8×8 pixel) have lost considerable facial information compared to their ground truth images, and it is difficult to restore precisely because learning ignores age related facial features (e.g. wrinkles). Figure 2 shows some typical example restored images from 8×8 pixel input images, which all have critical missing details, such as wrinkles. Hence the restored images look much younger than their ground truth. Figure 3 shows that the structural similarity index (SSIM), a measure of image similarity to ground truth, decreases relatively linearly with increasing subject (and hence face image) age, being considerably lower for 80 than 20 year old subjects. To solve this problem, we create a pipeline network learning model that can restore face images including age attributes by including age information in the pipeline network to help produce an appropriate facial image for their age.

The proposed pipeline network incorporates three stages.

1. Age estimation. We apply an Wide ResNet CNN to estimate age [RTVG15] from the original low resolution image. Young and old attributes are determined by this predicted age.

2. First restoration. 8×8 pixel images are restored to 128×128 pixel images using a wavelet based convolutional network (CNN) approach following [HHST17].

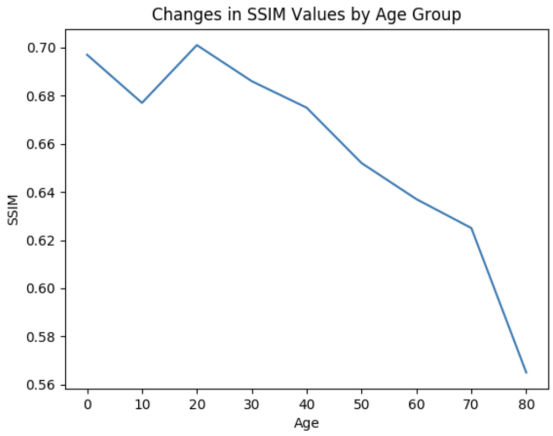

Figure 3: Structural similarity index (SSIM) for restored images with respect to subject age

3. Apply age attributes. We apply an aging general adversarial network (GAN) [ZSQ17] only for images with known age attributes to include subject's age related facial features (e.g. wrinkles).

We assume that restoring face image with facial features like wrinkles mitigates ill-posed problems by making it look similar to the original. Experimental results demonstrate that our pipeline network restores much more similarly to their original images especially for older people than other methods.

2 RELATED WORK

Many methods have been proposed for FH to solve the ill-posed problem. Yu et al. [YFG+18] restored facial images using a heatmap of the face, i.e., structural information, to solve the ill-posed problem caused by pose variation or misalignment. Chen et al. [CTL+18] used a super resolution encoder-decoder structure to generate facial images, creating more realistic images by employing a GAN structure with loss function being the difference between ground truth and generated images. Huang et al. [HHST17] converted the super resolution process to wavelet coefficient prediction for high resolution images. The high resolution image was then reconstructed using the predicted wavelet coefficients, considering texture and topology information from the original low resolution image. In addition to restoring low to high resolution images, recognition algorithms have been proposed to distinguish individuals by including facial features, such as eigenfaces [HYBK08].

Dahl et al. [DNS17] used pixelCNN to restore very low resolution images (8×8 pixel), employing a stochastic model network rather than linear interpolation to predict each pixel value in the high resolution images. However, each pixel RGB value must be separately trained in the model, requiring very high computational complexity. Therefore, restoring for scales above 4 is

very difficult. Yu et al. [YFHP18] proposed an FH method by adding facial attributes to maintain face detail. After restoring the low resolution image by interpolation, a facial attribute vector was inserted into the residual block and then trained. However, they only supported 8 scale restoration from 16×16 pixels, and restored the image using a random attribute rather than actual face image attributes. In contrast, the proposed approach supports 16 scale reconstruction for the challenging 8×8 pixel input resolution. We obtain the age attribute from the low resolution image and then use this in the pipeline to restore a facial image more likely reflecting the subject's own personality.

