Supplementary Material
Automatic Correction of Internal Units in Generative Neural Networks

1. Human Evaluations

1.1. Evaluation Results

We select 500 artifacts to measure the performance of corrections based on the criteria in each dataset. The evaluation process is two fold: (1) People re-label for the corrected generation and (2) Check the improvement of generation comparing than original one based on the criteria. The result is summarized in Table 1.

Dataset	Corrected (%)	Improved (%)
CelebA-HQ	53.00 (4.20)	96.00 (2.00)
LSUN-church	54.50 (0.90)	86.10 (6.30)
LSUN-bedroom	46.80 (8.60)	95.50 (1.10)

Table 1: Human evaluation results on corrected artifact generations. The number in the parentheses indicates the standard deviation over raters.

The stacked bar charts in Figure 1 shows the human evaluation results on each dataset. 'Strongly normal' refers to the portion of the unanimous agreement among 500 samples that the corrected image is normal, and 'Strongly artifact' refers to the portion of the unanimous agreement that corrected image is still an artifact. If it is not unanimous agreement, the sample falls into 'Neither normal nor aritfact' portion.

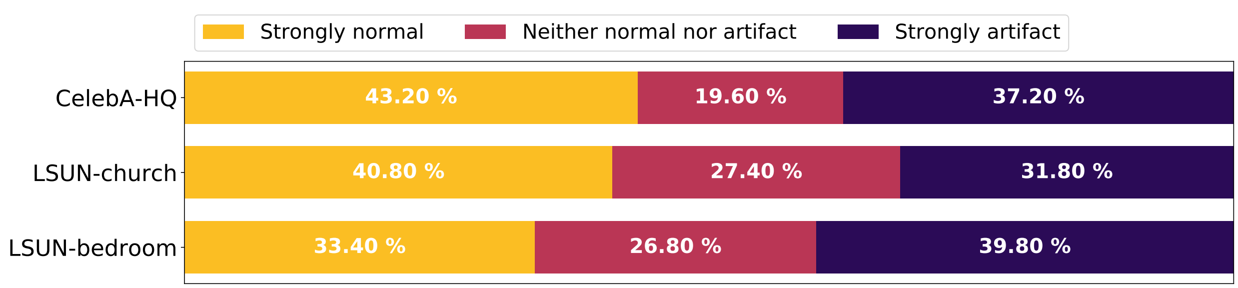

Figure 1: Detailed results of the re-labeling on the corrected images for CelebA-HQ, LSUN-church, and LSUN-bedroom datasets

1.2. Evaluation Criteria

In this section, we describe the criteria for human evaluations. At first, we set the criteria to classify each generation manually. Figure 2 indicates the examples of artifact depending on each criteria. We set different criteria for each network to handle the main artifact types for each dataset. The detailed descriptions of artifact types and the criteria are summarized in Table 2.

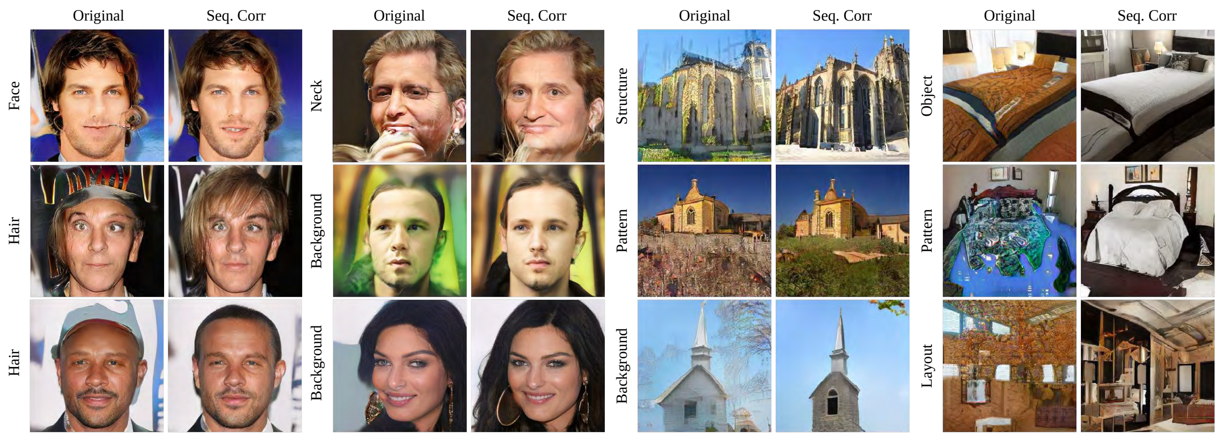

Figure 2: The examples of artifacts based on the descriptions for each dataset. The generation can be included in multiple criteria. For example, left-bottom generation satisfies the hair and background defectiveness at the same time.

Dataset	Type	Description
CelebA-HQ	Face	• There exist defective segments (i.e., eyes, nose, mouth) on the face. • The color-tone and texture of the face is not realistic.
	Hair	• The hair which is fused in the background. • The head is replaced with patterns. • There exist empty parts in hair.
	Neck	• The neck is missing. • Neck and shoulder line is not clear.
	Background	• There exist clear patterns in the background.
LSUN-church	Structural Defects	• Each part of church, i.e., roofs, windows, walls, doors and outlook are not separated each other or from the background. • Church is transparent or overlapped. • The structure of a church is not clear or impossible.
	Pattern	• There exists a clear pattern (i.e., checkerboard pattern) which is unnatural. • The color of an object is unnatural.
	Background	• There exists standing out, unrecognizable object. • There are transparent objects in the sky.
LSUN-bedroom	Object	• There is no bed in bedroom. • The structure or edge of an object is physically not feasible.
	Pattern	• The texture of the images is like a painting. • There is a transparent patterns. • There is a clear pattern crossing the boundary of an object.
	Layout	• The layout of the bedroom is not clear and realistic (i.e., The walls are not connected). • No objects or room at all.

