Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ25236

Resolving Facial Image Defects Through the Integration of Segmentation and Fusion Networks

Zishuo Xia[1*], Yin Ru[2], Yilei Yang[3], Kaiwen Xian[4]

[1]*School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China*
[2]*School of Computer Science, Gonzaga University, Spokane, USA*
[3]*Aberdeen School of Data Science and Artificial Intelligence, South China Normal University, Foshan, China*
[4]*Shandong Experimental High School, Jinan, China*
**Corresponding Author. Email: 954643526@qq.com*

Abstract. Generating images with dynamic expressions has become a key component of the social media industry. To enhance the realism, it is crucial to remove the potential distortion in the face image. Unfortunately, videos produced by PIrenderer often exhibit facial blur due to inadequate segmentation between foreground and background. In this paper, GrabCut and MODNet are used to post-process the videos generated by PIrenderer and then fuse them. Our proposed method can guarantee to reduce the influence of background on the face when generating dynamic expression videos. These post-processing steps optimize dynamic facial expression rendering, mitigate the face distortion problem, and ultimately produce more realistic video output.

Keywords: face-modeling, image generating, GrabCut, ModNet, postprocessing.

1. Introduction

With the continuous advancement of artificial intelligence (AI), the entertainment industry has entered a new era. Digital entertainment, including video games, streaming services, music platforms, and social media, has become an integral part of people's lives. Among them, video games dominate as the main form of entertainment. In games, it is an important step to generate videos of faces with changeable expressions. However, the existing models will produce facial distortions when processing images.

Several studies have contributed to the development of AI-driven facial modeling and image transformation techniques. Hertzmann et al. introduced "IMAGE ANALOGIES" a method that synthesizes unfiltered and filtered images to generate new target images, eliminating the need for users to manually adjust multiple filter settings [1]. Zhu et al. introduced it in ICCV 2017 paper of theirs [2]. (ICCV 2017) effectively removed the rain streaks in a single image with a solution called bi-layer optimisation model, to improve image clarity [3]. Bansal et al. developed a data-driven, unsupervised video re-targeting method that transfers content while preserving style characteristics. For instance, their model could transfer the content of John Oliver's speech to match the style of

DOI: 10.54254/2755-2721/2025.BJ25236

Stephen Colbert, Utilizing spatial and temporal data alongside adversarial loss for content transfer and style retention [4]. Wiles et al. explored both supervised and unsupervised methods for facial modeling. Supervised approaches rely on ground-truth data sets to learn variations such as lighting and pose, though they can be costly and subjective [4]. In contrast, self-supervised and unsupervised methods aim to automatically learn these variations, maximizing mutual information, or predicting future video frames. Additionally, CycleGAN has been used to transform images between domains while preserving semantic similarity [4]. Tewari et al. investigated the relationship between edit magnitude and loss, finding that smaller edits result in lower losses, whereas larger modifications increase loss values. Their research demonstrated that expression control networks perform better at maintaining other facial properties compared to control networks, which struggle with pose restoration [5].

However, current AI-generated facial models often suffer from two significant issues: dynamic blur and facial distortion. These problems can cause character expressions to appear unnatural or uncontrollable, negatively impacting user experience. By building upon these advancements, our project seeks to refine AI-driven face modeling, reducing issues such as blur and distortion while improving expression accuracy. Our ultimate aim is to create a system that delivers high-quality facial models tailored to user expectations.

To address this challenge, our team aims to develop a model that can generate realistic facial models more efficiently. While existing models offer similar functionalities, there is substantial room for improvement. Our goal is to enhance face modeling techniques to better align with user needs, ensuring greater realism and stability in facial expressions.

Our model is built upon PIrenderer [6], which relies on a 3DMM-based parametric design. This approach depends heavily on the accurate decoupling of 3DMM parameters and often suffers from suboptimal efficiency. To mitigate these issues, we introduce additional post-processing steps—specifically, segmentation and fusion—to address the shortcomings of PIrenderer. Experimental results demonstrate that these enhancements significantly improve the realism of the generated videos.

