FastAnim8: Efficient Facial Motion Transfer to Animated Character Images

Sahil Nagaralu
Dept. of Artificial Intelligence
Symbiosis Institute of Technology
Pune, India
sahil.nagaralu.btech2021@sitpune.edu.in

Sanskar Jadhav
Dept. of Artificial Intelligence
Symbiosis Institute of Technology
Pune, India
sanskar.jadhav.btech2021@sitpune.edu.in

Roshan Yadav
Dept. of Artificial Intelligence
Symbiosis Institute of Technology
Pune, India
roshan.yadav.btech2021@sitpune.edu.in

***Abstract*—This paper presents FastAnim8, an original algorithm for facial animation that addresses key challenges in existing techniques. FastAnim8 stands out for its lightweight model design and computational efficiency, making it highly accessible and versatile. Unlike traditional methods reliant on GPUs and extensive manual intervention, FastAnim8 streamlines the animation process, minimizing both computational resources and human effort. Through a combination of deep learning and computer vision, FastAnim8 accurately captures and transfers facial motion from human subjects to animated characters. Empirical evaluation demonstrates FastAnim8's superior performance and scalability across various datasets. By offering a more efficient and cost-effective solution, FastAnim8 aims to democratize facial animation technology, enabling broader adoption and innovation in the field.**

***Index Terms*—facial animation, lightweight model, motion transfer, deep learning, computer vision**

I. Introduction

Animation, an integral part of modern visual media, has long been challenged by the complexity of capturing and reproducing human facial expressions convincingly. While traditional methods rely heavily on manual frame-by-frame animation, recent advancements in deep learning and computer vision have paved the way for more automated, efficient techniques.

In the pursuit of lifelike animation, researchers have explored various approaches to bridge the gap between human motion and animated characters. Models like ControlNet in Stable Diffusion and WarpFusion have emerged as notable contenders, aiming to map the intricacies of human motion onto AI generated people. These methods, while promising, often require extensive manual intervention and lack the scalability required for real-world applications.

We propose FastAnim8, a novel algorithm designed to streamline the process of facial animation through the application of image processing and deep learning. FastAnim8 leverages specific neural networks and custom motion tracking algorithms to transfer facial motion from a human subject to an animated character's face image. By automating the mapping process, FastAnim8 eliminates the need for consistent manual intervention, offering a more efficient and scalable solution for facial animation. The development of FastAnim8 builds upon existing research in the field of computer vision and machine learning. Drawing inspiration from techniques such as facial landmark detection and optical flow estimation, FastAnim8 enhances the fidelity and expressiveness of animated characters by accurately capturing subtle facial movements. Moreover, by integrating iterative refinement processes and a feedback mechanism for evaluation, FastAnim8 can be trained to continuously improve its performance over time, adapting to a wide range of facial expressions and contexts.

Beyond its applications in entertainment, FastAnim8 holds promise in various domains, including virtual reality and human-computer interaction. By enabling the motion of lifelike avatars and virtual assistants, FastAnim8 facilitates more immersive and engaging user experiences, paving the way for new forms of interaction in the digital realm.

In this paper, we present a detailed analysis of FastAnim8, exploring its underlying principles, technical implementation, and practical implications. Through empirical evaluation and experimental demonstrations, we demonstrate the effectiveness of FastAnim8. Additionally, we discuss potential avenues for future research and development.

II. Literature Review and Research Gap

We went through existing literature as guidance for building our algorithm. Since our work required breaking down the final aim into fragments and each part is a unique problem statement, we reviewed literature relevant to both the individual problems and the combined general aim. The detailed literature analysis in this domain is discussed in Table 1.

DeepVideoPaint targets high-fidelity manipulation in videos [1]. It excels at replacing objects but struggles with complex interactions between those objects. FastAnim8 focuses on facial animation, achieving automatic and scalable motion transfer between a source subject and a target character's face image.

Like FastAnim8, the work by Liu et al. allows for complex object manipulation in videos [2]. However, their method relies on large training datasets and is computationally expensive. FastAnim8 prioritizes efficiency through image processing and deep learning techniques, making it more suitable for real-time applications.

