Large Reconstruction Model (LRM)

Introduction

Task: X to 3D

- Image-to-3D

- Text-to-3D

- Multi-view-to-3D, Video-to-3D, …

Results from [Xu & Shi et al. '24]

What is a Large Reconstruction Model?

Feed-Forward Network
trained on large data to generalize

Input:
single image,
text, …

Output:
Radiance field
representation,
e.g., NeRF, 3DGS

Large Reconstruction Model (LRM)

PixelNeRF

Input images

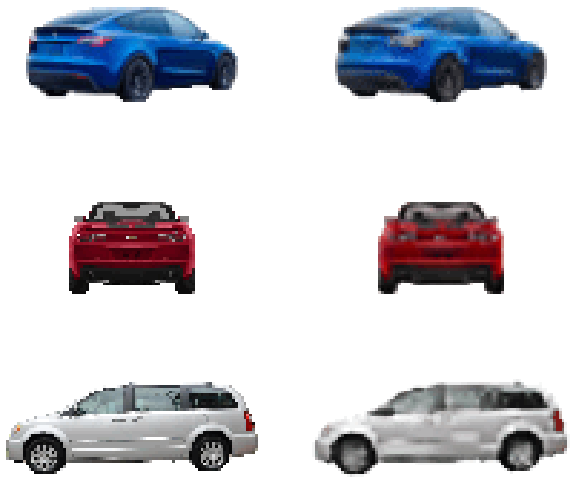
PixelNeRF NeRF

[Yu et al. '21] pixelNeRF: Neural Radiance Fields from One or Few Images

PixelNeRF

Input images

PixelNeRF

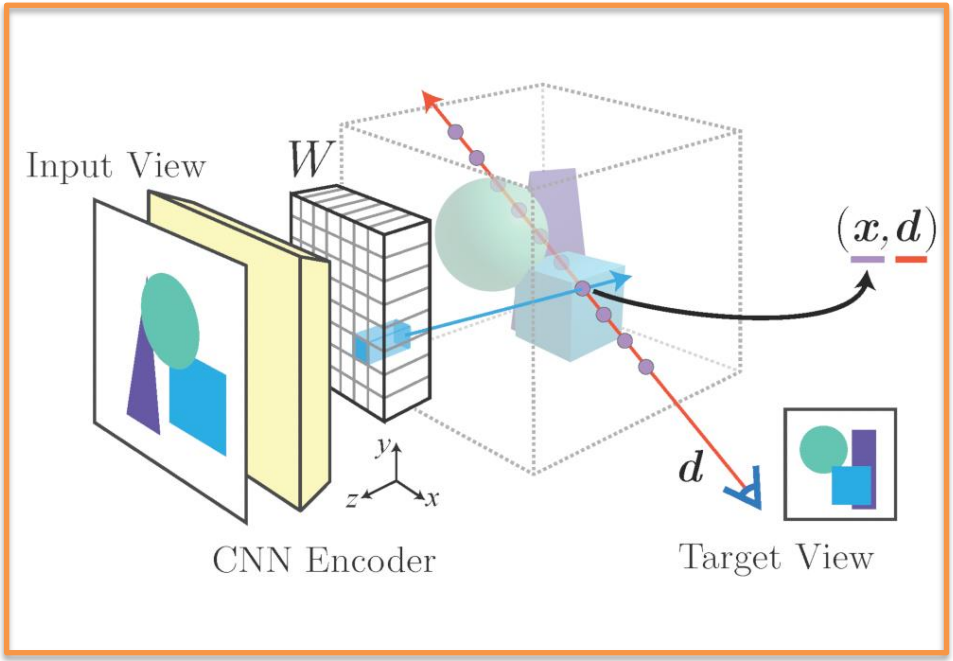


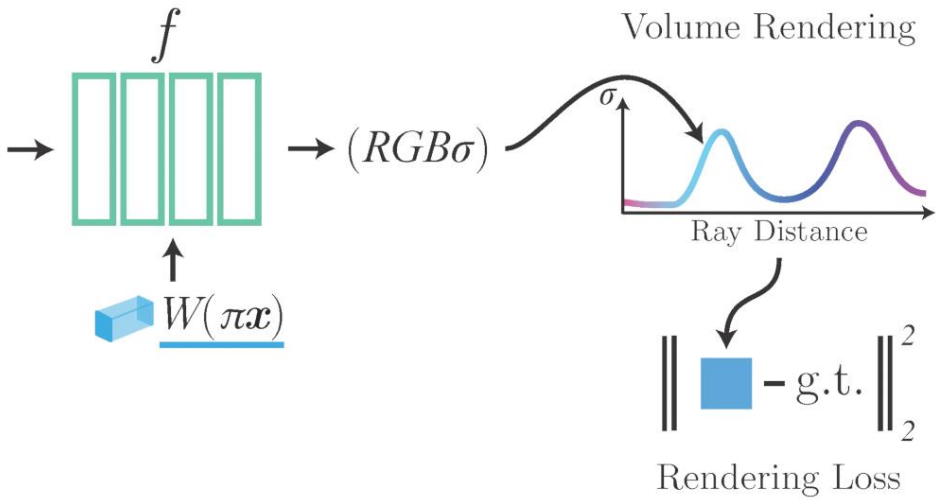
Input image

PixelNeRF

[Yu et al. '21] pixelNeRF: Neural Radiance Fields from One or Few Images

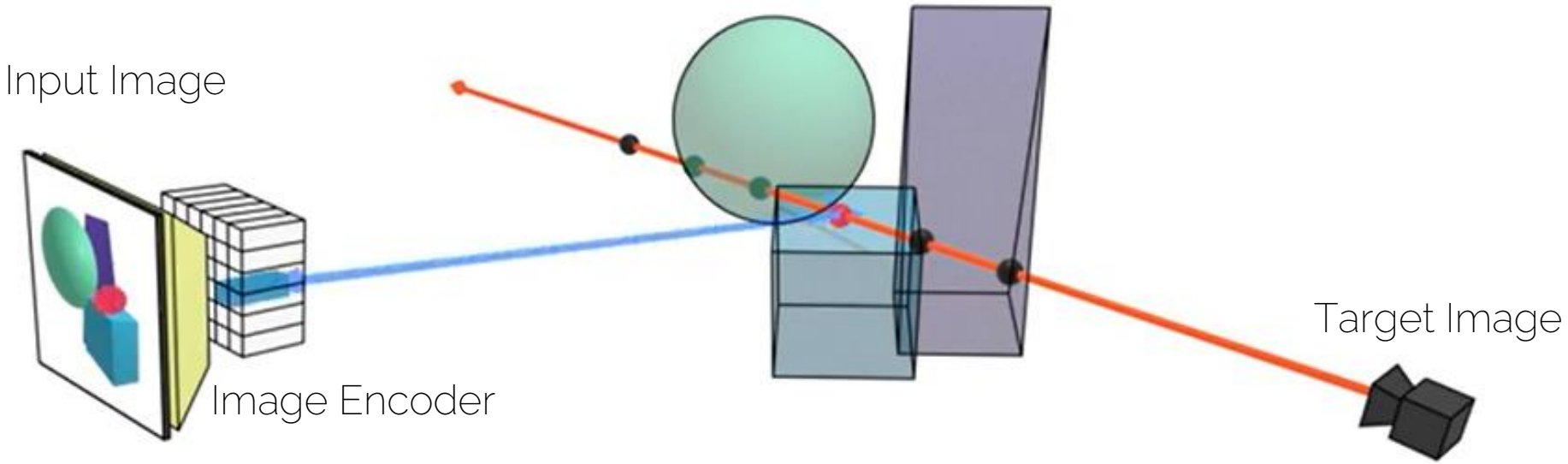
PixelNeRF – Single Image Input





[Yu et al. '21] pixelNeRF: Neural Radiance Fields from One or Few Images

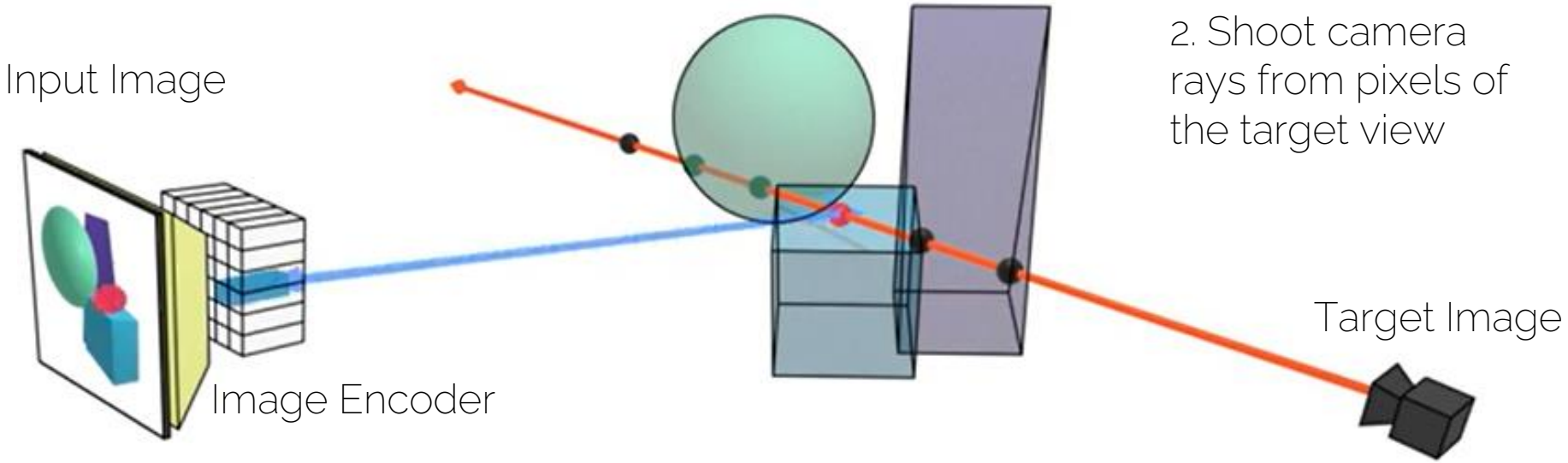
PixelNeRF – Single Image Input



1. Extract features from input image with CNN encoder

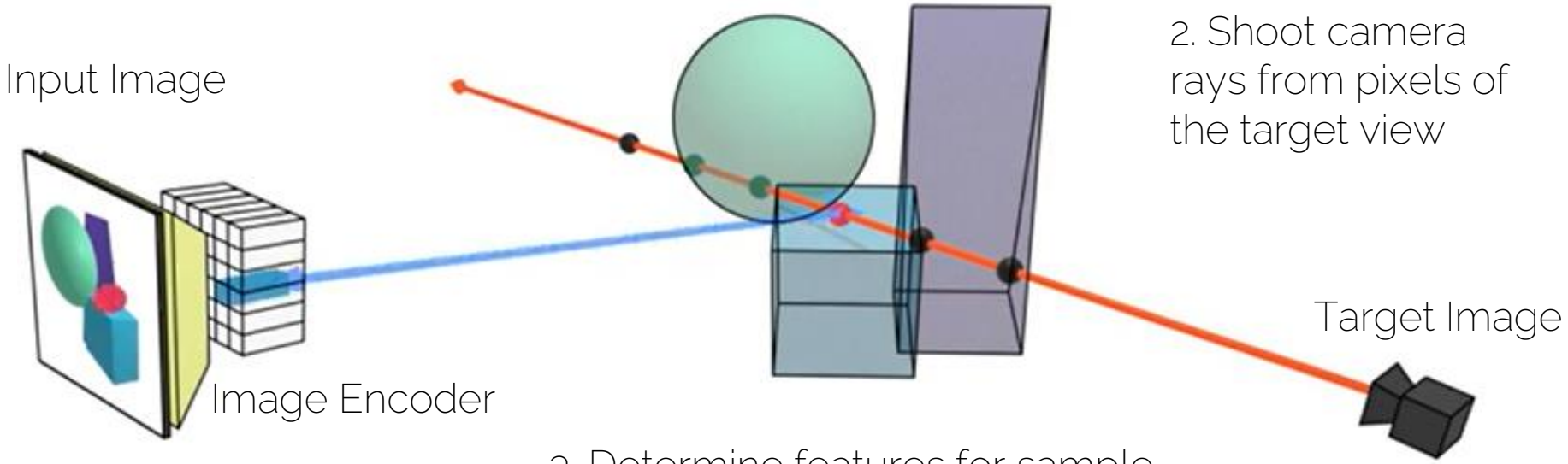
[Yu et al. '21] pixelNeRF: Neural Radiance Fields from One or Few Images

