High-Resolution Frame Interpolation with Patch-based Cascaded Diffusion

Junhwa Hur*, Charles Herrmann*, Saurabh Saxena, Janne Kontkanen, Wei-Sheng Lai,
Yichang Shih, Michael Rubinstein, David J. Fleet†, Deqing Sun

Google

Abstract

Despite the recent progress, existing frame interpolation methods still struggle with processing extremely high resolution input and handling challenging cases such as repetitive textures, thin objects, and large motion. To address these issues, we introduce a *patch-based* cascaded pixel diffusion model for high resolution frame interpolation, HiFI, that excels in these scenarios while achieving competitive performance on standard benchmarks. Cascades, which generate a series of images from low to high resolution, can help significantly with large or complex motion that require both global context for a coarse solution and detailed context for high resolution output. However, contrary to prior work on cascaded diffusion models which perform diffusion on increasingly large resolutions, we use a single model that always performs diffusion at the same resolution and upsamples by processing patches of the inputs and the prior solution. At inference time, this drastically reduces memory usage and allows a single model, solving both frame interpolation (base model's task) and spatial up-sampling, saving training cost as well. HiFI excels at high-resolution images and complex repeated textures that require global context, achieving comparable or state-of-the-art performance on various benchmarks (Vimeo, Xiph, X-Test, and SEPE-8K). We further introduce a new dataset, LaMoR, that focuses on particularly challenging cases, and HiFI significantly outperforms other baselines.

Project page — https://hifi-diffusion.github.io

Introduction

In a short amount of time, smartphone cameras have become both ubiquitous and significantly higher quality, capturing spatially higher resolution images and videos. However, the temporal resolution—*i.e.* video frame rate—of captured videos has lagged behind the spatial resolution, due to a combination of computational and memory costs and limited exposure time. The conflict between increased user interest in creative video content and technical limitations for capturing high frame-rate video has increased interest in techniques for high-resolution frame interpolation, which enables the synthesis of new frames between existing ones to enhance a video's frame rate. Despite the progress, the latest techniques struggle in the high resolution setting, where challenging cases such as repetitive textures, detailed or thin objects become more common place.

Existing methods often design models using strong domain knowledge, *e.g.*, correspondence matching (Ilg et al. 2017; Sun et al. 2018; Teed and Deng 2020) and synthesis based on warping (Jiang et al. 2018; Niklaus and Liu 2020; Park, Lee, and Kim 2021; Xue et al. 2019). Domain knowledge enables small models to perform well when trained on a small amount of data but may restrict their capabilities. For example, when motion cues are incorporated into the model, the final quality are bounded by the accuracy of the motion. This is particularly evident on high resolution inputs with large motion, repetitive texture, and thin structures, where motion estimation often struggles (see Fig. 1).

To address these challenges, we advocate a domain-agnostic diffusion approach, relying on model capacity and training data at scale for performance gains and generalization. Some recent work have explored diffusion for frame interpolation but towards generative aspect, *e.g.* better perceptual quality (Danier, Zhang, and Bull 2024) or complex and non-linear motion (Jain et al. 2024) between two frames very further apart in time. Their performance, however, falls behind in the classical setting which predicts an intermediate frame and evaluates its fidelity to the ground truth using standard metrics, *e.g.*, PSNR or SSIM.

We instead introduce a *patch-based* cascaded pixel diffusion approach for **Hi**gh resolution **F**rame **I**nterpolation, dubbed HiFI. HiFI generalizes across diverse resolutions up to 8K images, a wide range of scene motions, and a broad spectrum of challenging scenes. The diffusion framework allows us to scale both the model capacity and data size. We also show that our model can effectively utilize large-scale video datasets. While cascades offer significant benefits for processing diverse input resolutions with different levels of motion, standard cascades, which denoise the entire high-resolution image, often struggle with memory issues at very high resolutions such as 8K. To save memory during inference, we propose a new *patch-based* cascade for frame interpolation, which always denoises the same resolution but is applied to patches of high resolution frames. This also allows us to use one model for both base and super-resolution

[*] These authors contributed equally.

[†] DF is also affiliated with the University of Toronto and the Vector Institute.