The work presented by Zhou et al. shares the ability to control specific object animation [4]. However, it relies on

TABLE I
DETAILED LITERATURE REVIEW

Literature/Works	Methodology	Pros	Cons
DeepVideoPaint [1]	Uses Generative Adversarial Networks (GANs) to replace objects in real-time video.	Achieves high-quality results, good at maintaining temporal coherence.	Limited object variety, struggles with complex object interactions.
Spatio-Temporal Object Manipulation in Videos [2]	Leverages attention mechanisms and transformers to manipulate object appearance and motion in videos.	Handles complex object manipulations, allows for interactive control.	Requires large training datasets, high computational cost.
Neural Object Rewinding [3]	Employs a recurrent neural network (RNN) to generate frames where objects appear to rewind their motion.	Offers a unique visual effect, good for creative applications.	Limited object types, works best with simple motions.
Deep Exemplar-Based Image Animation [4]	Utilizes exemplar images to guide the animation of objects within a video sequence.	Enables control over specific object animations, handles partial occlusions.	Requires high-quality exemplar images, can be computationally expensive for long videos.
Monocular Articulated 3D Character Animation [5]	3D Body Model Generation, Pose Estimation and Tracking and Style Transfer Network	Animates entire body with style transfer for consistency and potentially handles complex body movements.	3D model creation might add complexity and requires more computational resources.
Semantic Image Editing with Image Inpainting [6]	Applies image inpainting techniques for object manipulation in static images, potentially adaptable to video.	Proven effective for basic object removal/replacement, conceptually applicable to live video.	Not specifically designed for video, may struggle with frame-to-frame coherence.
Deep Video Inpainting [7]	Investigates deep learning methods for video inpainting, potentially applicable to object manipulation.	Offers a foundation for future video object manipulation techniques.	Primarily focused on inpainting missing regions, requires further development for object manipulation.
Learning to Segment Moving Objects in Videos [8]	Develops algorithms for segmenting objects in videos, a crucial step for object manipulation.	Improves object segmentation accuracy, aiding future manipulation techniques.	Focused on segmentation, not manipulation itself, further research needed.
Mask R-CNN [9]	Proposes a deep learning model for object detection and segmentation, useful for identifying objects in live video.	Offers real-time object detection capabilities, valuable for preprocessing in manipulation.	Not specifically designed for manipulation, requires additional steps for animation.
FlowNet 2.0: Deep Learning for Optical Flow Estimation [10]	Introduces a deep learning architecture for estimating optical flow, essential for understanding object motion in video.	Improves accuracy of optical flow estimation, facilitating realistic object animation.	Focused on flow estimation, not manipulation itself, needs integration with other methods.
Thin-Plate Spline Motion Model for Image Animation [11]	Thin-Plate Splines & Multi-Resolution Masks for Unsupervised Image Animation	Enables flexible animation, handles occlusions, learns without needing labels	May struggle with large pose differences, increases model complexity, requires more training data.
Animating Arbitrary Objects via Deep Motion Transfer [12]	Deep Motion Transfer with Keypoint Detection & Heatmaps	Versatile object application, Motion capture, Independent motion control	Training data dependence, Computational cost, Black box nature
Latent Image Animator: Learning to Animate Images via Latent Space Navigation [13]	Variational Autoencoder (VAE) for style transfer between images and videos	Transfers animation style from videos to images, potentially applicable for creating anime-styled animations	May struggle with complex scenes and maintaining object consistency during style transfer
Structure-aware Video Style Transfer with Map Art [14]	Deep Neural Network with attention mechanism	The "map art" style transfer can provide a base for building anime-style visuals. Elements like flat colors and sharp lines are common in anime aesthetics.	While the technique alters the visual style, it doesn't directly address character animation, a crucial aspect of anime.
Anime-Like Motion Transfer with Optimal Viewpoints [15]	Anime-Like Motion Transfer with Optimal Viewpoints	Effectively extracts suitable poses for lower frame rates, Relieves redundancy in motions due to physical speed constraints	Inconsistent emphasis on speed, Difficulty in identifying character positions due to monotonous background

exemplar images and high-quality reference images representing the desired animation for the object. For example, if you want to animate a car turning a corner, you might provide an exemplar image of a car at each turn stage. However, FastAnim8 utilizes deep learning for automatic motion transfer, eliminating the need for high-quality exemplar images and reducing computational costs for long videos.

Video-In-Painting (VIP) by Liu et al. artistically transforms videos, but it lacks control over specific objects and may introduce artefacts in complex scenes [5]. Our work focuses on anime character facial animation while maintaining the original video style through careful design choices.