Table 2: Types of artifact generations and its description

2. Automatic Correction Results

We visualize the correction results for each method. FID and defective scores (DS) based correction methods are applied in layer 6. The sequential correction is applied from layer 1 to 6 (The layer 0 means the first dense layer with reshape.). Figure 3 indicates the correction result for random generations. Although we can identify that all of methods maintain information for the plausible regions, the sequential correction method shows effective repair for the defective regions (e.g. The stain on the forehead can be removed from the sequential correction method (Top-right in Figure 3.). Note that all of correction methods are in the global approach to improve the generator not the repair per query (given generation). The entire methods basically stop or reduce the generation information in the network (The indices of units are pre-defined based on FID score or DS.). Figure 4 - 9 indicate the correction results for each class in various dataset.

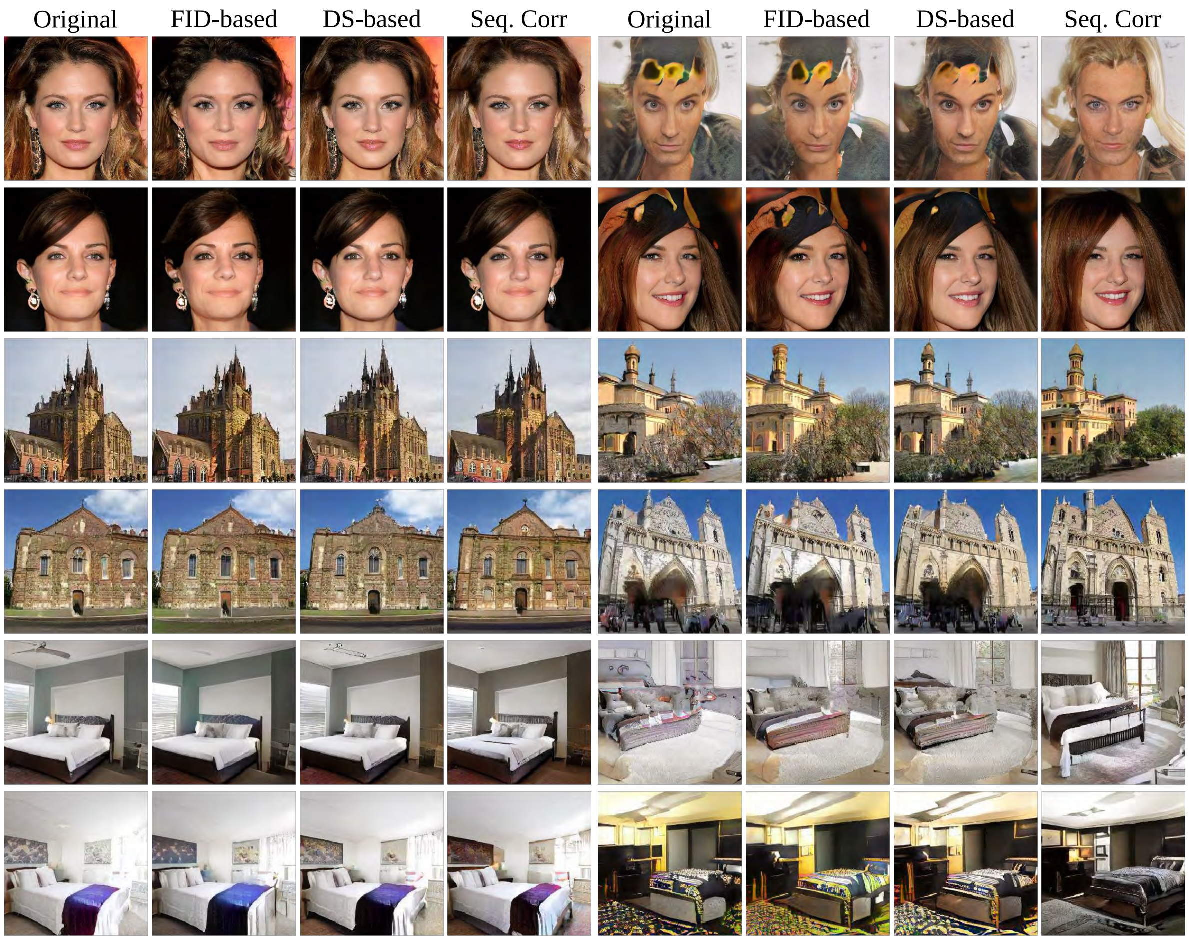

Figure 3: Correction result for each method in various dataset. Although the entire method can conserve the plausible regions, the proposed method can correct the defective regions effectively.

