

LADDER: An Efficient Framework for Video Frame Interpolation

Tong Shen¹, Dong Li¹, Ziheng Gao¹, Lu Tian¹, Emad Barsoum¹

¹ Advanced Micro Devices, USA
{t.shen, d.li, ziheng.gao, lu.tian, ebarsoum}@amd.com

Abstract

Video Frame Interpolation (VFI) is a crucial technique in various applications such as slow-motion generation, frame rate conversion, video frame restoration etc. This paper introduces an efficient video frame interpolation framework that aims to strike a favorable balance between efficiency and quality. Our framework follows a general paradigm consisting of a flow estimator and a refinement module, while incorporating carefully designed components. First of all, we adopt depth-wise convolution with large kernels in the flow estimator that simultaneously reduces the parameters and enhances the receptive field for encoding rich context and handling complex motion. Secondly, diverging from a common design for the refinement module with a UNet-structure (encoder-decoder structure), which we find redundant, our decoder-only refinement module directly enhances the result from coarse to fine features, offering a more efficient process. In addition, to address the challenge of handling high-definition frames, we also introduce an innovative HD-aware augmentation strategy during training, leading to consistent enhancement on HD images. Extensive experiments are conducted on diverse datasets, Vimeo90K, UCF101, Xiph and SNU-FILM. The results demonstrate that our approach achieves state-of-the-art performance with clear improvement while requiring much less FLOPs and parameters, reaching to a better spot for balancing efficiency and quality.

Introduction

Video Frame Interpolation (VFI) is an important task in video processing aiming to synthesize intermediate frames between existing frames. VFI can support many downstream tasks such as slow-motion generation (Jiang et al. 2018), video compression (Wu, Singhal, and Krähenbühl 2018), novel-view synthesis (Zhou et al. 2016) and video prediction (Wu, Qiang, and Qifeng 2022).

Various approaches have been proposed to tackle the problem, but it is still challenging to find an optimal balance spot between efficiency and quality. For example, VFIfomer (Lu et al. 2022), a previous state-of-the-art method, achieves PSNR of 36.5 dB on Vimeo90K (Xue et al. 2019a) dataset and outperforms another light-weight method, RIFE (Huang et al. 2022), by a large margin, which obtains PSNR of 35.61 dB. However, to compare the FLOPs of these two models, VFIfomer requires nearly 238 times more FLOPs than RIFE for the same inputs. Obviously, it is still worth ex-

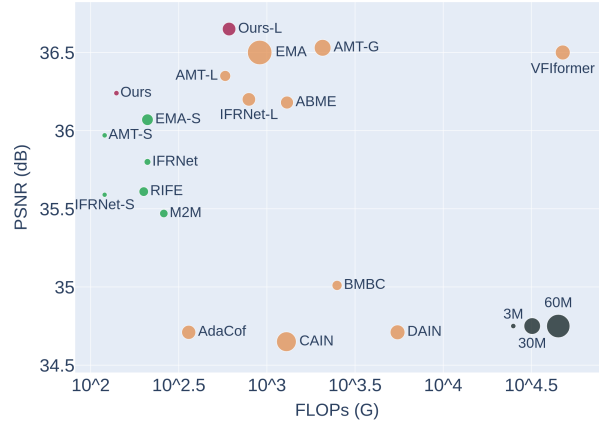


Figure 1: Model comparison on Vimeo90K dataset in terms of FLOPs, PSNR and number of parameters. We roughly divide the models into light-weight and large models, which are shown with green color and orange color respectively. The marker size indicates the number of parameters ranging from 1.4M to 60M. We present two versions of our method, which are colored red. For both light-weight and large groups, we achieve higher PSNR with much lower complexity, reaching to a better balancing spot.

ploring a better trade-off between efficiency and quality. To this end, we propose an efficient VFI framework with carefully designed components that achieves new state-of-the-art performance with clear improvement while requiring much less FLOPs and parameters (as shown in Figure 1).

A widely used form for flow-based methods mainly comprises three parts, a feature extractor, a flow estimator and a refinement module. Our general design philosophy lies in these components:

1) For the feature extractor, since good features not only provide good descriptors for flow estimation but also high-quality evidence for synthesizing frames, we adopt an effective feature extraction process inspired from (Zhang et al. 2023) that has convolution blocks for low-level features and light-transformer blocks for high-level features. Different from the original structure, we do not extract motion fea-

tures for simplicity.

2) For the flow estimator, this module is typically devised to operate in a coarse-to-fine fashion, progressively decoding and refining the flow estimation from an initial coarse approximation to the final output. A key motivation here is that flow estimation usually requires a large field of view (FoV) to capture large motions and rich context. Stacking up multiple 3×3 convolutions might be sufficient for low resolution feature maps, where one spatial feature vector corresponds to a large region in pixel space thanks to big down-sampling scales. However, it might not suffice for high resolution feature maps. Certainly, increasing kernel size or stacking more layers can further enlarge FoV, but it comes at the expense of largely increasing computational cost. Therefore, we propose to use large-kernel depth-wise convolution for those high resolution features. The idea of using large kernels has been investigated for classification, segmentation and object detection (Peng et al. 2017; Ding et al. 2022) but rarely studied in VFI, which motivates us to explore the effectiveness. Our carefully designed flow estimator actually consists of two types of encoders: **Low-res encoder** with normal 3×3 convolution for low resolution features and **High-res encoder** with large-kernel depth-wise convolution for high resolution features.

3) For the refinement module, existing approaches (Zhang et al. 2023; Lu et al. 2022; Huang et al. 2022) commonly adopt a UNet structure (Ronneberger, Fischer, and Brox 2015), featuring an encoder and a decoder. Our key observation is that the feature extractor has already encoded the images into high-quality features on different levels, which is already playing the role of the encoder, and it is not necessary to have another encoder inside of the refinement module. Therefore, diverging from the UNet-like design, we propose to use a decoder-only structure with only three levels that shares the computation and features with the feature extractor and directly estimates the final prediction from previously calculated results, which enables efficient and effective refinement process.

