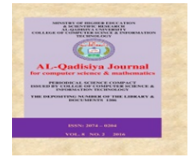


Available online at www.qu.edu.iq/journalcm

JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS

ISSN:2521-3504(online) ISSN:2074-0204(print)



Deep Video Spoof Generation Based on Matting Technique

Nabaa Maad Abd AL-Kareem¹, Saud Jamila H.²

¹ Department of Computer Science, College of Science, Al-Mustansiriyah University, Al-Bnook, Baghdad, Iraq, nabaamaad@uomustansiriyah.edu.iq

² Department of Computer Science, College of Science, Al-Mustansiriyah University, Al-Bnook, Baghdad, Iraq,

Dr.jamilahharbi@uomustansiriyah.edu.iq

ARTICLE INFO

Article history:

Received: 07 /03/2026

Rrevised form: 16 /05/2026

Accepted: 17 /05/2026

Available online: 30 /09/2026

Keywords:

Deep video spoof generation, Matting-based video generation, Face swapping, Image matting, Video matting, Background matting, Alpha matte, Trimap-free matting, Temporal consistency, Foreground-background compositing, Boundary refinement, Transformer-based matting, Diffusion models.

ABSTRACT

Deep video spoof generation has made significant progress by using well-designed face-swapping models with high-level identity transfer and visual realism. However, realistic deep video spoof generation from arbitrary videos remains difficult due to visible compositing artefacts both locally (e.g., hairline and semi-transparent regions) and temporally (flicker and drift). We consider deep video spoof generation, a focus on which is chiefly underpinned by matting, as an essential building block for boundary-accurate and temporally consistent insertion of manipulated foreground into real scenes with no requirement for the presence or recording of green screens. We highlight some key directions in image, background, and video matting: (1) background-referenced matting; (2) trimap-free and real-time portrait matting; and (3) temporally stable video matting methods that exploit spatio-temporal alignment as well as memory propagation. We also consider recent modelling trends, including transformer-based and diffusion-assisted matting, which provide better fine-detail reconstruction and interactive refinement. The analysed studies show that matting can alleviate key visible failure modes in spoof generation pipelines (which are improved alpha estimation quality, boundary sharpness, and temporal consistency) in all works we surveyed. Based on the observations, we finally summarise the open issues—robustness to occlusion and fast motion, generalisation from composited training to real-world scenes, computation overhead under high resolution, and unified evaluation protocols—and propose future directions for research towards more realistic and stable matting-based deep video spoof generation.

<https://doi.org/10.29304/jqcm.2026.18.32718>

1. Introduction

We witness the emergence of deep video spoof generation techniques progressing from primitive facial replacement to high-quality face-swapping pipelines, ensuring identity preservation and a naturalistic look. Recent methods focus on efficiency and visual quality, making it feasible to be used in practice for face swapping under various identities and conditions with high fidelity [1]. Meanwhile, the research community moved to higher resolutions – sometimes even megapixels – where natural imperfections can become quite obvious (i.e., are harder to hide) [2]. Meanwhile, surveys aggregate the status quo of face-swap systems and discuss the long-existing difficulties in realistic synthesis, particularly from controlled settings to unconstrained real videos [3], [4]. Taken all together, these works suggest that while identity transfer has grown strong, the realistic appearance of generated feigned

videos is still constrained by the level achieved in compositing, where the computer-generated face must be consistently composited into a target frame.