2.1. Artifacts in PGGAN with CelebA-HQ

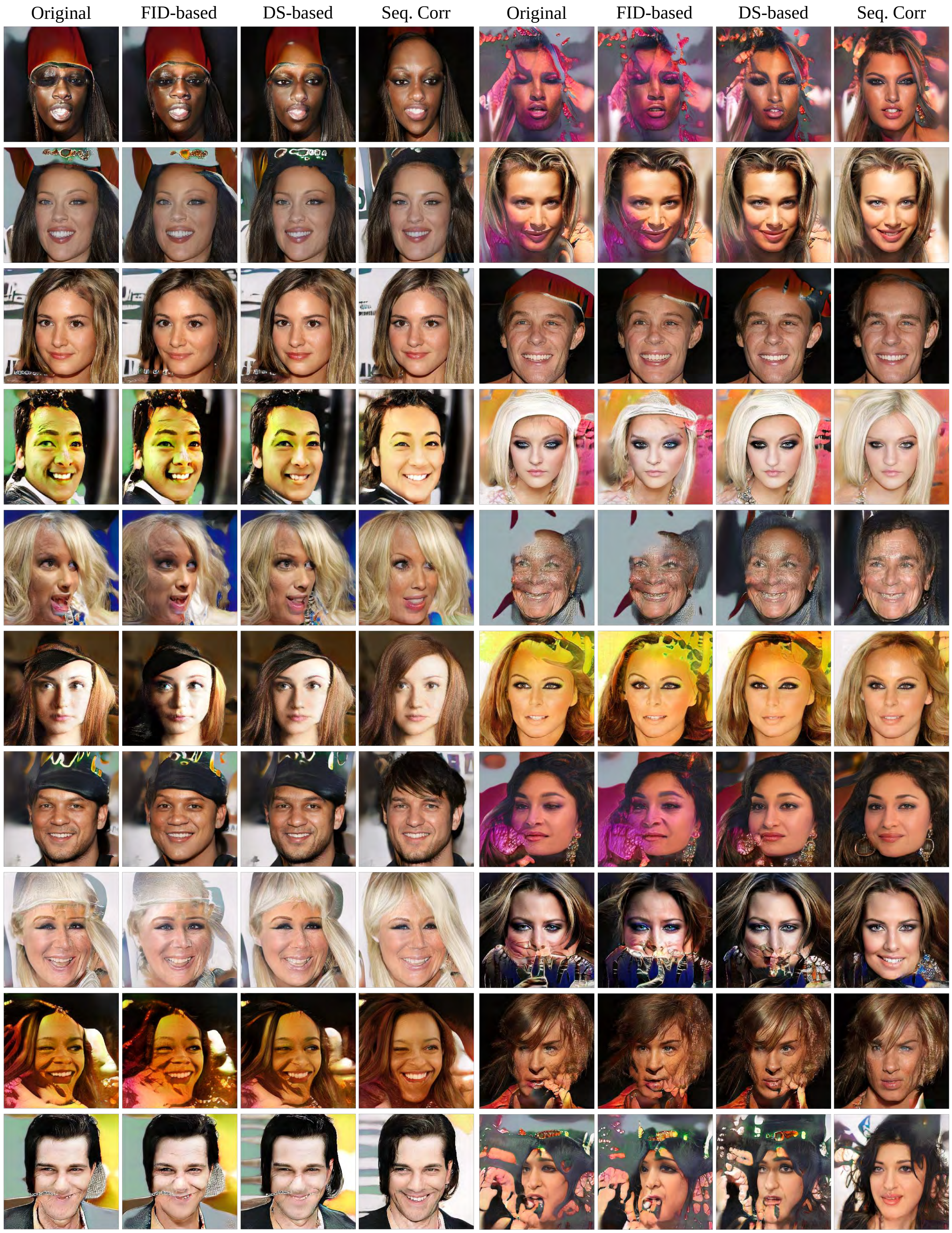

Figure 4: Correction results for the artifacts in PGGAN with CelebA-HQ.

2.2. Normal in PGGAN with CelebA-HQ

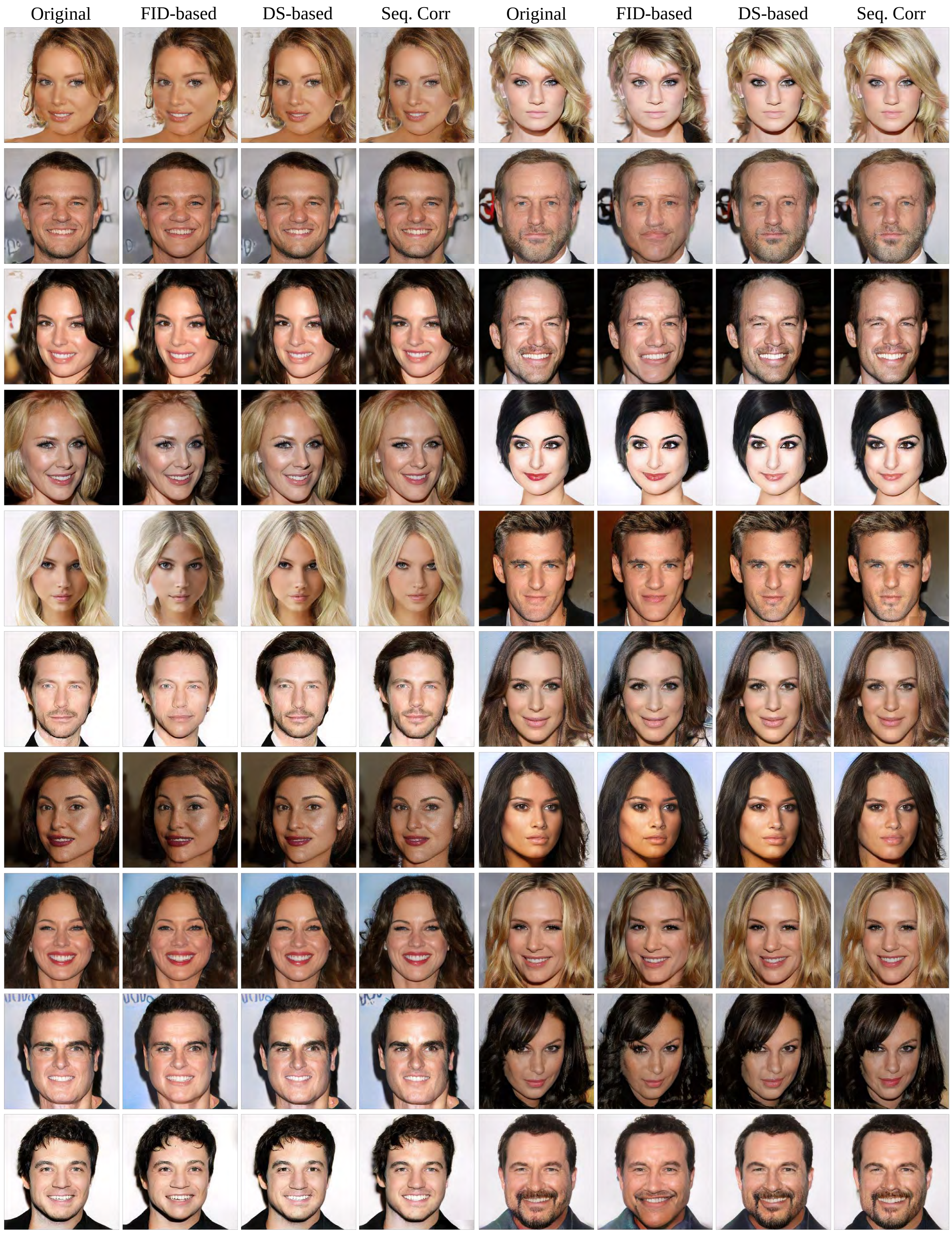

Figure 5: Correction results for the normal generations in PGGAN with CelebA-HQ.

