StereoSplat

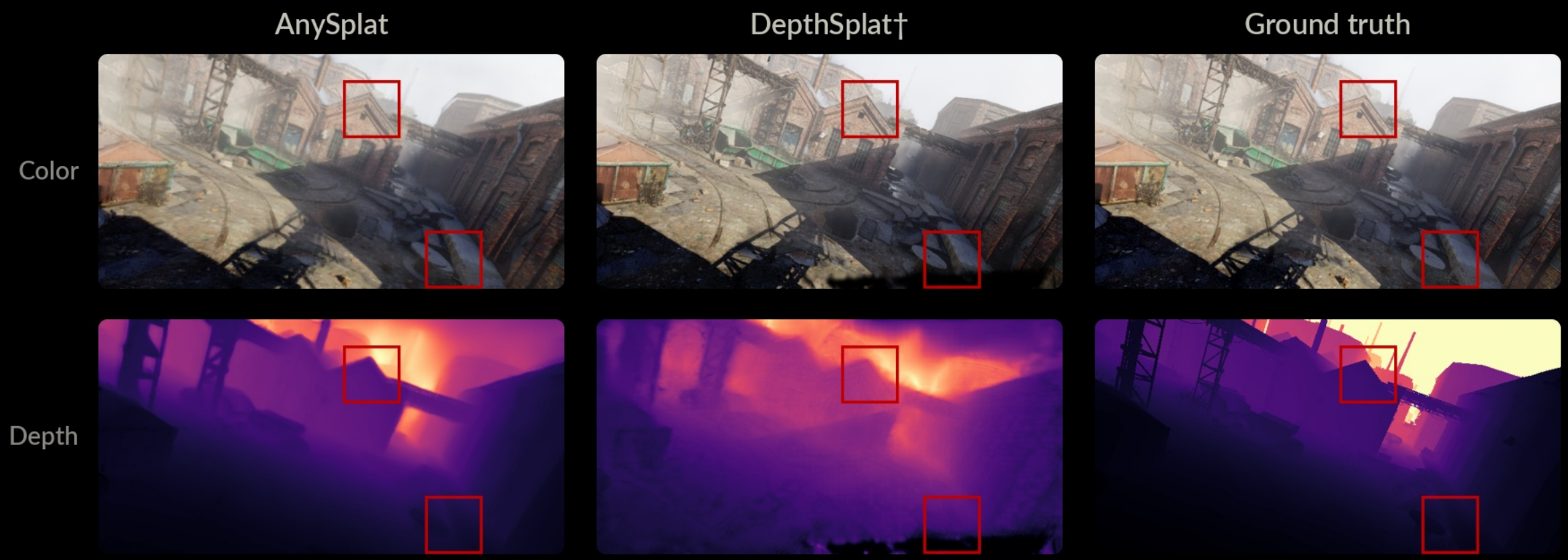
Metric-Scale Novel View Synthesis via Stereo-Grounded Gaussian Splatting

Vladimir Yugay[1] Denis Rozumny[2] Theo Gevers[1]
Elias Vansteenkiste[2] Martin R. Oswald[1] Diogo C. Luvizon[2]

