VIDEO-DRIVEN SKETCH ANIMATION VIA CYCLIC RECONSTRUCTION MECHANISM

*Zhuo Xie, Haoran Mo, Chengying Gao**

Sun Yat-sen University
{xiezh56,mohaor}@mail2.sysu.edu.cn, mcsgcy@mail.sysu.edu.cn

ABSTRACT

Considering the time-consuming manual workflow in 2D sketch animation production, we present an automatic solution by using videos as reference to animate the static sketch images. This includes motion extraction from the videos and injection into the sketches to produce animated sketch sequences in which appearance properties from the source sketches should be preserved. To reduce blurry artifact caused by complex motions and maintain stroke line continuity, we propose to incorporate inner masks of the sketches as an explicit guidance to indicate inner regions and ensure component integrality. Moreover, to bridge the domain gap between the video frames and the sketches when modelling the motions, we introduce a cyclic reconstruction mechanism to increase compatibility with different domains and improve motion consistency between the sketch animation and the driving video. Extensive results demonstrate the superiority of our method that outperforms existing methods both quantitatively and qualitatively.

***Index Terms*—** 2D Animation, Video-driven Animation, Motion Extraction, Motion Transfer

1. INTRODUCTION

2D animation is popular all over the world and still a mainstream form in commercial animation industry. Current workflow of 2D animation production relies heavily on drawing each keyframe manually, which is laborious and time-consuming, given that an animation clip may contain hundreds or thousands of keyframes [1]. Therefore, automatic techniques for assisting with the animation production are in great and urgent demand [2].

2D animation tends to start with a static character sketch manifesting the content, followed by injecting motions to enable dynamics. Considering that abundant videos from the Internet store a great diversity of motion information, it is intuitive to use videos as a kind of driven reference in an automatic 2D animation generation workflow. To this end, we propose a video-driven 2D sketch animation framework that extracts motions from the videos and transfers them to static sketches to produce animated sequences. The framework is applicable to both edge maps with fine-grained details and real sketches with sparse lines, as shown in Fig. 1. The generated animation frames exhibit appearance aligned with the given sketch, and motion consistent with the driven video.

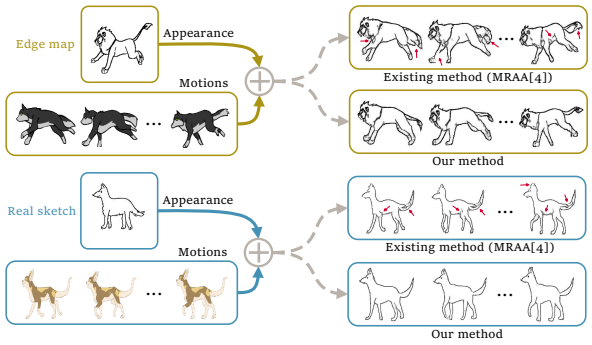

Fig. 1. Our approach extracts motions from videos and injects them into sketch images to produce 2D sketch animations with less blurry artifact and higher motion consistency.

There exist several works on image animation [3, 4, 5, 6], although they focus more on images with colors and textures. The two kinds of information help to identify inner regions of the objects, and thus maintain integrality of the components in the animated results even when complex motions (*e.g.*, occlusions) exist in the driving video. However, these approaches fail to work on sketch images with apparently fewer colors and textures, and generate blurry artifact that breaks apart the continuous strokes or the object components, as can be seen in Fig. 1. To overcome this issue, we propose to extract *inner masks* from the sketches, and inject them into the animation framework to explicitly indicate the inner regions of the line drawings. Such an explicit guidance largely reduces the blurry artifact and the broken strokes/components in the presence of complex motions.

