『2025 하계 세미나』

Motion modeling



Pusan National University
Dept. of Electronics Engineering
Computer Vision & Signal Processing Lab.



Presented By
문형주

Outline

- Motivation
- Generative Image Dynamics [CVPR 24 Best Paper Award]
- Motion-I2V: Consistent and Controllable Image-to-Video Generation with Explicit Motion Modeling [ACM SIGGRAPH 24]

Motivation

- WACV2026
  - Topic : 3D motion generation
    - 3D object에 대해서 자연스러운 motion을 생성하고자 함

∴ 3D motion field를 생성하기 위해 multi view의 2D motion map이 필요함



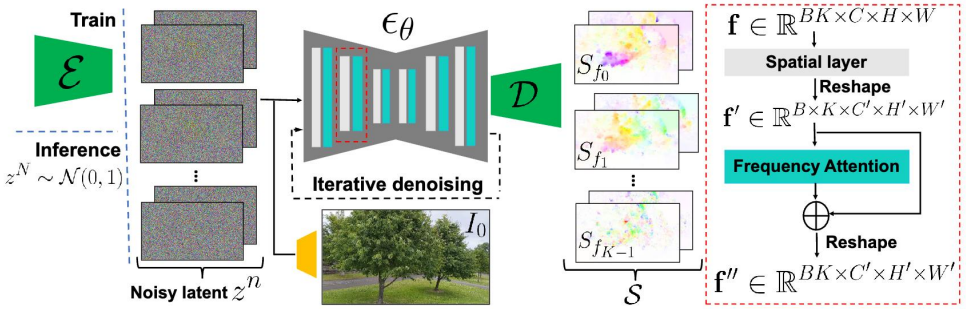
Generative Image Dynamics [CVPR 24 Best Paper Award]

Introduction

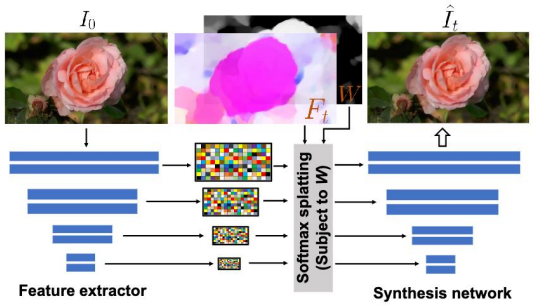
- Problem & Solution
  - Problem
    - Motion texture를 이용해 비디오를 생성하고자 할 때, 한 가지 선택지는 입력 이미지로부터 시간 영역의 motion texture를 직접 예측하는 것
      - T개의 출력 프레임을 생성한다면 T개의 displacement를 예측해야 함
      - 큰 출력 표현을 예측하는 것을 피하기 위해 많은 기존 애니메이션 기법들은 비디오 프레임들을 auto-regressively하게 생성하거나, 각 미래 프레임을 독립적으로 time embedding을 이용하여 예측
        - ✓두 전략 모두 생성된 비디오의 long-term temporal consistency를 완전히 보장하지 못 함
  - Proposed method
    - 자연스러운 움직임들은 서로 다른 주파수, 진폭, 위상을 가진 소수의 조화 진동자(harmonic oscillators)들의 중첩으로 설명될 수 있음
      - 비디오에서의 움직임을 효율적으로 표현하는 frequency space representation 인 spectral volume 을 채택
        - ✓spectral volume은 비디오에서의 픽셀별 trajectory를 temporal Fourier transform한 것임

Method

- Overview
  - Goal
    - 하나의 Static image가 주어졌을 때, 나무, 꽃과 같은 oscillatory motions가 포함된 비디오를 생성
    - 두 개의 모듈로 구성
      - Motion prediction module
      - Image-based rendering module



< Motion prediction module >



< Rendering module >

< Overview of Generative Image Dynamics >

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

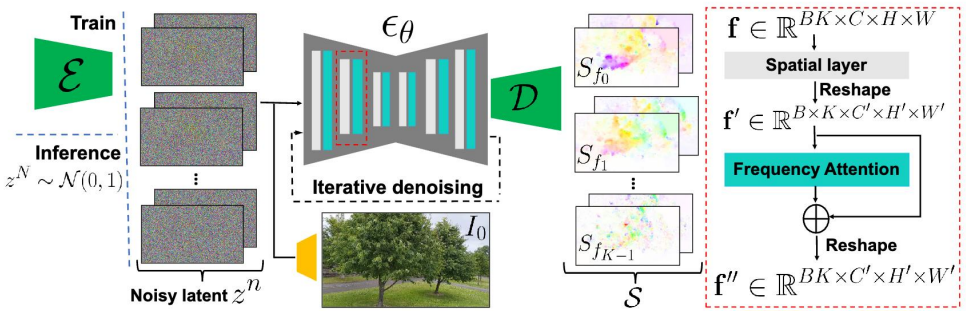
- Overview
  - Pipeline
    - 입력 image I_0에 대해 LDM(Latent Diffusion Model)을 사용하여 spectral volume S를 예측

$$S = (S_{f_0}, S_{f_1}, \dots, S_{f_{K-1}})$$

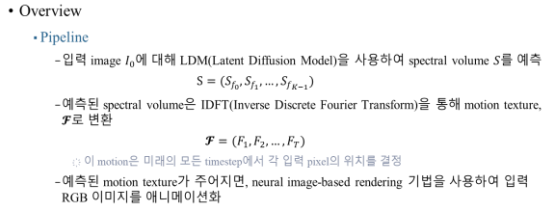
    - 예측된 spectral volume은 IDFT(Inverse Discrete Fourier Transform)을 통해 motion texture, $\mathcal{F}$로 변환

$$\mathcal{F} = (F_1, F_2, \dots, F_T)$$

      - 이 motion은 미래의 모든 timestep에서 각 입력 pixel의 위치를 결정
    - 예측된 motion texture가 주어지면, neural image-based rendering 기법을 사용하여 입력 RGB 이미지를 애니메이션화



< Motion prediction module >



< Rendering module >

< Overview of Generative Image Dynamics >

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

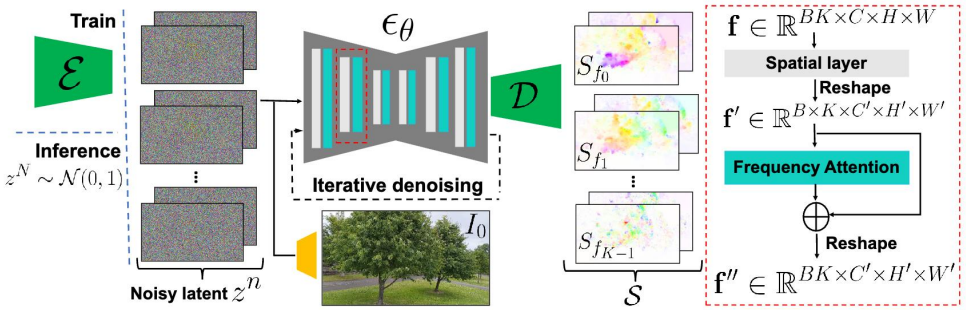
- Overview
  - Pipeline
    - 입력 image I_0에 대해 LDM(Latent Diffusion Model)을 사용하여 spectral volume S를 예측

$$S = (S_{f_0}, S_{f_1}, \dots, S_{f_{K-1}})$$

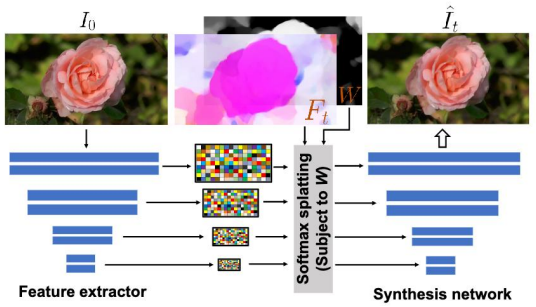
    - 예측된 spectral volume은 IDFT(Inverse Discrete Fourier Transform)을 통해 motion texture, $\mathcal{F}$로 변환

$$\mathcal{F} = (F_1, F_2, \dots, F_T)$$

      - 이 motion은 미래의 모든 timestep에서 각 입력 pixel의 위치를 결정
    - 예측된 motion texture가 주어지면, neural image-based rendering 기법을 사용하여 입력 RGB 이미지를 애니메이션화



< Motion prediction module >



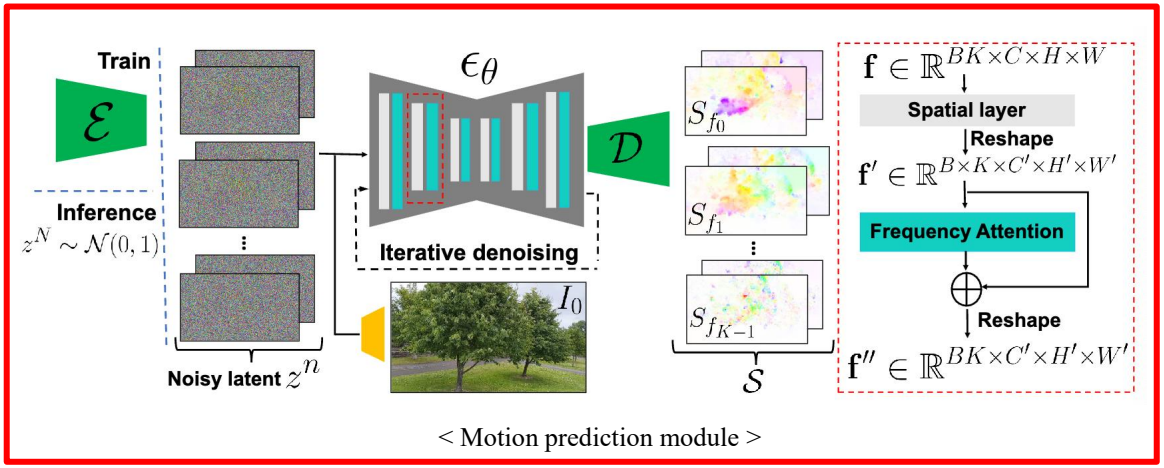
< Rendering module >

< Overview of Generative Image Dynamics >

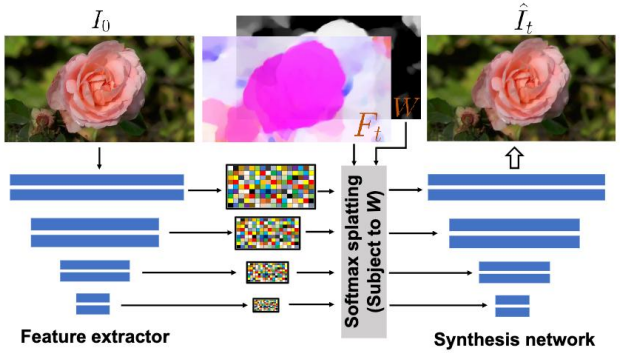
부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Stage 1



