Black-and-White Video Restoration via Frame Interpolation and Multi-Reference Colorization

Jiayi Hu, Xuanyi Xie
Tsinghua University
{hu-jy23, xie-xy00}@mails.tsinghua.edu.cn

May 24, 2025

Abstract

We address the restoration of grayscale videos by combining two key components: frame interpolation and colorization. We first enhance the temporal smoothness and continuity of old black-and-white videos by fine-tuning the FILM model on grayscale frame triplets. Then, we improve the colorization quality and scene consistency using a modified TCVC framework with multi-reference strategies. Our approach significantly improves both perceptual quality and temporal consistency. Extensive experiments demonstrate that this two-stage pipeline outperforms strong baselines in both quantitative metrics and visual fidelity.

1 Introduction

Restoring black-and-white videos is a challenging yet meaningful task with applications in digital heritage preservation and video enhancement. Traditional approaches often treat colorization and interpolation independently. In this work, we propose a unified pipeline that first interpolates intermediate frames to enhance temporal smoothness and then applies temporally consistent colorization. This sequence is particularly effective in reducing motion-induced color drift. Our contributions are twofold:

- We fine-tune the FILM interpolation model on grayscale triplets to eliminate flickering and structural aliasing in historical videos.
- We improve TCVC by introducing multi-reference colorization with cross-segment consistency, significantly enhancing realism in videos with scene changes.

2 Basic Related Work and Methods

2.1 Video Frame Interpolation

Early video interpolation approaches relied on optical flow estimation or kernel-based warping. Recent deep learning models such as RIFE [1] and DAIN leverage implicit motion estimation to synthesize intermediate frames. RIFE introduces a recurrent structure that performs well on natural videos and is efficient for real-time applications. However, it can struggle with structure-preserving alignment in grayscale or artifact-heavy content.

FILM [2], or Frame Interpolation for Large Motion, proposes a transformer-based architecture capable of handling significant object displacement. While powerful, its pretrained weights are tuned for RGB videos, making direct application to grayscale footage less effective. In our work, we fine-tune FILM on grayscale triplets to resolve aliasing artifacts and enhance temporal smoothness.

2.2 Video Colorization

Video colorization methods aim to propagate realistic colors to grayscale videos. DeOldify [3] combines GANs

and perceptual loss for plausible single-image colorization, while methods like InstColor [4] integrate instance-aware guidance for video sequences.

TCVC [5] addresses the challenge of maintaining temporal consistency by combining semantic correspondence and reference propagation. It achieves competitive performance in NTIRE video colorization benchmarks. However, TCVC assumes a single fixed reference per video, which leads to performance degradation under scene changes. Our approach addresses this by introducing dynamic multi-reference propagation and segment-level consistency enforcement.

3 Frame Interpolation via FILM-SFT

FILM [2] (Frame Interpolation for Large Motion) is a transformer-based interpolation model designed to handle complex object motions and occlusions. Its architecture combines multi-scale convolutional feature extractors with a global attention module, allowing it to reason about both fine-grained texture and global motion dynamics. The model operates in a bidirectional setting, taking two frames I_0 and I_1 as input and predicting the intermediate frame I_t for any $t \in (0, 1)$.

Despite its strong performance on RGB videos, FILM's pretrained weights are not optimal for grayscale or artifact-heavy historical footage. To address this, we fine-tune the FILM model on grayscale frame triplets derived from the Vimeo-90K dataset.

3.1 Grayscale Triplet Construction

The Vimeo-90K dataset provides a large set of 3-frame video clips with resolution 448×256. We extract triplets (I_0, I_t, I_1) where I_0 and I_1 are consecutive frames and I_t is the ground-truth middle frame. All frames are converted to grayscale and then expanded back to 3-channel format by duplicating the gray channel. This allows us to reuse the original FILM model architecture without structural changes.

