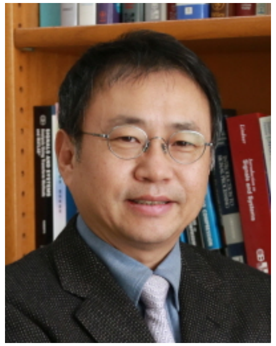
CVPR LONG BEACH CALIFORNIA June 16-20, 2019

LAF-Net: Locally Adaptive Fusion Networks for Stereo Confidence Estimation


Sunok Kim


Seungryong Kim


Dongbo Min


Kwanghoon Sohn

YONSEI UNIVERSITY

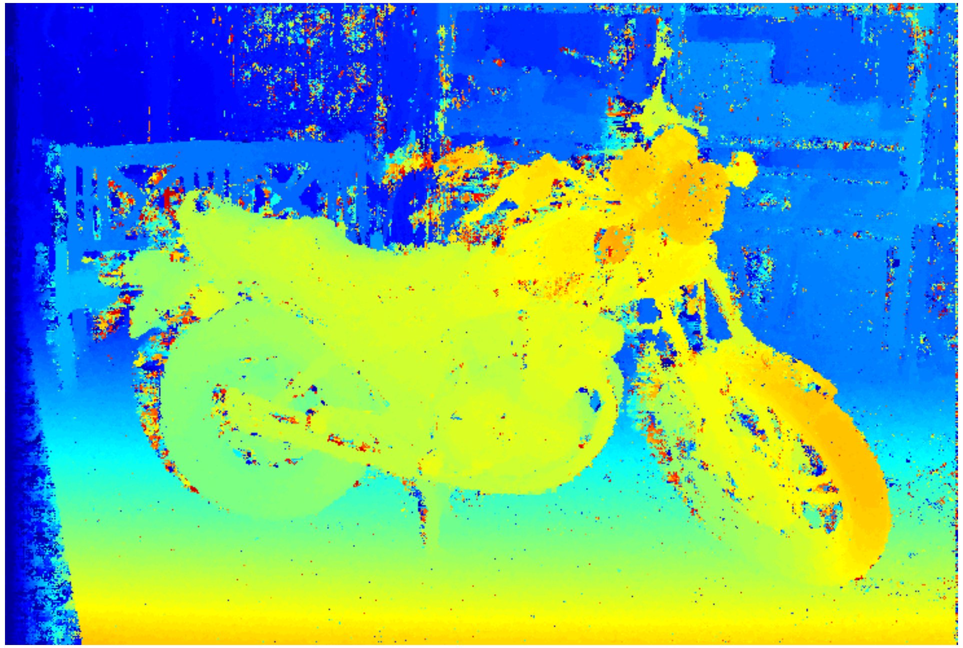
EPFL

이화여자대학교 EWHA WOMANS UNIVERSITY

What is Stereo Matching?

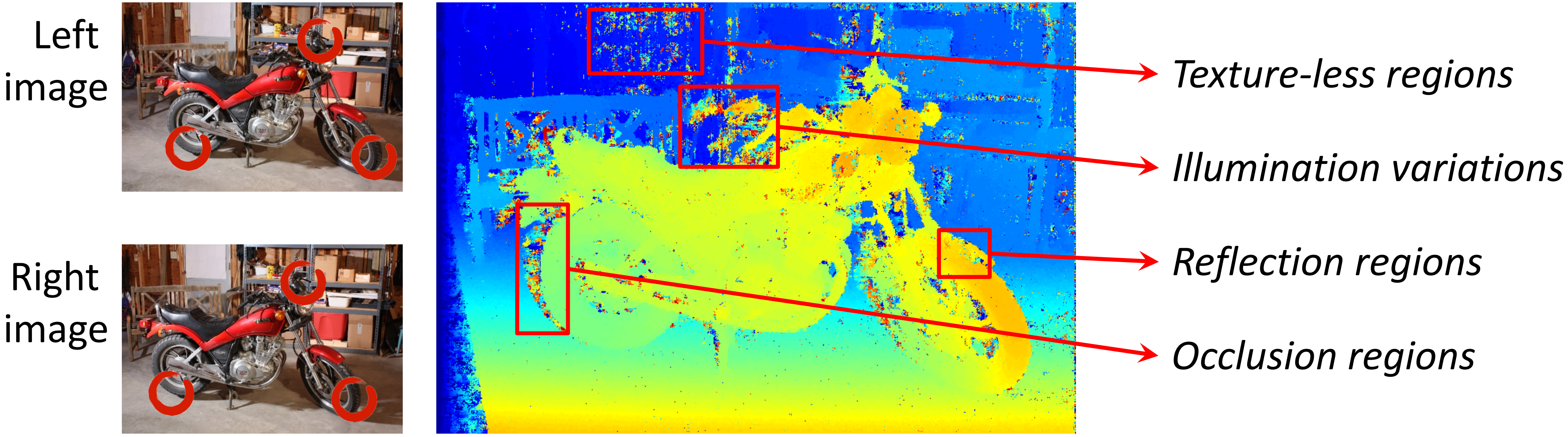
Left image

Right image



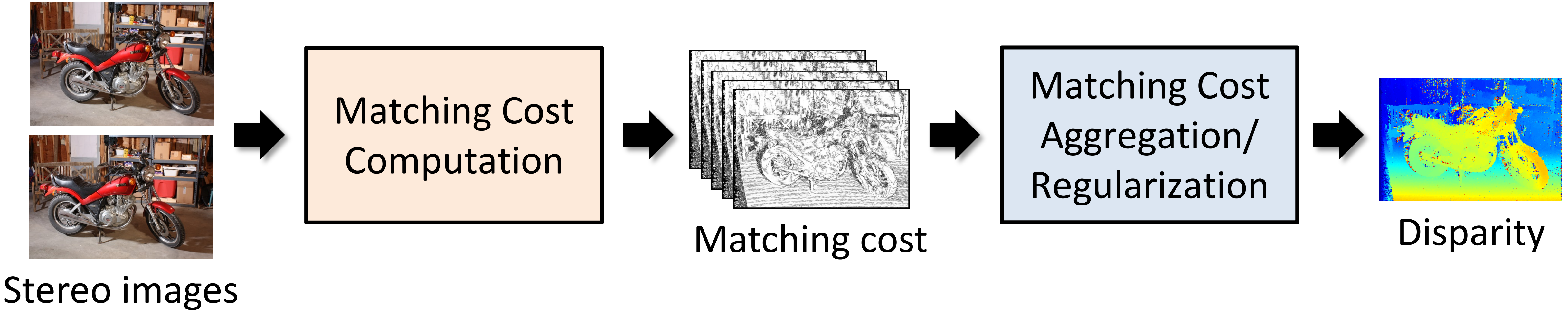
Disparity

What is Stereo Matching?