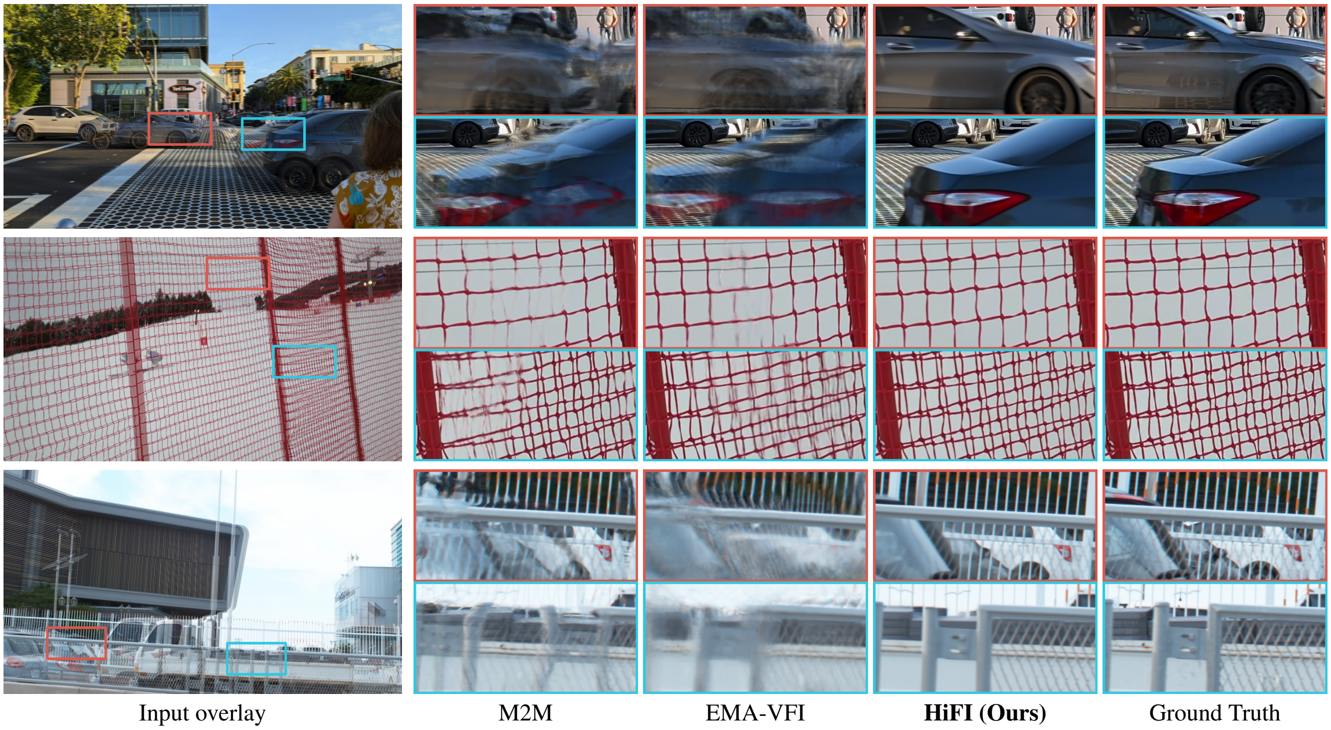

Figure 1: **Qualitative comparison on challenging cases** on our proposed LaMoR dataset (rows 1 and 2) and X-TEST (row 3). For challenging cases, such as large motion or repetitive textures, the proposed HiFI substantially outperforms other baselines.

tasks, saving time for training separate models for both tasks and disk space at inference time.

The proposed HiFI method achieves state-of-the-art accuracy on challenging high-resolution public benchmark datasets, Xiph (Niklaus and Liu 2020), X-TEST (Sim, Oh, and Kim 2021) and SEPE (Al Shoura et al. 2023), and demonstrates strong performance on challenging corner cases, *e.g.*, repetitive textures and large motion. We also introduce a new evaluation dataset, Large Motion and Repetitive texture (LaMoR), which specifically highlights these challenging cases and demonstrate that HiFI significantly outperforms existing baselines.

Related Work

Domain-specific architecture for interpolation. Motion-based approaches synthesize intermediate frames using estimated bi-directional optical flow between two nearby frames. These methods employ forward splatting (Hu et al. 2022; Jin et al. 2023; Niklaus and Liu 2018, 2020) or backward warping (Huang et al. 2022; Jiang et al. 2018; Park, Lee, and Kim 2021; Park et al. 2020; Sim, Oh, and Kim 2021), followed by a refinement module that improves visual quality. Performance is often bounded by motion estimation accuracy, as inaccuracies in the motion cause artifacts during the splatting or warping process. As a result, they struggle on inputs for which optical flow estimation is problematic, *e.g.*, large motion, occlusion, and thin objects.

Phase-based approaches (Meyer et al. 2015, 2018) propose to estimate an intermediate frame in a phase-based representation instead of the conventional pixel domain. Kernel-based approaches (Cheng and Chen 2020; Lee et al. 2020; Niklaus, Mai, and Wang 2021; Niklaus, Mai, and Liu 2017a,b) present simple single-stage formulations that estimate per-pixel $n \times n$ kernels and synthesize the intermediate frame using convolution on input patches. Both approaches avoid reliance on motion estimator, but they do not usually perform well on high resolution input with large motion, even with deformable convolution (Cheng and Chen 2020).

Generic architecture for interpolation. Some methods explore using a generic architecture without domain knowledge, such as attention (Choi et al. 2020), transformer (Shi et al. 2022), 3D convolution under multi-frame input setup (Kalluri et al. 2023; Shi et al. 2022). However, both attention and 3D convolution are computationally expensive and thus prohibitive at 4K or 8K resolution.

Diffusion models for computer vision. Recently diffusion models have demonstrated their strength on generative computer vision applications such as image (Ho et al. 2022a; Peebles and Xie 2023) and video generation (Blattmann et al. 2023; Ho et al. 2022b), image editing (Brooks, Holynski, and Efros 2023; Yang, Hwang, and Ye 2023), 3D generation (Qian et al. 2024), *etc*. Beyond generation, diffusion has also shown to be effective for dense computer vision tasks and has been become the state-of-the-art technique for classical problems such as depth prediction (Ke et al. 2024; Saxena et al. 2023), optical flow prediction (Saxena et al. 2023), correspondence matching (Nam et al. 2024), semantic segmentation (Baranchuk et al. 2022; Xu et al. 2023), *etc*.

Diffusion models for interpolation. Two recent works explore diffusion for video frame interpolation from a generative perspective. LDMVFI (Danier, Zhang, and Bull 2024) proposes using a conditional latent diffusion model and optimizes it for perceptual frame interpolation quality, but the PSNR or SSIM metric of the predicted frames tends to be lower than that by state of the art. VIDIM (Jain et al. 2024) uses a cascaded pixel diffusion model but focuses on a task closer to the conditional video generation. Given two temporally-far-apart frames, the method generates a base video of 7 frames at 64×64 resolution and then upsamples them to 256×256. It is unclear whether a diffusion-based approach can achieve competitive results on the classical frame interpolation problem, where the input frames come from a video with high FPS and can be up to 8K resolution.

Cascaded diffusion models. Beginning with CDM (Ho et al. 2022a), cascades have become standard for scaling up the output resolution of pixel diffusion models. Diffusion cascades consist of a "base" model for an initial low-resolution solution and a number of separate "super-resolution" models to produce a higher-resolution output conditioned on the low resolution output. While effective for high-resolution output, memory cost increases proportionally with resolution since each super-resolution model performs diffusion at its output resolution. Even with specialized super-resolution architectures (Ho et al. 2022a; Saharia et al. 2022), the memory problem still persists as the target resolution increases significantly, *e.g.* from 1K to 8K.

High resolution diffusion. Recent works in high-resolution image generation have introduced training-free approaches through merging the score functions of nearby patches (Bar-Tal et al. 2023; Liu et al. 2024b) or expanding the network (Shi et al. 2024; Kim, Hwang, and Park 2024). Other methods explicitly train models to denoise partitioned patches (or tiles) and then merge them into high-resolution output. Zheng et al. (2024) generates any-size high-resolution output by merging denoised non-overlapping tiles during sampling process. Ding et al. (2024) uses score value and feature maps to encourage consistency between denoised patches. Skorokhodov et al. (2024) uses a hierarchical patch structure for efficient video generation, but requires specialized modules for global consistency.

Unlike these efforts, we focus on frame interpolation, an estimation task, and improve inference memory efficiency at extremely high resolutions (4K or 8K). Generation tasks require inter-patch communication for coherence generation at high resolution. Estimation task, however, benefits from strong conditioning signals (input frames) that localize the problem at the patch level. This allows us to explore distinct architectural choices for frame interpolation.

