A Appendix

A.1 Choice of GAN inversion

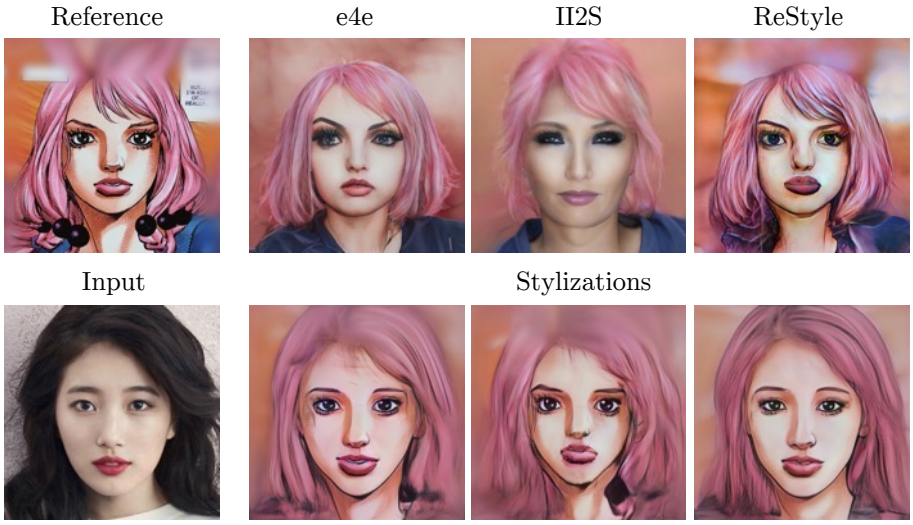

Fig. 14. The choice of GAN inversion matters. We compare JoJoGAN trained on e4e [34], II2S [42], and ReStyle [1] inversions. II2S gives the most realistic inversions leading to stylizations that preserves shapes and proportions of the reference. ReStyle gives the most accurate reconstruction leading to stylization that better preserves the features and proportions of the input.

JoJoGAN relies on GAN inversion to create a paired dataset. We investigate the effect of using 3 different GAN inversion methods, e4e [34], II2S [42], and ReStyle [1] in Figure 14.

Using e4e fails to accurately recreate the style reference and conveniently gives us a corresponding real face. On the other hand, ReStyle more accurately inverts the reference, giving a non-realistic face. II2S is a gradient-descent based method with a regularization term that allows us to map the style code to a higher density region in the latent space. The regularization term results in very realistic faces that are somewhat inaccurate to the reference.

The different inversions give us different JoJoGAN results. Training with ReStyle leads to clean stylization that accurately preserves the features and proportions of the input face. Training with II2S on the other hand leads to heavy stylization that borrows the shapes and proportions from the reference. However, this also leads to pretty heavy semantic changes from the input face and artifacts (note the change of identity and artifacts along the neck).

In practice, we blend the style codes from ReStyle and the mean face. For M, we borrow the style code from the mean face at layers 7, 9, and 11. This borrows

