

## Appendix A. Details of the datasets

We use five publicly available datasets to demonstrate the effectiveness of our Magic Mask VG-GFace2 (Cao et al., 2018) dataset, CelebA-HQ (Karras et al., 2018) dataset, FF++ dataset (Rossler et al., 2019), MPIE dataset (Gross et al., 2010), and LPFF dataset (Wu et al., 2023). The detailed information of the five dataset is as follows:

**VGGFace2** contains 3.31 million images of 9131 subjects (identities), with an average of 362.6 images for each subject. Images are downloaded from Google Image Search and have large variations in pose, age, illumination, ethnicity and profession (e.g. actors, athletes, politicians). The whole dataset is split to a training set (including 8631 identities) and a test set (including 500 identities).

**CelebA-HQ** is a visually enhanced version of the CelebFaces Attributes dataset (CelebA) (Liu et al., 2015), and it provides 30,000 images with  $1024 \times 1024$  resolution. FF++ dataset is one of the most popular benchmarks for evaluating face identity-swapping methods, and it contains 1,000 video sequences.

**FF++** is a forensics dataset consisting of 1000 original video sequences that have been manipulated with four automated face manipulation methods: Deepfakes, Face2Face, FaceSwap and NeuralTextures. The data has been sourced from 977 youtube videos and all videos contain a trackable mostly frontal face without occlusions which enables automated tampering methods to generate realistic forgeries. As we provide binary masks the data can be used for image and video classification as well as segmentation. In addition, we provide 1000 Deepfakes models to generate and augment new data.

**LPFF** comprises 19,590 high-quality, numerous identities, and extensive-pose diversity images. They firstly collect 155,720 raw portrait images from Flickr, then they remove all the raw images that already appeared in FFHQ (Kazemi and Sullivan, 2014). After that, they align the remaining facial images and remove low-resolution images as well as noisy and blurred images.

**MPIE** contains over 750,000 images of 337 individuals. Each subject was photographed under 15 poses and 19 illumination conditions while exhibiting a range of facial expressions.

Table 4 shows the URLs to download the datasets that we used for this paper.

Dataset	Public repository
VGGFace2 Cao et al. (2018)	<a href="https://www.robots.ox.ac.uk/~vgg/data/vgg_face2/">https://www.robots.ox.ac.uk/~vgg/data/vgg_face2/</a>
CelebA-HQ Karras et al. (2018)	<a href="https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html">https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html</a>
LPFF (Wu et al., 2023)	<a href="https://github.com/oneThousand1000/LPFF-dataset">https://github.com/oneThousand1000/LPFF-dataset</a>
MPIE (Gross et al., 2010)	<a href="https://www.kaggle.com/datasets/aliates/multi-pie">https://www.kaggle.com/datasets/aliates/multi-pie</a>

Table 4: The URLs of the datasets used for this paper.

## Appendix B. List of public repositories.

We provide URLs of the public repositories of the methods that we selected for the experiment for the performance comparison on extreme face pose cases. Table 5 shows the list of the public repositories.

Method	Public repository
FSGAN <a href="#">Nirkin et al. (2019)</a>	<a href="https://github.com/YuvalNirkin/fsgan">https://github.com/YuvalNirkin/fsgan</a>
SimSwap <a href="#">Chen et al. (2020)</a>	<a href="https://github.com/neuralchen/SimSwap">https://github.com/neuralchen/SimSwap</a>
BlendFace <a href="#">(Shiohara et al., 2023)</a>	<a href="https://github.com/mapoon/BlendFace">https://github.com/mapoon/BlendFace</a>
HifiFace <a href="#">(Wang et al., 2021b)</a>	<a href="https://github.com/maum-ai/hififace">https://github.com/maum-ai/hififace</a>
FaceDancer <a href="#">(Rosberg et al., 2023)</a>	<a href="https://github.com/felixrosberg/FaceDancer">https://github.com/felixrosberg/FaceDancer</a>

Table 5: Quantitative results of on the MPIE dataset. <sup>†</sup> denotes that we ran officially released source codes to obtain the results.