PixelNeRF – Single Image Input



[Yu et al. '21] pixelNeRF: Neural Radiance Fields from One or Few Images

PixelNeRF – Single Image Input

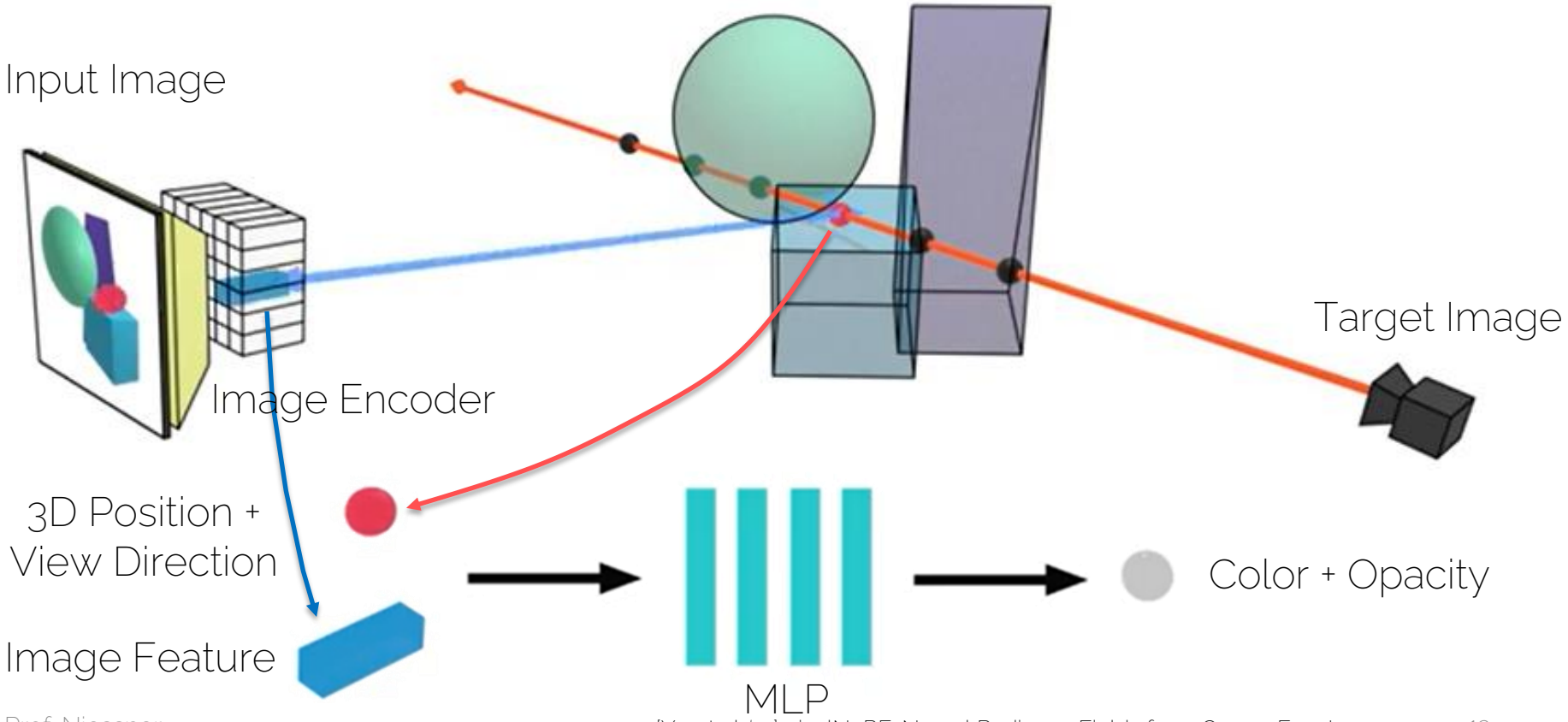


1. Extract features from input image with CNN encoder

3. Determine features for sample points along the rays:
- Project 3D points to image plane.
- Bilinearly interpolate image features

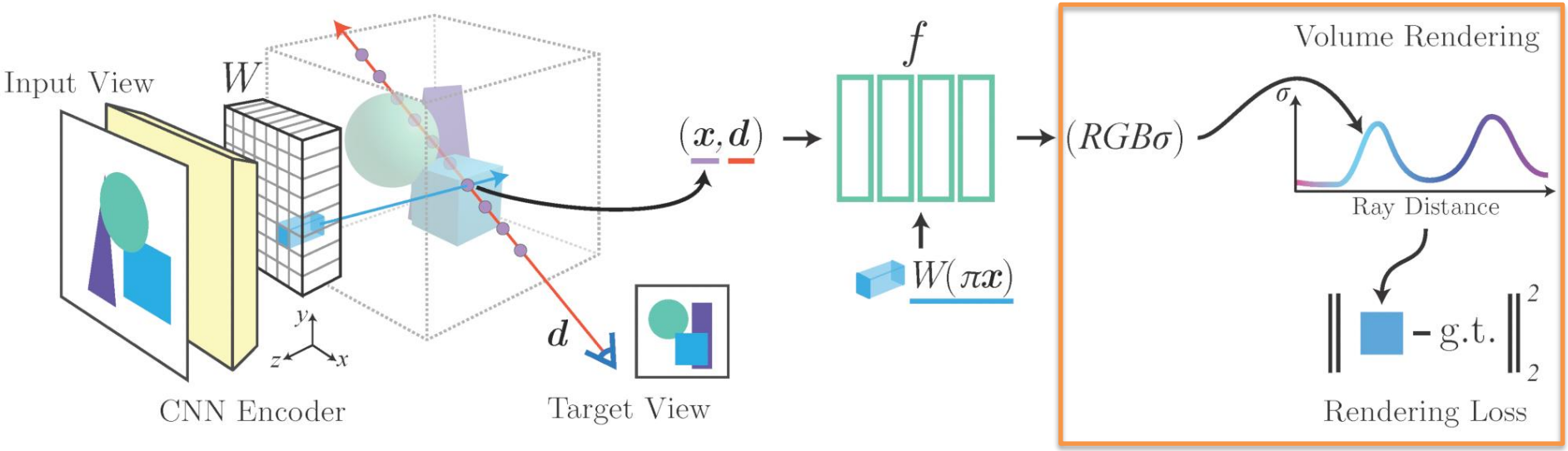
[Yu et al. '21] pixelNeRF: Neural Radiance Fields from One or Few Images

PixelNeRF – Single Image Input



[Yu et al. '21] pixelNeRF: Neural Radiance Fields from One or Few Images

PixelNeRF – Single Image Input

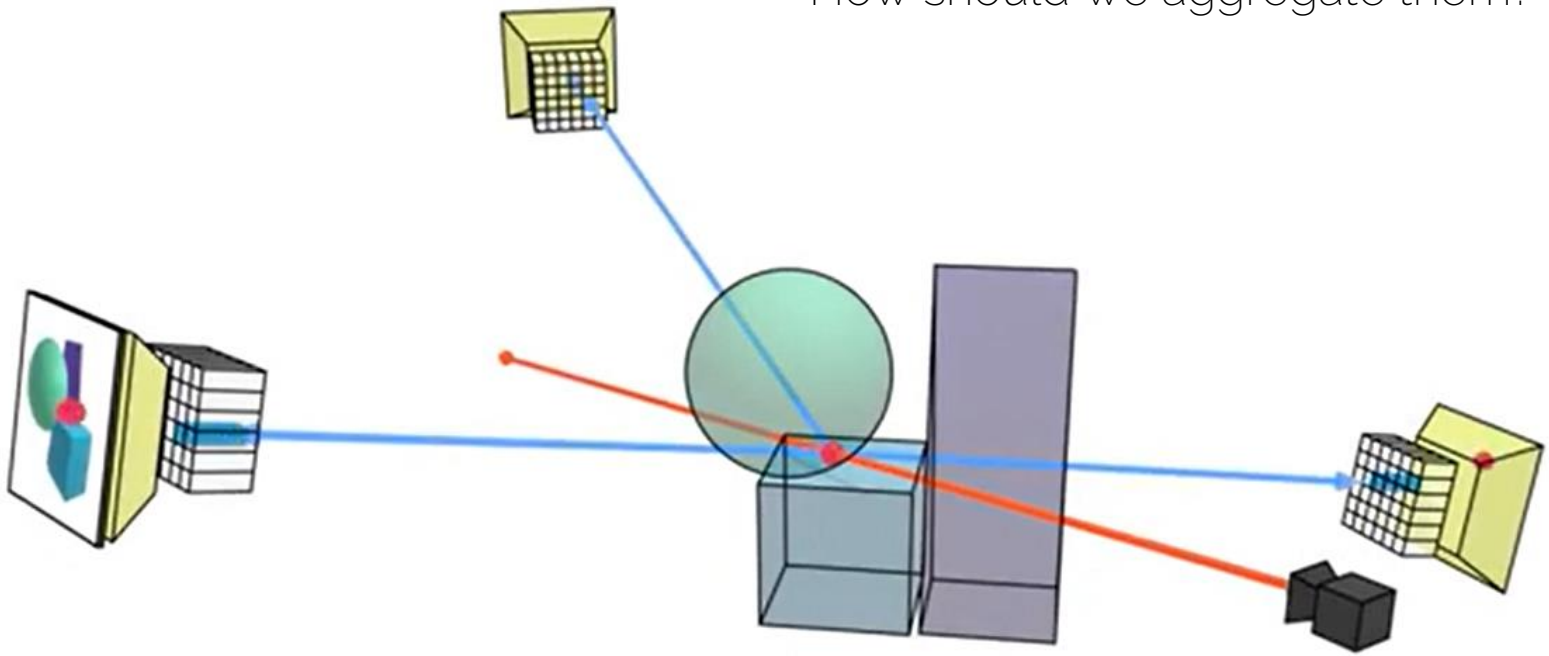


→ Volume rendering and supervision as in NeRF.

[Yu et al. '21] pixelNeRF: Neural Radiance Fields from One or Few Images

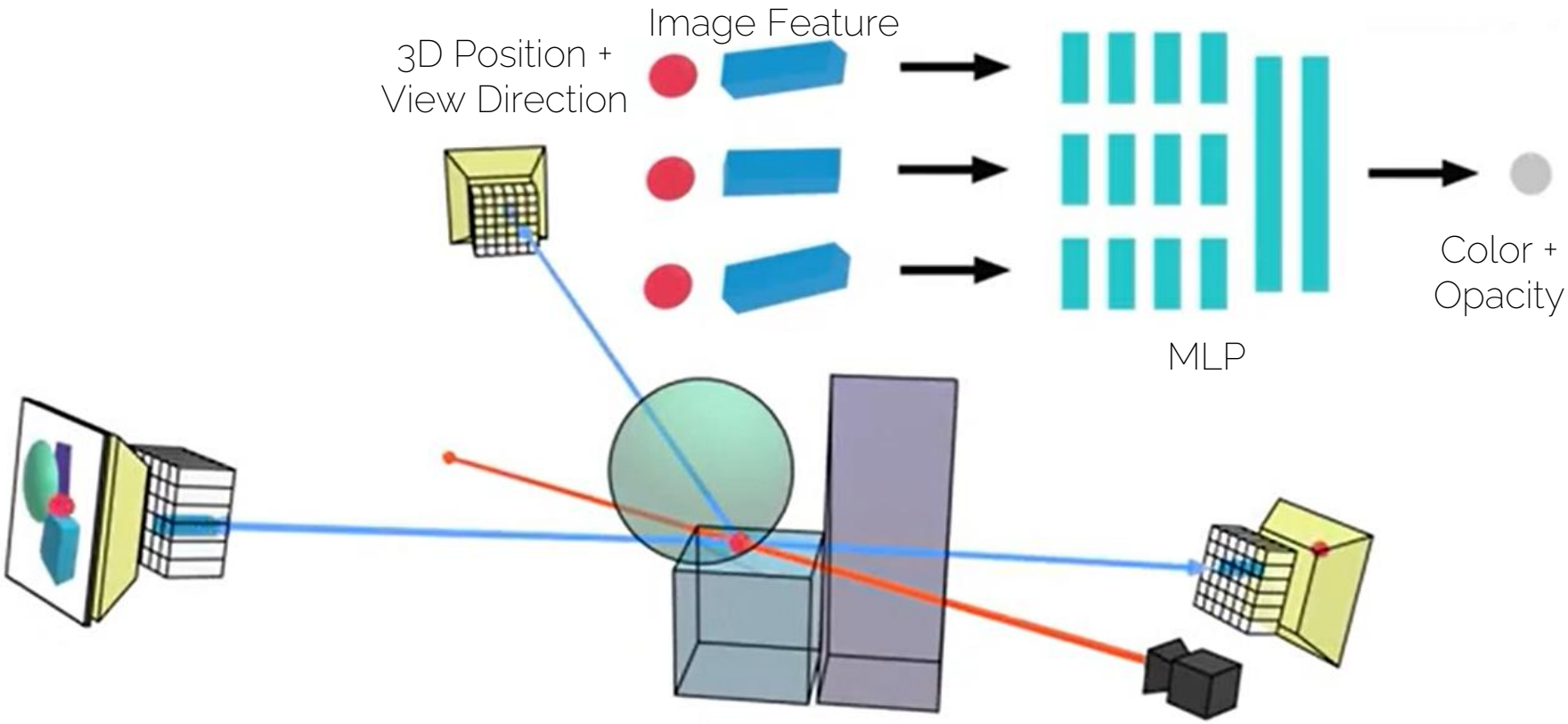
PixelNeRF – Multi-View Input

→ There are multiple image feature vectors. How should we aggregate them?



