2D diffusion based NeRF to Transformer based NeRF

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
이혜빈

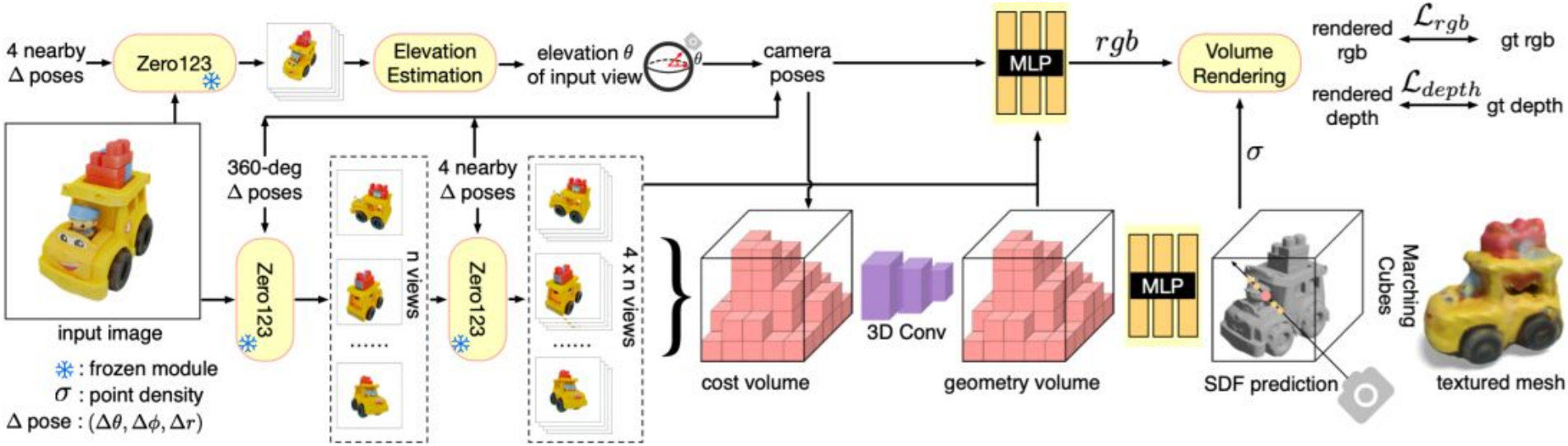
Outline

- Paper
  - One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization (NeurIPS 2023)
  - LRM: Large reconstruction model for single image to 3D (ICLR 2024)

One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization

Paper Review

- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Overview
    - 2D diffusion(각도별 이미지 생성) + 3D reconstruction module



Paper Review

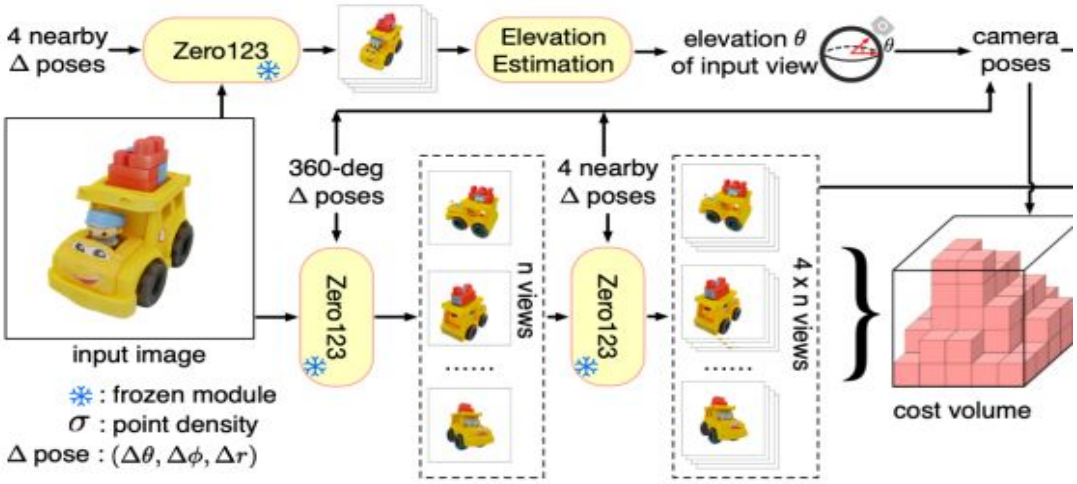
- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Intro
    - Single image to 3D reconstruction
    - 기존의 방법들
      - 2D diffusion model + 3D reconstruction module
      - Vision-language model + 3D reconstruction module
        - ✔ ex) DreamField, DreamFusion, Magic3D
    - 기존 방법들의 한계
      - optimization time 오래 걸림
      - 메모리 집약적
      - 3D 불일치 결과들(야누스 문제)
      - geometry 열악함

Paper Review

- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Intro
    - 논문이 제안하는 것
      - view-conditioned 2D diffusion model(Zero123) + 3D reconstruction module
        - ✔ multi-view 이미지들 생성 → 3D
        - ✔ 3D reconstruction module: SDF-based 생성 Neural Surface Reconstruction
      - 세 가지 모듈을 합성
        - ✔ multi-view 합성
        - ✔ elevation estimation
        - ✔ 3D reconstruction
      - 과도한 3D 인식 regularization, fine-tuning 없이 새로운 뷰에서 렌더링된 이미지, 실제 이미지 간 차이 최소화하며 학습

Paper Review

- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Method
    - Multi-views 생성



⁘ View-conditioned 2D diffusion model인 Zero123을 사용

✔ 2 stage 방법으로 Multi-view images 생성

⁘ Zero123의 input은 single image와 상대적인 camera transformation을 포함

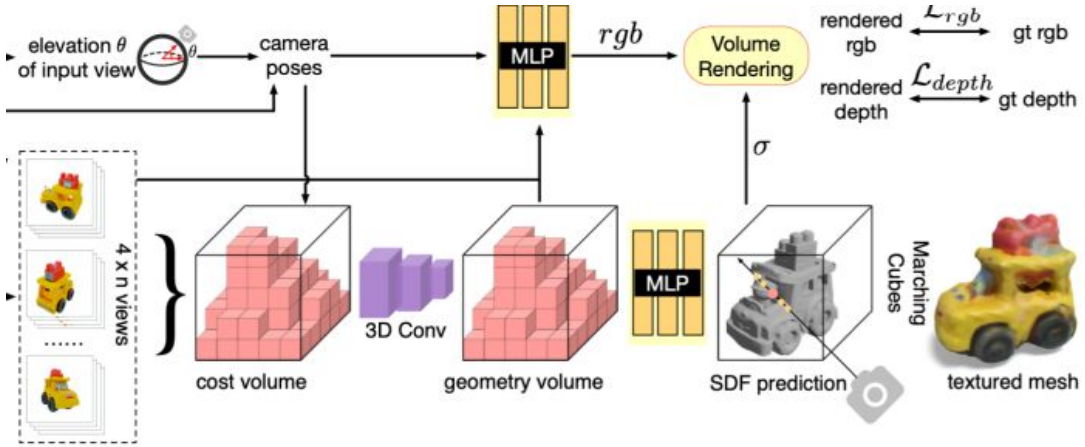
Paper Review

- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Method
    - Pose estimation

- Zero123에 의해 생성된 가까운 4개의 뷰들에 기초한 input 이미지의 elevation angle 측정
  - ✔ elevation angle: 해당 이미지에서 특정 지점이나 물체의 상대적인 높이를 나타내는 각도

Paper Review

- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Method
    - 3D reconstruction



multi-view posed images을 SDF 기반의 생성형 neural surface reconstruction module에 넣음

- ✔ SDF: 3D 공간에서 각 점까지의 부호 있는 거리 나타내는 함수
- ✔ 해당 방법은 SDF 함수를 사용하여 3D 객체의 표면 추정, 모델 재구성함

