Frame Interpolation via Multi-Directional Discrete Shift Modeling with Residual Learning U-Net Refinement

Huy Duc Vu
University of Engineering and Technology
23020380@vnu.edu.vn

Abstract

Frame interpolation aims to generate intermediate frames to increase frame rates and improve motion smoothness. While most existing methods rely on continuous optical flow estimation, we propose a lightweight alternative that represents motion using multi-directional discrete shifts.

To enhance interpolation quality and reduce structural artifacts, a U-Net-based refinement network is introduced and trained with residual learning. Inspired by denoising formulations in diffusion models, this module removes structured noise from the interpolated frames. The proposed method is evaluated against several frame interpolation approaches to demonstrate its effectiveness.

1 Introduction

Frame interpolation is a fundamental task in video processing, with applications in frame rate up-conversion, motion smoothing, and enhanced visual quality. The goal is to synthesize intermediate frames that preserve motion continuity while maintaining fine image details.

Despite recent advances, frame interpolation remains challenging due to complex motion patterns in real-world videos. Fast motion, occlusions, and motion boundaries often lead to structural distortions and artifacts such as ghosting or double-vision effects, requiring effective motion modeling and artifact correction.

Beyond visual quality, model complexity is also an important consideration. Many methods rely on large scale models, limiting their deployment in resource-constrained settings. As a result, developing lightweight approaches that achieve competitive interpolation performance remains an important research direction.

2 Proposed Method

2.1 Overall Architecture

Given two input frames, the model first extracts multi-scale features and estimates motion using a multi-directional discrete shift mechanism across different scales. The shifted frames from both temporal directions are aggregated using a visibility mask to produce an initial interpolated frame. Finally, a U-Net–based refinement network is applied to correct structural artifacts and improve the visual quality of the output frame.

2.2 Multi-Directional Discrete Shift Motion

Instead of estimating dense continuous optical flow at the pixel level, the proposed model represents motion using discrete shifts across multiple directions and scales, as illustrated in Fig. **1**. Inspired by the adaptive warping mechanism in AdaCoF [1], complex motion is approximated through a weighted combination of fixed spatial displacements, which simplifies motion estimation and reduces the complexity of the prediction branch.

A predefined set of shift directions, including axial, diagonal, and zero-displacement directions, is adopted at each scale. The shift flow module learns per-pixel weights for each direction, which are normalized using a softmax function to model their relative contributions.

Based on these learned weights, the input frames are shifted along each direction and scale and then aggregated in a weighted manner to generate warped frames from both temporal directions. The multi-scale design enables effective modeling of both small and large motions in video sequences.