The reason why they don't scale well is that compositing, in its very nature, is a matting task: obtaining a good-quality alpha matte and performing consistent blending at various locations such as hairline, boundary, semi-transparency region, and along motion. Background matting is proposed to treat real scenes as a "green screen" so that it can extract the front object for realistic composition without limitation of studio capture [5], and then we witness improvements in real-time, high-resolution background matting that makes it ready for practical deployment and has sharper boundaries [6]. Trimap-free portrait matting techniques are introduced to decrease the requirement of a trimap and result in cleaner alpha estimates in real-life situations [7]. When dealing with videos, due to the temporal coherence, frame-to-frame instability is effectively mitigated by spatio-temporal alignment and aggregation [8], while memory-based propagation stabilises mattes across longer sequences [9]. Resolution and motion complexity. In the recent past, sparsity-based methods enhanced the efficiency of ultrahigh-resolution image/video matting, as introduced in [10]; diffusion-based methods in favour of interactive matting and fine boundary recovery with strong priors are proposed in [11]. Recent transformer-based matting models further improve global context modelling and boundary detail, such as token-prior transformers and vision transformer matting formulations [12]–[14], as well as attention mechanisms and trimap generation techniques to improve hard edge regions [15]. Complementary journal works further promote real-world matting through refinement of high-resolution alpha [16], flexibility of user-guided trimap [17], and unsupervised background matting with strong priors [18] while closing the gap from composited training data to real-world cubic deployment [19]. More benchmarks and learning strategies, including RGB-D human matting datasets [20] and frequency-guided saliency-aware trimap-free training [21], are applied to ensure evaluation and generalisation. Other works, including multi-guided constraints and edge/boundary-aware formulations for refining fine structures [22], [23], and interactive paradigms, take advantage of the modern segmentation priors that make matting more controllable in practice [24]. A more in-depth survey merges these directions and explains how various matting methods are applied to practical compositing operations [25].

Putting these threads together, the proposed deep video spoof generation based on matting attempts to enhance the reality of the deep video spoof generation by explicitly optimising both the alpha boundary quality and temporal continuity of compositing rather than identity transfer only. In more detail, the face-swapping techniques deliver veritable identity manipulation [1]–[4], while matting solutions provide boundary-precise, temporally consistent foreground separation to layout spoofed content seamlessly in real scenes without visible seams, hairline artefacts, or flicker [5]–[25]. Accordingly, in this paper, we concentrate on how the modern matting (image, background, and video matting) can serve as a fundamental module for synthesising more visually adaptive deep video spoof generation in the wild.

1.1 Main Contribution

- This survey focuses on the connection between deep video spoof generation and matting-based compositing methods, rather than spoof detection
- The latest papers are grouped based on their practical capabilities, such as face swap and backdrop matting, as well as trimap-free portrait matting, video matting, high-resolution refinement matting, transformer-based active matting, and interactive or diffusion guidance-based image matting.
- It investigates how alpha matte accuracy, boundary refinement, illumination consistency, and temporal stability influence the realism of generated spoof videos.
- Specifically, it provides by identifying a number of important research gaps, such as Heavy Occlusion, Extra Alpha (Incorrect alpha estimation), Boundary Artefacts, Temporal Flickering, Auxiliary Inputs, and in Matting-based Generation, no standardized evaluation methodologies exist.
- It will explore future directions towards creating more robust and realistic deep video spoof generation systems based on matting in unconstrained real-world environments.

2. Related Work

2.1 Face-swapping-based deep video spoof generation

Deep video spoof generation is typically achieved via face swapping pipelines under the constraint that identity should be conserved but appearance, pose and overall visual realism of one target are retained. [Chen et al.2020] proposed SimSwap and demonstrated strong quantitative performance on identity and realism, such as ID similarity of around 0.8140 (LQ) and 0.8596 (HQ), with FID around 13.05 (LQ) and 6.03 (HQ), as well pose and expression errors (e.g., $\sim 4.6^\circ$ pose error and ~ 4.1 – 4.6 expression error, depending on setting) [1]. At higher fidelity outputs, [Zhu et al.2021] presented mega-pixel-wise one-shot face swapping on megapixels (MegaFS), where ID similarity = 0.5014, pose error = 3.58° , expression error = 2.87, and FID = 10.16 [2]. These observations indicate that though the identity manipulation quality is measurable and competitive, deep video spoof generation at higher resolution aggravates boundary imperfection (especially for hairlines and fine structures), which actuates an insight of using matting-based compositing.