4) Additionally, we notice that our model trained on Standard-Definition (SD) images (e.g. Vimeo90K (Xue et al. 2019a)) does not always give consistent increase of performance on High-Definition (HD) images (e.g. Xiph (Montgomery 1994) and SNU-FILM (Choi et al. 2020)) because of the large size of the image and the limited motion capture ability of the network. To address this issue, we also design a HD-aware augmentation strategy that enables the model’s ability to harness both the down-scaled flow and original flow, which leads to more stable enhancement in handling diverse image resolutions.

Combining above designs, we propose LADDER, a VFI framework with **L**arge-kernel **D**epth-wise Convolution and **D**Ecoder-only **R**efinement. Our method can be trained solely with image triplets without requiring extra supervisions or any pre-trained network (Park, Lee, and Kim 2021; Bao et al. 2019; Niklaus and Liu 2020).

To sum up, our contributions are:

- We propose LADDER for VFI task with a carefully designed flow estimator and refinement module. For the flow estimator, we leverage depth-wise convolution with

large kernels for reduced computational overhead and large field of view. The refinement module takes the form of a decoder-only structure, featuring efficiency and effectiveness.

- We also design a HD-aware augmentation strategy to address the “performance drop” issue on HD images.
- Extensive experiments are conducted on various datasets and our light-weight model, alongside with its more extensive counterpart, achieves state-of-the-art performance with clear improvement while demanding much less FLOPs and parameters.

Related Work

Video Frame Interpolation VFI methods can be roughly categorized into two groups. *Kernel-based methods* represent motion implicitly by using kernels, either a regular (Niklaus, Mai, and Liu 2017a,b; Peleg et al. 2019) or deformable kernel (Lee et al. 2020; Cheng and Chen 2020). The intermediate frame is generated by applying local convolution on the input frames. *Flow-based methods* formulate motion by optical flow, which is a common representation to describe motion between frames. Some approaches (Niklaus and Liu 2018a; Xu et al. 2019) rely on a pre-trained optical flow model for off-the-shelf flow estimation and further predict the intermediate frame by a synthesis network. After that, the concept of task-oriented flow is proposed (Xue et al. 2019b), defining a flow representation tailored for VFI task. Some methods directly predict intermediate flows instead of approximating them (Park et al. 2020; Park, Lee, and Kim 2021; Huang et al. 2022). With the increasing popularity of Transformer, there are works (Zhang et al. 2023; Lu et al. 2022; Li et al. 2023) incorporating Transformer structures into VFI models to model long-range correlation and capture larger motions.

Light-weight VFI. Some approaches aim at designing light-weight models for VFI task, which is also one of our goals in the work. (Huang et al. 2022) propose an efficient VIF pipeline termed IFNet that consists of multiple IFBlocks to estimate flows at different levels. Each block is designed for direct intermediate flow estimation, enabling both effectiveness and simplicity. A privileged distillation scheme is also introduced to further improve the performance. (Kong et al. 2022) present an efficient encoder-decoder based network, named IFRNet, for fast VFI. It employs a decoder to jointly refine the intermediate flow and generate a intermediate feature to produce the final synthesized frame. (Zhang et al. 2023) propose an efficient transformer-based network that leverages both inter-frame appearance features and motion features for VFI.

Method

Given two images ($\mathbf{I}_0, \mathbf{I}_1$) and a timestep t ($0 \leq t \leq 1$), the task is to synthesize an intermediate image as:

$$\hat{\mathbf{I}}_t = \mathcal{M}(\mathbf{I}_0, \mathbf{I}_1), \quad (1)$$

where $\mathcal{M}$ is our model. Since predicting the middle frame is the most common usage case, we focus on $t = 0.5$ here.

