Large-scale Outdoor RGB+D dataset and Its Application to Single Image Depth Estimation

Dongbo Min
Department of Computer Science and Engineering
Ewha Womans University, Korea
E-mail: dbmin@ewha.ac.kr

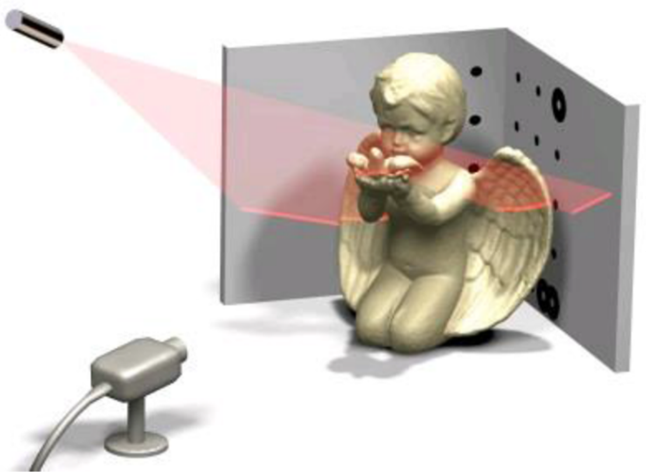
Contents

- Overview of single image depth estimation
- Constructing large-scale RGB+D dataset
  - DIML/CVL RGB+D dataset
- Stereo confidence estimation approach

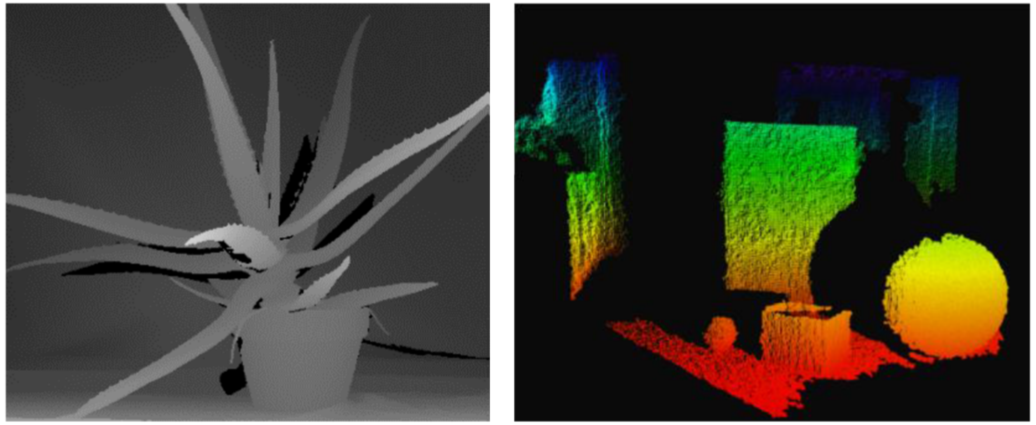
3D Sensing (Depth Estimation)

- 3D Sensing
  - Estimating depth or distance from a sensor to the scene surface, or complete 3D shape (structure) of the scene based on the geometrical and photometrical properties

1) 3D sensing with laser scanner

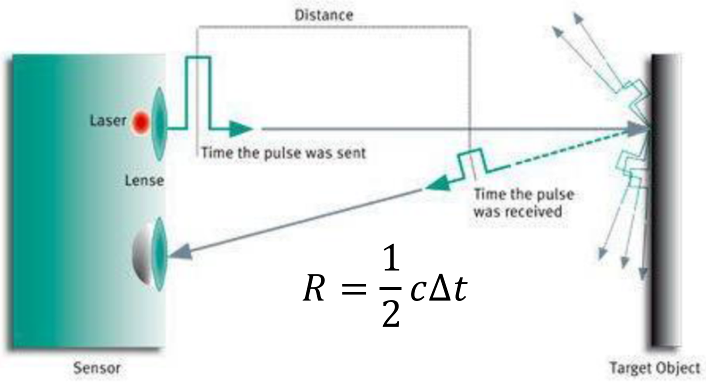


2) 3D sensing using stereo vision

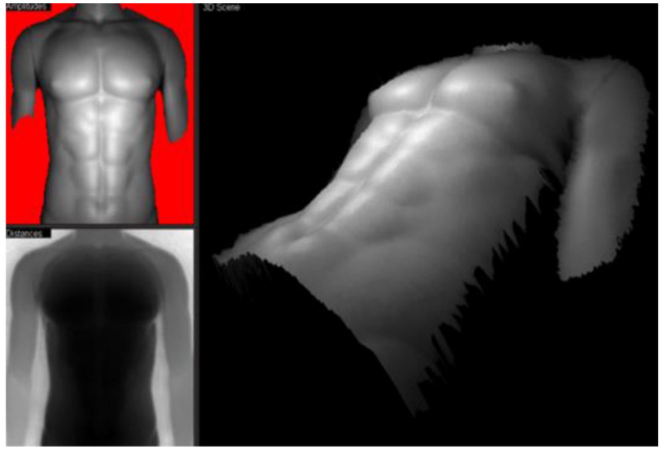


3D Sensing using Active Sensors

Time of Flight Sensor

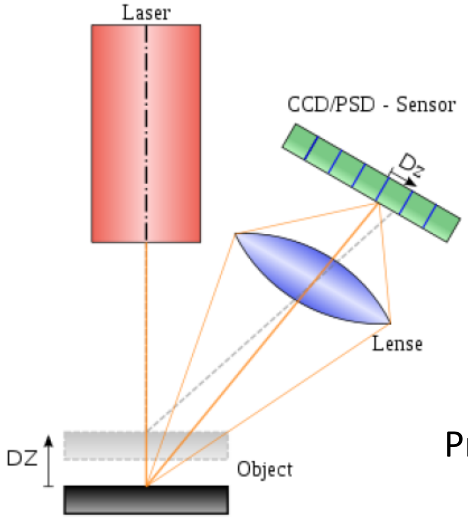


$$R = \frac{1}{2} c\Delta t$$

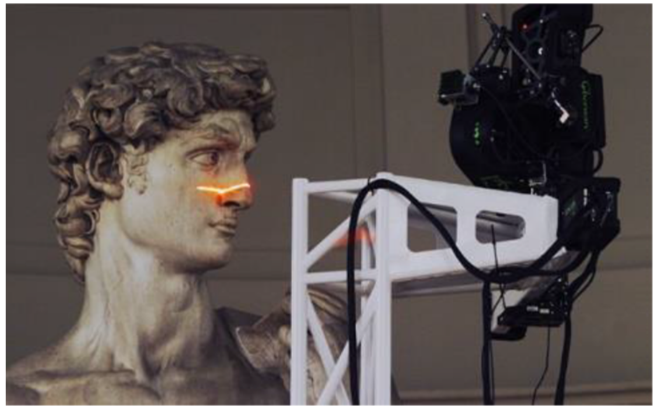


Principle of ToF sensors and acquired 3D data

Laser Scanner



Principle of Laser scanning



Digital Michelangelo Project
http://graphics.stanford.edu/projects/mich

3D Sensing using Shape-from-X

- Shape (Structure)-from-X
  - X: visual cue that can be extracted from images
    - Shading
    - Silhouette
    - Focus
    - Perspective effects
    - Occlusion
    - Motion
    - ***Stereo***

Pros

Applicable in general and relatively uncontrolled conditions

Large working ranges

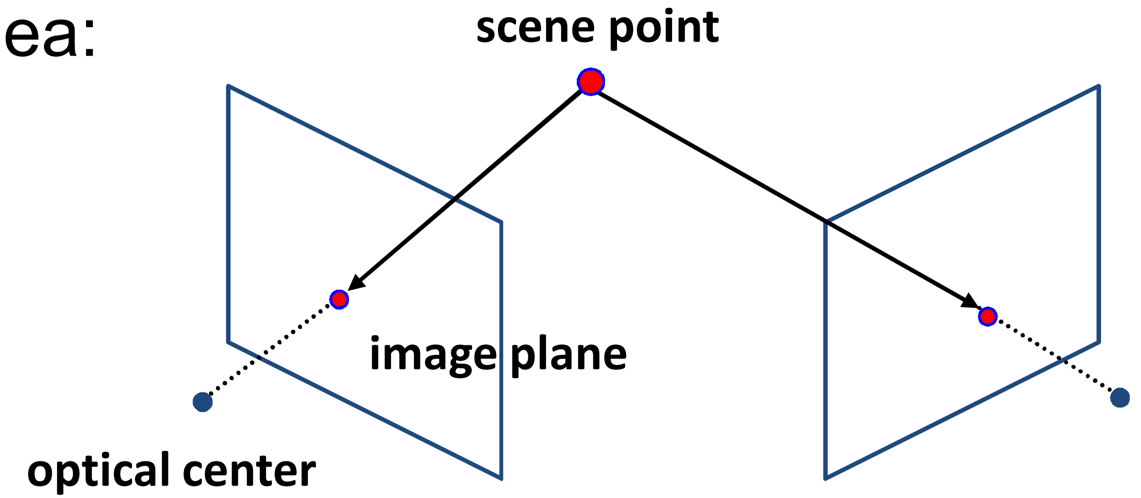
Cons

Low accuracy compared to the methods in metrology

3D Sensing using Shape-from-Stereo

- **Stereo**:
  - Shape from “motion” between two views
  - Infer 3D shape of scene from two (or multiple) images from different viewpoints