The work on semantic image editing with image inpainting by Iizuka et al. [6] does explore potential applications in video manipulation. However, it's not specifically designed for videos and may struggle with frame coherence.

While Mask R-CNN by He et al. [9] offers real-time object detection and FlowNet 2.0 by Ilg et al. [10] improves optical flow estimation, neither directly address animation. We utilized these techniques as building blocks to aid the core functionality of facial motion transfer.

III. Proposed Methodology

A. Background Substitution

Background substitution is a technique used in image and video editing to replace the background of an image or video with a different one. This process involves isolating the foreground subject (such as a person or object) from its original background and then inserting it into a new background scene.

1) Preprocessing: The preprocessing stage initiates with the application of histogram equalization to the image. This technique enhances image quality by adjusting the distribution of pixel intensities, thereby improving contrast and overall clarity. By standardizing the histogram, variations in brightness and contrast are mitigated, facilitating subsequent processing steps.

2) Edge Detection: Following preprocessing, the Canny edge detection algorithm is employed to identify significant edges within the image. This algorithm operates by detecting gradients in pixel intensity, identifying regions of abrupt change indicative of object boundaries. The resulting edge map provides a comprehensive outline of the character, crucial for subsequent segmentation and masking procedures.

3) Contour Masking: With the edge map established, contours are delineated around the character using the identified edges. These contours serve as a precise boundary delineation for the character, effectively separating it from the background. Through contour tracing, a cohesive mask is generated, isolating the character and facilitating accurate background removal in the next stage.

4) Background Replacement: The background replacement process commences with the creation of a binary mask encapsulating the character. This mask effectively masks out the character while preserving the background. With the inverse of this mask, the original background is retained while the character is removed. The resultant composite image combines the character with the desired background, yielding a cohesive and visually appealing composition.

B. Face Landmark Detection and Mapping

Face landmark detection and mapping is the process of identifying and locating key points on a human face, such as the eyes, nose, mouth, and other facial contours. These points, known as landmarks or keypoints, are used to describe the shape and structure of the face. Standard face recognition algorithms analyze images or video frames to automatically detect these landmarks.

1) Creation of GUI for Annotation: To facilitate the precise annotation of facial landmarks, a tailored Graphical User Interface (GUI) was developed. This GUI serves as a user-friendly tool for annotating key landmarks on animated faces using human face mesh landmarks as reference. Through the GUI, users can easily as many distinctive facial landmarks as they desire. For FastAnim8, we considered 81 landmarks available in dlib, a cross-platform machine learning toolkit. By standardizing the annotation process and providing intuitive controls to zoom in on selected areas, the GUI ensures consistency across annotations and minimizes potential discrepancies in landmark placement.

2) Landmark Mapping and Warping: After detecting and storing the positions of the landmarks on the human face in the video input, the next step involves mapping the landmarks from the first frame onto the corresponding positions of the given animated face. Then, thin plate spline warping techniques are employed to adjust the structure and proportions of the animated face in order to align it with the annotated landmarks from subsequent frames of the human face. This adjustment ensures that the animated face conforms to the desired facial expressions and poses as shown by the human. By leveraging advanced geometric transformations, namely affine and non-linear warping, the animated face is manipulated to match the spatial arrangement of landmarks from the human face, creating a smooth transition between different animation styles while preserving the integrity and anatomical accuracy of the facial features and avoiding the uncanny valley of irregular motion.

C. Aligning Frames for Conversion to Video

This refers to the process of ensuring consistency and continuity between individual frames of a video sequence. When creating animations or video effects involving multiple frames, it is crucial to align the pictures properly to maintain visual coherence and smooth transitions. This alignment process may involve adjusting the position, scale, rotation, and timing of individual frames to establish ideal integration and motion synchronization.

1) Placing Transparent Character Face: In this step, the transparent character face is overlaid onto the original video frames. This process involves precise alignment of the character's facial features with corresponding positions in the video frames.

2) Inpainting Black Regions: In cases where character movement leaves behind black regions or gaps in the video frames, inpainting techniques are employed to fill these areas. Inpainting algorithms analyze surrounding pixels and logically generate plausible replacements for the missing regions, resulting in smooth transitions between frames and preserving the visual continuity of the animation.

3) Edge Blurring for Motion Blur: To simulate motion blur and enhance the realism of character movement, the edges of the character's face are intentionally blurred. This blurring effect mimics the natural phenomenon of motion blur observed in fast-moving objects, adding a sense of dynamism and fluidity to the animation. By selectively blurring the edges of the character's face, the animation achieves a more lifelike appearance and captures the nuances of movement more convincingly.

