Delving Deep into Computer Vision

Caner Hazirbas

Machine Learning Meetup #1

Delving Deep into Computer Vision

FlowNet

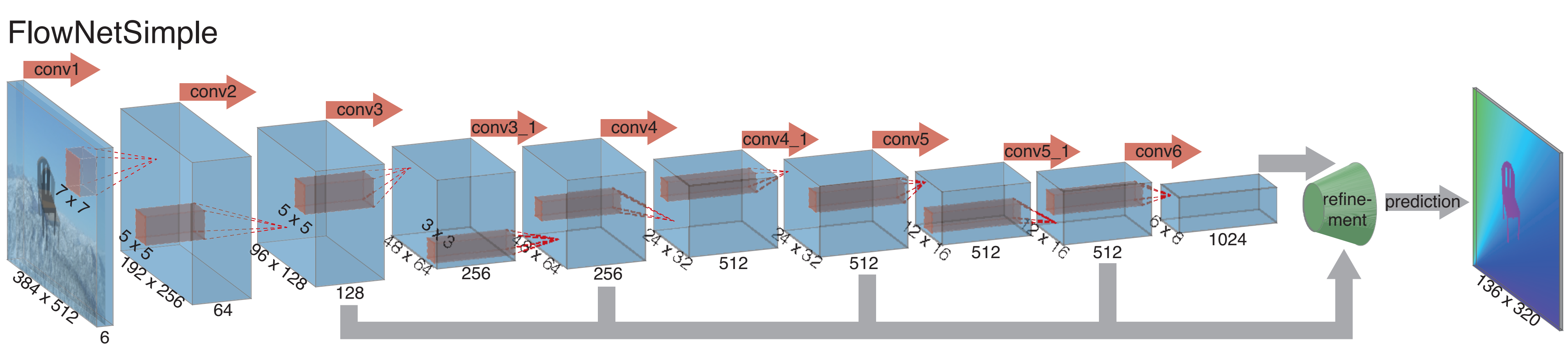
FuseNet

PoseLSTM

DDFF

Delving Deep into Computer Vision

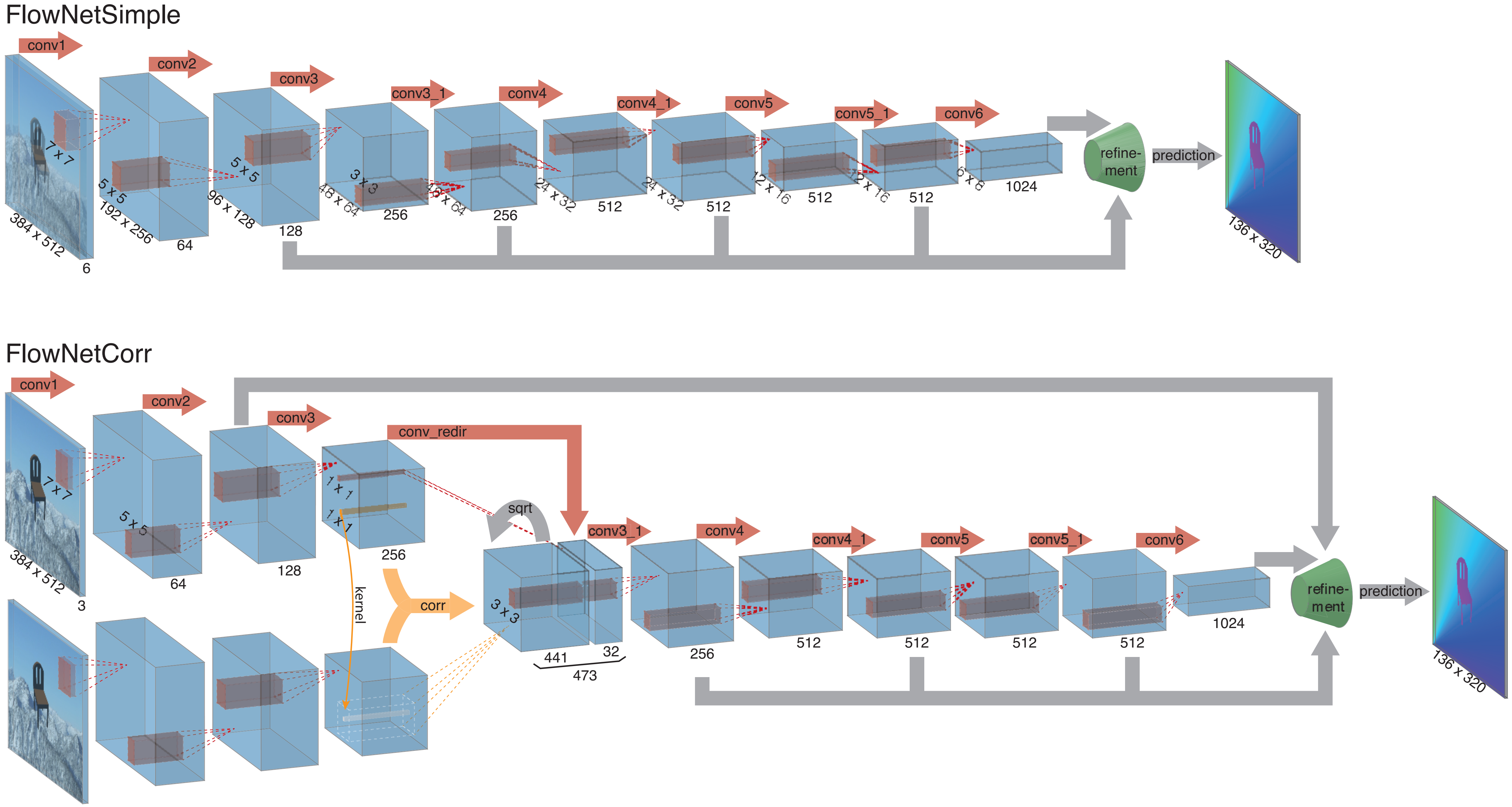
FlowNet



FlowNet

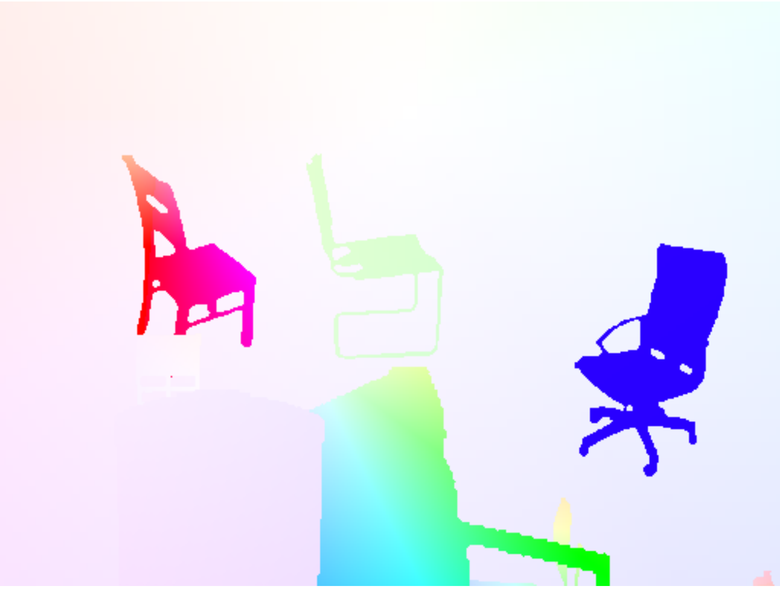
Learning Optical Flow with Convolutional Networks