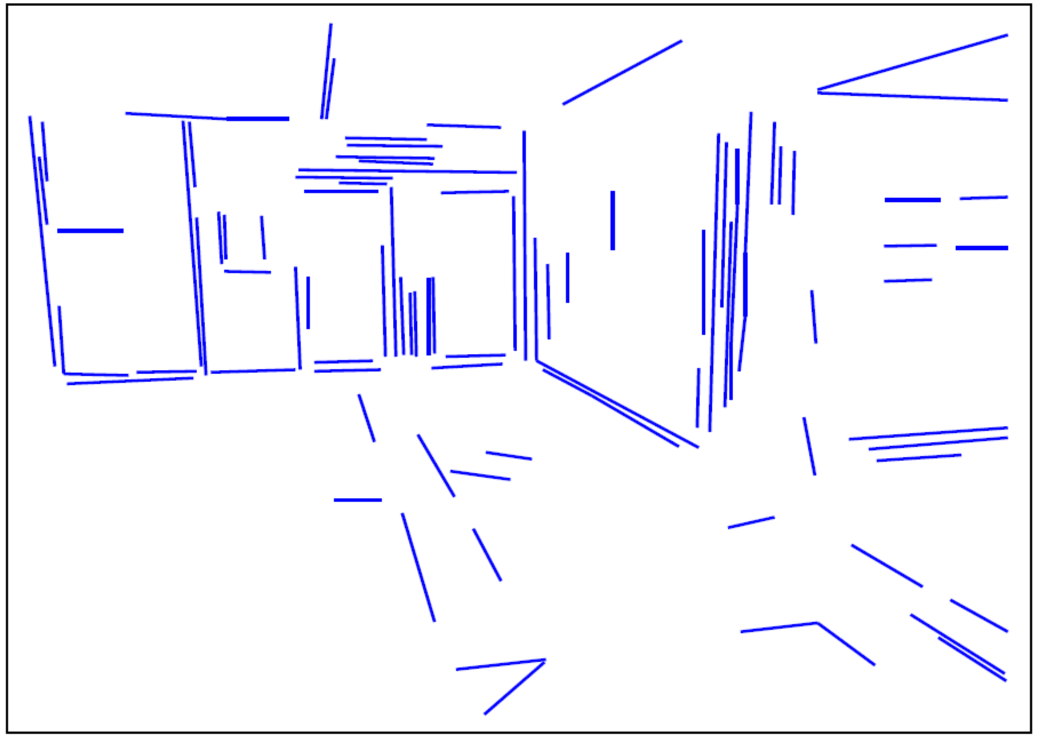
Main idea:



3D Sensing using Single (=Monocular) Image Only?

- **Goal**: Estimate 3D depth map from single image
  - Numerous approaches have been proposed using hand-crafted cues

    Ex) object contour, object segment, object motion and shading



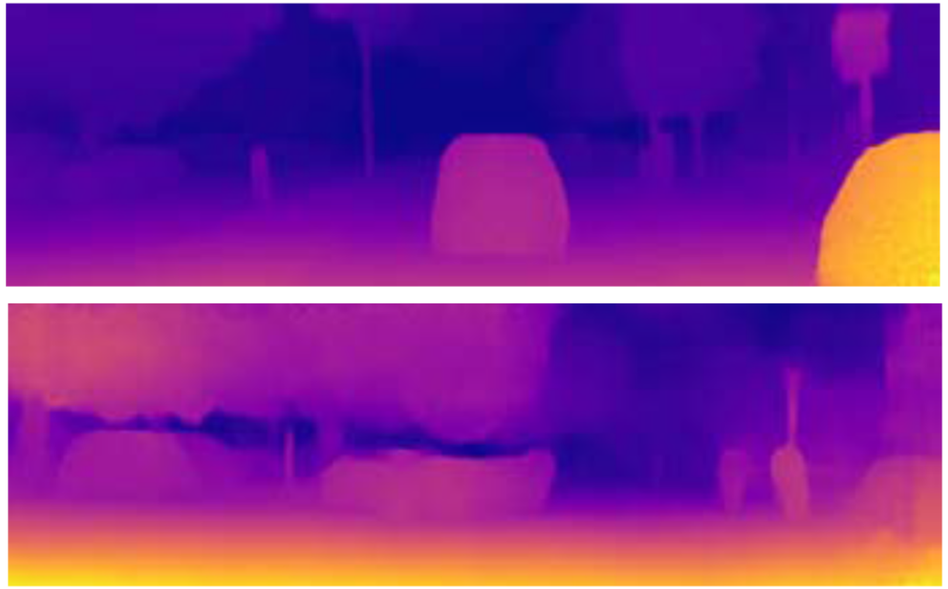
Such hand-crafted approaches often fail to capture plausible depth or work only at restricted environments

3D Sensing using Single (=Monocular) Image Only?

- Goal: Estimate 3D depth map from single image

Convolutional neural networks (CNNs) leads to a substantial improvement in 3D sensing using a single image

Input images



Depth maps from CNN-based monocular depth estimation approach

Overview of Monocular Depth Estimation for Deep Network

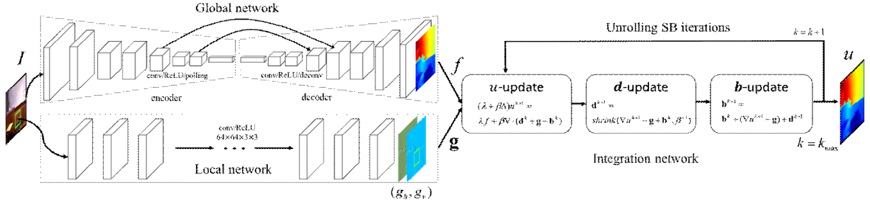
Training phase

1) Constructing Large-scale RGB+D Dataset

Large-scale RGB+D Database

3D depth sensing device

RGB Image



D

2) Deep learning model for single image depth estimation

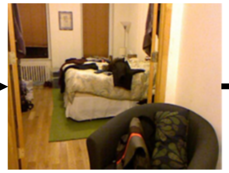
Testing phase

D : Ground truth depth maps from 3D depth sensing devices

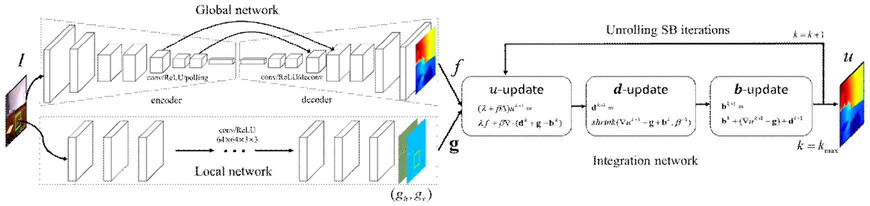
D^*: Depth map estimated using deep network

User

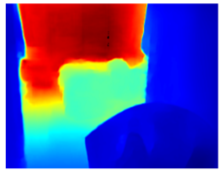
RGB Image



Depth estimation model for single image



D^*



Overview of Monocular Depth Estimation for Deep Networkn

- **Research items**
  1. Constructing large-scale RGB+D dataset
  2. Deep learning model for single image depth estimation

Related Project

High quality 2D-to-Multiview contents generation from large-scale-RGB+D database

Funding: Information and communications Technology Promotion (IITP)

Period: 2015.07 ~ 2017.08

Research Papers from the Project

- **Constructing large-scale RGB+D dataset**

  [1] DIML/CVL RGB+D dataset (1M outdoor dataset)

