Competitive Collaboration

Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation

Anurag Ranjan

Perceiving Systems

Max Planck Institute for Intelligent Systems





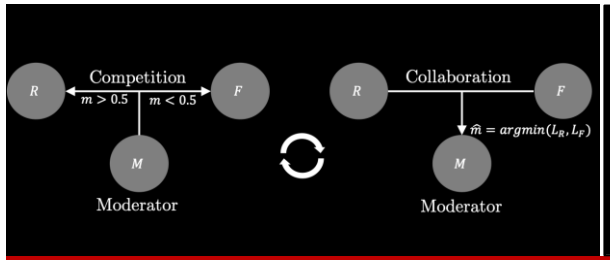

Varun Jampani Lukas Balles Deqing Sun Kihwan Kim Jonas Wulff Michael Black


Tübingen, Germany

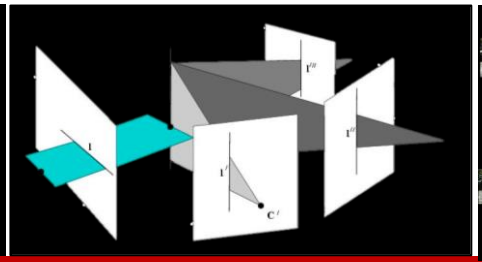
Outline

Motion and Optical Flow

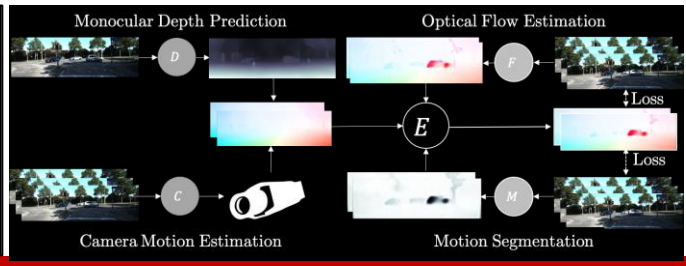
Deep Learning with Structure



Competitive Collaboration



Geometry



Unsupervised Learning of Everything

Supervised

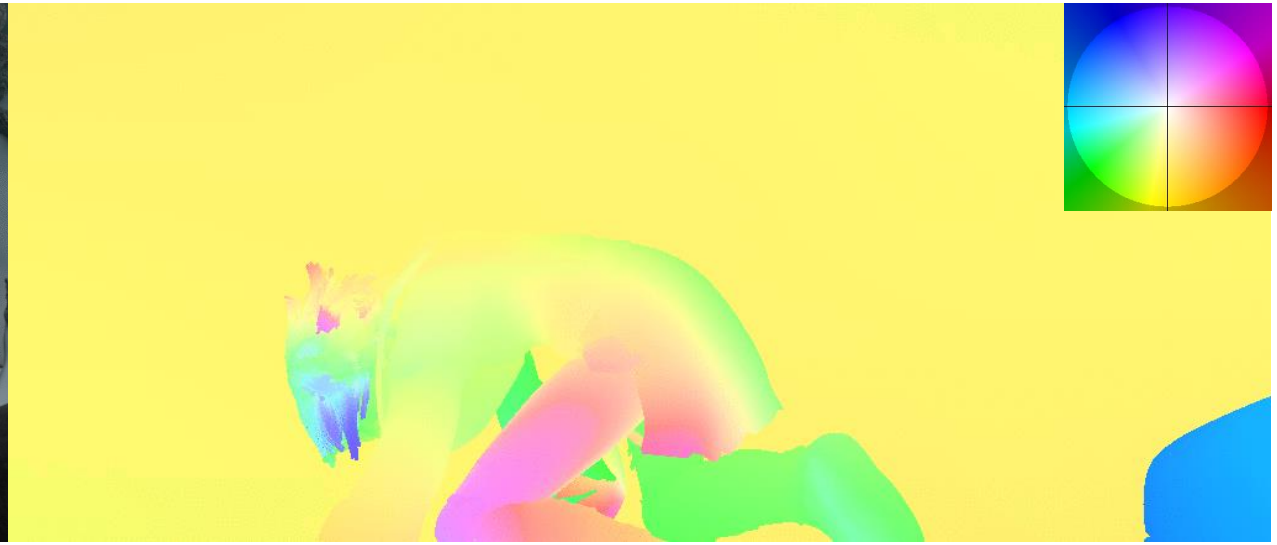
Unsupervised

Motion and Optical Flow

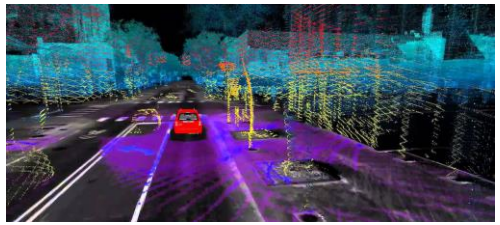
Optical Flow

2D velocity for all pixels between two frames of a video sequence.

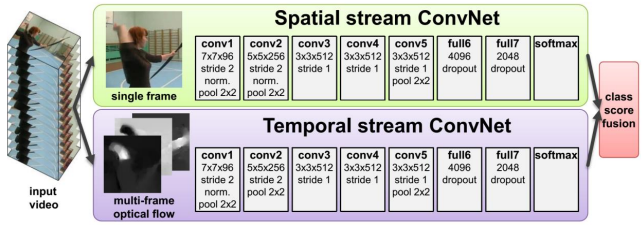
$$I(x, y, t-1) = I(x+u, y+v, t)$$



Why do we need Optical Flow



SLAM



Action Recognition

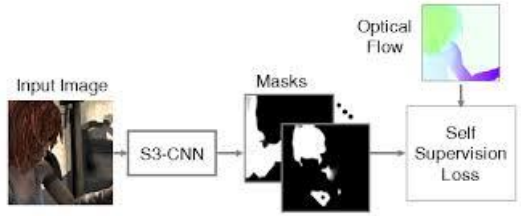
Super-resolution

Video Compression

Optical Flow

Slomo

VFX



Unsupervised Segmentation

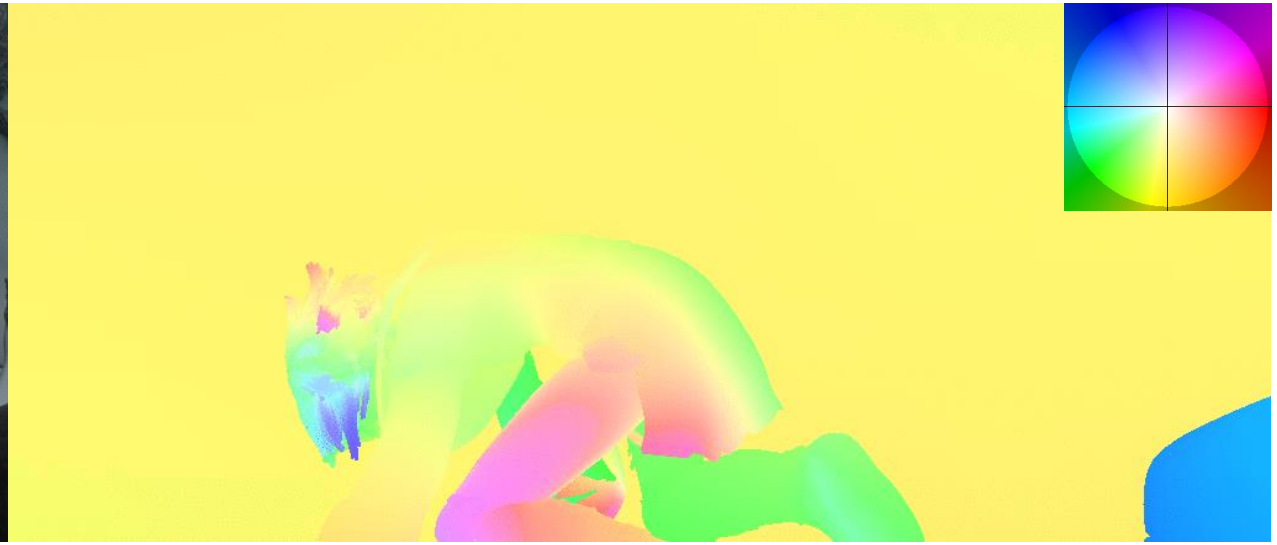
Motion Magnification

Unsupervised Segmentation: Mahendran et al., VFX: Black et al., Motion Magnification: Liu et al., Action Recognition: Simoyan et al.

Optical Flow

2D velocity for all pixels between two frames of a video sequence.

$$I(x, y, t-1) = I(x+u, y+v, t)$$



Estimating Optical Flow

$$I(x, y, t-1) = I(x+u, y+v, t)$$

$$\min_{u,v} \| I(x, y, t-1) - I(x+u, y+v, t) \|$$

$$\min_{u,v} \rho(I(t-1) - warp(I(t), u, v))$$

Photometric Loss

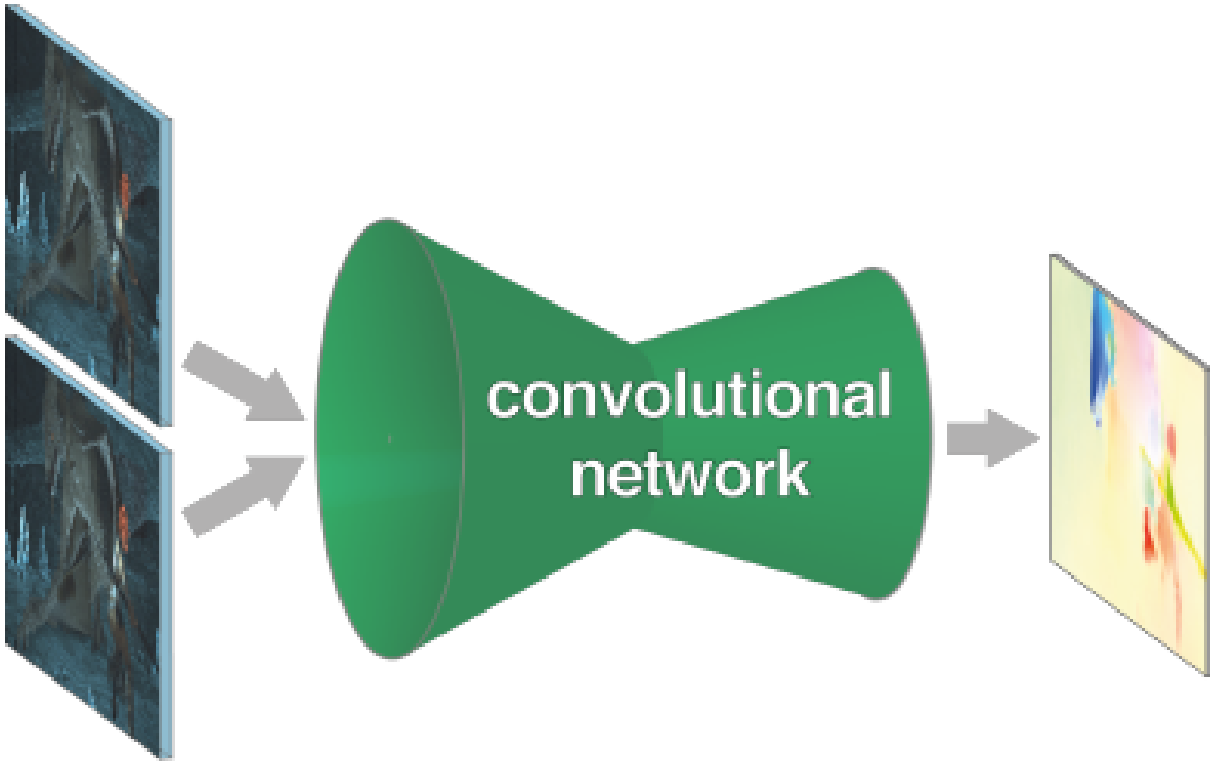
$$\min_{u,v} \rho(I(t-1) - warp(I(t), u, v))$$

Photometric Loss

No prior on structure

Can we learn from data?

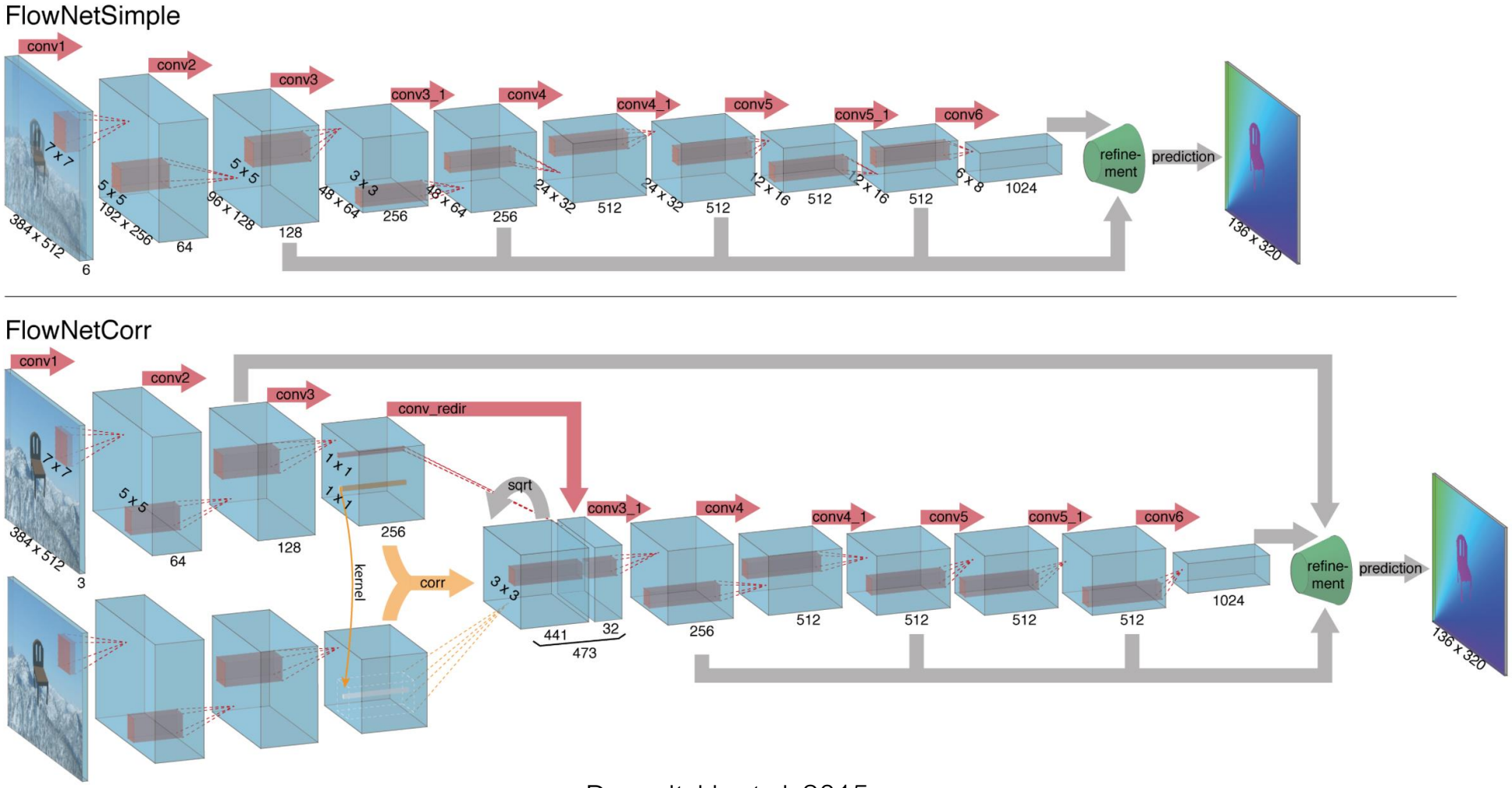
Optical Flow Estimation



$\in \mathbb{R}^{n \times n}$

Dosovitskiy et al. 2015

FlowNet



Dosovitskiy et al. 2015

Problem

FlowNet is too big. 33 M parameters.

