Novel View Synthesis from Dynamic Scenes

Jae Shin Yoon

University of Minnesota

CVPR 2020 Tutorial

Novel View Synthesis (NVS)

Srinivasan et al. "Pushing the boundaries of view extrapolation with multiplane image." CVPR 2019.

Goal

Goal

Goal

Goal


Viewpoint 1


1042
Viewpoint 2


Correspondences
1042
Viewpoint 1
Viewpoint 2


Occlusion
1042
Viewpoint 1
Viewpoint 2


Unseen Background
1042
Viewpoint 1
Viewpoint 2

Novel View Synthesis Pipeline

Images from multiple views
for **unseen background**

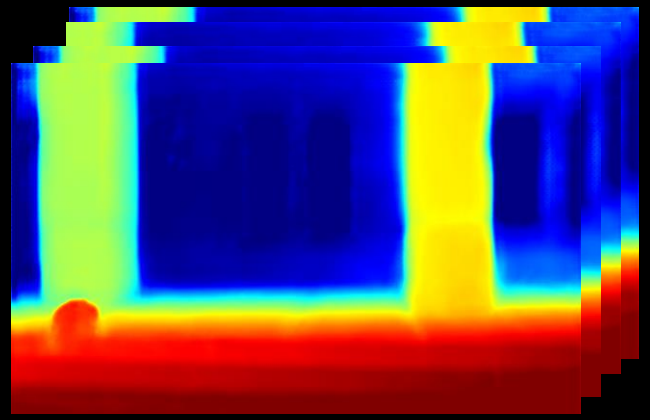
Novel View Synthesis Pipeline

Images from multiple views
for **unseen background**

More views provide the chances to see more scenes.

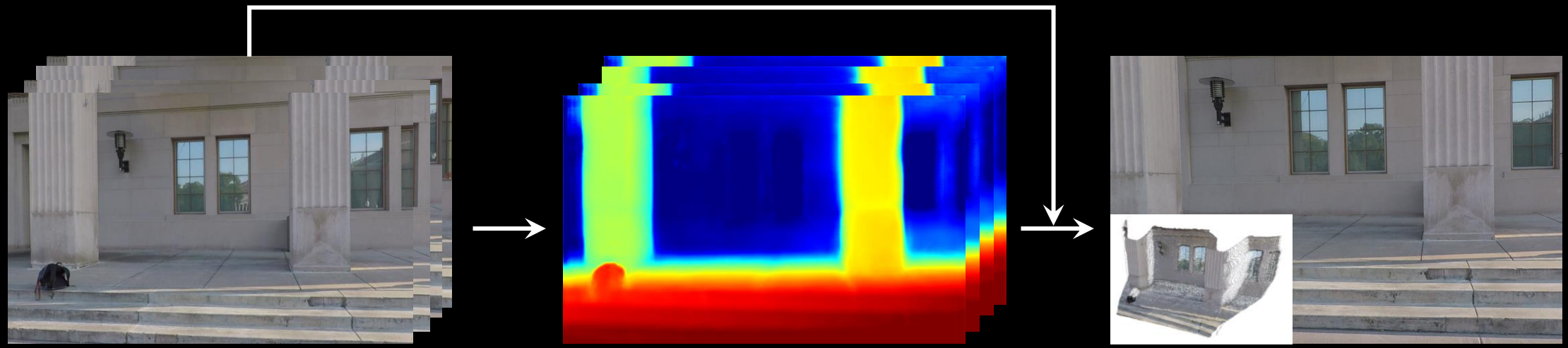
Novel View Synthesis Pipeline

Images from multiple views
for **unseen background**



Depth estimation
for **occlusion**

Novel View Synthesis Pipeline

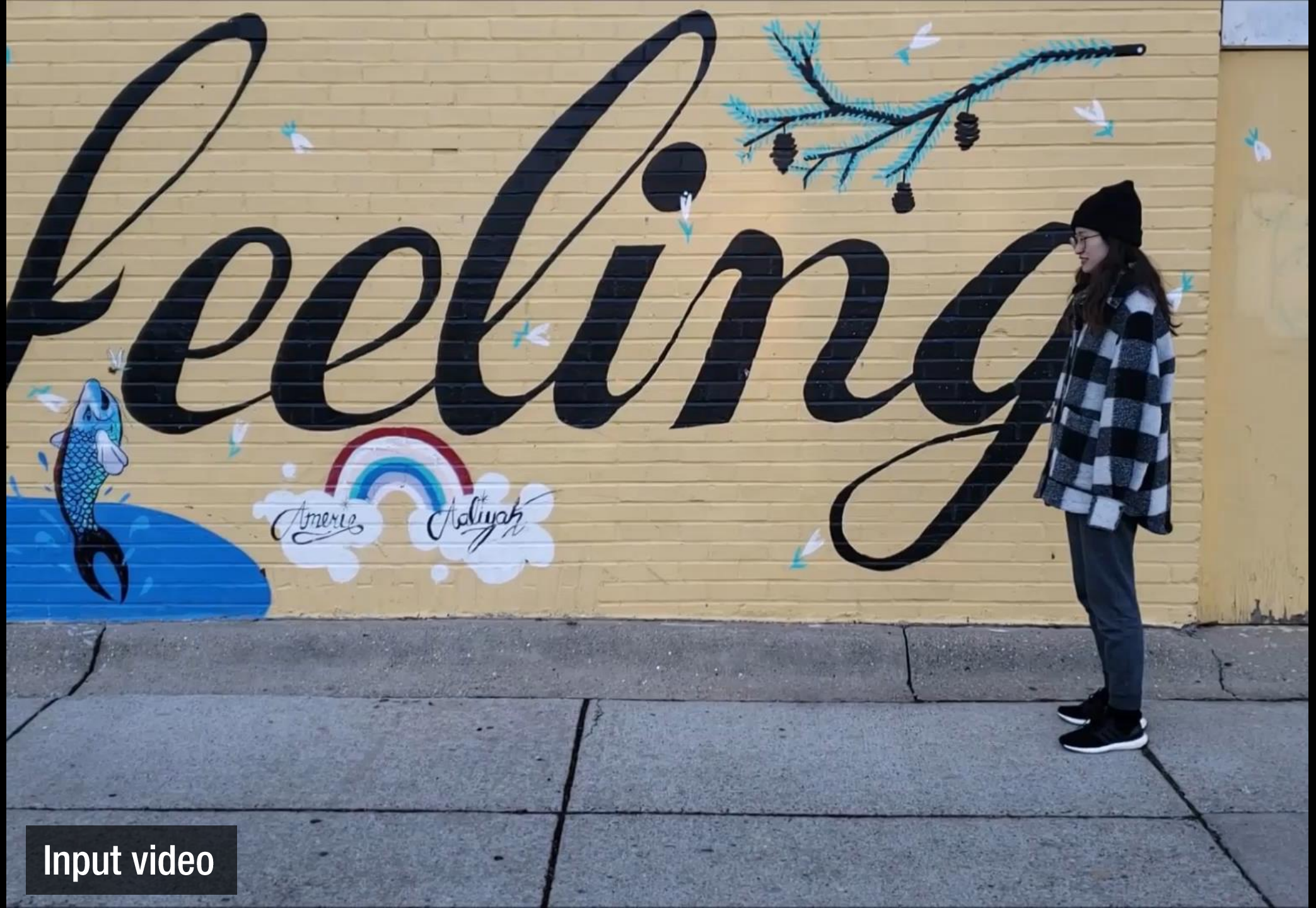


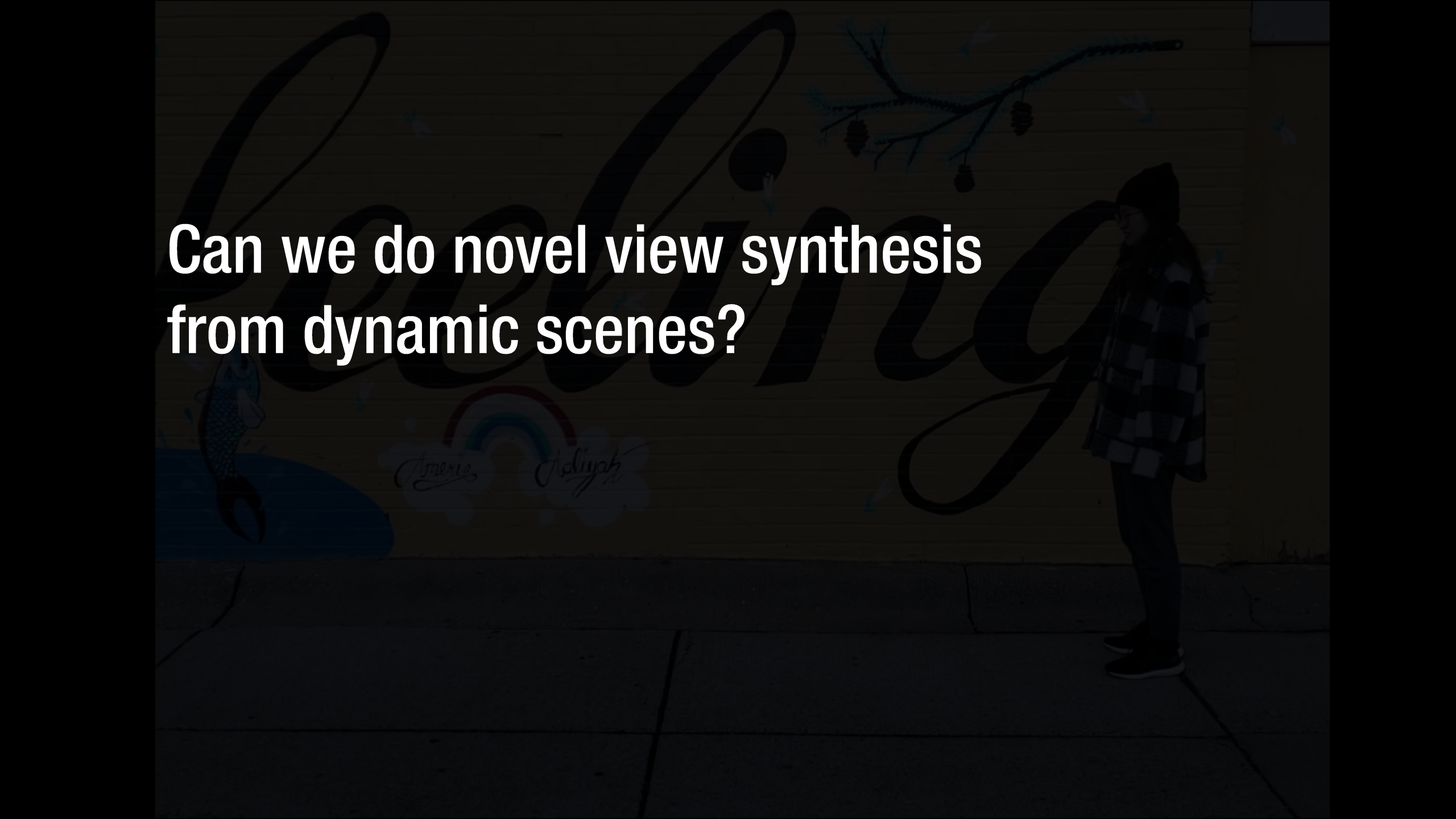
Images from multiple views
for **unseen background**

Depth estimation
for **occlusion**

Image warping via 3D geometry
for **correspondences**

Flynn et al. "DeepStereo: Learning to Predict New Views from the World's Imagery." CVPR 2016.
Kalantari et al. "Learning-Based View Synthesis for Light Field Cameras." SIGGRAPH 2016.
Zhou et al. "Stereo magnification: learning view synthesis using multiplane images." SIGGRAPH 2018.
Mildenhall et al. "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines." SIGGRAPH 2019.
Flynn et al. " DeepView: View synthesis with learned gradient descent." CVPR 2019.
Srinivasan et al. "Pushing the boundaries of view extrapolation with multiplane image." CVPR 2019.
Choi et al. "Extreme view synthesis." ICCV 2019.

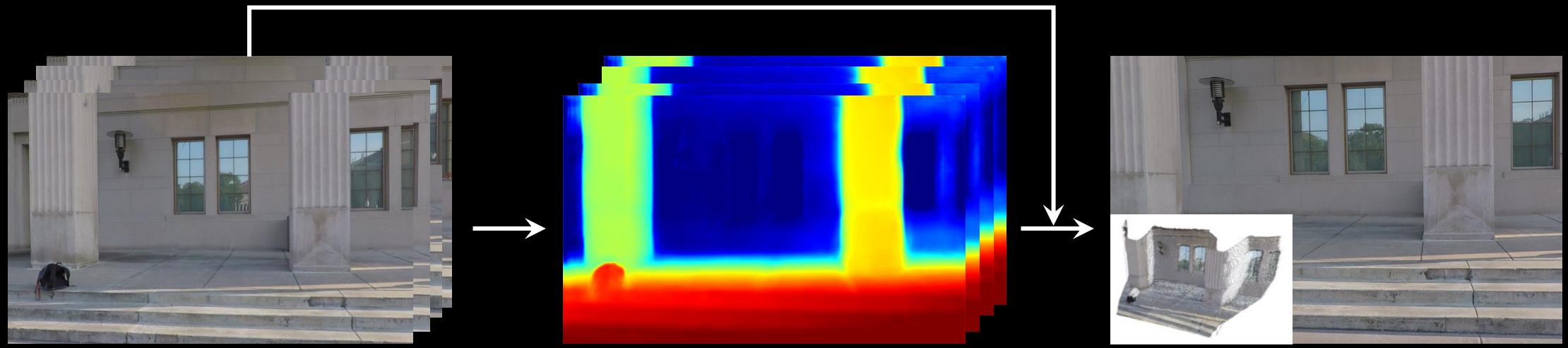

feeling
Amerie
Aaliyah
Input video


Can we do novel view synthesis
from dynamic scenes?

Can we do novel view synthesis from dynamic scenes?

Scenes that contain moving object(s).

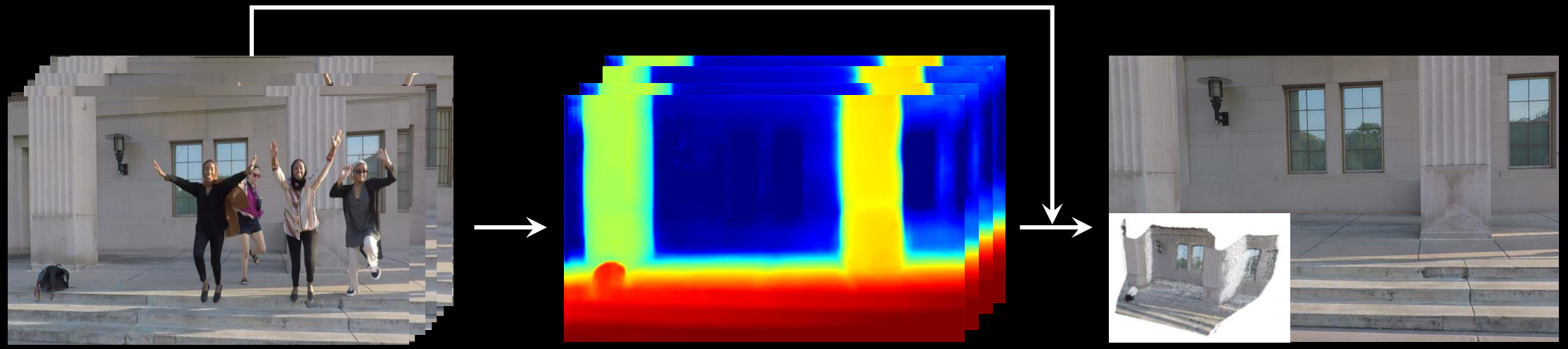
Novel View Synthesis Pipeline



Images from multiple views

Depth estimation

Image warping via 3D geometry

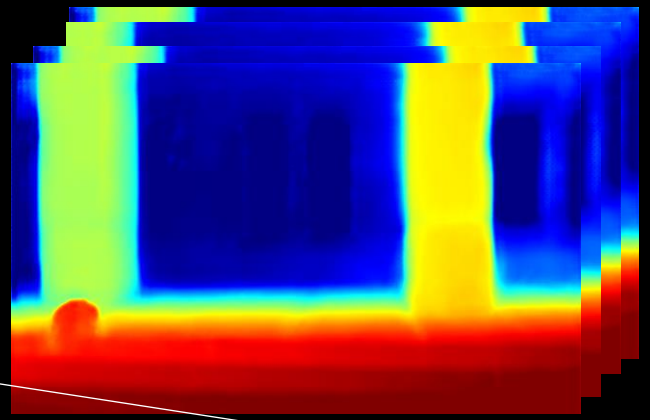


Images from multiple views

Depth estimation

Image warping via 3D geometry

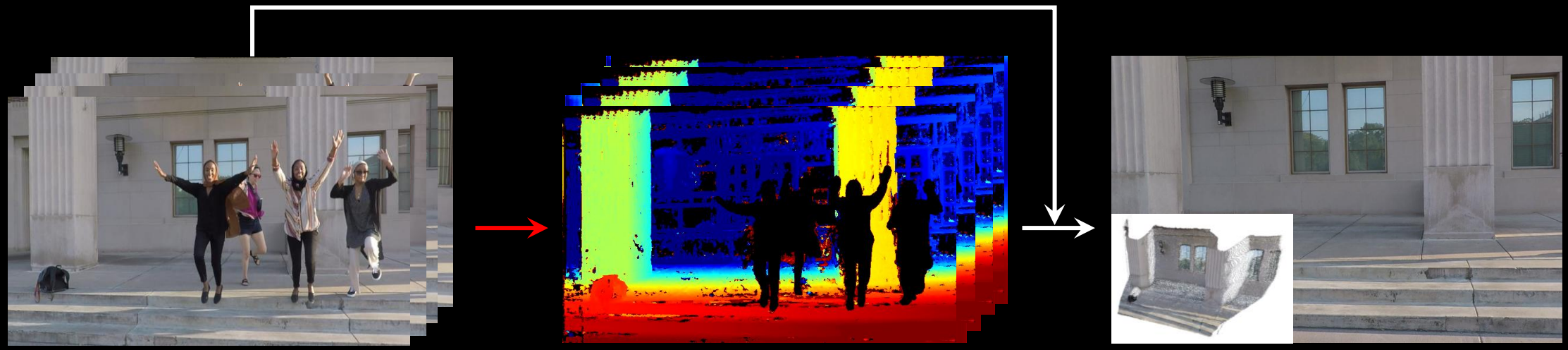
Images from multiple views



Depth estimation

Image warping via 3D geometry

Frame 1 Frame 2 Frame 3 Frame 4 Frame 5

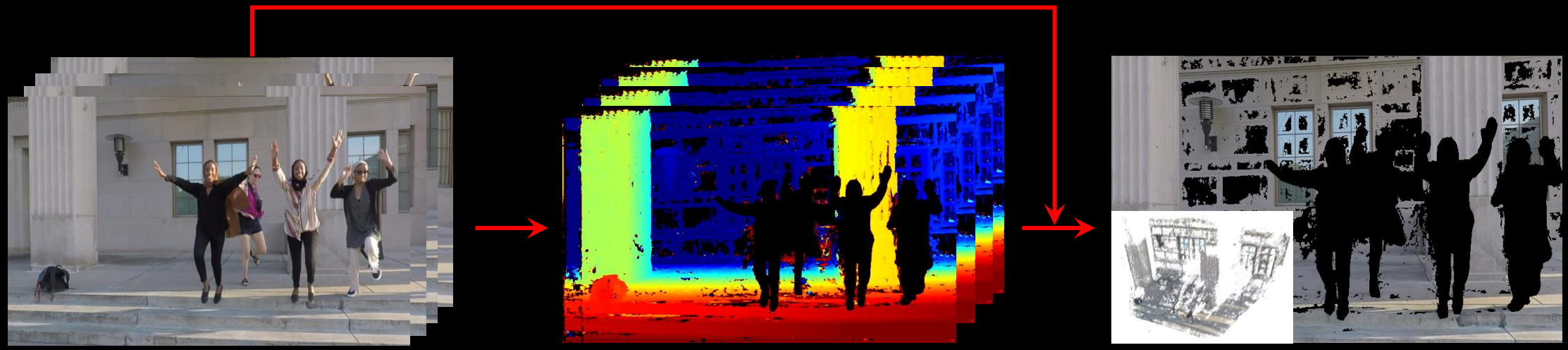


Images from multiple views

Depth estimation

Image warping via 3D geometry

The epipolar constraints do not apply in the dynamic region.



