Figure 1: Synthesized frontal images without using Convolutional Refinement Head.

1 Supplementary Material

1.1 Convolutional Refinement Head

A common limitation of Vision Transformer (ViT)–based image synthesis models is the presence of visible patch boundaries in the reconstructed outputs. Since ViTs operate on non-overlapping image patches, the direct reassembly of decoded patch embeddings often results in "blocky" artifacts or discontinuities at patch borders (see Fig. 1 for the synthesized test images with such artifacts). To address this issue, we introduce a convolutional refinement head that transforms the decoded transformer features into smooth, high-fidelity frontal face images.

1.2 Qualitative Results on Multi-Pie Dataset

Figures 2-5 present the outputs of the evaluated frontalization methods for images belonging to different individuals. The first row shows the non-frontal input images, while the second row presents the ground-truth masks used for evaluation. It should be noted that the CMU Multi-PIE evaluation protocol requires the use of frontal images in the neutral pose for evaluation.

For evaluation, CNN features are extracted from the synthesized frontal images and corresponding masks by using the ArcFace method. The cosine distance between these CNN features is then used for classification. As shown in the figures, the proposed method successfully synthesizes frontal face images under varying illumination conditions and poses with high accuracy. However, performance slightly degrades for images captured from a top-view perspective, as seen in the last three columns of the figures. In these cases, some identity information is lost. Since such samples are relatively limited in the training set, the proposed method struggles to accurately frontalize these types of extreme non-frontal images.

Among the compared methods, DR-GAN is also capable of synthesizing face images that resemble the original individuals. However, the remaining methods generally fail to preserve identity. For instance, pSp even fails to maintain gender information in some cases, as illustrated in Fig. 5. FFWM also struggles to generate proper frontal views and often produces distorted face images.

1.3 Tests on Images Collected from Web Environments

We also tested frontalization methods on high-resolution non-frontal images extracted from web environment. Fig. 6 presents qualitative frontalization results for several non-frontal face images collected from uncontrolled web environments. The first row shows the input images exhibiting significant pose variations. Since these images are obtained from web sources, no ground-truth

frontal images are available for direct comparison. The remaining rows illustrate the frontalized outputs produced by different methods, including FrontalViT, DR-GAN, FFWM, CR-GAN, and pSp.

The visual comparison highlights differences in the ability of the evaluated approaches to normalize pose while preserving identity and maintaining visual realism. The proposed FrontalViT method consistently generates visually plausible frontal faces with well-preserved facial structures and identity-related features. The synthesized images retain natural texture details and exhibit stable geometric consistency across different pose conditions. Similarly, DR-GAN produces relatively convincing frontal reconstructions and maintains a reasonable balance between pose correction and identity preservation.

In contrast, FFWM generates satisfactory frontal images for several examples but demonstrates inconsistent behavior when the pose deviation becomes more extreme. In particular, the method fails to properly frontalize some inputs, such as those shown in the 1st, 10th, 12th, and 13th columns, where noticeable artifacts or incomplete pose normalization can be observed. Furthermore, CR-GAN and pSp show clear limitations in preserving identity information. The generated images frequently exhibit distortions or identity inconsistencies, suggesting that these methods struggle to maintain discriminative facial characteristics under large pose variations.

Another important point to emphasize is that many of the evaluated frontalization methods, particularly FFWM and CR-GAN, fail to properly capture facial symmetry, which is a key characteristic of frontal face images. Although several existing approaches attempt to enforce symmetry through specially designed loss functions, these constraints are often insufficient to ensure consistent symmetric reconstruction. In contrast, the proposed method is able to better capture facial symmetry due to the self-attention mechanism of the transformer architecture, which allows the model to learn long-range relationships between symmetric facial regions. As illustrated in the figure, our method synthesizes largely symmetric frontal faces, with minor asymmetries appearing mainly in the hair region, where natural variability and occlusions are more common.

