Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/107/20241094

Advancing Image Animation: A Comparative Analysis of GAN-Based and Diffusion-Based Models

Yikai Sun

School of Computer Science, Middlebury College, VT, USA

yikais@middlebury.edu

Abstract. This paper provides an in-depth analysis of the latest advancements in image animation, focusing on two prominent models: Motion Representations for Articulated Animation (MRAA) and MagicAnimate. MRAA revolutionizes Generative Adversarial Networks (GANs)-based animation by employing regional descriptors instead of traditional key points, significantly enhancing the accuracy of motion capture and segmentation for complex articulated movements. MagicAnimate, on the other hand, utilizes a diffusion-based framework with temporal attention mechanisms, ensuring high fidelity and temporal consistency across animated sequences. The paper discusses the methodologies, datasets, and preprocessing techniques used in these models, offering a thorough comparison of their performance metrics, on various benchmark datasets. Through this comparative analysis, the paper highlights the strengths and limitations of these cutting-edge technologies, emphasizing MRAA's superior handling of complex movements and background dynamics, and MagicAnimate's excellence in identity preservation and temporal coherence. The study concludes by proposing future research directions, such as developing hybrid models that combine the advantages of GANs and diffusion techniques, to further enhance the realism, versatility, and control of image animation systems.

Keywords: MRAA, Generative Adversarial Networks (GANs), Regional Descriptors.

1. Introduction

Animation of static images, through advanced techniques like Generative Adversarial Networks (GANs) and diffusion models, has garnered significant attention due to its potential in creating animated movies, virtual avatars, and engaging social media content [1]. Enabled by the plentiful data available, data-driven approaches have made strides in enhancing the dynamic and interactive nature of visual content [2-4]. This evolution has minimized the need for costly traditional motion capture technologies, paving the way for lifelike digital representations.

Two dominant frameworks are utilized for human image animation: GAN-based and diffusion-based models. GAN-based methods [5-7] typically incorporate a warping network to adapt a reference image to a target pose, while employing GANs to address missing or obscured parts [7]. An exemplary GAN-based animation is the First Order Motion Model (FOMM) [5], which introduces a self-supervised key point detector for motion representation extraction and a dense motion network generating optical flow and occlusion maps. This model represents a remarkable breakthrough in object agnosticism and reduced reliance on extensive prior training.

DOI: 10.54254/2755-2721/107/20241094

Conversely, diffusion-based approaches [8] leverage appearance and pose conditions to generate the target image through pre-trained diffusion models [9-11]. MagicAnimate is a cutting-edge diffusion model that integrates video diffusion with temporal attention blocks, boosting temporal consistency and reference image fidelity [11]. It features a novel appearance encoder and ControlNet for seamless transitions in extended animations [10], delivering outstanding results on benchmarks like the TikTok dancing dataset.

Whereas headways in animation innovation have been noteworthy, there are still recognizable inconsistencies between created activities and genuine recordings. These contracts are credited to confinements in movement exchange capabilities in GAN-based models [12], as well as issues with temporal consistency in diffusion-based strategies [1]. Besides, the current need for controls for facial expressions and finger gestures includes these challenges. Future investigations in this field ought to center on coordination hybrid models that can address these restrictions successfully [11]. By improving control over complex motion dynamics, such as facial expressions and finger developments, analysts can progress the realism of automated animation systems. Furthermore, there's a requirement for advanced advancement to handle issues related to cross-identity animation scenarios, where exchanging movement between distinctive subjects remains difficult. Implementing more progressed temporal attention mechanisms seems to offer assistance in diminishing flashing and improving the smoothness of movement in longer groupings, pushing the boundaries of human picture animation. Hybrid models that combine the qualities of GANs and dissemination models offer promising arrangements, leveraging GANs' high-quality picture-era capabilities and dissemination models' capacity to deliver assorted and transiently steady yields. By investigating these roads, analysts can overcome current impediments and accomplish more reasonable and consistent activities.

