FAULT
JOURNAL

Open Access
Research Article
Received 1st May 2026
Accepted 6th May 2026

Exploring Edge-aware Loss in Sketch Frame Interpolation

AUTHOR NAME: Kong Liang[1,*]

[1] *Looking For Org*
* *Corresponding author: kongliangxm@foxmail.com*

Sketch frame interpolation presents unique challenges compared to natural video: sparse, high-frequency line structures demand precise edge fidelity rather than smooth photometric consistency. We investigate whether a Sobel-based edge-aware loss ($\mathcal{L}_{edge}$) can better serve this domain by explicitly supervising gradient similarity between predicted and ground-truth frames. Fine-tuning the AFI backbone on the STD-12K animation sketch dataset, we compare $\mathcal{L}_{edge}$ against pixel-regression ($\mathcal{L}_1$) and style loss ($\mathcal{L}_{style}$) across PSNR, SSIM, LPIPS, and Chamfer Distance. $\mathcal{L}_{edge}$ underperforms both baselines on all metrics, with qualitative results revealing ringing artifacts, inconsistent line weights, and overly thick strokes. We attribute this to the Sobel operator's indiscriminate response to all high-frequency transitions — including compression noise and soft shading — rather than the clean structural strokes that define sketch quality. We discuss more promising directions, including learned edge detectors conditioned on sketch structure and frequency-domain losses operating on stroke skeletons.

KEYWORDS: Deep Learning, Frame Interpolation, Sketch, Loss Function Design

1 Introduction

Animes, often hand-drawn at low frame rates, pose challenges for frame interpolation due to their non-linear motion and exaggeration. While deep learning excels in full anime keyframes interpolation, applying it to line-drawing frames has been less effective.

It is shown in AFI[3] and LineDiff[4] that loss functions can lead to significant differences in synthesis quality. We formulate Edge-aware Loss with the aim of emphasizing edge similarity to enhance perceptual quality for sketch frames.

2 Methodology

2.1 Model Overview

The overview of the testing backbone AFI[3] is shown in Figure 1. Given two input frames (F_1, F_3), the coarse optical flows $f_{1\to3}$ and $f_{3\to1}$ in both directions are estimated through the AnimeInterp's SGM module[5]. Then the coarse flows are set as the initialization and fed into the RFR module aggregated with the GMA module[2] to be gradually refined and eventually produced as the fine flow $f'_{1\to3}$ and $f'_{3\to1}$. Lastly, based on $f'_{1\to3}$ and $f'_{3\to1}$, the network trained with a ground-truth F_2 warps F_1 and F_3 to synthesize a mid-frame $\hat{F}_2\hat{F}2$ with time $t \in (0,1)$. During training and evaluation, we set $t = 0.5$ to comply with the training triplets.

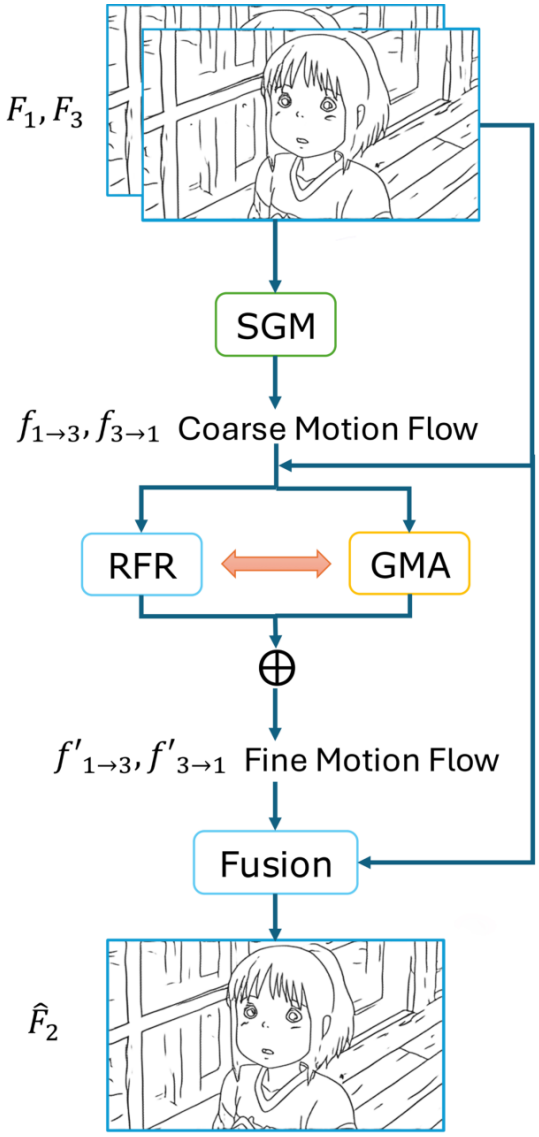

Figure 1: The overall pipeline of AFI[3].