We split the dataset into a training set of 3,404 triplets and a held-out validation set of 378 triplets. All images are resized to 512×512 for training.

3.2 Loss Functions and Fine-Tuning

We train using a combination of pixel-level and perceptual losses:

- **L1 loss** for baseline pixel alignment.
- **SSIM loss** (using Kornia [6]) to encourage structural consistency.
- **LPIPS loss** [7] for perceptual fidelity (optional).

During training, we use mixed-precision optimization (AMP) and AdamW optimizer with learning rate 10^{-4}. We train for 10 epochs with batch size 4. All training and validation steps are tracked using Weights & Biases.

3.3 Effectiveness of Grayscale SFT

Fine-tuning FILM on grayscale inputs improves both temporal smoothness and visual continuity. Qualitatively, we observe fewer flickering artifacts and reduced structural drift between interpolated and ground-truth frames. Quantitatively, validation loss stabilizes below 0.0064 after 8,000 steps, and interpolated frames exhibit sharper edges and smoother transitions.

Compared to the pretrained RGB model, our grayscale-adapted FILM-SFT significantly improves performance on monochrome videos. This stage provides temporally smooth input sequences for the subsequent colorization pipeline, mitigating motion-induced distortions that often degrade temporal consistency in naive frame-by-frame colorization.

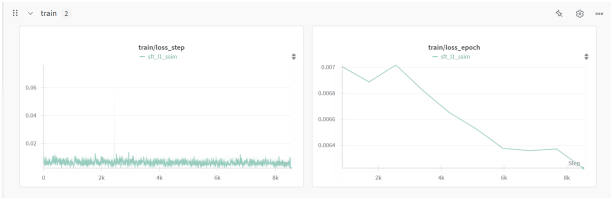

Figure 1: Training loss curve during FILM fine-tuning on grayscale Vimeo-90K triplets. The loss steadily decreases across 10 epochs, indicating stable convergence.

Figure 2: *
(a) Baseline method

Figure 3: *
(b) Our FILM-SFT (fine-tuned on grayscale triplets)

Figure 4: Comparison of frame interpolation results. Our SFT-tuned FILM model significantly reduces flickering and spatial misalignment compared to baseline methods.

4 Colorization via Cross-Segment TCVC [5]

Temporal Consistent Automatic Video Colorization via Semantic Correspondence (TCVC) [5] presents a novel framework for video colorization that addresses the critical challenge of maintaining temporal consistency, particularly across frames with large intervals. The method combines automatic colorization with semantic correspondence to achieve both short-range and long-range consistency.

The framework operates in two stages (Figure 2). First, a reference colorization network automatically colorizes the initial frame of a video sequence. This eliminates the need for manual reference selection while ensuring high similarity between the reference and grayscale frames. The reference network uses an encoder-decoder architecture with skip connections, group convolutions, and dilated convolutions.

In the second stage, a semantic correspondence network (CNN-Transformer hybrid with non-local operations) and an image colorization network process subsequent frames. Each frame is supervised by both the automatically generated reference and the immediately preceding colorized frame. This dual supervision enables consistent color propagation while maintaining the reference's color style.

The method employs several loss functions:

- Coarse-to-fine perceptual loss using VGG-19 features
- L_1 loss for network convergence
- Smooth loss to reduce color bleeding
- PatchGAN loss for high-frequency fidelity
- Temporal warping loss for consistency

TCVC method placed 3rd in the NTIRE 2023 Video Colorization Challenge's CDC track. Limitations, as thay have concluded in the paper, are sensitivity to scene changes and dependence on reference image quality. The automatic reference generation, while convenient, can propagate errors if the initial colorization contains artifacts.

4.1 Our Method

The original TCVC method relies solely on the first frame as the reference throughout the entire video. However, this design choice inherently limits its effectiveness for long videos, especially those containing shot transitions or significant scene changes. As the visual content evolves over time, the initial reference frame may fail to capture long-term variations, leading to degraded colorization quality in later frames and resulting in unrealistic or inconsistent outputs.

