Zero-Shot Head Swapping in Real-World Scenarios

Supplementary Material

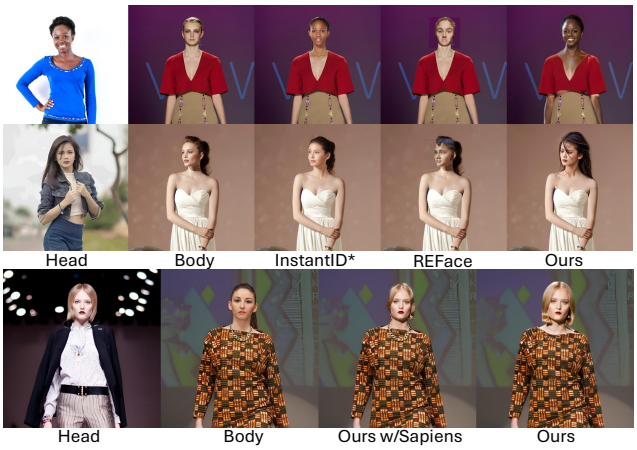

Figure 1. Additional qualitative comparisons between InstantID and Ours, as well as Ours with Sapiens mask and Ours.

1. Implementation Detail

1.1. Training

Training is implemented using the HuggingFace diffusers library [**?**] with data type casting to bfloat16 to enhance training efficiency. Additionally, it is conducted on 8 A100 40GB GPUs over 210,000 steps, using a batch size of 2 per GPU. To fuse the embeddings generated by the pretrained hair encoder, which are in the shape [18, 512], a 1×1 convolutional layer is employed to reshape them to [1, 512]. To train the Hair Fusion model and convolutional layer, the AdamW optimizer [**?**] is used with a constant learning rate of 0.00001 and a weight decay of 0.01.

1.2. IOMask

IOMask is used to extract precise masks for head generation. We set the threshold $\tau = 0.6$ and perform 40 denoising steps out of 50, corresponding to 80% noise levels. This approach ensures robust mask generation tailored to the head-swap task without requiring preprocessing steps like alignment or cropping.

2. Problem Definition

Face swapping can be defined in various ways, such as replacing only the eyes, nose, and mouth while preserving the original face shape. In our case, we define a "head swap" as the task of swapping only the head, which is considered part of the human body. Consequently, long beards should be reflected in the swap, whereas accessories should not. To ensure that accessories are not swapped, we refined the training dataset accordingly. As shown in Fig.7 in manuscript,

	Head (crop)		Hair (crop)		ID sim ↑	Head Pose ↓	Recon. ↑
	LPIPS↓	CLIP-I ↑	LPIPS↓	CLIP-I ↑			
REFace	0.4932	0.7430	0.3189	0.8312	0.5197	**20.33**	**0.6397**
InstantID*	0.4802	0.8090	0.3251	0.8470	0.2821	20.64	0.6308
Ours	**0.4567**	**0.8615**	**0.2966**	**0.8644**	**0.5546**	21.66	0.6286

Table 1. Quantitative comparison.

facial hair is partially preserved in the swap to a reasonable extent. Moreover, for a seamless transition, it is necessary to process not only skin of the neck area, which connects to the head, but also the hands. Our IOMask has the potential to enable this process as shown in Fig. 1, first row.

3. Analysis on Qualitative Results

The reason parts other than the head are sometimes altered, called hallucinations or artifacts, is that the mask leaves those areas uncovered. As mentioned in the Sec.1, if we want to generate a head independent of the original body's head, using predicted noise's difference methods are effective. However, these methods have a drawback: they often open up areas that do not need to be changed. IOMask helps minimize these issues while still preserving the advantages of such approaches as shown in Fig. 1, second and third row. Note that REFace is a method that uses the crop-and-paste-back approach, ensuring that other parts of the image remain unchanged which has its own limitations discussed in Sec.1. When using an off-the-shelf mask, it only fills the head region of I_b, which can lead to issues such as the inability to generate long hair or the appearance of sharp skin boundaries, as shown in Fig.1. In contrast, IOMask allows for more flexible and natural generation.

We use a dataset that includes the entire upper body. Therefore, rather than providing only head orientation information, we utilize OpenPose, which offers richer information to ensure better alignment with the body. It is also worth noting that, as shown in Fig.5, openpose has difficulty capturing fine details of hands [**?**]. Since our focus is on head swapping, this limitation lies outside the scope of our work. However, it is an important aspect to consider for future improvements, particularly in full-body swap scenarios.