Tamura et al. [TKM96] proposed a neural network that identifies gender with more than 90% accuracy at 8×8 pixels image, where the network learned image gender using face shapes, cheekbone shapes, and shading. Wang et al. [WCY+16] showed that recognition performance from low resolution images is much lower than ground truth images for face, digit and font recognition.

A key factor in identifying faces is the general patterns of aging facial marks (e.g. wrinkles). Jain et al. [JP09] attempted to increase recognition accuracy by detecting facial-marks such as wrinkles. Ling et al. [LSRJ07] showed that recognition error was higher when there was a large gap between actual and expected age.

3 APPROACH

Overview Structure

Figure 4 shows three stages of our pipeline network. Stage 1 (age estimation) predicts age attribute directly for the low resolution image (8×8 pixel), using an age classification network [RTVG15] with Wide ResNet [ZK16] structure. We only assigned age attributes to the image if estimated age is above 40. Stage 2 (restoration) uses a wavelet based CNN [HHST17] to restore the facial image. Finally, stage 3 (aging) applies the aging network [ZSQ17] for images with predicted age above 40. Predicted age from stage 1 and the restored image from stage 2 are input to the age GAN to produce the final restored image.

Dataset

We use the UTKFace dataset [ZSQ17] with age, gender, and race labels comprising 23,704 images, all aligned and cropped. This dataset has been used in many fields, including face detection, age progression/regression, and age estimation. However, more than half of this dataset subjects are in their 20-30s, which could produce age bias when estimating subject age for FH. Therefore, we constructed an age-balanced UTKFace dataset, with even distribution between young and old groups (less and more than 40 years old, respectively) by randomly removing some of the young group images. The age-balanced UTKFace dataset includes 13,656 images with 6,591 young and 7,065 older group images. Figure 5 compares the original and age-balanced UTKFace dataset distributions. The age-balance case is clearly more evenly distributed across the identified subject ages. Both datasets are employed separately for training and their results compared.

Age Estimation

The first stage in the pipeline estimates subject age from the low resolution images using Wide ResNet CNN [ZK16] and the UTKFace dataset. The age classification network comes from [RTVG15]. Wide ResNet solves the neural network depth problem by modifying the ResNet structure and increasing the number of CNN output channels rather than the number of CNN layers. We use the UTKFace images, scaled down to 8×8 pixel. First, the low resolution image is converted to 64×64 pixels by bicubic interpolation. Then image features are extracted using Wide ResNet CNN, and age values were extracted using the softmax activation function. We add data augmentation and dropout to improve estimation accuracy.

After predicting the age, we divide the dataset into young and old groups around the 40 year cutoff based on the predicted age. The aging network, in the final pipeline stage is only applied to images in the older group. Table 1 compares actual and predicted young and old attributes. Accuracy for subjects in their 40s and 50s is relatively low, since they are close to the cutoff, but overall matching rate is acceptable. The hit ratio are more important for 60+ aged subjects, where facial features such as wrinkles are very common. In particular, subjects in their 80s are quite accurately classified (80%).

type	hit	total	ratio
0s	2,967	3,059	96.99
10s	1,392	1,531	90.92
20s	6,712	7,343	91.41
30s	3,533	4,537	77.87
40s	857	2,245	38.17
50s	1,337	2,299	58.16
60s	893	1,318	**67.75**
70s	524	699	**74.96**
80s	407	504	**80.75**
all	18,749	23,704	79.1

Table 1: Age classification accuracy. (Notes: Hit = when predicted and actual age bracket agree.)

