Generating Diverse Structure for Image Inpainting With Hierarchical VQ-VAE (Supplementary Material)

Jialun Peng[1] Dong Liu[1] Songcen Xu[2] Houqiang Li[1]
[1] University of Science and Technology of China [2] Noah's Ark Lab, Huawei Technologies Co., Ltd.
pjl@mail.ustc.edu.cn, {dongeliu, lihq}@ustc.edu.cn, xusongcen@huawei.com

A. Architecture Hyperparameters and Training Details.

The hyperparameters used for training the hierarchical VQ-VAE are reported in Table 1. The hyperparameters used for training the diverse structure generator are reported in Table 2. As for the GAN-based texture generator, the hidden units of generator and discriminator are both 64. Our model is implemented in TensorFlow v1.12. Batch size is 8. We train the hierarchical VQ-VAE and the texture generator on a single NVIDIA 2080 Ti GPU, and train the diverse structure generator on two GPUs. Each part is trained for 10^6 iterations. Training the hierarchical VQ-VAE takes roughly 8 hours. Training the diverse structure generator takes roughly 5 days. Training the texture generator takes roughly 4 days. All the training time is independent of the data set and type of masks. Note that the diverse structure generator and the texture generator can be trained in parallel after the training of the hierarchical VQ-VAE.

B. Negative Log Likelihood and Reconstruction Error.

The VQ-VAE is inspired by lossy compression where performance is usually characterized with rate-distortion curves [5]. Our hierarchical VQ-VAE minimizes the mean-square-error (MSE) reconstruction error as the distortion metric, while our diverse structure generator minimizes the negative log likelihood (NLL) of global latent. As such, we report the distortion in MSE and the NLL of global latent (estimate of coding rate) in Table 3. We do not measure the NLL of local latent because we use a GAN rather than likelihood-based network for texture generation. Note that NLL values are only comparable between likelihood-based networks that use the same pre-trained VQ-VAE. Our diverse structure generator retains the advantages of likelihood-based methods, such as a clear objective to compare models, progress tracking, and measurement of overfitting and mode coverage (the properties that result in diverse samples).

C. Inference Time.

One advantage of GAN-based and VAE-based methods is their fast inference speed. We measure that FE [2] runs at 0.2 second per image on a single NVIDIA 1080 Ti GPU for images of resolution 256×256. In contrast, our model runs at 45 seconds per image. Naively sampling our autoregressive network is the major source of computational time. Fortunately, this time can be reduced by an order of magnitude using an incremental sampling technique [4] which caches and reuses intermediate states of the network. We may integrate this technique in the future.

D. More Visual Examples.

Following the recent inpainting methods, we use center masks or random masks to train our models on the CelebA-HQ, Places2, and ImageNet datasets. The center masks are 128×128 center holes in the 256×256 images. The random masks are rectangles and brush strokes with random positions and sizes, similar to those in [7]. We first show more results of our method using the center-mask models (see Figures 1, 2 and 3) and the random-mask models (see Figure 4). We also show some failure cases of our method (see Figure 5). Then we show that the degree of diversity is controlled by the condition, *i.e*. the available content (see Figures 6 and 7). The degree of diversity is also controlled by the location and size of the missing region (see Figure 8).

E. Discussions on Artifacts.

Although our method can generate more realistic results than prior works, some results still have noticeable artifacts. We analyze the reasons of these artifacts and find that most of them are due to the low quality of the generated structures. Note that we used a light-weight autoregressive network in the structure generator for the sake of computational efficiency, compared with the original, much more complex network in [5]. We anticipate that the results will be improved if using the complex network. In addition,

our texture generator also incurs some artifacts. We may improve it by integrating the new techniques proposed in the recent single-solution inpainting studies, such as feature discriminator [3], multi-scale discriminator [6], and multi-scale generator [2].

F. Discussions on Diversity.

In our method, the diversity is fully determined by the learned conditional distribution for structure generation (since the texture generation has no randomness). We visualize the pixel-wise entropy of the learned distribution to analyze the diversity. As shown in Figure 9, the training dataset and the complexity of incomplete image have impact on the entropy. Intuitively, higher entropy leads to higher diversity.

