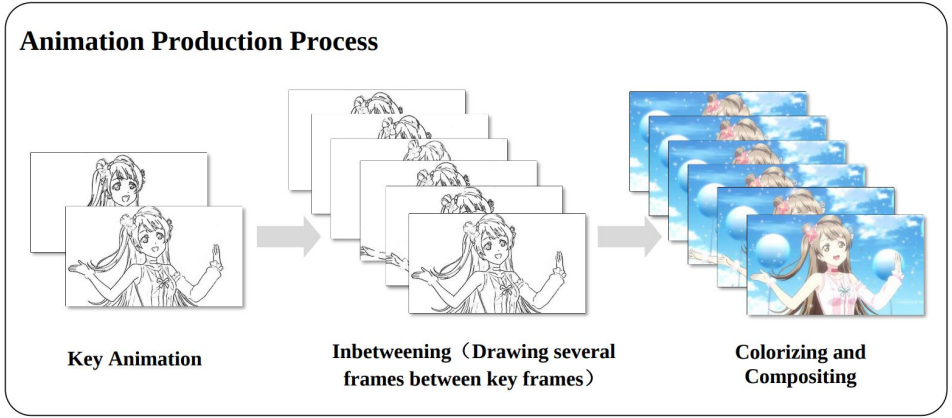
Thin-Plate Spline-based Interpolation for Animation Line Inbetweening

Tianyi Zhu, Wei Shang, Dongwei Ren* , Wangmeng Zuo
Harbin Institute of Technology, China

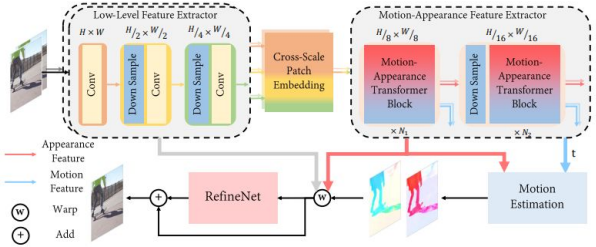
Abstract

- Animation line inbetweening is a crucial step in animation production aimed at enhancing animation fluidity by predicting intermediate line arts between two key frames.
- Chamfer Distance (CD) is commonly adopted for evaluating inbetweening performance. Despite achieving favorable CD values, existing methods often generate interpolated frames with line disconnections, especially for scenarios involving large motion.
- we propose a simple yet effective interpolation method for animation line inbetweening that adopts thin-plate spline-based transformation to estimate coarse motion more accurately by modeling the keypoint correspondence between two key frames, particularly for large motion scenarios.
- Building upon the coarse estimation, a motion refine module is employed to further enhance motion details before final frame interpolation using a simple UNet model.
- we refine the CD metric and introduce a novel metric termed Weighted Chamfer Distance, which demonstrates a higher consistency with visual perception quality. Additionally, we incorporate Earth Mover's Distance and conduct user study to provide a more comprehensive evaluation.

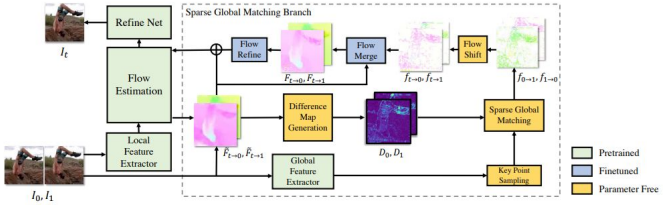


Related Work

- Video Frame Interpolation

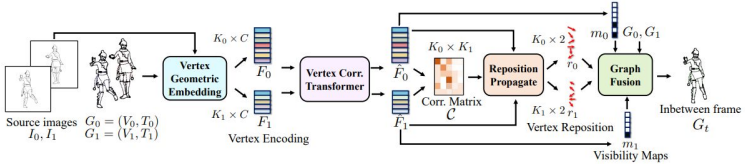


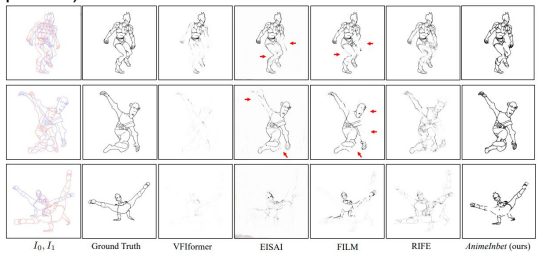
Extracting Motion and Appearance via Inter-Frame Attention for Efficient Video Frame Interpolation (cvpr2023)



Sparse Global Matching for Video Frame Interpolation with Large Motion (cvpr2024)