2. Related work

X2Face a self-supervised network architecture that manipulates the generated faces based on the audio input. However, this model does not make assumptions about the pose and identity of the person, which may lead to inaccurate and natural output results in some extreme cases [4]. Perceptually A3 fine-tune the renderer to improve the image performance. Although it solves part of the background interference, it still affects the quality of the final result [7]. First-order motion model for image animation. This model did not locate any keypoints in the video when the input video pose was significantly different from the initial pose. This indicates that the keypoint detection ability of the model is insufficient under complex actions [8]. Deferred Neural Rendering generates images based on modified pose, which is a very good idea, but it has limitations. This model is only applicable to a specific topic, and DNR needs to be trained independently to generate videos of different actors, which leads to the problem in different application scenarios. It may be necessary to adapt the model to new input data [8]. OneShotFN, the model implements partial free perspective synthesis, but based on the existing 2D perspective, OneShotFN only synthesizes from the original perspective, which cannot meet the user's demand for a new perspective [8]. VASA-1 deals only with the upper body of the human body, a limitation that affects its ability to apply to parts of the scene, and while the model uses The absence of a more explicit facial model in a 3D latent representation may result in visual anomalies [9]. PerceptualCh has achieved some success in

DOI: 10.54254/2755-2721/2025.BJ25236

generating dialogue head videos for the model. However, the renderer used by PerceptualCh has some defects in preserving character identity and resisting background interference, which leads to the lack of accuracy in the background of video generation [10]. Magic Animate performs relatively well in the foreground region when dealing with the area where the portrait is located, but its overall quality is affected due to its insufficient background information processing for the model [11].

3. Our model

We initially employ PIrenderer for video generation; however, its output suffers from significant facial distortions. To address these issues, we subsequently apply GrabCut and ModNet for video segmentation, and the fusion algorithm is uesd to fuse the segmented features. This fusion process considerably enhances the overall quality, effectively reducing both facial distortions and ghosting artifacts in the characters. In the following sections, we provide a detailed examination of each individual module.

3.1. Pirenderer

PIrenderer is a pretrained model which can parse out the 3DMM parameters given a driving face, using it to predict flow for a source face. Specifically, transferring only the expression from the driving face during inference by replacing the expression parameters on the source face with those on the driving one. However, the Method s mentioned above are sensitive to the accuracy of these 3DMMs [6].

3DMMs are known to be not particularly accurate for face reconstruction due to the limited number of Blend shapes. They have difficulty delineating facial details of the shape, eye, and mouth of the face, which may eventually have side effects on the synthetic results [12]. The details can be found in [6].

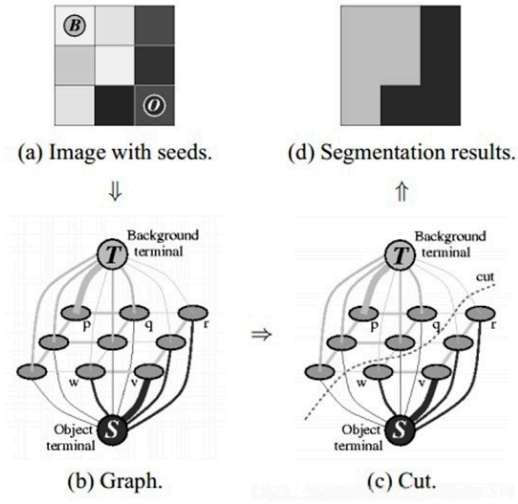

Figure 1. Principle of GrabCut [15]

3.2. Background segmentation

To address these limitations and improve the fidelity of our generated outputs, we integrate additional segmentation techniques that are better suited for capturing fine details.

DOI: 10.54254/2755-2721/2025.BJ25236

In this article, We use the grabCut [13] and ModNet [14] (Matting Objective Decomposition Network) Separate the foreground and background of the video generated by PIrenderer, grabCut is a method that uses Gaussian mixture model (GMM) [15]. As shown in Figure 1 MODNet provides a novel way to implement TRIMmapless portrait matting [15] to model references and generate, mark nodes, and cut if the background does not belong to the same terminal. To improve the performance of our portrait matting system.

