STRUCT GROUP SEMINAR

CVPR 2024 Best Paper

Generative Image Dynamics

Zhengqi Li Richard Tucker Noah Snavely Aleksander Holynski
Google Research

Presenter: Wenshuo Gao

2024.09.03

Outline

- **Author**
- **Background**
- **Method**
- **Experiments**

Outline

- Author
- **Background**
- Method
- Experiments

Background: Animating an Image

Task: Generate a video based on an input image

Method 1: Directly generate **raw RGB pixel volume**:

- Computationally expensive
- Inconsistency

Input Image

Result from Runway

Background: Animating an Image

Task: Generate a video based on an input image

Method 2: **Moving the image content around** according to motion:

- Since most pixel information are **shared** across the video
- Consistency
- Utilize **optical flow**

Input Image

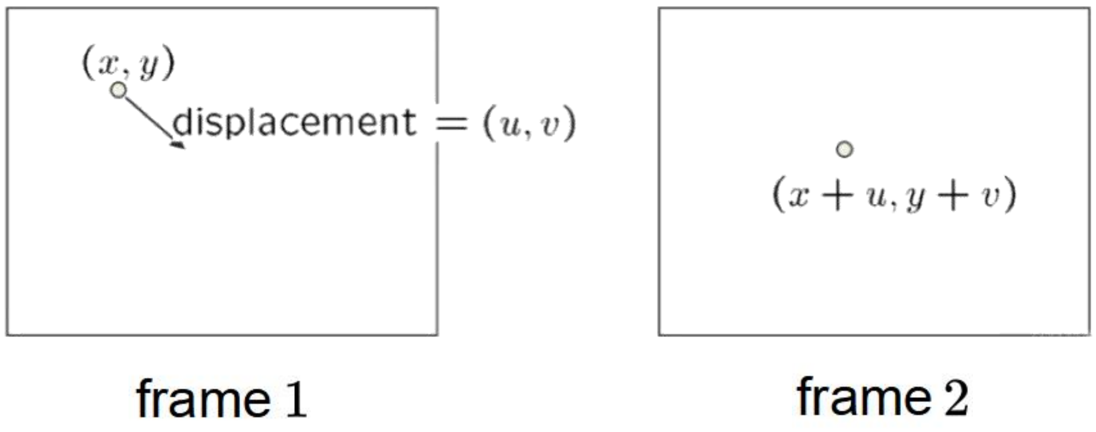
Result from Generative Image Dynamics

Background: Optical Flow

Optical Flow

- Description of displacement field
- $F(\mathbf{p}) : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is to describe the **relative position** of a pixel from frame 1 in position $\mathbf{p}$ to frame 2:

$$I_1(\mathbf{p}) = I_2(\mathbf{p} + F(\mathbf{p}))$$



Background: Optical Flow

Optical Flow

Background: Optical Flow

Estimation of Optical Flow

- Lucas-Kanade / Horn-Schunck method:
  Assume similar flows in nearby pixels
  Solve the equation for all $\mathbf{p}$:

$$I_1(\mathbf{p}) = I_2(\mathbf{p} + F(\mathbf{p}))$$

  (Details are shown in *Experiments* section)

- Machine learning method:
  Train models from video datasets

Background: Optical Flow

Recover Video from Optical Flow

$$I_1(\mathbf{p}) = I_2(\mathbf{p} + F(\mathbf{p}))$$

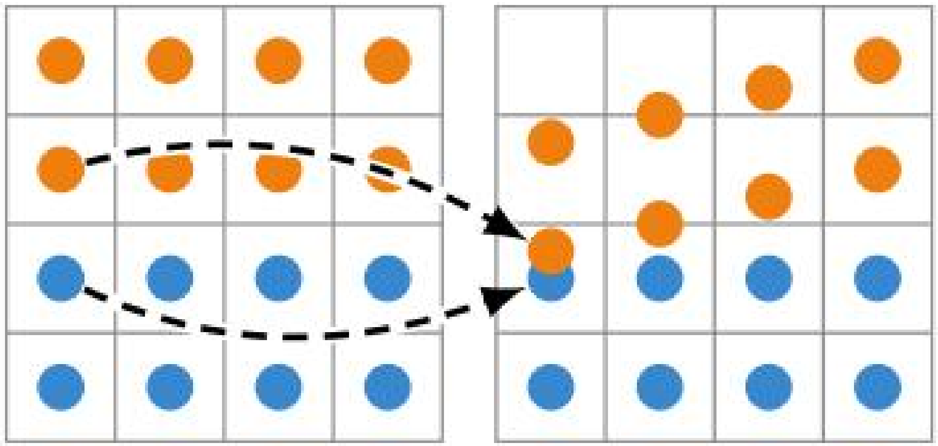
Handling conflicts

Solutions:

(a) Average splatting

(b) Linear splatting

(c) Softmax splatting



Background: Optical Flow

Recover Video from Optical Flow

Handling conflicts

Solutions:

(a) Average splatting:

- Directly calculate the average of colors
- Blend overlapping regions

(b) Linear splatting

(c) Softmax splatting

Background: Optical Flow

Recover Video from Optical Flow

Handling conflicts

Solutions:

(a) Average splatting

(b) Linear splatting:

- Calculate the **weighted** average
- High weight for **foreground** parts
- Low weight for **background** parts
- Require depth map

(c) Softmax splatting

Background: Optical Flow

Recover Video from Optical Flow

Handling conflicts

Solutions:

(a) Average splatting

(b) Linear splatting

(c) Softmax splatting:

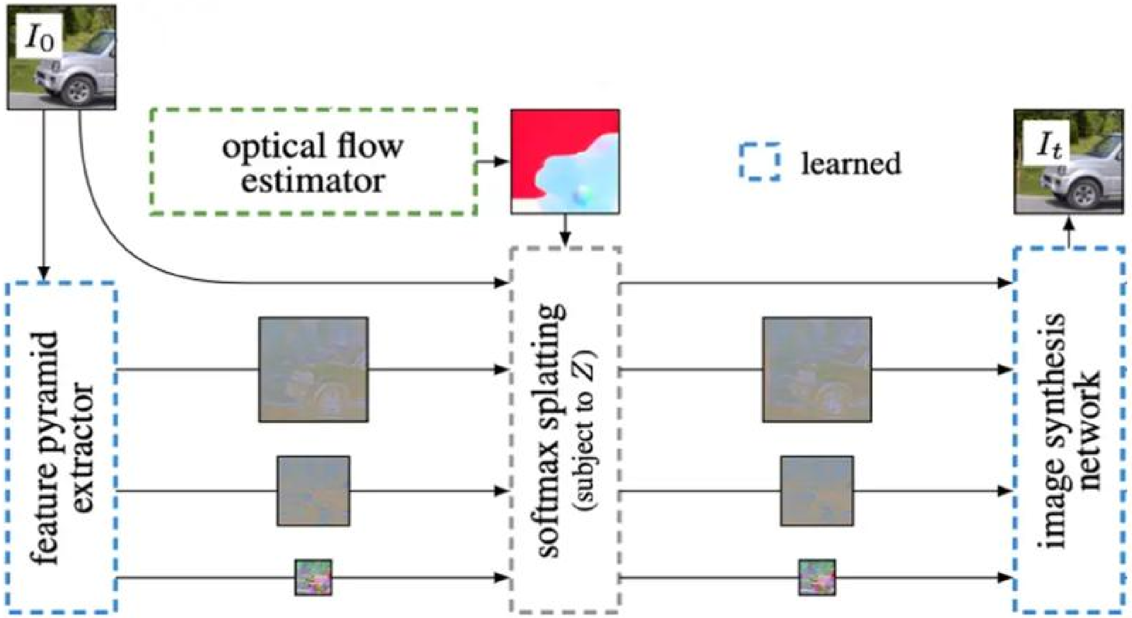
- Calculate the **weighted** average
- High weight for **moving** parts
- Low weight for **still** parts
- The weight function is trained in a network or computed from motion

