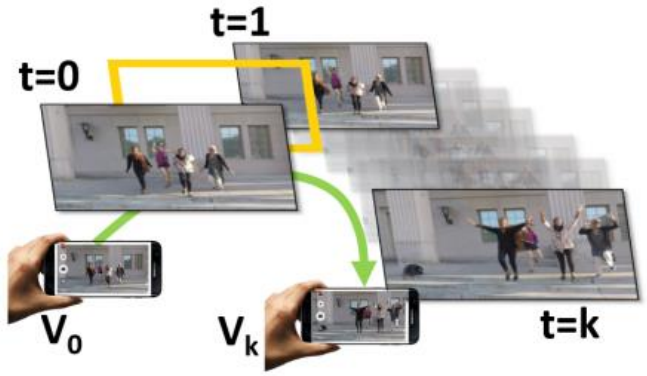
CVPR SEATTLE WASHINGTON JUNE 16-18 2020

# Novel View Synthesis of Dynamic Scenes with Globally Coherent Depths from a Monocular Camera

Jae Shin Yoon[1], Kihwan Kim[2], Orazio Gallo[2], Hyun Soo Park[1], and Jan Kautz[2]




1 University of Minnesota

2 NVIDIA

# Motivation

Input: Images captured from a dynamic scene

# Motivation

Input: Images captured from a dynamic scene

Output: Rendered image at arbitrary views and times


Bullet time effect


Space time navigation

Customized cinemagraph

# Related Work


Flynn at al. 2016


Kalantari et al. 2016

Zhou et al. 2018


Mildenhall et al. 2019


Penner et al. 2019


Mildenhall et al. 2019


Choi et al. 2019

Static scene assumption

# Related Work


Zitnick et al. 2004

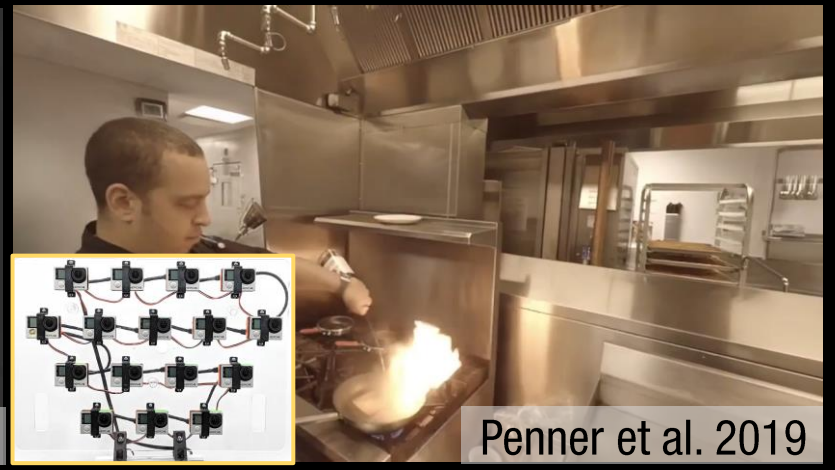

Lipski et al. 2009


Penner et al. 2019

# Related Work


Zitnick et al. 2004

Lipski et al. 2009



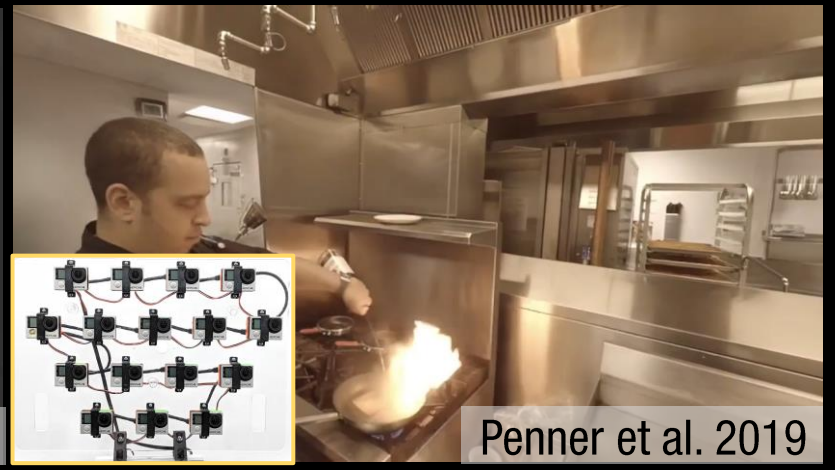
Penner et al. 2019

Requirement of multiview system

# Related Work

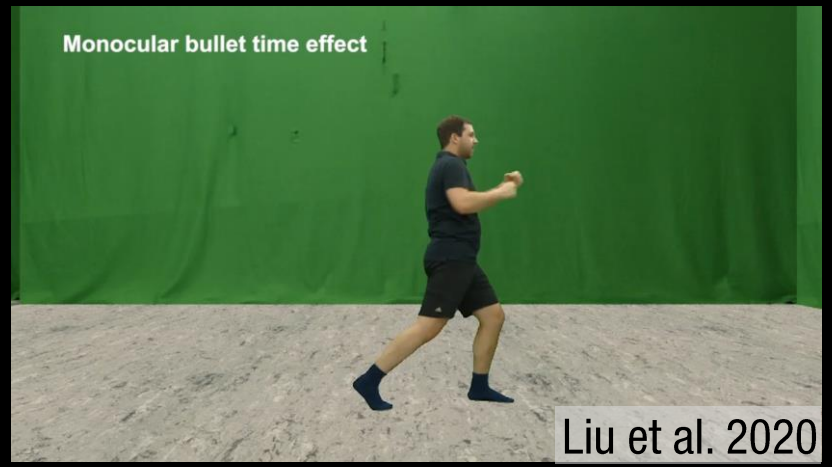
Zitnick et al. 2004


Lipski et al. 2009


Penner et al. 2019

Requirement of multiview system


Li et al. 2019



Liu et al. 2020

Requirement of human-specific priors

# Ours

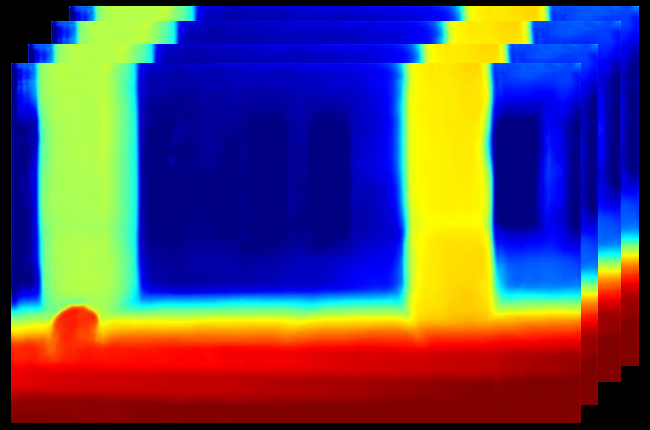

Novel view synthesis of class-agnostic dynamic scene using monocular camera

# Challenge

Images of a static scene

# Challenge

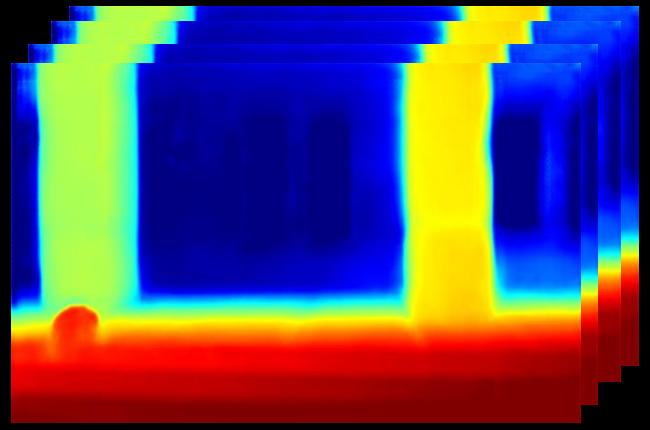
Images of a static scene



Depth estimation

# Challenge

Images of a static scene

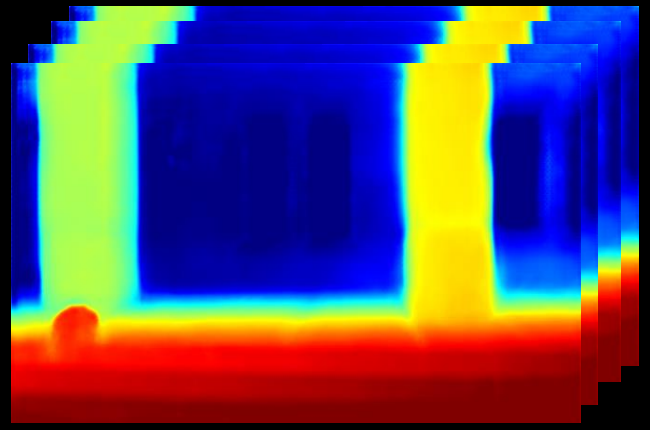


Depth estimation

Pixel transportation to a virtual view

# Challenge

Images of a static scene



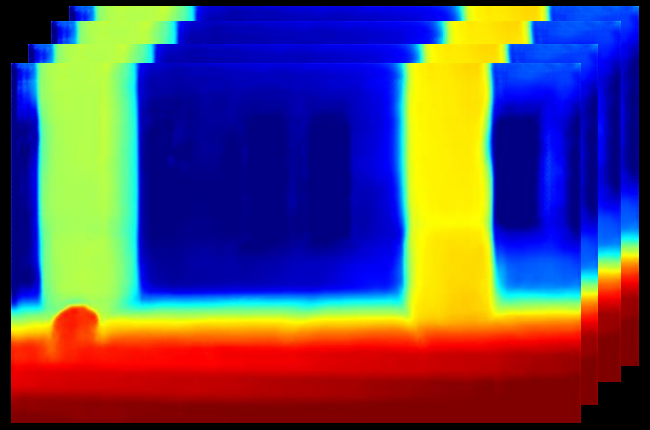
Depth estimation

Pixel transportation to a virtual view

Images of a dynamic scene

# Challenge

Images of a static scene



Depth estimation

Pixel transportation to a virtual view

Images of a dynamic scene

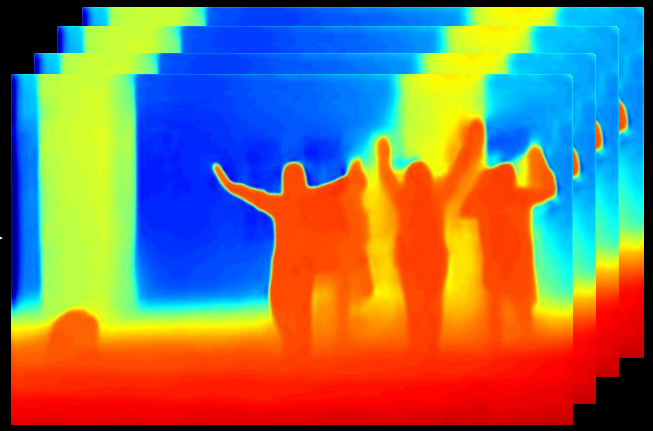
Depth estimation

Pixel transportation to a virtual view

# Challenge

How can we get the depth map that is complete and scale-invariant across the views from dynamic scene?

Images of a dynamic scene



Depth estimation

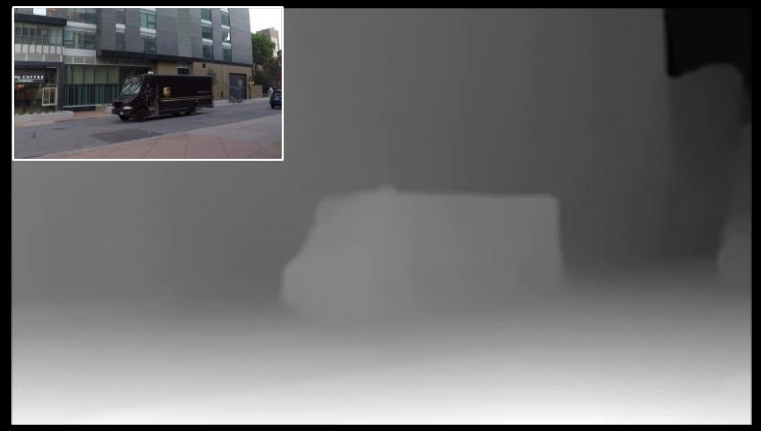
Pixel transportation to a virtual view

# Overview

Depth from multiview stereo
(scale-invariant, incomplete)

# Overview

Depth from multiview stereo
(scale-invariant, incomplete)



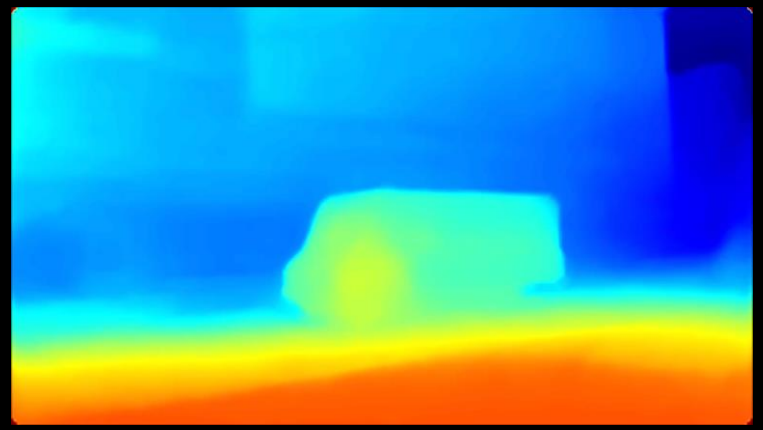
Depth from single view prediction
(scale-variant, complete)

# Overview

Depth from multiview stereo
(scale-invariant, incomplete)

