Latent Consistency Models for Video Frame Interpolation

Mohamed Ghanem [*1] **Rohith Nibhanupudi** [*1] **Faris Qadan** [*1]

Abstract

Video Frame Interpolation (VFI) is a challenging problem in video processing, where the goal is to generate intermediate frames between existing ones. Traditional VFI methods, based on optical flow models, often fall short in complex scenarios with large motions and dynamic textures. Recent advancements with Latent Diffusion Models (LDMs) have significantly enhanced visual quality in such challenging conditions. However, these improvements come at the cost of increased computational demands and slower generation times, presenting a notable trade-off in VFI technology. We introduce a novel approach using *Latent Consistency Models* (LCMs) to address this trade-off. LCMs leverage a single-step consistency function within a latent diffusion framework to ensure self-consistency along Probability Flow Ordinary Differential Equation (PF-ODE) trajectories. This method significantly simplifies the computational process compared to multi-step diffusion models, offering a substantial improvement in processing efficiency. Empirical results demonstrate that our LCM achieves parity with state-of-the-art models in both visual and statistical metrics, producing high-quality frame interpolations in just 25% of the time required by leading latent diffusion models. This advancement not only sets a new benchmark in VFI but also opens avenues for more efficient and rapid video processing in time-sensitive applications.

[*]Equal contribution [1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA. Correspondence to: Mohamed Ghanem <mghanem8@gatech.edu>, Rohith Nibhanupudi <rnibhanupudi3@gatech.edu>, Faris Qadan <fqadan3@gatech.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

1. Introduction

1.1. Prior Work

In prior work, optical flow models have been widely deployed for VFI (Lyasheva et al., 2020; Shi et al., 2023). The goal of optical flow is to determine the nature of motion between corresponding entities in consecutive frames and explicitly synthesize intermediate images to enhance the resulting video quality. This can be achieved either by assuming the type of motion present in the intermediate frame, or directly estimating intermediate flows. Flow-based methods utilizing deep learning or mathematical approaches have managed to achieve comparable or even superior results to CNN and GAN-based implementations with the ability outstand heavy computational requirements of deep learning methods on real-time benchmarks (Parihar et al., 2022). While flow-based models have been shown to perform well against Peak Signal-to-Noise Ratio (PSNR), they perform worse against more subjective human evaluations.

Diffusion models, a significant advancement in image processing, have substantially outperformed previous methodologies like flow-based models, GANs, VAEs, and CNNs in terms of image quality (Ho et al., 2020). These models function by iteratively denoising images initially distorted with Gaussian noise, effectively reversing a Markov chain process that incrementally introduces noise into a clean image.

A notable extension of this approach is the Latent Diffusion Model (LDM). Unlike traditional diffusion models that operate in the image space, LDMs conduct both forward and reverse diffusion processes in the image latent space. This shift to latent space processing, as highlighted in recent research (Rombach et al., 2021), results in the generation of high-quality samples in a markedly reduced timeframe compared to standard diffusion models.

When applied to the VFI problem, LDMs have demonstrated remarkable capabilities. One example is the state-of-the-art LDM-VFI model (Danier et al., 2023), which employs a vector-quantized GAN autoencoder, known as VQ-FIGAN.

However, despite their advancements, LDMs are not without limitations. The process of generating outputs in LDMs involves iterative sampling from the latent space, a method that, while effective, introduces significant delays in VFI

tasks. On average, the generation of a single frame using this model can take as long as 8.48 seconds, highlighting a critical bottleneck in the application of LDMs for real-time video processing.

1.2. Motivation for Latent Consistency Models

Consistency models (CMs) (Song et al., 2023) improve upon diffusion models by supporting single-step generation, introducing the notion of a trade-off between time and sample quality. CMs are built around a consistency function which seeks to map points on a trajectory of the probability-flow ordinary differential equation (PF-ODE) to that trajectory's origin (the solution of the PF-ODE). The PF-ODE smoothly maps from the data distribution to the noise distribution in continuous-time diffusion models. Additionally, consistency functions maintain self-consistency, meaning that all points on a PF-ODE trajectory map to the same origin. The consistency function call can be done in a one-step process, as opposed to the multi-step process required for sampling from a standard diffusion model. Furthermore, any diffusion model can be efficiently distilled into a consistency model given an efficient ODE solver (of which many exist (Song et al., 2021; Karras et al., 2022)). Thus, consistency models allow for the development of generative models that facilitate efficient, single-step generation without sacrificing important advantages of iterative sampling.

Latent Consistency Models (LCMs) (Luo et al., 2023) represent a significant advancement over both Latent Diffusion Models (LDMs) and traditional Consistency Models (CMs). LCMs implement consistency models within the latent space, thereby accelerating the generation of high-quality images while reducing computational overhead.