Background: Optical Flow

Recover Video from Optical Flow

Feature level softmax splatting

Render smoother results

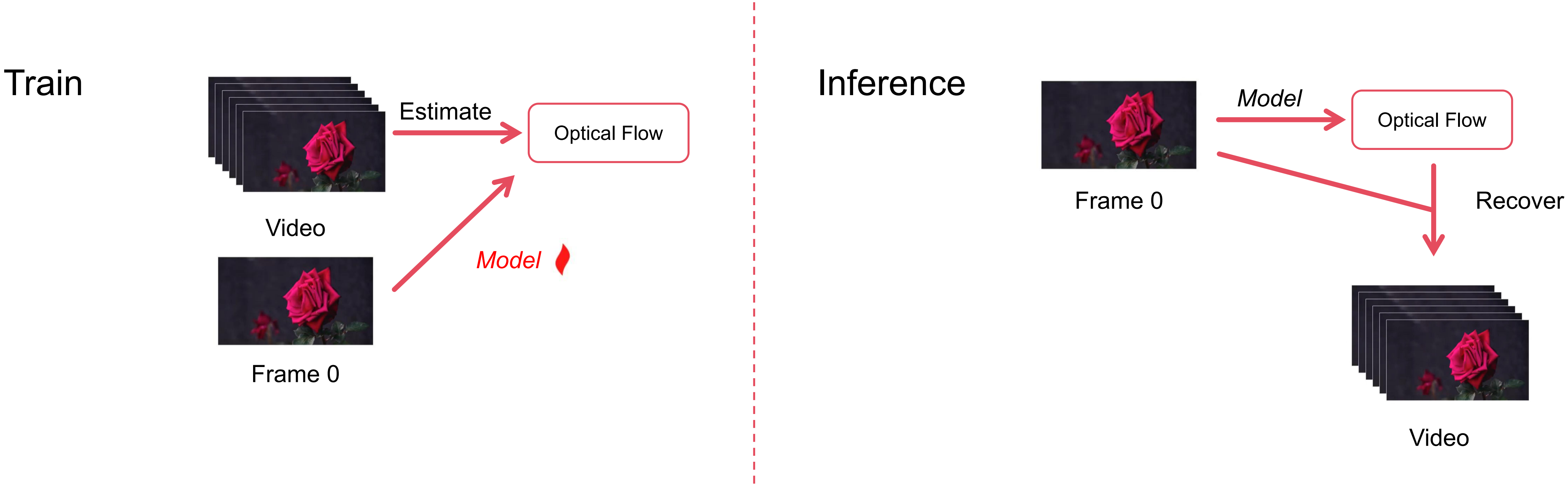


Softmax Splatting for Video Frame Interpolation. CVPR, 2020.

Background: Optical Flow

Generating Video using Optical Flow

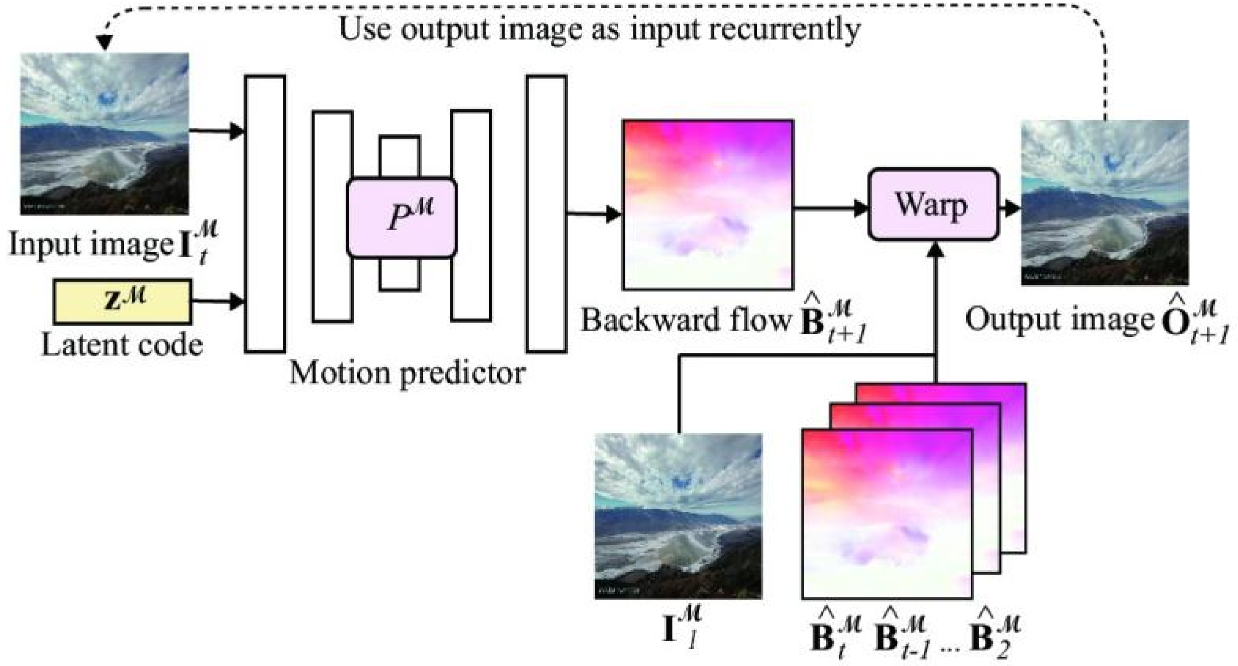
Using neural networks to predict optical flow from an image



Background: Optical Flow

Generating Video using Optical Flow

Using U-Net to predict optical flow

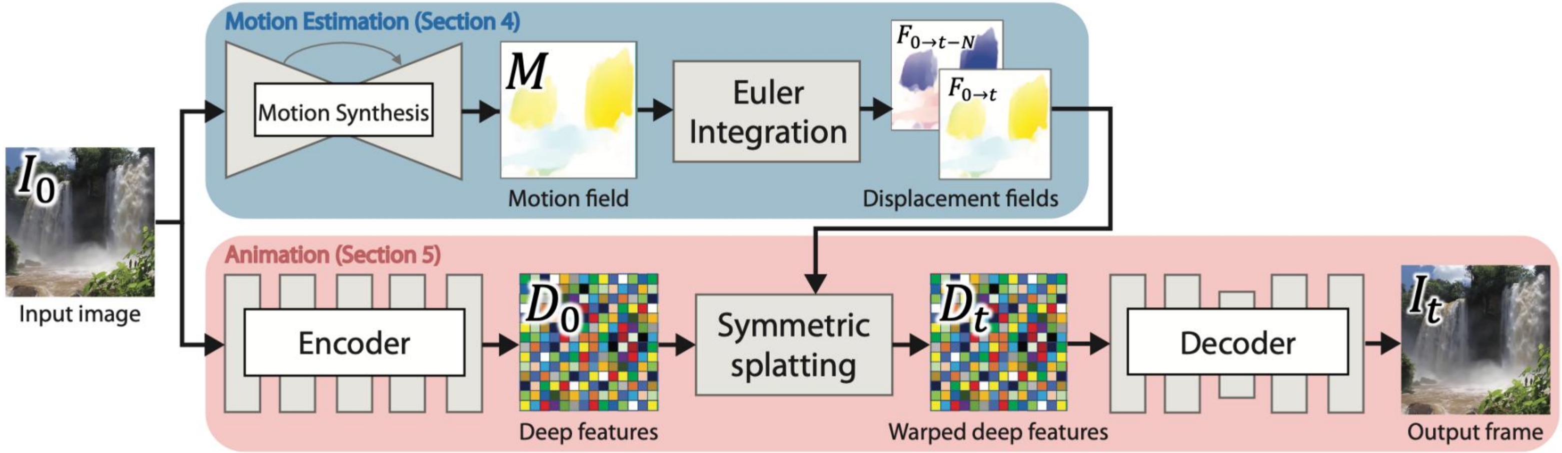


Animating Landscape: Self-Supervised Learning of Decoupled Motion and Appearance for Single-Image Video Synthesis. arXiv preprint, 2019.

Background: Optical Flow

Generating Video using Optical Flow

Feature level splatting



Animating Pictures with Eulerian Motion Fields. CVPR, 2021.