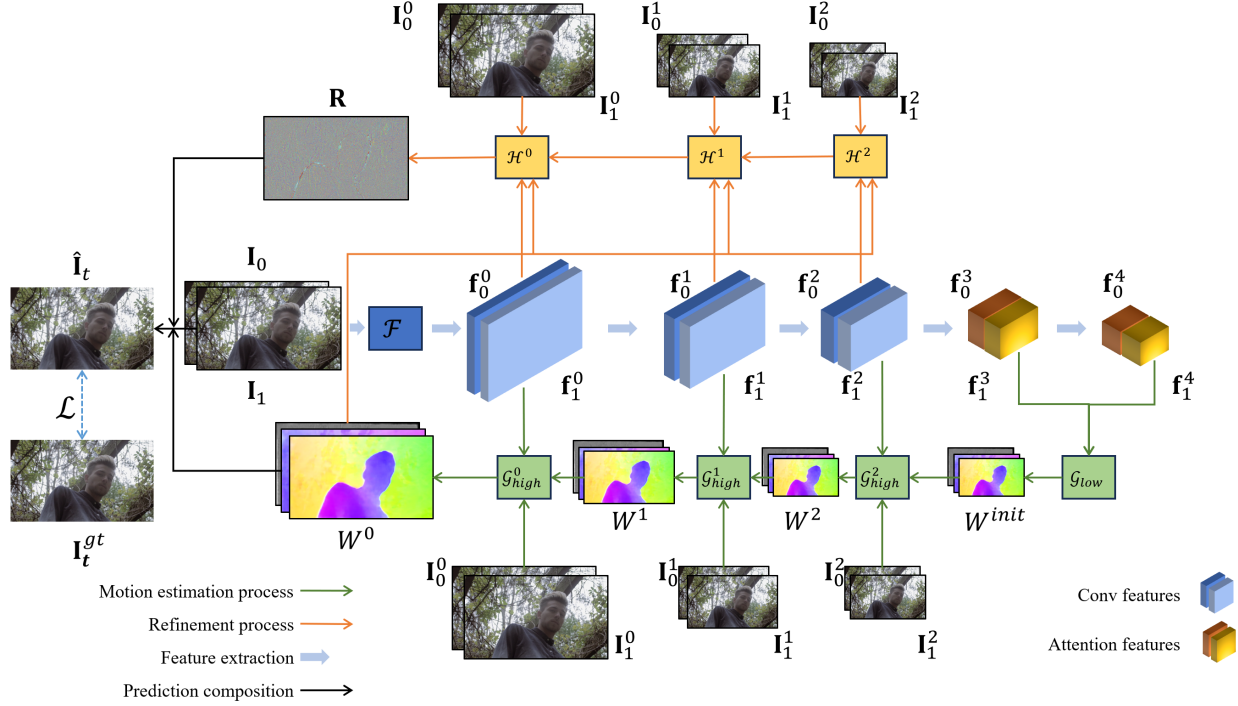


Figure 2: Illustration of our pipeline. Different processes are represented by colored lines. The images are first fed into the feature extractor $\mathcal{F}$ to produce five levels of features, where the first three levels are convolution-based features and the last two are attention-based features. These features are further processed by the flow estimator that consists of a low-res decoder $\mathcal{G}_{low}$ and three high-res decoders $\mathcal{G}_{high}^l$. This process, colored green, generates the motion flows and composition map, W^0 . The high-res features, the input images and the estimated motion are further processed by the refinement module to generate the residual term $\mathbf{R}$, colored orange. The final prediction, indicated by black lines, is the composition of the two input images plus the residual term.

Prediction of the intermediate frame is widely formulated as the composition of the two given images warped by motion flows plus a residual part, which is defined as:

$$\hat{\mathbf{I}}_t = \mathbf{m} \odot \overleftarrow{\mathcal{W}}(\mathbf{I}_0, \mathbf{F}_{t \rightarrow 0}) + (1 - \mathbf{m}) \odot \overleftarrow{\mathcal{W}}(\mathbf{I}_1, \mathbf{F}_{t \rightarrow 1}) + \mathbf{R}, \quad (2)$$

where $\mathbf{m}$ is a composition mask; $\odot$ denotes Hadamard product; $\overleftarrow{\mathcal{W}}$ is the backward warping operation; $\mathbf{F}_{t \rightarrow 0}$ and $\mathbf{F}_{t \rightarrow 1}$ represent motion flows between the intermediate frame and the given images; $\mathbf{R}$ is the residual part.

Our model $\mathcal{M}$ is composed of a feature extractor $\mathcal{F}$, a flow estimator $\mathcal{G}$, and a refinement module $\mathcal{H}$. Figure 2 illustrates the overall pipeline of our method. The input images are first fed to the feature extractor $\mathcal{F}$ to obtain features at five levels. The features are further passed to multiple decoders of the flow estimator $\mathcal{G}$ to estimate the motion flows and the mask. The residual part is then produced by $\mathcal{H}$ using the input images, image features and the estimated flow information. Eventually we obtain the final prediction of the frame $\hat{\mathbf{I}}_t$ by composition. Next we elaborate each part in detail.

Feature Extractor. The role of the feature extractor is to produce high quality features from the raw images, which are used for motion estimation and also image refinement. Pure convolution-based structures are a common choice

(Huang et al. 2022; Lee et al. 2020; Kong et al. 2022), and there are also approaches attempting to incorporate transformer-like structures where the attention mechanism enables long-range interactions within the frame or across frames (Lu et al. 2022; Zhang et al. 2023). We adopt a lightweight attention-based structure from (Zhang et al. 2023). For simplicity, we only borrow the appearance feature extraction part with inter-frame attention and the motion features are not incorporated.

More specifically, for the images with dimensions of $(H \times W)$, we extract five levels of features:

$$\{\mathbf{f}_0^l, \mathbf{f}_1^l, l \in [0, 4]\} = \mathcal{F}(\mathbf{I}_0, \mathbf{I}_1), \quad (3)$$

where each level of features has spatial dimensions of $(\frac{H}{2^l}, \frac{W}{2^l})$. The first three levels of features are produced by convolution blocks and the last two levels are processed by the inter-frame attention blocks.

Flow Estimator. The flow estimator produces the composition mask $\mathbf{m}$ and the motion flows, $\mathbf{F}_{t \rightarrow 0}$ and $\mathbf{F}_{t \rightarrow 1}$. However, directly computing the full resolution flows is challenging due to complex motions and diversity of scenes, therefore flows are usually estimated progressively in a coarse-to-fine fashion (Reda et al. 2022; Huang et al. 2022; Kong et al. 2022). In this process, a sufficient big FoV is critical

to encode rich context information and capture complex motions. For low-resolution features, stacking up multiple 3×3 convolutions might suffice, but for high-resolution features, it would be not sufficient. Simply enlarging the kernel size or adding more convolutions might not be an efficient solution for its largely increased computational cost. We propose to incorporate large-kernel depth-wise convolution to enable large FoV and at the same time reduce overhead. The idea of using large kernels has been investigated in other tasks such as image segmentation, detection and classification (Peng et al. 2017; Ding et al. 2022), but has rarely been studied for VFI, which motivates us to exploit the effectiveness. Depth-wise convolution has advantages in this specific case. For feature maps with 64 channels, without changing the channel number, compared with a normal 3×3 convolution layer, a depth-wise convolution layer with kernel size $k = 15$ consumes only half of FLOPs and uses half of parameters, but greatly enlarges FoV. Please also refer to the ablation experiments for more details.

Back to the design of our flow estimator, instead of using a uniform decoder block across all levels of features, we introduce a novel design featuring two types of decoders, low-res decoder $\mathcal{G}_{low}$ and high-res decoder $\mathcal{G}_{high}$, for efficient flow estimation. For low-resolution features, the low-res decoder, equipped with normal 3×3 convolution, is sufficiently effective, whereas for high-resolution features, the high-res decoders benefit from large-kernel depth-wise convolutions.