Wavelet based CNN

The second stage in the pipeline uses a wavelet based CNN [HHST17] to restore low resolution (8×8 pixel) to high resolution face (128×128 pixel) images. Rather

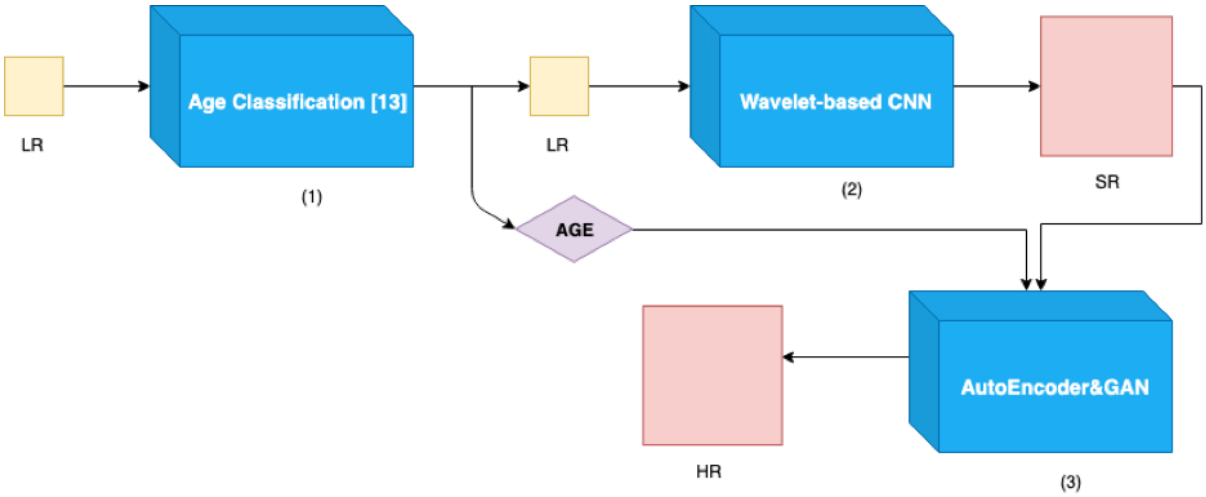

Figure 4: Proposed pipeline network

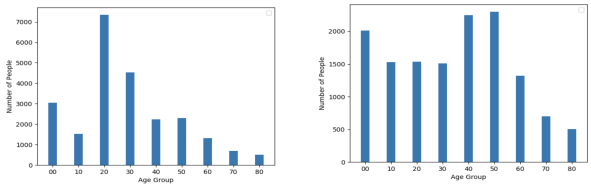

Figure 5: Population distribution by age group for (left) original UTKFace and (right) age-balanced UTKFace datasets

than simply reconstructing the high resolution image, wavelet coefficients for the high resolution image are predicted and restored while maintaining global topology and facial structural textual information.

Embedding Net

This is the process of mapping the low resolution image $(3 \times w \times h)$ to the feature map. We use 3×3 filters at 1 stride and output them to the final output $(N_e \times h \times w)$ without going through upsampling and downsampling. (N_e is the channel size of the last layer)

Wavelet Prediction Net

If n is the level of wavelet packet decomposition, then magnification r is $r = 2^n$ called the scaling factor and N_w is 4^n as the number of wavelet coefficients. The level of packet decomposition determines the scaling factor and the number of wavelet coefficients. Wavelet prediction net is composed with N_w parallel independent subnets that takes the output from the embedding net as input and predicts wavelet coefficients for high-resolution images. Each subnet learns the same image size as the embedding net used 3×3 filters with 1 stride ($3wh$). Predicting the wavelet coefficient values for all subnets will result in a wavelet coefficient of $N_w \times 3 \times w \times h$. Let $C = (c_1, c_2, ..., c_{N_w})$, $\hat{C} = (\hat{c}_1, \hat{c}_2, ..., \hat{c}_{N_w})$ be the ground-truth wavelet coefficients, predicted wavelet coefficients, respectively, and let $W = (\lambda_1, \lambda_2, ..., \lambda_{N_w})$ be the weight matrix to prioritize which wavelet coefficients are relatively important while restoration. As you can see the loss function below, wavelet loss function attempts to minimize the MSE loss of each wavelet coefficients.