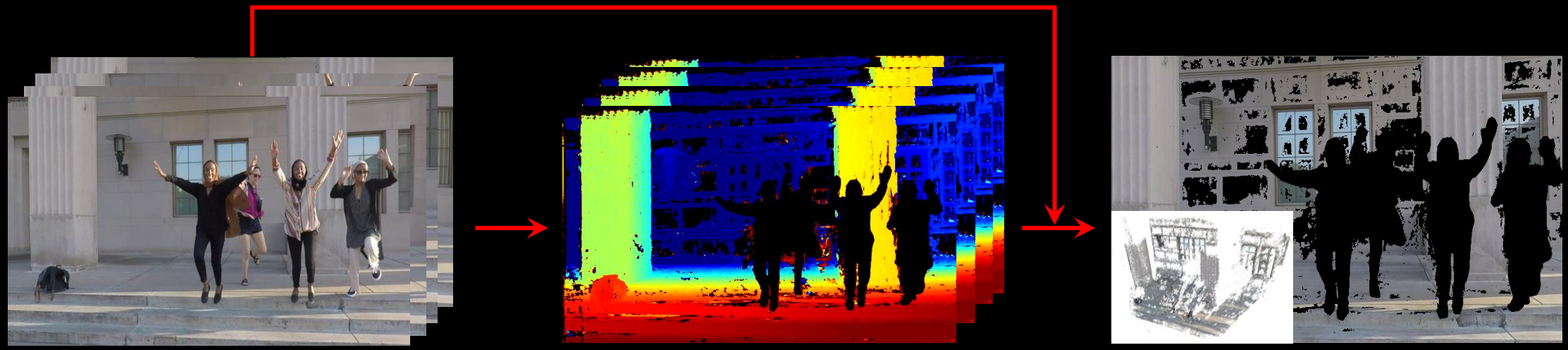
Images from multiple views

Depth estimation

Image warping via 3D geometry

The missing depth information translates to missing pixels in the synthesized views.

Challenge



Images from multiple views

Depth estimation

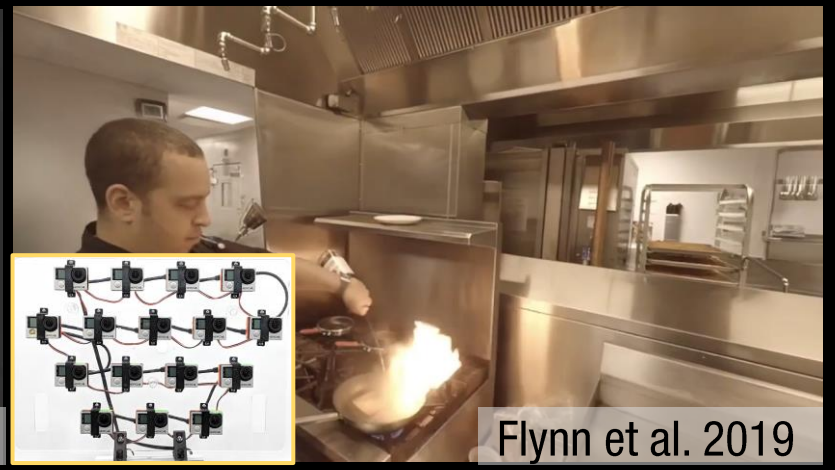
Image warping via 3D geometry

How to enable novel view synthesis when depth estimation is challenging?

Synchronized Multiview System

Zitnick et al. 2004


Lipski et al. 2009


Flynn et al. 2019

The principle of multi-view geometry can be applied to the synchronized images.

Zitnick et al. "High-quality video view interpolation using a layered representation.." SIGGRAPH 2004.

Lipski et al. " Virtual Video Camera: Image-Based Viewpoint Navigation Through Space and Time." Computer Graphics Forum 2010.

Flynn et al. " DeepView: View synthesis with learned gradient descent." CVPR 2019.

Synchronized Multiview System


Zitnick et al. 2004


Lipski et al. 2009

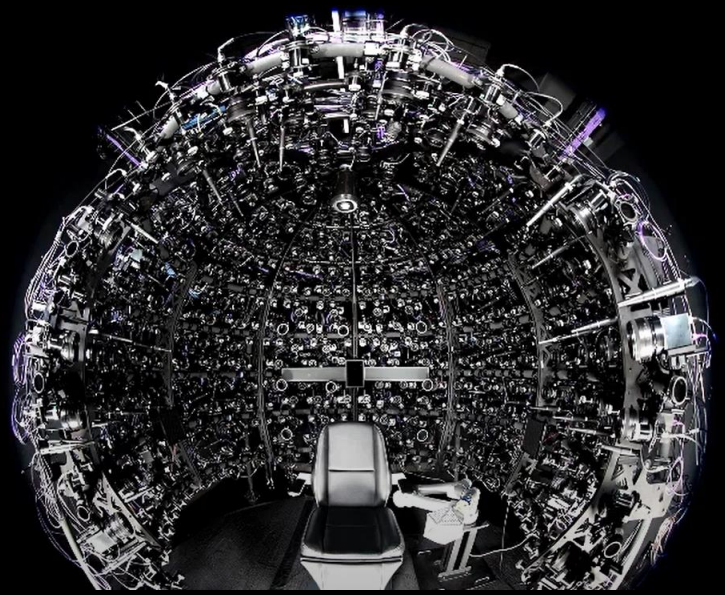

Flynn et al. 2019

Zitnick et al. "High-quality video view interpolation using a layered representation.." SIGGRAPH 2004.

Lipski et al. " Virtual Video Camera: Image-Based Viewpoint Navigation Through Space and Time." Computer Graphics Forum 2010.

Flynn et al. " DeepView: View synthesis with learned gradient descent." CVPR 2019.

Synchronized Multiview System



Large-scale multiview system

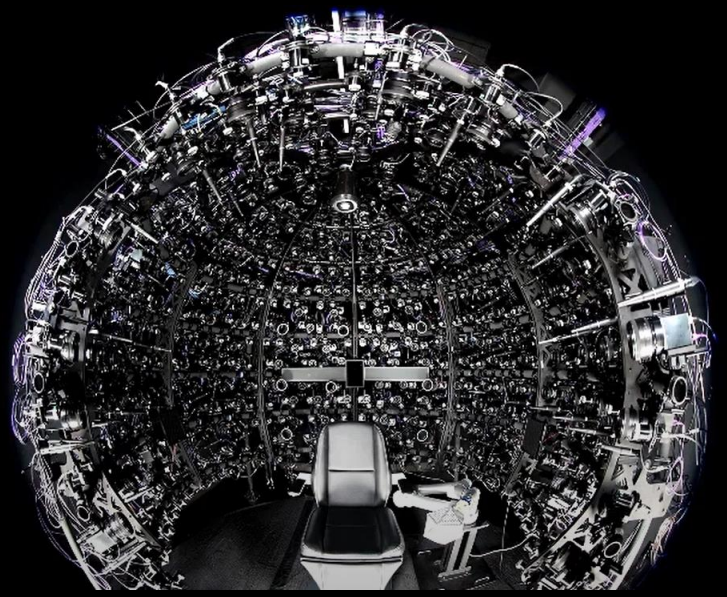
Multi-view image stream

Rendered novel views of high-fidelity face

More views provide the chances to see more scenes.

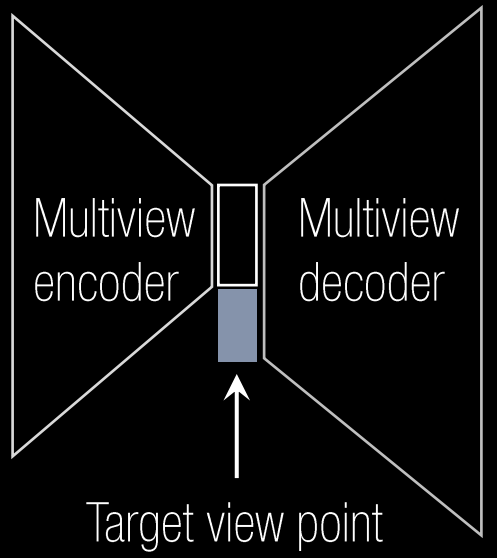
Lombardi et al. "Deep appearance models for face rendering." SIGGRAPH 2018.

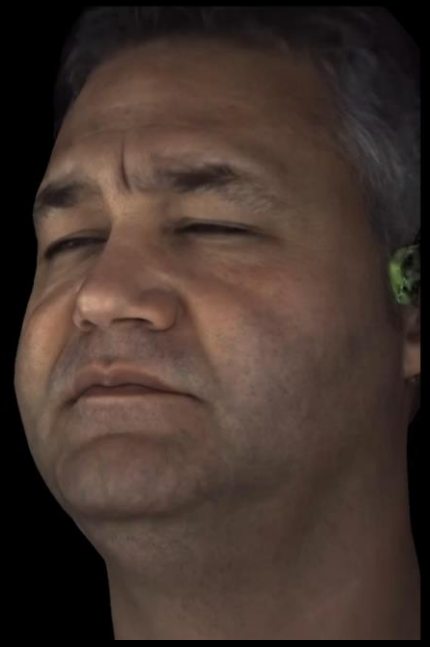
Synchronized Multiview System



Large-scale multiview system

Multi-view image stream

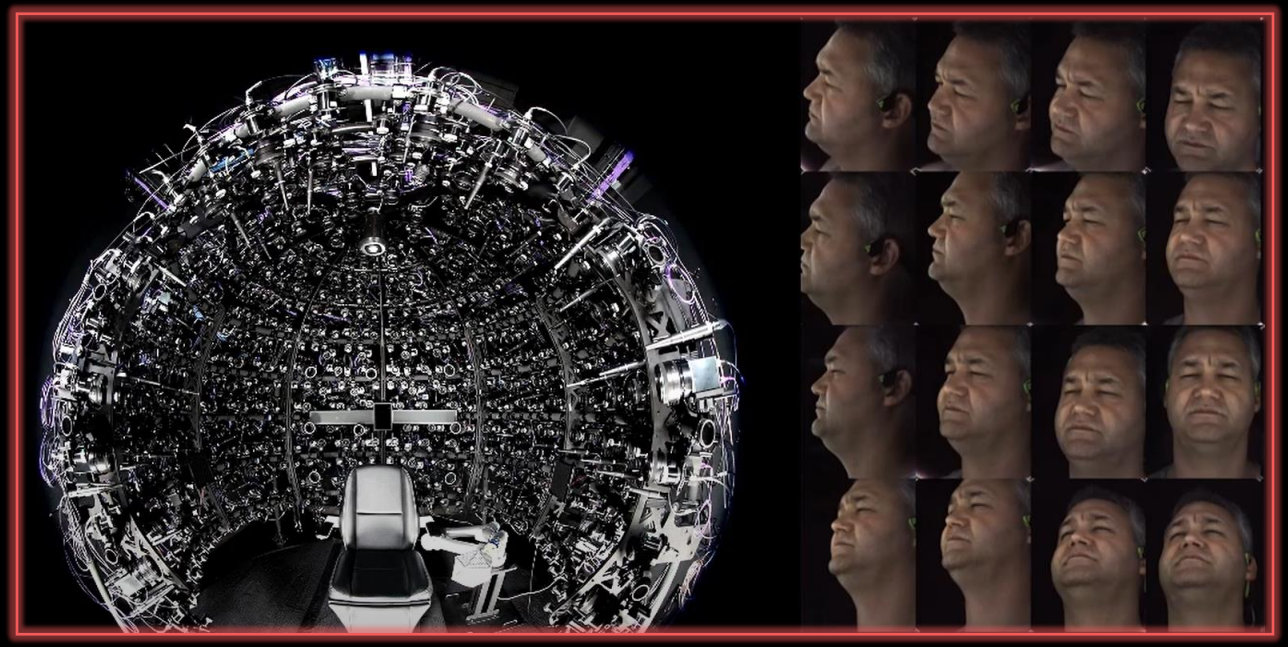




Rendered novel views of high-fidelity face

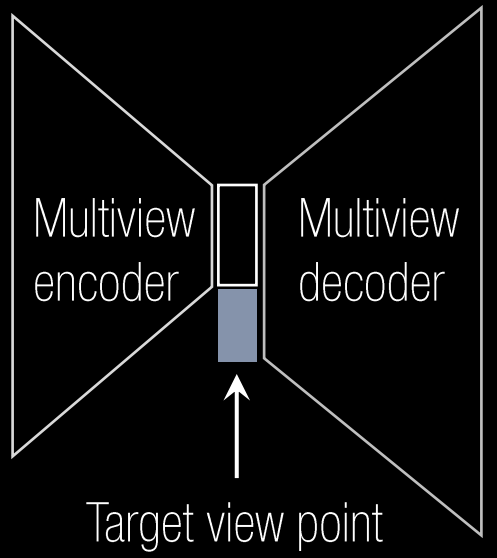
Lombardi et al. "Deep appearance models for face rendering." SIGGRAPH 2018.

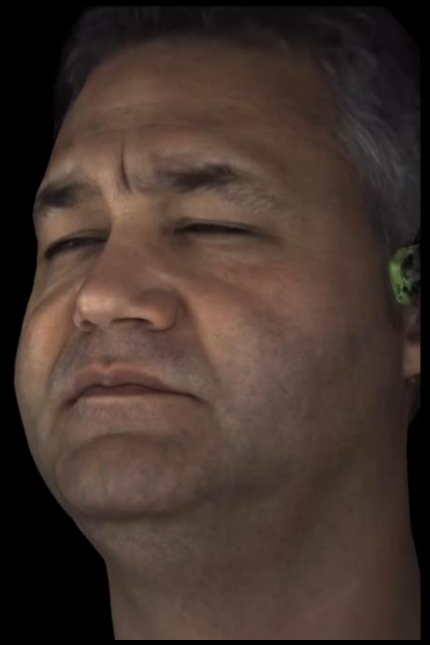
Synchronized Multiview System



Large-scale multiview system

Multi-view image stream



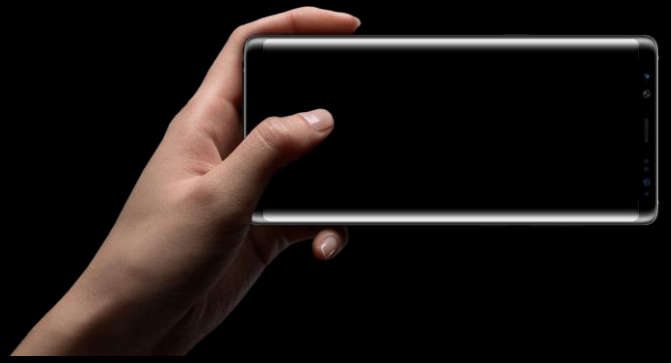


Rendered novel views of high-fidelity face

Using the multiview system is not feasible from our daily environmenet.

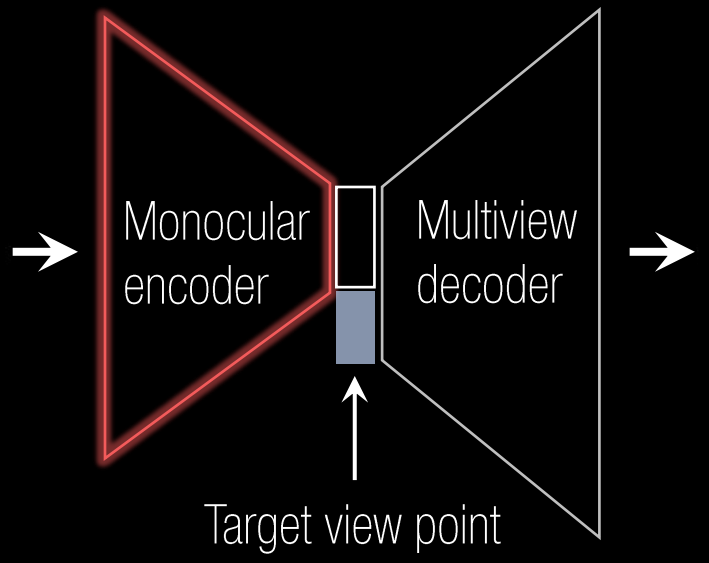
Lombardi et al. "Deep appearance models for face rendering." SIGGRAPH 2018.

Learning Model Prior



Cell-phone camera

Monocular video

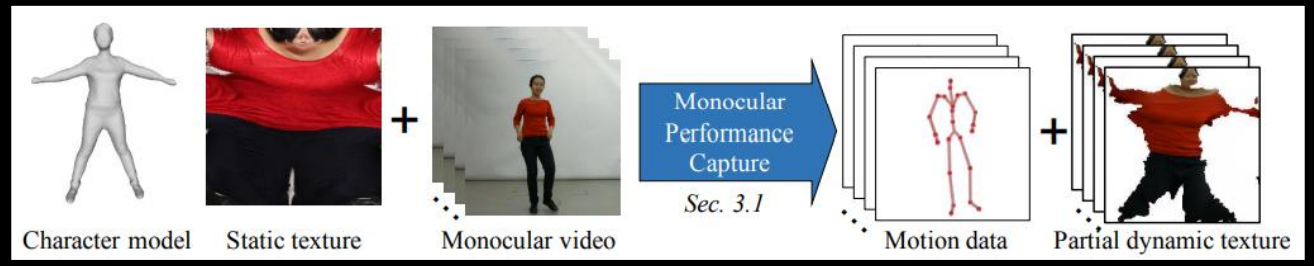


Rendered novel views of high-fidelity face

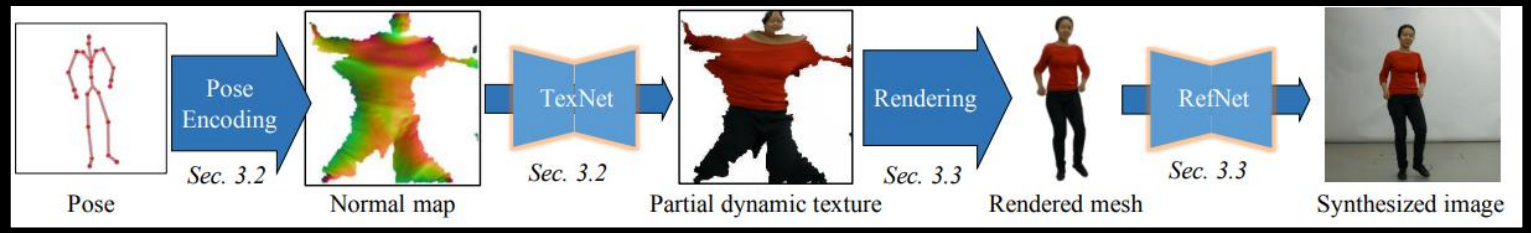
High-fidelity face model rendering from monocular camera.

Yoon et al., "Self-Supervised Adaptation of High-Fidelity Face Models for Monocular Performance Tracking" CVPR 2019

Learning Model Prior



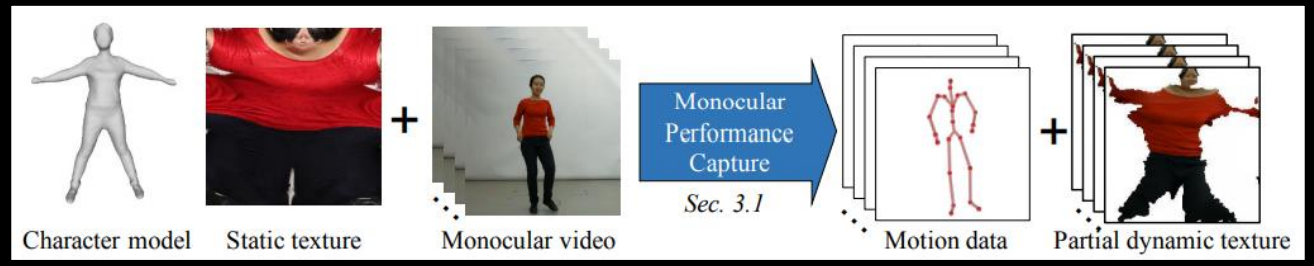
Training model specific geometry and texture



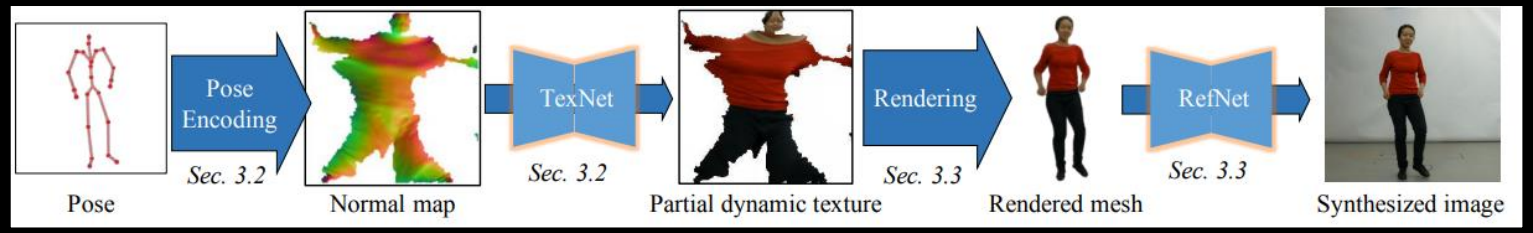
Synthesizing a person with model priors from a novel view

Liu et al. "Neural Human Video Rendering by Learning Dynamic Textures and Rendering-to-Video Translation." IEEE TVCG 2020.