단일 feed forward pass로 3D mesh를 재구성하도록 학습 가능한 Cost Volume 기반 neural surface reconstruction 모듈 사용했음

Paper Review

- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Method
    - 3D reconstruction
      - Cost Volume 기반 neural surface reconstruction 모듈
        - ✔ Cost Volume

          서로 다른 시점에서 촬영된 두 개의 이미지 비교하여 각 위치에 대한 대응점 찾는 데 필요한 정보를 저장, 관리하는 데 사용되는 데이터 구조

          1. 각 이미지 픽셀 위치에서 대응되는 깊이 후보들에 대한 일치도 계산 (각 위치에 대한 모든 가능한 깊이 값에 대한 일치도 나타내는 볼륨)
          2. Cost volume은 depth map을 생성하는 데 사용될 수 있음. 가장 낮은 비용을 갖는 깊이 선택하여 가능함
        - ✔ Neural Surface Reconstruction

          Cost Volume 활용하여 딥러닝 모델 학습. CNN이 사용되며 이 모델은 Cost Volume을 입력으로 받아 3D 공간에서의 표면을 예측하는 데 사용됨

Paper Review

- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Method
    - 제안된 방법과 다른 시도: Zero123 + NeRF 나 SDF 기반 방법 실험(최적화 기반)

Zero123

- ✔ Input: object의 하나의 RGB image, relative camera transformation
- ✔ Output: diffusion model을 컨트롤해서 변형된 카메라 뷰 조건의 새로운 이미지를 생성

TensoRF

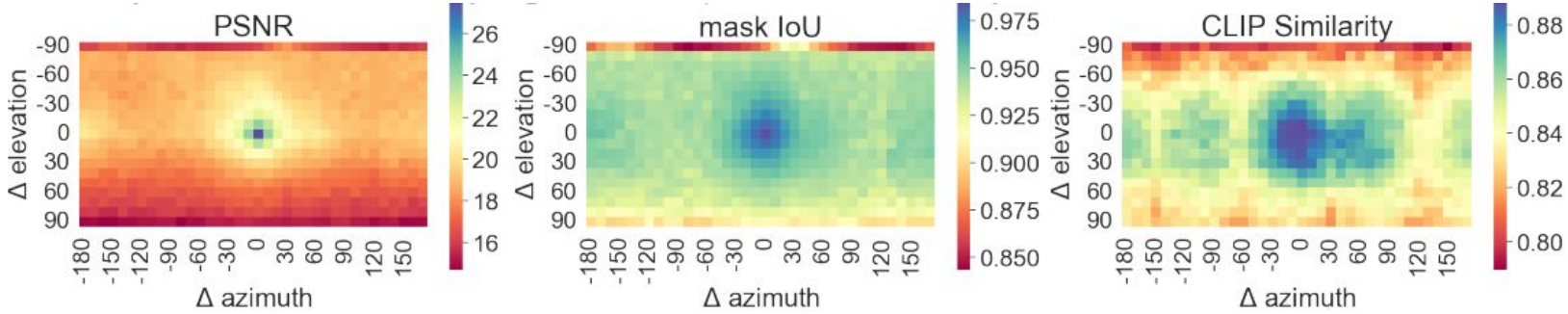
- ✔ single 이미지가 주어지면 3D로 reconstruction 해주는 NeRF 기반 모델

NeuS

- ✔ 특정 지점에서 표면까지의 거리와 방향을 포함하는 함수(SDF) 기반 3D reconstruction 모델

Paper Review

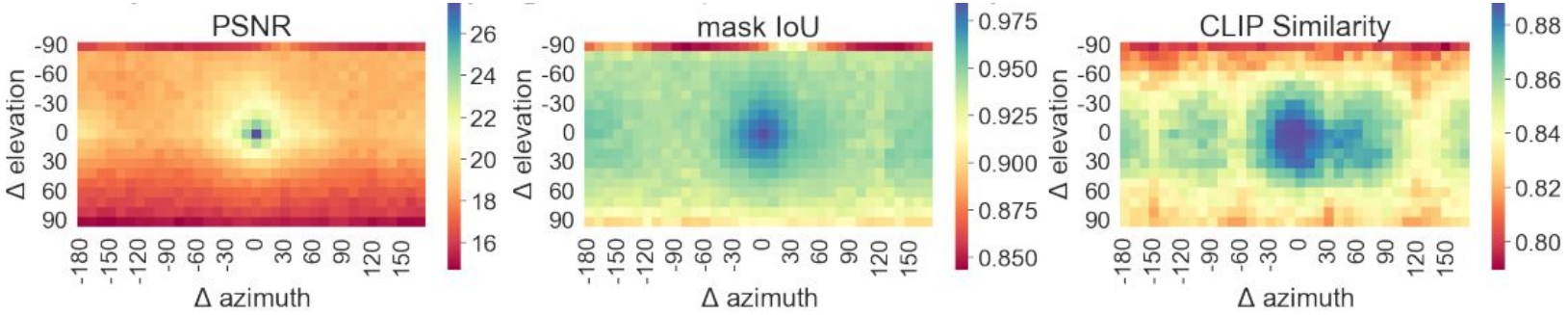
- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Method
    - 제안된 방법과 다른 시도: Zero123 + NeRF 나 SDF 기반 방법 실험(최적화 기반)



- View transformation마다 Objaverse dataset으로부터의 100개의 shapes의 PSNR, mask IoU, CLIP 유사도
- input relative pose가 크거나 target pose가 일반적이지 않은 위치들에 있을 때, 전체적인 PSNR이 높지 않음을 알 수 있음
  - ✔ PSNR이 높을수록 loss가 작고 퀄리티가 높음
  - ✔ Input view에서 각도가 커질수록 잘 안 됨

Paper Review

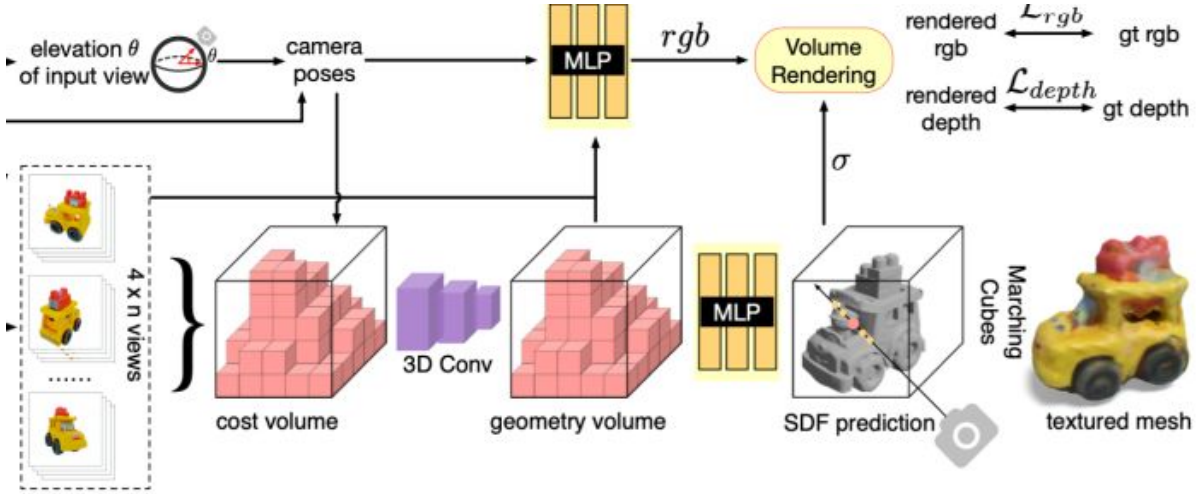
- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Method
    - 제안된 방법과 다른 시도: Zero123 + NeRF 나 SDF 기반 방법 실험(최적화 기반)



- 그러나 mask IoU와 CLIP 유사도는 상대적으로 좋음
  - ✔ Zero123가 gt와 지각적으로 유사하고 윤곽이나 경계가 유사한 예측을 생성
  - ✔ 그러나 픽셀 수준의 외관이 완전히 동일하지 않을 수 있음
- Source Views 간의 이러한 불일치는 이미 기존의 최적화 기반 방법에 치명적
  - ✔ 최적화에 시간이 많이 걸림