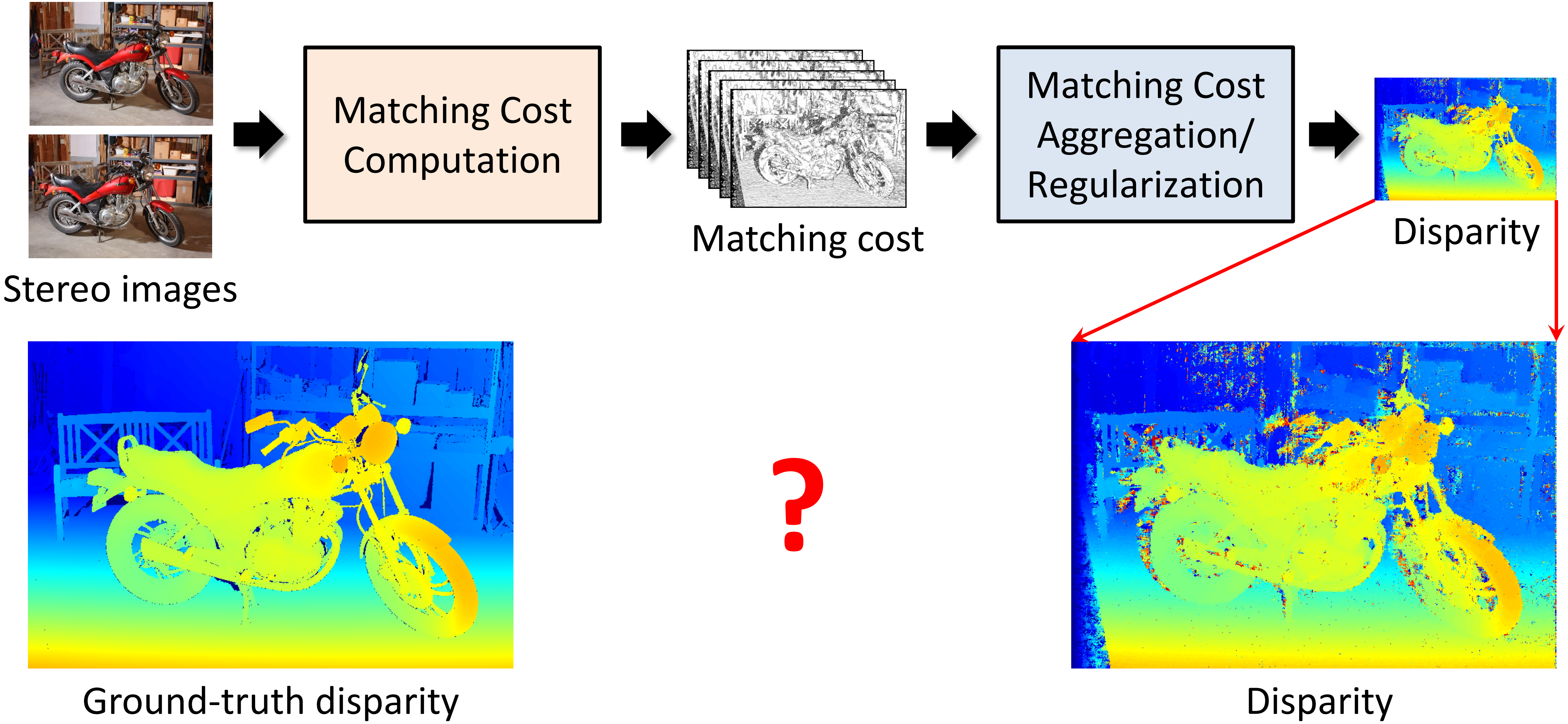


Disparity

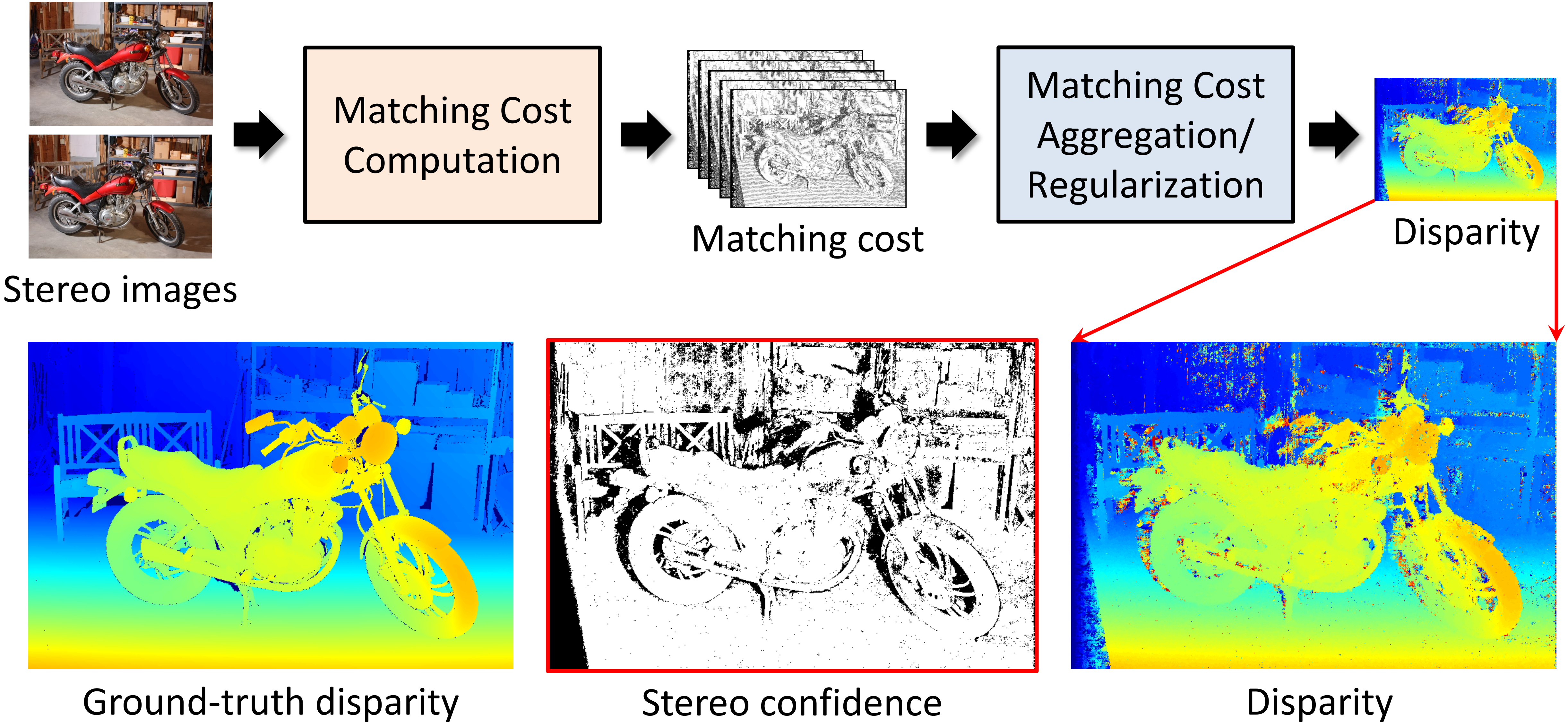
Stereo Matching Pipeline



Stereo Matching Pipeline



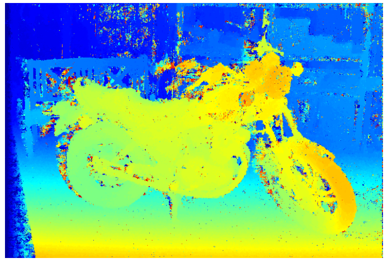
Stereo Matching Pipeline



Related Works

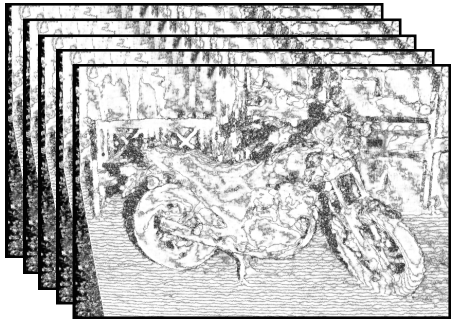
Initial Disparity Only

Poggi et al., BMVC’16, Seki et al., BMVC’16