Another challenge in our framework is how to bridge the domain gap between the video and the sketch image as they have significantly different visual characteristics. Existing methods [3, 4, 5, 6] designed for single-domain cases are shown to struggle with motion transfer between two different domains. We thus integrate the idea of cyclic processing com-

*The corresponding author is Chengying Gao. This work was supported by the Guangdong Basic and Applied Basic Research Foundation (Grant No.2022A1515011425).

monly used in the cross-domain scenario into our framework, and introduce a *cyclic reconstruction mechanism*. Formally, we first transfer the video motion to the sketch to obtain the resulting animation, from which we extract the motion again for reconstructing the input video by propagating that motion to one of the video frames. The derived reconstruction loss enforces that the motion from video to sketch animation should be equivalent to that from sketch animation to video, such that the motion extraction and injection processes in our framework learn to be compatible to images from different domains. As a result, the motion consistency between the generated sketch animation and the video can be preserved.

We compare our approach to state-of-the-art image animation algorithms through quantitative evaluations, qualitative comparisons and a user study. The results corroborate the superior effectiveness of our method in terms of preserving appearance properties, reducing blurry artifact and maintaining motion consistency. We also show that our framework, while trained on visually detailed edge maps, generalizes well to real sketches with fewer details and sparse strokes.

Our contributions can be summarized as follow:

- A video-driven 2D sketch animation framework that preserves original appearance properties of the sketches and reduces blurry artifact induced by occlusions in the video.
- A cyclic reconstruction mechanism for extraction and injection of cross-domain motions, which improves motion consistency between the generated sketch sequence and the input video.
- Comprehensive experiments comparing with the state-of-the-art methods demonstrate the superior performance of our proposed method.

2. RELATED WORK

2.1. Image Animation

The image animation approaches can be categorized into two lines: supervised and self-supervised methods. The supervised ones mainly focus on human body pose transfer [7, 8] and facial motion reenactment [9, 10]. They model the geometric structure through object-specific landmark detectors, which are usually pretrained on a large amount of labeled data. It is costly to obtain such a dataset and the pretrained model, and these approaches are limited to specific object types such as human body and face.

Self-supervised methods [3, 4, 5, 6] have been proposed to address the above issues. They typically leverage a large amount of unlabeled videos collected from the Internet and design reconstruction losses to model self-supervised motion representations (*e.g.,* keypoints [3, 4, 6] and local regions [5]). After extracting motion representations from the driving videos, they are used to estimate a dense motion flow through a motion model, such as first-order Taylor expansions [4] or Thin Plate Spline (TPS) transformations [6]. The motion flow is injected into the static source image to produce a dynamic sequence.

The methods above trained on data within a single domain tend to suffer from performance degradation when the source image and the driving video come from different domains, such as sketch images and videos with color frames in our task. We thus propose an animation framework designed for the cross-domain data, which produces high-quality 2D animation that preserves appearance properties of the source domain (*i.e.,* the sketches) and conforms to the motion in the target domain (*i.e.,* the videos).

2.2. Cycle Consistency in Cross-domain Tasks

The concept of cycle consistency is widely used in cross-domain tasks. CycleGAN [11] trains an unpaired image translation model between two domains. Due to the lack of ground-truth data, a cycle consistency loss that reverts a translated image back to its original domain is able to identify the key properties of each domain and improve the translation quality. Inspired by CycleGAN, Jeon et al. [12] propose a cross-identity training scheme that enables realistic motion transfer between subjects with obviously different appearances. Similarly, MAA [13] proposes a cyclic training pipeline for boosting the performance of cross-domain motion transfer. We also integrate the concept of cycle consistency into our framework, and introduce a cyclic reconstruction mechanism that bridge the gap between monochromatic sketches and color frames in the video. The mechanism noticeably enhances motion consistency in the resulting sketch animation.

3. METHOD

3.1. Sketch Animation Framework

Our framework extracts motion information from an input video and injects it into a static sketch image, and then generates a dynamic sketch sequence to form a 2D sketch animation. The approach processes the video frame-by-frame, as shown in Fig. 2. We propose a cyclic reconstruction mechanism for our training pipeline, which consists of a video frame-driven stage and a sketch frame-driven stage. It helps to enhance the motion consistency between the generated sketch sequence and the input video. During inference, we use the first stage only for sketch animation production.

