IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Unified Diffusion Models For Realistic And Efficient Face Swapping

Guided by :

Prof. Jose V Francis

Assistant Professor, CC

Presented by :

Sabarrinath S

S7, CC

Abstract

Face swapping has become an important area of research in computer vision. It is driven by uses in entertainment, privacy, virtual reality, and digital forensics. This review summarizes progress from ten recent studies focusing on GAN-based, diffusion-based, and hybrid models. These studies tackle issues like high fidelity, keeping identities intact, handling occlusions, pose invariance, and scalability.

Early GAN-based models such as SimSwap, FaceShifter, and FaceDancer introduced new techniques like identity injection, adaptive embedding fusion, and occlusion-aware refinement. These methods aim to balance transferring the source identity with preserving target attributes. High-resolution techniques like MegaFS used hierarchical latent encoding and StyleGAN2 synthesis for megapixel swaps. In contrast, low-resolution and occlusion-focused methods applied cross-resolution contrastive learning and special inpainting pipelines.

Diffusion-based techniques, including DiffSwap, DiffFace, and unified inpainting frameworks, have improved stability, control, and image realism. They achieved this through identity-guided sampling, 3D-aware masking, and CLIP feature disentanglement. Together, these advancements show significant development in creating realistic, high-resolution, and attribute-consistent swapped faces. However, there are still challenges with extreme pose variation, detailed preservation, and real-time efficiency. This review discusses the strengths and weaknesses of these approaches and suggests future research directions for creating reliable, ethical, and controllable face swapping systems.

List of Keywords

Face swapping, identity preservation, attribute preservation, generative adversarial networks (GANs), diffusion models, high-fidelity synthesis, occlusion handling, pose invariance, hierarchical latent encoding, 3D-aware masking, inpainting, high-resolution image synthesis, real-time processing, controllable face editing.

Introduction

Face swapping is the process of transferring a source identity onto a target face while keeping the target's pose, expression, lighting, and background intact. This technique has attracted significant attention in computer vision because of its uses in entertainment, augmented reality, privacy protection, and digital forensics. Recent progress in deep learning has led to big improvements in visual quality, identity preservation, occlusion handling, and scalability.

Generative Adversarial Network (GAN)-based methods like SimSwap and SimSwap++ have introduced feature-level identity injection and new loss functions for balanced identity and attribute preservation. FaceShifter has combined adaptive embedding integration with an occlusion-aware refinement network to enhance realism. FaceDancer has used adaptive feature fusion attention for single-stage high-quality synthesis. MegaFS has expanded face swapping to megapixel resolutions through hierarchical latent encoding and StyleGAN2-based generation. For tough real-world scenarios, low-resolution and occlusion-aware frameworks have applied cross-resolution contrastive learning and targeted inpainting to maintain identity under degraded input quality.

More recently, diffusion-based methods, including DiffSwap, DiffFace, and unified inpainting frameworks, have shown better stability, controllability, and detail preservation by using identity-guided sampling, 3D-aware masking, and CLIP-based feature disentanglement. This review looks at ten representative works, comparing methodologies, datasets, and performance measures, while also highlighting ongoing challenges. These include managing extreme pose variations, preserving fine details, and enabling real-time deployment.

Literature Review

1. Realistic and Efficient Face Swapping: A Unified Approach with Diffusion Models

The study by Sanoojan Baliah, Qinliang Lin, Shengcai Liao, Xiaodan Liang, Muhammad Haris Khan proposed face swapping aims to transfer the identity from a source face to a target image while keeping the target's pose, expression, and background intact. Early GAN-based methods like FSGAN [18] and HifiFace [30] improved realism through reenactment and structural guides. In contrast, SimSwap [3] and FaceShifter [15] developed designs that do not rely on priors, which helped with better attribute preservation. StyleGAN-based approaches [7, 17, 31, 36] provided high-resolution results but were still sensitive to pose, lighting, and occlusion, often resulting in artifacts.

Recent developments have focused on Diffusion Models due to their reliability and accuracy. DiffFace [12] incorporated identity embeddings with facial guidance but depended heavily on optimization during inference. This caused slow generation and occasional noise. DiffSwap [35] redefined face swapping as conditional inpainting, which enhanced shape consistency. However, it still executed key blending steps during inference, limiting efficiency and identity transfer.

Current unified methods, like the work by Baliah et al., tackle these challenges by redefining face swapping as a training-time inpainting task. This uses CLIP-based feature separation to maintain pose, expression, and lighting. It also introduces mask shuffling for various applications, including head swapping. By shifting blending to the training phase, these methods achieve faster inference, fewer artifacts, and better results under different conditions.