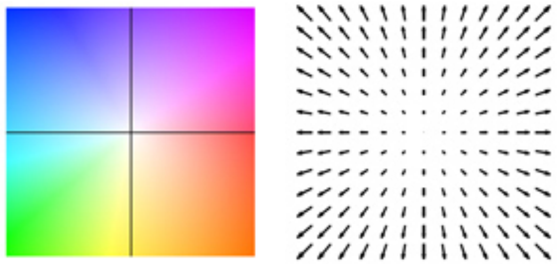
ICCV'15



FlowNet

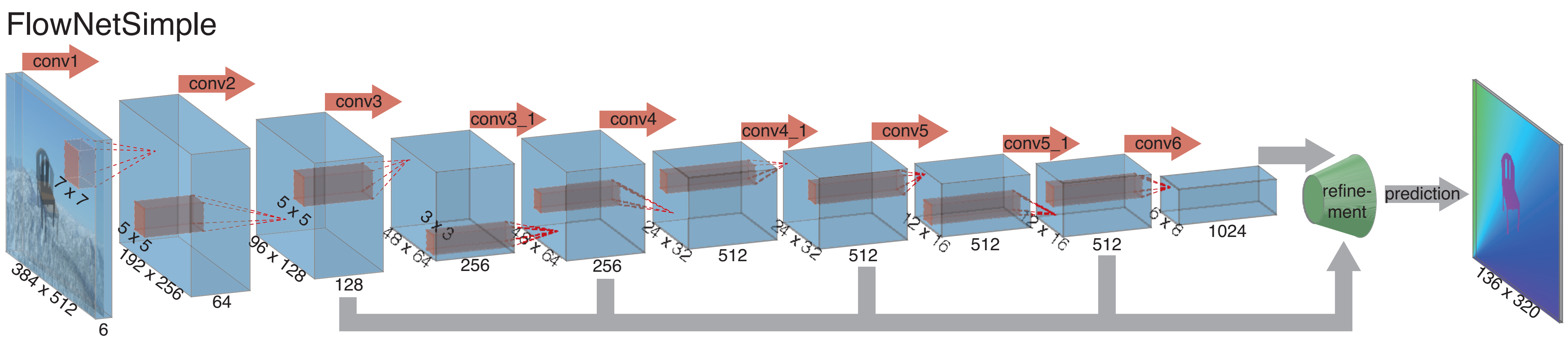
Flying Chairs

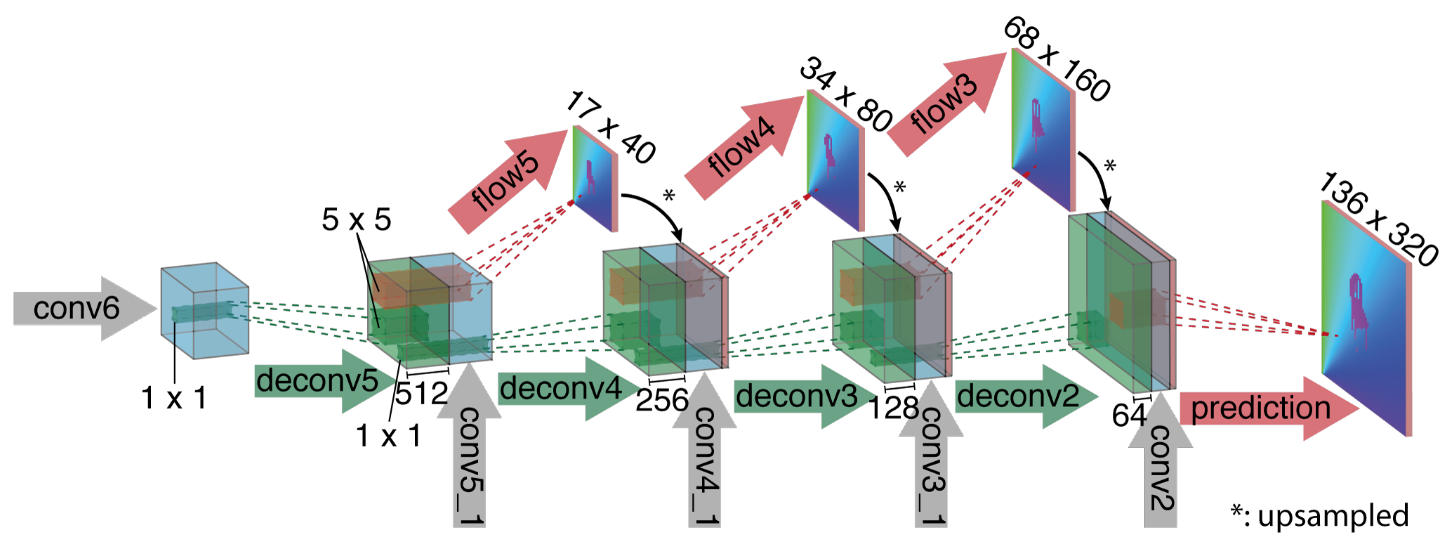




FlowNet

FlowNetSimple

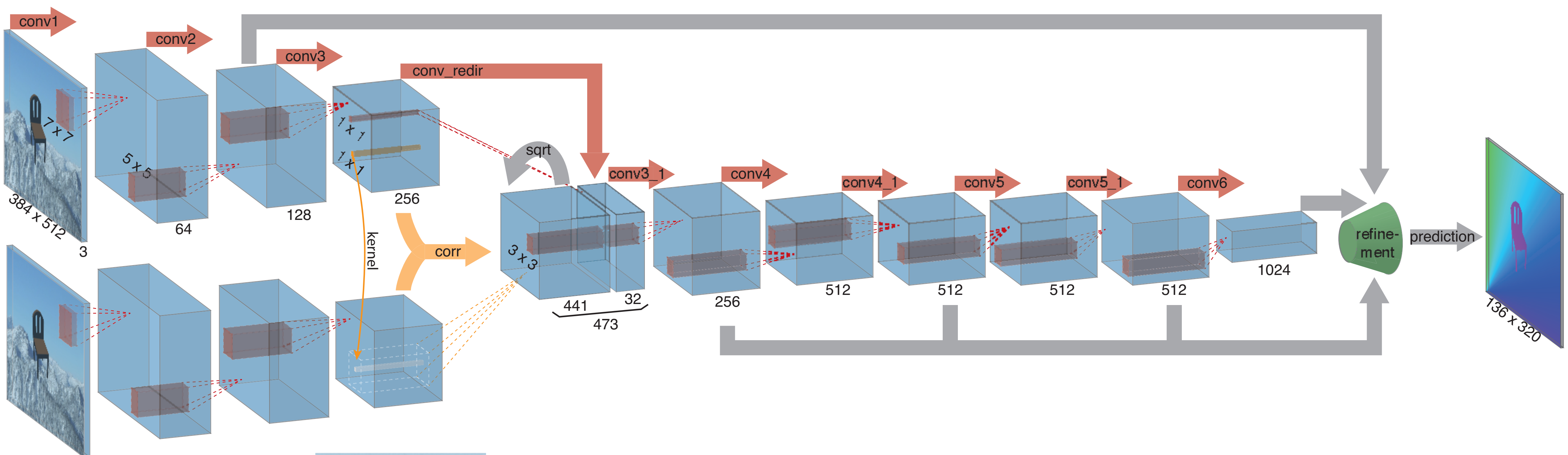


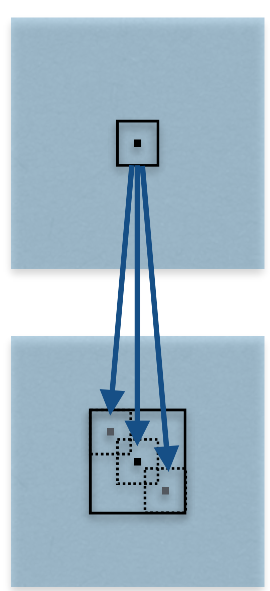


FlowNet

FlowNetCorr

FlowNetCorr





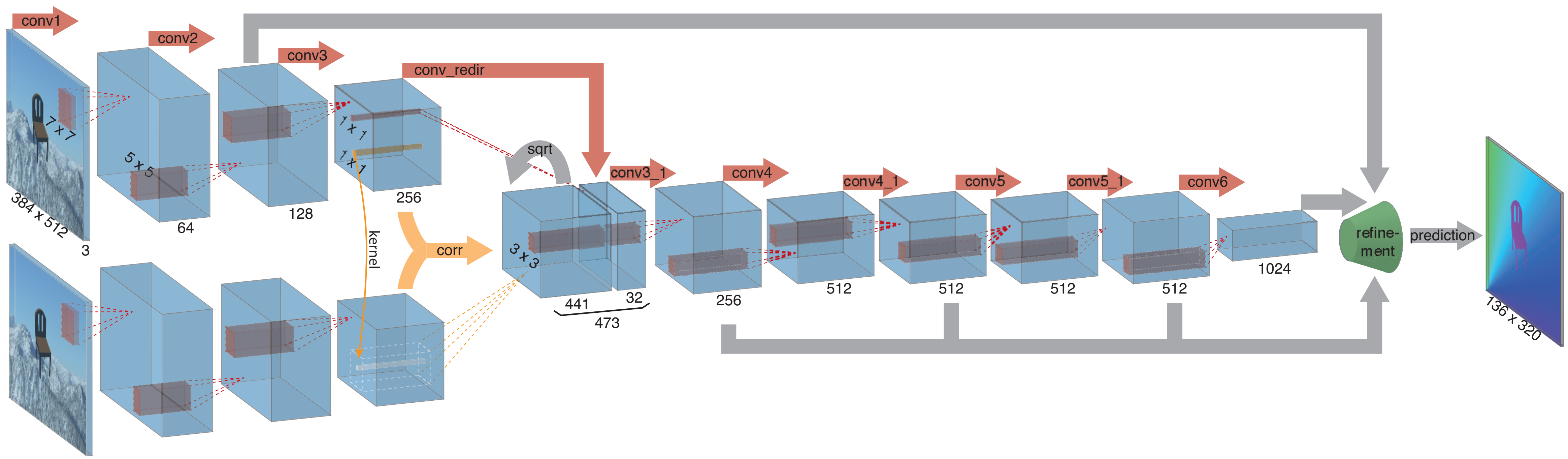
$$c(x_1, x_2) = \sum_{o \in [-k,k] \times [-k,k]} \langle f_1(x_1 + o), f_2(x_2 + o) \rangle \ ,$$