Paper Review

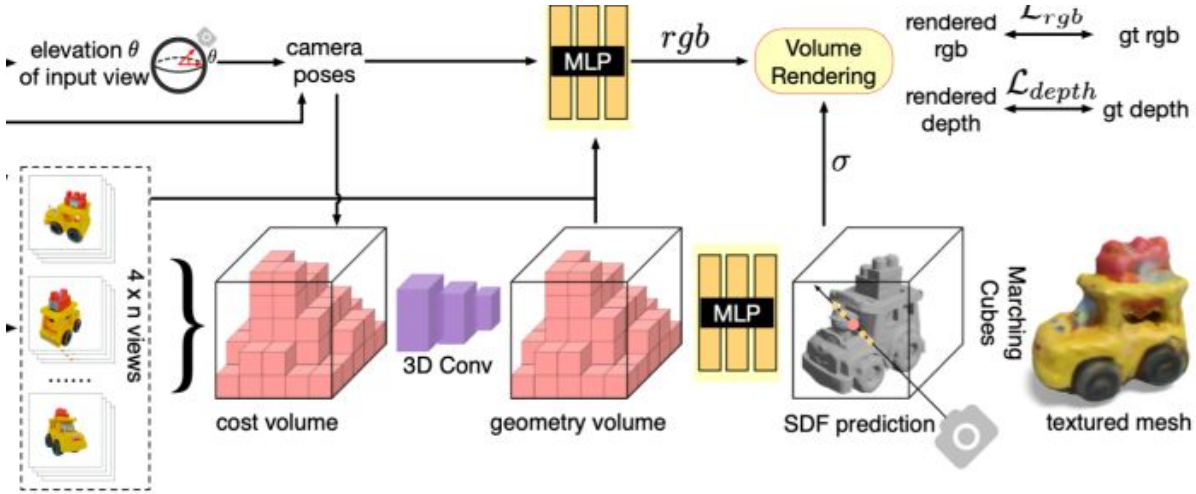
- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Method
    - 논문이 제안한 3D reconstruction 방식



- 일반화 가능한 SDF reconstruction 방법인 SparseNeuS를 기반
- Input: m개의 포즈 소스 이미지
- 2D feature 네트워크를 사용하여 m개의 2D feature map을 추출

Paper Review

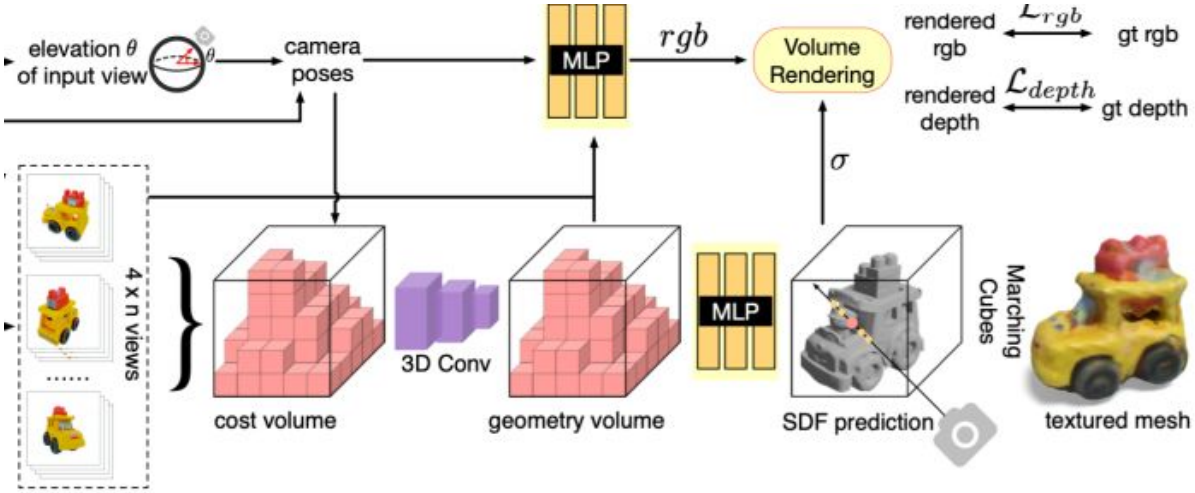
- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Method
    - 논문이 제안한 3D reconstruction 방식



- 각 3D voxel을 m개의 2D feature 평면에 project → m개의 project된 2D 위치에서 feature의 분산을 가져옴
- 콘텐츠가 계산되는 3D cost volume을 구축
- cost volume은 sparse 3D CNN을 사용하여 처리되어 입력 모양의 기본 geometry를 인코딩하는 geometry volume을 얻음

Paper Review

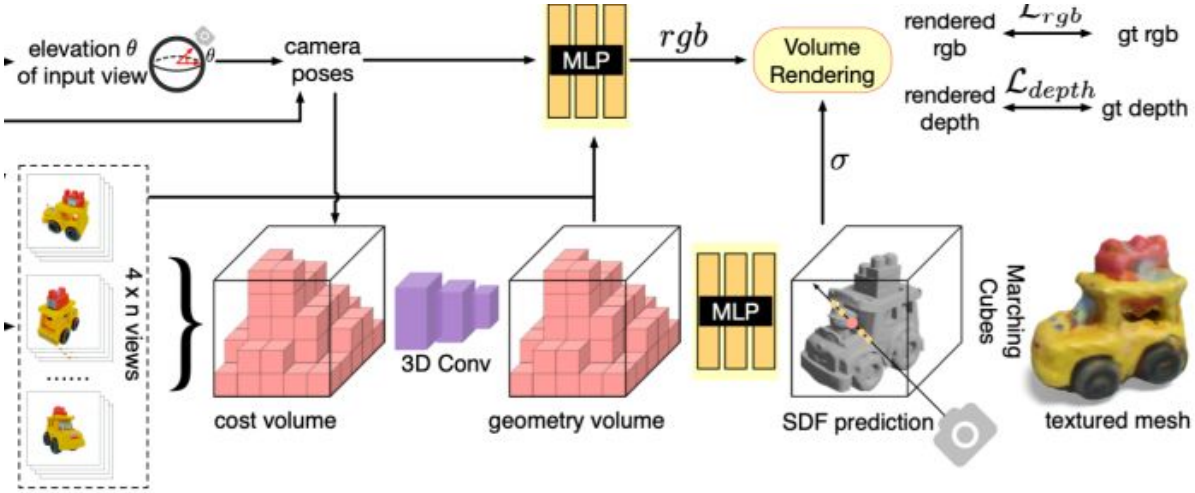
- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Method
    - 논문이 제안한 3D reconstruction 방식



- 임의 3D 포인트에서 SDF를 예측하기 위해 MLP 네트워크는 3D 좌표와 geometry volume에서 보간된 feature를 입력으로 사용
- 3D 포인트의 색상을 예측하기 위해 다른 MLP 네트워크는 project된 위치의 2D feature, geometry volume에서 보간된 feature, 소스 이미지의 뷰 방향을 기준으로 한 쿼리 광선의 뷰 방향을 입력으로 사용

Paper Review

- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Method
    - 논문이 제안한 3D reconstruction 방식



- 네트워크는 각 소스 뷰의 혼합 가중치를 예측하고, 3D 포인트의 색상은 project된 색상의 가중치 합으로 예측
- RGB 및 깊이 렌더링을 위해 두 개의 MLP 네트워크 위에 SDF 기반 렌더링 기술이 적용

