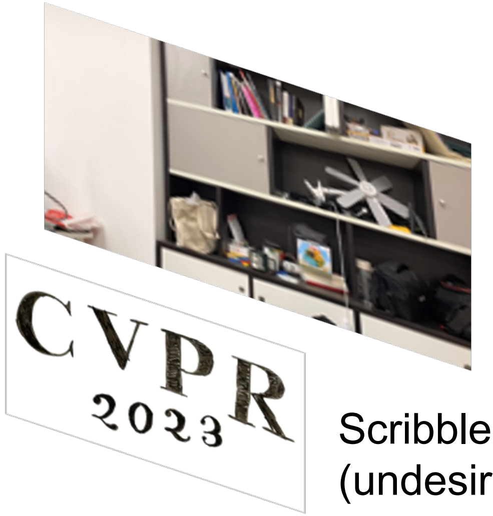
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Occlusion-Free Scene Recovery via Neural Radiance Fields

Chengxuan Zhu[1,2] Renjie Wan[3] Yunkai Tang[1,2] Boxin Shi[1,2]

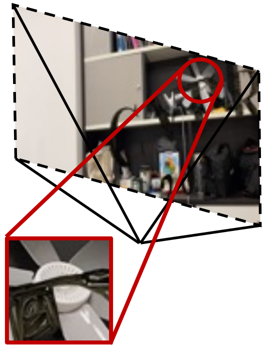
[1]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[3]Department of Computer Science, Hong Kong Baptist University

Occlusion Removal



Background (desired)

Scribbles, fences, waterdrops…
(undesired foreground occlusion)

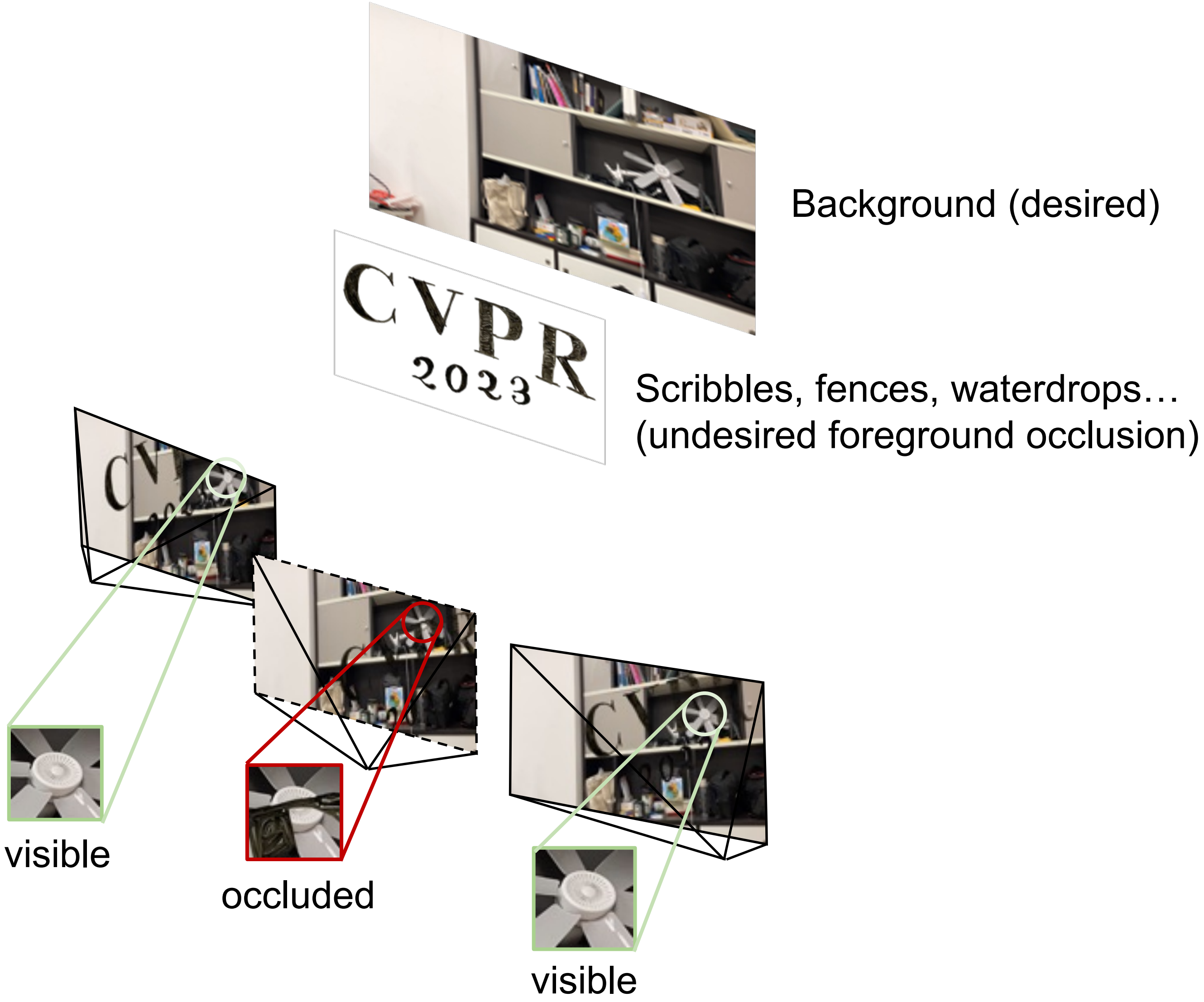


occluded

- Opaque occlusions can block useful information from reaching the camera
    - Structure-from-motion may fail
    - Occlusions cause trouble for downstream vision tasks

Occlusion-free scene representation?

Occlusion Removal



Background (desired)

Scribbles, fences, waterdrops…
(undesired foreground occlusion)

visible

occluded

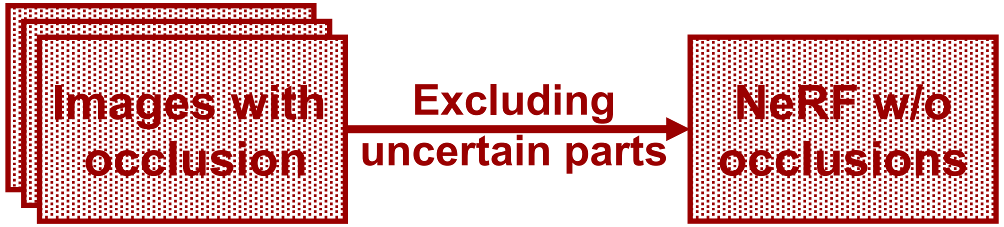
visible

- Opaque occlusions can block useful information from reaching the camera
  - Structure-from-motion may fail
  - Occlusions cause trouble for downstream vision tasks

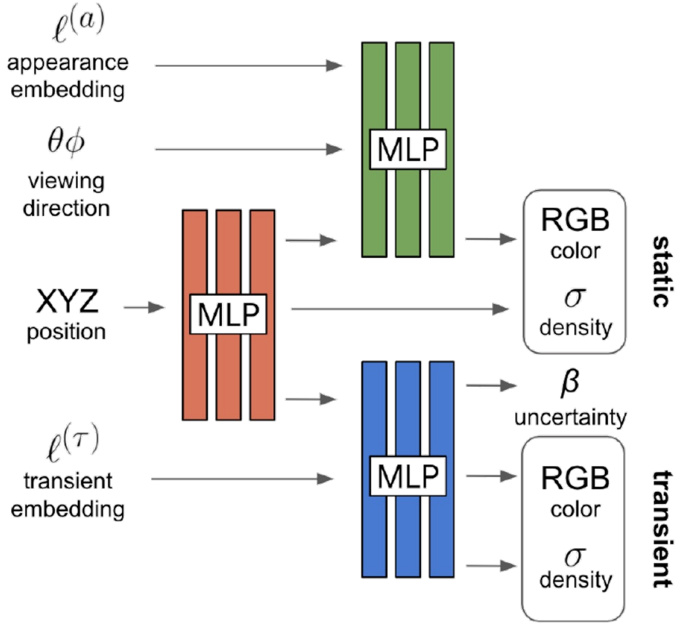
- Good news: We have multiple views!

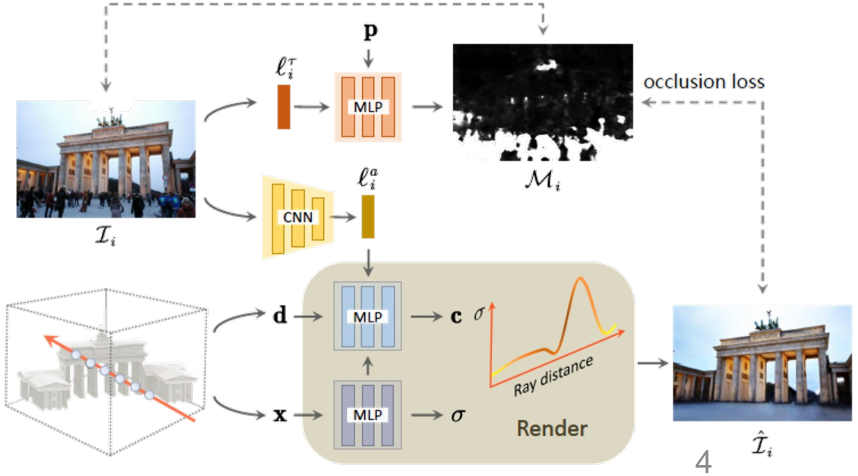
Occlusio Occlusion-free NeRF? tion?

