


Semi-Automatic Generation of Rotoscope Animation Using SAM and k-means Clustering

Mizuki Sakakibara¹ and Tomokazu Ishikawa^{1,2} 

¹*Toyo University, 1-7-11 Akabanedai Kita-ku, Tokyo, Japan*

²*Prometech CG Research, 3-34-3 Hongo Bunkyo-ku, Tokyo, Japan*
{s1F102001524, tomokazu.ishikawa}@iniad.org

Keywords: Rotoscope, Animation Techniques, Segmentation.

Abstract: This paper proposes a novel method for automating the rotoscoping process in anime production by combining SAM (Segment Anything Model) and k-means clustering. Traditional rotoscoping, which involves manually tracing live-action footage, is time-consuming and labor-intensive. Our method automatically generates line drawings and coloring regions suitable for anime production workflows through three main steps: line drawing creation using SAM2, shadow region generation using k-means clustering, and finishing with color design. Experimental results from 134 participants showed that our method achieved significantly higher ratings in both “rotoscope-likeness” and “anime-likeness” compared to existing methods, particularly in depicting complex human movements and details. The method also enables hierarchical editing of animation materials and efficient color application across multiple frames, making it more suitable for commercial anime production pipelines than existing style transfer approaches. While the current implementation has limitations regarding segmentation accuracy and line drawing detail, it represents a significant step toward automating and streamlining the anime production process.


1 INTRODUCTION

The anime production industry has been expanding its market scale in recent years and showing significant vitality. As viewing methods shift from television to internet streaming, Japanese anime maintains high popularity all over the world, increasing its importance as a content industry. Since the online distribution of animation has become mainstream, the quality of the work contributes to the number of views and is directly related to its reputation, so productions are spending more money on the drawing process so that we can see well-drawn animation every week. While approaches to improving animation quality vary depending on the work and direction, rotoscoping exists as one such technique.

Rotoscope was developed by Max Fleischer in 1915 as an aid to overcome awkward movement in animation caused by animators’ insufficient skills, involving tracing live-action footage. It continues to be used even in the past 20 years as animators’ skills have matured. Cultural factors include the proliferation of digital devices with cameras like smartphones

and easy reference material access via the internet. The universalization of the production technique of referencing live-action footage has had a major impact on the use of rotoscopes. Rotoscoping adoption is divided between animation assistance purposes and expressive technique purposes. The former is used to animate sophisticated real movements like walking, dancing, or musical instrument performance, while the latter is used when need to express the grotesqueness that emerges as a side effect of capturing reality without any stylization. It is also sometimes adopted as pre-visualization to visually express the final form in early production stages by determining layout and acting before animation, helping directors achieve their intended screen vision.

A problem when using rotoscoping is that manually tracing live-action footage is extremely time-consuming and labor-intensive. Since what needs to be drawn is predetermined, creativity is limited mostly to deciding what lines to take from the image subjects. For animators, this becomes repetitive manual labor, leading to decreased motivation. Therefore, this research aims to contribute to reducing animators’ burden by improving animation production efficiency through automating the rotoscoping work-

 <https://orcid.org/0000-0002-9176-1336>

flow. Specifically, we propose a method to automatically generate rotoscope animation by creating line drawings from live-action footage and separating materials while identifying line art expressions and coloring regions to facilitate incorporation into production. In the proposed system, rotoscope animation is generated through the following steps:

1. Line drawing creation using SAM
2. Creating shadow regions by reducing colors in basic coloring areas
3. Finishing (coloring) and compositing

The reason for identifying coloring regions is that generating only line drawings would require coloring work for each frame. By identifying coloring regions beforehand, colors can be applied to the entire animation at once. In anime production, cels are colored one by one in the finishing process while referring to color design, which specifies coloring instructions. In this research, if color design is available, the finishing process can be completed immediately. Afterwards, in the shooting process that composites multiple materials to create the final image or video, cels and backgrounds are composited and the screen is adjusted.

It is important to emphasize here that the proposed technology is not a mere style conversion method like Diffutoon (Duan et al., 2024) or DomoAI (DOMOAI PTE. LTD, 2024), but can also output intermediate data such as line drawings in accordance with the animation production process. The existence of intermediate data allows retakes and re-editing, thus replacing part of the existing animation production pipeline.