The diversity of the inpainting results depends on at least the following factors.

(1) *The training dataset.* Since our method learns a conditional distribution for diverse structure generation, it always benefits from diverse training data to enrich the learned distribution. This is evidenced by the experimental results that the resulting diversity on the face dataset is clearly less than that on the natural image datasets (Figure 1 vs. Figure 2); note that the face training images ($\sim 10^4$) are far less than the natural training images ($\sim 10^7$). We conjecture that using a larger dataset or performing training data augmentation may be helpful to increase diversity.

(2) *The incomplete image for inference.* As inpainting is a conditional generation task, the incomplete image acts as the condition or the constraint. The available content in the incomplete image, and the location/size of the missing region, both decide the diversity of the results to a large extent. The effect of the available content is shown in Figure 6 and Figure 7. The effect of the location/size of the missing region is shown in Figure 8.

(3) *The mask type.* We use center masks or random masks to train our models. We find that the models trained with random masks seem to have higher diversity, even for the same incomplete image (Figure 1 Row 3 vs. Figure 8 Row 1). We conjecture that using more random masks may be helpful to increase diversity.

(4) *The method itself.* Taking our method as example, we may improve the diversity by increasing the support of the conditional distribution (*e.g.* codebook size), using more sophisticated model for the distribution, adding regularization terms into the loss function (such as to increase the entropy of the distribution), etc.

(5) *Diversity-quality tradeoff.* We believe that there may be a tradeoff between diversity and quality in the inpainting task. If we pursue higher diversity, we may try to increase the entropy of the learned distribution (e.g. by adding regularization terms into the loss function); then, the quality may be deteriorated since the learned distribution is intentionally biased. Moreover, from a broader perspective, inpainting is a signal restoration task; in such tasks there are always different kinds of tradeoff, like the perception-distortion tradeoff [1]. We have interest to theoretically study the quality-diversity tradeoff in the future.

References

[1] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, pages 6228–6237, 2018. 2

[2] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *ECCV*, pages 725–741, 2020. 1, 2

[3] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *ICCV*, pages 4170–4179, 2019. 2

[4] Prajit Ramachandran, Tom Le Paine, Pooya Khorrami, Mohammad Babaeizadeh, Shiyu Chang, Yang Zhang, Mark A. Hasegawa-Johnson, Roy H. Campbell, and Thomas S. Huang. Fast generation for convolutional autoregressive models. *arXiv preprint arXiv:1704.06001*, 2017. 1

[5] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *NeurIPS*, pages 14866–14876, 2019. 1

[6] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H. Li, Shan Liu, and Ge Li. StructureFlow: Image inpainting via structure-aware appearance flow. In *ICCV*, pages 181–190, 2019. 2

[7] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *CVPR*, pages 1438–1447, 2019. 1

	E_{vq}-D_{vq}
Input size	256×256
Latent layers	32×32, 64×64
Commitment loss weight	0.25
Batch size	8
Hidden units	128
Residual units	64
Layers	2
Codebook size	512
Codebook dimension	64
Conv. filter size	3
Training steps	1,000,000
Polyak EMA decay	0.9997

Table 1. Hyperparameters of our hierarchical VQ-VAE.

	G_s
Input size	256×256
Latent layer	32×32
Batch size	8
Hidden units	128
Residual units	128
Conditioning hidden units	32
Conditioning residual units	32
Layers	20
Attention layers	4
Attention heads	8
Conv. Filter size	3
Dropout	0.1
Output stack layers	20
Training steps	1,000,000
Polyak EMA decay	0.9997

Table 2. Hyperparameters of our diverse structure generator.

	Training NLL	Validation NLL	Training MSE	Validation MSE
CelebA-HQ	1.180	1.243	0.0028	0.0033
Places2	0.969	0.952	0.0042	0.0042
ImageNet	1.127	1.113	0.0082	0.0084

Table 3. Quantitative results of negative log likelihood (NLL) and mean-squared-error (MSE) in the training and validation of our random-mask models. The reported results are evaluated on the CelebA-HQ, Places2, and ImageNet datasets.