$$l_{wavelet}(\hat{C}, C) = \|W^{1/2} \odot (\hat{C} - C)\|_F^2$$

$$= \sum_{i=1}^{N_w} \lambda_i \|\hat{c}_i - c_i\|_F^2$$

Reconstruction Net

The reconstruction net converts the full size of the wavelet image, $N_w \times 3 \times w \times h$, to the existing image size of $3 \times (r \times w) \times (r \times h)$. The deconvolution layer has a $r \times r$ filter and proceeds to r stride. The formula below shows the overall change in image resolution while learning with wavelet-based CNN. As mentioned before, embedding net maps low resolution images $(3 \times w \times h)$ to feature maps $(N_e \times h \times w)$. The feature maps are then divided into N_w wavelet coefficient images $(3 \times w \times h)$ in the wavelet prediction net. Since N_w is r^2, we can adjust the resolution of high resolution images $(3 \times (r \times h) \times (r \times w))$. The definition of wavelet based CNN networks [HHST17] can be formulated as follows:

$$\Psi : R^{3 \times h \times w} \rightarrow R^{N_e \times h \times w}$$

$$\varphi_i : R^{N_e \times h \times w} \rightarrow R^{3 \times h \times w}, i = 1, 2, ..., N_w$$

$$\phi : R^{N_w \times 3 \times h \times w} \rightarrow R^{3 \times (r \times h) \times (r \times w)}$$

Aging network

The aging network [ZSQ17] includes an encoder to convert face images into a latent vector and a GAN to generate face images from the vector. The encoder maps all images to latent vectors and collects them to form a latent space with age on the horizontal axis and personality on the vertical axis. We predict the latent

vector for the age progression/regression image in this latent space by moving the horizontal axis, and then put this latent vector into the GAN to generate the aged image. Traditional GAN uses latent vectors as random sampling, producing blurry images. However, this network creates facial images that retain specific personalities by inserting the latent vector that encodes the actual image into the GAN.

Discriminator on z

The discriminator on z (D_z) is trained to construct a prior distribution (uniform distribution) when mapping latent vectors z from the encoder into latent space. D_z is trained to discriminate z from the encoder and random sampling z^* from prior distributions. It selects a uniform distribution as prior and builds a competition with the encoder to ensure z is evenly distributed in latent space.

Discriminator on Face Image

The discriminator on facial images (D_{image}) is designed to produce realistic images similar to the discriminator used in traditional GANs. The generator takes a latent vector as input and created facial images to trick the discriminator, and the discriminator is trained to distinguish this generated image from the actual image. D_{image} also includes an age label to create a more realistic image for older subjects.

4 EXPERIMENTS

The proposed pipeline network is supported by Ubuntu 16.04 LTS. All pipeline stages are implemented using tensorflow and pytorch. We also use Anaconda V3 because each pipeline step uses a different python version, and Anaconda makes it easy to build other environments by putting development environments in separate containers. We use the age-balanced UTKFace and UTKFace datasets for training, and put the age attribute through the pipeline to restore low resolution facial images to high resolution facial images, while maintaining their personality. If the subject's predicted age (from stage 1 of the pipeline) exceeded 40 years, the old age attribute is assigned, and only those images go through the aging network. The second pipeline stage primarily restore face images. The final pipeline stage (aging network) is only performed on facial images assigned the old age attribute in the first step.

Qualitative Result

Figure 6 compares the proposed pipeline results qualitatively with current best-practice methods that support 16 scale from 8×8 pixel images. The proposed pipeline results are more similar to ground-truth image compared with the other methods, more successfully restoring subject personality, such as age-related wrinkles. Bicubic interpolation and SRCNN [DLHT15] fail to restore face shape, and SRGAN [LTH+17] creates blurry facial images due to the unstable GAN. Wavelet based CNN [HHST17] largely restores facial texture, but the over-smoothing causes subjects to look much younger than their actual age.