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

Depth Fusion

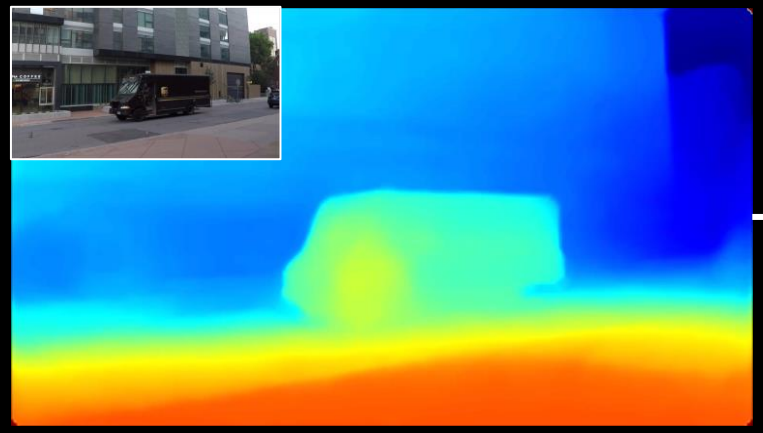


Fused depth
(scale-invariant, complete)

# Overview

Depth from multiview stereo
(scale-invariant, incomplete)

Depth Fusion



Fused depth
(scale-invariant, complete)

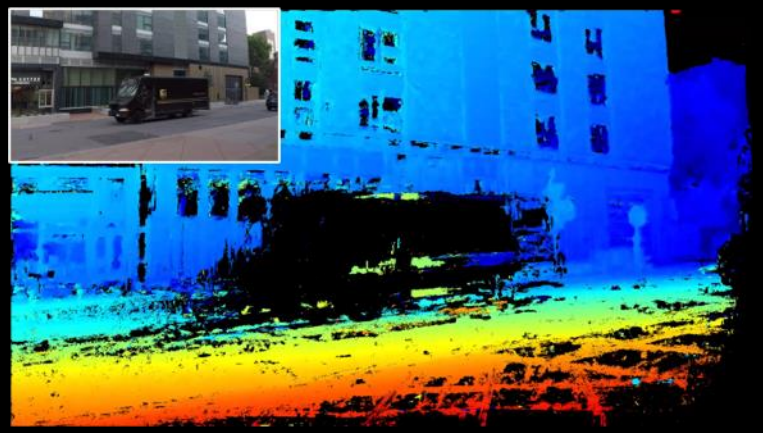
Image warping

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

View synthesis from a virtual view

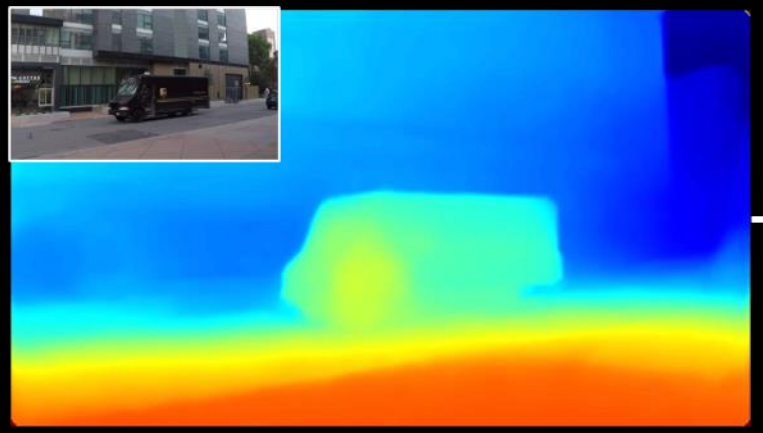
ImageBlender

# Overview



Depth from multiview stereo
(scale-invariant, incomplete)

Depth Fusion



Fused depth
(scale-invariant, complete)

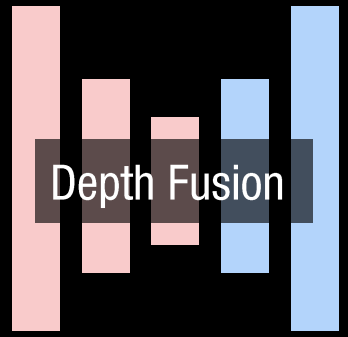
Image warping

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

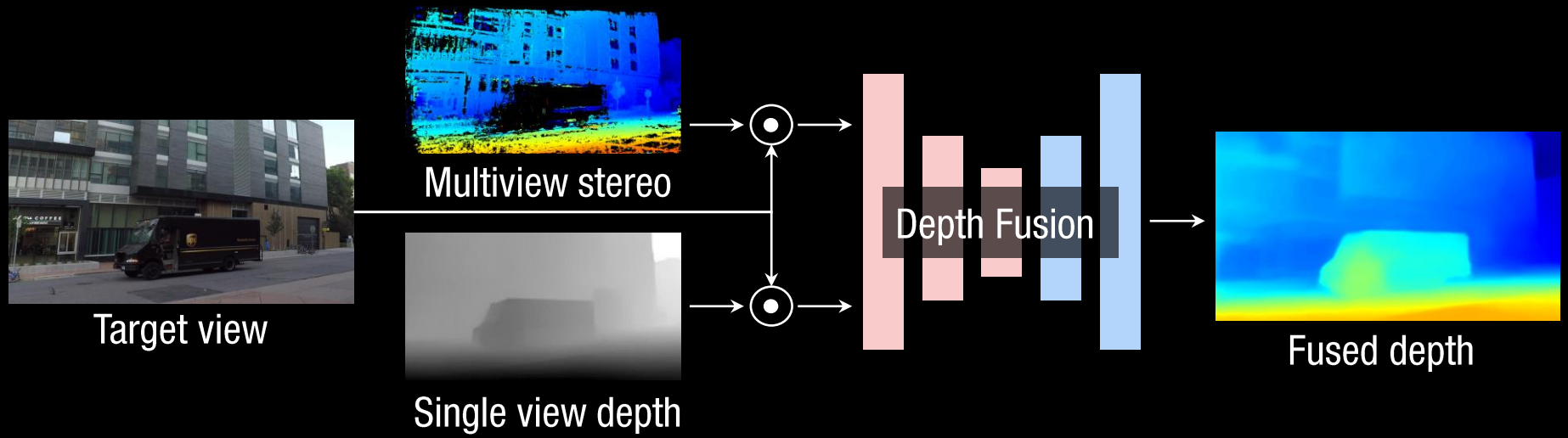
View synthesis from a virtual view

ImageBlender

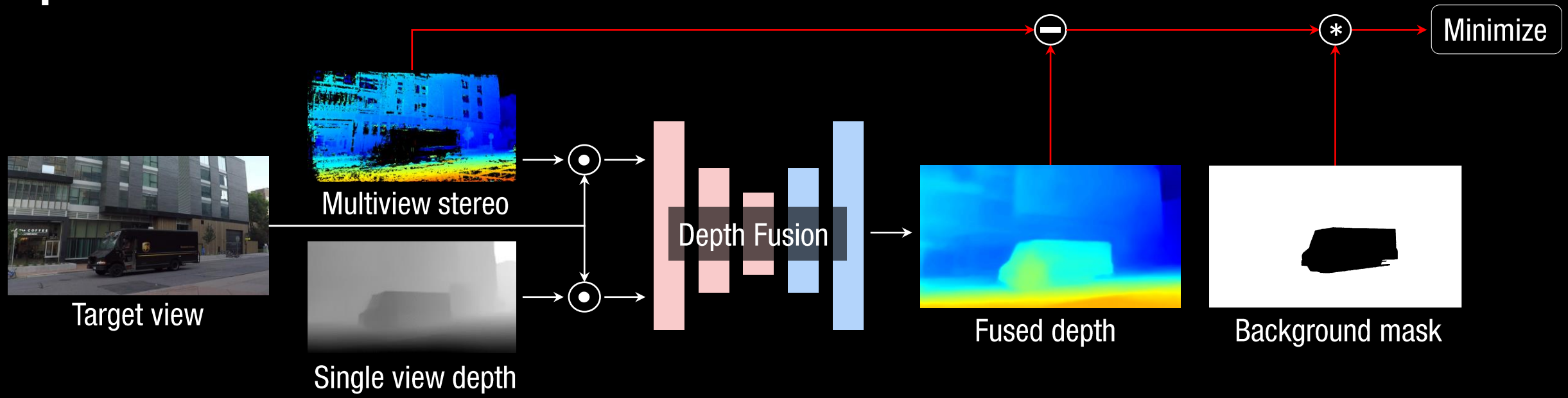
# Depth Fusion Network



# Depth Fusion Network

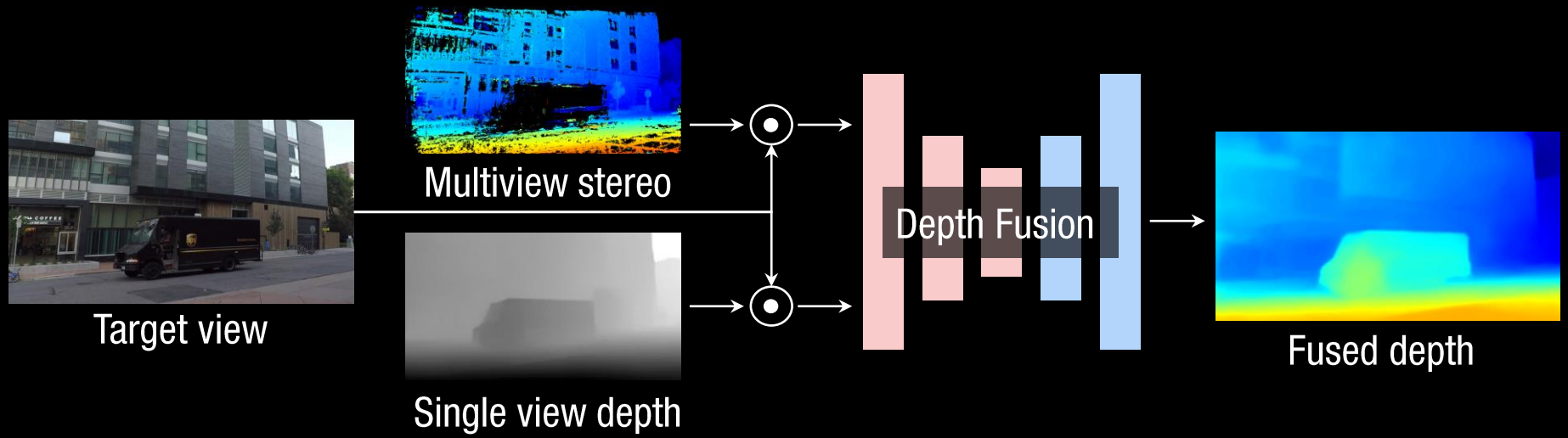


# Depth Fusion Network

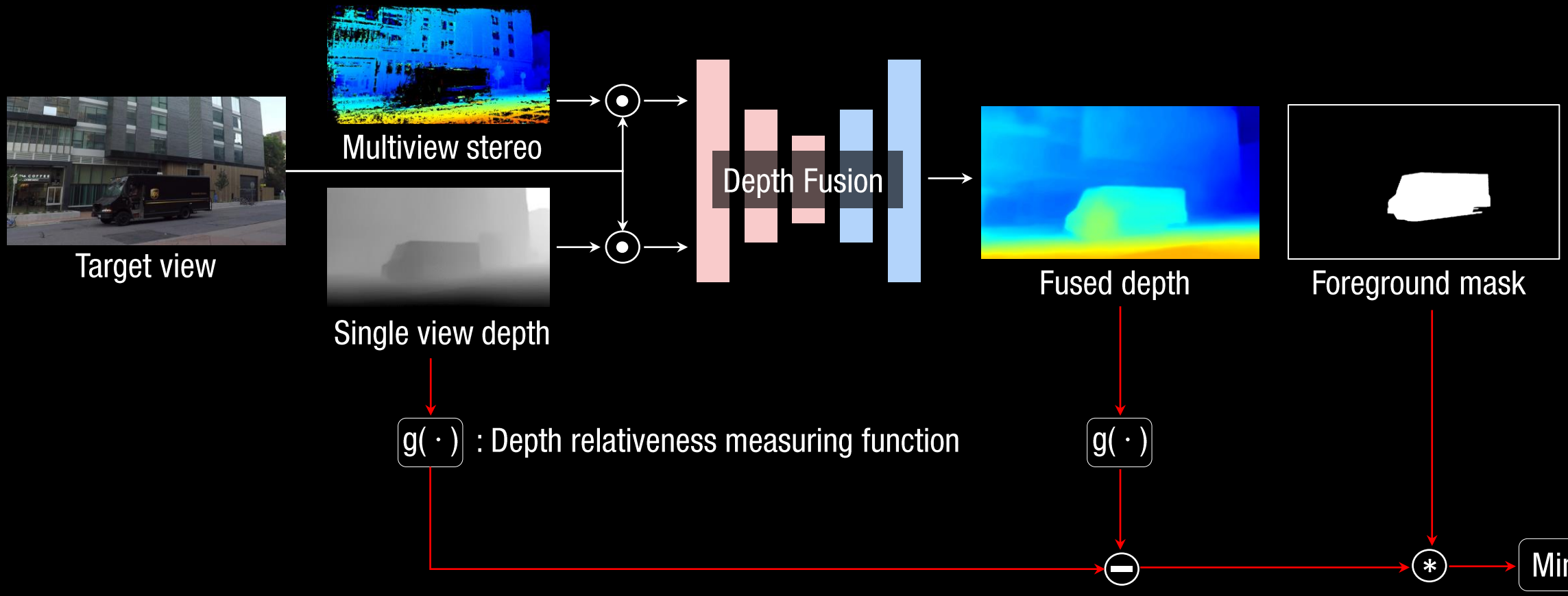



The estimated depth from static region must be aligned with multiview stereo depth.