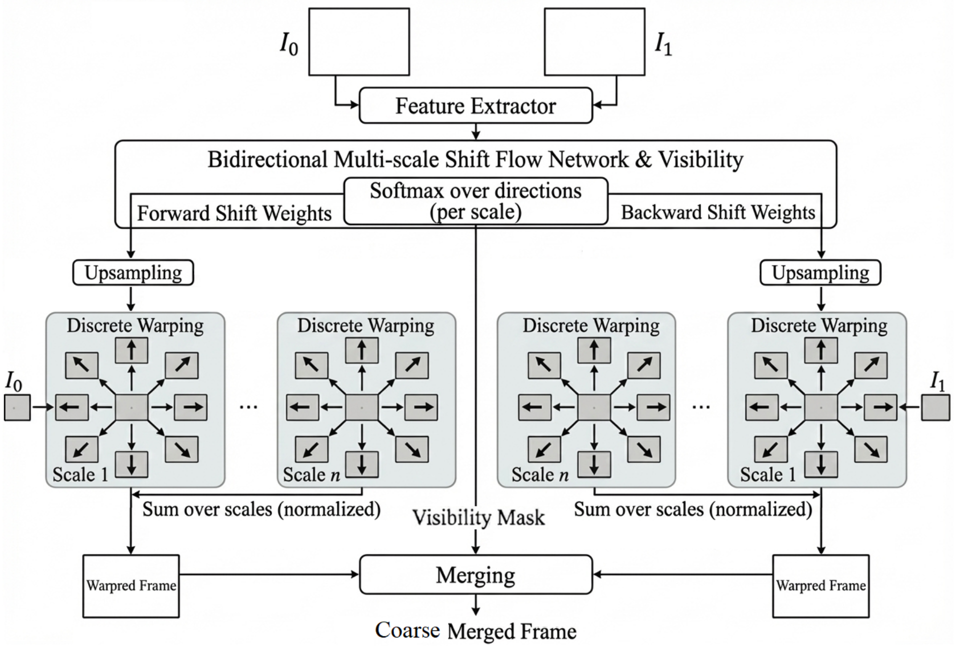

Figure 1: Overview of the Bidirectional Multi-scale Shift-based Frame Interpolation

2.3 Intermediate Frame Synthesis

Based on the learned shift weights, the model generates two warped frames corresponding to the two temporal directions, obtained by warping the preceding and succeeding input frames, respectively, as illustrated in Fig. **1**.

To fuse these warped frames, a single-channel visibility mask is jointly learned at each spatial location to model their relative reliability, allowing the model to handle occlusions and asymmetric motions.

The coarse merged frame is synthesized via a weighted linear combination of the two warped frames:

$$\hat{I} = V \odot I_0^w + (1 - V) \odot I_1^w, \tag{1}$$

where I_0^w and I_1^w denote the warped frames from the two input images, V is the visibility mask, and $\odot$ denotes element-wise multiplication.

The resulting frame $\hat{I}$ serves as a coarse intermediate representation and is subsequently refined by the refinement module.

2.4 Residual Refinement with Context-Aware U-Net

Although the coarse merged frame captures the overall motion structure, it may still contain artifacts and inaccuracies near motion boundaries or occluded regions. To address these issues, we employ a context-aware refinement network based on a U-Net [2] architecture that predicts a residual correction.

As illustrated in Fig. **2**, the refinement network takes as input a concatenation of the coarse merged frame, the warped frames from both input images, the visibility mask, and the learned shift flow weights. These inputs jointly encode appearance, motion, and confidence cues to facilitate effective artifact suppression.

The network follows an encoder–decoder structure with skip connections, where the encoder captures large-scale contextual information and the decoder progressively restores fine spatial details. Multi-scale context features are further injected into the encoder and bottleneck layers to improve robustness to complex motions and occlusions.

Instead of directly synthesizing the final frame, the network predicts a residual image that corrects errors in the coarse prediction. The refined output is obtained by adding this residual to the coarse merged frame, allowing the network to focus on refinement rather than full reconstruction.

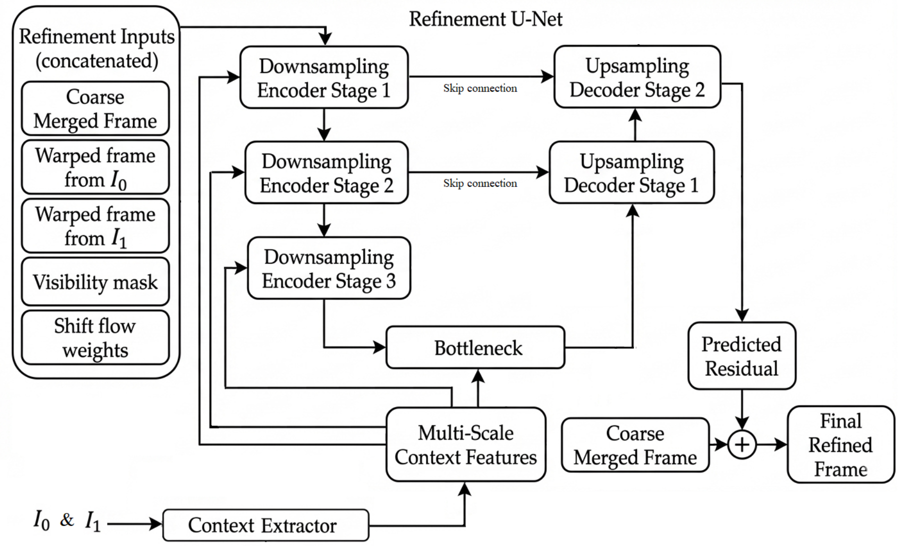

Figure 2: Overview of Residual Refinement with Context-Aware U-Net

3 Training Strategy

The proposed network is trained end-to-end using a composite loss function. The primary supervision is a Laplacian pyramid loss that enforces accurate reconstruction across multiple spatial scales. In addition, a VGG-based perceptual loss encourages semantic consistency, while a ternary census loss improves local texture preservation and robustness to illumination changes.