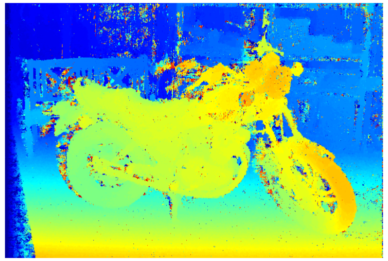
Matching Cost Only

Shaked et al., CVPR’17

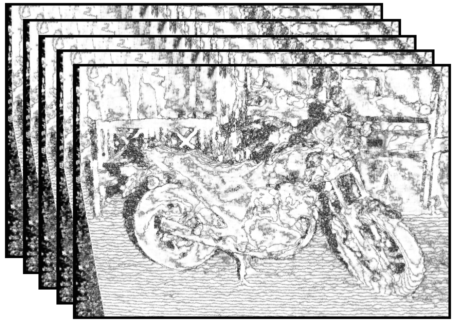


Disparity + Matching Cost

Kim et al., ICIP’17, Kim et al., TIP’19

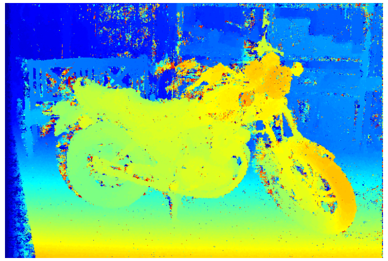


+



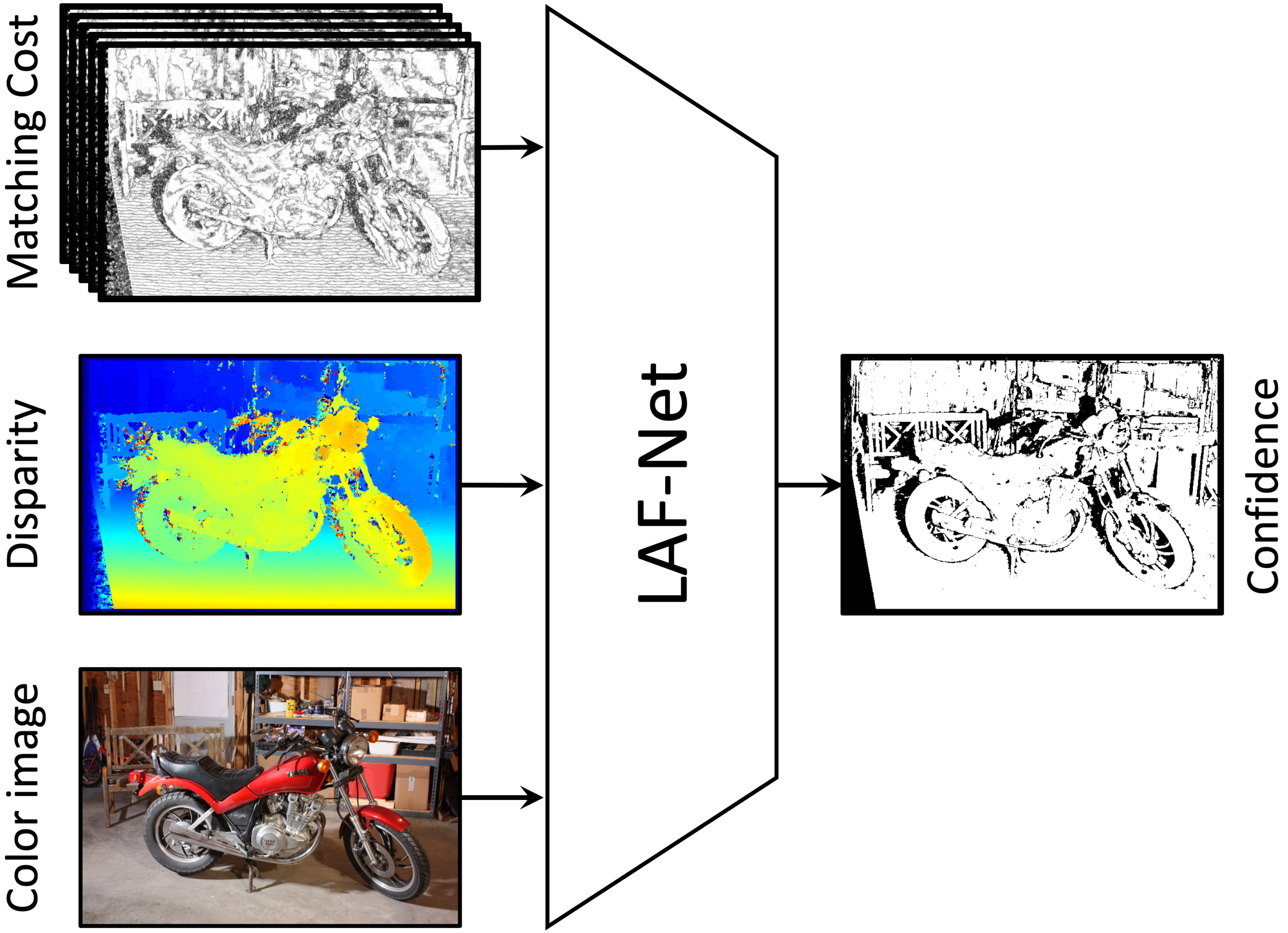
Disparity + Color

Fu et al., WACV’18, Poggi et al., ECCV’18

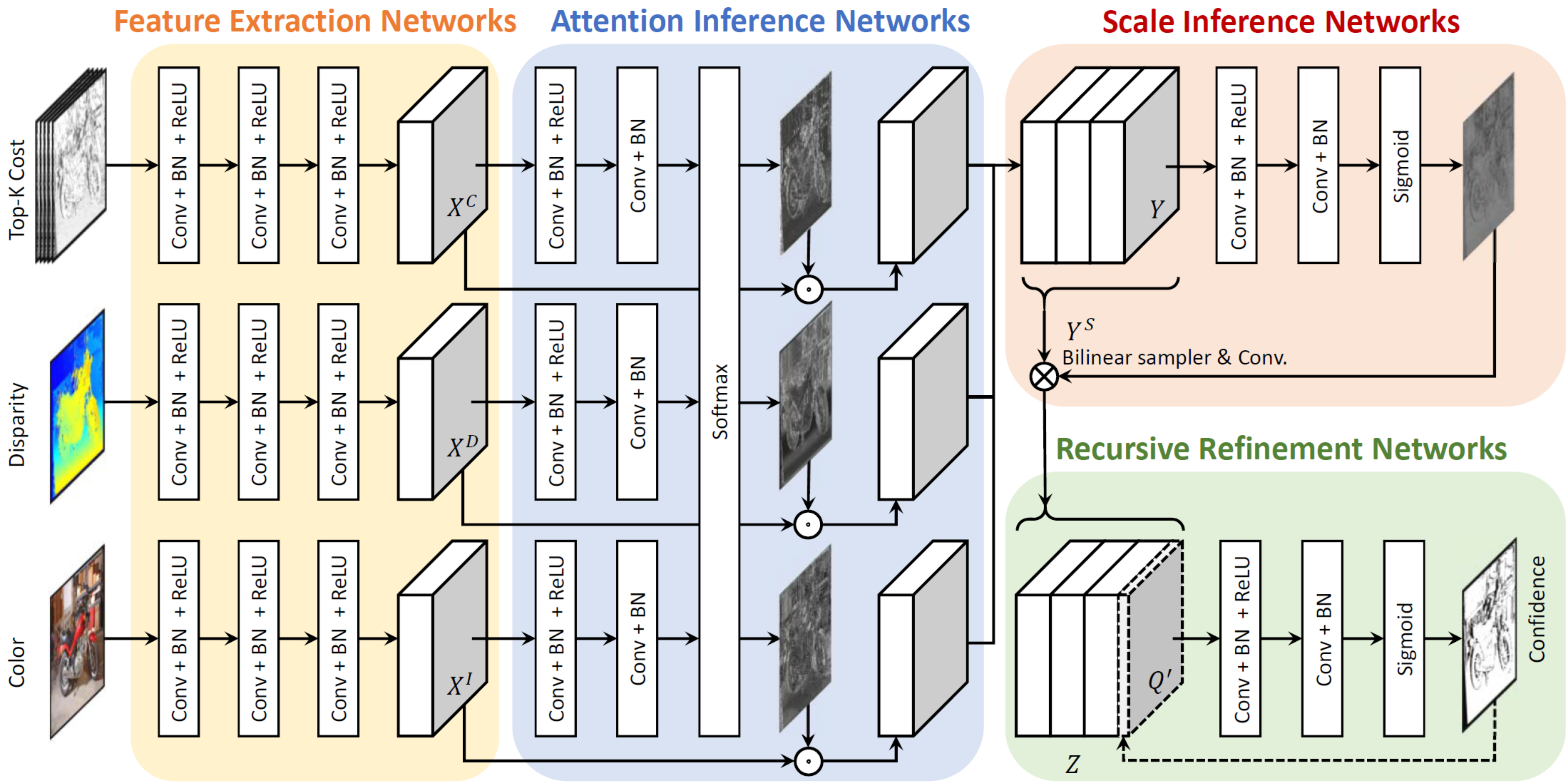