Overall, the qualitative results suggest that the proposed approach produces more stable and visually coherent frontal faces while better preserving identity information compared to the competing methods, even in the absence of ground-truth frontal references.

Figure 2: Visualization of frontalization results for selected non-frontal face images belonging a particular person. The first row presents the input images with pose variations, while the second row shows the corresponding ground-truth frontal references. The subsequent rows display the outputs produced by the evaluated methods.

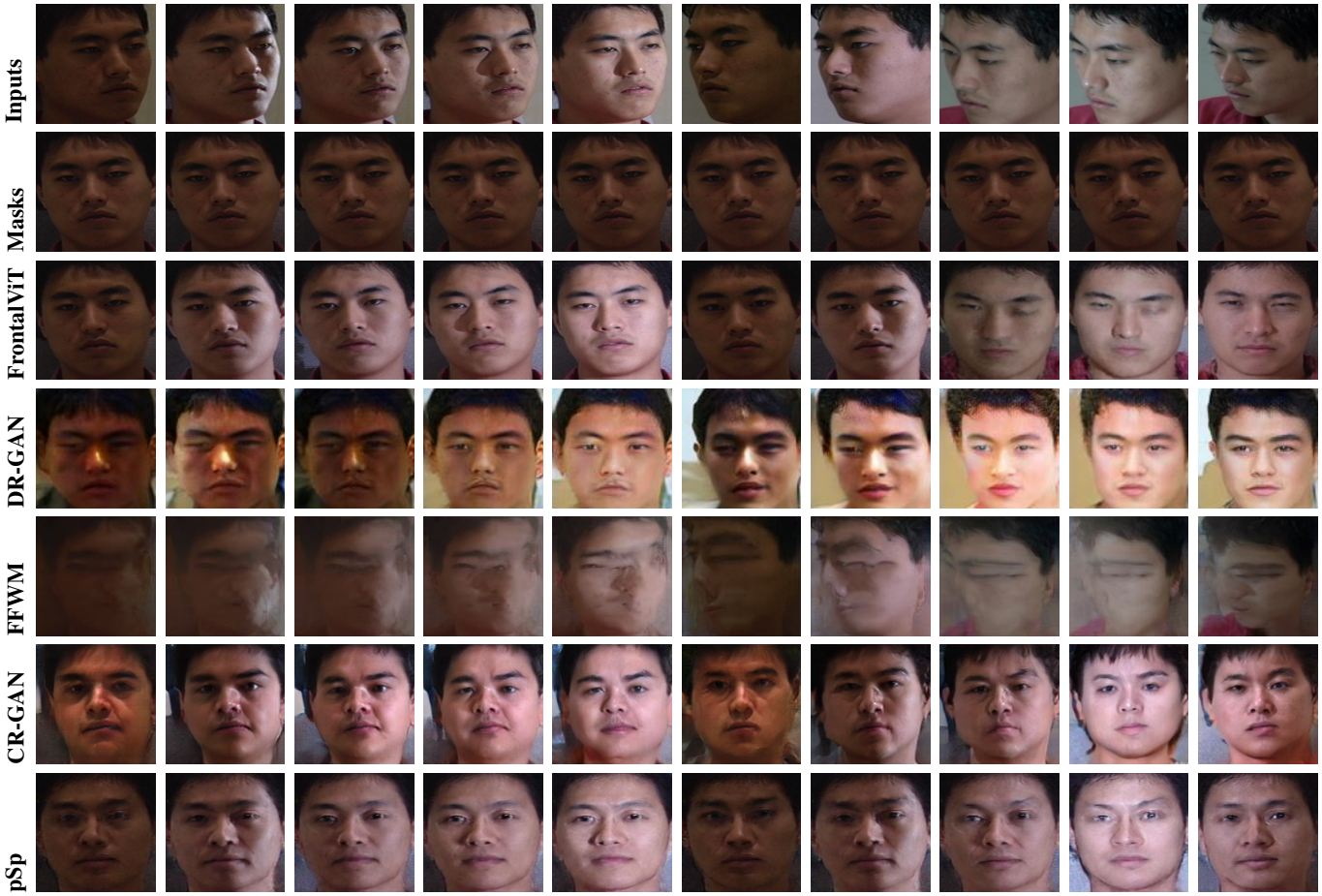

Figure 3: Visualization of frontalization results for selected non-frontal face images belonging a particular person. The first row presents the input images with pose variations, while the second row shows the corresponding ground-truth frontal references. The subsequent rows display the outputs produced by the evaluated methods.