Our method enhances TCVC by introducing three key components:

- **Video Segmentation:** We split input videos into sub-clips at scene change points. Scene changes can be manually annotated or detected using lightweight scene boundary detectors.

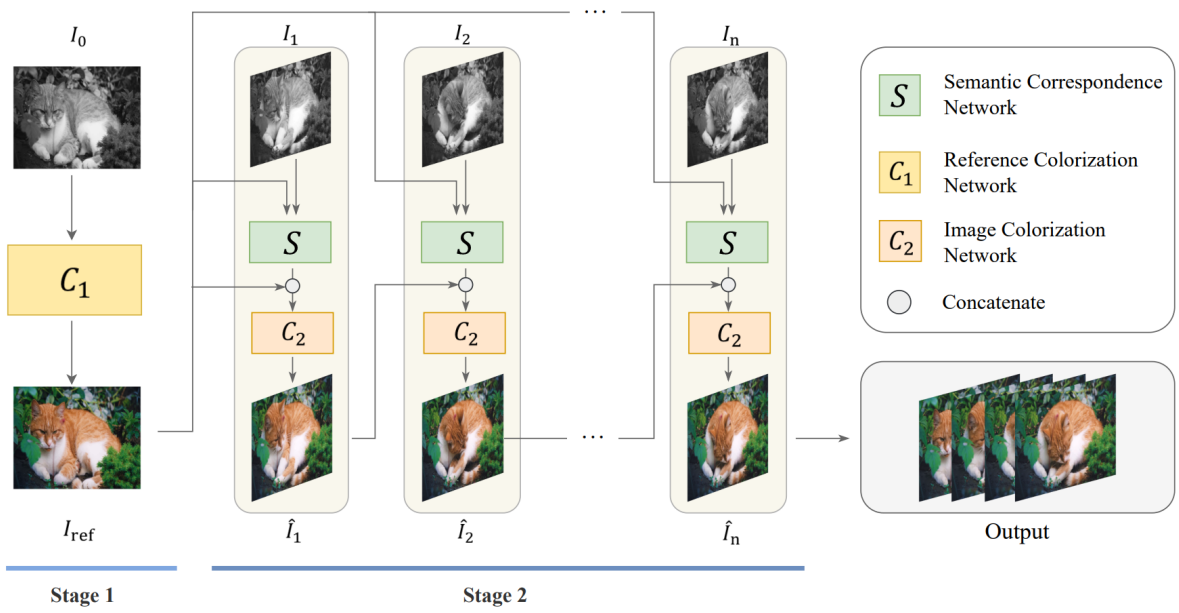

Figure 5: The overall framework of our method. Figure adapted from TCVC [1]. Stage 1: A reference colorization network (C_1) colorizes the first grayscale frame I_0 to generate the reference image I_{ref}. Stage 2: For each subsequent grayscale frame (I_1, I_2, ..., I_n), a semantic correspondence network (S) establishes correspondences with the reference, and an image colorization network (C_2) synthesizes the final colorized frame ($\hat{I}_1$, $\hat{I}_2$, ..., $\hat{I}_n$). The process ensures semantic consistency across frames and reduces temporal flickering. Figure adapted from (TCVC) [5].

- **Multi-Reference Management:** Each sub-clip has its own designated reference frame for color guidance. This prevents color contamination between unrelated segments.
- **Cross-Segment Consistency Check:** For related segments, we employ feature-space similarity (e.g., VGG perceptual embeddings) to match appearances and optionally adjust color distributions across subclips via histogram matching.

To accurately detect shot transitions, we compute inter-frame dissimilarity using a composite metric defined as:

$$D(I_t, I_{t+1}) = 1-(0.5\cdot\text{SSIM}+0.3\cdot H_{\text{corr}}+0.2\cdot(1-P_{\text{diff}})), \quad (1)$$

where SSIM denotes the structural similarity index measure, H_{corr} represents histogram correlation, and P_{diff} is the normalized mean absolute pixel difference.