Related Methods

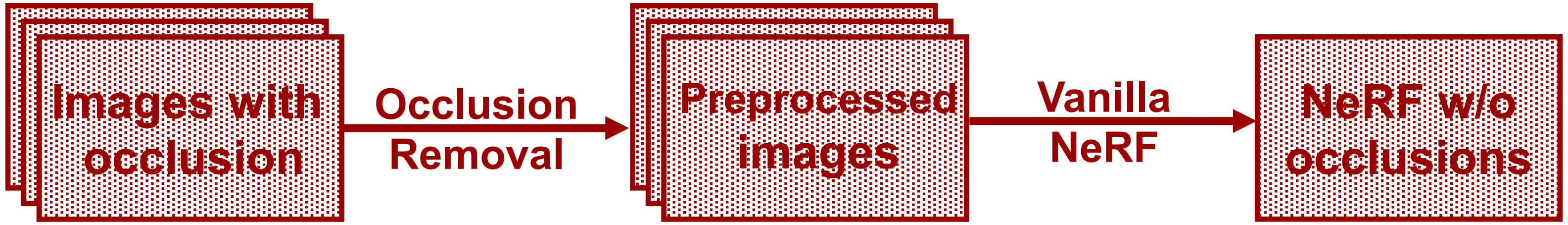


- Modeling a static scene and transient objects respectively (NeRF-W) [Martin-Brualla, CVPR’21]
- Exclude transient objects with a visible possibility map (Ha-NeRF) [Chen, CVPR’22]
- Problem: Reliant on pre-computed camera poses, and may not work with **static occlusions**





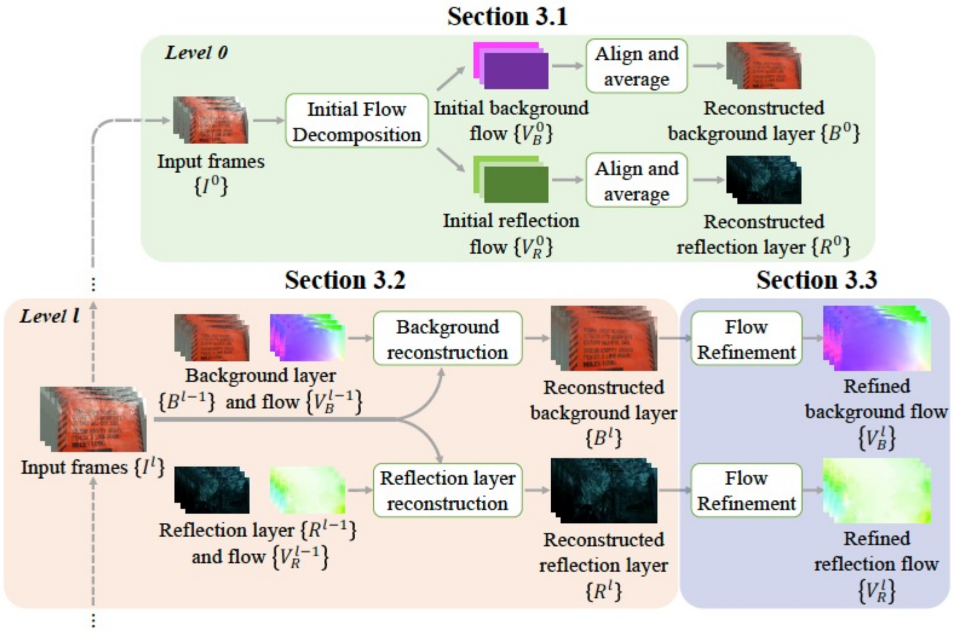
Related Methods



- Exploiting motion parallax estimated from multiple shots to separate different layers [Liu, CVPR'20]

- Inpainting based on single image is omitted due to its ill-posed nature and poor performance

- Problem: **Not 3D consistent** (NeRF may fail), and requires the input views to be close for flow estimation


Section 3.1
Level 0
Input frames $\{I^0\}$
Initial Flow Decomposition
Initial background flow $\{V_B^0\}$
Align and average
Reconstructed background layer $\{B^0\}$
Initial reflection flow $\{V_R^0\}$
Align and average
Reconstructed reflection layer $\{R^0\}$
Section 3.2
Section 3.3
Level l
Input frames $\{I^l\}$
Background layer $\{B^{l-1}\}$ and flow $\{V_B^{l-1}\}$
Background reconstruction
Reconstructed background layer $\{B^l\}$
Flow Refinement
Refined background flow $\{V_B^l\}$
Reflection layer $\{R^{l-1}\}$ and flow $\{V_R^{l-1}\}$
Reflection layer reconstruction
Reconstructed reflection layer $\{R^l\}$
Flow Refinement
Refined reflection flow $\{V_R^l\}$

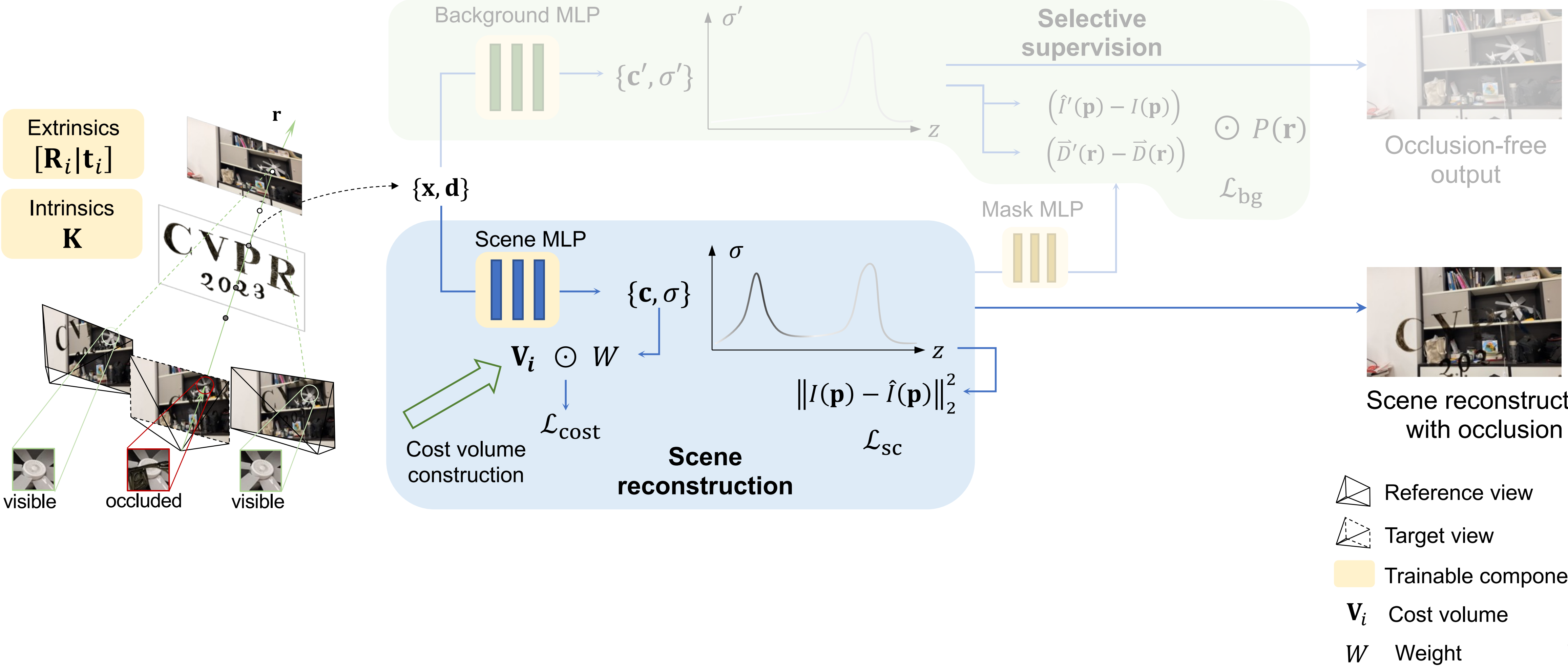
Occlusion Removal

- Joint optimization of **pose refinement** and **scene reconstruction** by effective multi-view feature fusion
- **Self-supervised** occlusion detection and occlusion-free scene recovery via NeRF

- Opaque occlusions can block useful information from reaching the camera
  - Structure-from-motion may fail
  - Occlusions cause trouble for downstream vision tasks
- Good news: We have multiple views!