< Motion prediction module >

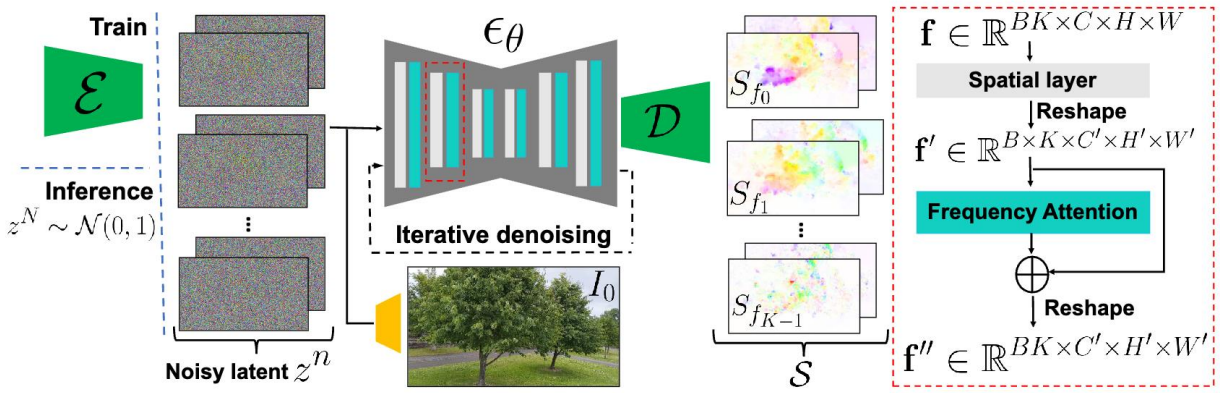


< Rendering module >

< Overview of Generative Image Dynamics >

Method

- Motion prediction module
  - Motion representation
    - LDM으로 $4K$-channel 2D motion spectrum map으로 구성된 spectral volume을 생성
      - K ($K \ll T$)는 모델링된 주파수의 수
        - ✓각 주파수마다 x-축과 y-축 방향에 대한 복소수 Fourier coefficients를 나타내기 위해 4개의 스칼라 값이 필요함
    - 주목할 점
      - 미래 time step에서의 한 픽셀의 모션 궤적 $\mathcal{F}(p)$와 그것의 spectral volume representation $S(p)$가 FFT로 연결되어 있음
        - ✓$S(p) = FFT(\mathcal{F}(p))$



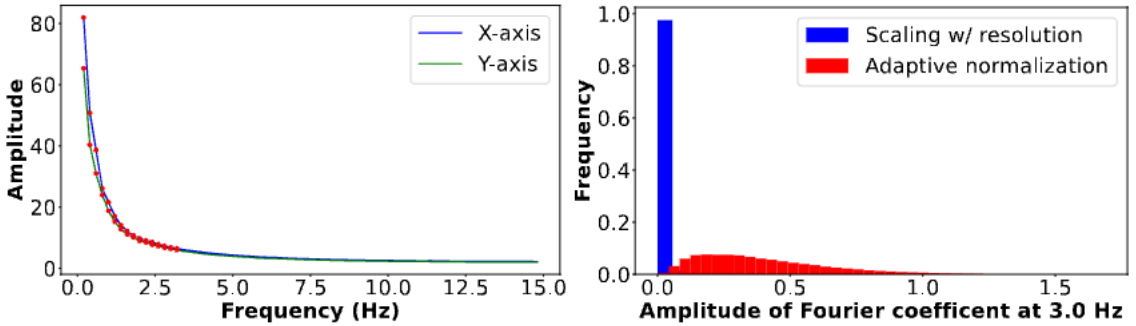
< Motion prediction module >

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Motion prediction module
  - Motion representation
    - 어떻게 K개의 출력 주파수를 선택해야 하는지?
      - 대부분의 자연스러운 진동 움직임이 주로 low-frequency 성분으로 구성되어 있음
      - 주파수 증가에 따라 전력 스펙트럼은 지수적으로 감소

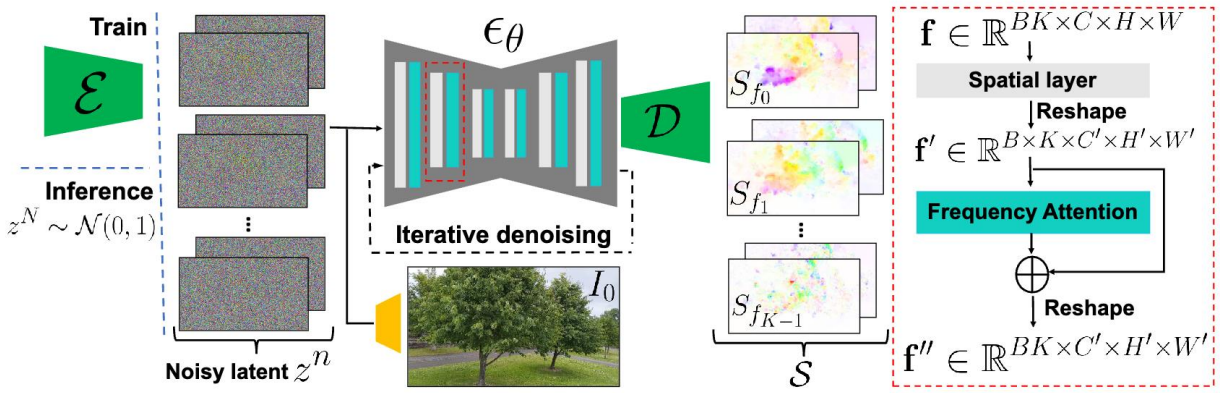
        ⇒ 대부분의 자연스러운 진동 움직임이 저주파항들로 잘 표현될 수 있음
      - 첫 $K = 16$개의 Fourier coefficient만으로도 다양한 실제 비디오와 장면에서 원래의 자연스러운 움직임을 사실적으로 재현할 수 있음을 발견



< Frequency Experiment >

Method

- Motion prediction module
  - Predicting motion with a diffusion model
    - 학습 과정 : RGB image가 아니라, 실제 비디오 시퀀스로부터 인코딩된 spectral volume에 적용됨
      - Spectral volume이 인코딩된 뒤, 미리 정의된 variance 스케줄에 따라 n step동안 noise가 추가되어 noisy latents z^n을 생성
      - 2D U-Net들은 각 step $n \in \{1, 2, \ldots, N\}$에서 noisy latent z^n을 denoise하기 위해 훈련됨
        - ✓Latent feature가 업데이트된 후 다시 디코딩되어 spectral volume이 재구성됨



< Motion prediction module >

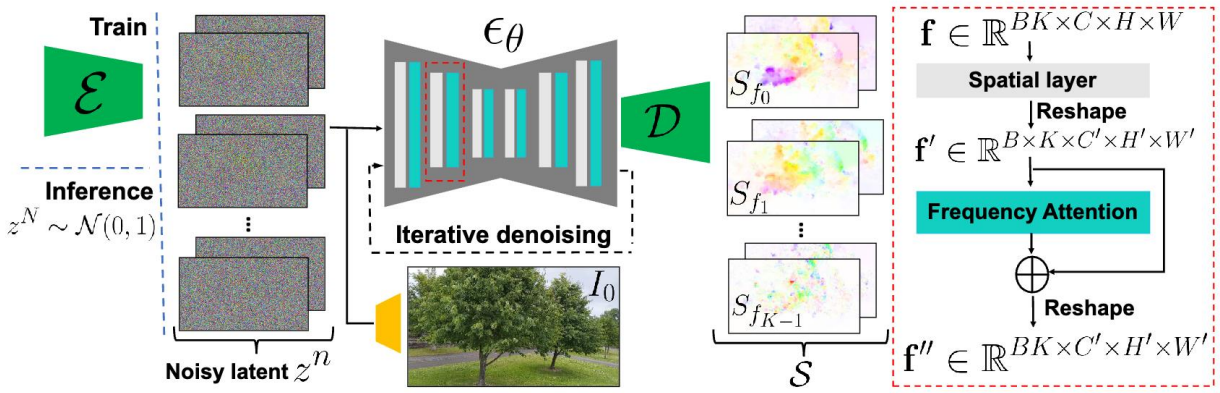
Method

- Motion prediction module
  - Predicting motion with a diffusion model
    - 학습 과정 : RGB image가 아니라, 실제 비디오 시퀀스로부터 인코딩된 spectral volume에 적용됨

LDM train loss

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathrm{n}\in u[1,\mathrm{N}],\epsilon\sim\mathcal{N}(0,1)}[\|\epsilon - \epsilon_\theta(z^n; n, c)\|^2]$$

c : text 혹은 train video 시퀀스의 첫 번째 프레임 I_0과 같은 condition 신호의 임베딩

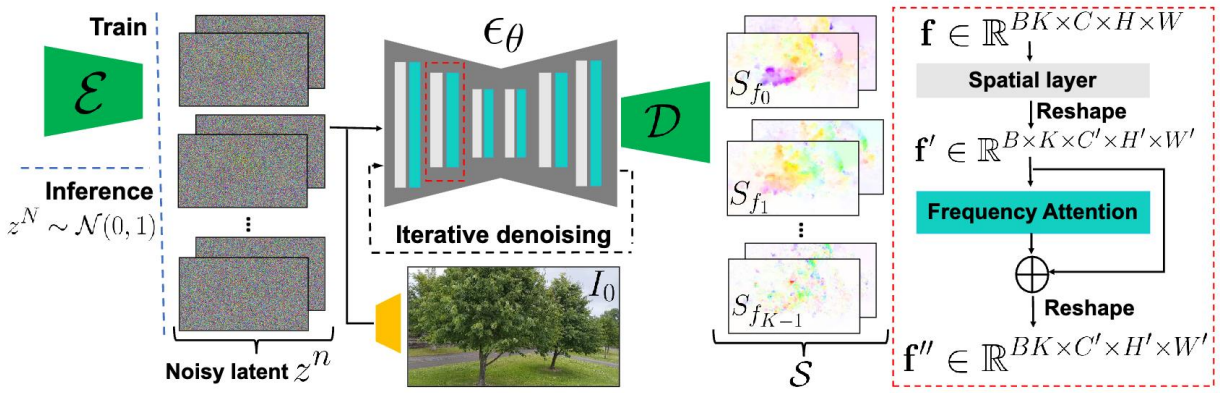


< Motion prediction module >

Method

- Motion prediction module
  - Frequency adaptive normalization
    - Motion textures가 주파수별로 특정한 분포 특성을 가짐
      - Diffusion model은 stable training 및 denoising을 위해 output의 절대값이 -1에서 1 사이에 있어야 함 → 실제 비디오로부터 추출한 S의 계수들을 학습 전에 정규화 해야 함
      - 이전 연구와 같이 input video의 크기에 기반하여 계수들의 크기를 [0,1] 범위로 스케일링하면, 높은 주파수에서 거의 모든 계수는 0에 수렴하게 됨

