Supplementary

1 Training Data Preparation

Our training and validation data contains real-world images. We collected 4037 images to create 19176 data triplets for training and 1009 for validation. All of the images were downloaded from Youtube with the diversity of image content. The image contents include cartoon, news, sports, etc.

In our design, RSFNets is modeled by three R-Encoder (as outlined in Fig.2 in our major manuscript). These weights are trained using a loss based on our prepared input. In our training, the input is a pair (Γ, κ), which is so-called *SE-input*. Here, Γ is a triplet of images and κ is considered as a consecutive signal. In the preparation of SE-input, a triplet Γ encompasses a pair of consecutive frames (x_i, x_j) and a "reference image" x_r. The reference image is the one that exists in same video clip with x_i and x_j. Supposed that $\mathcal{S}$ is the set of frames in a certain clip, x_r is defined as follows

$$x_r = \{x \in \mathcal{S} \setminus \{x_i, x_j\} s.t. \min_{k=1}^{n-2}(d_p(x, x_i)), \quad (1)$$

where n is the total number of frames in $\mathcal{S}$, and $d_p(.)$ is the PSNR measurement [1]. PSNR is a common measurement of image similarity, where the larger value implies the images are more similar. Consequently, we have a triplet $\Gamma = \{x_i, x_j, x_r\} \in \mathrm{R}^{H \times W}$ in which x_j is a "positive sample" while x_r is a "negative sample" of x_i. We feed the triplet Γ to the RSFNet. And the RSFNet shares the same weights. Let denote the latent vector of Γ as $\Gamma_v = \{R(x_i), R(x_j), R(x_r)\}$ with $R(.)$ represents RSFNet. Finally, they are normalized and fed to distance loss to estimate the latent distance.

2 Selection on Difference of Optical Flow

In this section, we discuss our motivation for the method of PMSM (*Pixel-wise Motion Similarity Measurement*) in our main manuscript.

When we watch a video, the regions that change drastically normally attract our attention, while those that change gently are overlooked easily. Thus, those pixels with the color difference of less than a threshold T_c two frames will be ignored. In this paper, T_c is set to 4 since it is easy to neglect the difference of color when it is just 4 or less.

For three consecutive frames $x_1, x_2, x_3 \in \mathrm{R}^{m \times n \times 3}$ in the result animation sequence, we hope the motion during these three frames should be as smooth as possible. As shown in Fig.1, for pixels p in frame x_1, the corresponding p' in frame x_2, and p'' in frame x_3, we want to minimize the angle between $\vec{pp'}$ and $\vec{p'p''}$. Therefore, a method is proposed to measure how drastic the change of the motion is.

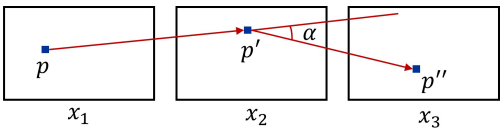

Fig. 1: For pixel p in the first frame and the corresponding pixels p' and p''. In the second and third frames, we consider the angle α between $\vec{pp'}$ and $\vec{pp''}$ as the smoothness of the motion. The smaller α is, the smoother the motion is.

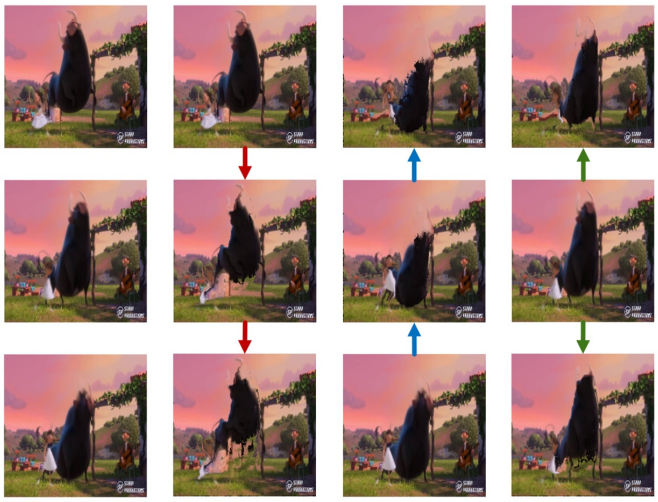

Fig. 2: The results of warping the image with optical flows from different starting frames. The first column is the ground-truth. The mean square error between the result and the ground truth is: red: 2485.8775. blue: 2808.1321, and green: 1970.9794