[Yu et al. '21] pixelNeRF: Neural Radiance Fields from One or Few Images

PixelNeRF – Multi-View Input



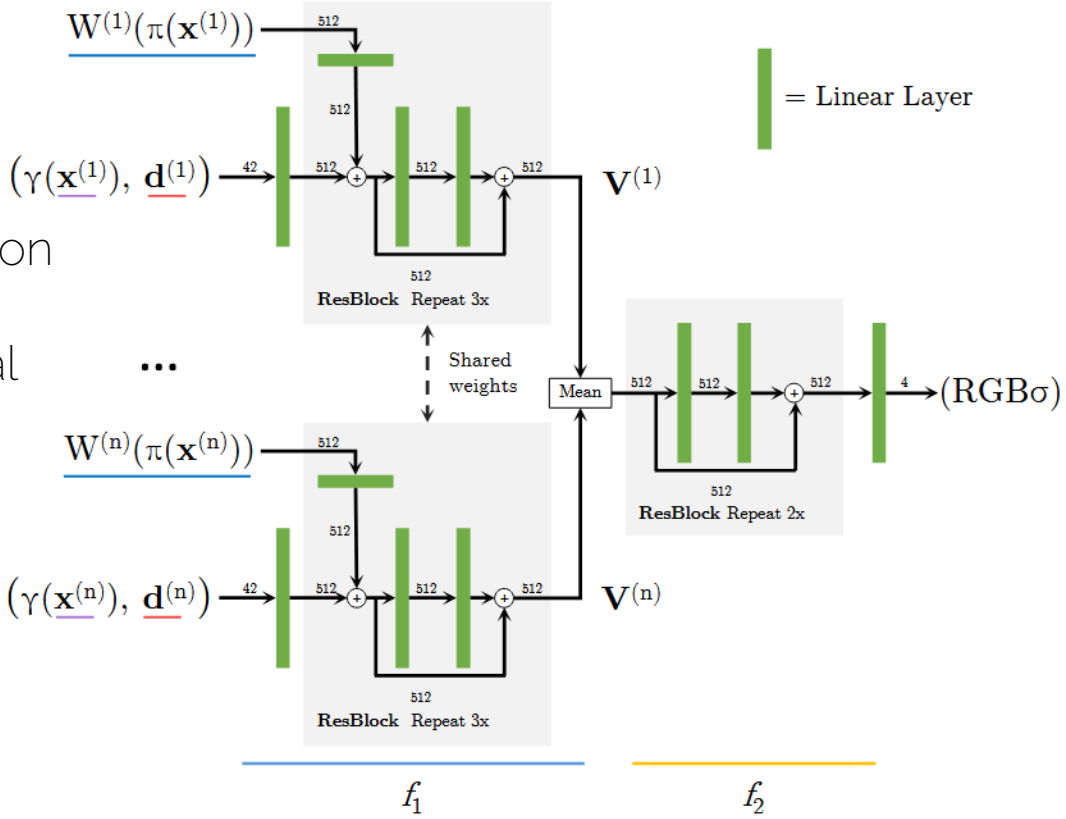
[Yu et al. '21] pixelNeRF: Neural Radiance Fields from One or Few Images

PixelNeRF – Architecture

Fully-connected ResNet architecture

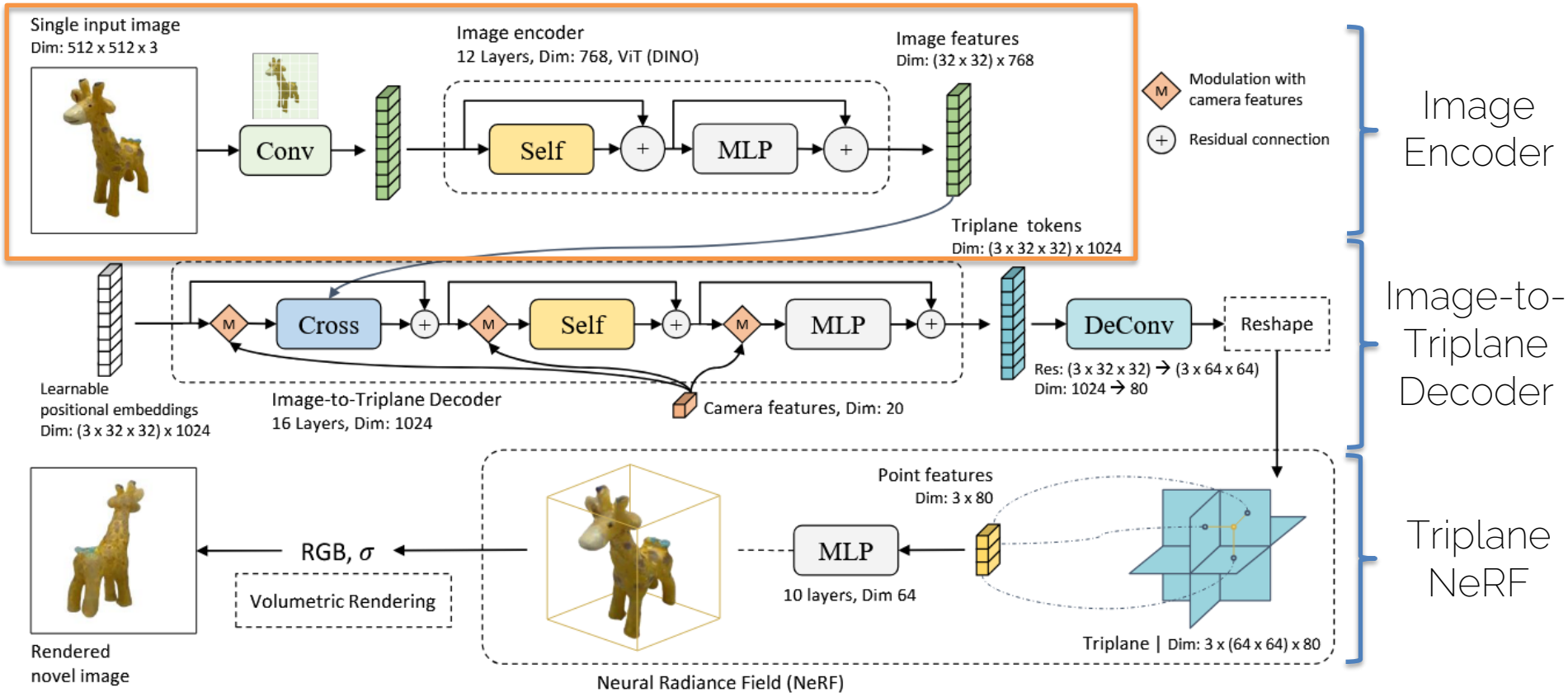
Important: Position & view direction are in input coordinate system.
→ View direction provides crucial information on relevance of a source view.

In single-image case, network boils down to 5 ResNet blocks.



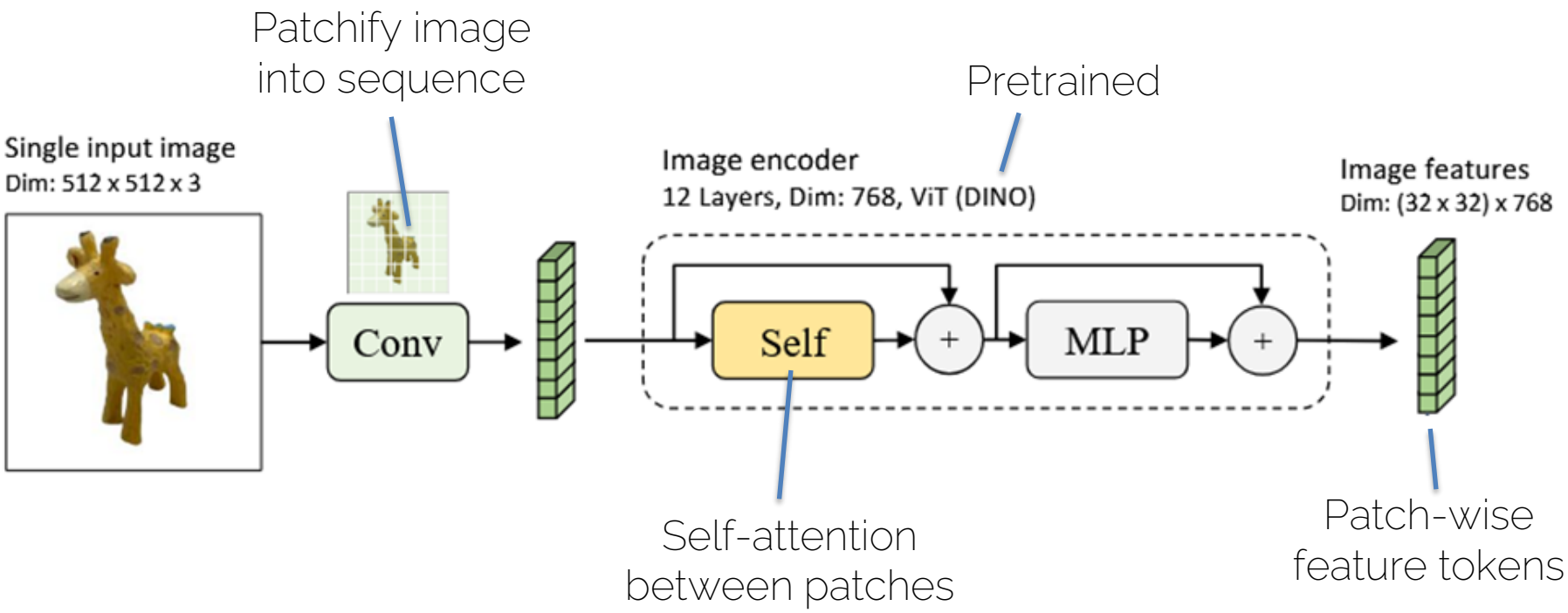
[Yu et al. '21] pixelNeRF: Neural Radiance Fields from One or Few Images

LRM: Large Reconstruction Model



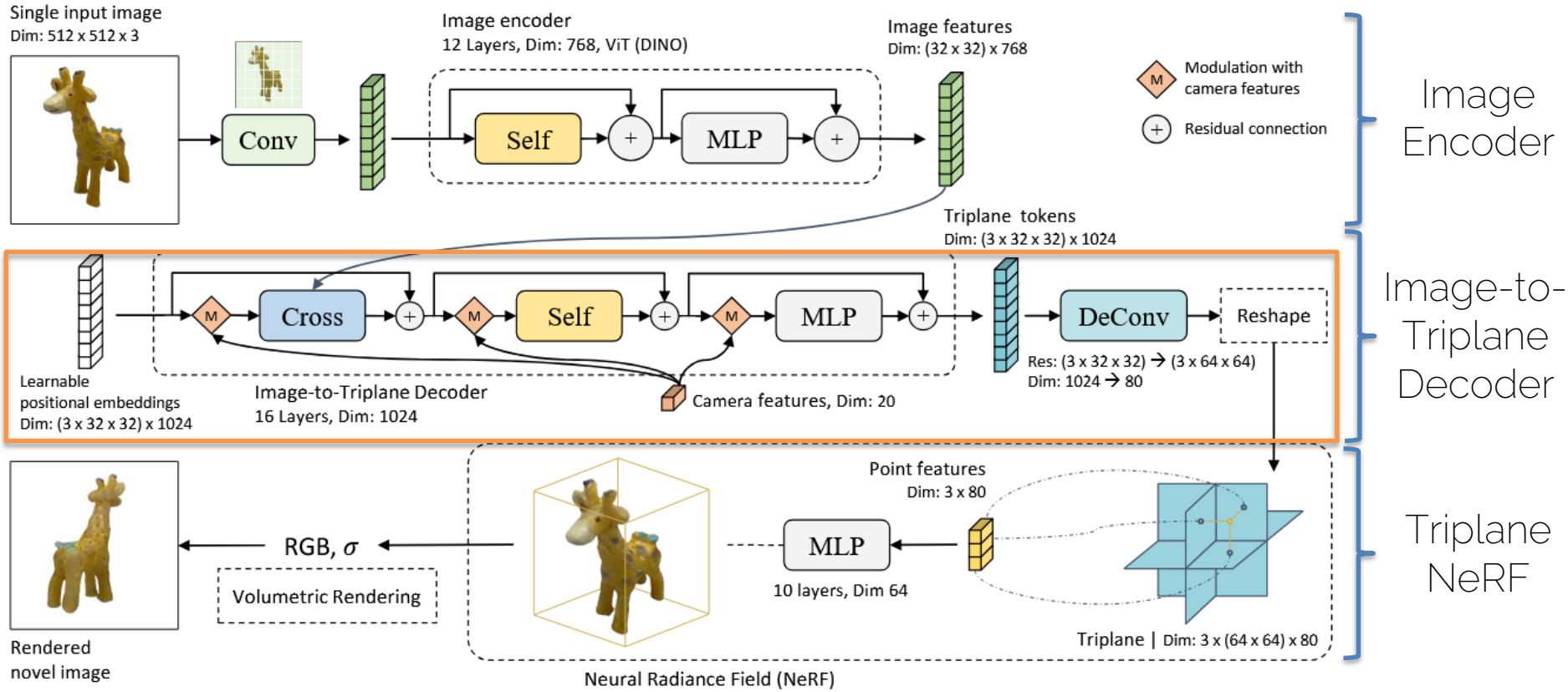
[Hong et al. '24] LRM: Large Reconstruction Model for Single Image to 3D