2.2 Matting for realistic compositing and boundary quality

Matting-based approaches directly optimize the alpha boundary necessary for a smooth composition, which is necessary for visually-realistic fake videos. [Sengupta et al.2020] proposed Background Matting for green-screen-like compositions with a reference background and reported SAD=1.72, MSE= 0.97×10^{-2} in SOTA on the standard matting evaluation [5]. To facilitate realistic high-resolution usage, [Lin et al.2021] presented real-time high-resolution background matting, achieving the real-time performance (e.g., approximately 30 fps at 4K and ~ 60 fps at HD) as well as quantitative matting accuracy, for example, SAD = 2.15, MSE = 0.005, Grad = 6.16 and Conn = 2.85 of their evaluation setting [6]. For the trimap free portrait matting, [Ke et al.2022] introduced MODNet, which reports MSE = 0.0044 and MAD = 0.0086, and also some real-time performance (at about 67 fps on a GTX1080Ti at some common input resolution [7]). These works exhibit measurable gains in boundary accuracy and efficiency, both of which are crucial when compositing a spoofed face onto authentic backgrounds.

2.3 Video matting and temporal stability

In video spoofing, matting should further be stable in time to avoid edge flicker. [Sun et al.2021] introduced Deep Video Matting and claimed for multi-metric results including SAD = 40.91, MSE = 0.014, Grad = 19.02, Conn = 40.58 and temporal stability, dtSSD = 15.11 using their evaluated setting [8]. For ultra-high-resolution cases, [Sun et al.2023] proposed a sparsity-based method as well, also on VideoMatte240K: MAD = 7.22, MSE = 2.01 and dtSSD = 2.27 were reported here, showing more temporal coherence and better quality under high resolution [10]. More recently, [Yang et al.2025] extended MatAnyone to enable uniform memory propagation, and showed that it is stable in practice (MAD = 0.14, MSE = 0.10 and dtSSD = 0.89) on a real benchmark test [9]. Collectively, these results reveal that current video matting more and more emphasizes capturing both spatial boundary fidelity and temporal consistency, which are the very bottlenecks responsible for most of the artifacts leading to degraded realism in spoof.

2.4 Image and Background Matting for Alpha Estimation and Refinement

Recent image and background matting methods refine the alpha estimate via sophisticated feature representations, leading to guidance, high-resolution refinement, and interactive prompting. Transformer-based methods such as MatteFormer, ViTMatte, and ProxyMatting boost global context modelling and boundary accuracy [12]–[14]. The development and dissemination of trimaps using attention mechanisms can diminish the necessity for manual trimaps; they remain susceptible to the quality of the generated guidance [15]. High-resolution refining techniques enhance matte detail derived from low-resolution alpha estimates [16]. User-guided and user-participatory methods can handle ambiguous boundaries by prompts or trimaps [17], [24]. Unsupervised background matting and real-world matting methods aim to bridge the gap between composited training data and real-world usage [18], [19]. These cues are critical for spoof generation, as they tackle the same weaknesses that are usually used to visually detect manipulated videos: sharp boundaries, irregular blending, and wrong foreground separation.

2.5 Comparative Summary and Interpretation

Table 1 provides a comparative overview of the most important studies reviewed in this work, describing their methodologies, datasets, performance metrics, and limitations.

Table 1. Comparative Summary of Deep Video Spoof Generation and Matting-Based Studies

Part A: Face-swapping, background matting, portrait matting, and video matting methods