Paper Review

- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Method
    - 2-Stage Source View Selection and Ground truth-Prediction Mixed Training
      - Zero123 을 freeze, 3D 객체 데이터셋에 대해 학습
      - 구형 카메라 모델 사용
      - n개의 카메라 포즈의 n개의 ground-truth RGB, depth 이미지 렌더링
      - Zero123으로 4개의 near view 예측
      - 학습 중 ground-truth pose가 포함된 $4 \times n$개의 예측 모두 reconstruction module에 input으로 넣음
      - n개의 ground-truth RGB 이미지 뷰 중 하나를 target view로 선택
      - 학습 중에 더 나은 supervision을 위해 첫 번째 단계에서 n개의 ground-truth 렌더링 사용 → depth loss 활성화
      - Inference 중에는 n개의 gt 렌더링을 zero123의 예측으로 대체 가능함
      - texture가 있는 mesh를 추출하려면 marching cube 사용→ SDF field에서 mesh 추출 → mesh 정점의 색상을 쿼리함
        - ✓ Marching cube: 3차원 공간 내 점에 대한 density → 3D mesh 생성

Paper Review

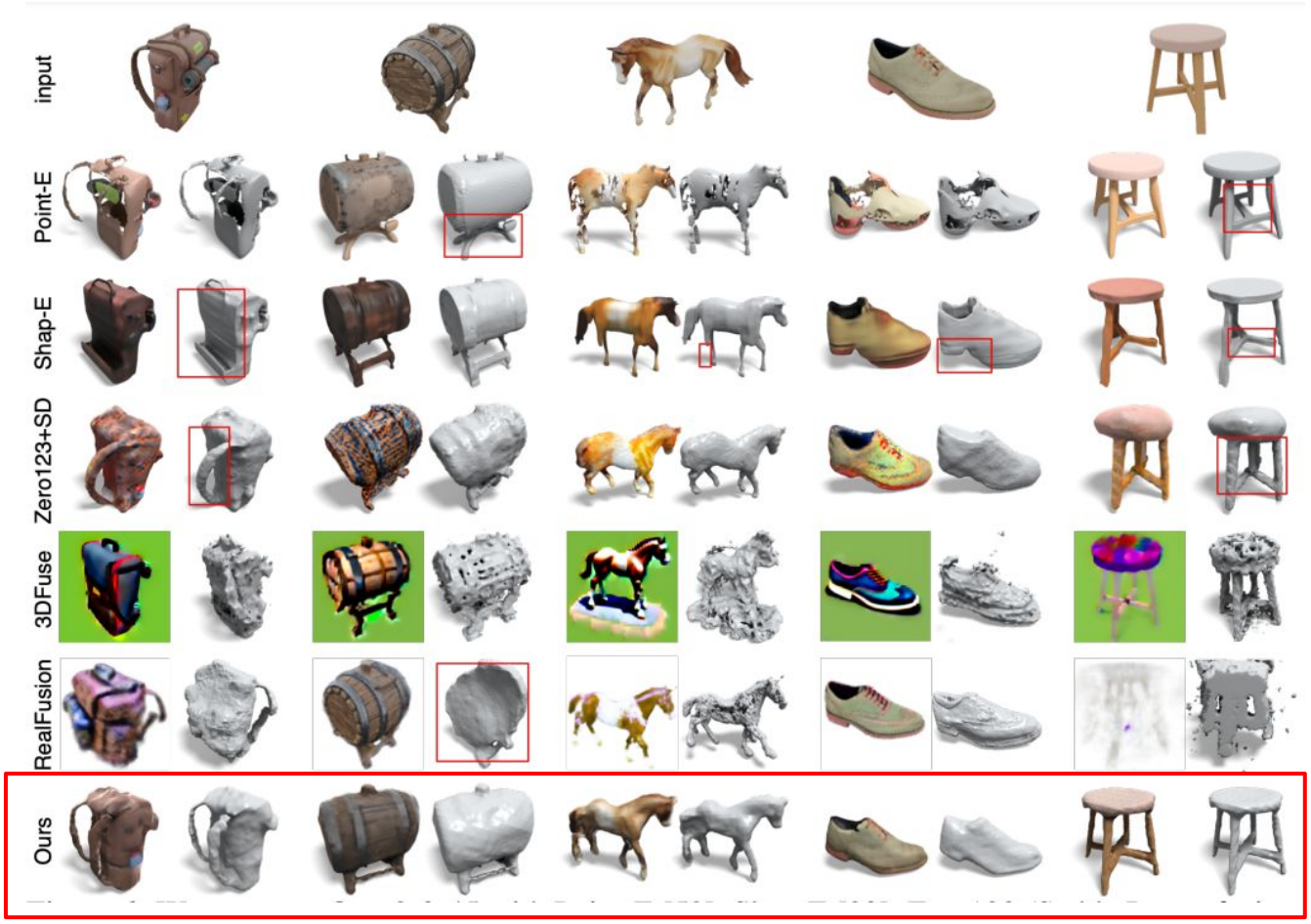
- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Method
    - Camera Pose Estimation
      - $4 \times n$ 개의 소스 이미지에 대한 카메라 포즈가 필요
      - 표준 구면 좌표계 (θ, ϕ, γ) 에서 카메라를 parameterize함
      - 카메라마다 구면 좌표를 가지고 있는데 적어도 하나의 카메라의 고도각이 절대값이어야 다른 값들을 알 수가 있음
      - 입력 이미지의 고도각을 추론하기 위한 고도각 추정 모듈을 제안 (elevation estimation) → 이미지, 카메라 포즈 간 reprojection loss 가장 적은 고도각 선택

Paper Review

- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Experiments
    - Dataset: Objaverse-LVIS
    - 3D models: 46k
    - 카테고리 수: 1,156
    - 방법:
      - Zero123 을 freeze하여 학습
      - 구의 표면에 위치한 camera poses를 선택함으로써 각각의 input image에 대해 8개의 이미지들을 생성함
      - 각 8개의 views에 대해 4개의 local images(10도 간격) 를 생성
    - gt RGB와 depth images를 render하기 위해 BlenderProc를 사용
      - BlenderProc: 사실적인 렌더링을 위한 프로시저 블렌더 파이프라인
    - Background 이미지들 때문에 SAM과 bounding-box prompts 를 사용
      - SAM: Segment Anything
        - ✔ Prompt Encoder + Image Encoder + lightweight mask Encoder 구조로 text로 이미지 Segment가 되게 하는 모델(https://arxiv.org/pdf/2304.02643v1.pdf)

Paper Review

- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Experiments: 다른 3D diffusion model 과의 정성적 비교



Paper Review

- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Experiments: 야누스 문제, 환각 문제 해결

  - 기존 diffusion model의 문제 원인
    - input 과 계속 가깝게 하려고 학습하다 보니 피규어 같은 경우 앞 뒤가 똑같이 구현되는 경우가 생김
      → 얼굴이 2개 → 야누스 문제
  - One-2-3-45
    - View Conditioned 2D diffusion model을 활용하여 본질적으로 3D 일관성을 향상시킴

Paper Review

- **One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization**
  - Experiments: F-Score