LRM – Image Encoder



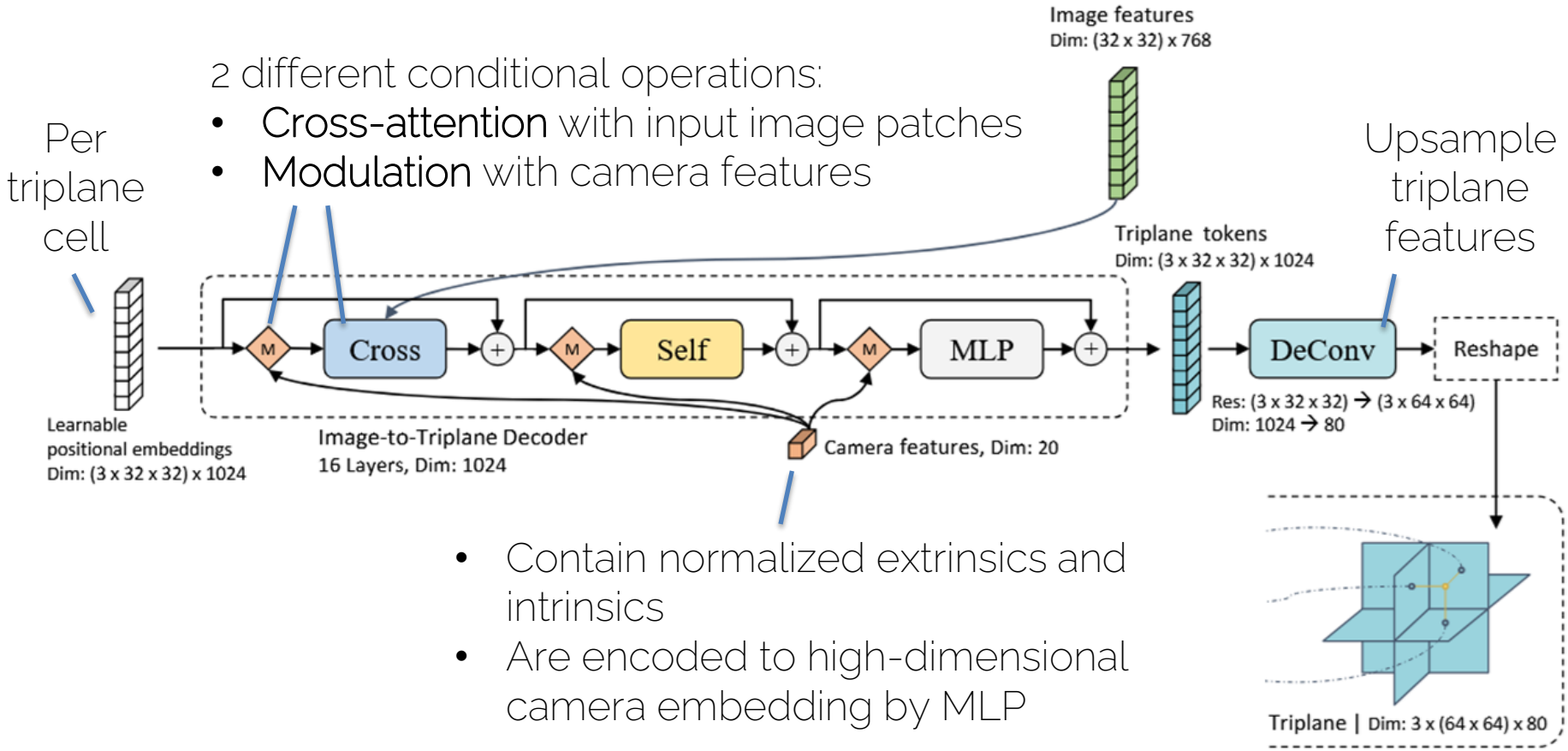
[Hong et al. '24] LRM: Large Reconstruction Model for Single Image to 3D

LRM: Large Reconstruction Model



[Hong et al. '24] LRM: Large Reconstruction Model for Single Image to 3D

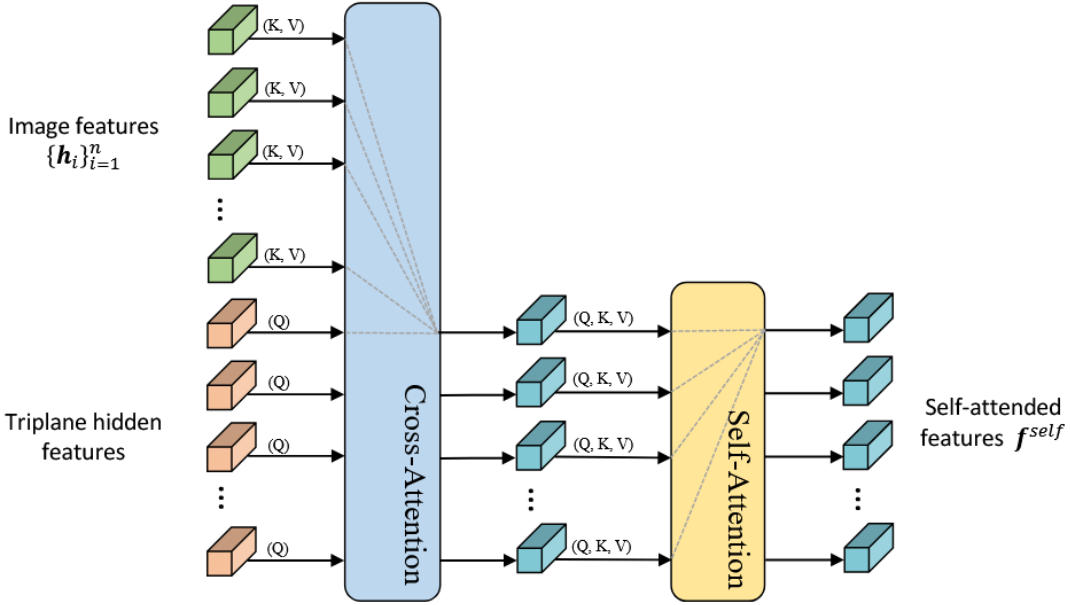
LRM – Image-to-Triplane Decoder



[Hong et al. '24] LRM: Large Reconstruction Model for Single Image to 3D

LRM – Image-to-Triplane Decoder

2 different conditional operations



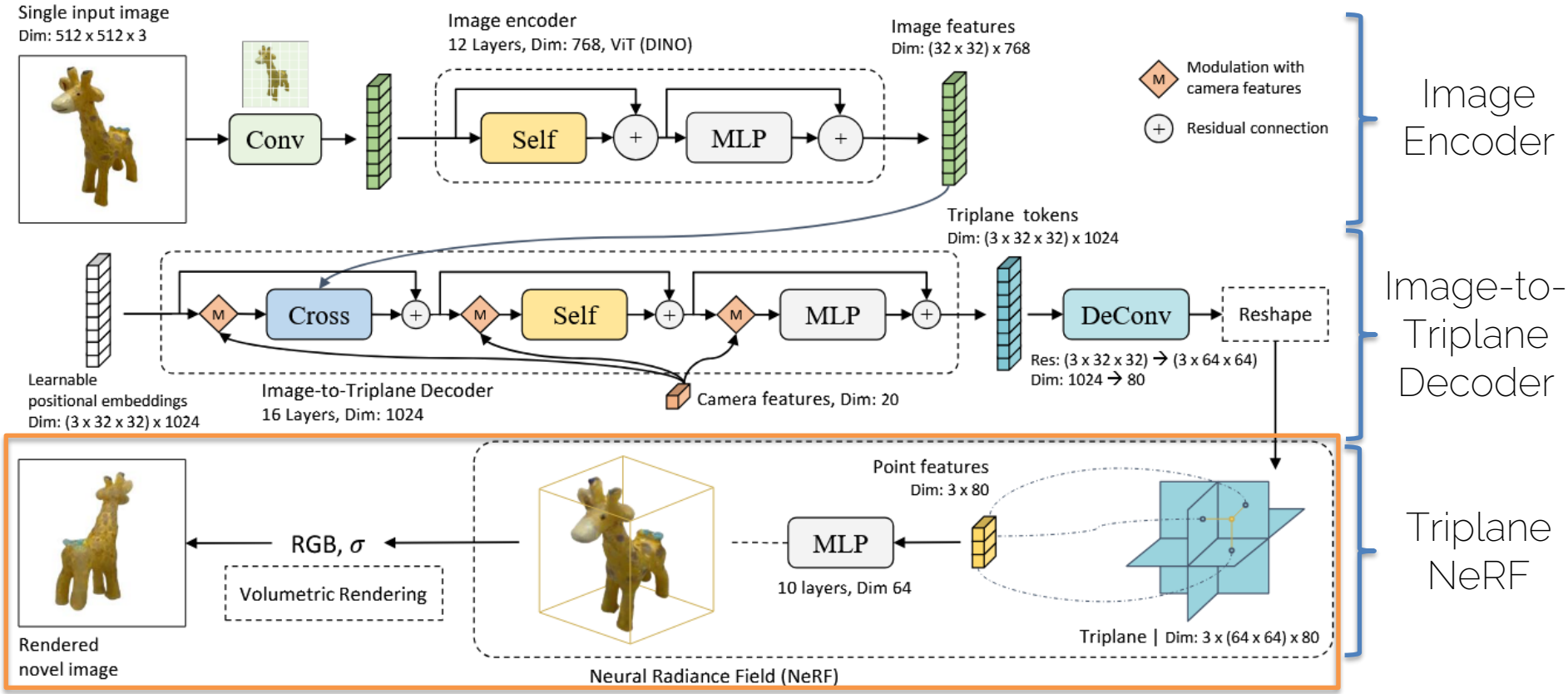
$$\gamma, \beta = \mathrm{MLP}^{\mathrm{mod}}(\tilde{\boldsymbol{c}})$$
$$\mathrm{ModLN_c}(\boldsymbol{f}_j) = \mathrm{LN}(\boldsymbol{f}_j) \cdot (1 + \gamma) + \beta$$

Modulation with camera features
→ control orientation and distortion of the whole shape

Cross-attention with input image patches
→ control fine-grained geometric and color information

[Hong et al. '24] LRM: Large Reconstruction Model for Single Image to 3D

LRM: Large Reconstruction Model



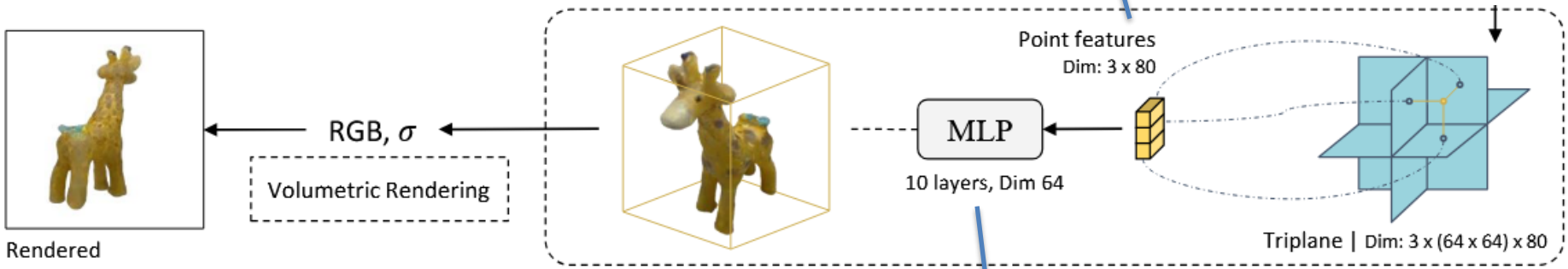
[Hong et al. '24] LRM: Large Reconstruction Model for Single Image to 3D

LRM – Triplane NeRF

Volume rendering as in NeRF; Training with color and LPIPS rendering losses

How to extract triplane features for a 3D point?