Author / Year	Title of Research	Methodology	Dataset	Performance	Limitation
Chen et al., 2020 [1]	SimSwap: An Efficient Framework for High Fidelity Face Swapping	Face swapping using identity injection and weak feature matching.	VGGFace2; FaceForensics++; CelebAMask-HQ	ID retrieval 92.83% on FaceForensics++; pose error 1.53 degrees.	Hairlines, occlusion, and boundary blending may still be imperfect; requires effective compositing.
Zhu et al., 2021 [2]	One Shot Face Swapping on Megapixels (MegaFS)	One-shot megapixel face swapping using StyleGAN2 latent space and hierarchical representation.	FaceForensics++; CelebA-HQ; FFHQ	FaceForensics++: ID retrieval 90.83; pose error 2.64 degrees; expression error 2.96. CelebA-HQ: ID similarity 0.5014; pose 3.58 degrees; expression 2.87; FID 10.16.	High-resolution outputs can make boundary artifacts more visible; depends on GAN prior and inversion quality.
Sengupta et al., 2020 [5]	Background Matting: The World Is Your Green Screen	Background-referenced matting to estimate foreground and alpha matte.	Adobe Matting Dataset	SAD 1.72; MSE 0.97 x 10 ⁻² .	Requires a background or reference image; performance may decrease with dynamic backgrounds or misalignment.
Lin et al., 2021 [6]	Real-Time High-Resolution Background Matting	Two-stage high-resolution background matting using base prediction and patch refinement.	VideoMatte240K; PhotoMatte13K; PhotoMatte85	Approximately 30 fps at 4K and 60 fps at HD; SAD 2.15; MSE 0.005; Grad 6.16; Conn 2.85.	Requires a stable reference background; success depends on alignment and background consistency.
Ke et al., 2022 [7]	MODNet: Real-Time Trimap-Free Portrait Matting via Objective Decomposition	Trimap-free portrait matting using semantic and detail branches.	PPM-100	MSE 0.0044; MAD 0.0086; approximately 67 fps on GTX1080Ti.	Optimized mainly for portraits; generalization to unfamiliar objects and transparency is more difficult.
Sun et al., 2021 [8]	Deep Video Matting via Spatio-Temporal Alignment and Aggregation	Video matting with spatio-temporal alignment, aggregation, and trimap propagation.	DVM composited test set	SAD 40.91; MSE 0.014; Grad 19.02; Conn 40.58; dtSSD 15.11.	Requires trimap guidance; may drift under long occlusion or fast motion; possible temporal flicker.
Sun et al., 2023 [10]	Ultrahigh Resolution Image/Video Matting with Spatio-Temporal Sparsity	Spatio-temporal sparsity for ultra-high-resolution image and video matting.	VideoMatte240K; AIM-Human; HHM2K	VideoMatte240K: MAD 7.22; MSE 2.01; dtSSD 2.27.	Computationally and memory intensive at ultra-high resolutions; affected by motion complexity and boundary detail.

Part B: Memory propagation, transformer-based matting, high-resolution refinement, and interactive matting methods

Author / Year	Title of Research	Methodology	Dataset	Performance	Limitation
Yang et al., 2025 [9]	MatAnyone: Stable Video Matting with Consistent Memory Propagation	Mask-guided video matting with consistent memory propagation.	VideoMatte; YouTubeMatte; real-world benchmark subset	Real-world benchmark: MAD 0.14; MSE 0.10; dtSSD 0.89.	Requires a target assignment mask at least for the first frame; guidance adds interaction overhead.
Yao et al., 2024 [13]	ViTMatte: Boosting Image Matting with Pre-Trained Plain Vision Transformers	Transformer-based matting using pretrained ViT features.	Composition-1K; Distinctions-646	Composition-1K: SAD 20.33; MSE 3.0; Grad 6.74; Conn 14.78. Distinctions-646: SAD 17.05; MSE 1.5; Grad 7.03; Conn 12.95.	Usually requires trimap guidance; computational cost is higher than lightweight CNN-based methods.
Zhou et al., 2023 [15]	Sampling Propagation Attention with Trimap Generation Network for Natural Image Matting	Trimap generation with propagation attention for trimap-free matting.	Composition-1K; Distinctions-646	Composition-1K: SAD 37.95; MSE 0.0045; Grad 16.82; Conn 34.76. Distinctions-646: SAD 38.79; MSE 0.0082; Grad 36.89; Conn 38.64.	Performance depends on generated guidance; errors in trimap generation propagate to alpha estimation.
Ye et al., 2025 [16]	High-Resolution Natural Image Matting by Refining Low-Resolution Alpha Mattes	High-resolution alpha refinement from low-resolution matte predictions.	Transparent-460	SAD 173.43; MSE 13.09; Grad 55.84; Conn 156.35.	Requires an initial alpha matte; final quality is constrained by the reliability of the initial estimate.
Xu et al., 2022 [18]	Unsupervised Deep Background Matting Using Deep Matte Prior	Unsupervised background matting with deep matte prior and reference background.	SC Adobe portrait and non-portrait settings	Portrait: SAD 1.54; MSE 1.7×10^{-3} . Non-portrait: SAD 7.30; MSE 10.0×10^{-3} .	Requires background reference; weak under background changes and misalignment.
Yao et al., 2024 [24]	Matte Anything: Interactive Natural Image Matting with the Segment Anything Model	Prompt-based interactive matting using segmentation-style guidance.	AIM-500; Composition-1K	AIM-500 zero-shot: SAD/MSE 27.8/9.4 overall and 110.5/35.3 for transparent regions; Composition-1K improved from 72.3 to 39.7 to 26.2 SAD across refinement settings.	Requires user prompts; output quality depends on prompt accuracy and interaction quality.