2 RELATED WORKS

We propose a novel automated rotoscoping method that automatically generates line drawings and coloring regions suitable for anime production. In this section, we classify related existing research from the following three perspectives and analyze their advantages and disadvantages. Then, we clarify the positioning and novelty of this research.

2.1 Conventional Rotoscoping Methods

Conventional rotoscoping has primarily been performed through manual line tracing. While line drawing extraction using edge detection like Canny method (Canny, 1986) has been studied, it faces challenges in generating closed regions necessary for coloring and cannot reproduce anime-specific line art expressions. Agarwala et al. proposed efficiency

improvements through keyframe interpolation (Agarwala et al., 2004), but the manual workload remains substantial. Adobe After Effects' Roto Brush tool specializes in silhouette extraction (Dissanayake et al., 2021; Torrejon et al., 2020) but is not suited for hierarchical generation of line drawings and coloring regions needed in anime production.

2.2 Image Anime-Stylization Using Deep Learning

GAN-based methods like CartoonGAN (Chen et al., 2018) and AnimeGAN (Chen et al., 2020), and Stable Diffusion-based methods (Rombach et al., 2022; Esser et al., 2024) can generate high-quality anime-style images. Nevertheless, these methods are not suitable for animation production as they cannot consider temporal coherence and shape consistency. Among these methods, the latest stylization techniques are Diffutoon (Duan et al., 2024) and DomoAI (DOMOAI PTE. LTD, 2024). These maintain general temporal consistency and demonstrate high quality as video generation AI. However, they cannot separately output line drawings, coloring regions, and shooting process effects, making integration into commercial anime production workflows difficult.

2.3 Segmentation Technology and Its Application to Anime

As emphasized by animator Tatsuyuki Tanaka, anime expression consists of "symbolic expressions of simple lines and color separation" (Tanaka, 2021), unlike realistic paintings. To capture these symbolic expressions, segmentation at the semantic level becomes crucial. In recent years, deep learning-based segmentation technology has rapidly advanced, enabling high-precision segmentation. The Segment Anything Model (SAM) (Kirillov et al., 2023) is a prime example, being a versatile model capable of accurately segmenting various objects at the pixel level.

In Tous's research (Tous, 2024), they use SAM to segment various visual features of characters (hair, skin, clothes, etc.) and combine it with a method called DeAOT to automatically generate retro-style rotoscope animations. However, it specializes in styles composed of limited colors and expressions like retro games, and since it does not consider general anime production or line drawing generation, it is not suitable for delicate expressions like Japanese anime created in a line-expression culture.

As shown in Table 1, existing methods do not adequately consider the hierarchical generation of line drawings and coloring regions necessary for anime

Table 1: Comparison with previous research.

Method	Input	Output	Technique	Symbolic Expression	Motion Expression	Hierarchical Editing
Agarwala et al.	Keyframes	Animation	Interpolation	High	High	Not Possible
CartoonGAN	Image	Anime-style image	GAN	High	Low	Not Possible
AnimeGAN	Image	Anime-style image	GAN	High	Low	Not Possible
Stable Diffusion	Text + image	Image	Diffusion Model	High	Low	Not Possible
Dissanayake et al.	Live-action video	Silhouette	Deep Learning	Low	Low	Not Possible
Tous et al.	Live-action video	Retro anime	SAM, DeAOT	Medium	High	Possible
Proposed Method	Live-action video	Line drawing + Coloring regions	SAM2, k-means	High	High	Possible

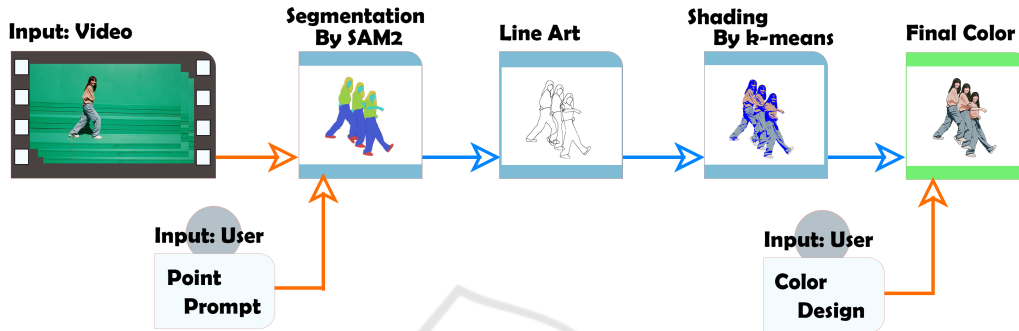


Figure 1: Processing flow of the proposed method.