1.3. Novelty and Approach

Previously, neither consistency models nor their latent counterparts have been applied for VFI. In this work, we explore the application of latent consistency models for VFI. We begin by distilling a pre-trained LDM-VFI model into an LCM using a specialized distillation algorithm. Our evaluation shows that this approach not only maintains state-of-the-art performance but also excels in scenarios demanding both time efficiency and computational effectiveness.

2. Methodology

Our approach focuses on distilling the pre-trained Latent Diffusion Model for Video Frame Interpolation (LDM-VFI) into its Latent Consistency Model (LCM) counterpart. This process involves adapting the consistency distillation algorithm to the frame interpolation task. By translating LDM-VFI into the LCM framework, we ensure that the distilled model not only retains the high-quality generation characteristics of the LDM-VFI but also benefits from the computational efficiency intrinsic to LCMs.

2.1. Latent Diffusion in Frame Interpolation

Latent diffusion models for VFI contain an image encoder $E : x \mapsto z$ that encodes an image x into a latent representation z, as well as a decoder $D : z \mapsto x$ that reconstructs the image x from the latent z. Between the encoder and decoder, a denoising U-Net model, ϵ_θ, is trained to predict the noise added to the latent at each time step t.

We utilize the encoder and decoder from VQ-FIGAN (Danier et al., 2023), a specialized autoencoder for VFI. It optimizes retention of high-frequency detail during encoding and decoding phases which is critical for perceived video quality. Its architecture, adapted from VQGAN, integrates features from neighboring frames using a frame-aided decoder, employs efficient MaxViT blocks for scalable self-attention, and utilizes deformable convolution kernels for enhanced interpolation. This design allows VQ-FIGAN to handle high-resolution videos effectively, making it a robust solution for VFI.

2.2. Running Inference on the Latent Diffusion Model

The algorithm for generating interpolated frames using latent diffusion is outlined in Algorithm 1.

Algorithm 1 Inference

Input: Original frames I^0, I^1, noise schedule $\{\beta_t\}_{t=1}^T$, maximum diffusion step T
Load: pre-trained denoising U-Net ϵ_θ, VQ-FIGAN encoder E and decoder D
Sample $z_T^n \sim \mathcal{N}(0, I)$
Encode $z^0 = E(I^0), z^1 = E(I^1)$ and store features ϕ^0, ϕ^1
for $t = T, \ldots, 1$ **do**
  Predict noise $\hat{z_t^n} = \epsilon_\theta(z_t^n, t, z^0, z^1)$
  Compute the mean μ_θ from $\hat{z_t^n}$
  Compute the standard deviation σ_t from β_t
  Sample z_{t-1}^n from $p_\theta(z_{t-1}^n | z_t^n) = \mathcal{N}(\mu_\theta, \sigma_t^2 I)$
end for
return $\hat{I}^n = D(z_0^n, \phi^0, \phi^1)$ as the interpolated frame

The VQ-FIGAN and denosing U-Net models are pre-trained by the original authors. To create the predicted interpolated frame $\hat{I}_n$ from input frames I^0 and I^1, sample Gaussian noise in the latent space, and then perform T steps of denoising using the U-Net network. The U-Net model and noise schedule provide a conditional probability distribution for sampling denoised latents at each step of denoising. Finally, the decoder D predicts the interpolated frame.

2.3. Latent Consistency Distillation

The LDM-VFI is converted to a LCM using the Latent Consistency Distillation (LCD) algorithm. Similar to how LDMs adopt diffusion models in the image latent space, LCMs adopt a consistency model in the image latent space. For improved latent diffusion performance, LCMs solve an augmented PF-ODE between the data and noise distributions that includes classifier-free guidance (Ho & Salimans, 2022).

Classifier-free guidance (CFG) in the reverse diffusion process, replaces the original noise prediction with a linear combination of conditional and unconditional noise. When sampling from the guided reverse process, the augmented consistency function approximates the augmented PF-ODE solution. The consistency loss is computed using the augmented consistency function f_θ as follows:

$$\mathcal{L}_{\text{CD}}(\theta, \theta^-; \Psi) = \mathbb{E}_{z,\omega,c,t}\left[d\left(f_\theta(z_{t+1}, \omega, c, t_{n+1}), f_{\theta^-}(\hat{z}_{t_n,\omega}, c, t_n)\right)\right]$$

The noise prediction $\hat{z}_{t_n}$ is generated by the denoising diffusion model $\tilde{\epsilon}_\theta$ as follows:

$$\hat{z}_{t,\omega} - z_{t+1} \approx (1+\omega)\Psi(z_{t+1}, t+1, t, c) - \omega\Psi(z_{t+1}, t+1, t, \varnothing)$$

Ψ is a PF-ODE solver, and we selected the DDIM solver described in (Song et al., 2020).