Table 1 shows three broad trends. The first aspect is that face-swapping research mostly measures identity preservation and visual realism but fails in the compositing step, which involves creating seamless hairlines, occluded areas, and maintaining fidelity to high-frequency edges. Second, backdrop and portrait matting techniques take advantage of alpha estimation and boundary quality, but state-of-the-art methods come with some additional inputs, such as reference backgrounds, trimaps, or human guidance. The third is that video matting methods focusing on temporal coherence with spatio-temporal alignment, sparsity, and memory propagation reduce flicker and boundary drift growth directly in manipulated videos. The research, which has been analysed, shows that matting is a key factor in the generation of deep video spoof generation for improving foreground-background integration; however, existing techniques face limitations on severe occlusion, fast motion with high computation cost (or processing time), and changing evaluation conditions. Thus, these constraints can be considered as the principal research gaps for upcoming matting-based deep video spoof generation systems.

3. Research Gaps and Challenges

Severe Occlusion: It is another serious drawback of existing matting models, as they struggle when parts of the face or foreground, due to hands, hair, objects, shadows, etc., are being occluded in real-time use. When severe occlusion occurs, there is an accuracy loss in estimating the alpha, and obvious artifacts would occur on the synthesized video.

Manipulated samples that typically showcase boundary artefacts and clearly outline edges like sharp foreground boundaries, halos, hairline distortions with bad blending; These edges are not actually detected edges, they need to be smoothed and added together.

Temporal Flickering: Because of the artifacts in matting between frames, flickering and shifting of boundaries occur. Due to the fact that one image might seem visually correct but incorrect in terms of time, this is a prominent problem in deep video spoof generation tasks.

Illumination and background mismatch: The edited foreground has to be consistent with the lighting, colour tone, and contextual environment of the target scene. Background variations and misinformation caused by registration errors are the weaknesses of reference-based matting methods.

Reliance on extra inputs: the majority of matting techniques make use of trimaps, reference backgrounds, masks, or user commands. These inputs optimize quality but can break complete automation and may be impractical.

Inconsistent evaluation protocol: Existing studies have used different datasets, metrics, and test settings. This makes it hard to compare them directly and also restricts the progress on standardised evaluation for matting-based deep video spoof generation.

4. Discussion

The analysed literature points out that the evaluation of deep video spoof generation must take into account not only identity matching and perception realism but also the quality of compositing. Matting techniques allow for the smooth integration of altered foreground areas in real video settings. As a consequence, SAD, MSE, MAD, Grad, Conn, and dtSSD are metrics that can be employed for both the alpha quality score and its boundary alignment, as well as stability in time. But, those metrics need to be used judiciously as they are usually calculated on different datasets and/or with different input assumptions. A method that does well with a reference background or trimap may perform poorly under the almost unconstrained conditions of a deep video spoof-generating task.