In the video frame-driven stage, a random frame from the video with posture information $m1$ is used as a driving frame D_{m1}^{video}. Another input is a static sketch image I_{m*}^{sketch} with an arbitrary posture $m*$. The model in this stage generates a sketch frame $\hat{D}_{m1}^{sketch}$ by extracting and transferring the posture $m1$ to the sketch I_{m*}^{sketch} while preserving its appearance properties.

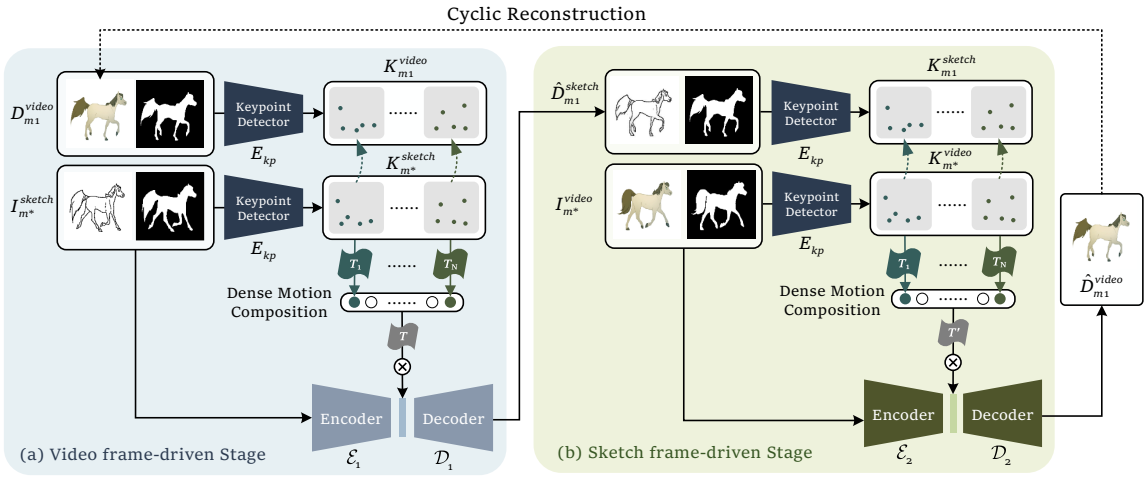

Fig. 2. The architecture of our video-driven 2D sketch animation framework that extracts posture information from each video frame and injects it into the sketch image to produce a resulting sketch frame with similar posture (a). We propose a cyclic reconstruction mechanism which makes the framework extract such a posture from the synthetic sketch and transfer it back to the video frame (b), in order to improve the motion consistency.

The main idea behind the sketch frame-driven stage is to extract the posture $m1$ from the synthetic sketch frame $\hat{D}_{m1}^{sketch}$ and transfer it back to the video frame so as to promote the cross-domain motion consistency. Thus, $\hat{D}_{m1}^{sketch}$ is used as a driving frame in this stage, and a random video frame I_{m*}^{video} with an arbitrary posture $m*$ is used as a source image. After the motion transfer, the synthetic video frame $\hat{D}_{m1}^{video}$ with posture $m1$ and the same identity as those in the video should be equivalent to D_{m1}^{video}, *i.e.*, $\hat{D}_{m1}^{video} \equiv D_{m1}^{video}$. This forms the cyclic reconstruction with a derived loss to ensure the cycle consistency of postures between two domains.