# Depth Fusion Network

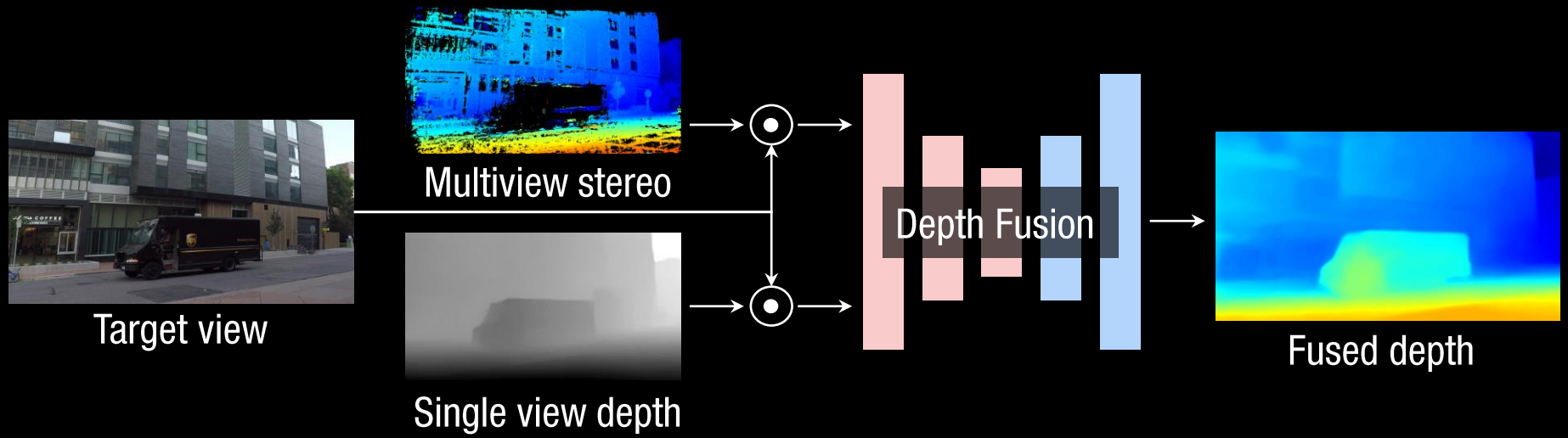


# Depth Fusion Network

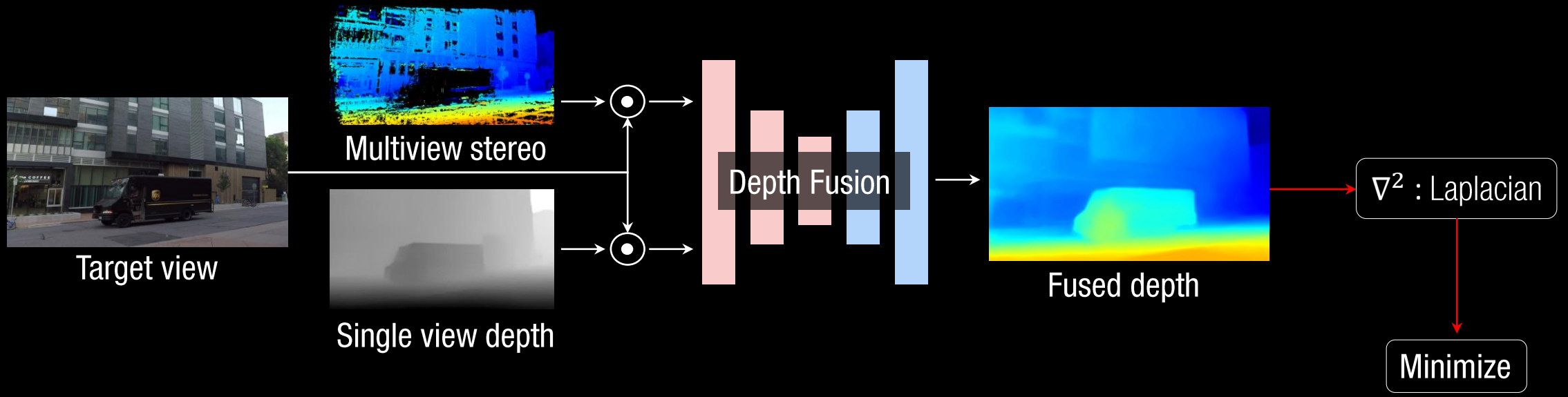



The depth relativeness of dynamic contents should be consistent with single view depth.

# Depth Fusion Network

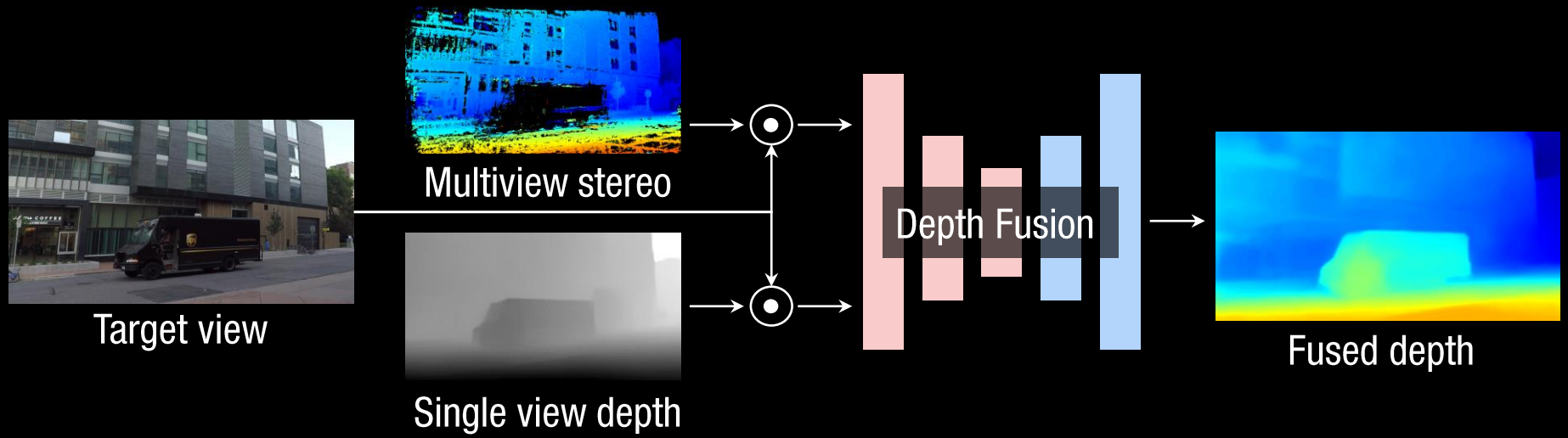


# Depth Fusion Network

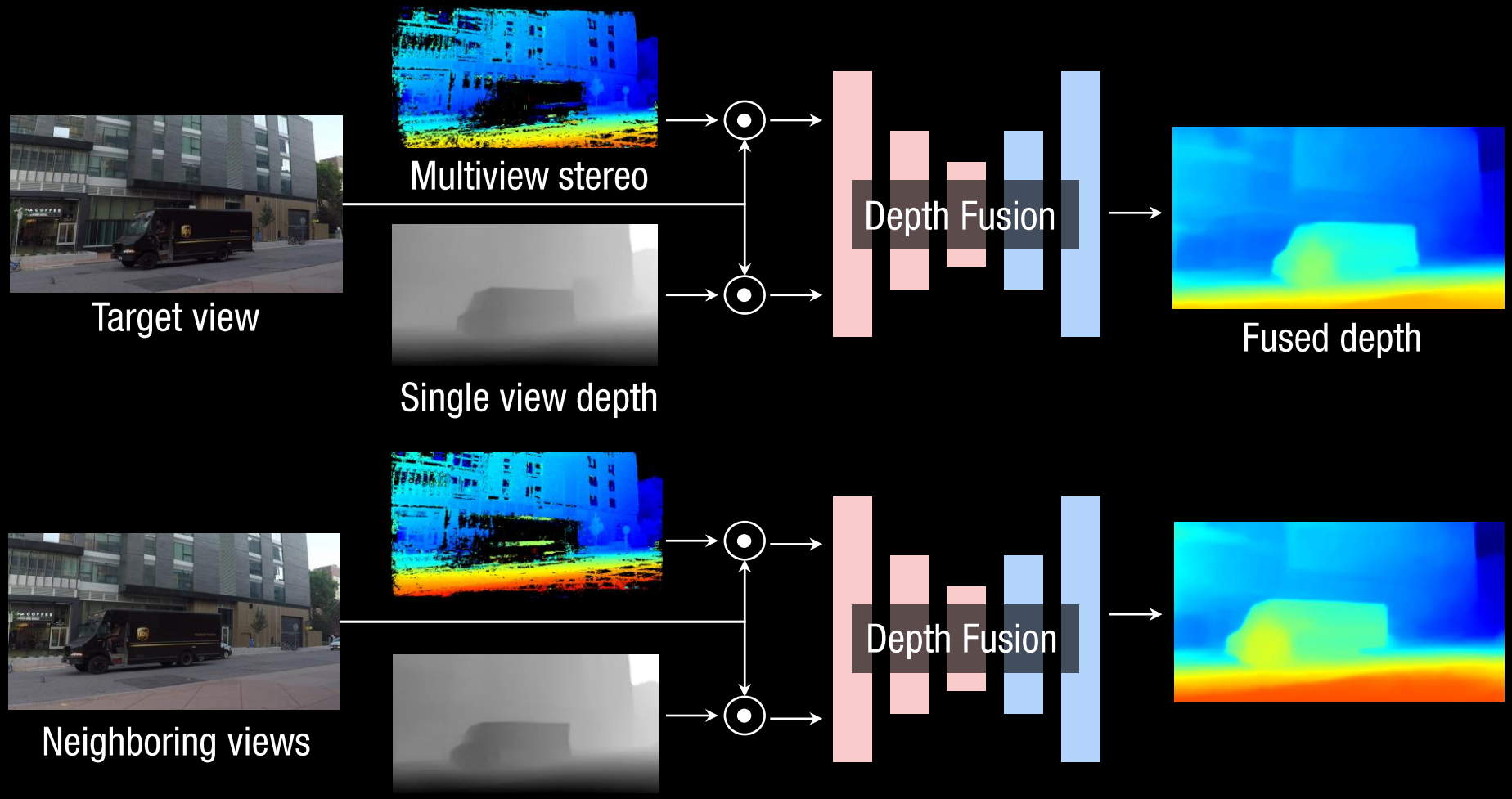



The output depth is spatially smooth.