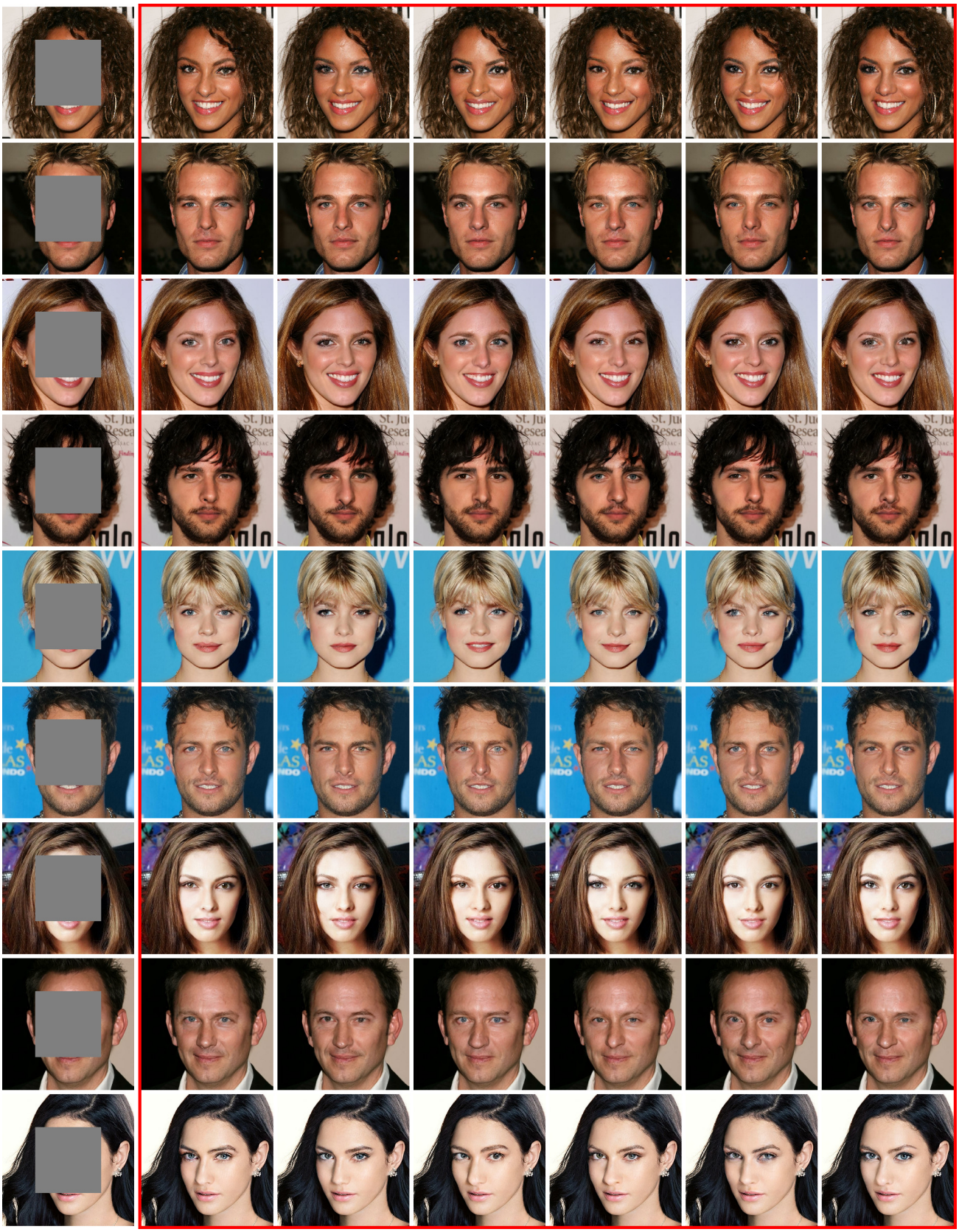

Figure 1. Additional results on the CelebA-HQ test set using the center-mask CelebA-HQ model.