[1] University of Amsterdam [2] Meta

NeurIPS 2026

Feed-forward Gaussian splatting today



† uses ground-truth poses

Calibrated stereo gives metric depth



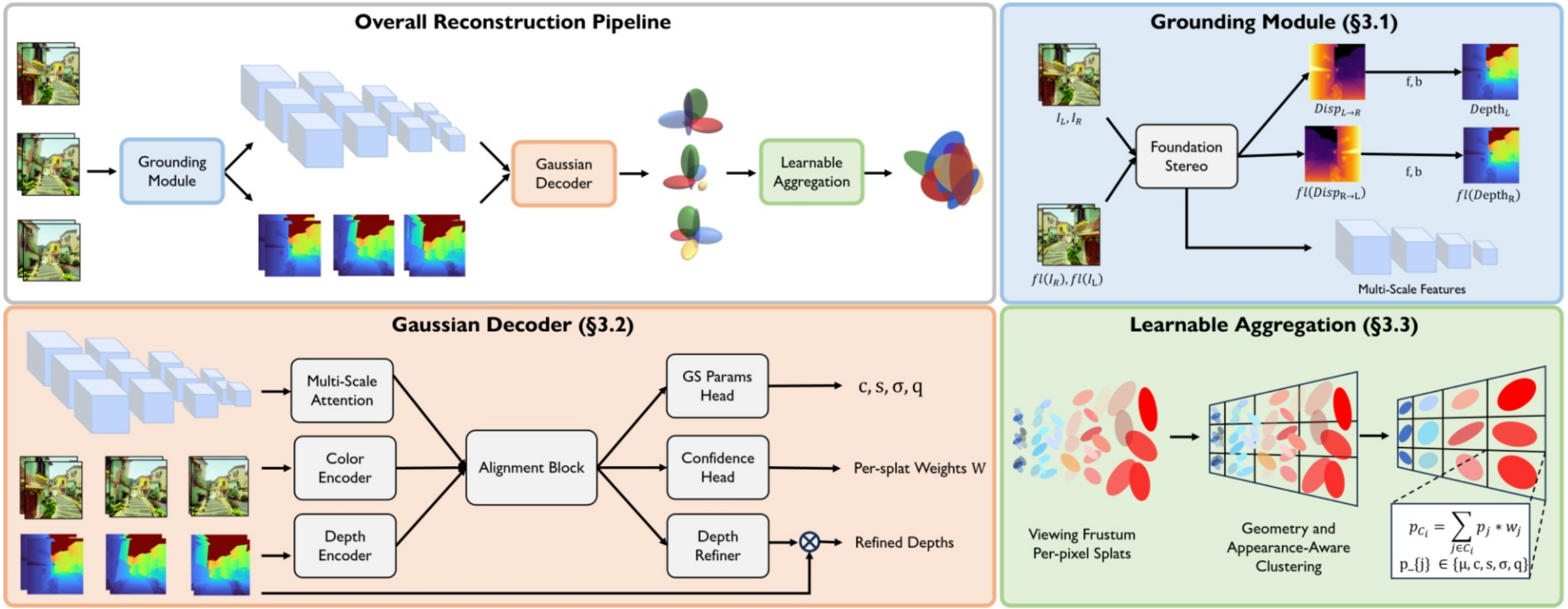
Stereo pair

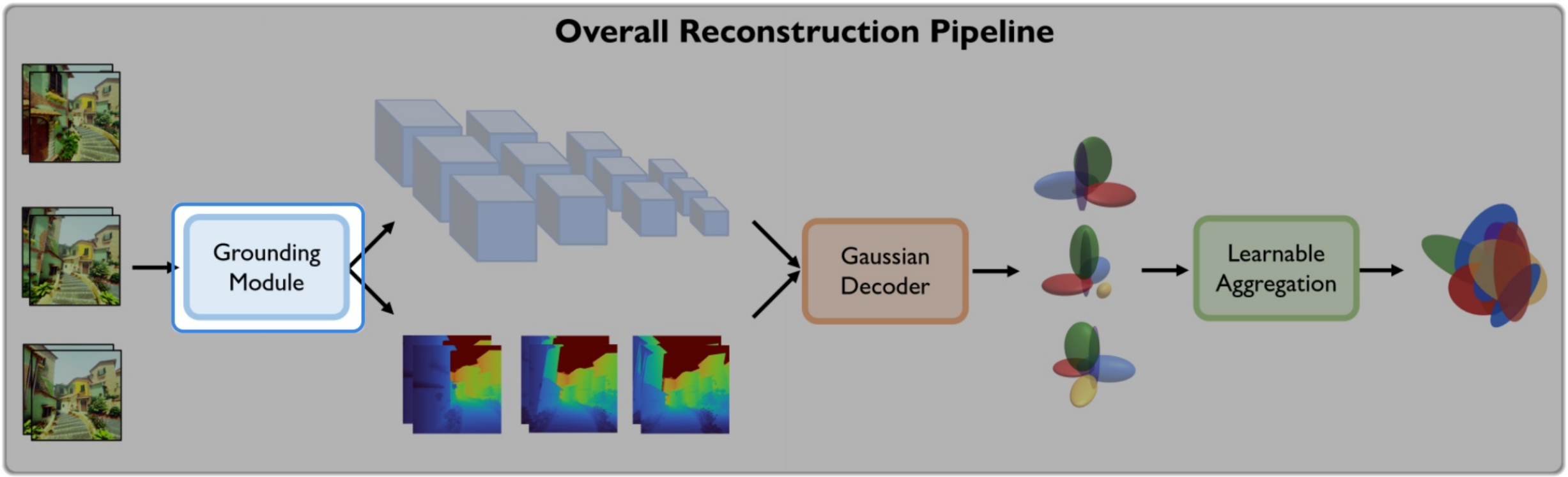


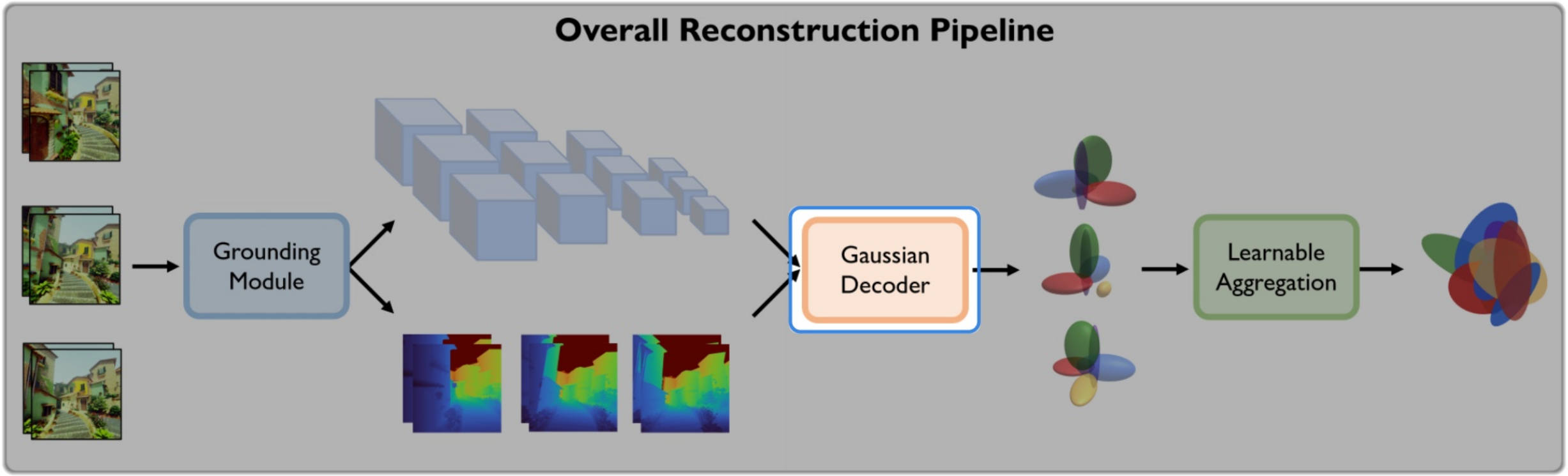
Metric depth

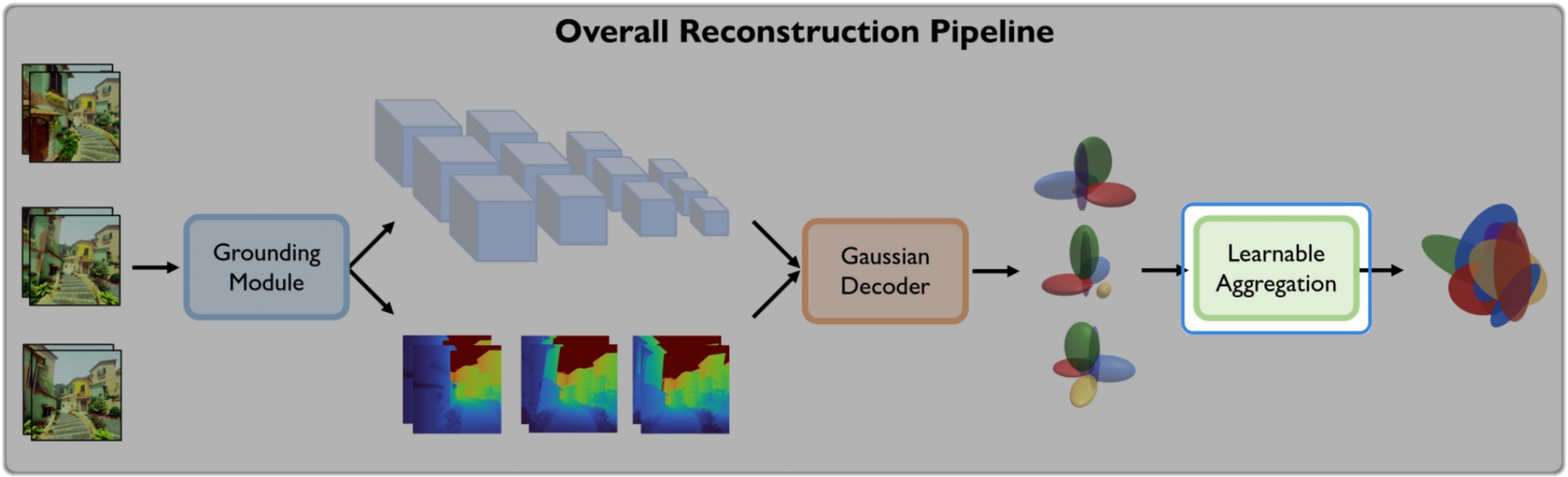
0 m 20 m 40 m 60 m

depth = focal × baseline / disparity


Overall Reconstruction Pipeline
Grounding Module
Gaussian Decoder
Learnable Aggregation
Grounding Module (§3.1)
I_L, I_R
Foundation Stereo
$Disp_{L\to R}$
f, b
$Depth_L$
$fl(Disp_{R\to L})$
f, b
$fl(Depth_R)$
$fl(I_R), fl(I_L)$
Multi-Scale Features
Gaussian Decoder (§3.2)
Multi-Scale Attention
Color Encoder
Depth Encoder
Alignment Block
GS Params Head
Confidence Head
Depth Refiner
c, s, σ, q
Per-splat Weights W
Refined Depths
Learnable Aggregation (§3.3)
Viewing Frustum Per-pixel Splats
Geometry and Appearance-Aware Clustering
$p_{C_i} = \sum_{j \in C_i} p_j * w_j$
p_{j} ∈ {μ, c, s, σ, q}


Overall Reconstruction Pipeline
Grounding Module
Gaussian Decoder
Learnable Aggregation


Overall Reconstruction Pipeline
Grounding Module
Gaussian Decoder
Learnable Aggregation


Overall Reconstruction Pipeline
Grounding Module
Gaussian Decoder
Learnable Aggregation

Grounding Module (§3.1)

