CanonSwap: High-Fidelity and Consistent Video Face Swapping via Canonical Space Modulation

Supplementary Material

A. Training Details

Our method is implemented in PyTorch and trained on two NVIDIA A6000 GPUs, with a batch size of 6 per GPU. We use the AdamW optimizer (weight decay = 1×10^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.999$) for both generator and discriminator and set the initial learning rate to 1×10^{-4}. The model is trained for 150k steps in total.

For discriminator, we adopt the same architecture as SPADE. During training, we introduce an additional gradient penalty loss, which enforces smooth decision boundaries by penalizing large gradients in the discriminator. This penalty stabilizes training and helps the discriminator better distinguish between real and generated samples.

B. Visualization of Canonical Space

To provide an intuitive illustration of how our canonical space appears after motion decoupling, we randomly select 10k frames from our CVF benchmark and apply a crop-and-align procedure to obtain *Align Set*. Next, we transform the images in *Align Set* into the canonical space, yielding *+Canonical Set*. We then use a face segmentation model to compute the average parsing map for each set, as well as individual nose, eyes, and mouth regions, and visualize the results in Fig. 8. As shown, the canonical space removes motion information, causing facial features to align almost perfectly. By contrast, the standard alignment method still contains motion, resulting in blurred parsing boundaries—particularly around the eyes, which can shift over a wide range.

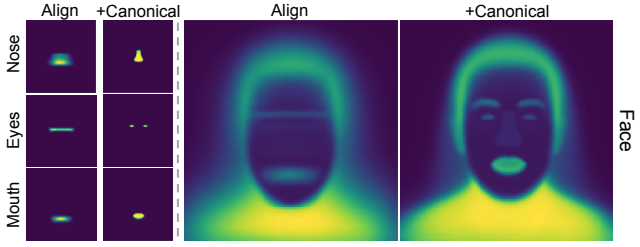

Figure 8. Comparison between traditional face alignment (left) and our canonical-space transformation (right), visualized by averaging segmentation maps across multiple samples. In traditional alignment, residual motion information causes blurred and inconsistent boundaries. By contrast, our canonical-space transformation effectively decouples motion, resulting in more uniform and clearly defined facial regions.

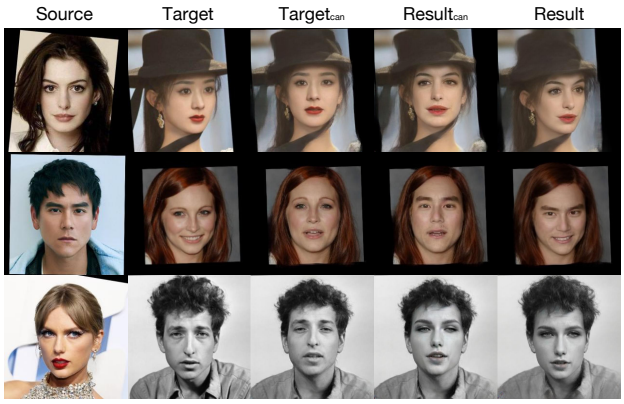

Figure 9. Visualization of outputs of each stage of CanonSwap.

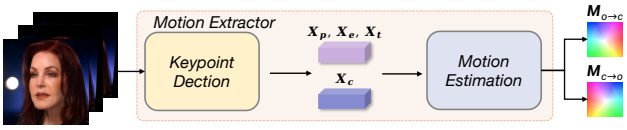

Figure 10. The components of our motion extractor.

Furthermore, we visualize the outputs of each stage of CanonSwap, as shown in Fig. 9.

C. Details of Motion Extractor

The details of motion extractor is shown in Fig. 10, specifically, for a frame of the target video in the original space V_o, we use an implicit keypoint detector to obtain the canonical keypoints $X_c \in \mathbb{R}^{n\times 3}$, along with motion deformations, which include pose rotation $X_p \in \mathbb{R}^{n\times 3}$, expression $X_e \in \mathbb{R}^{n\times 3}$, and translations $X_t \in \mathbb{R}^3$, where n denotes the number of keypoints. Using these components, the keypoints for the frame are computed as:

$$X = X_c X_p + X_e + X_t. \tag{12}$$

Then, we feed X and X_c into a motion estimation module $\mathcal{E}$ to estimate motion information. By swapping the order of X and X_c, we can simultaneously obtain the deformations from the original space to the canonical space $M_{o\to c}$, and from the canonical space back to the original space $M_{c\to o}$:

$$M_{o\to c} = \mathcal{E}(X, X_c), \quad M_{c\to o} = \mathcal{E}(X_c, X). \tag{13}$$

Figure 11. more qualitative results through a face matrix.