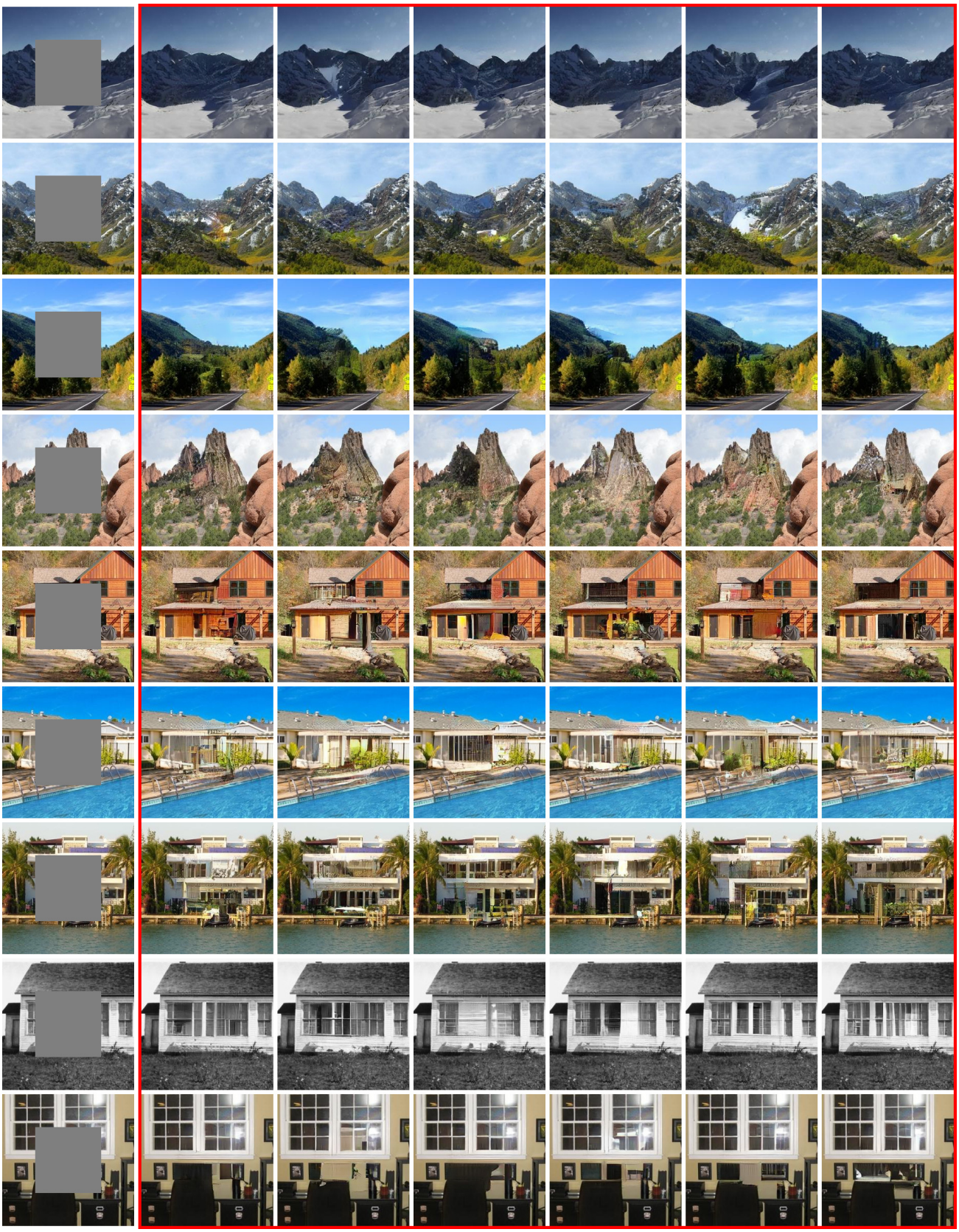

Figure 2. Additional results on the Places2 validation set using the center-mask Places2 model.

Figure 3. Additional results on the ImageNet validation set using the center-mask ImageNet model.