Needs to learn both large and small motions.

Does not perform well.

Approach

Image statistics are scale invariant.

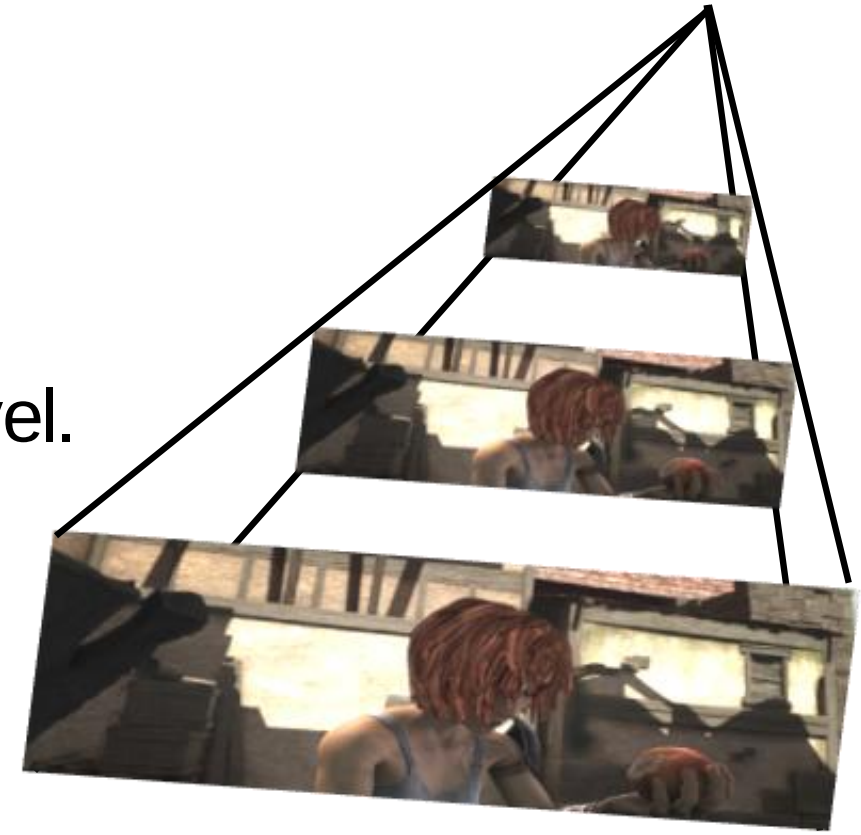
Use an image pyramid.

Train a small network for each pyramid level.

Compute residual flow at each level.

Network captures small displacements.

Pyramid captures large displacements.

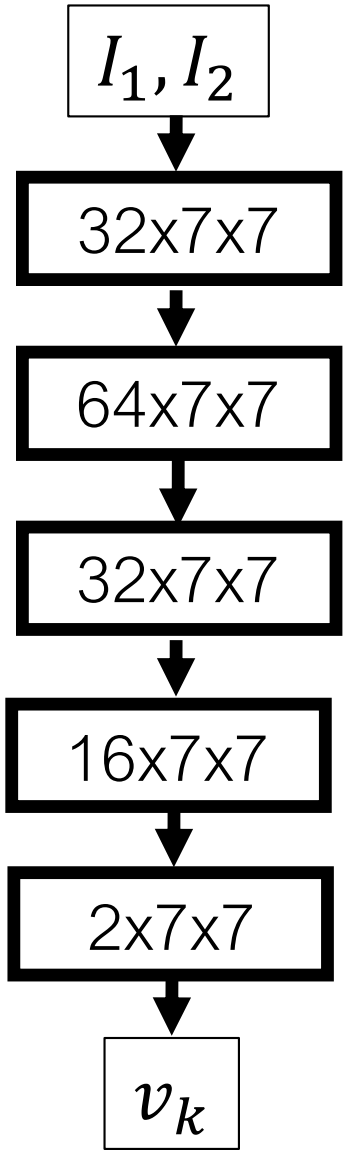


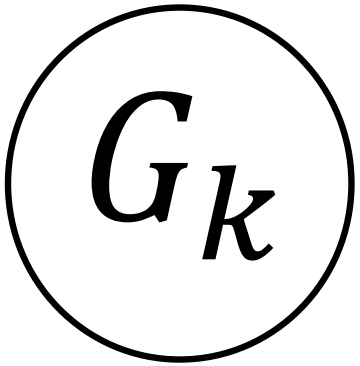
Burt and Adelson. The Laplacian pyramid as a compact image code. IEEE COM, 1983

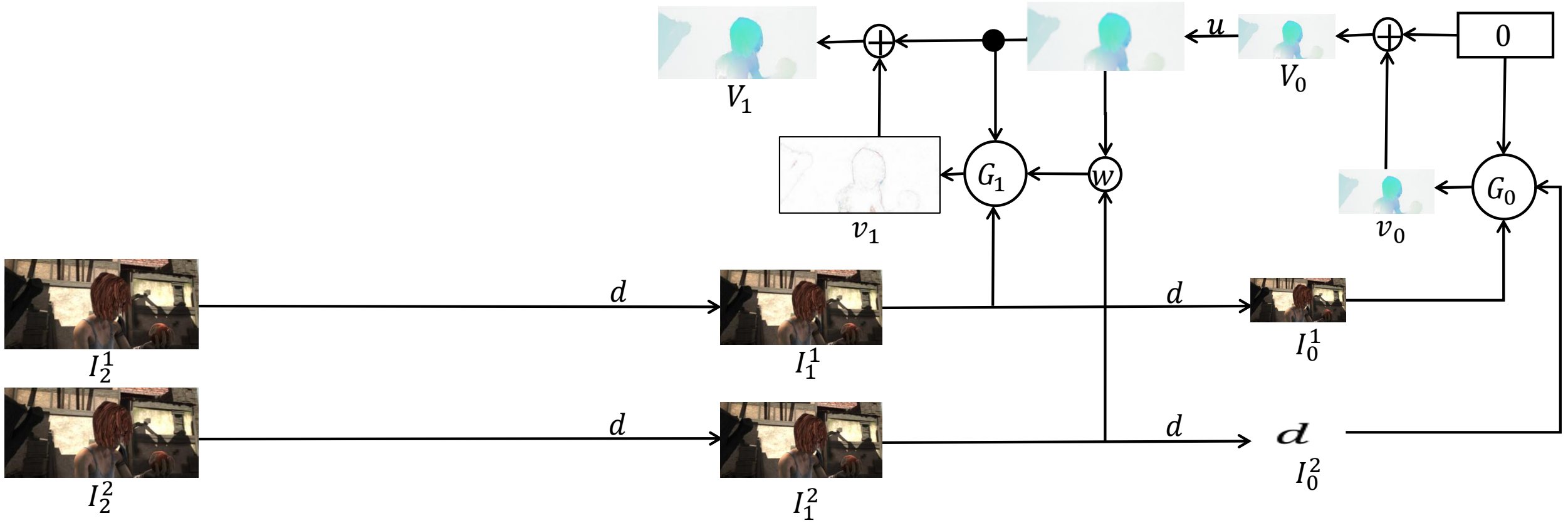
SPyNet

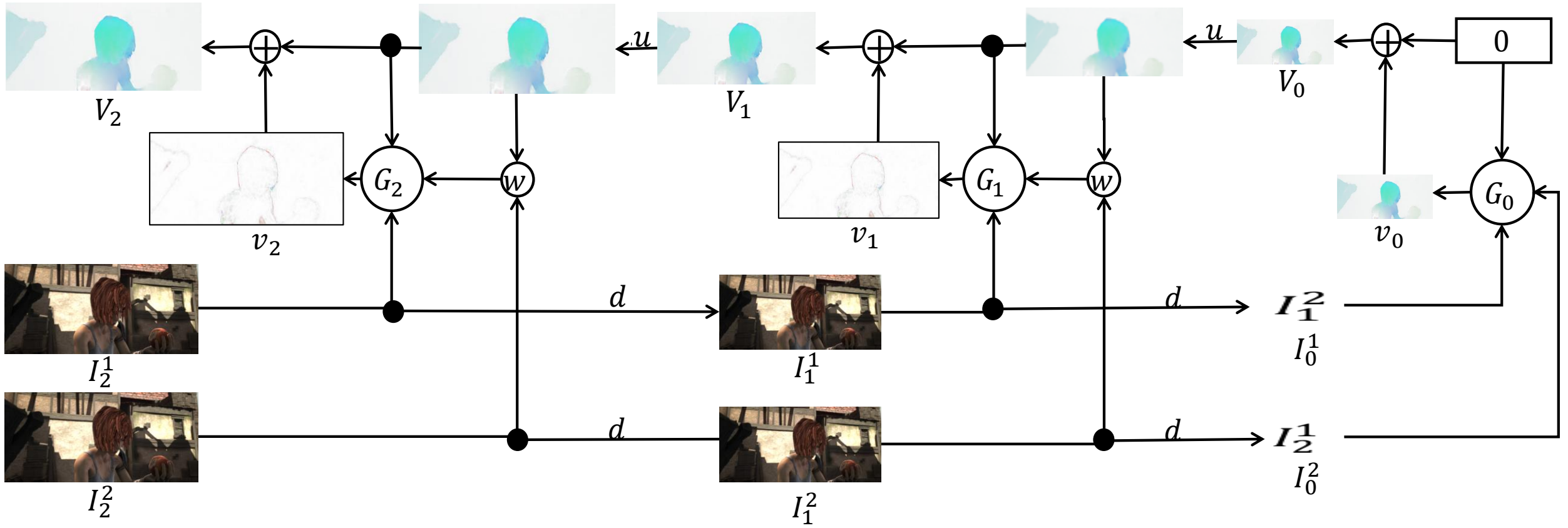
Spatial Pyramid Network for Optical Flow Estimation

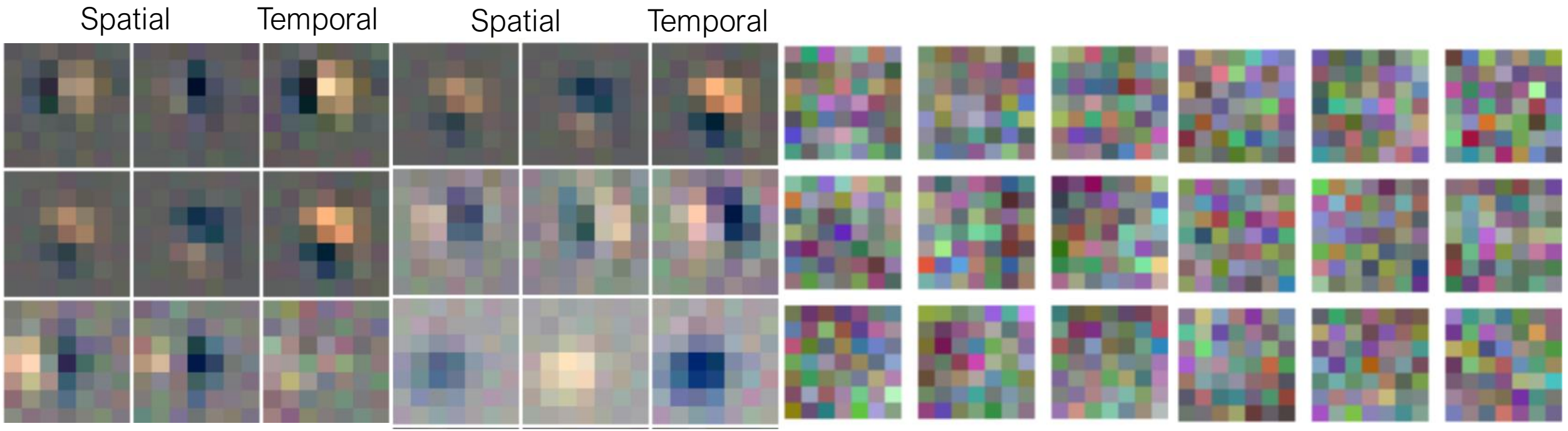
Ranjan et al. Optical Flow estimation using a Spatial Pyramid Network. CVPR 2017.


I_1, I_2
32x7x7
64x7x7
32x7x7
16x7x7
2x7x7
v_k


G_k


V_1
u
V_0
0
v_1
G_1
w
v_0
G_0
I_2^1
d
I_1^1
d
I_0^1
I_2^2
d
I_1^2
d
I_0^2

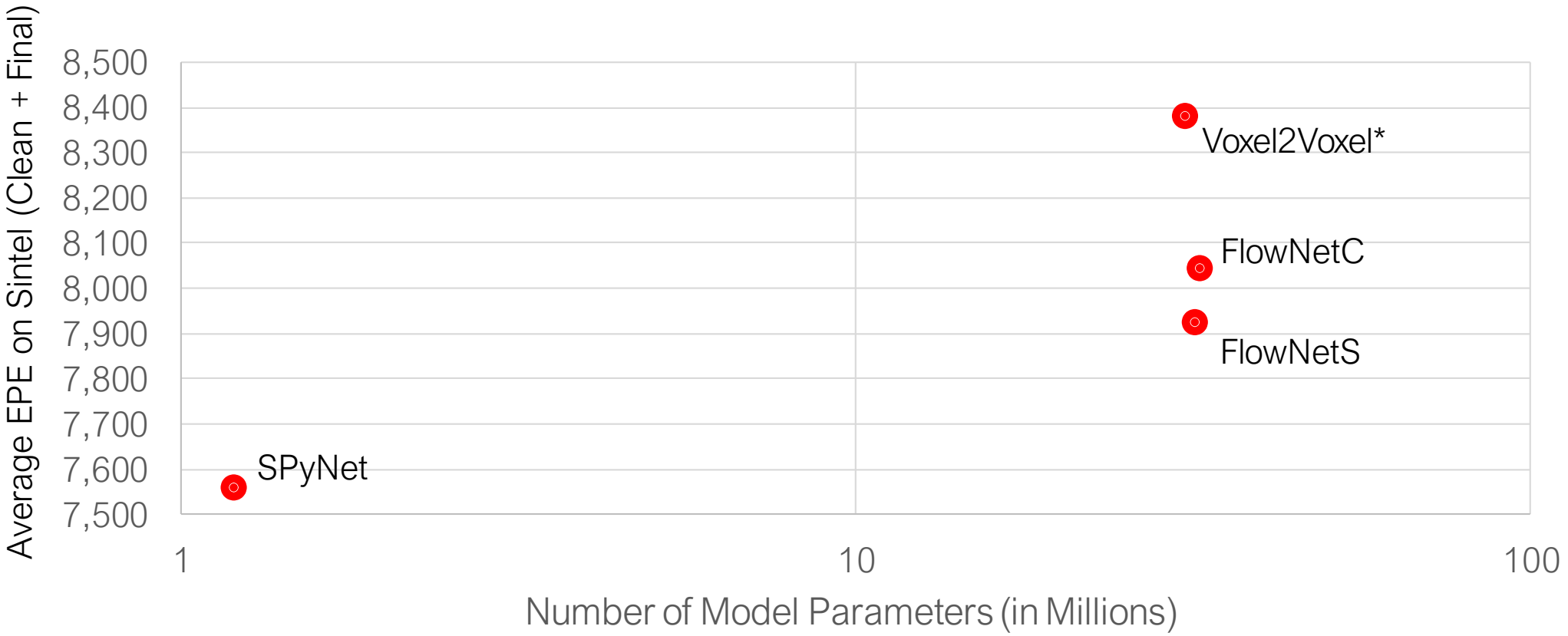

V_2
u
V_1
u
V_0
0
v_2
G_2
w
v_1
G_1
w
v_0
G_0
I_2^1
d
I_1^1
d
I_1^2
I_0^1
I_2^2
d
I_1^2
d
I_2^1
I_0^2