⇒ 이러한 데이터로 학습된 모델은 부정확한 움직임을 생성할 수 있으며,
추론 시에는 작은 예측오류가 큰 레이어 오류를 초래할 수 있음



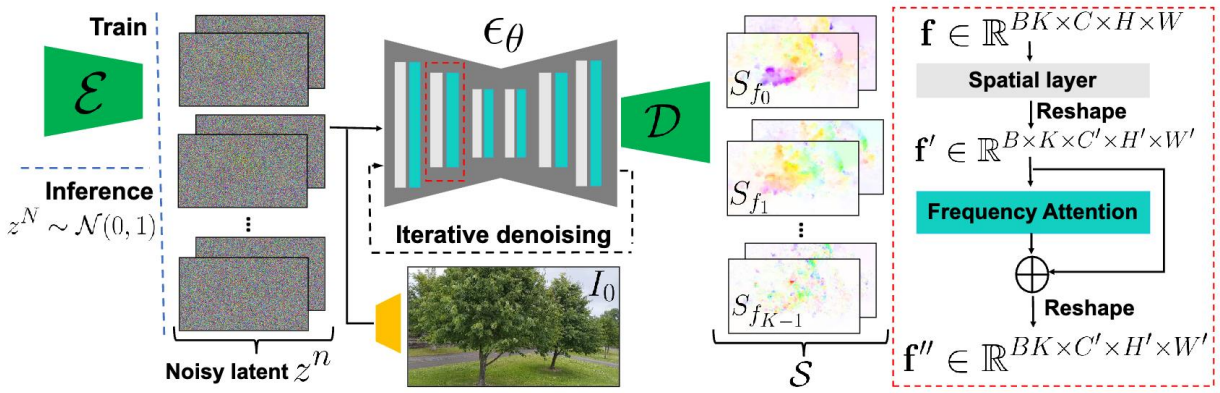
< Motion prediction module >

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Motion prediction module
  - Frequency adaptive normalization
    - Train set에서 계산된 통계량을 기반으로 각 주파수의 Fourier 계수들을 독립적으로 정규화 함
      - 각 주파수 f_j에 대해, 모든 입력 샘플에서 Fourier 계수 크기의 95번째 백분위수(95th percentile)를 계산하고, 이를 주파수별 스케일링 계수 s_{f_j}로 사용함
    - 각 Fourier 계수를 이 스케일로 나눈 후, 극단적인 값들을 피하기 위해 제곱근 변환(square root transform)을 적용

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



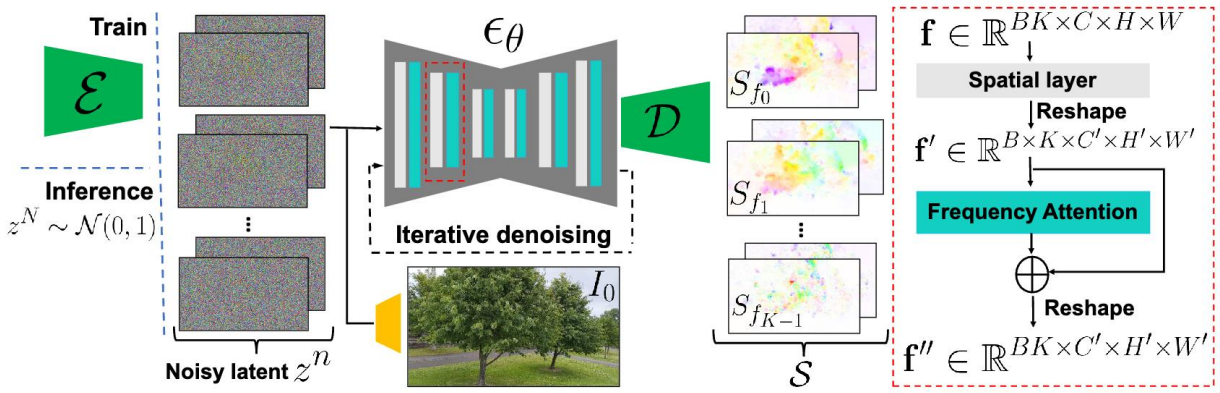
< Motion prediction module >

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Motion prediction module
  - Frequency-coordinated denoising
    - K개의 주파수 밴드를 가진 spectral volume S를 예측하기 위한 가장 단순한 방법
      - 단일 diffusion U-Net으로부터 $4K$ 채널의 텐서를 출력하는 것

⇒ 많은 수의 채널을 생성하도록 모델을 훈련할 경우,
과도하게 스무딩되고, 부정확한 출력이 발생할 수 있음



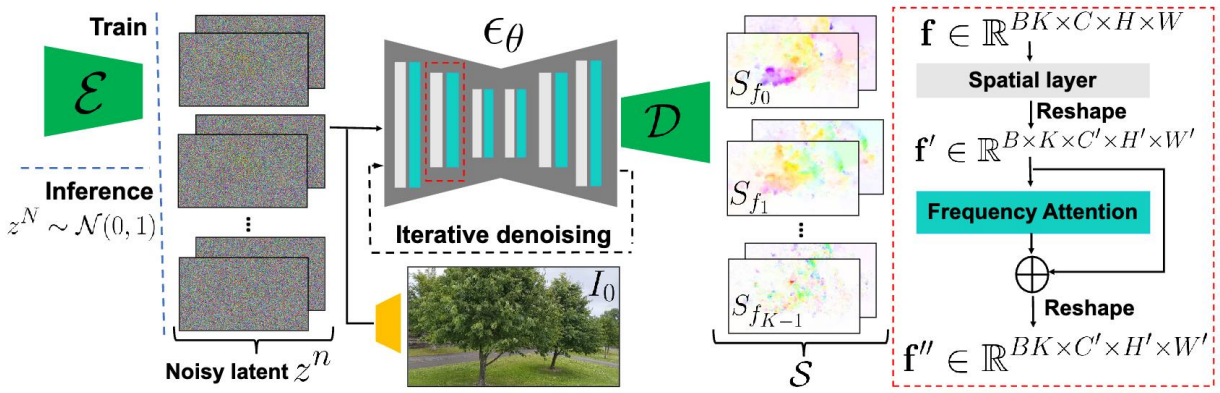
< Motion prediction module >

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Motion prediction module
  - Frequency-coordinated denoising
    - 각 개별 주파수 슬라이스를 독립적으로 예측하기 위해 LDM에 주파수 임베딩(frequency embedding)을 추가로 주입하는 것

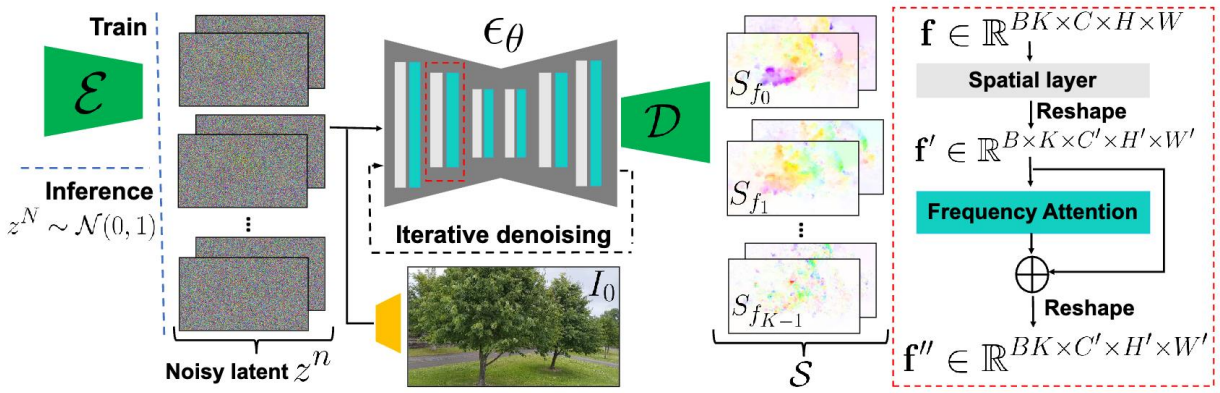
⇒ 이러한 설계 선택은 주파수 영역에서 상관되지 않은 예측을 초래하여 비현실적인 움직임을 만들 수 있음



< Motion prediction module >

Method

- Motion prediction module
  - Frequency-coordinated denoising
    - 최근 diffusion 연구에서 영감을 받아, Frequency-coordinated denoising 전략을 제안
      - 입력 이미지 I_0가 주어졌을 때, 먼저 LDM ϵ_θ를 훈련하여 spectral volume S의 단일 4채널 주파수 슬라이스 S_{f_j}를 예측하게 함
        - ✓여기서 time step 임베딩과 함께 추가 주파수 임베딩을 LDM에 주입함



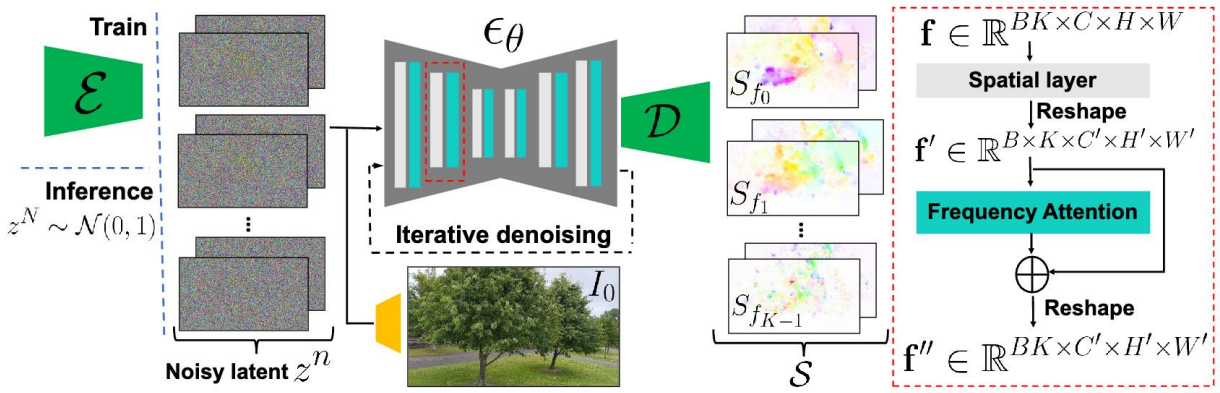
< Motion prediction module >

Method

- Motion prediction module
  - Frequency-coordinated denoising
    - 최근 diffusion 연구에서 영감을 받아, Frequency-coordinated denoising 전략을 제안
      - 이 LDM ϵ_θ의 파라미터를 고정하고, K개의 주파수 밴드 전체에서 ϵ_θ의 2D 공간 레이어들 사이에 attention 레이어를 삽입하고 fine-tune함
        - ✓배치사이즈 B에 대해 ϵ_θ의 2D 공간 레이어들은 $B \cdot K$개의 noisy latent feature들을 채널 사이즈 C로 가지는 독립 샘플로 처리함
          - Shape : $R^{(B\cdot K)\times C\times H\times W}$