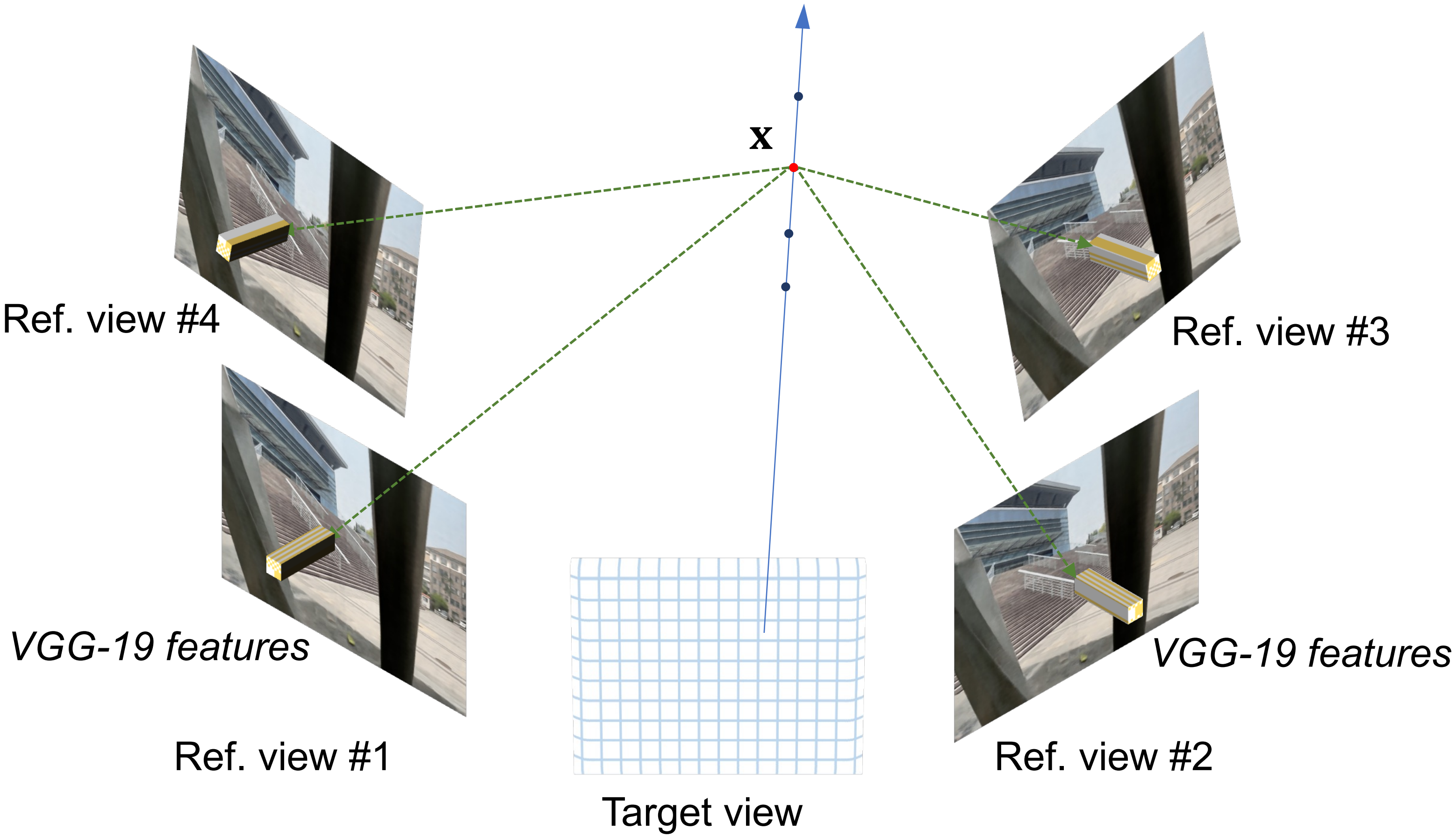
Occlusion-free NeRF?

Our Pipeline


Extrinsics
$[\mathbf{R}_i|\mathbf{t}_i]$
Intrinsics
$\mathbf{K}$
$\mathbf{r}$
visible
occluded
visible
$\{\mathbf{x}, \mathbf{d}\}$
Background MLP
$\{\mathbf{c}', \sigma'\}$
σ'
z
Selective supervision
$\left(\hat{I}'(\mathbf{p}) - I(\mathbf{p})\right)$
$\left(\vec{D}'(\mathbf{r}) - \vec{D}(\mathbf{r})\right)$
$\odot\ P(\mathbf{r})$
$\mathcal{L}_{\mathrm{bg}}$
Mask MLP
Occlusion-free output
Scene MLP
$\{\mathbf{c}, \sigma\}$
σ
z
$\mathbf{V}_i \odot W$
$\mathcal{L}_{\mathrm{cost}}$
Cost volume construction
$\|I(\mathbf{p}) - \hat{I}(\mathbf{p})\|_2^2$
$\mathcal{L}_{\mathrm{sc}}$
Scene reconstruction
Scene reconstruction with occlusion
Reference view
Target view
Trainable components
$\mathbf{V}_i$ Cost volume
W Weight