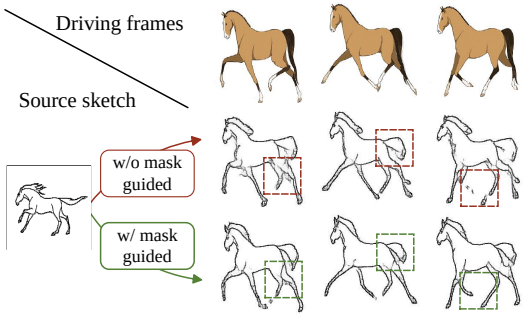

Fig. 3. Comparisons of sketches synthesized with or without cross-domain guidance from an inner mask.

Inner masks. With the pipeline above, the video-driven sketch animation generation is still challenging, due to the lack of color and texture information in sketches to distinguish the inner regions of the objects from background. This issue increases the difficulty of the task, especially in animations with dense motion and occlusions. Without the guidance of inner regions, continuous lines of the sketches tend to break apart or intersect with other lines, leading to blurry artifact as shown in Fig. 3.

To overcome the issue, we propose to incorporate inner masks into our framework, which play the role of color or texture for monochromatic sketches by indicating the inner regions of the foreground object. The masks are produced using a pretrained saliency detector U2-Net [14]. After extracting the masks, we concatenate them with the corresponding video frame/image or sketch image/frame, serving as joint inputs to the networks, as illustrated in Fig. 2.

The region masks help the models in identifying the unity of the lines and reducing their breakage and incorrect intersections, which results in less blurry artifact as demonstrated in Fig. 3.

3.2. Cyclic Reconstruction Mechanism

3.2.1. Video frame-driven stage

The model in this stage first extracts posture $m1$ from a video driving frame D_{m1}^{video}, and then transfers it to a source image I_{m*}^{sketch} (with an arbitrary posture $m*$) to produce a sketch frame $\hat{D}_{m1}^{sketch}$ with the posture $m1$ while maintaining its original appearance characteristics, as shown in Fig. 2-(a). The inner masks are attached to the driving frame and the source image as joint inputs, and we omit the notations of the inner masks for brevity in the following sections.

Motion Extraction. We represent motion with Thin Plate Spline (TPS) transformation [15] of keypoints for the objects. With the two inputs D_{m1}^{video} and I_{m*}^{sketch}, we utilize a ResNet-based Keypoint Detector E_{kp} to estimate a set of unsupervised keypoints $K_{m1}^{video} = E_{kp}(D_{m1}^{video})$ and $K_{m*}^{sketch} = E_{kp}(I_{m*}^{sketch})$. They describe the structural fea-

tures of the objects in the input images. Then, both the keypoints $K_{m1}^{video}, K_{m*}^{sketch}$ are equally divided into N groups (M keypoints in each group) to represent local motion. The keypoints for the driving video frame and the sketch are one-to-one corresponded through indexing, and thus we warp I_{m*}^{sketch} to D_{m1}^{video} with minimum distortion by using N TPS transformations [15] $\mathcal{T}_i (i = 1, 2, ..., N)$:

$$\min \iint_{\mathbf{R}^2} \left(\frac{\partial^2 \mathcal{T}_i}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 \mathcal{T}_i}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 \mathcal{T}_i}{\partial y^2}\right)^2 \mathrm{d}x\mathrm{d}y,$$
$$\text{s.t. } \mathcal{T}_i(K_{m*}^{sketch}(j)) = K_{m1}^{video}(j), j = 1, 2, ..., M. \quad (1)$$

The N warped sketches are combined and input to a Dense Motion Composition Module to estimate N contribution maps $\mathbf{C}_n \in \mathcal{R}^{H \times W} (n = 1, ..., N)$, where H and W are the height and width of the source sketch image. Then, these contribution maps are activated by a softmax operation to make them sum to 1 for each pixel position p:

$$\overline{\mathbf{C}}_n(p) = \frac{\exp\left(\mathbf{C}_n(p)\right)}{\sum_{k=1}^{N} \exp\left(\mathbf{C}_k(p)\right)}, n = 1, 2, ..., N. \quad (2)$$

The normalized contribution maps reflect the influence weight of the TPS transformations at each pixel position. Thus, we use them to combine the N TPS transformations to compute a dense motion flow $\mathcal{T} = \sum_{n=1}^{N} \overline{\mathbf{C}}_n \mathcal{T}_n$, which stores the global motion from I_{m*}^{sketch} to D_{m1}^{video}.