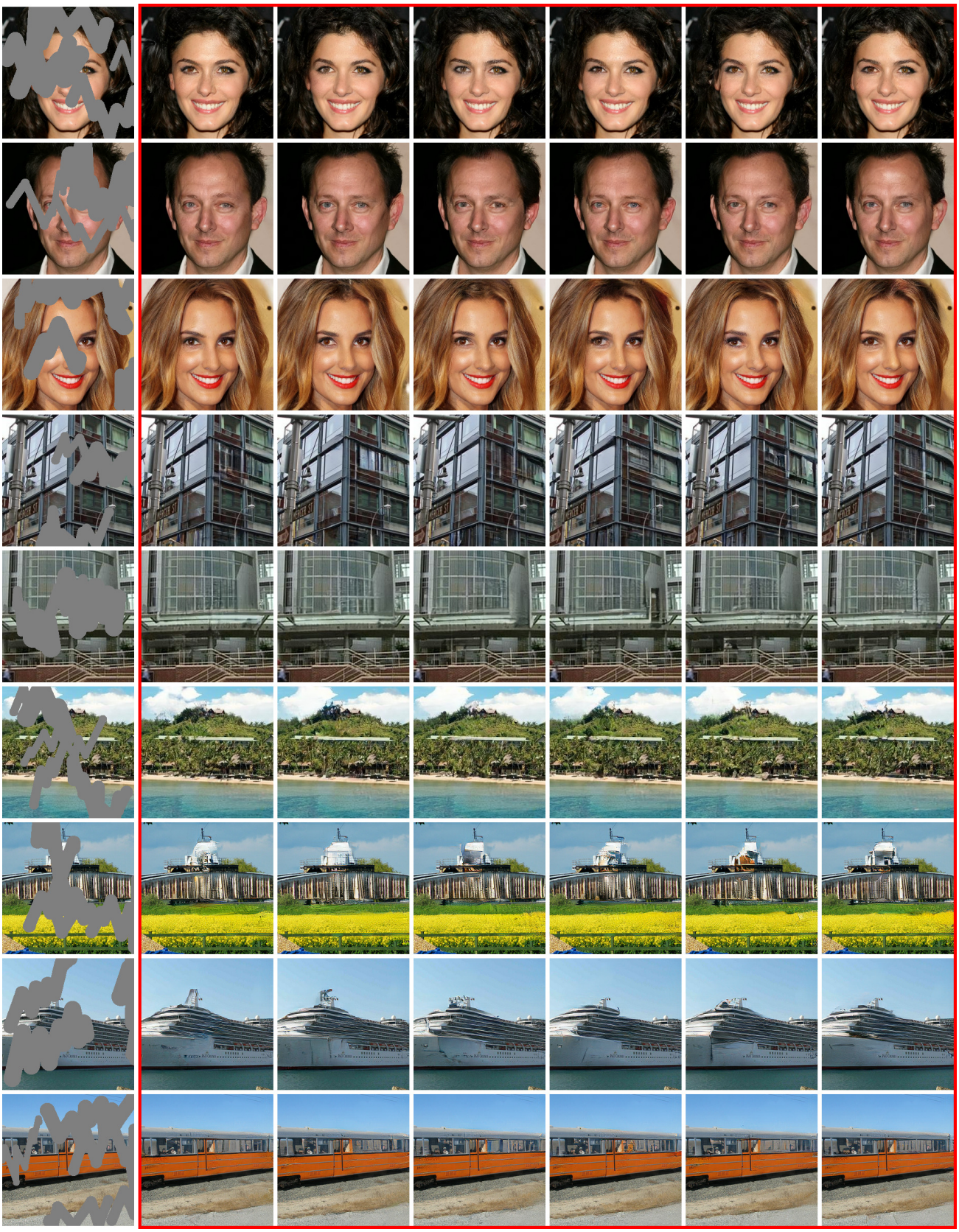

Figure 4. Additional results on the CelebA-HQ, Places2, and ImageNet test (or validation) sets using the random-mask models.