+

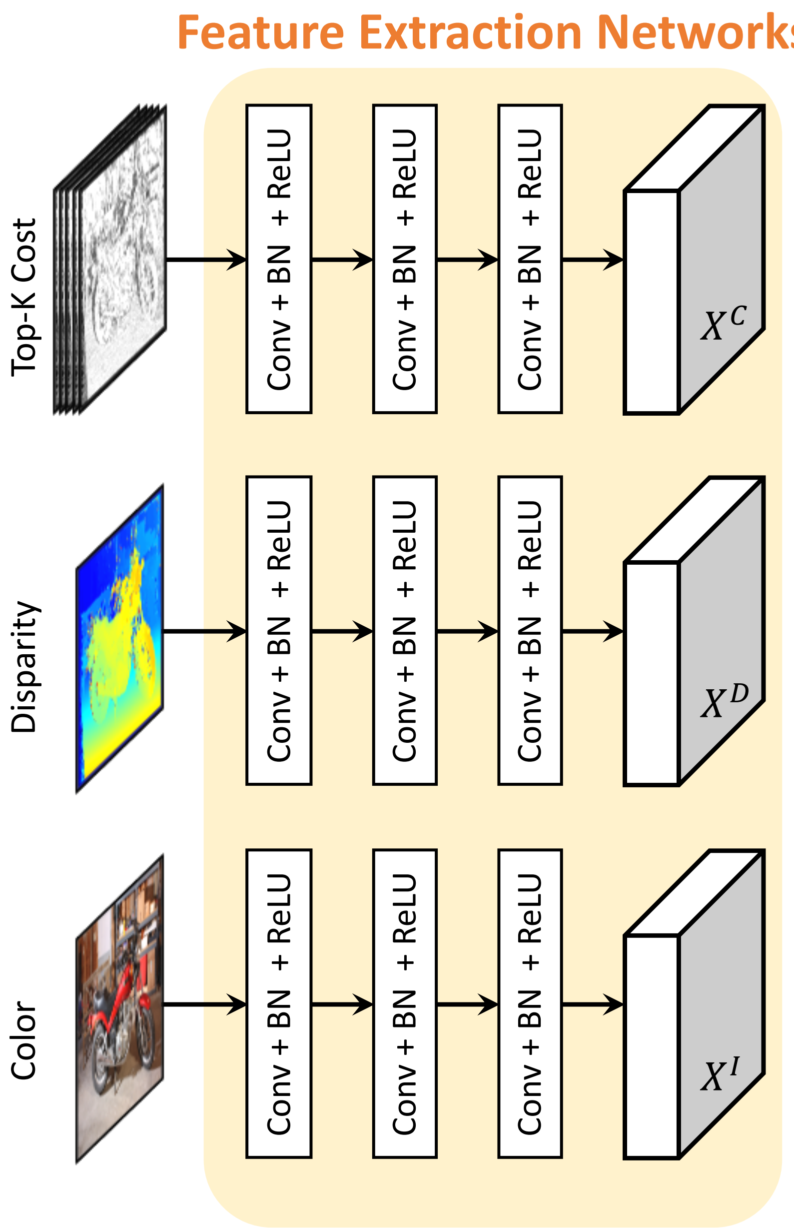
Locally Adaptive Fusion Networks (LAF-Net)



Locally Adaptive Fusion Networks (LAF-Net)



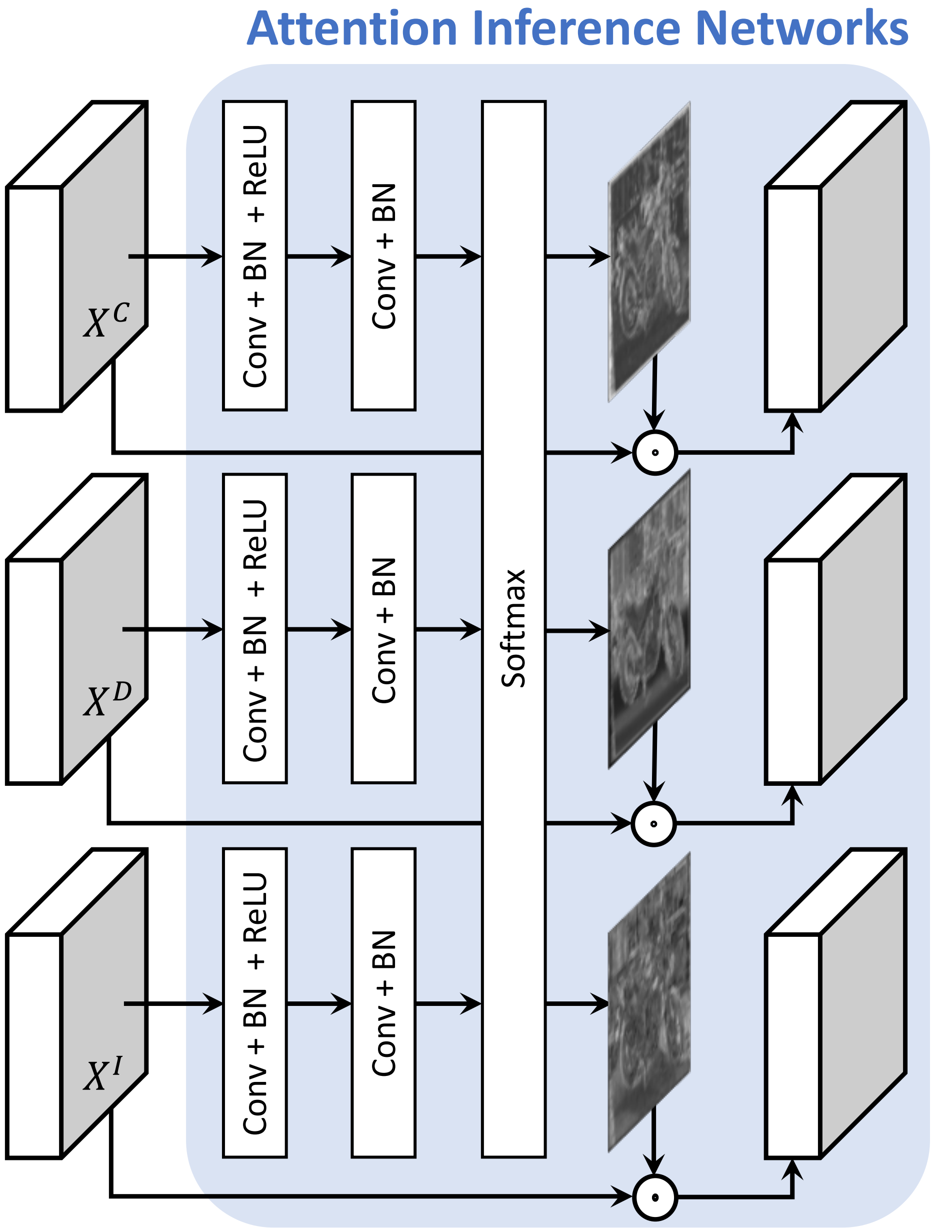
Feature Extraction Networks



I : Color
C : Matching cost
D : Initial disparity

X^I : Color features
X^C : Matching cost features
X^D : Initial disparity features