3.2.1. GrabCut

GrabCut [13] algorithm transforms the problem of image segmentation into the problem of distinguishing gray levels between colors, and its core is to use graph cut to find the best segmentation between foreground and background. GrabCut firstly divides the image into labels, and then distinguishes the gray levels of these divided images. After distinguishing different gray levels, it divides the foreground and background and cuts them as shown in Figure 1).

3.2.2. Architecture of MODNet

MODNet categorizes the trimap-free matting objective into three primary components: semantic estimation, detail prediction, and semantic-detail fusion. The components are concurrently optimized through three interrelated branches, as seen in Figure 2. The semantic estimation branch in MODNet is responsible for locating the depiction in the provided image I(the output image obtained by PIrenderer). It uses an encoder, especially the low-resolution branch S of MODNet, to extract high-level semantics. In our implementation, we adopt a backbone similar to MobileNetV2 for its efficiency in real-time applications.

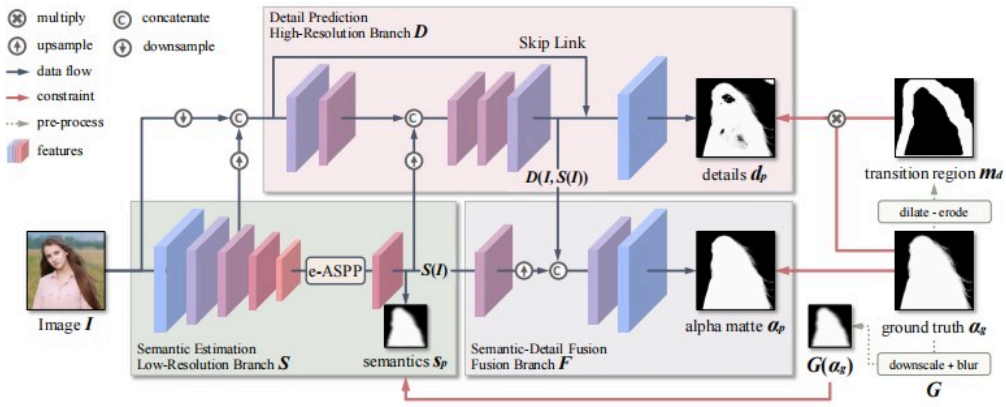

Figure 2. The specific operation flowchart of ModNet [14]

Semantic Mask Prediction

In order to predict the coarse semantic mask s_p , we input the output of encoder S(I) into the convolution layer activated by the Sigmaid function, reducing the number of channels to 1. The supervision of s_p comes from the down sampling thumbnail of the ground truth matte α_g. L2 loss function is formulated as:

$$\mathscr{L} = \frac{1}{2}\left|s_p - G\left(\alpha_g\right)\right|_2 \tag{1}$$

DOI: 10.54254/2755-2721/2025.BJ25236

where G represents a composite operation of downsampling followed by Gaussian blur. This operation removes fine-grained structures irrelevant to portrait semantics, ensuring the semantic mask focuses on the global portrait region.

Detail Prediction

The detail prediction branch D in MODNet focuses on processing the transition region around the foreground portrait. It takes three inputs:

• The original input image I,

• The output of the semantic estimation branch S(I),

• Low-level features extracted from encoder S.

By reusing the low-level features from S, this branch minimizes computational overhead while preserving spatial details.

The output of branch D is denoted as D(I,S(I)), explicitly indicating its dependency on high-level semantics S(I).D(I,S(I)) exports boundary details matte d_p and supervises through L1 loss function:

$$\mathscr{L} = m_d |d_p - \alpha_g|_1 \tag{2}$$

where m_d is a binary mask generated by applying morphological dilation and erosion to α_g. This mask constrains the loss calculation to the transition region:

$$m_d(x) = 1 \;\&\; \text{“}\{if\}\; x \in \text{“}\{transition\ region\}\text{”}$$

$$0 \;\&\; \text{“}\{otherwise\}\text{”}$$

The transition area is defined as the zone between the expanded and eroded real ground masks, effectively concentrating learning on the portrait boundaries.