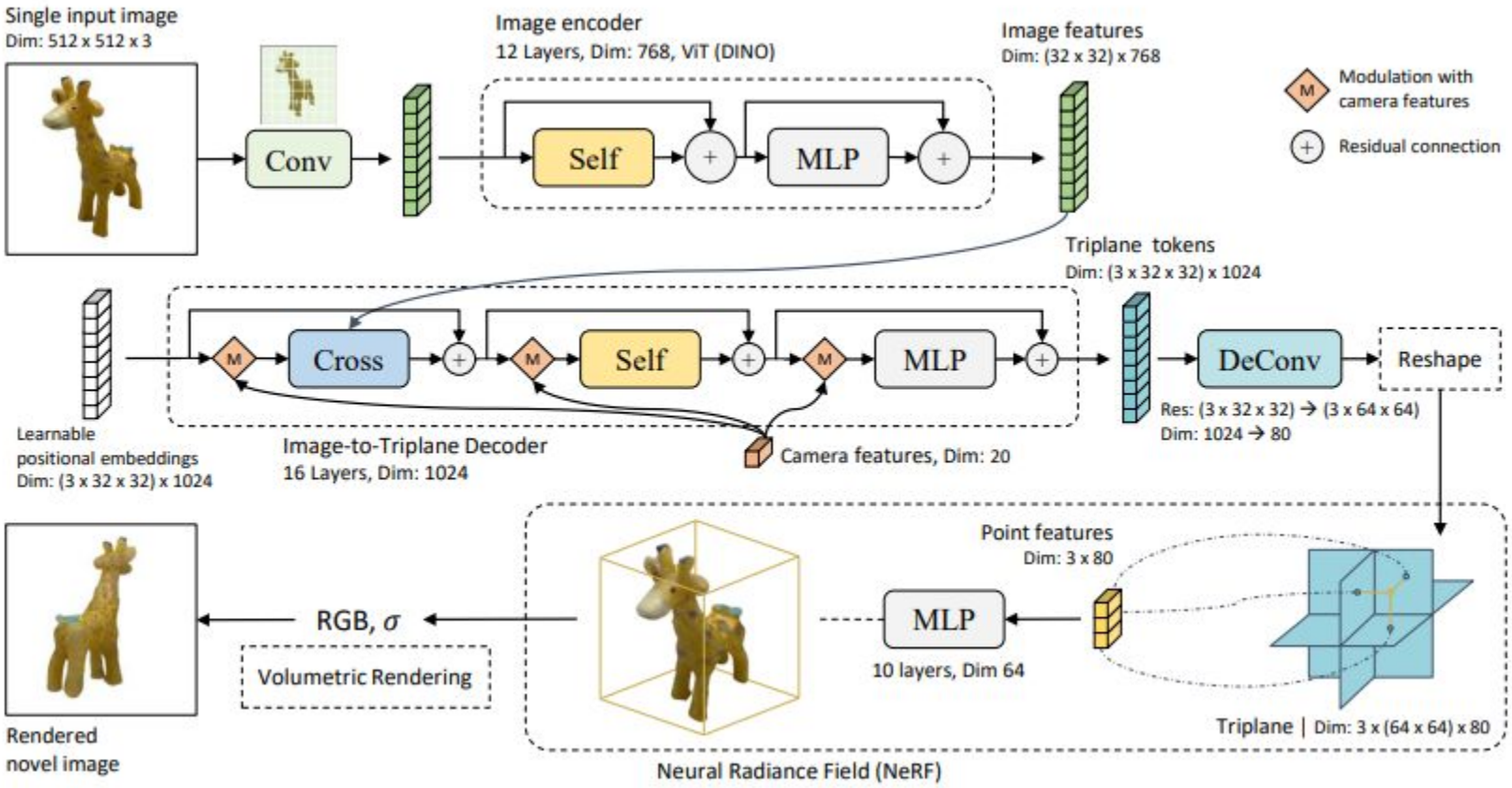
	Prior Source	F-Score GSO	F-Score Obj.	F-Score avg.	CLIP Similarity GSO	CLIP Similarity Obj.	CLIP Similarity avg.	Time
Point-E [52]	internal	81.0	81.0	81.0	74.3	78.5	76.4	78s
Shap-E [29]	3D data	83.4	81.2	82.3	**79.6**	**82.1**	**80.9**	27s
Zero123+SD [36]	2D diffusion models	75.1	69.9	72.5	71.0	72.7	71.9	∼15min
RealFusion [43]		66.7	59.3	63.0	69.3	69.5	69.4	∼90min
3DFuse [68]		60.7	60.2	60.4	71.4	74.0	72.7	∼30min
Ours		**84.0**	**83.1**	**83.5**	76.4	79.7	78.1	45s

  - F-Score:
    - 다른 모델들에 비해 가장 좋은 성능 보임
  - CLIP 유사도:
    - Geometry 유사도보다는 색상 유사도에 민감한 편이라 Shap-E에 비해 낮은 수치를 보임
  - Time:
    - One-2-3-45는 약 5초 만에 3D mesh를 재구성
    - 나머지 시간은 주로 A100 GPU에서 이미지당 약 1초가 걸리는 Zero123 예측에 사용됨

LRM: Large reconstruction model for single image to 3D

Paper Review

- **LRM: Large reconstruction model for single image to 3D**
  - Overview
    - Transformer + NeRF



Paper Review

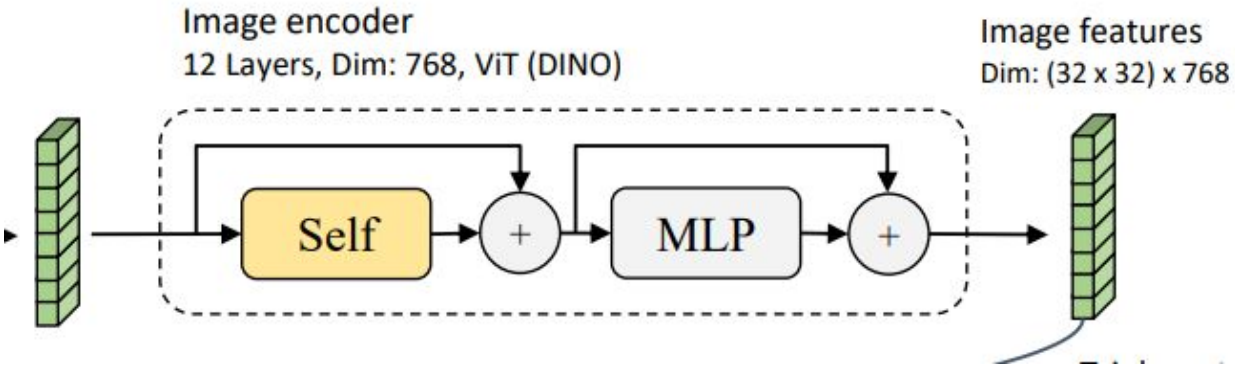
- **LRM: Large reconstruction model for single image to 3D**
  - Intro
    - 임의 개체의 single 이미지 → 3D 이미지
    - 기존의 방법들
      - 3D geometry의 ambiguity 로 인해 특정 class 범주 내에서 잘 수행되었음
      - 형태별 NeRF 최적화 → 일관된 geometry 구성
    - 기존 방법들의 한계
      - 비현실적임
      - pretrained model을 사용해왔음 → 새로운 type의 데이터의 경우 3D image 변환 불가능
      - 생성 모델, CLIP 같은 multi modal model에 의존

Paper Review

- **LRM: Large reconstruction model for single image to 3D**
  - Intro
    - 논문이 제안하는 것
      - Transformer Encoder-Decoder Architecture
      - Image Encoder: DINO
      - Decoder: Image-triplane Transformer Decoder
      - cross attention, self attention → 공간적으로 구조화된 3차원 토큰 간 관계 modeling
      - MLP 사용 → 각 점의 삼면 특성 디코딩 → 색상, 밀도 얻음 → 볼륨 렌더링 수행 → 이미지 렌더링
    - Data driven 방식, pretrained multi modal, diffusion model에 의존하지 않음
    - 과도한 3D 인식 regularization, fine-tuning 없이 새로운 뷰에서 렌더링된 이미지, 실제 이미지 간 차이 최소화하며 학습