This article offers an intensive investigation of two spearheading animation models: Motion Representations for Articulated Animation (MRAA) and MagicAnimate [11]. The consider dives into their models and functionalities, talking about their qualities, restrictions, and commitments to upgrading realism and flexibility in animation frameworks. It too proposes headings for future inquiries to address current challenges and progress in the field of vivified movement representations. This examination serves as a profitable asset for analysts and professionals pointing to move forward animation realism and flexibility. This article provides an in-depth analysis of two pioneering animation models, MRAA and MagicAnimate, highlighting their strengths, limitations, and contributions to enhancing realism and flexibility in animation systems. The study not only explores these models but also suggests future research directions to overcome current challenges. Its core significance lies in offering valuable insights for both researchers and practitioners aiming to advance the realism and adaptability of animated motion representations, driving future innovation in the field.

2. Methodology

2.1. Dataset description and preprocessing

A solid establishment of training datasets is basic for the improvement of compelling data-driven models in picture animation. Models like MagicAnimate [11] depend on datasets such as TikTok and Technology, Entertainment, Design talks (TED-talks) for preparation and assessment. Essentially, the Disentangled Control (DisCo) [1] and MRAA models utilize these datasets in their preparing forms. The TikTok dataset, for this case, contains an assorted extent of move recordings from the prevalent video-sharing stage, displaying people performing different move challenges in brief clips. These clips are particularly curated to include moves with direct developments, excluding those with over-the-top movement obscure to guarantee clarity and common sense for preparing purposes. By leveraging these comprehensive datasets, analysts can improve the execution and unwavering quality of data-driven models in picture animation, eventually progressing the capabilities of automated animation systems.

TED-talks dataset, sourced from YouTube, highlights clips of TED speakers, emphasizing upper body developments. Each clip, enduring at the slightest 64 outlines, grandstands controlled and

DOI: 10.54254/2755-2721/107/20241094

persistent development. Chosen for moderate developments, this dataset offers a controlled environment for preparing models in inconspicuous human expressions and signals.

The TaiChiHD dataset, crucial in assessing picture animation and video era models like Monkey-Net [7], FOMM [5], and MRAA [12], comprises 280 high-resolution Tai Chi recordings sourced from YouTube. Preprocessed into shorter clips resized to 256x256 and 512x512 pixels, these frame two subsets with lengths extending from 128 to 1024 outlines. The complex, non-rigid motions of Tai Chi work make this dataset an exceptional benchmark, posturing challenges and uncovering the vigor and viability of animation methods. See Figure 1, Figure 2 and Figure 3 for illustrative tests.

Figure 1. TikTok dataset samples.

Figure 2. TED-talks dataset samples.

Figure 3. TaiChiHD dataset samples.

DOI: 10.54254/2755-2721/107/20241094

2.2. Proposed approach

This article focuses on advanced image animation models, covering methodologies and experimental results. Section 1 traces the evolution of image animation tech, emphasizing the architectures and challenges of models like FOMM [5] and MagicAnimate [11]. Section 2 offers a comprehensive analysis of MRAA and MagicAnimate, outlining recent advancements in GAN-based and diffusion-based techniques. The "Results and Discussion" section presents a comprehensive evaluation of models using various metrics, comparing their performance across datasets. Insights into their strengths and limitations are provided. The conclusion highlights key advancements, suggests future research directions, and emphasizes the need for further improvements in image animation.

2.2.1. Background. Since their introduction by Ian Goodfellow and his colleagues in 2014, Generative Adversarial Networks (GANs) have revolutionized machine learning. Consisting of a generator (G) and a discriminator (D), GANs engage in a competitive game, where G produces data samples, such as images, designed to mimic authentic datasets, while D strives to distinguish between authentic and synthetic samples. The constant tussle between these two networks leads to increasingly realistic data generation, aiming to fool D. However, GANs are not without challenges, including training instability and mode collapse, which can restrict the diversity of their outputs. Stable Diffusion (SD), on the other hand, belongs to a class of Latent Diffusion Models (LDMs) that employ a sophisticated approach to image synthesis. LDMs shift the computationally expensive diffusion process from the high-dimensional pixel space to a lower-dimensional latent space, enabled by an autoencoder. This transformation significantly reduces computational costs while maintaining high visual fidelity. In this framework, an autoencoder efficiently encodes images into a compressed latent space. Then, a diffusion model operates within this space, progressively denoising the latent representations to reveal high-quality images. This method supports the generation of high-resolution images and offers flexibility for various applications like text-to-image synthesis, image inpainting, and super-resolution. Moreover, it is significantly more resource-efficient than traditional diffusion models.