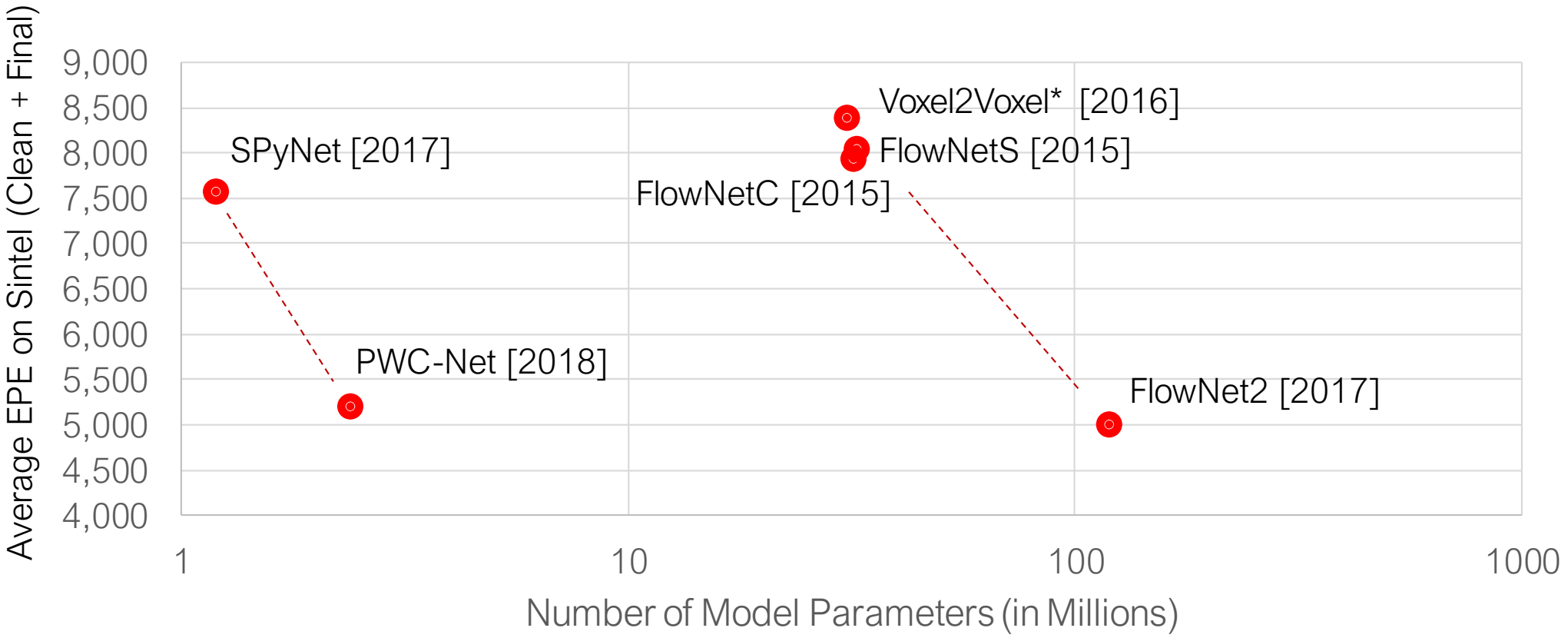
SPyNet

FlowNet


Frames
Ground Truth
FlowNetS
FlowNetC
SPyNet



*error metric not consistent with the benchmarks



*error metric not consistent with the benchmarks

Sintel Clean

	d0-10	d10-60	d60-140	s0-10	s10-40	s40+
SpyNet+ft	**5.501**	**3.122**	**1.719**	**0.832**	**3.343**	43.442
FlownetS+ft	5.992	3.561	2.193	1.424	3.815	40.098
FlownetC+ft	5.575	3.182	1.993	1.622	3.974	**33.369**

Sintel Final

	d0-10	d10-60	d60-140	s0-10	s10-40	s40+
SpyNet+ft	**6.694**	**4.368**	3.290	**1.395**	**5.534**	49.707
FlownetS+ft	7.252	4.610	**2.993**	1.873	5.826	43.236
FlownetC+ft	7.190	4.619	3.298	2.305	6.169	**40.779**

Distance from Motion Boundaries

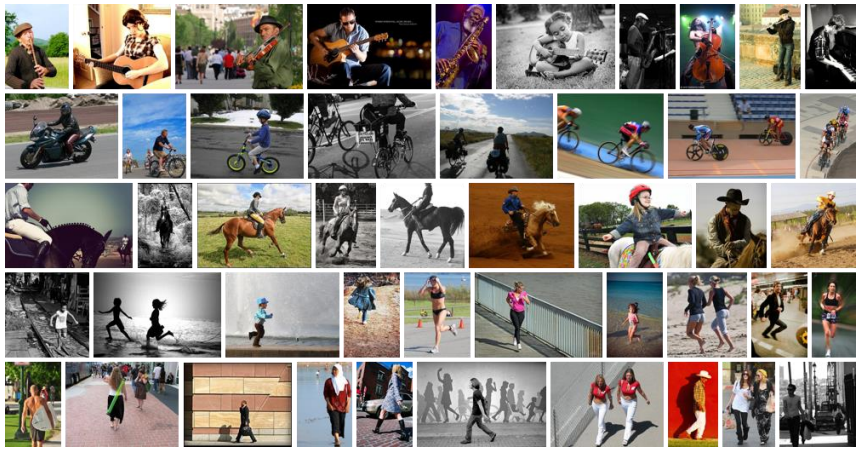
Average Displacement

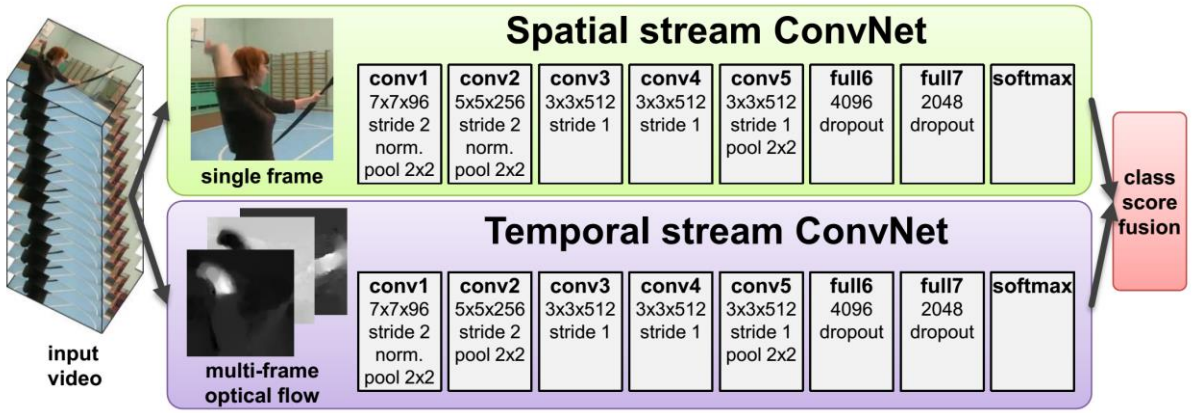
Problem

SPyNet [1]

[1] Ranjan et al. Optical Flow estimation using a Spatial Pyramid Network. CVPR 2017.

Why humans?



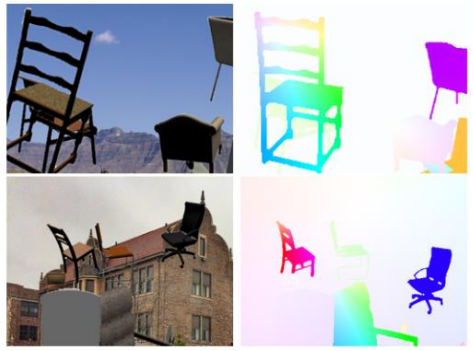


Scenes contain human actions.

- Useful for recognition problems.
- Two-stream architectures use fast classical optical flow methods.
- Deep Networks have massive GPU memory requirements.

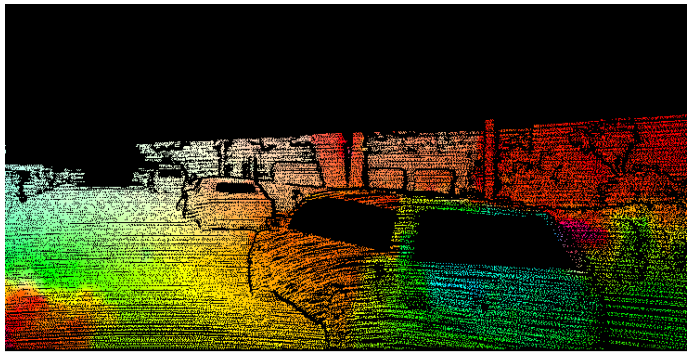
Left Image: Delaitre et al. Recognizing human actions in still images, BMVC 2010
Right Image: Simonyan et al. Two-stream convolutional networks for action recognition in videos. NIPS 2014.

Problem



Flying Chairs
[1]

MPI Sintel
[2]



KITTI [3]

No dataset for human optical flow for training neural networks.

[1] Dosovitskiy et al. Flownet: Learning optical flow with convolutional networks. ICCV 2015.
[2] Butler et al. A naturalistic open source movie for optical flow evaluation. ECCV 2012.
[3] Geiger et al. Vision meets robotics: The KITTI dataset. International Journal of Robotics Research 32.11 (2013): 1231-1237.

Idea

Create a new dataset for human optical flow.
Use it to train an existing fast and compact optical flow method.

Human Flow Dataset

Human Motion Capture data [1]

+

Realistic Human Body Model [2]

+

Environment [3]

+ Cloth texture, Lighting, Noise, Motion Blur, Camera Blur

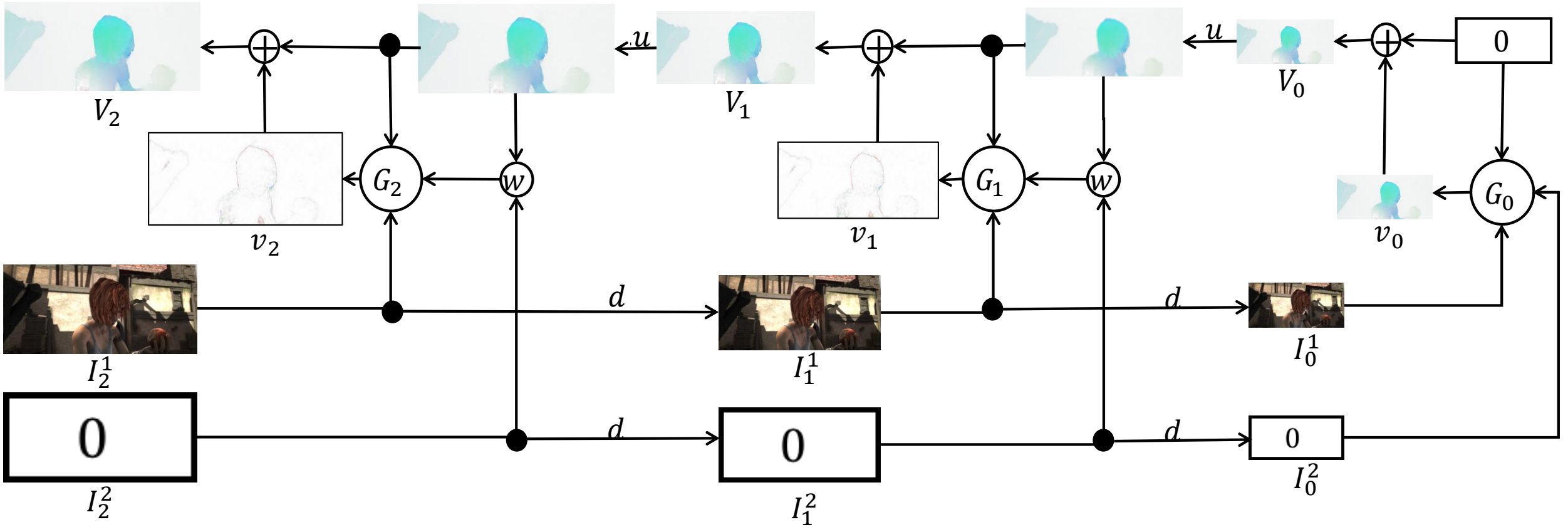
Blender

Simulate and Extract Motion Vectors

[1] Ionescu et al. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE PAMI 2014.
[2] Loper et al. MoSh: Motion and Shape Capture from Sparse Markers. SIGGRAPH Asia 2014.
[3] Yu et al. "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop." arXiv preprint arXiv:1506.03365 (2015).