Learning Model Prior



Training model specific geometry and texture



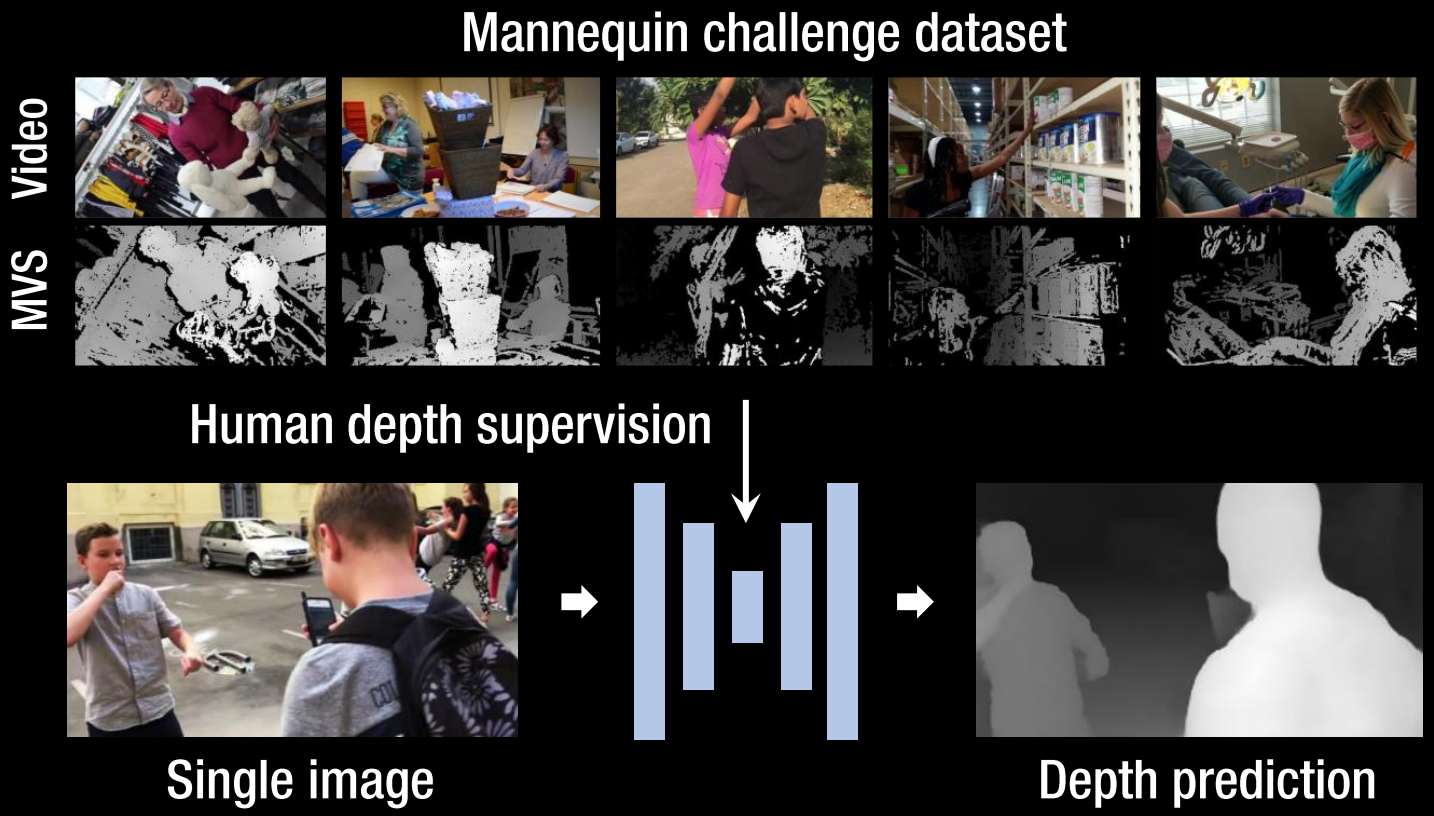
Synthesizing a person with model priors from a novel view

Monocular bullet time effect

The NVS application is limited to a person-specific model.

Liu et al. "Neural Human Video Rendering by Learning Dynamic Textures and Rendering-to-Video Translation." IEEE TVCG 2020.

Learning Human Depth Prior



Li et al. "Learning the depths of moving people by watching frozen people." CVPR 2019.

Learning Human Depth Prior

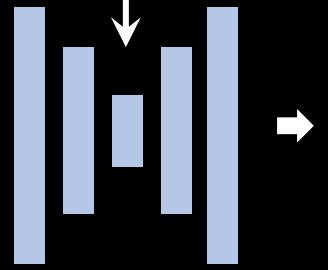
Mannequin challenge dataset

Video

MVS

Human depth supervision

Single image



Depth prediction

Novel view synthesis
from dynamic scenes with people

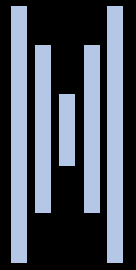
The NVS application is limited to the scenes with people.

Li et al. "Learning the depths of moving people by watching frozen people." CVPR 2019.

Generalized Single View Depth Estimation

Source camera

Class-agnostic depth prediction [Ranftl et al. 2019] [Niklaus et al. 2019]



Input image

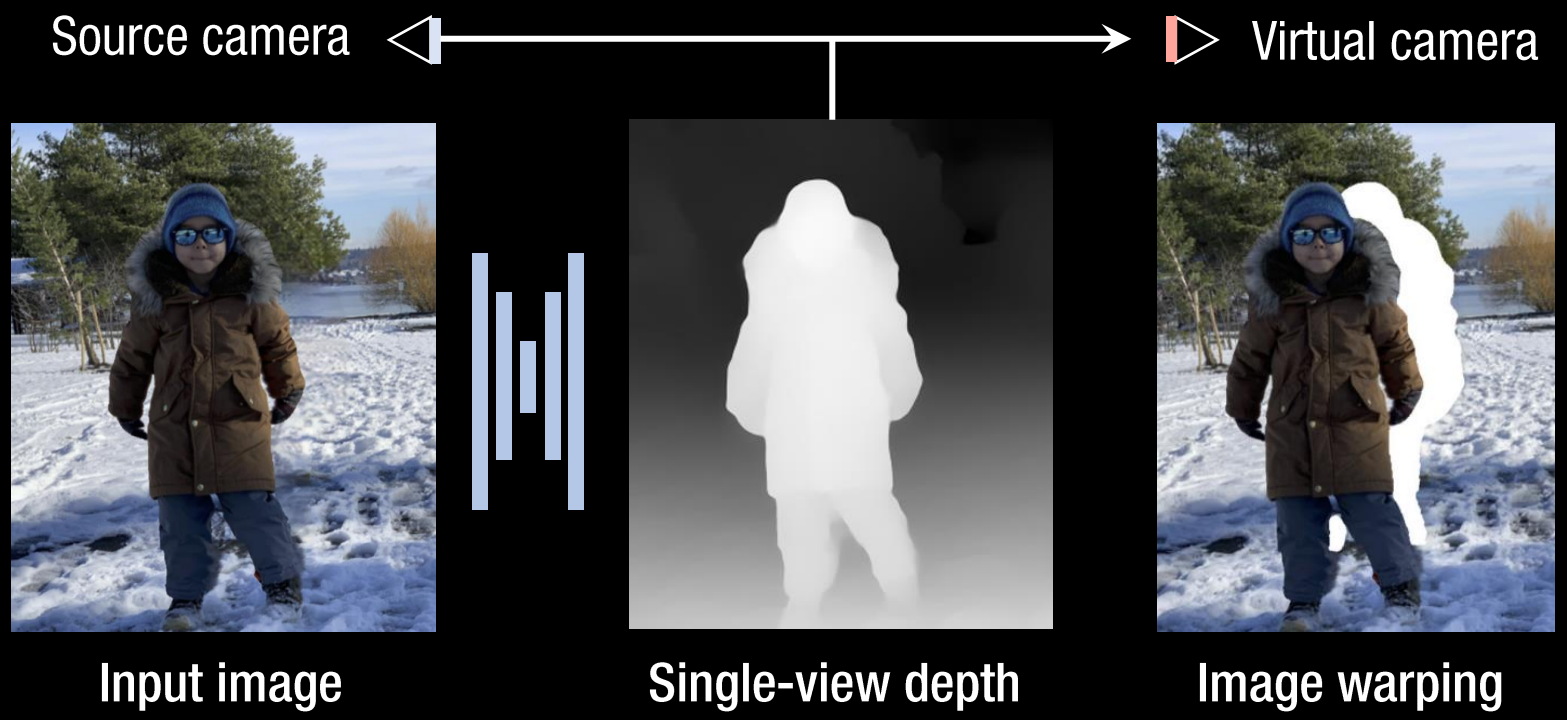
Single-view depth

Shih et al. "3D Photography using Context-aware Layered Depth Inpainting." CVPR 2020.

Niklaus et al. "3D Ken Burns Effect from a Single Image." SIGGRAPH 2019.

Ranftl et al. "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer." Arxiv 2019.

Generalized Single View Depth Estimation

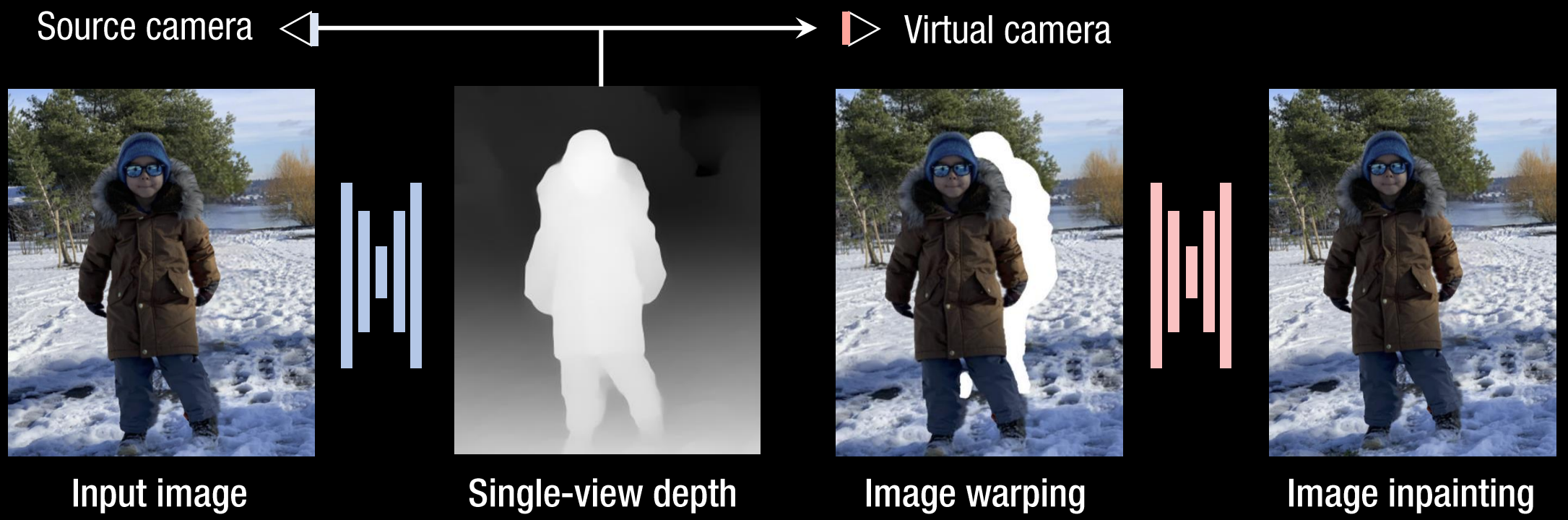


Shih et al. "3D Photography using Context-aware Layered Depth Inpainting." CVPR 2020.

Niklaus et al. "3D Ken Burns Effect from a Single Image." SIGGRAPH 2019.

Ranftl et al. "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer." Arxiv 2019.

Generalized Single View Depth Estimation



Shih et al. "3D Photography using Context-aware Layered Depth Inpainting." CVPR 2020.

Niklaus et al. "3D Ken Burns Effect from a Single Image." SIGGRAPH 2019.

Ranftl et al. "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer." Arxiv 2019.

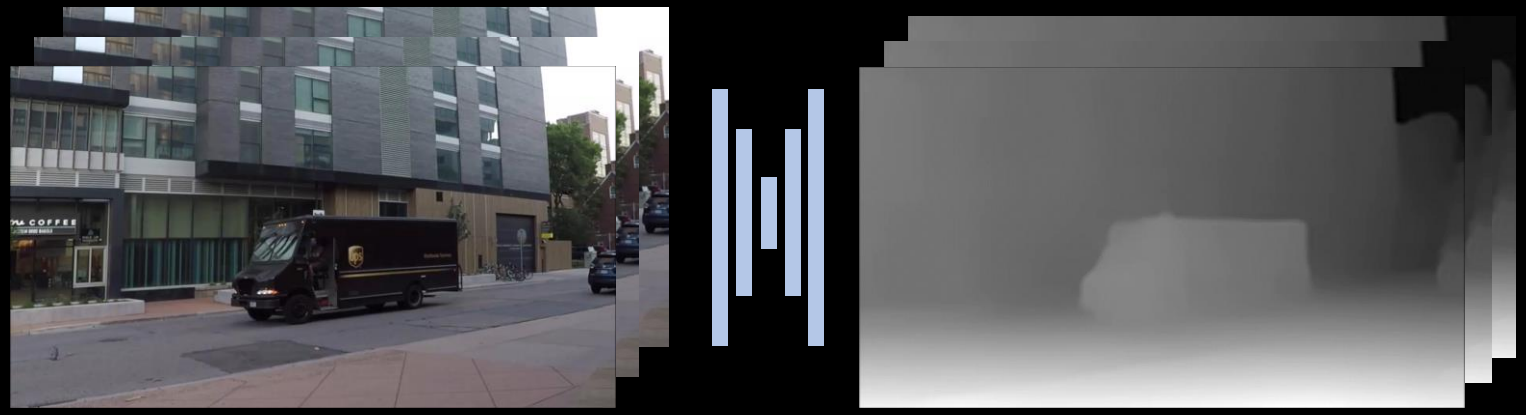
Generalized Single View Depth Estimation

The NVS application is limited to a small camera displacement and single time instance.

Shih et al. "3D Photography using Context-aware Layered Depth Inpainting." CVPR (2020).

Niklaus et al. "3D Ken Burns Effect from a Single Image." SIGGRAPH (2019).

Coherent Depth Estimation from Video

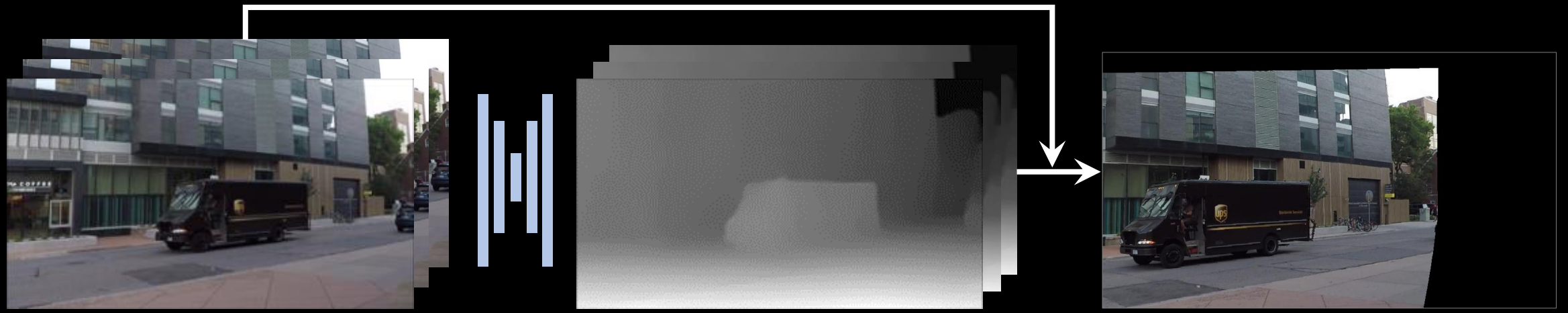


Images from multi-view

Single-view depth at each view

More views provide the chances to see more scenes and times.

Coherent Depth Estimation from Video

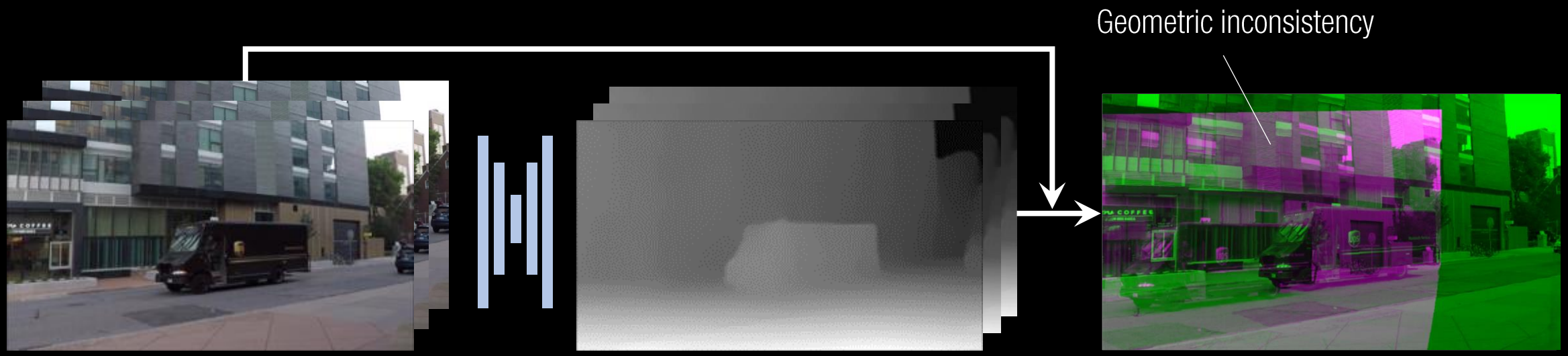


Images from multi-view

Single-view depth at each view

Image warping to other views

Coherent Depth Estimation from Video



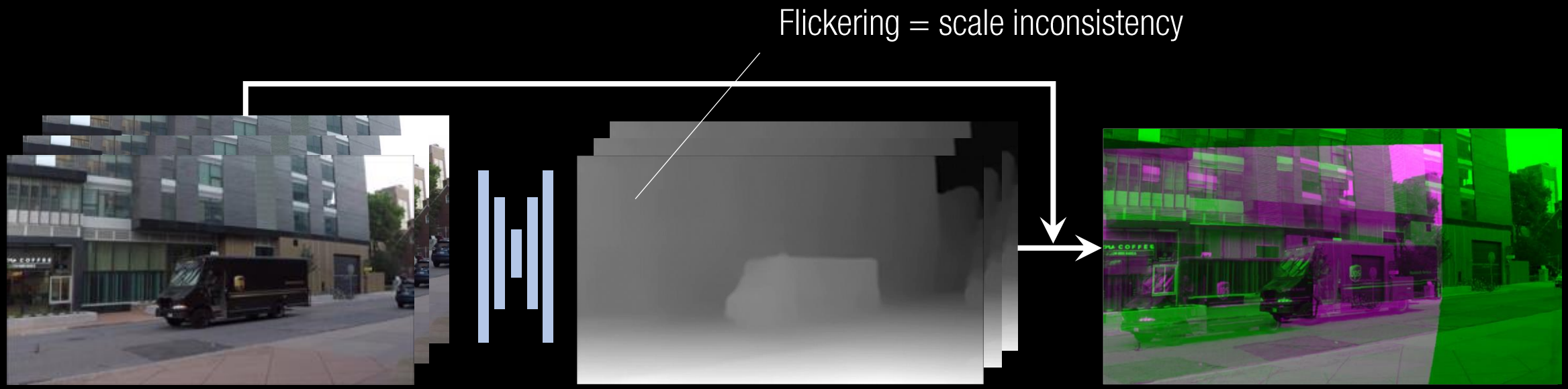
Images from multi-view

Single-view depth at each view

Overlay with ground-truth

The warping is geometrically incorrect.

Coherent Depth Estimation from Video



The scale consistent depth is the requirement to combine the pixels from all views.

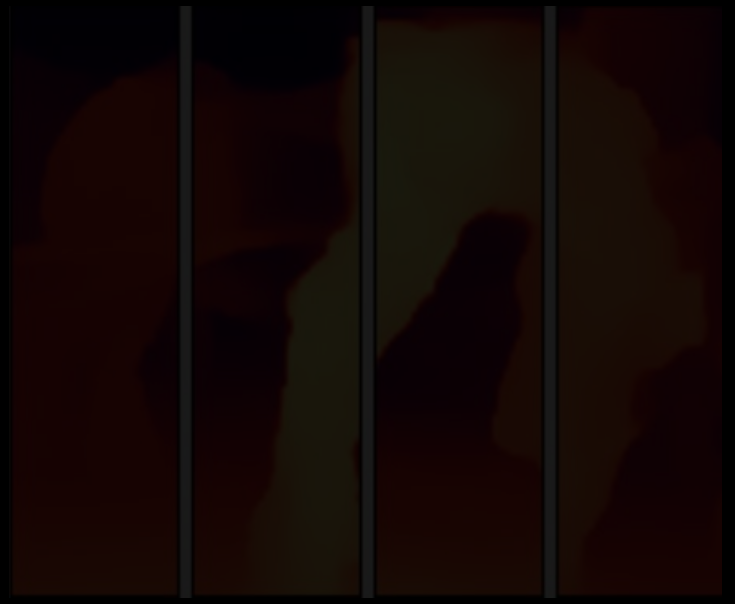
Coherent Depth Estimation from Video

Frame1 Frame2 Frame3 Frame4

Input multi-view images
from dynamic scene

Frame1 Frame2 Frame3 Frame4

Depth from multi-view stereo
(scale-invariant, incomplete)



Frame1 Frame2 Frame3 Frame4

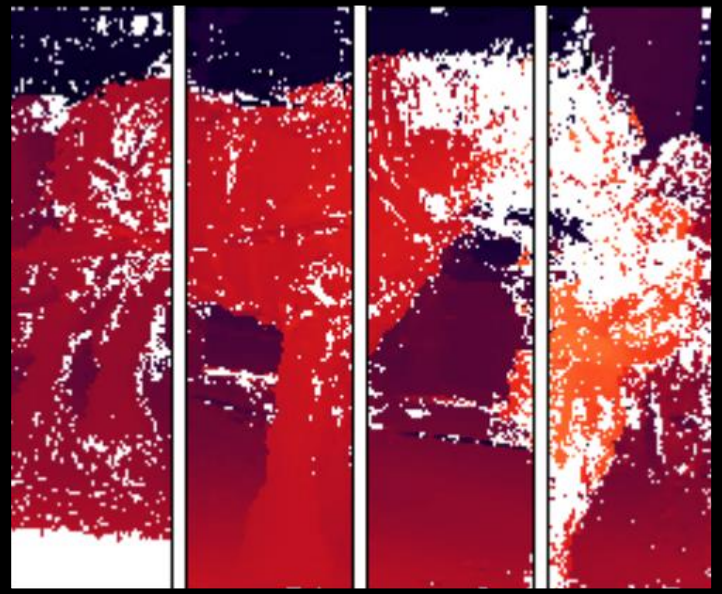
Depth from single-view prediction
(scale-variant, complete)

Luo et al. “Coherent Video Depth Estimation." SIGGRAPH 2020.