$$[B \cdot K, C, H, W]$$

K : 주파수에 대한 차원 없음
→ 주파수 간 정보교환 없음



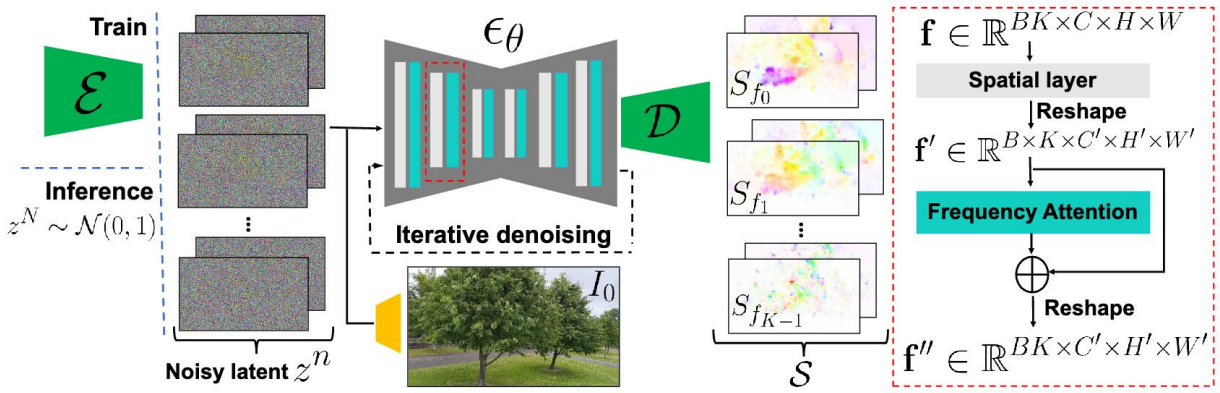
< Motion prediction module >

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Motion prediction module
  - Frequency-coordinated denoising
    - 최근 diffusion 연구에서 영감을 받아, Frequency-coordinated denoising 전략을 제안
      - 이 LDM ϵ_θ의 파라미터를 고정하고, K개의 주파수 밴드 전체에서 ϵ_θ의 2D 공간 레이어들 사이에 attention 레이어를 삽입하고 fine-tune함
        - ✓이전 2D 공간 레이어들로부터의 latent feature를 $R^{B\times K\times C\times H\times W}$로 reshape하여 attention 레이어에 입력하기 전에 전달함 ($[B, K, C, H, W]$)
        - ✓Attention 레이어는 이를 주파수 축을 가로지르는 연속적인 feature들로 해석함

⇒ frequency attention 레이어들은 모두 frequency slice들을 coordinate하도록 fine-tune되어, 일관된 spectral volume을 생성함

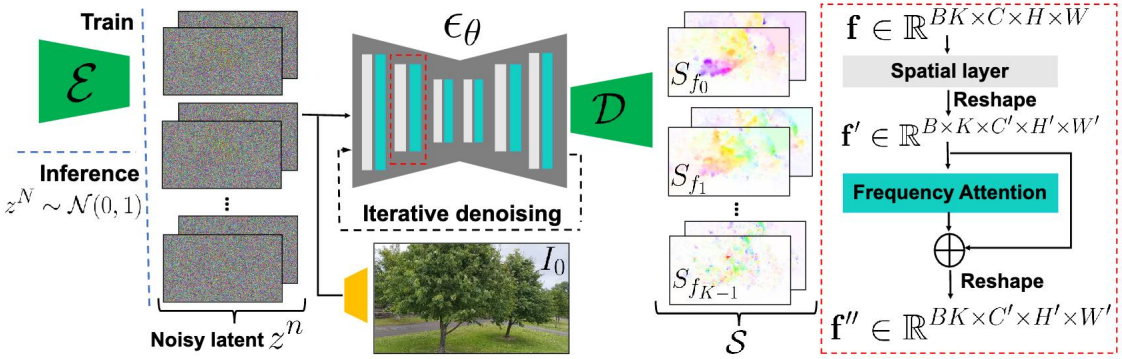


< Motion prediction module >

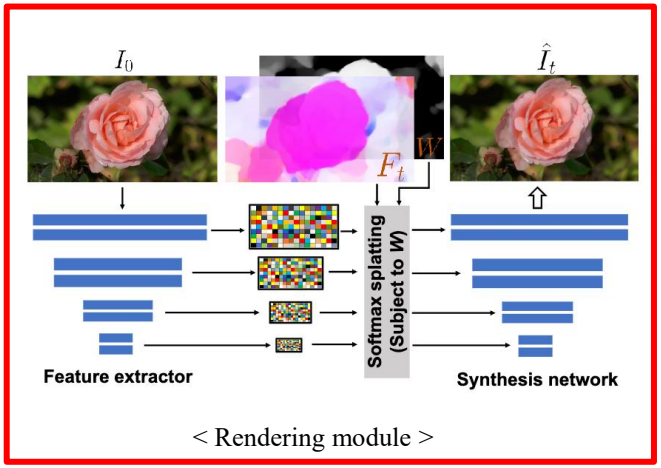
부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Stage 2



< Motion prediction module >

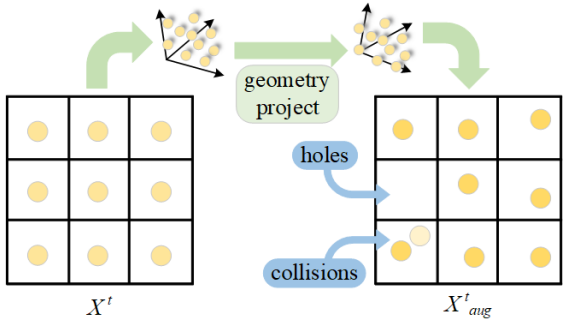


< Rendering module >

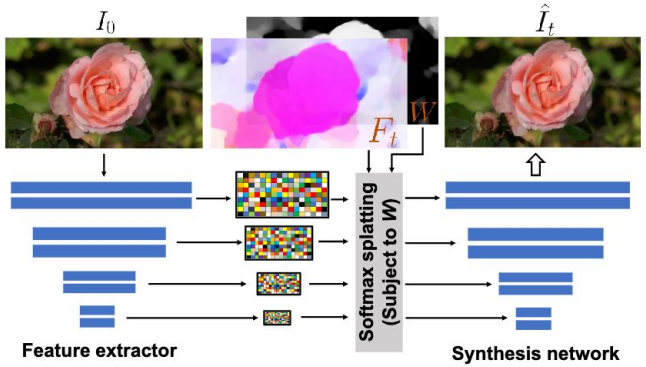
< Overview of Generative Image Dynamics >

Method

- Rendering module
  - Image-based rendering
    - 입력 이미지 I_0에 대해 예측된 Spectral volume S를 이용해 미래 프레임 $\hat{I}_t$를 렌더링
      - 각 픽셀에서 inverse FFT를 적용하여 시간 영역에서의 motion texture $\mathcal{F}(p) = \text{FFT}^{-1}(S(p))$를 유도
      - 미래 프레임 $\hat{I}_t$ 를 생성하기 위해, 딥 이미지 기반 렌더링 기법 채택, 예측된 motion field F_t를 사용하여 인코딩된 I_0를 forward-warping하여 splatting함



< Forward Warping >

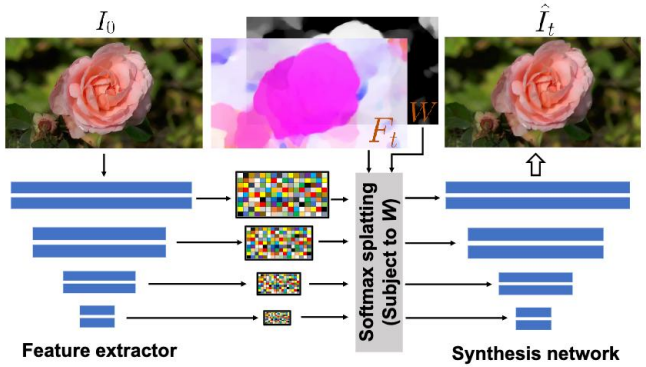


< Rendering module >

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Rendering module
  - Image-based rendering
    - Forward-warping은 hole을 유발할 수 있으며, 여러 소스 픽셀이 동일한 출력 2D 위치에 매핑될 수 있기 때문에, frame interpolation에 관한 기존 연구에서 제안된 feature pyramid softmax splatting 전략을 채택
      - feature extractor 네트워크를 통해 I_0를 인코딩하여 multi-scale feature map 을 생성
      - 각 스케일 j에서 개별 feature map에 대해, 예측된 2D 모션 필드 F_t를 해상도에 맞추어 리사이즈하고 스케일링 함
      - 예측된 flow 크기(predicted flow magnitude)를 깊이(depth)의 proxy 로 사용하여, 각 source 픽셀이 목표 위치(destination location)에 기여하는 가중치(contributing weight)를 결정

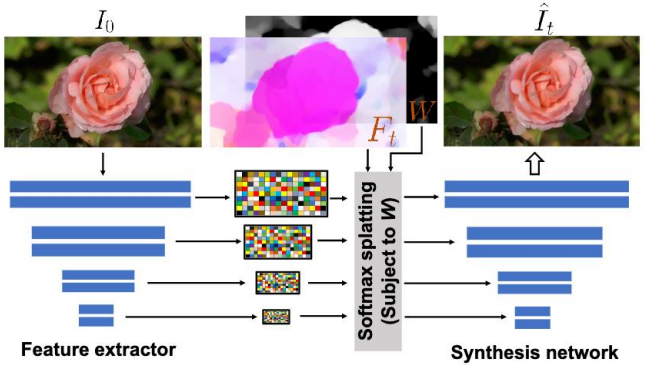


< Rendering module >

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

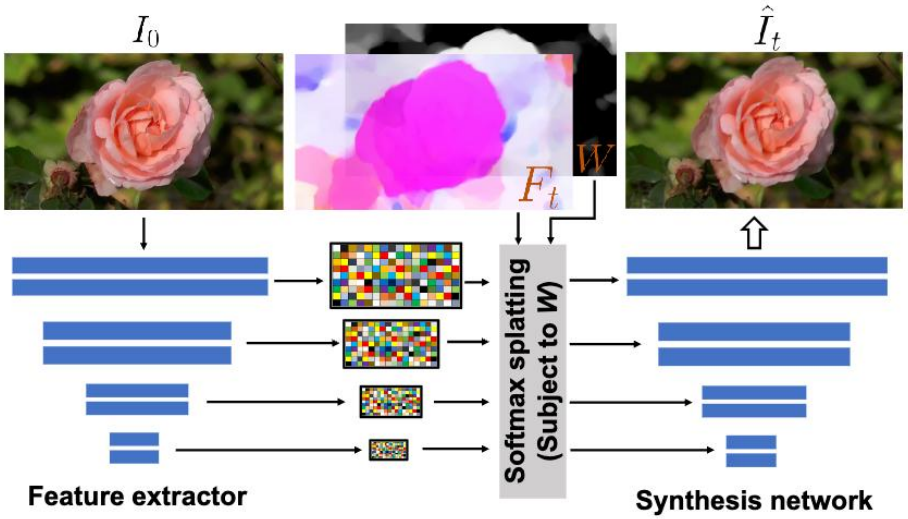
- Rendering module
  - Image-based rendering
    - Forward-warping은 hole을 유발할 수 있으며, 여러 소스 픽셀이 동일한 출력 2D 위치에 매핑될 수 있기 때문에, frame interpolation에 관한 기존 연구에서 제안된 feature pyramid softmax splatting 전략을 채택
      - 픽셀 단위 가중치 = 예측된 motion texture의 평균 크기
        - ✓$W(p) = \frac{1}{T}\sum_t \|F_t(p)\|_2$
        - ✓큰 움직임은 foreground object 에 대응하고, 작거나 영(0)의 움직임은 background 에 대응한다고 가정
      - learnable weights 대신 motion-derived weights 를 사용
        - ✓single-view 경우에 learnable weights 가 disocclusion ambiguities 해결에 효과적이지 않음