2. SimSwap++: Towards Faster and High-Quality Identity Swapping

The study by Xuanhong Chen , Bingbing Ni , Yutian Liu, Naiyuan Liu, Zhilin Zeng, and Hang Wang present the Face identity editing (FIE) is about transferring identity from a source image to a target while keeping non-identity features like pose, lighting, and expression. Early work in this area used 3D Morphable Models (3DMM) for geometry-based attribute transfer, as seen in Face2Face by Nirkin et al. This method created realistic reenactments but faced high computational costs and issues with facial reconstruction accuracy.

With the rise of Generative Adversarial Networks (GANs), methods like FSGAN, FaceShifter, and HifiFace began focusing on blending features. These methods improved realism but often needed large, complex architectures, which made them tough to use in real-time or on mobile devices. SimSwap provided a lighter option, introducing identity injection and weak feature matching to better preserve attributes without needing training on specific subjects. However, similar to many GAN-based methods, SimSwap had trouble generating high-resolution images because of the lack of suitable datasets and inefficiencies in computation.

High-resolution editing methods usually relied on StyleGAN priors, such as MegaFS and FSLSD-HiRes, which allowed for finer details but dealt with issues like latent space entanglement and poor generalization to real-world data. Dataset limitations also held back advancements; popular sets like VGGFace2 offered low-resolution, noisy images, while higher-quality datasets like CelebA-HQ and FFHQ lacked diversity in pose and expression.

3. SimSwap: An Efficient Framework For High Fidelity Face Swapping

The study by Renwang Chen, Xuanhong Chen, Bingbing Ni1, and Yanhao Ge about face swapping is the process of transferring the identity of a source face to a target face while keeping the target's pose, expression, lighting, and other features intact. Traditional methods depended on 3D Morphable Models (3DMMs) to align geometry and combine appearances (Blanz et al., 2004; Bitouk et al., 2008). However, these methods often struggled with accurately reproducing expressions, were costly in terms of computation, and had limited generalization.

The rise of Generative Adversarial Networks (GANs) transformed face swapping by allowing the creation of high-quality, data-driven images. Source-oriented methods like FSGAN (Nirkin et al., 2019) used a two-step process: face reenactment followed by inpainting to apply the source identity while matching the target features. Still, they were sensitive to the source image's posture and lighting. Target-oriented methods changed the deep features of the target image directly. DeepFakes made encoder-decoder architectures popular for identity-specific swapping, while IPGAN (Bao et al., 2018) and FSNet (Natsume et al., 2018) introduced ways to separate identity and attributes for more versatile swapping. FaceShifter (Li et al., 2019) achieved top results in identity transfer through a complex two-step approach. SimSwap showed better attribute preservation than FaceShifter and FSGAN while keeping a high level of identity accuracy (92.83% ID retrieval on FaceForensics++). Its effectiveness with various poses, lighting, and expressions emphasizes the need to balance identity changes with keeping attributes in modern face swapping systems.

4. One Shot Face Swapping on Megapixels

The study by Yuhao Zhu, Qi Li, Jian Wang, Chengzhong Xu, Zhenan Sun says that face swapping aims to transfer the identity of a source face to a target image while keeping the target's pose, expression, and background intact. Traditional methods focused on specific subjects, like DeepFakes and Disney Research's high-resolution swapping framework, need training on fixed source-target pairs. This requirement leads to limited generalization and high training costs. On the other hand, subject-agnostic methods such as FSNet, IPGAN, FSGAN, and FaceShifter work without needing to retrain for new

identities, which allows for broader use. Even though FaceShifter provides strong identity transfer with occlusion-aware designs, it still faces challenges in preserving high-resolution details and ensuring stable training.

By training these modules separately, MegaFS lowers GPU memory needs and keeps training stable, allowing for 1024×1024 resolution face swapping. This is a significant improvement over previous methods, which were usually limited to 256×256. Quantitative evaluations show that MegaFS offers better identity preservation and similar visual realism. It also has the added benefit of releasing the first megapixel face swapping dataset to aid DeepFake detection and image editing research.

5. FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping

The study by Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen and Fang Wen proposed a face swapping replaces a person's identity in a target image with that of a source image while maintaining target-specific details like pose, expression, lighting, and background. Early methods focused on direct replacement of the inner facial region but struggled with variations in pose and perspective. 3D-based methods used 3D Morphable Models (plate_number_1) to handle differences in pose and viewpoint; however, they often had issues with reconstruction accuracy, robustness in real-world situations, and capturing fine details like lighting and occlusions.