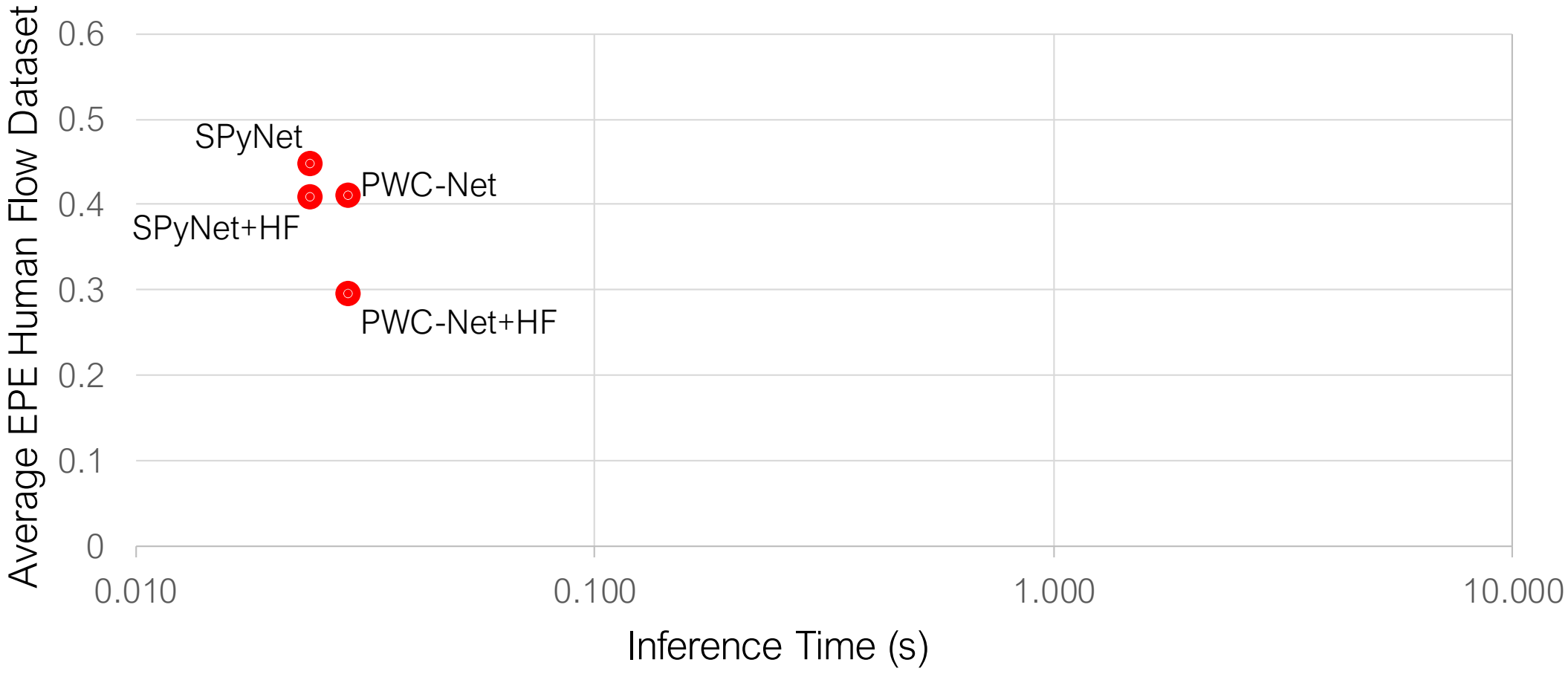
Human Flow Dataset

SPyNet

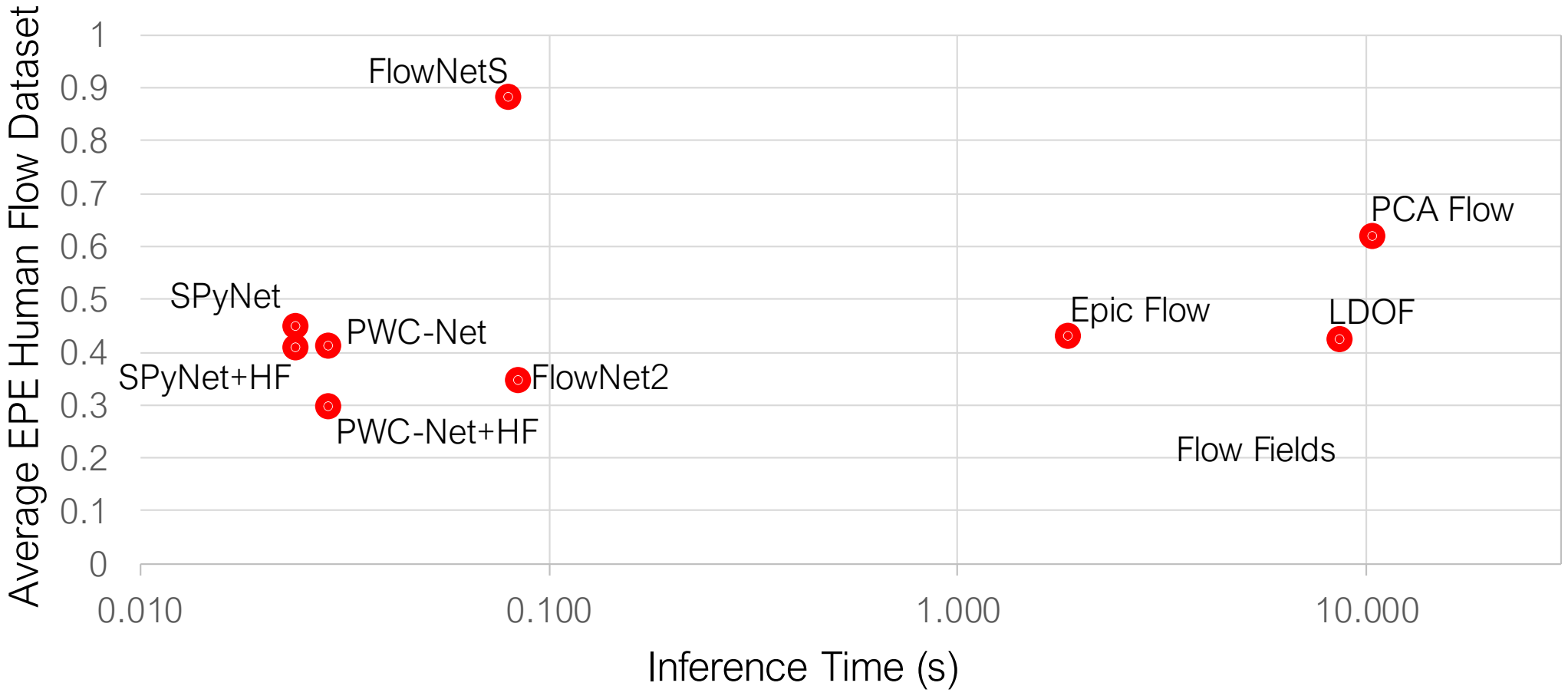


Ranjan et al. Optical Flow estimation using a Spatial Pyramid Network. CVPR 2017.

Evaluation of Optical Flow Networks



Evaluation of Optical Flow Networks



Visuals

Video | Ground Truth | Human Flow | SpyNet

Visuals

Video | Ground Truth | Human Flow | SpyNet

Visuals

Video | Ground Truth | Human Flow | SpyNet

Visuals

Video

Human Flow

SpyNet

Visuals

Video

Human Flow

SpyNet

Human Flow may not work on other parts of the scene.

Introduction to Scene Geometry

Motion of a Static Scene

For static scenes: Depth + Camera Motion = Optical Flow

Multi-view Geometry

Pinhole Camera Matrix

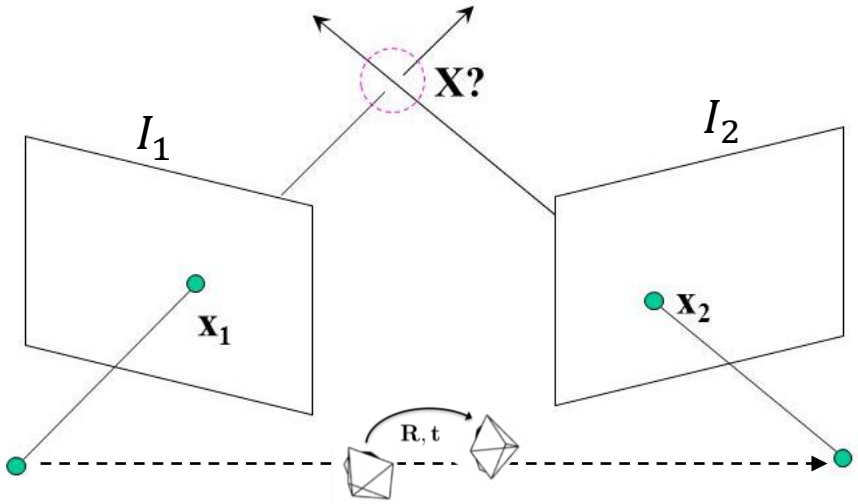
$$x_1 = KX, \qquad x_2 = K\begin{bmatrix} R \\ t \end{bmatrix} X,$$

$$X = \begin{bmatrix} \frac{d}{f} x_1 \\ d \end{bmatrix}$$

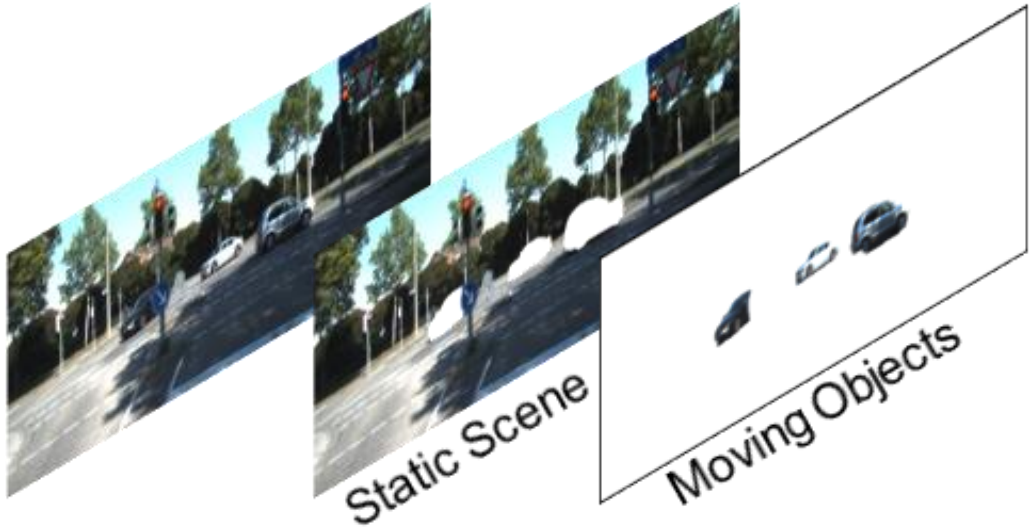
$$\| I_1(x_1) - I_2(x_2) \| = 0$$

$$\min_{R,t,d} \rho(I_1 - warp(I_2, R, t, d))$$

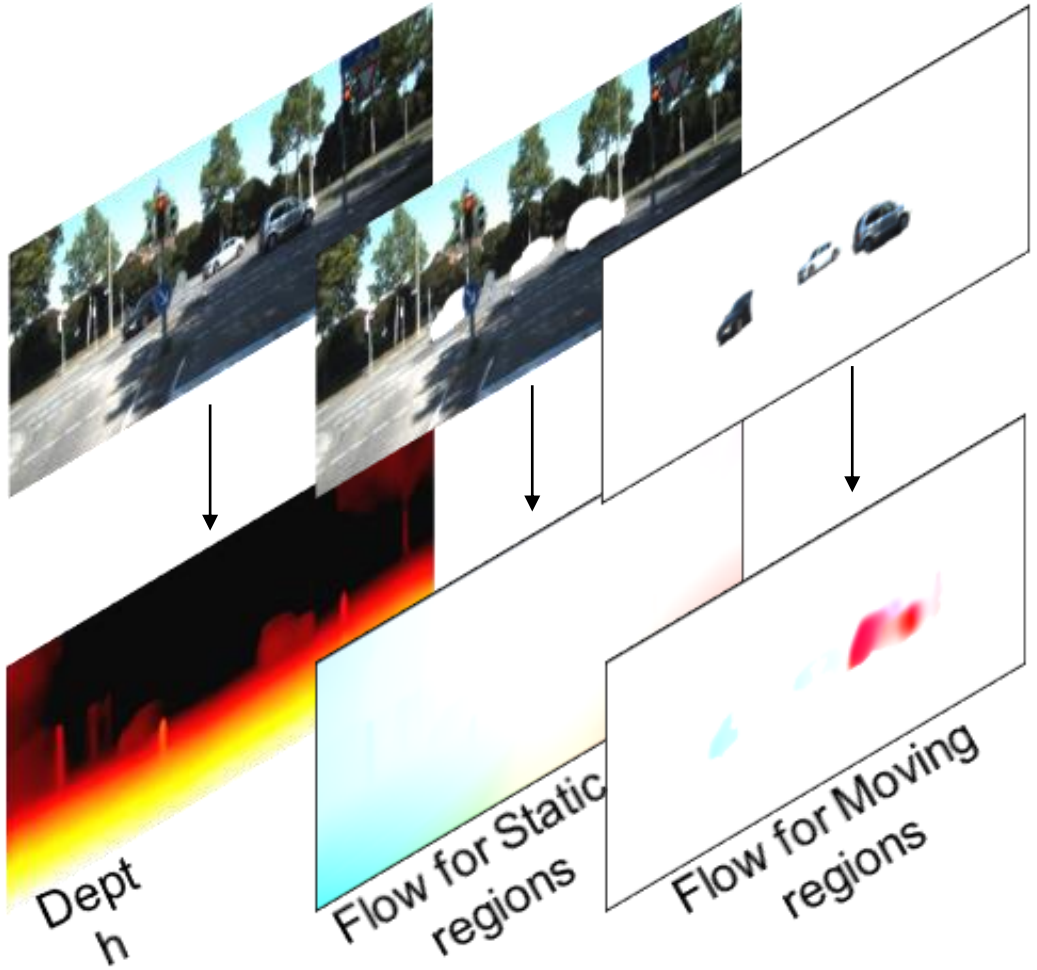
Photometric Loss



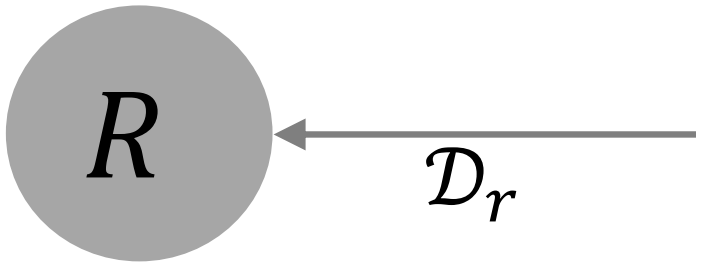
Static Scene and Moving Objects

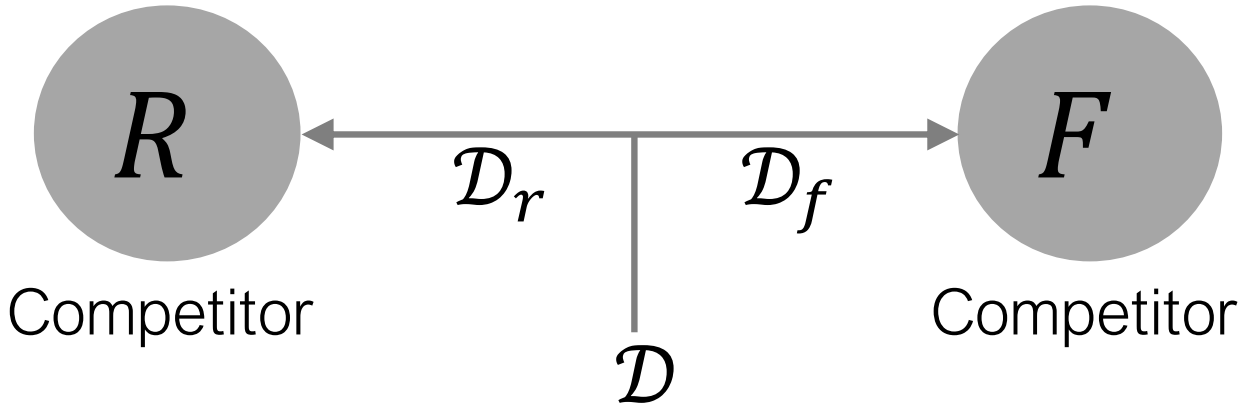


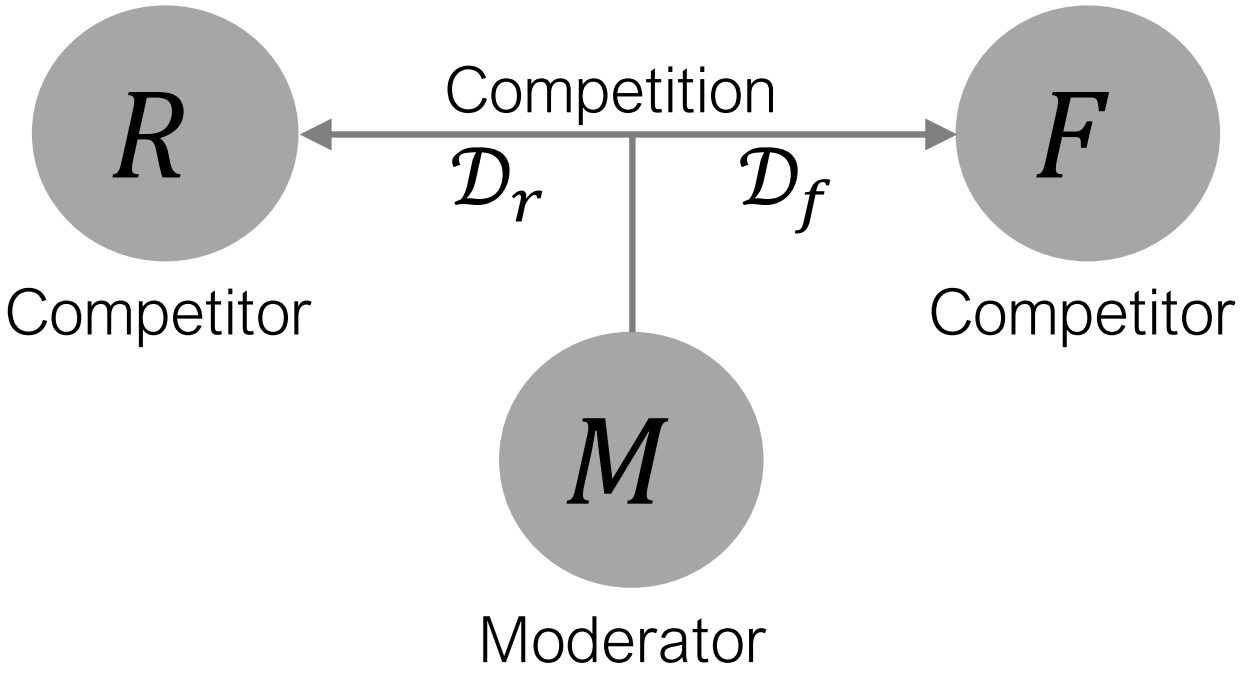
How to decompose a scene?