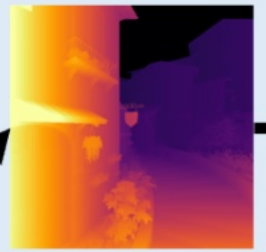


I_L, I_R



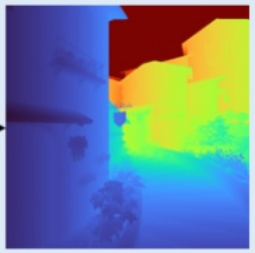
$fl(I_R), fl(I_L)$

Foundation Stereo

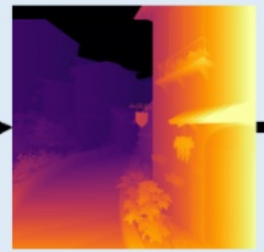


$Disp_{L\rightarrow R}$

f, b

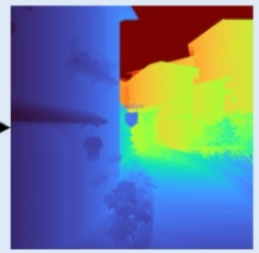


Depth_L

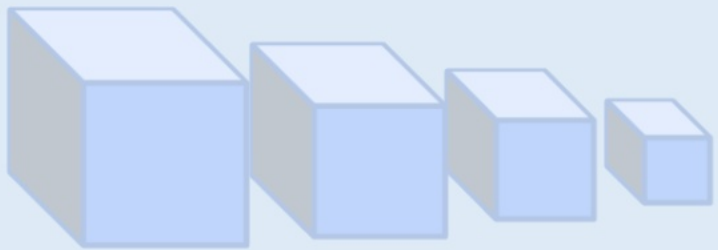


$fl(Disp_{\mathrm{R}\rightarrow\mathrm{L}})$

f, b

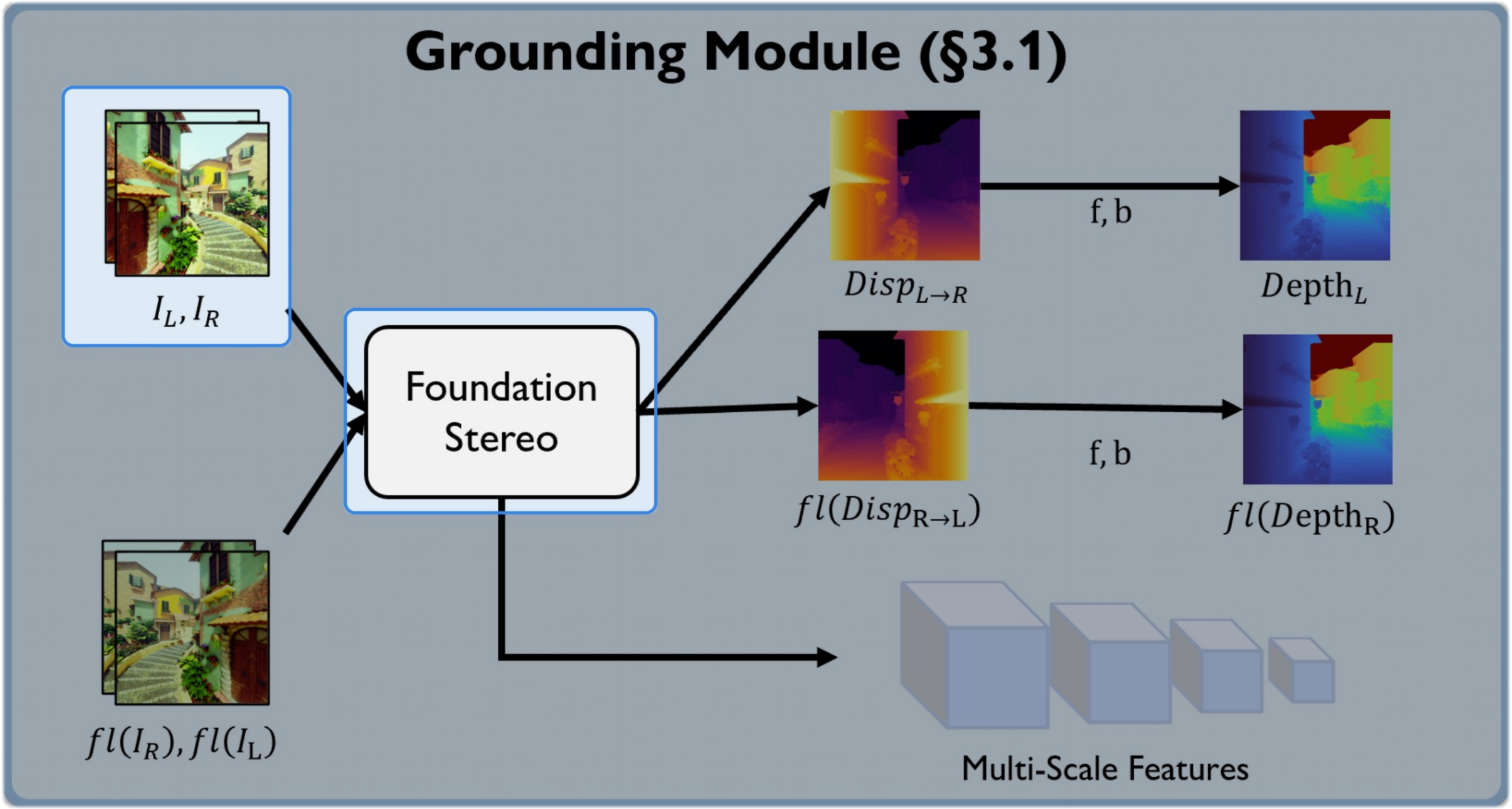


$fl(\mathrm{Depth}_{\mathrm{R}})$



Multi-Scale Features

Grounding Module (§3.1)

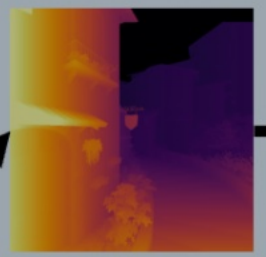


Grounding Module (§3.1)



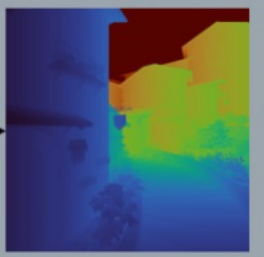
I_L, I_R

Foundation Stereo

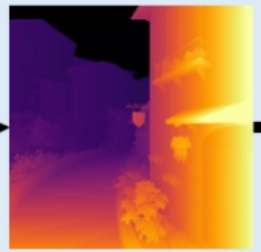


$Disp_{L\rightarrow R}$

f, b

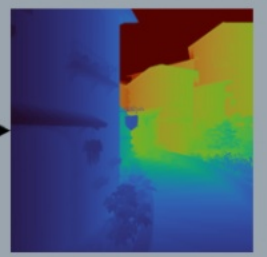


Depth_L



$fl(Disp_{\mathrm{R}\rightarrow\mathrm{L}})$

f, b



$fl(\mathrm{Depth}_{\mathrm{R}})$



$fl(I_R), fl(I_{\mathrm{L}})$



Multi-Scale Features

Grounding Module (§3.1)