Coherent Depth Estimation from Video

Frame1 Frame2 Frame3 Frame4

Input multi-view images
from dynamic scene



Frame1 Frame2 Frame3 Frame4

Depth from multi-view stereo
(scale-invariant, incomplete)

Frame1 Frame2 Frame3 Frame4

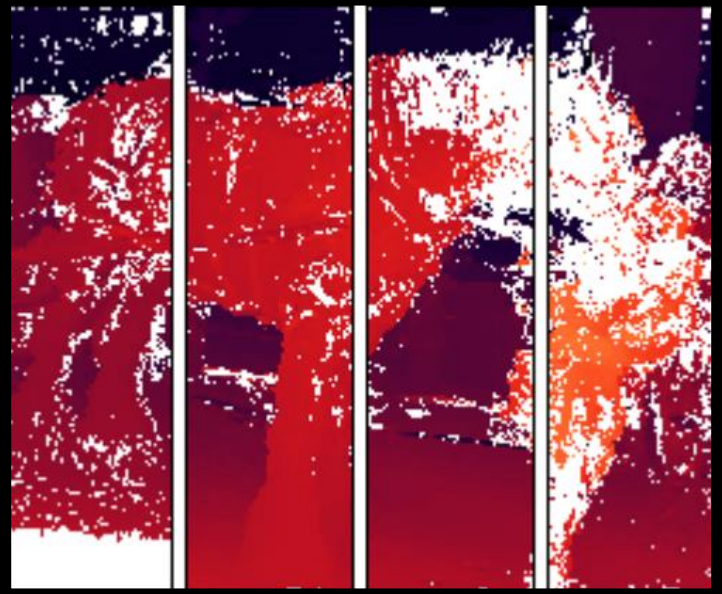
Depth from single-view prediction
(scale-variant, complete)

Luo et al. "Coherent Video Depth Estimation." SIGGRAPH 2020.

Coherent Depth Estimation from Video

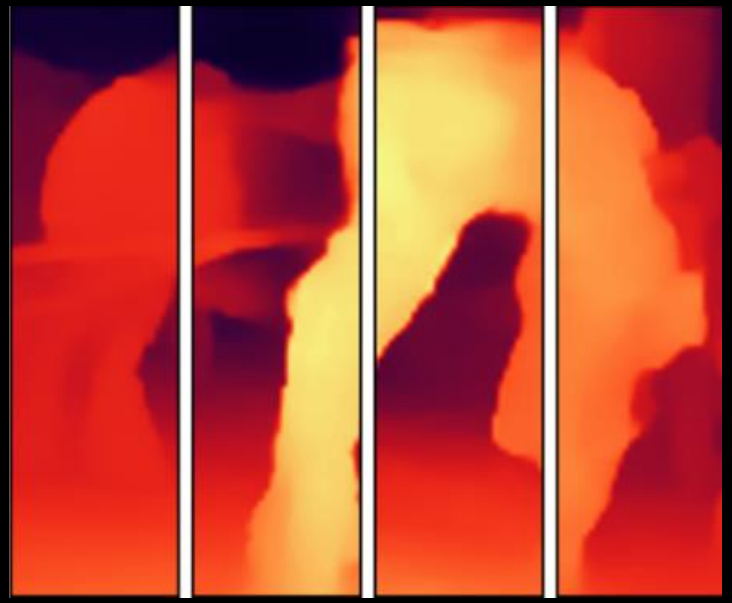
Frame1 Frame2 Frame3 Frame4

Input multi-view images
from dynamic scene



Frame1 Frame2 Frame3 Frame4

Depth from multi-view stereo
(scale-invariant, incomplete)



Frame1 Frame2 Frame3 Frame4

Depth from single-view prediction
(scale-variant, complete)

Luo et al. "Coherent Video Depth Estimation." SIGGRAPH 2020.

Coherent Depth Estimation from Video

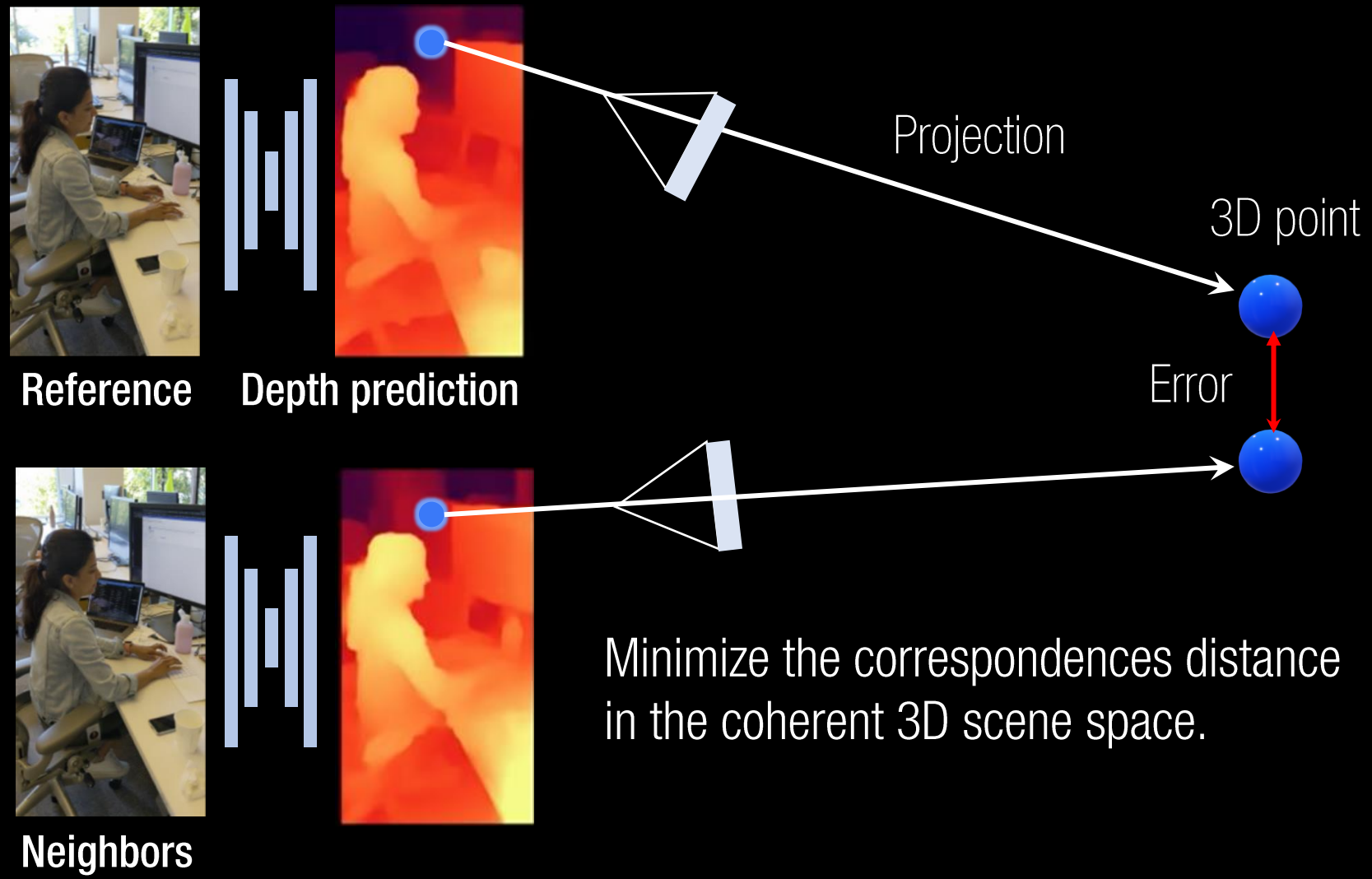
Reference

Depth prediction

Neighbors

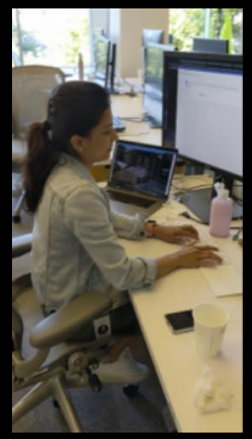
Luo et al. “Coherent Video Depth Estimation." SIGGRAPH 2020.

Coherent Depth Estimation from Video



Luo et al. "Coherent Video Depth Estimation." SIGGRAPH 2020.

Coherent Depth Estimation from Video



Reference

Depth prediction

Projection

3D point

Error

: optical flow

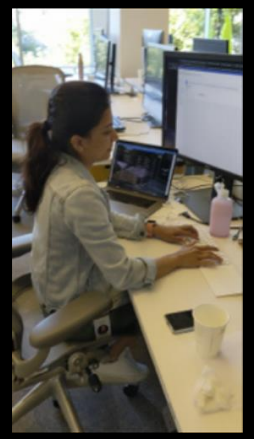
: Structure-from-motion

Neighbors

Minimize the correspondences distance
in the coherent 3D scene space.

Luo et al. "Coherent Video Depth Estimation." SIGGRAPH 2020.

Coherent Depth Estimation from Video



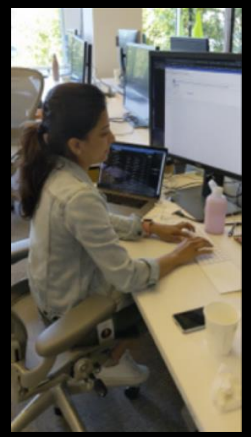
Reference

Depth prediction

Projection

3D point

Error



Neighbors

Minimize the correspondences distance in the coherent 3D scene space.

Cohereht video depth

Luo et al. "Coherent Video Depth Estimation." SIGGRAPH 2020.

Coherent Depth Estimation from Video

3D scene reconstruction

Depth-aware visual effet

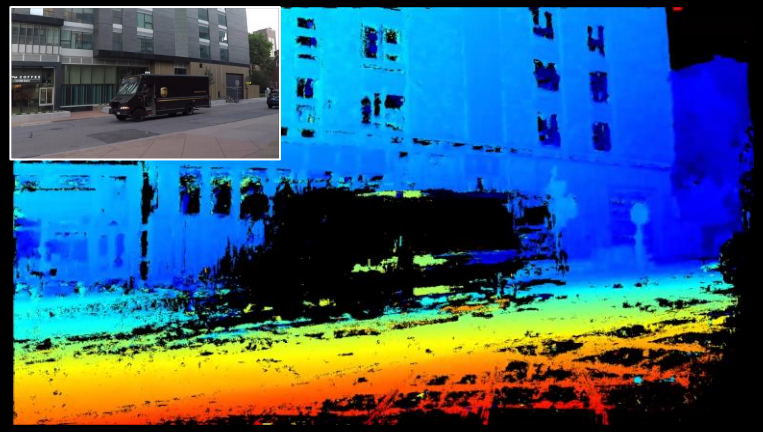
Augmented reality

Luo et al. “Coherent Video Depth Estimation." SIGGRAPH 2020.

Our Solution: Coherent Depth Estimation by Fusion

Yoon et al. “Novel View Synthesis of Dynamic Scenes with Globally Coherent Depths from a Monocular Camera." CVPR 2020.

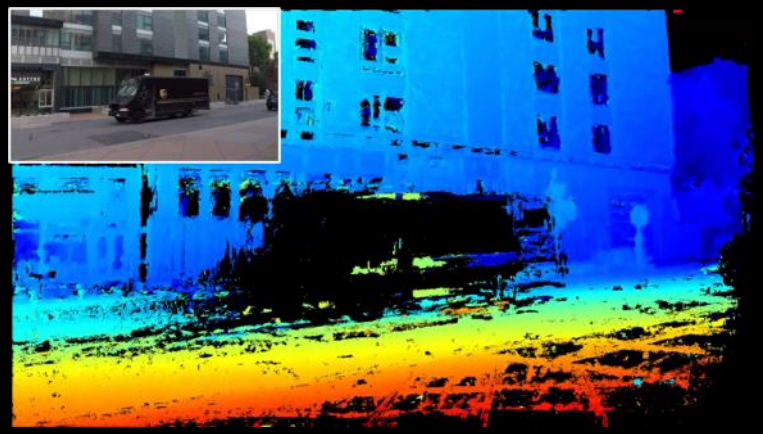
Our Solution: Coherent Depth Estimation by Fusion



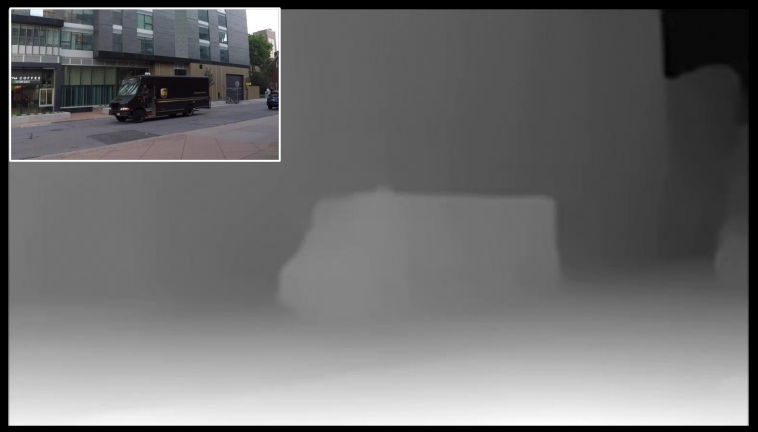
Depth from multi-view stereo
(scale-invariant, incomplete)

Yoon et al. “Novel View Synthesis of Dynamic Scenes with Globally Coherent Depths from a Monocular Camera." CVPR 2020.

Our Solution: Coherent Depth Estimation by Fusion



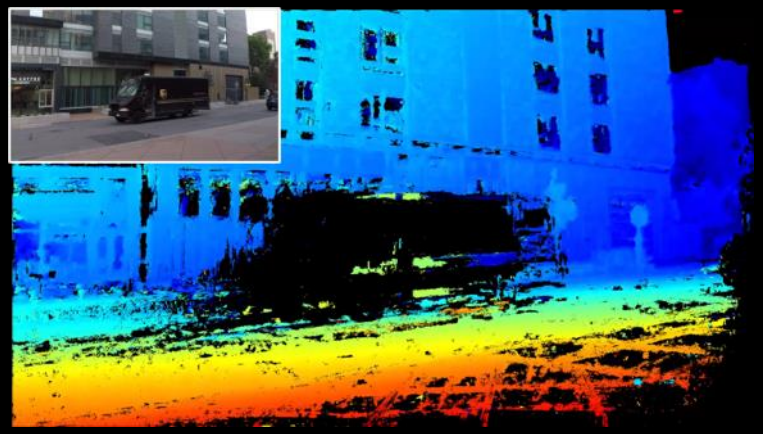
Depth from multi-view stereo
(scale-invariant, incomplete)



Depth from single-view prediction
(scale-variant, complete)

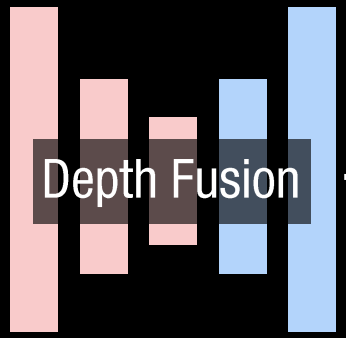
Yoon et al. “Novel View Synthesis of Dynamic Scenes with Globally Coherent Depths from a Monocular Camera." CVPR 2020.

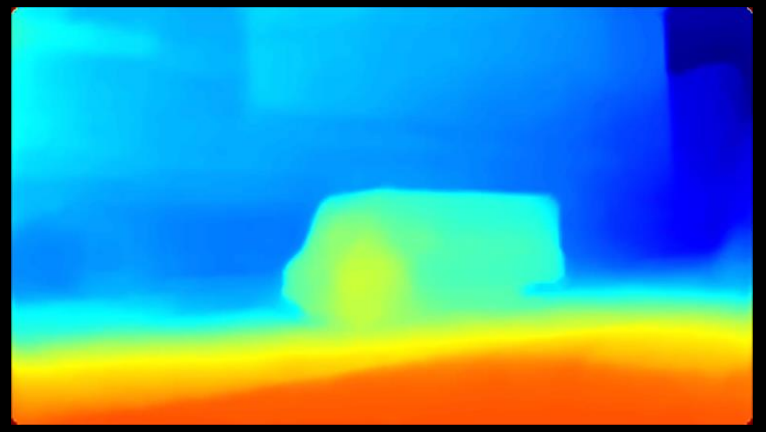
Our Solution: Coherent Depth Estimation by Fusion



Depth from multi-view stereo
(scale-invariant, incomplete)

Depth from single-view prediction
(scale-variant, complete)

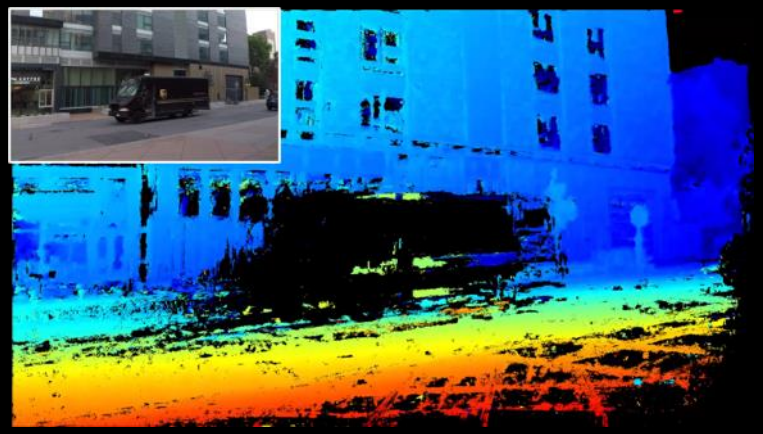




Fused depth
(scale-invariant, complete)

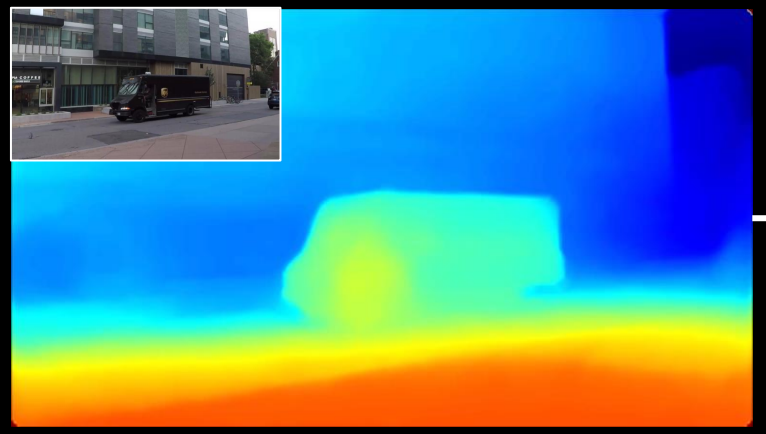
Yoon et al. "Novel View Synthesis of Dynamic Scenes with Globally Coherent Depths from a Monocular Camera." CVPR 2020.

Our Solution: Coherent Depth Estimation by Fusion



Depth from multi-view stereo
(scale-invariant, incomplete)

Depth Fusion



Fused depth
(scale-invariant, complete)

Image warping

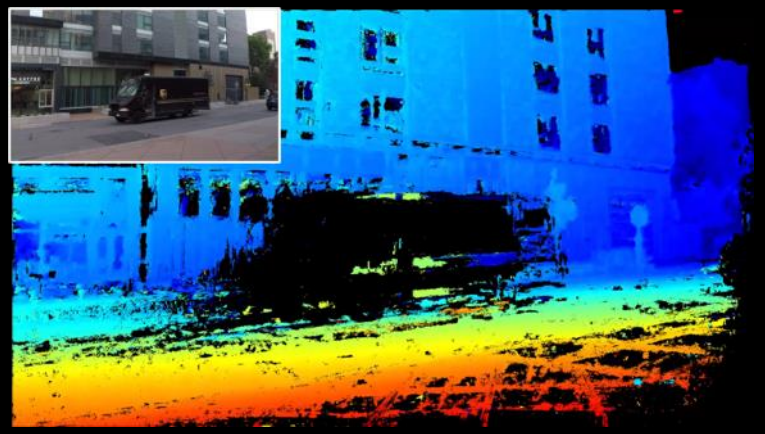
Depth from single-view prediction
(scale-variant, complete)

View synthesis from a virtual view

ImageBlender

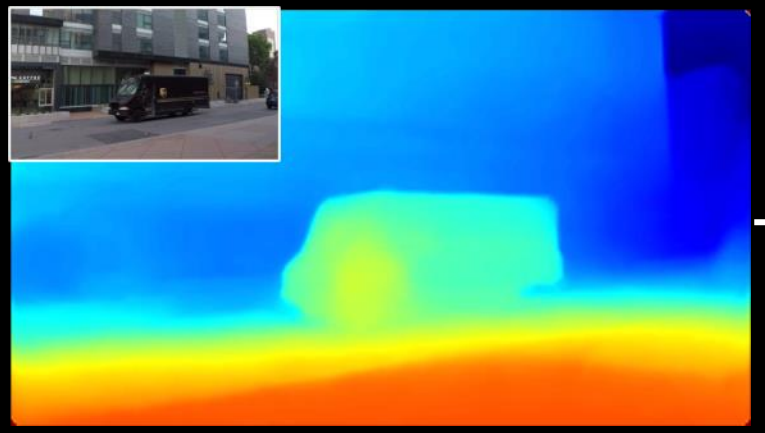
Yoon et al. "Novel View Synthesis of Dynamic Scenes with Globally Coherent Depths from a Monocular Camera." CVPR 2020.