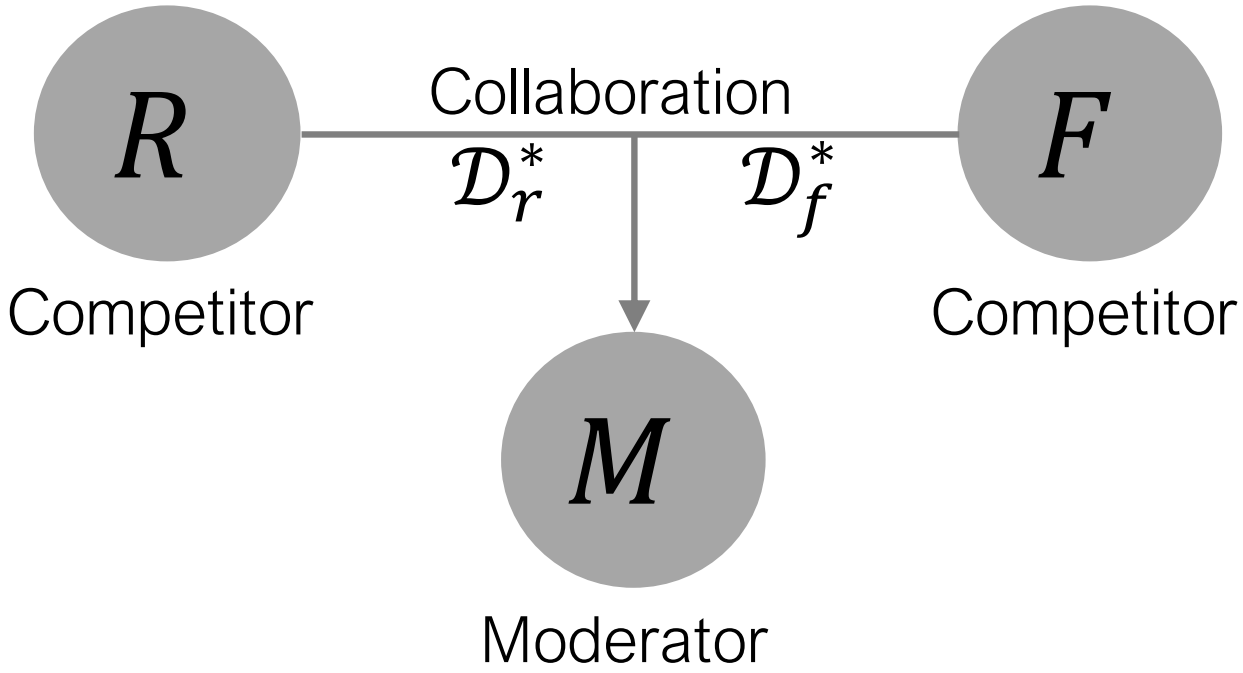



Competitive
Collaboration

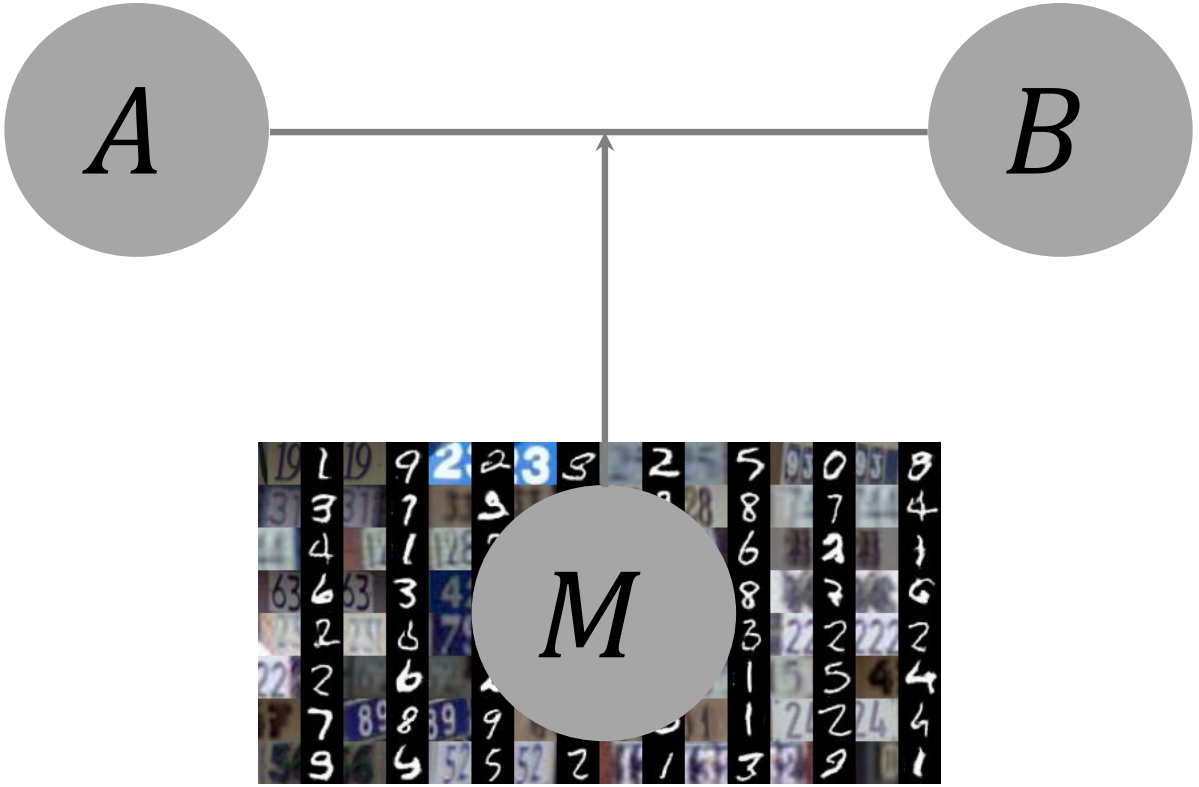

R
$\mathcal{D}_r$


R
F
$\mathcal{D}_r$
$\mathcal{D}_f$
Competitor
Competitor
$\mathcal{D}$


R
Competition
$\mathcal{D}_r$
$\mathcal{D}_f$
F
Competitor
Competitor
M
Moderator


R
Collaboration
$\mathcal{D}_r^*$
$\mathcal{D}_f^*$
F
Competitor
Competitor
M
Moderator

Mixed Domain Learning



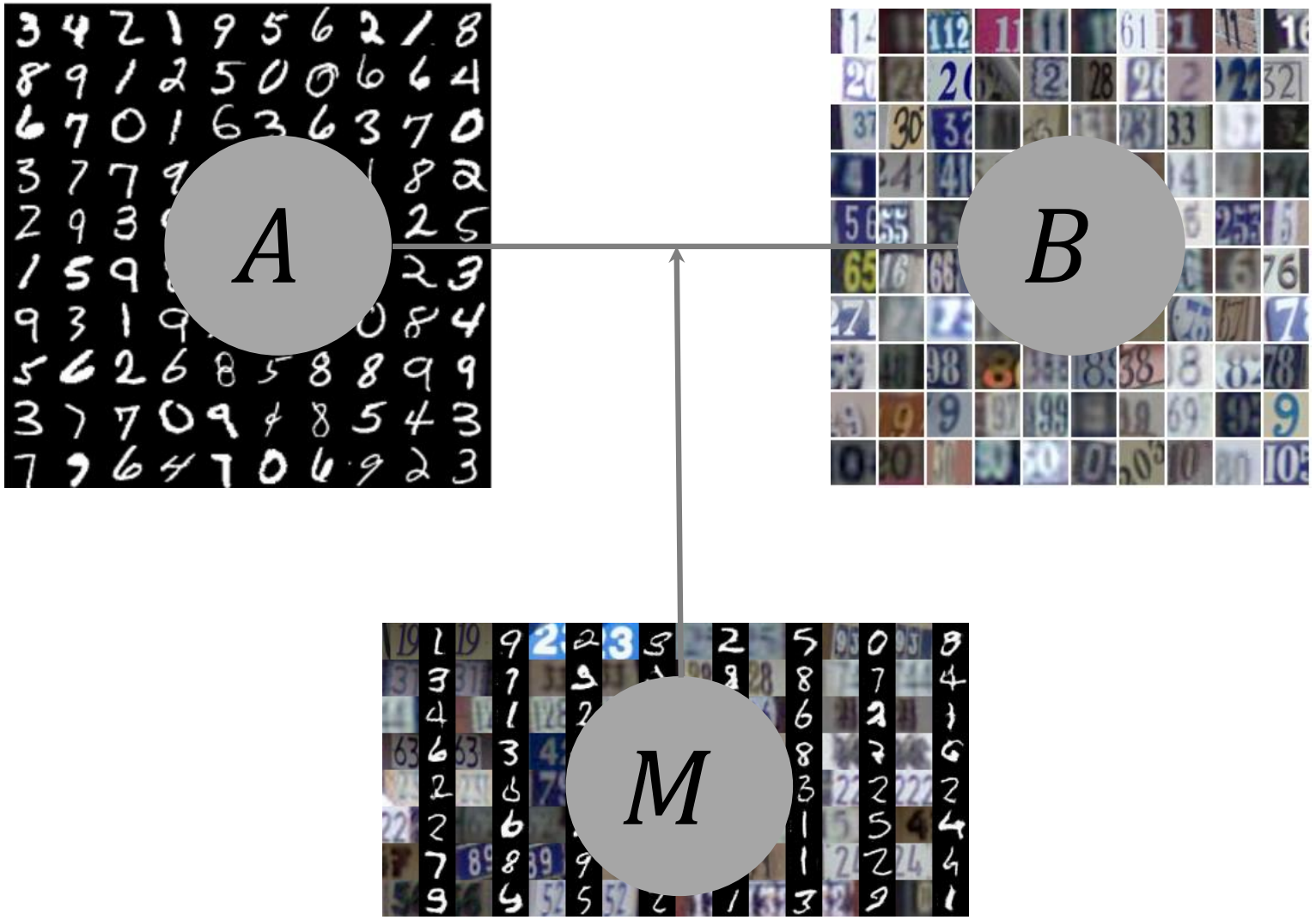
Competition Loss

$$E_{com} = m \cdot H(A(\boxed{5}), 5) + (1 - m) \cdot H(B(\boxed{5}), 5)$$

Collaboration Loss

$$E_{col} = E_{com} + \begin{cases} -\log(M(y) + \epsilon) & if\ E_A < E_B \\ -\log(1 - M(y) + \epsilon) & if\ E_A \geq E_B \end{cases}$$

$$E_A = H(A\ (\ \ ), 5)$$


A
B
M

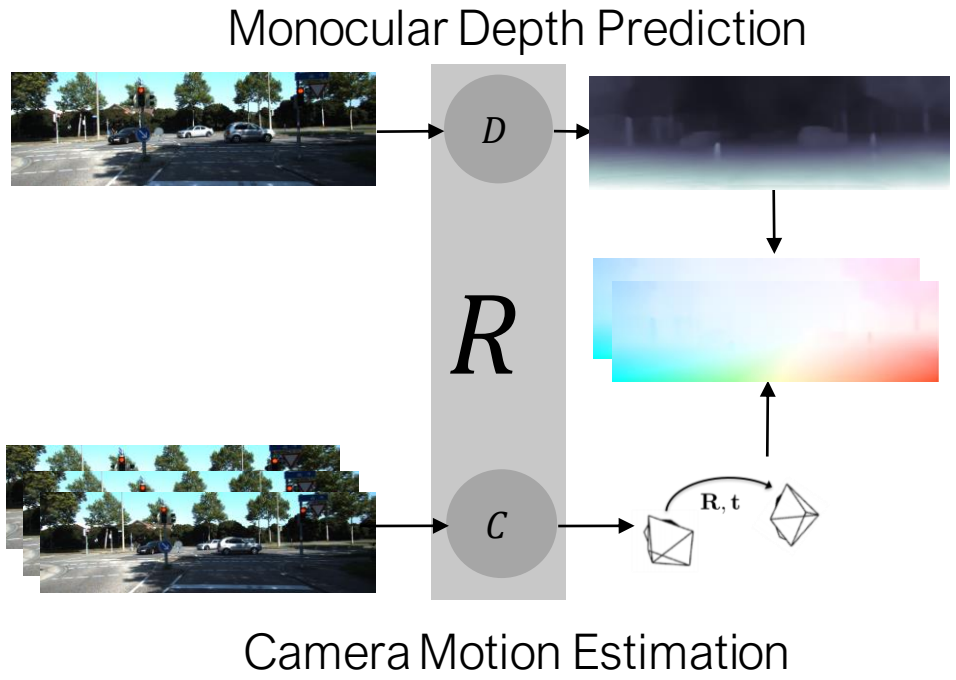
Accuracy

Model	Training	MNIST Error	SVHN Error	MNIST+SVHN Error
Alice	Basic	1.34	11.88	8.96
Alice	CC	1.41	**11.55**	8.74
Bob	CC	**1.24**	11.75	8.84
Alice+Bob+Mod	CC	**1.24**	**11.55**	**8.70**
Alice 3x	Basic	1.33	10.86	8.22