2.3. Artifacts in PGGAN with LSUN-church Outdoor

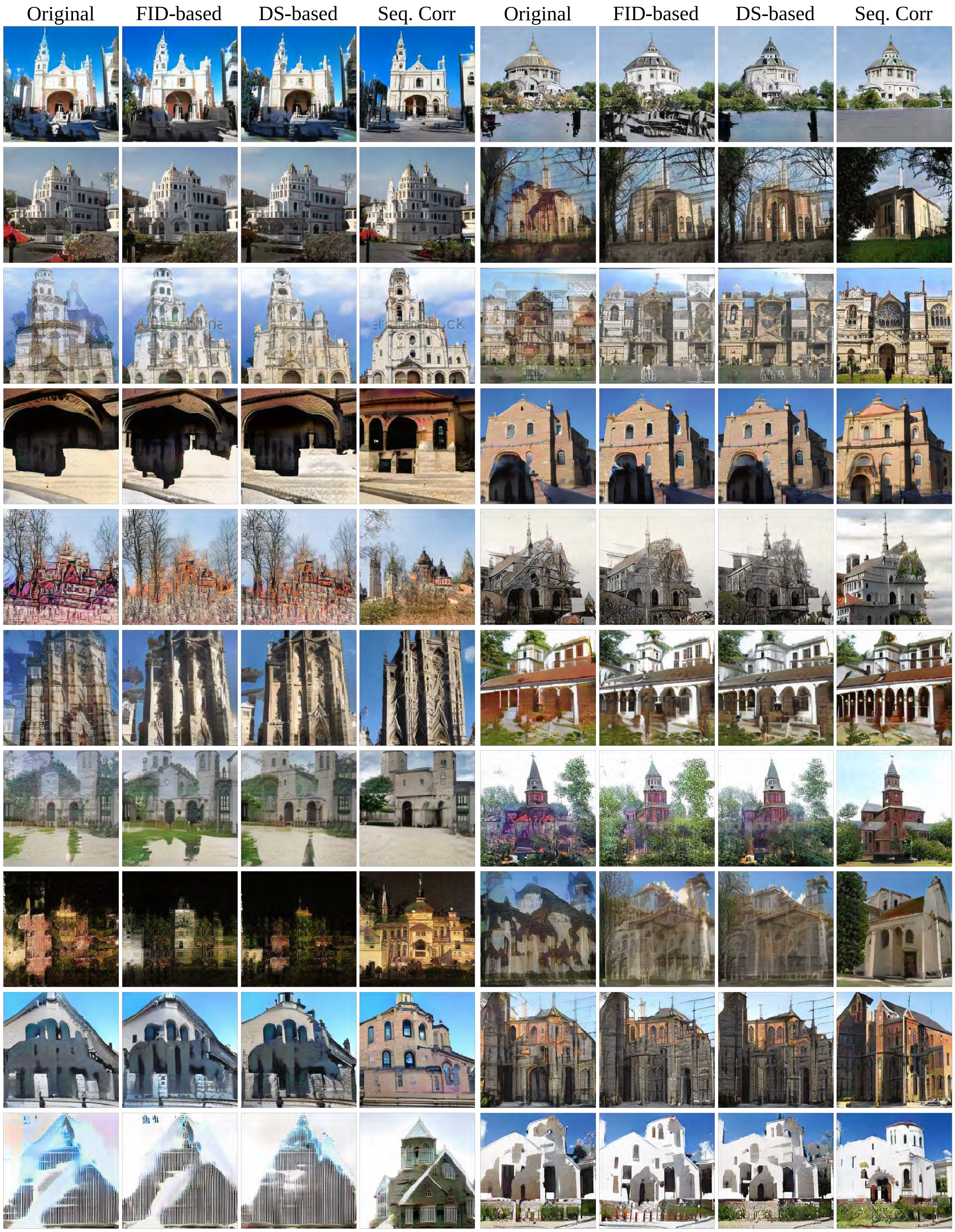

Figure 6: Correction results for the artifacts in PGGAN with LSUN-church Outdoor.