Our Solution: Coherent Depth Estimation by Fusion



Depth from multi-view stereo
(scale-invariant, incomplete)

Depth Fusion



Fused depth
(scale-invariant, complete)

Image warping

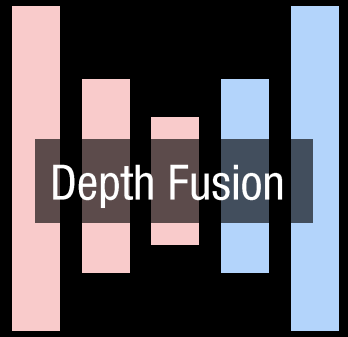
Depth from single-view prediction
(scale-variant, complete)

View synthesis from a virtual view

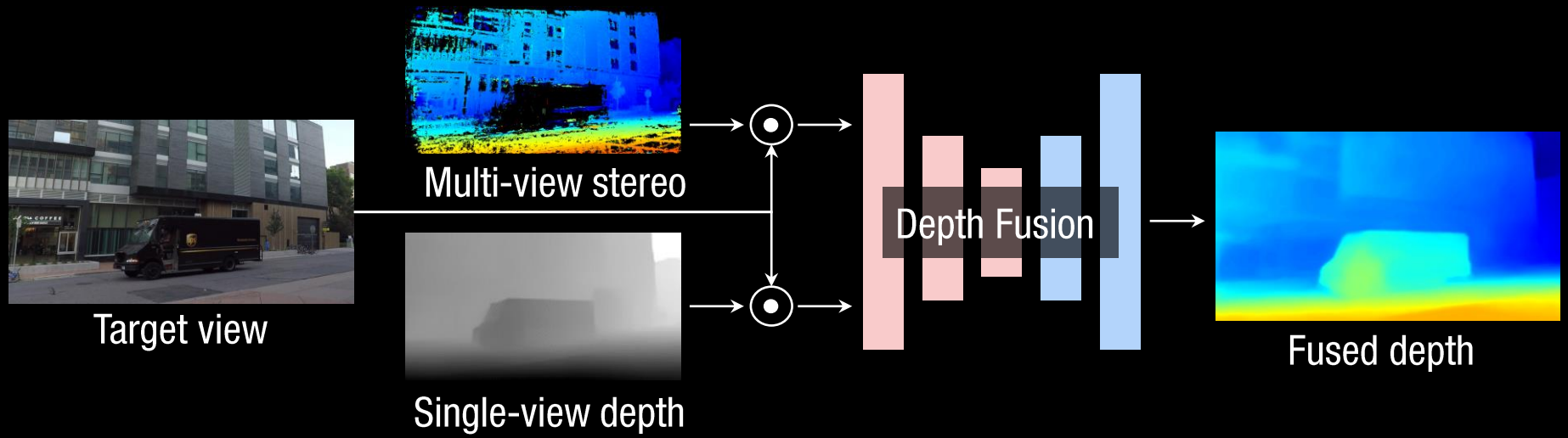
ImageBlender

Yoon et al. "Novel View Synthesis of Dynamic Scenes with Globally Coherent Depths from a Monocular Camera." CVPR 2020.

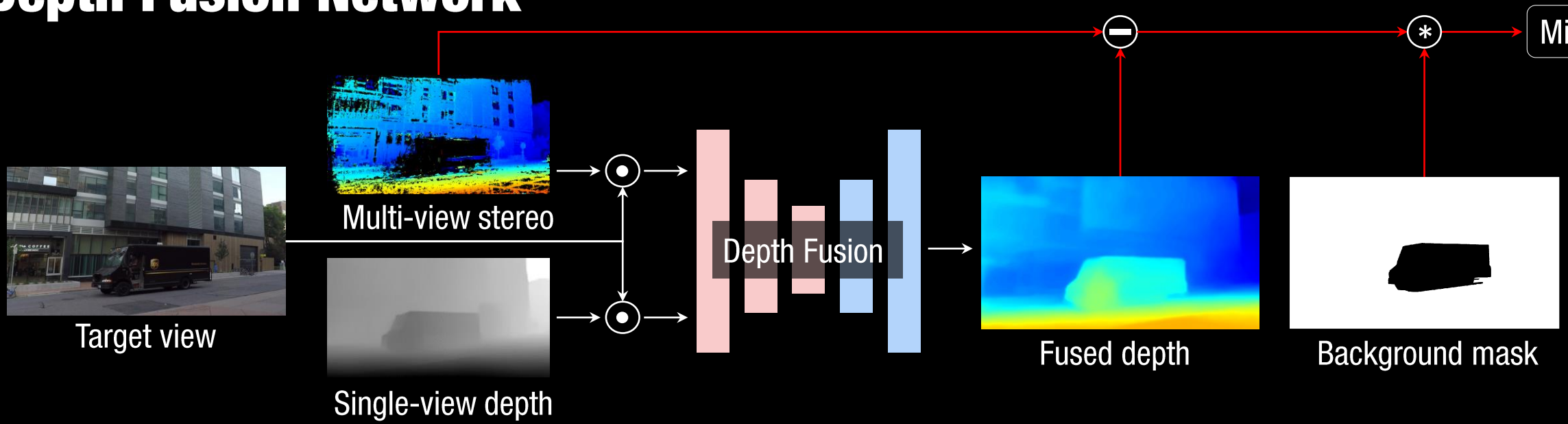
Depth Fusion Network



Depth Fusion Network

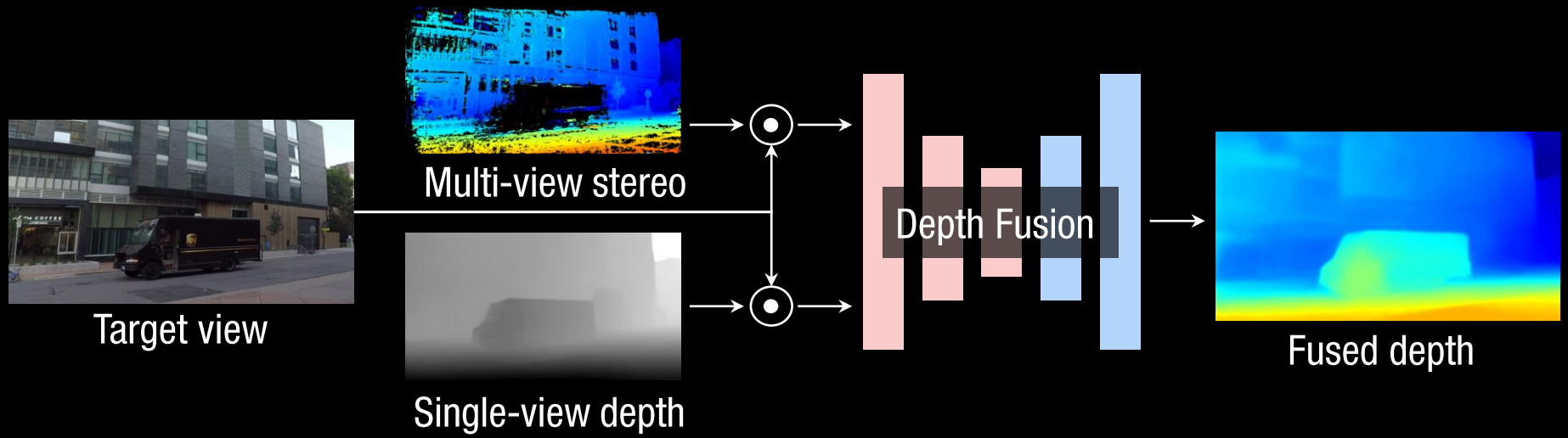


Depth Fusion Network

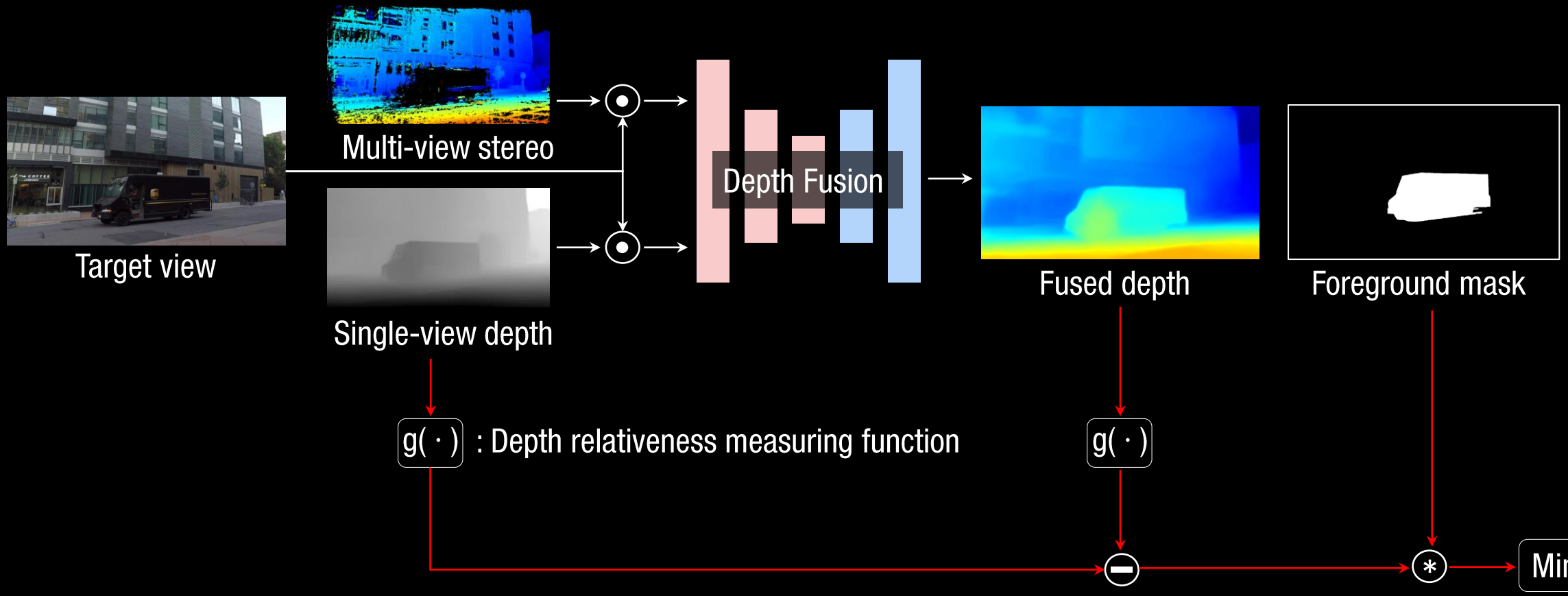


The estimated depth from static regions must be aligned with multiview stereo depth.

Depth Fusion Network

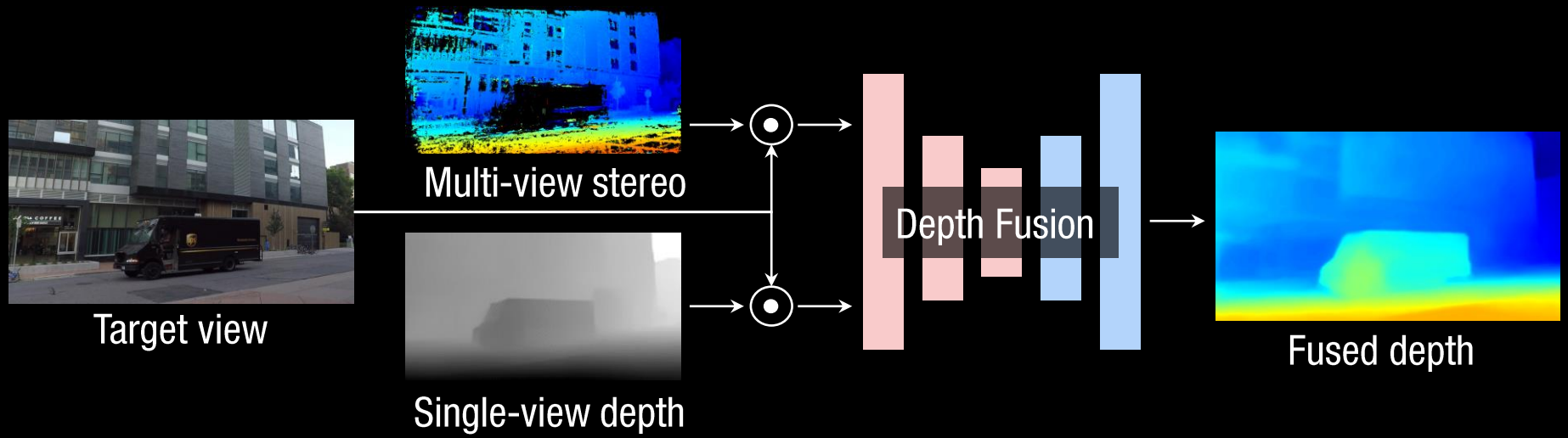


Depth Fusion Network

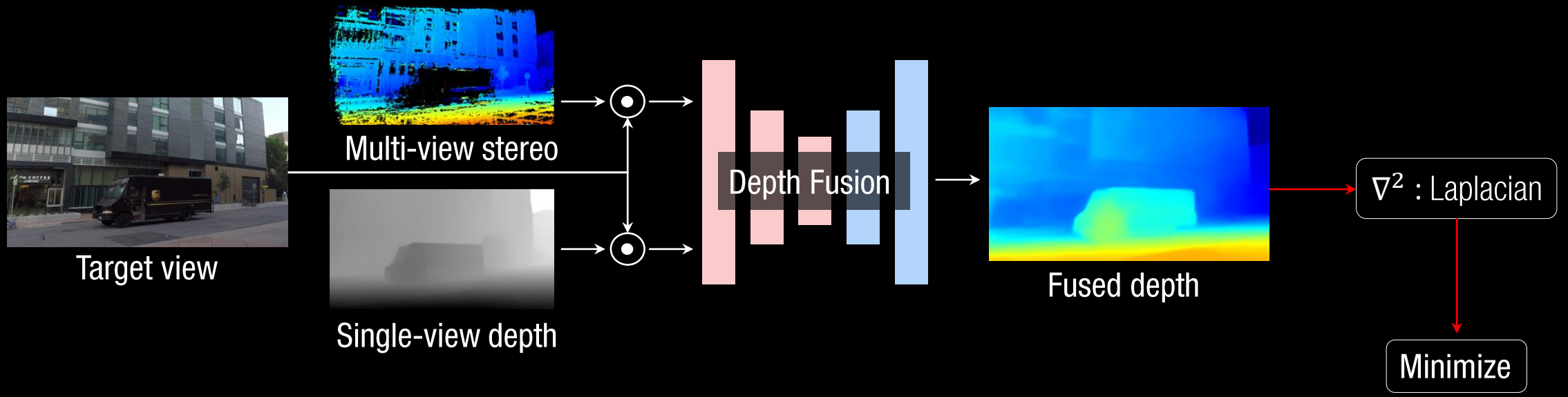


The relative depth of dynamic contents should be consistent with single view depth.

Depth Fusion Network

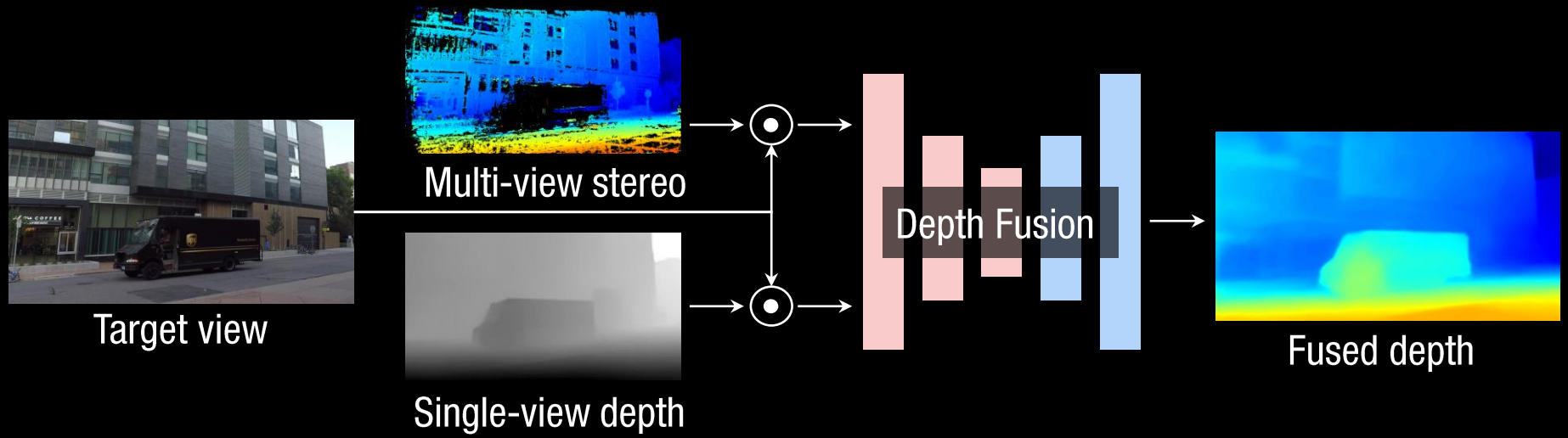


Depth Fusion Network

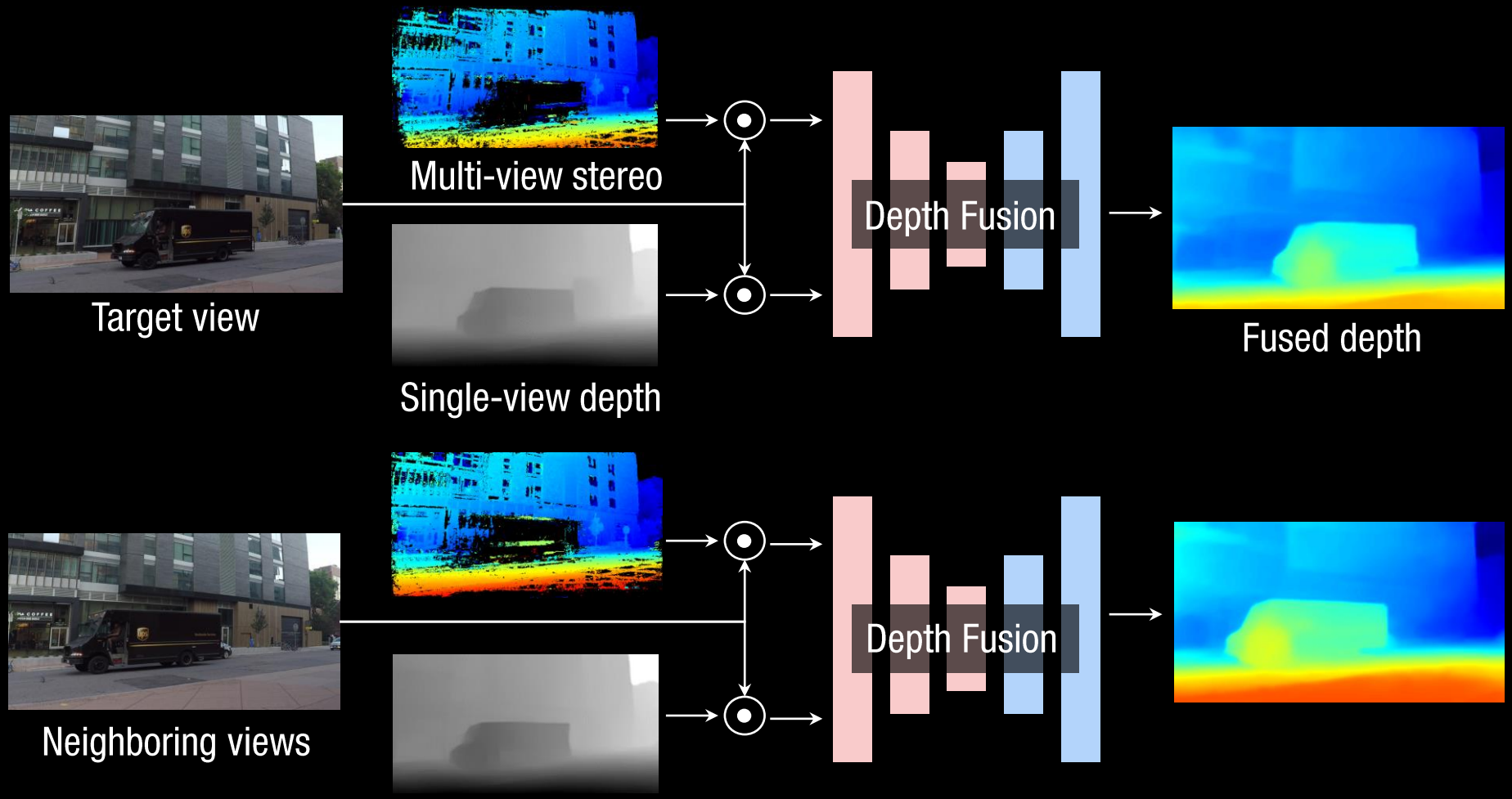


The output depth is spatially smooth.