Attention Inference Networks

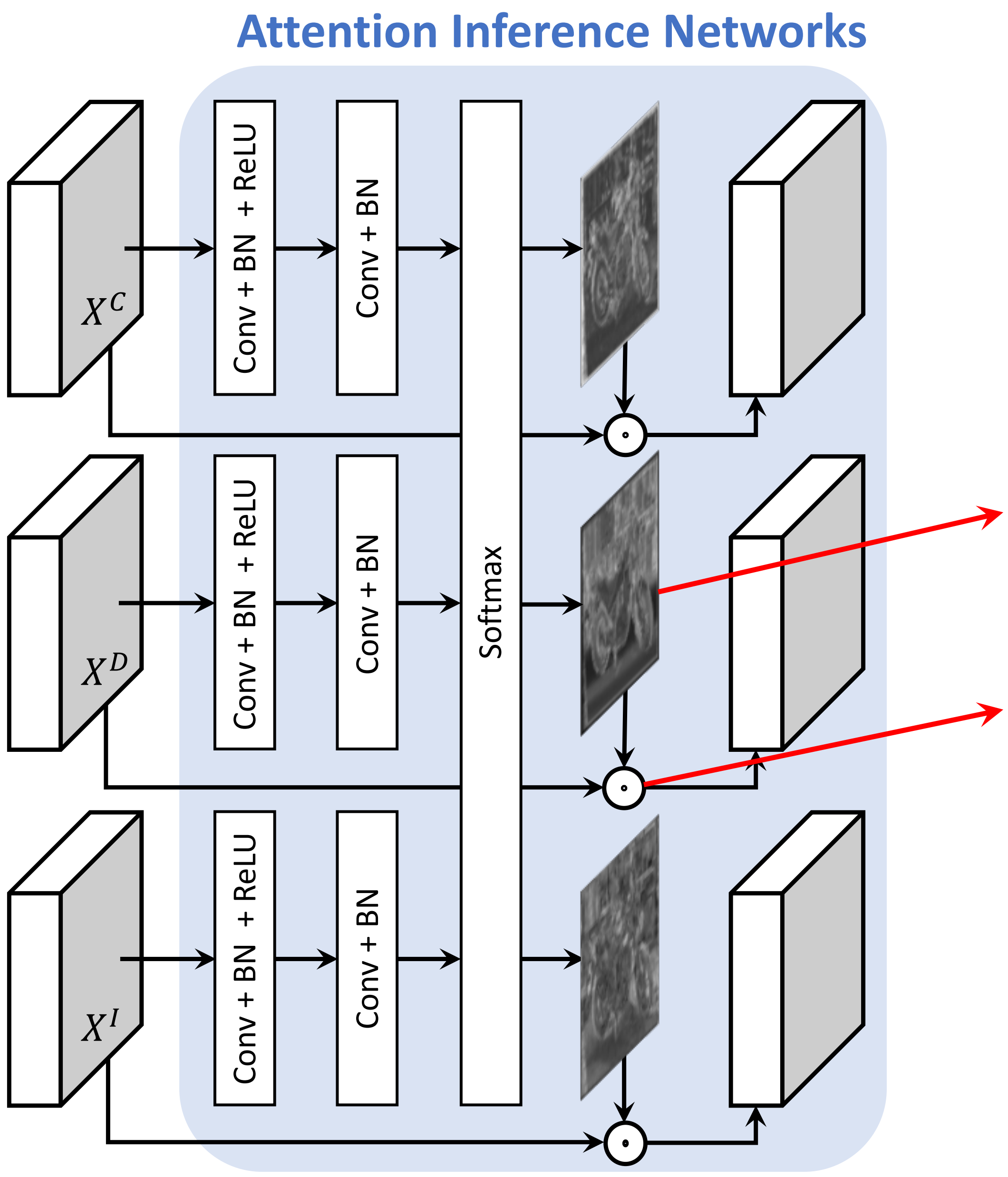


A^I : Attention for color

A^C : Attention for matching cost

A^D : Attention for initial disparity

Attention Inference Networks



A^I : Attention for color
A^C : Attention for matching cost
A^D : Attention for initial disparity

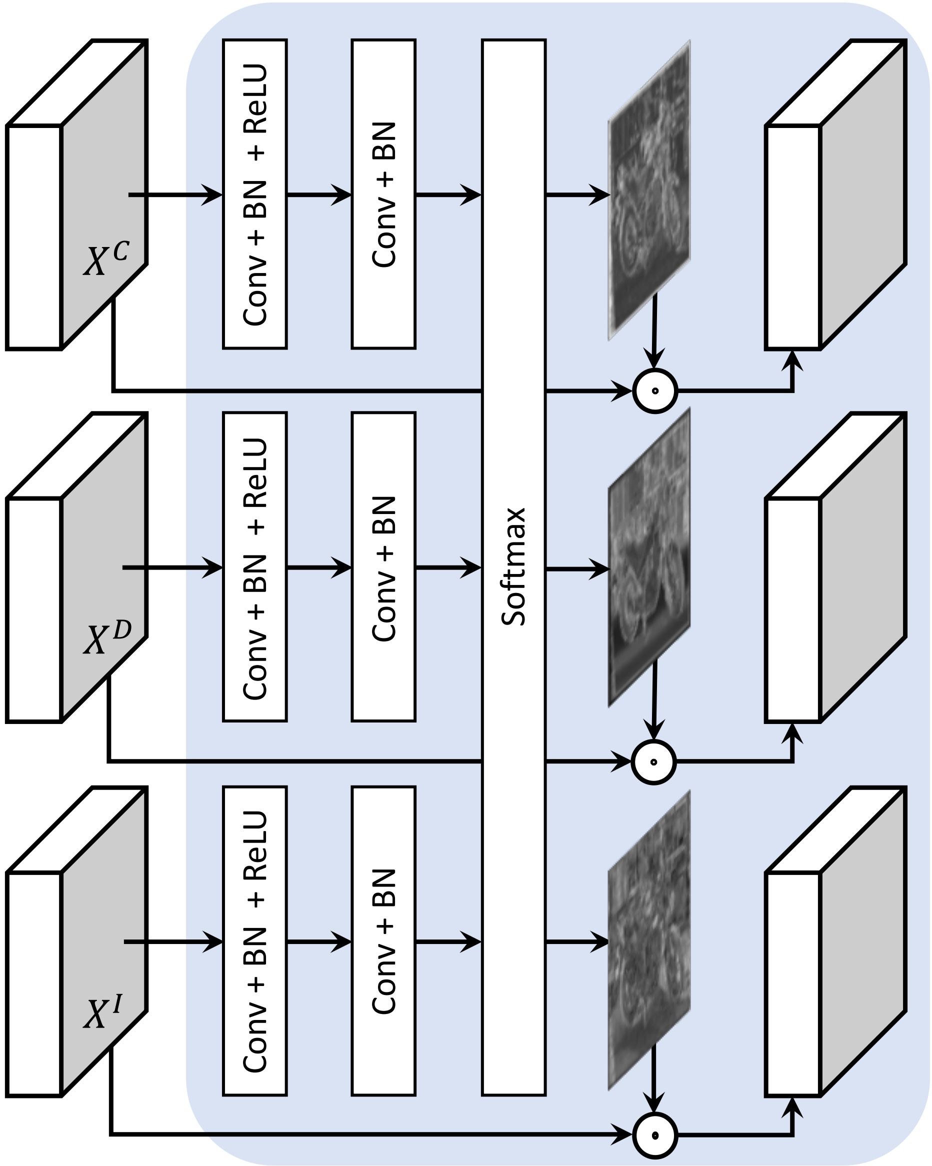
Locally-varying attention map

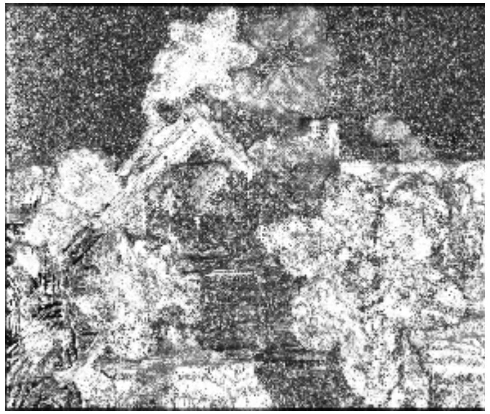
Element-wise multiplication of feature and attention map