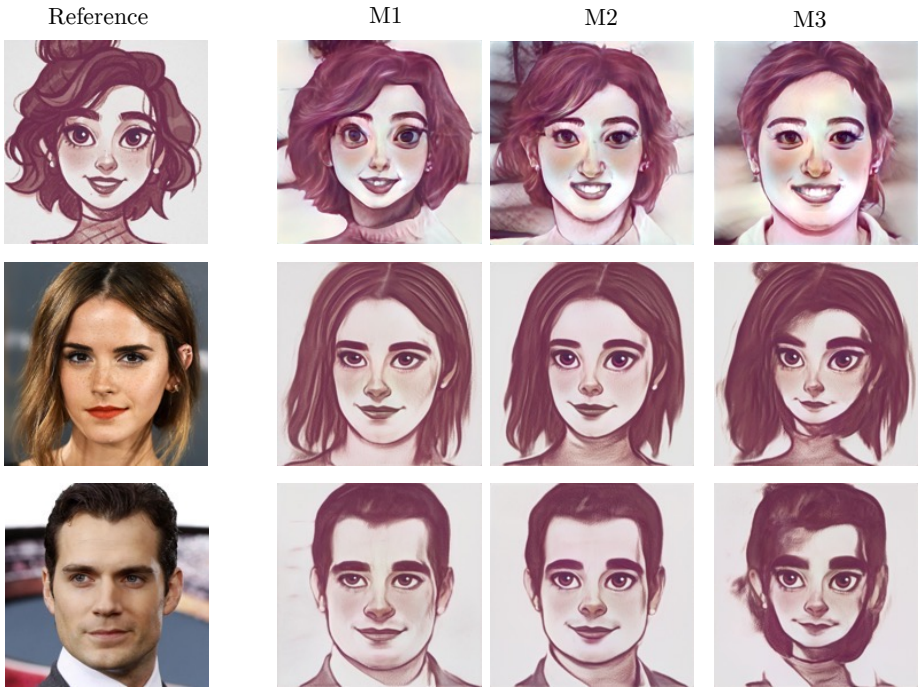

Fig. 15. The choice of M matters. M controls the blend between the inverted style with the mean style. $M1$ is the closest to the reference, leading to smaller features (e.g., eyes). $M3$ is the closest to a real face, leading to exaggerated features more like reference and also significant artifacts.

the facial features of the mean face to the inversion. However, it is impossible to only affect the proportions of the features by simply blending coarsely at a layer level. For example, naively blending the mean face can change the expression of the inversion, e.g. from neutral to smiling or introduce artifacts. We thus have to blend at a finer scale, which we are able to do so by isolating specific facial features in the style space using RIS [2]. Figure 15 compares the results of using different M for blending. Note that when the blended image is more face-like ($M3$), the exaggerated features of the reference is transferred. However, significant artifacts are introduced, see $M3$ row 2. By carefully selecting M, we can transfer the exaggerated features while avoiding artifacts, see $M2$.

A.2 Identity loss

Before computing identity loss, we grayscale the input images to prevent the identity loss from affecting the colors. The weight of the identity loss is reference-dependent, but we typically choose between 2×10^3 to 5×10^3.

A.3 Choice of style mixing space

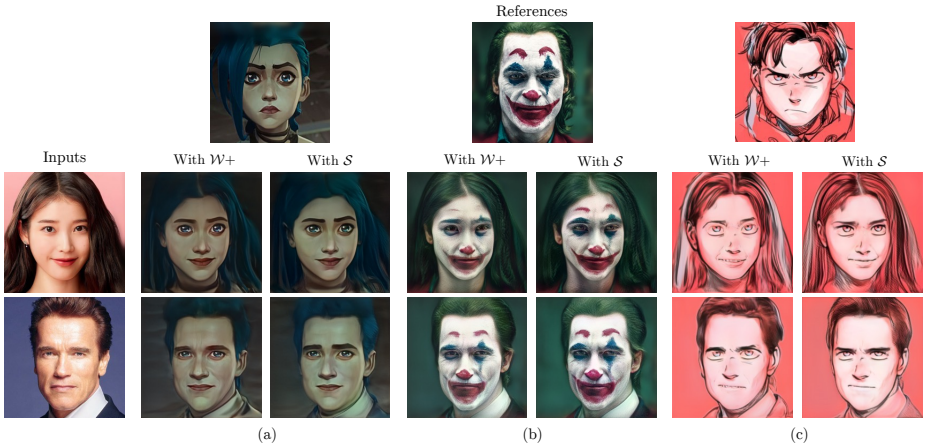

Fig. 16. We study how the choice of latent space to do style mixing affects JoJoGAN. Style mixing in $\mathcal{S}$ space gives more accurate color reproduction in (a) and (b) and better stylization effect (note the eyes) in (c).