Once we observed that, if the balancing weight for the ID loss is too low and lower than the reconstruction loss and the loss in charging of preserving the attribute, the model cannot generate an identity-swapped face well. Additionally, adversarial learning affects the pose and expression error. We can interpret that this circumstance has happened because the gradients of the total loss function lead the model to learn to regenerate the target attribute instead of enhancing the source identity representation. Additionally, adversarial learning impacts the quality of high-frequency details, detailed parts of the swept face.

## Appendix C. Balancing weight for the loss function

The setting for the balancing weight is crucial for achieving the best performance of our MagicMask. Our balancing weight settings (ID: 5.0, Recon: 2.0, AFAS 1.0) are based on various existing works ([Chen et al., 2020](#); [Shiohara et al., 2023](#); [Wang et al., 2021b](#)) which share similar architectural components and loss functions.

The balancing weight for ID loss is usually much higher than that for other losses (10 or 5), and the one for the reconstruction loss is between 5 and 1.0. Adversarial loss is usually 1.0. However, we couldn’t find suitable ablation studies about the balancing weight setting from those studies. We conducted ablation studies to capture the performance trend. Table 6 shows the performances of MagicMask depending on the setting of the balancing weights. The following results are the experimental data obtained to determine the values of the balancing weight. Those results are obtained from the MPIE dataset.

Loss type	Setting 1	Setting 2	Setting 3	Setting 4	Setting 5	Setting 6
ID loss	5.0	5.0	5.0	5.0	5.0	5.0
Recon loss	10	6.66	5.0	2.5	2.0	2.0
AFAS loss	0.0	0.0	0.0	1.0	0.0	1.0
Ratio (ID/Recon)	0.5	0.75 (Around)	1.0	2.0	2.5	2.5
CSIM	0.086	0.081	0.114	0.459	0.451	0.463
Pose error	4.12	4.11	4.15	3.36	4.07	3.35
Expr error	3.84	3.33	3.37	2.90	3.10	2.91

Table 6: Quantitative results of on the MPIE dataset. <sup>†</sup> denotes that we ran officially released source codes to obtain the results.

## Appendix D. Extended results on MPIE dataset

Figure 6 shows the extended results for face identity swapping on the MPIE dataset. These results are extended from the experiments on extreme poses described in Section 4. We can still observe that BlendFace generates some hallucinated faces that are totally mismatched with the actual face area. FSGAN results are significantly blurry but also sometimes do not change much. HifiFace’s results contain high contrast, making their swapped results totally disrupted. SimSwap achieves competitive performance, but in extreme poses, the boundaries between face and background are not clear enough.

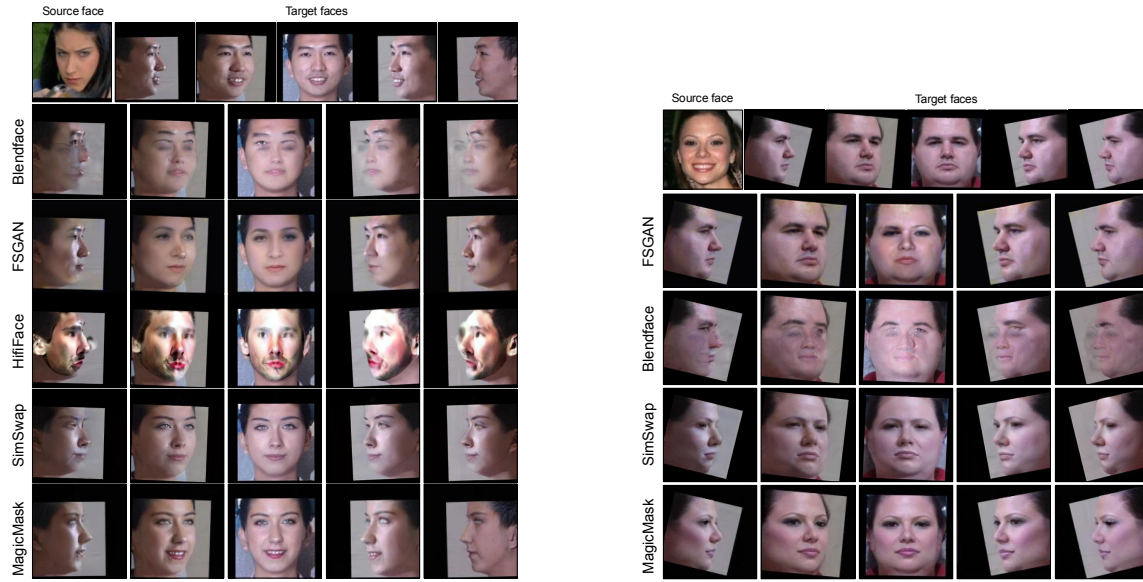


Figure 6: The face identity swapping result of the Magic Mask and other methods (Shiohara et al., 2023; Nirkin et al., 2019; Wang et al., 2021b; Chen et al., 2020) on MPIE dataset.