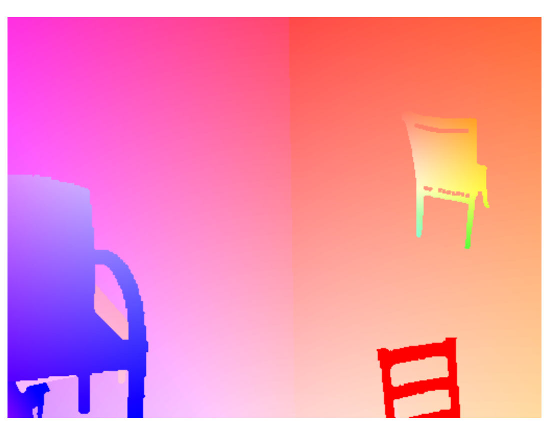
$$K := 2k + 1$$

FlowNet

Simple vs. Corr Flying Chairs

FlowNetCorr





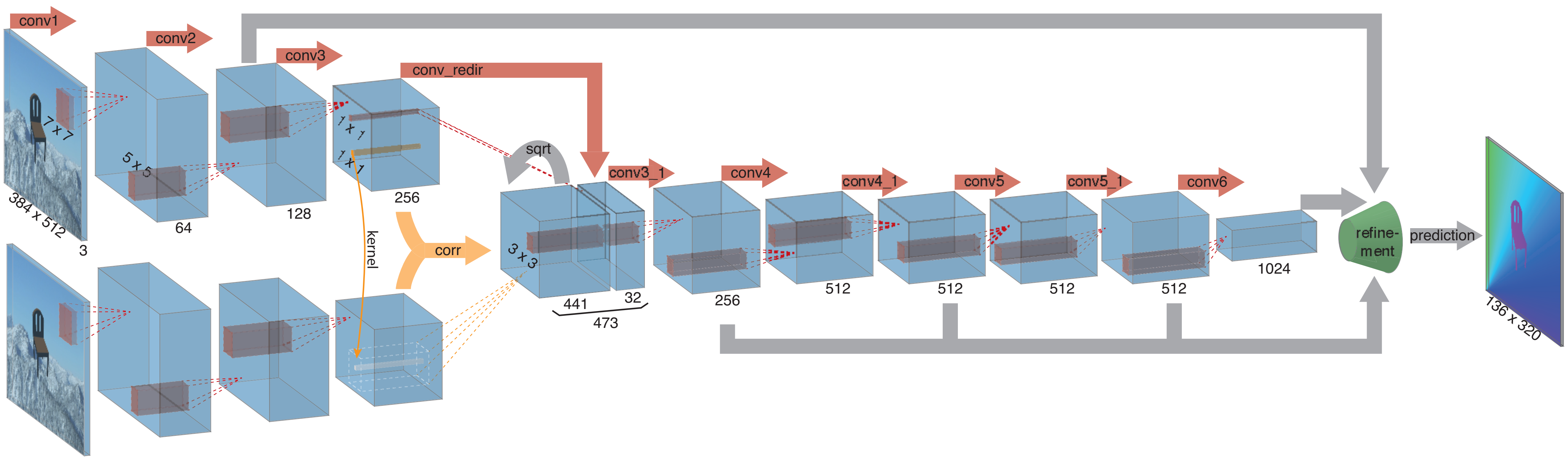
FlowNetS

FlowNetCorr

FlowNet

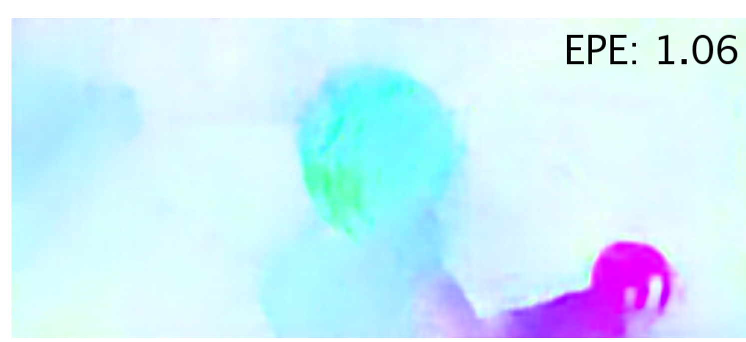
Simple vs. Corr
Sintel

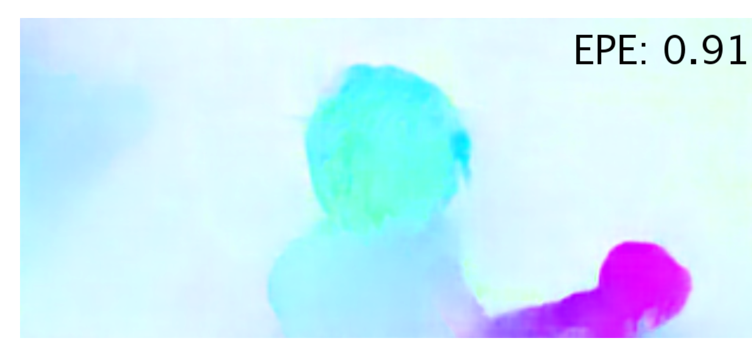
FlowNetCorr



FlowNetS

FlowNetCorr





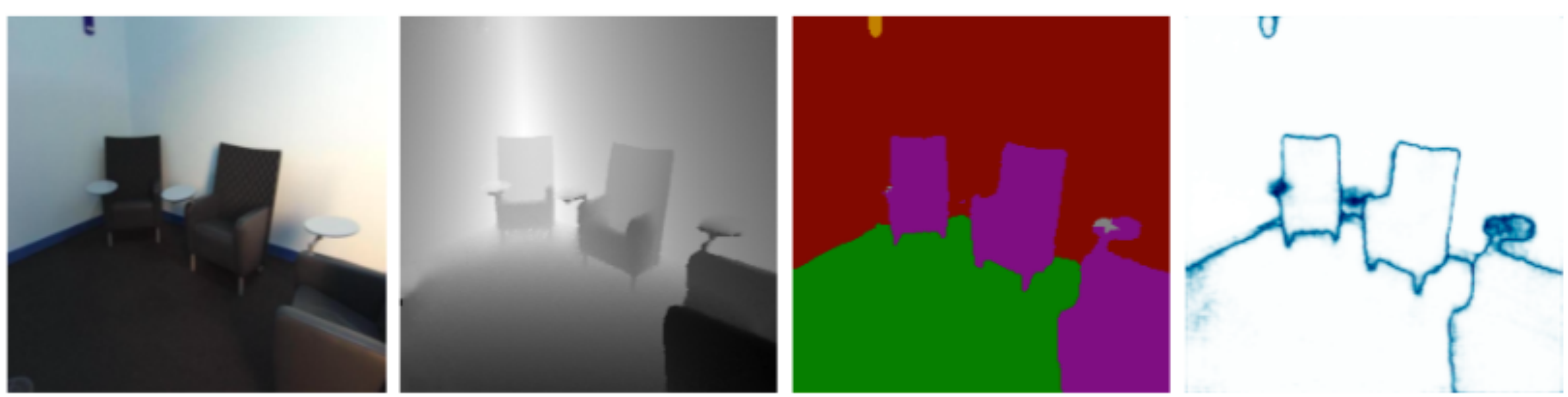
FlowNet

Learning Optical Flow with Convolutional Networks

Delving Deep into Computer Vision

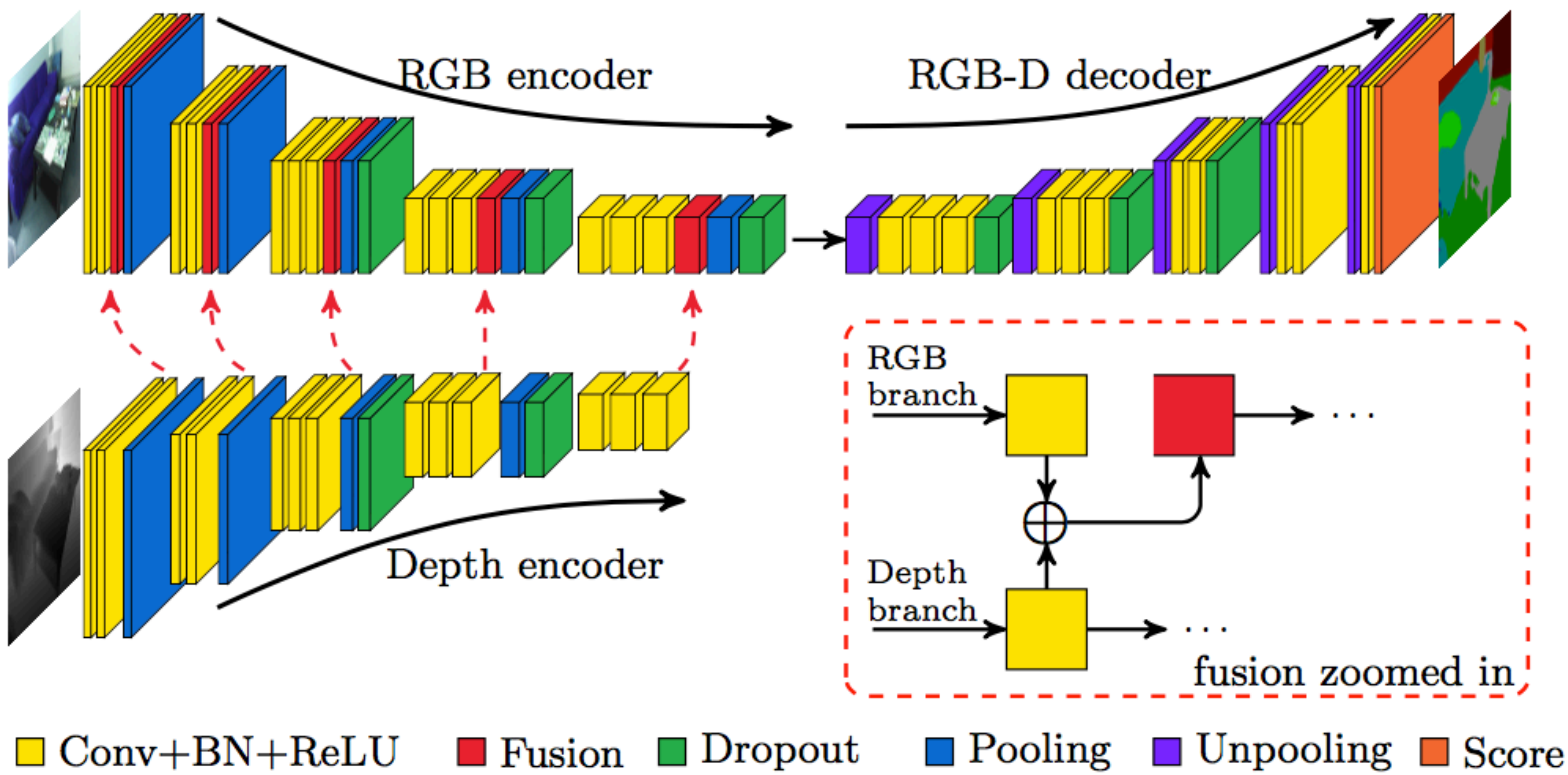
FlowNet

FuseNet



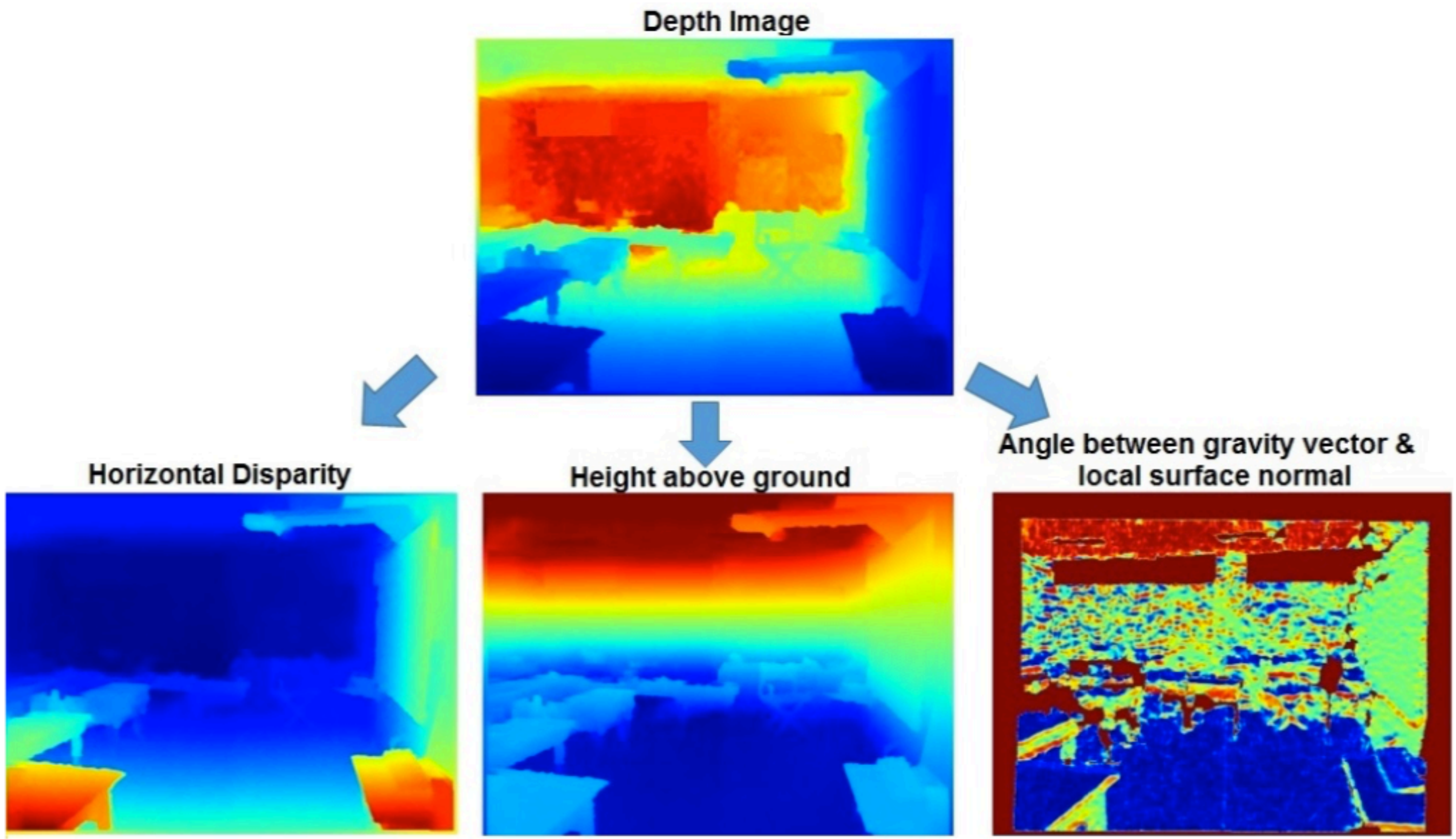
Incorporating Depth into Semantic Segmentation via Fusion-based CNN Architecture

ACCV'16



FuseNet

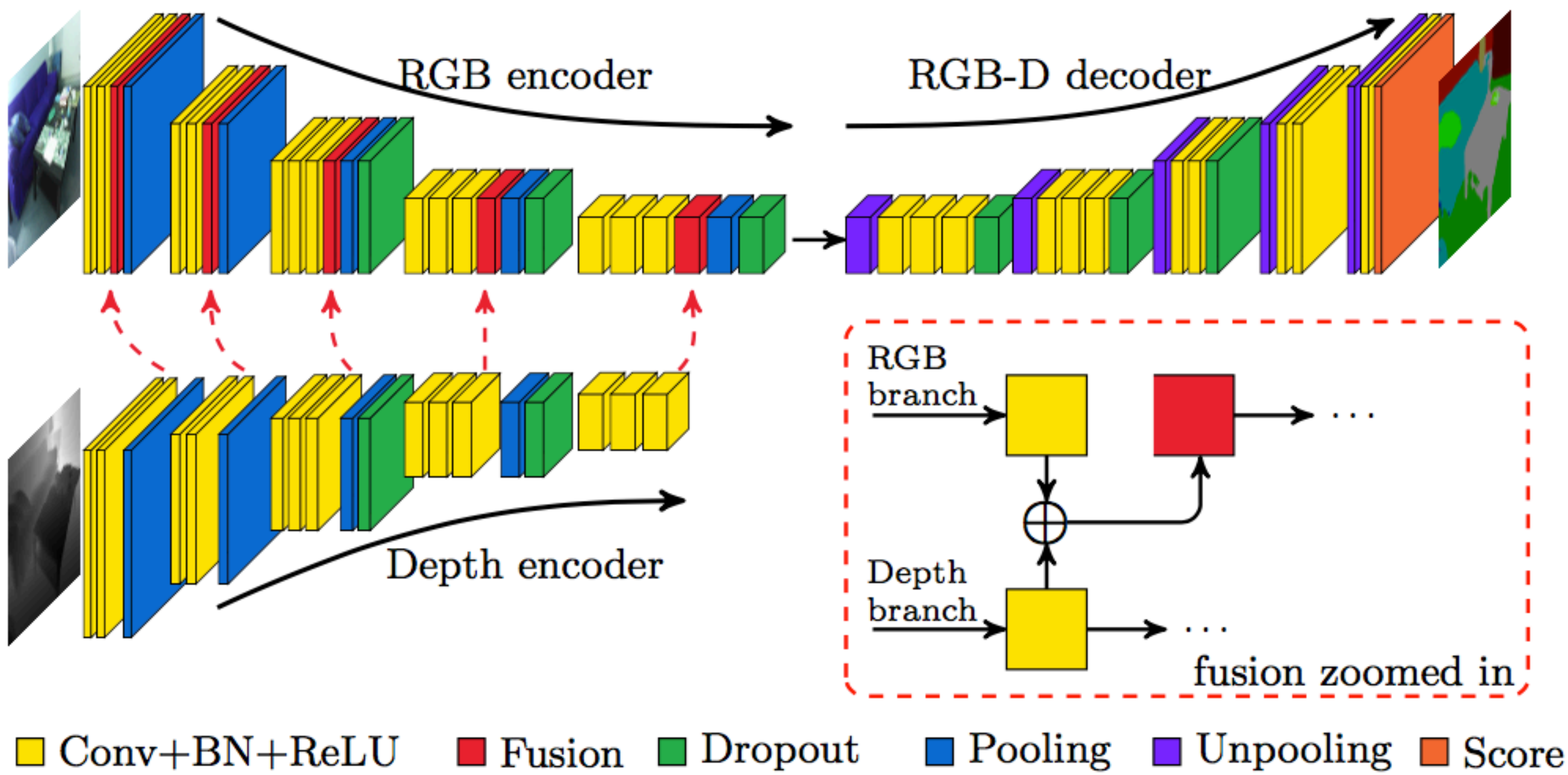
A conventional way: HHA



Multi-Scale Convolutional Architecture for Semantic Segmentation, Raj et al., Tech. Report, CMU-RI-TR-15-21,2015