1. Project point onto each plane
2. Trilinearly interpolate triplane features
3. Obtain one vector per triplane

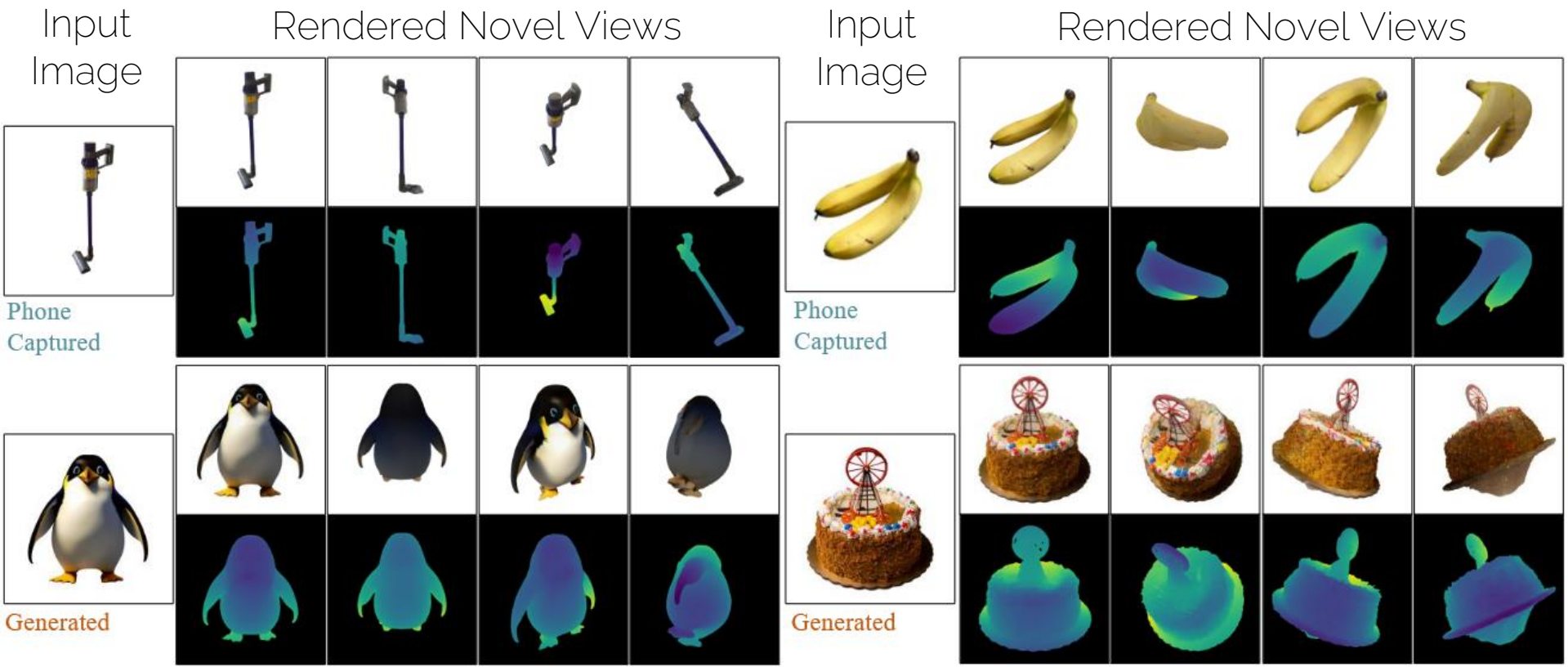


MLP maps features to color and density

3 axis-aligned feature planes

[Hong et al. '24] LRM: Large Reconstruction Model for Single Image to 3D

LRM – Results



[Hong et al. '24] LRM: Large Reconstruction Model for Single Image to 3D

GRM: Large Gaussian Reconstruction Model

Single image to 3D

Text to 3D

Multi-view to 3D

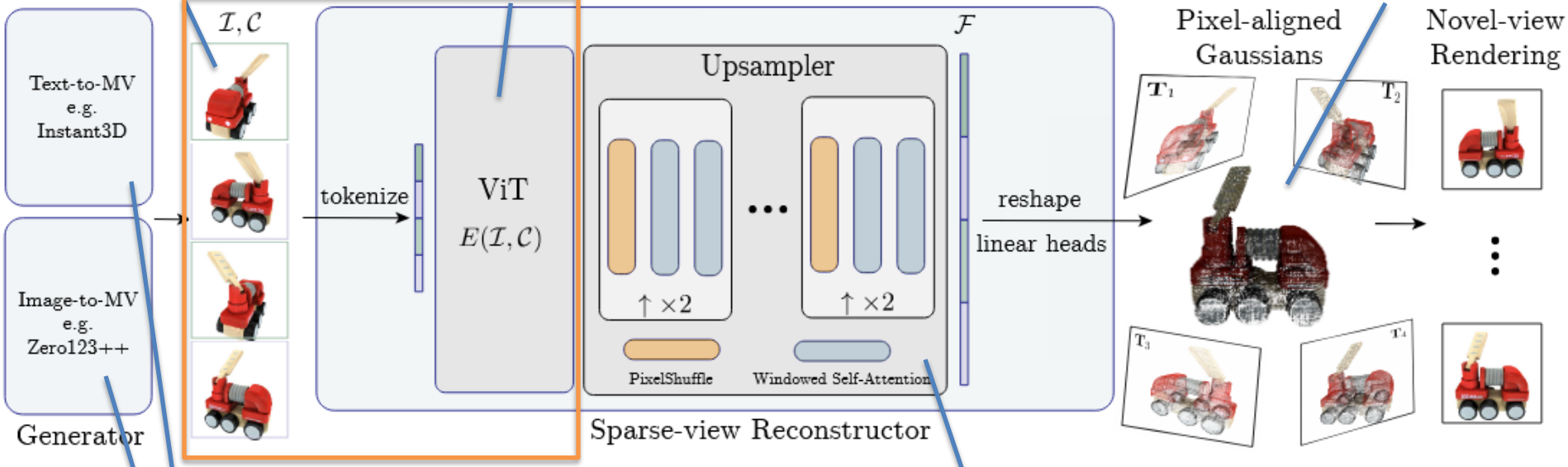
[Xu & Shi et al. '24] GRM: Large Gaussian Reconstruction Model for Efficient 3D Reconstruction and Generation

GRM: Large Gaussian Reconstruction Model

4 images and camera poses

Vision Transformer Encoder

Combine pixel-aligned Gaussians to single 3DGS representation

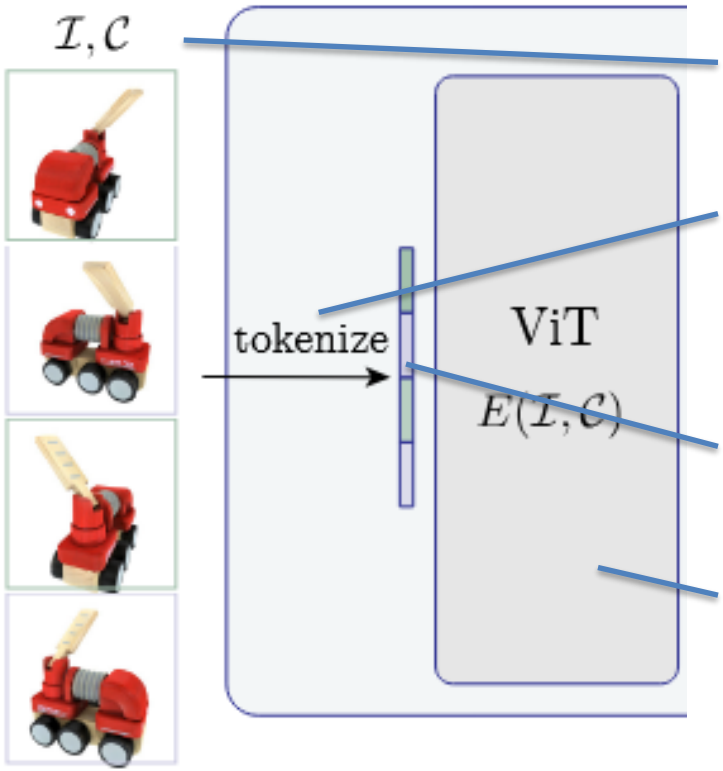


Facilitate text and single image input by using pretrained generators

Transformer-based Upsampler

[Xu & Shi et al. '24] GRM: Large Gaussian Reconstruction Model for Efficient 3D Reconstruction and Generation

GRM: ViT Encoder



1. Concatenate image with camera poses as Plücker rays
2. Encode with CNN tokenizer to create one sequence from all 4 images $4 \times {}^{H}/_{16} \times {}^{W}/_{16}$
3. Append learnable image position encodings.
4. Series of self-attention layers attending to all tokens across the input views.

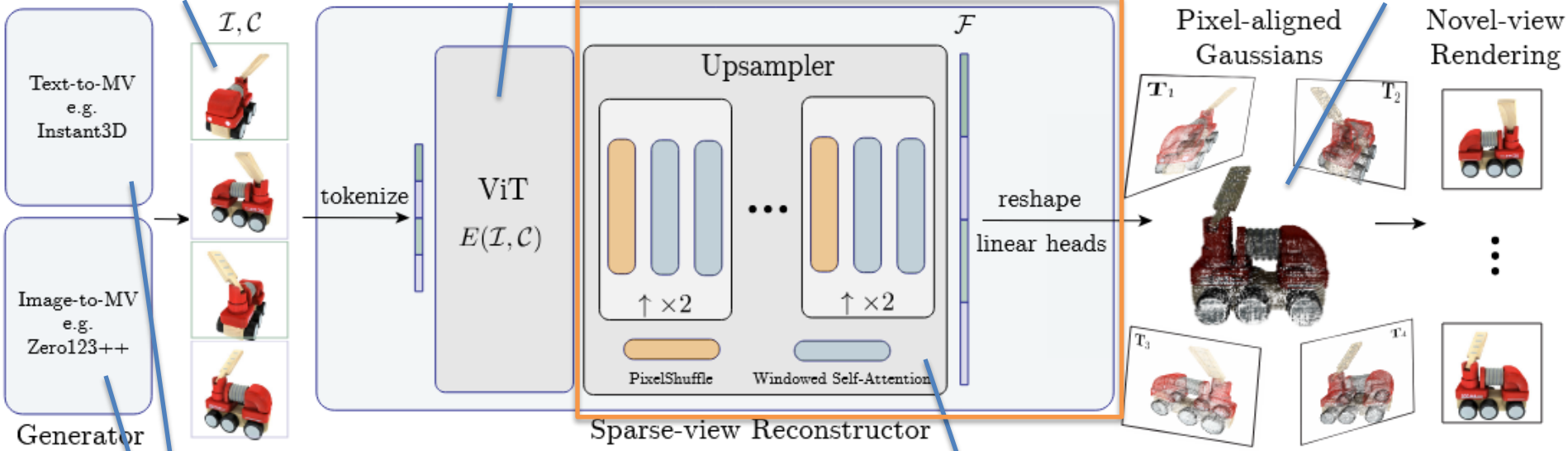
[Xu & Shi et al. '24] GRM: Large Gaussian Reconstruction Model for Efficient 3D Reconstruction and Generation

GRM: Large Gaussian Reconstruction Model

4 images and camera poses

Vision Transformer Encoder

Combine pixel-aligned Gaussians to single 3DGS representation



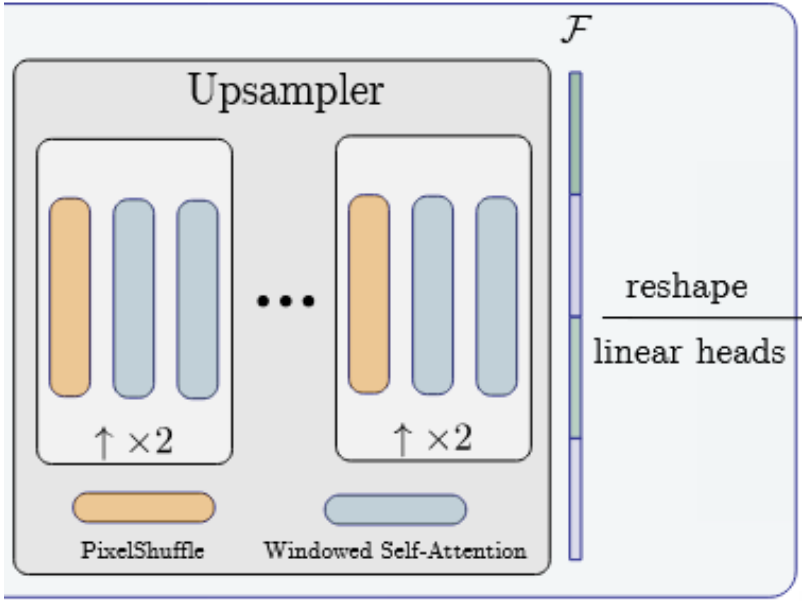
Facilitate text and single image input by using pretrained generators

Transformer-based Upsampler

[Xu & Shi et al. '24] GRM: Large Gaussian Reconstruction Model for Efficient 3D Reconstruction and Generation

GRM: Transformer-based Upsampler

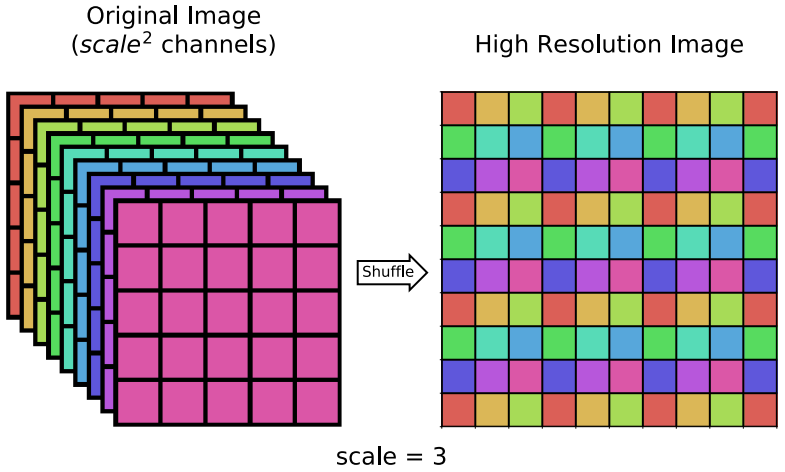
- Multiple upsample blocks progressively upsample by factor 2 until reaching $H \times W$
  - Quadruple number of channels
  - Double spatial dimension with PixelShuffle



[Xu & Shi et al. '24] GRM: Large Gaussian Reconstruction Model for Efficient 3D Reconstruction and Generation