2.4. Normal in PGGAN with LSUN-church Outdoor

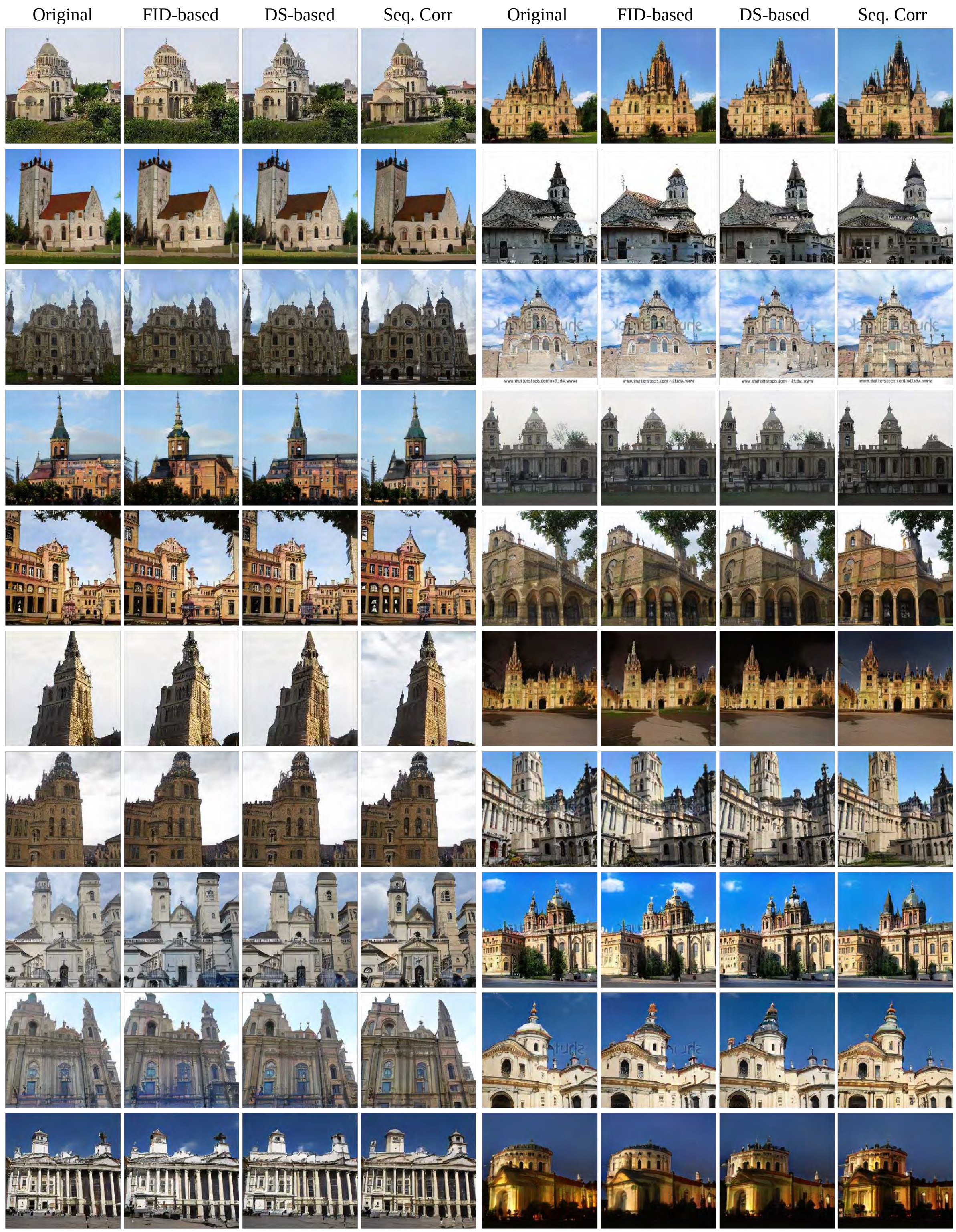

Figure 7: Correction results for the normal generations in PGGAN with LSUN-church Outdoor.

2.5. Artifacts in PGGAN with LSUN-bedroom

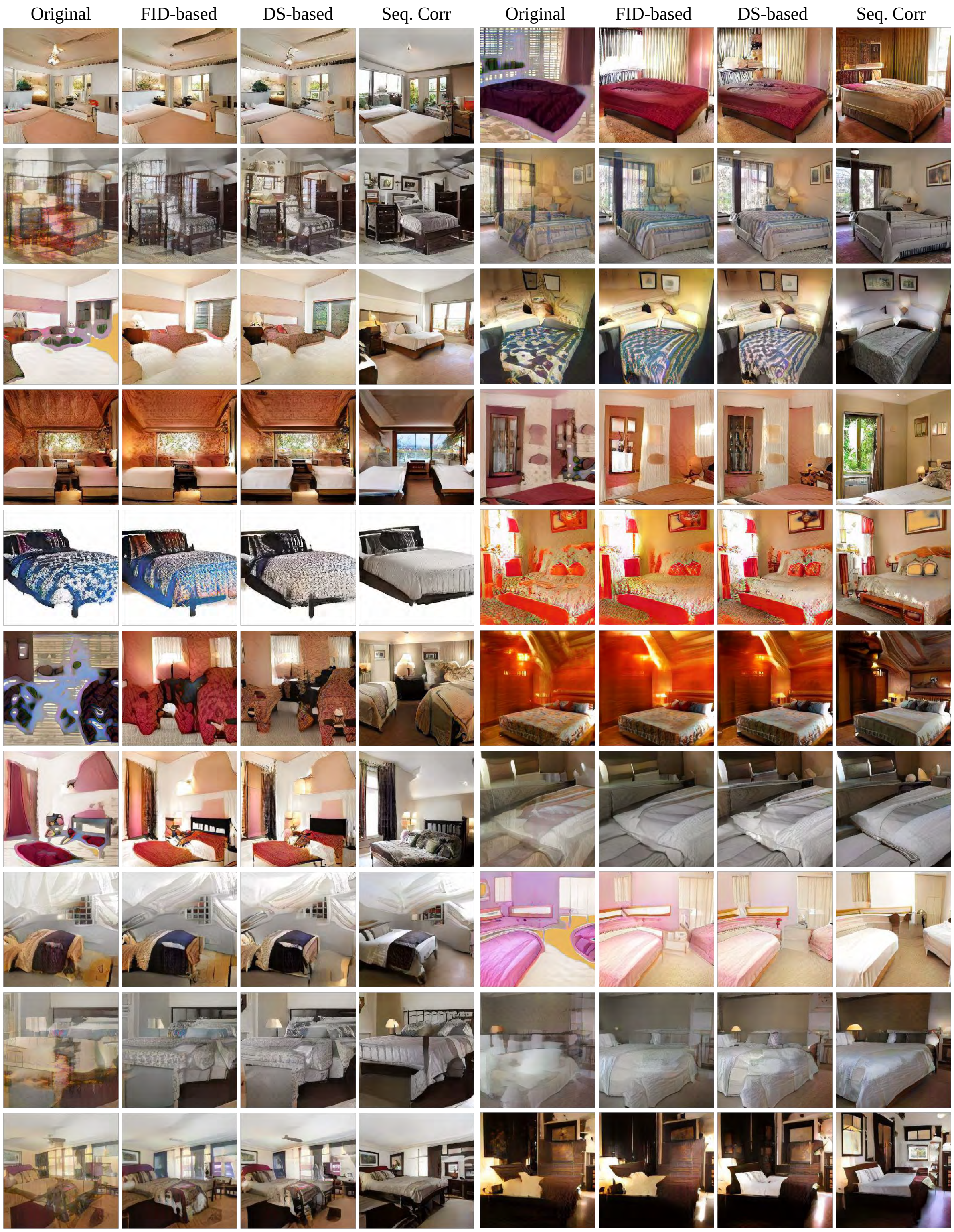

Figure 8: Correction results for the artifacts in PGGAN with LSUN-bedroom.