  [2][3][4][5] Stereo confidence estimation

- **Deep learning model for single image depth estimation**

  [6] Deep variational approach for single image depth estimation

[1] A Large RGB-D Dataset for Semi-supervised Monocular Depth Estimation, IEEE Trans. on Image Processing (under review)
[2] Feature Augmentation for Learning Confidence Measure in Stereo Matching, IEEE Trans. on Image Processing 2017
[3] Unified Confidence Estimation Networks for Robust Stereo Matching, IEEE Trans. on Image Processing 2019
[4] Learning Adversarial Confidence Measures for Robust Stereo Matching, IEEE Trans. on Image Processing (under review)
[5] LAF-Net: Locally Adaptive Fusion Networks for Stereo Confidence Estimation, IEEE CVPR 2019 (oral presentation)
[6] A Deep Variational Approach for Single Image Depth Estimation, IEEE Trans. on Image Processing, 2018

In This Talk

- **Constructing large-scale RGB+D dataset**
  - [1] DIML/CVL RGB+D dataset (1M outdoor dataset)
  - [2][3][4][5] Stereo confidence estimation

- **Deep learning model for single image depth estimation**
  - [6] Deep variational approach for single image depth estimation

[1] A Large RGB-D Dataset for Semi-supervised Monocular Depth Estimation, IEEE Trans. on Image Processing (under review)
[2] Feature Augmentation for Learning Confidence Measure in Stereo Matching, IEEE Trans. on Image Processing 2017
[3] Unified Confidence Estimation Networks for Robust Stereo Matching, IEEE Trans. on Image Processing 2019
[4] Learning Adversarial Confidence Measures for Robust Stereo Matching, IEEE Trans. on Image Processing (under review)
[5] LAF-Net: Locally Adaptive Fusion Networks for Stereo Confidence Estimation, IEEE CVPR 2019 (oral presentation)
[6] A Deep Variational Approach for Single Image Depth Estimation, IEEE Trans. on Image Processing, 2018

DIML/CVL RGB+D Dataset

1. How to acquire and process the RGB+D dataset

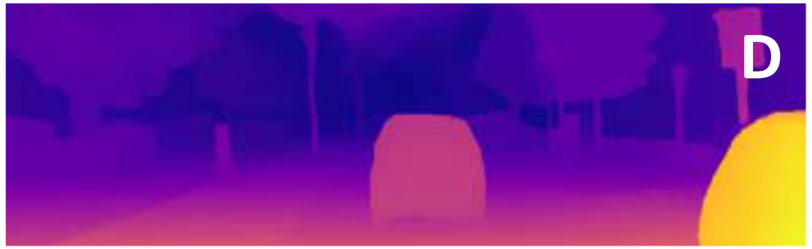
   DIML/CVL RGB-D Dataset: 2M RGB-D Images of Natural Indoor and Outdoor Scenes, Technical Report (http://diml.yonsei.ac.kr/DIML_rgbd_dataset/paper/technical_report.pdf)

2. Analyzing the RGB+D dataset

   A Large RGB-D Dataset for Semi-supervised Monocular Depth Estimation, IEEE Trans. on Image Processing (under review)

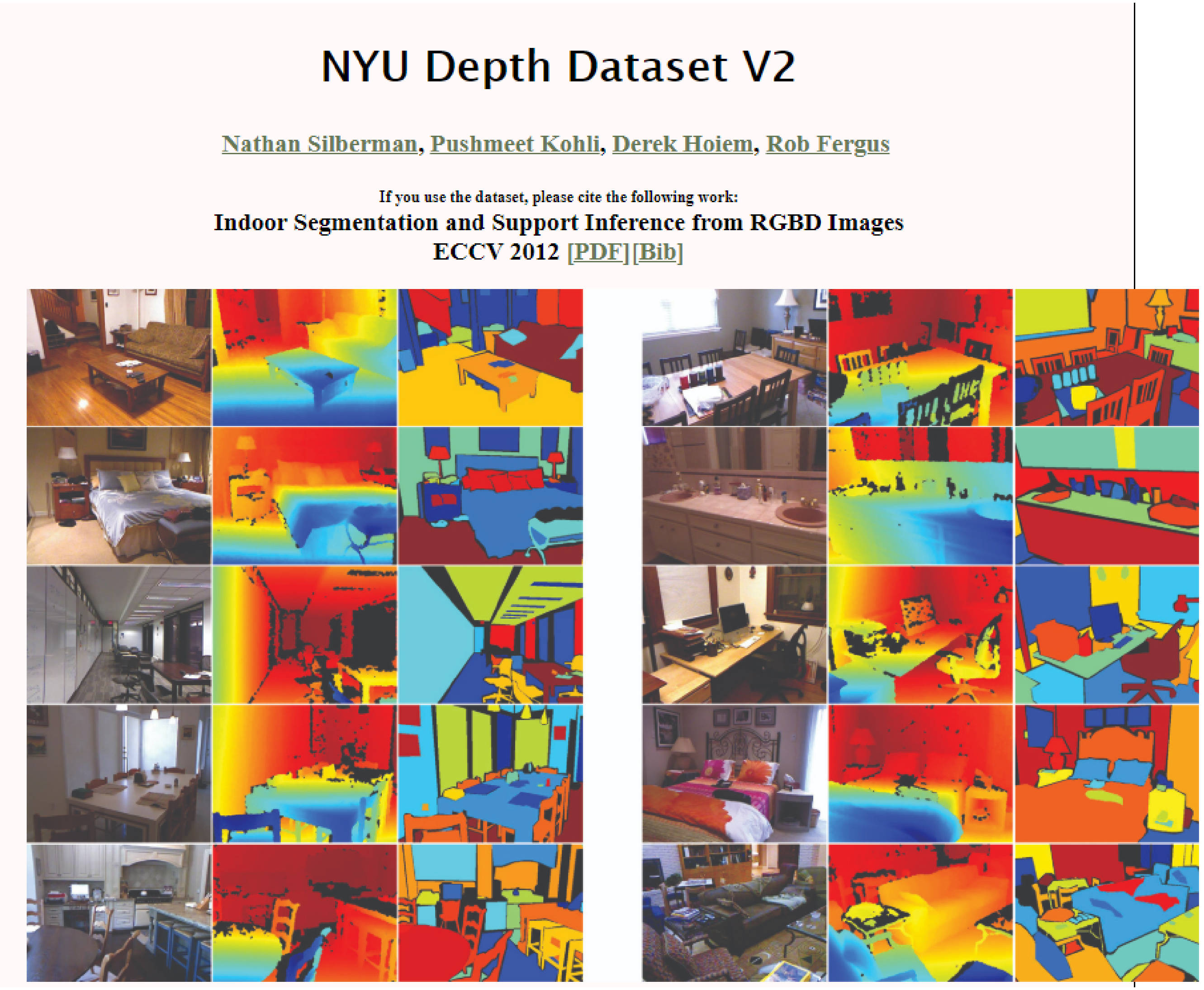
RGB+D Dataset

- RGB+D dataset
  - RGB (color image) + D (depth map)



RGB+D Dataset

- Many RGB+D datasets exist for *indoor* scenes



RGB+D Dataset

- But for *outdoor* scenes, a large-scale dataset is not ready yet due to the difficulty in obtaining depth maps!

Capturing devices/tools

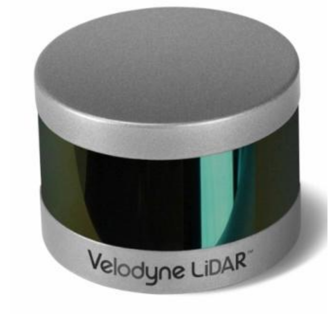
Structured light (Kinect)

Not applicable to outdoor scenes

Time-of-flight (Kinect v2)

Not applicable to outdoor scenes

Laser scanner (Velodyne LiDAR)



Accurate sensing results

Too sparse and expensive

(16 lines for vertical resolution)

RGB+D Dataset

- But for *outdoor* scenes, a large-scale dataset is not ready yet due to the difficulty in obtaining depth maps!

