COSE474-2024F: Final Project Proposal
“Fine-Tuning FILM Model for Frame Interpolation in Animated Content with ATD-12K Dataset”

Puti Nabilla Aidira

1. Introduction

The animation industry often relies on low frame rates due to the labor-intensive nature of hand-drawn frames. This has created a significant demand for automated interpolation techniques to enhance the visual fluidity of animated content. While recent technologies, such as AnimeInterp (Li et al., 2021) and ToonCrafter (Xing et al., 2024), have shown promising results in animated video interpolation, there is still potential for further exploration.

On the other hand, several effective models have been developed for interpolating real-world videos, one of the most recent being FILM: Frame Interpolation for Large Motion (Reda et al., 2022), which is trained on the Vimeo-90K (Xue et al., 2019) dataset. This project aims to fine-tune the pre-trained FILM model using the ATD-12K animated video frame dataset. The goal is to investigate whether fine-tuning the FILM model with this dataset can yield meaningful insights.

This project seeks to understand the effectiveness of this approach compared to existing state-of-the-art methods. The motivation behind this research is not merely to outperform previous models but to explore the FILM model’s capabilities in handling animated content. Such exploration is valuable for expanding the toolkit available for animated video interpolation, understanding the limitations of pre-trained models, and potentially discovering novel strategies that could inform future research and development.

2. Problem definition & challenges

This project focuses on adapting a pre-trained model designed for real-world video interpolation, FILM: Frame Interpolation for Large Motion, to the domain of animated content. The specific task at hand is generating an intermediate frame given two input frames from an animated video. While models like AnimeInterp and ToonCrafter have been specifically designed to tackle this challenge within the animation domain, FILM was originally trained on a real-world video dataset (Vimeo-90K) that contains natural motion and textures. The core problem is whether fine-tuning FILM with the animation-specific ATD-12K dataset can enable it to generate smooth and realistic intermediate frames, comparable to state-of-the-art animation-focused models.

Adapting a model trained on real-world video data to generate intermediate frames for animated videos introduces several non-trivial challenges. Some of them are the challenges introduced domain-specific to the animated content.

Lack of Texture. Animated frames often feature large, smooth regions with minimal texture. The lack of textures makes it challenging for interpolation models to accurately predict motion between two frames, as the models often struggle and fall into local minima during motion estimation.

Exaggerated and Non-linear Motions. Animation often involves exaggerated and non-linear motion, which is rare in natural videos. Models like FILM, trained on real-world footage, might struggle with extreme or sudden movements, making it difficult to accurately predict the correct position and appearance of objects in the intermediate frame between two given frames.

Handling of Occlusions and Drastic Scene Changes. Animated sequences frequently involve drastic scene changes or occlusions between two frames. A model pre-trained on real-world videos may not effectively handle such cases, where objects may disappear, change positions, or reappear with exaggerated perspectives. The challenge is ensuring that the fine-tuned FILM model can generate plausible intermediate frames under these complex conditions.

These challenges have been addressed by models explicitly crafted for animation video interpolation, such as AnimeInterp and ToonCrafter. However, they highlight the complexities involved in adapting a pre-trained model like FILM to generate intermediate frames for animated videos. In addition to the domain-specific challenges, fine-tuning FILM for animated videos presents its own optimization challenges.

Fine-Tuning Strategies. FILM was initially trained for real-world content, and adapting it to generate intermediate frames for animation without losing generalization requires careful tuning of parameters. Ensuring that the model adapts to animated data without overfitting specific cases in the

ATD-12K dataset is crucial. This includes making decisions about which layers to freeze or unfreeze during training, as well as determining the extent of fine-tuning needed for each layer.

Perceptual and Quantitative Evaluation. Evaluating the quality of generated intermediate frames in animated content is also a challenge. Metrics like structural similarity index measure (SSIM), typically used for real-world videos, may not fully capture perceptual quality in animations. Ensuring that the generated intermediate frame aligns with both the visual style and smooth motion between the two input frames is critical for evaluating success.

3. Related Works

AnimeInterp. AnimeInterp is a framework introduced in the paper ”Deep Animation Video Interpolation in the Wild.” (Li et al., 2021) This innovative approach addresses the unique challenges associated with interpolating animated videos, particularly the issues of texture lack and exaggerated motion. The framework features two main components: Segment-Guided Matching (SGM) and Recurrent Flow Refinement (RFR). The SGM module tackles the challenge posed by the absence of textures in animated frames by performing global matching among color segments that are coherent across frames. This allows the model to establish more accurate motion estimates. Using a transformer-like architecture, the RFR module further enhances interpolation by refining the initial optical flows through recurrent predictions, effectively addressing non-linear and large-motion challenges often encountered in animated sequences. Moreover, to facilitate comprehensive training and evaluation, the paper also introduced the ATD-12K dataset, a large-scale animation triplet dataset comprising 12,000 triplets with rich annotations.