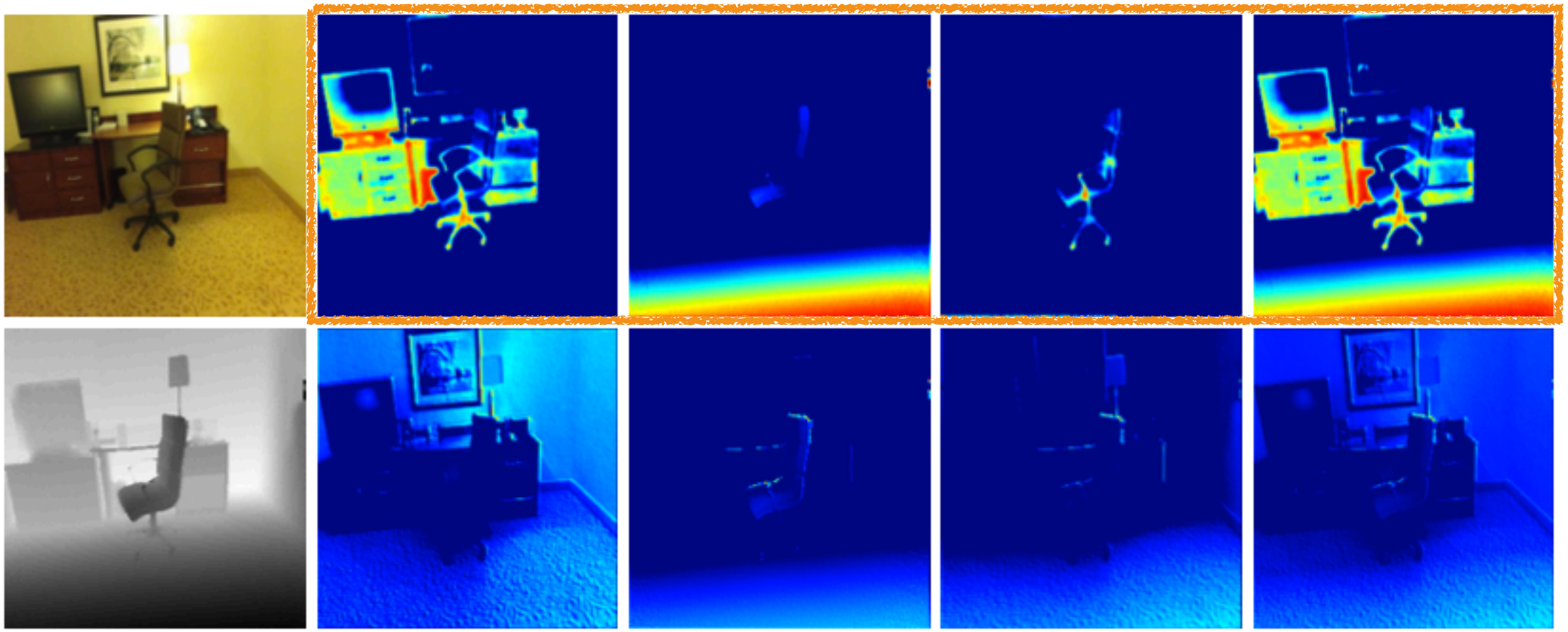
FuseNet

A deep way…



FuseNet

Why a second encoder for Depth input?



Input RGB-D | RGB branch | Depth branch | Sum before ReLU | Proposed

$$\max(0, f_c + f_d) \leq \max(0, f_c) + \max(0, f_d)$$

FuseNet

Are we any better than HHA?

- Proposed network improves all segmentation metrics

Input	Global	Mean	IoU
Depth	69.06	42.80	28.49
HHA	69.21	43.23	28.88
RGB	72.14	47.14	32.47
RGB-D	71.39	49.00	31.95
RGB-HHA	73.90	45.57	33.64
FusetNet-SF1	75.48	46.15	35.99
FusetNet-SF2	75.82	46.44	36.11
FusetNet-SF3	76.18	47.10	36.63
FusetNet-SF4	76.56	48.46	37.76
FusetNet-SF5	76.27	48.30	37.29

FuseNet

What about the others?

- Proposed network improves all segmentation metrics

	Global	Mean	IoU
FCN-32s [3]	68.35	41.13	29.00
FCN-16s [3]	67.51	38.65	27.15
Bayesian SegNet [14] (RGB)	71.2	45.9	30.7
LSTM [17]	-	48.1	-
Context-CRF [7] (RGB)	78.4	53.4	42.3
FuseNet-SF5	76.27	48.30	37.29
FuseNet-DF1	73.37	50.07	34.02

- Metrics

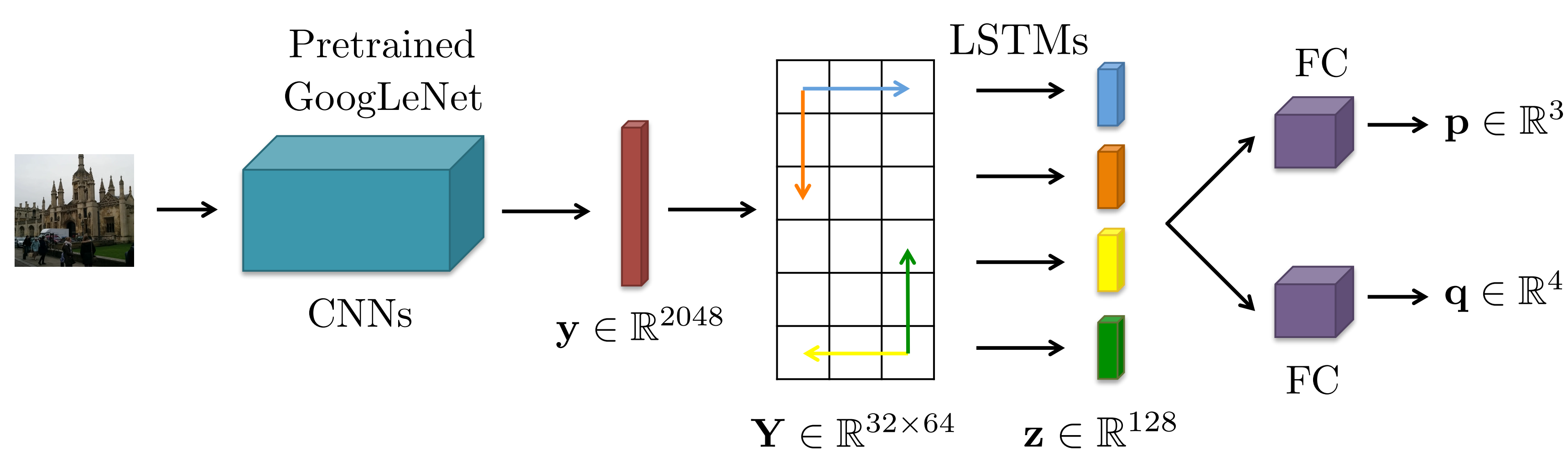
  Global: total number of correctly classified pixels
  Mean: average class accuracy
  IoU: average of intersection over union.

Delving Deep into Computer Vision

FlowNet

FuseNet

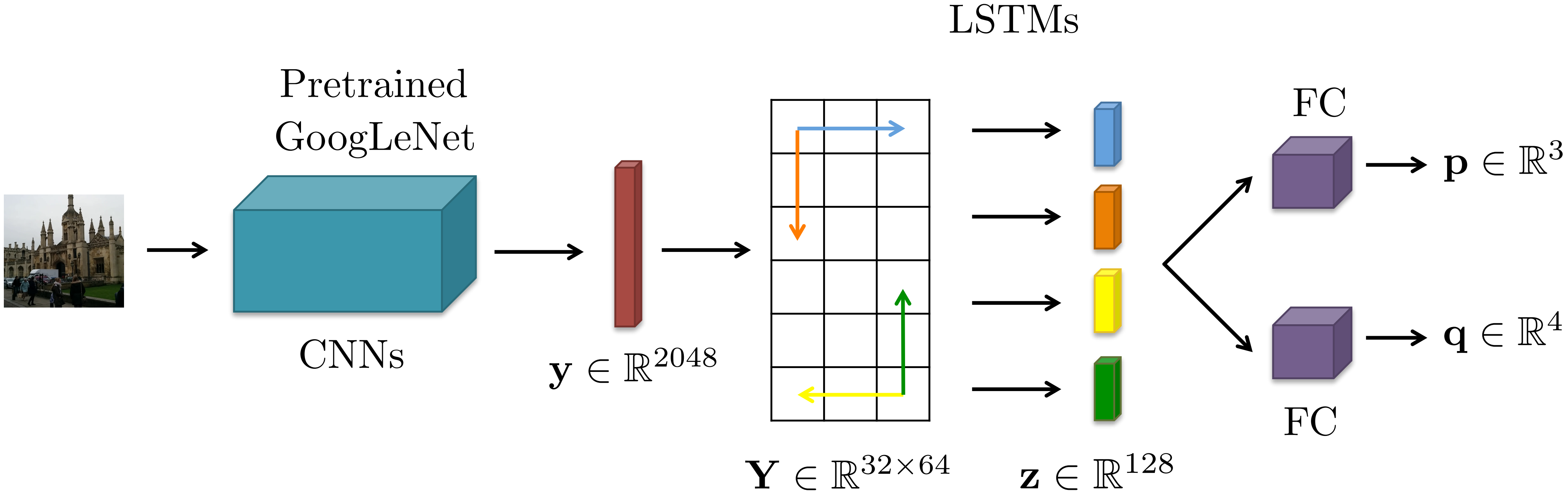
PoseLSTM



PoseLSTM

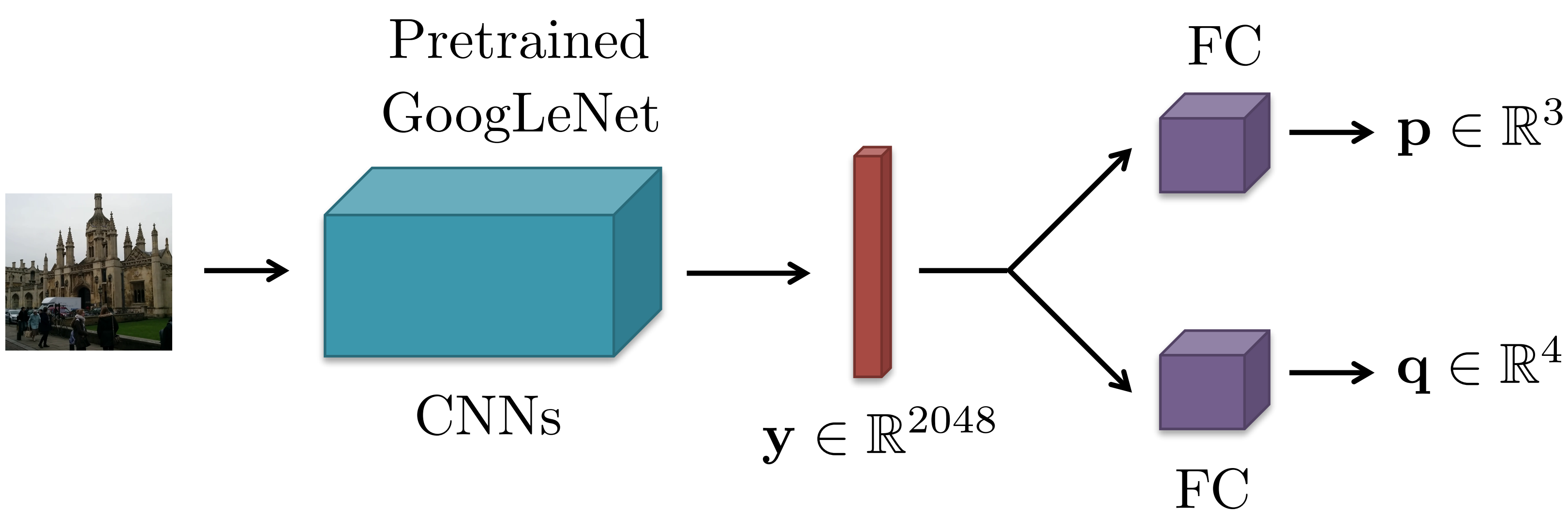
Image-based localization using LSTMs for structured feature correlation