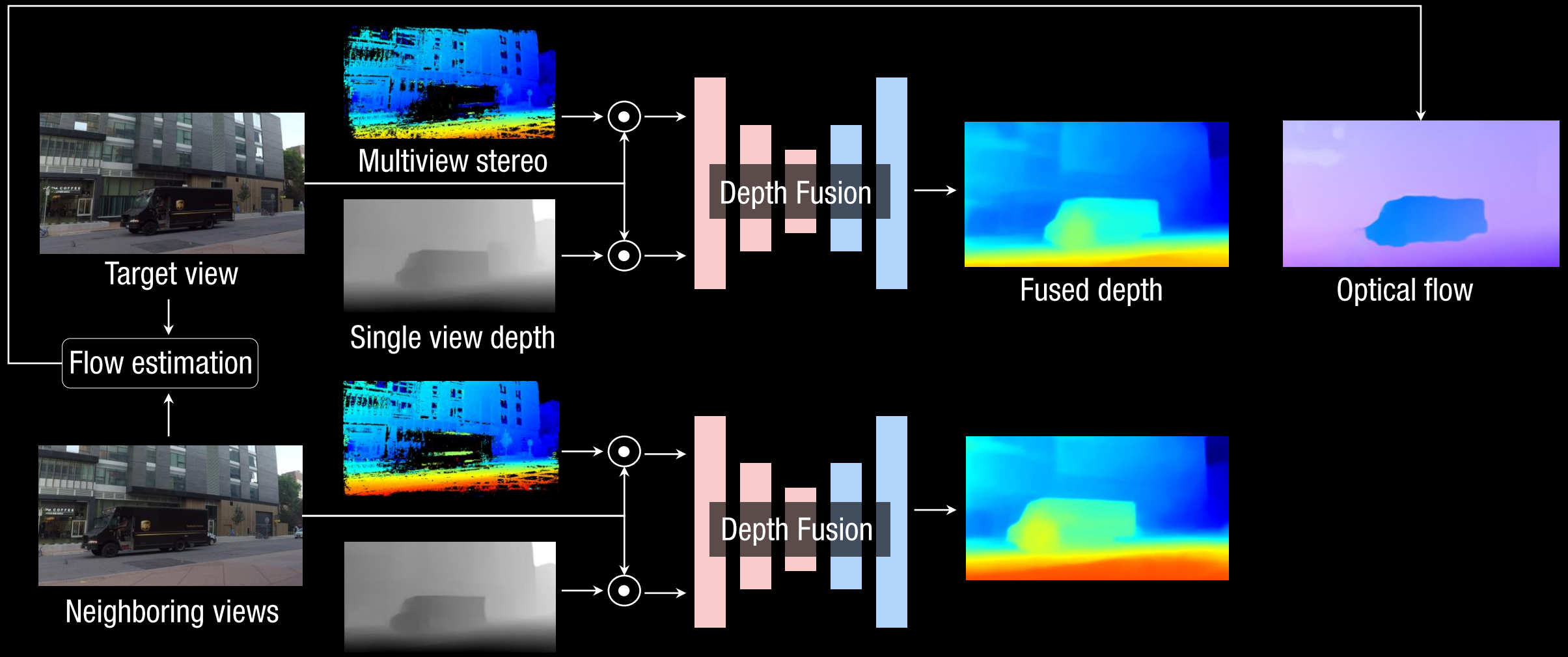
# Depth Fusion Network



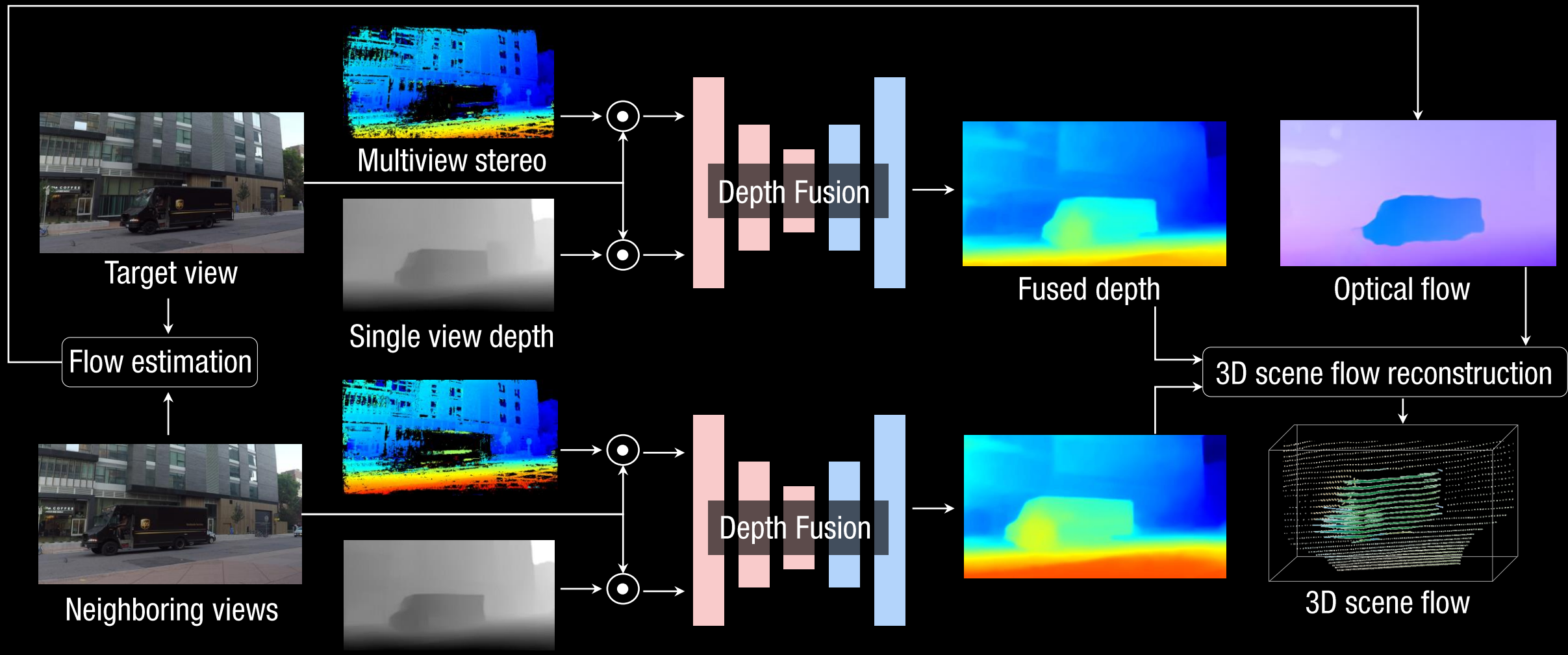
# Depth Fusion Network



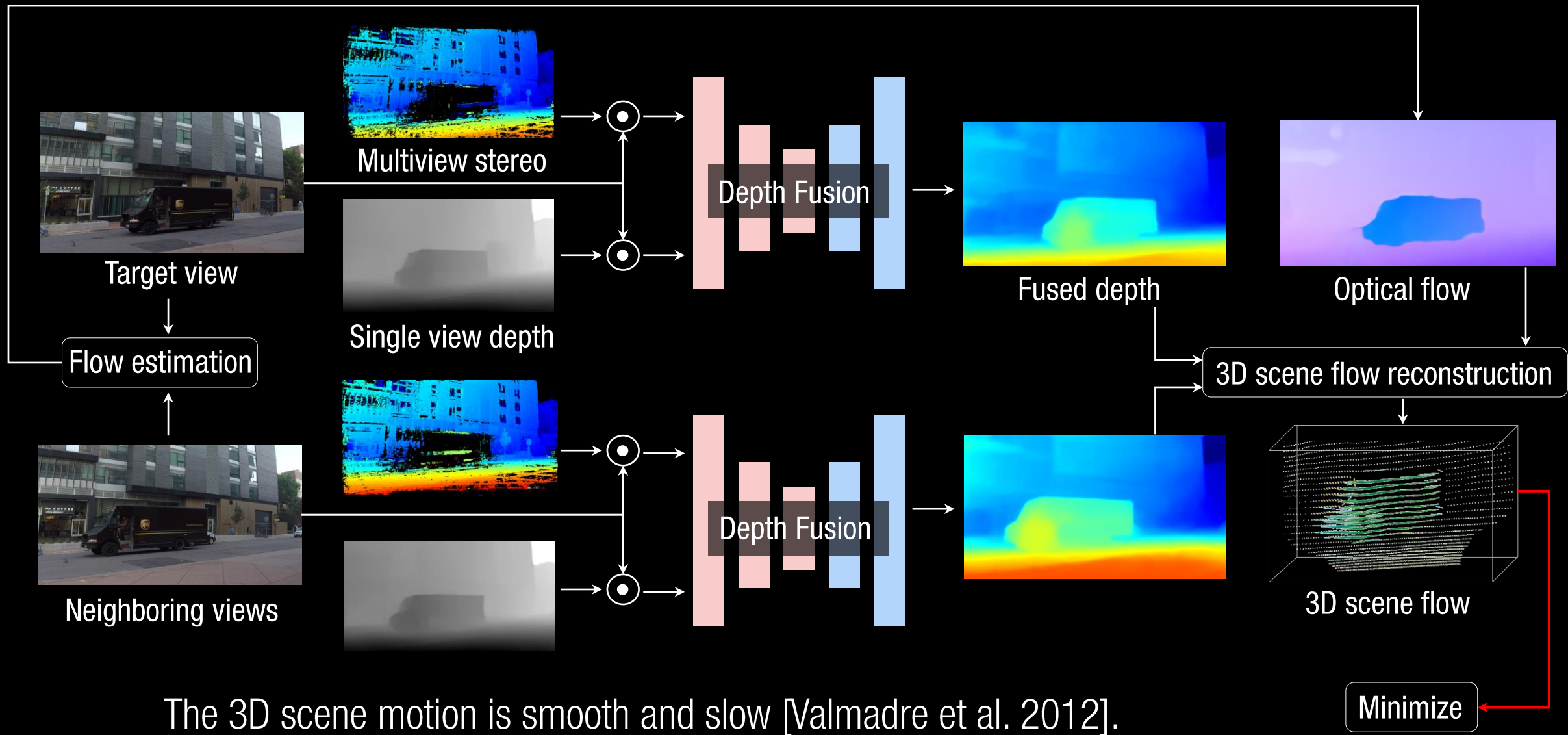
# Depth Fusion Network



# 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 ◁

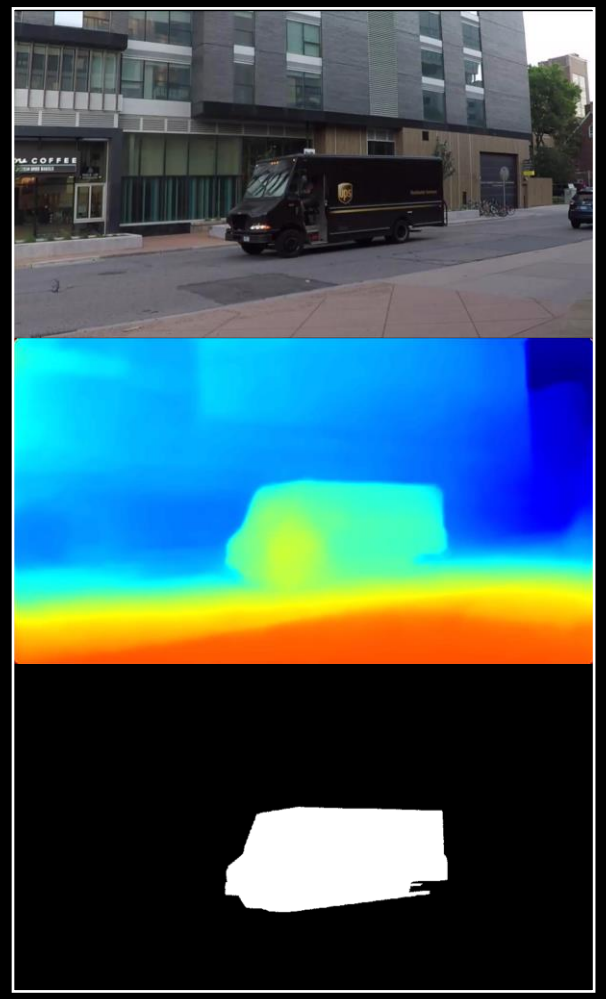


Image & mask & depth

# Novel View Synthesis of Dynamic Scenes

Source cameras

Virtual camera

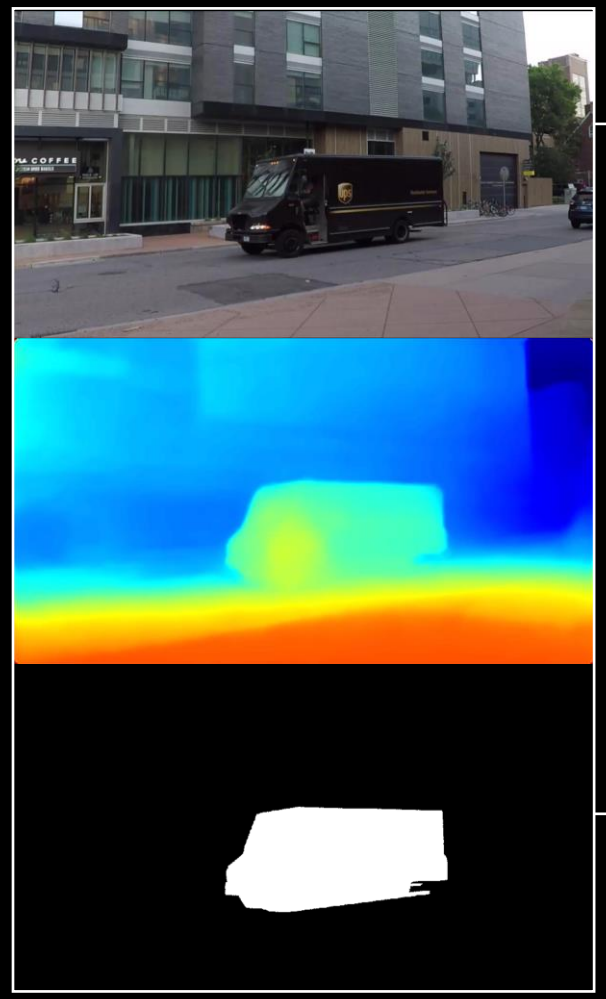


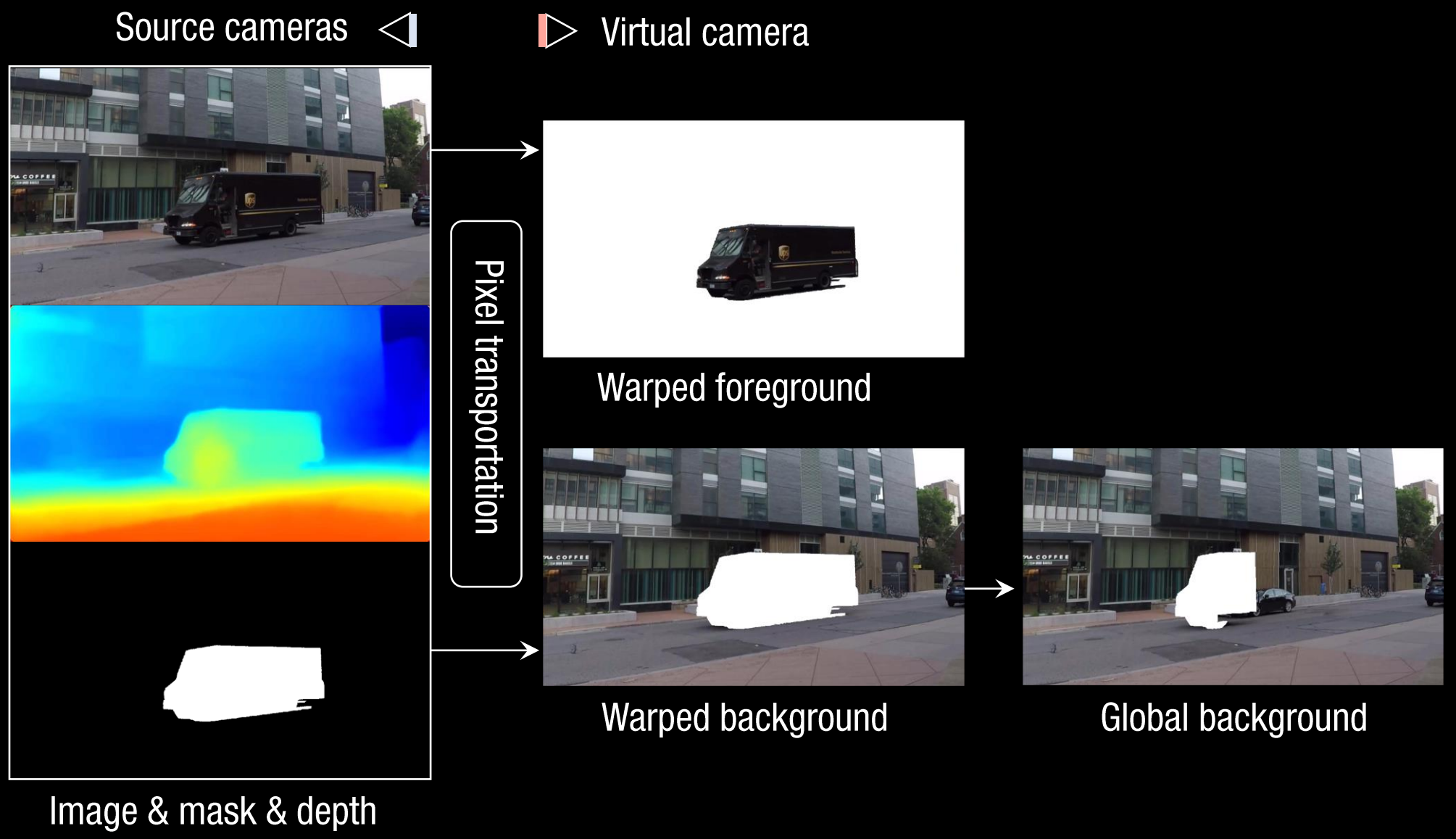
Image & mask & depth

Pixel transportation