< Rendering module >

Method

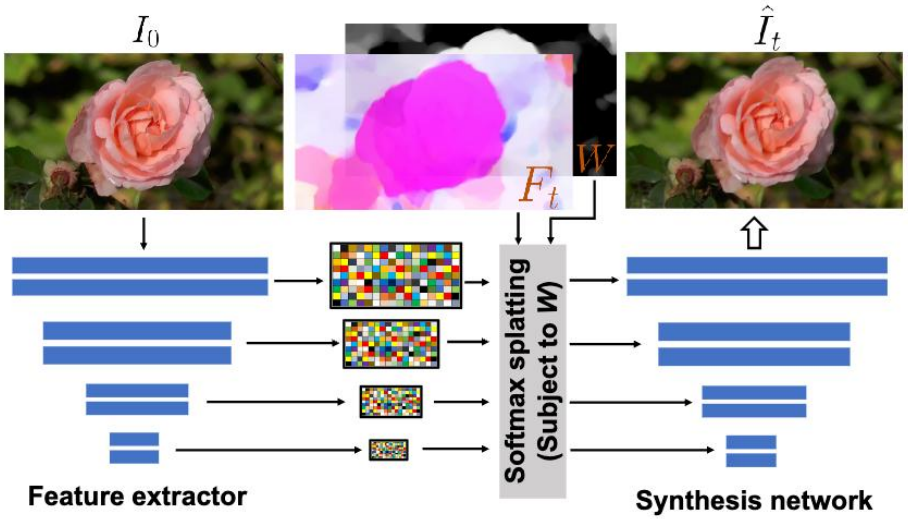
- Rendering module
  - Image-based rendering
    1. Motion field F_t와 weight W를 사용하여, softmax splatting을 적용해 각 스케일에서 feature map을 warp 하여 warped feature를 생성
    2. Warp 된 feature들은 이미지 합성 디코더(image synthesis decoder)의 각 블록에 삽입되어 최종 렌더링된 이미지 $\hat{I}_t$ 를 생성



< Rendering module >

Method

- Rendering module
  - Image-based rendering
    - Feature extractor와 합성 네트워크를 시작 프레임 I_0과 목표 프레임 I_t쌍을 사용하여 공동으로 학습
      - 학습은 실제 비디오로부터 임의로 샘플링된 (I_0, I_t)쌍을 사용하며, I_0로부터 I_t로의 estimated flow field 를 이용하여 I_0로부터 인코딩된 feature들을 warp하고, 예측된 $\hat{I}_t$ 를 VGG perceptual loss로 I_t와 비교하여 supervision



< Rendering module >

Experiments

- Data
  - 온라인 소스와 자체 촬영을 통해 진동 운동(oscillatory motion)이 있는 자연 장면의 비디오 3,015개를 수집 및 처리
    - 전체의 10%는 테스트용으로 보류하고 나머지는 학습에 사용
  - GT 모션 trajectory를 추출하기 위해, 선택된 시작 이미지와 이후 프레임 간에 coarse-to-fine optical flow 방법을 적용
    - 학습 데이터는 매 10번째 프레임을 입력 이미지로 사용하고, 이후 149개의 프레임 간 모션 궤적을 활용하여 대응되는 스펙트럼 볼륨 ground truth를 생성
  - 총 데이터는 150,000쌍 이상의 이미지-모션 쌍으로 구성

Experiments

- Quantiative Results
  - Image synthesis : 개별 합성 프레임의 품질을 평가
    - Metric
      - Fréchet Inception Distance (FID) : 실제 이미지와 생성된 이미지가 얼마나 유사한지 측정
      - Kernel Inception Distance (KID) : 생성 모델의 generation 성질을 생성한 데이터와 실제 데이터 feature의 moment의 차이를 측정
      - 두 metric을 활용하여 생성 프레임과 정답 프레임 분포 간의 평균 거리를 측정
  - Video synthesis : 합성된 비디오의 품질과 시간적 일관성을 평가
    - Metric
      - Fréchet Video Distance (윈도우 크기 16의 FVD, 윈도우 크기 32의 FVD32)
        - ✓Human Kinetics 데이터셋으로 학습된 I3D 모델 기반
      - Dynamic Texture Fréchet Video Distance (윈도우 크기 16의 DTFVD, 32의 DTFVD32)
        - ✓자연스러운 운동 텍스처가 중심인 Dynamic Textures Database 기반

Experiments

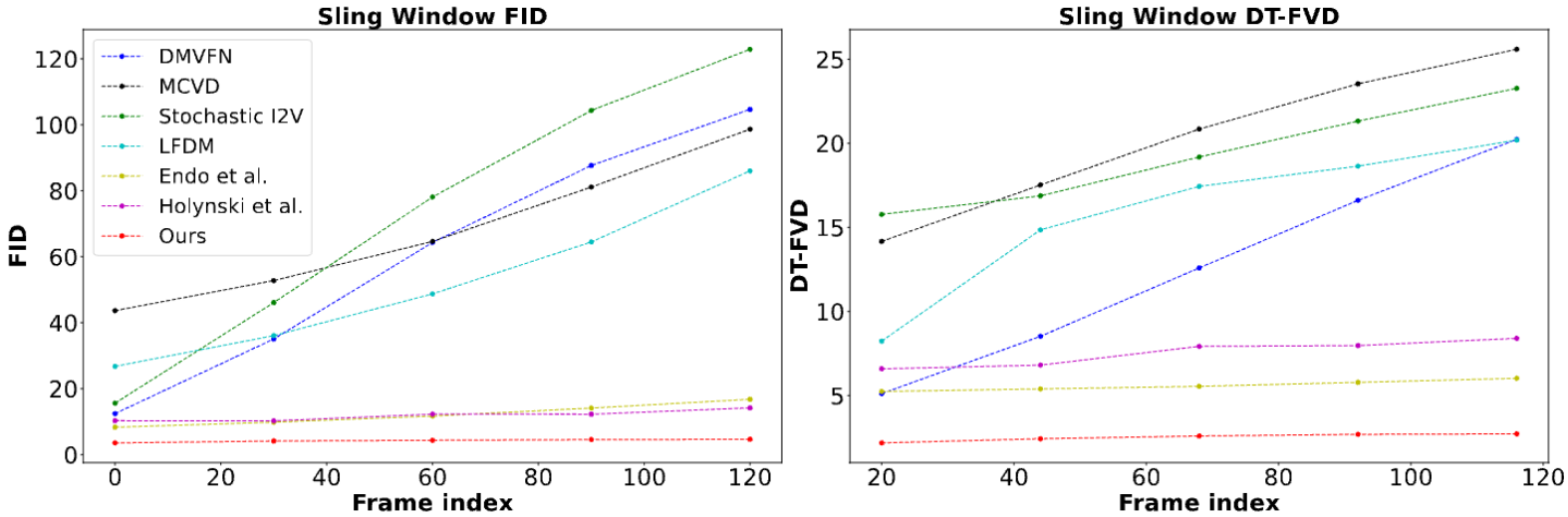
- Quantiative Results

Method	Image Synthesis FID	KID	Video Synthesis FVD	FVD_{32}	DTFVD	$DTFVD_{32}$
TATS [35]	65.8	1.67	265.6	419.6	22.6	40.7
Stochastic I2V [27]	68.3	3.12	253.5	320.9	16.7	41.7
MCVD [92]	63.4	2.97	208.6	270.4	19.5	53.9
LFDM [66]	47.6	1.70	187.5	254.3	13.0	45.6
DMVFN [48]	37.9	1.09	206.5	316.3	11.2	54.5
Endo *et al.* [29]	10.4	0.19	166.0	231.6	5.35	65.1
Holynski *et al.* [46]	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**

< Quantitative comparisons on the test set >

Experiments

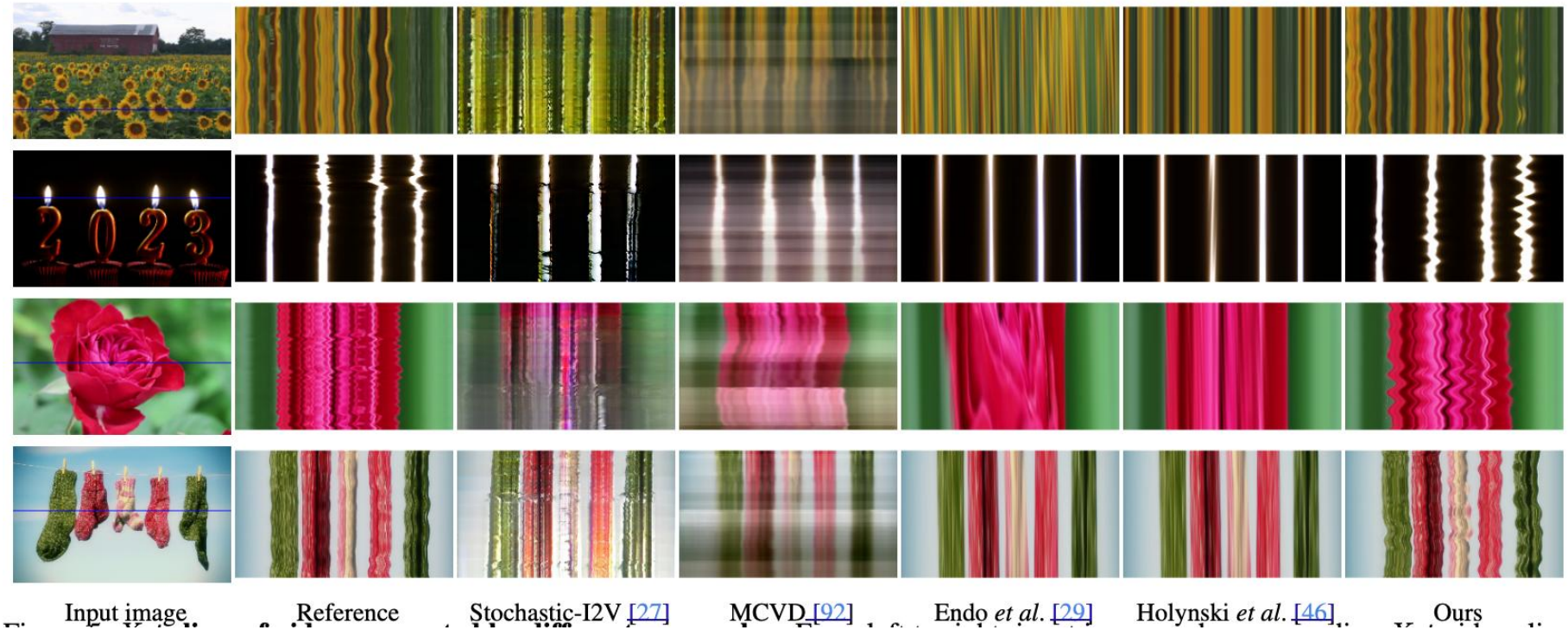
- Quantiative Results
  - Video synthesis : 비디오 품질이 시간이 지남에 따라 어떻게 저하되는지를 측정
    - Metric
      - 슬라이딩 윈도우 FID (30 프레임 크기)
      - 슬라이딩 윈도우 DTFVD (16 프레임 크기)