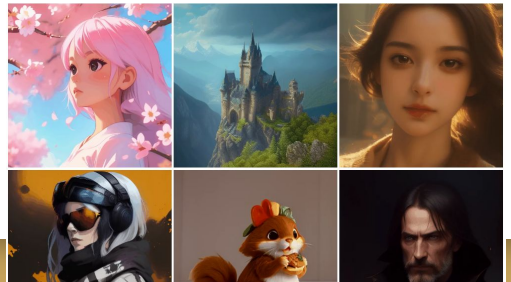
- Line Art Inbetweening





Deep Geometrized Cartoon Line Inbetweening(iccv 2023)

- Video Generation



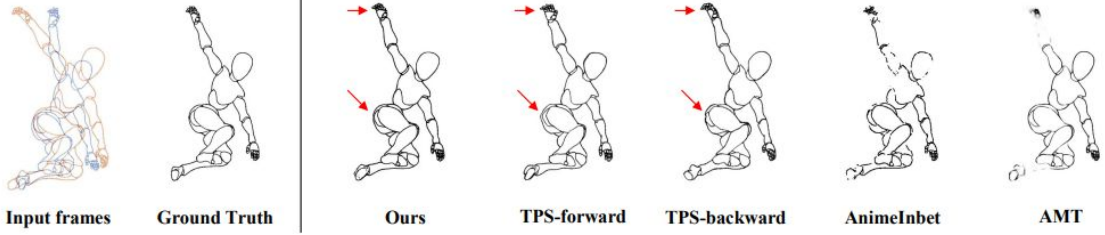
AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning (ICLR'24 spotlight)

Method

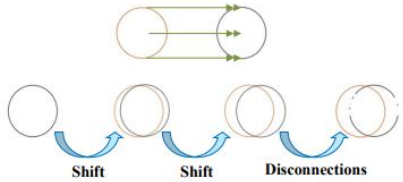
- Limitations in Existing Methods and Metric
  - Chamfer Distance (CD)

$$CD\left(\boldsymbol{b}_0, \boldsymbol{b}_1\right) = \frac{1}{2HWD} \sum \left(\boldsymbol{b}_0 \odot \mathcal{D}(\boldsymbol{b}_1) + \boldsymbol{b}_1 \odot \mathcal{D}(\boldsymbol{b}_0)\right), \tag{1}$$

  - improved metric dubbed Weighted Chamfer Distance (WCD)



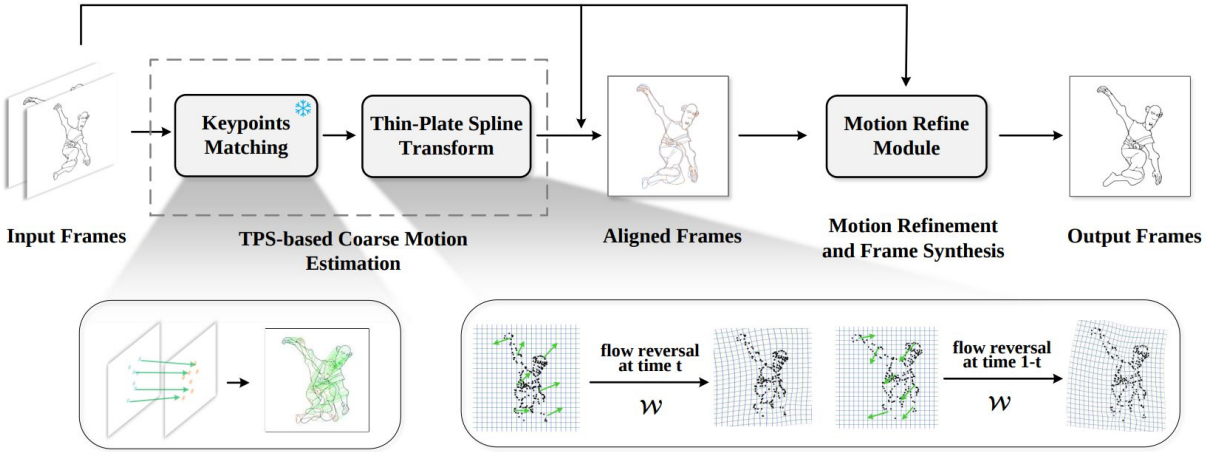
	Input frames	Ground Truth	Ours	TPS-forward	TPS-backward	AnimeInbet	AMT
CD (↓)			5.54	8.29	8.55	5.11	6.75
WCD (↓)			7.50	8.67	8.65	8.10	8.39