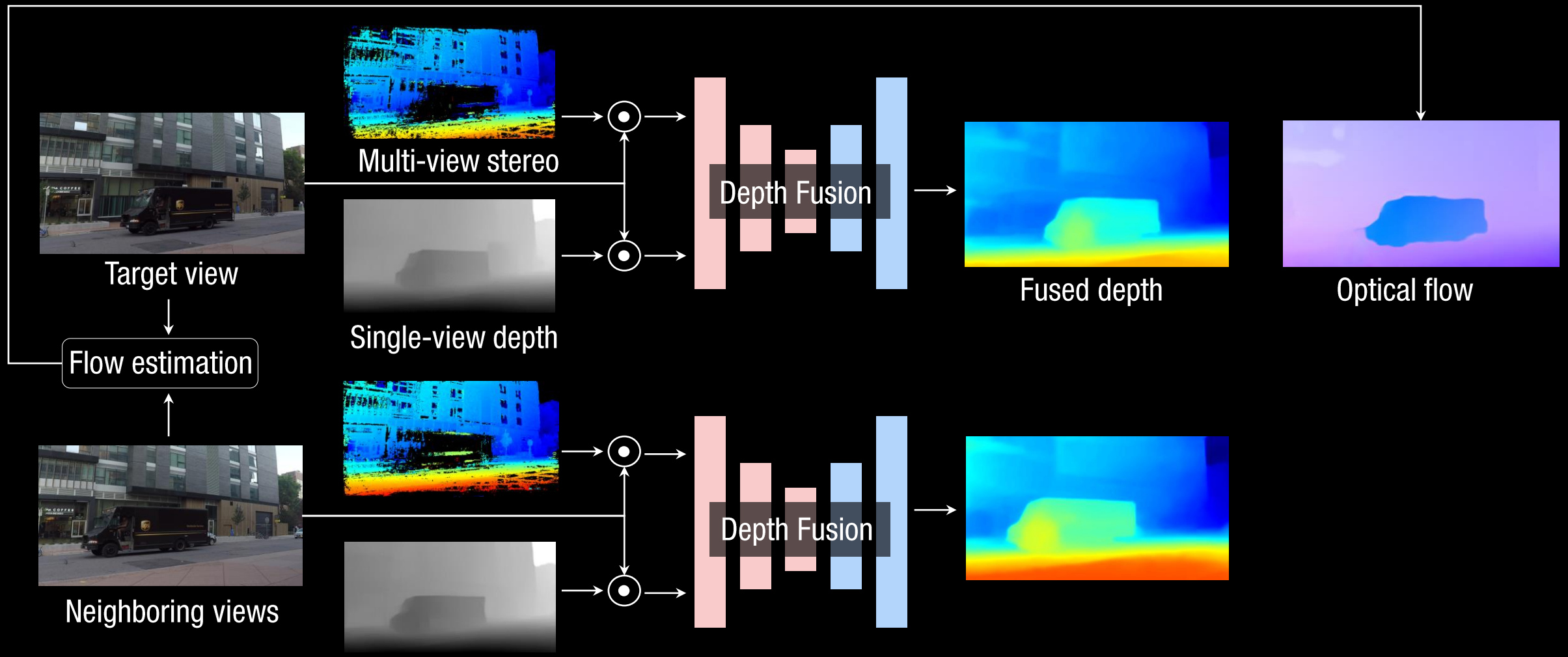
Depth Fusion Network



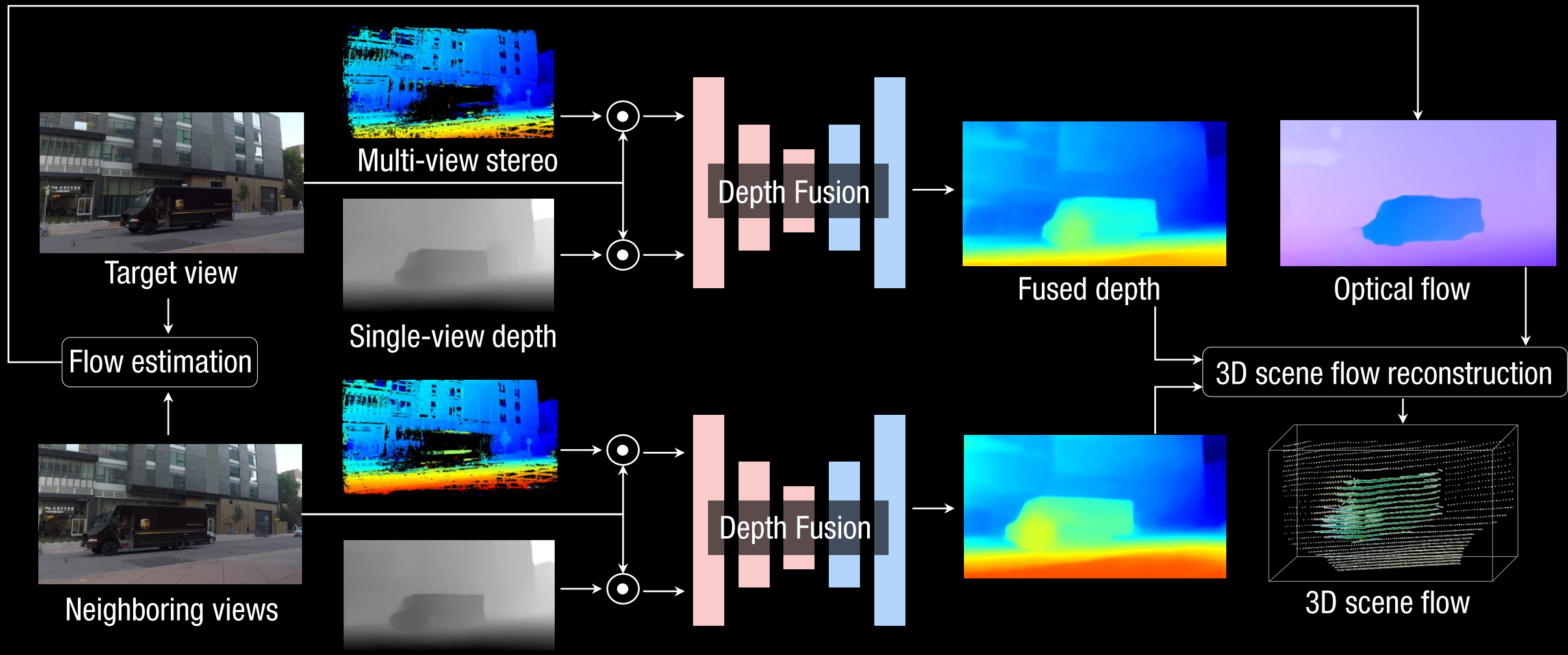
Depth Fusion Network



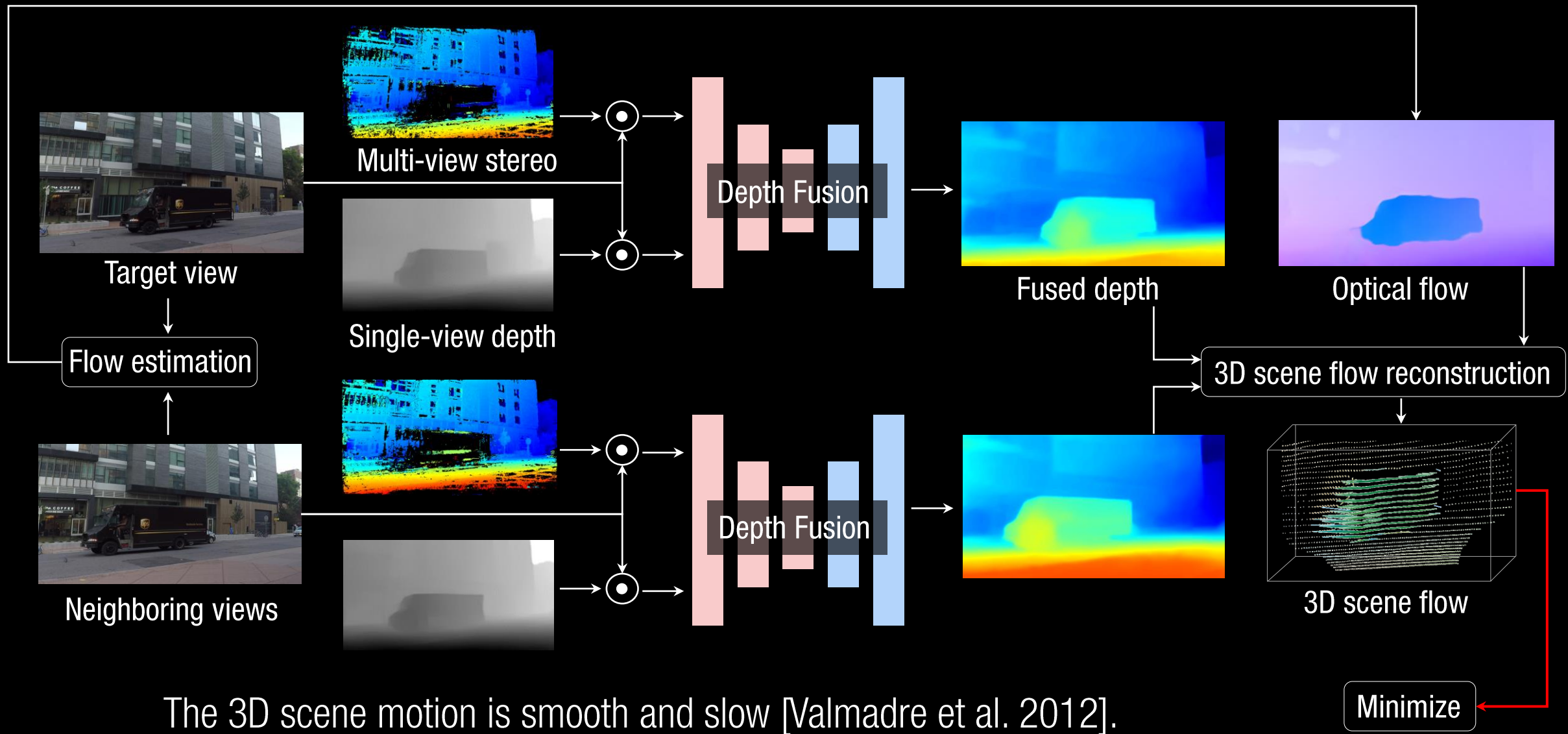
Depth Fusion Network



Depth Fusion Network



Depth Fusion Network



The 3D scene motion is smooth and slow [Valmadre et al. 2012].