Capturing devices/tools

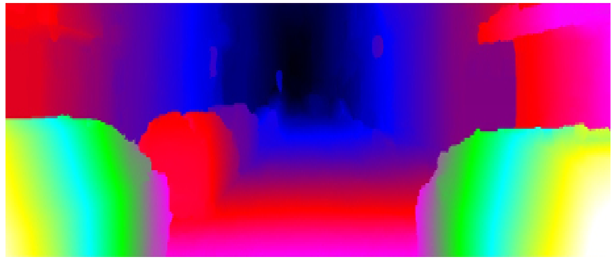
3D Graphic Rendering

Non-photorealistic

Manual Labeling

Cityscape data for semantic segmentation

Number of Semantic Labels: 30



KITTI data for depth map

Number of Depth Labels: >>1000

Manual labeling is impossible!

Our RGB+D Dataset: DIML/CVL Dataset

- **Our solution**

Stereo camera

Pros: High resolution and cheap

Cons: Stereo matching error

Stereo matching

+

Stereo confidence

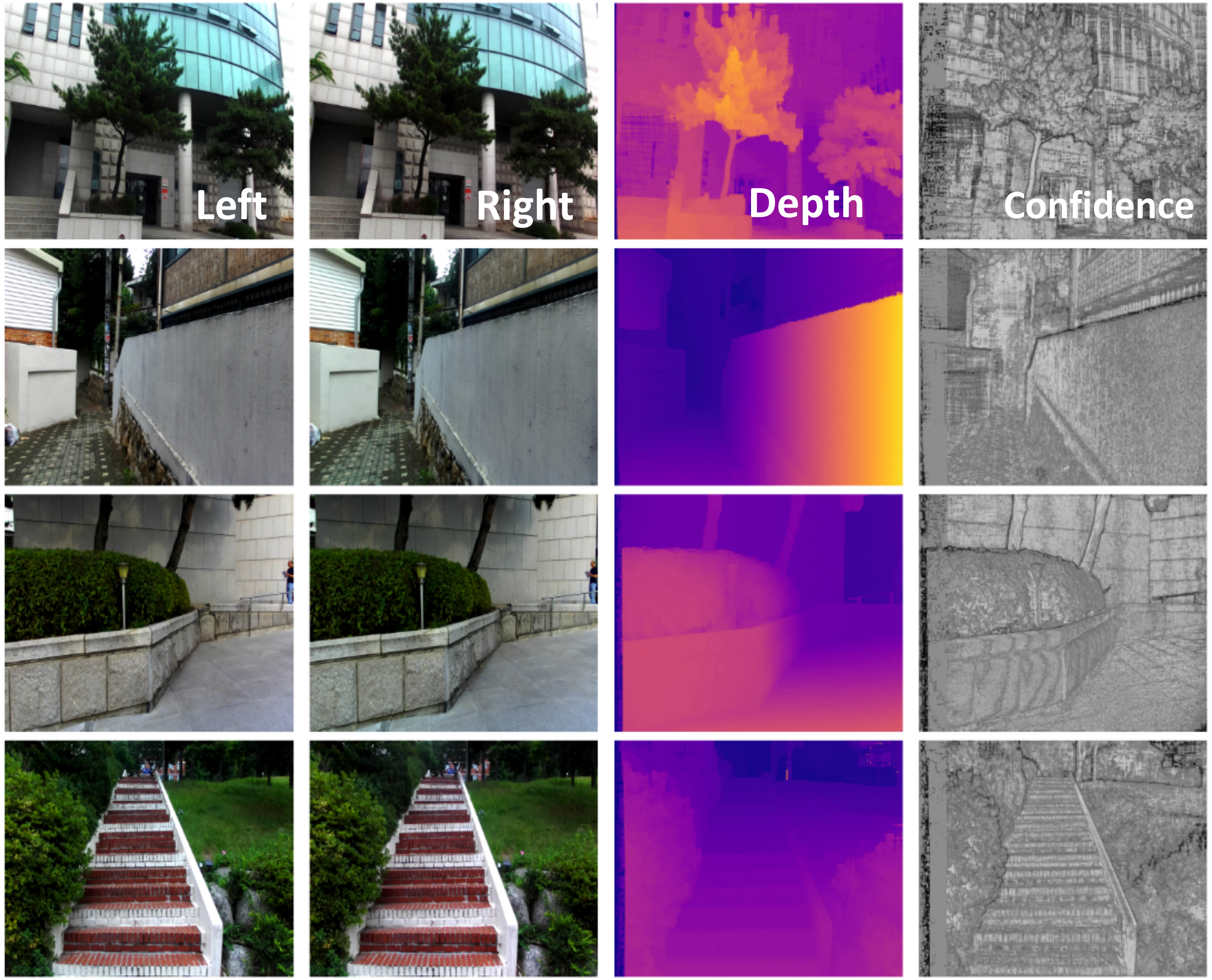
To compensate for erroneous depth estimates

- **DIML/CVL RGB+D dataset**
  - https://dimlrgbd.github.io/
  - Using stereo camera
  - Stereo matching for depth estimation
  - Confidence estimation of depth map

    Confidence map: indicates whether an estimated depth is reliable or not

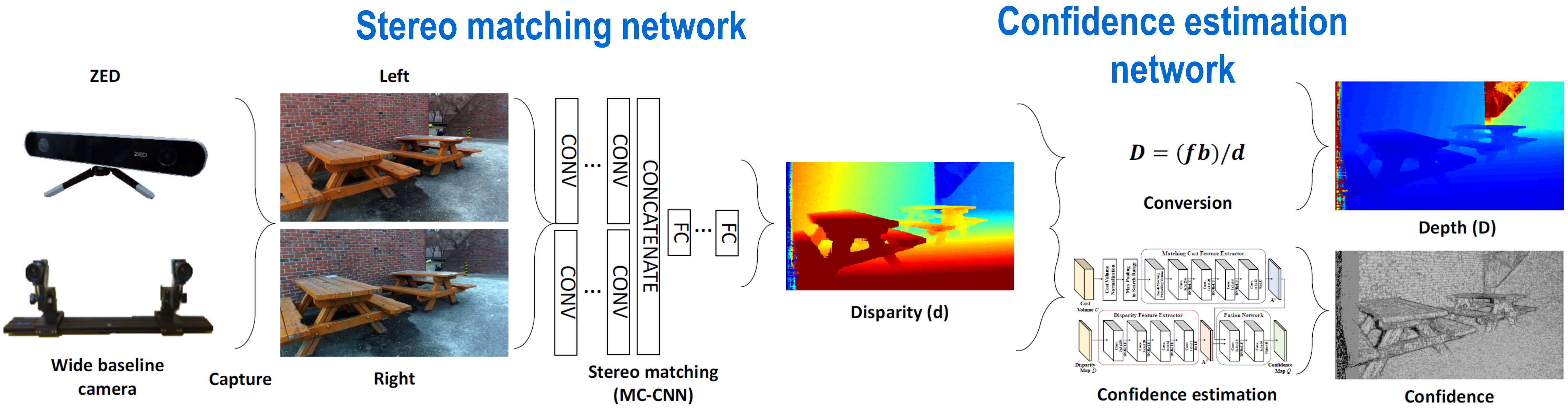
DIML/CVL RGB+D Dataset

Samples of our dataset



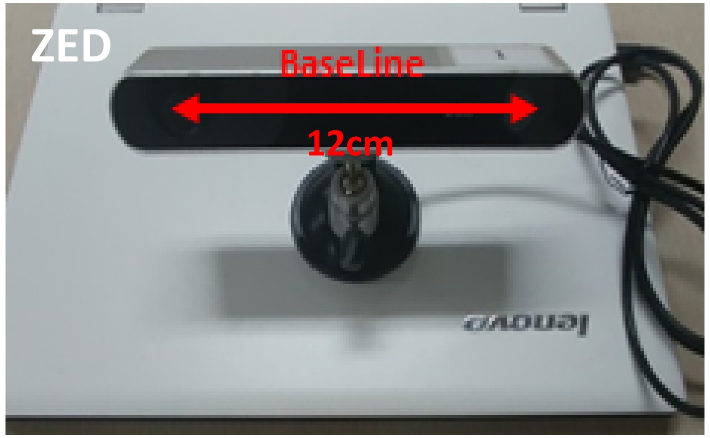
Confidence map: indicates whether an estimated depth is reliable or not
(0: unreliable <-> 1: reliable)

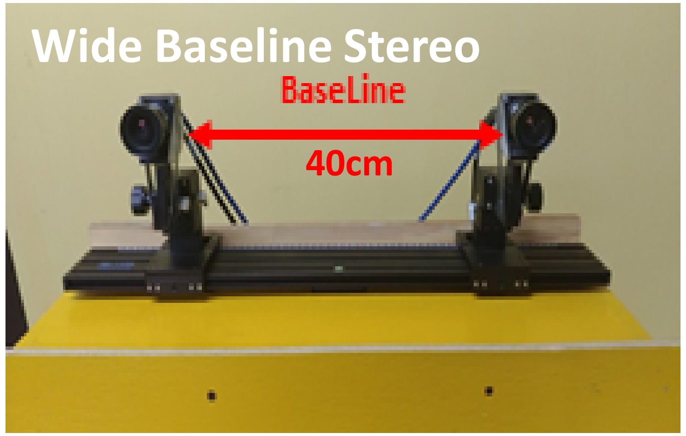
DIML/CVL RGB+D Dataset



Note) Any kind of stereo matching and confidence estimation approaches can be used here.