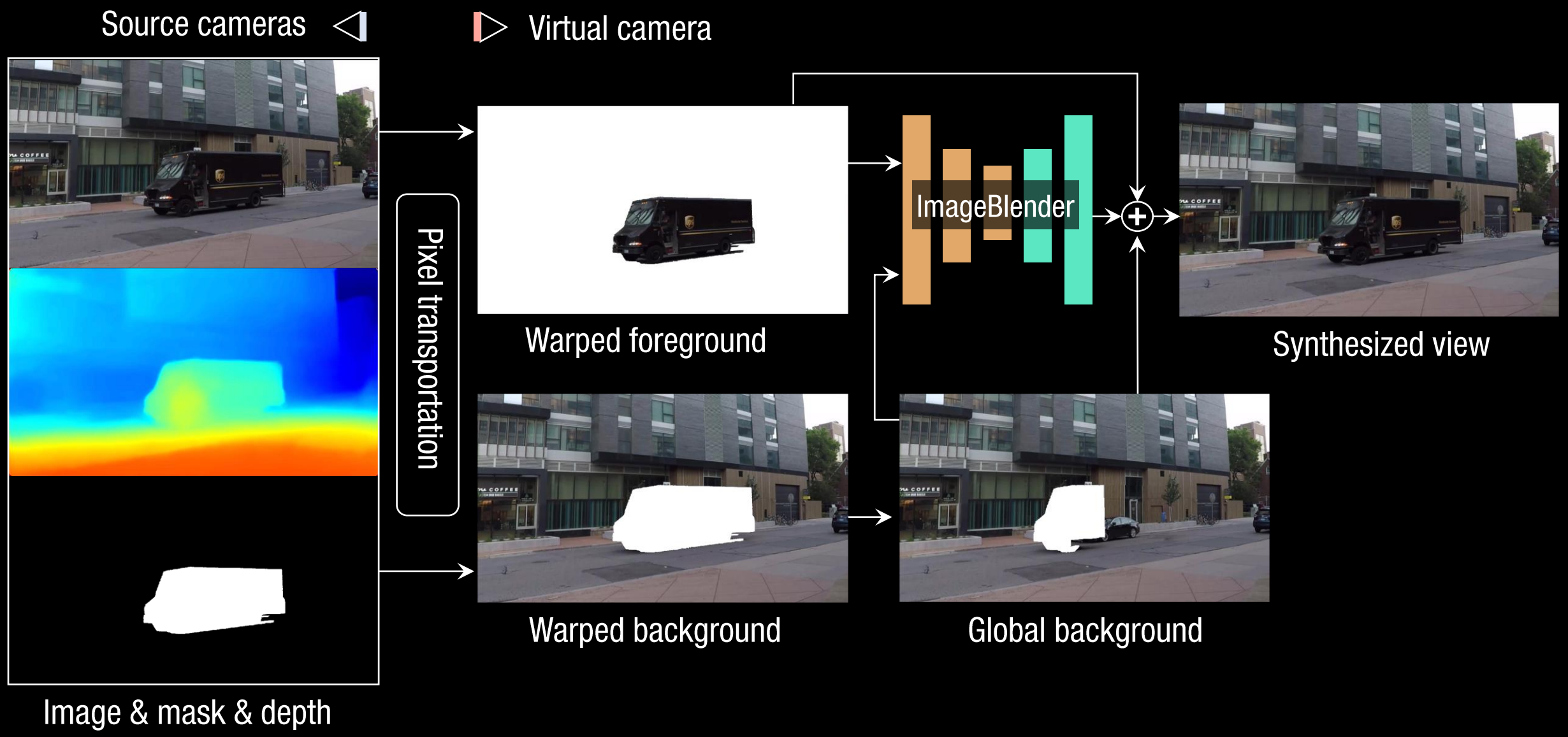
Warped foreground

Warped background

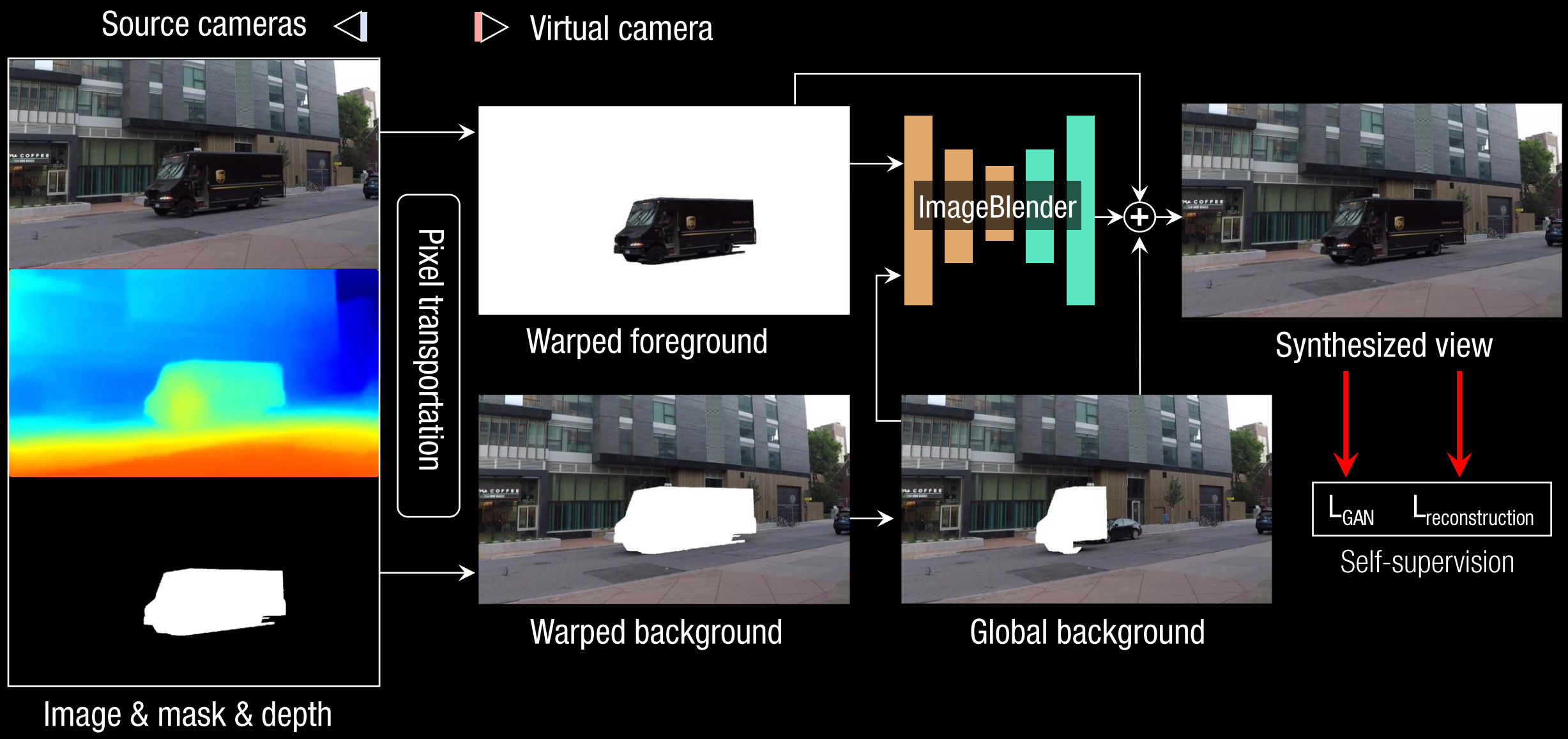
# Novel View Synthesis of Dynamic Scenes



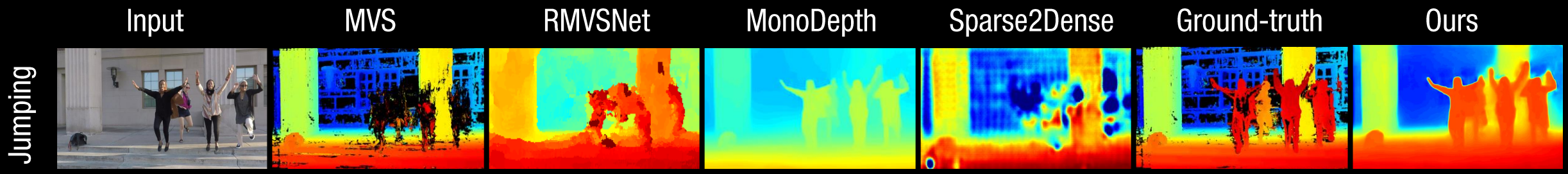
# Novel View Synthesis of Dynamic Scenes



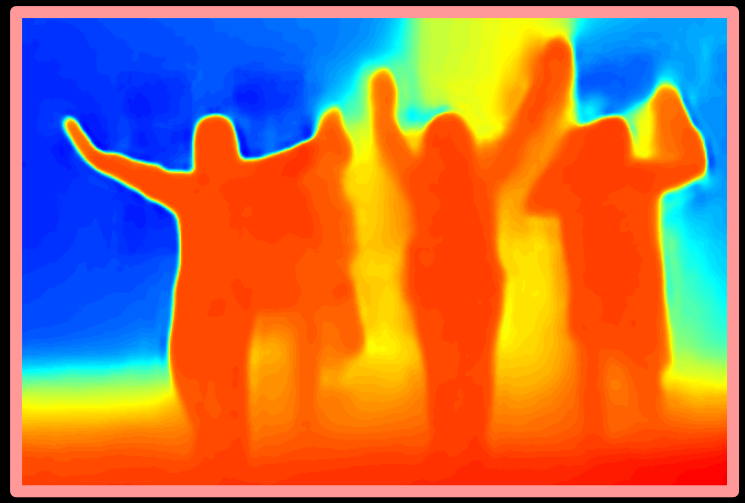
# Novel View Synthesis of Dynamic Scenes



# Experiments

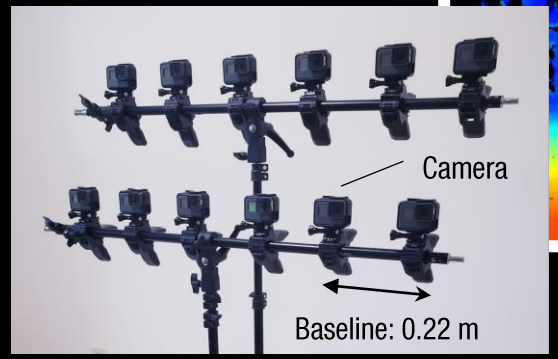


# Experiments

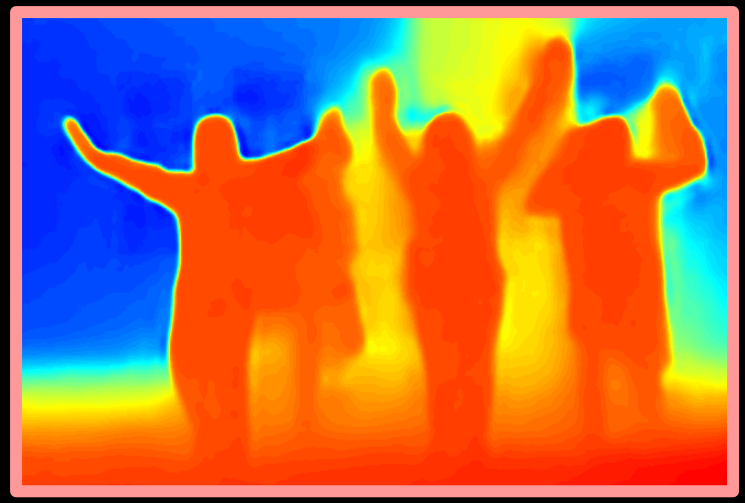



Ours (depth fusion)

# Experiments


Ground-truth

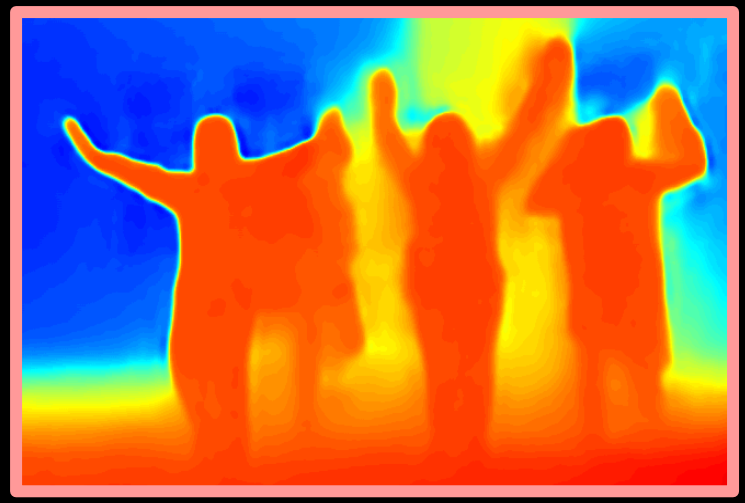


Ours (depth fusion)