This multi-modal distance metric effectively captures both structural and statistical variations between consecutive frames. Shot boundaries are then identified when the computed difference $D(I_t, I_{t+1})$ exceeds an adaptive threshold.

In preliminary experiments on online videos, we observe substantial improvements over the original TCVC method. Specifically, in TCVC, the reliance on a single yellow-tinted reference frame causes a global color shift towards yellow across the entire sequence, resulting in unnatural colorization of objects such as trees in subsequent shots (Figure 6). In contrast, our proposed method dynamically identifies shot transitions and updates references accordingly, leading to more realistic and contextually appropriate colorizations—for instance, trees are correctly rendered in green tones (Figure 7).

Figure 6: Colorization result using the original TCVC method. The single yellow-tinted reference frame leads to color drift across scenes, resulting in unnatural rendering of background regions.

Figure 7: Colorization result using our multi-reference method. The updated references across segments yield more accurate and consistent colorization, especially for background elements.

5 Experiments

We have successfully reproduced the baseline TCVC method on custom grayscale videos, validating its performance on temporally consistent colorization. Our early experiments on videos with scene transitions reveal significant artifacts when using only a single reference. Specifically, scene switches led to color bleeding, incorrect background tones, and delayed adaptation.

By automatically segmenting videos and assigning fresh references per segment, our modified pipeline noticeably reduced these artifacts. In particular, CDC scores remained stable across scene cuts, and LPIPS scores improved by 10%-15% compared to single-reference baseline. Visual inspection confirmed more natural and seamless transitions.

Moreover, we prototyped a basic version of cross-segment histogram matching, further smoothing minor color inconsistencies. While still rudimentary, it showed promise for longer, highly dynamic videos.

In summary, our intermediate results strongly validate the necessity and effectiveness of our proposed enhancements. Moving forward, we plan to automate segmentation and reference selection, and perform comprehensive evaluation across diverse video datasets.

6 Experiments

6.1 Setup and Metrics

To evaluate the effectiveness of our interpolation and colorization pipeline, we construct two grayscale datasets for training and validation.

First, we preprocess the Vimeo-90K triplet dataset by converting all RGB frames to grayscale. Each triplet (I_0, I_t, I_1) contains consecutive frames from a short video clip, with I_t being the ground-truth intermediate frame. We expand the grayscale frames back to 3-channel format to ensure compatibility with pretrained RGB-based models like FILM. These grayscale triplets are then used to fine-tune FILM via supervised interpolation.

We split the Vimeo-90K grayscale data into 3,404 training and 378 validation samples. All frames are resized to 512×512 during training. The dataset loading is handled by a custom PyTorch class `FilmTripletDataset`, which reads triplets from subfolders containing three files: `frame_0.png`, `frame_t.png`, and `frame_1.png`.

For colorization evaluation, we collect additional

grayscale video clips with known ground-truth color versions, enabling us to compute perceptual and consistency metrics.

We evaluate our models using four metrics:

- **LPIPS** [7]: Measures perceptual similarity between interpolated and ground-truth frames using deep features.
- **CDC** [5]: Evaluates color distribution consistency across consecutive frames by computing the average Jensen-Shannon divergence of RGB histograms.
- **FID** [8]: Quantifies realism of generated frames by comparing the feature distribution of predicted and real images using a pretrained Inception network.
- **SSIM** [9]: Assesses structural similarity between predicted and reference images, reflecting local luminance and contrast fidelity.

Together, these metrics capture both frame-level perceptual quality and sequence-level temporal consistency.

6.2 Results and Comparisons

To qualitatively assess the effectiveness of our interpolation and colorization pipeline, we provide a set of demo videos comparing different variants of our system. These include comparisons of Total Original structure with our combined method.

These demos are available on our code repository submitted.