Novel View Synthesis of Dynamic Scenes

Source cameras ◁

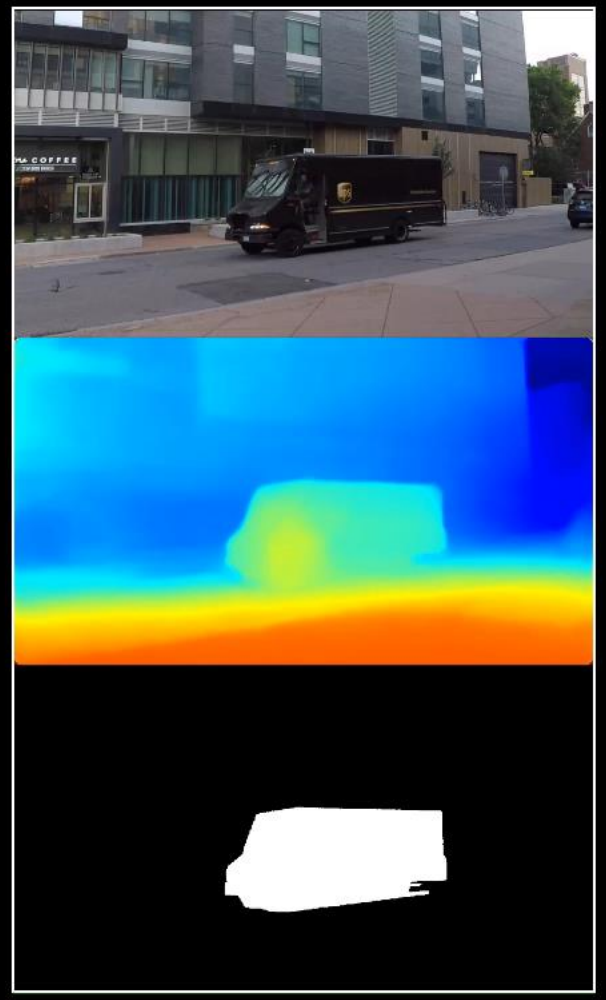


Image & depth & mask

Novel View Synthesis of Dynamic Scenes

Source cameras ◁ ▷ Virtual camera

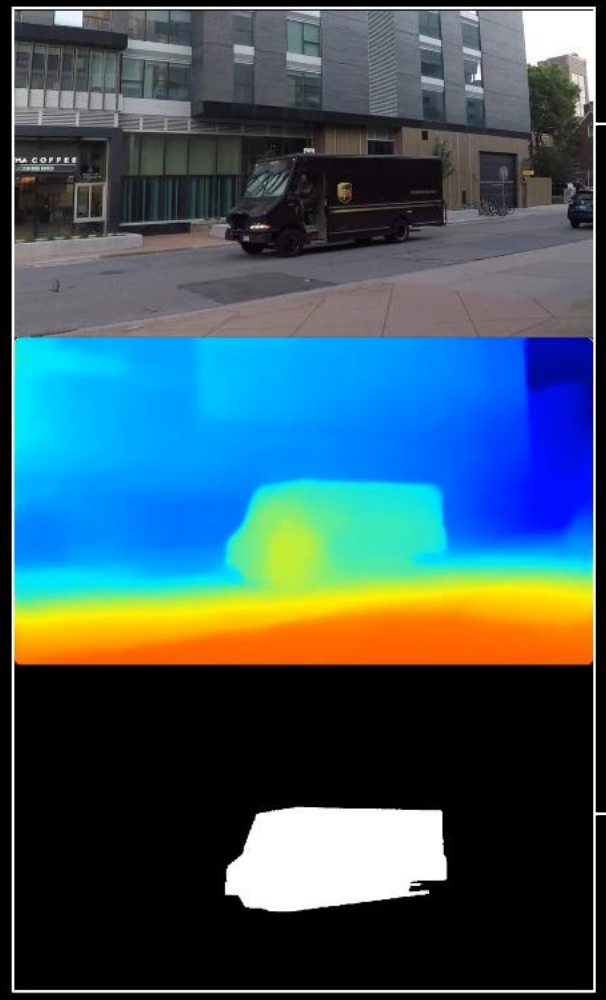


Image & depth & mask

Pixel transportation

Warped foreground

Warped background

Novel View Synthesis of Dynamic Scenes

Source cameras ◁ ▷ Virtual camera

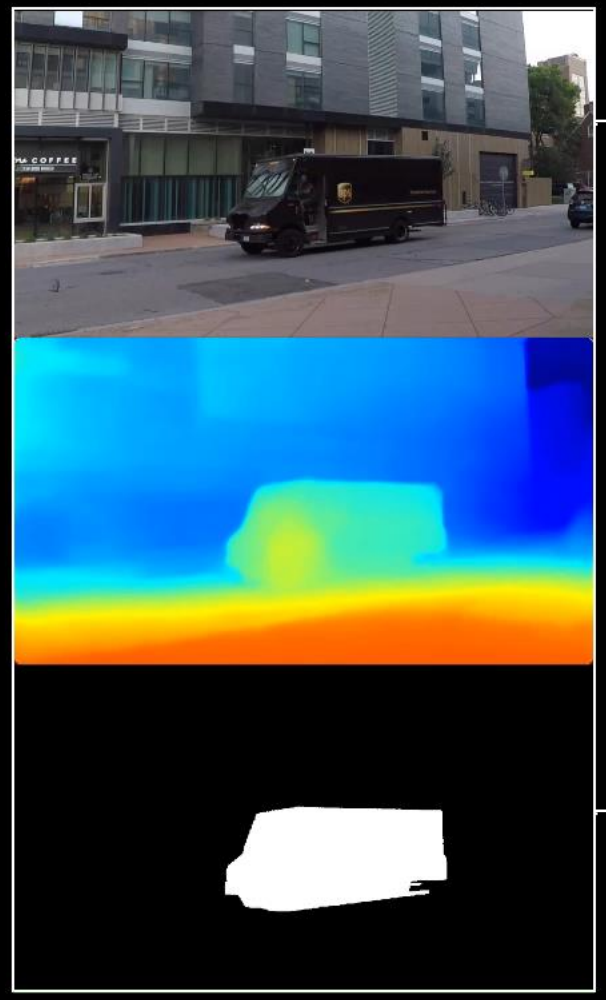


Image & depth & mask

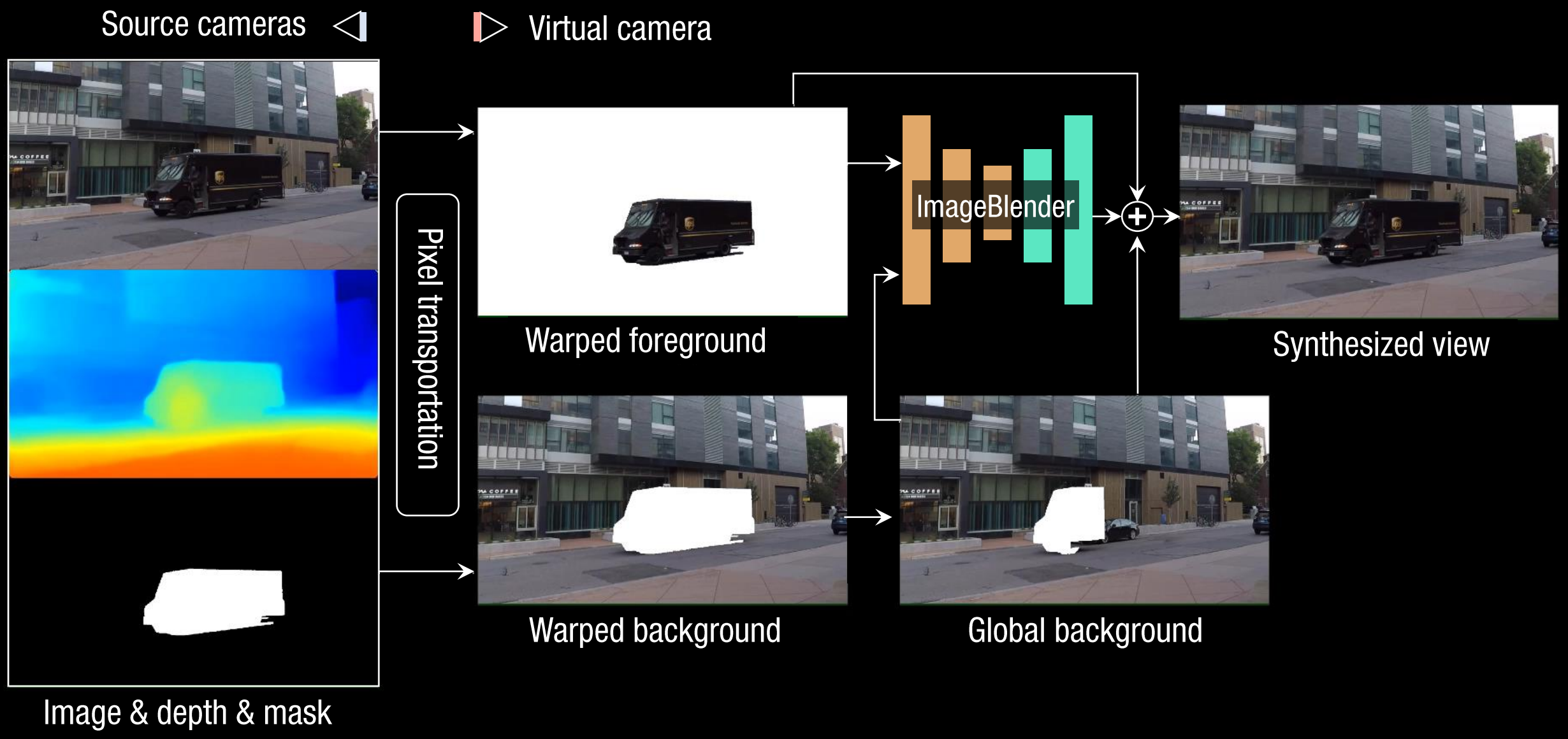
Pixel transportation

Warped foreground

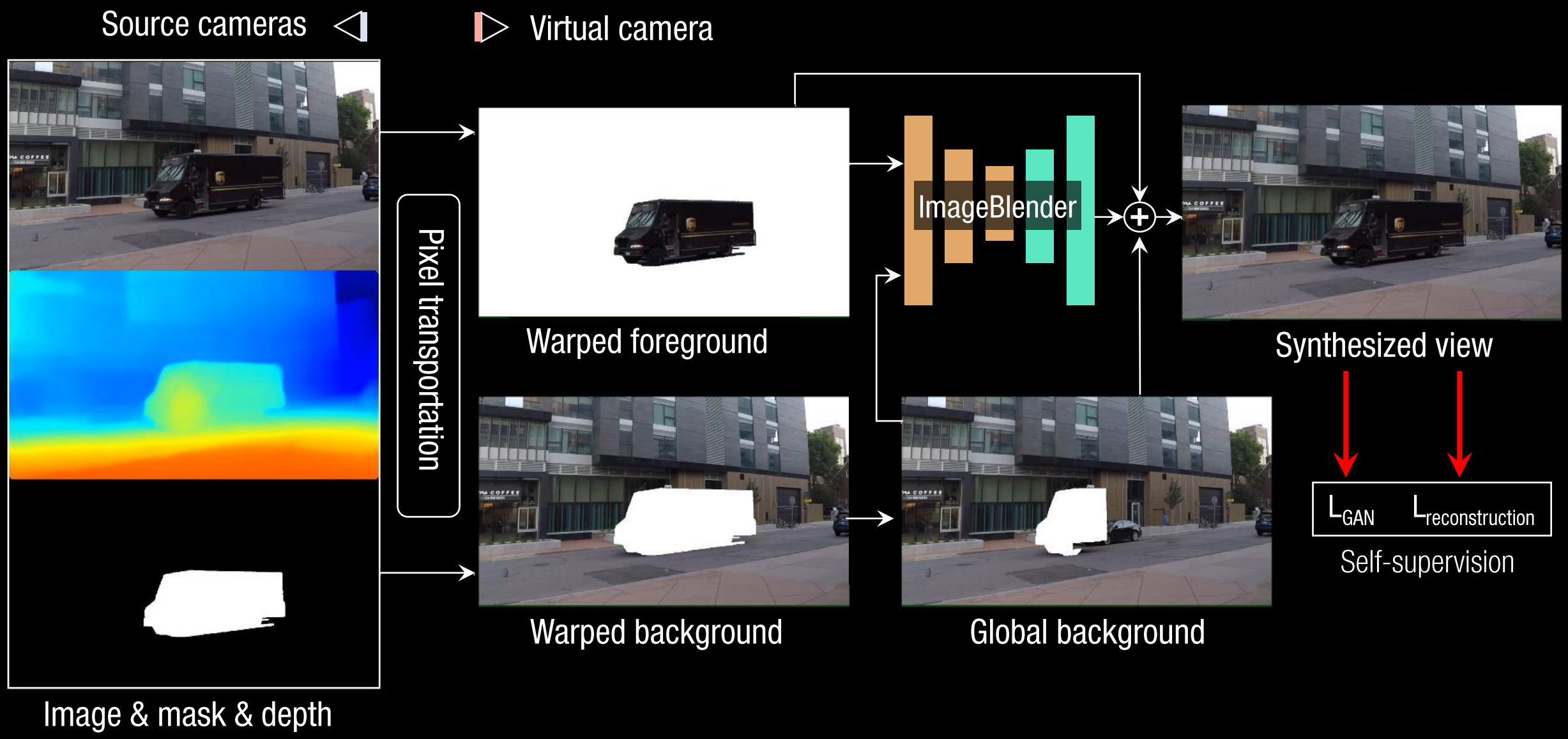
Warped background

Global background

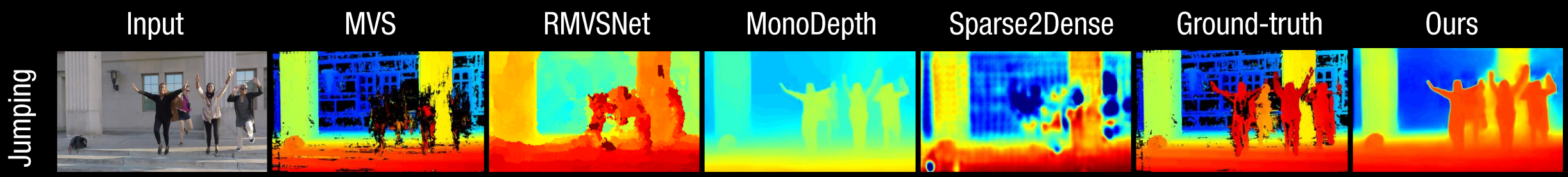
Novel View Synthesis of Dynamic Scenes



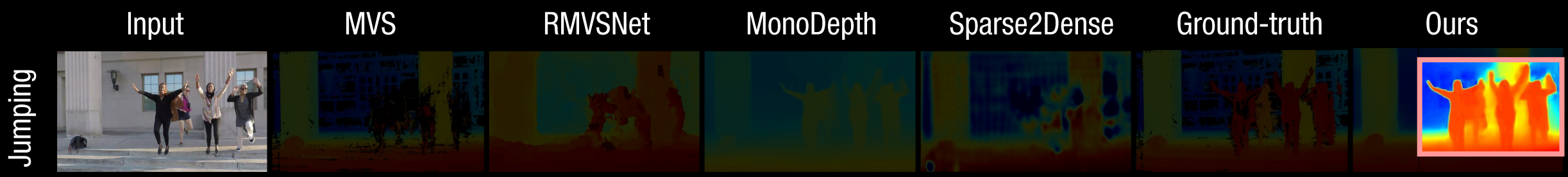
Novel View Synthesis of Dynamic Scenes

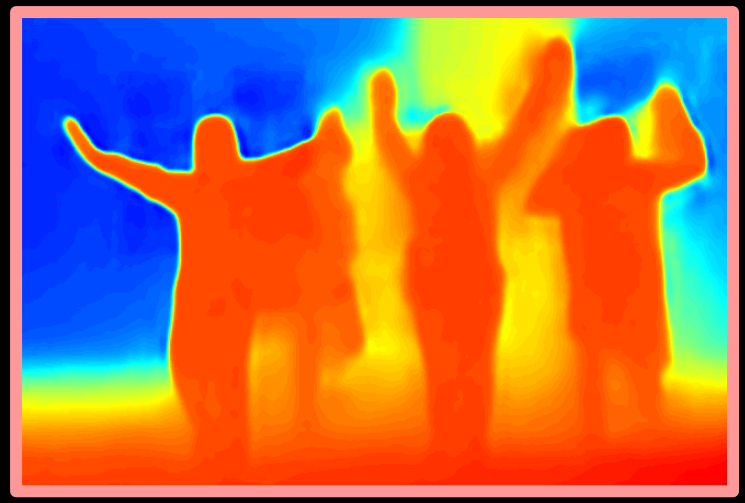


Experiments



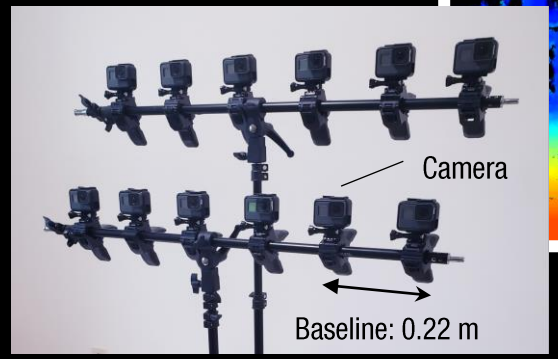
Experiments



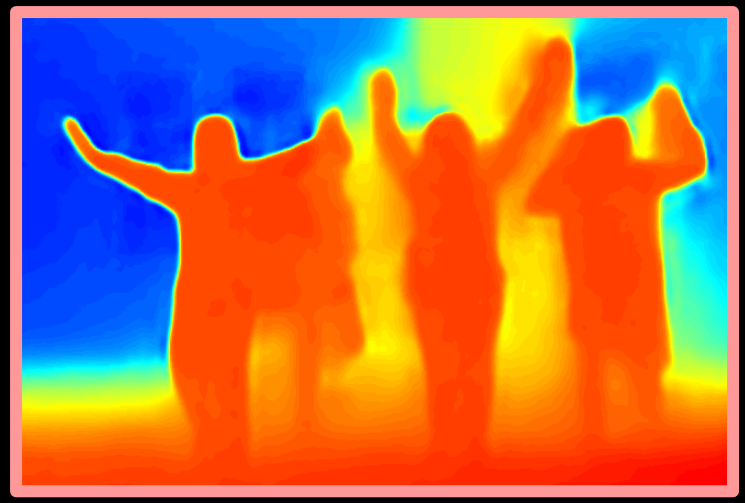


Ours (depth fusion)

Experiments

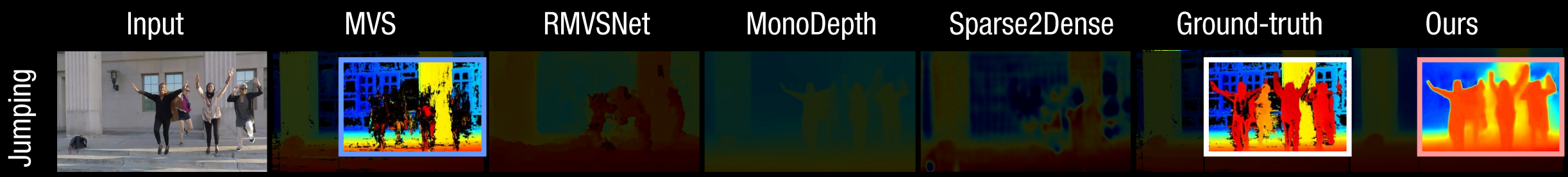


Ground-truth



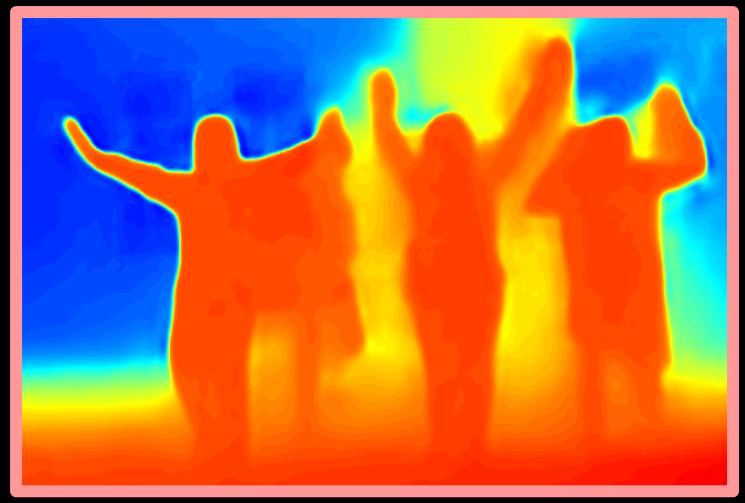
Ours (depth fusion)

Experiments



MVS

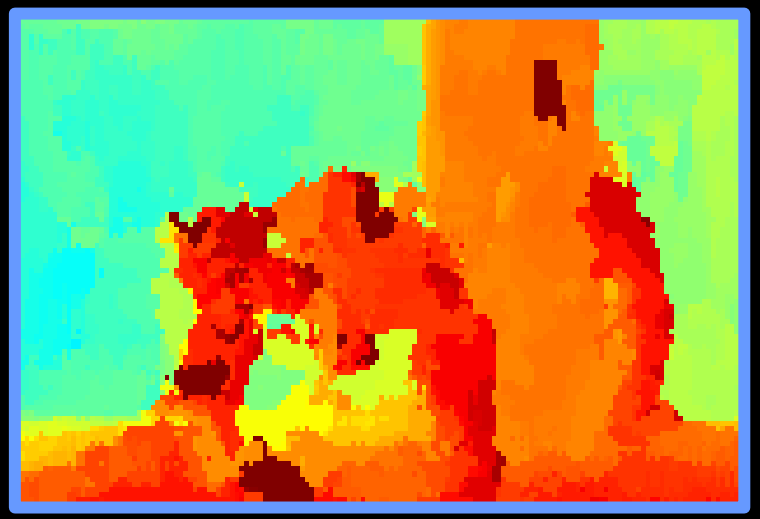
Ground-truth



Ours (depth fusion)

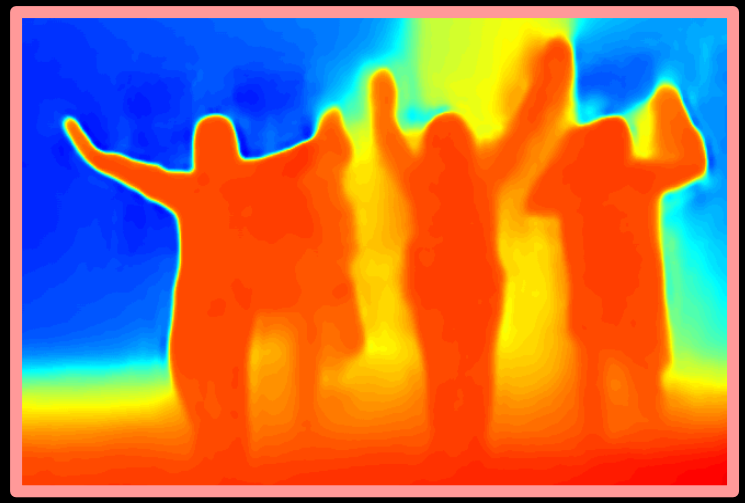
- MVS : Optimizaiton based multiview stereo [ECCV 2016 Schonberger et al.]

Experiments



RMVSNet

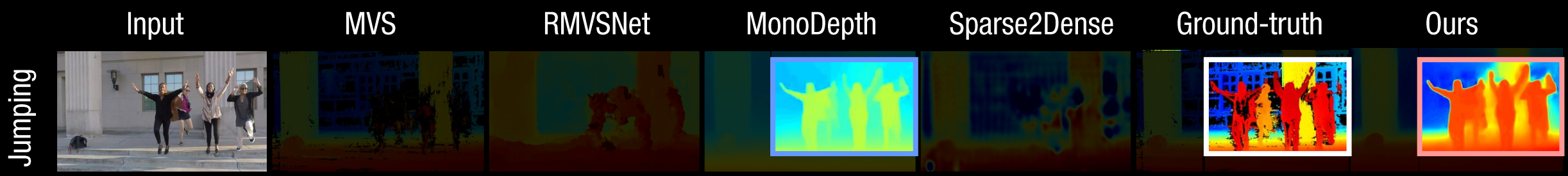
Ground-truth



Ours (depth fusion)

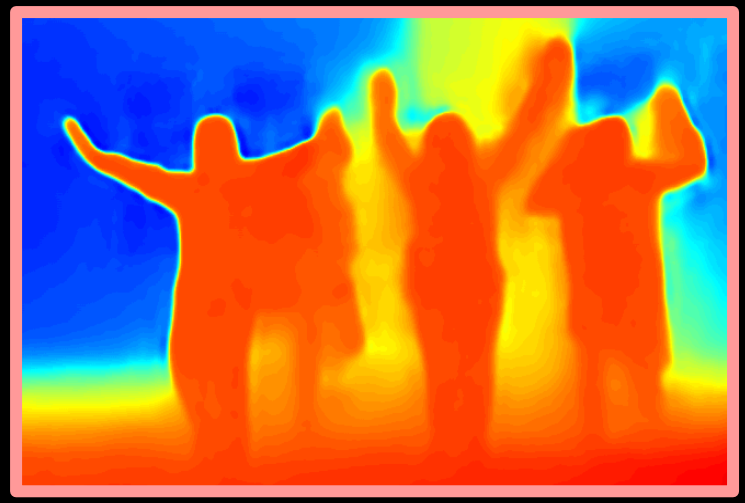
- RMVSNet : Learning based multiview stereo [CVPR 2019 Yao et al.]

Experiments



MonoDepth

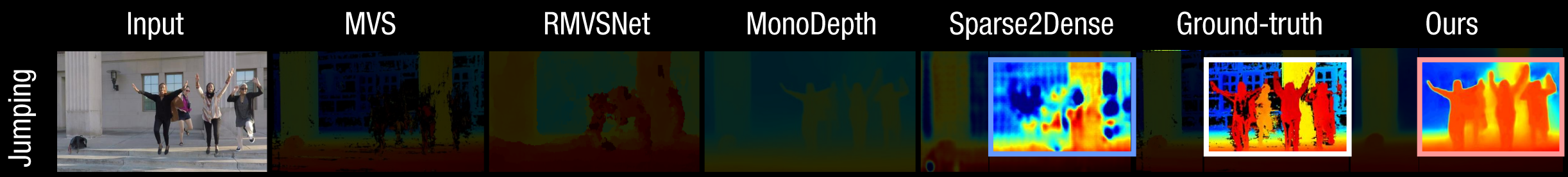
Ground-truth

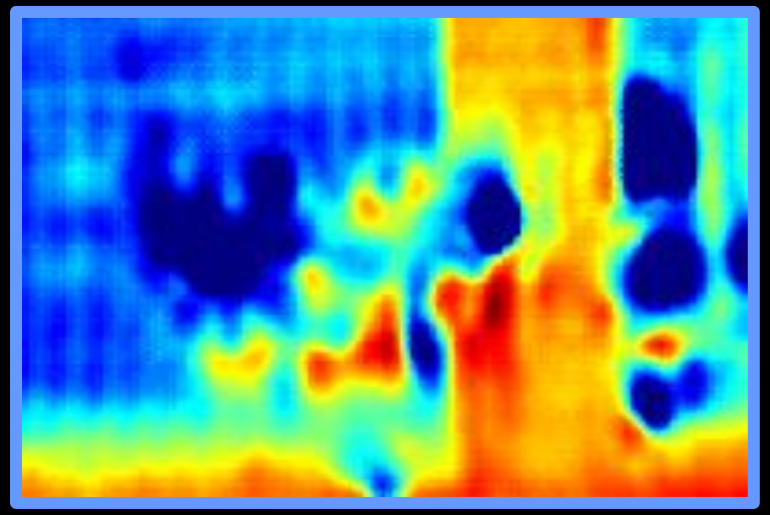


Ours (depth fusion)

- MonoDepth : Depth prediction from a single image [Arxiv 2019 Ranftl et al.]

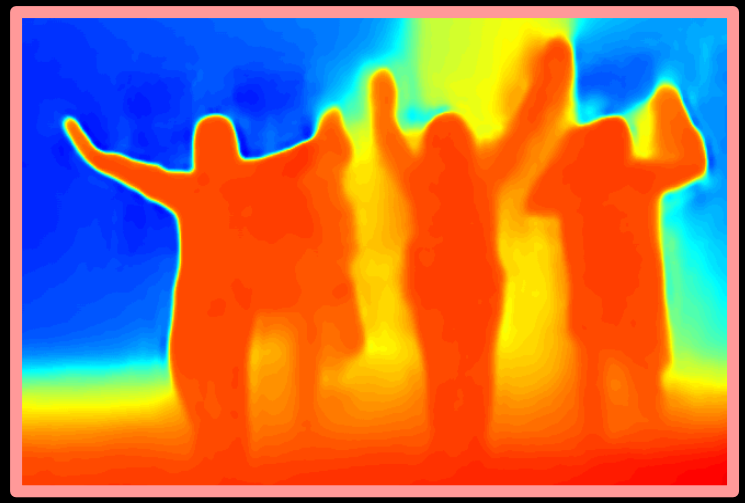
Experiments





Sparse2Dense

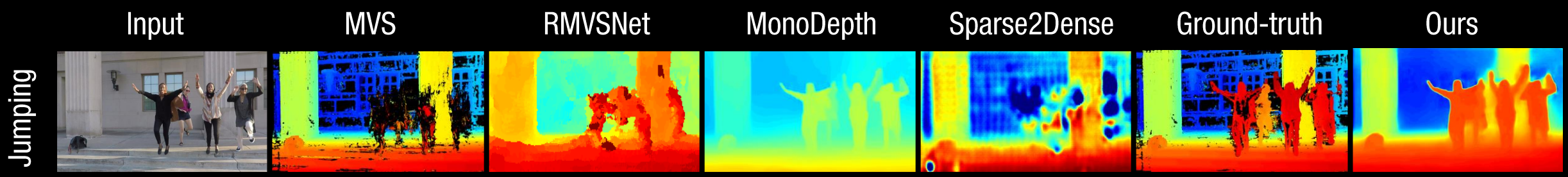
Ground-truth



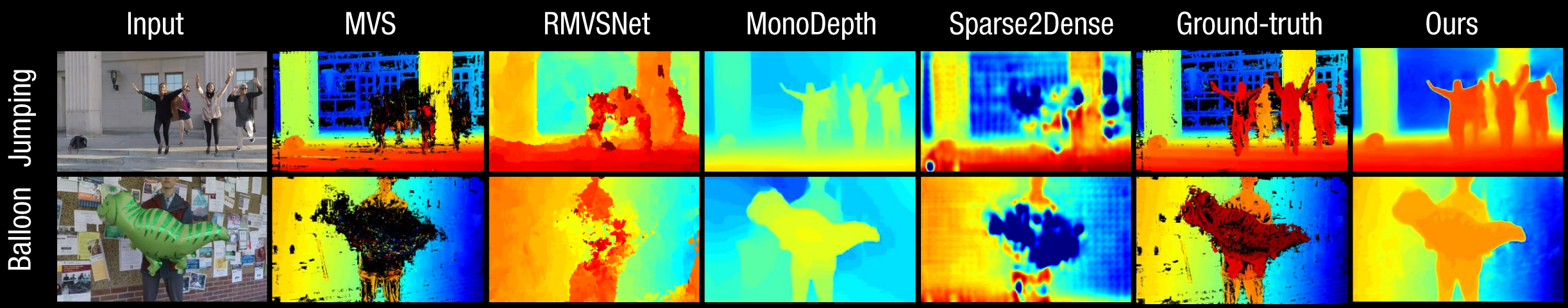
Ours (depth fusion)

- Spars2Dense : Depth completion from a sparse depth map [ICRA 2018 Mal et al.]