2.2 Edge-aware Loss

The edge-aware loss leverages the well-known Sobel operator[8], a discrete differentiation filter introduced by Sobel & Feldman in 1968. The Sobel operator applies two 3×3 convolution kernels to estimate horizontal and vertical image gradients:

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * I, G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} * I \quad (1)$$

$$\mathcal{L}_{edge}(\hat{I}_t, I_t) = \sqrt{\left\|G_x(\hat{I}_t) - G_x(I_t)\right\|^2 + \left\|G_y(\hat{I}_t) - G_y(I_t)\right\|^2} \quad (2)$$

The loss computes this Sobel-based gradient magnitude for both prediction and ground-truth images, and applies an $\mathcal{L}_2$ loss between these gradient maps. By emphasizing edge similarity, the model is guided to preserve line sharpness and structural fidelity—an especially useful property for sparse, high-frequency data like line drawings.

3 Experiments & Evaluation

The training pipeline utilizes the STD-12K dataset[5,7], which contains 12,000 animation sketch frame triplets. The dataset is pre-processed by resizing frames to 960 × 540 resolution and applying data augmentation techniques such as random cropping, flipping, and frame order reversal.

For ablation studies, we have compared three AFI models trained on L1 Loss, Style Loss[6] and Edge-aware

Loss. As mentioned in AnimeInterp[5], training the model from scratch requires tens of thousand of epochs using optical flow estimation and real-life video frame interpolation datasets. Due to time constraints, all the models have been trained from the pre-trained weights given in AFI[3]. From the pre-trained weights, each model was fine-tuned for 50 epochs.

3.1 Quantitative Results

The quantitative results on the test set of STD-12K[7] are shown below. We employ metrics as PSNR, SSIM, LPIPS[9], and chamfer distance (CD)[1] to quantify perceptual quality. Note that the LPIPS[9] computed is multiplied by 100 for readability. The best and runner-up values are bold and underlined, respectively.

From Table 1, $\mathcal{L}_s$ achieves the best perceptual quality with the lowest LPIPS and the highest SSIM, indicating that style-based supervision produces visually smoother and more coherent interpolations. $\mathcal{L}_1$ leads on PSNR and Chamfer Distance, suggesting it retains pixel-level accuracy and structural edge fidelity more faithfully. In contrast, $\mathcal{L}_{edge}$ underperforms on all four metrics, achieving

Method	LPIPS↓	Chamfer Distance↓	PSNR↑	SSIM↑
AFI - $\mathcal{L}_1$[3]	<u>12.247</u>	**12.280**	**19.505**	0.887
AFI - $\mathcal{L}_s$[3, 6]	**8.950**	<u>17.140</u>	19.461	**0.890**
AFI - $\mathcal{L}_{edge}$	12.444	28.866	<u>19.485</u>	<u>0.888</u>

Table 1: Quantitative Comparison

the worst Chamfer Distance of 28.866 and LPIPS of 12.444, which shows that Sobel-based gradient supervision alone does not effectively constrain structural sharpness. Notably, the gap in PSNR across all three variants is marginal (<0.05 dB), suggesting loss choice primarily affects perceptual and edge-level quality rather than overall pixel fidelity.