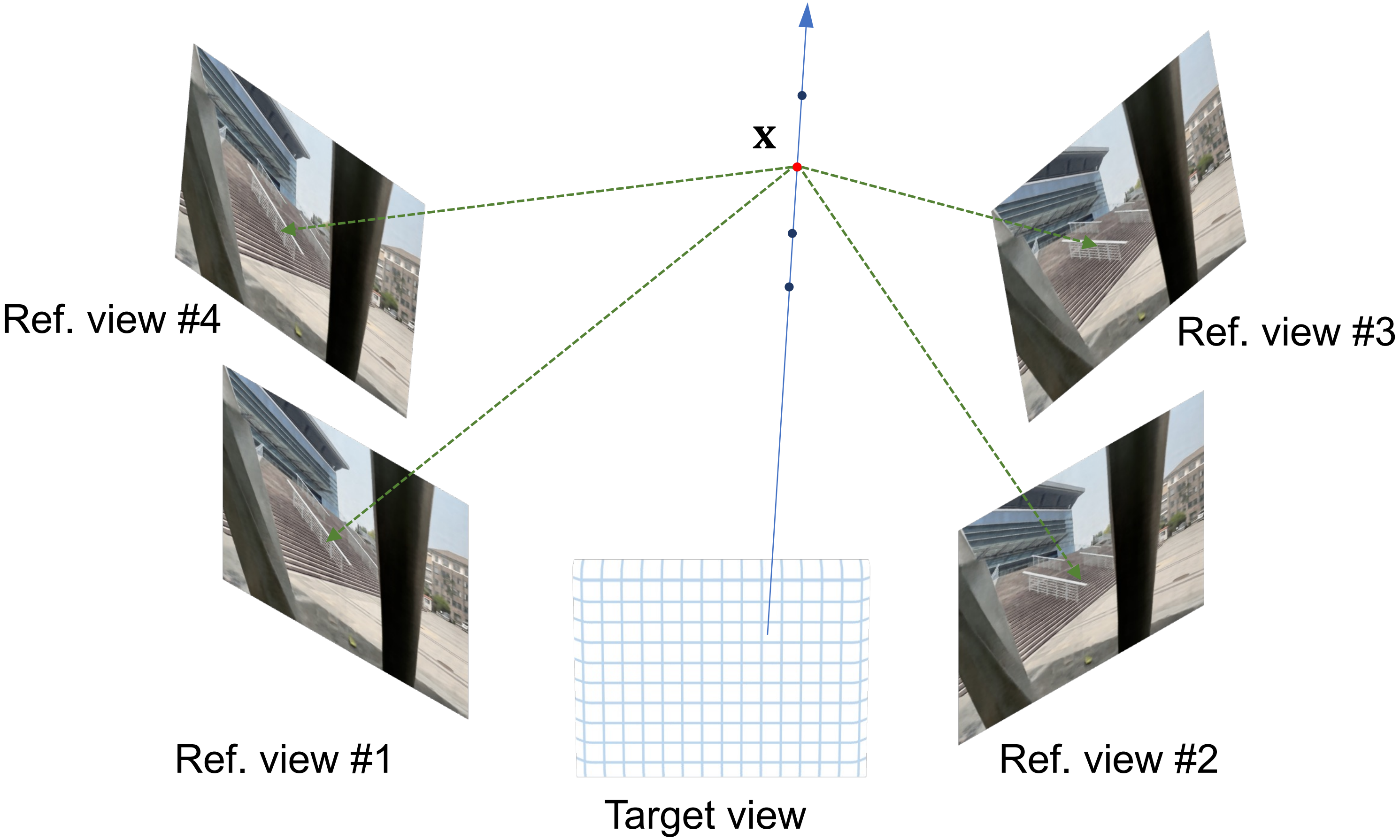
Joint Optimization

- Guided by **cost volume**



Joint Optimization

- Guided by **cost volume**

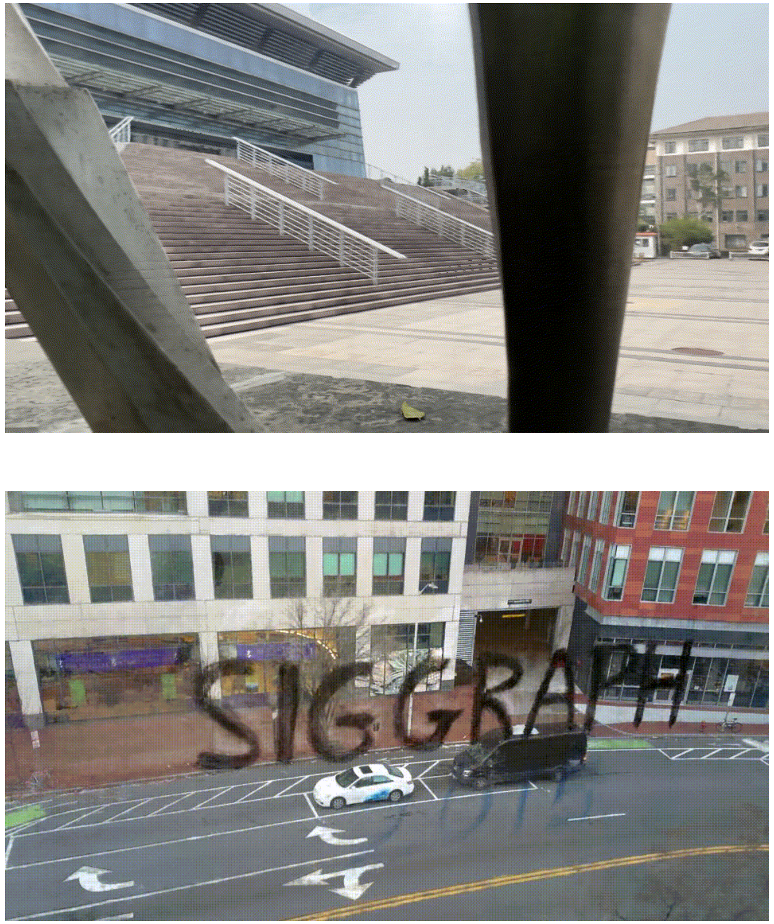


Intuition about cost volume

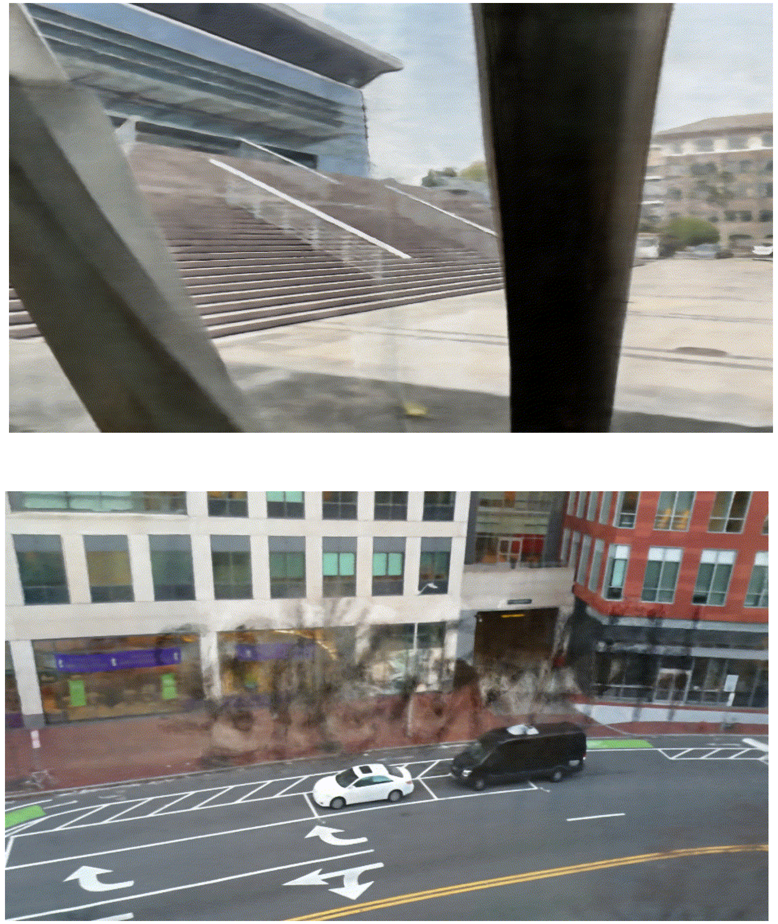
The larger Variance (▨),
the less probable a visible point is located at $\mathbf{x}$

We use a "scene MLP" to jointly reconstruct the scene and optimize camera parameters

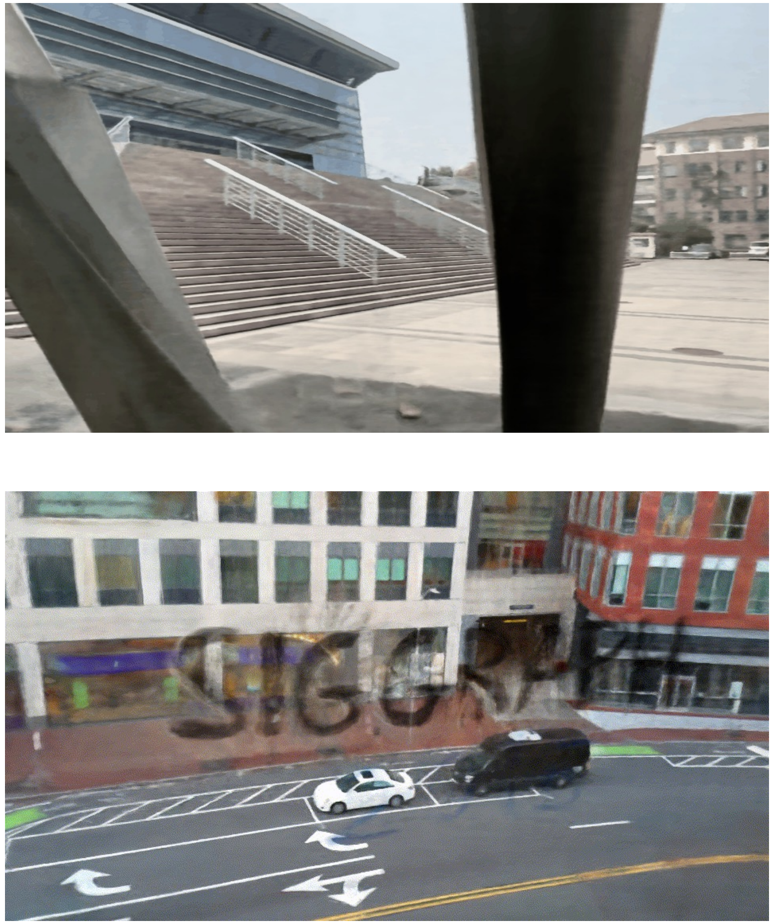
Joint Optimization



NeRF

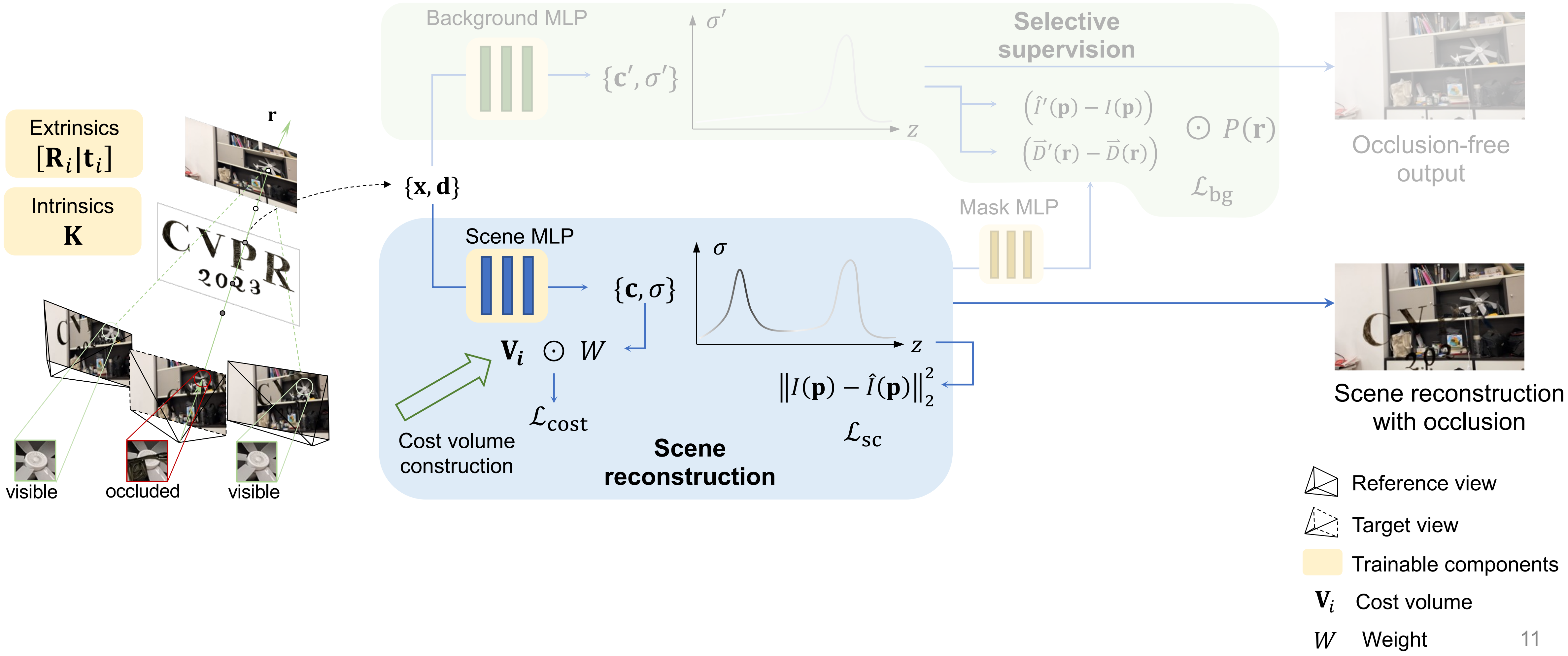


NeRF--
[Wang, ArXiv'21]