production, nor their integration with color design. In this research, we propose a novel method that combines SAM2 (Ravi et al., 2024) and k-means clustering to address these challenges. Leveraging SAM2’s high segmentation accuracy and interactive versatility, we accurately extract user-specified objects and generate line drawings and coloring regions. Furthermore, by using k-means clustering to reduce colors in coloring regions, we automatically separate basic color regions and shadow regions, facilitating color design. This significantly streamlines the rotoscoping process in anime production and supports high-quality animation production.

Our research aims to hierarchically generate line drawings and coloring regions that can be integrated into Japanese anime production workflows, achieving high-precision segmentation and efficient color design using SAM2 and k-means clustering.

3 PROPOSED METHOD

We propose the method to convert live-action video into drawable materials with symbolic expressions of simple lines and color separation, the processing flow of the proposed method is shown in Figure 1. The proposed method consists of a line drawing process using SAM2, shading using the k-means method, and finishing. Details of each process are described below.

3.1 Line Drawing Creation Using SAM2

We use Segment Anything Model 2 (SAM2) (Ravi et al., 2024) to generate line drawings and coloring regions. SAM2 is a segmentation model that can handle both images and videos, enabling high-precision and fast processing. It particularly excels at identifying spatial and temporal ranges of objects in videos, capable of high-precision segmentation even for fast-moving objects or partially occluded objects necessary for rotoscoping. Additionally, as it is designed for interactive interfaces, users can easily specify target objects with simple operations and obtain desired segmentation results. This interactive semantic segmentation capability combined with high-precision segmentation ability is the main reason we adopted SAM2 for this research. Specifically, we first extract frame images from the input video and use SAM2 to segment target objects (target silhouettes, hair, clothes, accessories, etc.). In this process, users instruct SAM2 on segmentation targets using prompts. In this case, we executed SAM2’s image segmentation using user clicks on frame images as input and used the obtained segmentation areas as prompts for video segmentation. For wide shots, we segment main parts such as hair, skin, clothes, and shoes, while for close-ups, we additionally segment details like eyes, mouth, and accessories. SAM2 generates segmentation areas for each object based on the specified prompts. The contours obtained by apply-

Table 2: Properties of videos used in the experiment.

Video ID	Contents	Resolution (W × H)	fps	Time
1	A woman dancing in a long shot	1920×1080	25	11s
2	A preening penguin in a long shot	1080×1920	25	10s

ing the Canny method to these segmentation areas are adopted as line drawings and used as the basis for generating coloring areas in subsequent processing.

3.2 Shading by Color Reduction in Basic Coloring Areas

In anime, shade is typically expressed as regions colored differently from the basic colors. To reproduce this characteristic, we perform color reduction of coloring regions using k-means clustering to extract basic color regions and shadow regions.

First, using the SAM2 segmentation areas obtained in Sec. 3.1, we extract only the target object regions from the original live-action video. We apply k-means clustering to this extracted image, dividing it into two clusters. This extracts high-brightness areas as basic color regions and low-brightness areas as shadow regions.

3.3 Final Coloring and Compositing

Using the line drawings and coloring regions (basic color regions, shadow regions) generated in Sec. 3.1 and Sec. 3.2, we generate the final animation materials. First, based on the basic color regions and shadow regions obtained in Sec. 3.2, users develop a color design and assign appropriate colors to each region by specifying colors in a palette.

Next, when creating animation, we use the line drawings obtained in Sec. 3.1 and the sequential images of colored basic color regions and shadow regions as materials. By compositing these materials with background images and applying shooting processes (effects, color correction, etc.) as needed, we complete the final animation.