For estimation tasks, the most similar to ours is DDVM (Saxena et al. 2023) which uses tiling for high-resolution inference. After the base model runs at a coarse solution, the output is upsampled by partially denoising tiles taken from the coarse solution and input frames. In the context of frame interpolation, we show that this tiling performs worse than our proposed patch-based cascade.

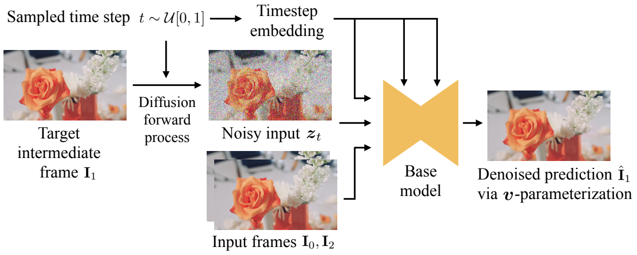
Sampled time step $t \sim \mathcal{U}[0,1]$
Timestep embedding
Target intermediate frame $\mathbf{I}_1$
Diffusion forward process
Noisy input $\boldsymbol{z}_t$
Input frames $\mathbf{I}_0, \mathbf{I}_2$
Base model
Denoised prediction $\hat{\mathbf{I}}_1$ via $\boldsymbol{v}$-parameterization

Figure 2: Our **base model** is conditioned on two input frames, $\mathbf{I}_0$ and $\mathbf{I}_2$, and predicts the intermediate frame $\mathbf{I}_1$. The model uses $\boldsymbol{v}$-parameterization (Salimans and Ho 2022; Saxena et al. 2024) for both model output and loss.

Diffusion for High-Resolution Frame Interpolation

We propose a pixel diffusion approach for frame interpolation. To enable high-resolution inference with low memory usage, we introduce a novel cascading strategy, patch-based, which performs diffusion on patches of high-resolution inputs. This cascade allows us to use the same model for both base estimation and upsampling.

Architecture

Our method adopts a conditional image diffusion framework. Given a concatenation of temporally nearby frames $\mathbf{I}_0$ and $\mathbf{I}_2$ as a conditioning signal, our model aims to estimate an intermediate frame $\hat{\mathbf{I}}_1$ as a reverse diffusion process in the pixel space, as illustrated in Fig. 2. We take a generic efficient U-Net architecture from DDVM (Saxena et al. 2023) with $\boldsymbol{v}$-parameterization (Salimans and Ho 2022; Saxena et al. 2024) for both model output and loss. The U-Net includes self-attention layers at two bottom levels. Given a noisy image $\boldsymbol{z}_t = \alpha_t \boldsymbol{x} + \sigma_t \boldsymbol{\epsilon}$ as an input, the network predicts $\hat{\boldsymbol{v}}$, where $\boldsymbol{x}$ is the target image (*i.e.*, $\mathbf{I}_1$), sampled random noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, sampled time step $t \sim \mathcal{U}[0,1]$, and $\alpha_t^2 + \sigma_t^2 = 1$. We directly apply L1 loss on $\boldsymbol{v}$ parameter space, *i.e.*, $||\hat{\boldsymbol{v}} - \boldsymbol{v}||_1$, where $\boldsymbol{v} = \alpha_t \boldsymbol{\epsilon} - \sigma_t \boldsymbol{x}$. The predicted image is recovered by $\hat{\boldsymbol{x}} = \alpha_t \boldsymbol{z}_t - \sigma_t \hat{\boldsymbol{v}}$, where $\hat{\boldsymbol{x}} = \hat{\mathbf{I}}_1$.

Patch-based cascade model

To handle extremely high resolutions (up to 8K), we propose a patch-based cascade approach that performs diffusion on patches of the input frames, keeping peak memory usage near constant at inference time. This allows us to use the same architecture for every upsample level and reuse the same model for both base and super-resolution settings, saving training time and disk space. On such high resolutions, standard cascades, which denoise the entire high resolution image directly, would require either a considerable amount of memory at inference time or a careful architecture search to reduce the memory cost.

Approach. Fig. 4 shows our overall inference strategy: we adopt the well-known coarse-to-fine idea for cascades and build an N-level image pyramid. Starting from the lowest scale s_{N-1} (where $s_n \equiv 1/2^n$), we downsample the input conditioning images by a factor of s_{N-1} (*i.e.*, $\mathbf{I}_0^{s_{N-1}}$ and

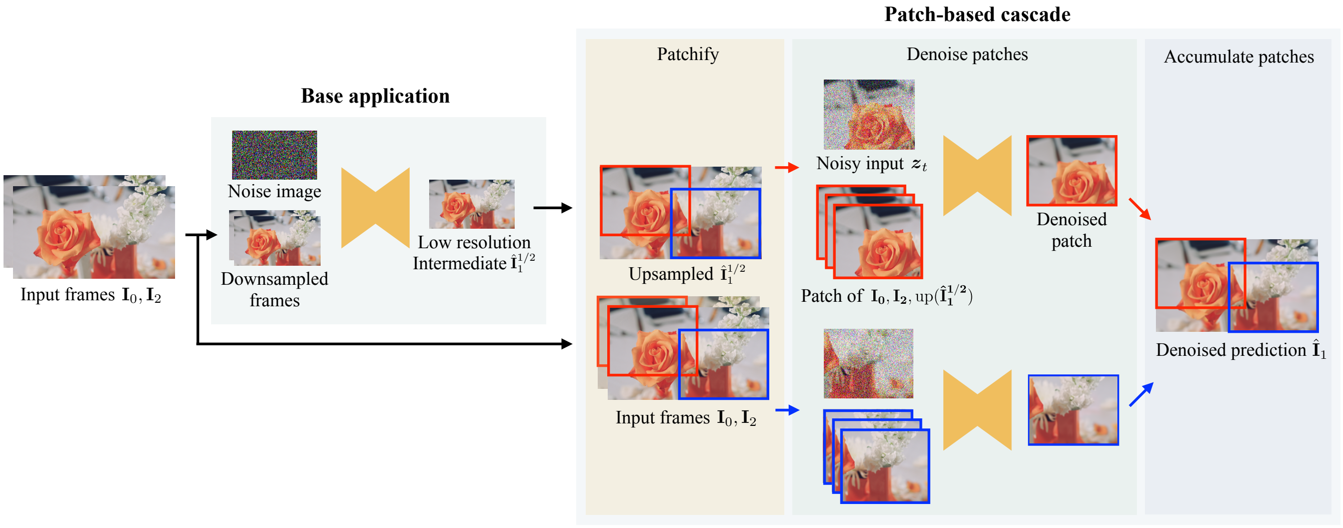
Patch-based cascade
Base application
Patchify
Denoise patches
Accumulate patches
Noise image
Downsampled frames
Low resolution Intermediate $\hat{\mathbf{I}}_1^{1/2}$
Input frames $\mathbf{I}_0, \mathbf{I}_2$
Upsampled $\hat{\mathbf{I}}_1^{1/2}$
Noisy input z_t
Patch of $\mathbf{I}_0, \mathbf{I}_2, \mathrm{up}(\hat{\mathbf{I}}_1^{1/2})$
Denoised patch
Denoised prediction $\hat{\mathbf{I}}_1$
Input frames $\mathbf{I}_0, \mathbf{I}_2$