ICCV'17

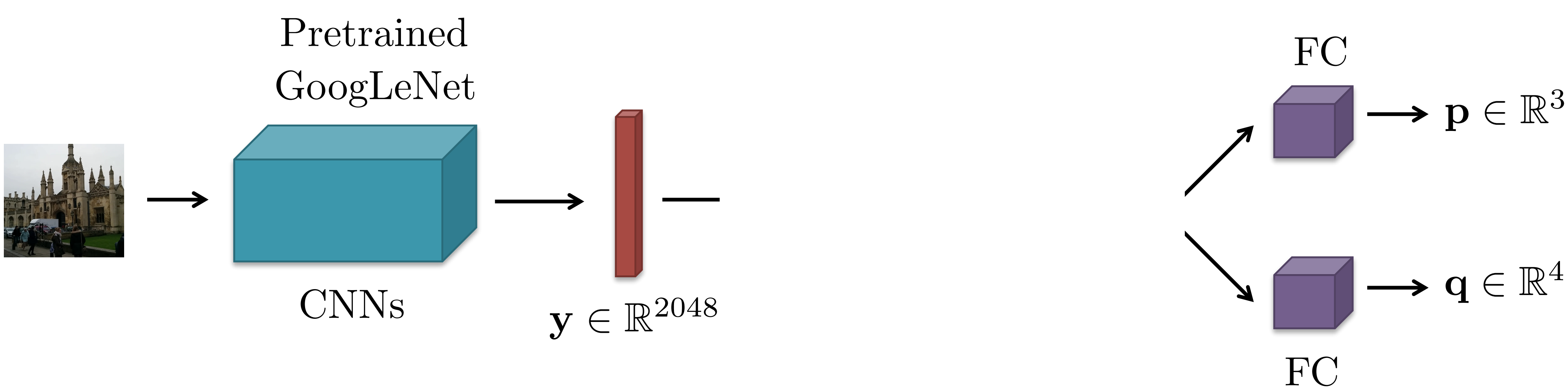


PoseLSTM

PoseNet



Structured Feature Correlation



PoseLSTM

Winner in Outdoor: SIFT

Scene	Area or Volume	Active Search (w/o) [5]	Active Search (w/)[5]	PoseNet[1]	Bayesian PoseNet[2]	Proposed + Improvement(pos,ori)	PoseNet with Geometric Loss[6]
King's College	5600 m^2	0.42 m, 0.55° (0)	0.57 m, 0.70° (0)	1.92 m, 5.40°	1.74 m, 4.06°	0.99 m, 3.65° (48,32)	0.88 m, 1.04°
Old Hospital	2000 m^2	0.44 m, 1.01° (2)	0.52 m, 1.12° (2)	2.31 m, 5.38°	2.57 m, 5.14°	1.51 m, 4.29° (35,20)	3.20 m, 3.29°
Shop Façade	875 m^2	0.12 m, 0.40° (0)	0.12 m, 0.41° (0)	1.46 m, 8.08°	1.25 m, 7.54°	1.18 m, 7.44° (19,8)	0.88 m, 3.78°
St Mary's Church	4800 m^2	0.19 m, 0.54° (0)	0.22 m, 0.62° (0)	2.65 m, 8.48°	2.11 m, 8.38°	1.52 m, 6.68° (43,21)	1.57 m, 3.32°
Average All	–	–	–	2.08 m, 6.83°	1.92 m, 6.28°	1.30 m, 5.52° (37,19)	1.63 m, 2.86°
Average by [5]	–	0.29 m, 0.63°	0.36 m, 0.71°	–	–	1.37 m, 5.52°	–
Chess	6 m^3	0.04 m, 1.96° (0)	0.04 m, 2.02° (0)	0.32 m, 8.12°	0.37 m, 7.24°	0.24 m, 5.77° (25,29)	0.13 m, 4.48°
Fire	2.5 m^3	0.03 m, 1.53° (1)	0.03 m, 1.50° (1)	0.47 m, 14.4°	0.43 m, 13.7°	0.34 m, 11.9° (28,17)	0.27 m, 11.3°
Heads	1 m^3	0.02 m, 1.45° (1)	0.02 m, 1.50° (1)	0.29 m, 12.0°	0.31 m, 12.0°	0.21 m, 13.7° (27,-14)	0.17 m, 13.0°
Office	7.5 m^3	0.09 m, 3.61° (34)	0.10 m, 3.80° (34)	0.48 m, 7.68°	0.48 m, 8.04°	0.30 m, 8.08° (37,-5)	0.19 m, 5.55°
Pumpkin	5 m^3	0.08 m, 3.10° (71)	0.09 m, 3.21° (68)	0.47 m, 8.42°	0.61 m, 7.08°	0.33 m, 7.00° (30,17)	0.26 m, 4.75°
Red Kitchen	18 m^3	0.07 m, 3.37° (0)	0.07 m, 3.52° (0)	0.59 m, 8.64°	0.58 m, 7.54°	0.37 m, 8.83° (37,-2)	0.23 m, 5.35°
Stairs	7.5 m^3	0.03 m, 2.22° (3)	0.04 m, 2.22° (0)	0.47 m, 13.8°	0.48 m, 13.1°	0.40 m, 13.7° (15,0.7)	0.35 m, 12.4°
Average All	–	–	–	0.44 m, 10.4°	0.47 m, 9.81°	0.31 m, 9.85° (29,5)	0.23 m, 8.12°
Average by [5]	–	0.05 m, 2.46°	0.06 m, 2.54°	–	–	0.30 m, 9.15°	–

SIFT-based

CNN-based

Where SIFT dies…

Area	# train/test	PoseNet [26]	Proposed
$5575\,\mathrm{m}^2$	875/220	$1.87\,\mathrm{m}, 6.14°$	$1.31\,\mathrm{m}, 2.79°$ (30,55)

The map cannot be reconstructed due to a lack of sufficient matches: repeated structures, textureless areas

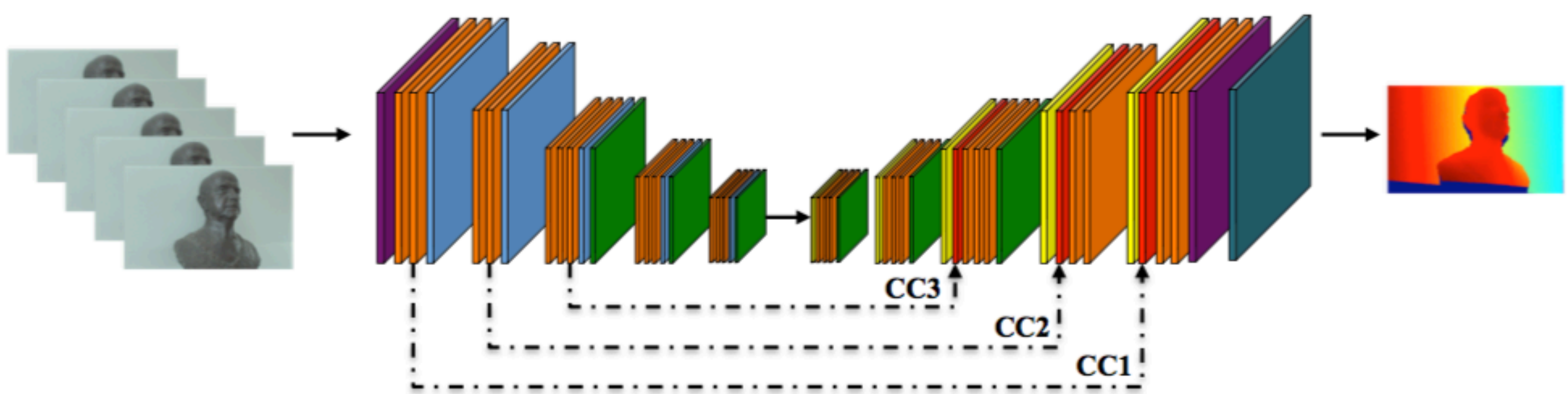
Delving Deep into Computer Vision

FlowNet

FuseNet

PoseLSTM

DDFF

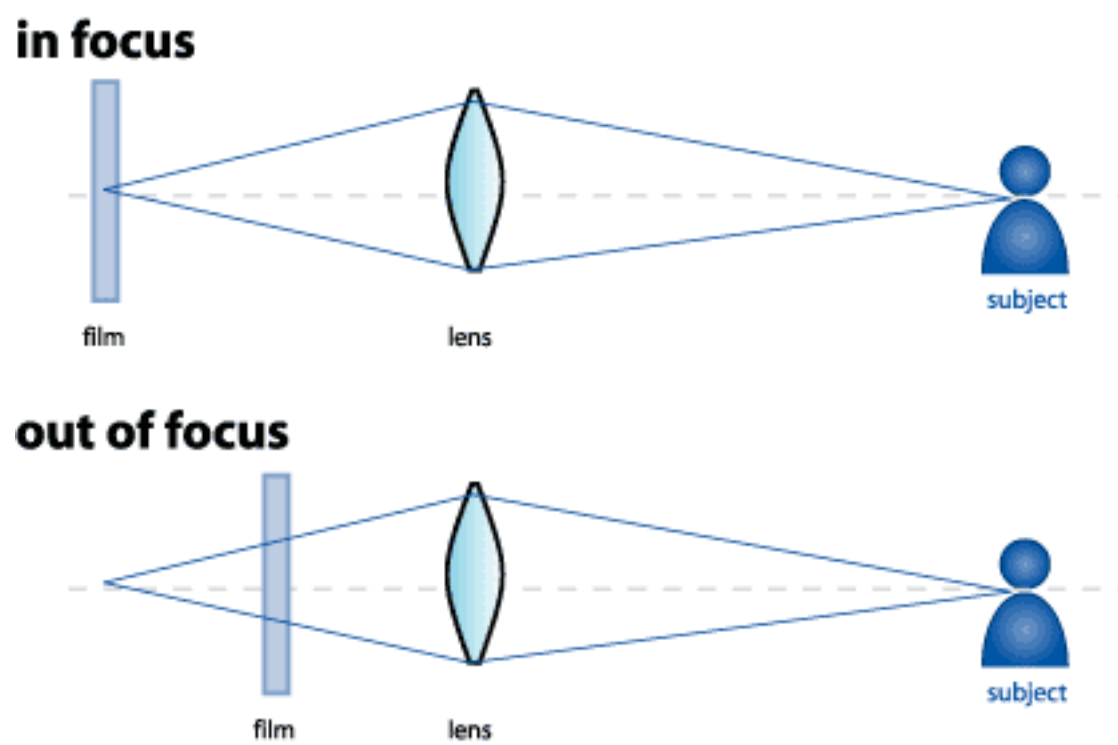


Caner Hazirbas | hazirbas@cs.tum.edu

DDFF

Deep Depth From Focus

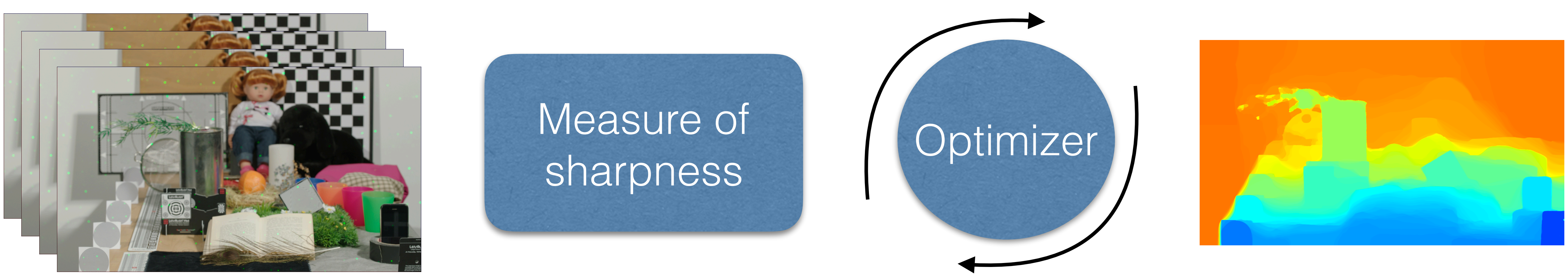
- Image of a point intersects the camera sensor when the point is in focus
- Therefore, sharpness determines the focused regions on the images



https://inst.eecs.berkeley.edu/~cs39j/sp02/session12.html

Conventional DFF methods

- Image of a point intersects the camera sensor when the point is in focus
- Therefore, sharpness determines the focused regions on the images
- Distance of a point from the camera can be formulated wrt. focus