DIML/CVL RGB+D Dataset





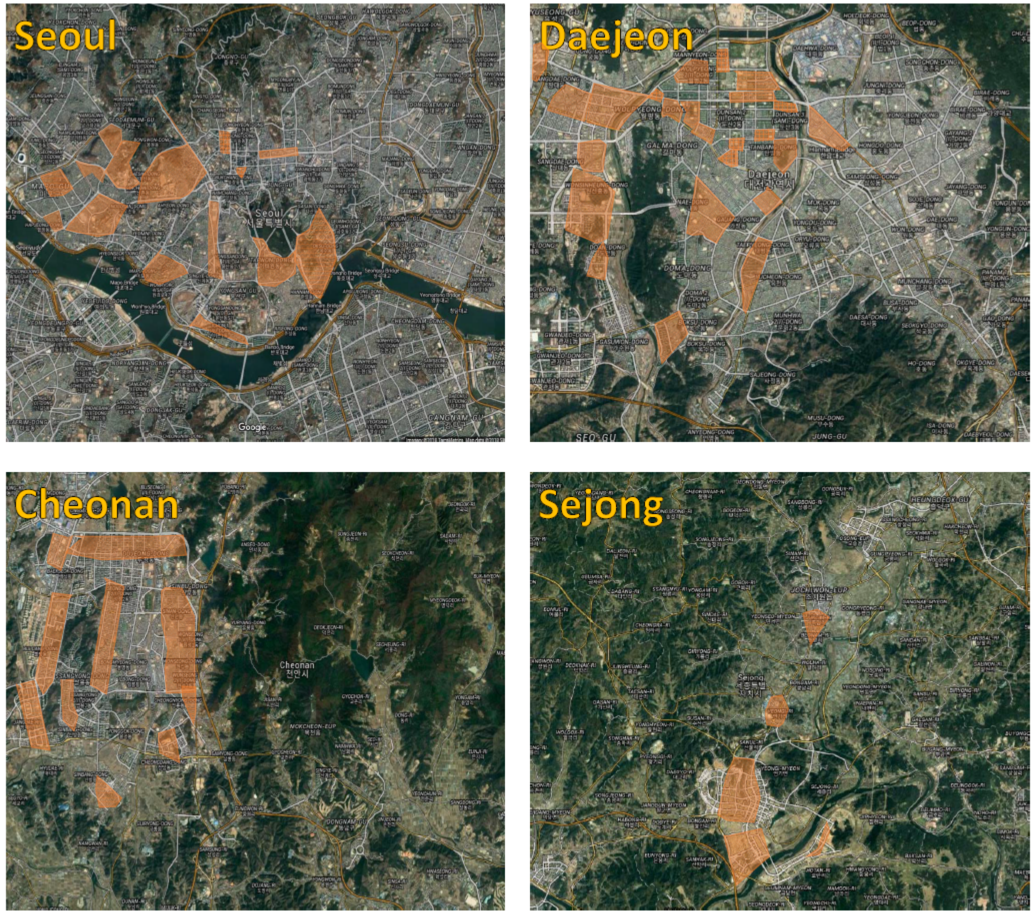
	ZED stereo	Built-in stereo
Color resolution	1920 x 1080 1280 x 720	1920 x 1080 1280 x 720
Depth resolution	1920 x 1080 1280 x 720	1920 x 1080 1280 x 720
Depth range	0.5 - 20 m	2 - 80 m
Baseline	12 cm	40 cm
Focal length	2.8 mm	3.5 mm

DIML/CVL RGB+D Dataset

	Outdoor dataset
Data acquisition	Stereo camera (ZED and built-in camera)
Data processing	• Calibration and rectification using Caltech toolbox • Stereo matching • Confidence estimation
Data format	**Color images** - Rectified left and right images **Disparity, depth, and confidence map** - Left disparity and depth map - Left Confidence map **Calibration parameters** - Intrinsic/extrinsic parameters for stereo camera

DIML/CVL RGB+D Dataset

- Shooting Location
  - Our dataset was acquired in various outdoor scenes including park, building, brook, road, apartment, and so on.
  - 4 different cities in South Korea: Seoul, Daejeon, Cheonan, Sejong



이화여자대학교
EWHA WOMANS UNIVERSITY

Scene category of our dataset

	Category	# of folders	# of files
	brook	1	4672
	building	22	58704
	construction	1	1871
	driveway	7	11114
	field	3	3039
	overpass	1	2794
	park	10	23384
	street	75	198097
	trail	9	18762

Our dataset

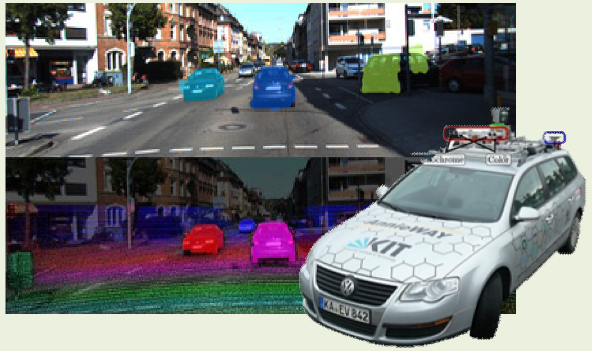
Non-driving scenes using hand-held stereo cameras (e.g., park, building, apartment, trail, and street)

Existing dataset

Driving scenes obtained from the depth sensor mounted on a vehicle (e.g., road and traffic scenes)

Comparison with Existing Outdoor Datasets

KITTI



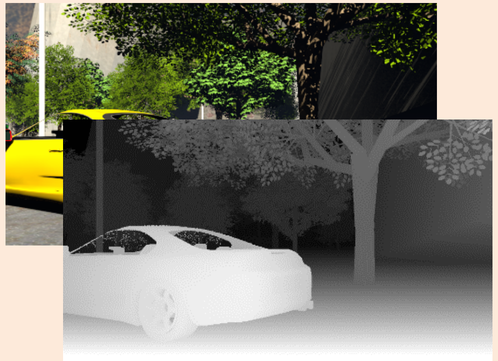
Depth map: Sparse LiDAR

RGB texturing: Real

# of RGB+D data: 40,000

Spatial resolution: 1242x375

DispNet



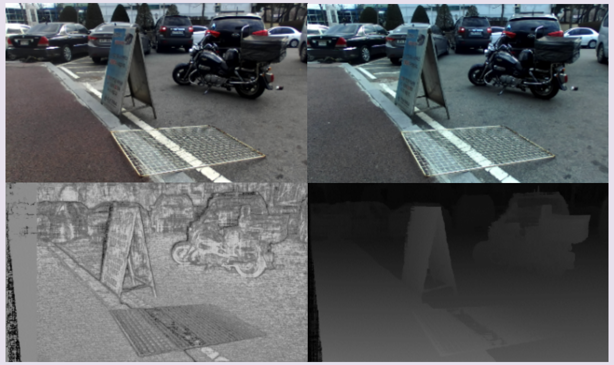
Depth map: Graphic rendering

RGB texturing: Synthetic

# of RGB+D data: 3,900

Spatial resolution: 960x540

DIML/CVL dataset



Depth map: Stereo matching + Confidence map

RGB texturing: Real

# of RGB+D data : 1,000,000

Spatial resolution: 1920x1080

Driving scenes

using LiDAR mounted on moving vehicle (40,000 data)

Graphic Data

Non-driving scenes

using hand-held stereo camera (1,000,000 data)