Moderator Behavior

	Alice	Bob
MNIST	0 %	100 %
SVHN	100 %	0 %

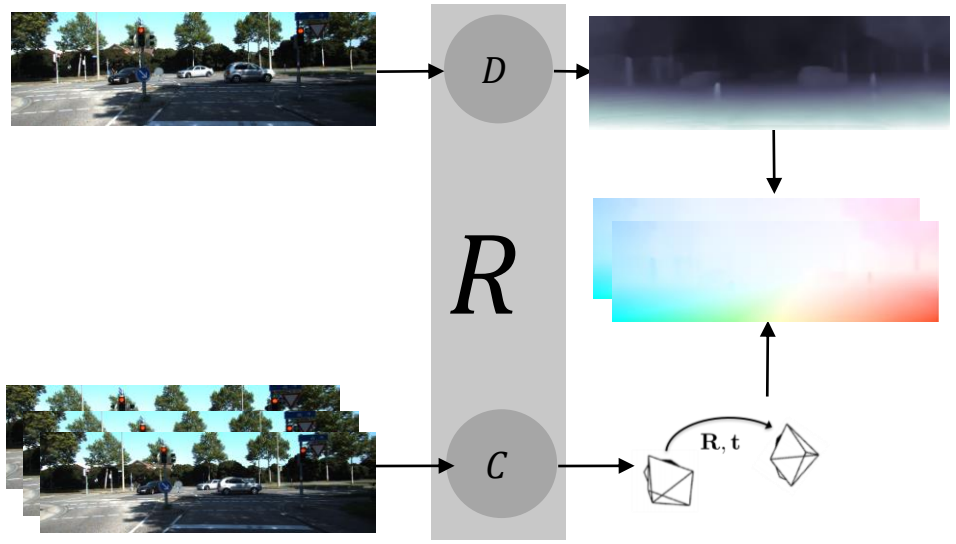
Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation



Zhou et al. CVPR 2017

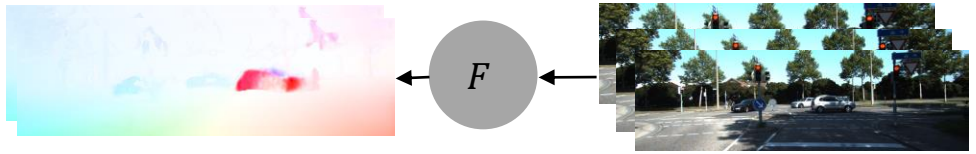
Meister et al. AAAI ‘18, Janai et al. ECCV ‘18

Monocular Depth Prediction

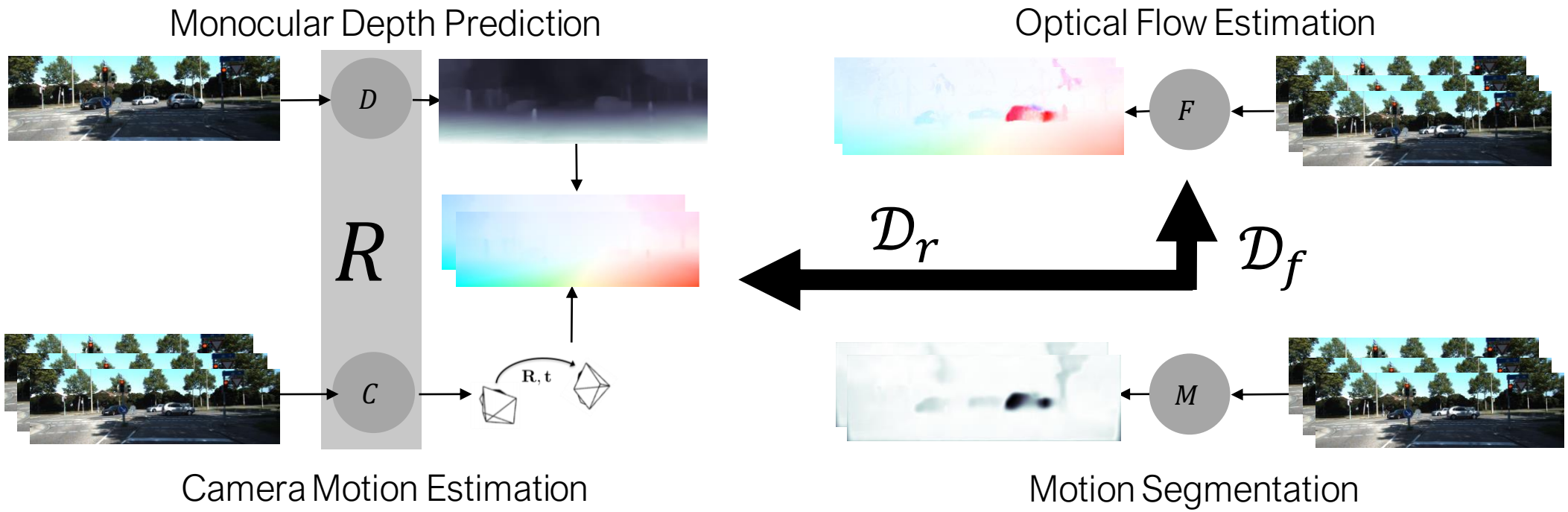


Camera Motion Estimation

Optical Flow Estimation



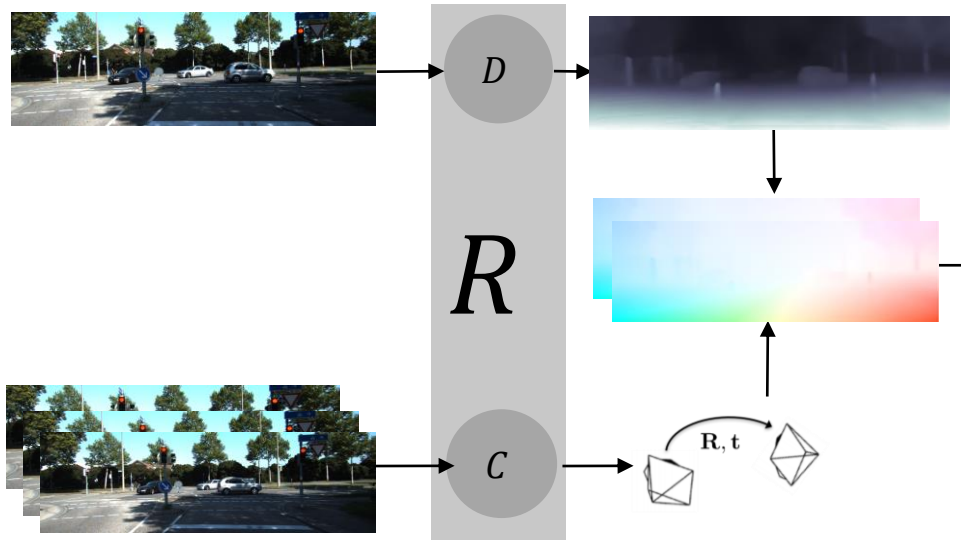
Zhou et al. CVPR 2017


Monocular Depth Prediction
D
R
C
R, t
Optical Flow Estimation
F
$\mathcal{D}_r$
$\mathcal{D}_f$
M
Camera Motion Estimation
Motion Segmentation

Photometric Loss

$$E_R = \rho(I, \text{warp}(I_+, c, d\,)) \cdot m$$

Monocular Depth Prediction

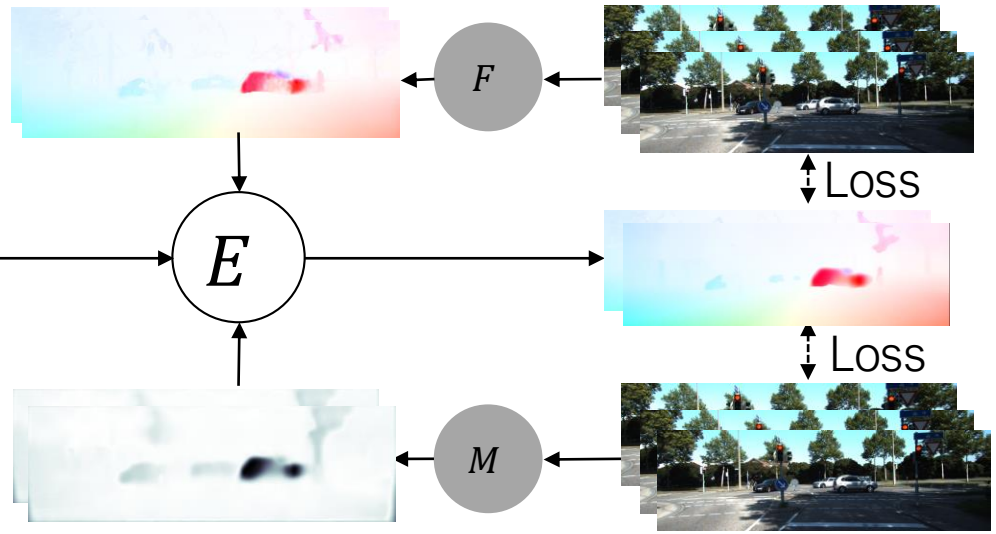


Camera Motion Estimation

Photometric Loss

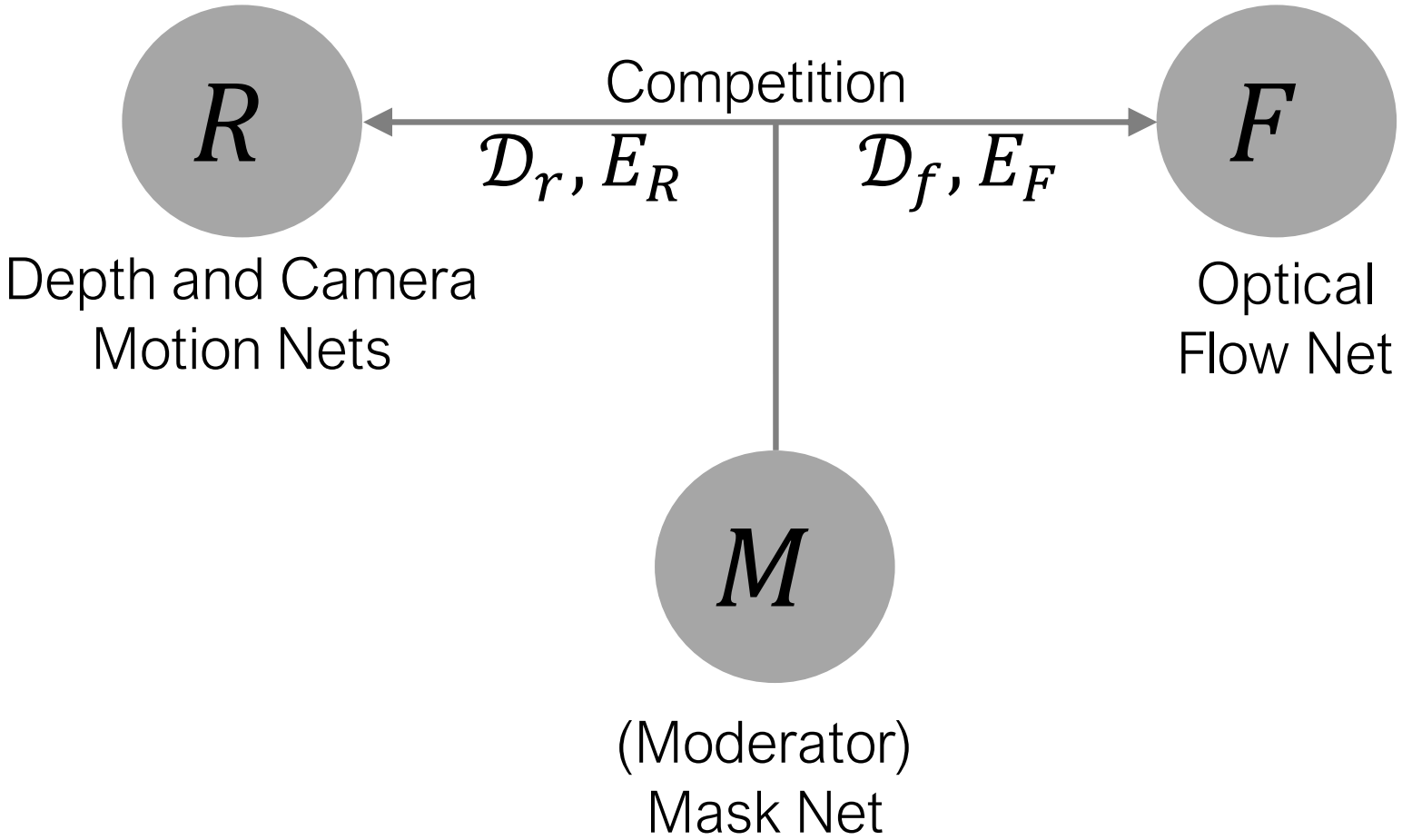
$$E_F = \rho(I, \text{warp}(I_+, u_+\,)) \cdot (1 - m)$$

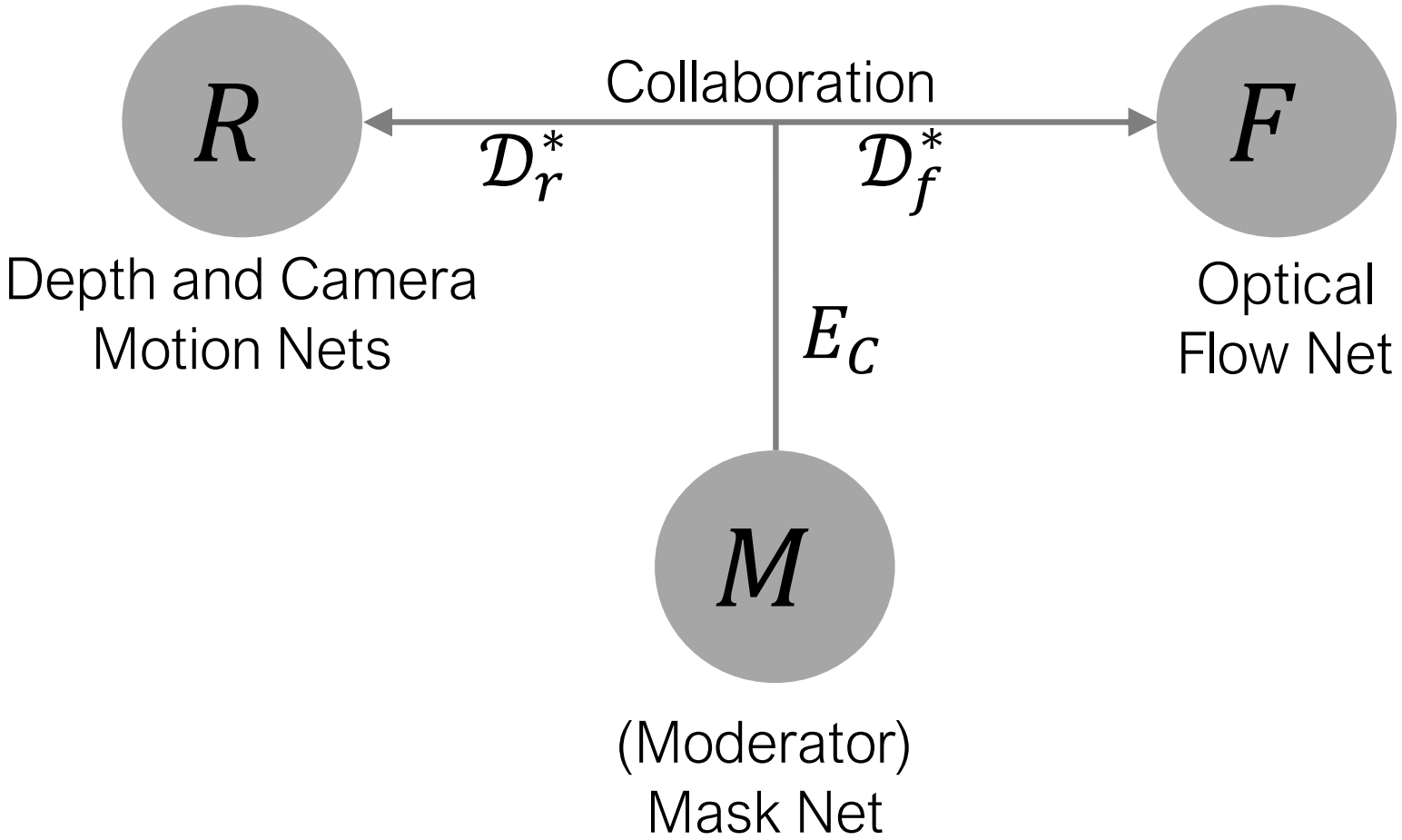
Optical Flow Estimation



Motion Segmentation

$$E_C = H(\boldsymbol{I}_{\|u_R - u_F\| < \lambda_c}, m)$$


R
Competition
$\mathcal{D}_r, E_R$
$\mathcal{D}_f, E_F$
F
Depth and Camera
Motion Nets
Optical
Flow Net
M
(Moderator)
Mask Net