Background: Optical Flow

Generating Video using Optical Flow

Limitation: Individual $t \in \{1, \ldots, T\}$ across video frames

- Computationally expensive
- Temporal inconsistency

Solution:

(1) Autoregressive

Using frame $t-3, t-2, t-1$ to predict frame t

(2) Timestep embedding

Using embedded t as input of model

Background: Optical Flow

Generating Video using Optical Flow

Limitation: Individual $t \in \{1, ..., T\}$ across video frames

- Computationally expensive
- Temporal inconsistency

Solution:

(3) **Spectral volume**

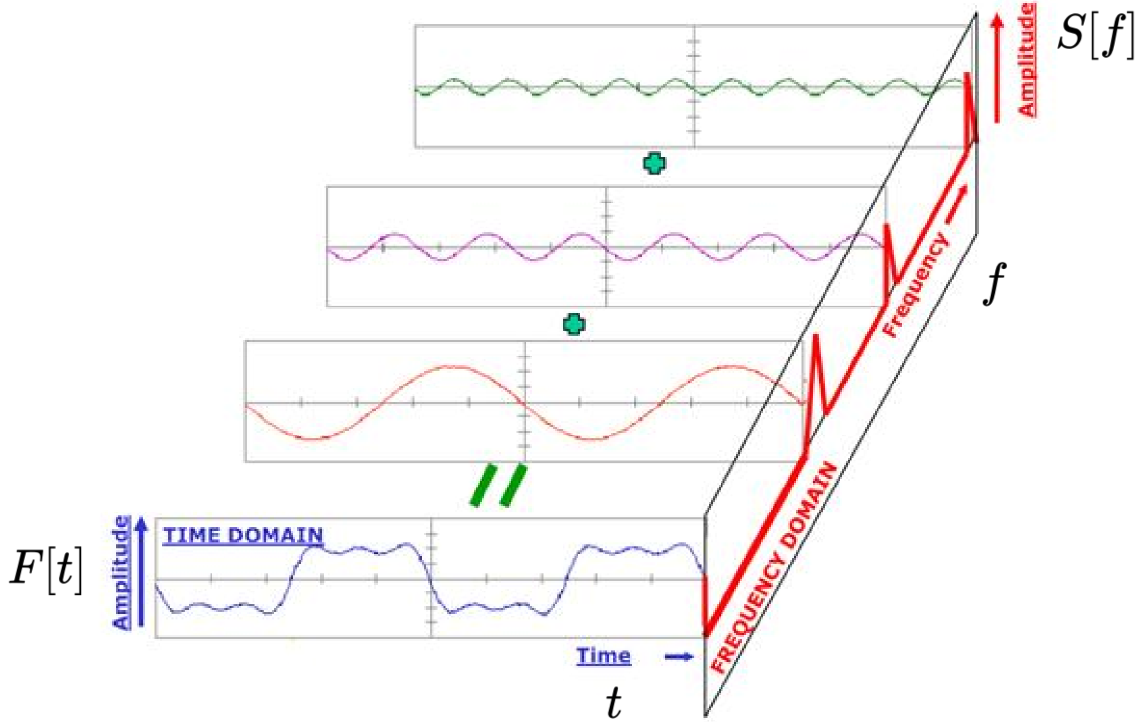
- The **frequency form** of motion
- The **Discrete Fourier Transform** of optical flow
- Capable of separating high-/low-frequency information
- Motion composed of summation of cosine curves → consistency

Background: Discrete Fourier Transform

Discrete Fourier Transform (DFT)

Decomposes functions into summation of **cosine curves**

Transforms **time-domain** data into **frequency-domain** information



$$F[t] = \frac{1}{T}\sum_{f=0}^{T-1} S[f]e^{i2\pi ft/T}$$

$$S[f] = \sum_{t=0}^{T-1} F[t]e^{-i2\pi ft/T}$$

$$t, f = 0, 1, \ldots, T-1$$

Background: Spectral Volume

Spectral Volume

For a T-frame video, optical flow: $\mathcal{F}(\mathbf{p}) = \{F_t(\mathbf{p}) | t = 1, \ldots, T\}$