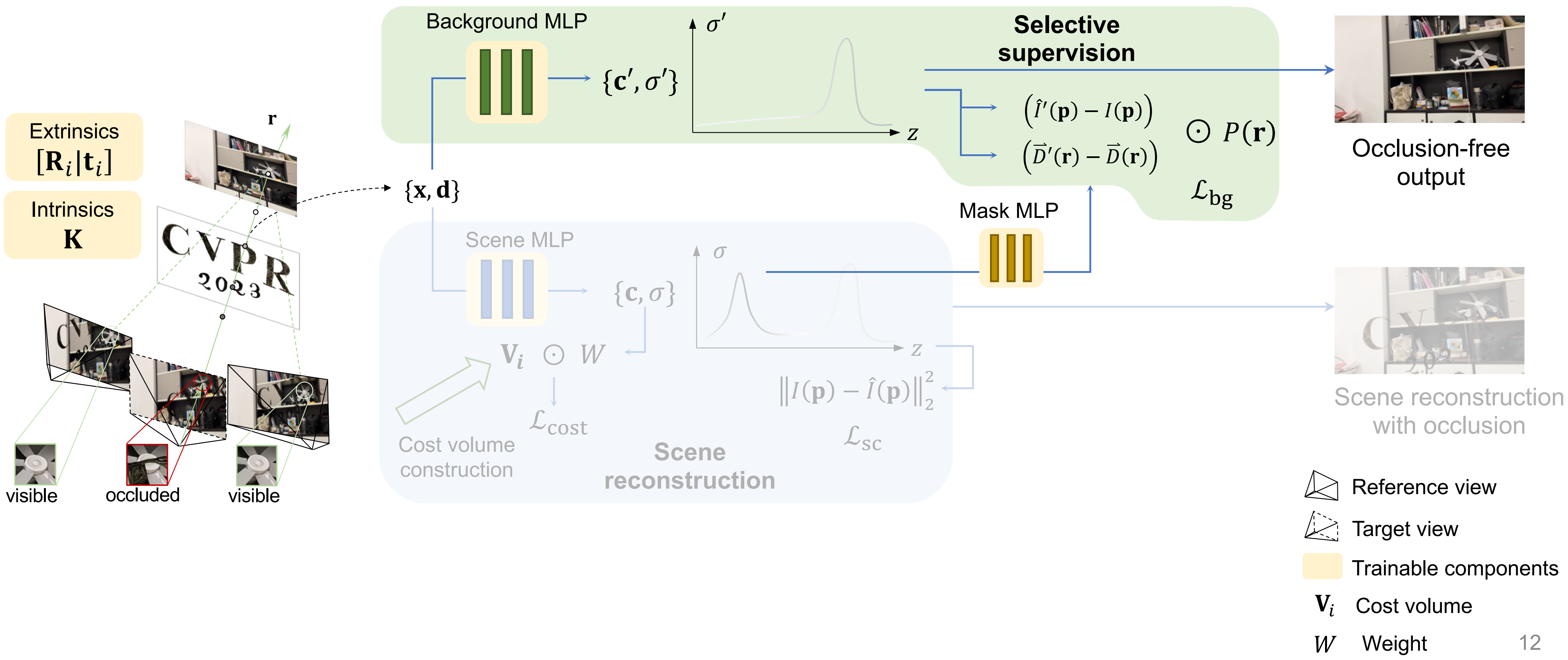


Ours, scene MLP

Our Pipeline

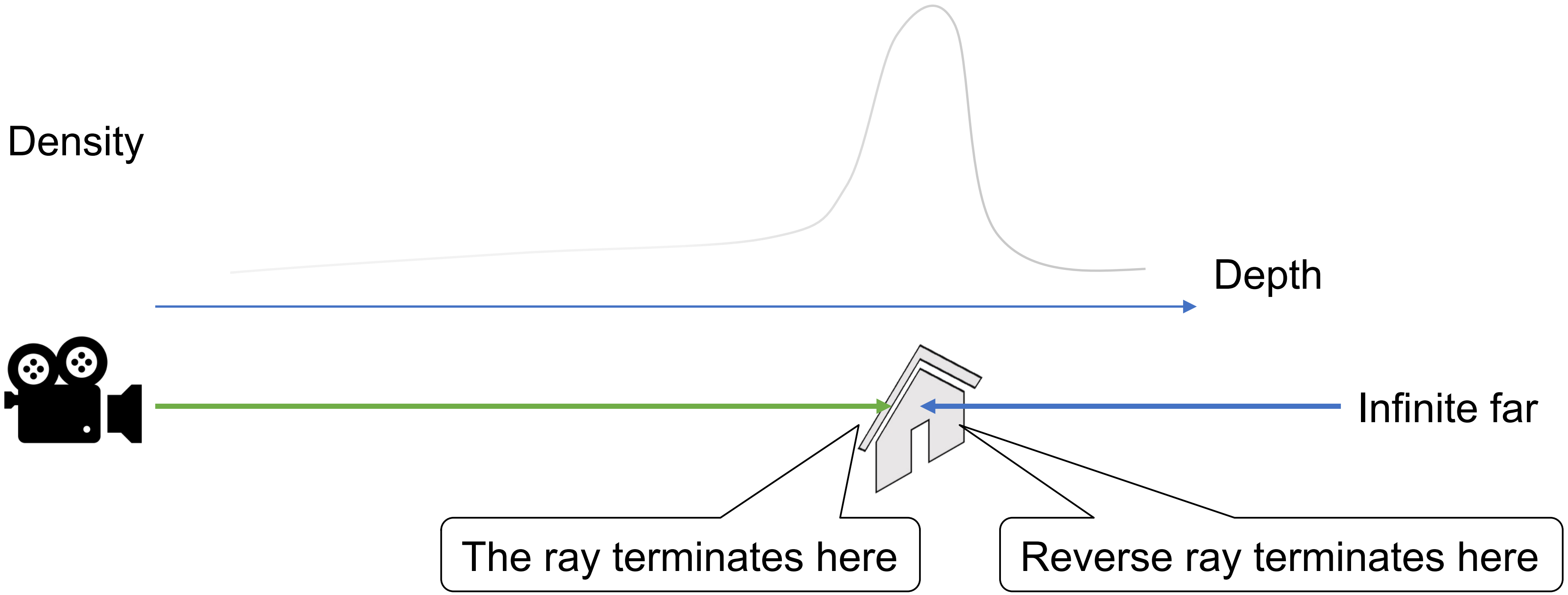

Extrinsics
$[\mathbf{R}_i|\mathbf{t}_i]$
Intrinsics
$\mathbf{K}$
$\mathbf{r}$
visible
occluded
visible
$\{\mathbf{x}, \mathbf{d}\}$
Background MLP
$\{\mathbf{c}', \sigma'\}$
σ'
z
Selective supervision
$\left(\hat{I}'(\mathbf{p}) - I(\mathbf{p})\right)$
$\left(\vec{D}'(\mathbf{r}) - \vec{D}(\mathbf{r})\right)$
$\odot\ P(\mathbf{r})$
$\mathcal{L}_{\mathrm{bg}}$
Mask MLP
Occlusion-free output
Scene MLP
$\{\mathbf{c}, \sigma\}$
σ
z
$\mathbf{V}_i \ \odot \ W$
$\mathcal{L}_{\mathrm{cost}}$
Cost volume construction
$\left\|I(\mathbf{p}) - \hat{I}(\mathbf{p})\right\|_2^2$
$\mathcal{L}_{\mathrm{sc}}$
Scene reconstruction
Scene reconstruction with occlusion
Reference view
Target view
Trainable components
$\mathbf{V}_i$ Cost volume
W Weight

Our Pipeline


Extrinsics
$[\mathbf{R}_i|\mathbf{t}_i]$
Intrinsics
$\mathbf{K}$
$\mathbf{r}$
visible
occluded
visible
$\{\mathbf{x},\mathbf{d}\}$
Background MLP
$\{\mathbf{c}',\sigma'\}$
σ'
z
Selective supervision
$\big(\hat{I}'(\mathbf{p})-I(\mathbf{p})\big)$
$\big(\vec{D}'(\mathbf{r})-\vec{D}(\mathbf{r})\big)$
$\odot\ P(\mathbf{r})$
$\mathcal{L}_{\text{bg}}$
Mask MLP
Occlusion-free output
Scene MLP
$\{\mathbf{c},\sigma\}$
σ
z
$\mathbf{V}_i\ \odot\ W$
$\mathcal{L}_{\text{cost}}$
Cost volume construction
$\|I(\mathbf{p})-\hat{I}(\mathbf{p})\|_2^2$
$\mathcal{L}_{\text{sc}}$
Scene reconstruction
Scene reconstruction with occlusion
Reference view
Target view
Trainable components
$\mathbf{V}_i$ Cost volume
W Weight

Selective Supervision

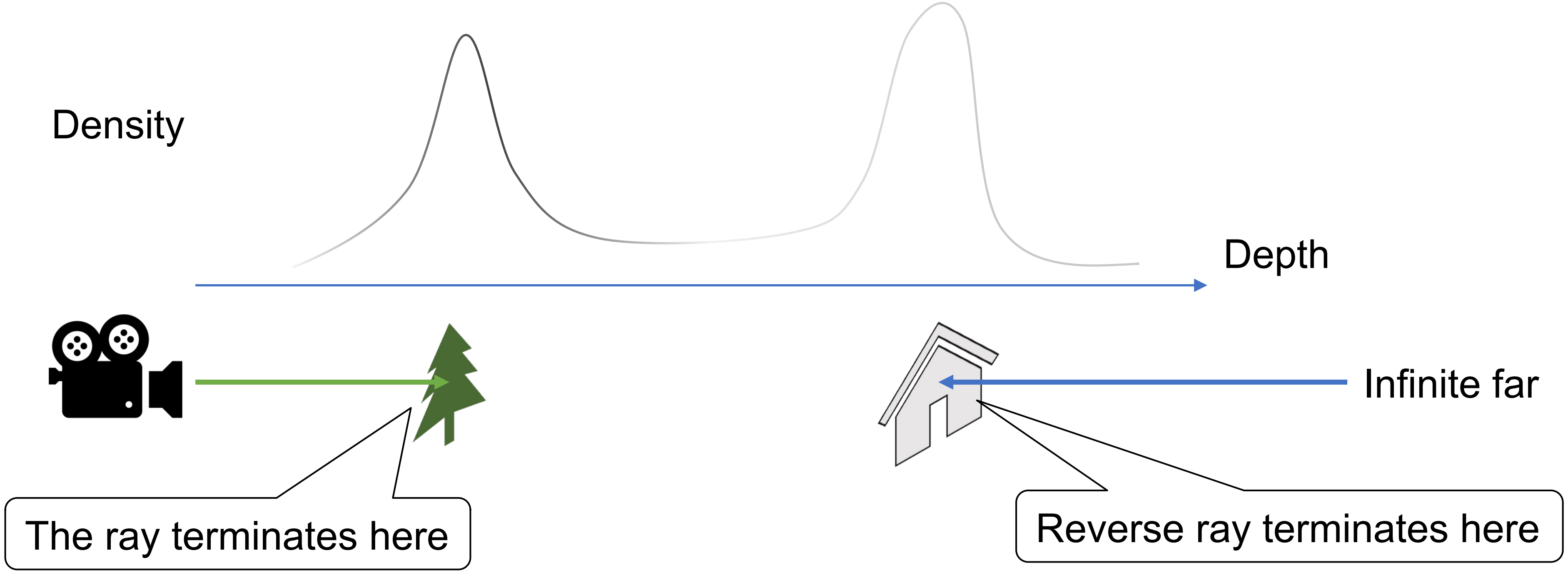
- Guided by **bidirectional depth inconsistency**



When bidirectional depth inconsistency is ***small***, there is no occlusion.