4) Testing Facial Feature Movement: Throughout the video, the movement of facial features is tested for accuracy and consistency. This involves analyzing the trajectory and behavior of individual facial features, such as the eyes, eyebrows, mouth, and nose, across consecutive frames. By ensuring that

facial features move in a realistic synchronized manner, the animation maintains integrity.

5) Setting a Bounding Box: To optimize tracking and rendering processes, a bounding box is strategically positioned near the character's face as the region of interest. This bounding box serves as a reference area for efficient tracking and rendering of facial movements. By focusing computational resources on the defined region, the animation workflow becomes more streamlined, resulting in faster processing times and improved performance.

D. Super Resolution of Each Frame in Video

1) Custom Super Resolution Wasserstein GAN: We utilize a custom Super Resolution Wasserstein Generative Adversarial Network (GAN) specifically trained on a curated dataset of selected anime images [16]. This GAN model is designed to enhance the resolution and quality of low-resolution images by learning the mapping between low-resolution and high-resolution images.

2) Resolution Enhancement Process: Each frame of the video undergoes an individual enhancement process using the trained GAN model. The low-resolution frames are inputted into the model, which then generates corresponding high-resolution versions. This process aims to improve the visual quality and fidelity of the video by enhancing details.

3) Compression into Video Format: Once the super resolution process is complete for all frames, the enhanced frames are compressed back into a video format. This compressed video retains the improved resolution and quality achieved through the super resolution process while ensuring efficient storage of the video data.

IV. Experimental Results

A. Background Substitution

Figure 1 displays the input character image and background image to be overlaid on. Figure 2 displays the output from Canny edge detection, contour masking, and background replacement. The goal was to isolate only the head and neck of the character. For this paper, we are considering one character for consistency throughout each subsection, but our algorithm is designed to work on any character.

It is important to note that these processes are only needed to be executed once, following which all successive frames will originate from this altered one.

Fig. 1. Input of (a) character face image (b) background image

Fig. 2. Output of (a) edge detection (b) contour masking (c) background substitution

B. Facial Landmark Mapping

Figure 3 shows an example of the 81 facial landmarks detection algorithm offered by dlib in Python. A real image of a human face, preferably smiling for visibility of all mouth landmarks, is considered as a reference, after aligning in orientation with the character image.

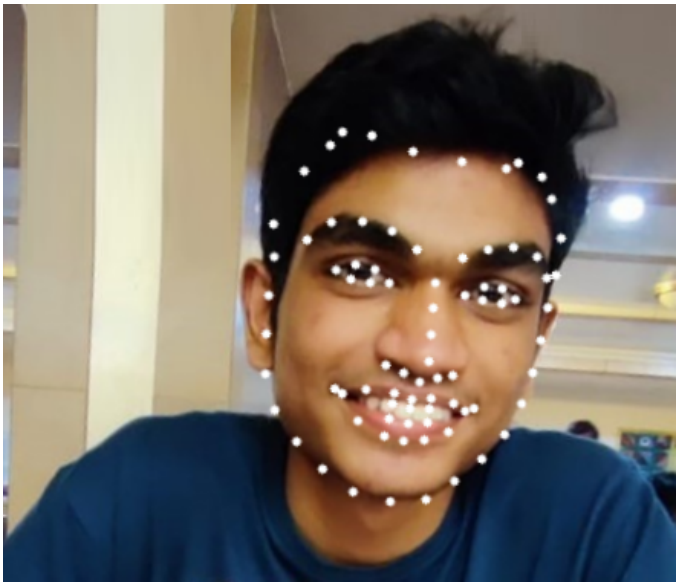

Fig. 3. 81 facial landmarks detected on a human face using dlib and OpenCV (image of author Sanskar Jadhav)

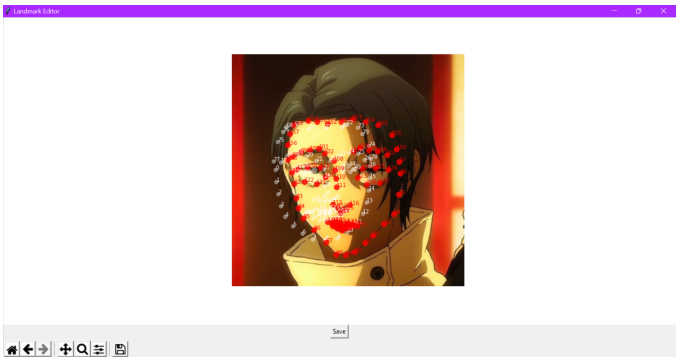

Fig. 4. Custom GUI for landmark annotation on character image with human facial landmarks for reference