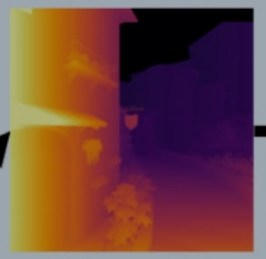


I_L, I_R



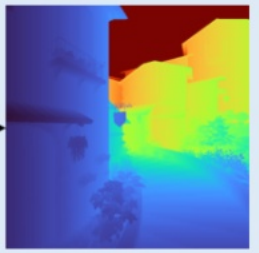
$fl(I_R), fl(I_L)$

Foundation Stereo

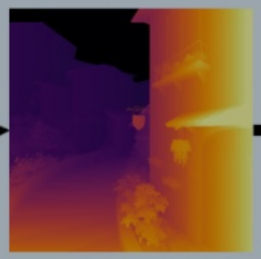


$Disp_{L\rightarrow R}$

f, b

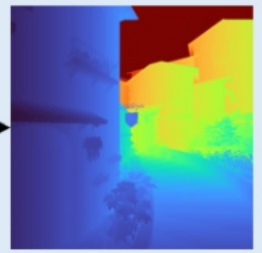


Depth_L

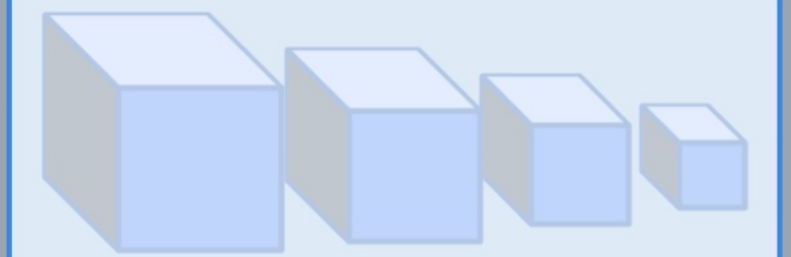


$fl(Disp_{\mathrm{R}\rightarrow\mathrm{L}})$

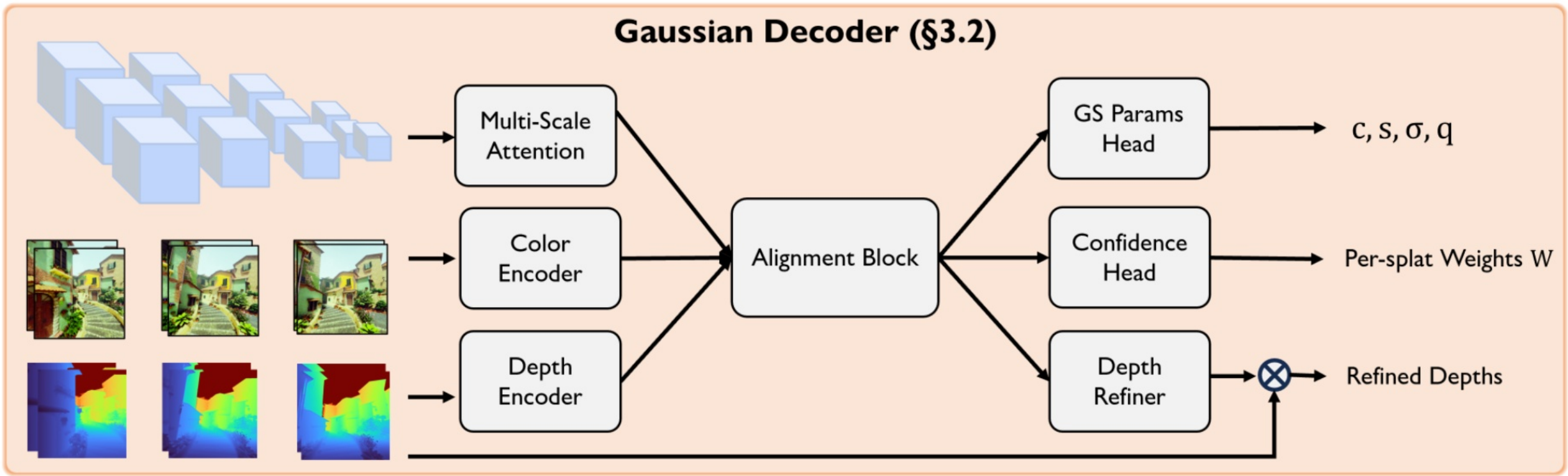
f, b

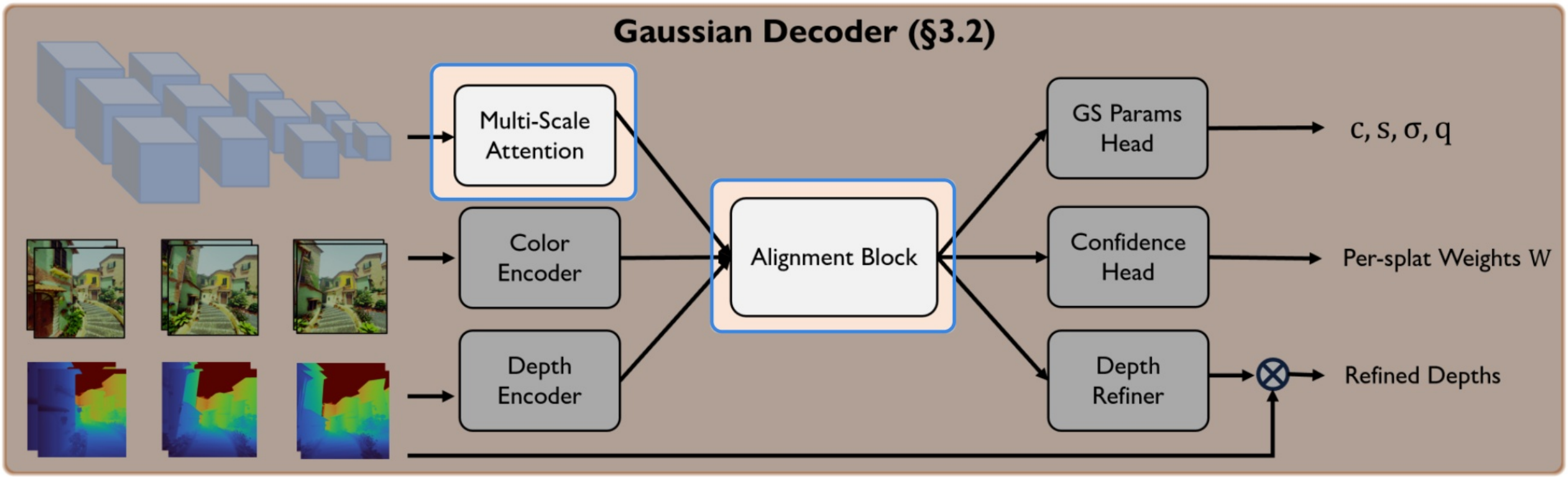


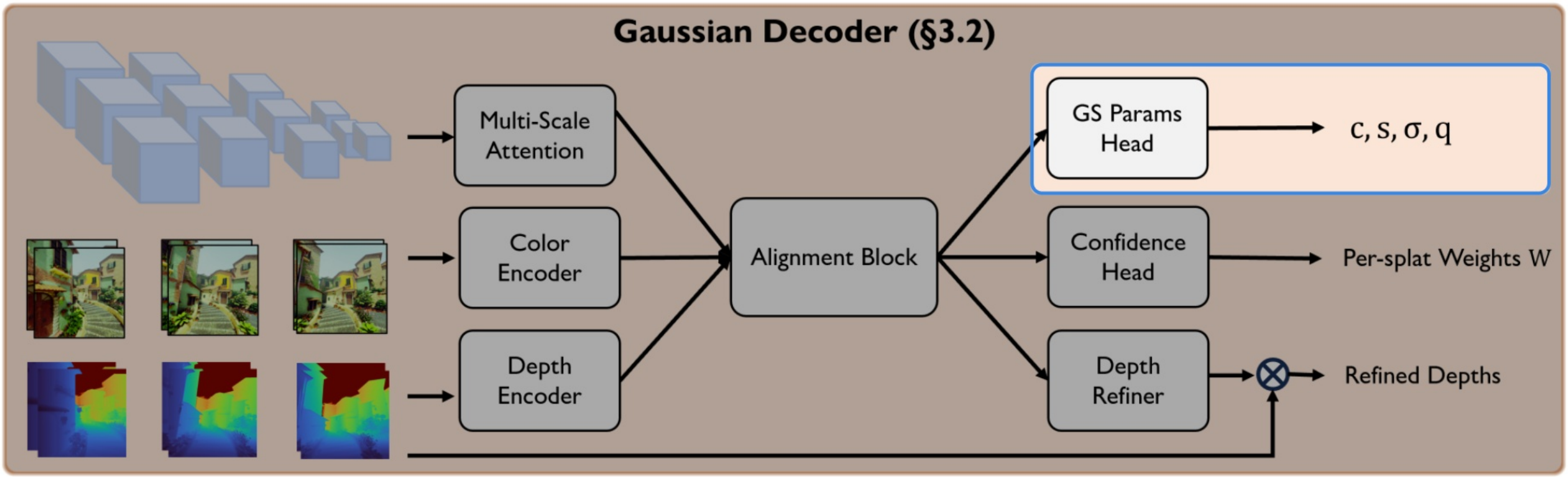
$fl(\mathrm{Depth}_{\mathrm{R}})$

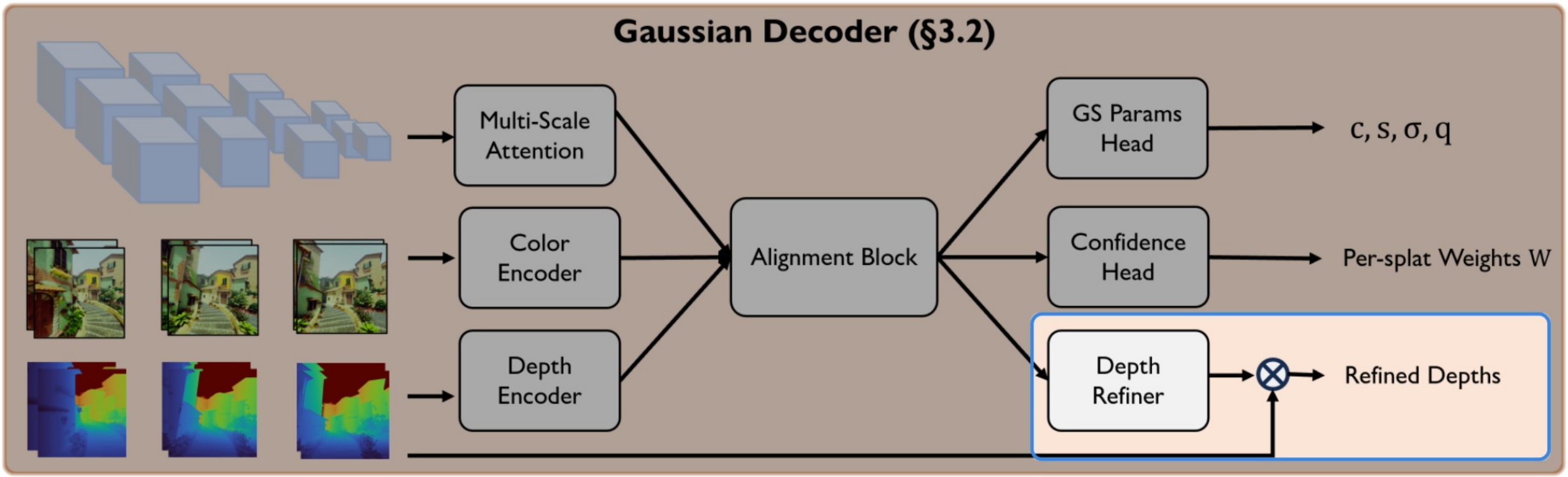


Multi-Scale Features

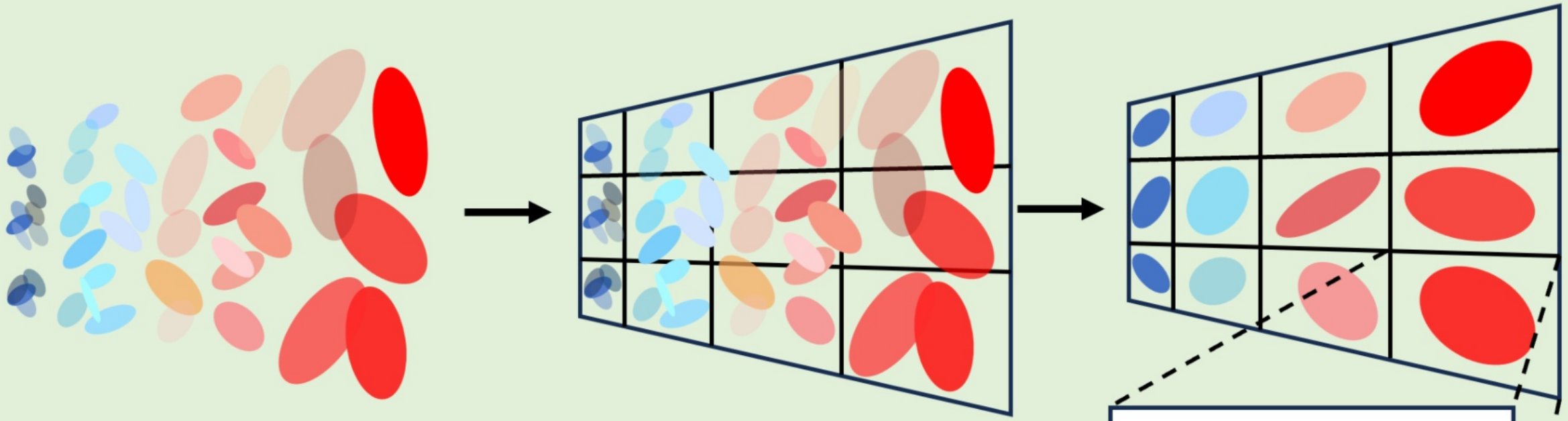

Gaussian Decoder (§3.2)
Multi-Scale Attention
Color Encoder
Depth Encoder
Alignment Block
GS Params Head
Confidence Head
Depth Refiner
c, s, σ, q
Per-splat Weights W
Refined Depths


Gaussian Decoder (§3.2)
Multi-Scale Attention
