CVSP 여름방학 세미나

4D Gaussian Editing using video editing model

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

Presented By
박사과정 정성훈

CVSP Lab
Computer Vision & Signal Processing

Outline

- Motivation
- Background
- InFusion: Inpainting 3D Gaussians via Learning Depth Completion from Diffusion Prior
- Señorita-2M: A High-Quality Instruction-based Dataset for General Video Editing by Video Specialists

Motivation

CVSP Lab
Computer Vision & Signal Processing

Motivation

* Kwon, Joohyun, Hanbyel Cho, and Junmo Kim. "Instruct-4DGS: Efficient Dynamic Scene Editing via 4D Gaussian-based Static-Dynamic Separation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

- 4D Gaussian Editing
  - 개요
    - 최근의 4D Gaussian에 대한 연구에서, 장면 전체 또는 객체 단위의 style을 편집하는 연구는 있으나, 객체를 새롭게 추가하는 연구는 없음
    - Video 관련 연구에서, 생성형 모델 (diffusion 등)을 활용하여 편집을 수행하는 연구가 다수 진행됨
    - 4D Gaussian에서 객체를 추가하기 위해 video editing 모델을 활용하고자 함



Original*



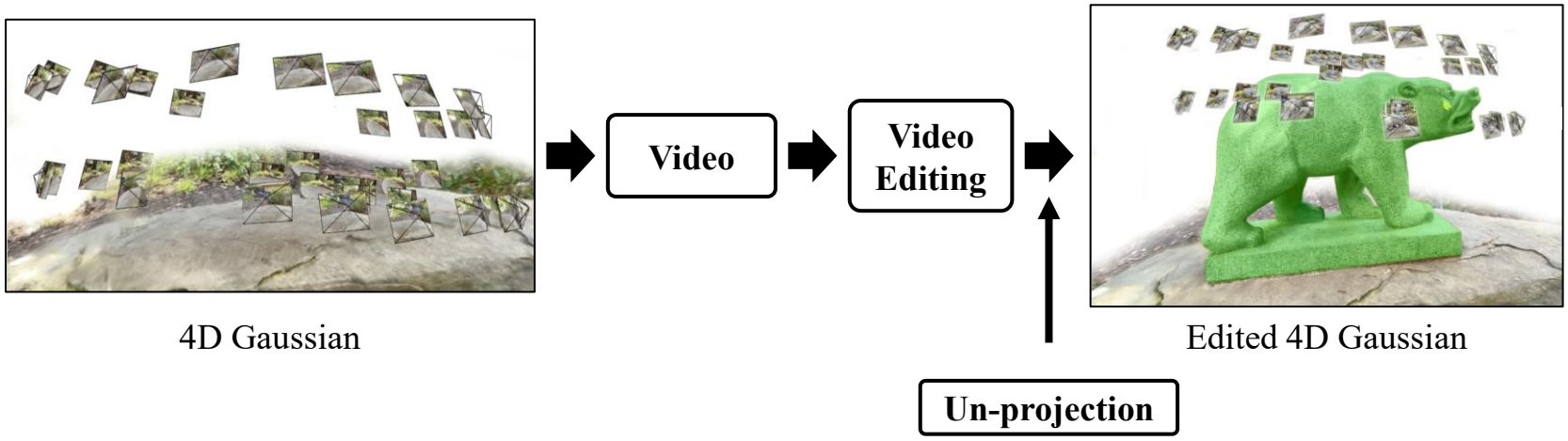
Van gogh*



Marble*

Motivation

- 4D Gaussian Editing
  - 개요
    - 최근의 4D Gaussian에 대한 연구에서, 장면 전체 또는 객체 단위의 style을 편집하는 연구는 있으나, 객체를 새롭게 추가하는 연구는 없음
    - Video 관련 연구에서, 생성형 모델 (diffusion 등)을 활용하여 편집을 수행하는 연구가 다수 진행됨
    - 4D Gaussian에서 객체를 추가하기 위해 video editing 모델을 활용하고자 함



4D Gaussian

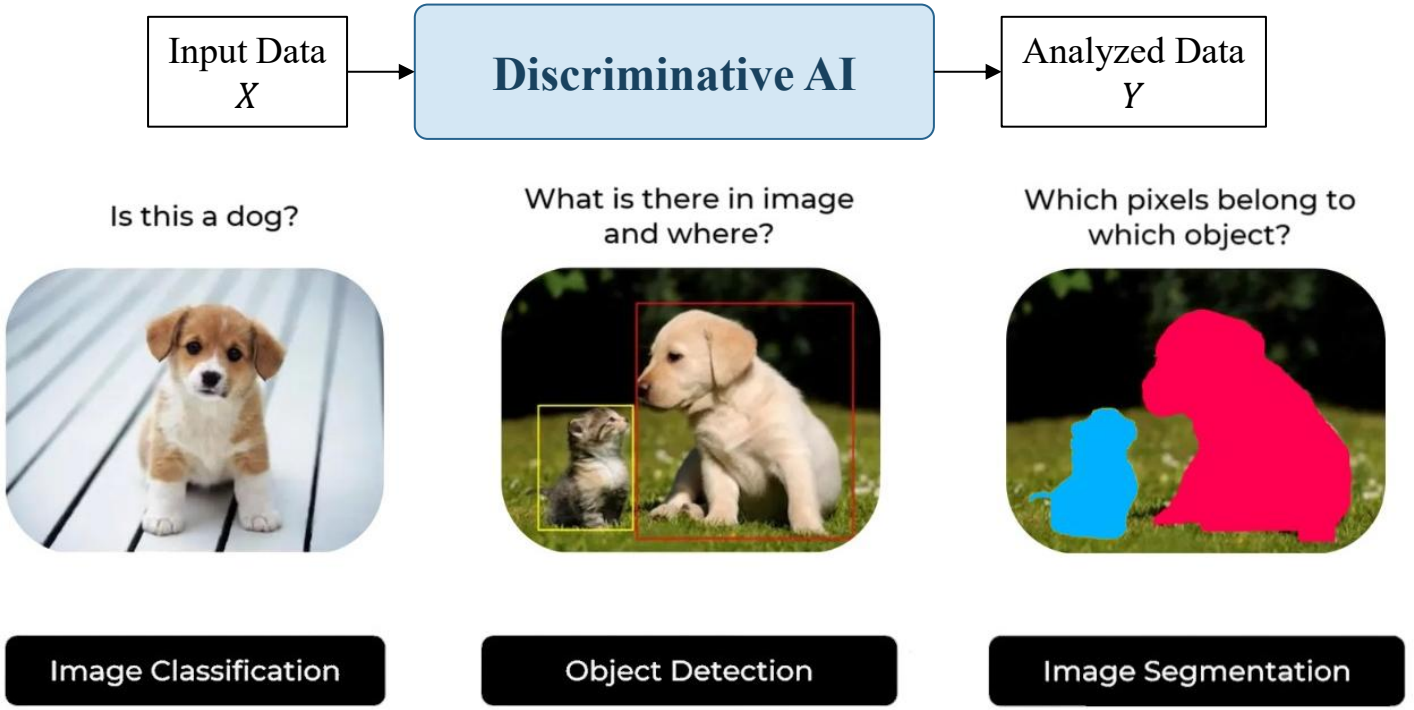
Edited 4D Gaussian

Background

CVSP Lab
Computer Vision & Signal Processing

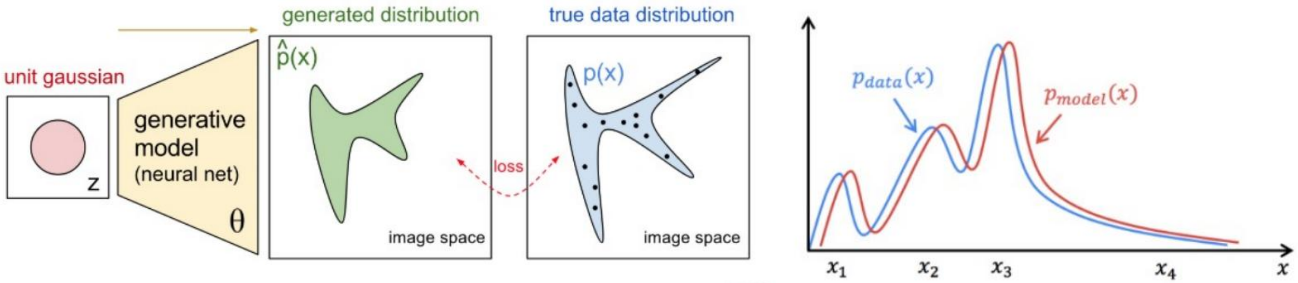
Background

- Generative AI
  - Discriminative AI vs. Generative AI
    - Discriminative AI
      - 주어진 데이터를 활용하여 예측 및 분류에 초점
      - Image Classification, Object Detection, Image Segmentation



Background

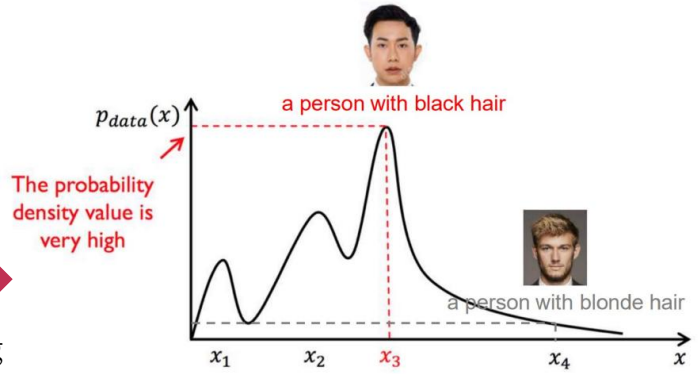
- Generative AI
  - Discriminative AI vs. Generative AI
    - Generative AI
      - 모델의 분포를 데이터 셋의 분포로 맞춰감으로써 학습 진행
    - GAN, Diffusion model



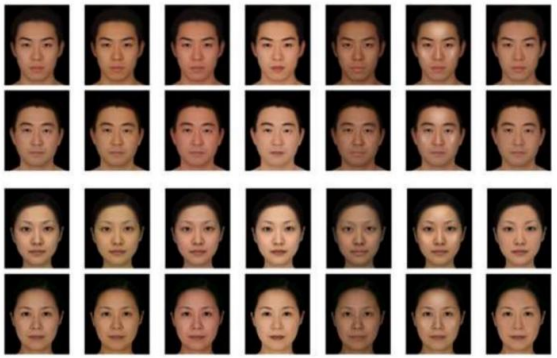


Random Distribution
(ex, Gaussian)

Matching



Distribution of Dataset



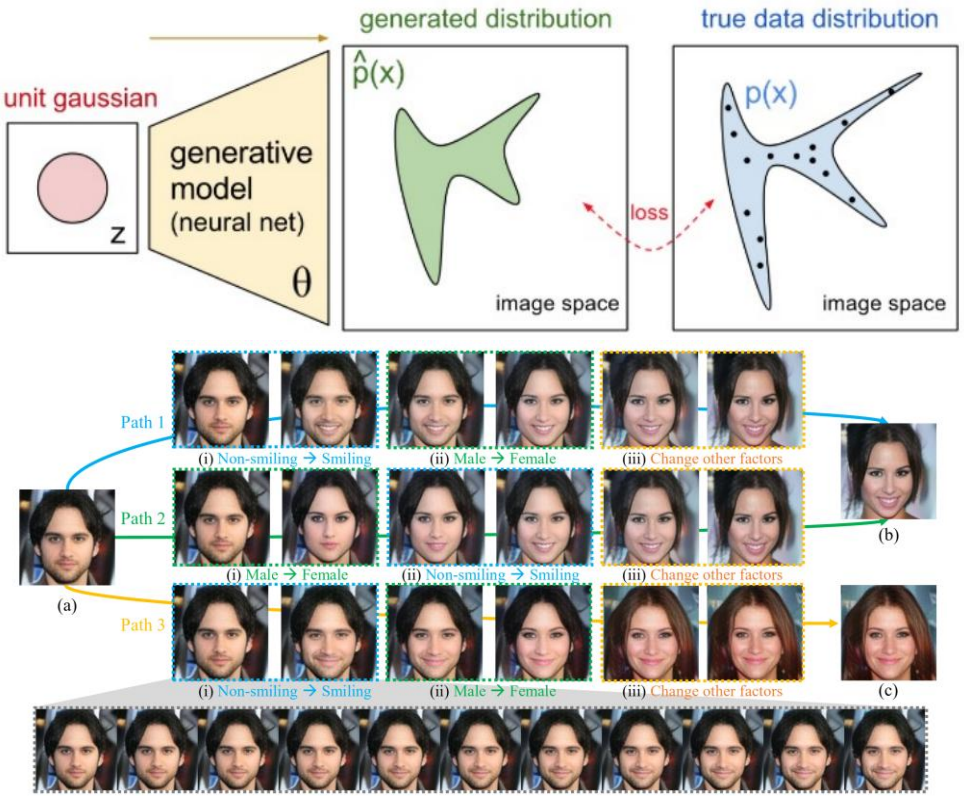
Dataset: Asian face (mostly black hair)

CVSP Lab
Computer Vision & Signal Processing

Background

* Chen, Ying-Cong, et al. "Homomorphic latent space interpolation for unpaired image-to-image translation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