More specifically, the low-res decoder receives features, $\{\mathbf{f}_0^4, \mathbf{f}_1^4, \mathbf{f}_0^3, \mathbf{f}_1^3\}$, as inputs and generate a good initial estimation of the flows and composition mask, defined as:

$$\{\mathbf{F}_{t \rightarrow 0}^{init}, \mathbf{F}_{t \rightarrow 1}^{init}, \mathbf{m}^{init}\} = \mathcal{G}_{low}(\mathbf{f}_0^4, \mathbf{f}_1^4, \mathbf{f}_0^3, \mathbf{f}_1^3) \quad (4)$$

Once the initial estimation is generated, we utilize three high-res decoders, $\mathcal{G}_{high}^l$ with $l \in [0, 1, 2]$, to progressively refine the results, which is defined as:

$$\Delta \mathbf{W}^l = \mathcal{G}_{high}^l(\tilde{\mathcal{W}}(\mathbf{f}_0^l, \mathbf{F}_{t \rightarrow 0}^{l+1}), \tilde{\mathcal{W}}(\mathbf{f}_1^l, \mathbf{F}_{t \rightarrow 1}^{l+1}), \mathbf{I}_0^l, \mathbf{I}_1^l, \tilde{\mathcal{W}}(\mathbf{I}_0^l, \mathbf{F}_{t \rightarrow 0}^{l+1}), \tilde{\mathcal{W}}(\mathbf{I}_1^l, \mathbf{F}_{t \rightarrow 1}^{l+1}), \mathbf{W}^{l+1}) \quad (5)$$

$$\mathbf{W}^l = (\Delta \mathbf{W}^l + \mathbf{W}^{l+1})_{\times 2} = \{\mathbf{F}_{t \rightarrow 0}^l, \mathbf{F}_{t \rightarrow 1}^l, \mathbf{m}^l\}, \quad (6)$$

where $\mathbf{W}^l$ represents the warping information, including motion flows and composition map; $(\cdot)_{\times 2}$ is a bi-linear up-sampler; $\mathbf{I}_1^l$ and $\mathbf{I}_0^l$ are images down-sampled to level l . $\mathbf{W}^0$ is the final output, and the bi-linear up-sampler is not applied.

For flow estimation, there are several design options in terms of the inputs. In (Reda et al. 2022), the decoder uses both the original image features and the warped features, where it requires a large channel width. In (Huang et al. 2022), the flow estimation is based on a so-called IFBlock, whose inputs are the down-scaled images, the warped images, and the flow information. Its input only takes 17 channels, which is a light-weight design. However it relies only on the raw image information and might not benefit from the high quality features provided by the extractor. Our design takes the advantage of both the rich features and raw image information.

Refinement Module.

The composition of the warped input images provides an initial prediction of the intermediate frame. However, the prediction highly relies on the given pixel information and motion quality. If the model fails to find proper information due to motion errors or occlusions, it will produce unsatisfactory results. To overcome this limitation, a refinement module is usually adopted to compute a residual $\mathbf{R}$ that compensates for the errors, as stated in Equation 2.

A common design for this refinement module is a UNet (Ronneberger, Fischer, and Brox 2015) structure that has an encoder with down-sampling blocks and a decoder with up-sampling blocks (Huang et al. 2022; Zhang et al. 2023; Lu et al. 2022). In (Park, Lee, and Kim 2021), a GridNet (Foufure et al. 2017) is used to perform refinement, which not only contains down-sampling and up-sampling blocks but also employs lateral blocks. We find this redundant in our framework since the feature extractor has already encoded frames into high-quality features on different levels, which make it unnecessary to employ another encoder in the refinement module. Therefore, we simplify this process by adopting a decoder-only structure, leading to effective and efficient refinement.

There are only three blocks in the refinement module $\mathcal{H}$ corresponding to high resolution features $\{\mathbf{f}_0^l, \mathbf{f}_1^l, l \in [0, 2]\}$. We find the features with inter-frame interactions $\{\mathbf{f}_0^l, \mathbf{f}_1^l, l \in [3, 4]\}$ do not provide boost in this stage, so we end up with this 3-level light-weight structure. Please refer to the ablation study where we compare different configurations of the refinement module.

For each level, it is defined as:

$$\mathbf{O}^l = \mathcal{H}^l(\tilde{\mathcal{W}}(\mathbf{f}_0^l), \tilde{\mathcal{W}}(\mathbf{f}_1^l), \mathbf{I}_0^l, \mathbf{I}_1^l, \tilde{\mathcal{W}}(\mathbf{I}_0^l), \tilde{\mathcal{W}}(\mathbf{I}_1^l), \tilde{\mathbf{W}}^l, \mathbf{O}^{l+1}), \quad (7)$$

where $\mathbf{I}_0^l$ and $\mathbf{I}_1^l$ are scaled images, $\tilde{\mathcal{W}}(\cdot)$ is a simplified representation for warping the image or features using $\mathbf{W}^0$ down-scaled to the corresponding size. $\mathbf{O}^l$ is the output for level l . For $l = 2$, $\mathbf{O}^{l+1}$ is “None”, and for $l = 0$, $\mathbf{O}^l$ is the final residual $\mathbf{R}$.

HD-aware Augmentation

A typical evaluation routine for VFI methods is training on Vimeo 90K dataset (Xue et al. 2019a), and evaluating on various datasets. We unavoidably face domain gaps between different sources of data. One of the issues is image resolution. Vimeo 90K dataset contains images with fixed size of 448×256 , while datasets such as Xiph and SNU-FILM (Montgomery 1994; Choi et al. 2020) feature HD images with size up to 1280×720 or 2048×1080 . Generally applying the model directly on full resolution HD images would not give satisfactory results since objects and motions are scaled up significantly compared to the SD images.

A widely used strategy is to use the down-scaled images to estimate motion flows and scale up the flows back to the original resolution and perform image synthesis (Kong et al. 2022; Li et al. 2023; Zhang et al. 2023; Lu et al. 2022). However, we find this post-training method not working consistently well, e.g. performance increasing on SD datasets does

not lead to corresponding improvement on HD datasets. In order to further stabilize the model performance on HD images, we incorporate this "synthesizing image with low-res flows" mechanism into the training process, enabling a more consistent ability to synthesize frames with both original and low-res flows, which we call HD-aware augmentation.

More specifically, after the standard training process, we further fine-tune the model by 5 epochs, where we randomly make the model utilize either the original flows estimated from the original size, or the down-scaled flows estimated from the down-sampled images. For simplicity, we only consider down-sampling scale of 0.5.