The optical flow describes where the pixels move, so we can obtain two vectors $\vec{pp'}$ and $\vec{p'p''}$ from optical flows between frames x_1, x_2, x_3. There are three ways to calculate the angle between these two vectors, that is to use the optical flows from three different starting frames. To choose the best approach, we warp the image from the optical flows and compute the error of three ways by the following equation:

$$Err = \frac{1}{W \times H \times C} \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(x_{w,h,c} - x'_{w,h,c}\right)^2, \quad (2)$$

where $x \in \mathrm{W} \times \mathrm{H} \times \mathrm{C}$ is the original frame and x' is the warped image.

Fig.2 shows the results of three different approaches. The column of red arrow is to start from x_1, and warp the image according to the optical flows $F_{x_1 \to x_2}$ and $F_{x_2 \to x_3}$. The column of blue arrow is the result from start frame

x_3 and the optical flows $F_{x_3 \to x_2}$ and $F_{x_2 \to x_1}$. The column of green arrow is the result from starting frame x_2 and the optical flows $F_{x_2 \to x_1}$ and $F_{x_2 \to x_3}$. We can see that the warped images from x_1 to x_3 cause severe consequences. The position of the third pixel will have more deviation due to the optical flow computation error Err added up during the calculation, and leads the corruption of the third image. Therefore, we use the third approach in the right column, which is to compute the optical flow from x_2 to x_3 and the optical flow from x_2 to x_1 to obtain the vectors of optical flow and measure the difference between them.

3 Human Perception-Based Evaluation

Here we detail how we conducted the user study to evaluate the meaningfulness of our generated sequences. The user study is conducted in two rounds. We also use the video data in Fig.5 of the manuscript but there are only seven videos are used in this section, *i.e.,* (A) - (G). It is because these cases have the small Δ_o, *i.e.,* significantly different with the input sequence. Besides, all the sequencing results in this evaluation are resquenced at frame 0, *i.e.,* starting at the same frame with the input clip. The first round is to generate stories of sequences. We recruit 14 users, nine of them are either content creators or have an interest in animation production. Our system allows each user to watch the video and write down the summarization of the video according to their perceptual feeling. We thus get two stories per video.

In the second round, a total of 11 participants are invited to join in this study. Their ages are in range of 20-35 and different backgrounds (five of them have graphics-related background). For each of the 7 videos in the test set, we ask 11 users to watch the video, read the two stories side by side, and rank the stories based on how well they describe the video. We ask the users two questions: *Question 1. Do you think the summarization can describe the video?*, and *Question 2. Do you think that the summarization is interesting/meaningful?* For each video, the participants answer two questions by voting in one of the following five levels: *strongly disagree, disagree, neutral, agree, strongly agree* which correspond to scores of 1, 2, 3, 4, and 5, respectively. Thereafter, we use Eq.(24) in the manuscript to define the meaningfulness degree or the sequence.

4 User Study

In addition to the evaluation metrics in our main manuscript, we conduct two user studies to further learn about human perception on the visual quality of our results and the effectiveness of our proposed system. Two user studies are conducted independently on two distinct groups of participants and designed with different goals. The first one, denoted as U-1, is to validate the visual quality of our rendered videos in terms of stability. The other one, denoted as U-2, is to measure the effectiveness of our system. We note here that the data we use in two user studies is the same set as the evaluation session in our manuscript (i.e., 12 videos in Fig.6).