2.2.2. GAN-based models. Siarohin et al.'s Monkey-Net may be a groundbreaking progression in deep learning and generative ill-disposed systems for enlivening a wide extend of objects. The key development of Monkey-Net lies in its object-agnostic system, which dispenses with the requirement for predefined models or explanations, extending its flexibility. By utilizing self-supervised key points extricated from a driving video to track developments, Monkey-Net can quicken objects in a source image by imitating the movement designs watched within the driving video. This approach empowers more energetic and versatile activities over different scenarios and sorts of objects, upgrading the adaptability and pertinence of automated animation systems. The inventive methods utilized by Monkey-Net have the potential to revolutionize the field of protest animation inside profound learning and GANs.

Extending on the spearheading work of Monkey-Net, the First Order Motion Model (FOMM) by Siarohin et al. presents progressed movement dealing with capabilities for quickening inactive pictures. FOMM joins nearby relative changes to offer exact control over movement representation, successfully overseeing complicated developments and closing key points. Moreover, FOMM highlights an occlusion-aware generator with an impediment veil, empowering consistent dealing with of ranges in activities that cannot be straightforwardly distorted from the source picture. This enhancement upgrades the by and large realism and quality of activities, particularly in scenarios including complex movements or critical occlusions. The developments showcased within the FOMM pipeline, as portrayed in Figure 4, build up FOMM as a modern apparatus in picture animation, giving unmatched realism and devotion to energized substance.

DOI: 10.54254/2755-2721/107/20241094

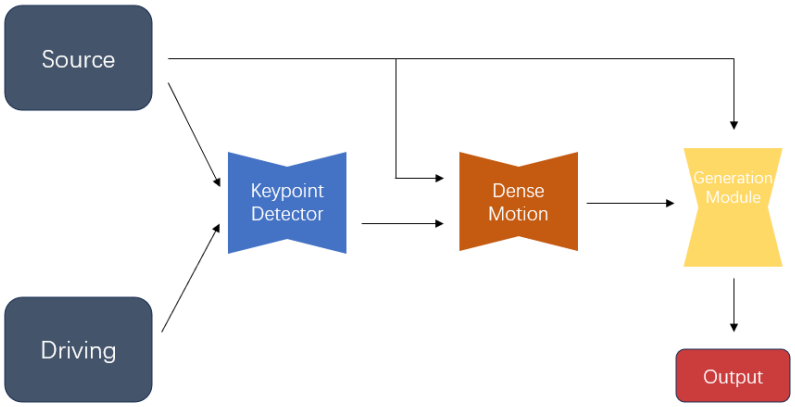

Figure 4. FOMM pipeline.

The MRAA introduces a special strategy of movement representation by utilizing territorial descriptors to degree first-order movement rather than depending on relapse procedures. This imaginative approach empowers the demonstration to capture the forms and shapes of personal protest parts more accurately, driving to move forward movement division. In differentiation from ordinary strategies, MRAA consolidates a component to handle foundation or camera developments. By anticipating parameters for a worldwide, relative change that depicts movements irrelevant to the question, the show successfully segregates the frontal area protest. This preparation stabilizes the focus of intrigued and improves demonstrate convergence by disposing of outside impacts, eventually refining the exactness and execution of movement representation in automated animation systems.

The method commences with the region's indicator, which makes heatmaps for each portion in both the source and driving pictures, outlined in Figure 5. These regions are at that point exchanged from the source picture to the driving picture. The changes in these locales and the foundation are combined by employing a pixel-wise stream expectation arrangement, associated with the one utilized by the First Order Motion Model (FOMM). Along these lines, the highlights from the source picture are encoded and distorted based on the calculated flow, eventually coming about within the era of the yield picture. This progressed technique not as it were raised the accuracy of movement capture but also boosts the solidness and viability of the demonstration in energetic settings, exhibiting critical changes in automated animation systems.