Selective Supervision

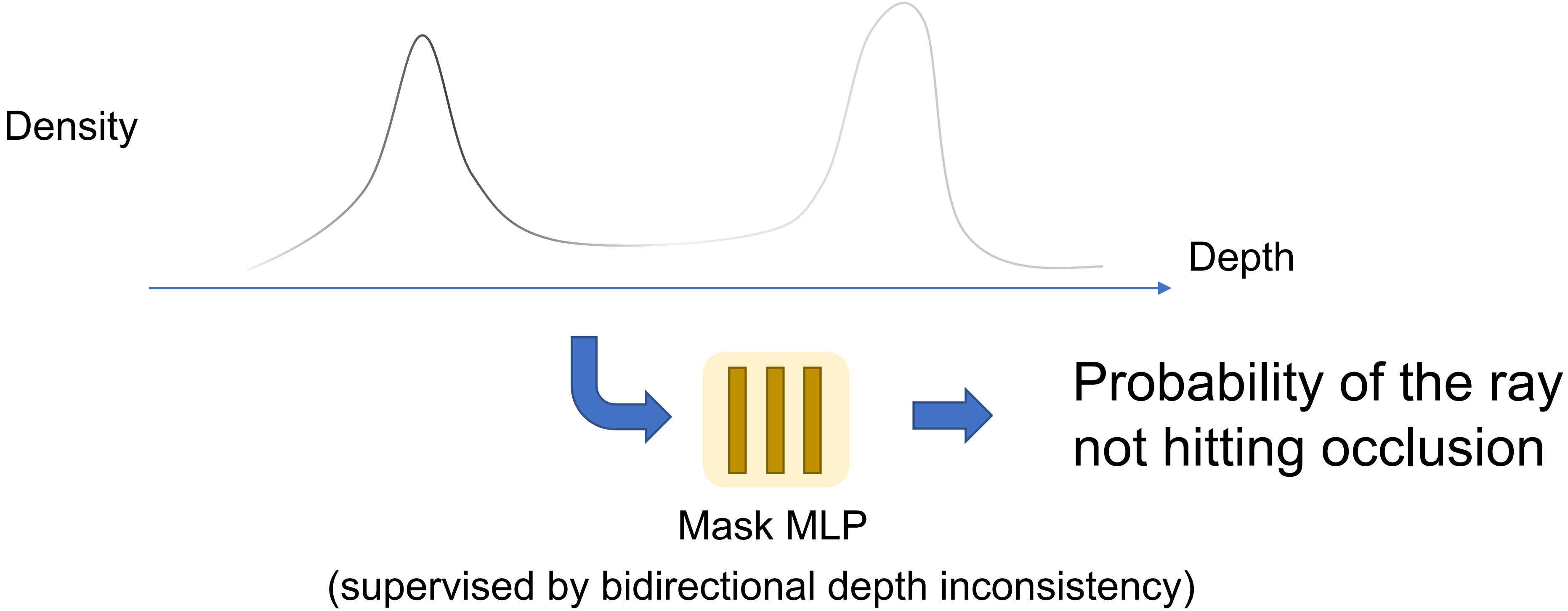
- Guided by **bidirectional depth inconsistency**



When bidirectional depth inconsistency is ***large***, there is occlusion.

Selective Supervision

- Guided by **bidirectional depth inconsistency**



Density

Depth

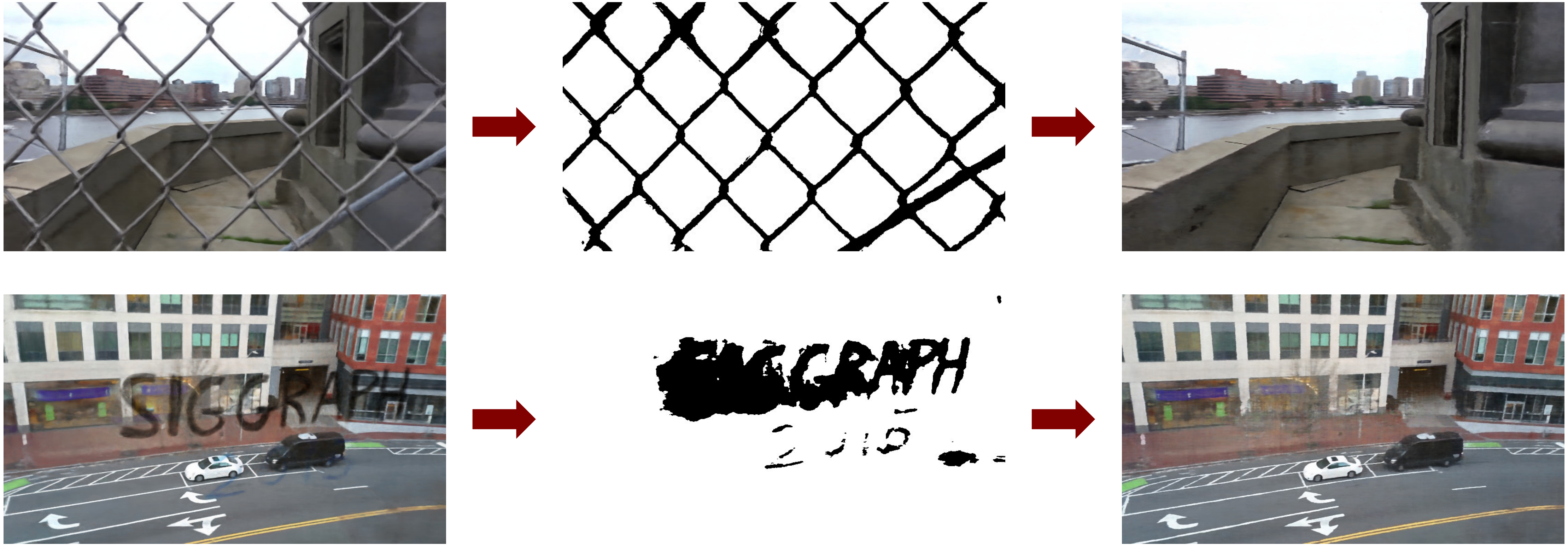
Mask MLP

(supervised by bidirectional depth inconsistency)

Probability of the ray not hitting occlusion

Selective Supervision

- Supervise the background MLP only where there is no occlusion

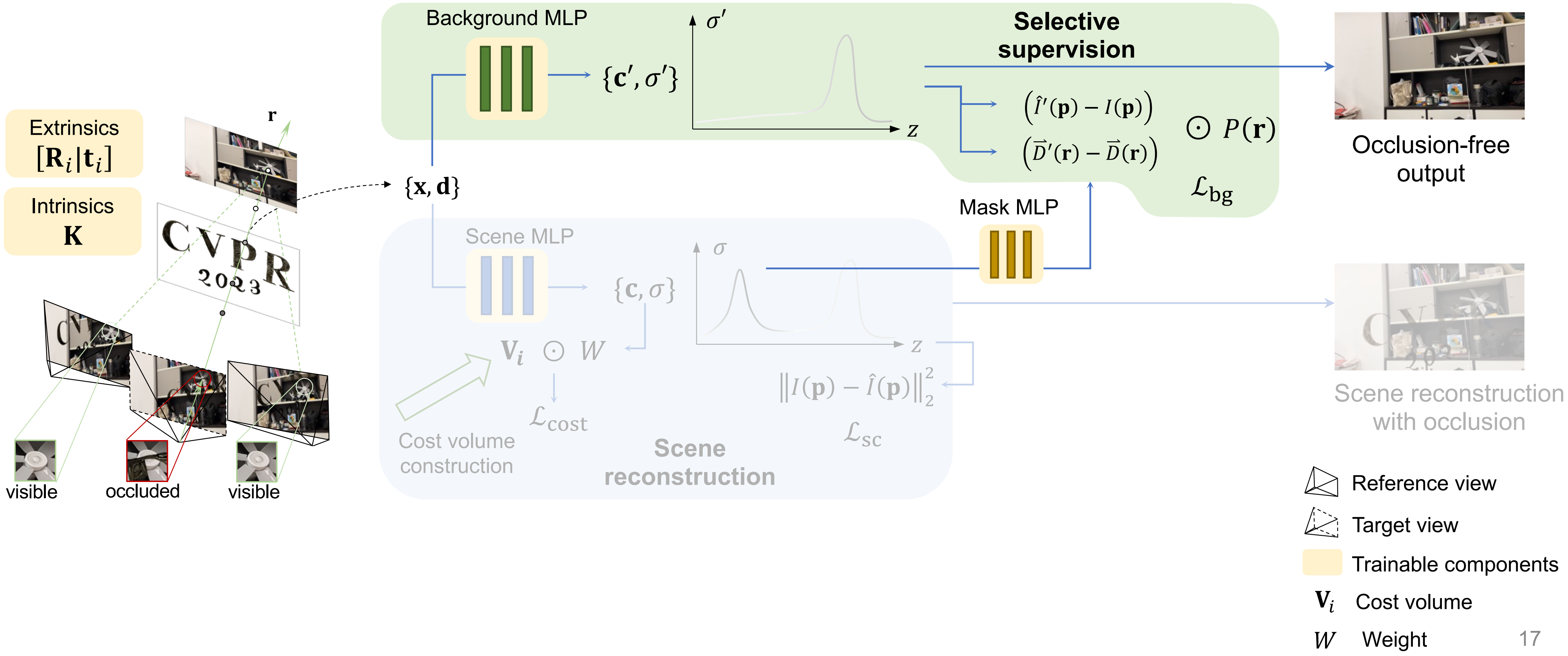


Input view

Mask MLP output

Ours, background MLP

Our pipeline



Extrinsics $[\mathbf{R}_i|\mathbf{t}_i]$

Intrinsics $\mathbf{K}$

$\mathbf{r}$

visible

occluded

visible

$\{\mathbf{x}, \mathbf{d}\}$

Background MLP

$\{\mathbf{c}', \sigma'\}$

σ'