With the development of GAN-based methods, face swapping quality improved significantly. Approaches like RSGAN, FSNet, and IPGAN aimed to separate identity and attributes into different latent representations for recombination. However, their compressed feature representations caused a loss of detail and reduced realism. FSGAN introduced a face segmentation network that considers occlusions to maintain target occlusions, but it did not fully capture other target-specific features like lighting or image resolution.

FaceShifter achieves top identity preservation with a 97.38% ID retrieval on FaceForensics++ and offers better visual realism compared to FaceSwap, Nirkin et al., DeepFakes, IPGAN, and FSGAN. By combining adaptive identity-attribute integration with strong occlusion recovery, FaceShifter establishes a new standard in high-fidelity, real-world face swapping.

6. FaceDancer: Pose- and Occlusion-Aware High Fidelity Face Swapping

The study by Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso-Fernandez, Cristofer Englund says that face swapping involves changing the identity of one face to that of another while keeping specific features of the original face, such as pose, expression, lighting, and occlusions. Traditional methods that focus on the source face, like 3D Morphable Model-based techniques (for example, Nirkin et al., 2018; Blanz et al., 2004) and FSGAN (Nirkin et al., 2019), first modify the source face to match the target's attributes before combining them. Although these methods work well in some situations, they often have difficulty with complex lighting, occlusions, and maintaining fine textures.

Methods that focus on the target image use its features directly, adjusting them with embeddings from the source identity for complete transfer. FaceShifter (Li et al., 2019) combines adaptive identity and attribute fusion with a two-step approach to handle occlusions. SimSwap (Chen et al., 2020) uses identity injection and weak feature matching to enhance attribute preservation. HifiFace (Wang et al., 2021) includes 3D shape priors for high-quality realism but is still sensitive to changes in pose.

Comprehensive evaluations on FaceForensics++ and AFLW2000-3D datasets show that FaceDancer outperforms leading methods in identity retrieval (up to 98.84%) and pose preservation. It also manages low-resolution and occluded inputs effectively. By combining adaptive feature fusion with clear attribute regularization, FaceDancer tackles persistent challenges in high-quality, real-world face swapping.

7. Face Swapping for Low-Resolution and Occluded Images In-the-Wild

The study by Jaehyun park, Wonjun Kang, Hyung il koo AND Nam ik cho proposed that Face swapping aims to replace the identity of a target face with that of a source face while keeping the target's pose, expression, lighting, and background intact. Early methods relied on 3D Morphable Models (3DMMs) to align geometry and appearance, but these were costly to compute and struggled with lighting, occlusion, and variations in expression. The introduction of GAN-based methods greatly improved realism. RSGAN performed swapping in latent feature space, while FSGAN used a two-stage process for subject-agnostic swapping. FaceShifter introduced an occlusion-aware refinement method (HEAR-Net) but still faced limitations with large or complex occlusions. High-resolution methods like MegaFS and FSLSD utilized StyleGAN2 for detailed synthesis but were less effective with low-quality, real-world faces due to gaps in the data.

Recent studies highlight the significance of face swapping for protecting privacy, particularly in surveillance and public images where faces are often low-resolution and occluded. Existing de-identification techniques mainly focus on removing identity. However, they struggle to maintain realism in such tough situations.

Experimental results show that the proposed method achieves better accuracy in identity retrieval and naturalness in both synthetic and real-world scenarios. It outperforms previous methods like SimSwap, FaceShifter, and HifiFace, especially in difficult low-resolution and occlusion-heavy conditions. This makes it a strong option for privacy-preserving applications in real-world video analysis.

8. DiffSwap: High-Fidelity and Controllable Face Swapping via 3D-Aware Masked Diffusion

The study by Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, Jiwen Lu present the face swapping aims to transfer the identity of a source face onto a target while keeping the target’s pose, expression, lighting, and background intact. Early methods based on 3D Morphable Models (3DMM) aligned facial geometry using structural guidelines, but often introduced artifacts and needed manual adjustments. GAN-based methods improved realism by combining source identity features with target features in a competitive setup. However, these methods often relied on several finely tuned loss functions. They struggled when the shapes of the source and target faces were quite different, and they lacked precise control over shape.