# Experiments


MVS

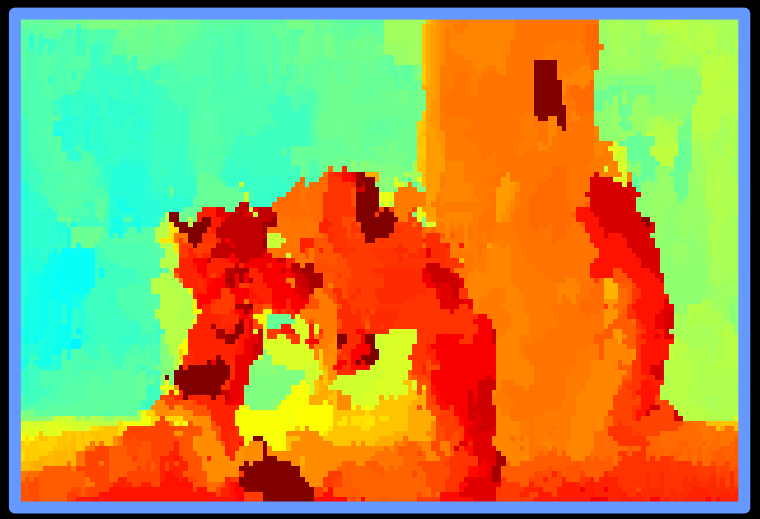
Ground-truth



Ours (depth fusion)

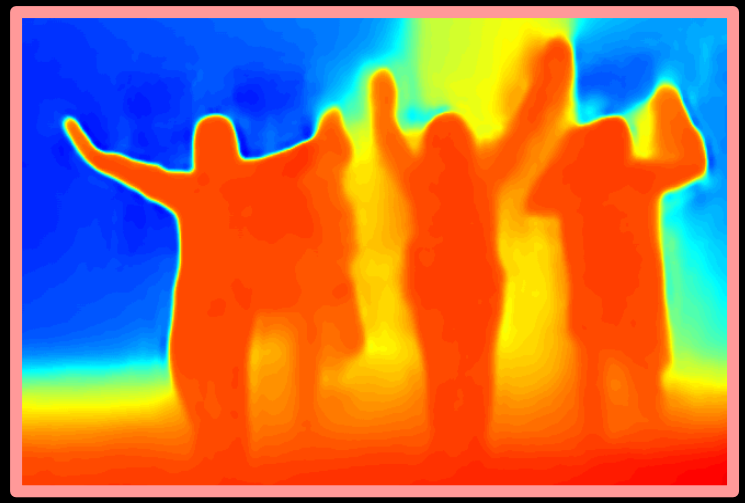
- MVS : Optimizaiton based multiview stereo [ECCV 2016 Schonberger et al.]

# Experiments

RMVSNet

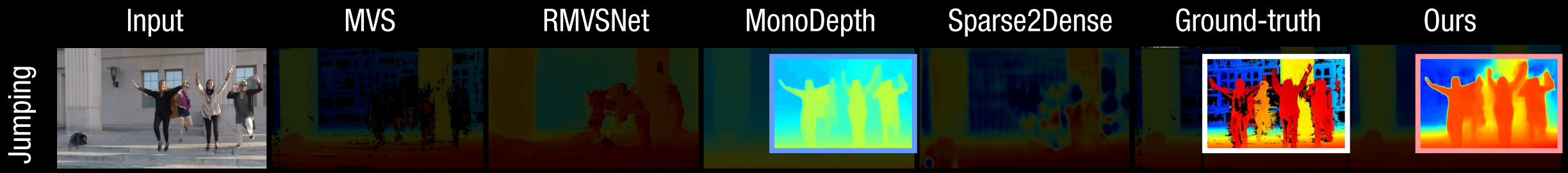
Ground-truth



Ours (depth fusion)

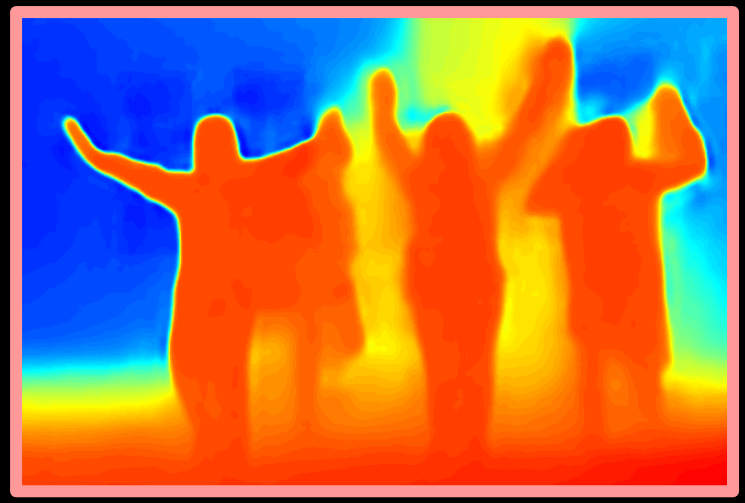
- RMVSNet : Learning based multiview stereo [CVPR 2019 Yao et al.]

# Experiments




MonoDepth

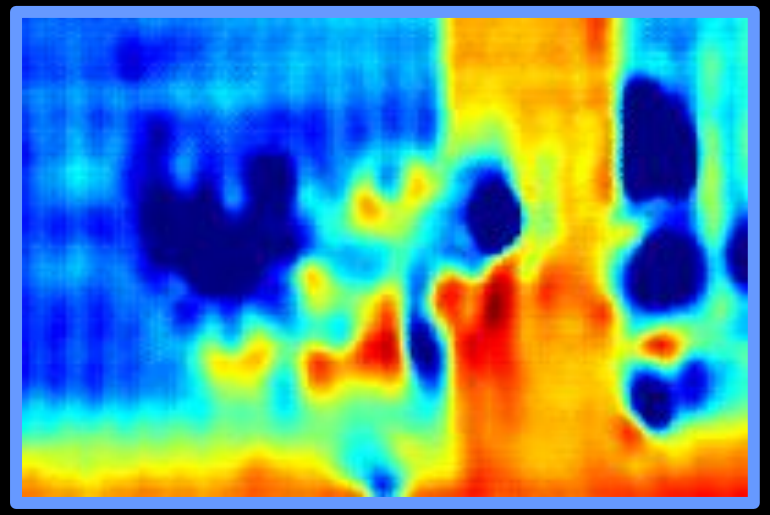
Ground-truth



Ours (depth fusion)

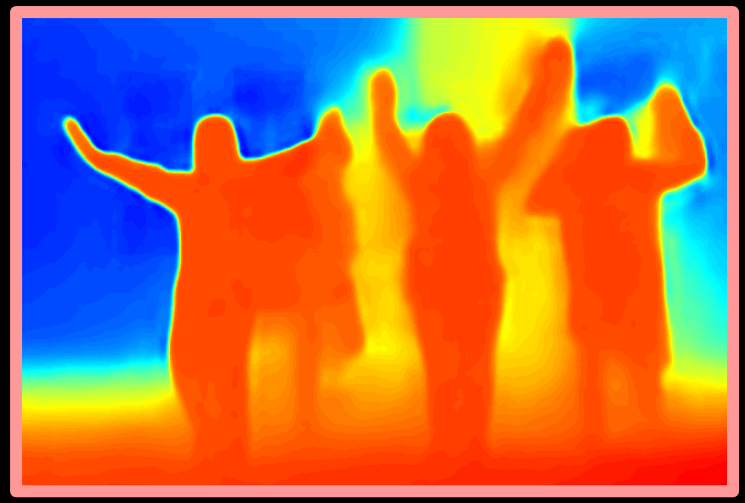
- MonoDepth : Depth prediction from a single image [Arxiv 2019 Ranftl et al.]

# Experiments

Sparse2Dense

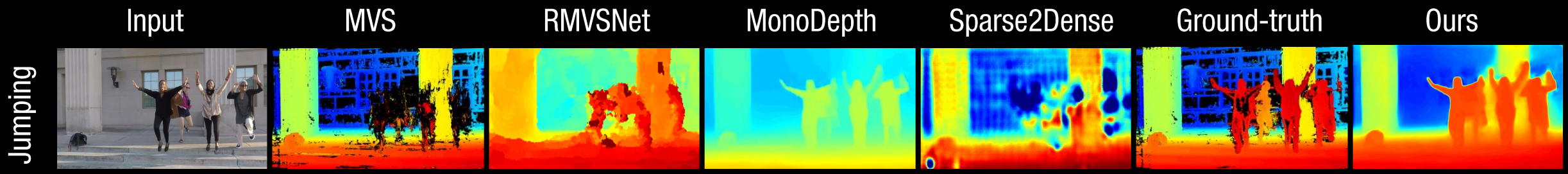
Ground-truth



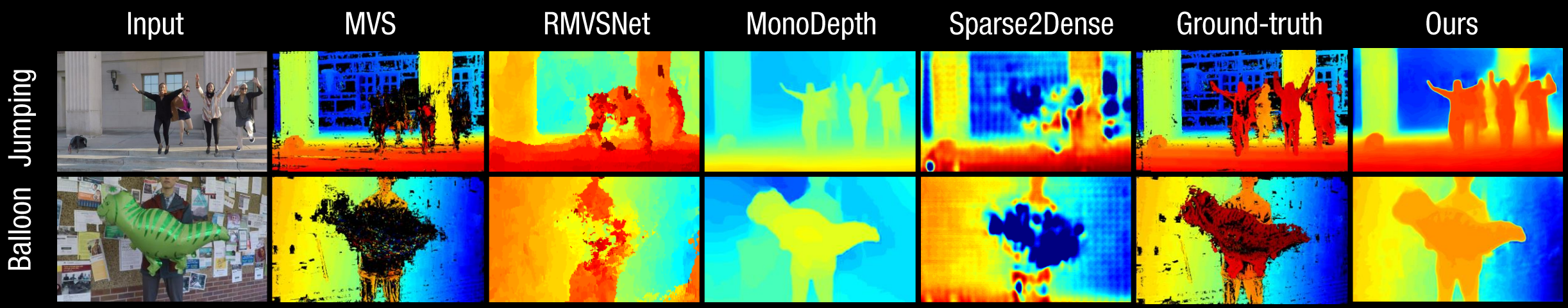
Ours (depth fusion)