Semantic and Detail Fusion

MODNet combines semantic information S(I) and boundary details D(I,S(I)) by fusing branches Fto predict the final alpha matte α_p. The process involves:

• Upsampling S(I) to match the spatial resolution of D(I,S(I))

• Concatenating the aligned features along the channel dimension

• Processing the concatenated features through F to generate α_p

The loss function for alpha matte prediction combines L1 error and compositional constraints:

$$\mathscr{L}_\alpha = |\alpha_p - \alpha_g|_1 + \mathscr{L} \tag{3}$$

where the compositional loss L_c is defined as:

$$\mathscr{L} = |I - (\alpha_p \odot F_g + (1 - \alpha_p) \odot B_g)|_1 \tag{4}$$

with F_g and B_g denoting the ground truth foreground and background respectively, and $\odot$ representing element-wise multiplication. MODNet [14] is optimized end-to-end through weighted summation of branch-specific losses:

$$\mathscr{L} = \lambda_s \mathscr{L} + \lambda_d \mathscr{L} + \lambda_\alpha \mathscr{L}_\alpha \tag{5}$$

DOI: 10.54254/2755-2721/2025.BJ25236

Hyperparameters $\lambda_s, \lambda_d, \lambda_\alpha$ balance the contributions:

- $\lambda_s = \lambda_\alpha = 1$ (semantic and fusion losses)
- $\lambda_d = 10$ (enhanced boundary detail supervision)

This configuration emphasizes boundary accuracy while maintaining global consistency.

3.3. Fusion

We employed a fusion model based on image segmentation results, aiming to integrate the advantages of reference images and generated images to obtain more desirable output results. The fusion process is mainly based on the following steps and formulas [14].

To make the segmentation results more stable and accurate, we calculated the median of consecutive frame segmentation results. Specifically, for each pixel point (x,y), the median segmentation result M_t^{med} (x,y) at time t is the median of the segmentation results of the corresponding pixel points in the previous five frames (including the current frame). Its mathematical expression is:

$$M_t^{med}(x,y) = \text{median}\{M_t(x,y), M_{t-1}(x,y), \cdots, M_{t-4}(x,y)\} \tag{6}$$

where $M_t(x,y)$ represents the pixel value at the position (x,y) in the segmentation result image at time t. By this means, noise and fluctuations in the segmentation results can be effectively reduced.

We extract the common background region from the reference image I_{ref} that matches the median segmentation result and fuse it with the generated image $\widetilde{I_{gen}}$. First, we obtain the intersection of the median segmentation result M_t^{med} and the background region mask M^{ref} in the reference image through the logical AND operation ($\cap$), and then smooth it using a Gaussian filter to obtain the fusion region mask M_t^{fusion} . Its expression is as follows:

$$M_t^{fusion} = \text{Gaussian}\left(M_t^{med} \cap M^{ref}\right) \tag{7}$$

The Gaussian filter (Gaussian) here uses a 7×7 kernel to smooth the boundaries of the fusion region and reduce splicing traces.

Finally, we fuse the generated image $\widetilde{I_{gen}}$ and the reference image I_{ref} according to the fusion region mask M_t^{fusion} to obtain the final fused image $\tilde{I}_t^{fusion}$. The fusion formula is:

$$\widetilde{I_t^{fusion}} = \left(1 - M_t^{fusion}\right) \odot \widetilde{I_{gen}} + M_t^{fusion} \odot I_{ref} \tag{8}$$

where $\odot$ represents element-by-element multiplication operation. The meaning of this formula is that for the part where the value of the fusion region mask is 0 (i.e., the region mainly belonging to the generated image), the pixel values of the produced image are preserved; for the part where the value of the fusion region mask is 1 (i.e., the region mainly belonging to the reference image), the pixel values of the reference image are adopted; and for the intermediate transition region, linear mixing is performed according to the value of the fusion region mask, thus achieving seamless fusion.