U-1 is conducted online. A total of 19 participants are invited to join in this user study. Twelve of them have image processing or graphics-related backgrounds. First, we show 12 sets to the participants. Each set consists of a source video and a rendered video by our system. The two videos are displayed in a random order and the participants are not provided any video information to prevent them from inferring the method and having bias in their perceptual feeling. Then, at the end of each set, we ask the participants to judge the stability of each video by voting in one of the following five levels: *1, 2, 3, 4, 5*, (1=bad, 5= very good). Thereafter, we compute the average score from 19 participants as the stable quality of our results. A higher score means better agreement for the good quality. Fig.3 shows the statistics results. The results reveal that the rendered clips by our method are judged to be as stable as the ground-truth. Especially in "I-River flow", our result receives the votes a bit higher than the ground truth. There are two cases, i.e., "B-Daffy Duck" and "D-Frog dance", scores of our result are relative lower than those in the corresponding video. However, they are still in acceptable rate (i.e., higher than 2.5).

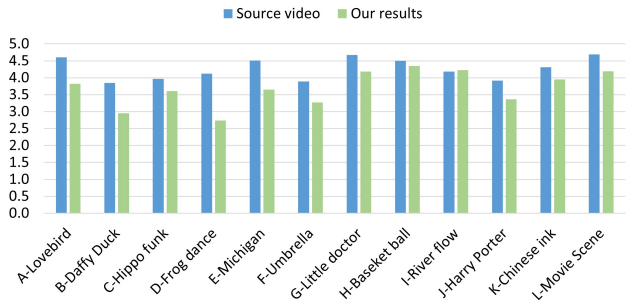

Fig. 3: Analysis result on user study U-1.

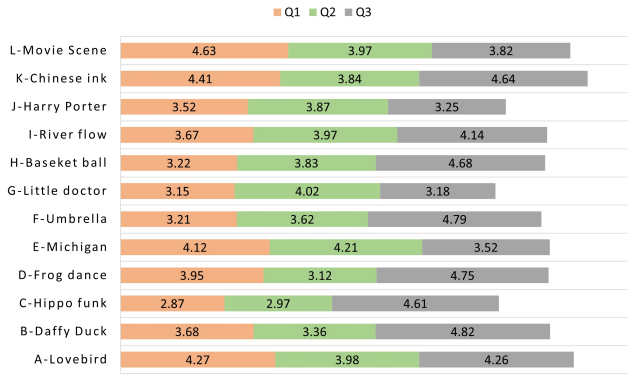

Fig. 4: Analysis result on user study U-2.

U-2 is conducted in-person. We invited totally 11 people to our Laboratory to experience our resequencing system. Eight of them have cartoon animation interest or graphics backgrounds. Unlike U-1, we only use the source videos in this user study and we let the participant generate results by themselves. That is, we let them choose the starting frame of the sequence that they are to make. Once they get the result, they give us the feed back on the result based on their perceptual feeling by answering four questions:

- *Q1. Do you think the resultant sequence is similar to your*

expectation?
- Q2. How do you think about the quality of the rendered video (e.g., any flicking artifact or noticeable discontinuity)?
- Q3. How do you think about the reasonability of the generated sequence?
The participants answer each question by voting in one of the following five levels: *strongly disagree*, *disagree*, *neutral*, *agree*, *strongly agree* which correspond to scores of 1, 2, 3, 4, and 5, respectively. Similar to U-1, we then compute the average score from 11 participants as the effectiveness of our proposed system. A higher score indicates better agreement. Fig.4 presents the analysis results, which demonstrates the evaluation by participants of our system based on the mentioned criteria. From the result, we can see that our proposed system receives positive feed back from participants. Most of the participants are satisfied with the sequences that our system creates with the starting frame given by them. For the quality of the results, our results receive the scores not highest but in the acceptable rate, i.e., from 2.97 to 4.21. In term of the reasonability of the sequence, most of the participants think that the results they gain from our system are reasonable.