Motion Transfer. With the extracted motion information $\mathcal{T}$, we fuse it with the source sketch image I_{m*}^{sketch} in a skip-connected hourglass network that contains an encoder $\mathcal{E}_1$ and a decoder $\mathcal{D}_1$. After encoding the sketch image into feature maps $\mathcal{E}_1(I_{m*}^{sketch}) \in \mathcal{R}^{h \times w}$, we rescale the dense motion flow $\mathcal{T}$ to size $h \times w$, and then apply a pixel-wise multiplication for them to obtain fused feature maps, which are fed to the decoder for a synthesized sketch frame $\hat{D}_{m1}^{sketch} = \mathcal{D}_1(\mathcal{E}_1(I_{m*}^{sketch}), \mathcal{T})$. The output frame is expected to exhibit posture from D_{m1}^{video} and appearance of I_{m*}^{sketch}.

3.2.2. Sketch frame-driven stage

This stage plays a fundamental role in our cyclic reconstruction mechanism, which shares the same pipeline as the video frame-driven one, except for the inputs, as shown in Fig. 2-(b). As a mirrored stage, we use the synthetic sketch image $\hat{D}_{m1}^{sketch}$ as the driving frame, and a random video frame I_{m*}^{video} with an arbitrary posture $m*$ as the source image. The Keypoint Detectors E_{kp} in this stage share network weights from the video frame-driven stage. Therefore, by using them to extract the posture information from $\hat{D}_{m1}^{sketch}$ and inject it back to the video frame in a cyclic manner, we are able to encourage them to accommodate images from different domains and capture the motion regardless of appearance variations. This helps to improve motion consistency when processing cross-domain images.

Formally, the Keypoint Detectors estimate unsupervised keypoints $K_{m_1}^{sketch} = E_{kp}(\hat{D}_{m1}^{sketch})$ and $K_{m*}^{video} = E_{kp}(I_{m*}^{video})$. Then, a dense motion flow $\mathcal{T}'$ storing motion from I_{m*}^{video} to $\hat{D}_{m1}^{sketch}$ is produced and injected into a generative network similarly. We use a Dense Motion Composition Module and an encoder-decoder network with the same architectures as those in the video frame-driven stage, but they are trained separately. Finally, we have the output video frame $\hat{D}_{m1}^{video} = \mathcal{D}_2(\mathcal{E}_2(I_{m*}^{video}), \mathcal{T}')$, which is expected to exhibit posture $m1$ and the appearance properties of the video frame. That is, $\hat{D}_{m1}^{video}$ should be equivalent to the original driving frame D_{m1}^{video} in the video frame-driven stage. This enables us to enforce a cyclic reconstruction constrain between them, which boosts the performance of our framework in motion extraction and transfer.

3.2.3. Losses for Cyclic Reconstruction

In the video frame-driven stage, we produce a sketch frame $\hat{D}_{m1}^{sketch}$ by injecting posture $m1$ into the sketch. During training, we can obtain its ground truth by extracting the edge map of the input D_{m1}^{video} as they share the same identity and posture $m1$. We denote the ground truth as $\overline{D}_{m1}^{sketch}$. A perceptual loss is adopted to measure their difference:

$$\mathcal{L}_{\text{perc}} = \frac{1}{L} \sum_{l=1}^{L} \left\| F_l(\hat{D}_{m1}^{sketch}) - F_l(\overline{D}_{m1}^{sketch}) \right\|_1, \quad (3)$$