Figure 3: **Patch-based cascade model**. Given a low-resolution intermediate from the previous level, patch-based cascade creates patches from bi-linearly upsampled low-resolution intermediate and two input frames and uses these patches as conditioning for a diffusion process. It then combines denoised patches to form the whole image. At inference time, only a single weight-shared model is recursively used across different image scales as in Fig. 4. Two-stage cascade is shown for simplicity.

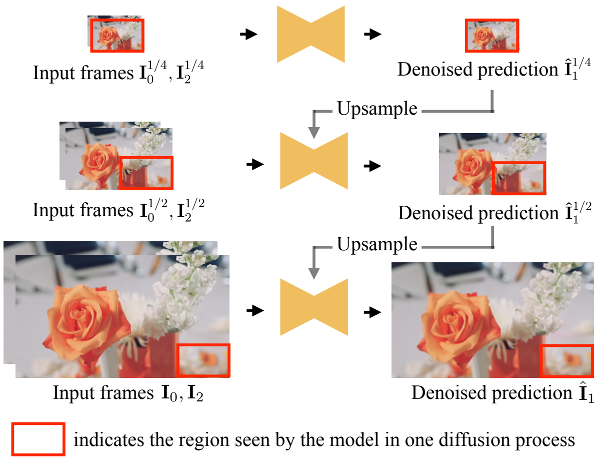
Input frames $\mathbf{I}_0^{1/4}, \mathbf{I}_2^{1/4}$
Denoised prediction $\hat{\mathbf{I}}_1^{1/4}$
Upsample
Input frames $\mathbf{I}_0^{1/2}, \mathbf{I}_2^{1/2}$
Denoised prediction $\hat{\mathbf{I}}_1^{1/2}$
Upsample
Input frames $\mathbf{I}_0, \mathbf{I}_2$
Denoised prediction $\hat{\mathbf{I}}_1$
indicates the region seen by the model in one diffusion process

Figure 4: **Upsampling strategy**. Like a standard cascade, we process the image from coarse to fine, but we always denoise at the same resolution, as indicated by the red box. Details on each step of the cascade are in Fig. 3.

$\mathbf{I}_2^{s_{N-1}}$) and predict an intermediate image $\hat{\mathbf{I}}_1^{s_{N-1}}$ at the same scale s_{N-1}. We then apply $2\times$ bilinear upsampling to this intermediate image $\mathrm{up}(\hat{\mathbf{I}}_1^{s_{N-1}})$ and use it as a conditioning signal for a denoising process at the scale s_{N-2}.

At each pyramid level, we upscale prediction via patch-based cascade, as shown in Fig. 3. For refinement at scale s_{N-2}, we take the prediction from the prior scale and then upsample it to match s_{N-2}. At each level, we perform three stages: *(i)* patchify, where we crop overlapping patches from the upsampled intermediate prediction and the input at that scale, *(ii)* denoise patches, where we run diffusion to obtain the prediction for each patch, and *(iii)* accumulate patches, where we use MultiDiffusion (Bar-Tal et al. 2023) to merge results from different patches for the prediction $\hat{\mathbf{I}}_2^{s_{N-2}}$. Here, we merge denoised patches at every denoising step. We then upsample $\hat{\mathbf{I}}_2^{s_{N-2}}$ to level s_{N-3} and repeat this process until $n = 0$, the original input scale.

Training setup. For the patch-based cascade model, we want to train a diffusion model that is conditioned on a pair of input images and a half resolution representation of the target we aim to predict. We first predict the intermediate frame at a half resolution $\hat{\mathbf{I}}_1^{1/2}$ by feeding downsampled inputs to a pre-trained base model (Fig. 2) computed offline. This intermediate frame is then upsampled to the original scale and used as a conditioning image, along with the original inputs, for training the patch-based cascade model using standard diffusion. This inference step is performed offline to improve training efficiency.

Single model for all stages. By conditioning on the low resolution estimate but using dropout 50% during training time, we can use the same model for all cascade stage, including base and super-resolution; base generation is done by passing zeros as the low resolution condition. While similar to CFG (Ho and Salimans 2022), we do not combine unconditional and conditional estimations at inference. Empirically we find that a single shared model for all stages performs slightly better than having a dedicated super-resolution model. It also substantially reduces training time (training one model instead of multiple separate ones) and disk space at inference time (saving only one model). Interestingly, we observe that a dedicated super-resolution model trained without dropout on the coarse estimation does not work since it takes the shortcut of upsampling the low resolution instead of attending to the high resolution inputs.

Experiments

Implementation details. Similar to previous diffusion-based methods (Jain et al. 2024; Danier, Zhang, and Bull