D. Advantages of the PIM.

Our PIM module addresses a key drawback of traditional AdaIN/modulation-based methods—their global application alters identity-irrelevant regions, which is suboptimal for face swapping. This often leads to (1) **visible artifacts** and (2) **unstable training from conflicting losses**, the latter often overlooked. We compare AdaIN, global modulation, and PIM under the same setting. As shown in Fig. 12, PIM converges faster and alleviates the conflict between identity loss and perceptual loss (lowest ID loss and lowest perceptual loss), resulting in better overall performance and a higher optimization ceiling.

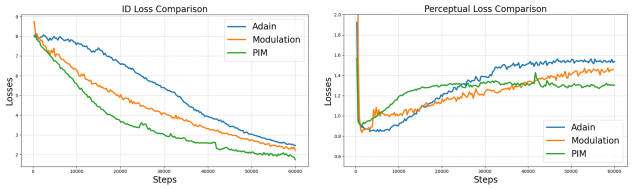

Figure 12. Training loss curves on the same dataset, our PIM achieves the fastest convergence rate and demonstrates lower ID Loss and Perceptual Loss, effectively mitigating the adversarial relationship between losses and achieving a higher performance ceiling.

E. Computational Efficiency

We evaluate inference efficiency by comparing our method with existing approaches, as shown in the table below (FPS). Our methods is faster than Diffusion/StyleGAN-based methods.

Metrics	Simswap	FSGAN	E4S	Diffswap	FaceAdapter	REFace	Ours
FPS	16	21	4	0.11	0.35	0.21	14

F. Face Swapping and Animation

To achieve face swapping and animation, we need to change the warping back deformation $M_{c\to o}$ in Eq. 13. Specifically, we obtain X_c, X_p, X_e, and X_t from the target frame,

Figure 13. By exchanging canonical keypoints, our method can also achieve shape transfer to some extent.

Figure 14. Qualitative comparison in large pose variation situation.

and also extract the source's expression from the source frame. During the transformation from the original space to the canonical space, we follow the procedure described in the main text. In the warp-back stage, we compute a new keypoint X' as

$$X' = X_p X_c + X_e^s + X_t, \tag{14}$$

where X_e^s denotes the source's expression. We then use the motion estimator to obtain a new warp deformation,

$$M'_{c\to o} = \mathcal{E}\big(X_c, X_2\big), \tag{15}$$

and apply it to warp back, thereby transferring the source expression to the target image.

G. More Qualitative Results

To demonstrate the robustness of our model, we conducted a matrix swap, and the results are shown in Fig. 11. Furthermore, compared to existing face swapping methods, our approach can leverage powerful animation priors to maintain robust performance under large pose variations. Moreover, by replacing the target's canonical keypoints with those of the source, the facial geometry can be adaptively aligned to match the source's structure to some extent, which is shown in Fig. 13. We also conduct an evaluation in large pose variation situation, which is shown in Fig. 14. Warping-based animation (e.g., talking head) may struggle with extreme pose variations due to insufficient target-pose features. In contrast, CanonSwap performs face swapping in a canonical pose and warps back to the original pose while preserving the original pose features. This enables robust performance under large pose variation. As shown in Fig. (a), CanonSwap outperforms prior methods in such scenarios, where SimSwap typically fails to handle large pose differences.

H. Ethical Considerations

This research is conducted solely for academic purposes and to advance the video face swapping technology. We use publicly available datasets and adhere to ethical guidelines in our experimentation. While our work aims to improve the fidelity and temporal consistency of face swapping, we acknowledge the potential for misuse in applications such as deepfakes and identity manipulation. We strongly advocate for responsible use of this technology and caution against applications that may infringe on privacy, consent, or intellectual property rights. Researchers and practitioners are encouraged to consider the ethical implications and to implement safeguards to prevent harmful or deceptive uses of our methods.