Attention-boosted features:
$Y = \prod(X^I \odot A^I, X^C \odot A^C, X^D \odot A^D)$

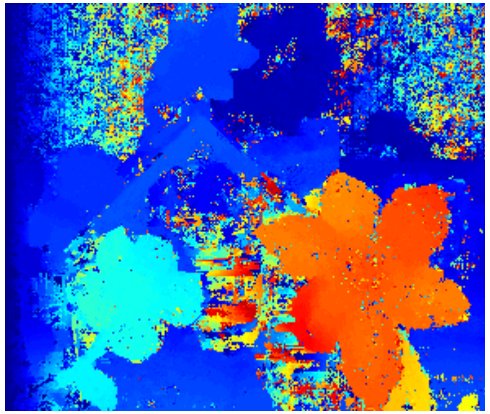
Attention Inference Networks

Attention Inference Networks



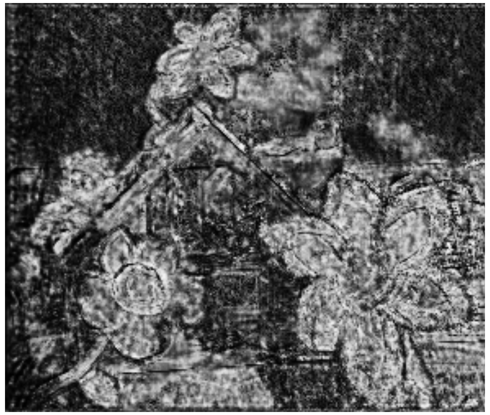


Matching cost

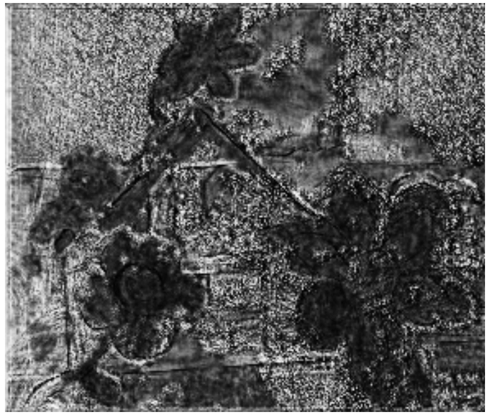


Disparity

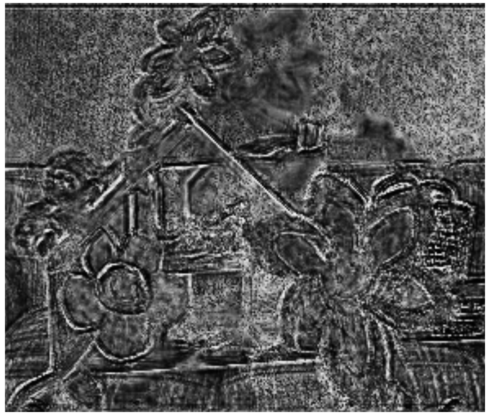
Color image



Attention of matching cost

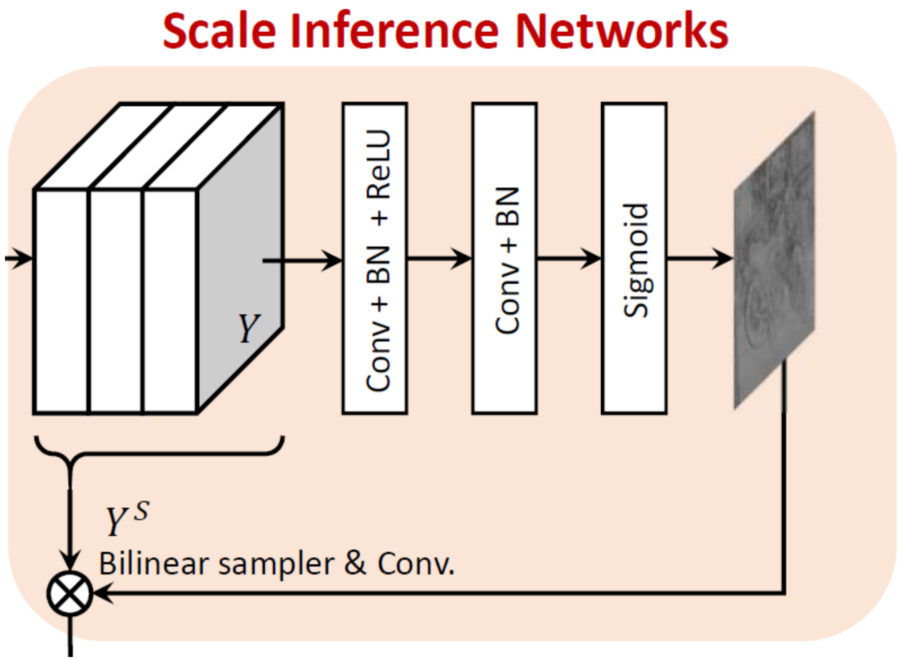


Attention of disparity



Attention of color image

Scale Inference Networks

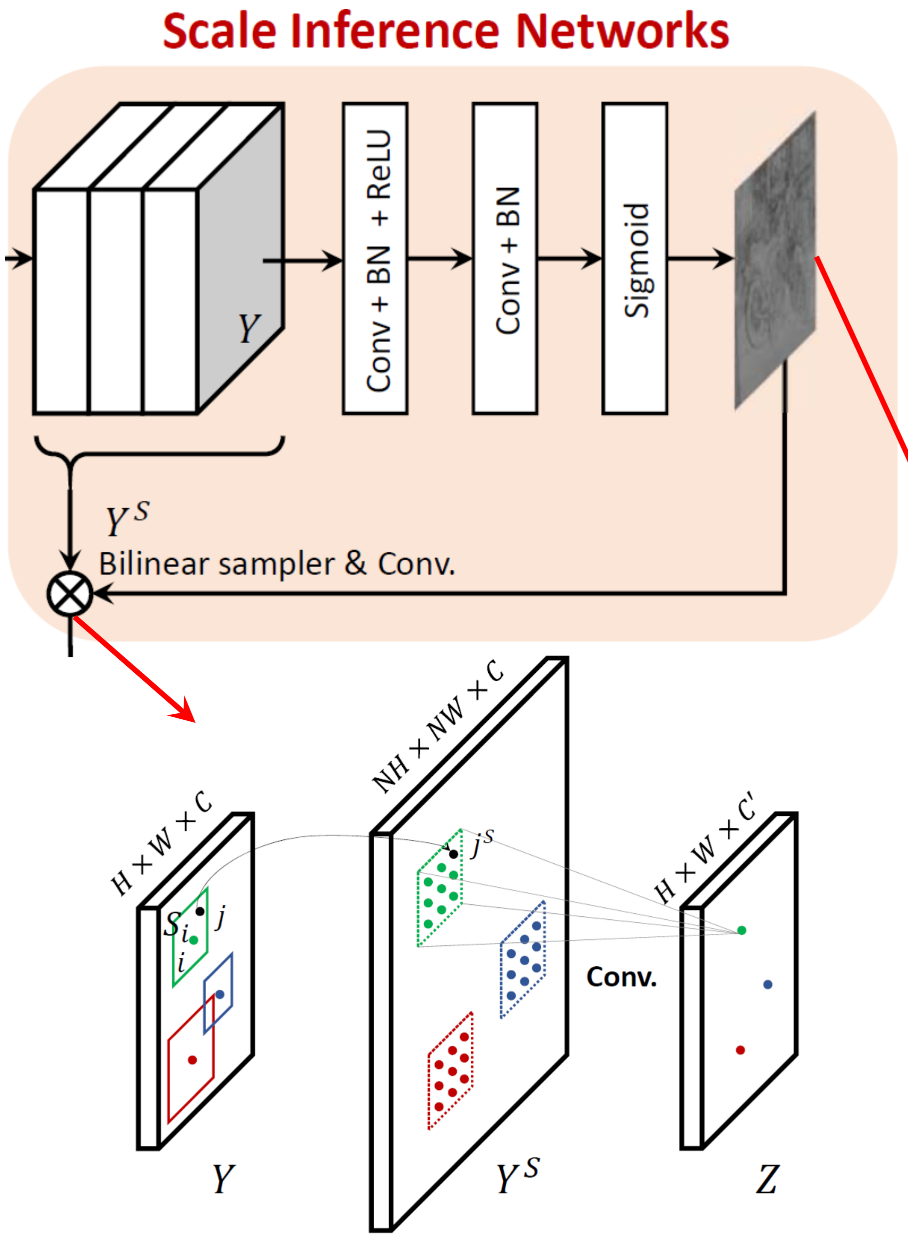


Y : Attention-boosted features

Y^S : Warped attention-boosted features

Z : Scale-adaptive features

Scale Inference Networks

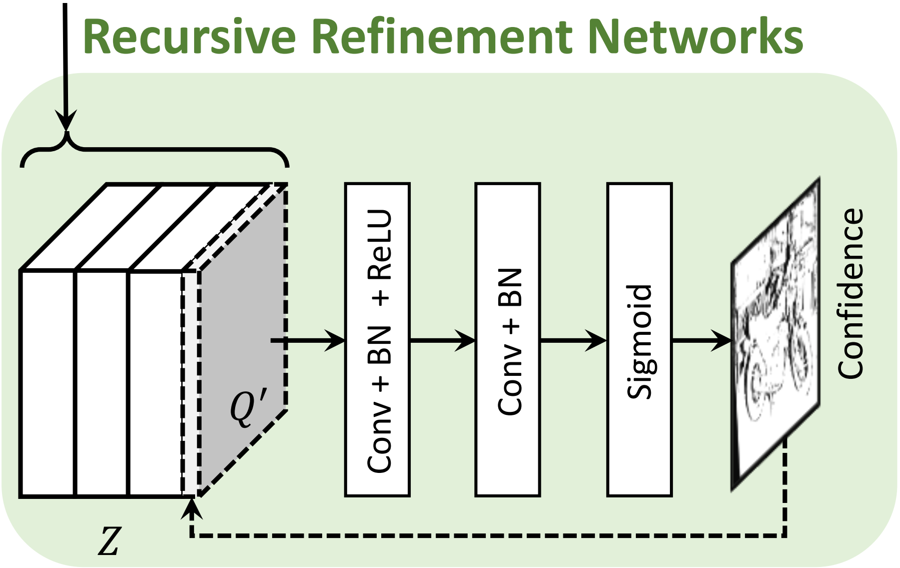


Y : Attention-boosted features
Y^S : Warped attention-boosted features
Z : Scale-adaptive features

Optimal scale is inferred for each pixel

$(Y \rightarrow Y^S)$ Using locally-varying sampling grid, the convolution activation Y are resampled into Y^S
$(Y^S \rightarrow Z)$ Convolution is applied

Recursive Refinement Networks

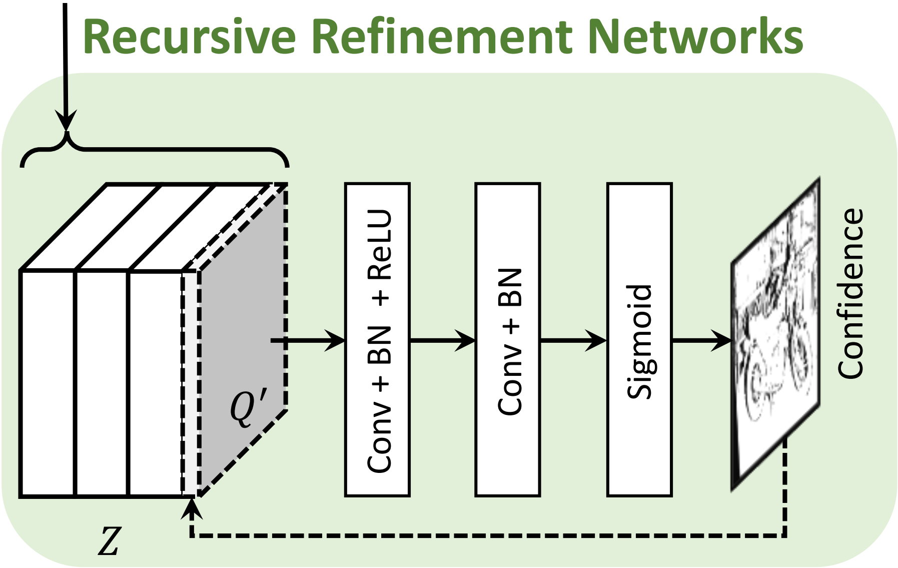