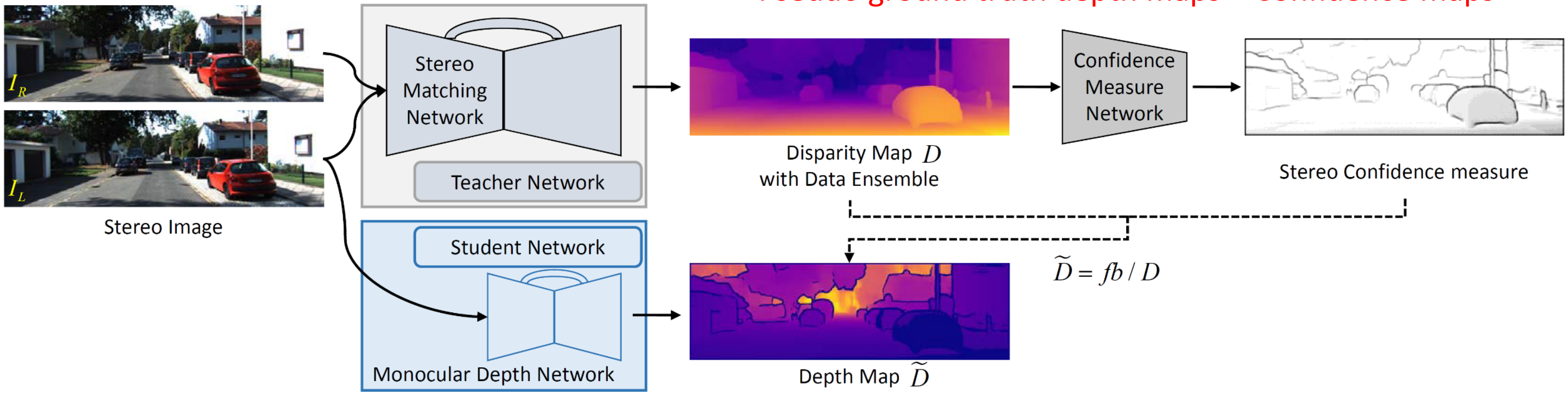
Our Approach using DIML/CVL Dataset

Our approach is based on 'Student-Teacher strategy'

Teacher network: stereo matching & confidence measure networks
(stereo images -> depth map & confidence map)

Student network: monocular depth network (single image -> depth map)

Pseudo ground truth depth maps + Confidence maps



Our Approach using DIML/CVL Dataset

Teacher network: RGB+D data generation

Training & Test: Left & Right image -> Left depth map & Confidence map

Student network

Training: Left image -> Left depth map (assisted by confidence map)

Test: Left image -> Left depth map

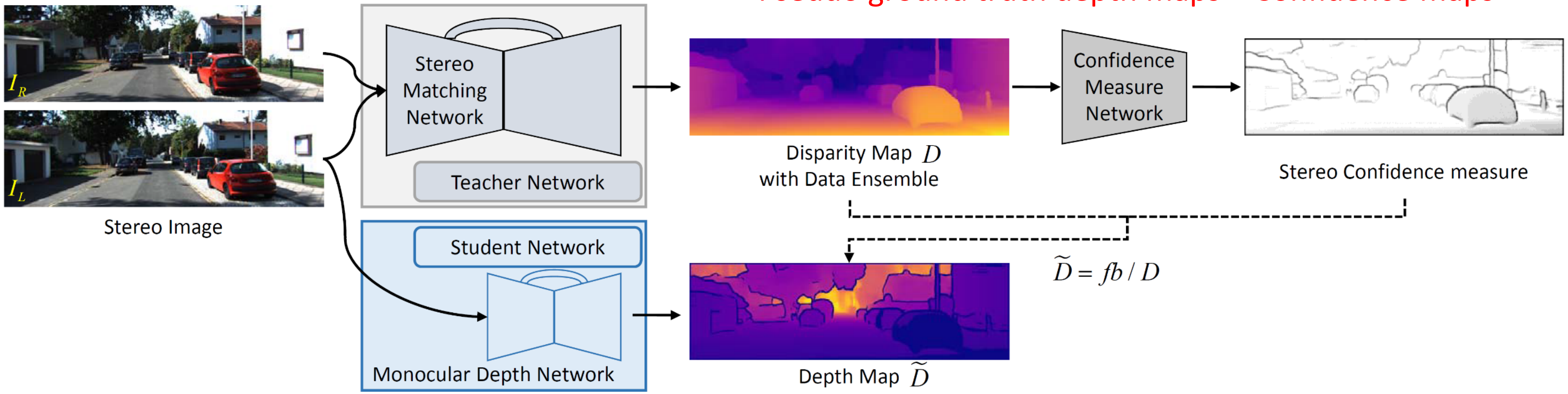
Loss function for student network

$$\mathcal{L}_c = \frac{1}{\sum_p M_p} \sum_p M_p \cdot \left| \hat{D}(p) - \tilde{D}(p) \right|_1 ,$$

$$M_p = \begin{cases} 1, & \text{if } C(p) \geq \tau \\ 0, & \text{if } C(p) < \tau \end{cases} .$$

C(p): confidence at pixel p

Pseudo ground truth depth maps + Confidence maps

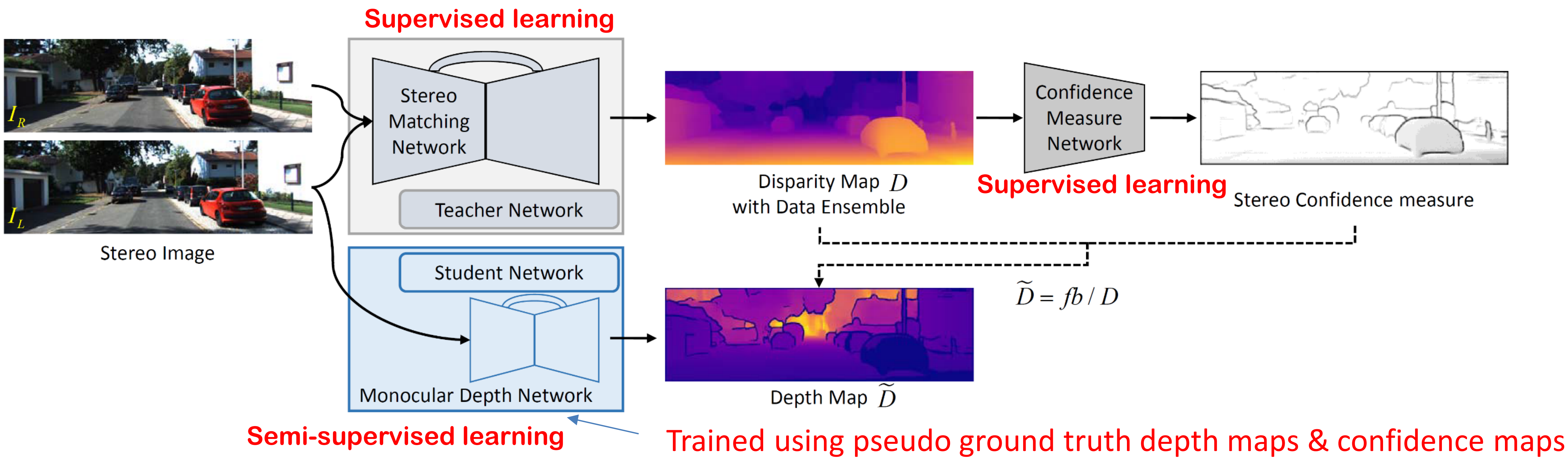


Our Approach using DIML/CVL Dataset

Our method is a semi-supervised approach

- **Teacher network** (stereo matching and confidence measure networks) are trained in a supervised manner, but no massive training data is not needed.