- Generative AI
  - Why is it called '**Generative**' AI?
    - 모델이 입력 데이터셋으로부터 내재된 분포를 학습하기 때문에 원본 데이터에는 없었던 새로운 샘플들을 생성할 수 있음



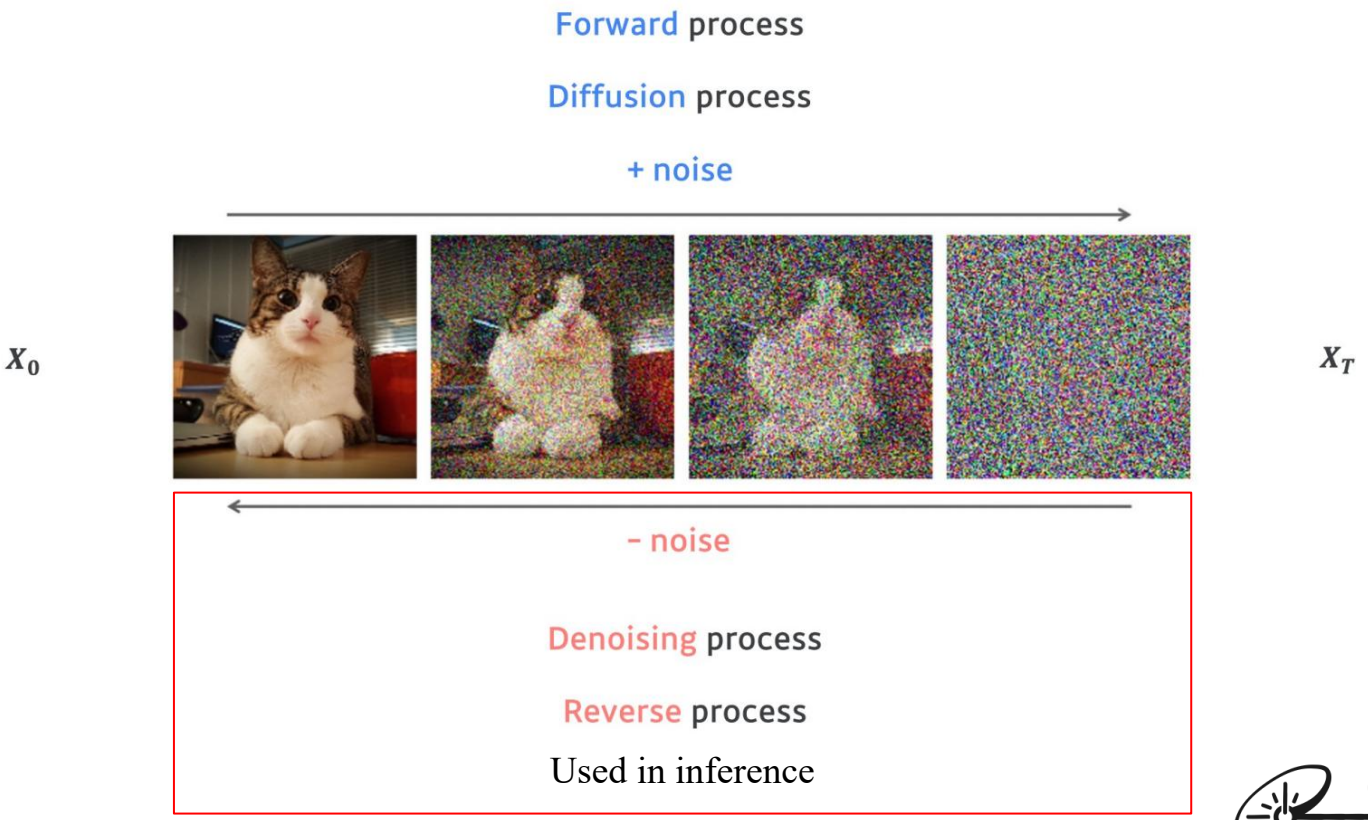
Example points interpolated between existing samples*

CVSP Lab
Computer Vision & Signal Processing

Background

* Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

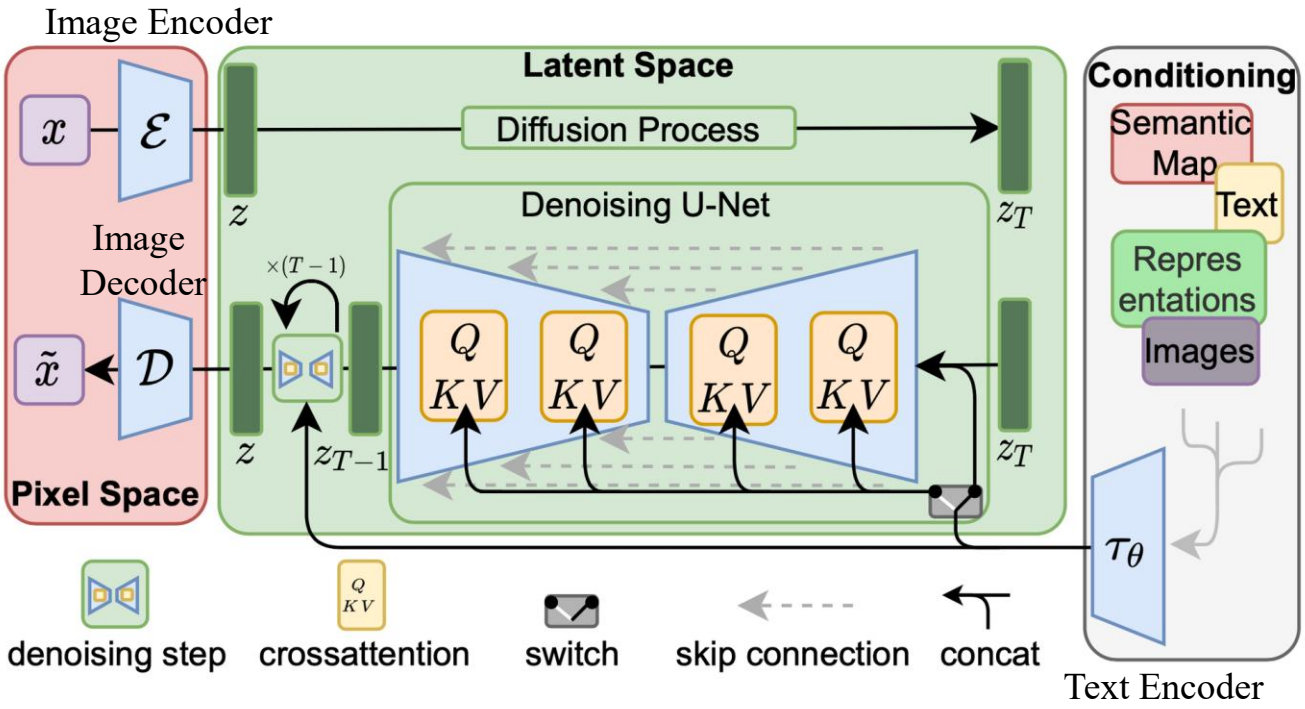
- Diffusion Model*
  - A Powerful Generative Model
    - 노이즈를 순차적으로 추가 및 제거하여 새로운 샘플을 생성하도록 학습하는 생성 모델
    - 학습된 모델은 역방향 과정에서 제거될 노이즈를 예측



Background

* Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

- Latent Diffusion Models* (Stable Diffusion)
  - A diffusion-based generative model that performs **high-resolution image generation in latent space**, thereby achieving both computational and memory efficiency
    - Can generate images as desired by conditioning on various inputs
    - Encoder: Compresses the high-resolution image (or text) x in pixel space into a low-dimensional latent vector z
    - Decoder: Reconstructs the high-resolution image $\tilde{x}$ from the denoised latent vector $\tilde{z}$.



Latent vector $z \approx$ feature
t: Time step (50)

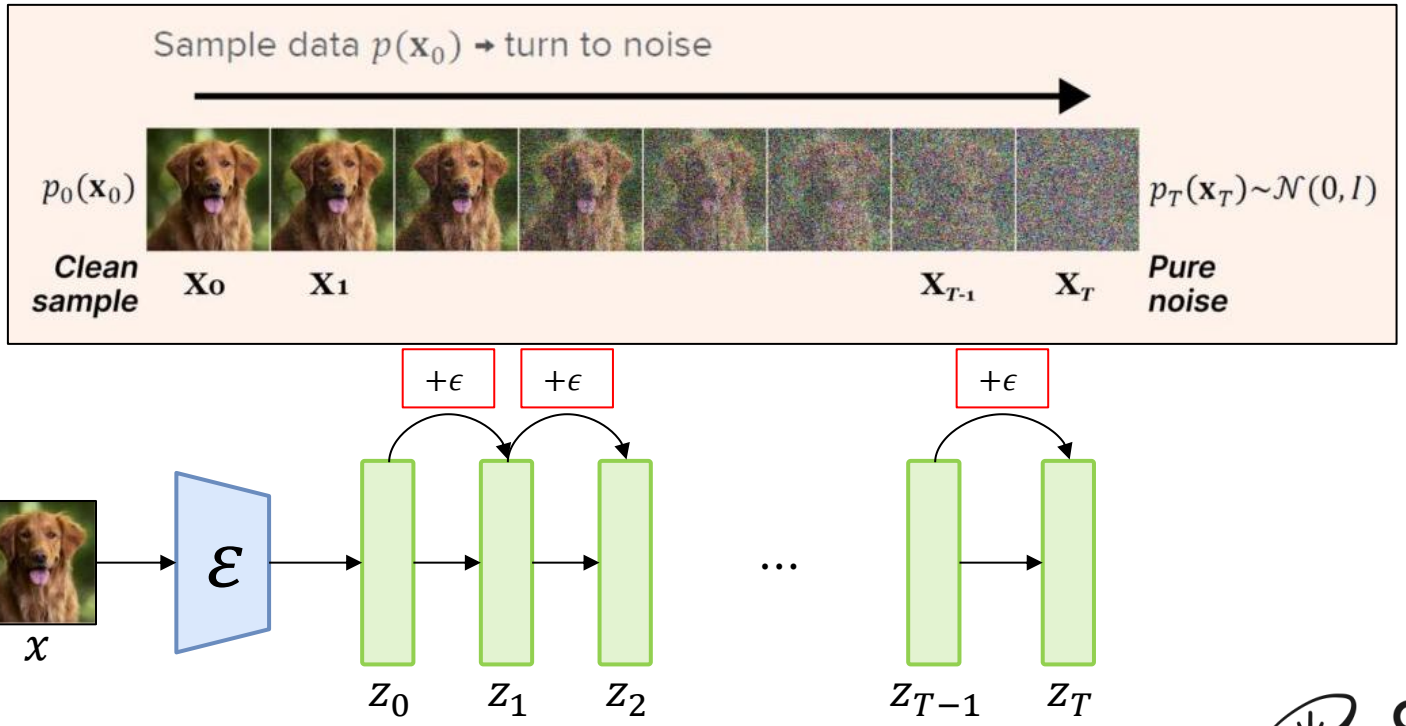
CVSP Lab
Computer Vision & Signal Processing

Background

* Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

- Latent Diffusion Models* (Stable Diffusion)
  - Train – Diffusion Process
    - A clean sample x_0 is gradually corrupted by adding Gaussian noise over T steps, forming a Markov chain
    - The original image is first encoded into a latent representation z_0, and then noise $\epsilon \sim N(0, I)$ is added step-by-step to reach a fully noisy latent vector z_T

Diffusion Process

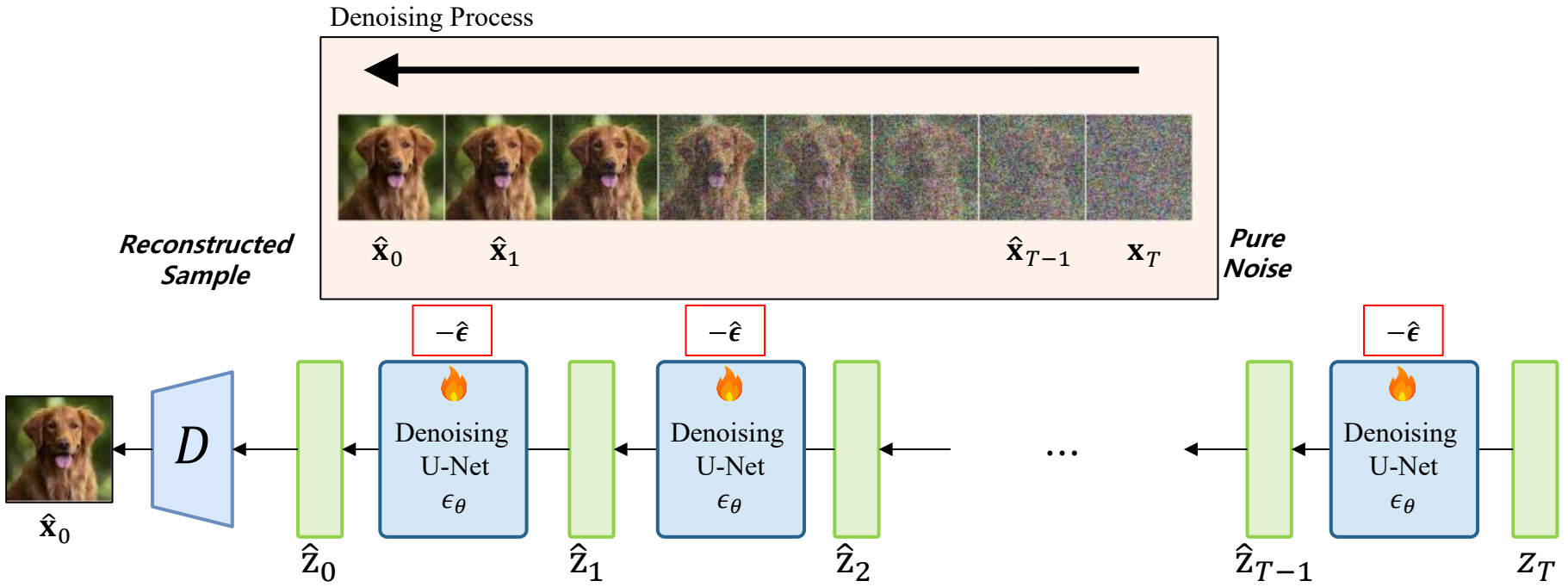


CVSP Lab Computer Vision & Signal Processing

Background

* Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

- Latent Diffusion Models* (Stable Diffusion)
  - Train – Denoising Process
    - The denoising process learns to reverse the forward diffusion by iteratively removing noise from a sampled latent vector $z_T \sim N(0, I)$
    - A denoising U-Net ϵ_θ predicts the added noise at each timestep, and subtracts it from z_T to obtain $\hat{z}_{T-1}$
    - This process continues until we obtain a clean latent representation $\hat{z}_0$, which is then decoded by D to reconstruct the image $\hat{\mathbf{x}}_0$



CVSP Lab Computer Vision & Signal Processing

*Kerbl, Bernhard, et al. "3D Gaussian splatting for real-time radiance field rendering." ACM Trans. Graph. 42.4 (2023): 139-1.