Quantitative Result

Tables 2 and 3 compare SSIM by age group for the age-balanced UTKFace and UTKFace datasets, respectively. The proposed pipeline generally achieves higher SSIM than the other methods for subjects older than 40 years that had the age attribute set.

Age	SRCNN	SRGAN	Wavelet	Ours
0s	0.513	0.696	0.675	0.698
10s	0.447	0.653	0.651	0.664
20s	0.435	0.667	0.666	0.68
30s	0.442	0.664	0.652	0.667
40s	0.443	0.65	0.644	**0.659**
50s	0.435	0.632	0.625	**0.652**
60s	0.444	0.634	0.622	**0.652**
70s	0.43	0.609	0.598	**0.629**
80s	0.398	**0.579**	0.551	0.577

Table 2: Structural similarity index scores for the considered image restoration methods on the age-balanced UTKFace dataset

Age	SRCNN	SRGAN	Wavelet	Ours
0s	0.486	0.706	0.697	0.721
10s	0.418	0.66	0.677	0.69
20s	0.408	0.676	0.701	0.714
30s	0.41	0.671	0.686	0.698
40s	0.421	0.665	0.675	**0.684**
50s	0.418	0.649	0.652	**0.66**
60s	0.414	0.641	0.637	**0.642**
70s	0.417	0.623	0.625	**0.65**
80s	0.373	0.582	0.565	**0.612**

Table 3: Structural similarity index scores for the considered image restoration methods on the UTKFace dataset

5 CONCLUSION

We propose a pipeline network to restore very low resolution images to be similar to ground truth by including an age attribute. The proposed pipeline includes three stages: age estimation, wavelet based CNN, and aging network. We construct an age-balanced UTKFace dataset for training to avoid biases due to UTKFace dataset images being weighted toward a certain age.

We compare restored face images qualitatively with current best practice methods, and confirm that the proposed pipeline network produce facial images that restore subject personality due to age, such as wrinkles. Quantitatively, the proposed pipeline generally achieves superior SSIM scores than the other considered methods.