## Appendix E. Extended results on LPFF dataset

Figure 7 shows the extended results for face identity swapping on the LPFF dataset. These results are extended from the experiments on extreme poses described in Section 4.

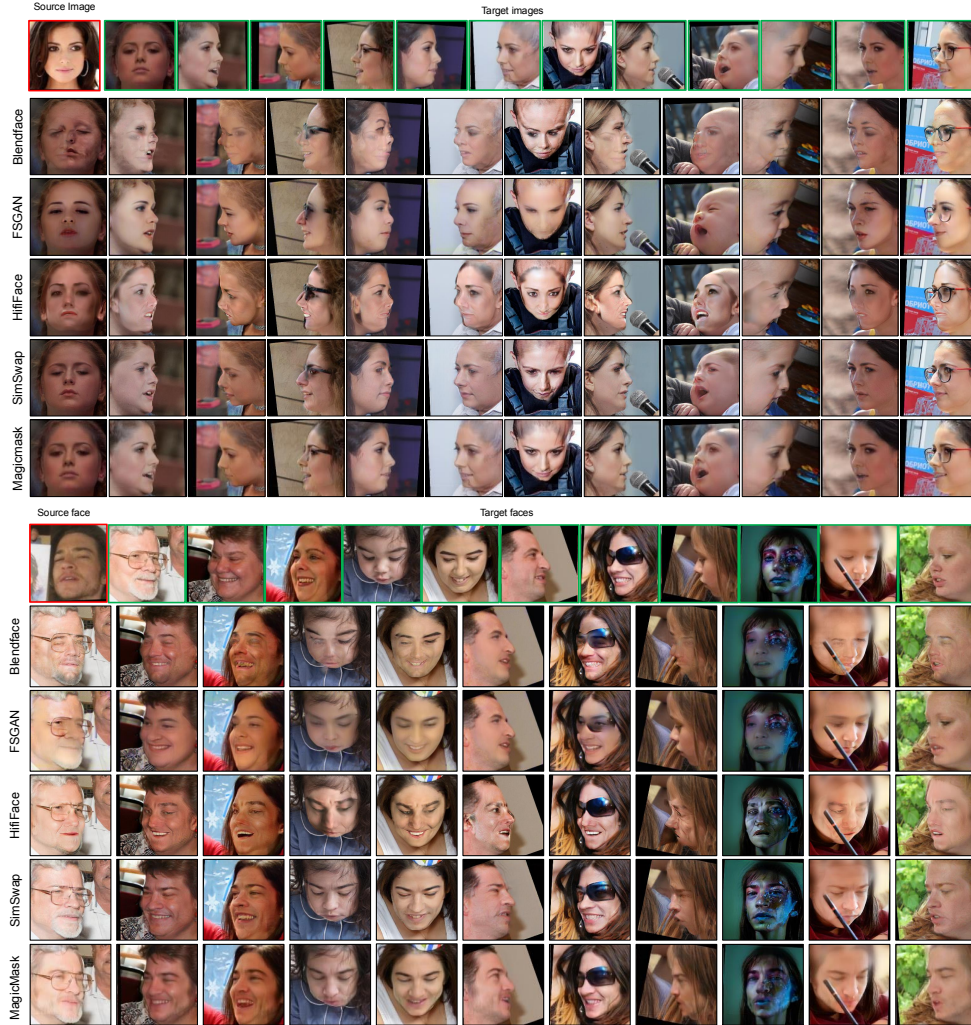


Figure 7: The face identity swapping result of the Magic Mask and other methods (Shiohara et al., 2023; Nirkin et al., 2019; Wang et al., 2021b; Chen et al., 2020) on LPFF dataset.