Color Encoder
Depth Encoder
Alignment Block
GS Params Head
Confidence Head
Depth Refiner
c, s, σ, q
Per-splat Weights W
Refined Depths


Gaussian Decoder (§3.2)
Multi-Scale Attention
Color Encoder
Depth Encoder
Alignment Block
GS Params Head
c, s, σ, q
Confidence Head
Per-splat Weights W
Depth Refiner
Refined Depths


Gaussian Decoder (§3.2)
Multi-Scale Attention
Color Encoder
Depth Encoder
Alignment Block
GS Params Head
Confidence Head
Depth Refiner
c, s, σ, q
Per-splat Weights W
Refined Depths

Learnable Aggregation (§3.3)



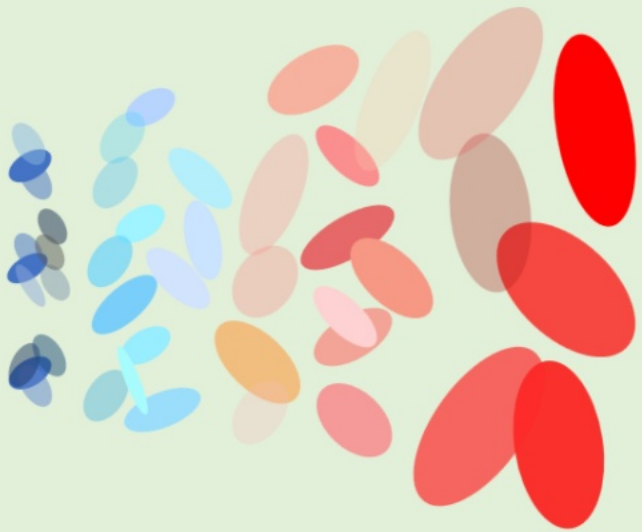
Viewing Frustum
Per-pixel Splats

Geometry and
Appearance-Aware
Clustering

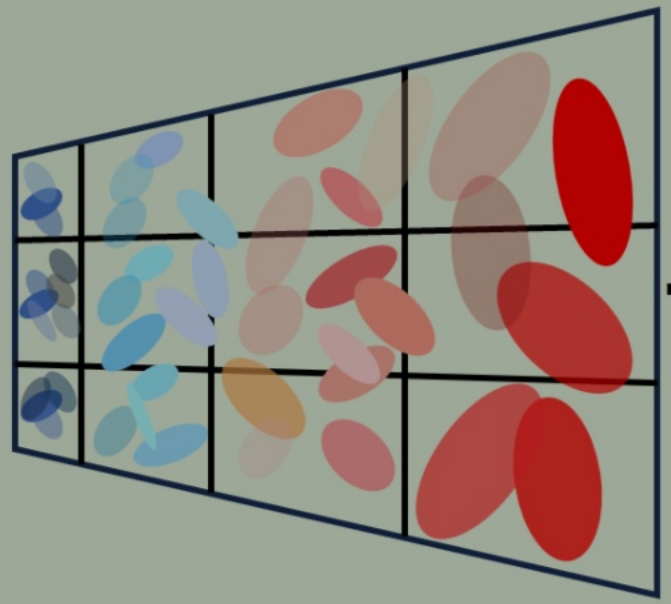
$$p_{C_i} = \sum_{j \in C_i} p_j * w_j$$

p_{j} $\in \{\mu, c, s, \sigma, q\}$

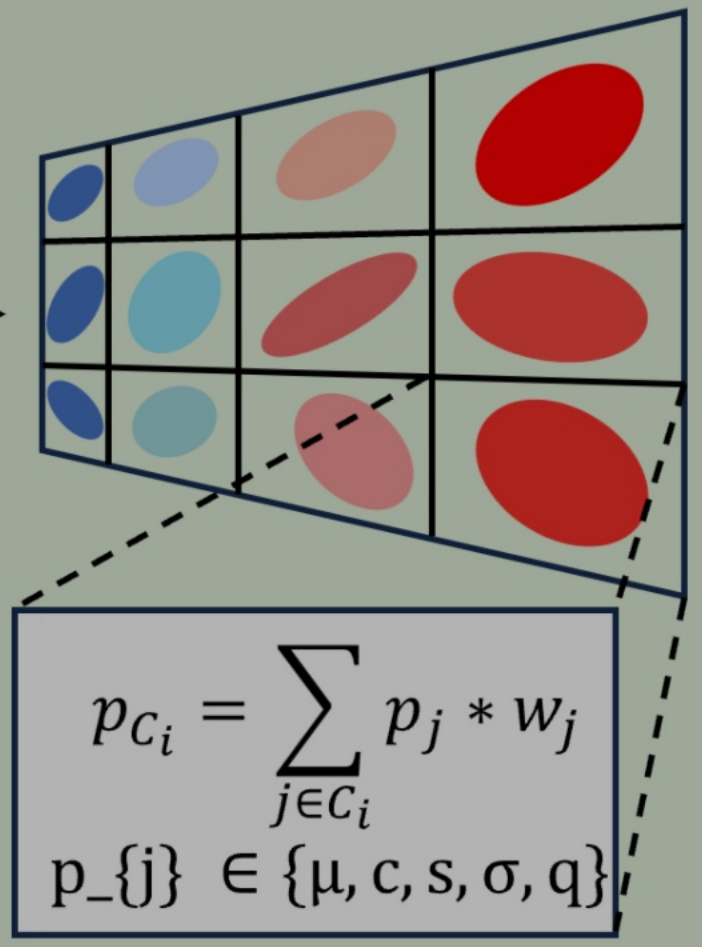
Learnable Aggregation (§3.3)