The realistic matting-based deep video spoof generation framework needs to combine accurate alpha estimation, robust boundary refining, temporal consistency modelling, and colour or illumination harmonisation. It should also be tested in challenging real-world scenarios where major occlusion, rapid motion, and complex backgrounds exist, or green-screen capture is impractical. This approach may help combine the swapping models with real-composed video.

5. Conclusion

5.1 Summary of Findings

Our review examined the latest advances in deep generative models for both spoof generation and matting-based composition of the final visualisation. Our studies suggest that while state-of-the-art face swapping models have achieved substantial improvements in identity transfer and visual quality, the realism of spoofed videos remains highly dependent on foreground extraction, alpha matte accuracy, boundary refinement, illumination consistency, and temporal stability. In the specific subfields of image matting, backdrop matting, trimap-free portrait matting, high-resolution refinement, transformer-based matting, and diffusion-assisted matting/memory-guided video matting have all contributed to an increase in realism when compositing.

5.2 Principal Contributions

One of the most important contributions of this survey is to articulate deep video spoof generation and matting within a cohesive framework. Matting is vital for creating realistic spoofs, as it minimizes boundary artefacts, provides alpha estimation, and stabilises foreground/background blending between consecutive video frames. Apart from that, we categorize relevant studies by their methods, datasets, performance metrics, and limitations to clarify the relationship between spoof generation and matting-based compositing.

5.3 Recognised Research Deficiencies

Numerous research gaps persist. SOTA methods still struggle with heavy occlusion, inaccurate alpha predictions, hard fringe artifacts, temporal flickering, lighting inconsistency issues, and need additional inputs such as trimaps, masks, prompts, or other reference backgrounds. Moreover, there is currently no standardised evaluation methodology for matting-based deep video spoof generation, which makes comparing approaches difficult.

5.4 Future Directions

Future work can be directed along the path of effectively designing matting-based deep video spoof generation systems for uncontrolled real-world data. Important guidelines include improving occlusion handling, reducing temporal flickering, designing robust boundary refinement modules, synchronizing the illumination of foregrounds and backgrounds of a video, minimizing reliance on manual or auxiliary inputs (e.g., alpha trees), establishing standardized datasets for evaluating matting-based video spoof generation, as well as evaluation metrics that are consistent with authentication tasks. The task of producing realistic-looking outputs without controlled green-screen capture represents one particular practical avenue in that wide-ranging area of research.