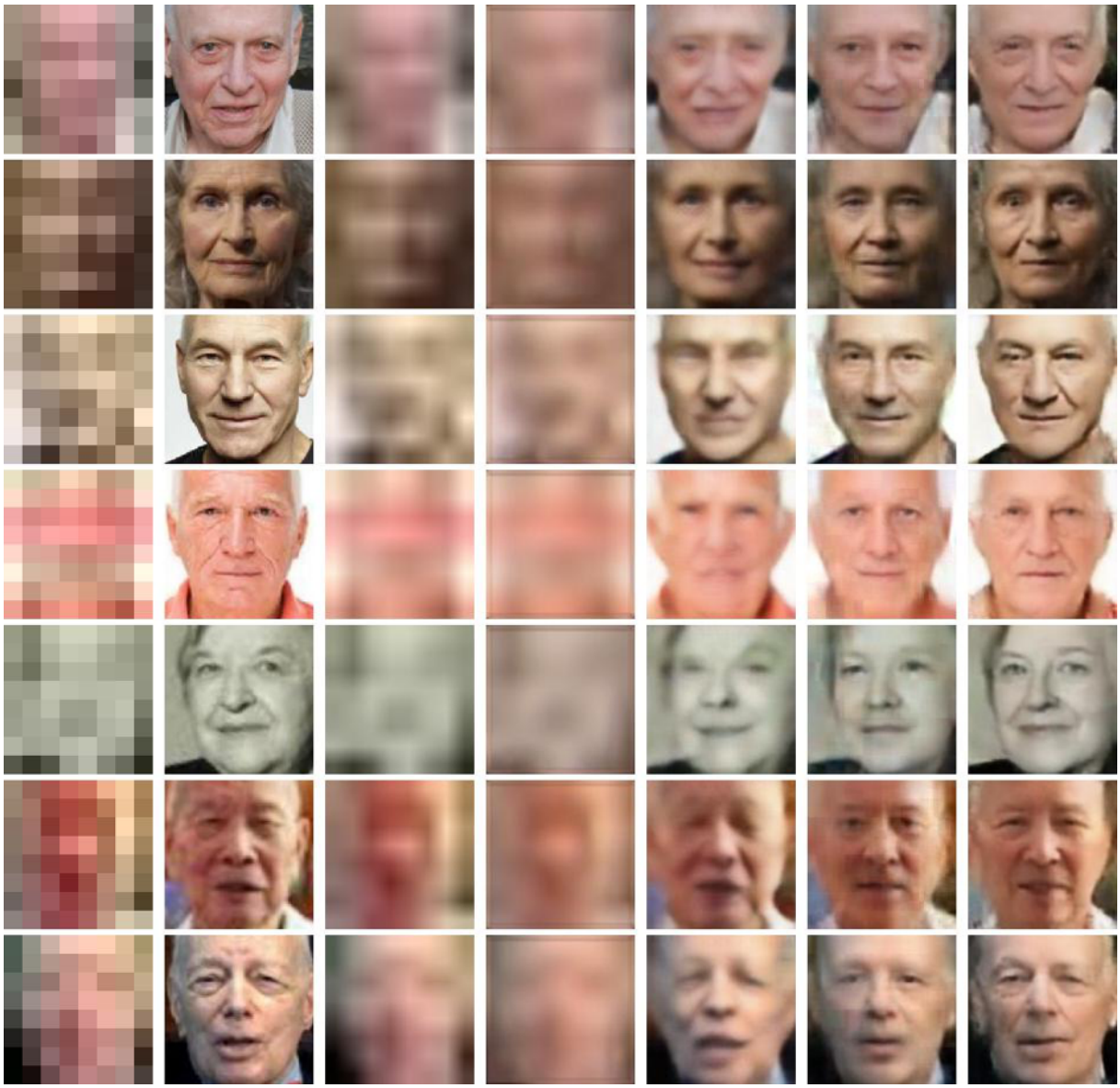

Figure 6: Image restoration outcomes for example 8×8 pixel input images. From left: input, ground truth, bicubic interpolation, SRCNN, SRGAN, Wavelet based CNN, proposed pipeline.

Future study will expand the image restoration pipeline to include not only age but also factors representing individual's facial features, such as gender, race, and facial shape. We will also reduce the pipeline process by restoring images with the age attribute in the upscaling process.

6 ACKNOWLEDGEMENTS

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-01158, Development of a Framework for 3D Geometric Model Processing)

7 REFERENCES

[CTL+18] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2492–2501, 2018.

[DLHT15] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[DNS17] Ryan Dahl, Mohammad Norouzi, and Jonathon Shlens. Pixel recursive super

resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5439–5448, 2017.

[HHST17] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1689–1697, 2017.

[HYBK08] Pablo H Hennings-Yeomans, Simon Baker, and BVK Vijaya Kumar. Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[JP09] Anil K Jain and Unsang Park. Facial marks: Soft biometric for face recognition. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 37–40. IEEE, 2009.

[LSRJ07] Haibin Ling, Stefano Soatto, Narayanan Ramanathan, and David W Jacobs. A study of face recognition as people age. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[LTH+17] 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[RTVG15] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 10–15, 2015.

[TKM96] Shinichi Tamura, Hideo Kawai, and Hiroshi Mitsumoto. Male/female identification from 8× 6 very low resolution face images by neural network. *Pattern recognition*, 29(2):331–335, 1996.

[WCY+16] Zhangyang Wang, Shiyu Chang, Yingzhen Yang, Ding Liu, and Thomas S Huang. Studying very low resolution recognition using deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4792–4800, 2016.

[YFG+18] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 217–233, 2018.

[YFHP18] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 908–917, 2018.

[ZK16] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.

[ZSQ17] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5810–5818, 2017.