Training Objectives

Our model is trained solely on ground-truth frames without requiring any other pre-trained networks or extra supervisions. Our model is supervised by three terms.

$$\mathcal{L} = \lambda_{ch}\mathcal{L}_{ch} + \lambda_{lap}\mathcal{L}_{lap} + \lambda_f\mathcal{L}_f \quad (8)$$

The first term is a image reconstruction loss defined as:

$$\mathcal{L}_{ch} = \sqrt{(\hat{\mathbf{I}}_t - \mathbf{I}_t^{gt})^2 + \epsilon^2}, \quad (9)$$

where ϵ is a small constant set to $1e^{-6}$. It is called Charbonnier penalty function (Charbonnier et al. 1994) that serves as a better reconstruction loss than the pure L1 loss.

The second term is a Laplacian pyramid loss (Niklaus and Liu 2018b) defined as:

$$\mathbf{L}_{lap} = \sum_{i=1}^5 2^{i-1} \left\| L^i(\hat{\mathbf{I}}_t) - L^i(\mathbf{I}_t^{gt}) \right\|_1 \quad (10)$$

The $\mathbf{L}_{lap}$ is capable of retaining fine details, but focusing less on the low-frequency content such as color information. Therefore $\mathcal{L}_{ch}$ combined with $\mathbf{L}_{lap}$ provides better results.

We also add another frequency loss term $\mathcal{L}_f$, which is applied to image super resolution task (Fuoli, Gool, and Timofte 2021). The loss is defined as:

$$\mathcal{L}_f = \frac{1}{2} \left\| \mathcal{T}_{fft}^{|\cdot|}(\hat{\mathbf{I}}_t) - \mathcal{T}_{fft}^{|\cdot|}(\mathbf{I}_t^{gt}) \right\|_1 + \frac{1}{2} \left\| \mathcal{T}_{fft}^{\angle}(\hat{\mathbf{I}}_t) - \mathcal{T}_{fft}^{\angle}(\mathbf{I}_t^{gt}) \right\|_1, \quad (11)$$

where $\mathcal{T}_{fft}^{|\cdot|}(\cdot)$ and $\mathcal{T}_{fft}^{\angle}(\cdot)$ are amplitude and phase of the frequency components calculated by fast Fourier transform (FFT). This loss supervises the prediction in the frequency domain, providing supervision from a global holistic perspective in addition to the local spatial information.

Experiments

Implementation Details

Model Configuration. We use C to describe the base-width of feature channels. For the feature extractor, feature channels for each level are $[C, 2C, 4C, 8C, 16C]$. For the decoders, the low-res decoder has internal feature number of $2C$. The three high-res decoders work on feature channels of $4C$, $2C$ and C respectively. For the image refinement module, the internal channel widths for the three blocks are

$8C$, $4C$ and $2C$ respectively. We present two versions of our framework. For a light-weight one, we set $C = 16$ and use two attention blocks for each level. For a relatively larger model, we have $C = 32$ and assign four attention blocks.

Datasets. We conduct experiments on four widely used datasets, Vimeo90K (Xue et al. 2019a), UCF101 (Soomro, Zamir, and Shah 2012), SNU-FILM (Choi et al. 2020) and Xiph (Montgomery 1994), covering SD and HD scenarios and rich types of scenes.

Training details. The model is trained solely on Vimeo-90 training set and evaluated on all four datasets. Images are randomly cropped to a patch of 256. In addition, we apply data augmentation methods: random vertical and horizontal flipping, random scale with $[1, 2]$, random rotation with $[-45^\circ, 45^\circ]$ and triplet reversing. The model is optimized by AdamW (Loshchilov and Hutter 2019) with batch size of 32, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of $1e - 4$. The learning rate is decreased by the cosine annealing schedule from $2e - 4$ to $2e - 5$.

We adopt a two-stage training scheme, training the flow estimator first and continuing training the whole framework by including the refinement module. Each stage is performed with the same routine. The only difference is that the first stage only considers composition of the two warped images, whereas the second stage involves the residual prediction.

Comparisons with Other Methods

We use PSNR and SSIM to perform quantitative comparison between different methods, including M2M (Hu et al. 2022), IFRNet (Kong et al. 2022), RIFE (Huang et al. 2022), AMT (Li et al. 2023), EMA (Zhang et al. 2023), ToFlow(Xue et al. 2019b), SepConv (Niklaus, Mai, and Liu 2017b), AdaCof (Lee et al. 2020), CAIN (Choi et al. 2020), DAIN (Bao et al. 2019), BMBC (Park et al. 2020), SoftSplat (Niklaus and Liu 2020), ABME (Park, Lee, and Kim 2021), and VFIfomer (Lu et al. 2022).

Table 1 shows the experimental results and comparison between methods. We divide the models into two groups based on computational overhead (FLOPs) and number of parameters. For the light-weight models, ours obtains the best PSNR for all the entries. Besides, compared with the overall second best method, EMA-S, our model requires 33% less FLOPs and 79% less parameters. For large models, we also achieve clear improvement over the previous best models for almost all the entries except for SSIM on Vimeo90k and PSNR on UCF101, where ours is very close to the best one, differing only in the last decimal digit. Additionally, our large model again shows clear advantage in terms of FLOPs and model size, requiring 70% and 35% less FLOPs and parameters than the second best, AMT-G. Figure 3 illustrates some visual results for a qualitative comparison.

Ablation Study

We further conduct experiments with different configurations to verify effectiveness of the components, which include loss functions, block configurations for the flow estimator, structures of the refinement module, pre-training and HD augmentation.