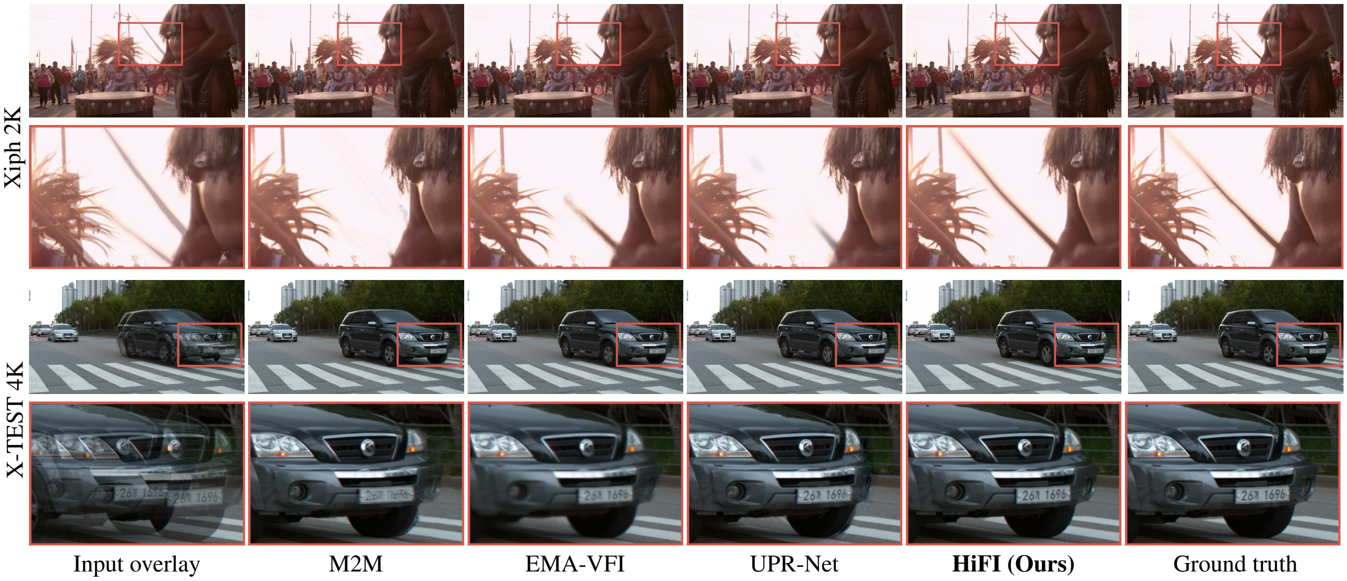

Figure 5: **Qualitative examples for public datasets.** Our method performs well even in cases of large motion and complex textures such as a thin object on the top and the plate number at the bottom.

2024), we utilize a large-scale video dataset for training, to test the scalability of the diffusion model better. The dataset contains up to 30 M videos with 40 frames, collected from the internet and other sources with licenses permitting research uses. We first train our base model on the dataset, and then we additionally include Vimeo-90K triplet (Xue et al. 2019) and X-TRAIN (Sim, Oh, and Kim 2021) to finetune the cascade model. For fair comparison, we also prepare a model trained on Vimeo-90K and X-TRAIN only from scratch. We use a mini-batch size of 256 and train the base model for 3 M iteration steps and the patch-based cascade model for 200 k iteration steps. We use the Adam optimizer (Kingma and Ba 2014) with a constant learning rate $1e^{-4}$ with initial warmup. For inference, we use 3-stage patch-based cascade setup with a patch size of 512×768, averaging 4 samples estimated via 4 sampling steps.

Our data augmentation includes random crop and horizontal, vertical, and temporal flip with a probability of 50%. We use a crop size of 352×480 for large-scale base model training and 224×288 for the cascade model training. We use a multi-resolution crop augmentation that crops an image patch with a random rectangular crop size between the original resolution and the final crop size and then resize it to the final crop size. While commonly used, we find random 90° rotation augmentation and photometric augmentation to be less effective, so we opt not to use them.

More details are in the supplementary material.

Public benchmark evaluation

We first evaluate HiFI on three popular benchmark datasets, Vimeo-90K triplet (Xue et al. 2019), Xiph (Niklaus and Liu 2020), and X-TEST (Sim, Oh, and Kim 2021) in Table 1, as well as an 8K dataset, SEPE (Al Shoura et al. 2023).

Vimeo-90K. The low-resolution (256×448) Vimeo-90K is one of the most heavily studied benchmark, where numbers are highly saturated among different methods. HiFI achieves competitive accuracy with a generic training procedure. Please view the supplementary for further discussion.

Xiph and X-TEST. Both Xiph and X-TEST have high resolution (2K and 4K). The motion of X-TEST can be over 400 pixels at the 4K resolution, particularly challenging for existing methods. For X-TEST, we follow the evaluation protocol discussed in (Sim, Oh, and Kim 2021) that interpolates 7 intermediate frames. When trained on a combination of Vimeo and X-TRAIN, HiFI performs favorably against state of the art on Xiph and X-TEST datasets, both in 2K and 4K resolutions. Pre-training on a large video dataset significantly boosts the performance of HiFI on Xiph and X-TEST, setting a new state of the art. Visually, HiFI can better interpolate fine details with large motion at high resolution, as shown in Fig. 5. We will analyze key components that contribute to the performance in the ablation study below.

SEPE. We also test HiFI on SEPE that includes 8K resolution videos. Most methods we tested ran out of memory except M2M (PSNR 28.34 (dB) and SSIM 0.883) and SGM-VFI (Liu et al. 2024a) (PSNR 28.43 (dB) and SSIM 0.880), compared with PSNR 29.78 (dB) and SSIM 0.900 by HiFI. Please view the supplementary for visual comparison.

Large Motion and Repetitive texture dataset

Public benchmark datasets, while diverse, do not fully capture the failure modes of current methods, especially large motion or repetitive texture cases common in real-world videos. To better evaluate existing methods and further innovation, we introduce a Large Motion and Repetitive texture (**LaMoR**) dataset that includes such 19 challenging examples at 4K resolution in both portrait and landscape modes, as shown in Fig. 6. As in Table 2 and Fig. 7, HiFI substantially outperform all state of the arts on challenging cases of

Method	Training dataset	Vimeo-90K		Xiph				X-TEST			
				2K		4K		2K		4K	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
M2M (Hu et al. 2022)	Vimeo	35.47	0.978	36.44	0.967	33.92	0.945	32.07	0.923	30.81	0.912
FILM (Reda et al. 2022)	Vimeo	36.06	0.970	36.66	0.951	33.78	0.906	31.61	0.916	26.98	0.839
AMT (Li et al. 2023)	Vimeo	36.53	**0.982**	36.38	0.941	34.63	0.904	-	-	-	-
UPR-Net (Jin et al. 2023)	Vimeo	36.42	**0.982**	-	-	-	-	-	-	30.68	0.909
FITUG (Plack et al. 2023)	Vimeo	36.34	0.981	-	-	-	-	-	-	-	-
TCL (Zhou et al. 2023)	Vimeo	**36.85**	**0.982**	-	-	-	-	-	-	-	-
IQ-VFI (Hu et al. 2024)	Vimeo	36.60	**0.982**	36.68	0.942	34.72	0.905	-	-	-	-
EMA-VFI (Zhang et al. 2023)	Vimeo (+septuplet for X-TEST)	36.64	**0.982**	36.90	0.945	34.67	0.907	32.85	0.930	31.46	0.916
XVFI (Sim, Oh, and Kim 2021)	Vimeo / X-TRAIN	35.07	0.976	-	-	-	-	30.85	0.913	30.12	0.870
BiFormer (Park, Kim, and Kim 2023)	Vimeo + X-TRAIN	-	-	-	-	34.48	0.927	-	-	31.32	0.921
HiFI (Ours)	Vimeo + X-TRAIN	35.70	0.979	36.64	0.967	34.45	0.948	33.03	0.927	32.03	0.918
	Vimeo + X-TRAIN + Raw videos	36.12	0.980	**37.36**	**0.969**	**35.40**	**0.953**	**33.94**	**0.941**	**32.92**	**0.931**