where $F_l(\cdot)$ is feature map computed by the l^{th} layer of a pre-trained VGG-19 network [16]. $\overline{D}_{m1}^{sketch}$ can be treated as a warped version of I_{m*}^{sketch} with posture $m1$, so we following TPSMM [6] and enforce an auxiliary constraint for the encoder based on the warped encoder features of I_{m*}^{sketch} and the encoder features of the warped sketch $\overline{D}_{m1}^{sketch}$:

$$\mathcal{L}_{\text{warp1}} = \sum_i \left\| \mathcal{T}\left(\mathcal{E}_1^{(i)}(I_{m*}^{sketch})\right) - \mathcal{E}_1^{(i)}(\overline{D}_{m1}^{sketch}) \right\|_1, \quad (4)$$

where $\mathcal{E}_1^i$ is the i^{th} layer of the encoder $\mathcal{E}_1$. This loss improves the ability of the encoder in fusing motion into the source image.

The subsequent sketch frame-driven stage is designed for the cyclic reconstruction mechanism, which encourages the cycle consistency between the synthesized video frame $\hat{D}_{m1}^{video}$ and the input video frame D_{m1}^{video} in the first stage. Thus, we use the perceptual loss again as the cyclic reconstruction loss:

$$\mathcal{L}_{\text{recon}} = \frac{1}{L} \sum_{l=1}^{L} \left\| F_l(D_{m1}^{video}) - F_l(\hat{D}_{m1}^{video}) \right\|_1. \quad (5)$$

The auxiliary constraint for the encoder is also adopted via the dense flow $\mathcal{T}'$:

$$\mathcal{L}_{\text{warp2}} = \sum_i \left\| \mathcal{T}'\left(\mathcal{E}_2^{(i)}(I_{m*}^{video})\right) - \mathcal{E}_2^{(i)}(D_{m1}^{video}) \right\|_1, \quad (6)$$

where $\mathcal{E}_2^i$ is the i^{th} layer of the encoder $\mathcal{E}_2$.

Table 1. Quantitative comparisons on MGif dataset.

Method	$L_1(\downarrow)$	LPIPS($\downarrow$)	SWD($\downarrow$)
FOMM [4]	6.86	11.84	6.91
MRAA [5]	6.75	11.20	6.61
TPSMM [6]	6.42	10.75	6.12
Ours (w/o cyclic)	6.34	10.72	6.09
Ours (w/o mask)	6.30	10.61	6.10
Ours (full)	**6.27**	**10.50**	**5.91**

3.3. Training Objective

Besides the losses above, we also define constraints for the Keypoint Detector E_{kp} following previous works [4, 5, 6], in order to improve its robustness. We first create a nonlinear transformation $\mathcal{T}_r$ with random parameters. Assuming that keypoints of a warped frame by $\mathcal{T}_r$ are equivalent to warped keypoints of that frame, we define the loss in both stages:

$$\begin{aligned}\mathcal{L}_{\text{eq1}} &= \left\| E_{kp}(\mathcal{T}_r(D_{m1}^{video})) - \mathcal{T}_r(E_{kp}(D_{m1}^{video})) \right\|_1, \\ \mathcal{L}_{\text{eq2}} &= \left\| E_{kp}(\mathcal{T}_r(I_{m*}^{video})) - \mathcal{T}_r(E_{kp}(I_{m*}^{video})) \right\|_1.\end{aligned} \tag{7}$$

The full objective function of our end-to-end training pipeline is:

$$\mathcal{L} = \mathcal{L}_{\text{perc}} + \mathcal{L}_{\text{warp1}} + \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{warp2}} + \mathcal{L}_{\text{eq1}} + \mathcal{L}_{\text{eq2}}. \tag{8}$$

3.4. Testing Phase

During testing, we use only the video frame-driven stage in Fig. 2-(a) for sketch animation generation. Given a source sketch I_{m*}^{sketch} and a driving video $\{D_{m_t}^{video}\}(t = 1, 2, ..., T)$ with T frames, we generate a sequence of $\{\hat{D}_{m_t}^{sketch}\}$ to form the dynamic sketch animation.