Viewing Frustum
Per-pixel Splats



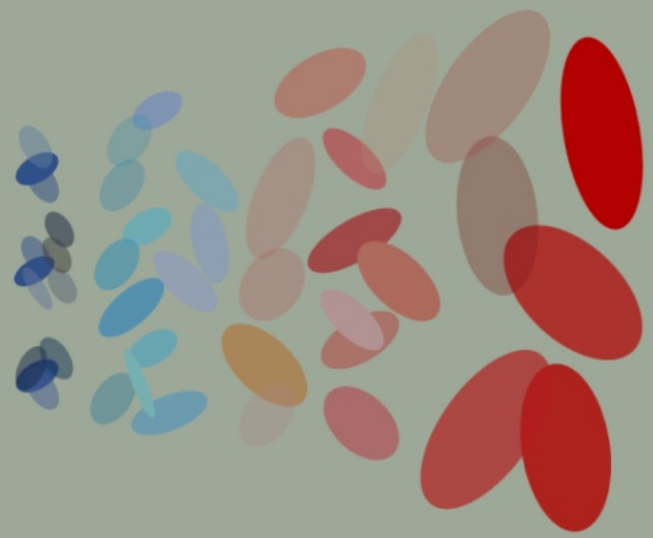
Geometry and
Appearance-Aware
Clustering



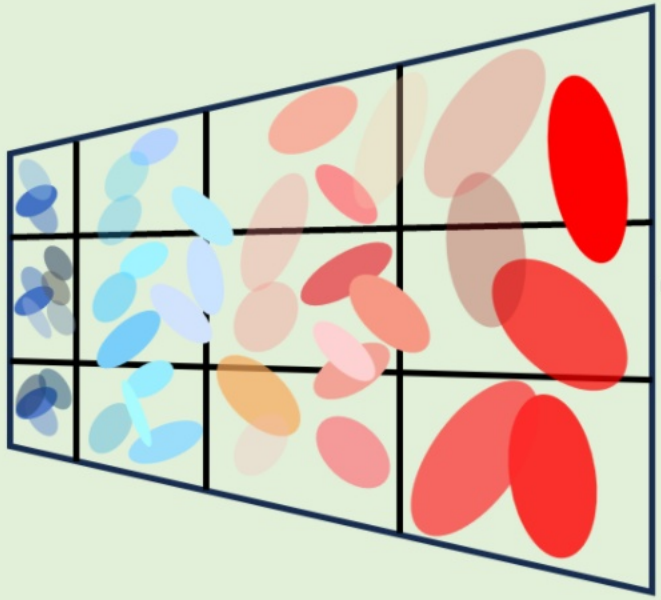
$$p_{C_i} = \sum_{j \in C_i} p_j * w_j$$

p_{j} $\in \{\mu, c, s, \sigma, q\}$

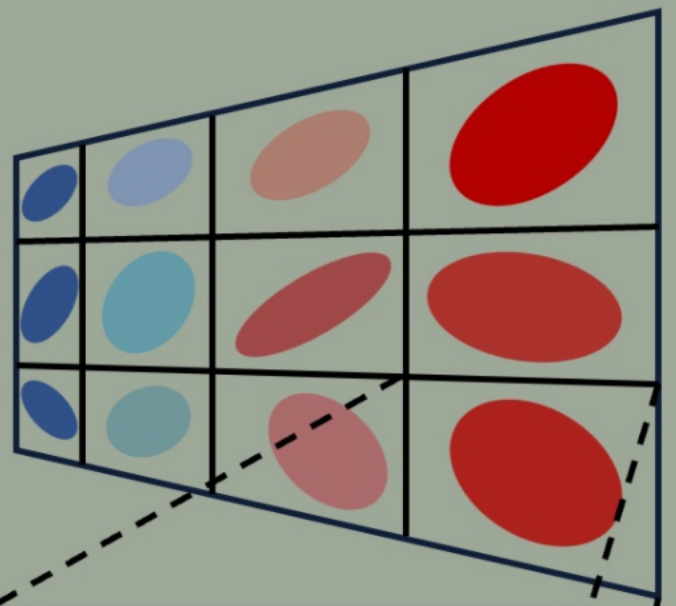
Learnable Aggregation (§3.3)



Viewing Frustum
Per-pixel Splats



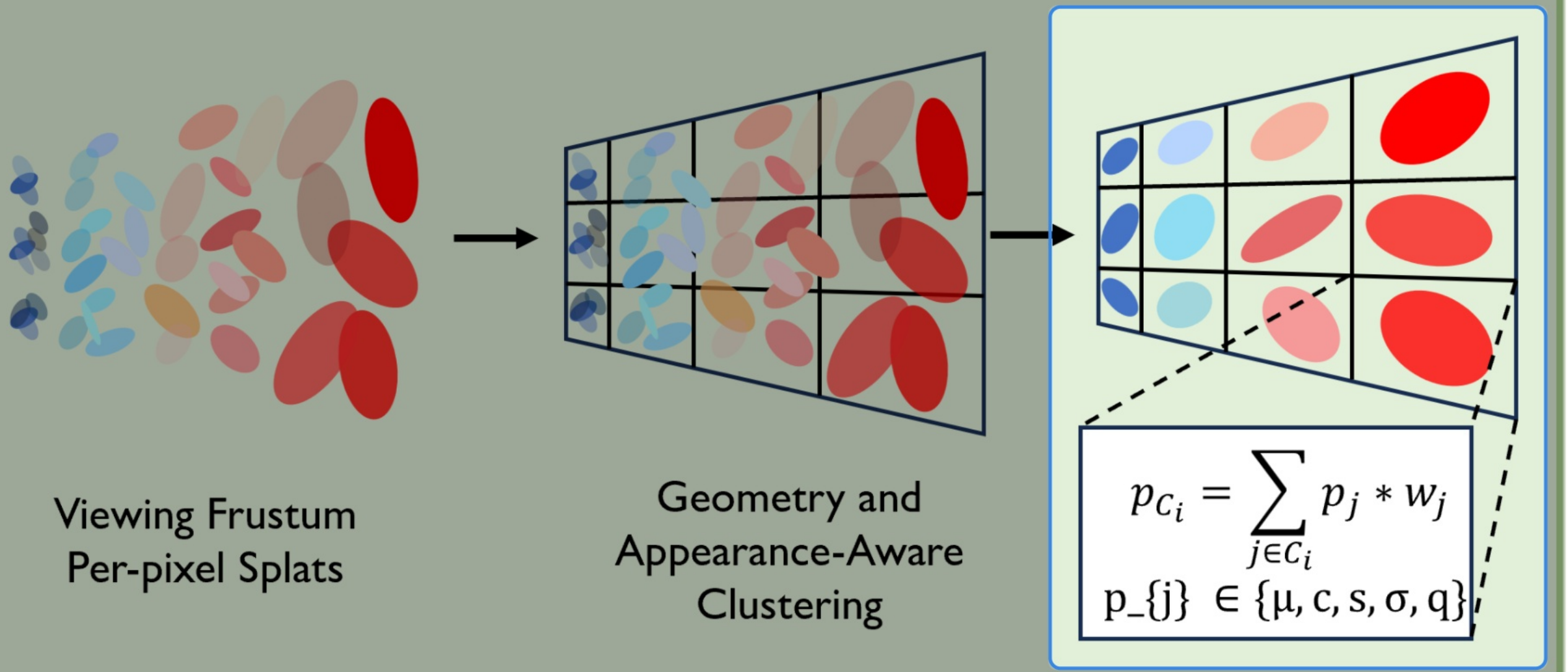
Geometry and
Appearance-Aware
Clustering