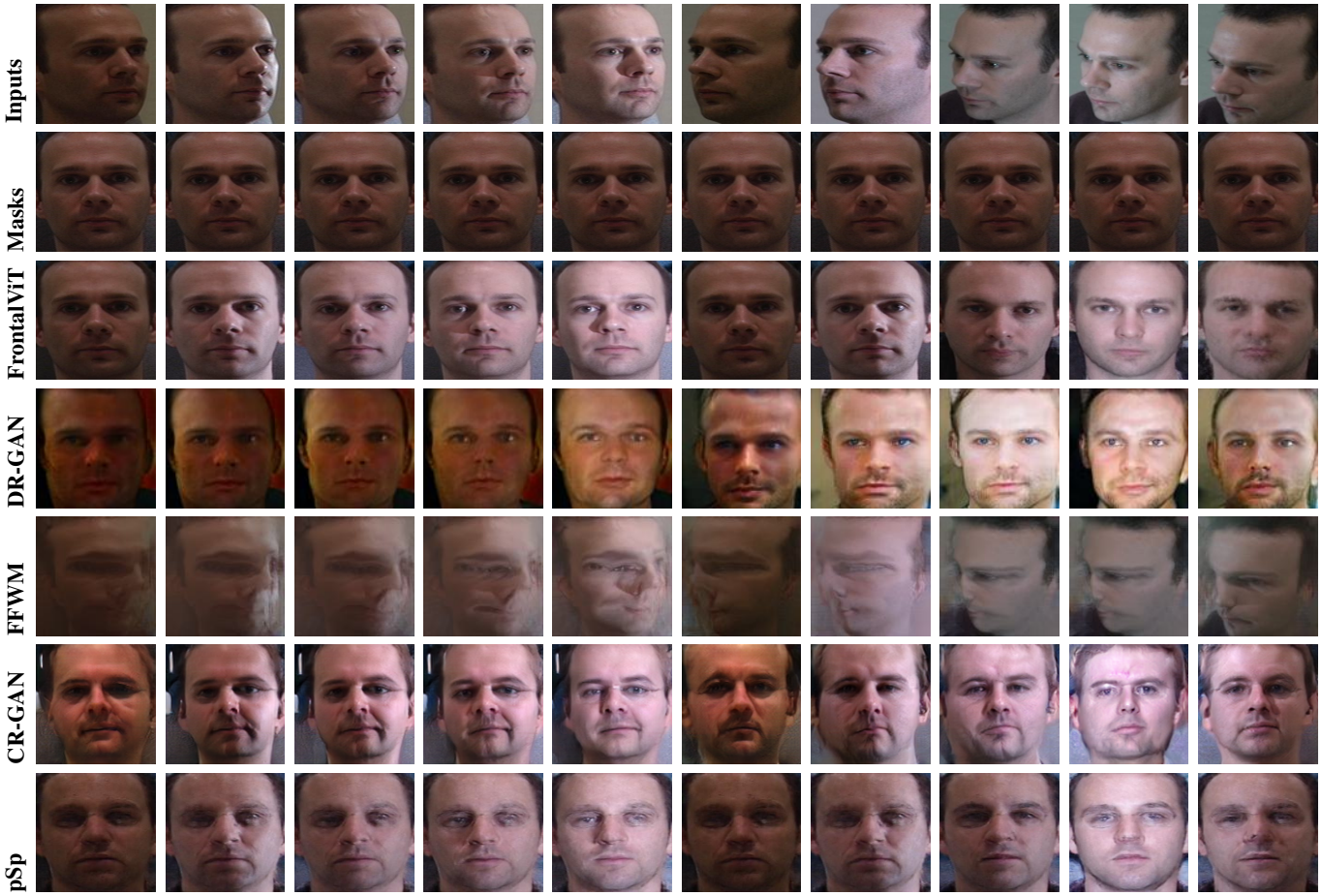

Figure 4: Visualization of frontalization results for selected non-frontal face images belonging a particular person. The first row presents the input images with pose variations, while the second row shows the corresponding ground-truth frontal references. The subsequent rows display the outputs produced by the evaluated methods.