< Sliding window FID and DTFVD >

Experiments

- Qualitative Results



< X-t slices of videos generated by different approaches >

Experiments

- Quantiative Results
  - Ablation study

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 [46]	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

< Ablation study >

Motion-I2V: Consistent and Controllable Image-to-Video Generation with Explicit Motion Modeling

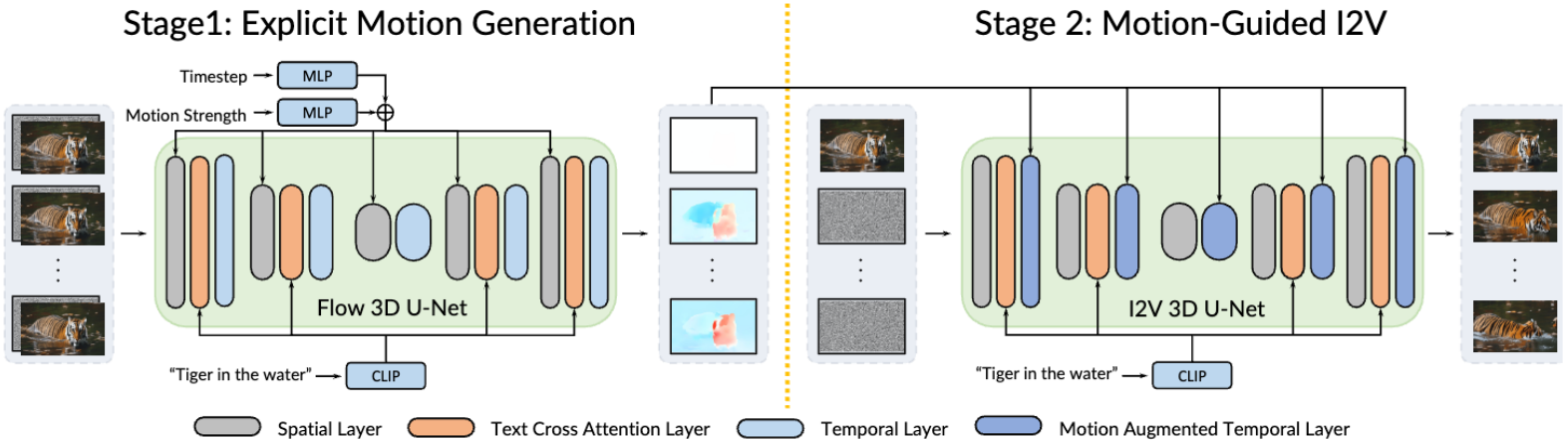
[ACM SIGGRAPH 24]

Introduction

- Problem & Solution
  - Existing I2V
    - 복잡한 시공간 priors를 모델링해야 하며, 1-D temporal attention의 좁은 시간 수용 범위는 큰 움직임이 있는 경우 특히 생성된 비디오의 temporal consistency를 확보하는 데 어려움을 줌
    - 제어 가능성이 제한적임
  - Proposed method
    - I2V 생성을 두 단계로 분해
      - Stage 1 : 픽셀 단위의 trajectories 형태로 가능한 움직임을 예측
      - Stage 2 : explicit modeling한 motion을 기반으로 일관된 애니메이션 생성

Method

- Overview
  - Explicit Motion Generation
    - Reference image 픽셀의 trajectory를 추론하는데 중점을 둔 diffusion-based motion field predictor
  - Motion-Guided I2V
    - Video latent diffusion models의 제한적인 1-D temporal attention을 강화하기 위해 motion-augmented temporal attention

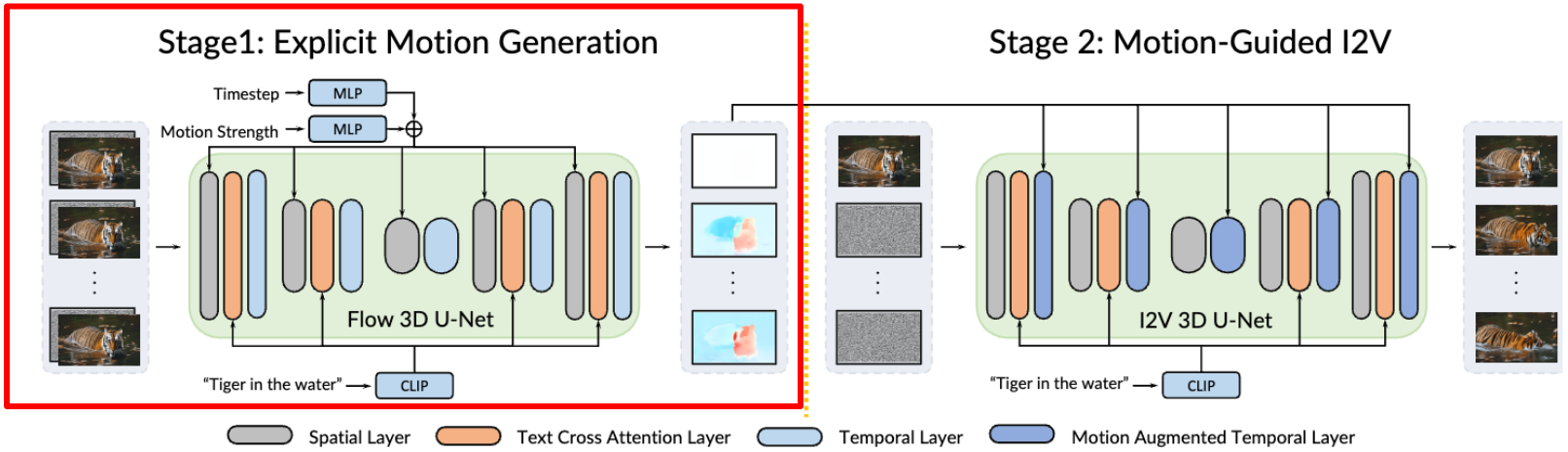


< Motion-I2V Framework>

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Motion Prediction with Video Diffusion Models
  - Motion fields modeling
    - 예측 목표
      - Reference image를 자연스럽게 애니메이션화하는 motion fields
    - Motion fields
      - 2D displacement maps의 시퀀스 $\{f_{0\to i} | i = 1, \dots, N\}$으로 표기
        - ✓각 $f_{0\to i} \in \mathbb{R}^{2\times H\times W}$는 기준 프레임과 시간 i의 미래 프레임 사이의 optical flow를 나타냄
        - ✓N은 합성된 프레임들의 개수

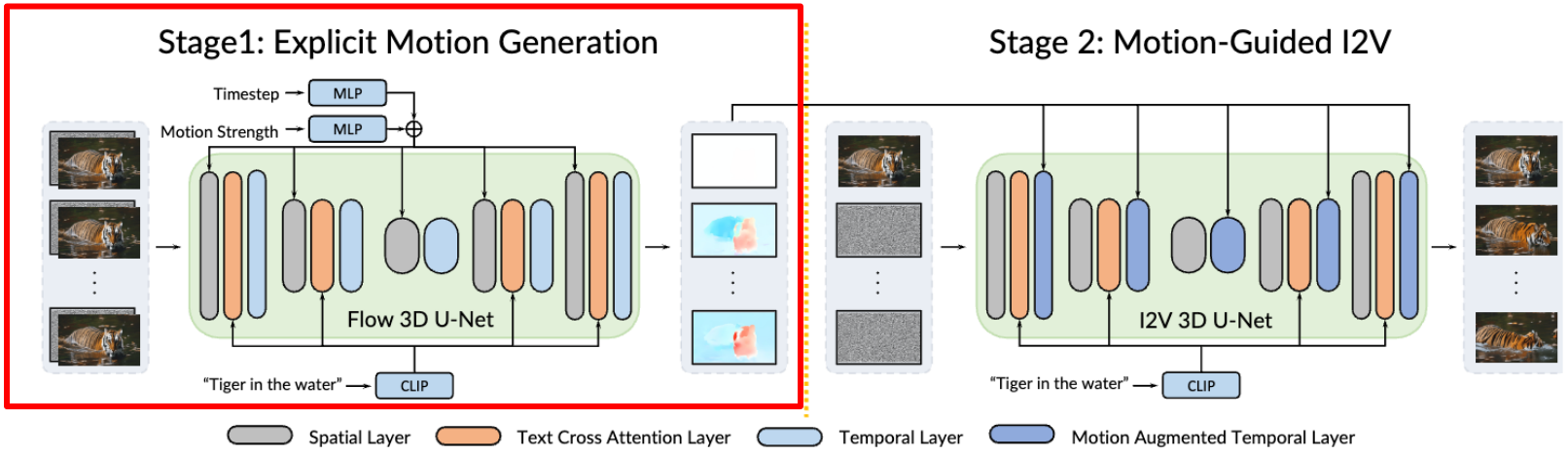


< Motion-I2V Framework>

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Motion Prediction with Video Diffusion Models
  - Motion fields modeling
    - Motion fields
      - 기준 이미지 I_0의 각 source pixel $p \in \mathbb{I}^2$에 대해, time step i에서의 target image I_i상의 대응좌표 $p_i' = p + f_{0\to i}(p)$를 결정 가능