3.2 Qualitative Analysis

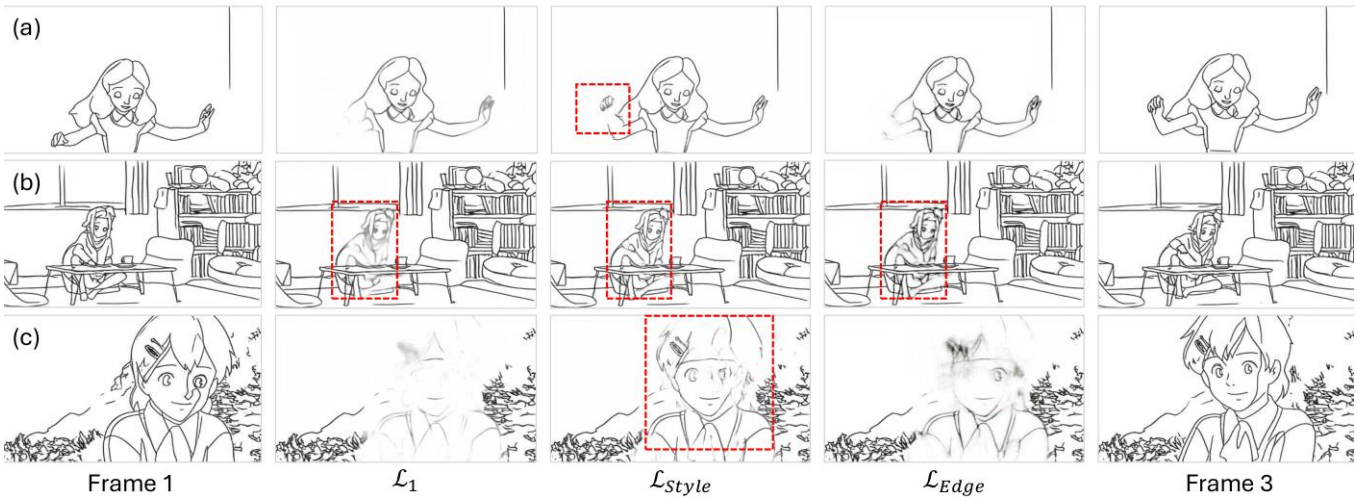

Figure 2: Visual Comparison

Figure 2 shows three representative scenes from the STD-12K[5, 7] test set. Each row displays Frame 1 and Frame 3 as inputs, along with the mid-frame predictions from each loss variant.

In Scene (a), the character performs a fast arm swing. The $\mathcal{L}_1$ result produces a blurred arm region with smeared outlines, while $\mathcal{L}_{style}$ recovers sharper contours and cleaner stroke edges in the highlighted area. $\mathcal{L}_{edge}$ also recovers some edge structure, but introduces visible ringing artifacts and inconsistent line weights along the character silhouette.

In Scene (b), a small-motion scene involving a room environment, all three models output the full outline of the character. $\mathcal{L}_1$ reconstructs the character outlines conservatively but leaves blurry fill areas. $\mathcal{L}_{style}$ better

inpaints the structural lines of the person, with the just right line thickness consistent with the sketch style. $\mathcal{L}_{edge}$ struggles here, producing overly thick lines, leading to its poor Chamfer Distance score.

In Scene (c), a largfe-motion scene with a shaded character, all three models cease to produce satisfactory results. But $\mathcal{L}_{style}$ again preserves the most consistent stroke density and facial feature fidelity (highlighted in red). $\mathcal{L}_{edge}$ perserves slightly more details than $\mathcal{L}_1$.

Overall, the qualitative results corroborate the quantitative findings: $\mathcal{L}_{style}$ yields the most perceptually faithful interpolations for sketch content, while $\mathcal{L}_{edge}$ in its current form amplifies structural errors in complex or disoccluded regions.

4 Conclusion

This work set out to explore whether an edge-aware loss, grounded in Sobel gradient supervision, could improve perceptual quality for sketch frame interpolation. The answer, at least in the formulation tested here, is no — and we report that result openly.

Across all four metrics on the test set, the edge supervision tended to over-emphasise high-contrast boundaries, producing ringing artifacts, inconsistent line weights, and overly thick strokes — the very pathologies it was designed to suppress.

We believe the root cause is a mismatch between the Sobel operator and the nature of sketch edges. Sobel gradients respond to all high-frequency transitions, including compression artifacts and soft shading boundaries, rather than the clean structural strokes that define sketch quality. Without a mechanism to distinguish semantic edges from noise, the loss penalises correct low-contrast detail and rewards spurious high-contrast error.

This negative result is nonetheless informative. It rules out naive gradient-matching as a standalone objective for this domain, and points toward more targeted directions: learned edge detectors conditioned on sketch structure, multi-scale Laplacian supervision, or frequency-domain losses that operate on the stroke skeleton rather than raw pixel gradients. We hope that documenting this false start helps future work avoid the same detour and builds toward a loss formulation that genuinely respects the geometry of hand-drawn lines.

References

[1] Shuhong Chen and Matthias Zwicker. Improving the perceptual quality of 2d animation interpolation. In European Conference on Computer Vision, 2021.

[2] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard I. Hartley. Learning to estimate hidden motions with global motion aggregation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 9752–9761, 2021.

[3] Liang Kong and Suguru Saito. Deep animation video interpolation with global motion aggregation and style loss. In 2025 Nicograph International (NICOInt), pages 88–95, 2025.

[4] Liang Kong and Suguru Saito. Linediff: Semi-generative line-drawing frame interpolation. In Proceedings of the SIGGRAPH Asia 2025 Posters, SA Posters '25, New York, NY, USA, 2025. Association for Computing Machinery.

[5] Siyao Li, Shiyu Zhao, Weijiang Yu, Wenxiu Sun, Dimitris N. Metaxas, Chen Change Loy, and Ziwei Liu. Deep animation video interpolation in the wild. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6583–6591, 2021.

[6] Fitsum A. Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. ArXiv, abs/2202.04901, 2022.

[7] Jiaming Shen, Kun Hu, Wei Bao, Chang Wen Chen, and Zhiyong Wang. Bridging the gap: Sketch-aware interpolation network for high-quality animation sketch inbetweening. In Proc. of ACM International Conference on Multimedia (MM'24), 2024.

[8] Irwin Sobel. An isotropic 3x3 image gradient operator. Presentation at Stanford A.I. Project 1968, 02 2014.

[9] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 586–595, 2018.