4. EXPERIMENTS

4.1. Comparisons with Baseline Methods

We compare our approach with three advanced methods, namely FOMM [4], MRAA [5] and TPSMM [6] on MGif dataset [3] with animated frames of cartoon characters.

Quantitative Evaluation. We perform this evaluation via a cross-domain video reconstruction task, where we animate a source sketch with a series of video frames with the same identity. We follow the existing methods above and use the following metrics: (1) L_1 *distance* which measures visual similarity between the animated sketch frames and their ground truth derived from the extracted edge maps of the video frames, (2) *LPIPS score* for perceptual and structural similarity, and (3) *Sliced Wasserstein Distance (SWD)* [17] which reflects the difference of distributions of two videos in our context.

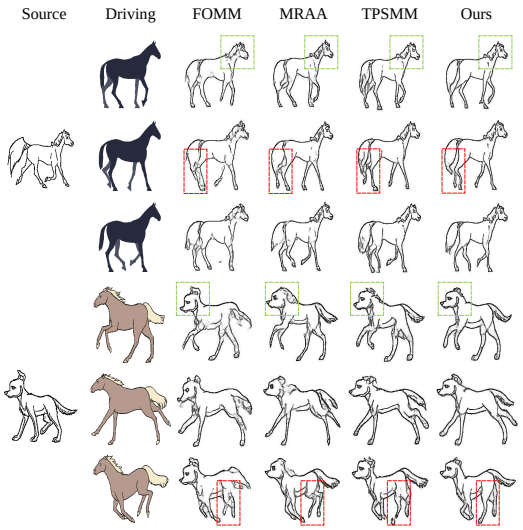

Fig. 4. Comparisons with baseline methods. Please refer to the supplemental video for more results on real sketches and animated sequences.

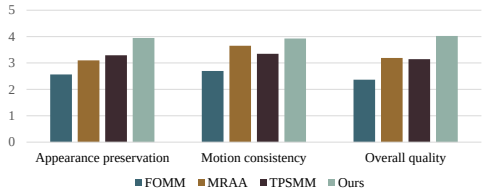

Fig. 5. User study of video-driven sketch animation.

As shown in Table 1, FOMM and MRAA perform poorly on all metrics, probably because their rigid motion representation is not suitable for sketch data. TPSMM works better than them, but is still inferior to ours due to the lack of considering cross-domain adaption. In contrast, our framework exhibits the best performance, indicating that the proposed inner mask injection and cyclic reconstruction mechanism help to ensure appearance preservation and cross-domain motion consistency.

Qualitative Evaluation. We demonstrate the results of sketch animation on both edge maps and real sketches in Fig. 1 and Fig. 4, where we select three representative frames in the driving video. FOMM [4] and MRAA [5] work poorly in preserving the appearance properties (*e.g.,* the head of the horse and the dog). What's worse, their results exhibit undesired blurry artifact that is detrimental to the visual quality. While TPSMM [6] is able to maintain the identities of the objects, it generates blurry artifact and incontinuous stroke lines likewise. Motion inconsistency also occurs in some cases (*e.g.,* the second row of the horse). In sharp contrast, our approach shows superior performance in preserving both the appearance properties and motion consistency in presence

of large motion, and alleviating the blurry issue. Although trained solely with edge maps, our framework generalizes well to real sketches with sparse lines.

User Study. We invite 40 participants for a user study to assess the performance of our method and the baselines. We randomly choose 15 sketches and 15 driven videos from the test set to generate the sketch animations. 3 random frames in each group are selected for the study. Each participant is assigned 5 groups of results, and required to score 1 to 5 for each method according to three aspects, namely appearance preservation, motion consistency, and overall quality. As can be seen in Fig. 5, the averaged scores suggest that our approach has the best performance in these criteria.