2.6. Normal in PGGAN with LSUN-bedroom

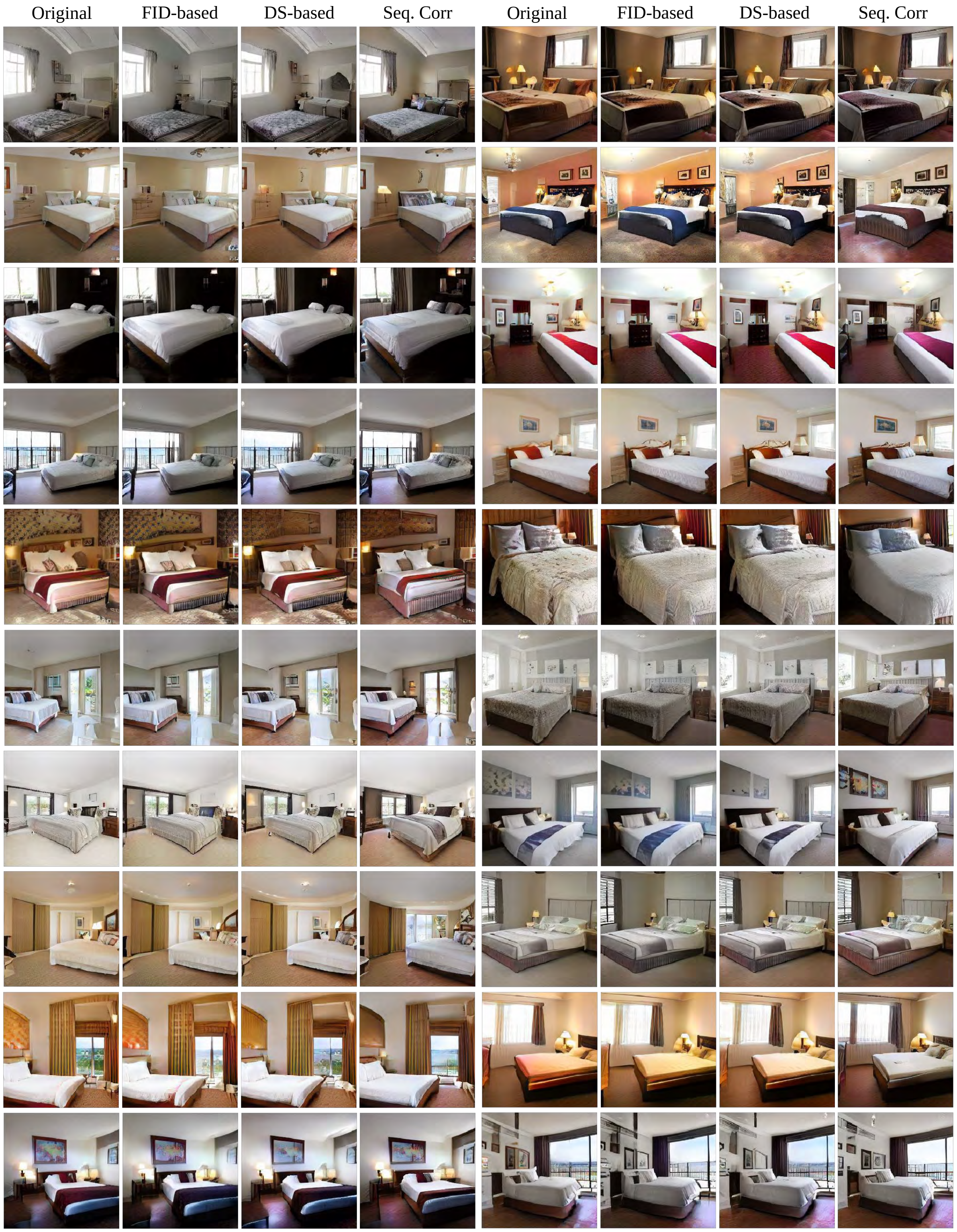

Figure 9: Correction results for the normal generations in PGGAN with LSUN-bedroom.

3. Sequential Correction in StyleGAN v2

Due to the structural difference between PGGAN and StyleGAN v2(e.g. stochastic variation induced at each convolutional layer), the global unit correction is not trivial. However, we modified our method to use each generation artifact map obtained by GradCAM to derive the sequential correction. After a generation is classified as artifact, we obtain a mask for defective region. Note that the classifier is only trained once using the PGGAN generated annotated images on CelebAHQ. Similar to our global unit identification for PGGAN, we can assign relative defective score using GradCAM mask. Figure 10 depicts the result of sequential local region correction for images generated by StyleGAN v2 trained on FFHQ dataset.

Figure 10: Correction result for the generations in StyleGAN v2 with FFHQ.

4. Sequential Correction in U-net GAN

As described in Section 5.3 in the paper, we can apply our method with some modification to other sate-of-the-art generation models. For this purpose, we select U-net GAN which the generator module is same with BigGAN with some modifications. Specifically, U-Net GAN generator feed the same input latent vector to BatchNorm. Furthermore, they introduced an unconditional generator by replacing class-conditional BatchNorm in BigGAN with self-modulation. We followed the same procedure as StyleGAN v2 for sequential correction.

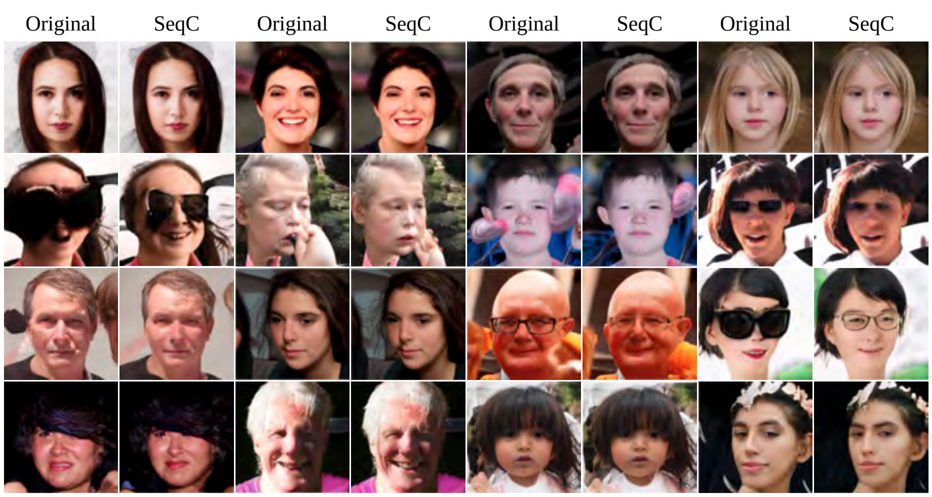

Figure 11: Correction result for the generations in U-net GAN with FFHQ.