For knowledge distillation, the student model is guided by a stronger teacher that differs only by having access to ground-truth frames, enabling more reliable motion estimation and effective supervision for shift-weight distillation.

The total training loss is a weighted sum of all components. Optimization is performed using AdamW with cosine learning rate scheduling and warm-up, along with mixed-precision training and gradient clipping for stability. The model is trained on the Vimeo90K triplet dataset [3] for 300 epochs with a batch size of 256, an initial learning rate of 1.2×10^{-3} decayed to 1.2×10^{-5}, weight decay 8×10^{-4}, and 2000 warm-up steps. Training is conducted on a single NVIDIA H200 SXM5 GPU for approximately 30 hours.

4 Result

4.1 Model inference result

Figure **3** qualitatively illustrates the intermediate representations and the final interpolated frame produced by the proposed method with scales of $[1, 2, 4, 8, 16, 32]$. Total parameters of the model is $3.1M$ based on the number of scales.

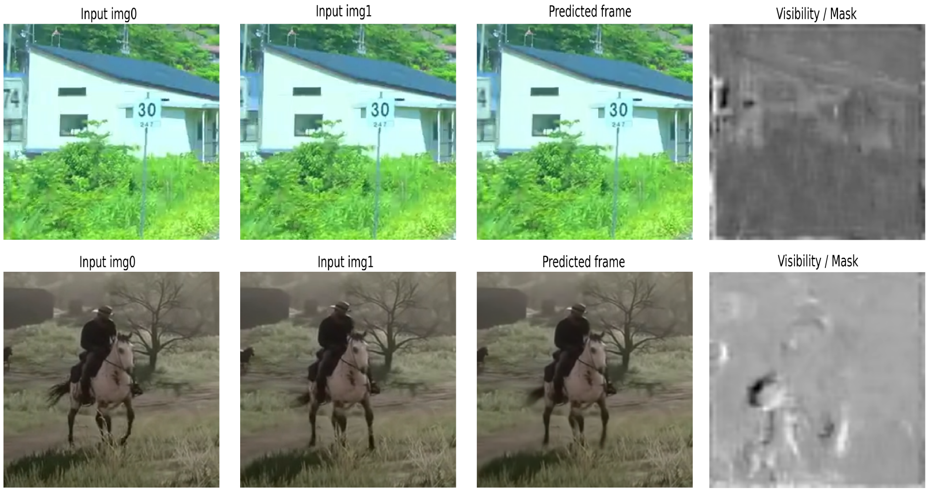

Figure 3: Qualitative results of the proposed interpolation framework, including input frames (I_0, I_1), predicted intermediate frame and the estimated visibility mask

The example in Figure **3** - upper images illustrates a left-to-right motion, where we can see the ”74” sign is visible in I_0 but largely occluded in I_1. The predicted frame partially recovers the sign based on the estimated visibility mask, indicating effective handling of motion-induced occlusions.

The example in Figure **3** - lower images illustrates a horse walking sequence where two input frames are temporally distant. Our model successfully predicts the intermediate frame, preserving motion continuity and fine structural details of the legs and body.

4.2 Limitations

Figure **4** illustrates a challenging scenario involving complex and fast human motion. Due to large non-linear displacements between the two input frames, the model produces ghosting and double-vision artifacts, leading to noticeable blur all around the moving subject.