Z : Attention for color

Q^t : Confidence at t^{th} iteration

Q^{t-1} : Confidence at $t-1^{th}$ iteration

Recursive Refinement Networks



Z : Attention for color

Q^t : Confidence at t^{th} iteration

Q^{t-1} : Confidence at $t-1^{th}$ iteration

Recursive confidence estimation:

$Q^t = F(Z, Q^{t-1})$

Final confidence:

$Q = Q^{\max}$

Experimental Results

Ablation study of input tri-modal data

Match. cost	✓		✓		✓
Disparity		✓		✓	✓
Color			✓	✓	✓
MID 2006	0.0431	0.0392	0.0381	0.0375	**0.0364**
MID 2014	0.0762	0.0703	0.0687	0.0685	**0.0683**
KITTI 2015	0.0347	0.0245	0.0237	0.0231	**0.0225**

Ablation study of three sub-networks

Attention	✓			✓	✓
Scale		✓		✓	✓
Recursive			✓		✓
MID 2006	0.0374	0.0375	0.0372	0.0371	**0.0364**
MID 2014	0.0686	0.0688	0.0685	0.0685	**0.0683**
KITTI 2015	0.0235	0.0236	0.0231	0.0229	**0.0225**

Experimental Results

Ablation study of input tri-modal data

Match. cost	✓		✓		✓
Disparity		✓		✓	✓
Color			✓	✓	✓
MID 2006	0.0431	0.0392	0.0381	0.0375	**0.0364**
MID 2014	0.0762	0.0703	0.0687	0.0685	**0.0683**
KITTI 2015	0.0347	0.0245	0.0237	0.0231	**0.0225**