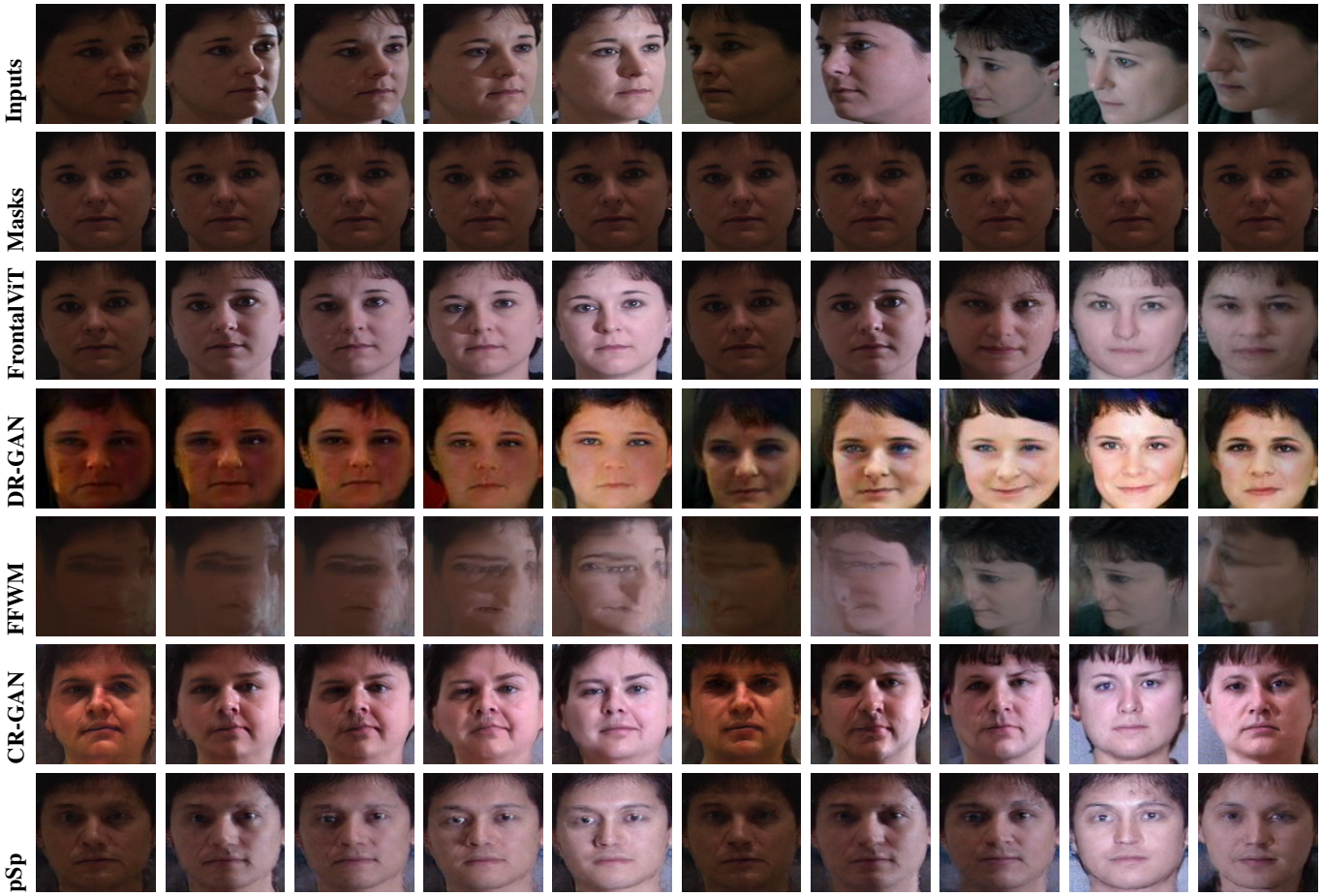

Figure 5: Visualization of frontalization results for selected non-frontal face images belonging a particular person. The first row presents the input images with pose variations, while the second row shows the corresponding ground-truth frontal references. The subsequent rows display the outputs produced by the evaluated methods.