4. Additional Quantitative Comparisons

We additionally evaluated our method using four metrics: ID similarity with the face recognition model [**?**] head pose L2 distance with the head pose estimation network, HopeNet [**?**], LIPIS and CLIP-I for cropped images to compare with REFace, as well as a reconstruction metric, MSE, to assess the degree of preservation outside of the

head area. Furthermore, we conducted an additional comparison with SDXL Inpainting InstantID [**?**], using Sapiens [**?**] mask, with openpose ControlNet. InstantID not only struggles to preserve identity compared to our method but also fails to transfer hair. Our method outperformed in all aspects except for head pose and reconstruction. The weaknesses observed in head pose and reconstruction can be attributed to the trade-off between flexibility and preservation, which arises from the different masking methods.

5. Additional Qualitative Results

We provide additional qualitative results generated by our proposed approach.

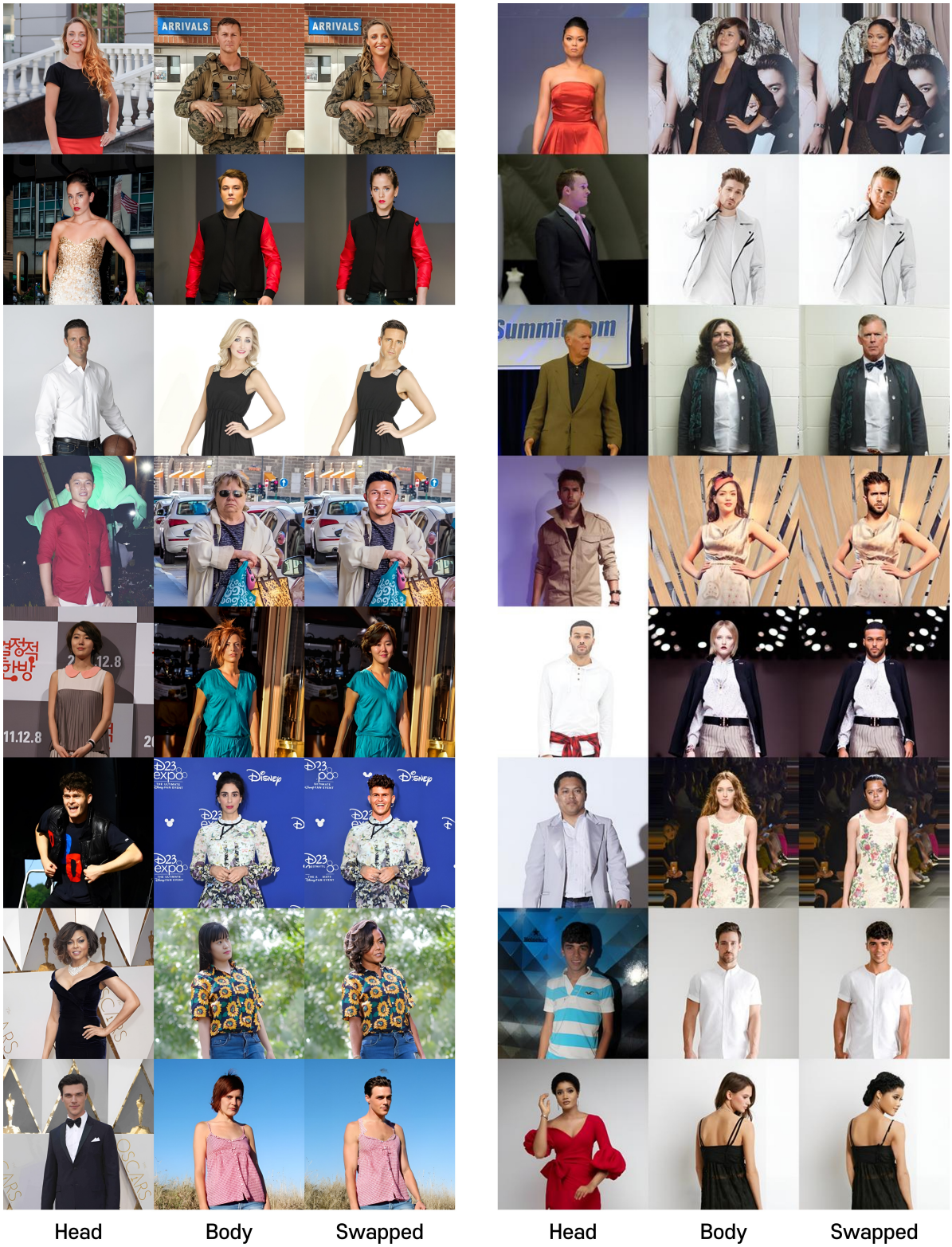

Figure 2. **Additional qualitative results.** The head in the images of *Head* column is seamlessly combined with the body in the images of *Body* column by the proposed method, HID, resulting in head-swapped images in the *Swapped* column.