GRM: Transformer-based Upsampler

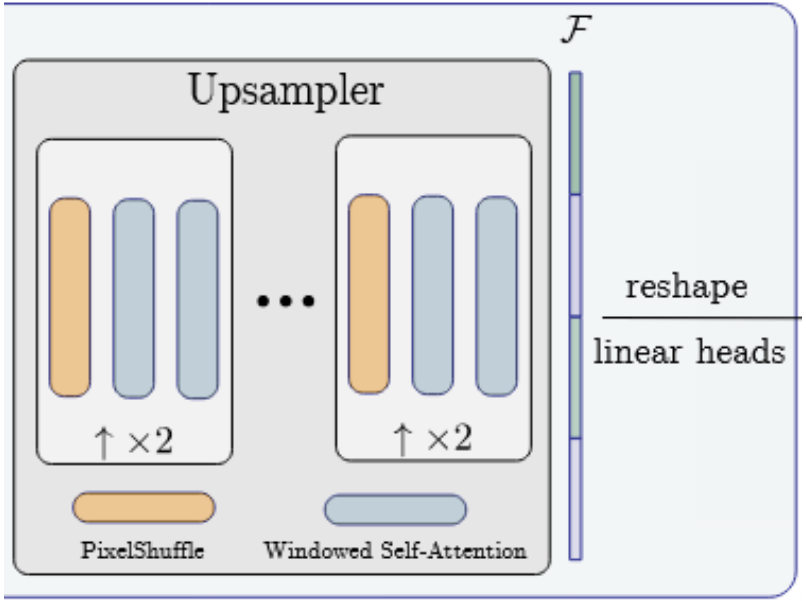
- Multiple upsample blocks progressively upsample by factor 2 until reaching $H \times W$
  - Quadruple number of channels
  - Double spatial dimension with **PixelShuffle**



https://nico-curti.github.io/NumPyNet/NumPyNet/layers/pixelshuffle_layer.html

GRM: Transformer-based Upsampler

- **Windowed Self-Attention**: balance between need for non-local multi-view information aggregation and feasible computation cost
- Separate linear heads produce Gaussian features

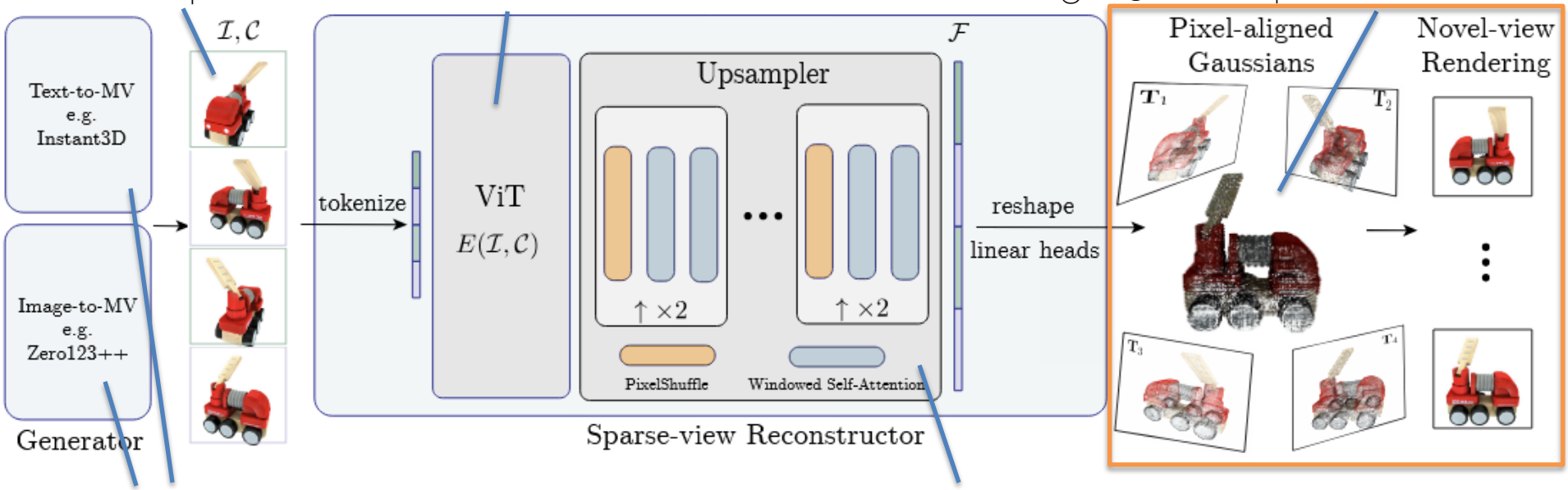


GRM: Large Gaussian Reconstruction Model

4 images and camera poses

Vision Transformer Encoder

Combine pixel-aligned Gaussians to single 3DGS representation



Facilitate text and single image input by using pretrained generators

Transformer-based Upsampler

[Xu & Shi et al. '24] GRM: Large Gaussian Reconstruction Model for Efficient 3D Reconstruction and Generation

GRM: Pixel-aligned 3D Gaussians

- Predict one Gaussian per pixel, i.e., one Gaussian attribute map $H \times W \times C$ per input image, $C = 12$, depth (1), rotation (4), scale (3), opacity (1), rgb (3).

→ Pixel-aligned Gaussians are easier to learn

→ 3D Gaussians render in **real-time**

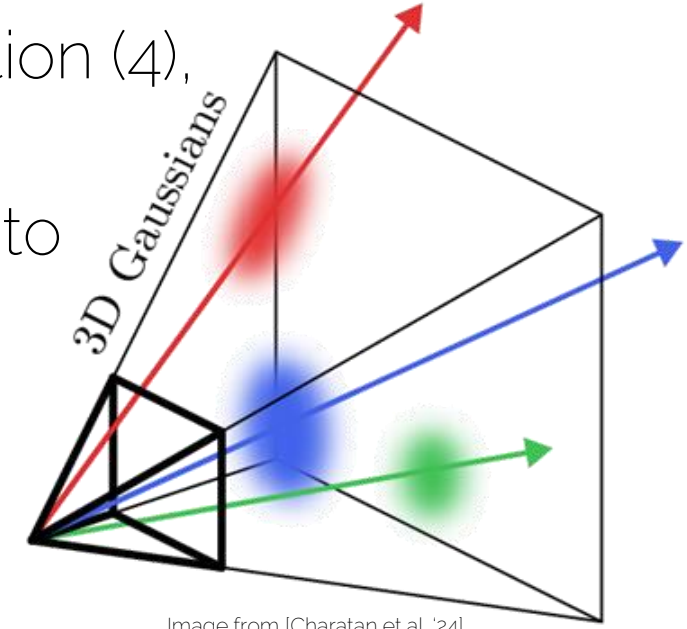
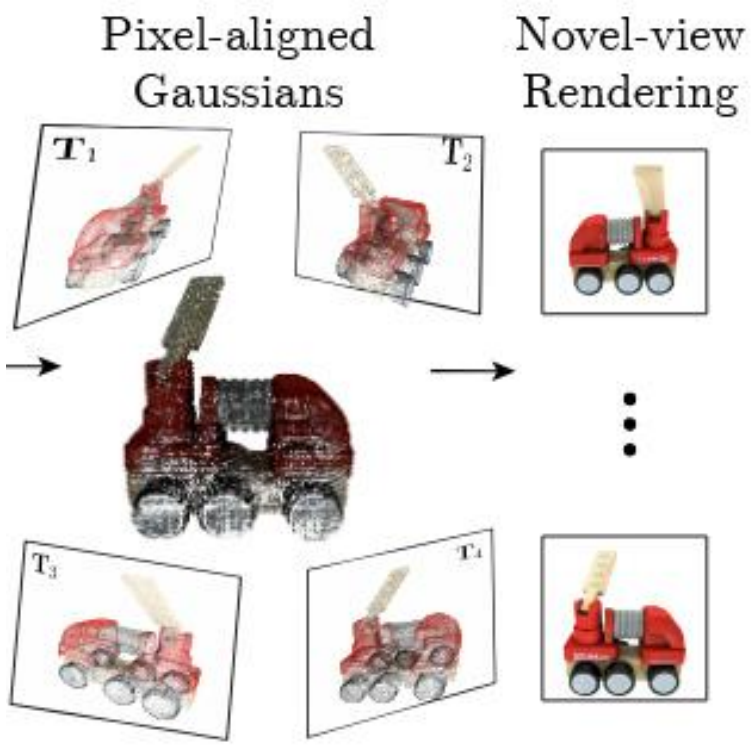


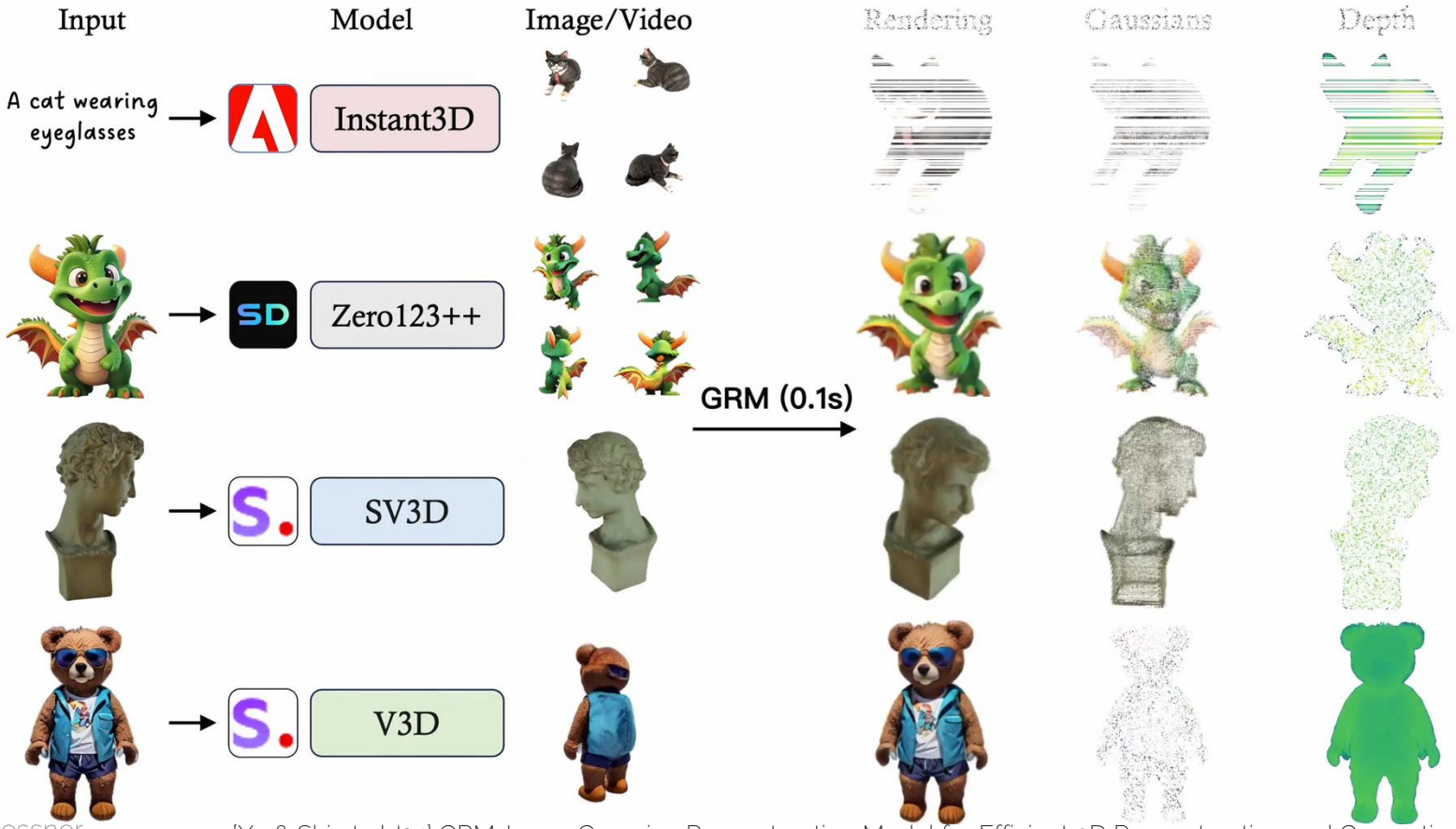
Image from [Charatan et al. '24]

GRM: Pixel-aligned 3D Gaussians

1. Backproject pixel-aligned Gaussians to single 3DGS representation
2. Render as in 3DGS
3. Supervise novel views with image (color & perceptual) and mask loss



GRM - Results



[Xu & Shi et al. '24] GRM: Large Gaussian Reconstruction Model for Efficient 3D Reconstruction and Generation