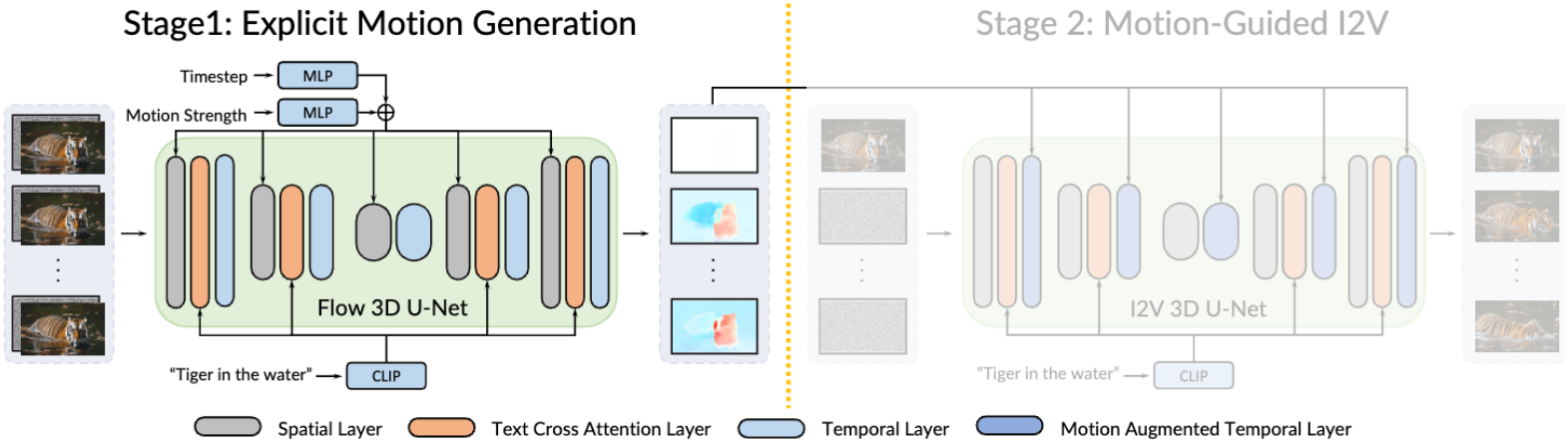


< Motion-I2V Framework>

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Motion Prediction with Video Diffusion Models
  - Motion fields modeling
    - Training a motion field predictor
      - VLDM을 학습하기 위해, 세 단계의 fine-tuning 전략 제안
        1. Reference image와 text prompt에 condition된 단일 displacement field를 예측하도록 사전 학습된 LDM을 tuning
        2. Tuning된 LDM 파라미터를 고정, vanilla temporal modules를 통합하여 VLDM을 구성해 학습
        3. Temporal modules 학습이 끝난 후, 전체 VLDM 모델을 fine-tune하여 최종 motion fields predictor 얻음

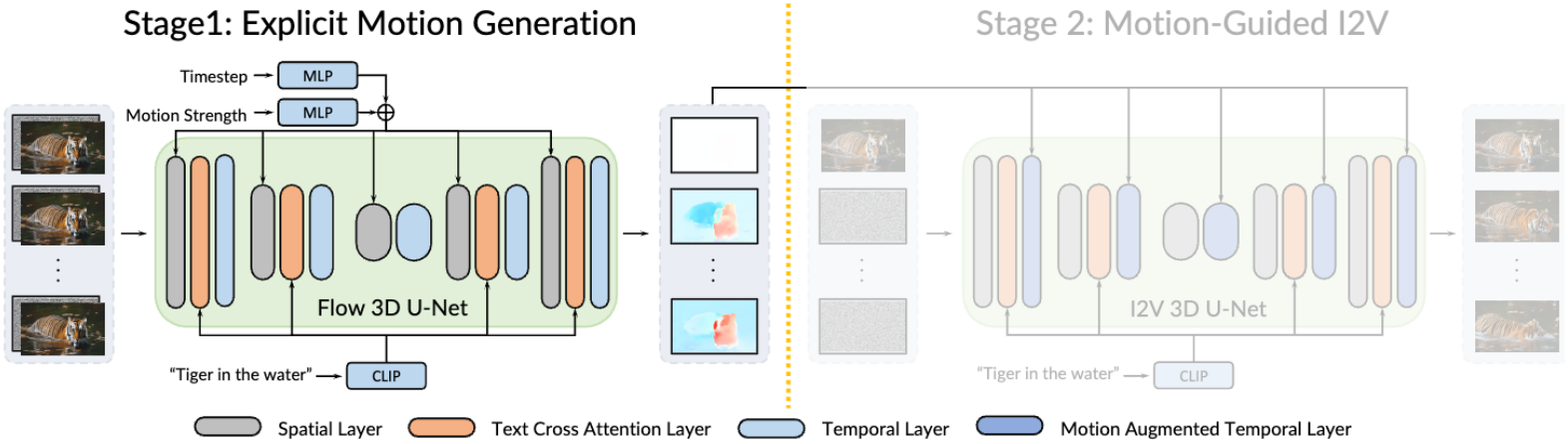


< Motion-I2V Framework>

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Motion Prediction with Video Diffusion Models
  - Motion fields modeling
    - Training a motion field predictor
      - FlowFormer++와 DOT를 사용하여, 각각 optical flow와 multi-frame trajectories를 추정하고, 이들을 학습 중 super-vision으로 활용



< Motion-I2V Framework>

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

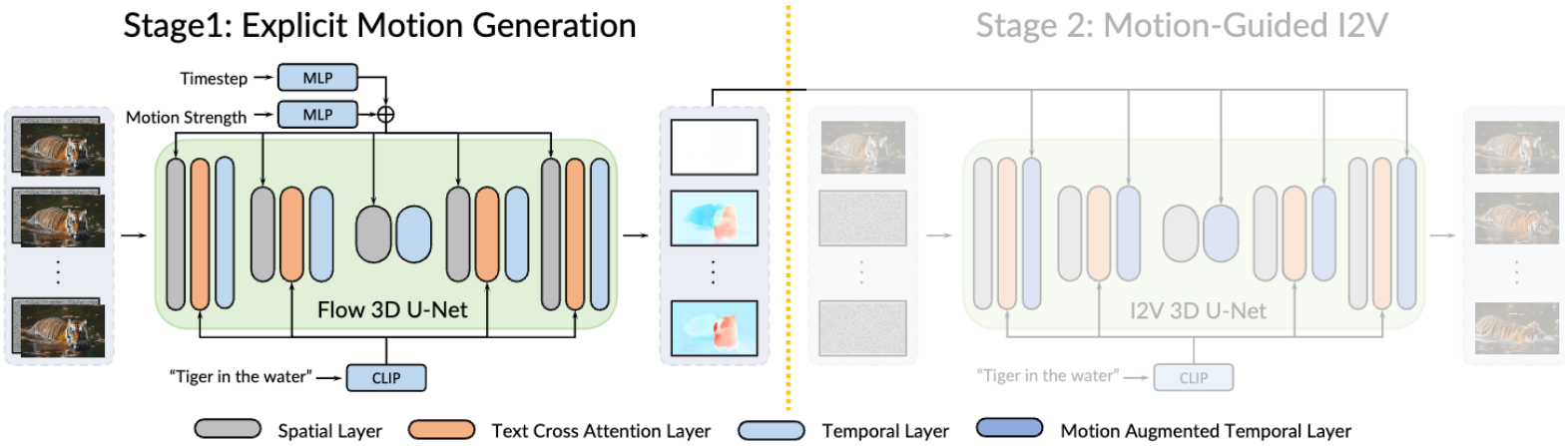
- Motion Prediction with Video Diffusion Models
  - Motion fields modeling
    - Encoding motion fields and conditional image

각 flow map $f_{0\to i} \in \mathbb{R}^{2\times H\times W}$를 optical flow VAE encoder를 사용하여 latent representation $z_{0\to i,0} = \varepsilon_{flow}(f_i) \in \mathbb{R}^{4\times h\times w}$로 인코딩

✓$h = \frac{H}{8}, w = \frac{W}{8}$

✓Optical flow autoencoder는 LDM 이미지 autoencoder의 구조를 그대로 따름

- 3채널 RGB image 대신 2채널 optical flow map을 입력으로 받고 출력함

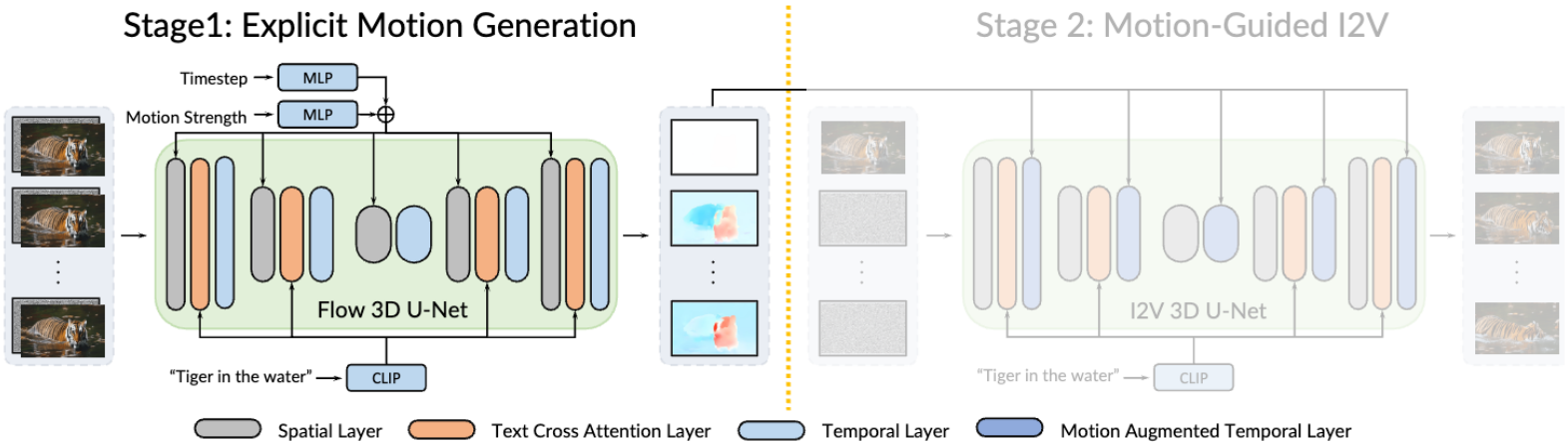


< Motion-I2V Framework>

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Motion Prediction with Video Diffusion Models
  - Motion fields modeling
    - Encoding motion fields and conditional image
      - 각 flow map $f_{0\to i} \in \mathbb{R}^{2\times H\times W}$를 optical flow VAE encoder를 사용하여 latent representation $z_{0\to i,0} = \varepsilon_{flow}(f_i) \in \mathbb{R}^{4\times h\times w}$로 인코딩
        - ✓깨끗한 reference image I_0의 latent code $\varepsilon(I_0) \in \mathbb{R}^{4\times h\times w}$를 채널 차원으로 concatenate함