DFT transforms optical flow into spectral volume with K frequencies

$$\mathcal{S}(\mathbf{p}) = \{S_{f_k}(\mathbf{p}) | k = 0, \ldots, K-1\}$$

where $\mathcal{S}(\mathbf{p}) = DFT(\mathcal{F}(\mathbf{p}))$

Note that if $K << T$, the motion is stored in **less parameters**

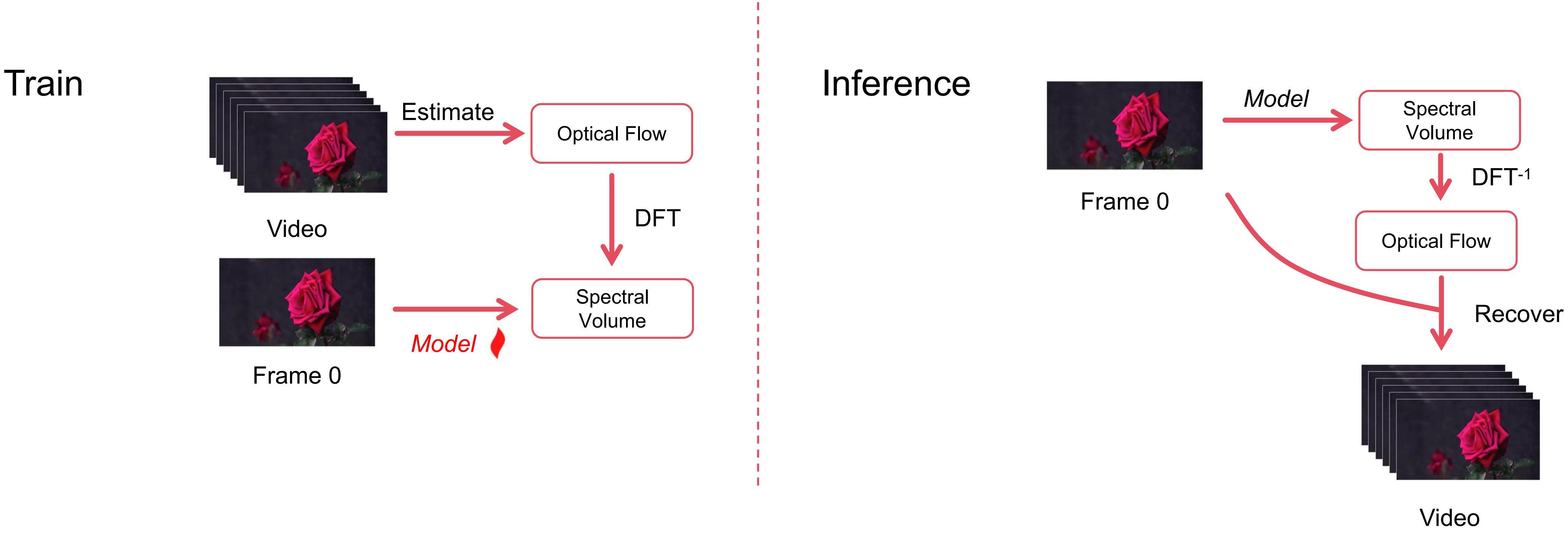
Outline

- Author
- Background
- **Method**
- Experiments

Method

How to Generate Video using Spectral Volume

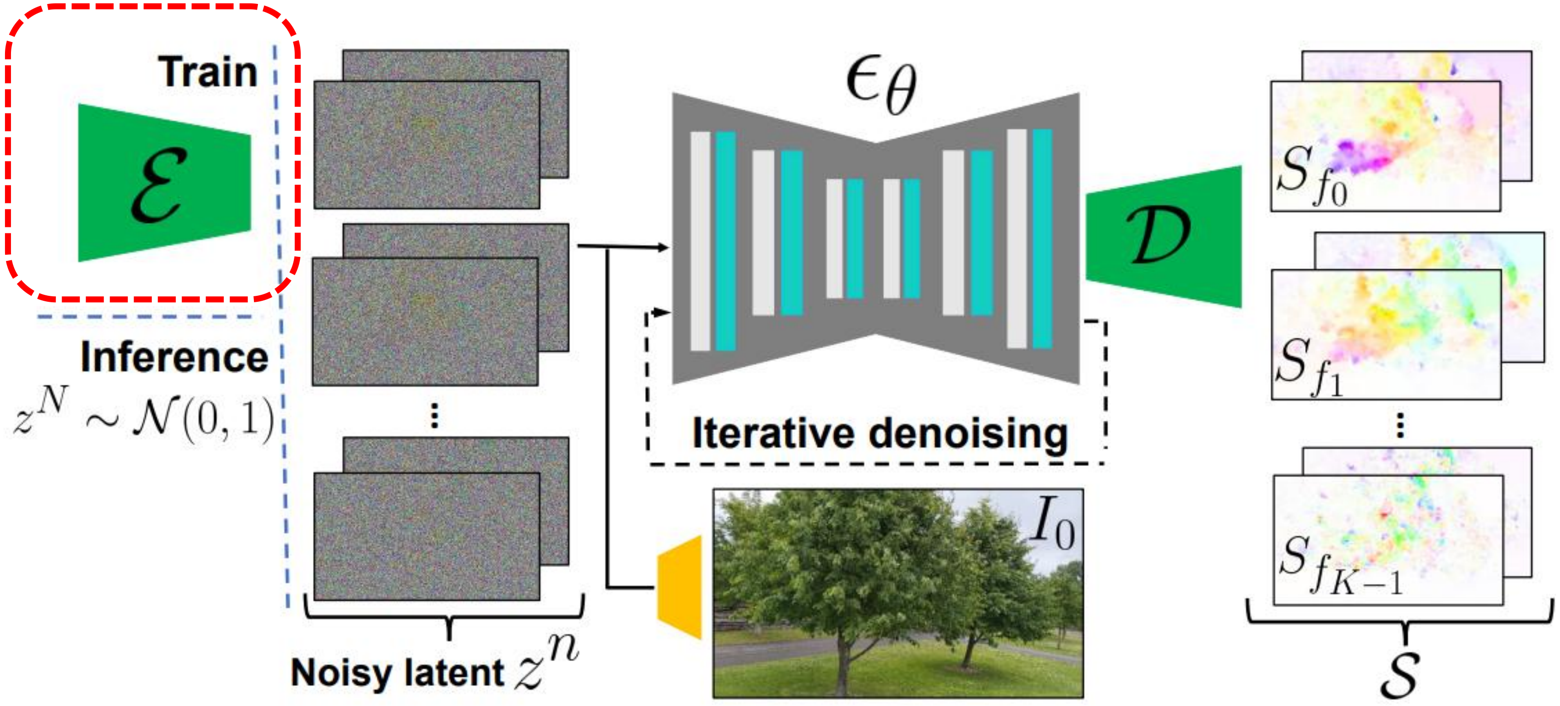
Using neural networks to predict spectral volume from an image



Method

Predict Spectral Volume by Latent Diffusion Model (LDM)

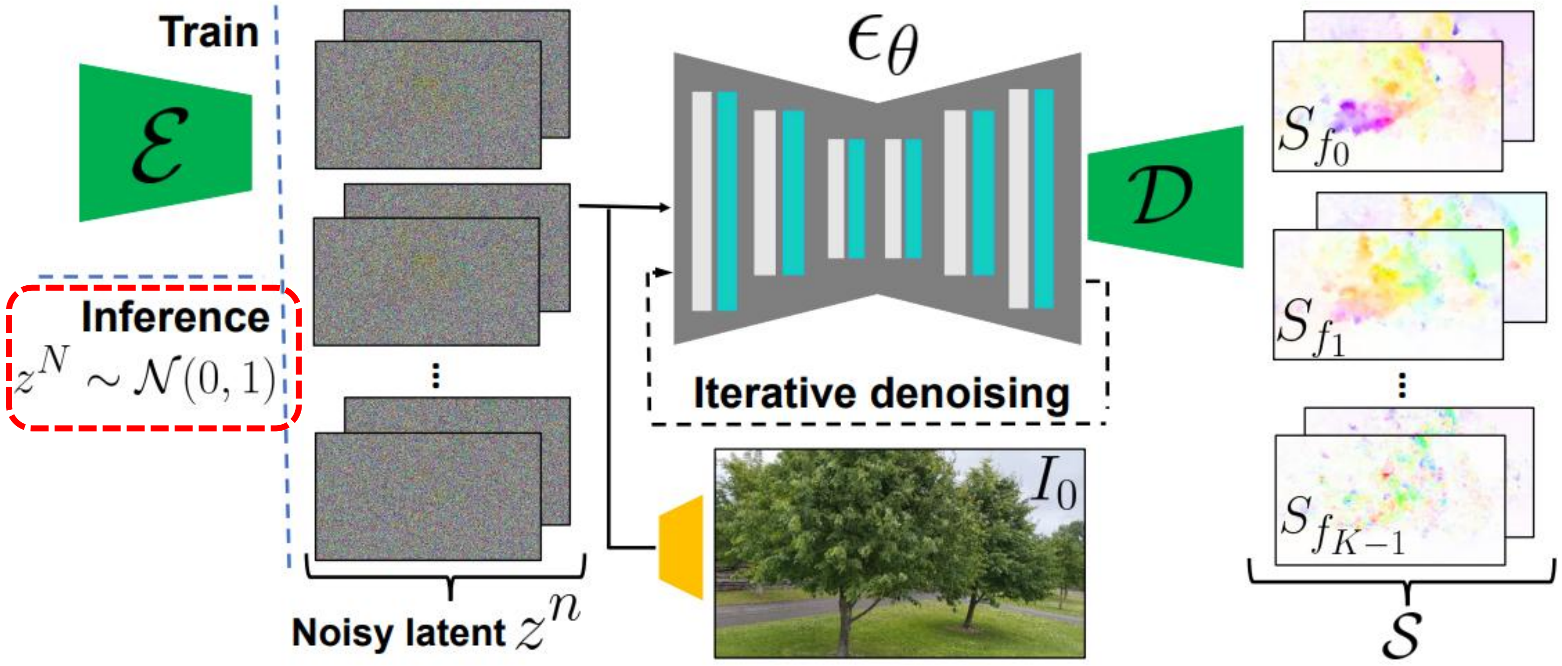
Input during training: Noisy latent features encoded from GT spectral volume



Method

Predict Spectral Volume by Latent Diffusion Model (LDM)

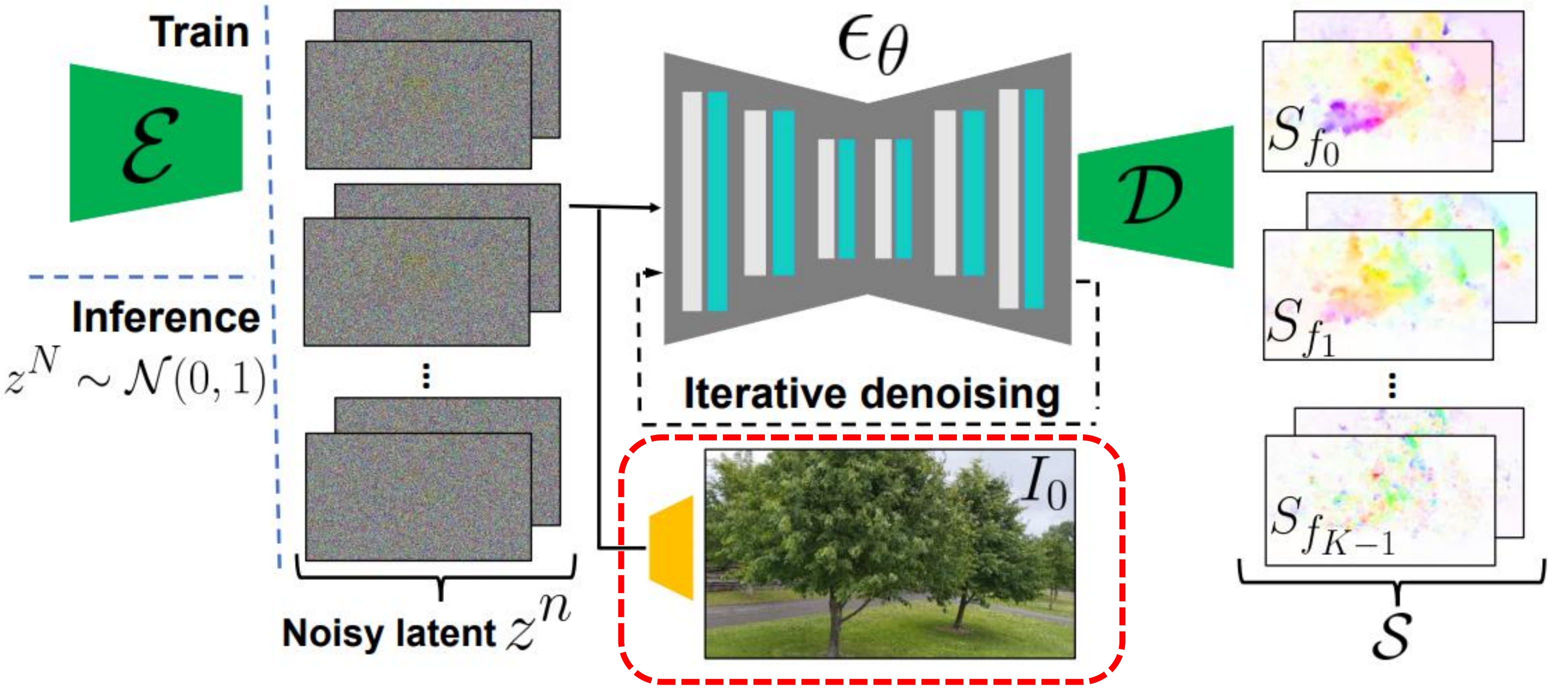
Input during inferencing: Gaussian noise