Figure 4 exhibits the GUI built for placing each corresponding landmark of 81 in total onto the character's face. The white dots represent the human face landmarks to be used as reference when placing the red dots i.e. the character face landmarks. Once the Save button is clicked, the pixel coordinates of each landmark on the image plane is saved in a CSV file.

C. Creation of Successive Animated Frames

The input video of a human speaking was first split into individual frames and passed through the dlib facial detection algorithm to obtain the landmark positions, filtering out any noisy or blurry frames which cannot be accurately annoted with landmarks. For testing our model's performance, we used the VoxCeleb2 dataset [17]. It is an audio-visual dataset consisting of short clips of human speech in square resolution, extracted from interview videos uploaded to YouTube. The square resolution was crucial as we needed to work with bounding boxes for focusing on our region of interest i.e. the face and head movement.

Once the landmark positions for each frame were recorded and stored in CSVs, the scaled difference between x and y coordinates in the image plane for all landmarks in consecutive frames was calculated and the same transformation was replicated onto the landmark positions of the animated character's face.

Figure 5 illustrates an example of the landmark transformation performed on the original character face. Specifically affine transformations and thin plate spline warping was done to match the positions of all 81 face landmarks in order to recreate the same facial expression and orientation.

Fig. 5. Warping of animated character's face to new landmark positions

Now, having obtained the warped character images, the next phase involved replacing the original character image. Simply overlaying the new image would not work as warping and affine transformations also change the relative size of the image, leading to inconsistencies in image positioning upon overlaying, as seen in Figure 6. Hence, we created a bounding box for easier positioning since we get a fixed area for all movement. The margins of this box can be tuned with respect to the size of the animated character's head and face. For our testing on the VoxCeleb2 dataset, square video resolutions allowed us to encompass the entire area as a bounding box.

Fig. 6. Direct overlaying of warped image onto original without positioning

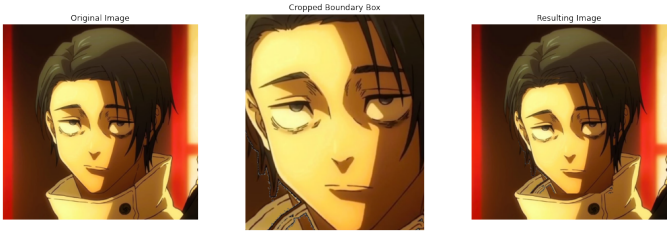

Fig. 7. Overlaying of warped image using bounding box for positioning

As can be seen in Figure 7, the results from the bounding box positioning were more appealing and thus, this method was continued and propagated through all frames in the video.

D. Final Result

We tested the algorithm with a sample from the VoxCeleb2 dataset and then with videos of ourselves speaking. The results are depicted in Figures 8 and 9. Note that the model output is a video, and the below results are of individual frames in the output video created.

Fig. 8. Model output on a sample video from VoxCeleb2 dataset

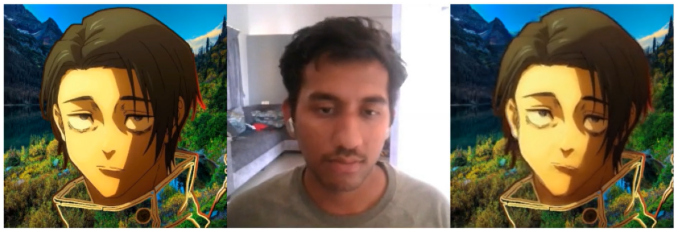

Fig. 9. Model output on a recorded video of author Sahil Nagaralu

With the addition of our custom trained SRGAN, we could also animate the mouth specifically despite the original image of the character having a closed mouth. The results from the

implementation of SRGAN on the output video can be seen in Figure 10.