Style mixing in Equation (1) allows us to generate more paired datapoints. It is reasonable to map faces with slight differences in textures and colors to the same reference. As such it is pertinent that while we style mix to generate different faces, we need certain features such as identity, face pose, etc to remain the same. We study how the choice of latent space to do style mixing affects the stylization. In Figure 16 we see that style mixing in $\mathcal{S}$ gives better color reproduction and overall stylization effect. This is because $\mathcal{S}$ is more disentangled [36] and allows us to more aggressively style mix without changing the features we want intact.

A.4 Varying dataset

Using $\mathcal{C}$ and $\mathcal{X}$ gives different stylization effects. Finetuning with $\mathcal{X}$ accurately reproduces the color profile of the reference while $\mathcal{C}$ tries to preserve the input color profile. However, this is insufficient to fully preserve the colors as we see in Figure 17. Grayscaling the images before computing the loss in Equation (2) in addition to finetuning with $\mathcal{C}$ gives us stylization effects without altering the color profile. We show that it is necessary to use both $\mathcal{C}$ and grayscaling to achieve this effect and using $\mathcal{X}$ and grayscaling is insufficient.

A.5 Feature matching loss

For discriminator feature matching loss, we compute the intermediate activations after resblock $2, 4, 5, 6$.

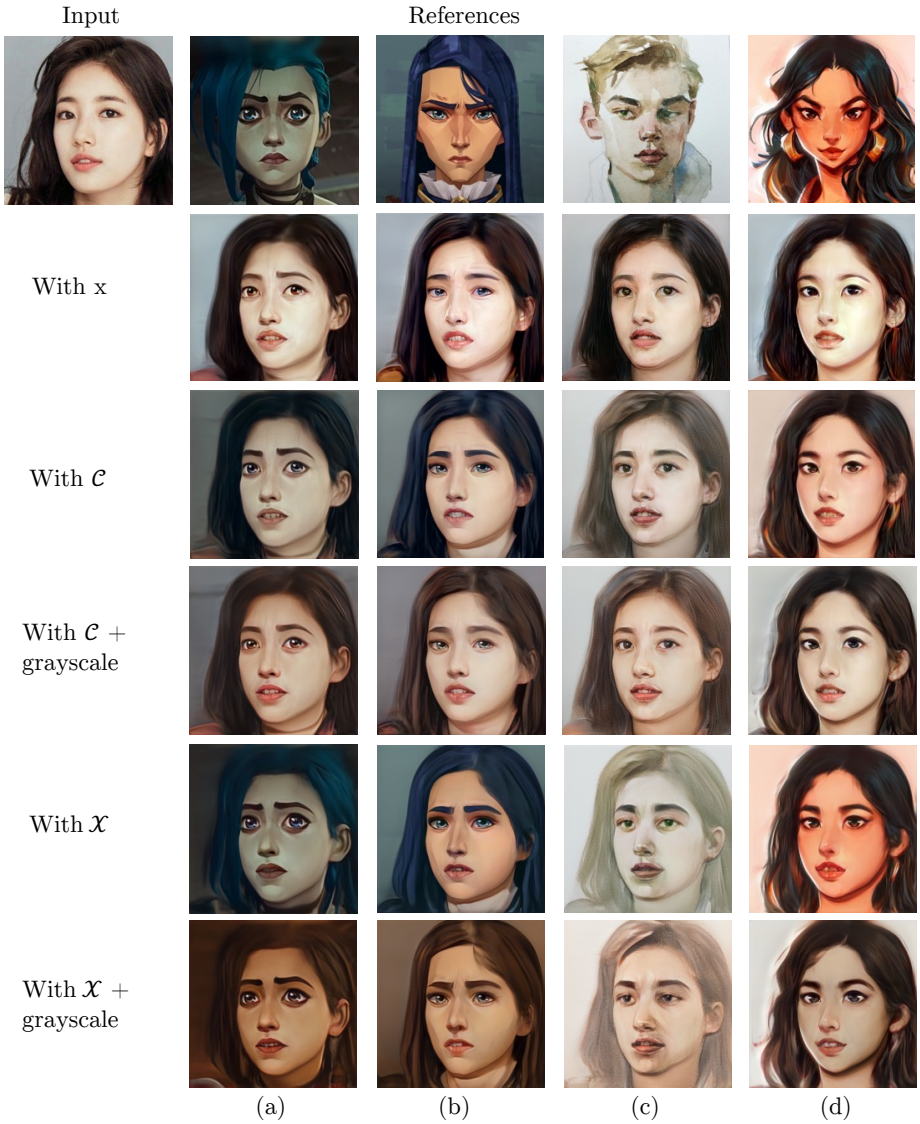

Fig. 17. The choice of training data has an effect. **First row:** when there is just one example in $\mathcal{W}$, JoJoGAN transfers relatively little style, likely because it is trained to map "few" images to the stylized example. **Second row:** same training procedure as in Figure 8, using $\mathcal{C}$. **Third row:** same training procedure as second row but with grayscale images for Equation (2). **Fourth row:** same training procedure as in Figure 8, using $\mathcal{X}$. **Fifth row:** same training procedure as Fourth row but with grayscale images for Equation (2).