5 More visualization

We visualize the general effectiveness of RSFNet in motion distance though data in Fig.6. Assuming that we are to find the adjacent frame of the transition from frame 45. After the first distillation layer, we have a set of 10 frames (S_1) which is defined as content-correlation with frame 45. We compute the motion distance between two optical flows by Eq.(15) of the manuscript in two ways: (a) without RSFNet and (b) with RSFNet. Eq(15) in these two ways respectively appears as:

$$\delta(FC, FK) = - \parallel \mathbf{X}_c^p - \mathbf{X}_k^p \parallel_2, \tag{3}$$

$$\delta(FC, FK) = - \parallel \mathcal{R}(\mathbf{X}_c^p) - \mathcal{R}(\mathbf{X}_k^p) \parallel_2, \tag{4}$$

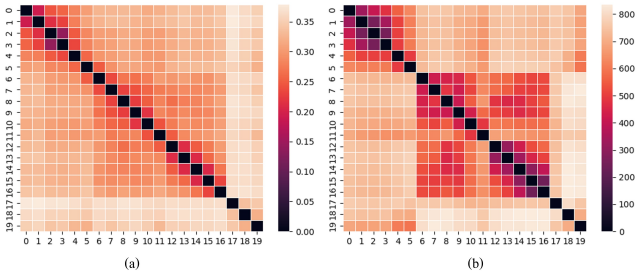

Fig. 5: Heat maps of distance metric calculated by Euclidean distance (a) and our learning-based Euclidean distance (b). The experiment is conducted on Daffy Duck clips, and a segment with 20 frames are picked out here.

The heat maps of the distance metrics shown in the figure reveal that they yield different effects. More specifically, if we denote A and B are the set of candidates can guarantee smooth transitions in (a) and (b) respectively, A={index 4, index 5, index 8} and B = {index 3, index 4, index 6, index 9}. We infer these indices to their visual appearance in frame form and it could be observed that the gesture

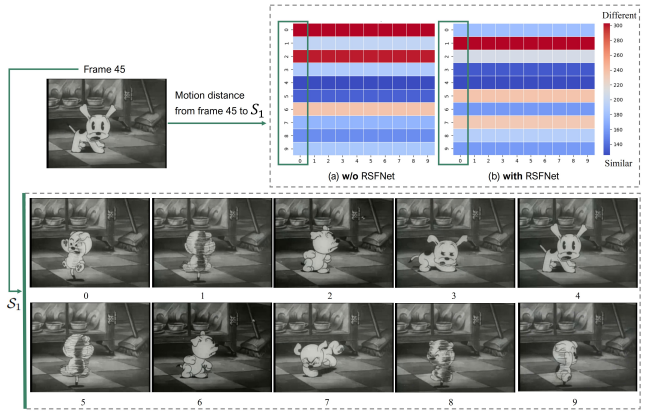

Fig. 6: Demonstrates the effectiveness of RSFNet in motion distance.

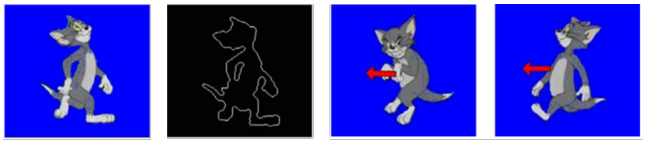

Fig. 7: The characters are obtained from Yang et al. [2]'s figures, Fig.4 and Fig.6. (a) Cartoon character is extracted from the frame, (b) corresponding gesture of (a), (c) the same moving direction is defined by the similar gestures.

s

of the dog in all of the frames in set B could be plausible transitions from frame 45. However, the frame index 8 in set A is obviously a low-degree smooth transition. Besides, the frame index 0, 2, and 6 are defined as different in motion with frame 45. But, we can observe that the gesture in these cases could result in smooth transitions. In contrast, the heat map of (b) shows the more accurate results. Therefore, we utilize RSFNet to boost the accuracy of the motion distance. Video results of these cases can be seen here http://graphics.csie.ncku.edu.tw/SDPF/Transitions.mp4

Fig.5 shows the heat maps of Fig.12 in the manuscript, and Fig.7 depicts the visualization of instance in Yang et al. [2], such as extracted cartoon character, gestures, etc.

References

[1] A. Hore and D. Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

[2] Y. Yang, Y. Zhuang, D. Tao, D. Xu, J. Yu, and J. Luo. Recognizing cartoon image gestures for retrieval and interactive cartoon clip synthesis. *IEEE transactions on circuits and systems for video technology*, 20(12):1745–1756, 2010.