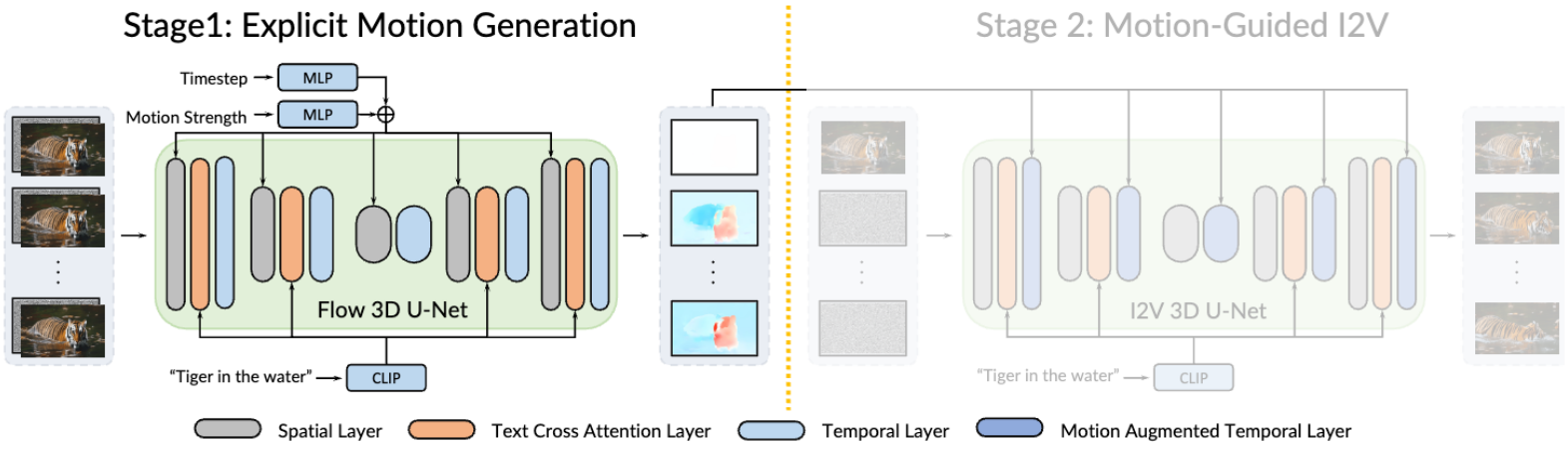


< Motion-I2V Framework>

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Motion Prediction with Video Diffusion Models
  - Motion fields modeling
    - Encoding motion fields and conditional image
      - 각 flow map $f_{0\to i} \in \mathbb{R}^{2\times H\times W}$를 optical flow VAE encoder를 사용하여 latent representation $z_{0\to i,0} = \varepsilon_{flow}(f_i) \in \mathbb{R}^{4\times h\times w}$로 인코딩
        - ✓SD 1.5 checkpoint 의 모든 LDM 가중치를 초기화하고, 새로 추가된 4개의 입력 채널에 대한 가중치는 0으로 설정 (더 큰 시간 간격이 더 큰 움직임을 가진다고 가정)
        - ✓네트워크가 시간 간격을 인식하도록 하기 위해, 프레임 stride i 는 두 개의 레이어로 구성된 MLP 로 임베딩되며, time embeddings 에 추가되어 motion strength condition으로 작용



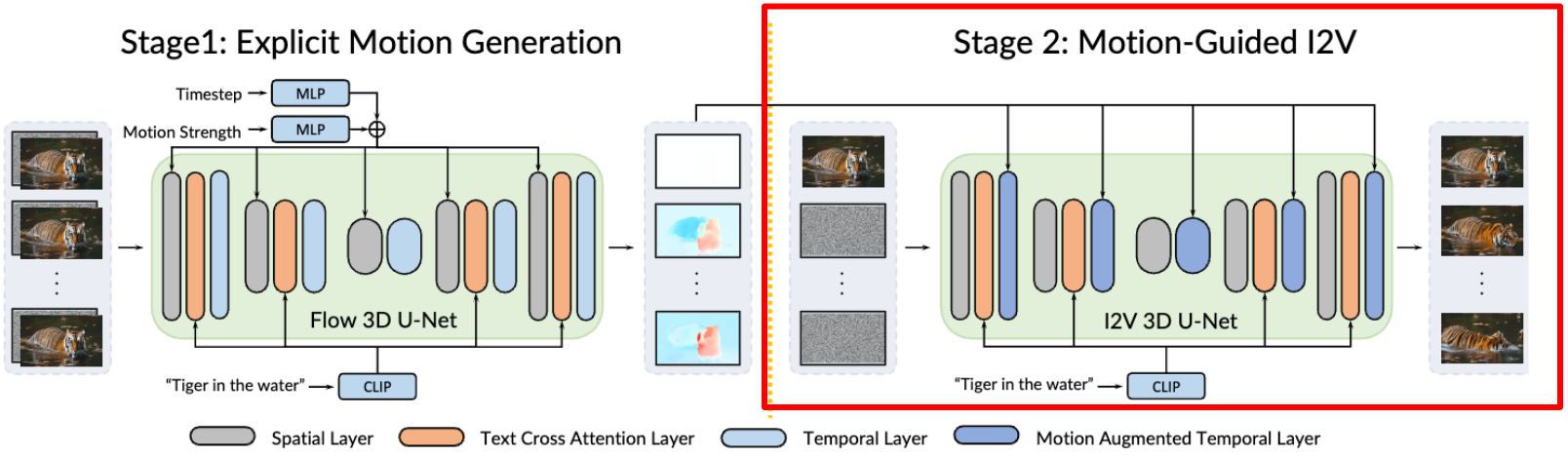
< Motion-I2V Framework>

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- Video Rendering with Predicted Motion
  - Motion-augmented temporal attention
    - 목표
      - Reference image의 내용을 stage 1에서 예측된 Motion field에 따라 합성된 프레임으로 전달하여 이미지의 fidelity와 시간적 일관성을 유지
      - 한 가지 단순한 대안 : Reference image의 feature와 후속 프레임들의 feature를 시간 차원에서 concatenate하고, reference image의 내용을 1-D temporal attention을 통해 전파하는 것

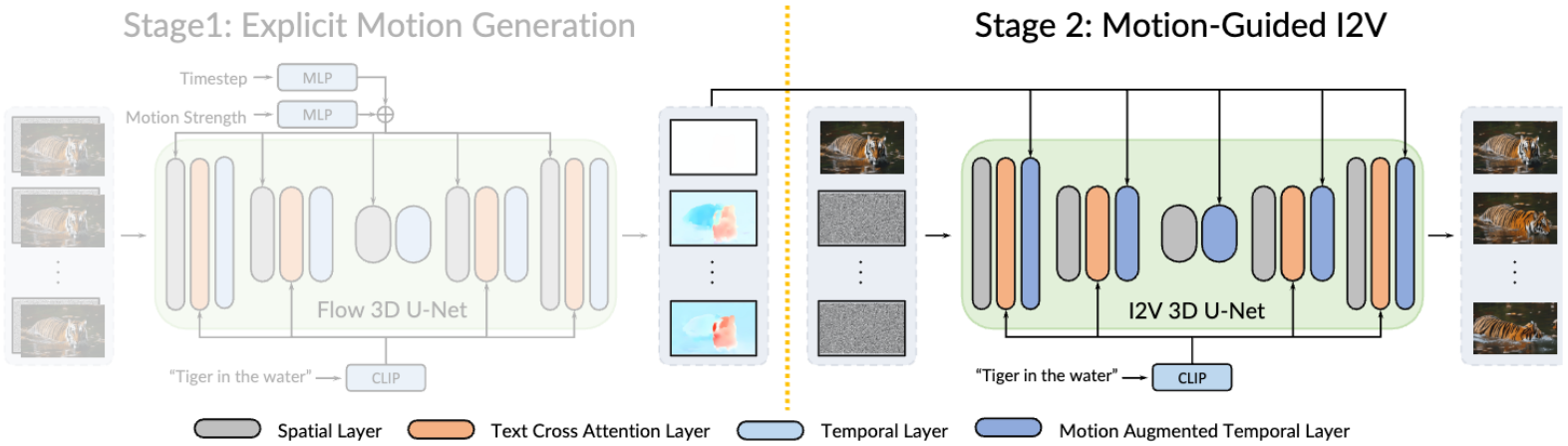
⇒ 1-D temporal attention의 receptive field가 좁아 한계가 있음



< Motion-I2V Framework>

Method

- Video Rendering with Predicted Motion
  - Motion-augmented temporal attention
    - VLDM의 기본 1-D temporal attention을 motion-augmented temporal attention으로 향상시키고, 다른 모듈들은 그대로 유지
      - 예측된 motion field를 기반으로 reference image의 feature를 warping
      - 변형된 feature를 각 프레임마다 cross-attention
    - 단순한 시간 attention보다 훨씬 더 넓은 시간 정보를 반영할 수 있음
    - Image 한 장만으로도 일관된 motion을 가진 비디오 생성 가능



< Motion-I2V Framework>

Experiment

- Train
  - Stage 1 : LDM model
    - Stable Diffusion v1.5
  - Stage 2 : VLDM model
    - AnimateDiff v2
  - Train dataset
    - WebVid-10M : large scale text-video dataset
  - Sampling
    - Stride 8로 16프레임짜리 비디오 클립을 무작위로 샘플링
  - Optimizer
    - 일정한 학습률을 사용하는 AdamW optimizer 사용

Experiment

- Quantitative Results
  - Dataset
    - 인간 활동, 동물, 차량, 자연 장면, AI 생성 이미지 등 다양한 범주를 포함한 테스트 세트를 자체적으로 구축
    - Pixabay에서 다운로드한 저작권 무료 이미지 80개로 구성
  - Prompt
    - 이미지 내용과 가능한 동작에 대한 프롬프트 생성을 위해 ChatGPT-4V를 사용
  - Metric
    - 프롬프트와 생성된 프레임 간의 일치도를 측정하기 위해 CLIP 텍스트-이미지 logit을 사용
    - 연속 생성 프레임 간의 CLIP 임베딩의 코사인 유사도를 계산하여 temporal consistency를 평가
    - 최초 프레임과 이후 프레임 간의 optical flow를 추정하여 생성된 동작의 크기를 평가

Experiment

- Quantitative Results
  - Dataset
    - 인간 활동, 동물, 차량, 자연 장면, AI 생성 이미지 등 다양한 범주를 포함한 테스트 세트를 자체적으로 구축
    - Pixabay에서 다운로드한 저작권 무료 이미지 80개로 구성
  - Prompt
    - 이미지 내용과 가능한 동작에 대한 프롬프트 생성을 위해 ChatGPT-4V를 사용
  - Metric
    - Prompt와 생성된 프레임 간의 일치도를 측정하기 위해 CLIP text-image logit을 사용
    - 연속 생성 프레임 간의 CLIP 임베딩의 코사인 유사도를 계산하여 temporal consistency를 평가
    - 최초 프레임과 이후 프레임 간의 optical flow를 추정하여 생성된 동작의 크기를 평가

Method	Prompt Consistency ↑	Frame Consistency ↑	Average Displacement
VideoComposer	32.62	0.9393	67.15
I2VGen-XL	33.69	0.9650	17.70
DynamiCrafter	34.60	0.9860	3.31
Ours	34.86	0.9871	20.06

< Quantitative comparison with SOTA approaches >

Experiment

- Quantiative Results
  - Ablation study

Stage 1?	Fusion Type	Prompt Consistency ↑	Frame Consistency ↑	Average Displacement
✗	-	32.95	0.9505	66.44
✓	Addition	33.99	0.9542	48.91
✓	Attention	34.86	0.9871	20.06

< Ablation study >

- Qualitative Results
  - Project page

감사합니다

supplementary

- Q1 Motion I2V 논문에서 stage 2에서도 결국 diffusion model을 사용하는데 굳이 stage를 나눠서 모션을 만들고 다시 diffusion 모델을 사용해야하는지
  - 큰 motion이나 viewpoint variation이 있는 경우에도 더 일관성 있는 비디오를 생성하기 위함도 있지만 가장 큰 motivation은 사용자가 소수의 trajectory 점과 영역만으로도 정밀하게 motion trajectory 및 영역을 제어할 수 있음
    - 단순히 텍스트 조건에만 의존하는 기존 방식보다 훨씬 더 강력한 제어 가능성을 제공함