Paper Review

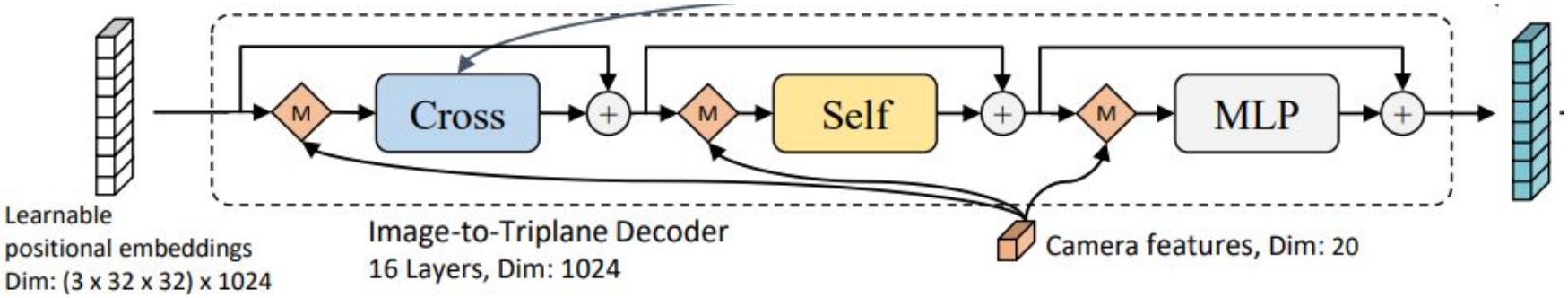
- **LRM: Large reconstruction model for single image to 3D**
  - Method
    - Encoder: DINO (ViT/B-16)



- input: RGB 이미지
- 패치별 feature token으로 인코딩
- 이미지에서 두드러진 구조, 질감에 대해 해석 가능한 attention을 학습하는 self-diffusion model DINO 사용
- DINO > ResNet, CLIP

Paper Review

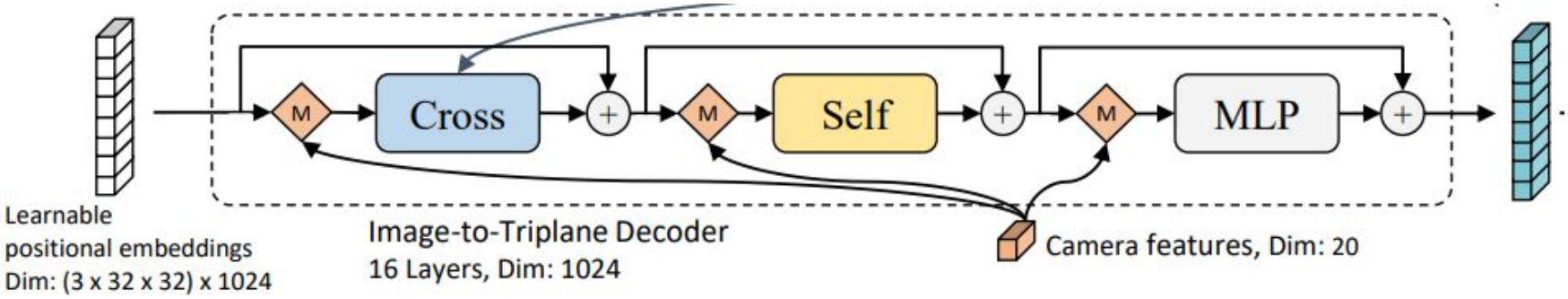
- **LRM: Large reconstruction model for single image to 3D**
  - Method
    - Decoder: Image-To-Triplane Decoder



- 2D image feature, camera feature → 학습 가능한 임베딩 space에 투영
- cross-attention: image features, camera features 간 관계 모델링
- Camera Features
  - ✔ 2D → 3D projection 위해서는 camera features를 이미지와 잘 매칭시켜야 함
  - ✔ intrinsic parameters, extrinsic parameters
  - ✔ 주요 layer 앞단에서 modulation 연산

Paper Review

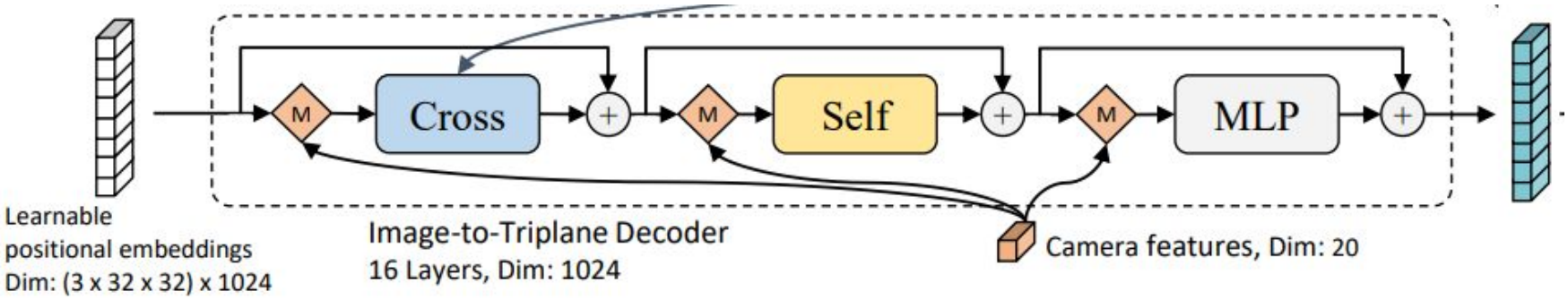
- **LRM: Large reconstruction model for single image to 3D**
  - Method
    - Decoder: Image-To-Triplane Decoder



- Triplane Representation
  - ✔ 3차원 Transformer Decoder에서 각 레이어마다 cross-attention, modulation 통해서 2d 이미지 feature를 3D 이미지 feature로 변환
  - ✔ image feature와 camera feature를 사용하는 이유
    - camera feature:카메라 전체 모양, 방향, 왜곡 제어
    - image feature: 색상 정보, 기하학적 정보
    - 두 정보를 통해 3D projection을 하기 위함

Paper Review

- **LRM: Large reconstruction model for single image to 3D**
  - Method
    - Decoder: Image-To-Triplane Decoder



⁘ Modulation with Camera features

✔ DiT로부터 영감을 받음

- DiT = Transformer + Diffusion + AdaLN

⁘ ModLN

$$\gamma, \beta = \mathrm{MLP}^{\mathrm{mod}}(\tilde{\boldsymbol{c}})$$
$$\mathrm{ModLN}_{\mathrm{c}}(\boldsymbol{f}_j) = \mathrm{LN}(\boldsymbol{f}_j) \cdot (1 + \gamma) + \beta$$

✔ camera features를 LayerNorm, MLP 연산을 통해 output 을 γ, β로 받음

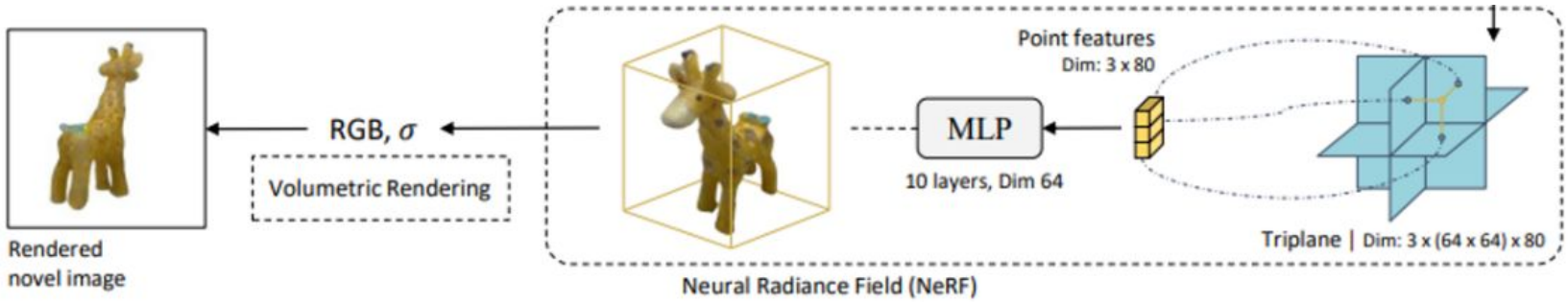
✔ r은 scale, b는 shift factor

Paper Review

- **LRM: Large reconstruction model for single image to 3D**
  - Method
    - Decoder: Image-To-Triplane Decoder
      - ModLN
        - ✔ AdaLN의 방식 + Modulation
        - ✔ AdaLN (Adaptive Layer Normalization)
          - DiT 같은 이미지 생성 모델에서 사용하는 Normalization 기법
          - 각 채널에 독립적으로 학습 가능한 scale, shift parameter 도입
          - 입력 데이터의 특성에 따라 동적으로 조절 가능
          - 생성된 이미지의 품질과 안정성 향상시킴