References

- [1] R. Chen, X. Chen, B. Ni, and Y. Ge, "SimSwap: An efficient framework for high fidelity face swapping," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2020, pp. 2003–2011, doi: 10.1145/3394171.3413630.
- [2] X. Zhu et al., "One-shot face swapping on megapixels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, doi: 10.1109/CVPR46437.2021.00480.
- [3] R. Dhanyalakshmi, G. Stoian, D. Danculescu, and D. J. Hemanth, "A survey on face-swapping methods for identity manipulation in deepfake applications," *IET Image Processing*, vol. 19, no. 1, art. no. e70132, 2025, doi: 10.1049/ipr2.70132.
- [4] S. A. R. Syed Abu Bakar, S. Waseem, Z. Omar, et al., "Exploring the advancements and challenges of deepfake face-swap: A survey," *Multimedia Tools and Applications*, vol. 85, art. no. 14, 2026, doi: 10.1007/s11042-026-21285-8.
- [5] S. Sengupta, V. Jayaram, B. Curless, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Background matting: The world is your green screen," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 2291–2300, doi: 10.1109/CVPR42600.2020.00236.
- [6] S. Lin, L. Chen, Y. Wang, Z. Luo, and Y. Tai, "Real-time high-resolution background matting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 8762–8771, doi: 10.1109/CVPR46437.2021.00866.
- [7] Z. Ke et al., "MODNet: Real-time trimap-free portrait matting via objective decomposition," in *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, vol. 36, no. 1, 2022, pp. 1140–1147, doi: 10.1609/aaai.v36i1.19979.
- [8] J. Sun, J. Wang, Z. Tang, H. Guo, and C.-K. Tai, "Deep video matting via spatio-temporal alignment and aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 6975–6984, doi: 10.1109/CVPR46437.2021.00690.
- [9] P. Yang, S. Zhou, J. Zhao, Q. Tao, and C. C. Loy, "MatAnyone: Stable video matting with consistent memory propagation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 7299–7308.
- [10] J. Sun et al., "Ultrahigh resolution image/video matting with spatio-temporal sparsity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 14112–14121, doi: 10.1109/CVPR52729.2023.01356.
- [11] L. Huang et al., "SDMatte: Grafting diffusion models for interactive matting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025, pp. 15229–15239, doi: 10.48550/arXiv.2508.00443.
- [12] G. Park et al., "MatteFormer: Transformer-based image matting via prior-tokens," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 11696–11706, doi: 10.1109/CVPR52688.2022.01140.
- [13] J. Yao et al., "ViTMatte: Boosting image matting with pre-trained plain vision transformers," *Information Fusion*, vol. 103, art. no. 102091, Mar. 2024, doi: 10.1016/j.inffus.2023.102091.
- [14] J. Li et al., "ProxyMatting: Transformer-based image matting via region proxy," *Knowledge-Based Systems*, vol. 310, art. no. 112911, Feb. 2025, doi: 10.1016/j.knsys.2024.112911.
- [15] Y. Zhou et al., "Sampling propagation attention with trimap generation network for natural image matting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5828–5843, Oct. 2023, doi: 10.1109/TCSVT.2023.3260025.
- [16] X. Ye, Y. Liang, M. Tan, F. Feng, L. Wang, and H. Huang, "High-resolution natural image matting by refining low-resolution alpha mattes," *IEEE Transactions on Image Processing*, vol. 34, pp. 3323–3335, 2025, doi: 10.1109/TIP.2025.3573620.
- [17] X. Fang, S.-H. Zhang, T. Chen, X. Wu, A. Shamir, and S.-M. Hu, "User-guided deep human image matting using arbitrary trimaps," *IEEE Transactions on Image Processing*, vol. 31, pp. 2040–2052, 2022, doi: 10.1109/TIP.2022.3150295.
- [18] Y. Xu, B. Liu, Y. Quan, and H. Ji, "Unsupervised deep background matting using deep matte prior," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4324–4337, Jul. 2022, doi: 10.1109/TCSVT.2021.3132461.
- [19] Y. Wang et al., "From composited to real-world: Transformer-based natural image matting," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2097–2111, Apr. 2024, doi: 10.1109/TCSVT.2023.3300731.
- [20] B. Peng et al., "RGB-D human matting: A real-world benchmark dataset and a baseline method," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 4041–4053, Aug. 2023, doi: 10.1109/TCSVT.2023.3238580.

- [21] L. Dai et al., "Enabling trimap-free image matting with a frequency-guided saliency-aware network via joint learning," *IEEE Transactions on Multimedia*, vol. 25, pp. 4868-4879, 2023, doi: 10.1109/TMM.2022.3183403.
- [22] G. Yao and A. Sun, "Multi-guided-based image matting via boundary detection," *Computer Vision and Image Understanding*, vol. 243, art. no. 103998, Jun. 2024, doi: 10.1016/j.cviu.2024.103998.
- [23] A. Sun, J. Chang, and G. Yao, "DiffMatter: Different frequency fusion for trimap-free image matting via edge detection," *Computer Vision and Image Understanding*, vol. 259, art. no. 104424, Sep. 2025, doi: 10.1016/j.cviu.2025.104424.
- [24] J. Yao, X. Wang, L. Ye, and W. Liu, "Matte anything: Interactive natural image matting with segment anything model," *Image and Vision Computing*, vol. 147, art. no. 105067, Jul. 2024, doi: 10.1016/j.imavis.2024.105067.
- [25] D. C. Lepcha et al., "Image Matting: A comprehensive survey on techniques and applications," *International Journal of Image and Graphics*, 2023, doi: 10.1142/S0219467823500110.