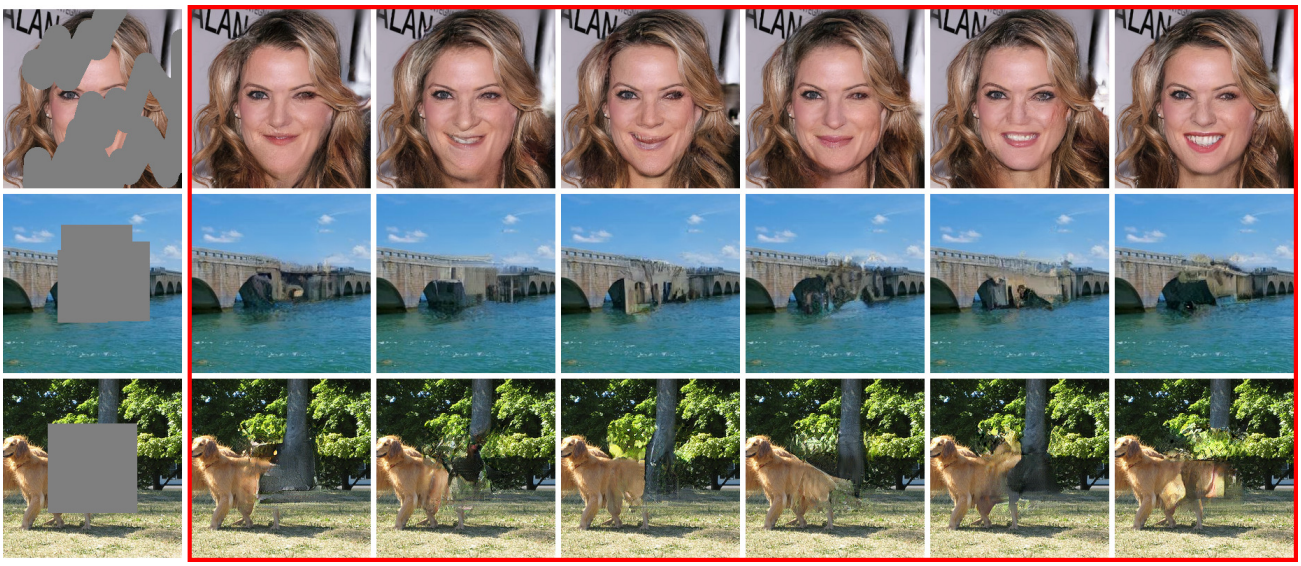

Figure 5. Failure cases of our method on the CelebA-HQ, Places2, and ImageNet test (or validation) sets. (Top) Face image inpainting with big holes. The results are of low quality, *e.g*. distorted faces, asymmetric eyes, missing nostrils, and blurry teeth. (Middle) Scene image inpainting with a complex structure. The results cannot reconstruct the bridge architecture and the bridge holes of varying sizes. (Bottom) Natural image inpainting with a hole of mixed foreground (*i.e*. dog) and background (*i.e*. trees). A large portion of foreground is lost. And the generated results have unsatisfactory structures and blurry textures.

Figure 6. Additional results of our method with low diversity. The degree of diversity is limited when the available content has a simple structure and a plain texture.

Figure 7. Additional results of our method with high diversity. The degree of diversity is high when the available content has a complex structure and an intricate texture.