DOI: 10.54254/2755-2721/2025.BJ25236

4. Experimental results

In this section, our model processes raw videos from PIrender [6] using pretrained ModNet [14] and GrabCut [13] respectively, employing these two segmentation methods for foreground separation and background composition to generate new videos. We denote the strategy of employing ModNet for foreground background segmentation and fusion within PIRender as PIRender-ModNet, while the approach utilizing GrabCut for equivalent segmentation and fusion is referred to as PIRender-GrabCut. (results shown in Figures 3 and Figure 4) We adopted the Learned Perceptual Image Patch Similarity (LPIPS) [16] metric, an image similarity evaluation metric based on deep neural network feature space distance, used to quantify the perceptual differences between two images, to evaluate the generated results (containing five different models/subjects). A lower LPIPS metric indicates better generation results. The quantitative assessment demonstrates that GrabCut achieves marginally superior LPIPS values compared to ModNet in the current test cases. However, beyond the LPIPS metric, we also conducted qualitative evaluations through visual observation of video outputs.And case analysis reveals three distinct scenarios:

PIRender-GrabCut superiority (e.g., Model 5 in Figure 3): The third image generated by PIRender-ModNet exhibits abnormal dark artifacts along hair boundaries (indicated by red arrows in the Figure 3), likely due to suboptimal segmentation and fusion in ModNet's processing, whereas PIRender-GrabCut produces artifact-free results. PIRender-ModNet superiority (e.g., Model 1 in Figure 4): The fourth image from PIRender-GrabCut displays noticeable artifacts near the microphone region (indicated by red arrows in the Figure 4), which are absent in PIRender-ModNet's output. Comparable performance (e.g., Model 3 and 4): No discernible differences were observed between the outputs of both methods. However, while PIRender-ModNet demonstrated superior performance in a small subset of specific test cases compared to PIRenderGrabCut, it is worth noting that this advantage comes with prerequisite conditions. Unlike GrabCut, a traditional graph-cut based segmentation algorithm that requires no training, ModNet necessitates substantial computational resources and time investments for model training and parameter optimization prior to implementation. Thus, our implementation utilizes a pre-trained ModNet deep learning model for automated foreground-background segmentation and composition. This fundamental distinction in operational requirements positions GrabCut as a more accessible solution for scenarios demanding rapid deployment, particularly in resource-constrained environments.

Quantitative Analysis of LPIPS Metrics (results shown in Table 1): The experimental results demonstrate significant improvements through post-processing strategies: PIRender-GrabCut achieves an average LPIPS score of 0.0276 across five subjects, PIRender-ModNet is 0.0330. This represents a 16.36% relative improvement for GrabCut over ModNet. Both post-processed methods substantially outperform the baseline PIRender (baseline LPIPS = 0.0355): GrabCut shows 22.25% improvement, ModNet achieves 7.04% improvement.

These quantitative findings confirm the effectiveness of our proposed postprocessing framework in enhancing the original PIRender outputs.

Quantitative and qualitative evaluations demonstrate that PIRender-GrabCut generally achieves superior performance in both LPIPS metrics and visual inspection for most test cases, with limited exceptions showing comparable or slightly inferior results. Crucially, both post-processing strategies (GrabCut and ModNet implementations) exhibit significant improvements over the baseline PIRender output across all evaluation metrics.

