Towards Coding for Human and Machine Vision: A Scalable Image Coding Approach

Supplementary Material

As supplementary material of our paper, we present the following contents:

- Detailed network architectures. (Fig. 1)
- Comparison with JPEG on human vision and machine vision tasks. (Figs. 2−3)
- Quantitative evaluation (user study). (Table 1)

1. Detailed network architectures

Our generator G utilizes the fully convolutional U-Net architecture as in pix2pix [1]. The discriminator D follows the SN-PatchGAN [3] for stable and fast training. The detailed network architectures are shown in Fig. 1, where "C" denotes a Convolution layer, "CSN" denotes a Convolution-SpectralNorm layer [2], "CB" denotes a Convolution-BatchNorm layer, the prefix "U" denotes an Upsampling layer, the suffix "R" and "LR" denote ReLU and LeakyReLU layers, respectively. We use "$k * k * c/s$" to indicate that the convolutional layer has c filters with a spatial size of $k * k$ and stride s.

Layer	Parameters
CLR	3*3*64/1
CBLR	3*3*128/2
CBLR	3*3*256/2
CBLR × 5	3*3*512/2
CR	4*4*512/2
UCBR × 5	3*3*512/2
UCBR	3*3*256/2
UCBR	3*3*128/2
UCBR	3*3*64/2
C+Tanh	3*3*3/1

G

Layer	Parameters
CSNLR	4*4*64/2
CSNLR	4*4*128/2
CSNLR	4*4*256/2
CSNLR × 2	4*4*512/2
CSNLR	4*4*512/1
CSN	4*4*64/1

D

skip connection

Figure 1. Overview of network architectures.

2. Comparisons with State-of-the-Art Methods

Visual quality evaluation. In Fig. 2, we present a visual comparison of the proposed method with JPEG compression under different quality parameters (qp).