Method

Predict Spectral Volume by Latent Diffusion Model (LDM)

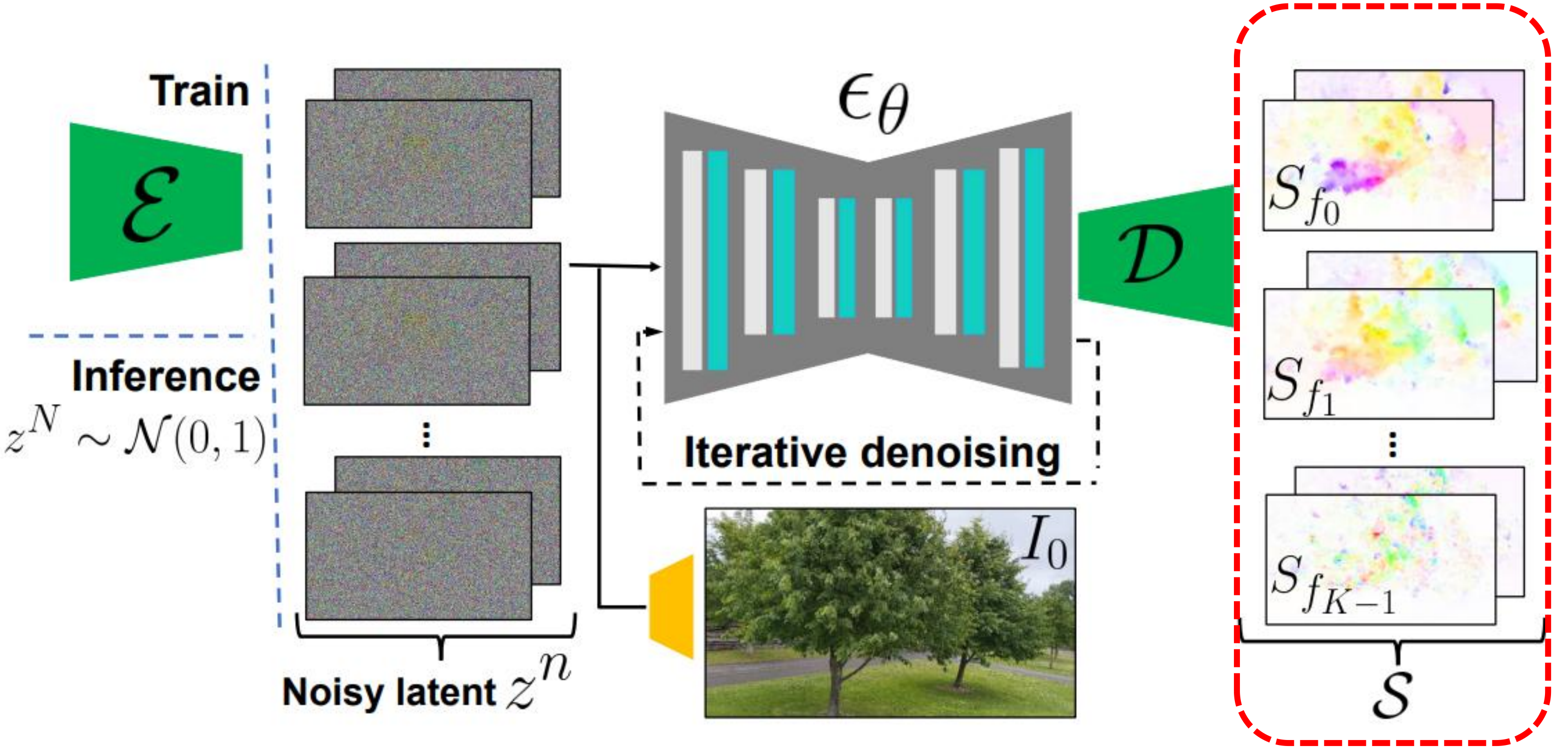
Denoising: Downsampled initial frame **as condition**



Method

Predict Spectral Volume by Latent Diffusion Model (LDM)

Output: Denoised features decoded to produce 4K-channel spectral volume



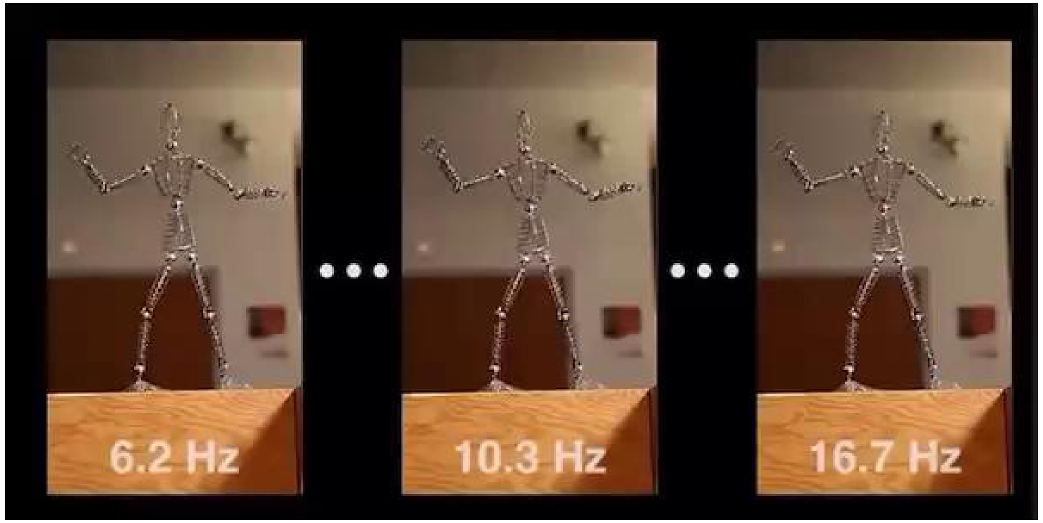
Method

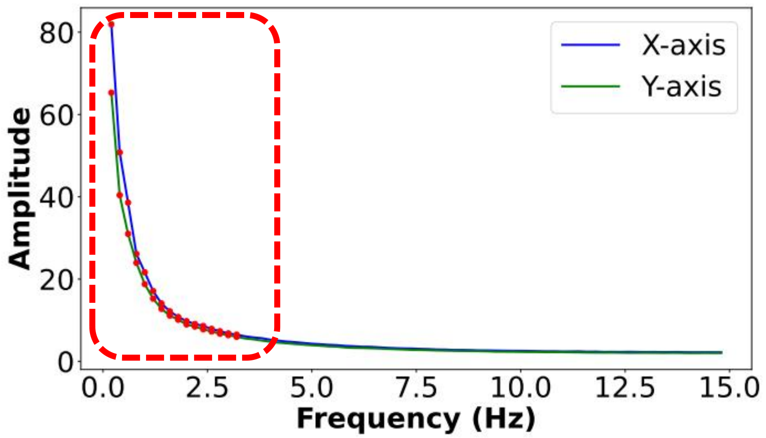
Predict Spectral Volume by Latent Diffusion Model (LDM)

How to choose frequencies?