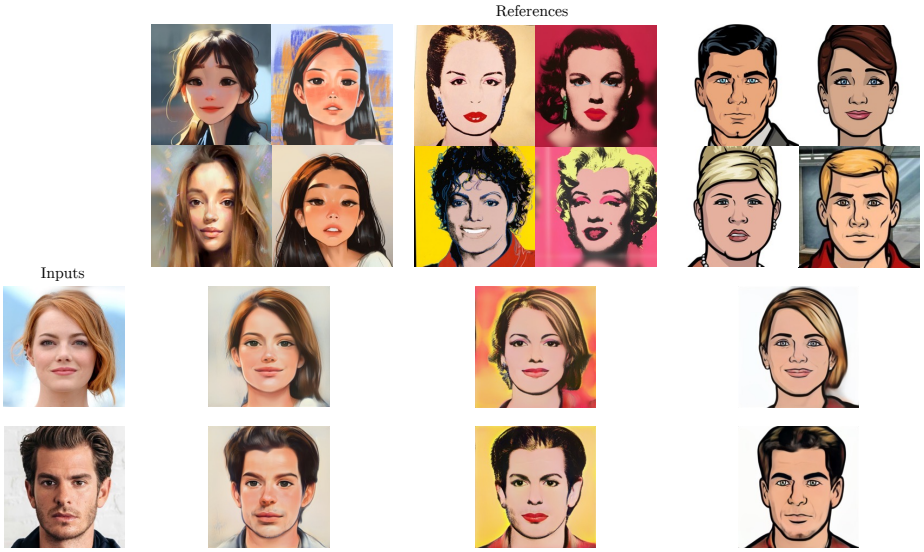

Fig. 18. More multi-shot examples

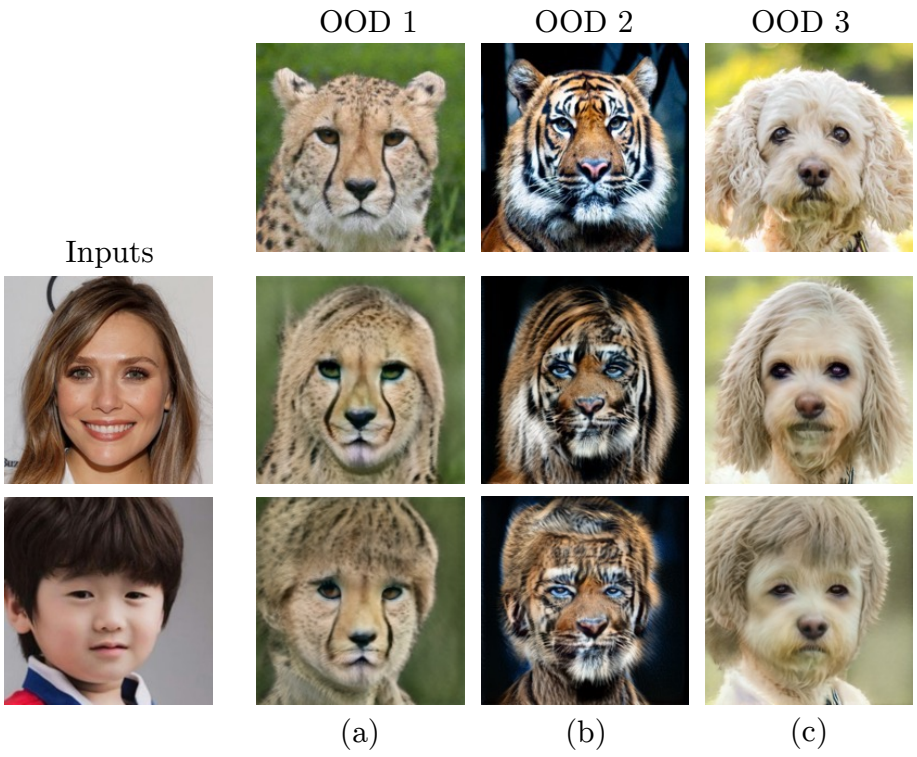

Fig. 19. JoJoGAN produces unsatisfactory style transfers on OOD cases, producing human-animal hybrids.

Fig. 20. We compare with Zhu *et al*. [41] on all examples for references used in their paper and described as hard cases there. For each reference, the top