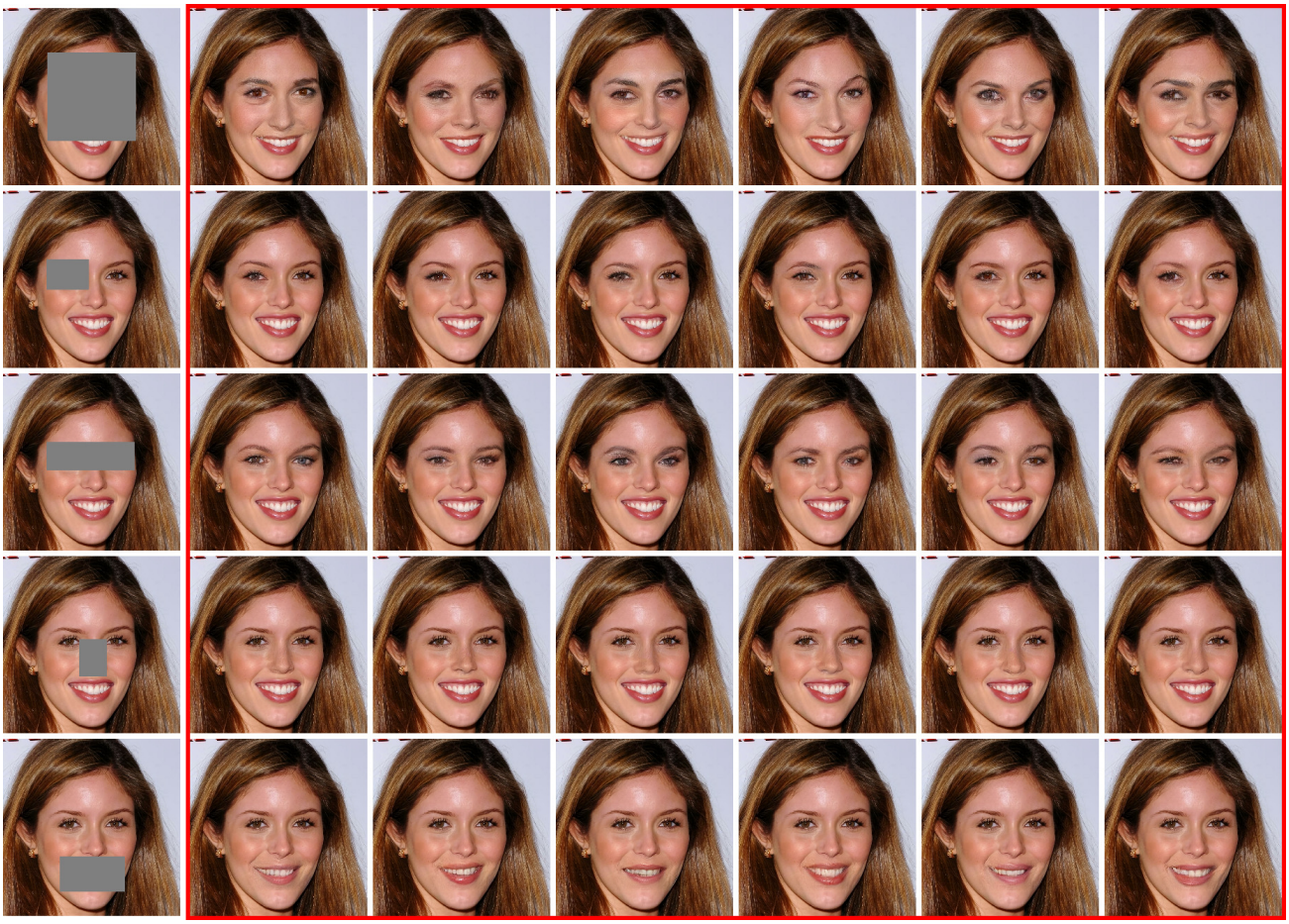

Figure 8. Additional results on one CelebA-HQ test image with different holes using the random-mask CelebA-HQ model. For different rows, the degree of diversity is controlled by the location and size of the missing region.

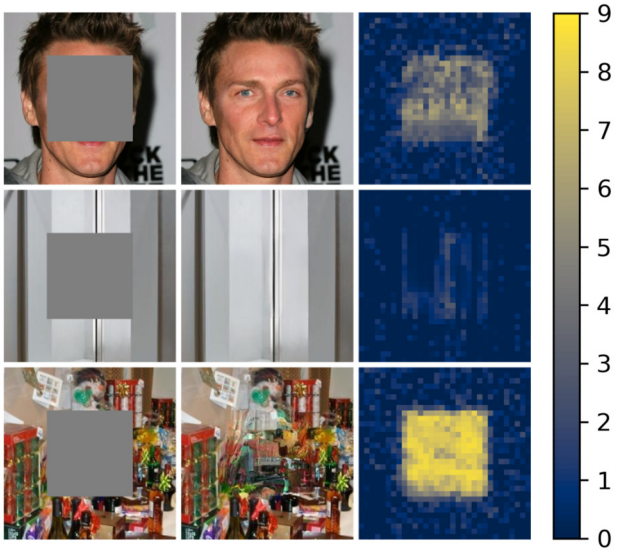

Figure 9. Visualization results of the entropy of learned distribution. For each row, from left to right, the pictures are: incomplete image, one result of our method, and the corresponding visualized entropy. The maximum entropy is 9 because the codebook size is $K = 2^9$.