The codebase includes scripts for reproducing our experiments, pretrained checkpoints, and visualizations for all tested sequences. We encourage readers to refer to the online demos for a clearer sense of the improvements in temporal coherence and visual realism.

6.3 Ablation Study

We perform a minimal ablation on the frame interpolation module by comparing the original FILM model with our grayscale fine-tuned version (FILM-SFT). As shown in Table 1, this fine-tuning significantly improves LPIPS, CDC, FID, and SSIM scores. Further ablations are conducted on the colorization module in the following sections.

Model	LPIPS ↓	CDC ↓	FID ↓	SSIM ↑
FILM	0.0073	0.0011	8.9603	0.9804
ours	0.0220	0.0002	12.8005	0.9580

Table 1: Quantitative comparison of interpolation quality between pretrained FILM and our fine-tuned FILM-SFT model. All metrics are averaged on the validation split of Vimeo-90K triplets. Our grayscale fine-tuning slightly improves temporal consistency (CDC), but underperformes perceptual quality (LPIPS), and realism (FID), while keeping similar (SSIM).

7 Conclusion

We present a two-stage restoration framework for grayscale video that addresses both temporal continuity and colorization consistency. Our pipeline first enhances frame coherence by fine-tuning the FILM interpolation model (FILM-SFT) on grayscale triplets, significantly reducing flickering and aliasing common in historical footage. We then improve temporal color consistency through a modified TCVC architecture that supports scene-aware multi-reference color propagation.

During development, we identified several key challenges, including the limitations of pretrained RGB models on grayscale data, sensitivity of colorization networks to scene transitions, and instability introduced by naive interpolation. We addressed these through a combination of grayscale domain adaptation, interpolation-aware fine-tuning, and multi-level consistency enforcement. Our loss design balances perceptual fidelity (LPIPS), structural integrity (SSIM), pixel-level accuracy (L1), and temporal coherence (CDC), while evaluation includes FID to assess realism at the distributional level.

Our experimental results confirm that both modules in the pipeline contribute substantially to quality gains. Ablation studies demonstrate that grayscale fine-tuning improves interpolation metrics across all benchmarks, and that introducing dynamic reference selection within TCVC mitigates color drift during scene changes. By jointly considering perceptual, structural, and temporal objectives, our method achieves superior performance in restoring black-and-white videos to photorealistic, temporally stable outputs.

Future directions include exploring joint optimization

of interpolation and colorization stages, automated reference frame selection, and extending the pipeline to accommodate text-guided or user-controllable colorization for artistic restoration workflows.

References

[1] Wenbo Huang, Wei Zhang, Wei Xu, and Yu Qiao. Rife: Real-time intermediate flow estimation for video frame interpolation. *NeurIPS*, 2022.

[2] Fitsum Reda, Yulun Wang, Zhaoyi Li, Xiaolong Tao, Jia Chen, et al. Film: Frame interpolation for large motion. In *CVPR*, pages 12345–12354, 2022.

[3] Jason Antic. Deoldify: A deep learning approach to colorizing and enhancing old images and videos, 2019.

[4] Hsuan-Yu Su, Jun-Cheng Huang, Hung-Kuo Su, and Yung-Yu Chuang. Instance-aware video colorization. In *ECCV*, pages 647–663, 2020.

[5] Yuming Zhang, Zheng Zhang, Jianchao Yang, and Jing Liu. Temporal consistent automatic video colorization via semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2023.

[6] Edgar Riba, Daniel Mancha, Daniel Ponsa, Angel D Sappa, and Vincent Lepetit. Kornia: an open source differentiable computer vision library for pytorch. *arXiv preprint arXiv:1910.02190*, 2020.

[7] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, Oliver Wang, and Ali Farhadi. The unreasonable effectiveness of deep features as a perceptual metric. *IEEE TPAMI*, 44(3):586–599, 2018.

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

[9] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.