row is JoJoGAN while the second row is Zhu *et al*. Note how their method distorts chin shape, while JoJoGAN produces strong outputs.

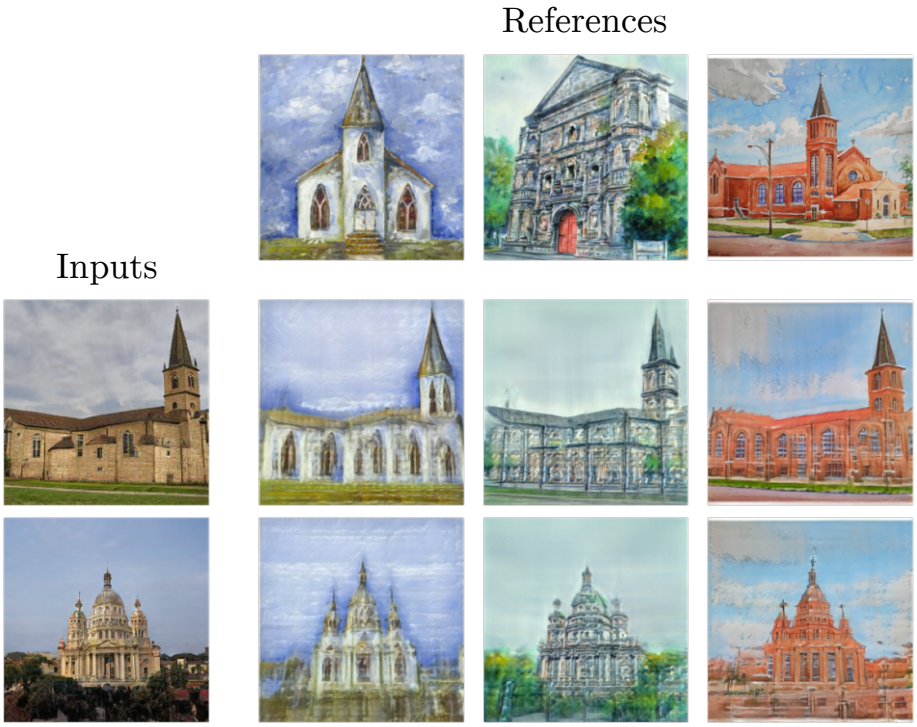

Fig. 21. JoJoGAN is a method to benefit from what a StyleGAN knows, and so should apply to other domains where a well-trained StyleGAN is available. Here we demonstrate JoJoGAN applied to LSUN-Churches.