Natural oscillations are composed mainly of **low-frequency** components

Keep the lowest 16 frequencies ($K = 16$) is sufficient



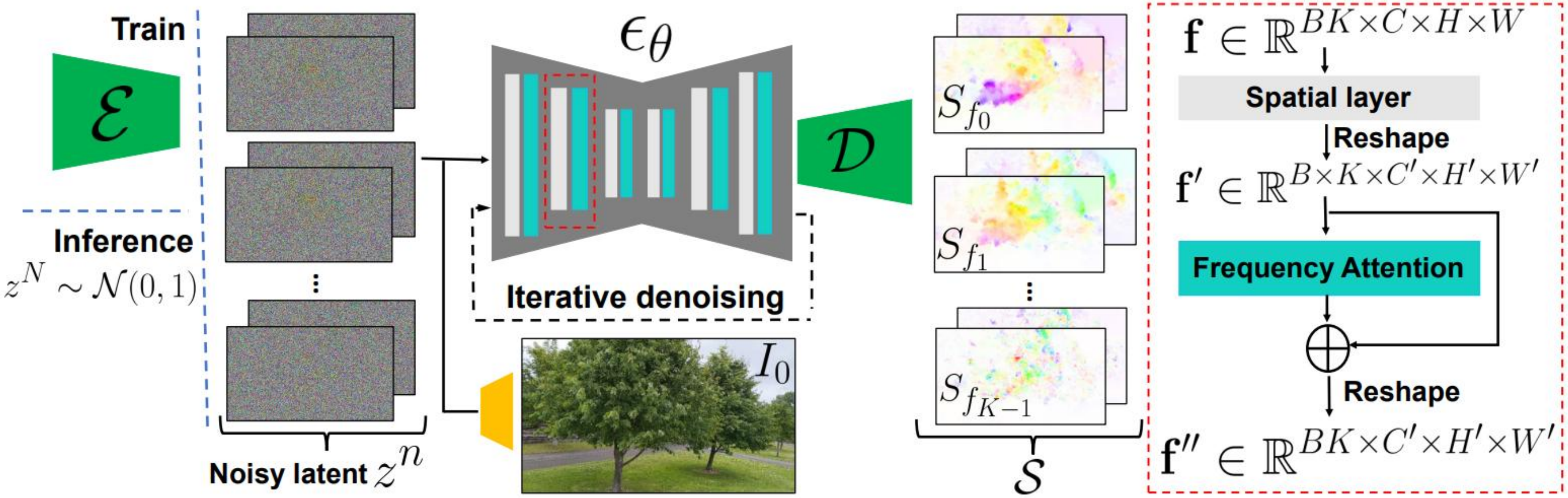


Method

Predict Spectral Volume by Latent Diffusion Model (LDM)

Directly predict 4K-channel spectral volume: computational expensive / inconsistency

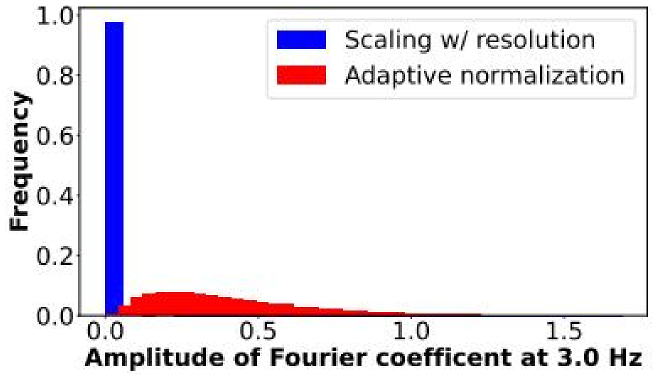
Solution: frequency embedding, **as condition** (cross attention)



Method

Predict Spectral Volume by Latent Diffusion Model (LDM)

During denoising, data should be ranged in $[-1, 1]$



Solution:

(1) Directly scaling according to resolution:

- Coefficients at higher frequencies close to 0

(2) Adaptive normalization:

- Normalizes by using statistics (like the 95th percentile) from training data
- Coefficients distribute more evenly

$$S'_{f_k}(\mathbf{p}) = \sqrt{\left|\frac{S_{f_k}(\mathbf{p})}{s_{f_k}}\right|}$$

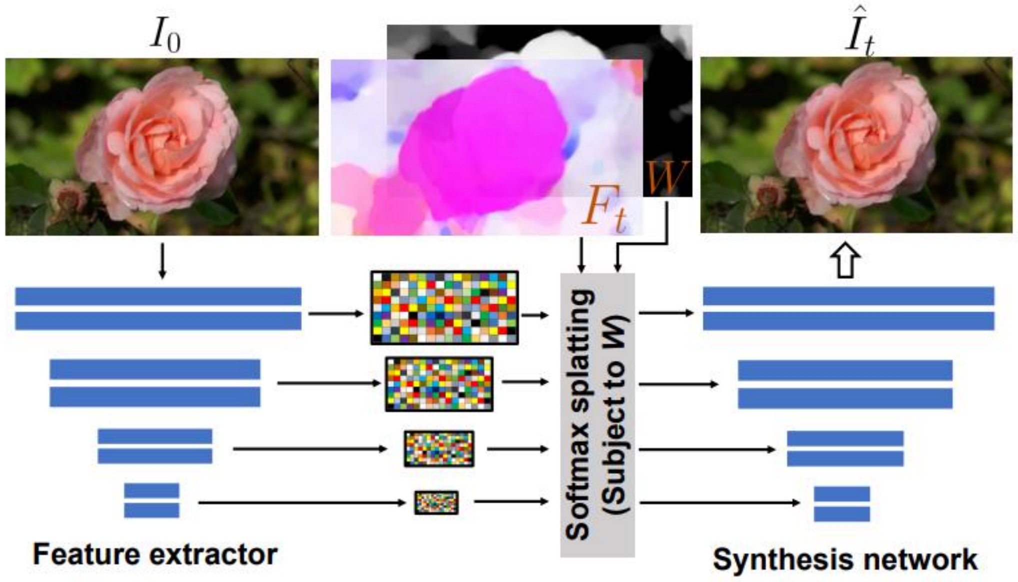
Method

Recover Video from Spectral Volume

Calculate optical flow:

$$\mathcal{F}(\mathbf{p}) = DFT^{-1}(\mathcal{S}(\mathbf{p}))$$

Recover video from optical flow using softmax splatting:

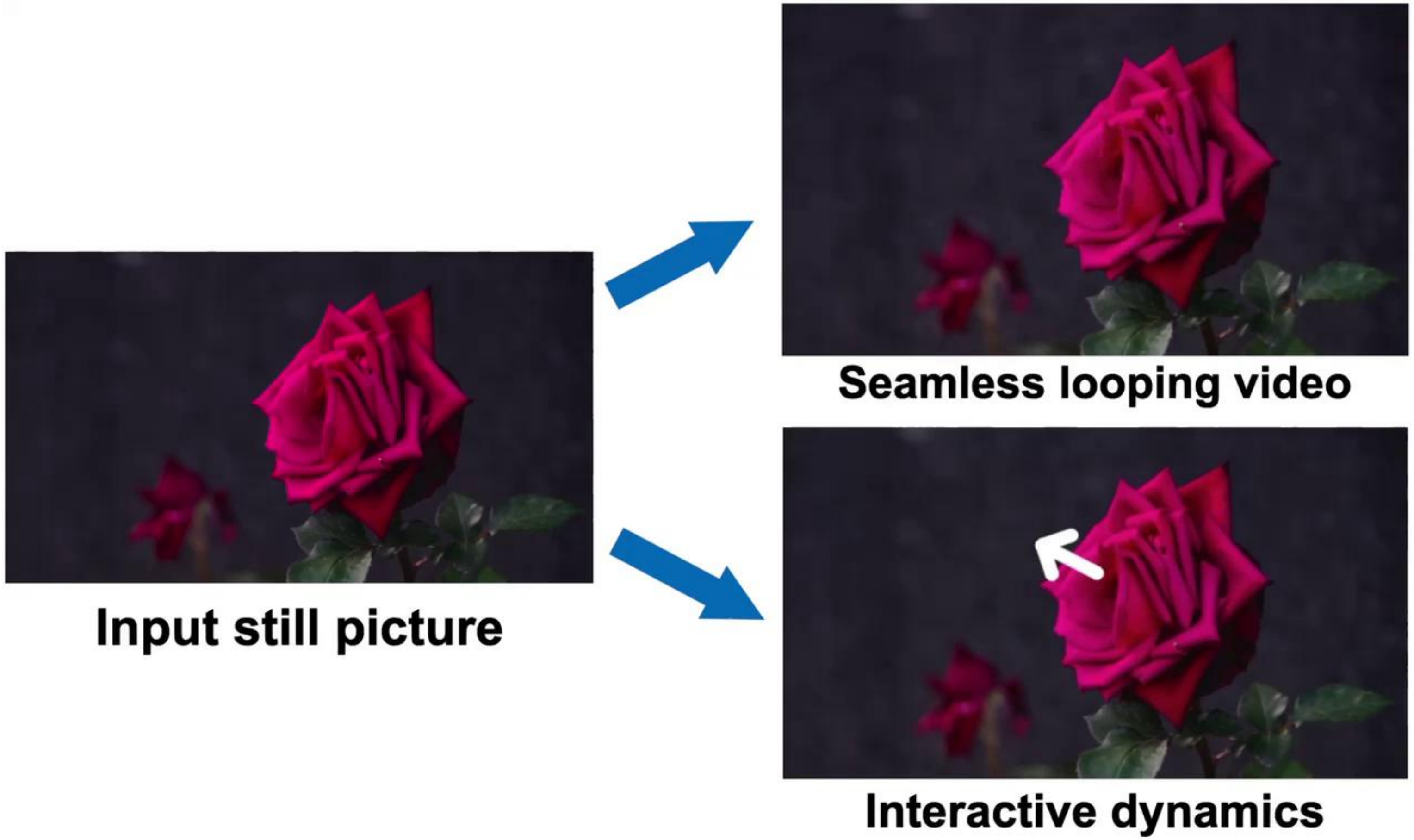


The weight function is calculated by:

$$W(\mathbf{p}) = \frac{1}{T} \sum_t ||F_t(\mathbf{p})||_2$$

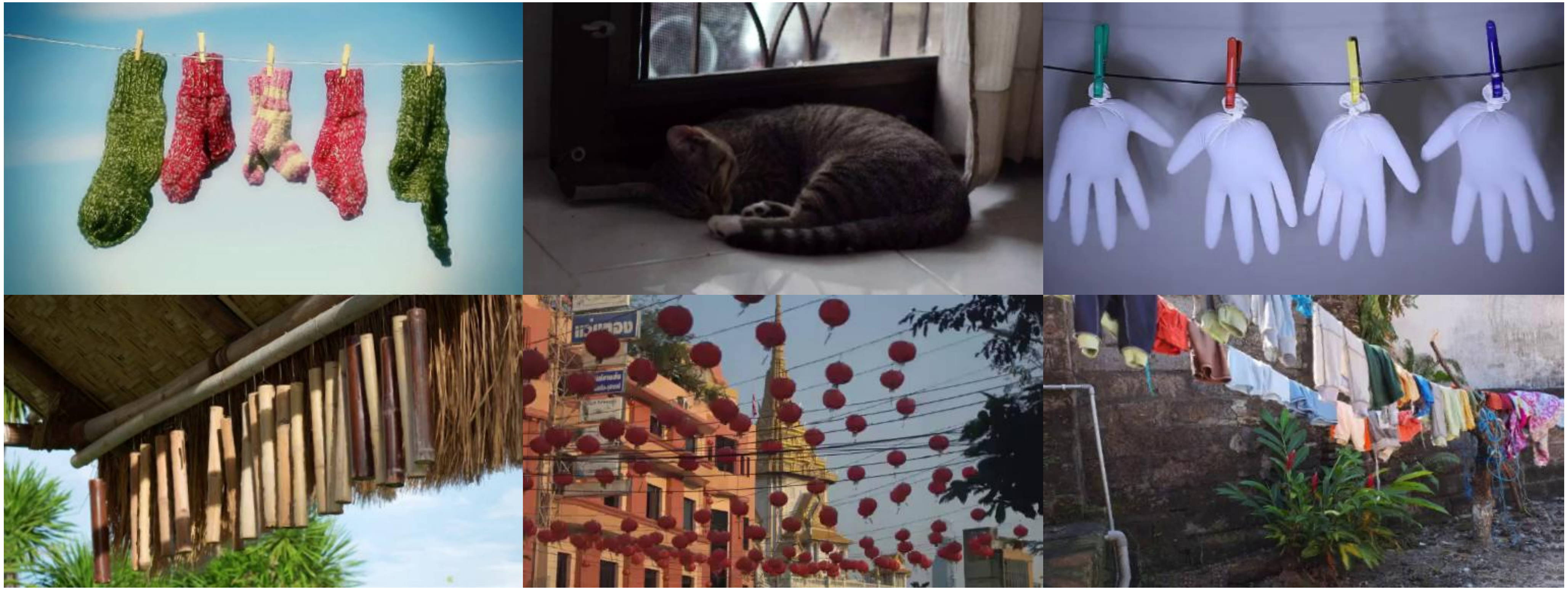
(a) Average-splat (b) Learned W (c) W from motion

Application



Application: Seamless Looping Video

Application: Seamless Looping Video



Application: Interactive Dynamics

$$||\mathbf{q}_{f_j}(0)|| = ||\frac{\mathbf{f}(0)}{||\mathbf{f}(0)||_2} \cdot S_{f_j}||_2 \qquad \phi_{\text{drag}}(\mathbf{q}_{f_j}(0)) = -\phi(\frac{\mathbf{f}(0)}{||\mathbf{f}(0)||_2} \cdot S_{f_j})$$

See supplementary material p.1

Outline

- Author
- Background
- Method
- **Experiments**

Experiments: Data

Collected 3000+ natural scenes exhibiting oscillatory motions

Extracted GT motions from a classical flow method

(DL-based flow method: too smooth)

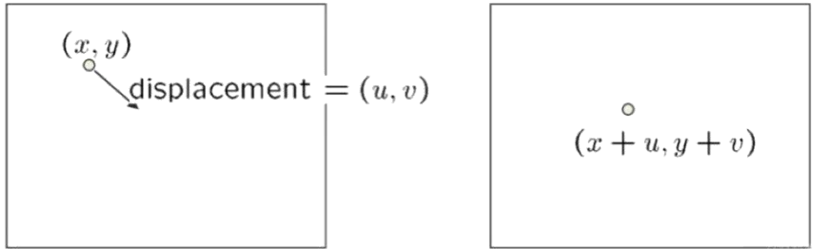
$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t)$$

$$I(x + \Delta x, y + \Delta y, t + \Delta t) \approx I(x, y, t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t$$

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0$$

$$I_x u + I_y v + I_t = 0$$

$$E(u, v) = \iint \left[(I_x u + I_y v + I_t)^2 + \alpha^2 \left(|\nabla u|^2 + |\nabla v|^2 \right) \right] dx\, dy$$



Experiments: Quantitative

Metrics

- **Frechet Inception Distance** (FID)
- **Kernel Inception Distance** (KID)

  distance between the distributions of generated frames and GT frames
- **Frechet Video Distance** (FVD, FVD_{32})
- **Dynamic Texture Frechet Video Distance** (DTFVD, $DTFVD_{32}$)

  reflect synthesis quality for the natural oscillation motions

	Image Synthesis		Video Synthesis			
Method	FID	KID	FVD	FVD_{32}	DTFVD	$DTFVD_{32}$
TATS	65.8	1.67	265.6	419.6	22.6	40.7
Stochastic I2V	68.3	3.12	253.5	320.9	16.7	41.7
MCVD	63.4	2.97	208.6	270.4	19.5	53.9
LFDM	47.6	1.70	187.5	254.3	13.0	45.6
DMVFN	37.9	1.09	206.5	316.3	11.2	54.5
Endo *et al.*	10.4	0.19	166.0	231.6	5.35	65.1
Holynski *et al.*	11.2	0.20	179.0	253.7	7.23	46.8
Ours	**4.03**	**0.08**	**47.1**	**62.9**	**2.53**	**6.75**

Experiments: Quantitative

Ablation

Retaining of frequencies

Method	Image Synthesis FID	KID	Video Synthesis FVD	FVD_{32}	DTFVD	DTFVD_{32}
Repeat I_0	-	-	237.5	316.7	5.30	45.6
$K = 4$	3.92	0.07	60.3	78.4	3.12	8.59
$K = 8$	3.95	0.07	52.1	68.7	2.71	7.37
$K = 24$	4.09	0.08	48.2	65.1	2.50	6.94
w/o adaptive norm.	4.53	0.09	62.7	80.1	3.16	8.19
Independent pred.	4.00	0.08	52.5	71.3	2.70	7.40
Volume pred.	4.74	0.09	53.7	71.1	2.83	7.79
Baseline splat	4.25	0.09	49.5	66.8	2.83	7.27
Full ($K = 16$)	4.03	0.08	47.1	62.9	2.53	6.75