The LCD is also accelerated by skipping time steps. This is done by ensuring consistency between the current time step and k time steps away given by $t + k \mapsto t$, with a value of k = 20 selected as chosen in (Luo et al., 2023). Thus, the consistency distillation loss and augmented PF-ODE solver are modified to ensure consistency from t_{n+k} to t_n.

From these expressions, we present the pseudo-code for distilling the LDM-VFI model with LCD, CFG, and skipping time steps in Algorithm 2:

Algorithm 2 Latent Consistency Distillation (LCD)

Input: dataset $\mathcal{D}$, initial LDM-VFI parameter θ, learning rate η, ODE solver Ψ, distance metric d, EMA rate μ, noise schedule $\alpha(t)$, $\sigma(t)$, guidance scale $[\omega_{\min}, \omega_{\max}]$, skipping interval k, and VQ-FIGAN encoder E
Encoding training data into latent space: $\mathcal{D}_z = \{(z, c)|z = E(x), (x, c) \in \mathcal{D}\}$
Initialize: $\theta^- \leftarrow \theta$
repeat
  Sample $(z, c) \sim \mathcal{D}_z$,
  Sample $n \sim \mathcal{U}[1, N-k]$
  Sample $\omega \sim \mathcal{U}[\omega_{\min}, \omega_{\max}]$
  Sample $z_{t_{n+k}} \sim \mathcal{N}(z_{t_{n+k}}; z, \sigma^2(t_{n+k})I)$
  $\tilde{z}_{t_n,\omega} \leftarrow z_{t_n} + (1+\omega)(z_{t_{n+k}} - z_{t_n})$
  $L(\theta, \theta^-; \Psi) \leftarrow d(f_\theta(z_{t_n,\omega}, c, t_n), f_{\theta^-}(\tilde{z}_{t_n,\omega}, c, t_{n+k}, t_n))$
  $\theta \leftarrow \theta - \eta\nabla_\theta L(\theta, \theta^-)$
  $\theta^- \leftarrow \text{stopgrad}(\mu\theta^- + (1-\mu)\theta)$
until convergence

2.4. Training Dataset

The training set is outlined in (Danier et al., 2023), and it consists of 64,612 frame septuplets from Vimeo90k. The model trains on frame triplets (I^0, I^n, I^1), and so, for each septuplet, only the three frames in the center (3, 4, and 5) are grabbed. For data augmentation, we randomly crop 256 × 256 patches from the triplets, and perform random flipping and temporal order reversing.

2.5. Distributed Training

To train the model, we employed distributed training techniques for efficient multi-GPU training. Models were trained with and without CFG to investigate its effect on interpolation quality.

Overall, training took 5 hours for each model configuration with 2 NVIDIA V100 GPUs utilized for each run.

2.6. Evaluation Dataset and Procedure

We then evaluated the fully-trained LCM on the most commonly used VFI benchmarks, including Middlebury (Baker et al., 2007), UCF-101 (Soomro et al., 2012), DAVIS (Pont-Tuset et al., 2017), and SNU-FILM (Choi et al., 2020) datasets.

Following the training phase, we proceeded to visualize interpolated samples on the previously mentioned evaluation datasets. Additionally, we performed VFI with all models to upscale selected video clips from 30 fps to 60 fps. To assess the quality of the generated samples, we employed metrics such as PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index). The metrics obtained from the LCM were benchmarked against those from a pre-trained LDMVFI provided by the paper authors. These experiments revealed that with low-shot sampling, our approach yielded interpolations comparable in quality to those produced by the vanilla LDM-VFI model at significantly higher-shot sampling.

Results are quantitatively compared using PSNR and SSIM, as well as qualitatively evaluated by team members using overlays in video editing software (iMovie, Microsoft Clipchamp). These results are summarized in the Table 1.

3. Results, Discussions, and Findings

Results are quantitatively compared using PSNR and SSIM, as well as qualitatively evaluated by team members using overlays in video editing software (iMovie, Microsoft

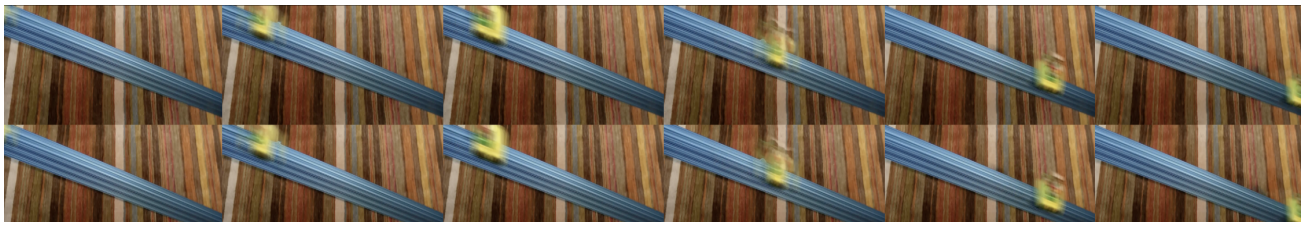

Figure 1. Visualization of video frame interpolations produced by our model at 50 DDIM sampling steps with those generated by a pre-trained LDMVFI model at 200 DDIM sampling steps.