z

Selective supervision

$\left(\hat{I}'(\mathbf{p}) - I(\mathbf{p})\right)$

$\left(\vec{D}'(\mathbf{r}) - \vec{D}(\mathbf{r})\right)$

$\odot \; P(\mathbf{r})$

$\mathcal{L}_{\text{bg}}$

Mask MLP

Occlusion-free output

Scene MLP

$\{\mathbf{c}, \sigma\}$

σ

z

$\mathbf{V}_i \odot W$

$\mathcal{L}_{\text{cost}}$

Cost volume construction

$\left\|I(\mathbf{p}) - \hat{I}(\mathbf{p})\right\|_2^2$

$\mathcal{L}_{\text{sc}}$

Scene reconstruction

Scene reconstruction with occlusion

Reference view

Target view

Trainable components

$\mathbf{V}_i$ Cost volume

W Weight

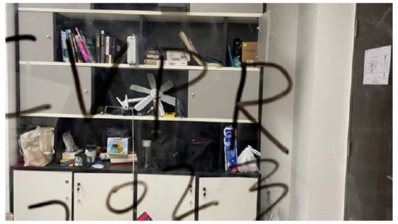
Dataset

- An evaluation dataset containing 10 different scenes
- Covering various types of occlusions

(a) FENCE1

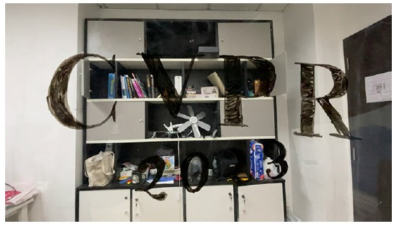
(b) FENCE2

(c) FENCE3



(d) SCRIBBLE1

(e) SCRIBBLE2

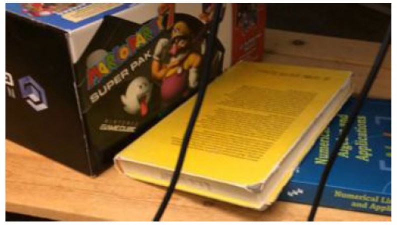


(f) SCRIBBLE3

(g) RAINDROP

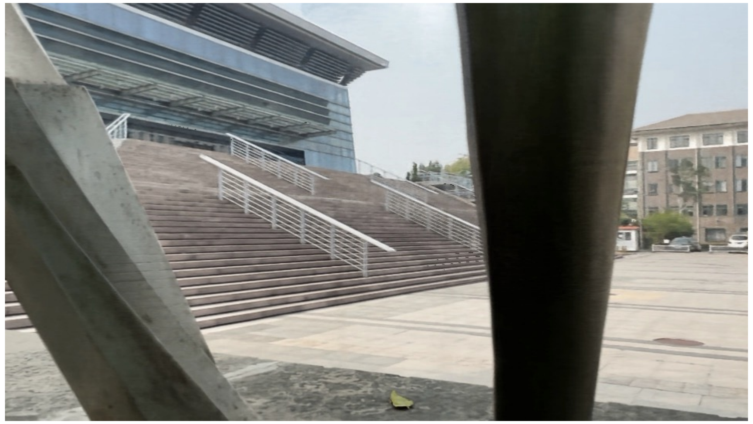
(h) STATUE

(i) WIRE1

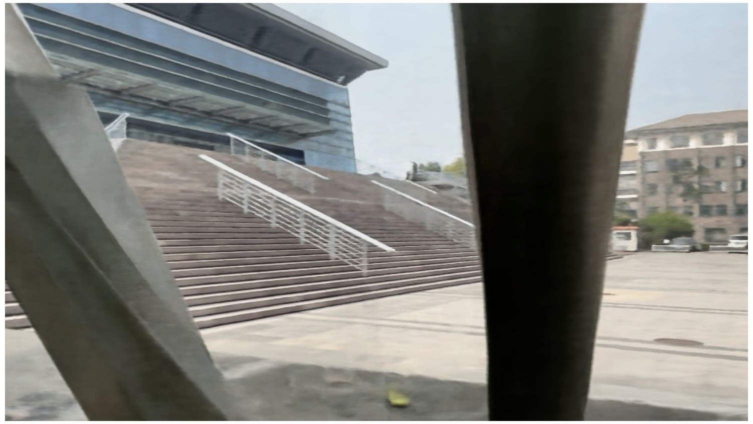


(j) WIRE2