DOI: 10.54254/2755-2721/2025.BJ25236

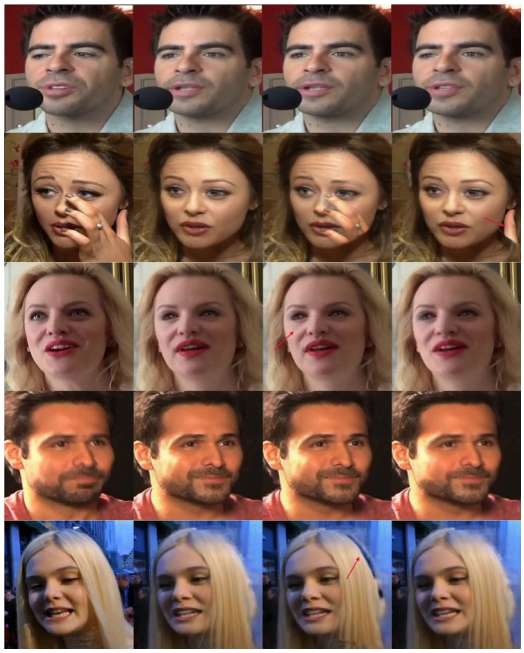

Figure 3. Selected comparative results where PIRender-GrabCut outperforms PIRender-ModNet (The four columns represent: Reference image, PIRender output, PIRender-GrabCut result, and PIRender-ModNet result respectively) slightly inferior results. Crucially, both post-processing strategies (GrabCut and ModNet implementations) exhibit significant improvements over the baseline PIRender output across all evaluation metrics

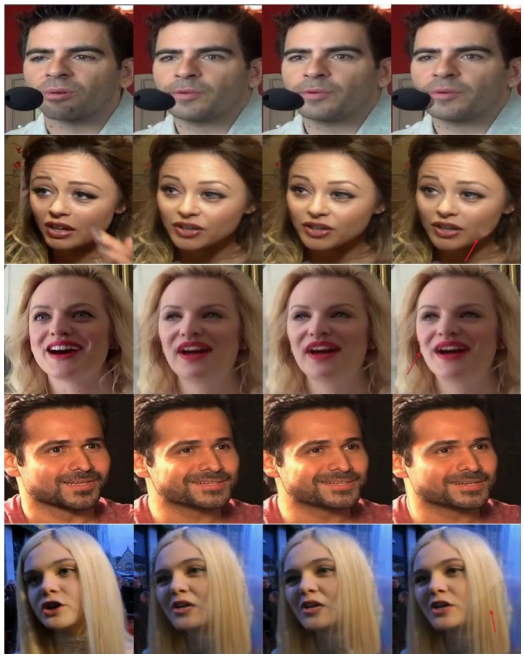

Figure 4. Selected comparative results where PIRender-ModNet outperforms PIRender-GrabCut (The four columns represent: Reference image, PIRender output, PIRender-GrabCut result, and PIRender-ModNet result respectively)

DOI: 10.54254/2755-2721/2025.BJ25236

Table 1. Quantitative analysis of LPIPS metrics (the best statistics of experimental results are bold in the table)

Method	Model 1	Model 2	Model 3	Model 4	Model 5	Models average
Pirender	0.0194	0.0371	0.0319	0.0147	0.0743	0.0355
Pirender-GrabCut	0.0165	0.0245	0.0252	0.0115	0.0604	0.0276
Pirender-ModNet	0.0190	0.0335	0.0301	0.0131	0.0693	0.0330

5. Discussion

Our experimental results consistently demonstrate that the proposed post-processing pipeline, leveraging either GrabCut or MODNet for segmentation followed by fusion, effectively mitigates the key limitations—facial distortion, and background interference—inherent in videos generated by PIrenderer. The significant quantitative improvements in LPIPS scores (averaging a 22.25% improvement with GrabCut and 7.04% with MODNet over the baseline, Table 1) and qualitative reduction in visual artifacts (Figures 3 & 4) confirm the efficacy of this approach. This success can be primarily attributed to the core function of segmentation and fusion: by decoupling the foreground (subject's face) from the potentially noisy or inadequately rendered background in PIrenderer's output, and subsequently fusing it with a stable background reference, we directly address the root causes of the observed distortions and blurring. This process alleviates the negative influence of background artifacts on the foreground and potentially recovers finer facial details that may be lost due to the limitations of the 3DMM-based parameter decoupling in PIrenderer [6].