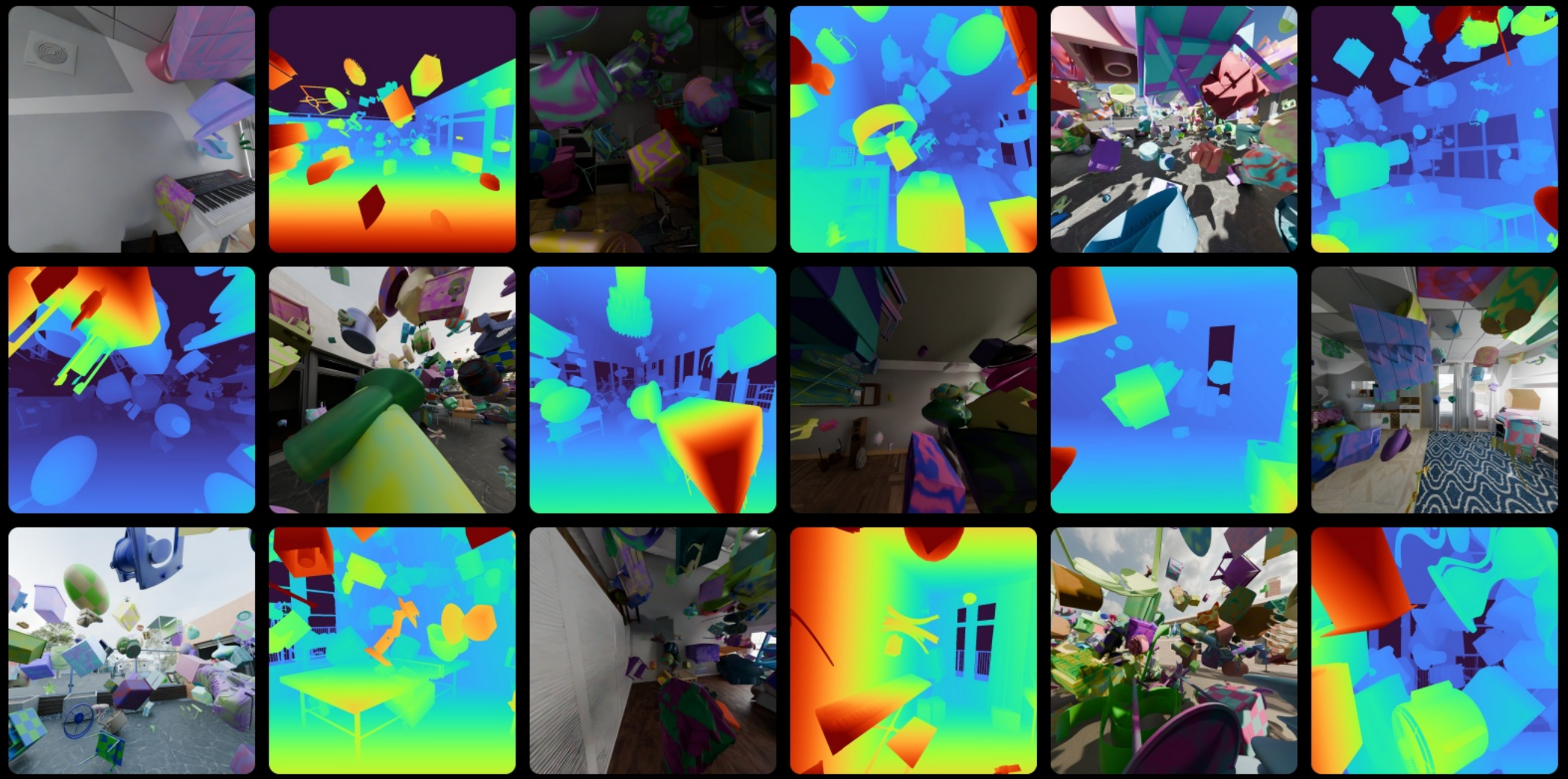
$$p_{C_i} = \sum_{j \in C_i} p_j * w_j$$

p_{j} $\in \{\mu, c, s, \sigma, q\}$

Learnable Aggregation (§3.3)



XRStereo: synthetic stereo video for training

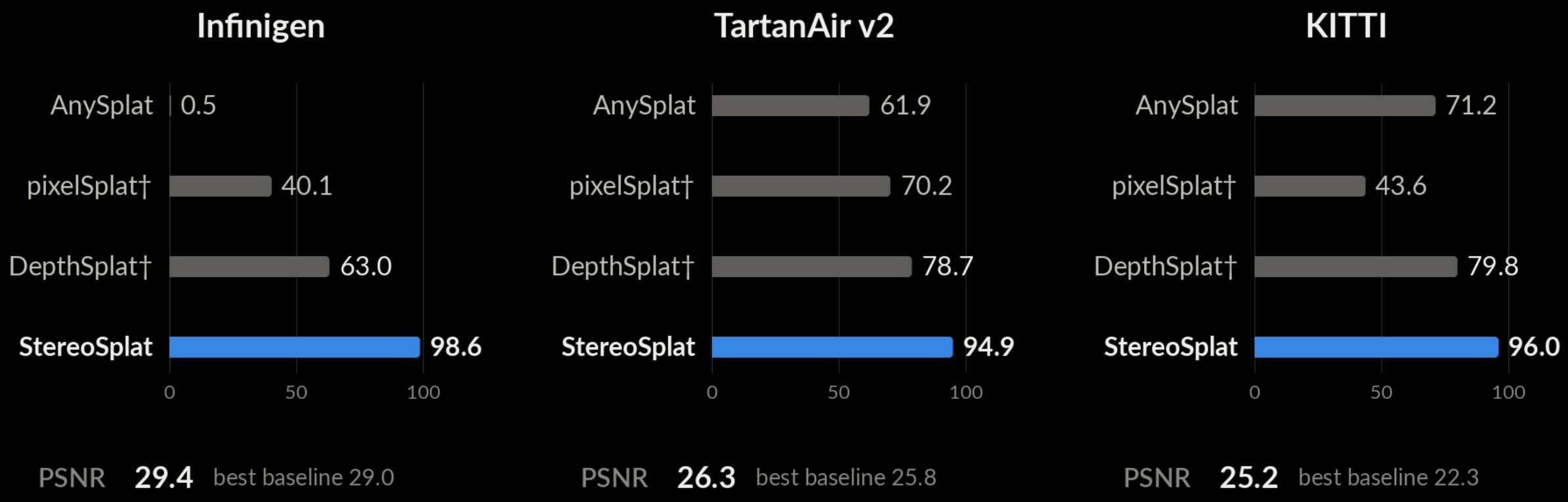


Field of view 50°–130°

Baseline 4–12 cm

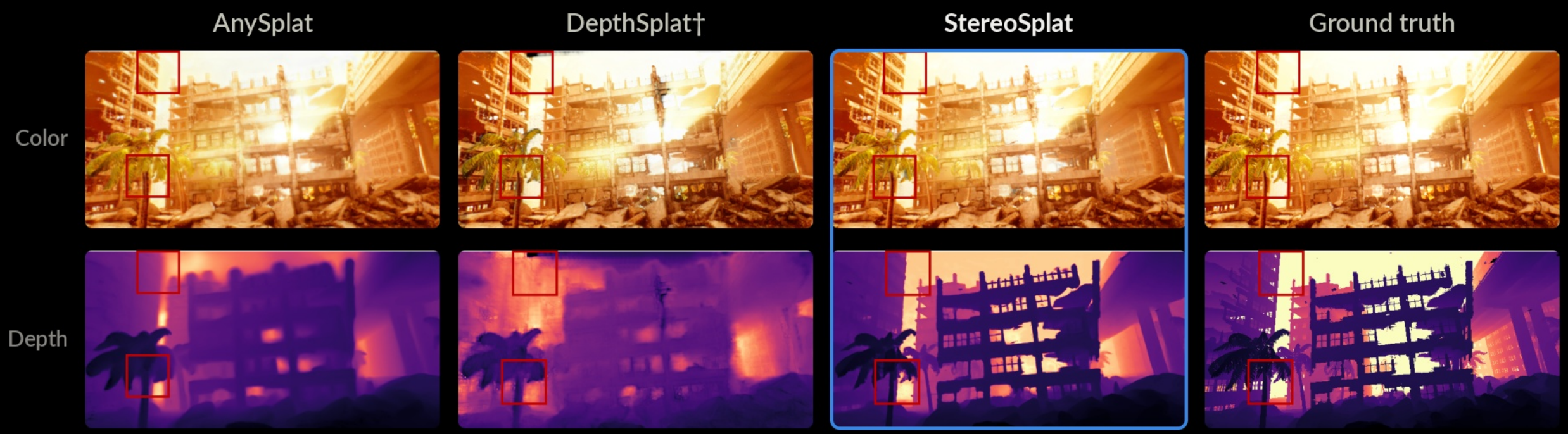
Rectified pairs + disparity

Novel-view depth accuracy, δ_1 (%)



† ground-truth poses. StereoSplat uses visual odometry poses.