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

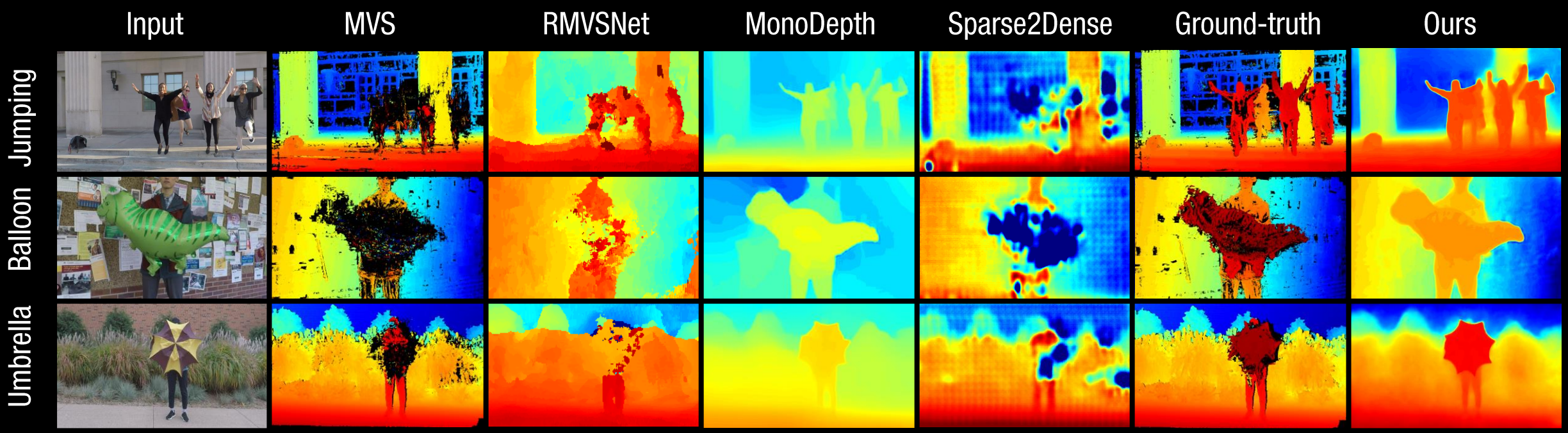
# Experiments



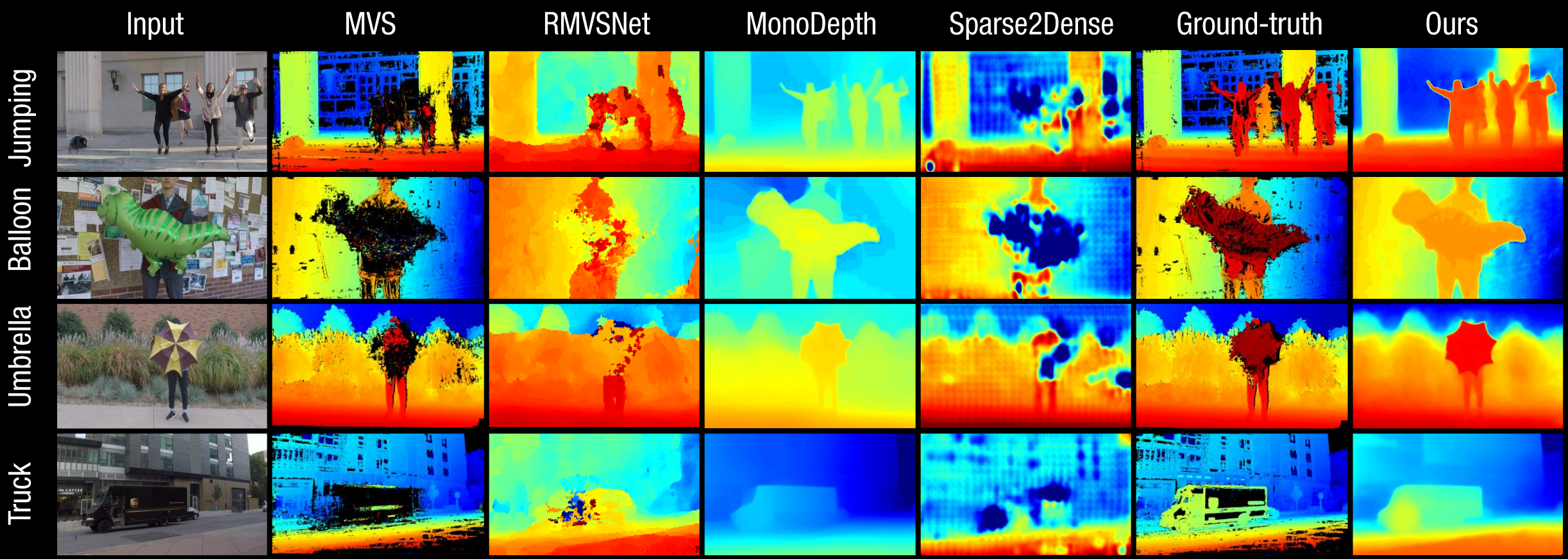
# Experiments



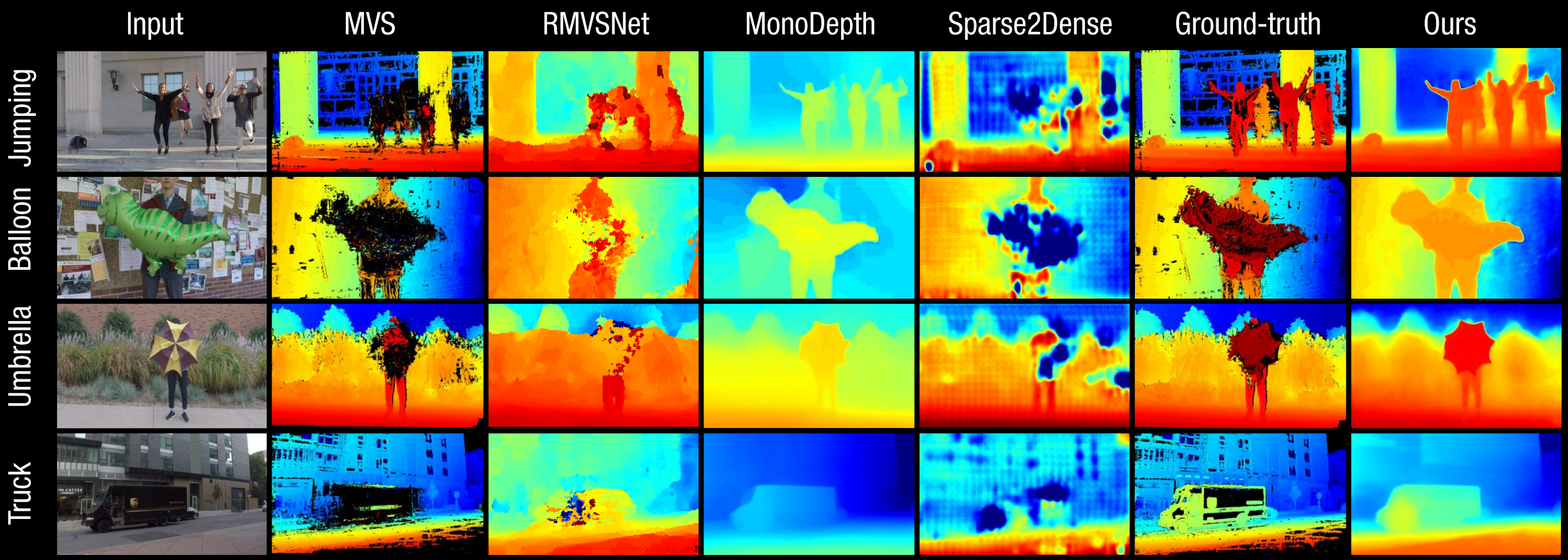
# Experiments

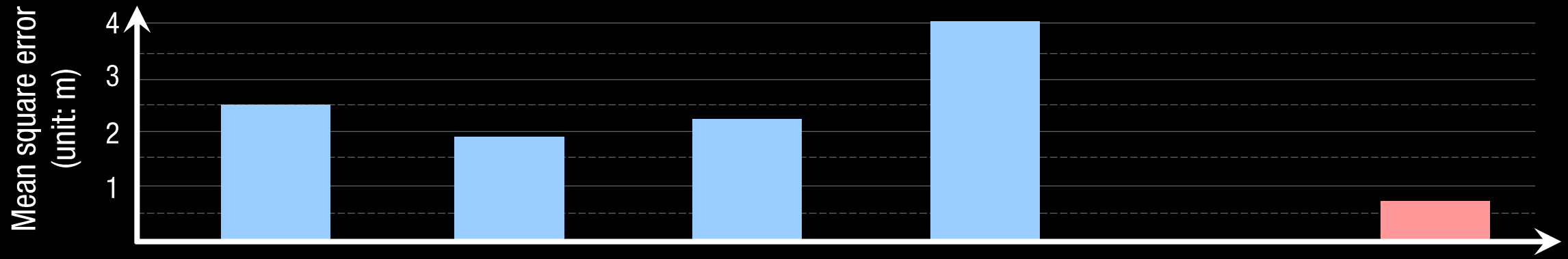


# Experiments



# Experiments



# Experiments

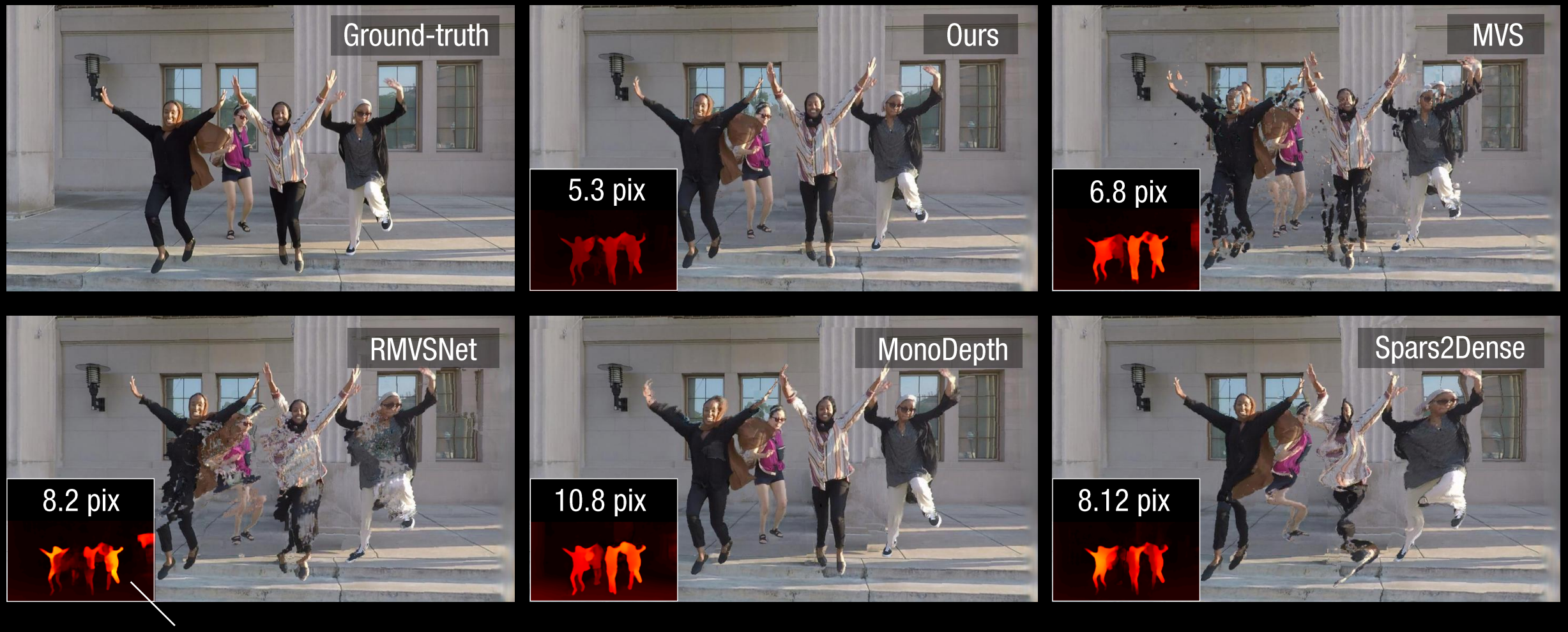





# Experiments


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

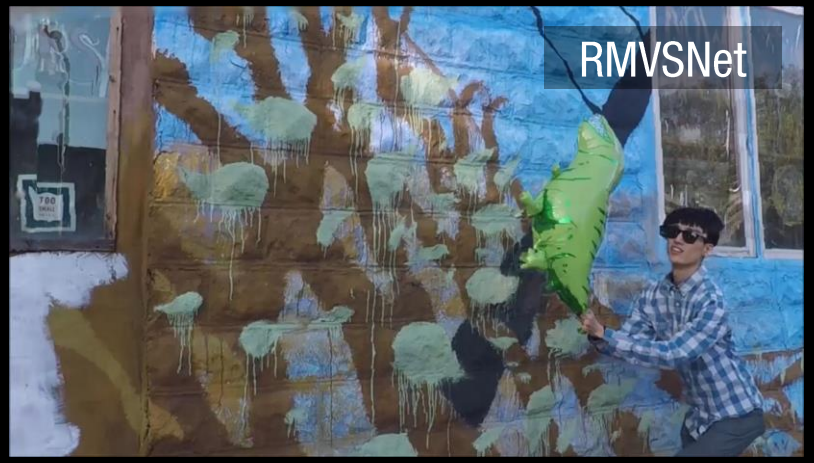
# Experiments

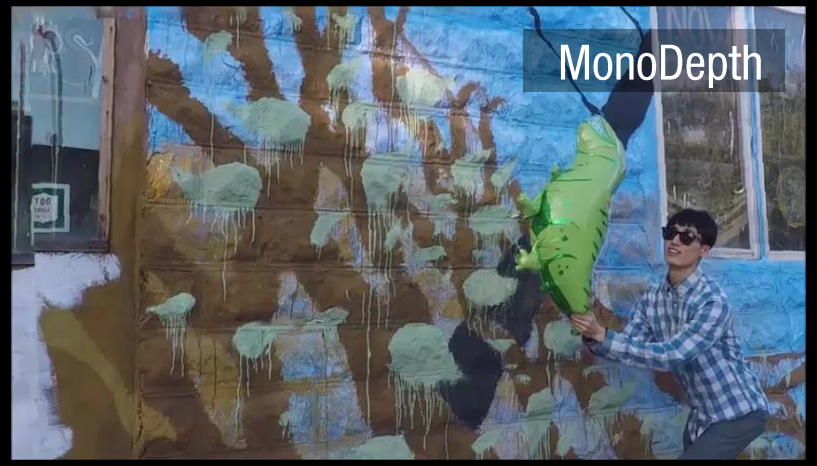





# Experiments




# Limitation

Input

Output

Erroneous mask
(fragmentation, afterimage)

# Limitation

Input

Output

Erroneous mask
(fragmentation, afterimage)

Cluttered scene

# Limitation

Input

Output

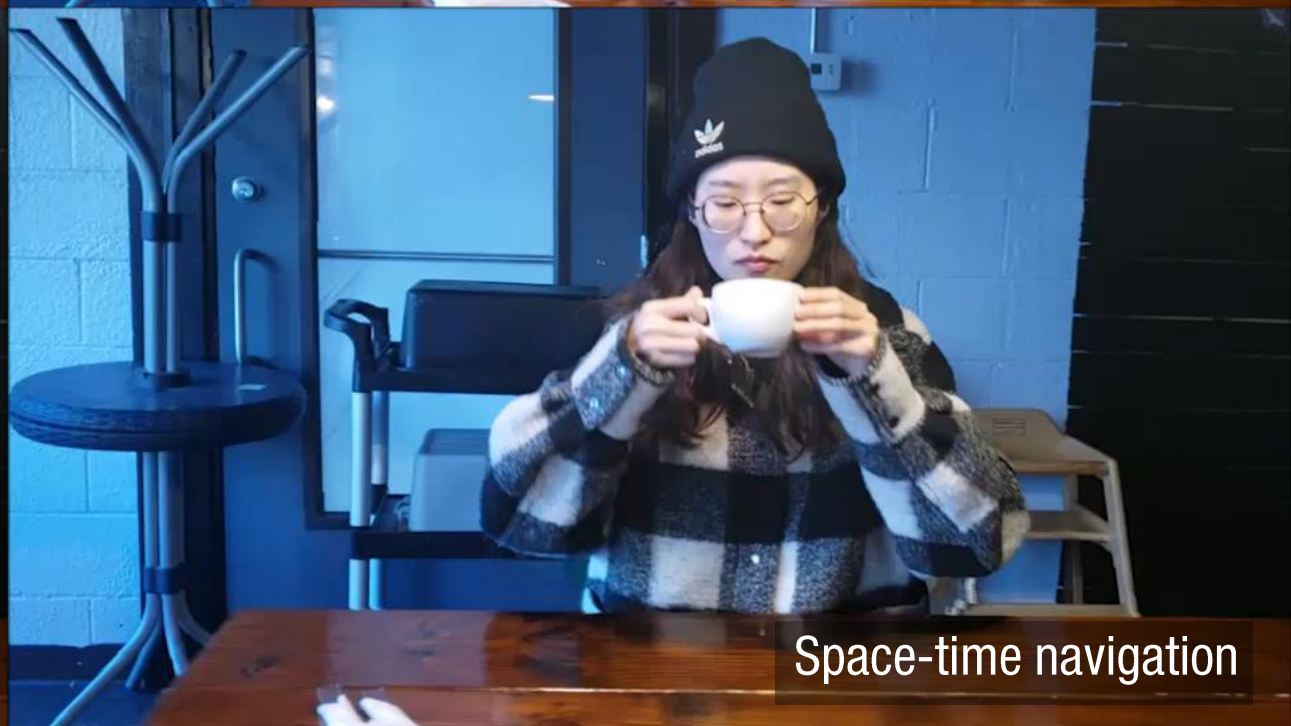
Erroneous mask
(fragmentation, afterimage)