Background

- 3D Gaussian Splatting*
  - 개요
    - 3D Gaussian을 활용하여 고품질의 다각도 이미지 렌더링이 가능함
    - 기존 NeRF 기반의 방법론에 비해 압도적으로 빠른 렌더링 속도를 제공



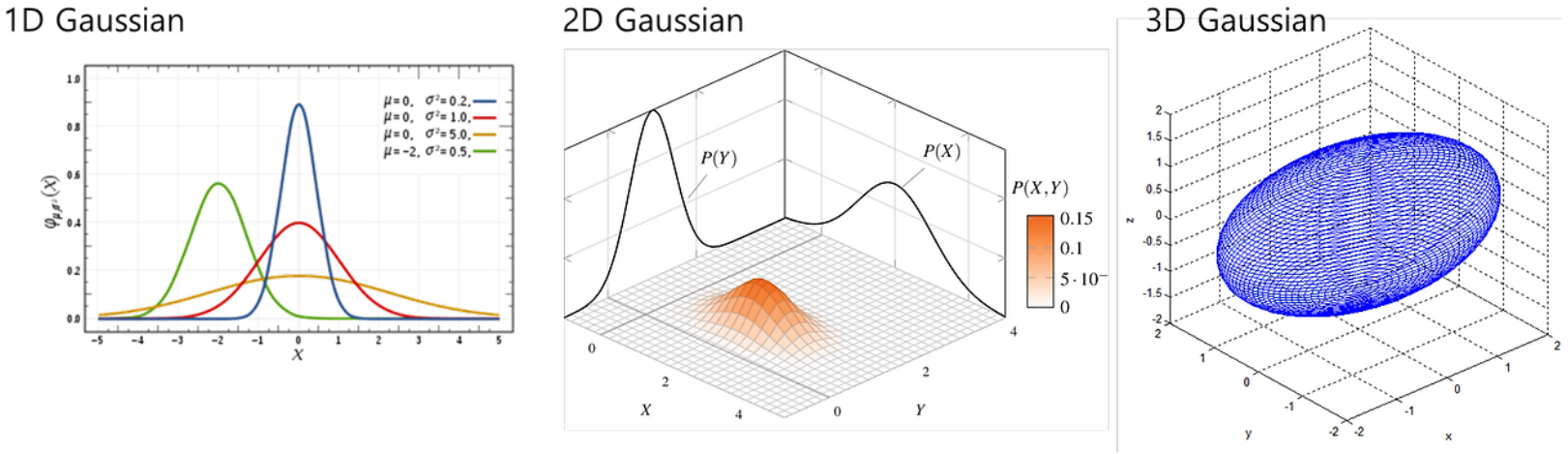
Reconstructed Scene

CVSP Lab
Computer Vision & Signal Processing

*Kerbl, Bernhard, et al. "3D Gaussian splatting for real-time radiance field rendering." ACM Trans. Graph. 42.4 (2023): 139-1.

Background

- 3D Gaussian Splatting*
  - 1D to 3D Gaussian
    - Gaussian은 확률 분포 함수로, 1D → 2D → 3D로 차원 확장이 가능함
    - 3D Gaussian은 위치, 크기, 방향을 포함하는 공간 분포체로, 복잡한 3차원 장면을 효과적으로 표현할 수 있음



1D to 3D Gaussian

CVSP Lab
Computer Vision & Signal Processing

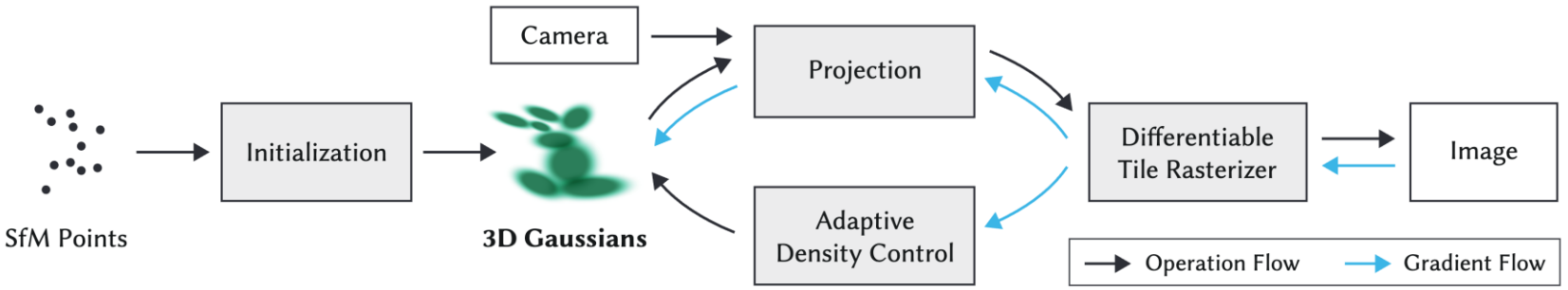
*Kerbl, Bernhard, et al. "3D Gaussian splatting for real-time radiance field rendering." ACM Trans. Graph. 42.4 (2023): 139-1.

Background

- 3D Gaussian Splatting*
  - Initialization
    - SfM Points는 입력 이미지로부터 추출된 포인트 클라우드로 3차원 좌표 정보가 포함
    - 이를 개별 Gaussians들의 위치 정보로 초기화
  - Projection & Differentiable Tile Rasterization
    - 3차원에서 2차원으로 투영시키는 과정으로, GT Image에 대응하는 카메라 파라미터로 Gaussian들의 픽셀 색상을 계산
  - Optimization
    - Rasterization 된 image와 GT image간의 L1, D-SSIM loss를 사용하여 손실 계산

$$\mathcal{L} = (1-\lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{D-SSIM}$$

    - 이를 3D Gaussians로 backpropagate하여 parameter update

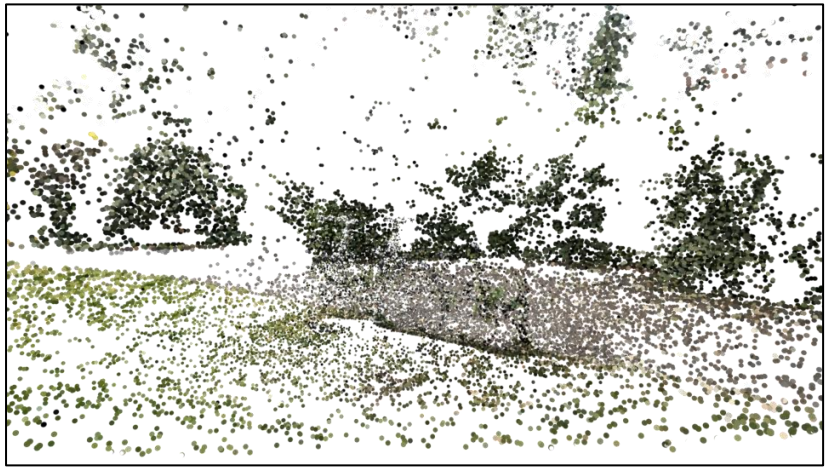


Overall Framework

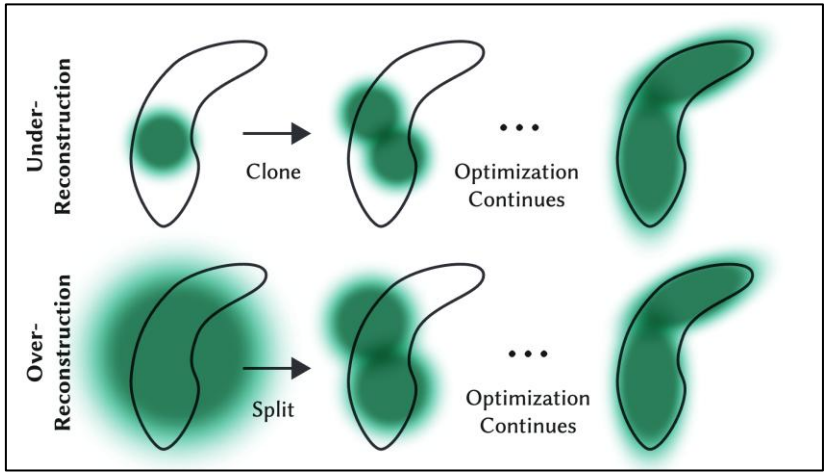
CVSP Lab
Computer Vision & Signal Processing

Background

- 3D Gaussian Splatting*
  - Adaptive Density Control
    - Sparse한 SfM Point로부터 초기화하기 때문에 Gaussians의 수를 늘려가며 복잡한 영역을 표현함
    - Clone & Split
      - gradient를 기준으로 Under-, Over- Reconstruction 영역을 탐지함
      - Scale 값을 기준으로 작은 Gaussian은 clone, 큰 경우 split 함



SfM Points



Adaptive Density Control

CVSP Lab
Computer Vision & Signal Processing

Background

* Wu, Guanjun, et al. "4d gaussian splatting for real-time dynamic scene rendering." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

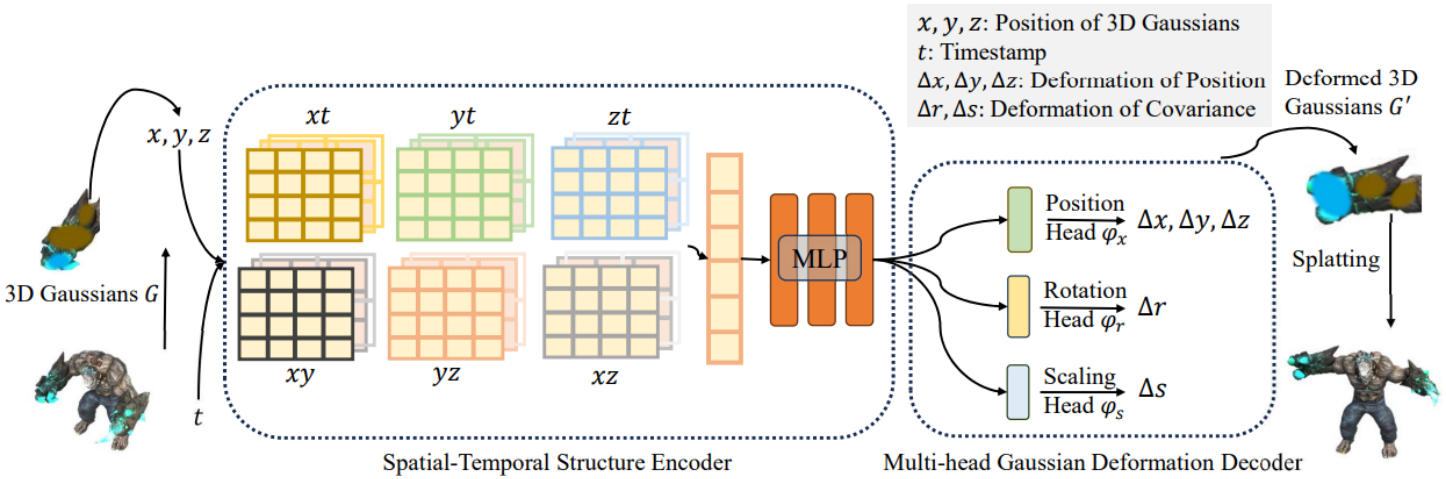
- Dynamic Gaussian Splatting*
  - 개요
    - 기존 3D Gaussian Splatting 기술을 시간 축까지 확장하여, 동적 장면의 4차원 복원을 가능하게 한 기술
    - 다각도 비디오를 입력으로 받아, 각 프레임에 대응하는 **3D Gaussian**의 **시간적 변형** 정보를 함께 예측



Background

* Wu, Guanjun, et al. "4d gaussian splatting for real-time dynamic scene rendering." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