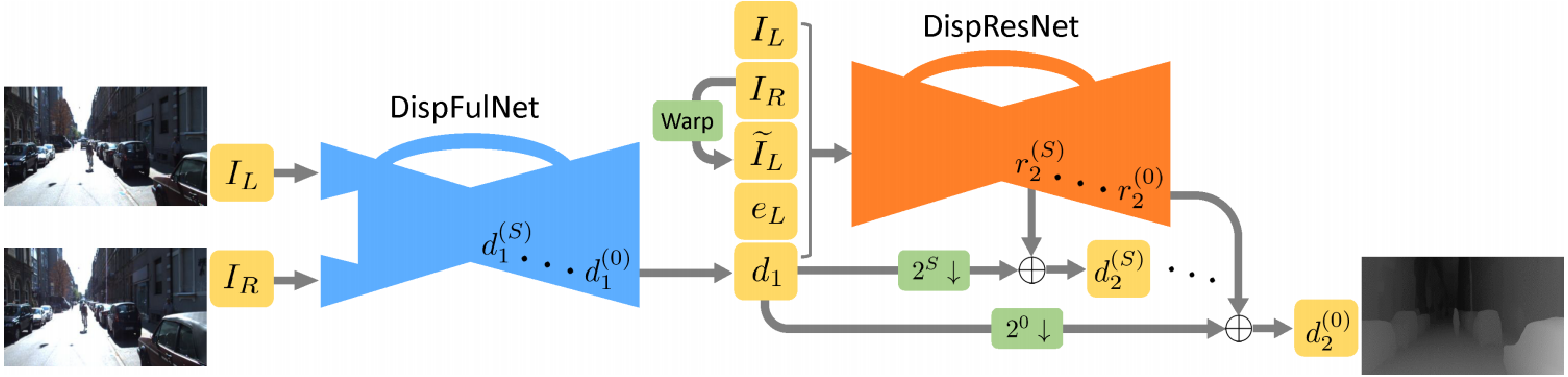
- **Student network** (monocular depth network) are trained using pseudo ground truth depth maps and confidence maps obtained from the teacher network



Our Approach using DIML/CVL Dataset

- **Teacher network: Stereo matching network [32]**

  Note) Any kind of stereo matching approaches can be used here.

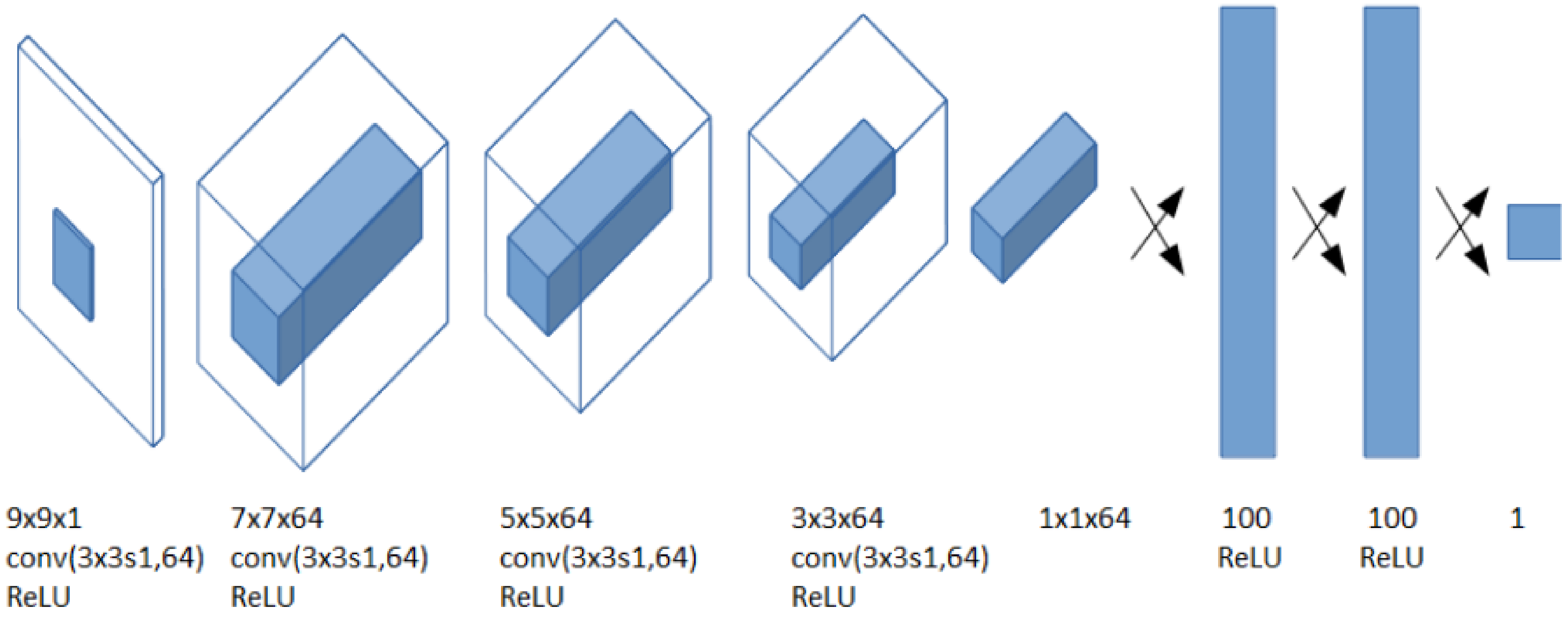


[32] J. Pang, W. Sun, JSJ. Ren, C. Yang, and Q. Yan, "Cascade Residual Learning: A Two-Stage Convolutional Neural Network for Stereo Matching," ICCV 2017

Our Approach using DIML/CVL Dataset

- **Teacher network: Confidence measure network [26]**

  Note) Any kind of stereo confidence approaches can be used here.



[26] Learning from scratch a confidence measure, BMVC 2016

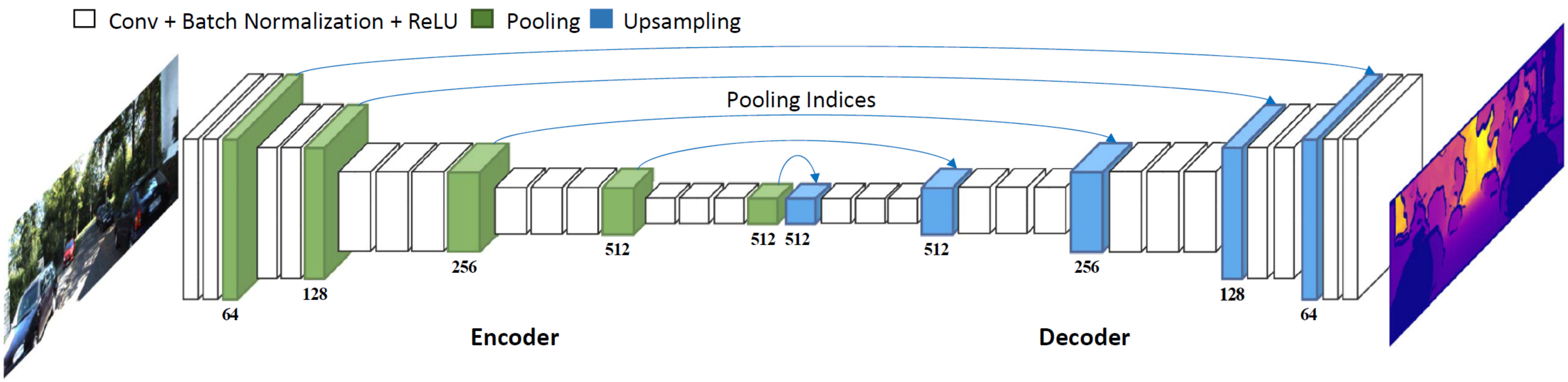
Our Approach using DIML/CVL Dataset

- **Student network: Monocular depth network**

A variant of U-Net architecture [1]

[1] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” MICCAI, 2015.

Note that this baseline architecture is very simple, when compared to state-of-the-arts for monocular depth estimation.
Even with such a simple baseline architecture, our method outperforms the state-of-the-arts.



Analysis: Why is DIML/CVL Dataset Useful?

1. **LiDAR** vs. **Stereo depth map**
   - Easy to solve *the domain adaption problem* with stereo depth maps

2. **Stereo image** vs. **Stereo depth map**

   Unsupervised approach << Semi-supervised approach

   (using stereo image) (using **stereo depth map**)

Analysis: Why is DIML/CVL Dataset Useful?

3. Effect of confidence map

– How much does the confidence map C(p) have on the final performance?

4. Why do we choose a semi-supervised approach?

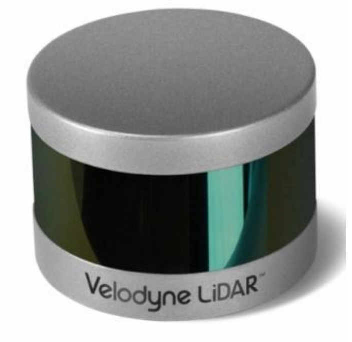
– May the teacher network (stereo matching and confidence measure networks) have the domain adaptation issue?

1. LiDAR vs. Stereo Depth Map: Scene Diversity

Domain adaptation problem

- Diverse scenes must be provided as training data.