Method	Vimeo90k	UCF101	SNU-FILM				Xiph		FLOPs (T)	Params (M)
			Easy	Medium	Hard	Extreme	2K	4K		
M2M	35.47/0.9778	35.28/0.9694	39.66/0.9904	35.74/0.9794	30.30/0.9360	25.08/0.8604	36.44/ 0.943	33.92/0.899	0.26	7.6
IFRNet-S	35.59/0.9786	35.28/0.9691	39.96/0.9905	35.92/0.9795	30.36/0.9357	25.05/0.8582	35.87/0.936	33.80/0.891	0.12	2.8
RIFE	35.61/0.9779	35.28/0.9690	39.80/0.9903	35.76/0.9787	30.36/0.9351	25.27/0.8601	36.19/0.938	33.76/0.894	0.20	9.8
IFRNet	35.80/0.9794	35.29/0.9693	<u>40.03/0.9905</u>	35.94/0.9793	30.41/0.9358	25.05/0.8587	36.00/0.936	33.99/0.893	0.21	5.0
AMT-S †	35.97/ <u>0.9800</u>	<u>35.35/0.9697</u>	39.95/0.9905	<u>35.98/0.9796</u>	<u>30.60/0.9369</u>	25.30/0.8625	36.14/0.940	<u>34.32/0.902</u>	0.12	3.0
EMA-S	<u>36.07/0.9797</u>	35.34/0.9696	39.81/0.9906	35.88/0.9795	<u>30.69/0.9375</u>	<u>25.47/0.8632</u>	<u>36.55/0.942</u>	<u>34.25/0.902</u>	0.21	14.5
Ours	36.24/0.9804	35.40/0.9698	40.16/0.9908	36.14/0.9799	30.77/0.9375	25.54/0.8633	36.60/0.943	34.35/0.903	0.14	3.1
ToFlow	33.73/0.9682	34.58/0.9667	39.08/0.9890	34.39/0.9740	28.44/0.9180	23.39/0.8310	33.93/0.922	30.74/0.856	0.62	1.4
SepConv	33.79/0.9702	34.78/0.9669	39.41/0.9900	34.97/0.9762	29.36/0.9253	24.31/0.8448	34.77/0.929	32.06/0.880	0.38	21.7
AdaCoF	34.47/0.9730	34.90/0.9680	39.80/0.9900	35.05/0.9754	29.46/0.9244	24.31/0.8439	34.86/0.928	31.68/0.870	0.36	21.8
CAIN	34.65/0.9730	34.91/0.9690	39.89/0.9900	35.61/0.9776	29.90/0.9292	24.78/0.8507	35.21/0.937	32.56/0.901	1.29	42.8
DAIN	34.71/0.9756	34.99/0.9683	39.73/0.9902	35.46/0.9780	30.17/0.9335	25.09/0.8584	35.95/0.940	33.49/0.895	5.51	24.0
BMBC	35.01/0.9764	35.15/0.9689	39.90/0.9902	35.31/0.9774	29.33/0.9270	23.92/0.8432	32.82/0.928	31.19/0.880	2.50	11.0
SoftSplat	36.10/0.9802	35.39/0.9697	39.88/0.9897	35.68/0.9772	30.19/0.9312	24.83/0.8500	<u>36.62/0.944</u>	33.60/0.901	1.00	7.7
ABME	36.18/0.9805	35.38/0.9698	39.59/0.9901	35.77/0.9789	30.58/0.9364	25.42/0.8639	<u>36.53/0.944</u>	33.73/0.901	1.30	18.1
IFRNet-L	36.20/0.9808	35.42/0.9698	40.10/0.9906	<u>36.12/0.9797</u>	30.63/0.9368	25.27/0.8609	36.21/0.937	34.25/0.895	0.79	19.7
AMT-L †	36.35/0.9813	35.42/0.9699	39.95/0.9904	36.09/0.9797	30.75/0.9377	25.41/0.8633	36.31/0.941	34.51/0.904	0.58	12.9
EMA-L *	36.50/0.9800	<u>35.42/0.9700</u>	39.58/0.9890	35.86/0.9790	30.80/0.9380	<u>25.59/0.8640</u>	<u>36.74/0.944</u>	34.55/0.906	0.91	65.6
VFIformer	36.50/0.9816	<u>35.43/0.9700</u>	<u>40.13/0.9907</u>	<u>36.09/0.9799</u>	30.67/0.9378	<u>25.43/0.8643</u>	<u>36.82/0.946</u>	<u>34.43/0.907</u>	47.71	24.1
AMT-G †	<u>36.53/0.9819</u>	<u>35.45/0.9700</u>	39.88/0.9905	<u>36.12/0.9798</u>	<u>30.78/0.9379</u>	25.43/0.8640	36.42/0.942	<u>34.65/0.905</u>	2.07	30.6
Ours-L	36.65/0.9818	35.44/0.9702	40.17/0.9909	36.22/0.9803	30.87/0.9384	25.71/0.8658	36.89/0.946	34.88/0.908	0.61	20.0

Table 1: Quantitative comparison between VFI methods. We divide the methods into two groups based on the FLOPs and model size. For each group, the best result is shown in **bold** and the second best is underlined. * indicates the result without test-time augmentation for fair comparison. †represents our duplicated results using the official code.

$\mathcal{L}_{ch}$	$\mathcal{L}_{lap}$	$\mathcal{L}_f$	PSNR (dB)
✓	✗	✗	36.1
✗	✓	✗	15.04
✓	✓	✗	36.17
✓	✓	✓	36.24

Table 2: Ablation study on different losses.

Block conf. in $\mathcal{G}_{high}^l$	PSNR (dB)	FLOPs (G)	Params (M)
Conv k=[3, 3, 3]	36.11	150.74	3.21
DSCConv k=[5, 5, 5]	36.12	127.67	3.03
DSCConv k=[7, 7, 7]	36.13	130.65	3.05
DSCConv k=[7, 15, 15]	36.24	139.73	3.08

Table 3: Ablation study on high-res decoder settings.

Loss functions. We study the effect of the three loss functions in the training. As described in Table 2, training the model purely with $\mathcal{L}_{ch}$ gives us a baseline score of 36.1 dB on PSNR. The model trained only with $\mathcal{L}_{lap}$ does not produce a plausible score, only 15.04dB, but interestingly the training loss decreases normally. The reason is that $\mathcal{L}_{lap}$ is capable of capturing high-frequency details but might lose some low-frequency information such as color. Therefore $\mathcal{L}_{lap}$ and $\mathcal{L}_{ch}$ are a good combination for such case, reaching PSNR of 36.17dB. Furthermore, with the help of $\mathcal{L}_f$, the

$\mathcal{H}$ conf.	PSNR (dB)	FLOPs (G)	Params (M)
UNet	36.21	207.27	14.49
5L decoder-only	36.20	179.57	9.49
4L decoder-only	36.22	159.68	4.37
3L decoder-only	36.24	139.73	3.08
2L decoder-only	36.13	119.63	2.76

Table 4: Results of different structures of refinement module.