6. Conclusion

This work presents a hybrid methodology that integrates the GrabCut and MODNet algorithms to enhance background segmentation in portrait video creation. GrabCut offers high user controllability and requires no pre-trained models, making it lightweight and flexible. On the other hand, MODNet provides a high degree of automation and delivers superior segmentation accuracy through its end-to-end trimap-free matting architecture. By integrating the strengths of both methods, our model aims to mitigate the common issues found in existing models, such as facial distortion and background interference. While this dual-method approach enhances segmentation performance, it also introduces challenges, such as increased data processing requirements and dependency on the initial input quality. Additionally, the computational demands for training and deployment are relatively high. Overall, the proposed system demonstrates that a post-processing pipeline combining traditional and deep learning-based segmentation techniques can significantly enhance the visual realism and stability of face generation systems. And our work has surely brought improvements to the original module. Future work will focus on optimizing model efficiency and extending its adaptability to more diverse and dynamic real-world scenarios.

References

[1] Aaron Hertzmann et al. "Image analogies". In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '01. New York, NY, USA: Association for Computing Machinery, 2001, pp. 327–340. isbn: 158113374X. doi: 10.1145/383259.383295. url: https: //doi.org/10.1145/383259.383295.

[2] Lei Zhu et al. "Joint Bi-Layer Optimization for Single-Image Rain Streak Removal". In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Oct. 2017.

DOI: 10.54254/2755-2721/2025.BJ25236

[3] Aayush Bansal et al. “Recycle-GAN: Unsupervised Video Retargeting”. In: European Conference on Computer Vision. 2018. url: https: //api. semanticscholar.org/CorpusID: 51987197.
[4] Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. “X2Face: A network for controlling face generation by using images, audio, and pose codes”. In: ArXiv abs/1807.10550 (2018). url: https: //api.semanticscholar. org/CorpusID: 51866642.
[5] Ayush Tewari et al. “StyleRig: Rigging StyleGAN for 3D Control over Portrait Images, CVPR 2020”. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. June 2020.
[6] Yurui Ren et al. “PIRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering”. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021), pp. 13739–13748. url: https: //api.semanticscholar.org/CorpusID: 237562793.
[7] Lee Chae-Yeon et al. “Perceptually Accurate 3D Talking Head Generation: New Definitions, Speech-Mesh Representation, and Evaluation Metrics”. In: 2025. url: https: //api.semanticscholar.org/CorpusID: 277345403.
[8] Aliaksandr Siarohin et al. “First Order Motion Model for Image Animation”. In: Neural Information Processing Systems. 2020. url: https: //api.semanticscholar.org/CorpusID: 202767986.
[9] Sicheng Xu et al. “Vasa-1: Lifelike audio-driven talking faces generated in real time”. In: Advances in Neural Information Processing Systems 37 (2024), pp. 660–684.
[10] Ai-Mei Huang, Zhewei Huang, and Shuchang Zhou. “Perceptual Conversational Head Generation with Regularized Driver and Enhanced Renderer”. In: Proceedings of the 30th ACM International Conference on Multimedia (2022). url: https: //api.semanticscholar.org/CorpusID: 250072874.
[11] Zhongcong Xu et al. “MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model”. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2024, pp. 1481–1490.
[12] Youxin Pang et al. “DPE: Disentanglement of Pose and Expression for General Video Portrait Editing”. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2023, pp. 427–436.
[13] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. “GrabCut”: interactive foreground extraction using iterated graph cuts”. In: ACM Trans. Graph. 23.3 (Aug. 2004), pp. 309–314. issn: 0730-0301. doi: 10.1145/1015706.1015720. url: https: //doi.org/10.1145/1015706. 1015720.
[14] Zhanghan Ke et al. “MODNet: Real-Time Trimap-Free Portrait Matting via Objective Decomposition”. In: AAAI Conference on Artificial Intelligence. 2020. url: https: //api.semanticscholar.org/CorpusID: 246295022.
[15] Department of Computer Science Rhodes University. Research Project G02M1682. https: //www.cs.ru.ac.za/research/g02m1682/. Accessed: 2025-04-08.
[16] Richard Zhang et al. “The unreasonable effectiveness of deep features as a perceptual metric”. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, pp. 586–595.