Table 1: **Results on public benchmark datasets**: HiFI performs favorably on the highly-saturated Vimeo-90K (Xue et al. 2019) and is substantially more accurate than existing two-frame methods on high-resolution Xiph (Niklaus and Liu 2020) and X-TEST (Sim, Oh, and Kim 2021) datasets. **Best** and second-best are highlighted in color.

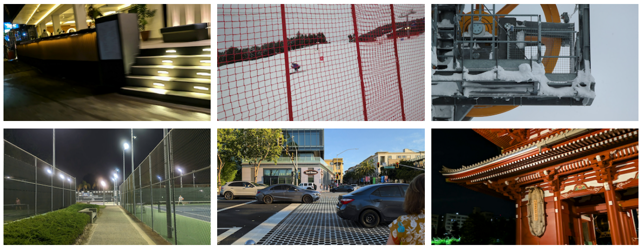

Figure 6: **A few examples from our LaMoR dataset** that includes challenging scenes, such as repetitive texture and large motion where typical methods fail.

Method	PSNR	SSIM
LDMVFI (Danier, Zhang, and Bull 2024)	21.952	0.828
EMA-VFI (Zhang et al. 2023)	22.327	0.845
M2M (Hu et al. 2022)	24.995	0.884
SGM-VFI (Liu et al. 2024a)	25.122	0.894
UPR-Net (Jin et al. 2023)	25.856	0.892
BiFormer (Park, Kim, and Kim 2023)	26.330	0.893
HiFI (Ours)	**28.141**	**0.912**

Table 2: Results on **LaMoR.** HiFI is significantly more accurate than state-of-the-art methods.

repetitive textures and large motion.

Ablation study

Dedicated upsample model vs single model. In Table 3, we compare the accuracy of the base model, two distinct models for base and upsample, and our final setting of using the same model for base and upsample. Both cascade strategies are effective for handling large motion, substantially improving accuracy on X-TEST. Using the same model for both base and upsample performs on-par or even better than having a dedicated upsample model, especially on challenging X-TEST. This validates the strength of re-using the same model for both over the more expensive dedicate model setup. Increasing the number of upsample stages improves the accuracy but saturates over three.

Method	Model size	Vimeo-90K		X-TEST 2K		X-TEST 4K	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Base	647 M	35.44	0.978	30.32	0.879	28.57	0.876
Standard two-stage cascade	1294 M	**36.12**	**0.980**	33.86	0.939	32.48	0.926
Patch-based self-cascade × 2	647 M	**36.12**	**0.980**	33.93	**0.941**	32.77	0.930
Patch-based self-cascade × 3	647 M	**36.12**	**0.980**	**33.94**	**0.941**	**32.92**	**0.931**

Table 3: **Base *vs.* cascade models.** Our patch-based self-cascade formulation substantially outperforms the base with the same number of parameters. Through model sharing, our self-cascade generalizes better on the X-TEST dataset than the standard cascade but with half of the parameters.

Comparison with coarse-to-fine refinement. Coarse-to-fine tiling refinement from DDVM (Saxena et al. 2023) first predicts the target at low resolution, bilinearly upsamples it to the target resolution, and refines it from an intermediate sampling step in a patch-wise manner. Our patch-based cascade performs consistently better than the coarse-to-fine tiling refinement on the X-TEST benchmark; 32.92 (dB) vs 32.54 (dB) on 4K, and 33.94 (dB) *vs.* 33.03 (dB) on 2K.

Architecture. In Table 4, we analyze where the major gain originates from by ablating attention layers or diffusion process in the base model, given the same training assets (*e.g.*, datasets, computations, *etc.*). Using attention layers brings about moderate performance gains on both the small (Vimeo) and large (X-TEST) motion datasets. We find the attention layers help with handling large motion and repetitive textures, enabling the accurate interpolation of frames by capturing the global context of these textures. Removing the diffusion process also leads to significant performance degradation. We also test one widely-used traditional