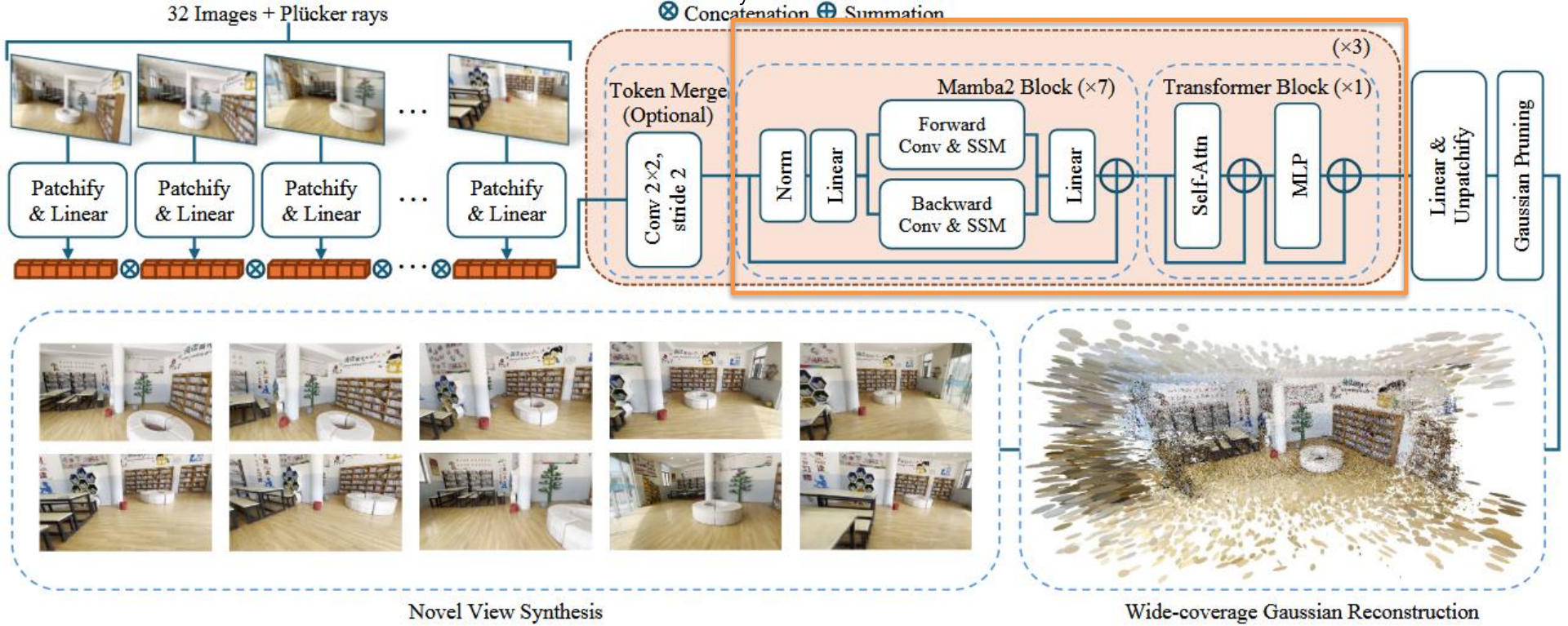
Long-LRM: Long-sequence Large Reconstruction Model

→ from 32 images in 1.3s

[Chen et al. 2024] Long-LRM: Long-sequence Large Reconstruction Model for Wide-coverage Gaussian Splats

Long-LRM: Long-sequence Large Reconstruction Model

Hybrid Blocks: Mamba2 + Transformer

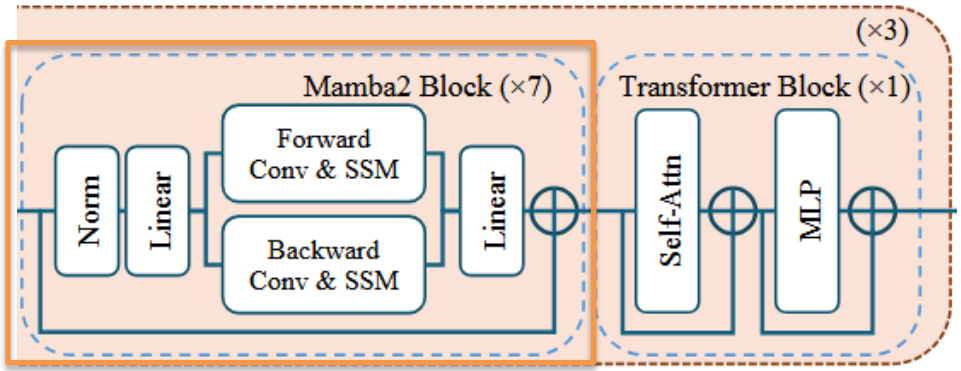


[Chen et al. 2024] Long-LRM: Long-sequence Large Reconstruction Model for Wide-coverage Gaussian Splats

Long-LRM – Hybrid Block

- Combine Mamba2 blocks & transformer blocks → better scalability to higher resolution & denser views
- Hybrid block = 7 Mamba2 blocks + 1 transformer block

L = Sequence Length O(L) O(L^2)

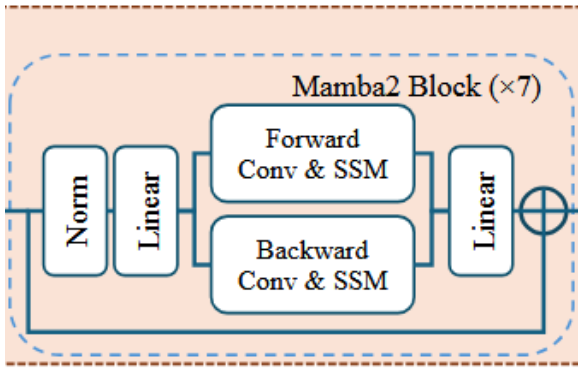


Due to global self-attention

[Chen et al. 2024] Long-LRM: Long-sequence Large Reconstruction Model for Wide-coverage Gaussian Splats

Long-LRM – Mamba2 Block

- Mamba2 block [Dao & Gu ,24] is designed for language tasks → scans through sequence in one direction → suboptimal for images.
- Vision Mamba [Zhu et al. '24] take bi-directional scans over the concatenated token sequence



Hidden State

Input Token

$$h_t = \mathbf{A}h_{t-1} + \mathbf{B}x_t$$

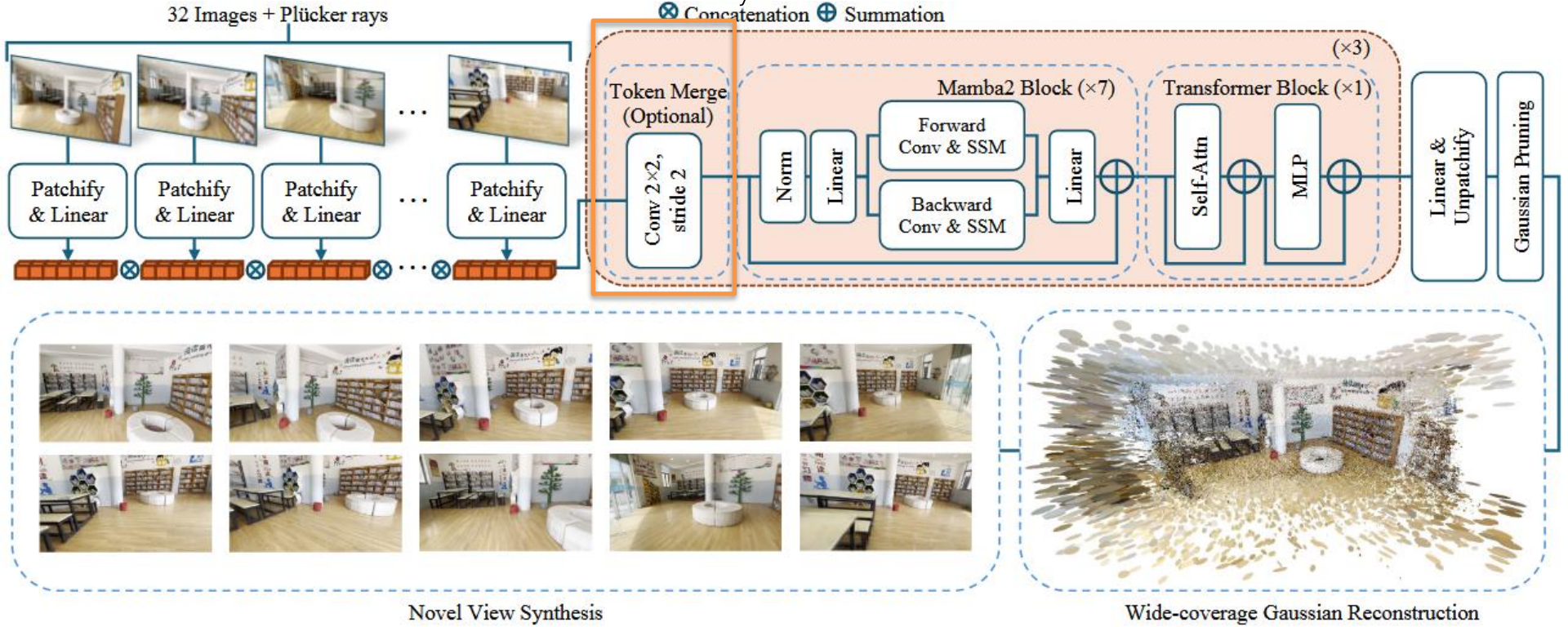
$$y_t = \mathbf{C}h_t$$

Output Token

Compute state params from input using linear layer

Long-LRM: Long-sequence Large Reconstruction Model

Hybrid Blocks: Mamba2 + Transformer

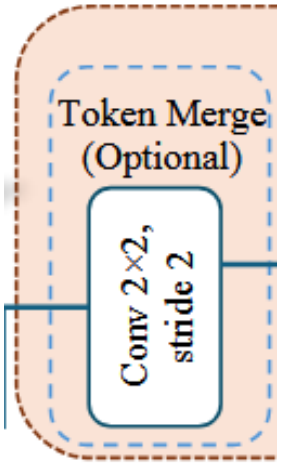


[Chen et al. 2024] Long-LRM: Long-sequence Large Reconstruction Model for Wide-coverage Gaussian Splats

Long-LRM – Token Merge

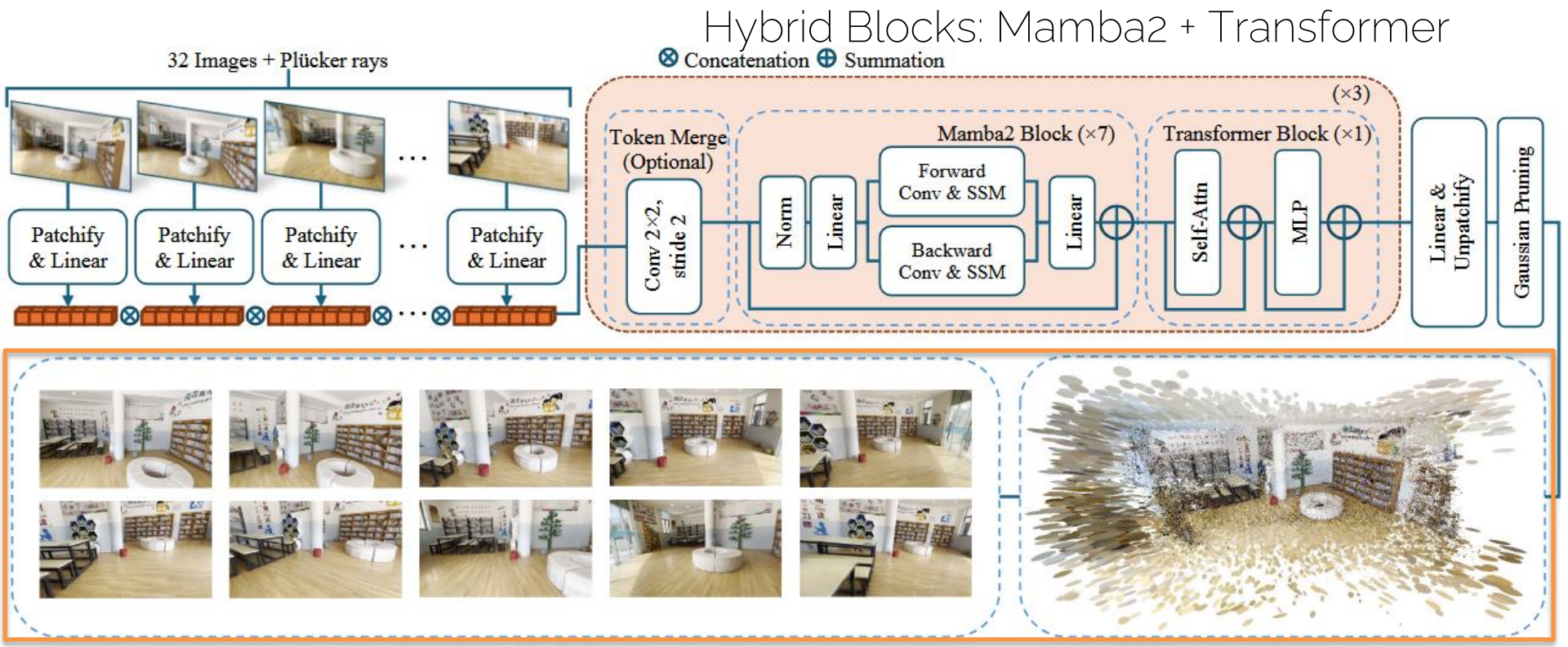
1. Reshape input sequence $L \times D$ back to $N \times \frac{H}{p} \times \frac{W}{p} \times D$ where N = #images, p = patch size
2. Apply 2D convolution, kernel size 2, stride 2, resulting in $N \times \frac{H}{2p} \times \frac{W}{2p} \times D'$
3. Resape back to $\frac{L}{4} \times D'$

→ Reduces sequence length to 1/4



[Chen et al. 2024] Long-LRM: Long-sequence Large Reconstruction Model for Wide-coverage Gaussian Splats

Long-LRM: Long-sequence Large Reconstruction Model



Long-LRM – Decode to Gaussians

- Decode output tokens to per-pixel Gaussian parameters
- At training-time prune to fixed number of Gaussians, at test-time prune by opacity → improve efficiency at high resolution and increased views

Novel View Synthesis

Wide-coverage Gaussian Reconstruction

Long-LRM – Training Objective

$$\mathcal{L} = \mathcal{L}_{\text{image}} + \lambda_{\text{opacity}} \cdot \mathcal{L}_{\text{opacity}} + \lambda_{\text{depth}} \cdot \mathcal{L}_{\text{depth}}$$

$$\mathcal{L}_{\text{image}} = \frac{1}{M} \sum_{i=1}^{M} \left(\text{MSE}\left(\mathbf{I}_i^{\text{gt}}, \mathbf{I}_i^{\text{pred}}\right) + \lambda \cdot \text{Perceptual}\left(\mathbf{I}_i^{\text{gt}}, \mathbf{I}_i^{\text{pred}}\right) \right)$$

$$\mathcal{L}_{\text{depth}} = \frac{1}{M} \sum_{i=1}^{M} \text{Smooth-L1}\left(\mathbf{D}_i^{\text{da}}, \mathbf{D}_i^{\text{pred}}\right)$$