- Dynamic Gaussian Splatting*
  - 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering
    - 하나의 Canonical 3D Gaussian을 유지하며 예측된 시간적 변형 정보를 적용하여 동적 장면을 표현함
    - Timestamp와 3D 위치 정보를 기반으로, **MLP 기반** 구조에서 **deformation을 예측**함으로써 구현됨

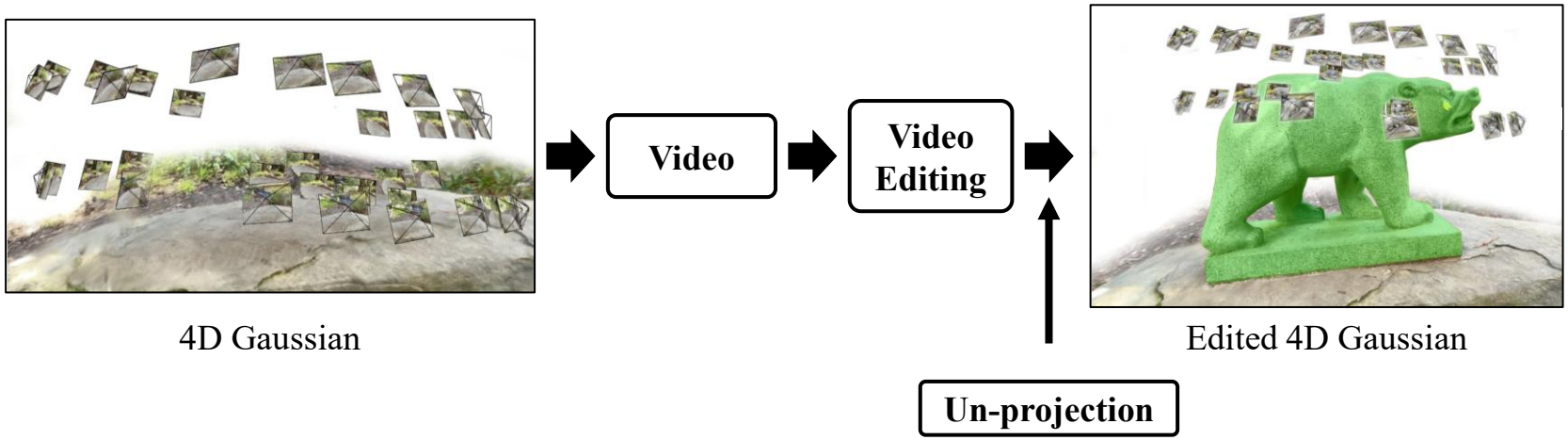


Overall Framework

CVSP Lab
Computer Vision & Signal Processing

Motivation

- 4D Gaussian Editing
  - 개요
    - 최근의 4D Gaussian에 대한 연구에서, 장면 전체 또는 객체 단위의 style을 편집하는 연구는 있으나, 객체를 새롭게 추가하는 연구는 없음
    - Video 관련 연구에서, 생성형 모델 (diffusion 등)을 활용하여 편집을 수행하는 연구가 다수 진행됨
    - 4D Gaussian에서 객체를 추가하기 위해 video editing 모델을 활용하고자 함



4D Gaussian

Edited 4D Gaussian

InFusion: Inpainting 3D Gaussians via Learning Depth Completion from Diffusion Prior

Introduction

* Wang, Junjie, et al. "Gaussianeditor: Editing 3d gaussians delicately with text instructions." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.
** Chen, Yiwen, et al. "Gaussianeditor: Swift and controllable 3d editing with gaussian splatting." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

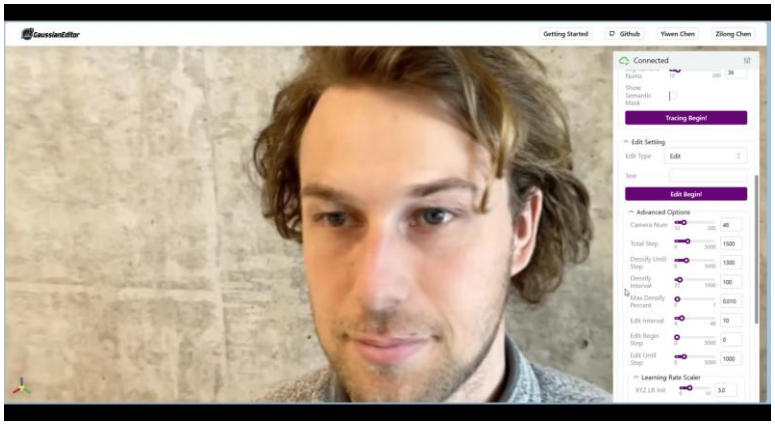
- Motivation
  - 기존 3D gaussian editing method의 한계
    - Inpainting 되지 않은 영역의 경계에서 gaussian을 성장시키는(growing) 방식 활용*
    - 2D view image를 먼저 editing 한 후, 3D gaussian들을 직접적으로 update 하는 방법**
      - Gaussian 생성할 때, 2D view image와의 불일치로 인해 blurred texture를 생성
      - 또한 gaussian 생성 및 growth 과정이 느림



Original view

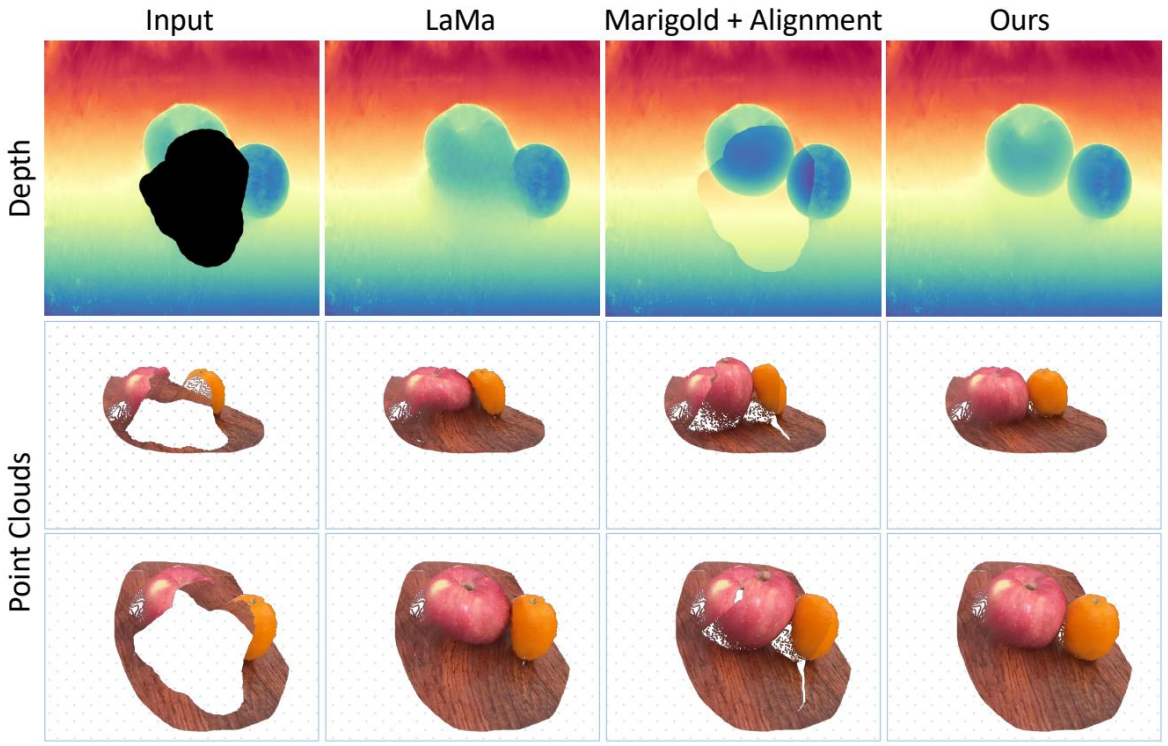


"Turn him into a clown"



Introduction

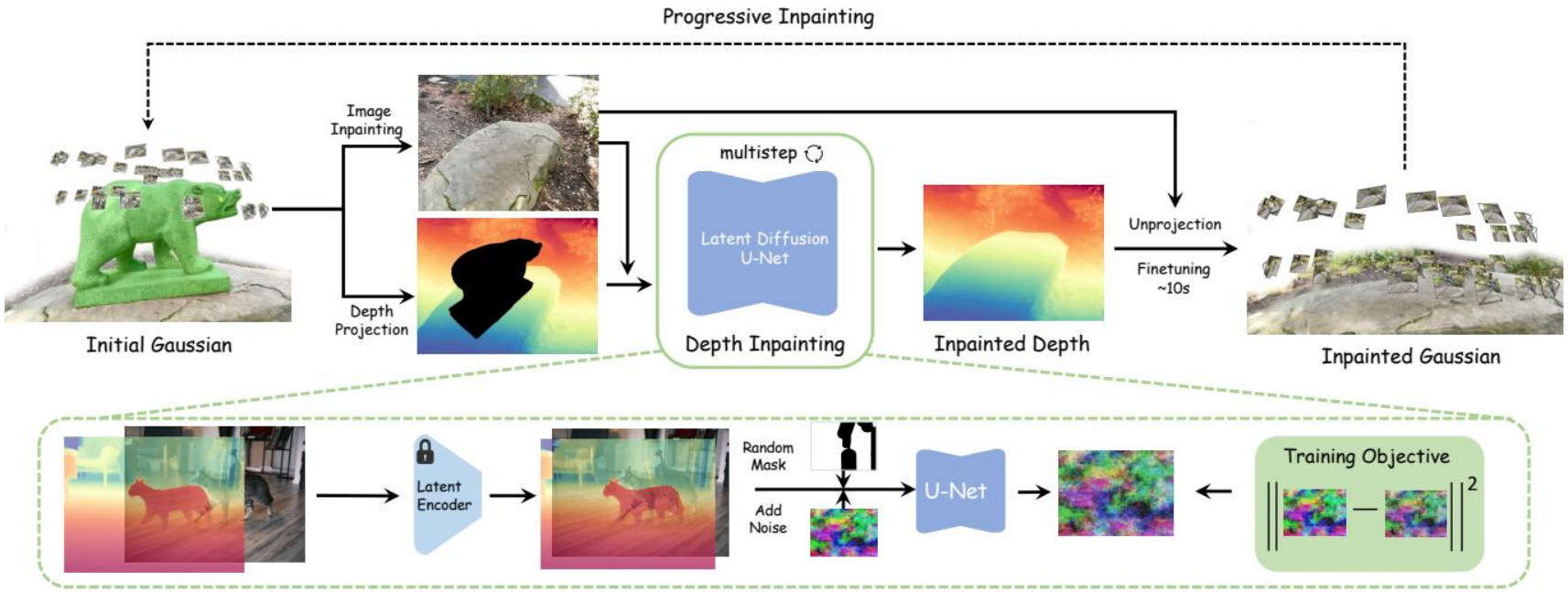
- Motivation
  - Depth 정보를 활용한 3D inpainting
    - Editing 영역에 initial points를 추가하여 editing 된 2D view image와 일치시키는 방법이 효율적임
      - Initial points의 position 정확도에 따라 quality가 결정됨
      - 이는 editing 된 2D view image의 depth의 quality에 따라 결정됨



CVSP Lab
Computer Vision & Signal Processing

Introduction

- Proposed Method
  - Depth inpainting using Diffusion prior
    - Initial point의 정확한 배치를 위해 diffusion model을 활용한 depth inpainting model 제안
    - High quality depth map을 통해 initial point를 edited image에 맞춰 정확하게 배치 가능
    - 매우 적은 반복 (약 100회)의 fine-tunning으로 3D inpainting 수행 가능



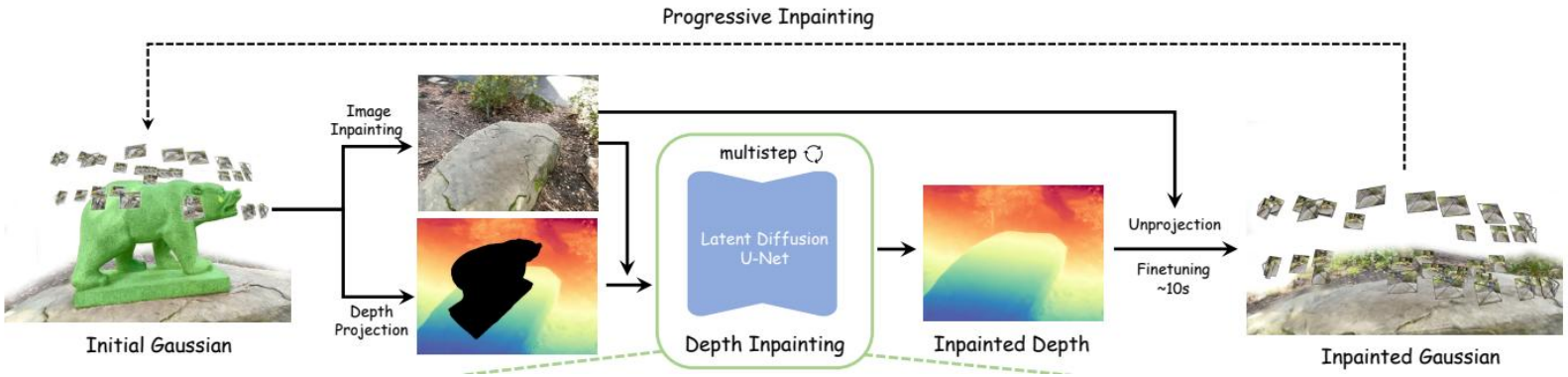
Method

- Overview
  - Observation
    - Novel view synthesis에서, 3D gaussian의 reconstruction quality는 **initial point의 위치에 민감**함
    - Inpainting된 depth map을 활용하여 빈 공간으로 나타나는 editing 영역의 gaussian들을 initialization하는 것은 3D gaussian inpainting에 유리함
    - Depth estimation에 diffusion prior를 적용하는 것은 정확도를 현저히 향상시킴
  - Diffusion Models for Depth Completion
    - 기존의 depth inpainting model은 inpainting되지 않은 영역과 일치하지 않는 depth를 생성
      - 이러한 불일치는 3D gaussian inpainting의 품질을 저하시킴
    - 이를 해결하기 위해 diffusion process의 강점을 활용하는 generalized depth inpainting model을 개발