Training config in $\mathcal{G}_{high}^l$	PSNR (dB)
One-stage training	36.01
One-stage training with extra sup.	36.05
Two-stage training	36.24

Table 5: Ablation study on pre-training.

performance is further enhanced to 36.24 dB.

Block configurations for high-res decoders. We investigate different settings for the high-res decoders. As shown in Table 3, using only three 3×3 convolution layers only achieves PSNR of 36.11dB. By replacing all the normal convolutions to the depth-wise separable convolutions, we get slightly better performance with reduced FLOPs (127.67 vs 150.74) and parameters (3.03 vs 3.21). Increasing all kernel sizes to 7 further slightly improves the score with limited overhead. Our final design adopts the setting of [7, 15,

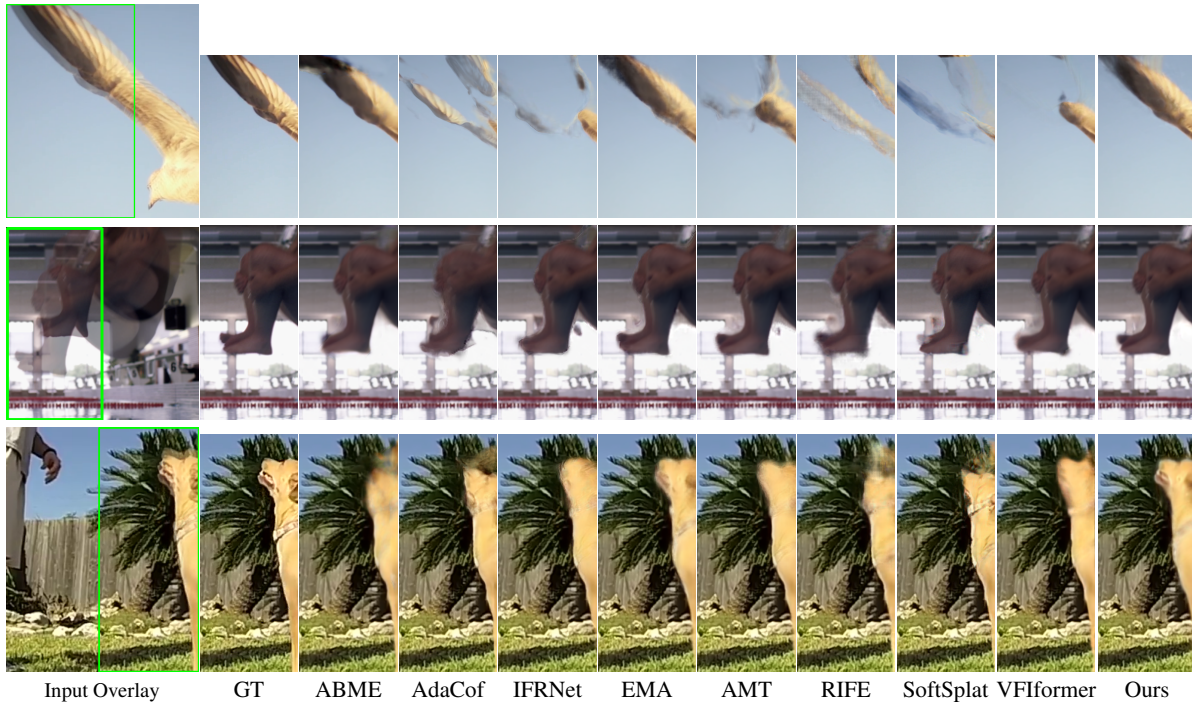


Figure 3: Qualitative comparison.

Method	Easy	Medium	Hard	Extreme
With orig. flow	40.11	35.86	30.23	25.02
With ds. flow	39.86	35.98	30.71	25.5
HD-aware aug	40.16	36.14	30.77	25.54

Table 6: Ablation study on HD augmentation.

15], which boosts the PSNR by 0.13dB compared with the normal-convolution version, while requires less FLOPs (7% less) and parameters (4% less).

Structures of refinement module. To verify the effectiveness of our decoder-only refinement module, we conduct experiments on various structure settings, shown in Table 4. XL means how many levels of features we feed to the refinement module. Compared with our optimal structure with 3 levels, the UNet structure does not give better performance while largely increasing FLOPs and model size. For decoder-only structures, using inter-frame attention features (4L and 5L structures) does not boost the performance and on the contrary, using only 2 levels (2L structure) of convolution features is clearly not sufficient. The ablation study indicates that fully utilizing all three convolution-based features maximizes the performance.

Pre-training. We adopt an effective two-stage training scheme in our framework, where the flow-estimator is firstly trained to predict plausible flow information and the refinement is further added for the second stage training. We find the two-stage training practically useful and we study different training strategies in Table 5. Directly training the whole

framework from scratch only achieves PSNR of 36.01dB. By applying extra supervision to the intermediate flow estimation process, we obtain slightly better performance of 36.05dB. Taking the two-stage training boosts the performance to 36.24dB.

HD-aware augmentation. We also investigate effectiveness of our HD-aware augmentation strategy. As shown in Table 6, the model performance with the original flow (calculated by original inputs) is acceptable on the Easy set but deteriorates on the remaining three sets. In Contrast, the utilization of down-sampled flow (calculated by the down-sampled inputs) improves the performance on those three sets but fails to retain the score on the Easy set. Applying our HD-aware augmentation strategy yields consistent improvement over all four sets.