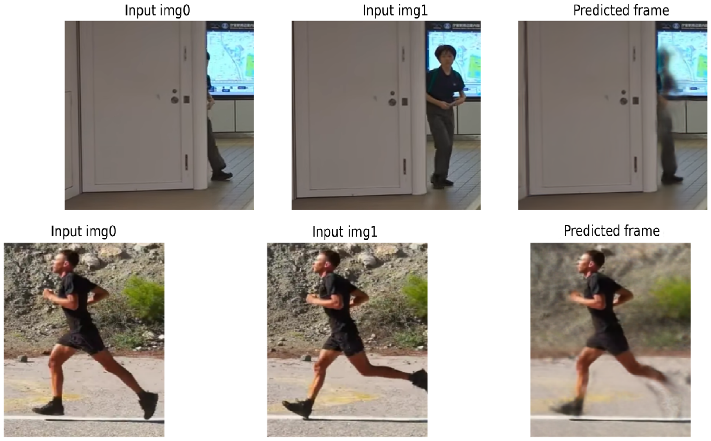

Figure 4: Complex human motion, showing ghosting and blur artifacts in the interpolated frame

In addition to motion-related challenges, the proposed model still contains several computationally suboptimal operations. Specifically, the multi-scale shift-based warping relies on Python-level loops over motion directions and scales, as well as repeated tensor shifting and interpolation operations. These operations are not fully optimized for GPU parallelism and introduce additional memory overhead, limiting inference efficiency despite the lightweight network design.

4.3 Comparison

Table 1: Quantitative comparison with several existing methods on UCF101 ,Vimeo90K and MiddleBury.

Method	# Parameters (Million)	Runtime (ms)	UCF101[4]		Vimeo90K [3]		M.B. IE
			PSNR	SSIM	PSNR	SSIM	
DVF*	**1.6**	80	34.92	0.968	34.53	0.973	2.47
RIFE*	9.8	**16**	35.28	0.969	35.61	0.978	1.96
AdaCoF*	21.8	34	34.91	0.968	34.27	0.971	2.31
VFIformer†	24.1	365	**35.43**	**0.970**	**36.50**	**0.981**	**1.82**
Ours*	3.1	92	35.28	**0.970**	35.08	0.974	2.13

* DVF, RIFE, AdaCoF and Ours are evaluated using the original evaluation code provided in the RIFE paper [5]; the evaluation code for Ours is adapted from the RIFE benchmark for compatibility.

† The reported results of VFIformer are taken directly from its original paper [6].

Bold indicates the best performance.

Underline indicates the second-best performance.

Our method achieves competitive interpolation quality with a compact model size, at the cost of higher runtime due to current Python implementation inefficiencies as mentioned earlier.

5 Conclusion

This paper presents a lightweight frame interpolation method based on multi-directional discrete shift modeling with residual refinement using a U-Net architecture. By replacing dense optical flow with discrete shift representations, the proposed approach reduces model complexity while maintaining competitive interpolation quality.

Rather than aiming to outperform optical flow-based methods, this work explores an alternative formulation that demonstrates competitive performance can be achieved with substantially fewer parameters. The residual refinement strategy focuses on correcting structured artifacts in the coarse prediction, improving visual quality near motion boundaries and occlusions. Experimental results show that the proposed model achieves comparable performance to existing methods with only 3.1M parameters.

Future work will focus on optimizing multi-scale shift operations for better GPU parallelism and extending the motion representation to handle complex non-linear motions and fast-moving objects.

References

[1] H. Lee, T. Kim, and I. S. Kweon, “Adacof: Adaptive collaboration of flows for video frame interpolation,” in *CVPR*, 2020.

[2] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[3] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, “The vimeo-90k dataset: Large-scale video frame interpolation benchmark,” in *CVPR*, 2019.

[4] K. Soomro, A. R. Zamir, and M. Shah, “Ucf101: A dataset of 101 human action classes from videos in the wild,” University of Central Florida, 2012. Used for frame interpolation evaluation.

[5] Z. Huang, Q. Liu, T. Wang, and X. Wang, “Rife: Real-time intermediate flow estimation for video frame interpolation,” in *ECCV*, 2020.

[6] Z. Shi, X. Chen, Y. Chen, and X. Zhang, “Vfiformer: Video frame interpolation with transformer,” in *CVPR*, 2022.