CVSP Lab
Computer Vision & Signal Processing

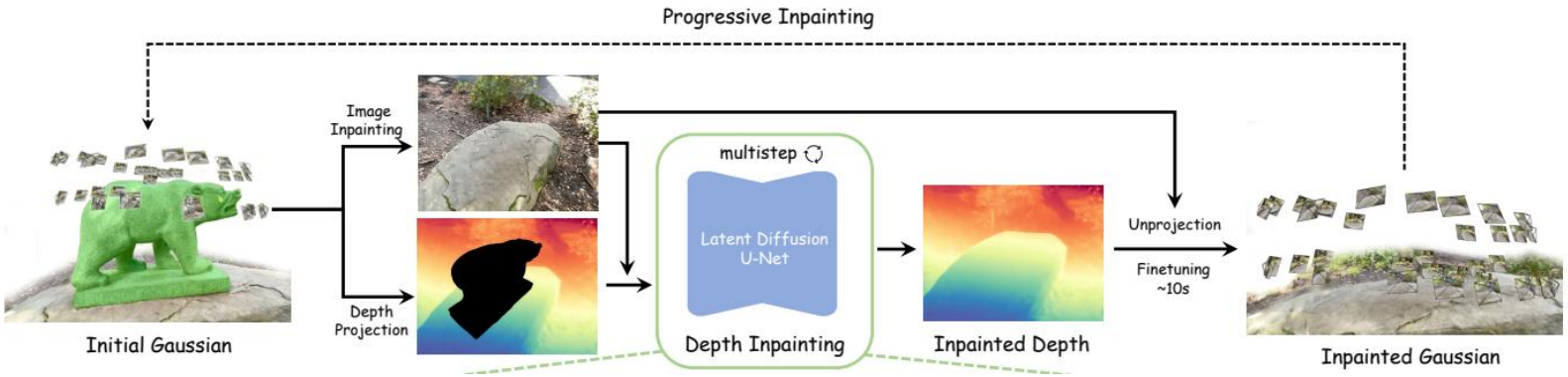
Method

- 3D Inpainting
  - Process
    1. 초기 설정 및 reference view 선택
       - 최적화된 기존의 3D gaussian에서 inpainting할 영역 mask로 지정
       - 원치 않는 gaussian 분할 및 제거
    2. 2D Image Inpainting
       - 선택된 reference view를 2D image로 projection, mask와 결합
       - Stable diffusion 기반 2D inpainting 모델 적용하여 이미지에 대해 inpainting 수행



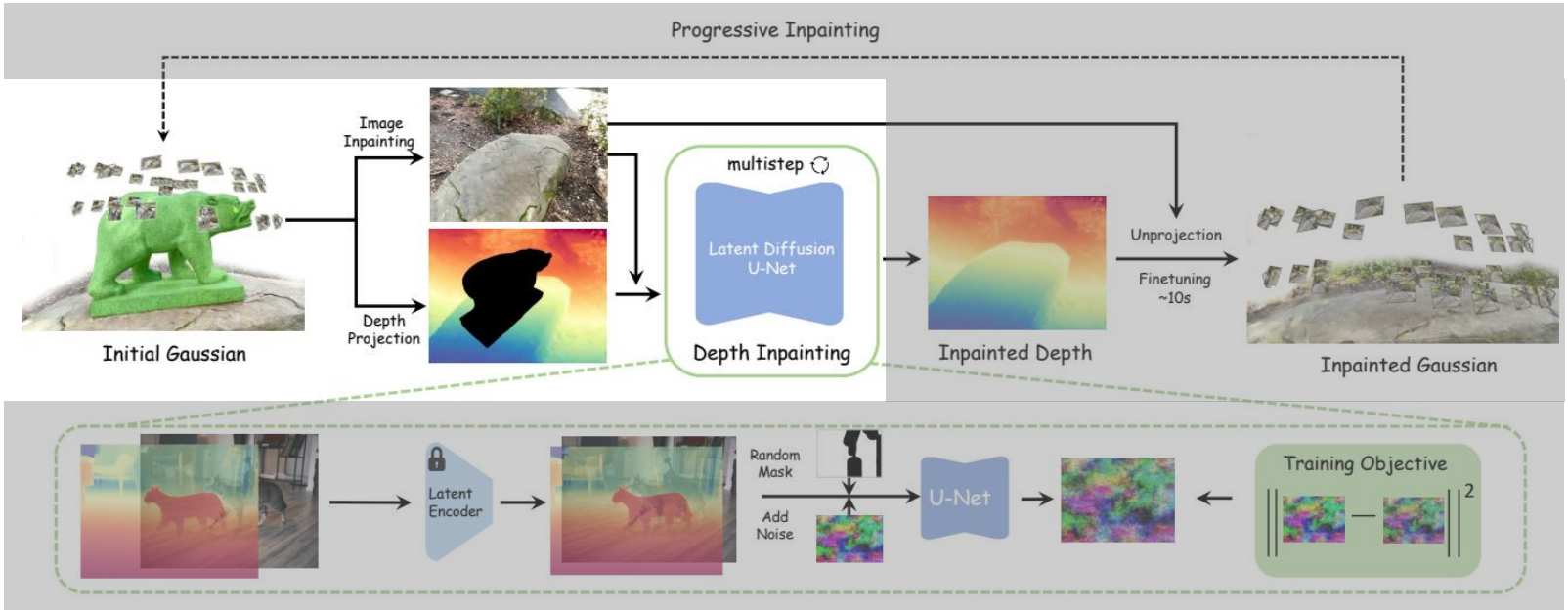
Method

- 3D Inpainting
  - Process
    3. 원치 않는 gaussian이 분할 및 제거된 3D gaussian에 대해 depth 추출
    4. **Diffusion Models for Depth Completion**을 통해 depth inpainting
    5. 3D un-projection 및 fine-tuning



Method

- Diffusion Models for Depth Completion
  - 개요
    - Latent diffusion model (LDM)을 활용
    - Input
      - Depth map d
      - Color image I
      - Mask m



Method

* Ke, Bingxin, et al. "Repurposing diffusion-based image generators for monocular depth estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

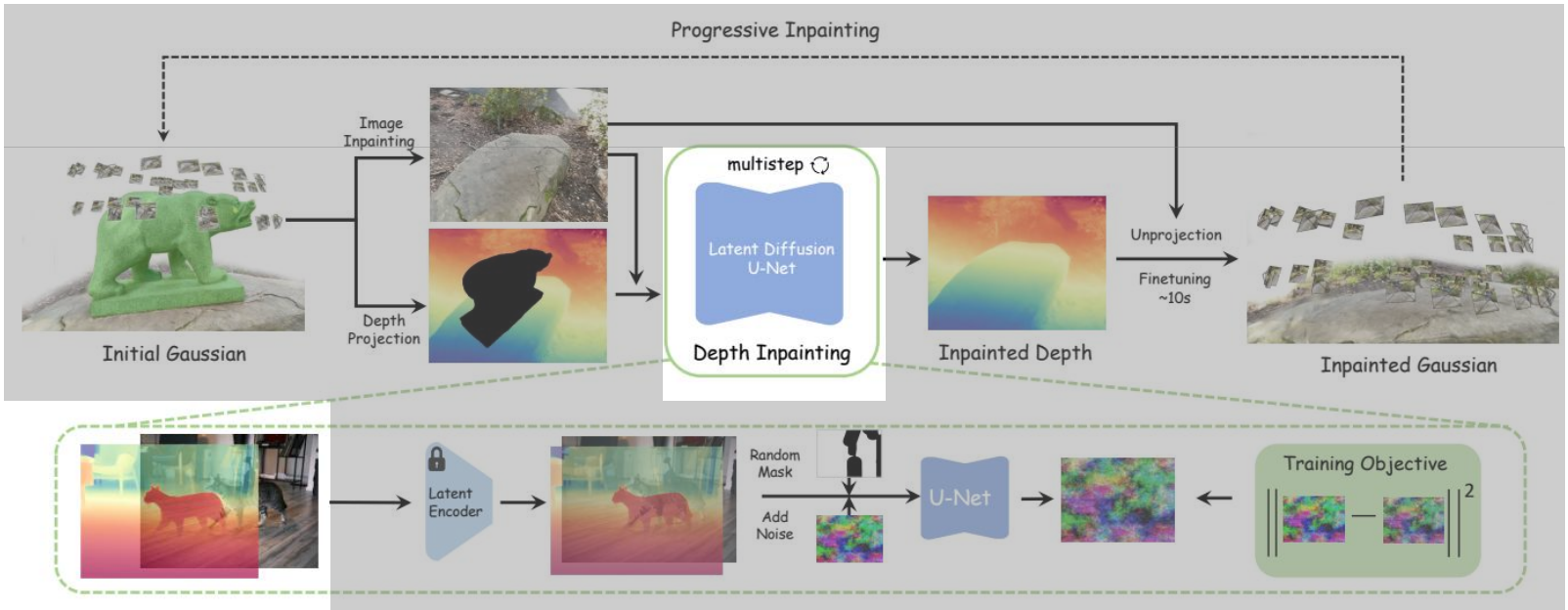
- Diffusion Models for Depth Completion
  - Training
    - 먼저 depth 정규화 수행
    - 3 channel depth → single channel depth map
      - Marigold*에서 활용하는 linear normalization 적용 → [-1, 1]

$$d' = \frac{d - d_2}{d_{98} - d_2} \times 2 - 1,$$

Marigold
- LDM 기반 monocular depth estimation model

d_2, d_{98}
- 개별 depth map의 percentile
- 극단적인 이상치를 제외하기 위해 설정

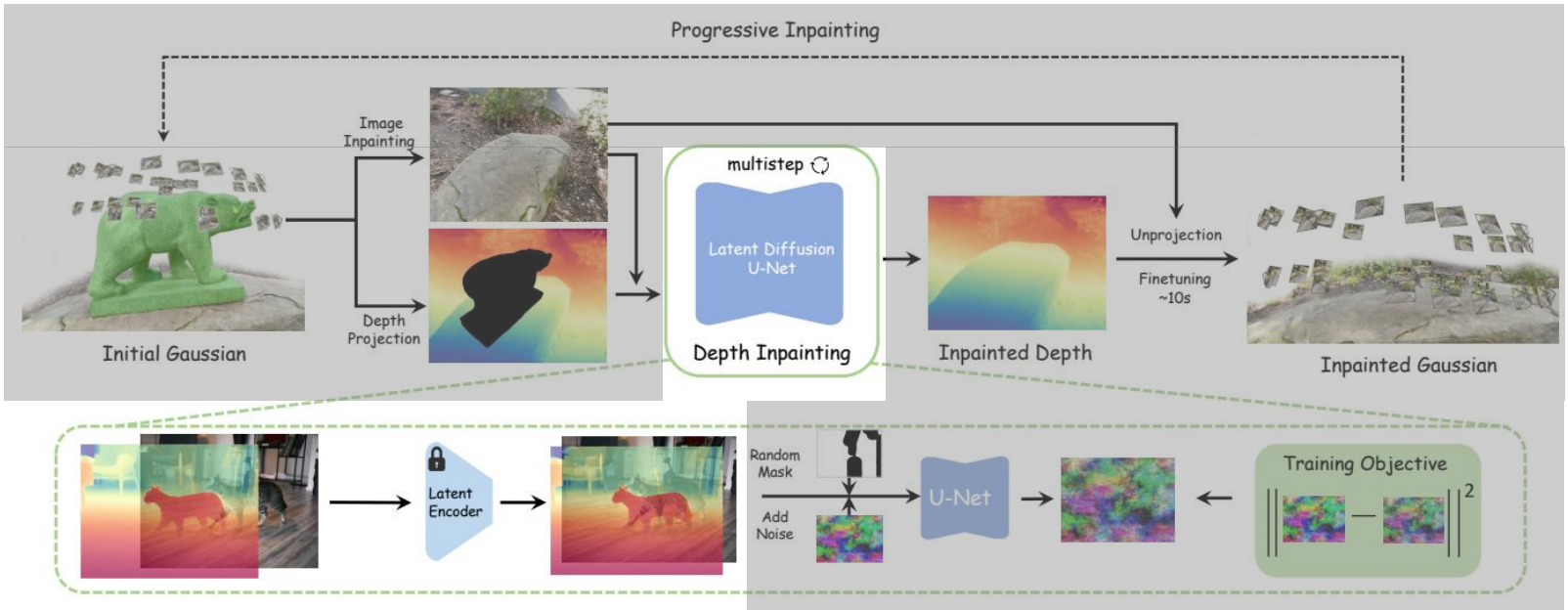
Linear Normalization 수행 이유
- VAE의 입력 범위를 맞추기 위해