Experiments: Quantitative

Ablation

Scaling according to resolution

Method	Image Synthesis FID	KID	Video Synthesis FVD	FVD_{32}	DTFVD	$DTFVD_{32}$
Repeat I_0	-	-	237.5	316.7	5.30	45.6
$K = 4$	3.92	0.07	60.3	78.4	3.12	8.59
$K = 8$	3.95	0.07	52.1	68.7	2.71	7.37
$K = 24$	4.09	0.08	48.2	65.1	2.50	6.94
w/o adaptive norm.	4.53	0.09	62.7	80.1	3.16	8.19
Independent pred.	4.00	0.08	52.5	71.3	2.70	7.40
Volume pred.	4.74	0.09	53.7	71.1	2.83	7.79
Baseline splat	4.25	0.09	49.5	66.8	2.83	7.27
Full ($K = 16$)	4.03	0.08	47.1	62.9	2.53	6.75

Experiments: Quantitative

Ablation

No frequency embedding

Method	Image Synthesis FID	KID	Video Synthesis FVD	FVD_{32}	DTFVD	$DTFVD_{32}$
Repeat I_0	-	-	237.5	316.7	5.30	45.6
$K = 4$	3.92	0.07	60.3	78.4	3.12	8.59
$K = 8$	3.95	0.07	52.1	68.7	2.71	7.37
$K = 24$	4.09	0.08	48.2	65.1	2.50	6.94
w/o adaptive norm.	4.53	0.09	62.7	80.1	3.16	8.19
Independent pred.	4.00	0.08	52.5	71.3	2.70	7.40
Volume pred.	4.74	0.09	53.7	71.1	2.83	7.79
Baseline splat	4.25	0.09	49.5	66.8	2.83	7.27
Full ($K = 16$)	4.03	0.08	47.1	62.9	2.53	6.75

Experiments: Quantitative

Ablation

No latent

Method	Image Synthesis FID	KID	Video Synthesis FVD	FVD_{32}	DTFVD	DTFVD_{32}
Repeat I_0	-	-	237.5	316.7	5.30	45.6
$K = 4$	3.92	0.07	60.3	78.4	3.12	8.59
$K = 8$	3.95	0.07	52.1	68.7	2.71	7.37
$K = 24$	4.09	0.08	48.2	65.1	2.50	6.94
w/o adaptive norm.	4.53	0.09	62.7	80.1	3.16	8.19
Independent pred.	4.00	0.08	52.5	71.3	2.70	7.40
Volume pred.	4.74	0.09	53.7	71.1	2.83	7.79
Baseline splat	4.25	0.09	49.5	66.8	2.83	7.27
Full ($K = 16$)	4.03	0.08	47.1	62.9	2.53	6.75

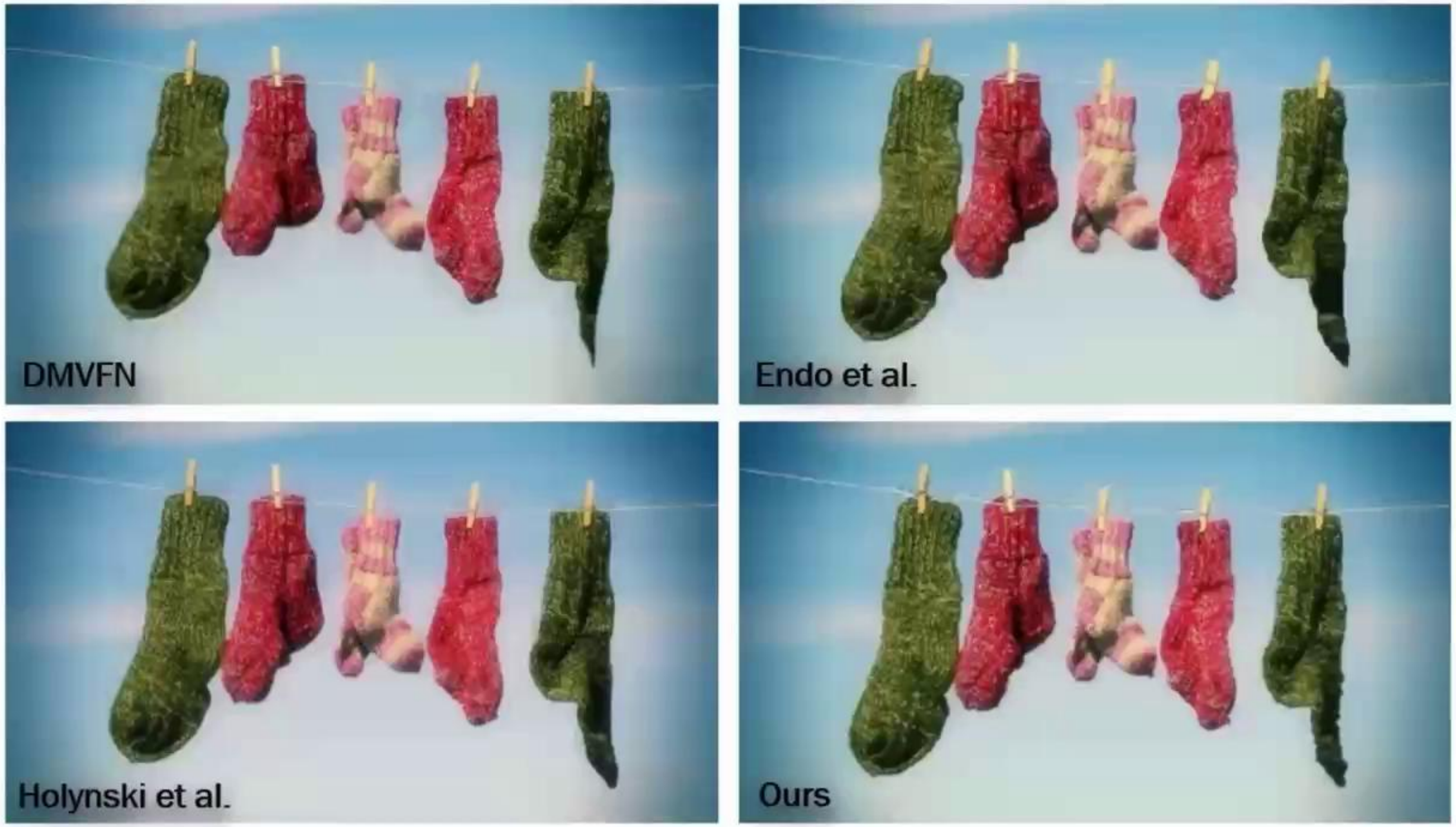
Experiments: Quantitative

Ablation

Learnable weights in softmax splatting

Method	Image Synthesis FID	KID	Video Synthesis FVD	FVD_{32}	DTFVD	DTFVD_{32}
Repeat I_0	-	-	237.5	316.7	5.30	45.6
$K = 4$	3.92	0.07	60.3	78.4	3.12	8.59
$K = 8$	3.95	0.07	52.1	68.7	2.71	7.37
$K = 24$	4.09	0.08	48.2	65.1	2.50	6.94
w/o adaptive norm.	4.53	0.09	62.7	80.1	3.16	8.19
Independent pred.	4.00	0.08	52.5	71.3	2.70	7.40
Volume pred.	4.74	0.09	53.7	71.1	2.83	7.79
Baseline splat	4.25	0.09	49.5	66.8	2.83	7.27
Full ($K = 16$)	4.03	0.08	47.1	62.9	2.53	6.75

Experiments: Qualitative



Conclusion

1. A new approach for modeling natural oscillation dynamics from a single still picture
2. Produces photo-realistic animations from a single picture and significantly outperforms prior baselines
3. Demonstrates potential to enable several downstream applications such as creating seamlessly looping or interactive image dynamics

Conclusion

Limitation:

The model is not capable of generating:

(a) non-oscillating motions

(b) high-frequency oscillations (only low-frequencies were kept)

(c) contents not covered by dataset

Discussion

1. Creative combination of existing works

    Require broad foundations and insights

2. Fancy results

3. Interesting downstream applications

    Interactive image dynamics

Thanks for listening!