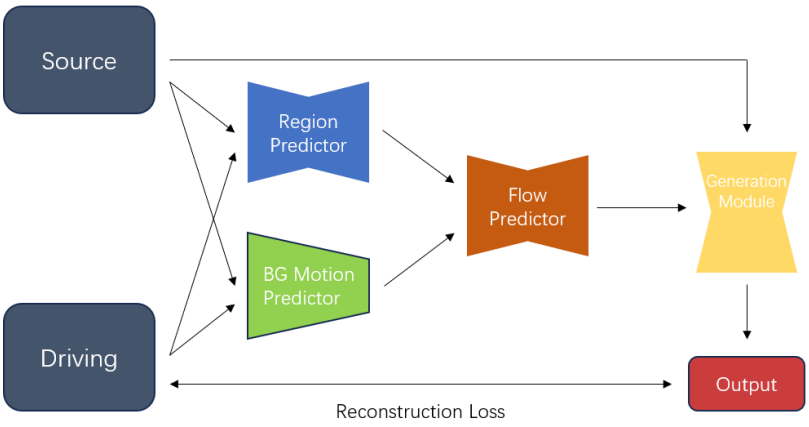

Figure 5. MRAA pipeline.

DOI: 10.54254/2755-2721/107/20241094

2.2.3. Diffusion-based models. The DisCo [1] demonstrates that could be a groundbreaking approach for making true human move developments inside a diffusion system. Disco stands out by advertising partitioned control over posture, human appearance, and foundation components, encouraging the generation of exact move groupings that are relevantly pertinent over different scenarios. By leveraging pre-trained dissemination models and coordination ControlNet to condition the era handle based on particular movement arrangements like move postures, DisCo overcomes common challenges in move animation. The outlined structure in Figure 6 empowers DisCo to exceed expectations in assignments requiring exact movement exchange and keeping up constancy in appearance and movement all through produced groupings. This inventive demonstration sets an unused standard for producing reasonable human move developments, exhibiting remarkable capability in capturing perplexing movement groupings and protecting genuineness in both appearance and development.

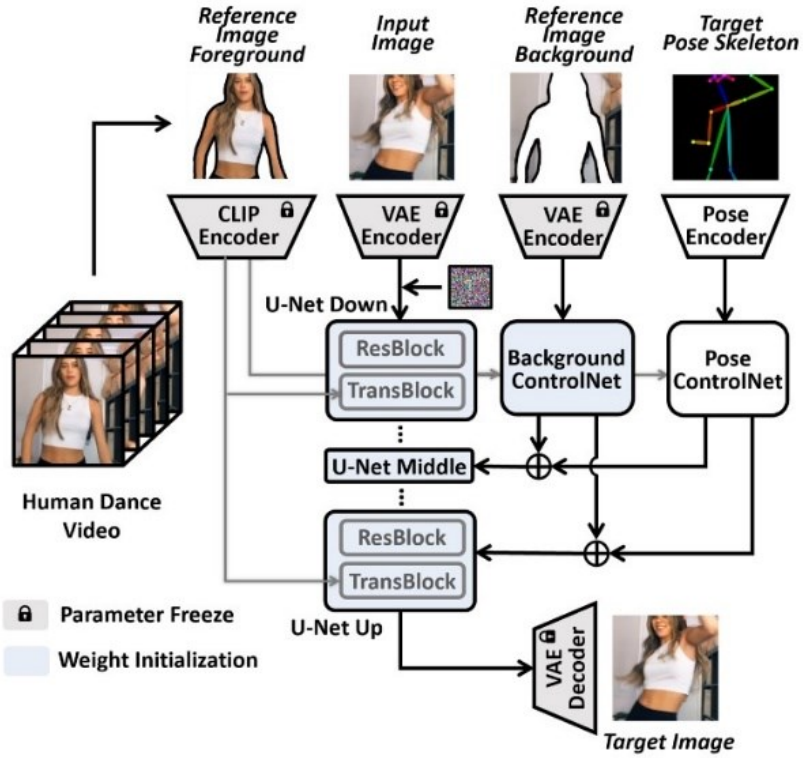

Figure 6. DisCo Pipeline [1].

MagicAnimate speaks to an inventive diffusion-based system custom-made for making human picture movements with worldly coherence. Not at all like routine approaches that handle video outlines autonomously, MagicAnimate consolidates temporal attention blocks to guarantee smooth moves and consistency between outlines. It presents a special appearance encoder that captures the complexities of the reference picture, counting personality and foundation subtle elements, upgrading the devotion of person outlines and generally transient coherence of the animation. Furthermore, MagicAnimate utilizes a straightforward video fusion method in induction to encourage consistent moves in extended liveliness. The MagicAnimate pipeline, outlined in Figure 7, illustrates how this show leverages progressed strategies to convey high-quality, coherent human picture movements with liquid and steady movement moves.