Method

- Diffusion Models for Depth Completion
  - Training
    - 정규화된 depth d', color image I는 VAE $\mathcal{E}$를 통해 latent space로 encoding

$$z^{d'} = \mathcal{E}(d'), \qquad z^{I} = \mathcal{E}(I)$$



Method

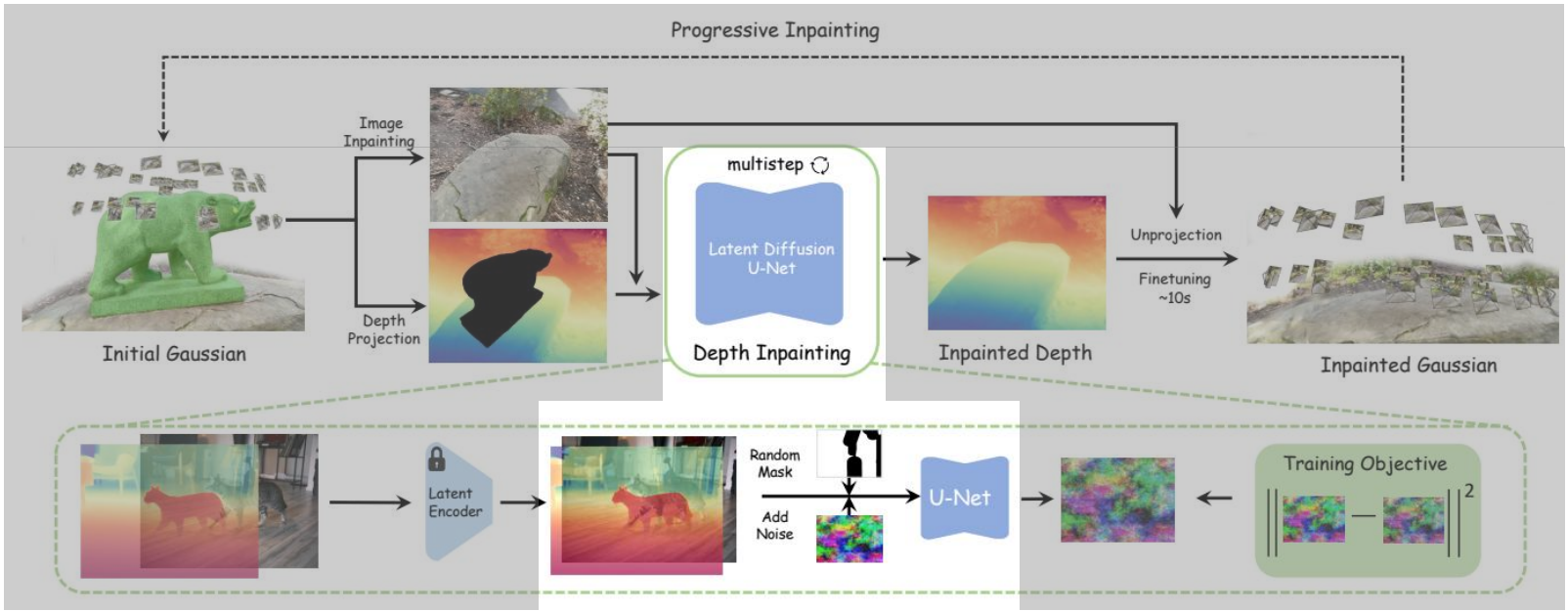
- Diffusion Models for Depth Completion
  - Training
    - U-Net 학습을 위해, mask m을 무작위로 생성, 개별 sample들의 전체 영역 중 30%만 적용
    - U-Net input

Inference 시 mask
- SAM-Track* 활용

$$z_t = \text{cat}\left(z_t^{d'}, z_m^{d'}, z^I, m'\right) \Rightarrow 4+4+4+1$$

✓$z_t^{d'}$: noise가 적용된 depth의 latent feature, $z_m^{d'}$: noise가 없는 depth feature에 mask 적용

✓encoder를 통과한 이미지 z^I, Downscaled mask m'



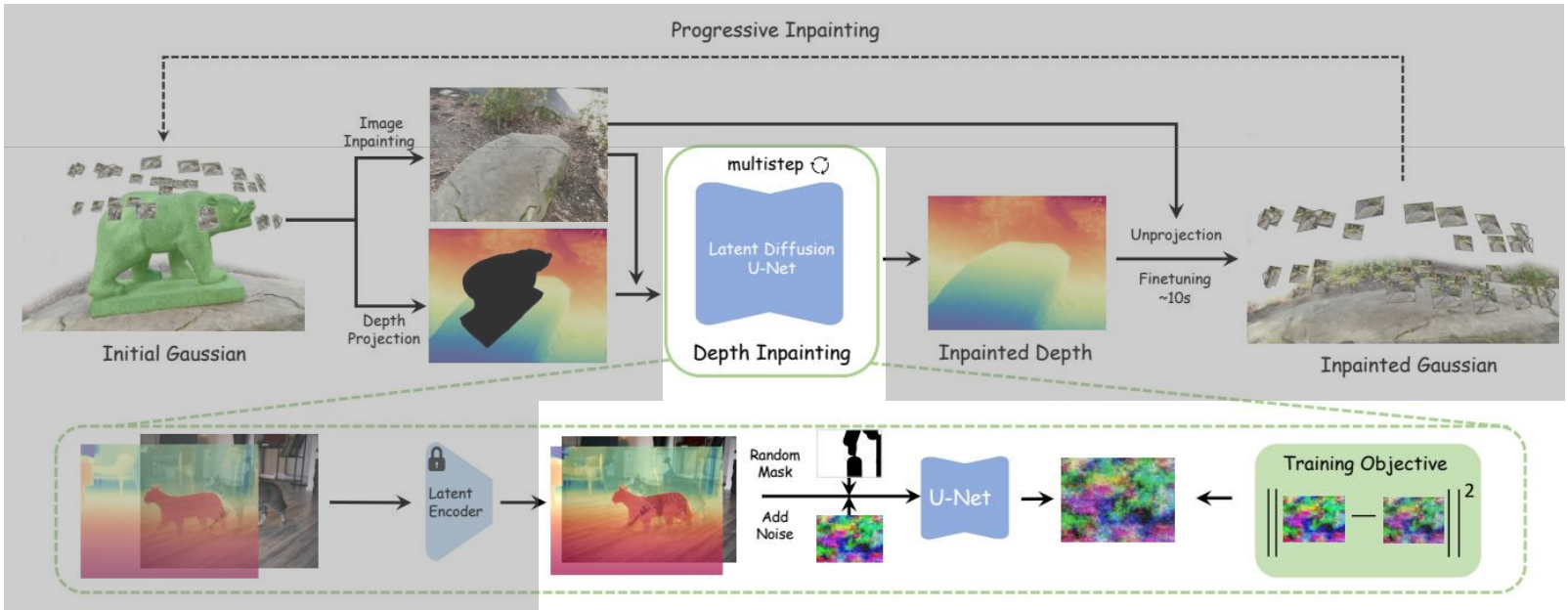
Method

* Ke, Bingxin, et al. "Repurposing diffusion-based image generators for monocular depth estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

- Diffusion Models for Depth Completion
  - Training
    - U-Net 학습 시, U-Net은 noise가 추가된 latent feature z_t를 입력으로 받아 원본 noise ϵ를 예측하도록 학습

$$\mathcal{L} = \mathbb{E}_{d,I,t,\epsilon\sim N(0,1)} \|\epsilon - \epsilon_\theta(z_t)\|_2^2$$

      - d, I, t: depth map, color image, time step
      - ϵ: 시간 t에서 z_t에 추가된 noise
      - ϵ_θ: U-Net 모델 θ가 z_t를 입력으로 받아 예측한 noise



* https://lmb.informatik.uni-freiburg.de/resources/datasets/SceneFlowDatasets.en.html

Experiments

- Diffusion Models for Depth Completion
  - Training
    - Dataset: SceneFlow dataset*
      - 100,000개 이상의 프레임과 GT depth 정보를 포함
    - LDM은 Marigold에서 얻은 pretrained depth estimation weight로 초기화
    - 200 epoch, 8개의 A100 GPU, one day



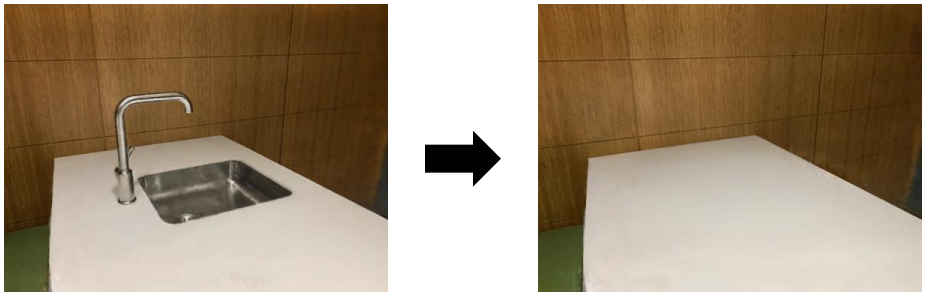
Scene Flow dataset 일부

CVSP Lab
Computer Vision & Signal Processing

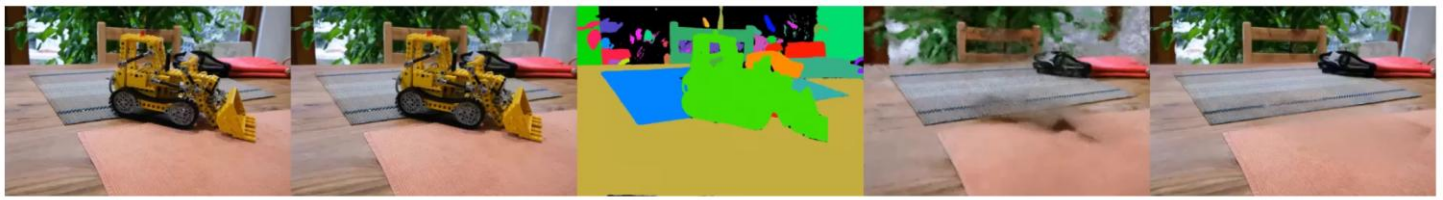
Experiments

* Mirzaei, Ashkan, et al. "Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
** Ye, Mingqiao, et al. "Gaussian grouping: Segment and edit anything in 3d scenes." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

- Diffusion Models for Depth Completion
  - Baseline
    - Spin-NeRF*
      - NeRF(Neural Radiance Fields) 기반의 3D inpainting 방법
    - Gaussian-grouping**
      - 3D Gaussian Splatting (3DGS)을 활용하는 3D inpainting 방법
      - Inpainting을 위해 diffusion model 활용, depth가 아닌 segmentation 활용



Spin-NeRF



Gaussian-grouping

CVSP Lab
Computer Vision & Signal Processing

Experiments

- Diffusion Models for Depth Completion
  - Qualitative Results
    - 기존의 방법들 보다 시각적으로 향상된 결과
      - 1개의 reference view에 대해서만 inpainting 및 un-projection을 수행했을 때에도 정확한 depth 추정으로 인해 inpainting 영역 외의 gaussian과 consistency가 일치



Infusion



CVSP Lab
Computer Vision & Signal Processing

Experiments

- Diffusion Models for Depth Completion
  - Quantitative Results
    - LPIPS (Learned Perceptual Image Patch Similarity)
      - 인간의 지각(Perception)과 유사하게 이미지의 유사성을 측정하도록 학습된 지표
      - Pretrained model (AlexNet, VGG, SqueezeNet 등)의 feature extraction layer를 사용
      - 비교할 두 이미지(원본 이미지와 생성된 이미지)를 이 모델에 통과시켜 각 레이어에서 feature map을 추출
      - 추출된 feature map 들 간의 Euclidean distance를 계산하고, 이 거리에 학습된 가중치를 곱하여 최종 LPIPS 점수를 산출
        - ✓가중치는 인간의 유사성 판단과 일치하도록 별도로 학습

	SPIn-NeRF	Gaussians Grouping	Ours
LPIPS ↓	0.465	0.454	**0.421**
FID ↓	156.64	123.48	**92.62**
Time ↓	5h	20min	**40s**

Experiments

- Diffusion Models for Depth Completion
  - Quantitative Results
    - FID (Fréchet Inception Distance)
      - 생성된 이미지의 품질과 다양성을 평가하는 데 사용되는 지표
      - 생성 모델이 만들어낸 이미지 분포가 실제 이미지의 분포와 얼마나 유사한지를 측정
        - ✓사전 훈련된 Inception-v3 모델을 feature extractor로 사용
        - ✓실제 이미지 데이터셋과 생성된 이미지 데이터셋을 Inception-v3 모델에 통과시켜 각 이미지의 feature vector를 추출
        - ✓각 데이터셋의 특징 벡터들이 이루는 분포의 평균과 공분산을 계산
        - ✓두 gaussian 분포 간의 Fréchet 거리(Fréchet distance)를 계산

	SPIn-NeRF	Gaussians Grouping	Ours
LPIPS ↓	0.465	0.454	**0.421**
FID ↓	156.64	123.48	**92.62**
Time ↓	5h	20min	**40s**

CVSP Lab
Computer Vision & Signal Processing

Experiments

- Diffusion Models for Depth Completion
  - Quantitative Results
    - LPIPS, FID metric에서 더 낮은 score 달성
      - 모델이 생성한 3D inpainting area을 2D image로 렌더링한 결과물과 해당 영역의 원본 2D image를 비교
    - Fine-tuning 속도는 제안하는 방법이 압도적으로 낮음
      - 약 100회의 반복으로도 정확한 reconstruction 수행 가능

	SPIn-NeRF	Gaussians Grouping	Ours
LPIPS ↓	0.465	0.454	**0.421**
FID ↓	156.64	123.48	**92.62**
Time ↓	5h	20min	**40s**

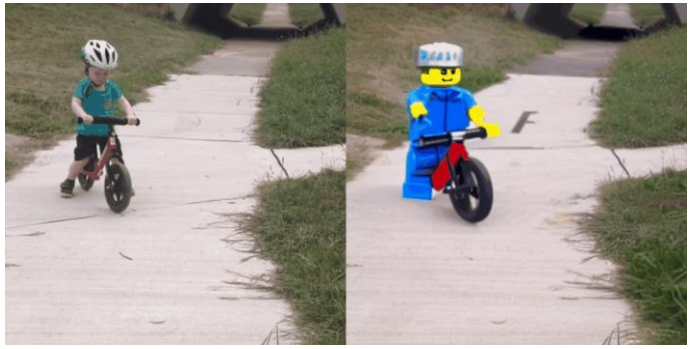
CVSP Lab
Computer Vision & Signal Processing

Señorita-2M: A High-Quality Instruction-based Dataset for General Video Editing by Video Specialists

Introduction

* Liu, Shaoteng, et al. "Video-p2p: Video editing with cross-attention control." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
** Cheng, Jiaxin, Tianjun Xiao, and Tong He. "Consistent Video-to-Video Transfer Using Synthetic Dataset." In *International Conference on Learning Representations* (ICLR), 2024.