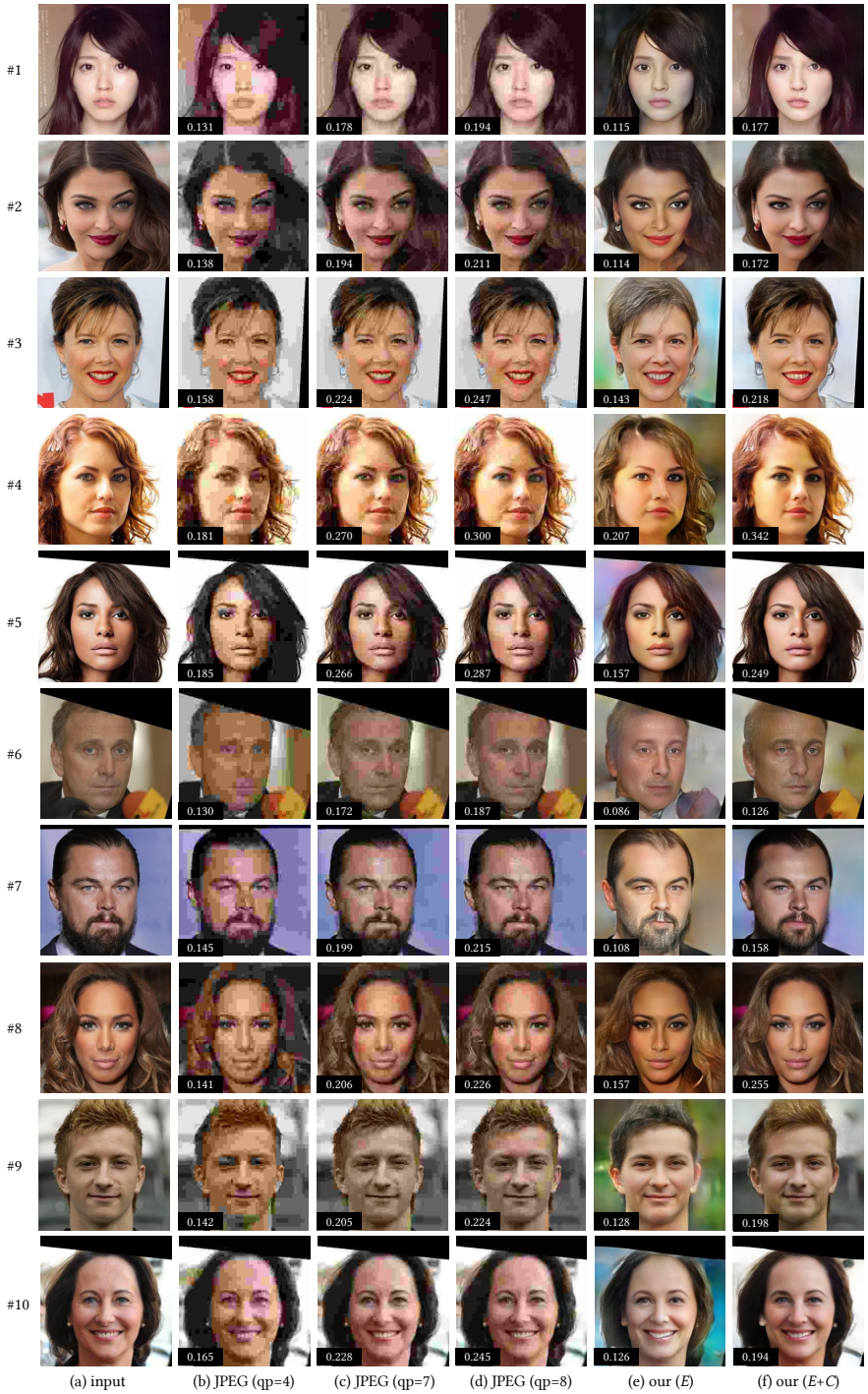

Figure 2. Visual comparison with JPEG compression. (a) Input image. (b)-(d) Images compressed by JPEG using quality parameter of 4, 7 and 8, respectively. (e) Our decoded images using the encoded edge representations. (f) Our decoded images using both the encoded edge representation and color representation. For each reconstructed image, its bit-rate (bit per pixel, bpp) is shown in the lower left black box.

Landmark detection. In Fig. 3, we present a comparison of the proposed method with JPEG compression under different quality parameters (qp) on facial landmark detection. We select 10 cases from the testing set with the detected landmarks plotted as white circles for better visual comparison.