4 EXPERIMENTS AND RESULTS

We describe the experimental conditions of the proposed method and the results and their evaluation. The evaluation of the proposed method was conducted using both quantitative and qualitative assessments. For quantitative evaluation, we measured the processing time of each processing step. For qualitative evaluation, we conducted a subjective evaluation

Table 3: Experimental environment.

OS	Windows 11 Home 64bit
CPU	Intel Core i7-13700KF @3.0GHz
RAM	32GB
GPU	NVIDIA GeForce RTX 4070Ti 12GB
Software	Adobe After Effects

experiment with 134 participants, assessing two criteria: "rotoscope-likeness" and "anime-likeness."

4.1 Experimental Conditions

To verify the effectiveness of our proposed method, we conducted experiments using two types of videos containing diverse movements and subjects. Table 2 describes the content and conditions of the videos used. The experimental environment is as shown in Table 3. We used default values for all parameters of SAM2 and the Canny method. For k-means clustering, we specified the number of clusters as 2.

4.2 Result Images

Line drawings, coloring area, and still images of the final animation generated using the proposed method are shown in Figures 2.

Limitations: Accurate segmentation cannot be performed when the segmentation target is blurred or un-

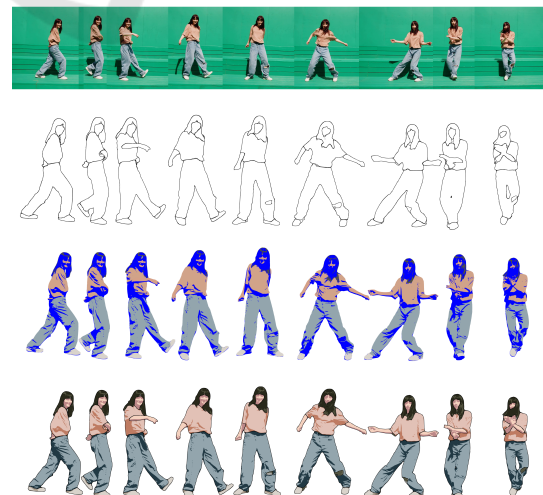


Figure 2: The result of processing on long shot video of a dancing woman. From top to bottom: live-action video, generated line drawing, after shading, and after coloring.

clear. The method needs to be applied to videos where the rotoscoping targets are clearly visible.

4.3 Processing Time

Using the experimental environment shown in Table 3, we measured the processing time for two steps: segmentation using SAM2, and line drawing extraction using the Canny method combined with shadow region generation using k-means clustering. The user prompt input time, while not included in these measurements, averaged approximately 1 minute per video for initial segmentation setup. This operation time is proportional to the number of segmentation regions. This initial investment significantly reduces manual rotoscoping time, which typically requires 15-20 minutes per frame for traditional methods. The average FPS for the segmentation portion of the process using SAM2 was 3.01 fps, and the average FPS per segmentation area was 1.16 fps. The number of segmented regions for Video 1 was 6, Video 2 was 7. On the other hand, the processing using the Canny method and k-means clustering for line drawing and shadow generation achieved an average FPS of 3.69 fps, with an average FPS per segmentation area of 0.63 fps. The results suggested that the segmentation processing using SAM2 is significantly affected by the increase in the number of segmentation areas.

4.4 Quality Evaluation

To evaluate the quality of rotoscope animation produced by our proposed method, we conducted a subjective evaluation experiment with 134 participants, both male and female university students in their 20s. Participants were shown videos processed using three different methods of converting live-action footage to anime - our proposed method (Ours), the cartoon effect in Aftereffects (Cartoon), and k-means clustering + Canny method (k-means) - for each of 2 different video sections (Video 1 and 2). Since the k-means clustering + Canny method is a classic method of transforming cartoon-like images, it was used as the comparison target in this experiment. In this method, let k be 6. They were asked to evaluate two criteria, "rotoscope-likeness" and "anime-likeness," on a 5-point scale (1: Strongly disagree to 5: Strongly agree).

To help them judge the rotoscope-like nature of the project, we explained what rotoscope was beforehand and had them watch line drawings and animations created using rotoscope. The order in which the methods were presented was randomized in each video section. Figures 3 and 4 show the input video

and the results of image processing by each method.