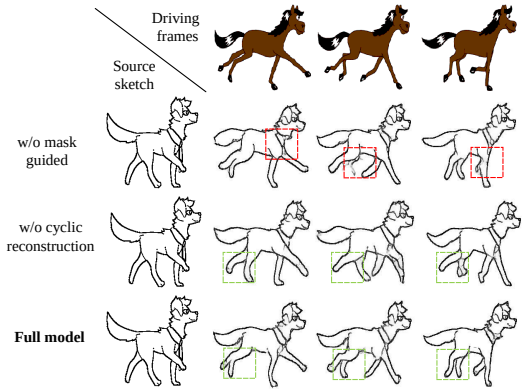

Fig. 6. Qualitative results of ablation study.

4.2. Ablation Study

We evaluate our inner mask injection method and cyclic reconstruction mechanism via ablation studies. For the former, we delete the masks in the pipeline. For the latter, we remove the sketch frame-driven stage in our framework (Fig. 2) to disable the cyclic reconstruction. The quantitative results are shown in Table 1, where the performance degrades without any of them. From visual results in Fig. 6, we see that the model without inner mask guidance ("w/o mask-guided") tends to produce blurry artifact where complicated motion or occlusion occurs, which breaks apart the continuous stroke lines. The model without the cyclic reconstruction mechanism ("w/o cyclic reconstruction") fails to adhere to motions in the driving video frames in most cases (*e.g.,* the legs). Our approach with the two technical components addresses the issues and generates visually appealing animated frames.

5. CONCLUSION

We propose a video-driven 2D sketch animation framework that extracts the motion information from the video and transfers it to the sketch image to create an animated sketch sequence. An inner mask injection strategy and a cyclic reconstruction mechanism are introduced to preserve visual properties of the sketch and ensure motion consistency with the video. While producing visually appealing sketch animations, our approach might still generate blurry strokes in cases with complicated motions. To address this issue, vector-level animation techniques that operate on parameterized strokes could be incorporated into our approach, which is a promising future direction.

6. REFERENCES

[1] Jie Jiang, Hock Soon Seah, Hong Ze Liew, and Quan Chen, "Challenges in designing and implementing a vector-based 2d animation system," in *The Digital Gaming Handbook*. 2020.

[2] Haoran Mo, Chengying Gao, and Ruomei Wang, "Joint stroke tracing and correspondence for 2d animation," *ACM Transactions on Graphics*, vol. 43, no. 3, 2024.

[3] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, et al., "Animating arbitrary objects via deep motion transfer," in *CVPR*, 2019.

[4] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, et al., "First order motion model for image animation," in *NeurIPS*, 2019.

[5] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, et al., "Motion representations for articulated animation," in *CVPR*, 2021.

[6] Jian Zhao and Hui Zhang, "Thin-plate spline motion model for image animation," in *CVPR*, 2022.

[7] Caroline Chan, Shiry Ginosar, Tinghui Zhou, et al., "Everybody dance now," in *ICCV*, 2019.

[8] Wen Liu, Zhixin Piao, et al., "Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis," in *ICCV*, 2019.

[9] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *ICCV*, 2019.

[10] Kuangxiao Gu, Yuqian Zhou, et al., "Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis," in *AAAI*, 2020.

[11] Jun-Yan Zhu, Taesung Park, et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[12] Subin Jeon, Seonghyeon Nam, et al., "Cross-identity motion transfer for arbitrary objects through pose-attentive video reassembling," in *ECCV*, 2020.

[13] Borun Xu, Biao Wang, Jinhong Deng, Jiale Tao, et al., "Motion and appearance adaptation for cross-domain motion transfer," in *ECCV*, 2022.

[14] Xuebin Qin, Zichen Zhang, et al., "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern recognition*, vol. 106, pp. 107404, 2020.

[15] Fred L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 11, no. 6, pp. 567–585, 1989.

[16] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *ICLR*, 2018.