Novel views on TartanAir v2



† uses ground-truth poses

What each component adds

PSNR gain over the model without it

+2.4 dB

Cross-view fusion

PSNR 22.8 → 25.2

+1.9 dB

Learned aggregation

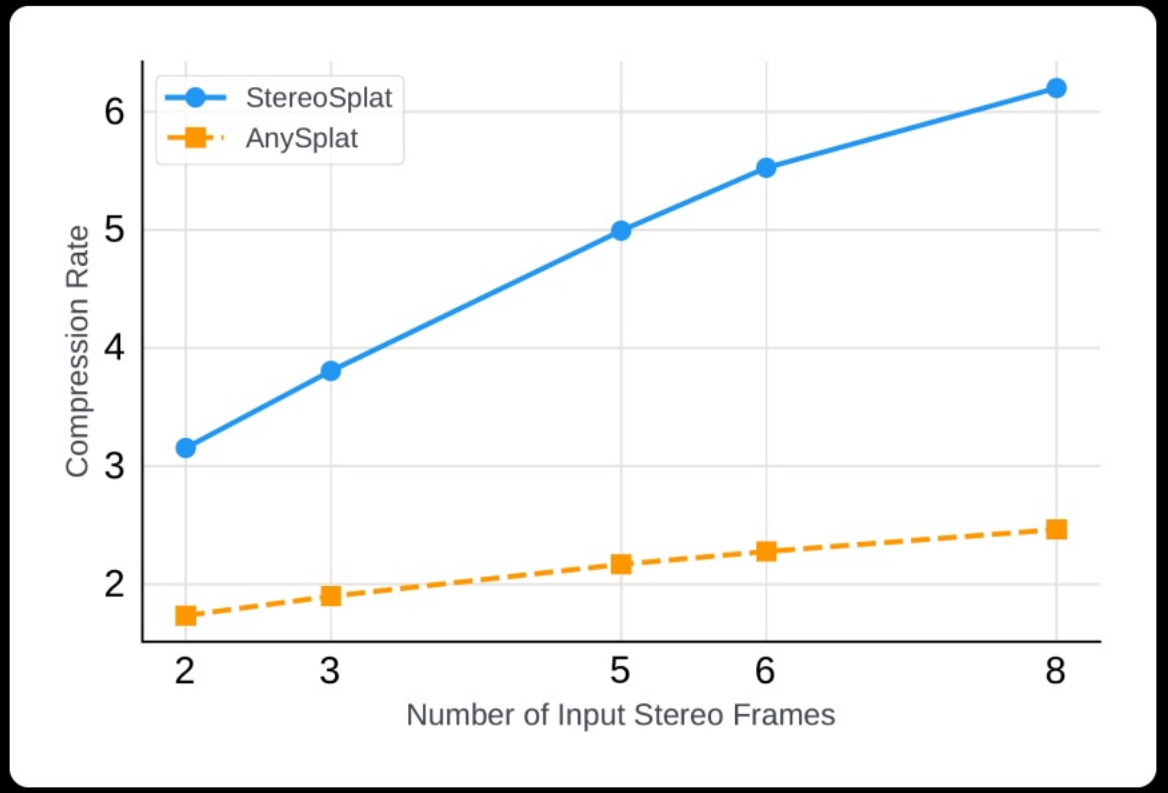
PSNR 23.3 → 25.2 · 1.6× fewer Gaussians

+1.3 dB

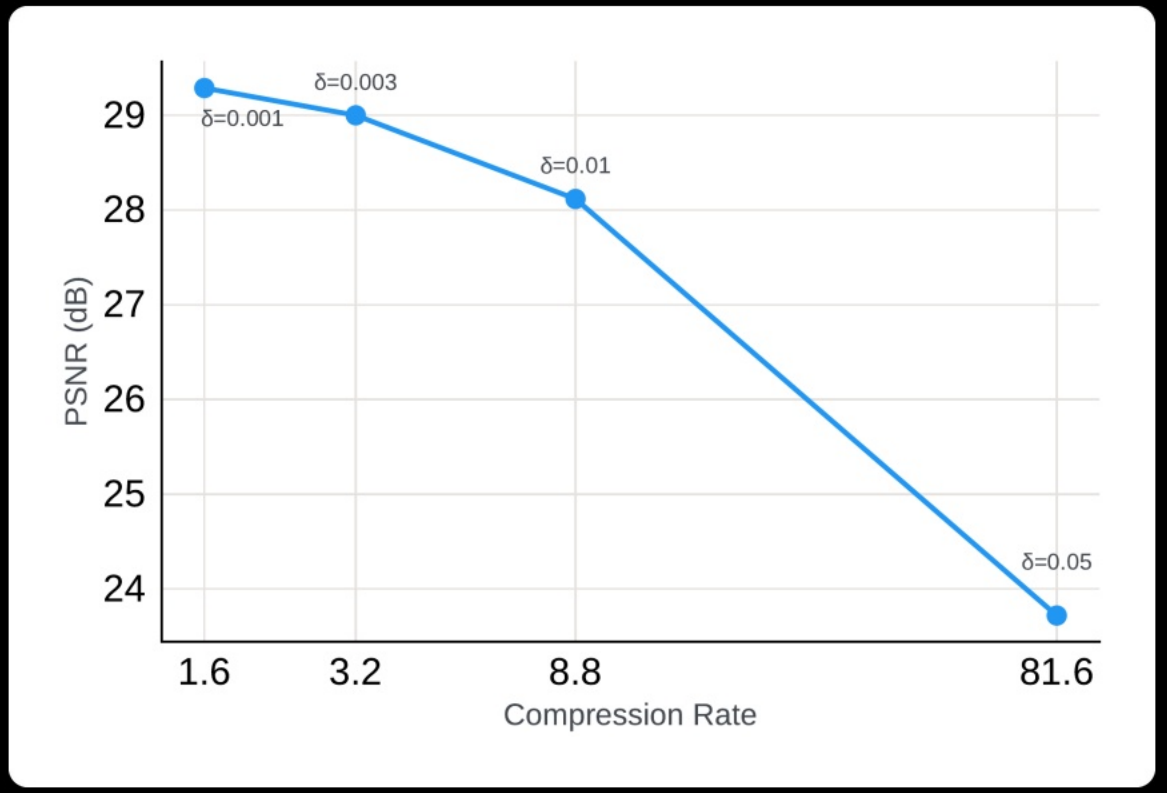
Gated depth refinement

PSNR 23.9 → 25.2

Learned aggregation keeps scenes compact



Compression rate vs. input stereo frames

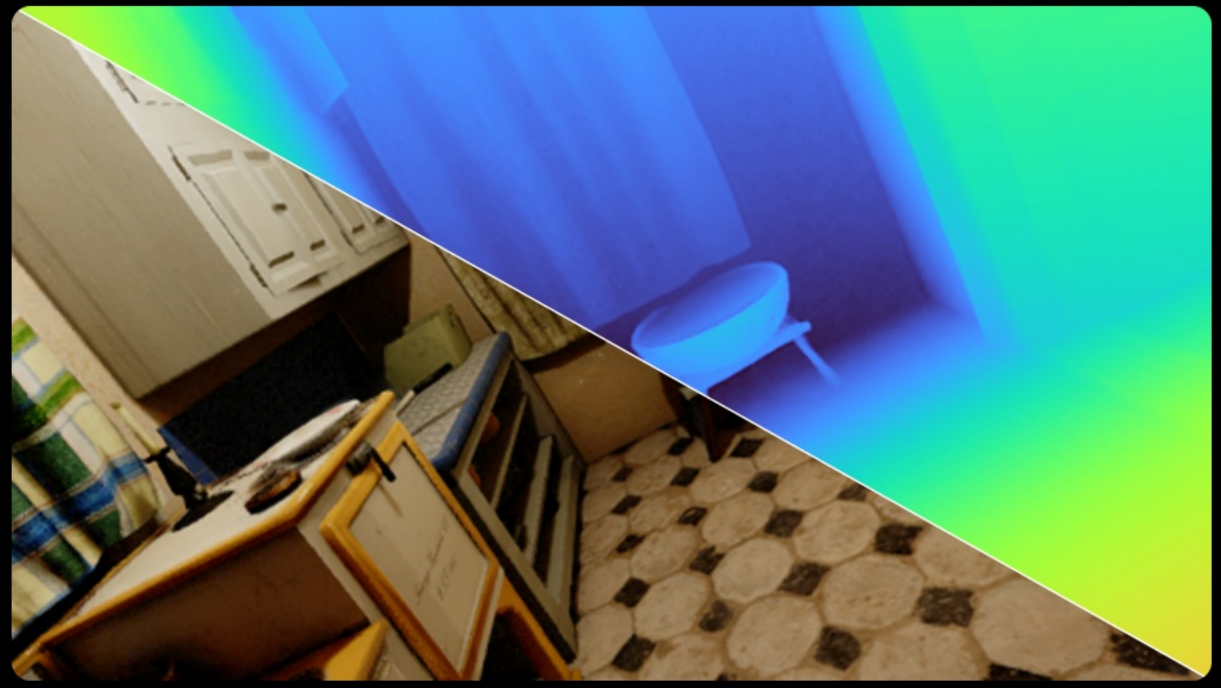


PSNR vs. compression rate (voxel size δ)

Novel views and metric depth from three stereo pairs



Outdoor: Apocalyptic



Indoor: Country House

TartanAir v2. Each half of a frame shows color, the other half rendered depth.

StereoSplat

Metric-scale novel view synthesis from stereo video

Project page	vladimiryugay.github.io/stereosplat
Code	github.com/facebookresearch/stereo-splat
Weights	huggingface.co/facebook/stereo-splat
XRStereo	huggingface.co/datasets/facebook/xrstereo

NeurIPS 2026