CD (↓)	0	3.34	6.28	5.29

Method

- Thin-Plate Spline-based Inbetweening Method



Real-Time Intermediate Flow Estimation for Video Frame Interpolation (eccv 2022)

Method

- Loss Function

$$\mathcal{L} = \mathcal{L}_{dt} + \lambda_{cnt}\mathcal{L}_{cnt} + \lambda_{bi}\mathcal{L}_{bi} + \lambda_{lpips}\mathcal{L}_{lpips}, \qquad (13)$$

dt: the difference between the distance transfor map of predicted frames and the ground truth.
cnt: better inbetweening results usually yield a sum of effective pixels roughly equivalent to that of the ground truth.
bi: To make the resulting line art clearer, we add binarization loss Lbi to encourage the network to generate black or white pixels
lpips: we adopt LPIPS loss Llpips to improve the perceptual quality of line art.

Experiments

- Implementation Details
  - GlueStick [19] as our keypoints matching model.

  - MixiamoLine240 Dataset

  - We set the frame gap N = 5 during training, and tested on the test set with the gaps N = 1, 5, 9.
  - we applied random temporal flipping
  - NVIDIA RTX A6000 GPU

Experiments

- Evaluation Metrics
  - PSNR and SSIM [32] are common evaluation metrics in video interpolation tasks but they often fall short when applied to animation videos
  - Previous works have adopted CD as the evaluation metric
  - To address the shortcomings of the inappropriate metric, we introduce an improved metric based on CD for evaluating line art visual quality, called WCD
    - a weighted factor H(b0, b1) ∈ [0.5, 1] derived from the difference between the effective pixels of the two line art images. Furthermore, we apply a non-linear softplus function mapping G(·) to the values obtained from the distance transform, imposing a greater penalty on values with larger distance.
  - EMD [1, 16], a metric commonly used in 3D point cloud representation, to assist in the assessment of line art quality.

$$WCD\left(\boldsymbol{b}_0, \boldsymbol{b}_1\right) = \frac{\mathcal{H}(\boldsymbol{b}_0, \boldsymbol{b}_1)}{HWD} \sum (\mathcal{G}\left(\boldsymbol{b}_0 \odot \mathcal{D}(\boldsymbol{b}_1)\right) + \mathcal{G}\left(\boldsymbol{b}_1 \odot \mathcal{D}(\boldsymbol{b}_0)\right)), \tag{14}$$

$$\mathcal{H}(\boldsymbol{b}_0, \boldsymbol{b}_1) = \text{Sigmoid}\left(\frac{||\boldsymbol{b}_0|_n - |\boldsymbol{b}_1|_n|}{\min\left(|\boldsymbol{b}_0|_n, |\boldsymbol{b}_1|_n\right)}\right) \tag{15}$$

Experiments

- Comparison with State-of-the-arts

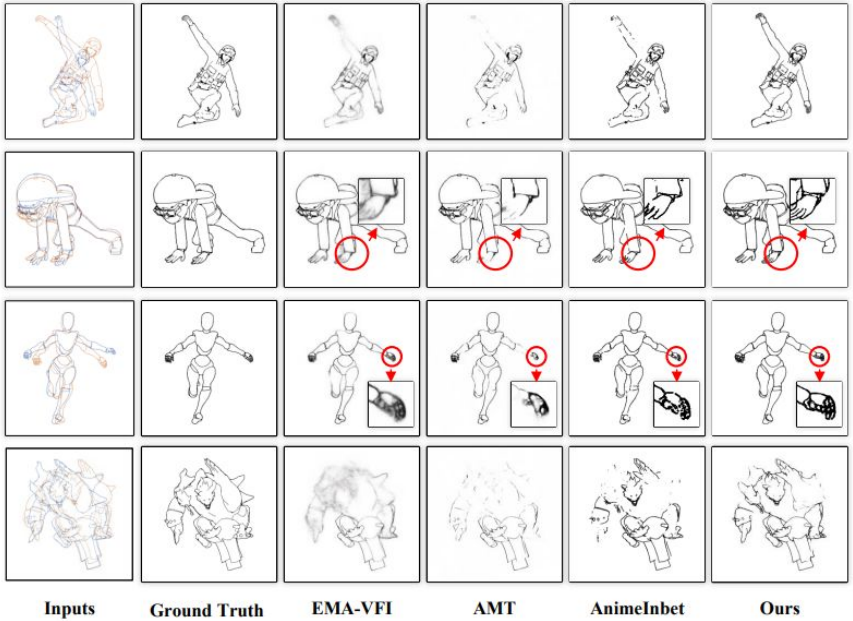


Table 1. Quantitative comparison (CD↓ / WCD↓ / EMD↓) with state-of-the-art methods on different frame gaps. The first place and runner-up are highlighted in bold and underlined, respectively. CD is scaled by 10^5, WCD by 10^4, EMD by 10^3.