R
Collaboration
F
$\mathcal{D}_r^*$
$\mathcal{D}_f^*$
Depth and Camera
Motion Nets
E_C
Optical
Flow Net
M
(Moderator)
Mask Net

Best amongst **Unsupervised Methods** on
Single View Depth Prediction
Camera Motion Estimation
Optical Flow

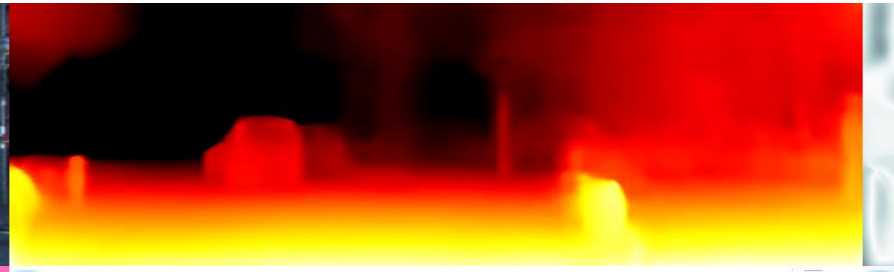
Only Network that does
Unsupervised Motion Segmentation

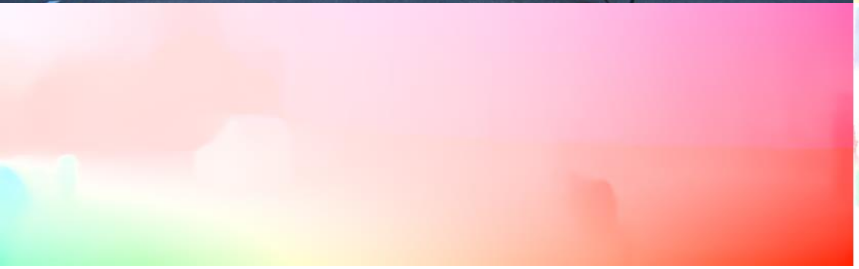
Results

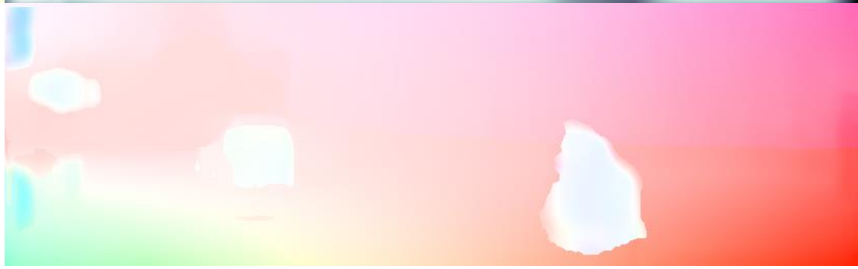
Image

Depth

Segmentation







Static Flow

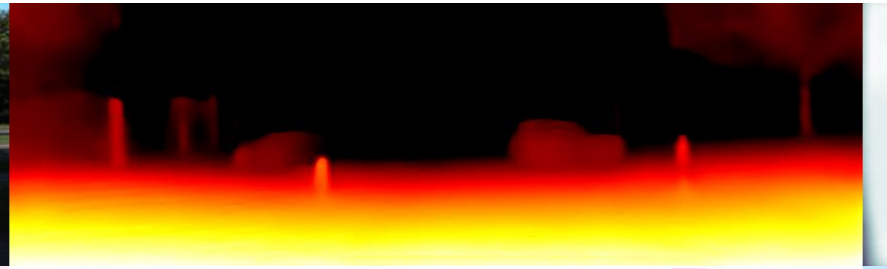
Segmented Dynamic Flow

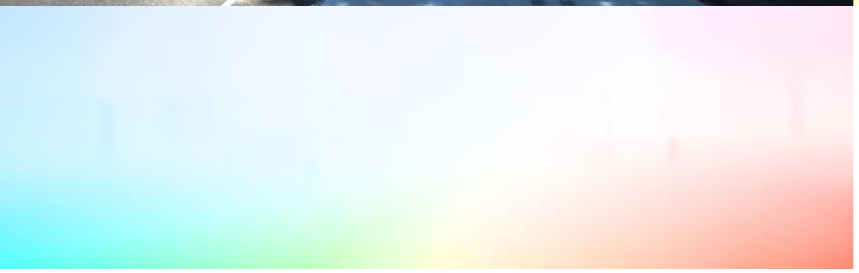
Full Flow

Image

Depth

Segmentation





Static Flow

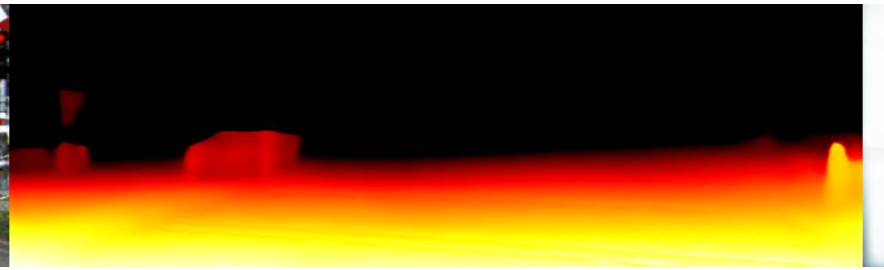
Segmented Dynamic Flow

Full Flow

Image

Depth

Segmentation



Static Flow

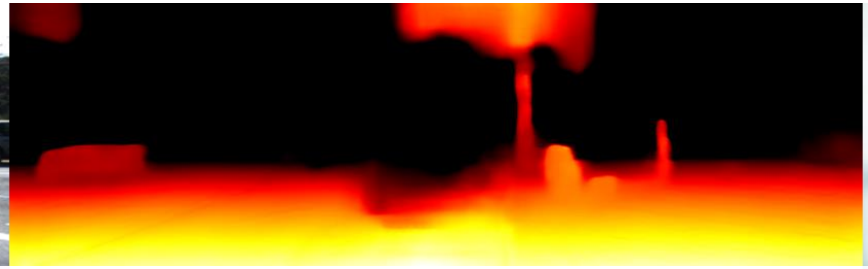
Segmented Dynamic Flow

Full Flow

Image

Depth

Segmentation



Static Flow

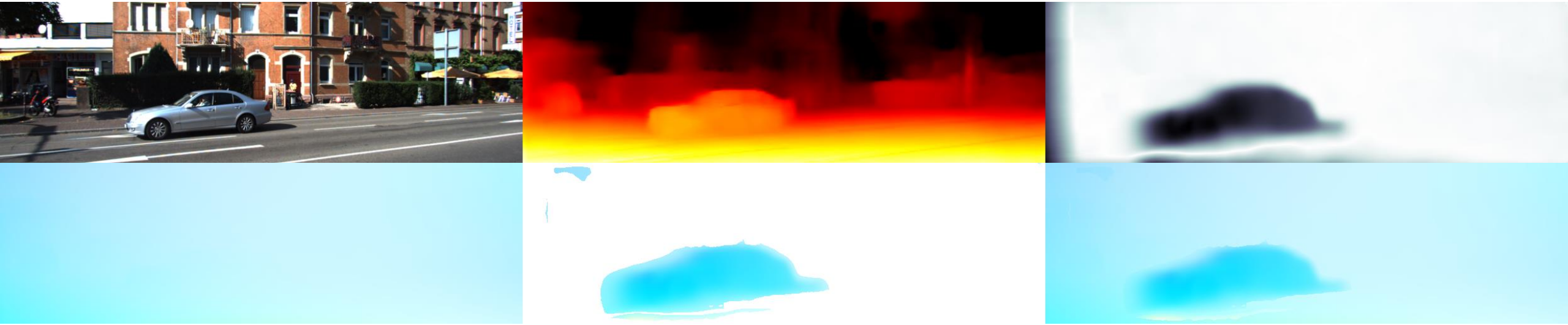
Segmented Dynamic Flow

Full Flow

Image

Depth

Segmentation



Static Flow

Segmented Dynamic Flow

Full Flow

Depth Evaluation

Model	Dataset	AbsRel	SqRel	RMS	RMSlog
Eigen et al. 2014	KITTI	0.203	1.548	6.307	0.282
Zhou et al. 2017	KITTI	0.183	1.595	6.709	0.270
Geonet 2018	KITTI	0.155	1.296	5.857	0.233
DF-Net 2018	KITTI	0.150	1.124	5.507	0.223
Ours	KITTI	**0.140**	**1.070**	**5.326**	**0.217**
Zhou et al. 2017	CS+KITTI	0.198	1.836	6.565	0.275
Geonet 2018	CS+KITTI	0.153	1.328	5.737	0.232
DF-Net 2018	CS+KITTI	0.146	1.182	5.215	**0.213**
Ours	CS+KITTI	**0.139**	**1.032**	**5.199**	**0.213**
Godard et al.	CS+KITTI+S	0.114	0.991	5.029	0.203

[1] Zhou et al. Unsupervised learning of depth and ego-motion from video. CVPR 2017.
[2] Yin et at. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. CVPR 2018..
[3] Zou et al. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. ECCV 2018.

Depth Ablation

Model	Dataset	Net D	Net F	AbsRel	SqRel	RMS	RMSlog
Basic	KITTI	DispNet	-	0.168	1.396	6.176	0.244
CC	KITTI	DispNet	FlowNetC	0.148	1.149	5.464	0.226
CC	KITTI	DispResNet	FlowNetC	0.144	1.284	5.716	0.226
CC	KITTI	DispResNet	PWC Net	0.140	1.070	5.326	0.217
CC	CS+KITTI	DispResNet	PWC Net	**0.139**	**1.032**	**5.199**	**0.213**

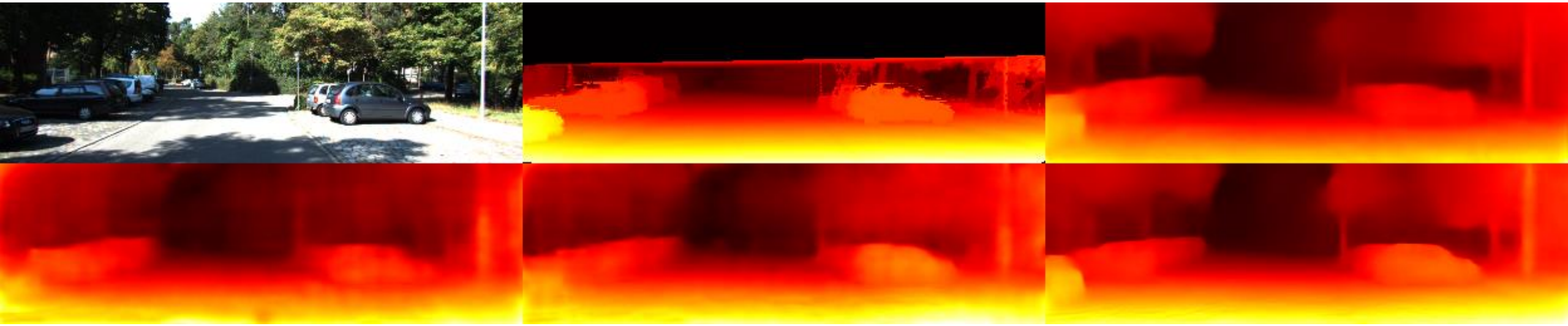
DispResNet > DispNet
PWC Net > FlowNetC

Depth Visuals

Image

Ground Truth

CC (CS+K)



Basic

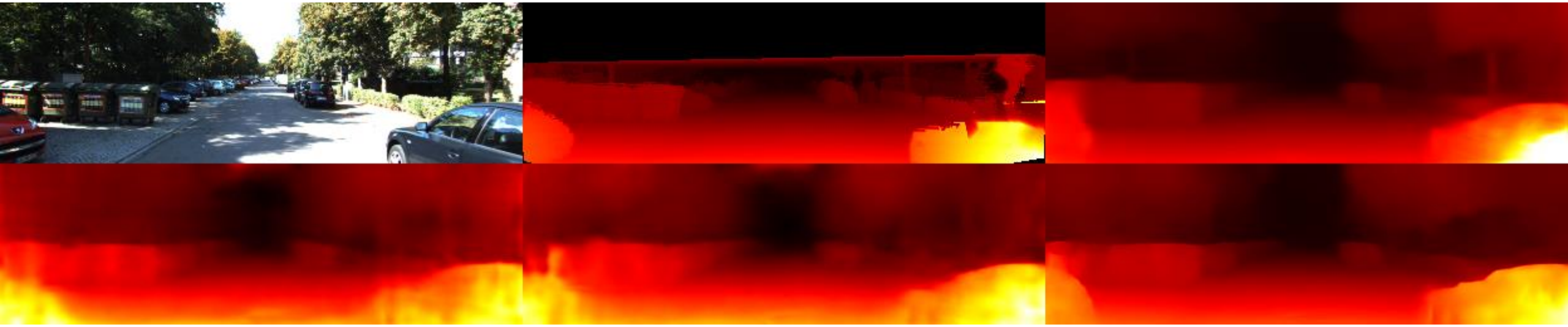
Basic+ssim

CC (K)

Image

Ground Truth

CC (CS+K)