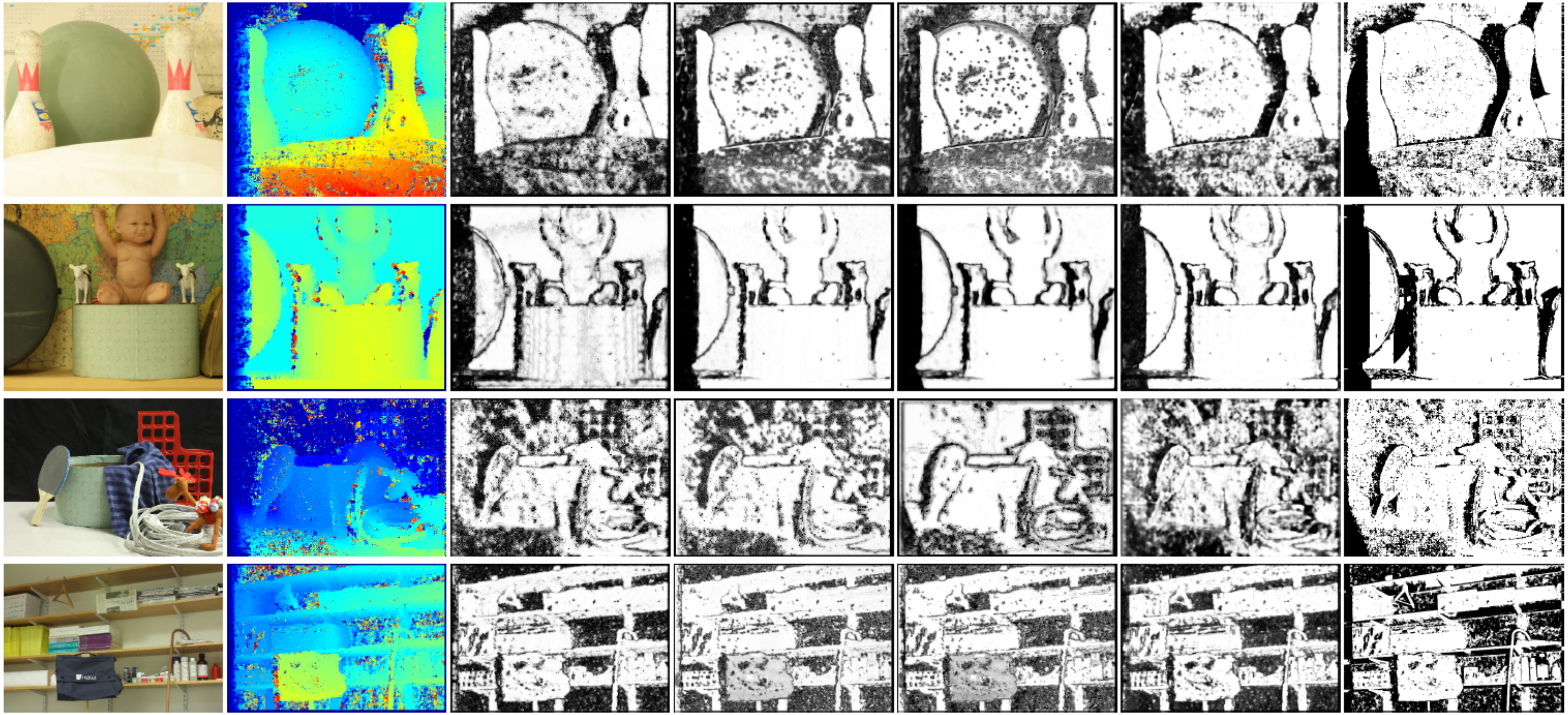
Ablation study of three sub-networks

Attention	✓			✓	✓
Scale		✓		✓	✓
Recursive			✓		✓
MID 2006	0.0374	0.0375	0.0372	0.0371	**0.0364**
MID 2014	0.0686	0.0688	0.0685	0.0685	**0.0683**
KITTI 2015	0.0235	0.0236	0.0231	0.0229	**0.0225**

Using tri-modal inputs and three sub-networks leads to substantial performance gain!

Experimental Results

Qualitative Evaluation



Color Disparity Kim et al. [21] LFN [7] LGC-Net [39] Ours GT

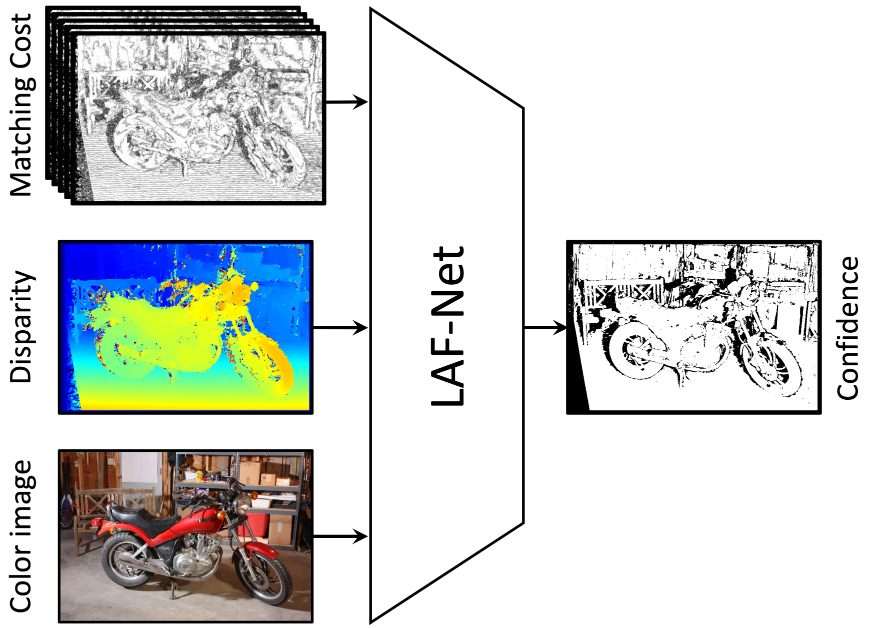
Experimental Results

Quantitative Evaluation (Average AUC)

- Middlebury 2006 (MID 2006), Middlebury 2014 (MID 2014), KITTI 2015

Datasets	MID 2006 [34]		MID 2014 [33]		KITTI 2015 [24]	
	Census-SGM	MC-CNN	Census-SGM	MC-CNN	Census-SGM	MC-CNN
Haeusler et al. [8]	0.0454	0.0417	0.0841	0.0750	0.0585	0.0308
Spyropoulos et al. [38]	0.0447	0.0420	0.0839	0.0752	0.0536	0.0323
Park and Yoon [27]	0.0438	0.0426	0.0802	0.0734	0.0527	0.0303
Poggi et al. [29]	0.0439	0.0413	0.0791	0.0707	0.0461	0.0263
Kim et al. [20]	0.0430	0.0409	0.0772	0.0701	0.0430	0.0294
CCNN [30]	0.0454	0.0402	0.0769	0.0716	0.0419	0.0258
PBCP [36]	0.0462	0.0413	0.0791	0.0718	0.0439	0.0272
Shaked et al. (Conf) [37]	0.0464	0.0495	0.0806	0.0736	0.0531	0.0292
Kim et al. (conf) [21]	0.0419	0.0394	0.0749	0.0694	0.0407	0.0250
LFN [7]	0.0416	0.0393	0.0752	0.0692	0.0405	0.0253
ConfNet [39]	0.0451	0.0428	0.0783	0.0721	0.0486	0.0277
LGC-Net [39]	0.0413	0.0389	0.0735	0.0685	0.0392	0.0236
LAF-Net	**0.0405**	**0.0364**	**0.0718**	**0.0683**	**0.0385**	**0.0225**
Optimal	0.0340	0.0323	0.0569	0.0527	0.0348	0.0170

Concluding Remarks



- **Using tri-modal input leads to a substantial performance gain**
  - Matching cost, disparity, and color image
- **Attention and scale inference networks are used to fuse heterogeneous tri-modal input**
- **Recursive refinement networks improves the accuracy**

Thank you!

Poster 80 @Tuesday, Session 1.1

Seungryong Kim
Post-Doctoral Researcher
School of Computer and Communication Sciences (IC)
École Polytechnique Fédérale de Lausanne (EPFL)
E-mail: seungryong.kim@epfl.ch
Homepage: http://seungryong.github.io