Recent developments in Diffusion Models (DMs) offer better stability, high-quality synthesis, and controllable generation, but using them for face swapping comes with challenges. These include the lack of ground truth swapped pairs and high inference costs due to multi-step sampling. To overcome this, DiffSwap redefines face swapping as a conditional inpainting task, utilizing both identity embeddings and facial landmarks as inputs.

Evaluations on FaceForensics++ and FFHQ show that DiffSwap achieves top-notch identity preservation (98.54% ID retrieval) and an excellent FID score (2.16). It also demonstrates strong resilience in terms of pose, shape, and high-resolution generation (up to 512×512). Its controllable and scalable design positions it as a promising option for future face swapping systems.

9. DiffFace: Diffusion-based Face Swapping with Facial Guidance

Face swapping involves transferring the features of a source face onto a target face while keeping the target’s pose, expression, gaze, lighting, and background intact. Traditional methods using 3D Morphable Models (3DMMs) [4, 5, 24] allowed for geometric alignment but had issues with reconstruction and maintaining fine details. The introduction of GAN-based approaches [6, 8, 17, 21, 23, 29, 35, 36] significantly improved realism through adversarial training. Notable works like FSGAN [35], FaceShifter [24], SimSwap [8], and HifiFace [52] used identity embedding and handled occlusions to accomplish

facial swapping without specific subjects. However, these models faced problems with unstable training, complex hyperparameter tuning, and balancing identity preservation with attribute retention.

Recently, there has been a focus on Diffusion Models (DMs) [9, 11, 15, 26, 42] as a substitute for GANs because of their training stability, high-quality image generation, and better control. Diffusion models create images through a process of iterative denoising, which allows for the integration of external facial expert models for attribute direction. Despite these benefits, earlier work had not directly applied diffusion models to face swapping until the introduction of DiffFace.

This framework allows for flexible control over the identity and attribute balance by adjusting mask thresholds and guidance weights, enabling both full-face and specific region swapping. Experiments on FaceForensics++ show that DiffFace outperforms GAN-based models in identity similarity (ArcFace: 0.620) while maintaining pose, shape, and gaze. Additionally, it shows strong performance on out-of-domain datasets, including artistic and cartoon faces.

10. ArcFace: Additive Angular Margin Loss for Deep Face Recognition

The study by Jiankang Deng, Jia Guo, Niannan Xue and Stefanos Zafeiriou presented that Deep face recognition has progressed through the use of Deep Convolutional Neural Networks (DCNNs). Designing effective loss functions is central to improving feature discrimination. Traditional softmax loss offers good separation but falls short on explicit intra-class compactness and inter-class margin.

To tackle this issue, Centre Loss introduced Euclidean distance minimization to improve intra-class compactness but faced scalability challenges. SphereFace enhanced discriminative power by using a multiplicative angular margin, but it needed complex approximations, which led to unstable training. CosFace made this simpler with an additive cosine margin, resulting in better performance and easier implementation.

Building on these ideas, ArcFace presents an additive angular margin that relates directly to geodesic distance on a hypersphere. This approach offers a clear geometric interpretation, stable training, and superior performance across various tests, establishing a new benchmark in face recognition.

No.	Method	Model Type	Resolution	Key Features	Strengths	Limitations
1	**ArcFace**	Face Embedder	N/A	Additive Angular Margin Loss on hypersphere	Best identity features for swapping, strong open-set recognition	Not a swapping method; used as backbone in others.
2	**FaceShifter**	Two-Stage GAN	High	AEI-Net, AAD, HEAR-Net for occlusion recovery	Occlusion-aware, strong identity preservation, realistic lighting	Two-stage complexity
3	**SimSwap**	GAN-based	Medium-High	Identity Injection Module, weak feature matching loss	Good generalization, balance of identity and attributes	Minor quality degradation in extreme conditions
4	**MegaFS**	StyleGAN2-based	Very High (1024×1024)	Hierarchical encoder, nonlinear latent manipulation (FTM)	Megapixel output, preserves fine details, efficient on GPUs	Demands careful training structure
5	**FaceDancer**	Single-Stage GAN	Low to Medium	AFFA (adaptive fusion), IFSR for feature regularization	Works well under occlusions/poses, no facial mask needed	Color defects if AFFA used at high-res layer
6	**DiffFace**	Diffusion (Guided)	High	Facial guidance: identity, semantics, gaze + blending	High-quality, flexible control, adaptable to out-of-domain data	Output instability from stochastic sampling
7	**DiffSwap**	Diffusion + 3D	High (512×512)	Midpoint estimation, 3D-aware masks, region-wise swapping	High accuracy (98.52%), fast (81.4 FPS), low compute cost	Loses fine target details, unstable output possible
8	**SimSwap++**	GAN + Custom Convs	High (512×512)	CD-Conv, MKD, VGGFace2-HQ dataset	High fidelity, shape control, region-level customization	Slightly complex training with distillation setup