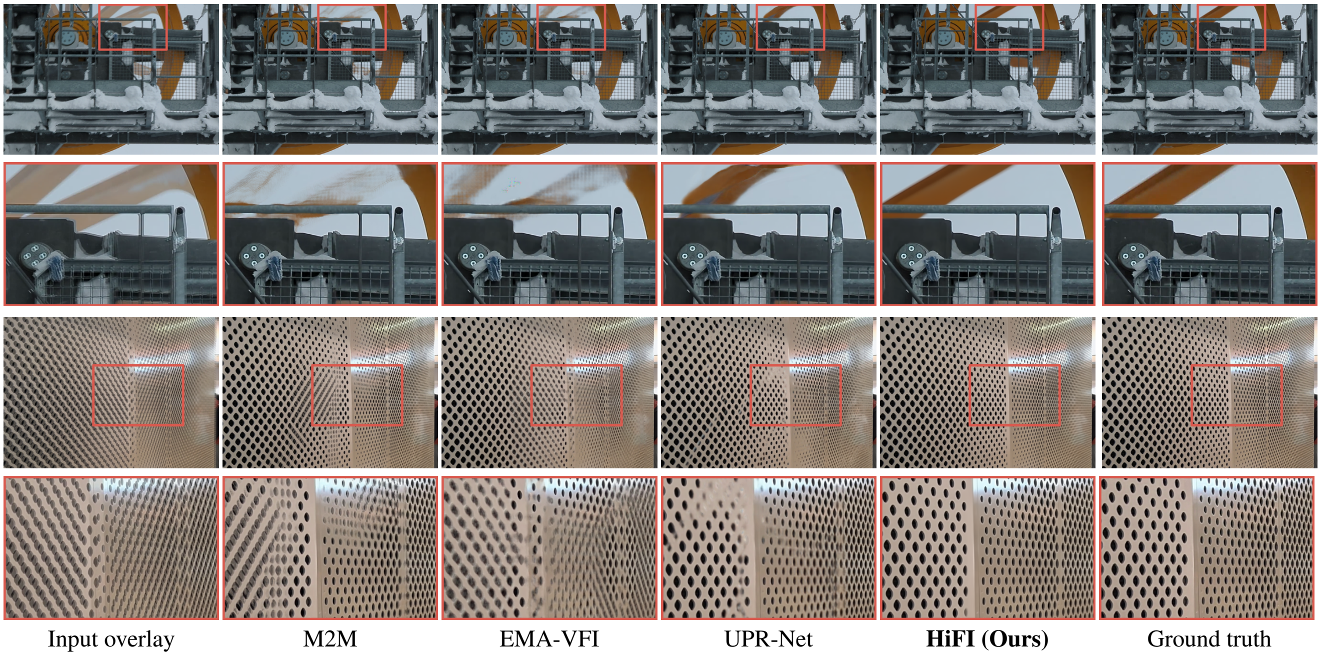

Figure 7: **Qualitative comparison on LaMoR.** The proposed HiFI is particularly effective at very challenging cases including repetitive textures and large motion.

Method	Vimeo-90K		X-TEST 2K		X-TEST 4K	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Ours, base model	**35.44**	**0.978**	**30.32**	**0.879**	**28.57**	**0.876**
w/o attention layers	35.13	0.977	29.73	0.861	27.75	0.854
w/o diffusion	33.78	0.965	28.05	0.852	27.56	0.861
FILM (Reda et al. 2022)	34.02	0.970	28.15	0.854	27.24	0.856

Table 4: **Architecture analysis.** Both attention layers and diffusion process contribute to substantial accuracy gain. Comparing to a domain-specific architecture, FILM, our approach scales up better when training on the same large-scale video dataset.

method FILM (Reda et al. 2022), which relies on a "scale-agnostic" motion estimator to handle large motion. FILM trained on the same large dataset is substantially worse than HiFI, suggesting that traditional, hand-designed methods do not scale up well *w.r.t.* data.

Number of sampling steps. The optimal number of sampling steps also differ between the base and the patch-based cascade model. In general, we find that the model needs more sampling steps for large motion (*e.g.*, X-TEST) than for small motion (*e.g.*, Vimeo-90K (Xue et al. 2019)). However, the patch-based cascade model is able to achieve better numbers across different datasets with fewer sampling steps.

Discussions. Despite the performance gains on standard benchmarks, some extremely complicated motion types, *e.g.*, fluid dynamics, are still challenging for HiFI. Furthermore, diffusion models are computationally heavy and need distillation (Salimans and Ho 2022) for applications with a limited computational budget.

Steps	Vimeo-90K	X-TEST 4K
1	34.87	27.95
2	35.37	27.92
4	**35.44**	28.57
8	35.21	29.67
16	34.58	30.34
32	33.53	**30.40**
64	32.68	30.02

(a) Base model

Steps	Vimeo-90K	X-TEST 4K
1	36.13	32.32
2	**36.15**	32.83
4	36.12	**32.92**
8	36.06	**32.92**
16	35.98	32.84
32	35.92	32.68
64	35.86	32.64

(b) Patch-based cascade

Table 5: **Effect of the sampling steps** on PSNR for the base model and patch-based cascade model. More steps tends to be better for higher resolution and large motion datasets.

Conclusion

We have introduced a diffusion-based method for high resolution frame interpolation, named HiFI. Our proposed patch-based cascade achieves state-of-the-art performance on several high-resolution frame interpolation benchmarks up to 8K resolution, while improving efficiency for training and inference. We also establish a new benchmark, LaMoR, which focuses on challenging cases, *e.g.* large motion and repeated textures at high resolution. Our method substantially outperforms all methods on the benchmark as well.

Acknowledgements

We thank Tianhao Zhang, Yisha Sun, Tristan Greszko, Christopher Farro, Fuhao Shi for their help in collecting the dataset, and Yifan Zhuang, Ming-Hsuan Yang, David Salesin, and the anonymous reviewers for their constructive feedback and discussions.

References

Al Shoura, T.; Dehaghi, A. M.; Razavi, R.; Far, B.; and Moshirpour, M. 2023. SEPE Dataset: 8K Video Sequences and Images for Analysis and Development. In *Conference on ACM Multimedia Systems*, 463–468.

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *ICML*.

Baranchuk, D.; Voynov, A.; Rubachev, I.; Khrulkov, V.; and Babenko, A. 2022. Label-Efficient Semantic Segmentation with Diffusion Models. In *ICLR*.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; Jampani, V.; and Rombach, R. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv:2311.15127 [cs.CV]*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. InstructPix2Pix: Learning to follow image editing instructions. In *CVPR*, 18392–18402.

Cheng, X.; and Chen, Z. 2020. Video frame interpolation via deformable separable convolution. In *AAAI*, 10607–10614.

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*, 10663–10671.

Danier, D.; Zhang, F.; and Bull, D. 2024. LDMVFI: Video frame interpolation with latent diffusion models. In *AAAI*, 1472–1480.

Ding, Z.; Zhang, M.; Wu, J.; and Tu, Z. 2024. Patched denoising diffusion models for high-resolution image synthesis. In *ICLR*.

Ho, J.; Saharia, C.; Chan, W.; Fleet, D. J.; Norouzi, M.; and Salimans, T. 2022a. Cascaded diffusion models for high fidelity image generation. *JMLR*, 23(47): 1–33.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video diffusion models. *NeurIPS*, 35: 8633–8646.

Hu, M.; Jiang, K.; Zhong, Z.; Wang, Z.; and Zheng, Y. 2024. IQ-VFI: Implicit Quadratic Motion Estimation for Video Frame Interpolation. In *CVPR*, 6410–6419.

Hu, P.; Niklaus, S.; Sclaroff, S.; and Saenko, K. 2022. Many-to-many splatting for efficient video frame interpolation. In *CVPR*, 3553–3562.

Huang, Z.; Zhang, T.; Heng, W.; Shi, B.; and Zhou, S. 2022. Real-time intermediate flow estimation for video frame interpolation. In *ECCV*, 624–642.

Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; and Brox, T. 2017. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*.