The mean evaluation values for each method in each video are shown in Figures 5 and 6. For each video and evaluation criterion, we used Friedman's test to determine whether there were significant differences among the methods. Subsequently, we conducted Wilcoxon signed-rank tests with Bonferroni correction as post-hoc tests to perform pairwise comparisons between methods. Figures 5 and 6 are marked with * for pairs with p-values below the 5% significance level, ** for pairs with p-values below the 1% significance level, and *** for pairs below the 0.1% significance level, respectively. Additionally, we calculated Cliff's delta as an effect size measure and evaluated its magnitude in four levels: negligible, small, medium, and large.

These results showed that our proposed method received significantly higher evaluations in both "rotoscope-likeness" and "anime-likeness" compared to the other two methods in most cases. The superiority of our proposed method was particularly notable in Video 1 (woman dancing), where the effect size is large. The reason why Video 2's video evaluation was comparable to the cartoon filter might be because the penguin's inherently limited color palette already made it work as a symbolic expression of simple lines and color separation. These results suggest that our proposed method can effectively express both rotoscope-likeness and anime-likeness when generating animation materials from live-action footage, particularly in depicting human movements and details.

5 CONCLUSIONS AND FUTURE WORK

In this research, we proposed a novel method that automates the rotoscoping process in anime production by combining SAM2 and k-means clustering. The experimental results suggested that our proposed method could achieve superior results in both rotoscope-likeness and anime-likeness compared to conventional methods. This effect was particularly notable in depicting complex human movements and details.

Future challenges include improving segmentation accuracy, enhancing line drawing expression, and accommodating various anime styles. Improving line drawing expression involves enhancing line drawing details. In this study, we gave up on adding appropriate lines for finger shapes and clothing wrinkles. This was because it was difficult to add symbolic lines as intended by humans, and because k-means clustering-based shadowing increased the information content of

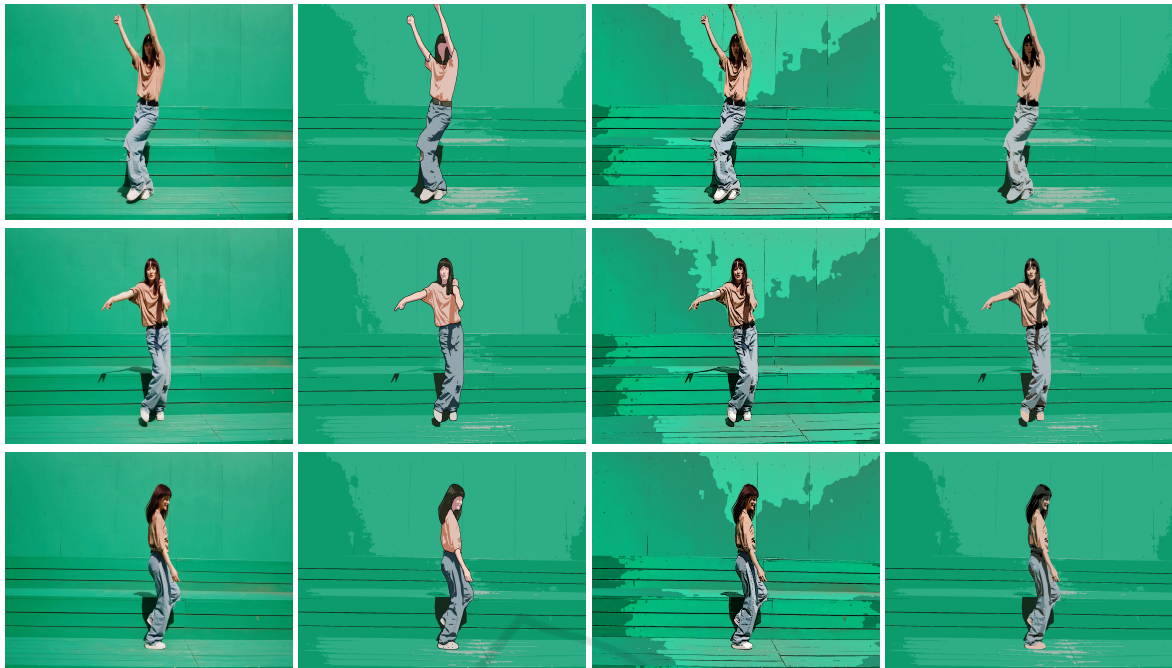


Figure 3: Comparison of the input video (Video 1) and the results after each image processing. From left: input video, ours, comic, k-means.

