Automatic Correction of Internal Units in Generative Neural Networks

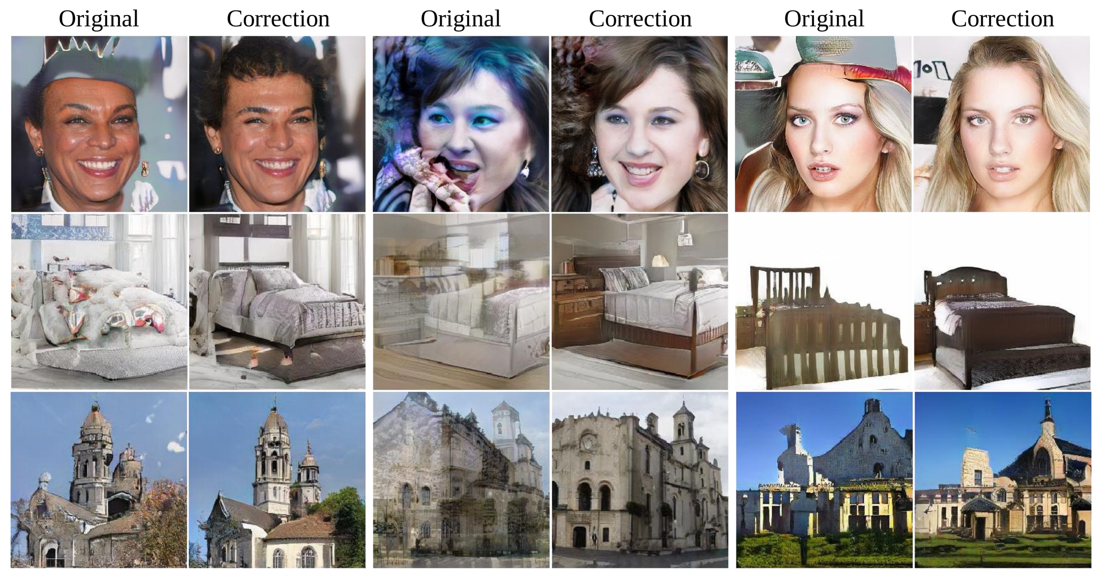
Ali Tousi[1,*], Haedong Jeong[1,2,*], Jiyeon Han[1], Hwanil Choi[1] and Jaesik Choi[1,3]

[1]KAIST, [2]UNIST, [3]INEEJI

(* indicates equal contribution)

GAN Correction

Goal: Correction of artifact generations without re-training the generator.



Motivations:

- Existing Generative Adversarial Networks (GANs) generate low visual fidelity images known as artifacts.

Key Contributions:

- Identifying internal defective units in GANs.
- An artifact removal method by globally ablating defective units.
- Generalization for various structure of generator.

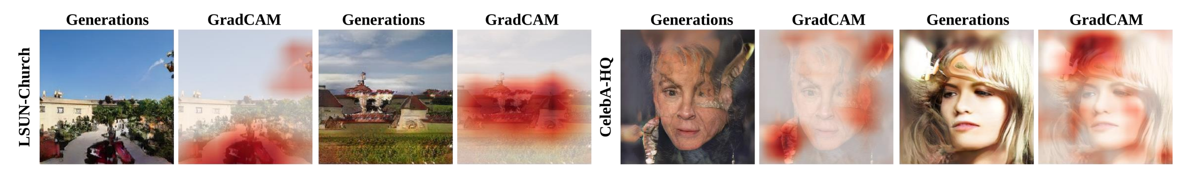
Artifact Unit Identification

FID-Based Artifact Unit Identification:

- In existing research [1], artifact units are identified based on Fr´echet Inception Distance (FID).
- However, the FID-based identification misjudge some units.

Classifier-Based Artifact Unit Identification:

- Train a classifier with hand-labeled generations.
- Apply GradCAM [2] to obtain artifact mask.
- Define defective score (DS) based on Intersection of Unions between internal featuremaps and Grad-CAM mask.