Jain, S.; Watson, D.; Tabellion, E.; Hołyński, A.; Poole, B.; and Kontkanen, J. 2024. Video Interpolation with Diffusion Models. In *CVPR*.

Jiang, H.; Sun, D.; Jampani, V.; Yang, M.-H.; Learned-Miller, E.; and Kautz, J. 2018. Super SloMo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*, 9000–9008.

Jin, X.; Wu, L.; Chen, J.; Chen, Y.; Koo, J.; and Hahm, C.-h. 2023. A unified pyramid recurrent network for video frame interpolation. In *CVPR*, 1578–1587.

Kalluri, T.; Pathak, D.; Chandraker, M.; and Tran, D. 2023. FLAVR: Flow-agnostic video representations for fast frame interpolation. In *WACV*, 2071–2082.

Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; and Schindler, K. 2024. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*.

Kim, Y.; Hwang, G.; and Park, E. 2024. DiffuseHigh: Training-free Progressive High-Resolution Image Synthesis through Structure Guidance. *arXiv preprint arXiv:2406.18459*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. In *ICLR*.

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 *CVPR*, 5316–5325.

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 *CVPR*, 9801–9810.

Liu, C.; Zhang, G.; Zhao, R.; and Wang, L. 2024a. Sparse Global Matching for Video Frame Interpolation with Large Motion. In *CVPR*, 19125–19134.

Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2024b. SyncDreamer: Generating multiview-consistent images from a single-view image. In *ICLR*.

Meyer, S.; Djelouah, A.; McWilliams, B.; Sorkine-Hornung, A.; Gross, M.; and Schroers, C. 2018. PhaseNet for video frame interpolation. In *CVPR*, 498–507.

Meyer, S.; Wang, O.; Zimmer, H.; Grosse, M.; and Sorkine-Hornung, A. 2015. Phase-based frame interpolation for video. In *CVPR*, 1410–1418.

Nam, J.; Lee, G.; Kim, S.; Kim, H.; Cho, H.; Kim, S.; and Kim, S. 2024. Diffusion Model for Dense Matching. In *ICLR*.

Niklaus, S.; and Liu, F. 2018. Context-aware synthesis for video frame interpolation. In *CVPR*, 1701–1710.

Niklaus, S.; and Liu, F. 2020. Softmax splatting for video frame interpolation. In *CVPR*, 5437–5446.

Niklaus, S.; Mai, L.; and Liu, F. 2017a. Video frame interpolation via adaptive convolution. In *CVPR*, 670–679.

Niklaus, S.; Mai, L.; and Liu, F. 2017b. Video frame interpolation via adaptive separable convolution. In *ICCV*, 261–270.

Niklaus, S.; Mai, L.; and Wang, O. 2021. Revisiting adaptive convolutions for video frame interpolation. In *WACV*, 1099–1109.

Park, J.; Kim, J.; and Kim, C.-S. 2023. BiFormer: Learning bilateral motion estimation via bilateral transformer for 4K video frame interpolation. In *CVPR*, 1568–1577.

Park, J.; Ko, K.; Lee, C.; and Kim, C.-S. 2020. BMBC: Bilateral motion estimation with bilateral cost volume for video interpolation. In *ECCV*, 109–125.

Park, J.; Lee, C.; and Kim, C.-S. 2021. Asymmetric bilateral motion estimation for video frame interpolation. In *ICCV*, 14539–14548.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *ICCV*, 4195–4205.

Plack, M.; Briedis, K. M.; Djelouah, A.; Hullin, M. B.; Gross, M.; and Schroers, C. 2023. Frame Interpolation Transformer and Uncertainty Guidance. In *CVPR*, 9811–9821.

Qian, G.; Mai, J.; Hamdi, A.; Ren, J.; Siarohin, A.; Li, B.; Lee, H.-Y.; Skorokhodov, I.; Wonka, P.; Tulyakov, S.; and Ghanem, B. 2024. Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors. In *ICLR*.

Reda, F.; Kontkanen, J.; Tabellion, E.; Sun, D.; Pantofaru, C.; and Curless, B. 2022. FILM: Frame interpolation for large motion. In *ECCV*, 250–266.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 36479–36494.

Salimans, T.; and Ho, J. 2022. Progressive Distillation for Fast Sampling of Diffusion Models. In *ICLR*.

Saxena, S.; Herrmann, C.; Hur, J.; Kar, A.; Norouzi, M.; Sun, D.; and Fleet, D. J. 2023. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. In *NeurIPS*.

Saxena, S.; Hur, J.; Herrmann, C.; Sun, D.; and Fleet, D. J. 2024. Zero-Shot Metric Depth with a Field-of-View Conditioned Diffusion Model. In *ECCVW*.

Shi, S.; Li, W.; Zhang, Y.; He, J.; Gong, B.; and Zheng, Y. 2024. ResMaster: Mastering High-Resolution Image Generation via Structural and Fine-Grained Guidance. *arXiv:2406.16476 [cs.CV]*.

Shi, Z.; Xu, X.; Liu, X.; Chen, J.; and Yang, M.-H. 2022. Video frame interpolation transformer. In *CVPR*, 17482–17491.

Sim, H.; Oh, J.; and Kim, M. 2021. XVFI: extreme video frame interpolation. In *ICCV*, 14489–14498.

Skorokhodov, I.; Menapace, W.; Siarohin, A.; and Tulyakov, S. 2024. Hierarchical Patch Diffusion Models for High-Resolution Video Generation. In *CVPR*, 7569–7579.

Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2018. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 8934–8943.

Teed, Z.; and Deng, J. 2020. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 402–419.

Xu, J.; Liu, S.; Vahdat, A.; Byeon, W.; Wang, X.; and De Mello, S. 2023. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2955–2966.

Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019. Video enhancement with task-oriented flow. *IJCV*, 127: 1106–1125.

Yang, S.; Hwang, H.; and Ye, J. C. 2023. Zero-shot contrastive loss for text-guided diffusion image style transfer. In *ICCV*, 22873–22882.

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 *CVPR*, 5682–5692.

Zheng, Q.; Guo, Y.; Deng, J.; Han, J.; Li, Y.; Xu, S.; and Xu, H. 2024. Any-size-diffusion: Toward efficient text-driven synthesis for any-size HD images. In *AAAI*, 7571–7578.

Zhou, K.; Li, W.; Han, X.; and Lu, J. 2023. Exploring motion ambiguity and alignment for high-quality video frame interpolation. In *CVPR*, 22169–22179.