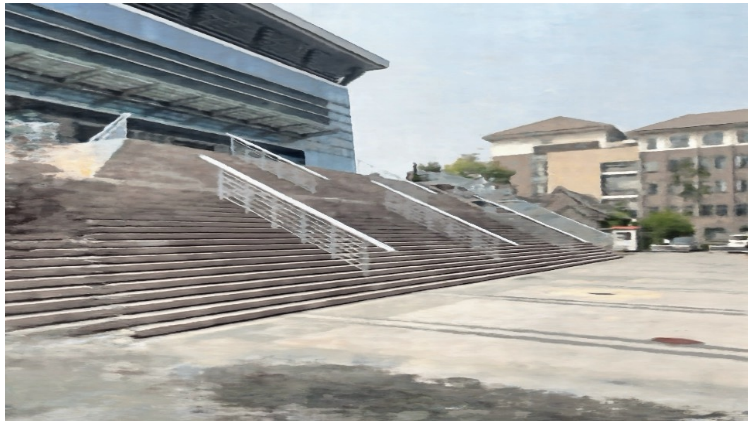
Results



NeRF

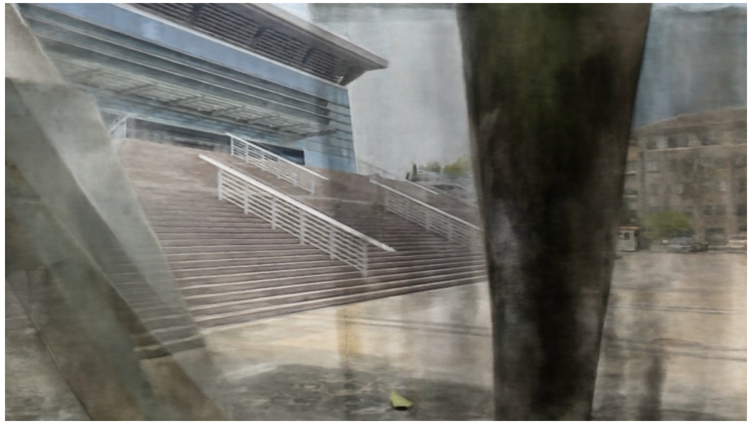


Ours, Scene MLP

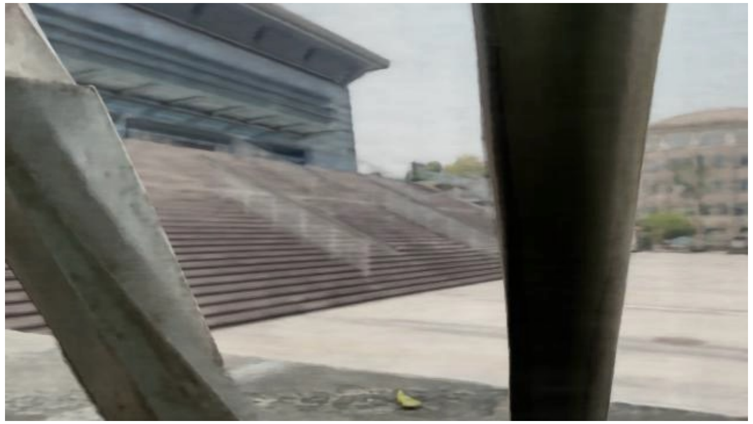


Ours, Background MLP

Ha-NeRF

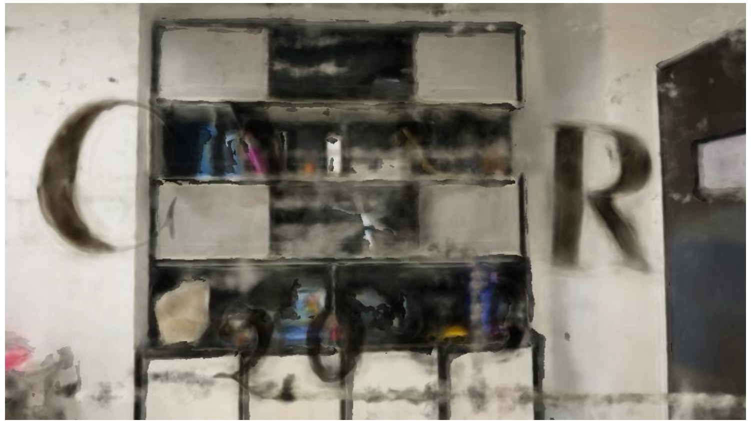


PWC-Net + NeRF

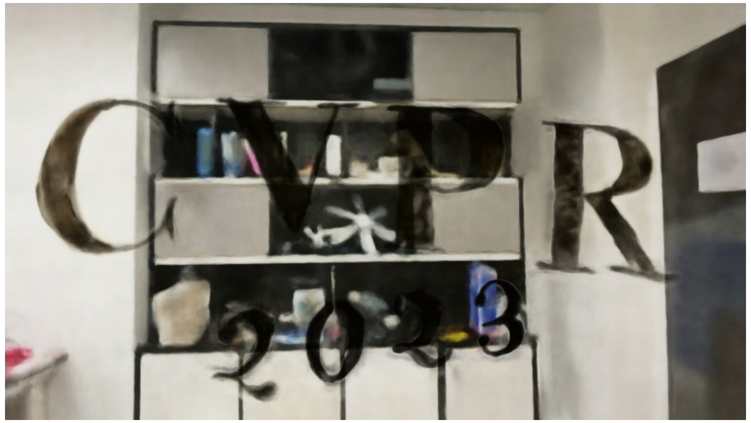


NeRF-W

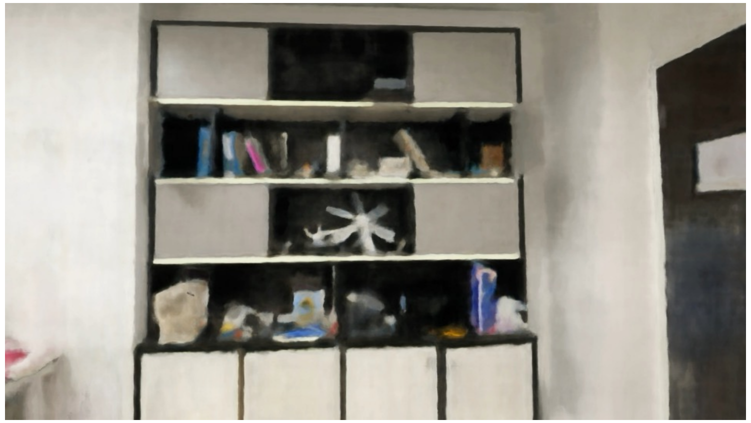
Results



NeRF



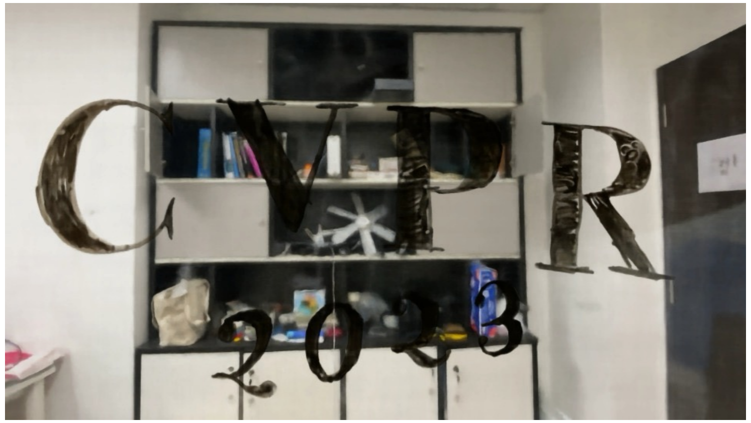
Ours, Scene MLP



Ours, Background MLP

Ha-NeRF

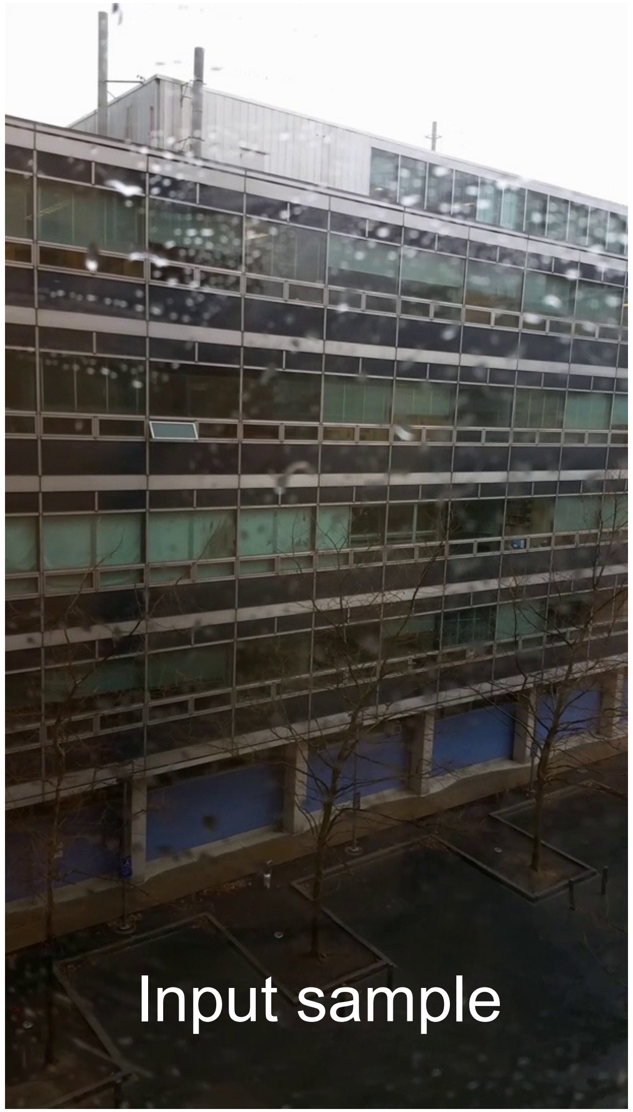
PWC-Net + NeRF



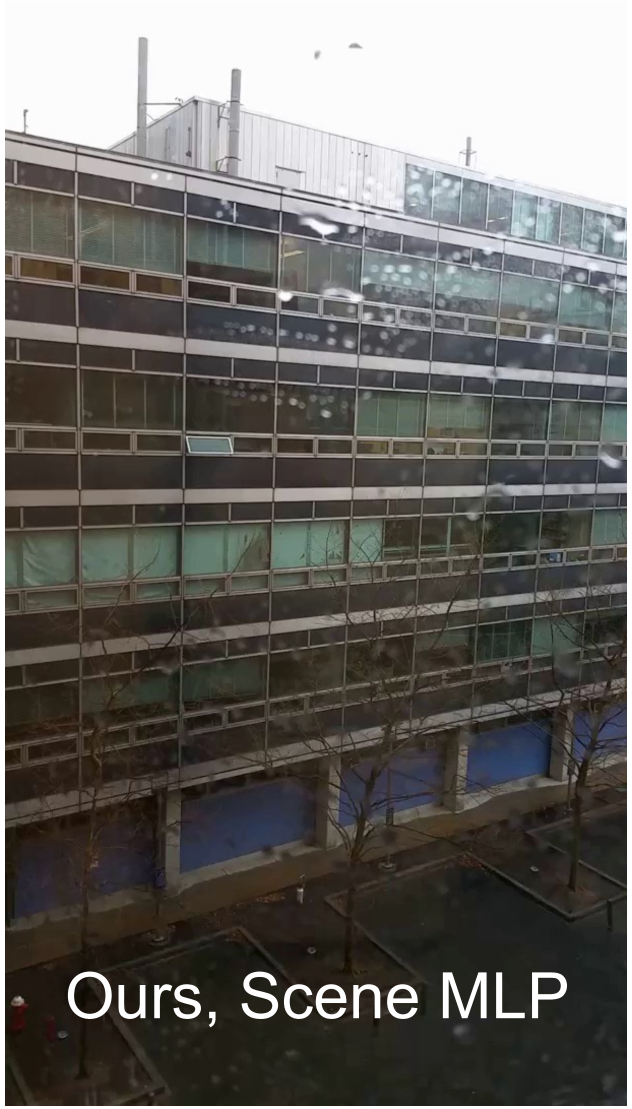
NeRF-W

Results

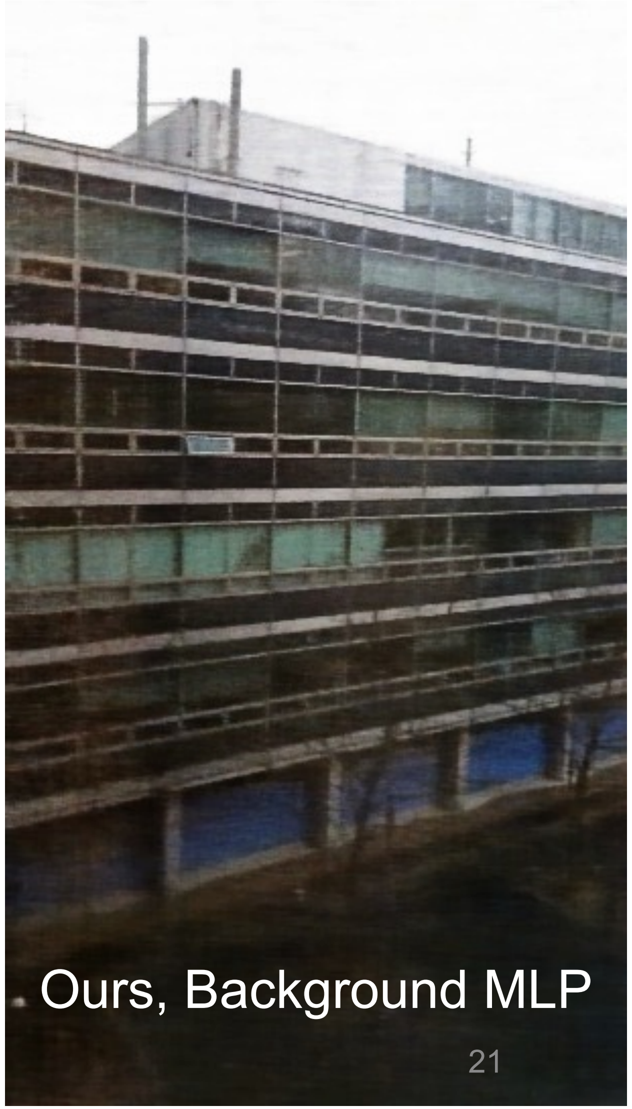
* other baselines are omitted due to **COLMAP failure**



Input sample



Ours, Scene MLP



Ours, Background MLP

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Occlusion-Free Scene Recovery via Neural Radiance Fields

Chengxuan Zhu[1,2] Renjie Wan[3] Yunkai Tang[1,2] Boxin Shi[1,2]

[1]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[3]Department of Computer Science, Hong Kong Baptist University