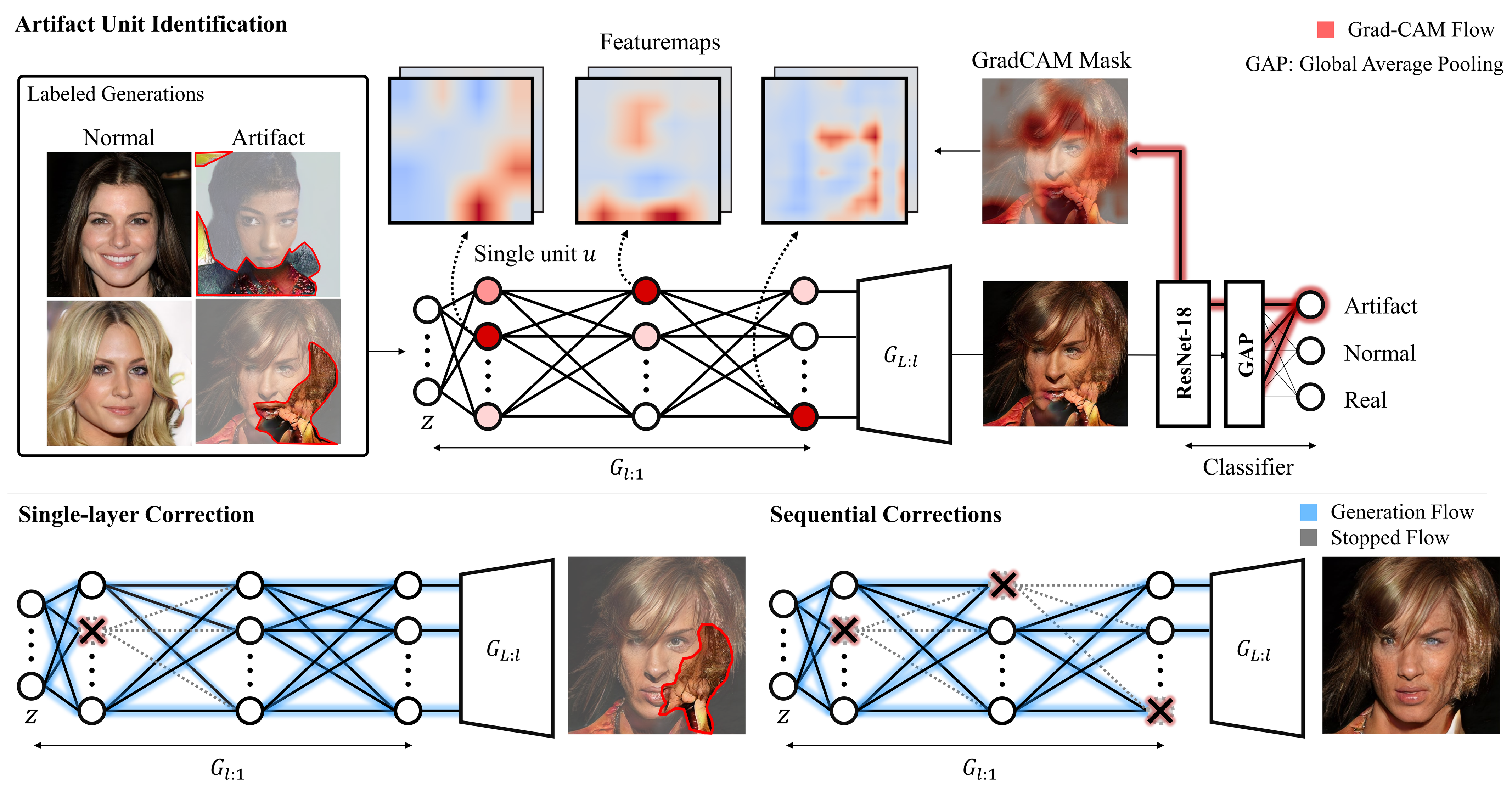


Trade-off in Single Layer Ablation:

- Although increasing the number of ablation units can correct the artifact region, it may degrade the quality at the same time.

Automatic Correction of Internal Units

Identification of the artifact units for each layer (top) and the generation flow for two correction method (bottom).



Experiments & Results

Algorithm 1 Sequential Correction

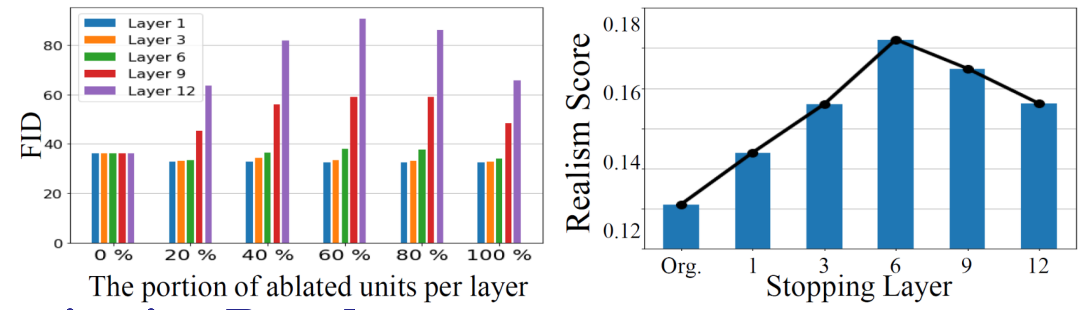
Input: z_0: a query, $G(.) = f_{L:1}(.)$: a generator, l: a stopping layer, $DS_{l:1,a}$: normalized defective scores for each layer, λ: a scaling factor, n: the number of ablated units

Output: X: the corrected generation

1: $h_0 = z_0$
2: **for** $k \leftarrow 0$ to l **do**
3: $\quad h_{k+1} = f_{k+1:k}(h_k)$
4: $\quad$ **for** $j \leftarrow$ Top 1 to Top n **do**
5: $\qquad h_{k+1,j} = \lambda(1 - DS_{k+1,j,a})h_{k+1,j}$
6: $\quad$ **end for**
7: **end for**
8: $X = f_{L:l+1}(h_{l+1})$
9: **return** X

Analysis for hyper parameters:

- FID and Realism score [3] for various hyper parameters.