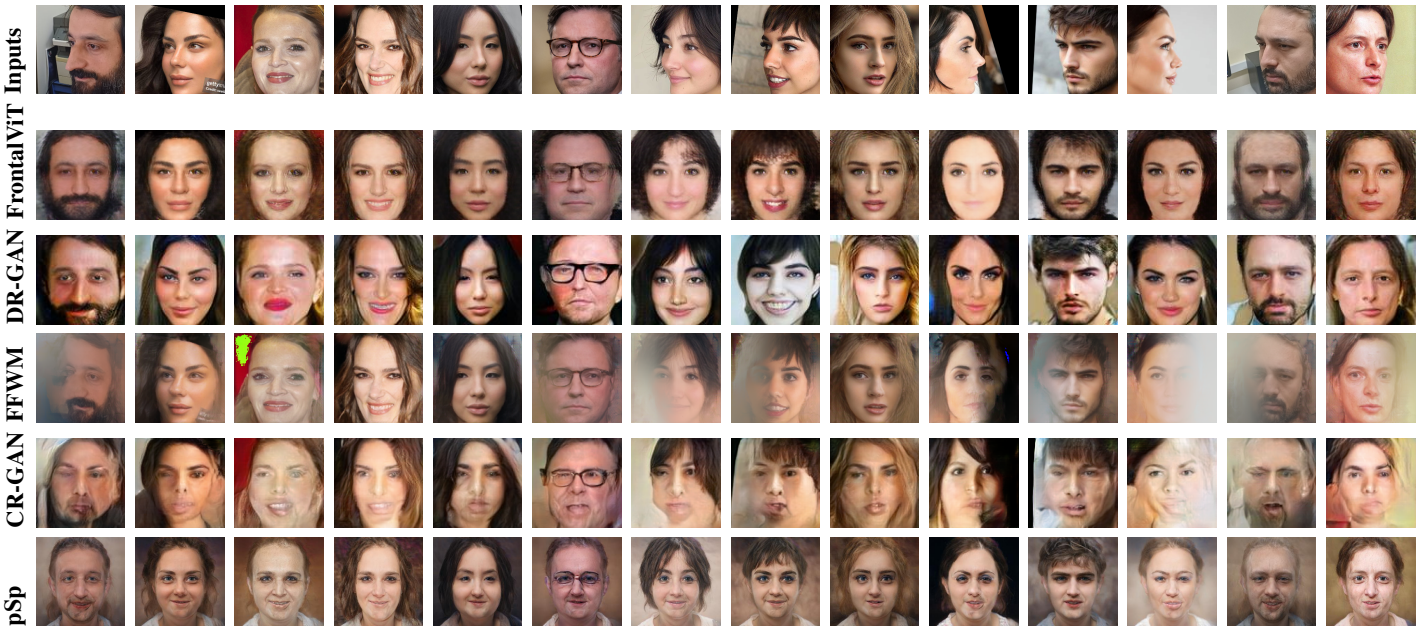

Figure 6: Visualization of frontalization results for several non-frontal face images collected from uncontrolled web environments. The first row presents the input images with pose variations, while the remaining rows display the outputs produced by the evaluated methods.