Cluttered scene

Large viewing angle

# More Results

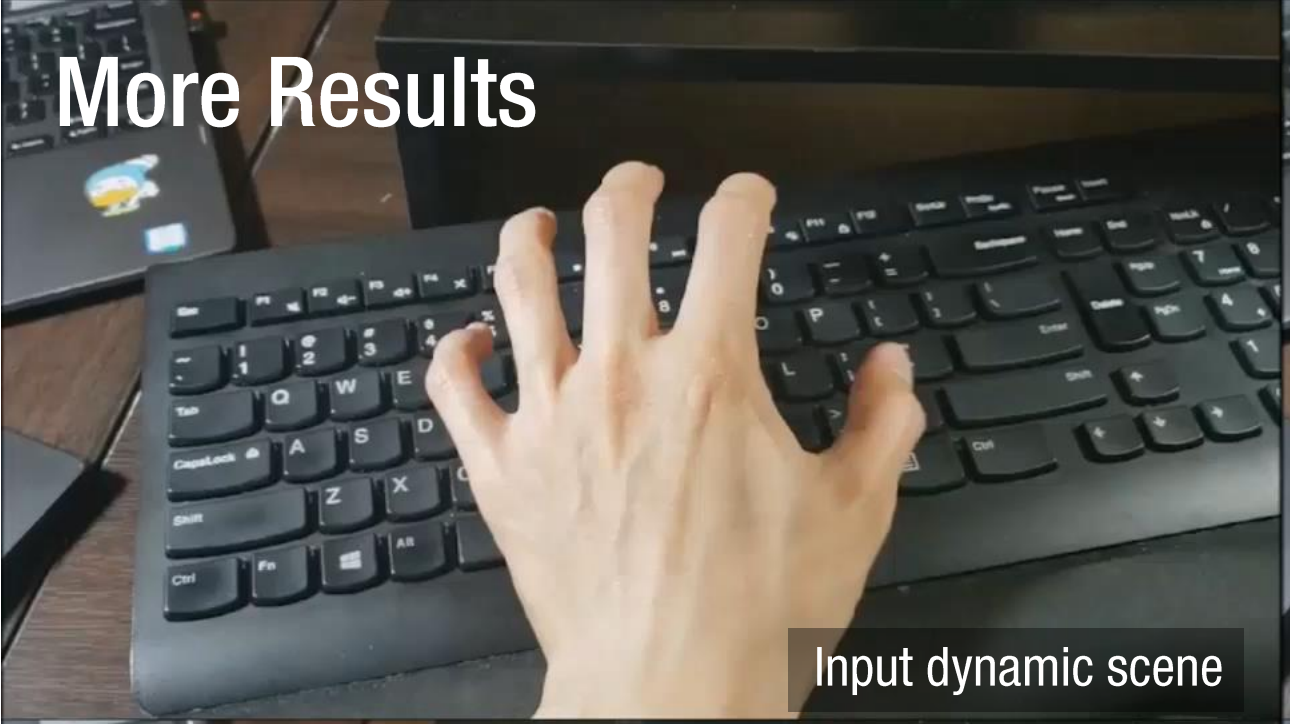


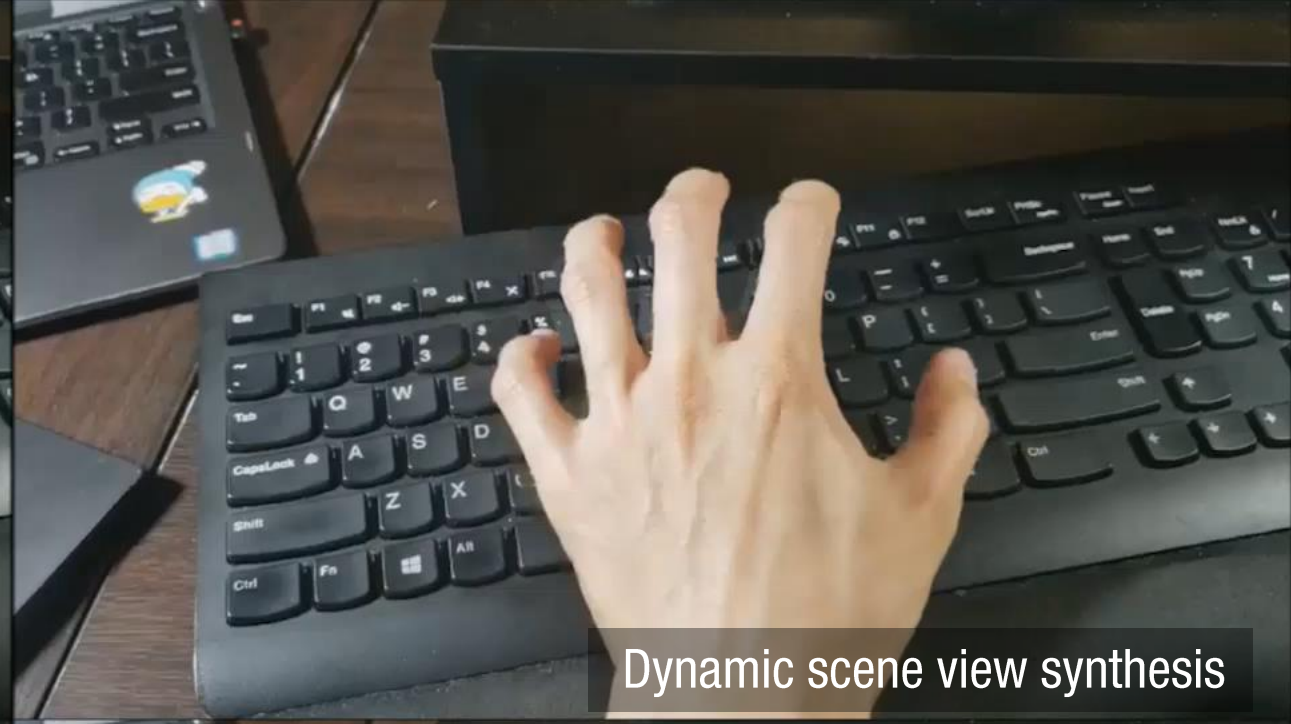



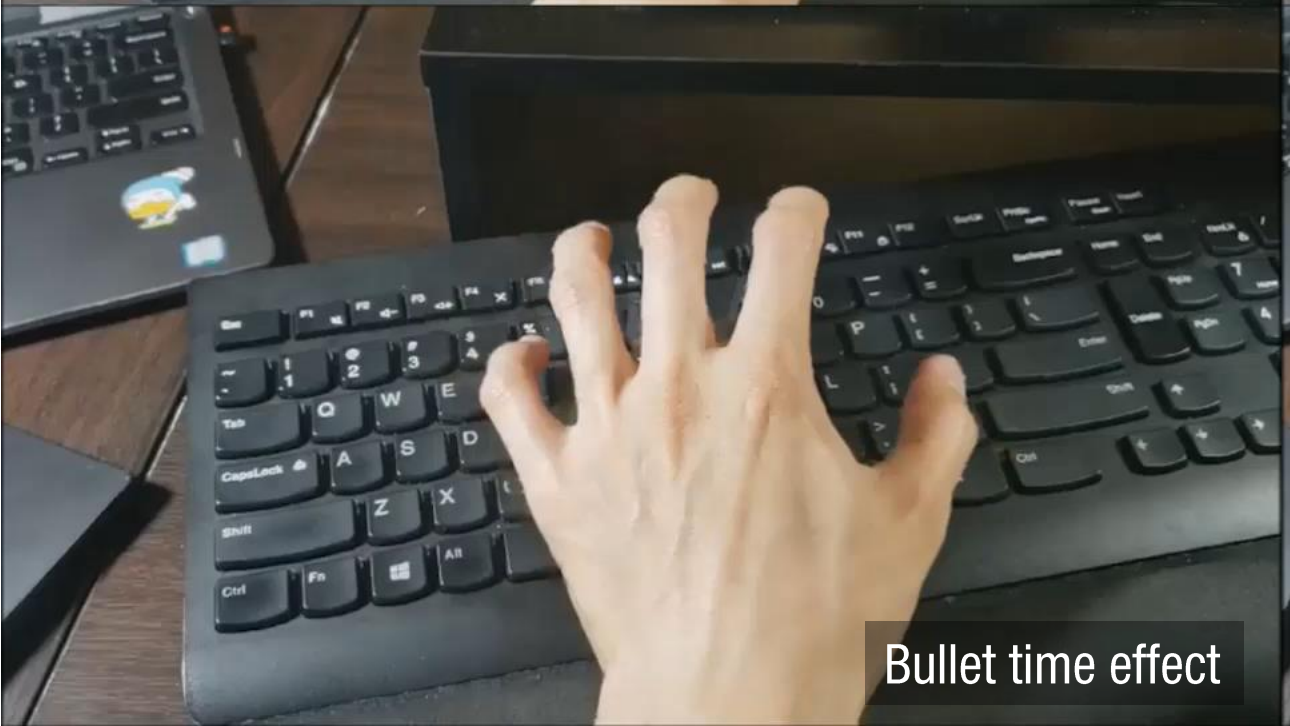
# More Results

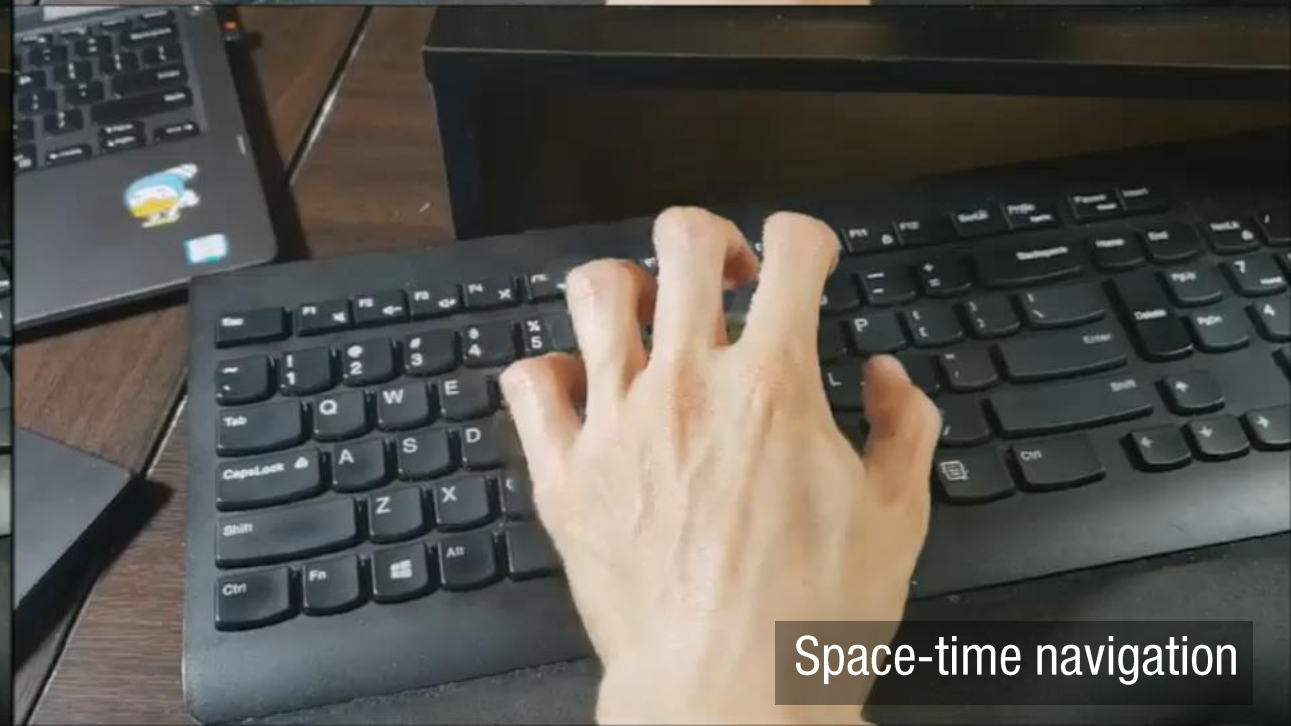



# More Results



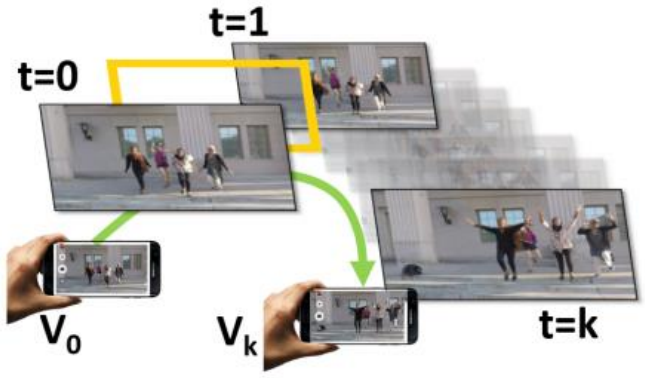



CVPR SEATTLE WASHINGTON JUNE 16-18 2020

# Novel View Synthesis of Dynamic Scenes with Globally Coherent Depths from a Monocular Camera

Jae Shin Yoon[1], Kihwan Kim[2], Orazio Gallo[2], Hyun Soo Park[1], and Jan Kautz[2]




1 UNIVERSITY OF MINNESOTA

2 NVIDIA