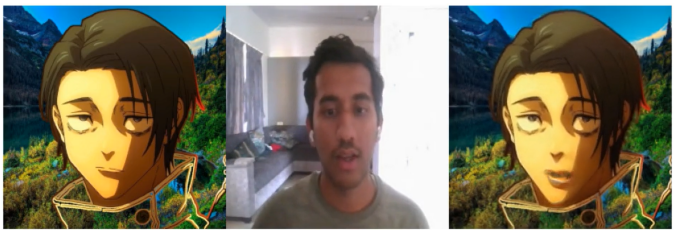

Fig. 10. Model output with SRGAN implemented on a recorded video of author Sahil Nagaralu

V. Conclusion and Future Prospects

FastAnim8 stands to be a groundbreaking advancement in facial animation, pioneering a computationally inexpensive deep learning and computer vision driven approach. Its lightweight design stands in stark contrast to traditional methods, characterized by resource-intensive GPU processing and substantial manual intervention throughout the animation process. FastAnim8 minimizes the need for manual input by maintaining the same character design throughout all frames, thus simplifying the animation process while simultaneously broadening accessibility for a wider range of users.

Our evaluation with existing and custom datasets of videos has demonstrated the effectiveness of FastAnim8 in accurately capturing and transferring facial motion from human subjects to animated characters.

Looking ahead, there are several avenues for future research and development. Through our literature review, it was brought to our attention that research in this domain is limited due to a lack of datasets for facial landmark detection on animated character faces and for similar character designs in alternate perspectives. We aim to further refine performance by exploring larger facial landmark sets and training a model to learn facial recognition on animated character faces.

Current datasets of animated faces tend to either possess images with minimal variation or drastic variation, both of which are undesirable. We aim to contribute to the creation of animated character datasets by working with LoRAs to generate custom animated character images with similar design to our training dataset of human-like characters, with variation mainly in perspective rather than details.

Additionally, we plan to investigate its applicability in other domains beyond entertainment, such as in education or medicine, where an animated character can interact with children, increasing their interest and attention span in education and relieving them of stress or fear in the hospital. For now, the videos generated do not have any audio, but a potential future improvement includes working with AI voice simulators to recreate the animated character's voice using samples recorded from the respective cartoon show, anime, or movie to have the character speak custom text in their voice.

References

[1] Gao, Y., Cao, Y., Kou, T., Sun, W., Dong, Y., Liu, X. & Zhai, G. (2023). Vdpve: Vqa dataset for perceptual video enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1474-1483).

[2] Ding, S., Zhao, P., Zhang, X., Qian, R., Xiong, H., & Tian, Q. (2023). Prune spatio-temporal tokens by semantic-aware temporal accumulation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 16945-16956).

[3] Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y. & Wang, J. (2023). NeuralDome: A neural modeling pipeline on multi-view human-object interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8834-8845).

[4] Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., ... & Wen, F. (2023). Paint by example: Exemplar-based image editing with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18381-18391).

[5] Kappel, M., Golyanik, V., Elgharib, M., Henningson, J. O., Seidel, H. P., Castillo, S., ... & Magnor, M. (2021). High-fidelity neural human motion transfer from monocular video. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1541-1550).

[6] Yeh, R. A., Chen, C., Yian Lim, T., Schwing, A. G., Hasegawa-Johnson, M., & Do, M. N. (2017). Semantic image inpainting with deep generative models. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5485-5493).

[7] Kim, D., Woo, S., Lee, J. Y., & Kweon, I. S. (2019). Deep video inpainting. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5792-5801).

[8] Tokmakov, P., Schmid, C., & Alahari, K. (2019). Learning to segment moving objects. International Journal of Computer Vision, 127, 282-301.

[9] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[10] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., & Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2462-2470).

[11] Zhao, J., & Zhang, H. (2022). Thin-plate spline motion model for image animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3657-3666).

[12] Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & 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).

[13] Wang, Y., Yang, D., Bremond, F., & Dantcheva, A. (2022). Latent image animator: Learning to animate images via latent space navigation. arXiv preprint arXiv:2203.09043.

[14] Le, T. N. H., Chen, Y. H., & Lee, T. Y. (2023). Structure-aware Video Style Transfer with Map Art. ACM Transactions on Multimedia Computing, Communications and Applications, 19(3s), 1-25.

[15] Koroku, Y., & Fujishiro, I. (2022). Anime-Like Motion Transfer with Optimal Viewpoints. In SIGGRAPH Asia 2022 Posters (pp. 1-2).

[16] M. Ando, A. D. Kaplan, and A. K. Sims, "Danbooru2021: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2021 (pp. 4910-4917).

[17] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019 (pp. 2327-2335).