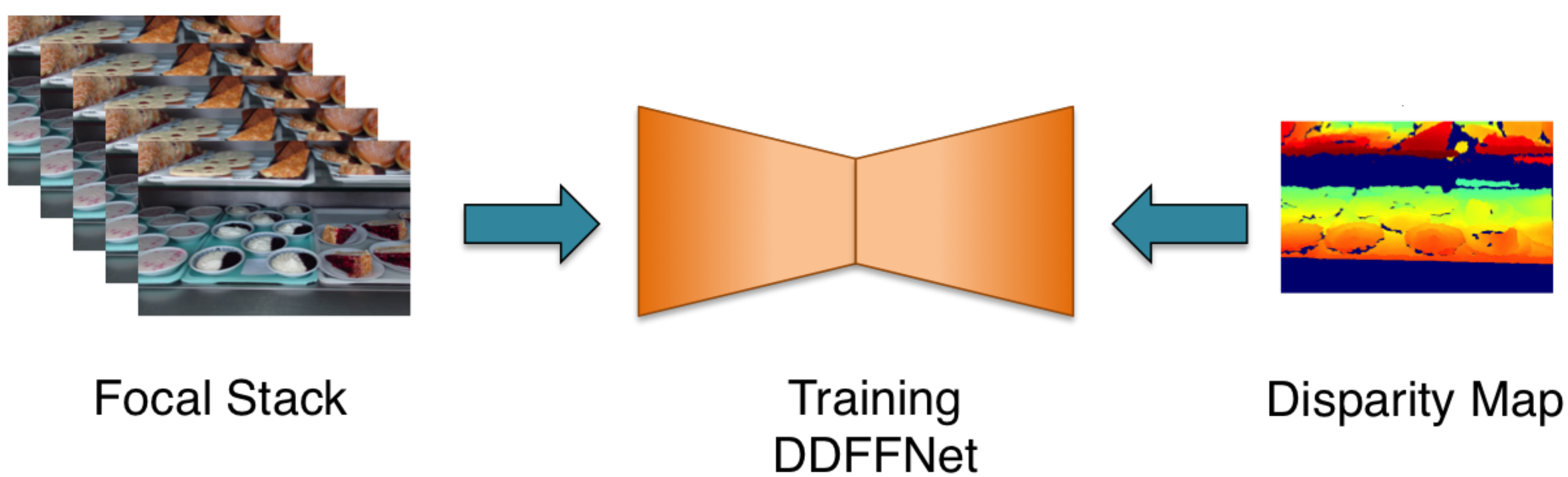


[Pertuz et al.] [Moeller et al.]

Deep Depth From Focus

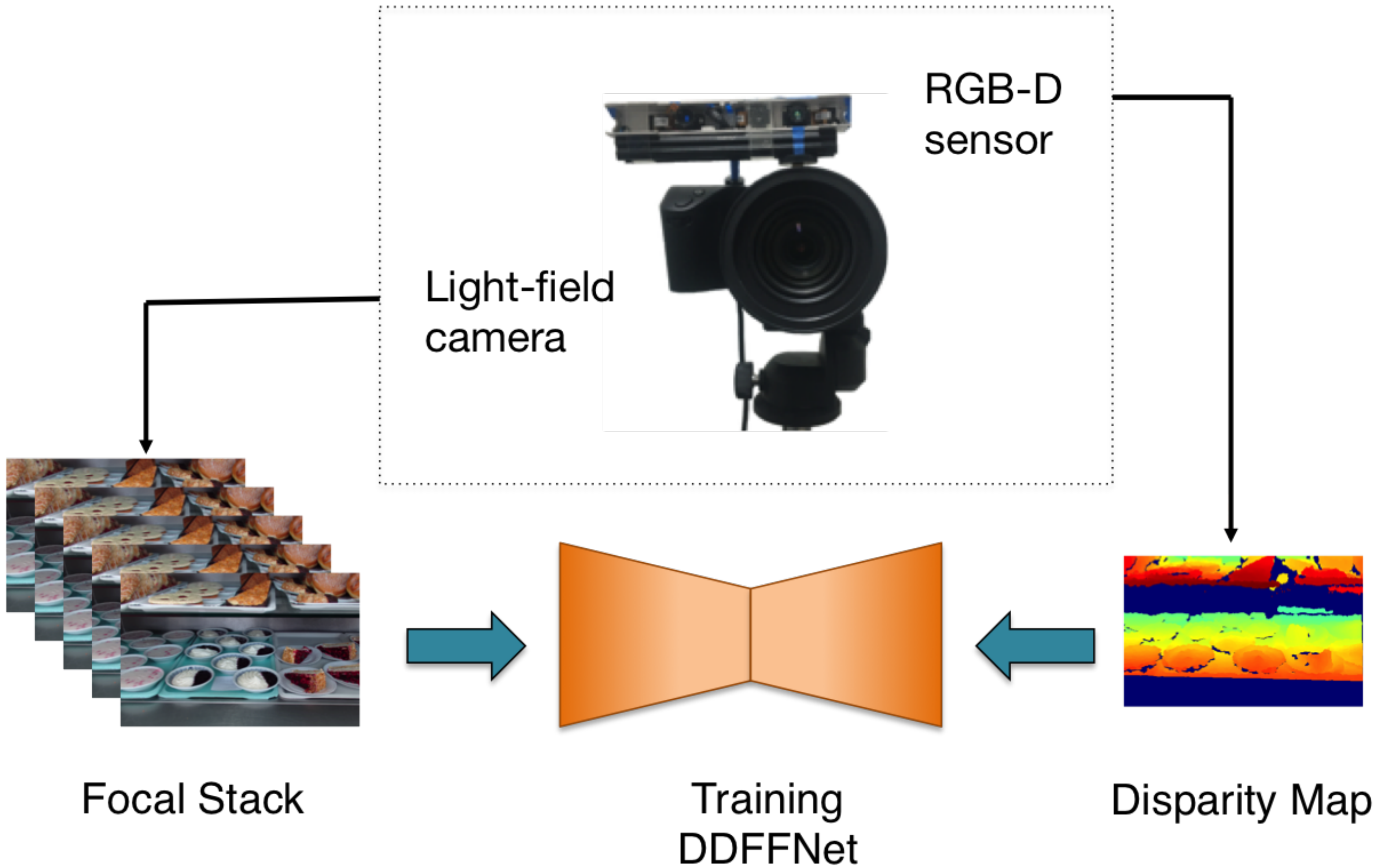
DDFF

- Focus gradually changes on each image in the stack
- End-to-end trained convolutional auto-encoder
- Depth (disparity) from focal stack

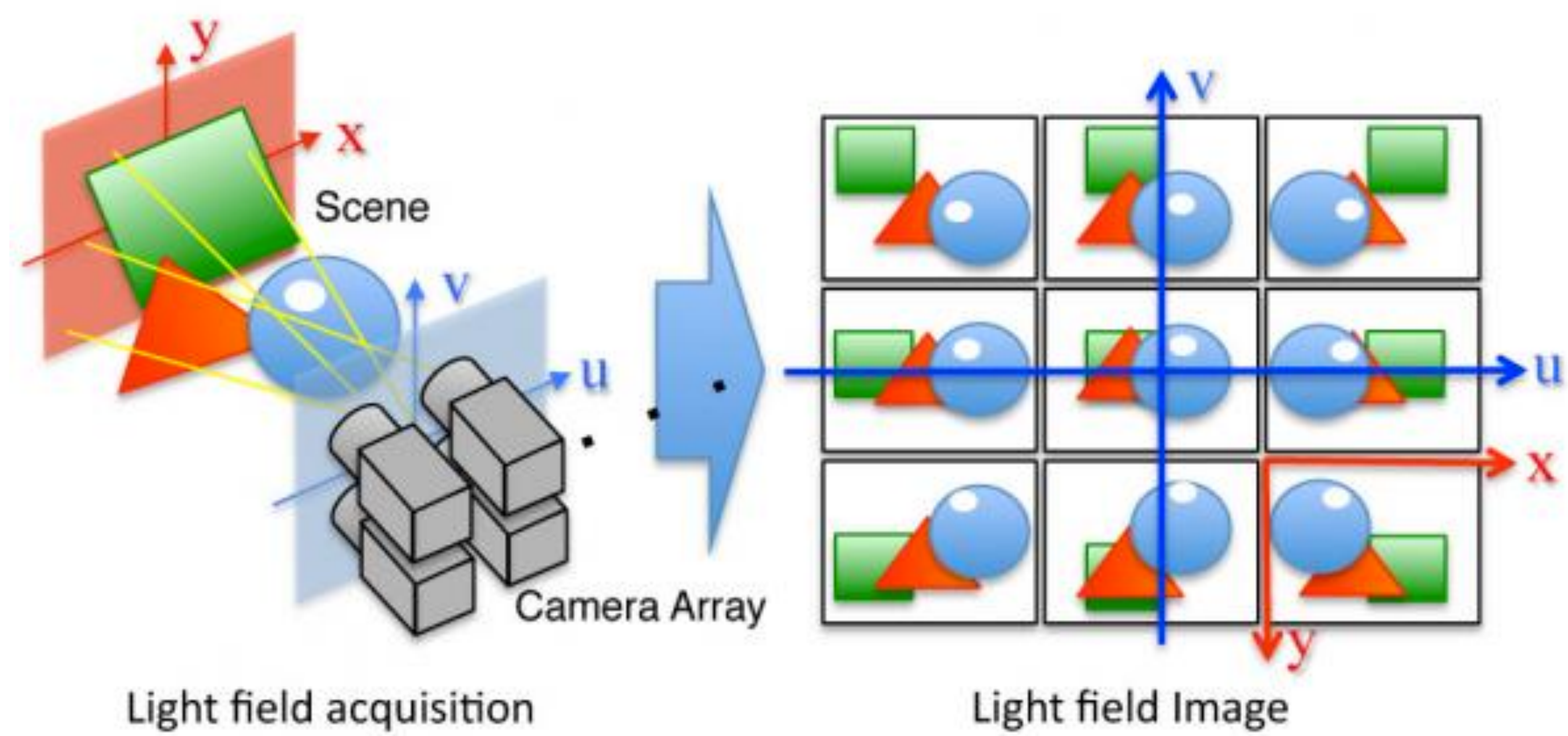


DDFF

How to get data?