5. Representative generation for each unit

In this section, we visualize the the representative generation and highly activated generations for selected units in layer 6 of PGGAN generator with various dataset to validate the selected units generation concepts. Figure 13 - 15 indicate the visualization results from top 15 units for each scoring method (left: FID scores and right: defective score (DS)).

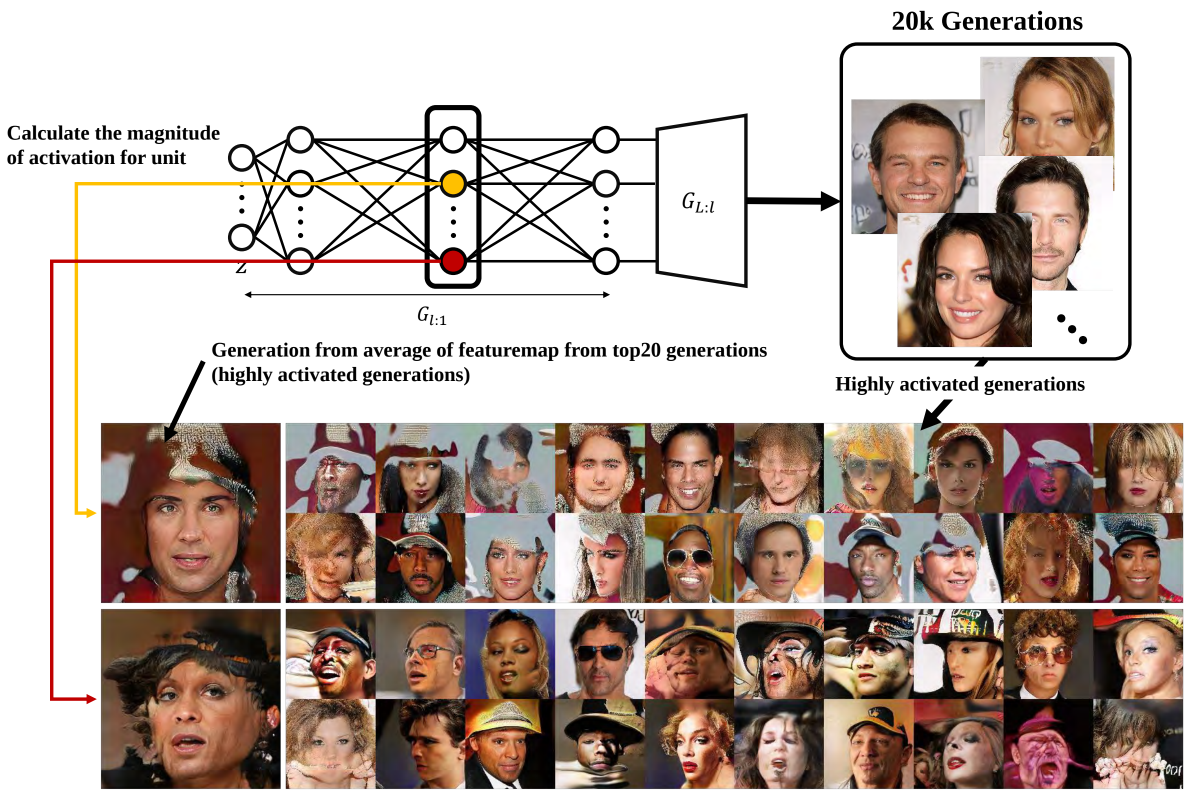

Figure 12: How to select 20 generations that maximize each featuremap unit

Figure 13: Representative generation for each unit in layer 6 of PGGAN with CelebA-HQ. (Left) FID score based unit selections. (Right) DS based unit selections.