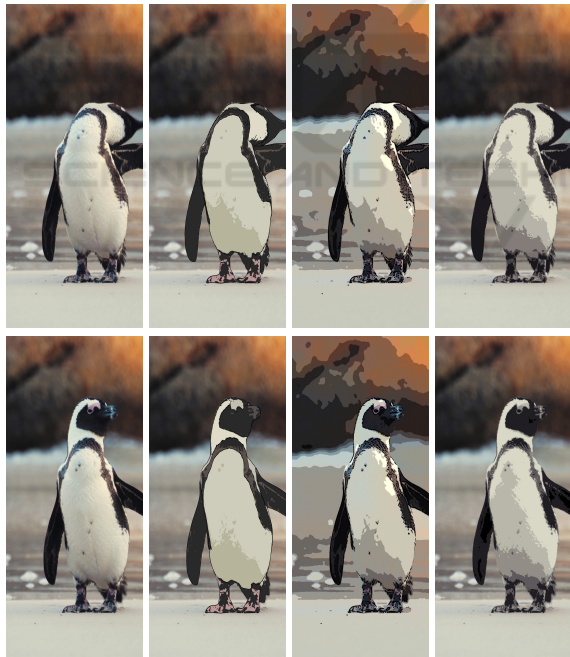


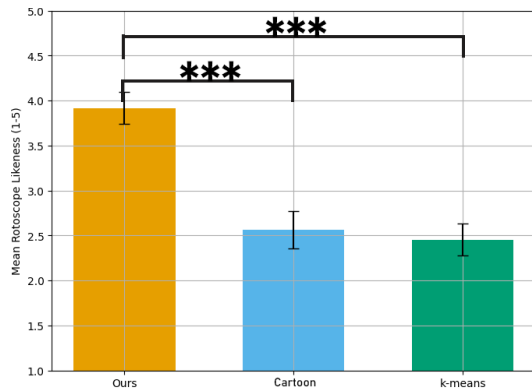
Figure 4: Comparison of the input video (Video 2) and the results after each image processing. From left: input video, ours, comic, k-means.

cells, making it better not to add lines when considering information control. While k-means clustering-based shadowing is highly rated for its accurate edge-line shadowing and three-dimensional shadow-

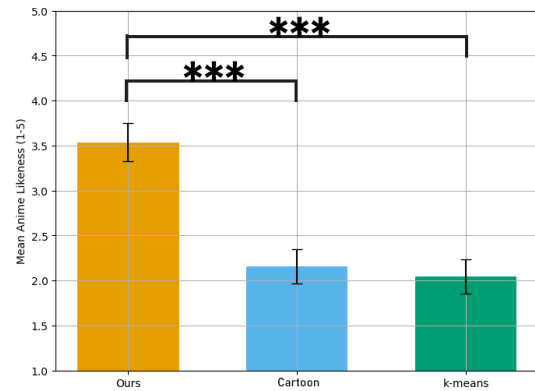
ing necessary for anime drawing, it appears stiff as it fails to appropriately reduce the image information content. To make it more anime-like, we could consider moderately enhancing line drawing details while maintaining appropriate and accurate shadowing.

Additionally, having options to change only faces or clothing in footage with other elements would make it more manageable in production. It would also be beneficial to be able to change the lighting of subjects. We expect image generation AI to provide these two options. We also envision adapting to other styles, such as creating two shadow regions instead of one and adding highlights to the drawings.