Conclusion

In this paper, we introduce an efficient framework for VFI with carefully designed components, which achieves State-of-the-art performance with much less FLOPs and parameters. More specifically, we first propose a flow estimator using depth-wise convolution with large kernels, enabling effective handling of large motions with reduced overhead. Furthermore, our proposed refinement module, featuring a decoder-only structure, provides an efficient process of refinement. Additionally, we introduce an HD-aware augmentation strategy to address "performance drop" issue on HD images. We hope this would establish a solid baseline for future research of the community.

References

- Bao, W.; Lai, W.-S.; Ma, C.; Zhang, X.; Gao, Z.; and Yang, M.-H. 2019. Depth-Aware Video Frame Interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; and Barlaud, M. 1994. Two deterministic half-quadratic regularization algorithms for computed imaging. In *ICIP*, volume 2, 168–172 vol.2.
- Cheng, X.; and Chen, Z. 2020. Video Frame Interpolation via Deformable Separable Convolution. In *AAAI*.
- Choi, M.; Kim, H.; Han, B.; Xu, N.; and Lee, K. M. 2020. Channel Attention Is All You Need for Video Frame Interpolation. In *AAAI*.
- Ding, X.; Zhang, X.; Zhou, Y.; Han, J.; Ding, G.; and Sun, J. 2022. Scaling Up Your Kernels to 31x31: Revisiting Large Kernel Design in CNNs. *arXiv preprint arXiv:2203.06717*.
- Fourure, D.; Emonet, R.; Fromont, E.; Muselet, D.; Trémeau, A.; and Wolf, C. 2017. Residual Conv-Deconv Grid Network for Semantic Segmentation. In *Proceedings of the British Machine Vision Conference, 2017*.
- Fuoli, D.; Gool, L. V.; and Timofte, R. 2021. Fourier space losses for efficient perceptual image super-resolution. In *ICCV*.
- Hu, P.; Niklaus, S.; Sclaroff, S.; and Saenko, K. 2022. Many-to-many Splatting for Efficient Video Frame Interpolation.
- Huang, Z.; Zhang, T.; Heng, W.; Shi, B.; and Zhou, S. 2022. Real-Time Intermediate Flow Estimation for Video Frame Interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Jiang, H.; Sun, D.; Jampani, V.; Yang, M.-H.; Learned-Miller, E.; and Kautz, J. 2018. Super slo-mo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*.
- Kong, L.; Jiang, B.; Luo, D.; Chu, W.; Huang, X.; Tai, Y.; Wang, C.; and Yang, J. 2022. IFRNet: Intermediate Feature Refine Network for Efficient Frame Interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Lee, H.; Kim, T.; Chung, T.-y.; Pak, D.; Ban, Y.; and Lee, S. 2020. AdaCoF: Adaptive Collaboration of Flows for Video Frame Interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Li, Z.; Zhu, Z.-L.; Han, L.-H.; Hou, Q.; Guo, C.-L.; and Cheng, M.-M. 2023. AMT: All-Pairs Multi-Field Transforms for Efficient Frame Interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Loshchilov, I.; and Hutte, F. 2019. Decoupled weight decay regularization. In *ICLR*.
- Lu, L.; Wu, R.; Lin, H.; Lu, J.; and Jia, J. 2022. Video Frame Interpolation with Transformer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Montgomery, C. 1994. Xiph.org video test media (derf's collection). In *Online*, <https://media.xiph.org/video/derf/>.
- Niklaus, S.; and Liu, F. 2018a. Context-aware Synthesis for Video Frame Interpolation. In *CVPR*.
- Niklaus, S.; and Liu, F. 2018b. Context-aware synthesis for video frame interpolation. In *CVPR*.
- Niklaus, S.; and Liu, F. 2020. Softmax Splatting for Video Frame Interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Niklaus, S.; Mai, L.; and Liu, F. 2017a. Video Frame Interpolation via Adaptive Convolution. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Niklaus, S.; Mai, L.; and Liu, F. 2017b. Video Frame Interpolation via Adaptive Separable Convolution. In *IEEE International Conference on Computer Vision*.
- Park, J.; Ko, K.; Lee, C.; and Kim, C.-S. 2020. BMBC: Bilateral Motion Estimation with Bilateral Cost Volume for Video Interpolation. In *European Conference on Computer Vision*.
- Park, J.; Lee, C.; and Kim, C.-S. 2021. Asymmetric Bilateral Motion Estimation for Video Frame Interpolation. In *International Conference on Computer Vision*.
- Peleg, T.; Szeliski, P.; Sabo, D.; and Sendik, O. 2019. Im-net for high resolution video frame interpolation. In *CVPR*.
- Peng, C.; Zhang, X.; Yu, G.; Luo, G.; and Sun, J. 2017. Large Kernel Matters – Improve Semantic Segmentation by Global Convolutional Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Reda, F.; Kontkanen, J.; Tabellion, E.; Sun, D.; Pantofaru, C.; and Curless, B. 2022. FILM: Frame Interpolation for Large Motion. In *European Conference on Computer Vision (ECCV)*.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N.; Hornegger, J.; III, W. M. W.; and Frangi, A. F., eds., *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, 234–241. Springer.
- Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *CoRR*, abs/1212.0402.
- Wu, C.-Y.; Singhal, N.; and Krähenbühl, P. 2018. Video Compression through Image Interpolation. In *ECCV*.
- Wu, Y.; Qiang, W.; and Qifeng, C. 2022. Optimizing Video Prediction via Video Frame Interpolation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Xu, X.; Siyao, L.; Sun, W.; Yin, Q.; and Yang, M.-H. 2019. Quadratic video interpolation. In *NeurIPS*.
- Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019a. Video Enhancement with Task-Oriented Flow. *International Journal of Computer Vision (IJCV)*, 127(8): 1106–1125.
- Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019b. Video Enhancement with Task-Oriented Flow. *International Journal of Computer Vision (IJCV)*, 127(8): 1106–1125.

Zhang, G.; Zhu, Y.; Wang, H.; Chen, Y.; Wu, G.; and Wang, L. 2023. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5682–5692.

Zhou, T.; Tulsiani, S.; Sun, W.; Malik, J.; and Efros, A. A. 2016. View Synthesis by Appearance Flow. In *European Conference on Computer Vision*.