Depth regularizer to avoid "floaters"

$$\mathcal{L}_{\text{opacity}} = \frac{1}{N} \sum_{i=1}^{N} |o_i|$$

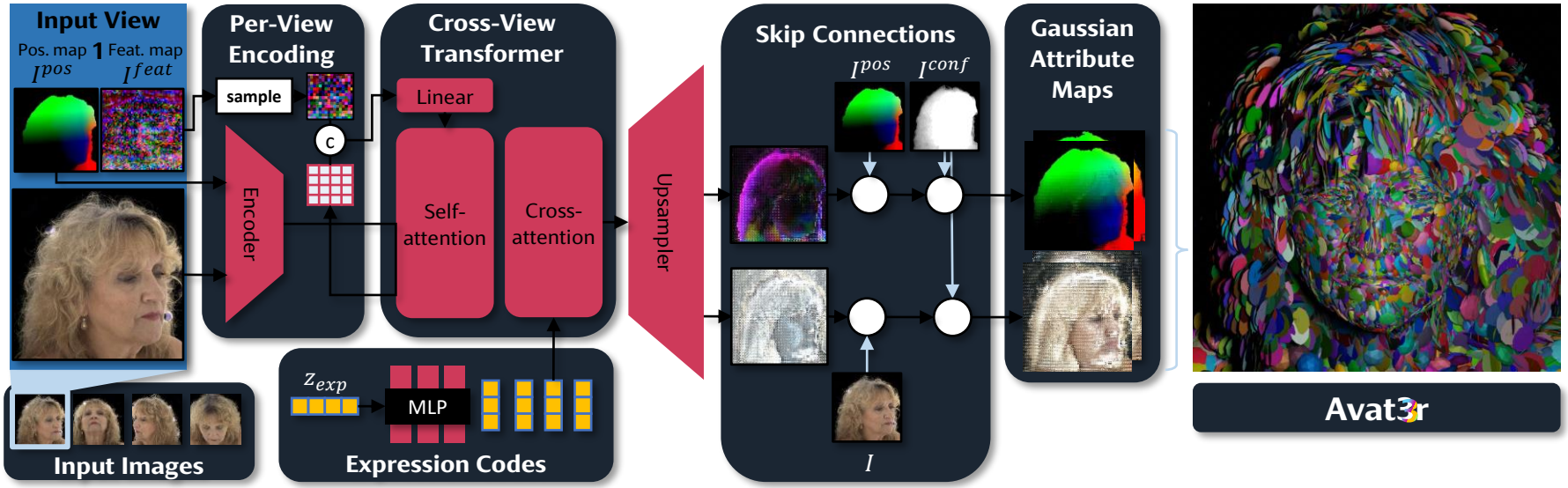
Opacity regularizer to reduce the number of relevant Gaussians

[Chen et al. 2024] Long-LRM: Long-sequence Large Reconstruction Model for Wide-coverage Gaussian Splats

Long-LRM – Results

[Chen et al. 2024] Long-LRM: Long-sequence Large Reconstruction Model for Wide-coverage Gaussian Splats

Avat3r: Generalizable 3D Head Avatars



Avat3r: Generalizable 3D Head Avatars

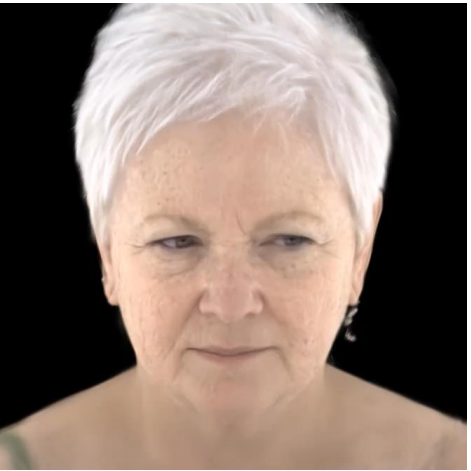
arXiv'25 [Kirschstein et al.] Avat3r

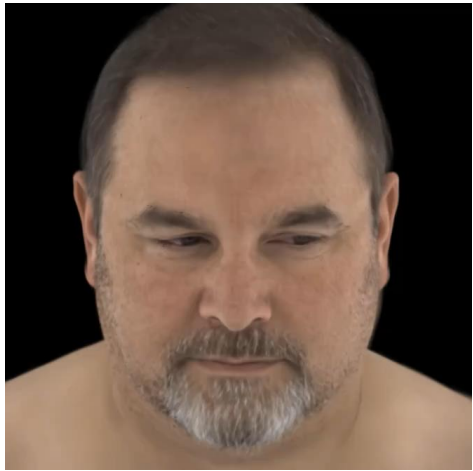
Avat3r

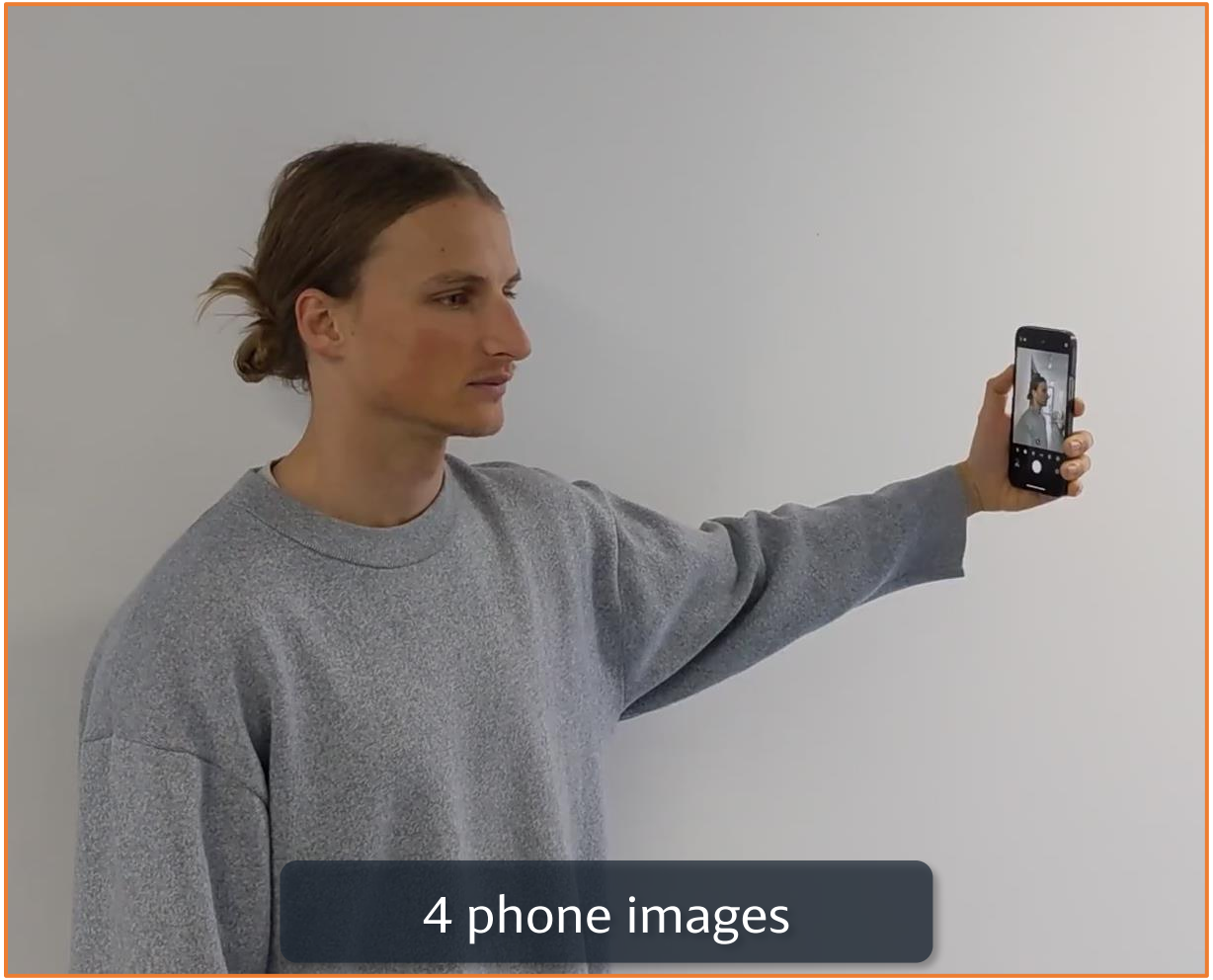
Large Animatable Reconstruction Model for High-fidelity 3D Head Avatars

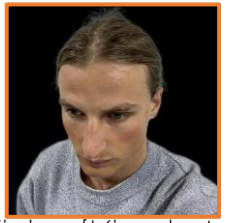
Tobias Kirschstein[1,2] - Javier Romero[2] - Artem Sevastopolsky[1,2] - Matthias Nießner[1] - Shunsuke Saito[2]

[1]Technical University of Munich [2]Meta Reality Labs

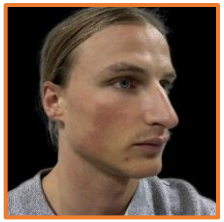


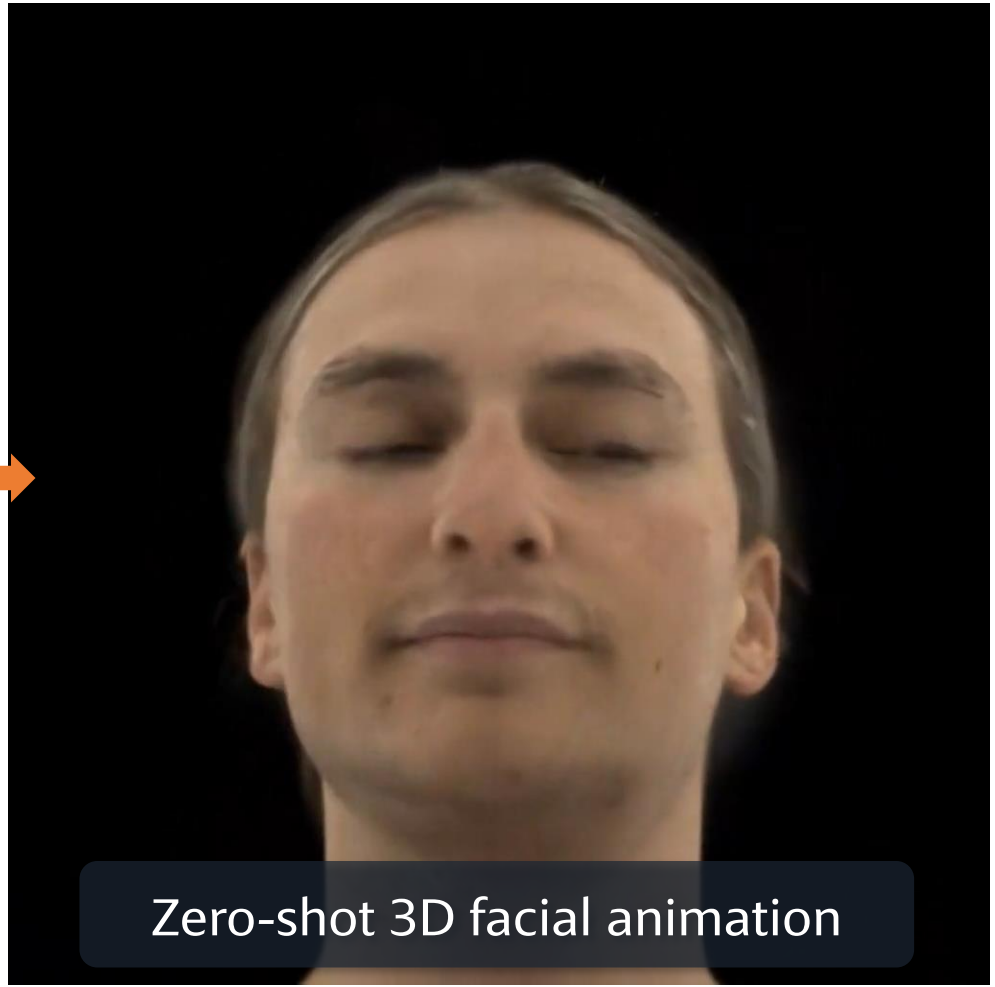




arXiv'25 [Kirschstein et al.] Avat3r



Avat3r



Reading Homework

- [Hong et al. '24] LRM: Large Reconstruction Model for Single Image to 3D
  - https://arxiv.org/pdf/2311.04400

Literature

- [Yu et al. ‘21] pixelNeRF: Neural Radiance Fields from One or Few Images
- [Hong et al. ‘24] LRM: Large Reconstruction Model for Single Image to 3D
- [Xu & Shi et al. ‘24] GRM: Large Gaussian Reconstruction Model for Efficient 3D Reconstruction and Generation

Literature

- [Chen et al. 2024] Long-LRM: Long-sequence Large Reconstruction Model for Wide-coverage Gaussian Splats
- arXiv '25 [Kirschstein et al.] Avat3r

Thanks for watching!