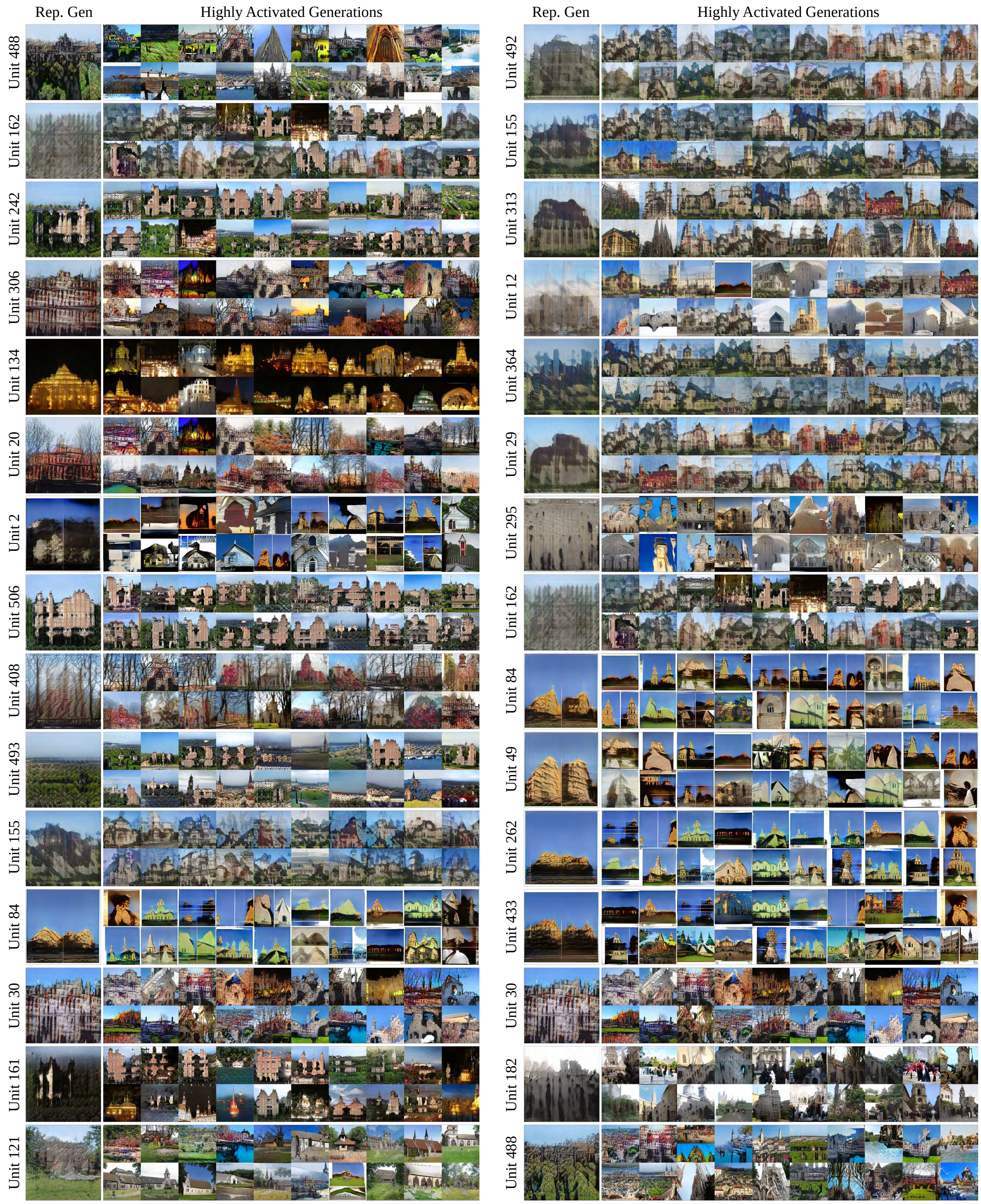

Figure 14: Representative generation for each unit in layer 6 of PGGAN with LSUN-church Outdoor. (Left) FID score based unit selections. (Right) DS based unit selections.

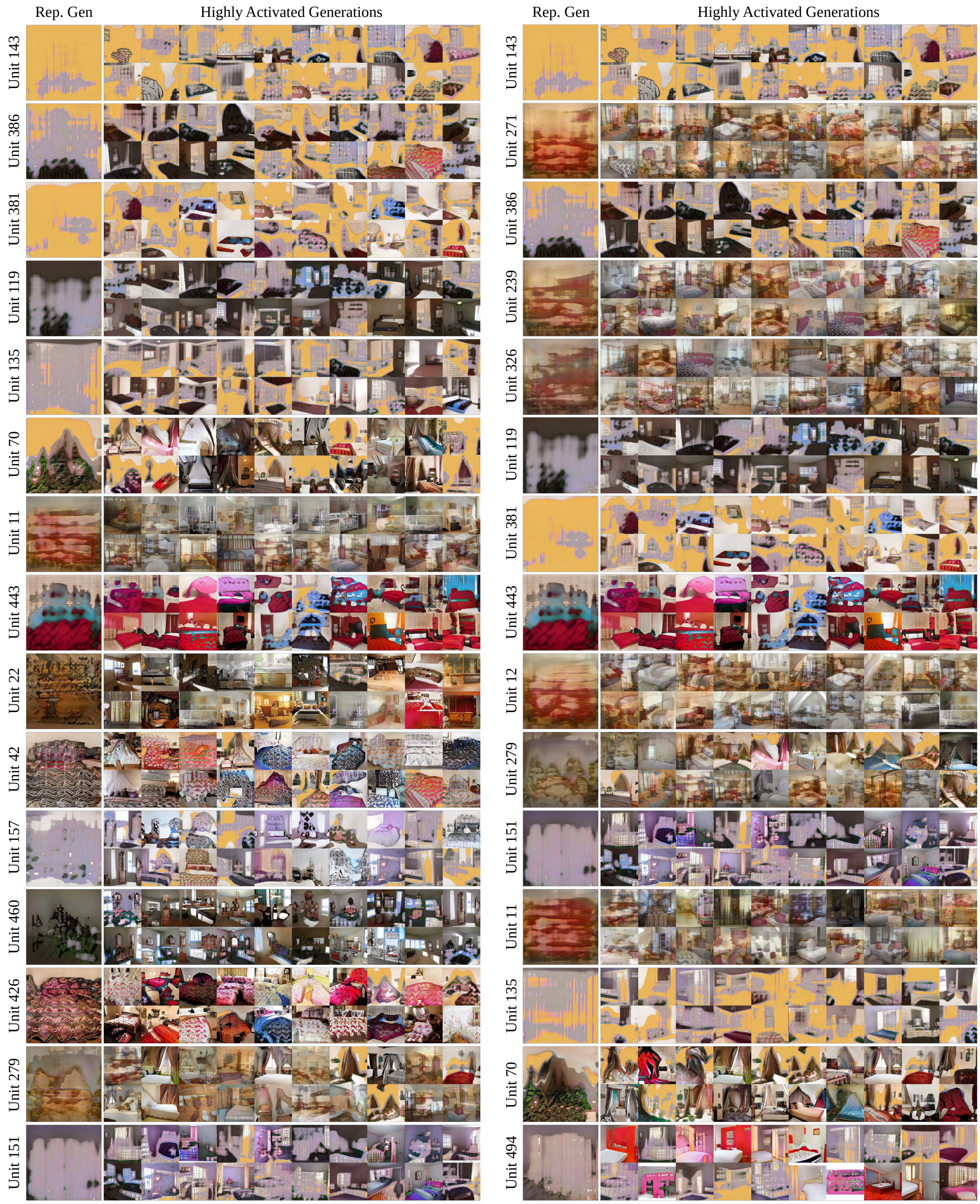

Figure 15: Representative generation for each unit in layer 6 of PGGAN with LSUN-bedroom. (Left) FID score based unit selections. (Right) DS based unit selections.