- Motivation
  - 기존 video editing 방법의 한계
    - Inversion-based method
      - 학습이 필요없지만, inference 시간이 오래 걸리고, 세밀한 편집을 수행하기 어려움
      - Artifact 및 jitter를 생성하는 경향이 있음
      - Frame consistencty를 유지하기 어려움
    - End-to-End method
      - 편집된 비디오 쌍을 학습에 사용하며 더 빠른 inference 속도를 제공하지만, 고품질 학습 비디오 쌍의 부족으로 인해 종종 만족스럽지 못한 편집 결과를 산출



Inversion-based Video Editing Method
- VideoP2P*



End-to-End Video Editing Method
- InsV2V**

CVSP Lab
Computer Vision & Signal Processing

Introduction

- Proposed Method
  - Señorita-2M dataset 구축
    - End-to-end video editing method를 위한 고품질 비디오 editing dataset 구축
      - 약 2백만 개의 video editing paire로 구성
      - 고품질 학습 video pair 부족을 직접적으로 해소하여, 모델이 효과적으로 학습할 수 있는 기반 제공



Add a hat on girl's head.



Make it anime style.



Make it watercolor style.

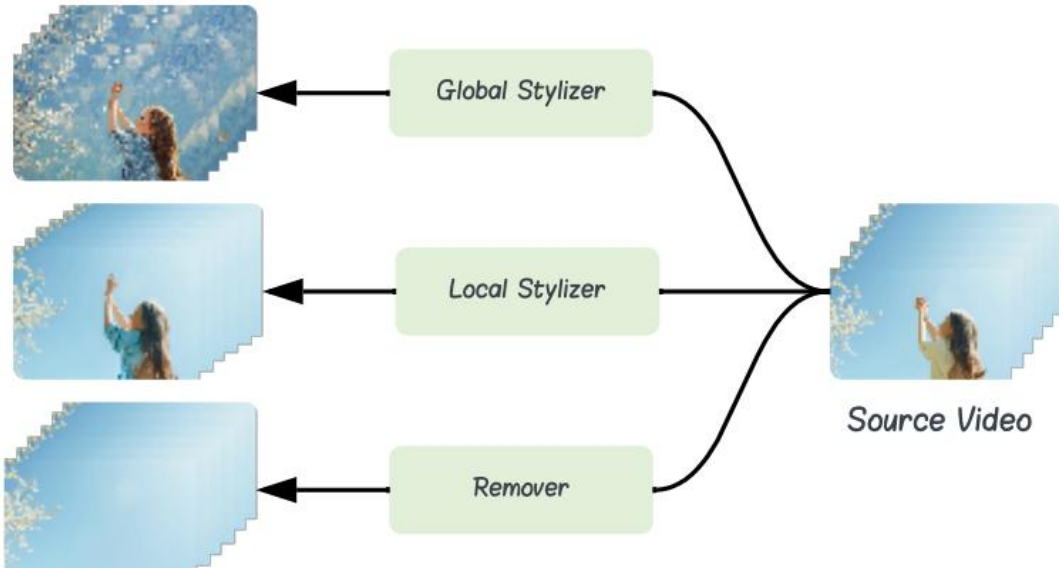


Add a rainbow.

CVSP Lab
Computer Vision & Signal Processing

Introduction

- Proposed Method
  - High Quality Video Editing Experts
    - Señorita-2M 데이터셋 구축 및 Quality를 보장하기 위해, 4가지 High Quality의 expert video editing model을 직접 제작하고 학습
      - Global Stylizer, Local Stylizer, **Object Remover**, Object Swapper



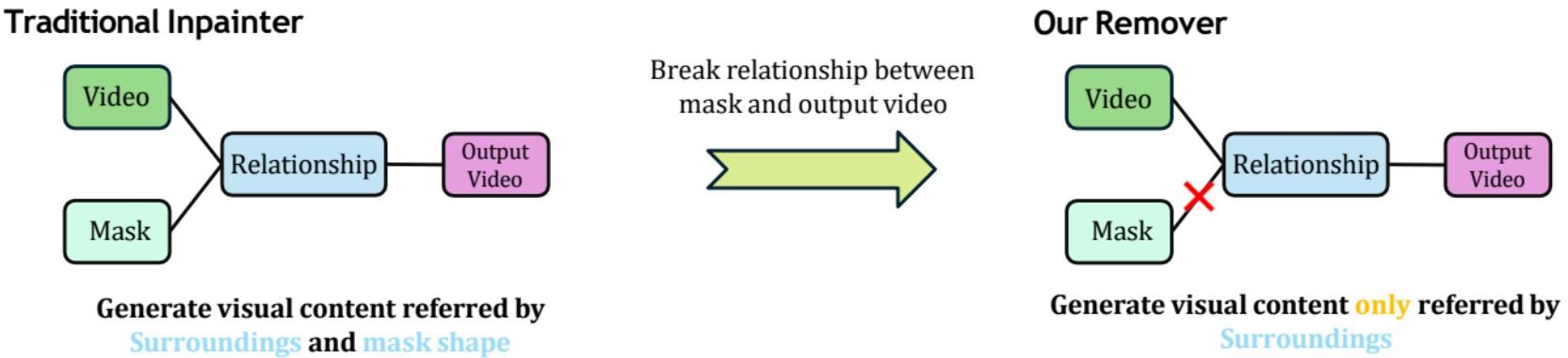
CVSP Lab
Computer Vision & Signal Processing

Method

* Zhou, Shangchen, et al. "Propainter: Improving propagation and transformer for video inpainting." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

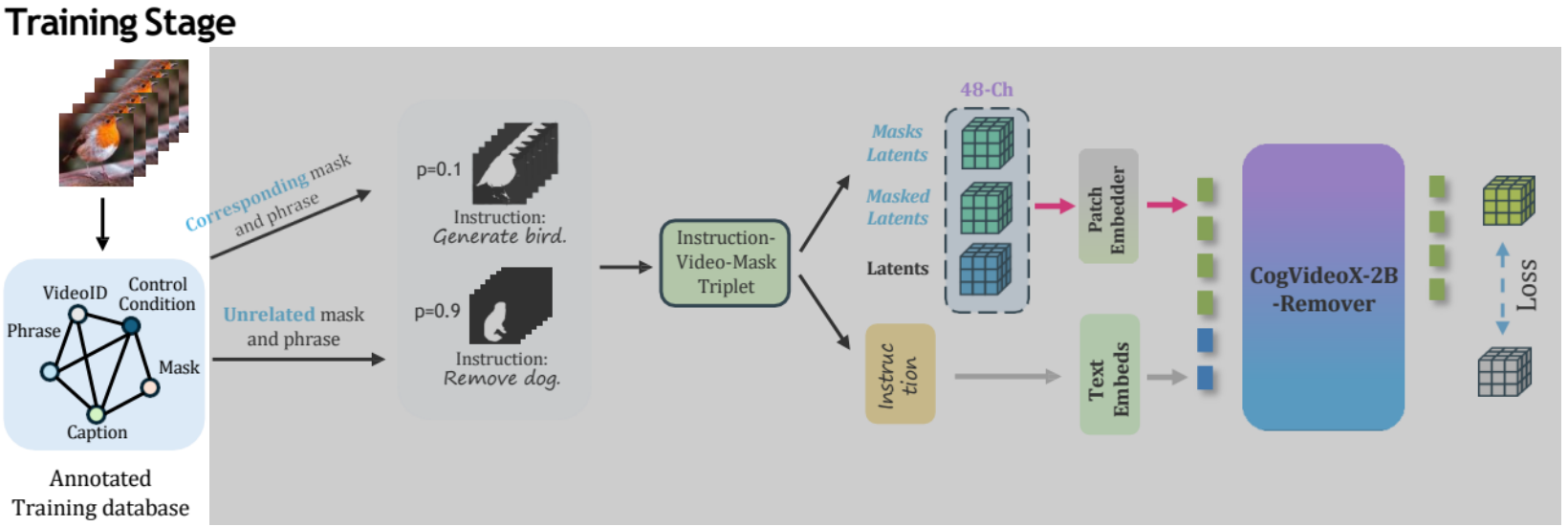
- Proposed Method
  - Object Remover
    - 기존 inpainting model은 object를 제거할 때, blur가 생성되고 mask 모양에 맞춰 객체를 생성하려는 경향이 있음
    - CogVideoX-2B 기반 video object remover 제안
      - 생성된 콘텐츠와 마스크 모양 간의 상관 관계를 끊기 위한 학습 전략 제안

CogVideoX-2B
- ZhiPu AI에서 출시한 오픈 소스 비디오 생성 모델
- 3D VAE



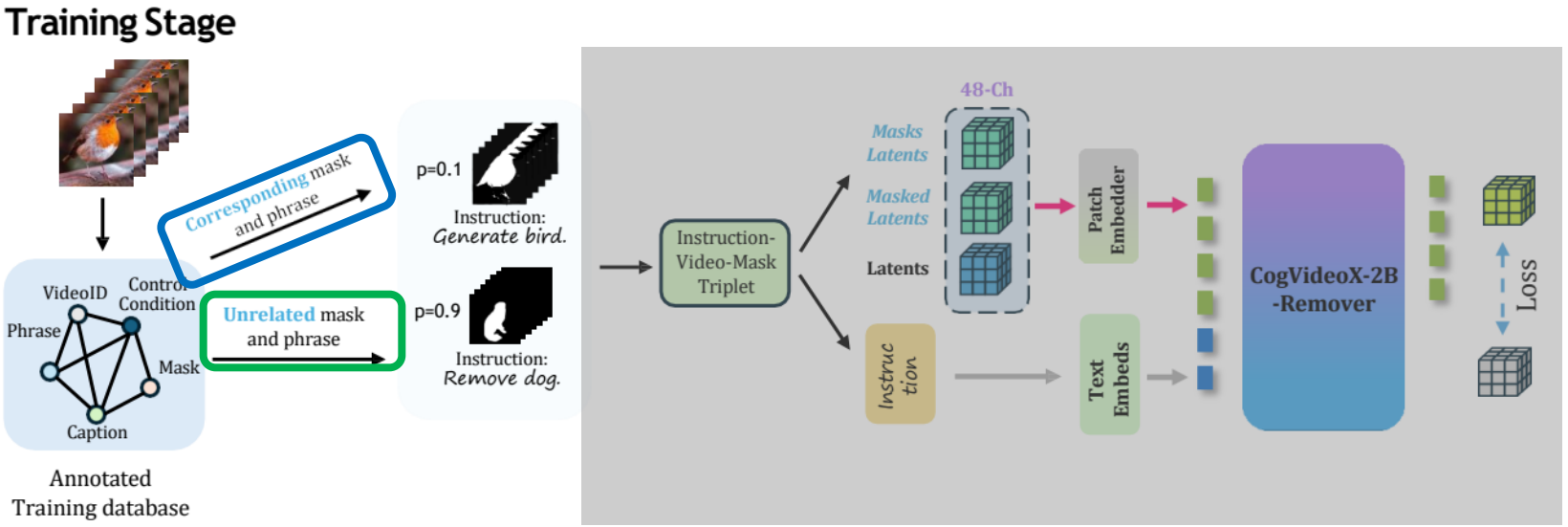
Method

- Proposed Method
  - Object Remover Training Stage
    - Annotated Training database
      - VideoID: 각 비디오의 고유 식별자 (label)
      - Phrase: 비디오 내용 또는 editing instruction과 관련된 텍스트 구문 (prompt)
      - Caption: 비디오에 대한 설명
      - Mask: 비디오 내 특정 영역을 나타내는 mask
      - Control Condition: 비디오 편집에 사용될 수 있는 추가 제어 정보 (Canny Edges, HED, Depth 등)



