Response to reviewers (*Submission ID #256*)

We thank all reviewers for their efforts and constructive comments. Below we give our detailed responses.

To Reviewer #1: A1. The "test time" of our model. Thanks for your comments. ***Training settings:*** Our model is trained on a large corpus and it can be applied to the identities that are not presented in the training set. Principally, the larger training set, the better generalization. ***Whether the target image is required:*** Yes, the target image is leveraged as a condition to realize the face swapping. ***Retrain or not:*** We do not retrain any modules in the testing phase. These details will be updated in the revised version.

A2. Limitations and failure cases. If a face swapping backbone cannot handle occlusion cases, the results produced by our model might have artifacts (see Fig. 1). These will be added in the revised version.

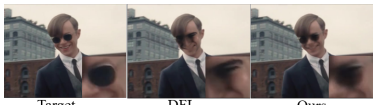

Figure 1: Failure cases.

A3. Minor flaws. Thanks for your careful review. We have revised all the typos and we will invite a native speaker to proofread our manuscript. The related references have been added in the revised version. The scores in Table 2 (b) represent "the rate at which users picked the result of the presented method" as you inferred. We will make it clear in the revised version.

To Reviewer #2: A1. The differences between Eq. (3) and Eq. (4). Eq.(4) is derived from $W = \sup_{\Psi \in \mathcal{F}^1}[\mathbb{E}_{v_r \sim \mathbb{V}_r}(\Psi(v_r)) - \mathbb{E}_{v_t \sim \mathbb{V}_t}(\Psi(v_t))]$, which is the dual form of Eq. (3). Hence, optimizing Eq. (4) will minimize the transportation cost implicitly. The detailed proof of the equality of Eq. (3) and Eq. (4) will be provided in the revised version. Additionally, Ω serves as the generator in WGAN. In WGAN's setting, both [1] and [2] agree that the generator aims to solve the optimal transportation. Finally, we also visualize the transportation as shown in Fig. 6.

A2. Relation between the "mixup" and the MSD. The "mixup" is a part of MSD. It is denoted as "MixLayer" in Fig. 2. As a data augmentation technique, mixup is originally proposed for empirical risk minimization. Differently, we use the "mixup" operation to mix two images, forcing the generator to produce more realistic results (see R2A3).

A3. Experiments on 'verification', 'pose', 'landmarks'. We have conducted the relative experiments, as reported in Table 1. Our method achieves comparable performance compared with pre-stage face swapping methods. It verifies that, as a post-processing step for face swapping, our method does not degrade the performance of previous methods.

A4. Empirical comparison against WGAN. MSD is leveraged to distinguish which part of an image is "fake", providing fine-grained information to guide the generator. The generator can produce smooth and well-blended results under this supervision. However, WGAN takes the whole image as a condition, thus the detailed information may be neglected. Experimental results in Table 2 verify the superiority of our method over WGAN.

A5. Time complexity. Please refer to supplementary material (C.4).

To Reviewer #3: A1. Compare to other fully automatic or end-to-end methods. Firstly, FSGAN is a fully automatic method, and we have already compared with it (Fig. 3, Fig. 4, and Table 1). In addition, we cannot provide the comparison with the other SOTA end-to-end method (FaceShifter [3]), since the official code is unavailable.

Table 1: Results of verification, pose, and landmark on both FF++ and DPF-1.0.

	verification↓		pose↓		landmark↓	
	FF++	DPF	FF++	DPF	FF++	DPF
DFL	0.231	0.243	3.161	3.940	3.073	3.289
+Ours	0.237	0.249	3.159	3.953	3.070	3.316
FSGAN	0.314	0.392	2.631	2.767	2.823	2.412
+Ours	0.317	0.389	2.638	2.762	2.831	2.409

A2. Why a two-stage computation is preferred. Due to the difficulties of modeling the complex appearance mapping directly, one-stage methods, which are required to process expressions, poses, and appearances simultaneously, tend to fail under complex lighting conditions. Therefore, a refinement is necessary. Furthermore, as emphasized by R1, as a post-processing model, our model can be easily concatenated with other reenactment models to form a fully automatic face swapping pipeline.

A3. Justify our model does not worsen the results from previous methods. The proposed perceptual encoder leverages 3D information to encode the semantic and geometric information, which is important to the generation. Also, a UNet-like architecture can preserve the earlier information with the skip connections. Furthermore, the proposed MSD can also improve synthesis performance. Although our model produces the blurry cases under some specific conditions (Please refer to R1A2), extensive quantitative and qualitative results (Fig. 3, Fig. 4, and Table 1) on a large-scale dataset verify the improvements over previous methods.

To Reviewer #4: A1. Analyses on identity preservation. As a post-processing stage for face swapping, our method mainly focuses on refining the reenactment results while not changing the generated identities of previous methods. Hence, identity preservation mainly relies on previous reenactment methods. Experimental results of "verification" on Table 1 verify that our method does not degrade the performance of the backbone methods.

Table 2: The evaluation on FF++.

	SSIM-e	SSIM-w	gram loss
WGAN	0.8089	0.7315	0.013667
Ours	**0.8301**	**0.7810**	**0.003578**

A2. Name of style loss. We will use "appearance loss" in the revised version.

A3. More challenging examples (such as cross-gender). Please refer to the supplementary material Fig. S5 (bottom row) and Fig. S6 (top row) for cross gender results. These figures also contain some results with large poses.

[1] Gulrajani et al. Improved Training of Wasserstein GANs, NeurIPS, 2017

[2] Lei et al. A Geometric View of Optimal Transportation and Generative Model, CAGD, 2019

[3] Li et al. FaceShifter: Towards High Fidelity and Occlusion Aware Face Swapping, CVPR, 2020