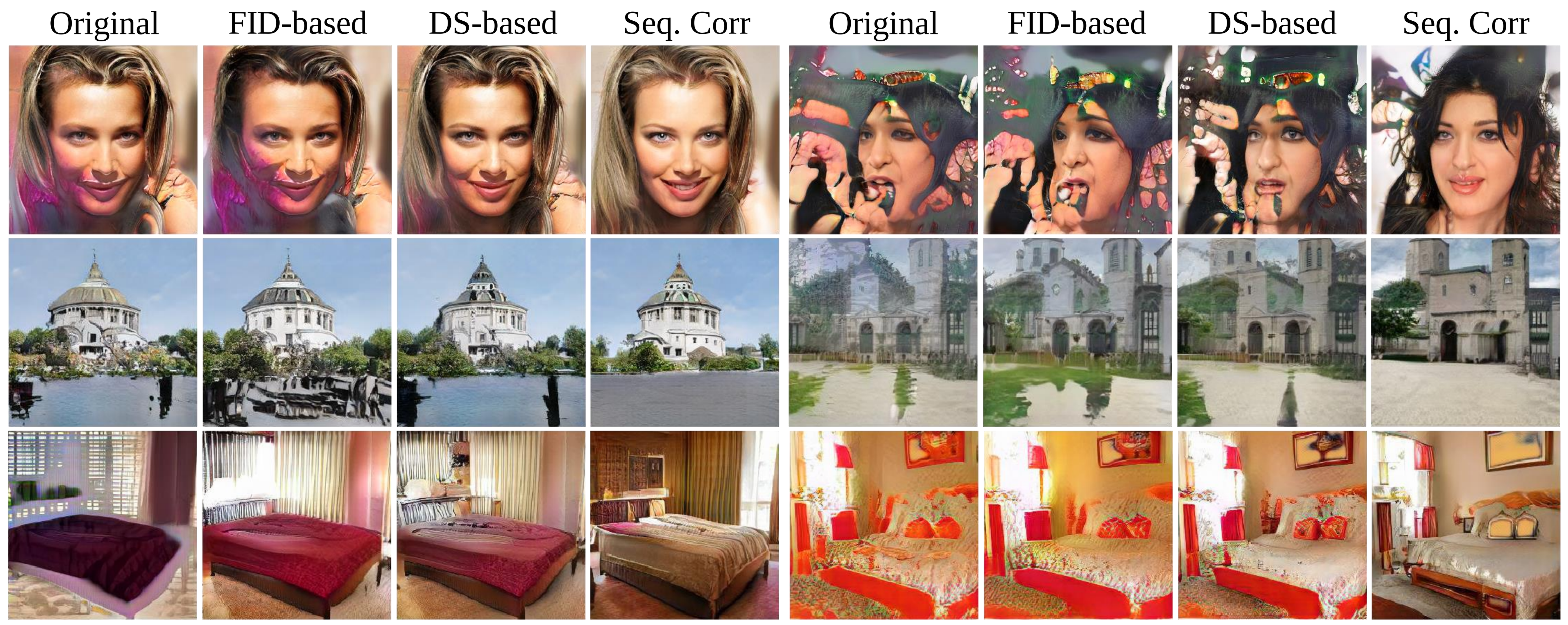


Quantitative Results:

- FID scores of corrected artifact generations for PG-GAN with various dataset.

Correction	LSUN-Church	LSUN-Bedroom	CelebA-HQ
Random	53.43	42.10	67.46
FID	40.66	44.37	48.48
DS	32.82	34.71	44.93
Seq. Corr	**23.96**	**34.71**	**40.71**

Qualitative Results:



Generalization:

- The proposed method with minor modification can be generalized for the various structure of generator.
- In StyleGAN v2 and U-net GAN which is a variant of BigGAN, the correction performance is validated.

Discussion:

- Sequential correction method that requires no additional retraining.
- Plausible correction performance and generalization for various recent generator models.
- Illustrated below are some failure cases which the original structure was changed after correction.

Reference

[1] David Bau, et al. Gan dissection: Visualizing and understanding generative adversarial networks. ICLR, 2019.
[2] Ramprasaath R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. ICCV, 2017.
[3] Tuomas Kynk¨a¨anniemi, et al. Improved precision and recall metric for assessing generative models. CoRR, abs/1904.06991, 2019.