ToonCrafter. ToonCrafter is a pioneering framework introduced in the paper titled ”ToonCrafter: Generative Cartoon Interpolation,” (Xing et al., 2024) which aims to address the limitations of traditional correspondence-based methods for interpolating cartoon videos. These conventional techniques often rely on assumptions of linear motion and struggle with complex phenomena such as dis-occlusion, particularly when dealing with the exaggerated non-linear motions and occlusions that are characteristic of animated content. To overcome these challenges, ToonCrafter leverages live-action video motion priors within a generative framework, significantly enhancing the interpolation process for cartoons. The framework employs a toon rectification learning strategy that adapts these live-action motion priors to the cartoon domain, effectively addressing the domain gap and mitigating content leakage issues. Additionally, ToonCrafter features a dual-reference-based 3D decoder that compensates for detail loss often encountered in compressed latent prior spaces, thereby preserving fine visual details in the generated frames. Furthermore, a flexible sketch encoder is also included, enabling user interaction and control over the interpolation results, allowing for a more tailored and creative output. Comprehensive experiments demonstrate that ToonCrafter not only produces visually convincing and natural dynamics but also excels in handling dis-occlusion.

4. Datasets

For this project, the ATD-12K dataset, introduced in the paper ”Deep Animation Video Interpolation in the Wild,” serves as a primary resource. This large-scale animation triplet dataset is specifically designed to facilitate the training and testing of the AnimeInterp model. The dataset comprises 12,000 triplets, divided into 10,000 training samples and 2,000 testing samples, all of which have been meticulously inspected to ensure quality and consistency. Each triplet in the dataset contains three frames, providing the necessary input and target for frame interpolation tasks. Notably, the test subset of 2,000 triplets includes rich annotations, which are invaluable for evaluating the model’s performance. These annotations include levels of difficulty, Regions of Interest (RoIs) on movements, and tags categorizing different types of motion. The dataset was compiled from 30 series of animated movies, reflecting a wide range of styles and production backgrounds. This diverse collection encompasses over 25 hours of footage and features a total of 101 clips, available in two resolutions: 1920×1080 and 1280×720. Varied animation styles from different producers are also included, enhancing the generalization capabilities of models trained on this dataset as they are exposed to different artistic choices and visual dynamics. However, it is important to note that some triplets within the dataset may have issues such as subtitles and watermarking.

5. State-of-the-art methods and baselines

In addition to AnimeInterp and ToonCrafter, various other state-of-the-art (SOTA) methods have been developed for video interpolation. One notable example is DAIN (Depth-Aware Video Frame Interpolation) (Bao et al., 2019), which employs a depth estimation module to enhance frame interpolation in real-world videos. DAIN uses a two-step approach that involves estimating optical flow and synthesizing intermediate frames based on depth information. While DAIN excels in handling real-world scenarios, its reliance on depth information may not translate well to animated videos, where depth cues can be less pronounced or absent. This highlights a significant limitation when applying traditional video interpolation methods to the unique characteristics of animation.

Another significant contribution to the field is the Super

SloMo framework (Jiang et al., 2018), which focuses on high-quality slow-motion video generation through interpolation. By leveraging a generative adversarial network (GAN) (Goodfellow et al., 2014) approach, Super SloMo produces intermediate frames that exhibit fluid motion, particularly in live-action footage. However, similar to DAIN, its architecture and assumptions about motion may not effectively address the exaggerated and stylized movements commonly found in animated content. Furthermore, the requirement for extensive training data from real-world footage may limit the framework's adaptability to the animation domain, underscoring the need for specialized models that cater to the unique characteristics of animated videos.

Lastly, recent advancements have also paved the way for transformer-based architectures. Models such as VQ-VAE-2 (Razavi et al., 2019) utilize hierarchical latent representations and attention mechanisms to capture long-range dependencies in video sequences. While these methods show promise in improving the quality of generated frames, they often rely on extensive computational resources and large datasets, which may not always be available for animation-specific tasks. As a result, despite the progress made by these SOTA methods, the challenges associated with animated video interpolation remain largely unaddressed, emphasizing the significance of developing frameworks like AnimeInterp and ToonCrafter that are specifically designed for the unique demands of animated content.

6. Schedule

The schedule plan for this project is shown in Table 1.

Weeks	Tasks
Week 1-2	Literature review and setup of the environment with FILM and ATD-12K.
Week 3-4	Fine-tuning FILM on the ATD-12K dataset.
Week 5	Evaluation and comparison with AnimeInterp and ToonCrafter using SSIM.
Week 6	Prepare final report and presentation.

Table 1. Project Schedule Plan

References

Bao, W., Lai, W.-S., Ma, C., Zhang, X., Gao, Z., and Yang, M.-H. Depth-aware video frame interpolation. *arXiv preprint arXiv:1904.00830v1*, 2019.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *arXiv preprint arXiv:1406.2661v1*, 2014.

Jiang, H., Sun, D., Jampani, V., Yang, M.-H., Learned-Miller, E., and Kautz, J. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. *arXiv preprint arXiv:1712.00080v2*, 2018.

Li, S., Zhao, S., Yu, W., Sun, W., Metaxas, D., Loy, C. C., and Liu, Z. Deep animation video interpolation in the wild. *arXiv preprint arXiv:2104.02495*, 2021.

Razavi, A., van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. *arXiv preprint arXiv:1906.00446v1*, 2019.

Reda, F., Kontkanen, J., Tabellion, E., Sun, D., Pantofaru, C., and Curless, B. Film: Frame interpolation for large motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Google Research, University of Washington, 2022.

Xing, J., Liu, H., Xia, M., Zhang, Y., Wang, X., Shan, Y., and Wong, T.-T. Tooncrafter: Generative cartoon interpolation. *arXiv preprint arXiv:2405.17933*, 2024.

Xue, T., Chen, B., Wu, J., Wei, D., and Freeman, W. T. Video enhancement with task-oriented flow. *arXiv preprint arXiv:1711.09078v3*, 2019.