Paper Review

- **LRM: Large reconstruction model for single image to 3D**
  - Method
    - Decoder: Triplane-NeRF



3차원 representation T 에서 query된 point features로부터의 RGB, density를 예측

Paper Review

- **LRM: Large reconstruction model for single image to 3D**
  - Method
    - Training Objectives

$$\mathcal{L}_{\text{recon}}(\boldsymbol{x}) = \frac{1}{V}\sum_{v=1}^{V}\left(\mathcal{L}_{\text{MSE}}(\hat{\boldsymbol{x}}_v, \boldsymbol{x}_v^{GT}) + \lambda\mathcal{L}_{\text{LPIPS}}(\hat{\boldsymbol{x}}_v, \boldsymbol{x}_v^{GT})\right)$$

$\hat{x}_v$: rendering 된 뷰들

x_v^{GT} : input view들과 side view들

pixel 간의 L2 Loss + LPIPS Loss

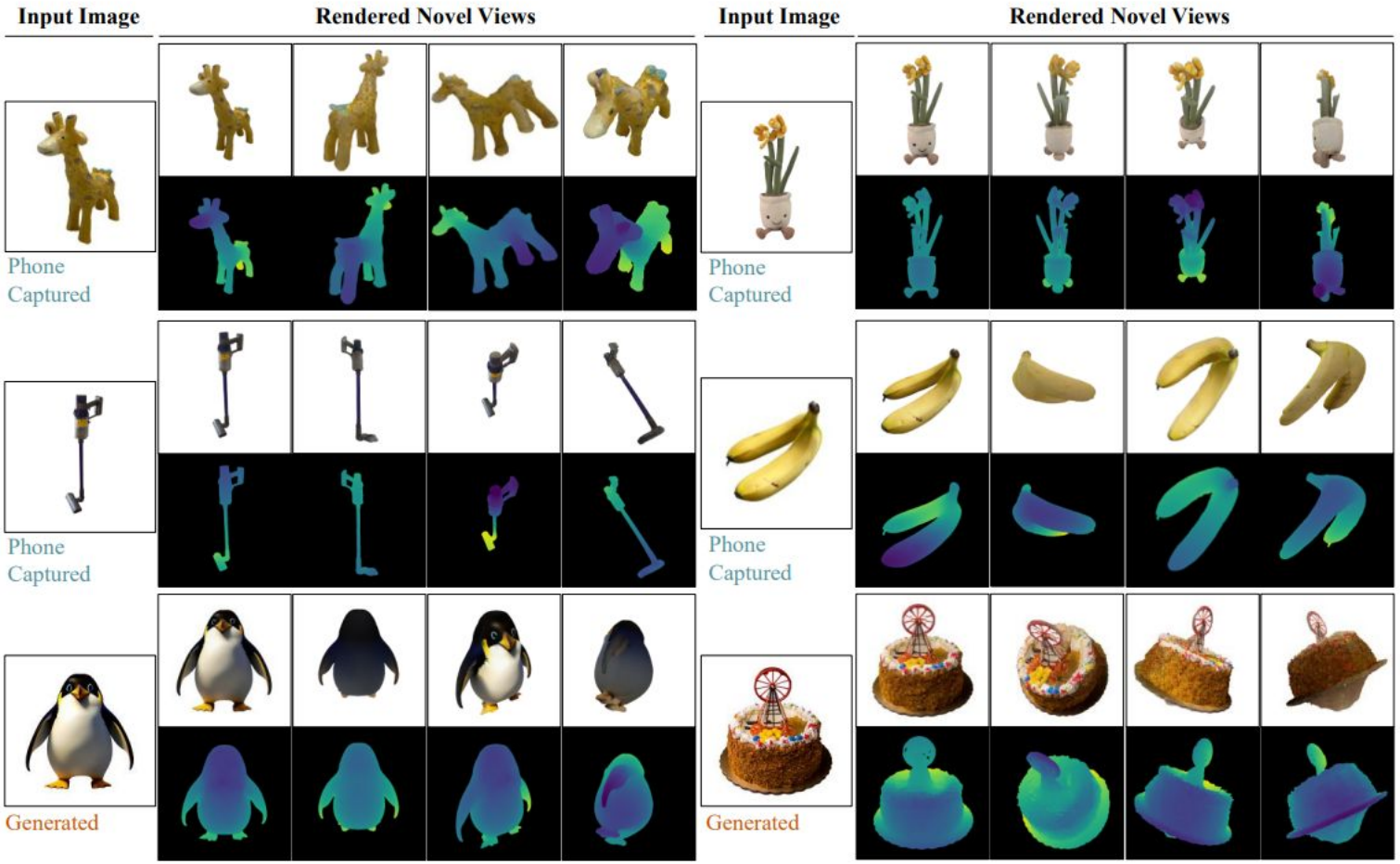
✔ LPIPS Loss: 인지된 이미지 패치 유사도에 대한 Metric

Paper Review

- **LRM: Large reconstruction model for single image to 3D**
  - Experiments
    - Training Objectives
      - Dataset: Objaverse, MVImgNet (video)
      - GPU: batch 1024, 128개의 A100
      - epoch: 30
      - 학습시간: 3일
      - optimizer: AdamW
      - cosine scheduler 사용
      - learning rate: 4×10^{-4}
    - Inference
      - GPU: A100
      - Camera feature: Objaverse 로 학습할 때의 feature
      - input single image → mesh (5초 미만)