DOI: 10.54254/2755-2721/107/20241094

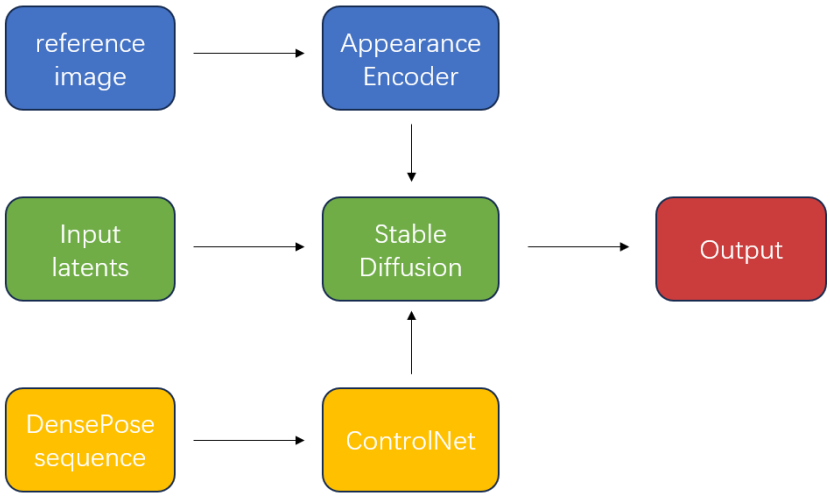

Figure 7. Illustration of MagicAnimate pipeline.

3. Result and Discussion

3.1. Evaluation Metrics

L1 error, a key metric in machine learning, assesses regression to demonstrate execution by measuring the absolute contrasts between predicted and genuine values. Calculated employing a direct equation, it evaluates the exactness of expectations within the show.

$$L1 \quad Error = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{1}$$

The Average Keypoint Distance (AKD) assesses the accuracy of anticipated key points compared to ground truth key points by averaging the Euclidean separations between them.

$$AKD = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2} \tag{2}$$

The Missing Keypoint Rate (MKR) measures the event of truant or undetected anticipated key points within the yield created by a posture estimation demonstration.

$$MKR = \frac{Number \quad of \quad missed \quad keypoints}{Total \quad number \quad of \quad keypoints} \tag{3}$$

Average Euclidean Distance (AED) extends on AKD by calculating the normal distance in Euclidean space between the anticipated focuses and their comparing ground truth focuses.

$$AED = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{j=1}^{d}(p_{ij} - \hat{p}_{ij})^2} \tag{4}$$

DOI: 10.54254/2755-2721/107/20241094

Peak Signal-to-Noise Ratio (PSNR) assesses the recreation quality in lossy compression codecs for pictures and recordings by differentiating the most noteworthy conceivable flag control with the control of interferometer clamor, communicated as:

$$MSE = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}[I(i,j) - K(i,j)]^2 \tag{5}$$

$$PSNR = 10 \cdot \log_{10}(\frac{MAX_I^2}{MSE}) \tag{6}$$

The Structural Similarity Index (SSIM) may be a metric created to compare the similitude between two pictures while upgrading conventional measures like peak signal-to-noise ratio (PSNR) and cruel squared mistake (MSE). It considers seen changes in structural information, giving a more precise portrayal of visual picture debasement.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{7}$$

Learned Perceptual Image Patch Similarity (LPIPS) evaluates the perceptual refinement between two pictures through deep learning characteristics. In differentiation from ordinary strategies centering on pixel-wise errors, LPIPS leverages neural arrange highlights to assess varieties taken after human visual elucidation.

Fréchet Inception Distance (FID) surveys likeness by comparing vector measurements from genuine and created pictures utilizing the InceptionV3 demonstration. A lower FID score indicates negligible contrasts and higher picture quality.

$$FID = \left\|\mu_r - \mu_g\right\|^2 + Tr(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r\Sigma_g}) \tag{8}$$

Fréchet Video Distance (FVD), akin to FID for pictures, is custom-made for recordings to gauge transient and spatial disparities by differentiating the highlight vectors extricated from genuine and created video clips.