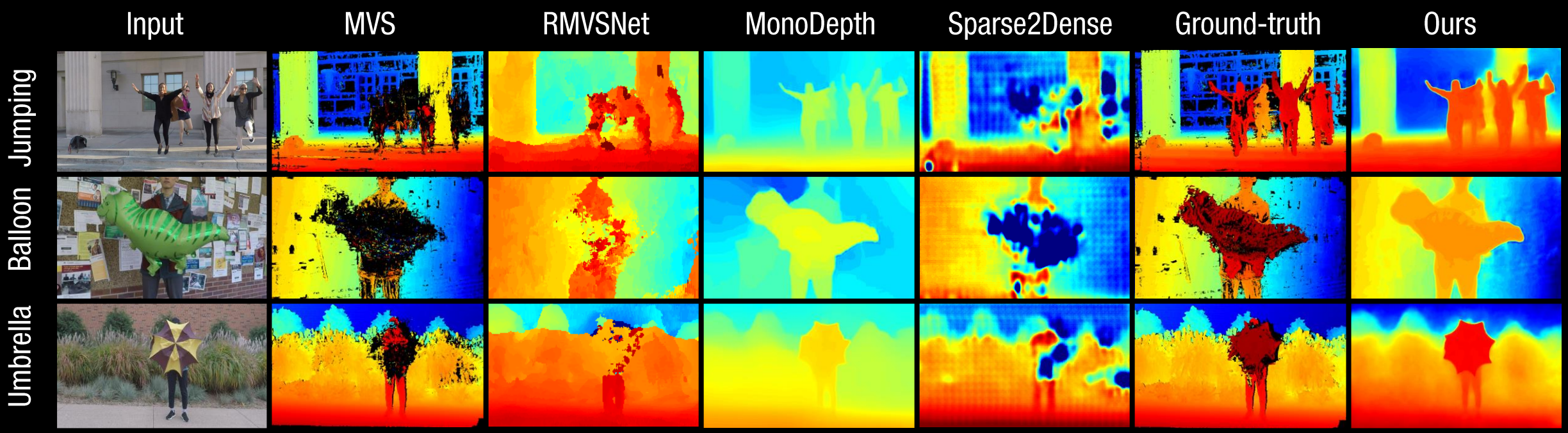
Experiments



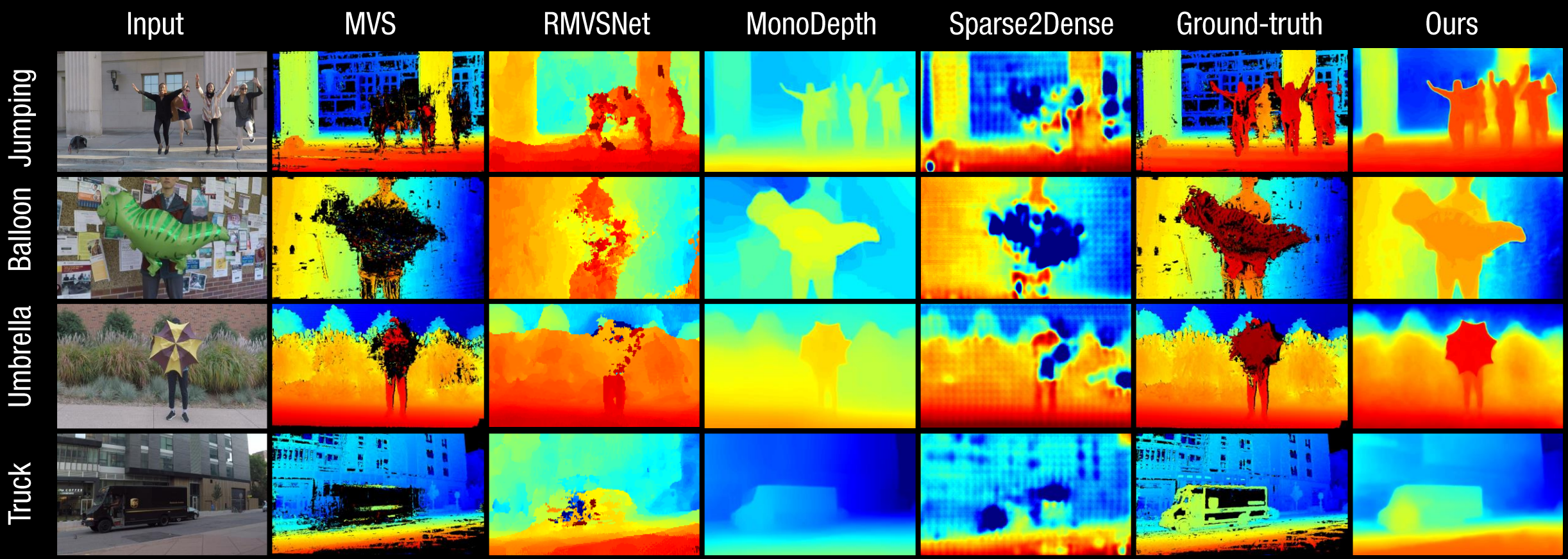
Experiments



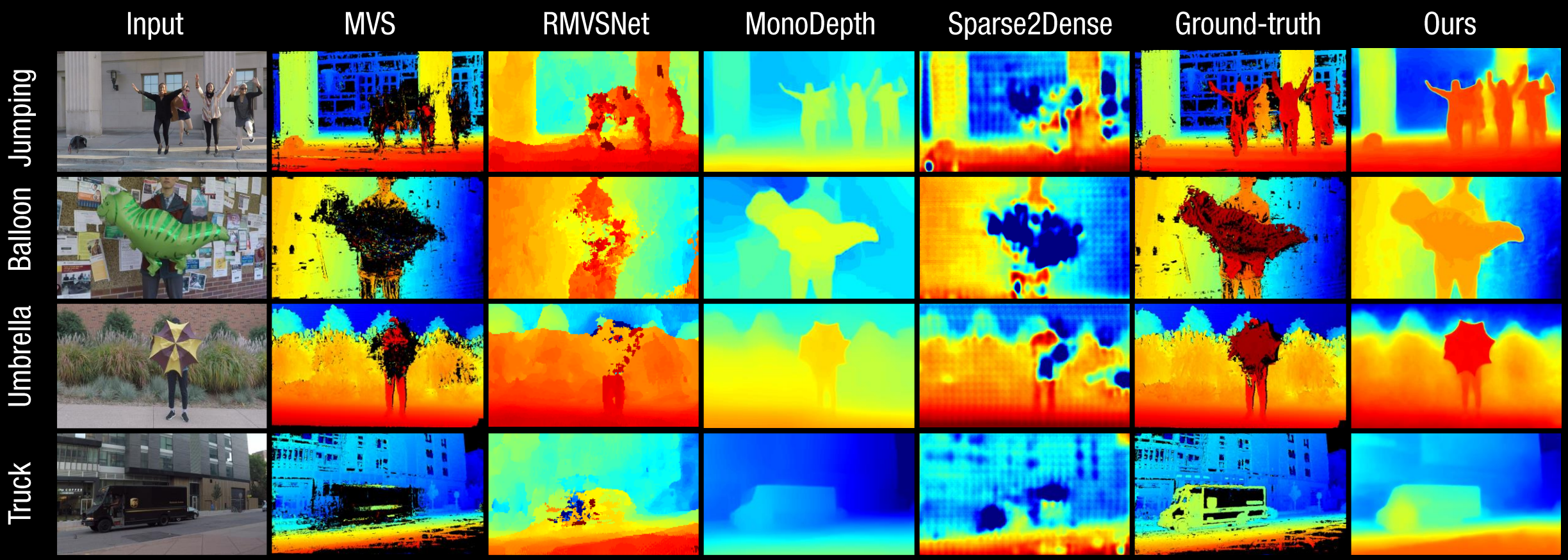
Experiments

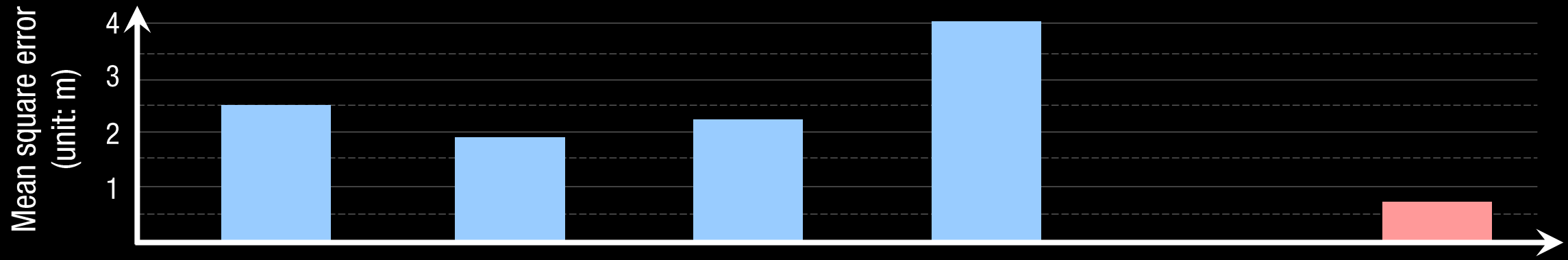


Experiments



Experiments



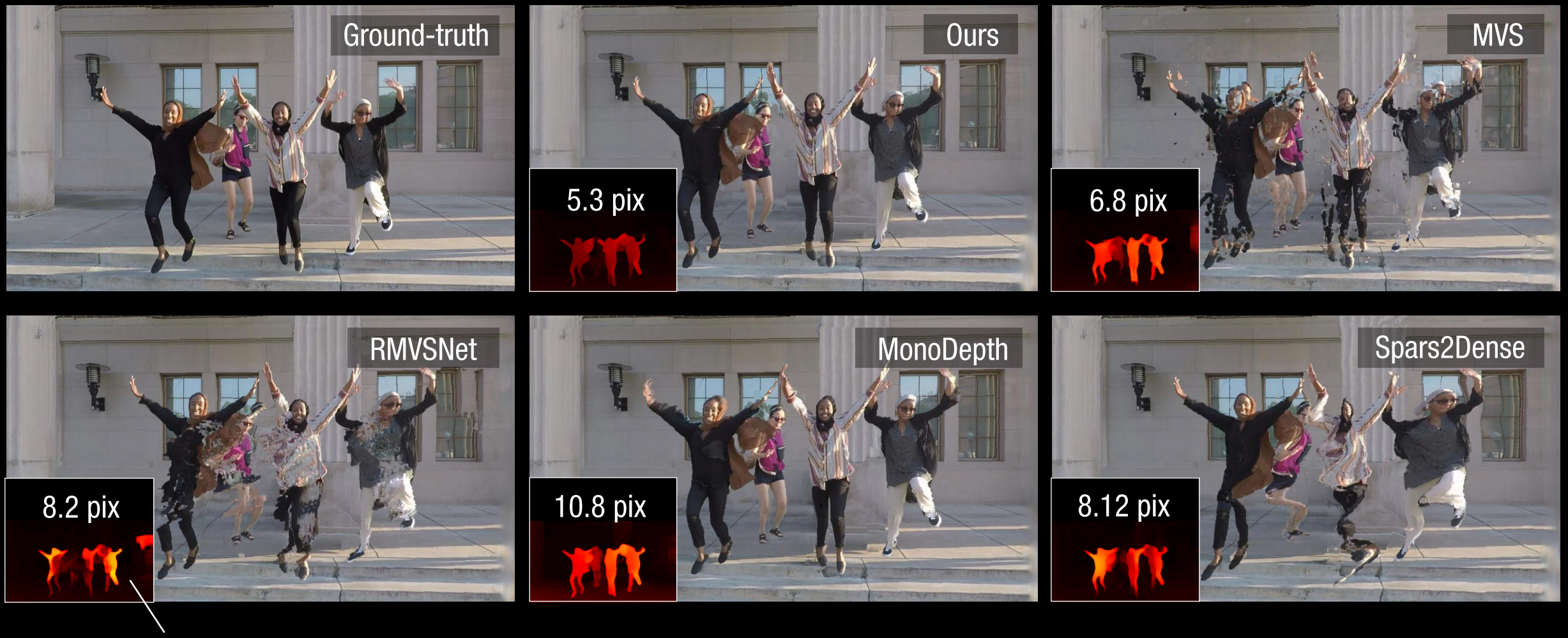


Experiments

Experiments

The magnitude of optical flow from the ground-truth to the synthesized image. (unit: pixel) 0 50

Experiments



The magnitude of optical flow from the ground-truth to the synthesized image. (unit: pixel) 0 50

Experiments

Experiments


Bullet time effect


Space time navigation

Customized cinemagraph

More Results

Yoon et al. “Novel View Synthesis of Dynamic Scenes with Globally Coherent Depths from a Monocular Camera." CVPR 2020.

Limitations

Input

Output

Erroneous mask
(fragmentation, afterimage)

Limitations

Input

Output

Erroneous mask
(fragmentation, afterimage)

Cluttered scene

Limitations

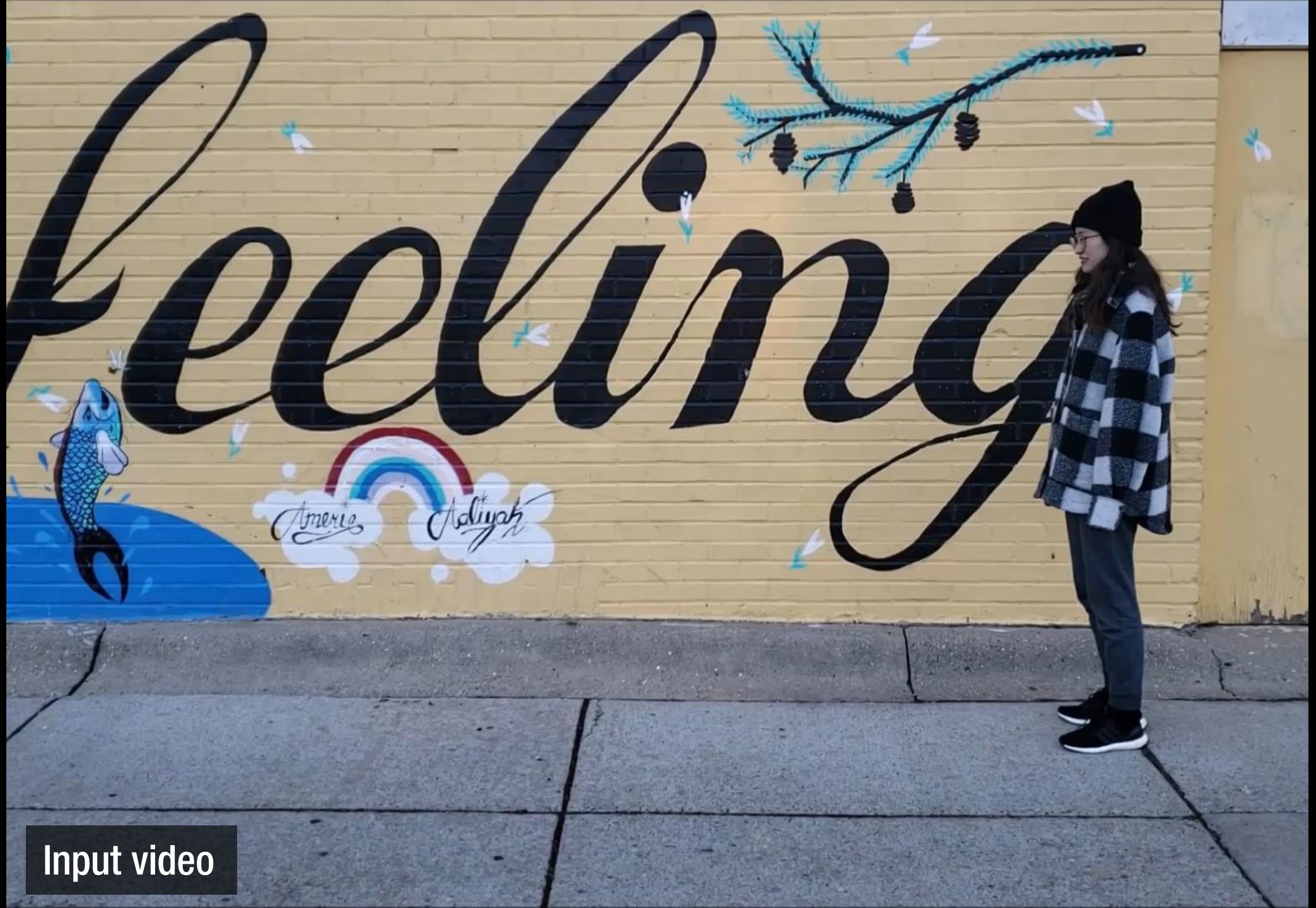
Input

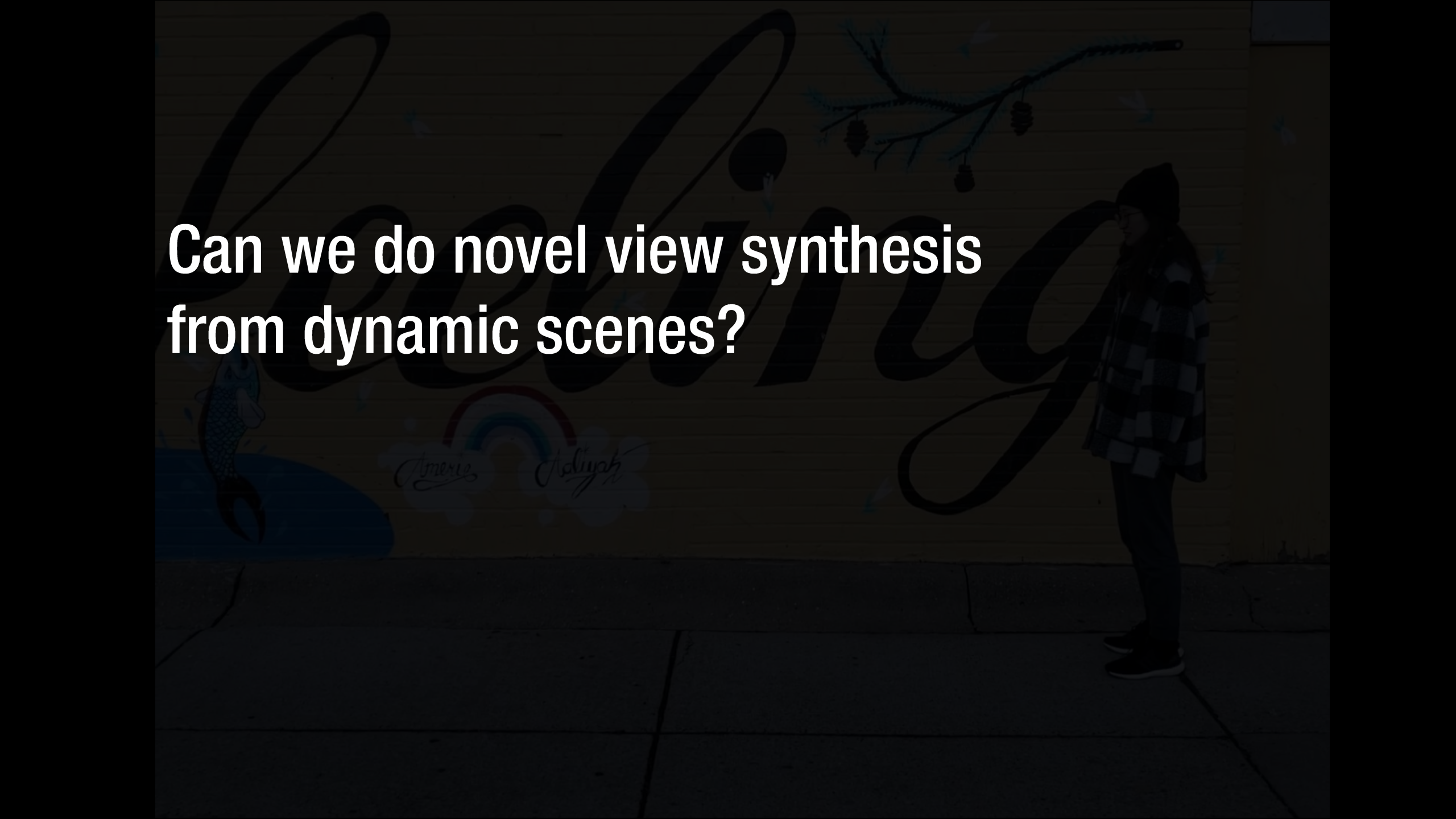
Output

Erroneous mask
(fragmentation, afterimage)

Cluttered scene

Large viewing angle


feeling
Amerie
Aaliyah
Input video


Can we do novel view synthesis
from dynamic scenes?


Input dynamic scene
Dynamic scene view synthesis
Bullet time effect
Space-time navigation

Thank you

Jae Shin Yoon

University of Minnesota

CVPR 2020 Tutorial

Poster 2.1-#48, Wed, 10:00, 22:00

Novel View Synthesis of Dynamic Scenes With Globally Coherent Depths From a Monocular Camera

Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, Jan Kautz