Method

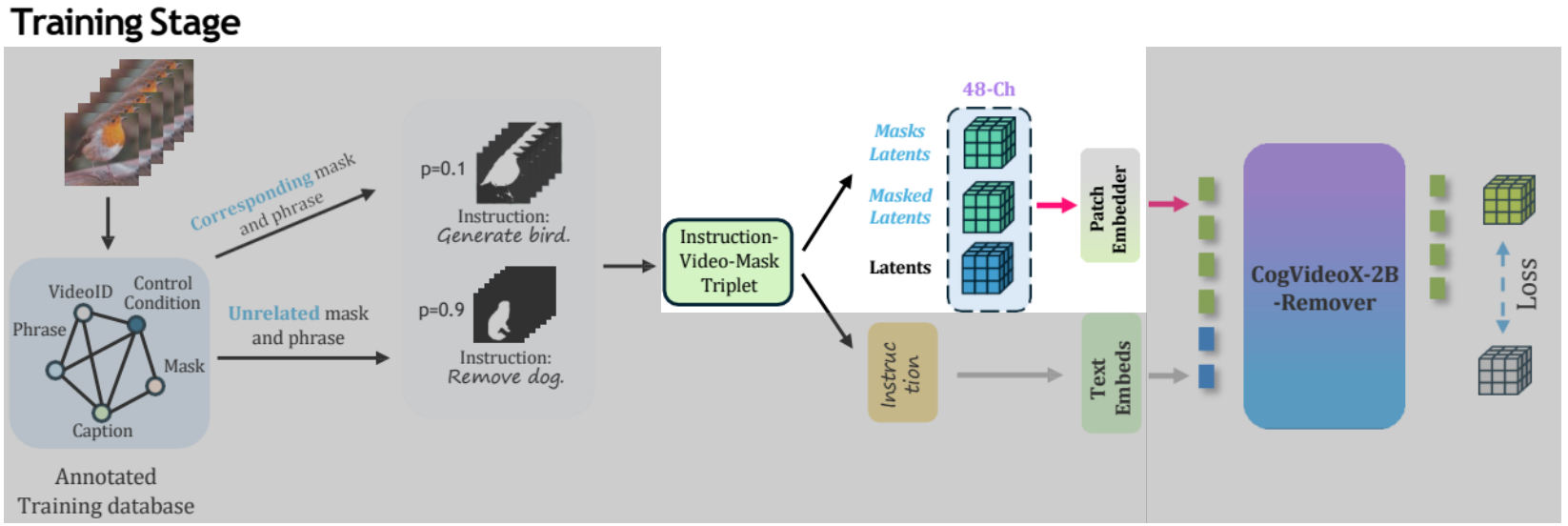
- Proposed Method
  - Object Remover Training Stage
    - Mask Selection and Assigning Instructions
      - Corresponding mask and phrase: 비디오 콘텐츠와 **일치하는** mask와 phrase가 사용됨
        - ✓비디오에 없는 객체를 추가하는 능력(객체 추가)을 모델이 학습하도록 유도
      - Unrelated mask and phrase: 비디오 콘텐츠와 **관련 없는** mask와 phrase가 사용됨
        - ✓모델이 masked 영역을 mask의 모양에 구애받지 않고 텍스트 phrase에 따라 자연스럽게 채울 수 있도록 학습



CVSP Lab
Computer Vision & Signal Processing

Method

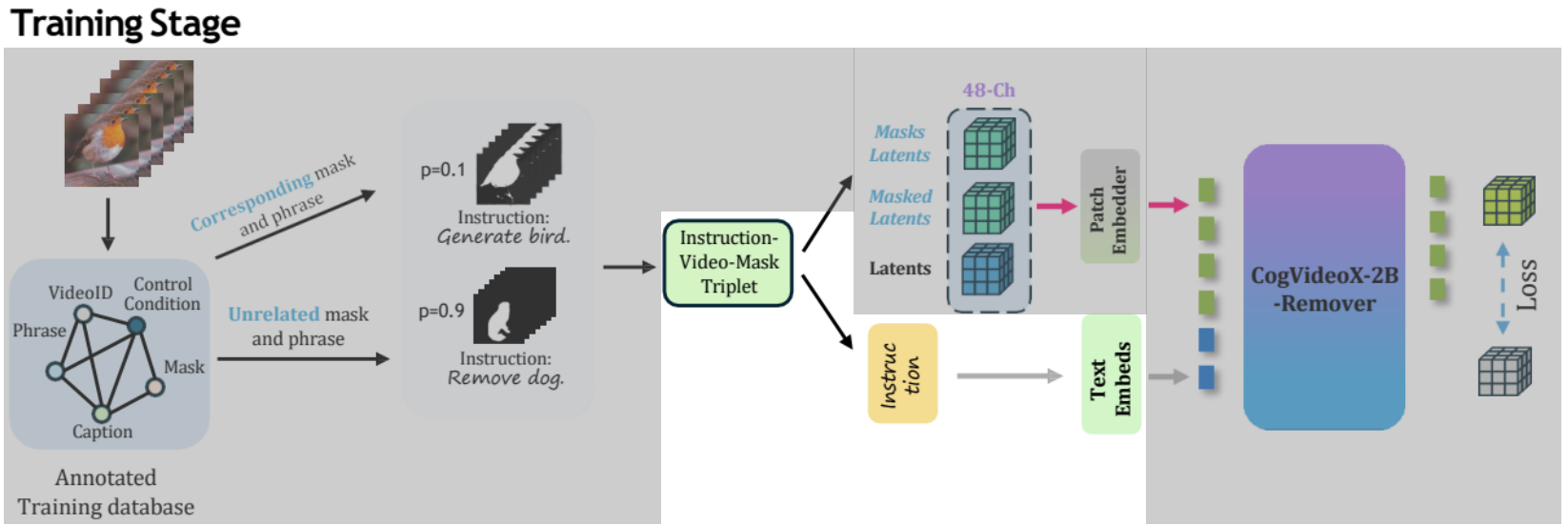
- Proposed Method
  - Object Remover Training Stage
    - Training
      - Instruction-Video-Mask Triplet은 CogVideox-2B-Remover 모델의 입력으로 변환됨
        - ✓Masks Latents: mask 자체의 latent feature
        - ✓Masked Latents: 원본 비디오에 mask가 적용된 부분의 latent feature
        - ✓Latents: 원본 비디오 프레임의 latent feature
        - ✓세 가지 입력은 채널 차원(총 48채널)을 따라 결합되어 Patch Embedder로 전달



CVSP Lab
Computer Vision & Signal Processing

Method

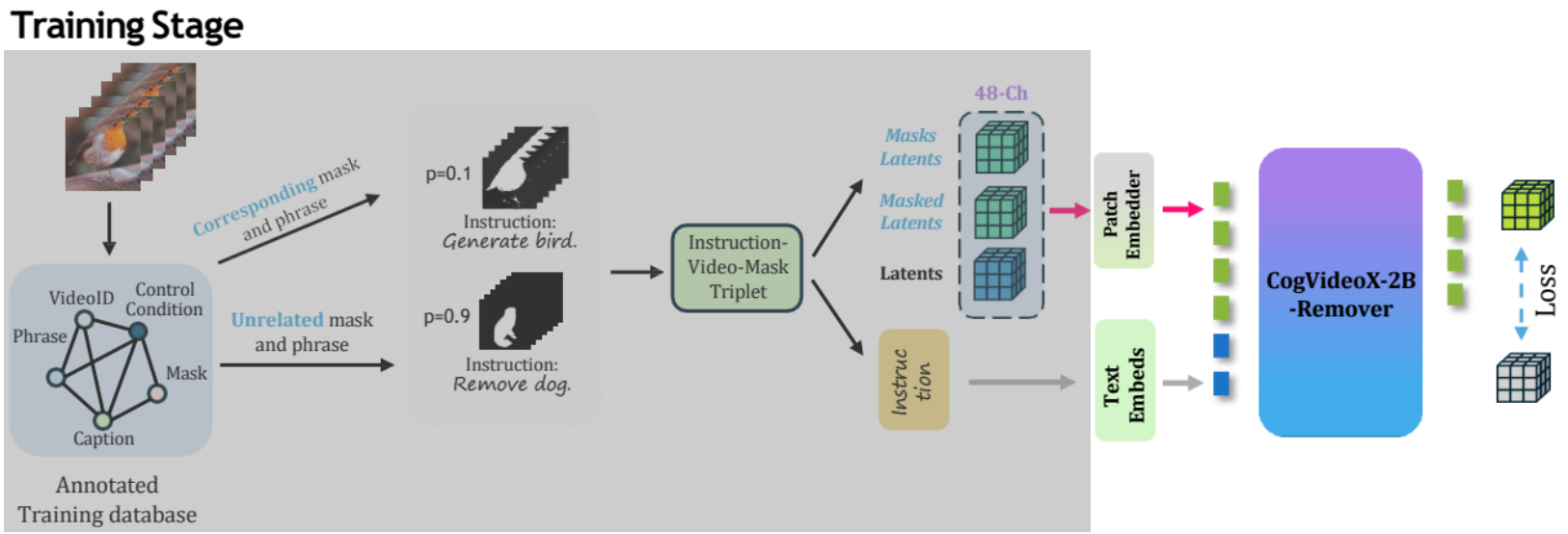
- Proposed Method
  - Object Remover Training Stage
    - Training
      - Instruction-Video-Mask Triplet은 CogVideox-2B-Remover 모델의 입력으로 변환됨
        - ✓Text instruction ("Generate bird.", "Remove dog." 등)은 Text Embedding으로 변환되어 모델에 입력됨



Method

* Hong, Wenyi, et al. "Cogvlm2: Visual language models for image and video understanding." arXiv preprint arXiv:2408.16500 (2024).
** Ravi, Nikhila, et al. "Sam 2: Segment anything in images and videos." arXiv preprint arXiv:2408.00714 (2024).

- Proposed Method
  - Object Remover Training Stage
    - Training
      - Model: CogVideox-2B-Remover
        - ✓CogVideox-2B 기반
        - ✓시각적 latent feature와 text embedding을 받아 edited video를 생성
      - Dataset: WebVid-10M
      - Caption: CogVLM2* 사용 / Mask: Grounded-SAM2**



CVSP Lab
Computer Vision & Signal Processing

Experiments

- Qualitative Results






CVSP Lab
Computer Vision & Signal Processing

* Lai, Wei-Sheng, et al. "Learning blind video temporal consistency." Proceedings of the European conference on computer vision (ECCV). 2018.

Experiments

- Quantitative Results
  - Metric
    - Ewarp* ↓
      - Warping 아티팩트 또는 변형 왜곡을 측정
    - Relevance ↓
      - 제거 instruction과 편집된 비디오 간의 CLIP 유사도를 측정
    - Temp-Cons ↑
      - 비디오 프레임 간의 시간적 일관성(Temporal Consistency)을 측정

Table 8. Quantitative Comparison of Object Removal. To assess the performance of object removal, we calculate the CLIP similarity between the removal instruction and the edited video, denoted as **relevance**. A lower relevance score indicates better removal performance. The best results are **blodfaced**.

Methods	Ewarp(10^{-3}) (↓)	Relevance (↓)	Temp-Cons (↑)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	MSE (↓)
Tokenflow	16.34	0.1597	0.9786	18.38	0.8395	0.1639	1095.06
Flatten	11.18	0.2194	0.9759	18.87	0.8367	0.1529	1088.33
InsV2V	6.67	0.2134	0.9747	22.27	0.9187	0.0648	563.17
AnyV2V	13.14	0.1774	0.9765	19.80	0.8825	0.1290	800.56
Propainter	4.93	0.1685	0.9862	**36.87**	**0.9978**	**0.0081**	**16.37**
Our Expert	**4.21**	**0.1554**	**0.9864**	29.16	0.9863	0.031	89.62

CVSP Lab
Computer Vision & Signal Processing

Experiments

- Quantitative Results
  - Metric
    - PSNR ↑
      - 원본 비디오와의 픽셀 수준 유사성을 나타내는 지표
    - SSIM ↑
      - 원본 비디오와의 구조적 유사성을 나타내는 지표
    - LPIPS ↓
      - 지각적 거리(Perceptual Distance)를 측정
    - MSE ↓
      - 평균 제곱 오차(Mean Squared Error)로 픽셀 값의 차이를 측정

Table 8. Quantitative Comparison of Object Removal. To assess the performance of object removal, we calculate the CLIP similarity between the removal instruction and the edited video, denoted as **relevance**. A lower relevance score indicates better removal performance. The best results are **blodfaced**.

Methods	Ewarp(10^{-3}) (↓)	Relevance (↓)	Temp-Cons (↑)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	MSE (↓)
Tokenflow	16.34	0.1597	0.9786	18.38	0.8395	0.1639	1095.06
Flatten	11.18	0.2194	0.9759	18.87	0.8367	0.1529	1088.33
InsV2V	6.67	0.2134	0.9747	22.27	0.9187	0.0648	563.17
AnyV2V	13.14	0.1774	0.9765	19.80	0.8825	0.1290	800.56
Propainter	4.93	0.1685	0.9862	**36.87**	**0.9978**	**0.0081**	**16.37**
Our Expert	**4.21**	**0.1554**	**0.9864**	29.16	0.9863	0.031	89.62

CVSP Lab
Computer Vision & Signal Processing