Method	$N = 1$ (validation / test)		$N = 5$ (validation / test)		$N = 9$ (validation / test)	
EMA-VFI	3.01/8.62/5.84	3.70/8.59/6.16	15.21/12.04/11.1	17.47/12.00/11.31	28.44/14.94/13.95	31.30/14.86/14.05
RIFE	3.43/8.28/5.48	2.93/8.39/3.81	14.14/11.27/8.37	15.82/11.15/9.63	25.70/13.67/11.61	28.27/13.59/11.77
AMT	2.43/7.33/2.46	2.84/7.31/2.34	14.34/9.58/3.83	16.65/10.13/4.07	22.39/11.62/5.02	25.77/12.21/5.24
PerVFI	2.88/7.91/2.98	3.32/7.79/2.78	11.90/9.23/3.29	12.62/9.05/3.37	19.80/10.20/3.42	21.38/10.09/3.65
EISAI	3.58/8.91/6.26	4.02/8.56/5.86	13.06/9.80/6.29	14.46/9.57/5.80	22.41/10.69/6.27	24.46/10.54/5.79
AnimeInbet	2.51/7.54/3.02	3.07/7.93/3.74	9.33/8.53/3.51	**10.74**/9.84/4.91	16.08/10.01/4.36	**17.76**/11.54/5.74
Ours	**2.34/7.32/1.01**	**2.77/7.28/2.24**	**9.24/7.99/1.34**	10.79/**8.39/2.88**	**15.89/8.85/1.69**	18.09/**9.48/3.37**

Experiments

- User Study
  - we conducted a user study involving 20 participants. Each participant was presented with 50 sets of images randomly selected from the results of EMA-VFI, AnimeInbet, and our method.
  - The ratio of frame gap choices was 20% for 1, 60% for 5, and 20% for 9.

Table 2. User study on evaluating inbetweening performance of competing methods.

	Ours	AnimeInbet	EMA-VFI
$N = 1$	57.50%	31.50%	11.00%
$N = 5$	75.17%	16.33%	8.50%
$N = 9$	69.50%	16.50%	14.00%
All	**70.50%**	19.40%	10.10%

Table 3. User study on the consistency of visual perception and evaluation metrics.

	CD	WCD	EMD
$N = 1$	55.07%	60.67%	52.25%
$N = 5$	57.56%	77.41%	66.57%
$N = 9$	28.66%	68.90%	59.15%
All	51.74%	**72.50%**	61.73%

Experiments

- Ablation Study

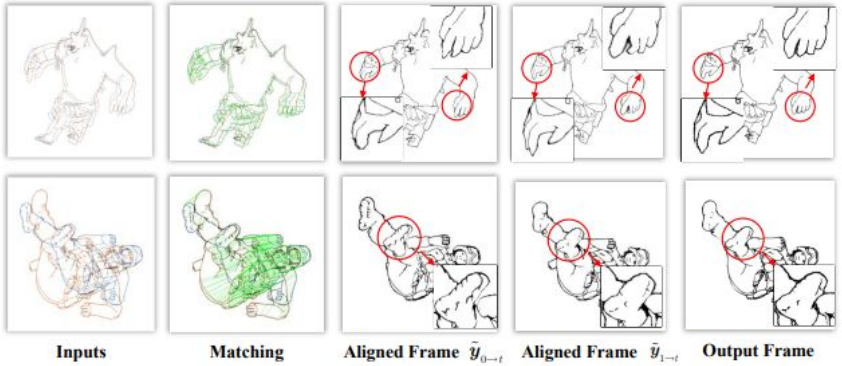


Figure 6. Visual comparison of ablation study on modules, where TPS module can well predict poses in intermediate frames, while FlowNet and UNet bring benefits to detail enhancement.

Table 4. Ablation study on modules.

	CD	WCD	EMD
w/o MRM	11.800	8.850	4.233
w/o flow refine	9.314	8.015	2.625
full model	**9.235**	**7.994**	**1.343**

Conclusion

- we propose a novel framework for addressing the animation line inbetweening problem based on raster images.
- leveraging thin-plate spline-based transformation, more accurate estimates of coarse motion can be obtained by modeling point correspondence between key frames
- we introduce an improved metric called WCD that aligns more closely with human visual perception
- Our quantitative and qualitative evaluations demonstrate the superiority of our method over existing approaches, highlighting its potential impact in relevant fields.