Light-field Imaging

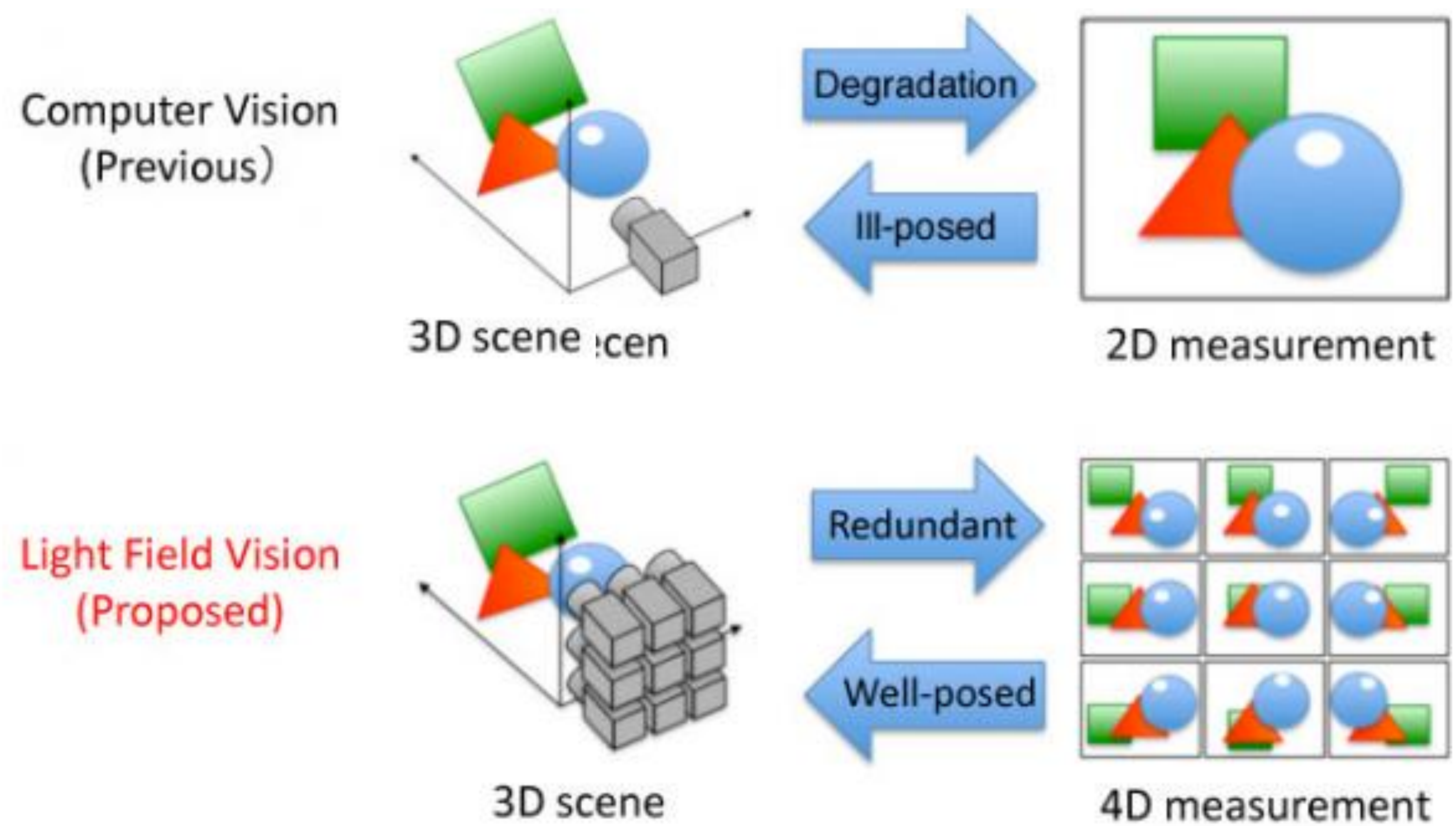


$$I(x,y) = \int_u \int_v L(u,v,x,y)\ \partial u\, \partial v$$

http://limu.ait.kyushu-u.ac.jp/e/project/project003.html

DDFF

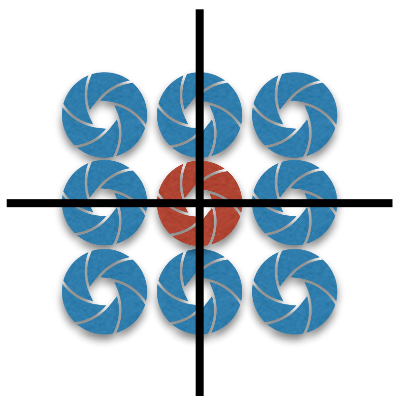
Light-field Imaging



http://limu.ait.kyushu-u.ac.jp/e/project/project003.html

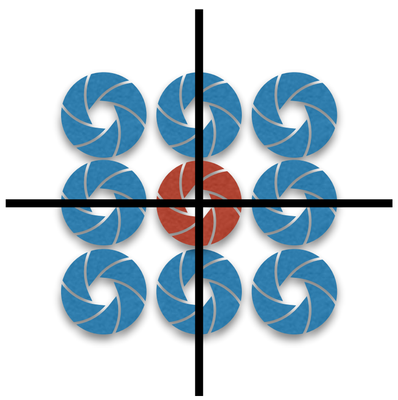
DDFF

Digital Refocusing



$$I'(x,y) = \int_u \int_v L(u, v, x + \Delta_x(u), y + \Delta_y(v))\, \partial u\, \partial v$$

Digital Refocusing



$$I'(x,y) = \int_u \int_v L(u, v, x + \Delta_x(u), y + \Delta_y(v))\, \partial u\, \partial v$$

DDFF

Digital Refocusing

$$\begin{pmatrix}\Delta_x(u)\\ \Delta_y(v)\end{pmatrix} = \underbrace{\frac{\text{baseline}\cdot f}{Z}}_{\text{disparity}} \cdot \begin{pmatrix}u_{center} - u\\ v_{center} - v\end{pmatrix}$$

- Z: any arbitrary depth
- baseline: distance between adjacent sub-apertures
- f: focal length of the micro-lenses
- $(u\ v)^T$: spatial location of the sub aperture in the camera plane

$$I'(x,y) = \int_u \int_v L(u, v, x + \Delta_x(u), y + \Delta_y(v))\, \partial u\, \partial v$$

DDFF

DDFF 12-Scene dataset

- 720 recorded light-field depth pairs
- collected in 12 different scenes
- each of 6-scene has 100, each of 6-scene 20

DDFF

First Challenge

- Micro disparity (270 micrometer = 27e-5 m) between sub-apertures results in sub-pixel shift
- Therefore, focus is not observable by human eyes
- Shift the sub-apertures using phase-shift algorithm

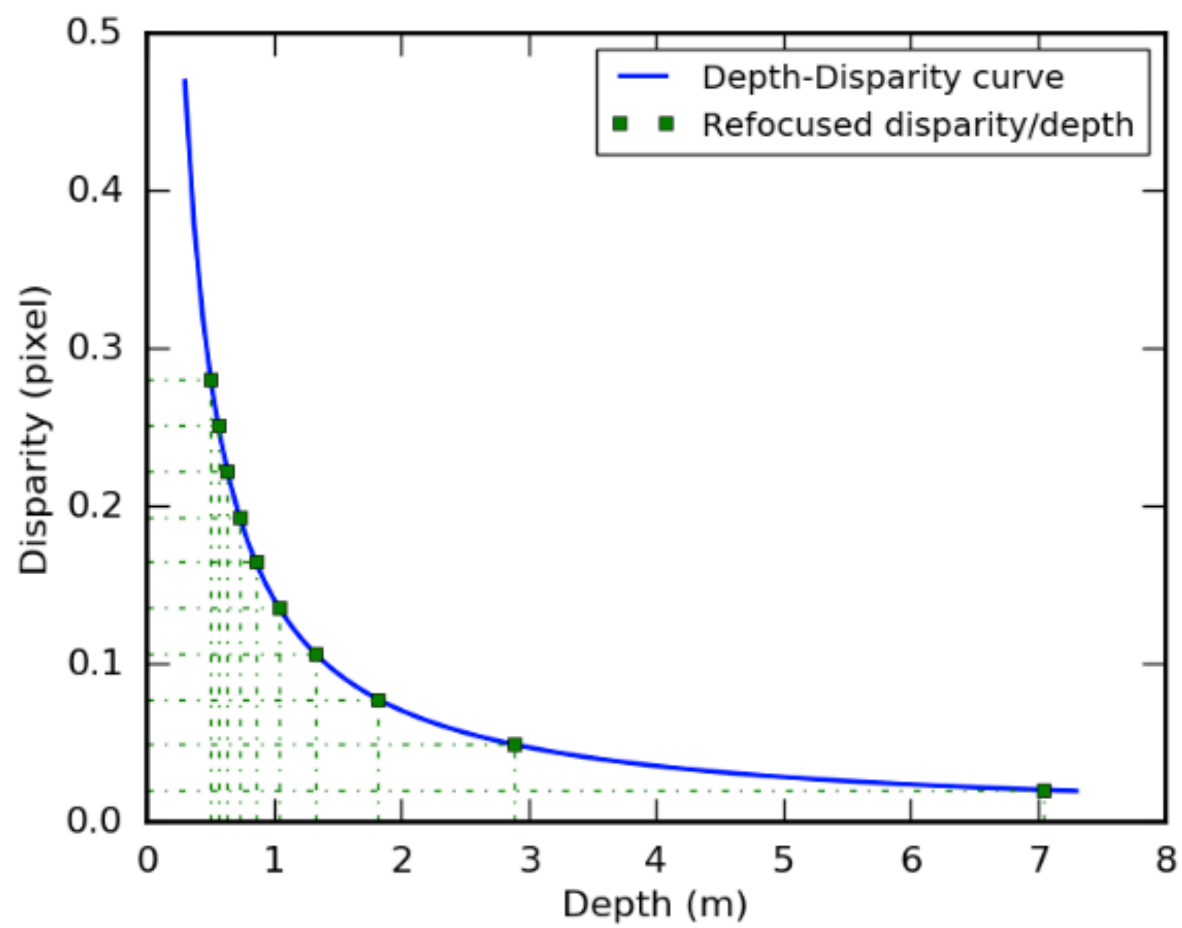
$$\mathcal{F}\{I'(x + \Delta_x(u))\} = \mathcal{F}\{I(x)\} \cdot \exp^{2\pi i \Delta_x(u)}$$

[Jeon et al.]

DDFF

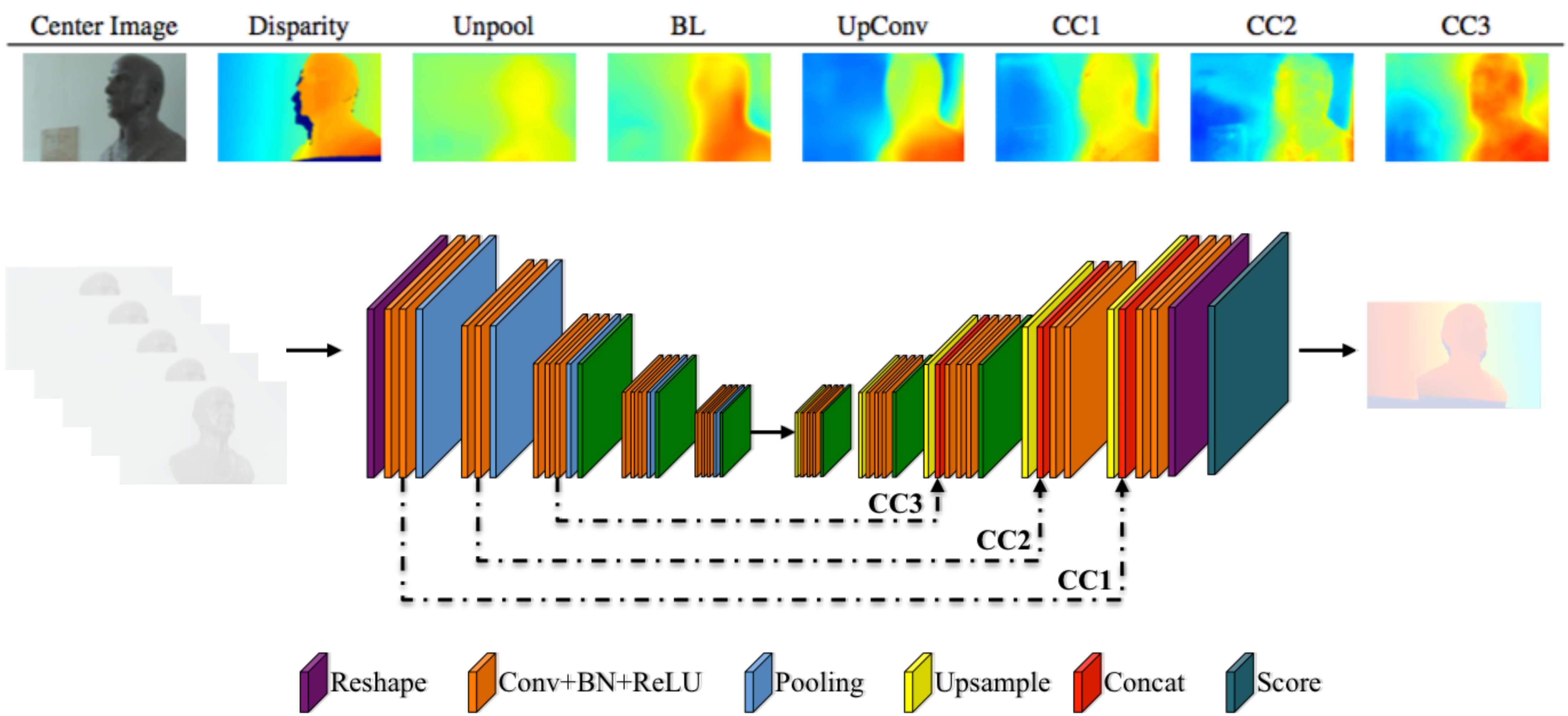
Focal Stack

- 10 refocused images in between 50cm to 7m
- Linear change of focus (disparity) in the stack



Second Challenge

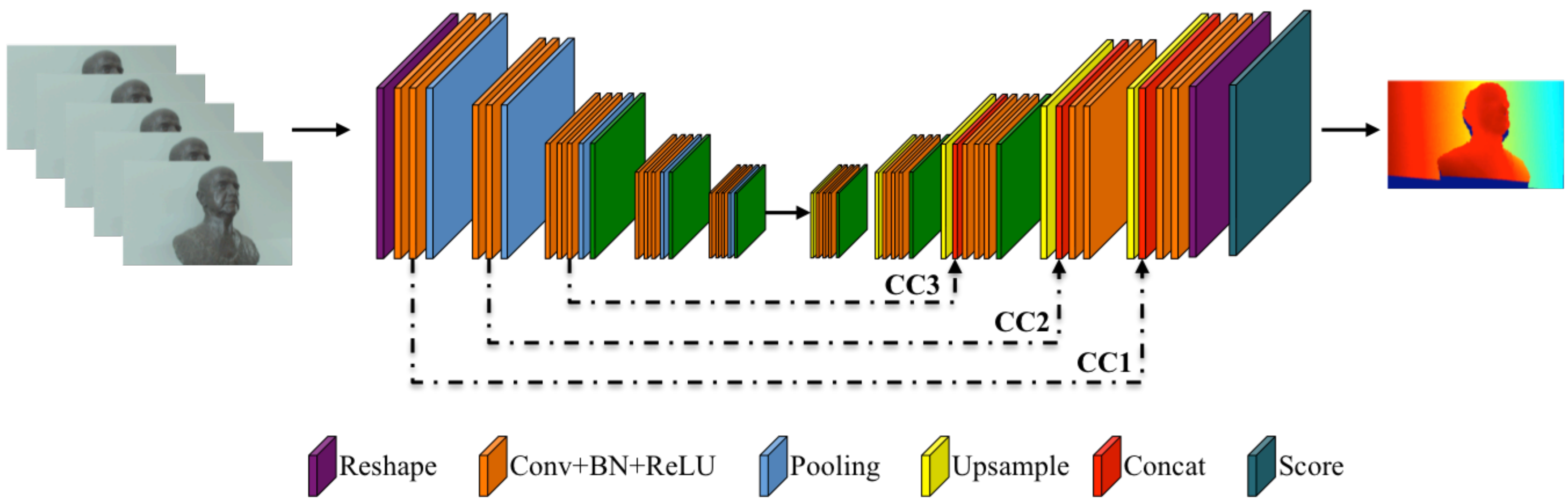
- What network to choose?