3.2. Comparison of Monkey-Net, FOMM, and MRAA in quantitative assessments

The assessment of Monkey-Net, FOMM, and MRAA's performance in reconstruction and animation tasks reveals MRAA's superiority. Analyzing the results in table 1, MRAA demonstrates the lowest L1 error, AKD, MKR, and AED scores on TaiChiHD and TED-talks datasets, showcasing superior accuracy and fidelity in reconstructions. In contrast, Monkey-Net records the highest metric values, indicating substantial reconstruction errors. MRAA's consistent strong performance across various resolutions and region counts underscores its scalability and effectiveness.

MRAA showcases significant enhancements in animating intricate limb movements, as evidenced in a user preference study with 50 videos evaluated by 50 users each. Results detailed in table 2 illustrated a clear preference for MRAA over FOMM in articulating human bodies. FOMM faced challenges in accurately animating essential body parts like hands, whereas MRAA excelled in faithfully recreating these elements according to driving poses, resulting in heightened user satisfaction.

DOI: 10.54254/2755-2721/107/20241094

Table 1. Comparison of Monkey-Net, FOMM, and MRAA in video reconstruction on three datasets [12].

Method	TaiChiHD (256)			TaiChiHD (512)			TED-talks		
	L1	(AKD, MKR)	AED	L1	(AKD, MKR)	AED	L1	(AKD, MKR)	AED
Monkey-Net	0.077	(10.80,0.059)	0.228	-	-	-	-	-	-
FOMM	0.056	(6.53,0.033)	0.172	0.075	(17.12,0.066)	0.203	0.033	(7.07,0.014)	0.163
MRAA	0.047	(5.58,0.027)	0.152	0.064	(13.86,0.043)	0.172	0.026	(3.75,0.007)	0.114

Table 2. User preference study favoring MRAA over FOMM summarized [12].

Dataset	MRAA vs FOMM (%)
TaiChiHD (256)	83.7%
TED-talks	96.6%

3.3. Comparative analysis of MRAA, DisCo, and MagicAnimate quantitatively assessed

Table 3 outlines a comprehensive comparison of MagicAnimate with baseline models DisCo and MRAA on TikTok and TED-talks datasets. MagicAnimate outperforms in reconstruction metrics like L1, PSNR, SSIM, and LPIPS on TikTok, showing significant enhancements—6.9% better SSIM and 18.2% in LPIPS—compared to DisCo. It also excels in video fidelity, with a superior FVD score, demonstrating its strong ability in maintaining video quality.

MagicAnimate demonstrates superiority on the TED-talks dataset, with the highest FVD score compared to MRAA. It excels in single-frame fidelity with the best FID score. However, MRAA outperforms in L1 error due to better background motion control. Nevertheless, MagicAnimate shines in AKD, MKR, and AED metrics, showcasing its exceptional ability in identity preservation and complex movement animation.

Table 3. Comparison of MRAA, DisCo, and MagicAnimate on two benchmarks [11].

(a) Numerical analyses conducted on the TikTok dataset.

Model	L1	PSNR	SSIM	LPIPS	FID	FVD
MRAA [12]	4.61E-04	28.39	0.646	0.337	85.49	468.66
DisCo [1]	3.78E-04	29.03	0.668	0.292	30.75	292.80
MagicAnimate [11]	3.13E-04	29.16	0.714	0.239	32.09	179.07

(b) Detailed comparisons on the TED-talks dataset.

Model	AKD	MKR	AED	L1	FID	FVD
MRAA [12]	4.37	0.024	0.246	1.61E-04	35.75	182.78
DisCo [1]	2.96	0.019	0.253	2.07E-04	27.51	195.00
MagicAnimate [11]	2.65	0.013	0.204	2.92E-04	22.78	131.51

3.4. Discussion

MRAA showcases expertise in articulating complex movements, especially with its advanced region representation, particularly in limb animations [13]. Its specialized background motion predictor contributes to improved reconstruction accuracy. MagicAnimate impresses in identity preservation and fidelity, indicating suitability for high-quality human figure animation. Nonetheless, both models have areas for enhancement, with MRAA possibly needing refinement in single-frame fidelity and