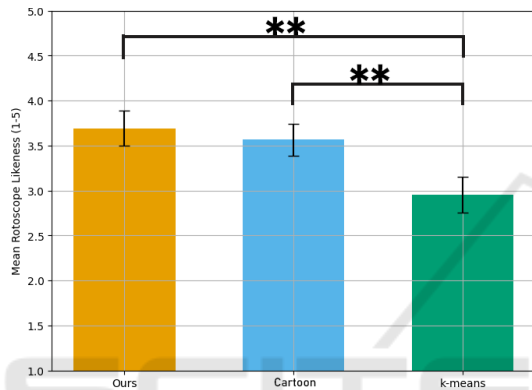
We believe that there should be an option to remove the number of frames in the video appropriately. The advantage of rotoscoping is that it enables anime creation with frame-by-frame shooting that fills all frames within a second. However, if it's possible to select frames capturing lively subjects from the footage, reducing frames might improve quality. Since there are a number of existing studies (Koroku and Fujishiro, 2022; Miura et al., 2014) on frame dropping, We believe that introducing them into the proposed system will help create more animated images. By addressing these challenges from this research, our proposed method is expected to significantly contribute to the efficiency of animation production.



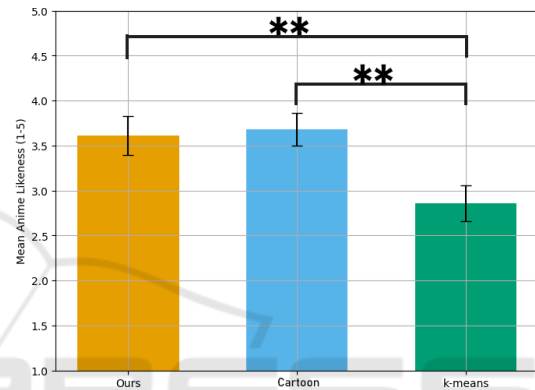
(a) Video 1



(a) Video 1



(b) Video 2



(b) Video 2

Figure 5: Results of subjective evaluation experiment (rotoscope-likeness).

Figure 6: Results of subjective evaluation experiment (anime-likeness).

ACKNOWLEDGEMENTS

This work was supported by Toyo University Top Priority Research Program.

REFERENCES

- Agarwala, A., Hertzmann, A., Salesin, D. H., and Seitz, S. M. (2004). Keyframe-based tracking for rotoscoping and animation. *ACM Trans. Graph.*, 23(3):584–591.
- Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698.
- Chen, J., Liu, G., and Chen, X. (2020). Animegan: A novel lightweight gan for photo animation. In *Artificial Intelligence Algorithms and Applications: 11th International Symposium, ISICA 2019, Guangzhou, China, November 16–17, 2019, Revised Selected Papers 11*, pages 242–256. Springer.
- Chen, Y., Lai, Y.-K., and Liu, Y.-J. (2018). CartoonGAN:

Generative adversarial networks for photo cartoonization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9465–9474.

- Dissanayake, S., Ayoob, M., and Vekneswaran, P. (2021). Autoroto: Automated rotoscoping with refined deep masks. In *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, pages 1–6.
- DOMOAI PTE. LTD (2024). DomoAI. <https://domoai.app/>. (Accessed on 11/20/2024).
- Duan, Z., Wang, C., Chen, C., Qian, W., and Huang, J. (2024). Diffutoon: High-resolution editable toon shading via diffusion models. In Larson, K., editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7645–7653. International Joint Conferences on Artificial Intelligence Organization.
- Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. (2024). Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C.,

- Lo, W.-Y., Dollár, P., and Girshick, R. (2023). Segment anything. *arXiv:2304.02643*.
- Koroku, Y. and Fujishiro, I. (2022). Anime-like motion transfer with optimal viewpoints. In *SIGGRAPH Asia 2022 Posters, SA '22*, New York, NY, USA. Association for Computing Machinery.
- Miura, T., Kaiga, T., Katsura, H., Tajima, K., Shibata, T., and Tamamoto, H. (2014). Adaptive keypose extraction from motion capture data. *Journal of Information Processing*, 22(1):67–75.
- Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., and Feichtenhofer, C. (2024). Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.
- Tanaka, T. (2021). *Ani Man LLust – Tatsuyuki Tanaka Art Techniques*. SMIRAL Co.,Ltd. p. 4, line 15.
- Torrejon, O. E., Peretti, N., and Figueroa, R. (2020). Rotoscope automation with deep learning. *SMPTE Motion Imaging Journal*, 129(2):16–26.
- Tous, R. (2024). Lester: Rotoscope animation through video object segmentation and tracking. *Algorithms*, 17(8):330.