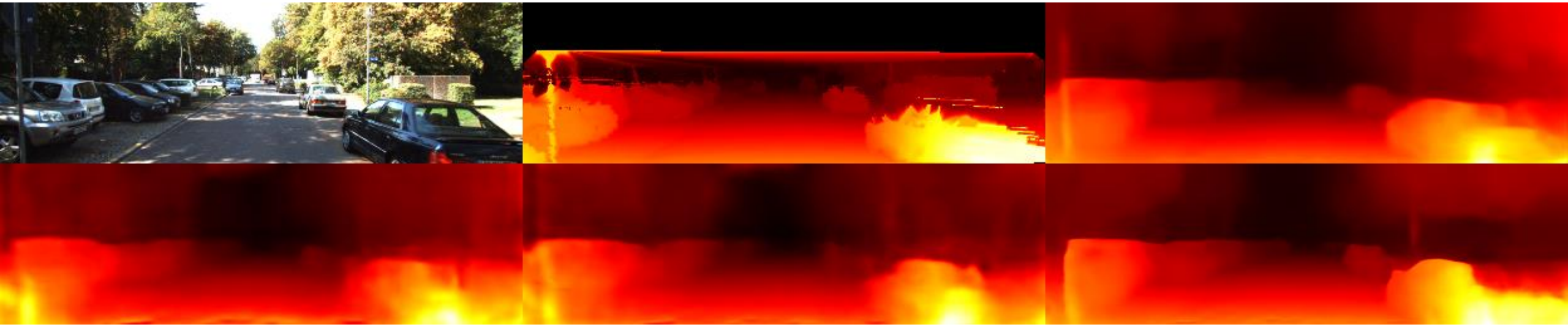


Basic

Basic+ssim

CC (K)

Image Ground Truth CC (CS+K)

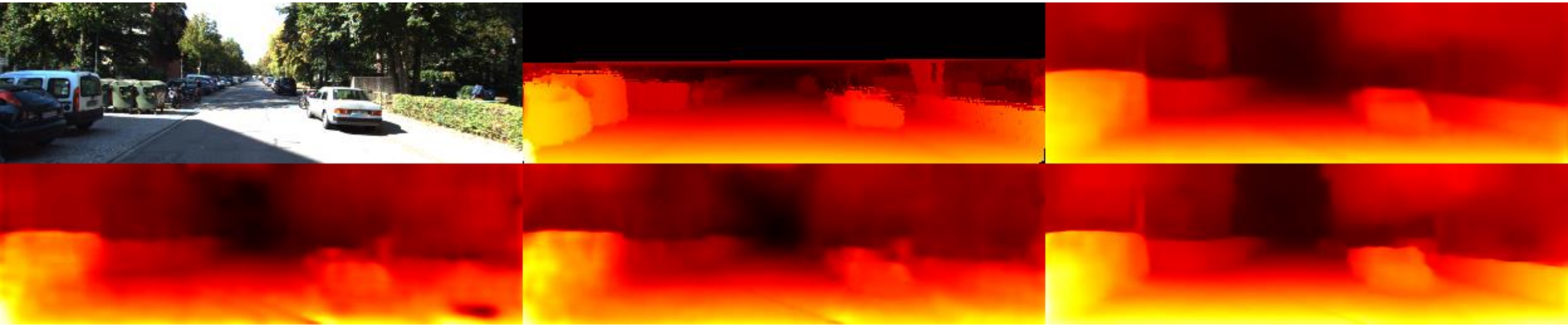


Basic Basic+ssim CC (K)

Image

Ground Truth

CC (CS+K)



Basic

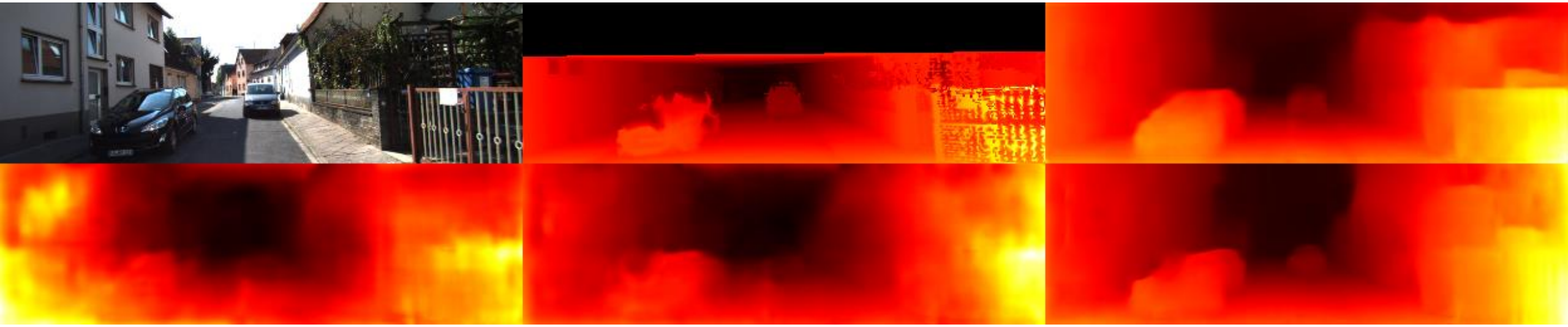
Basic+ssim

CC (K)

Image

Ground Truth

CC (CS+K)



Basic

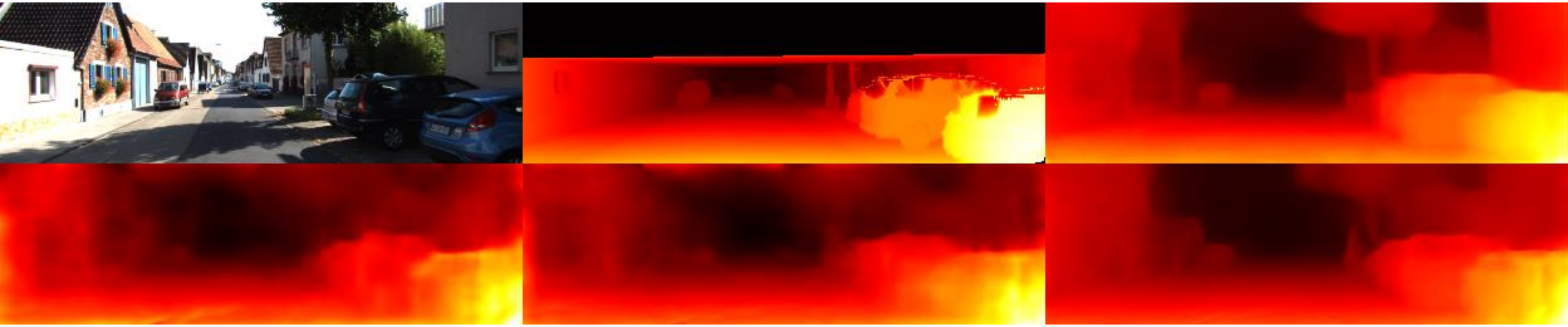
Basic+ssim

CC (K)

Image

Ground Truth

CC (CS+K)



Basic

Basic+ssim

CC (K)

Pose Evaluation

Model	Sequence 09	Sequence 10
ORB-SLAM	0.014 ± 0.008	0.012 ± 0.011
Zhou et al. 2017	0.016 ± 0.009	0.013 ± 0.009
Geonet 2018	**0.012 ± 0.007**	0.012 ± 0.009
DF-Net 2018	0.017 ± 0.007	0.015 ± 0.009
Ours	**0.012 ± 0.007**	**0.012 ± 0.008**

[1] Zhou et al. Unsupervised learning of depth and ego-motion from video. CVPR 2017.
[2] Yin et at. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. CVPR 2018..
[3] Zou et al. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. ECCV 2018.

Flow Evaluation on KITTI

Model	EPE	Fl	Test Fl
UnFlow-CSS 2018	8.10	23.27 %	-
Back2Future 2018	6.59	24.21 %	**22.94 %**
Geonet 2018	10.81	-	-
DF-Net 2018	8.98	26.41 %	25.70 %
Ours	**5.66**	**20.93 %**	25.27 %
PWC-Net 2018	10.35	33.67%	-
PWC-Net+ft 2018	(2.16)	(9.80%)	9.60%

[1] Janai et al. Unsupervised learning of multi-frame optical flow with occlusions. ECCV 2018.
[2] Yin et at. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. CVPR 2018..
[3] Zou et al. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. ECCV 2018.

Flow Visuals

Image 1

Ground Truth

UnFlow-CSS

GeoNet

DF-Net

Ours

[1] Meister et al. Unsupervised learning of optical flow with a bidirectional census loss. AAAI 2018.
[2] Yin et at. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. CVPR 2018..
[3] Zou et al. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. ECCV 2018.

Image 1

Ground Truth

UnFlow-CSS

GeoNet

DF-Net

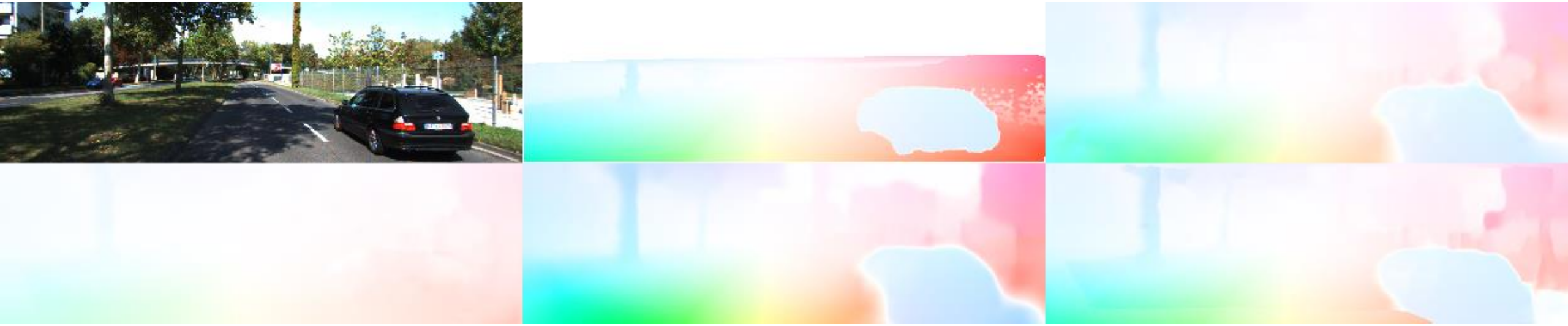
Ours

[1] Meister et al. Unsupervised learning of optical flow with a bidirectional census loss. AAAI 2018.
[2] Yin et at. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. CVPR 2018..
[3] Zou et al. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. ECCV 2018.

Image 1

Ground Truth

UnFlow-CSS



GeoNet

DF-Net

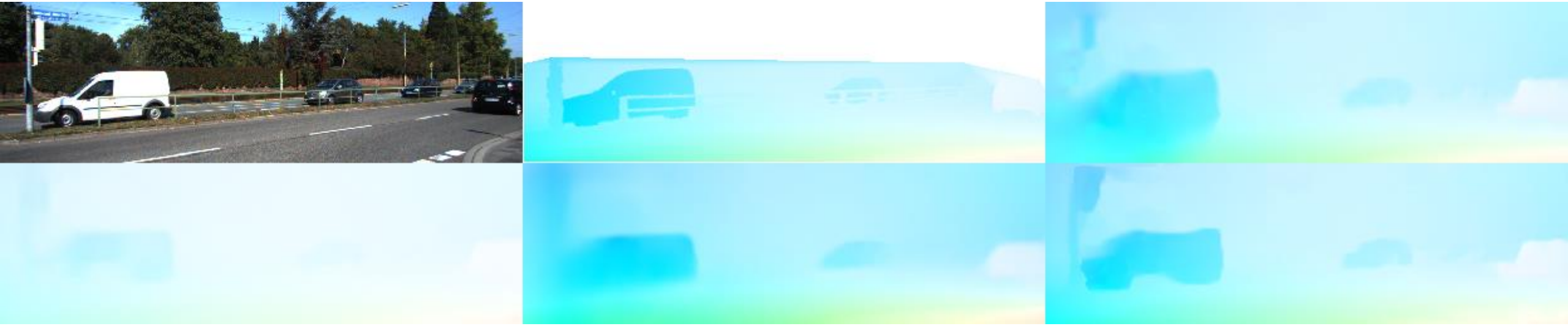
Ours

[1] Meister et al. Unsupervised learning of optical flow with a bidirectional census loss. AAAI 2018.
[2] Yin et at. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. CVPR 2018..
[3] Zou et al. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. ECCV 2018.

Image 1

Ground Truth

UnFlow-CSS



GeoNet

DF-Net

Ours

[1] Meister et al. Unsupervised learning of optical flow with a bidirectional census loss. AAAI 2018.
[2] Yin et at. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. CVPR 2018..
[3] Zou et al. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. ECCV 2018.

Image 1

Ground Truth

UnFlow-CSS

GeoNet

DF-Net

Ours

[1] Meister et al. Unsupervised learning of optical flow with a bidirectional census loss. AAAI 2018.
[2] Yin et at. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. CVPR 2018..
[3] Zou et al. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. ECCV 2018.

[1] Meister et al. Unsupervised learning of optical flow with a bidirectional census loss. AAAI 2018.
[2] Yin et at. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. CVPR 2018..
[3] Zou et al. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. ECCV 2018.

What's next?

supervision

Future Goal

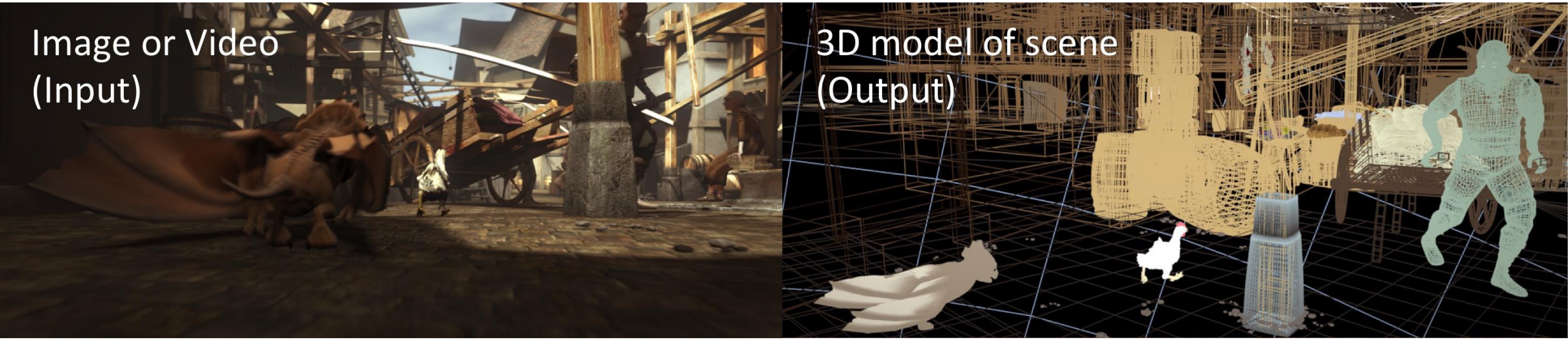
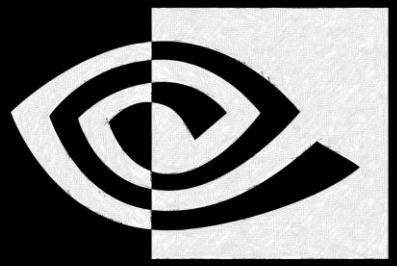


Image courtesy: https://ps.is.tuebingen.mpg.de/research_fields/inverse-graphics



github.com/anuragranj

Michael Black (MPI), Jonas Wulff (MIT), Timo Bolkart (MPI), Siyu Tang (MPI), Joel Janai (MPI), Deqing Sun (NVIDIA),
Fatma Güney (Oxford), Varun Jampani (NVIDIA), Andreas Geiger (MPI), Clément Pinard (ENSTA),
Soubhik Sanyal (MPI), Yiyi Liao (MPI), George Pavlakos (UPenn), Kihwan Kim (NVIDIA),
Lukas Balles (MPI), Frederick Künstner (MPI), Dimitris Tzionas (MPI),
David Hoffmann (MPI)