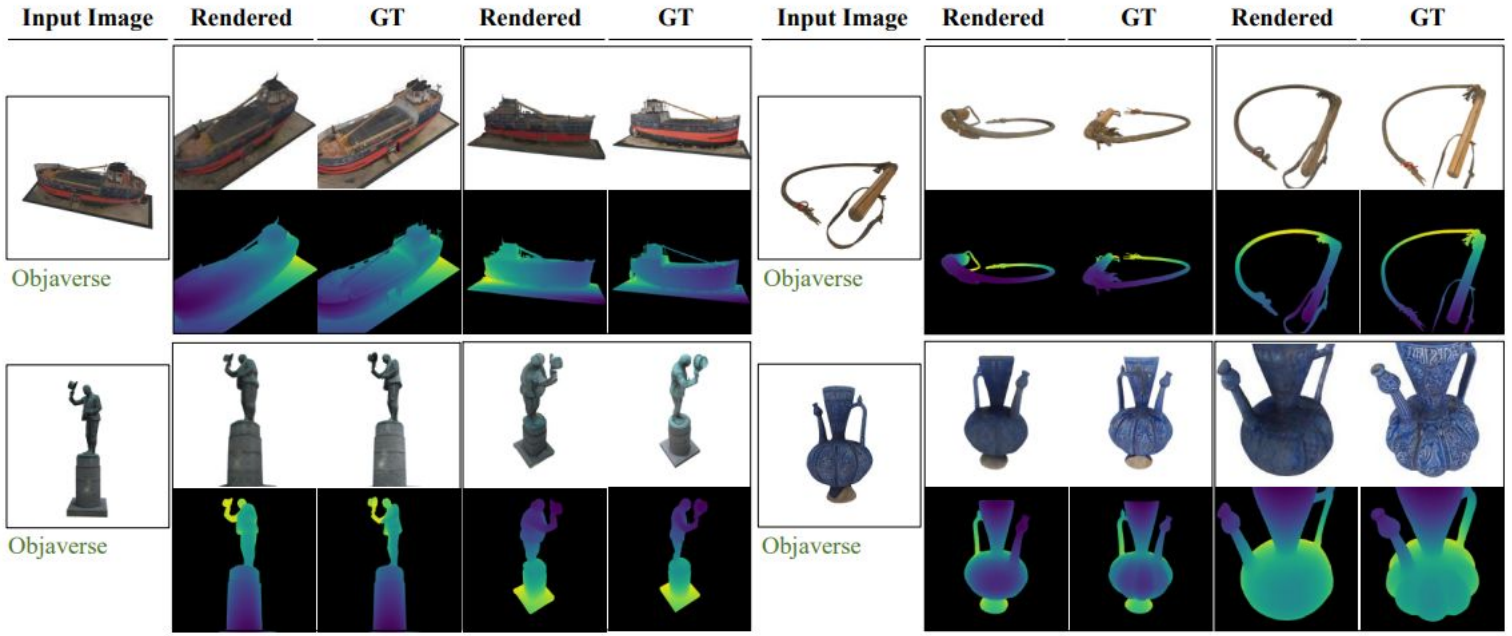
Paper Review

- **LRM: Large reconstruction model for single image to 3D**
  - Experiments



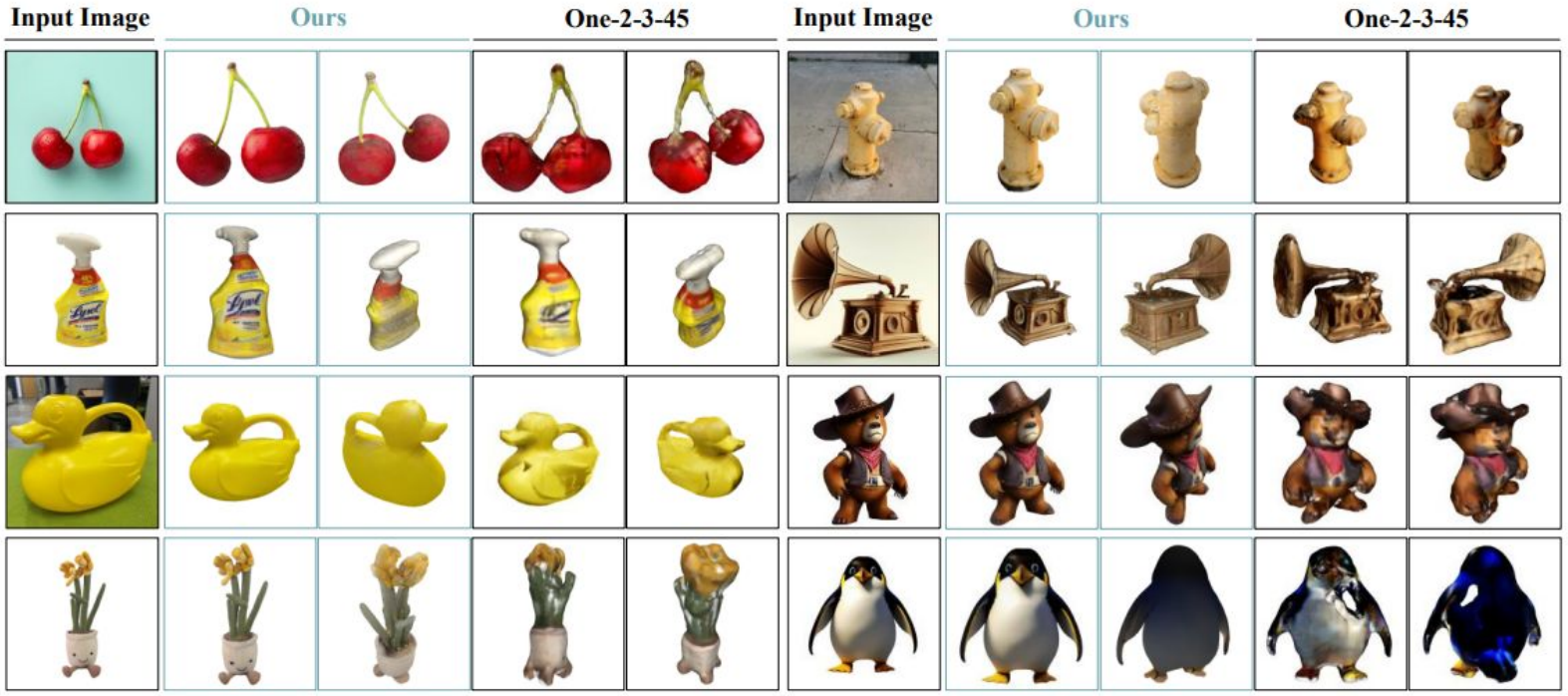
Paper Review

- **LRM: Large reconstruction model for single image to 3D**
  - Experiments



Paper Review

- **LRM: Large reconstruction model for single image to 3D**
  - Experiments

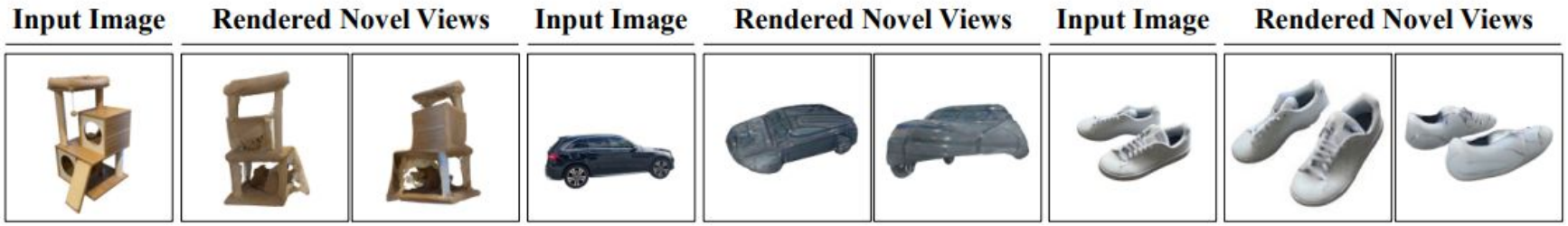


- One-2-3-45와의 비교

  ∴LRM이 훨씬 더 선명한 디테일, 일관된 표면 생성함

Paper Review

- **LRM: Large reconstruction model for single image to 3D**
  - Limitation



- 폐색된 영역에 대해 흐릿한 텍스처 생성하는 경향
  - 원인: single image → 3D 자체가 확률론으로의 접근
    - ✔ 평균에 가까운 영역들만 잘 표현하고 그렇지 않을 경우 생성이 잘 안 됨
- 추론 시간 동안 고정된 camera intrinsic, extrnsic parameter set을 테스트 이미지에 할당함
  - 카메라 매개변수가 실제와 일치가 잘 되지 않게 됨 → 왜곡된 형태 재구성
- 배경이 없는 사물의 이미지만을 다룸
- 빛나는 금속, 세라믹 같은 뷰 의존적 외관 충실하게 재구성 못함
  - NeRF의 뷰 종속 모델링을 생략함 (기존에 NeRF는 이걸 잘 해서 정반사되는 사물들도 잘 표현했었음)

Paper Review

- **LRM: Large reconstruction model for single image to 3D**
  - Limitation
    - 개인적으로 생각했던 추가적인 문제점:
      - input image와 camera feature modulation
        - ✔ input 이미지, 카메라 feature 간 연산으로 구도가 바뀌었을 때의 이미지는 잘 rendering됨
        - ✔ 그러나 광원의 위치에 따른 색상 변화가 학습되지 않아서 input image의 색상과 거의 똑같이 이미지 색상이 결정됨
          - 그로 인해 자세한 feature들이 잘 반영되지 않음

Reference

- One-2-3-45 논문 리뷰
  - https://kimjy99.github.io/%EB%85%BC%EB%AC%B8%EB%A6%AC%EB%B7%B0/one-2-3-45/
- One-2-3-45 논문
  - https://arxiv.org/pdf/2306.16928v1.pdf
- LRM 논문
  - https://arxiv.org/pdf/2311.04400.pdf
- Cost Volume 관련
  - https://velog.io/@chunjakim/Cost-Volume
- Marching Cube 관련
  - https://xoft.tistory.com/47

감사합니다😀