Table 1. Model Evaluation Results on the Middlebury Optical Flow Dataset

	Beanbags		Dimetrodon	
Model	PSNR	SSIM	PSNR	SSIM
LDM-200	27.532	0.961	38.729	0.997
LCM-200	27.520	0.960	38.729	0.995
LCM-10	26.423	0.931	36.348	0.955

Clipchamp). These results are summarized in the Table 1.

For the LCM at 50 DDIM steps compared to the LDM at 200 DDIM steps, image quality metrics like PSNR and SSIM were all within 2%. The most important takeaway is our model generates similar quality frames with significantly less timesteps. Generating a frame at 50 steps takes approximately 2s compared to the 8 seconds required for high quality sampling from the LDMVFI, decreasing generation time by 400%.

4. Conclusions

In this work, we introduced the application of consistency models in video frame interpolation by distilling the state-of-the-art VFI-specific latent diffusion model into a latent consistency model.

We showed that, with close to 5 hours of V100 GPU training, our model was able to produce near-identical results to LDM-VFI, both visually and numerically, with 25% of the generation time.

This work has the potential to greatly enhance the application of powerful latent consistency models in frame interpolation and other time or resource-critical tasks.

5. Future Work

With additional time, we would have extended evaluation on more advanced visual quality metrics such as LPIPS, and a bespoke VFI metric introduced in LDM-VFI, called FloLPIPS. According to the authors, these metrics have shown superior correlation with perceptual image quality compared to PSNR and SSIM. Since the PSNR and SSIM metrics as well as the extracted visualizations were within a 2% margin of error, additional metrics would have greatly assisted in defining the exact performance of our LCM.

In addition, we would have investigated more difficult interpolation scenarios through a custom dataset. For instance, triplet sequences of the Vimeo90K dataset feature limited motion diversity and per pixel magnitude variation. Training and evaluating on large frame gaps (> 10 frames) may improve the model's flexibility to large motional and visual differences.

6. Specific Contributions Per Team Member

Table 2. Contributions Per Team Member

Team Member	**Contributions**
Mohamed Ghanem	Consistency distillation code, model training, interpolation evaluation, paper methodology and results
Rohith Nibhanupudi	Consistency distillation code, distributed training, metric evaluation, paper methodology and results
Faris Qadan	Code compilation, literature review, presentation, paper abstract, introduction, and conclusion

References

Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M., and Szeliski, R. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92:1–31, 01 2007. doi: 10.1007/s11263-010-0390-2.

Choi, M., Kim, H., Han, B., Xu, N., and Lee, K. M. Channel attention is all you need for video frame interpolation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:10663–10671, 04 2020. doi: 10.1609/aaai.v34i07.6693.

Danier, D., Zhang, F., and Bull, D. Ldmvfi: Video frame interpolation with latent diffusion models. *arXiv preprint arXiv:2303.09508*, 2023.

Ho, J. and Salimans, T. Classifier-free diffusion guidance, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=k7FuTOWMOc7.

Luo, S., Tan, Y., Huang, L., Li, J., and Zhao, H. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023.

Lyasheva, S., Rakhmankulov, R., and Shleymovich, M. Frame interpolation in video stream using optical flow methods. *Journal of Physics: Conference Series*, 1488 (1):012024, 2020. doi: 10.1088/1742-6596/1488/1/012024. URL https://dx.doi.org/10.1088/1742-6596/1488/1/012024.

Parihar, A. S., Varshney, D., Pandya, K., and Aggarwal, A. A comprehensive survey on video frame interpolation techniques. *Vis. Comput.*, 38(1): 295–319, jan 2022. ISSN 0178-2789. doi: 10.1007/s00371-020-02016-y. URL https://doi.org/10.1007/s00371-020-02016-y.

Pont-Tuset, J., Perazzi, F., Caelles, S., Arbelaez, P., Sorkine-Hornung, A., and Gool, L. V. The 2017 DAVIS challenge on video object segmentation. *CoRR*, abs/1704.00675, 2017. URL http://arxiv.org/abs/1704.00675.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2021.

Shi, C., Liu, H., Jin, J., Li, W., Li, Y., Wei, B., and Zhang, Y. Ido-vfi: Identifying dynamics via optical flow guidance for video frame interpolation with events, 2023.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *CoRR*, abs/2010.02502, 2020. URL https://arxiv.org/abs/2010.02502.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Soomro, K., Zamir, A. R., and Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. URL http://arxiv.org/abs/1212.0402.