9	**Low-Res Occluded Swapping**	Hybrid Pipeline	Low (wild images)	CRCL, occlusion parser, inpainting, post-refinement	Privacy-focused, realistic results on surveillance-type images	Limited to occlusion-heavy, low-res scenarios
10	**Diffusion-based Unified Approach**	Diffusion (Inpainting)	High (512×512)	DDIM sampling, CLIP disentanglement, mask shuffling	Fast inference, artifact-free, supports head swapping	Challenged by extreme poses/expressions

Conclusion

Face swapping has changed a lot over the past few years. It has moved from simple GAN-based systems to high-resolution, diffusion-driven methods. This review looked at ten leading techniques, each making distinct contributions to identity preservation, attribute control, occlusion handling, and real-time efficiency.Early models like SimSwap and FaceShifter made important strides in identity injection and occlusion-aware synthesis. Later models, such as SimSwap++ and FaceDancer, further advanced the field by adding dynamic convolution, attention-based fusion, and feature similarity regularization. This led to more realistic and adaptable outputs. At the same time, high-resolution models like MegaFS enabled megapixel-level face swapping with remarkable detail.

Diffusion models, seen in DiffSwap, DiffFace, and the unified inpainting-based diffusion approach, have transformed the face swapping landscape. They provide better realism, control, and stability. These models not only deliver higher identity and perceptual quality but also introduce features like 3D-aware swapping, facial guidance, and target-preserving blending. Moreover, methods tailored for low-resolution and occluded images offer practical benefits for surveillance and privacy protection in real-life situations.

Despite these improvements, challenges still exist in managing extreme facial variations, occlusions, and achieving consistent results in diffusion models. Future research should aim to enhance robustness, clarity, and ethical protections, given the potential for misuse of this technology.In summary, face swapping has grown into an effective, flexible, and highly controllable technique, supported by various innovative structures and training methods. As this field develops, it will be important to combine high-resolution synthesis, real-time performance, and ethical considerations for responsible and meaningful use.

References

1. Sanoojan Baliah, Qinliang Lin, Shengcai Liao, Xiaodan Liang, Muhammad Haris Khan, "Realistic and Efficient Face Swapping: A Unified Approach with Diffusion Models," IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2025. Available : https://ieeexplore.ieee.org/document/10579803.

2. Xuanhong Chen, Bingbing Ni, Yutian Liu, Naiyuan Liu, Zhilin Zeng, and Hang Wang "SimSwap++: Towards Faster and High-Quality Identity Swapping," IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL.46, NO.1, JANUARY 2024. Available : https://ieeexplore.ieee.org/document/10225678

3. JAEHYUN PARK , WONJUN KANG, HYUNG IL KOO AND NAMIKCHO " Face Swapping for Low-Resolution and Occluded Images In-the-Wild," In IEEE Access, vol. 12, pp. 91383-91395, 2024. Available: https://ieeexplore.ieee.org/document/10579803

4. Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge."Simswap: An efficient framework for high fidelity face swapping". In Proceedings of the 28th ACM International Conference on Multimedia, pages 2003–2011, 2020. 1, 2, 3, 8.

5. Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. "Arcface: Additive angular margin loss for deep face recognition". In Proceedings of the IEEE/CVF con ference on computer vision and pattern recognition, pages 4690–4699, 2019. 5, 6.

6. Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. "Diffface: Diffusion-based face swapping with facial guidance". arXiv preprint arXiv:2212.13344, 2022. 1, 2, 3, 8.

7. Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. "Faceshifter: Towards high fidelity and occlusion aware face swapping". arXiv preprint arXiv:1912.13457, 2019. 1, 2.

8. Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso Fernandez, and Cristofer Englund. "Facedancer: pose-and occlusion-aware high fidelity face swapping". In Proceedings of the IEEE/CVF winter conference on applications of com puter vision, pages 3454–3463, 2023.

9. Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. "Diffswap: High-fidelity and con trollable face swapping via 3d-aware masked diffusion". In Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition, pages 8568–8577, 2023. 1, 3, 7, 8.

10. YuhaoZhu, QiLi, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun." One shot face swapping on megapixels". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4834–4844, 2021. 2, 8.