DOI: 10.54254/2755-2721/107/20241094

computational efficiency, while MagicAnimate could improve background handling and reduce sensitivity to minor variations between frames. Streamlining the model's complexity may also assist in handling speed without compromising yield quality [14,15]. Both models are perfect for applications in virtual reality, film production, and video games, emphasizing the significance of high-quality human figure animation. Future improvements for MRAA may include consolidating versatile foundation modeling strategies to address different scenarios and improve single-frame exactness. Making strides in the model's effectiveness seems hoist its convenience in real-time frameworks. This study critically evaluates the strengths and limitations of MRAA and MagicAnimate, offering constructive insights into their potential future enhancements. While MRAA excels in detailed motion articulation and accuracy, improvements in single-frame fidelity and computational efficiency are suggested. Similarly, MagicAnimate's strength in identity preservation could benefit from enhanced background handling and robustness to frame variations. By proposing refinements in model complexity and efficiency, the study paves the way for optimizing these models in real-time applications, ensuring their continued relevance in virtual reality, film production, and video game animation.

4. Conclusion

This paper delved into recent advances in the field of image animation, with a particular focus on MRAA and MagicAnimate models. The MRAA model innovates the motion representation based on GAN by introducing the region representation method, so as to better enhance complex joint movements such as limb movement. The MagicAnimate model uses a diffusion-based framework and temporal attention mechanism to ensure the high fidelity and coherence of the animated sequences. Evaluations on datasets such as TaiChiHD and TED-talks show that MRAA models excel at handling complex actions and background dynamics, while MagicAnimate models excel at maintaining identity consistency and temporal consistency. The performance of these models in improving image animation technology is verified by SSIM and LPIPS. Future research should focus on developing hybrid models that integrate GANs and diffusion methods to further improve the accuracy of motion transfer and animation. At the same time, improving background modeling technology and strengthening the control of complex motion dynamics such as facial expressions and gestures are also key directions for future development. These advances will greatly enhance the realism and applicability of image animation in digital media and interactive environments, pushing existing technologies forward.

References

[1] Wang T Li L Lin K 2024 Disco: Disentangled control for realistic human dance generation In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp 9326-9336

[2] Chan C Ginosar S 2019 Everybody dance now In Proceedings of the IEEE/CVF international conference on computer vision pp 5933-5942

[3] Geng Z Cao C Tulyakov S 2019 3d guided fine-grained face manipulation In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp 9821-9830

[4] Xu Z Zhang J Liew J H 2024 Xagen: 3d expressive human avatars generation Advances in Neural Information Processing Systems 36

[5] Siarohin A Lathuilière S Tulyakov S Ricci E and Sebe N 2020 First order motion model for image animation. Advances in neural information processing systems 32

[6] Wang T C Mallya A and Liu M Y 2021 One-shot free-view neural talking-head synthesis for video conferencing In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp 10039-10049

[7] Siarohin A Lathuilière S Tulyakov S Ricci E and Sebe N 2019 Animating arbitrary objects via deep motion transfer In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp 2377-2386

DOI: 10.54254/2755-2721/107/20241094

[8] Karras J Holynski A Wang T C and Kemelmacher-Shlizerman I 2023 Dreampose: Fashion image-to-video synthesis via stable diffusion In 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp 22623-22633
[9] Radford A et al. 2021 Learning transferable visual models from natural language supervision In International conference on machine learning pp 8748-8763
[10] Zhang L Rao A and Agrawala M 2023 Adding conditional control to text-to-image diffusion models In Proceedings of the IEEE/CVF International Conference on Computer Vision pp 3836-3847
[11] Xu Z et al. 2023 Magicanimate: Temporally consistent human image animation using diffusion model In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp 1481-1490
[12] Siarohin A Woodford O 2021 Motion representations for articulated animation In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp 13653-13662.
[13] Huang J Yan M Chen S 2024 MagicFight: Personalized Martial Arts Combat Video Generation In ACM Multimedia pp 202
[14] Wei Y Shan W Zhang Q 2024 Real-Time Interaction with Animated Human Figures in Chinese Ancient Paintings In 2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW) pp 1-6
[15] Hou J Lu Y Wang M 2024 A Markov Chain approach for video-based virtual try-on with denoising diffusion generative adversarial network Knowledge-Based Systems vol 300 p 112233