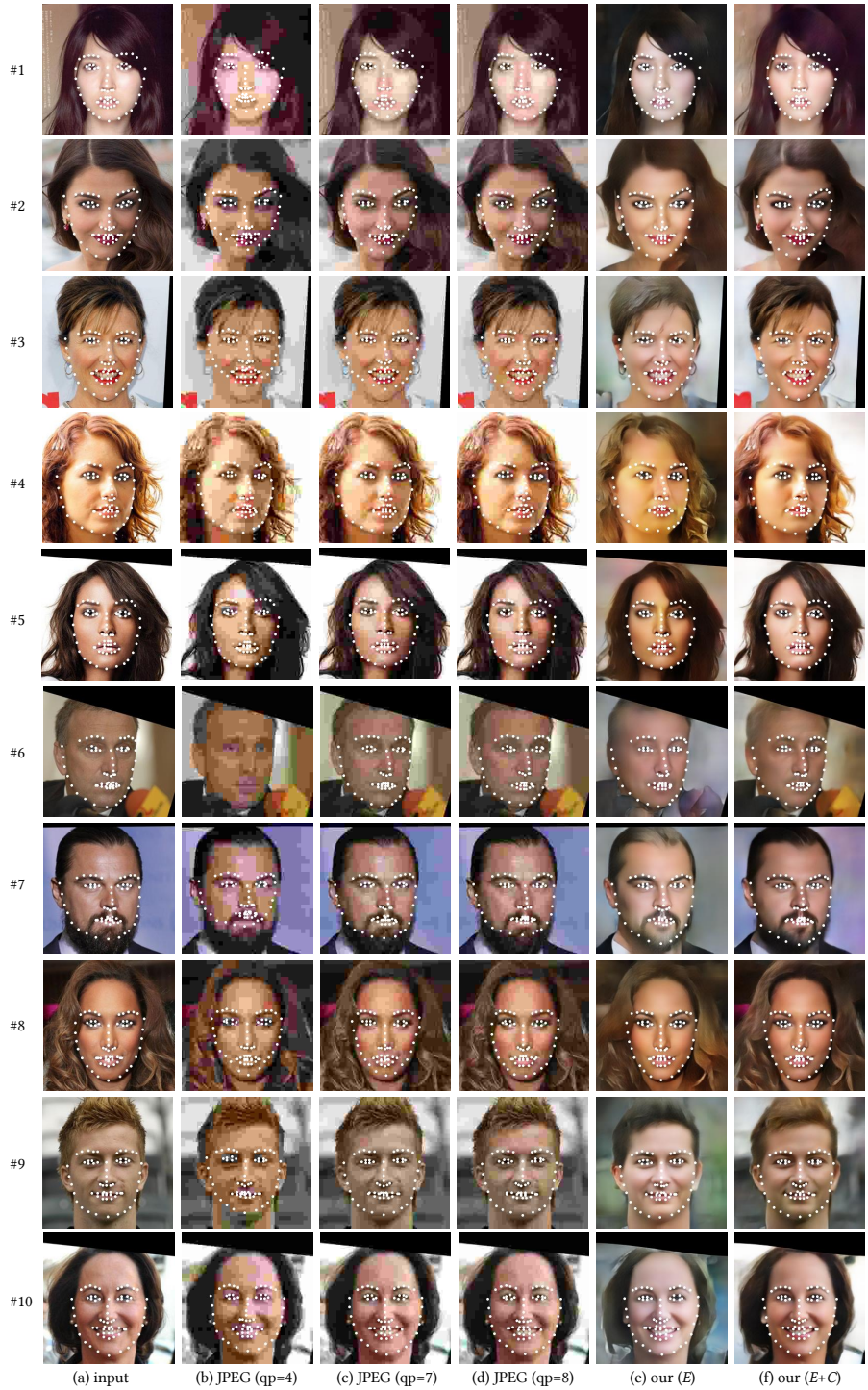

Figure 3. Comparison with JPEG compression on facial landmark detection. (a) Input image. (b)-(d) Images compressed by JPEG using quality parameter of 4, 7 and 8, respectively. (e) Our decoded images using the encoded edge representations. (f) Our decoded images using both the encoded edge representation and color representation. The detected landmarks are shown as white circles.

3. Quantitative Evaluation

To better understand the performance of the compared methods, we perform user studies for quantitative evaluations. Participants are shown the 10 cases in Figs. 2. Each subject is asked to select one from the five results that best matches the original image (**Fidelity**) and has the best visual quality (**Aesthetics**). To ensure the fairness, the orders of five results randomly change every round. A total of 10 subjects participate in this study and a total of 200 selections are tallied. The preference ratio is used as the evaluation metrics. It is calculated by:

$$\text{preference ratio of Method } A = \frac{\text{The total number of times Method } A \text{ was selected}}{\text{The total selection number}}. \quad (1)$$

According to the definition, if Method A performs significantly better than all other methods, its mean preference ratio can reach 1.0. As shown in Table 1, the proposed structure-color-hybrid method obtains the best average preference ratio of 0.90 and 0.73 for both the fidelity and aesthetics, respectively, outperforming JPEG compression under the similar bit-rate. The user study quantitatively verifies the superiority of our method.

Table 1. Quantitative evaluations

Fidelity					
ID	JPEG ($qp = 4$)	JPEG ($qp = 7$)	JPEG ($qp = 8$)	our (E)	our ($E + C$)
#1	0.00	0.10	0.00	0.00	**0.90**
#2	0.00	0.00	0.00	0.00	**1.00**
#3	0.00	0.00	0.00	0.00	**1.00**
#4	0.00	0.10	0.10	0.00	**0.80**
#5	0.00	0.00	0.10	0.20	**0.70**
#6	0.00	0.00	0.10	0.00	**0.90**
#7	0.00	0.00	0.10	0.00	**0.90**
#8	0.00	0.00	0.00	0.20	**0.80**
#9	0.00	0.00	0.00	0.00	**1.00**
#10	0.00	0.00	0.00	0.00	**1.00**
Average	0.00	0.02	0.04	0.04	**0.90**
Aesthetics					
ID	JPEG ($qp = 4$)	JPEG ($qp = 7$)	JPEG ($qp = 8$)	our (E)	our ($E + C$)
#1	0.00	0.00	0.10	0.40	**0.50**
#2	0.00	0.00	0.00	**0.50**	**0.50**
#3	0.00	0.00	0.00	0.00	**1.00**
#4	0.00	0.00	0.00	0.30	**0.70**
#5	0.00	0.10	0.00	0.00	**0.90**
#6	0.00	0.00	0.00	0.20	**0.80**
#7	0.00	0.00	0.00	0.20	**0.80**
#8	0.00	0.00	0.00	0.40	**0.60**
#9	0.00	0.00	0.00	0.20	**0.80**
#10	0.00	0.00	0.10	0.20	**0.70**
Average	0.00	0.01	0.02	0.24	**0.73**

References

[1] Phillip Isola, Jun Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 5967–5976, 2017. 2

[2] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *Proc. Int'l Conf. Learning Representations*, 2018. 2

[3] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proc. Int'l Conf. Computer Vision*, 2019. 2