DDFF

Second Challenge

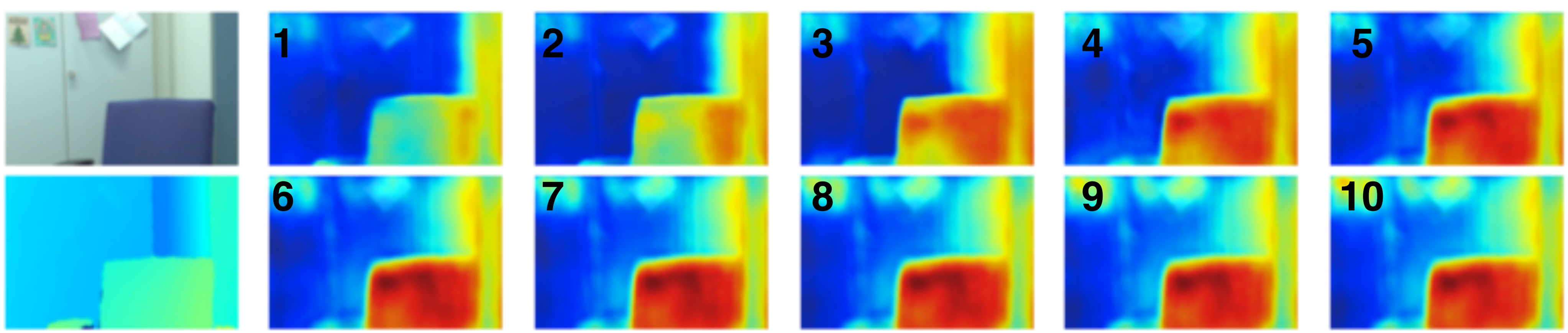
- What network to choose?
- How to process the stack through the network?



DDFF

Second Challenge

- What network to choose?
- How to process the stack through the network?
- What to expect from the network to learn?

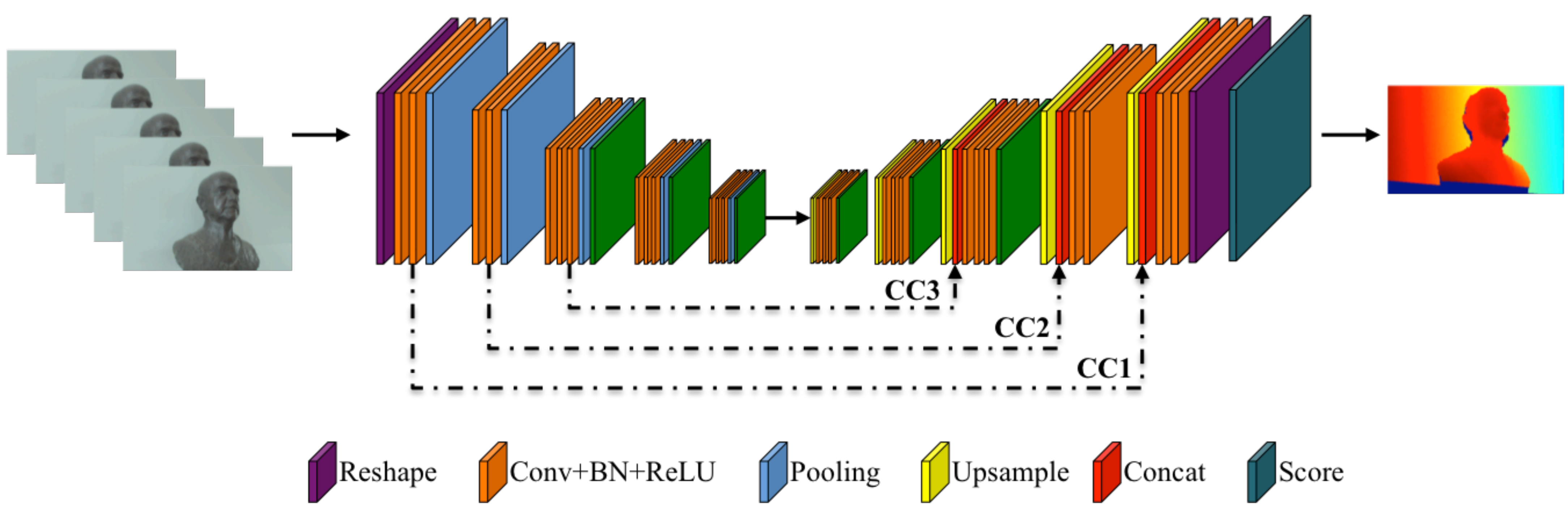


DDFF

Training

- Loss: missing depth/disparity values are ignored

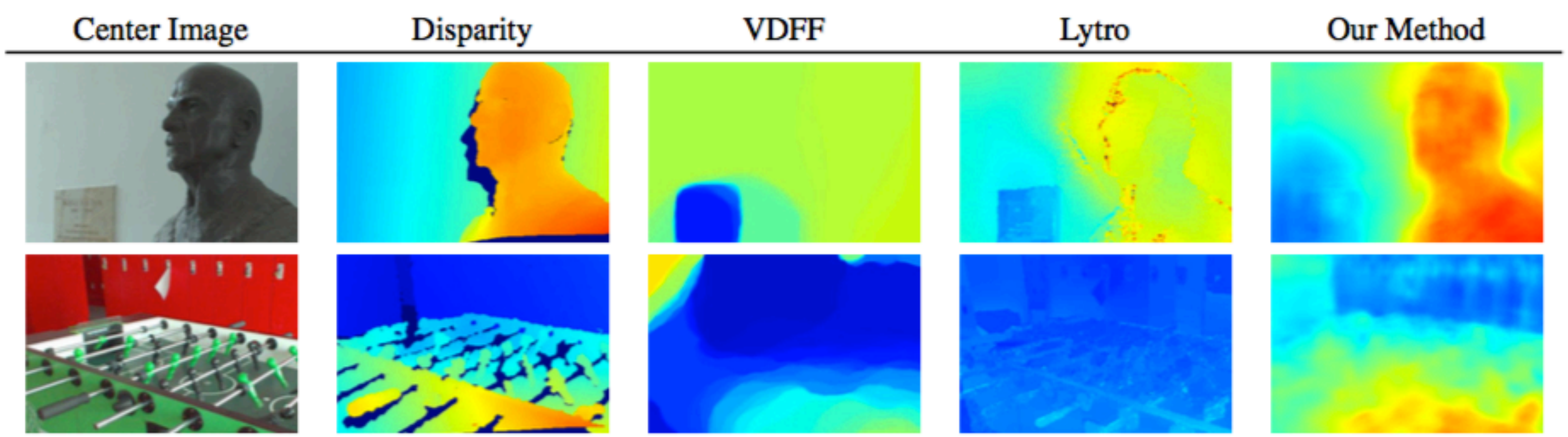
$$\mathcal{L} = \sum_{p}^{HW} \mathcal{M}(p) \cdot \left\| f_{\mathbf{W}}(\mathcal{S}, p) - D(p) \right\|_2^2 + \lambda \|\mathbf{W}\|_2^2$$



DDFF

Evaluation

- DDFFNet reduces the depth error by 75% respect to VDFF



- Best scaling factor for VDFF and Lytro:

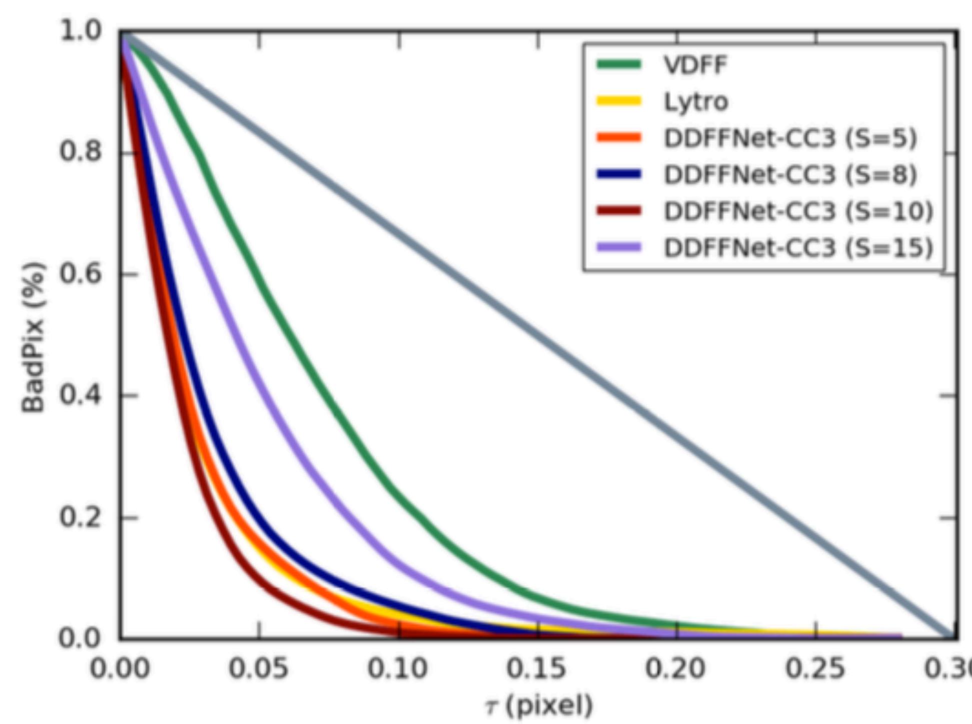
$$k^* = \arg\min_k \sum_p \|k \cdot \tilde{Z}_p - Z_p\|_2^2$$

DDFF

Evaluation

- DDFFNet-CC3 (S=10) has the least badpix and depth error

	Method	Runtime (s.)	Depth (m.)
DDFFNet	Unpool	0.55	1.40
DDFFNet	BL	0.43	1.10
DDFFNet	UpConv	0.50	0.58
DDFFNet	CC1	0.60	0.79
DDFFNet	CC2	0.60	0.86
DDFFNet	CC3	0.58	0.86
	DFLF	0.59	1.50
	VDFF	2.83	8.90
	Lytro	25.26 (CPU)	0.99



DDFF

What's Next?

- More analyses of DDFFNet
  - sharpness in DDFFNet
  - non-linearly refocused stack

- DDFF 12-Scene dataset
  - ***refocusing***,
  - DLF,
  - 3D reconstruction

Delving Deep into Computer Vision

FlowNet

FuseNet

PoseLSTM

DDFF

References

- FlowNet: Learning Optical Flow with Convolutional Networks
  A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, T. Brox, ICCV'15

- FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-based CNN Architecture
  C. Hazirbas, L. Ma, C. Domokos, D. Cremers, ACCV'16

- Image-based localization using LSTMs for structured feature correlation
  F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, D. Cremers, ICCV'17

- Deep Depth From Fous
  C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, D. Cremers, ArXiv'16