KITTI



LiDAR

Sparse resolution

Hard to capture various scenes

→ Does NOT scale well in obtaining massive training data consisting of diverse scenes (**Domain adaptation problem**)

DIML/CVL dataset

Stereo camera

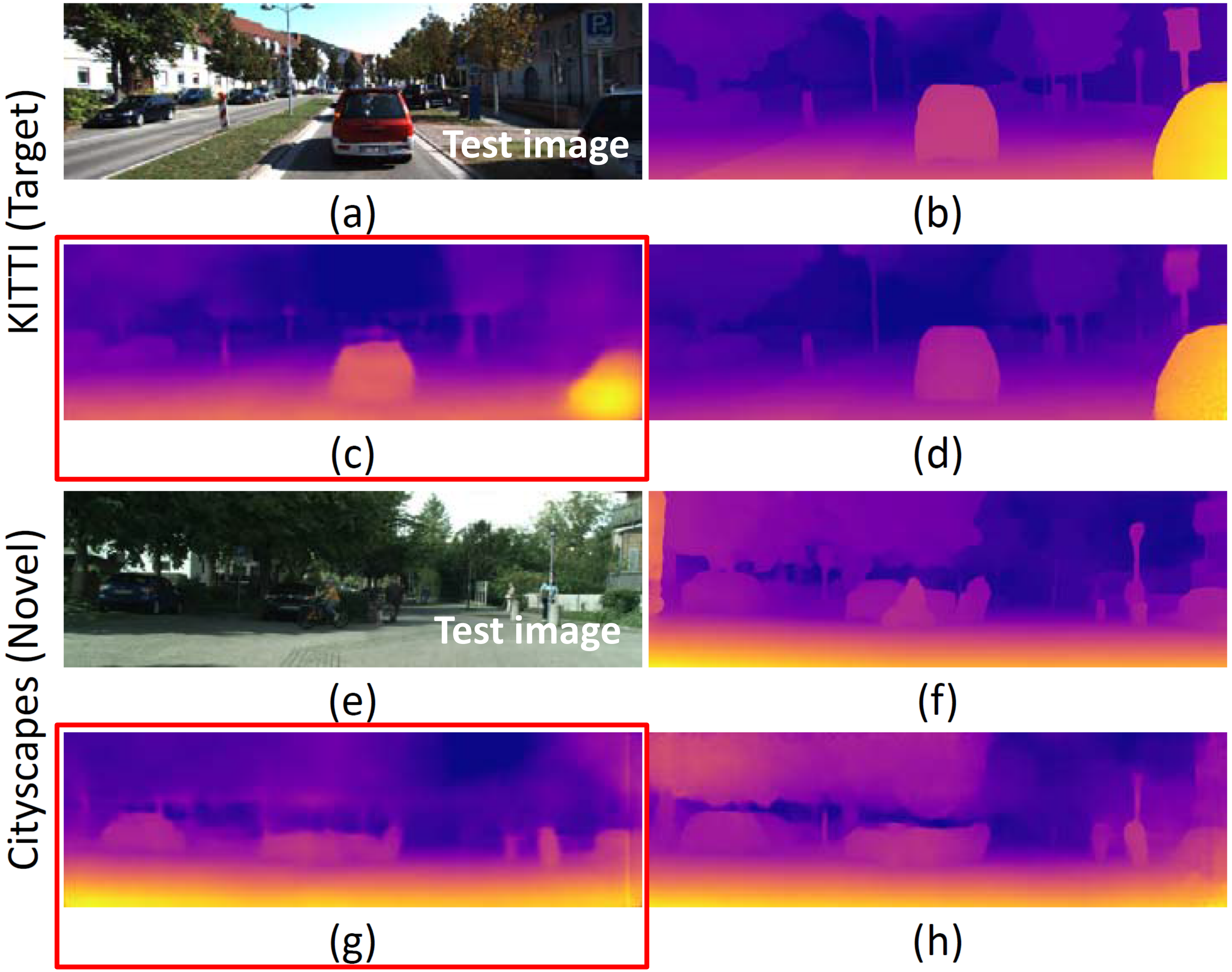
Stereo matching

+

Stereo confidence

Easy to capture various scenes

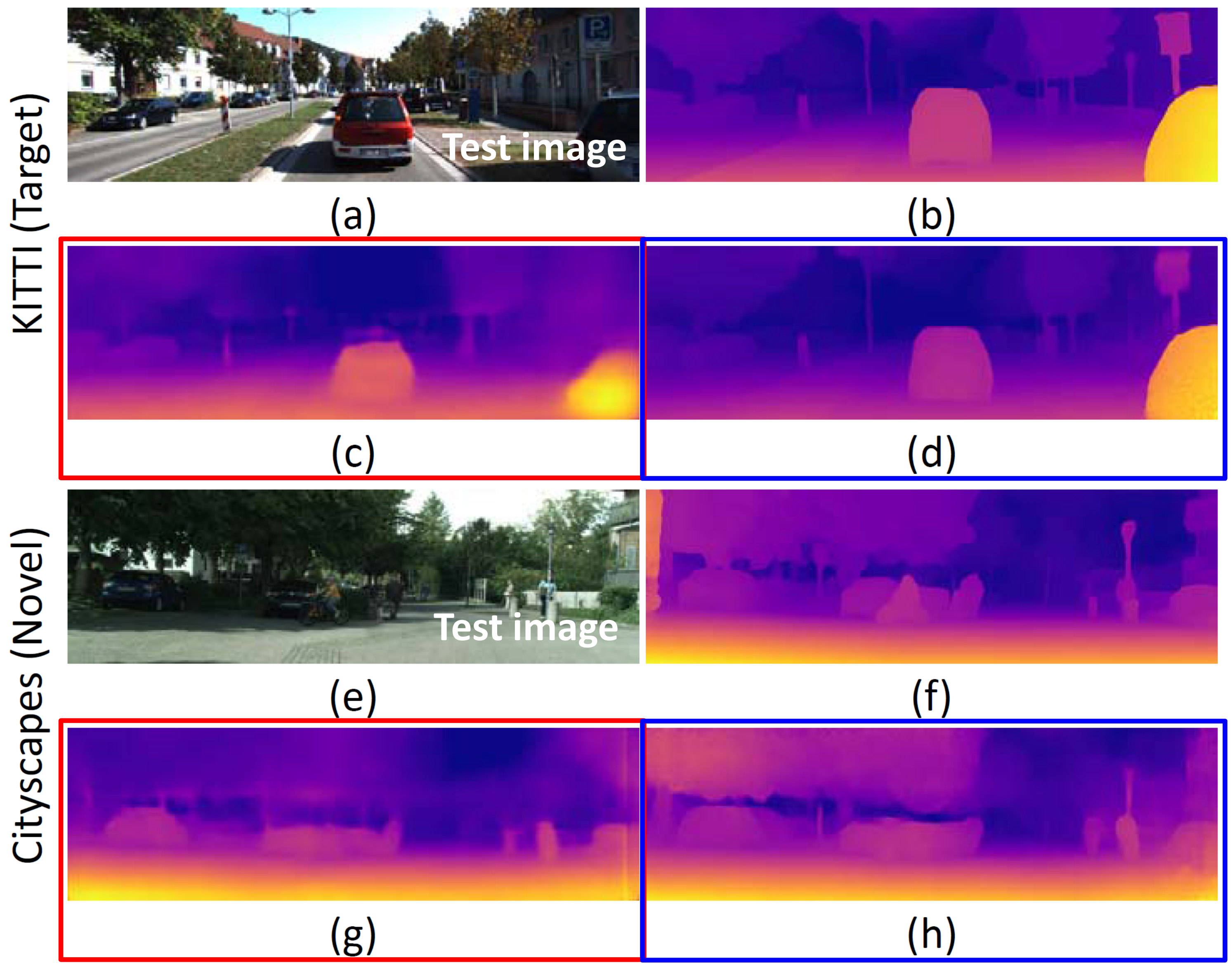
→ Appropriate to obtain massive training data consisting of diverse scenes



(c) and (g) Depth maps of state-of-the-art monocular depth estimation network [22]
Training data: KITTI LiDAR
Test data: KITTI, Cityscape

Though both the KITTI and Cityscapes datasets contain driving scenes, a severe domain adaptation problem occurs.

[22] Y. Kuznietsov, J. Tsai, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," CVPR, 2017.



(c) and (g) Depth maps of state-of-the-art monocular depth estimation network [22]

Training data: KITTI LiDAR

Test data: KITTI, Cityscape

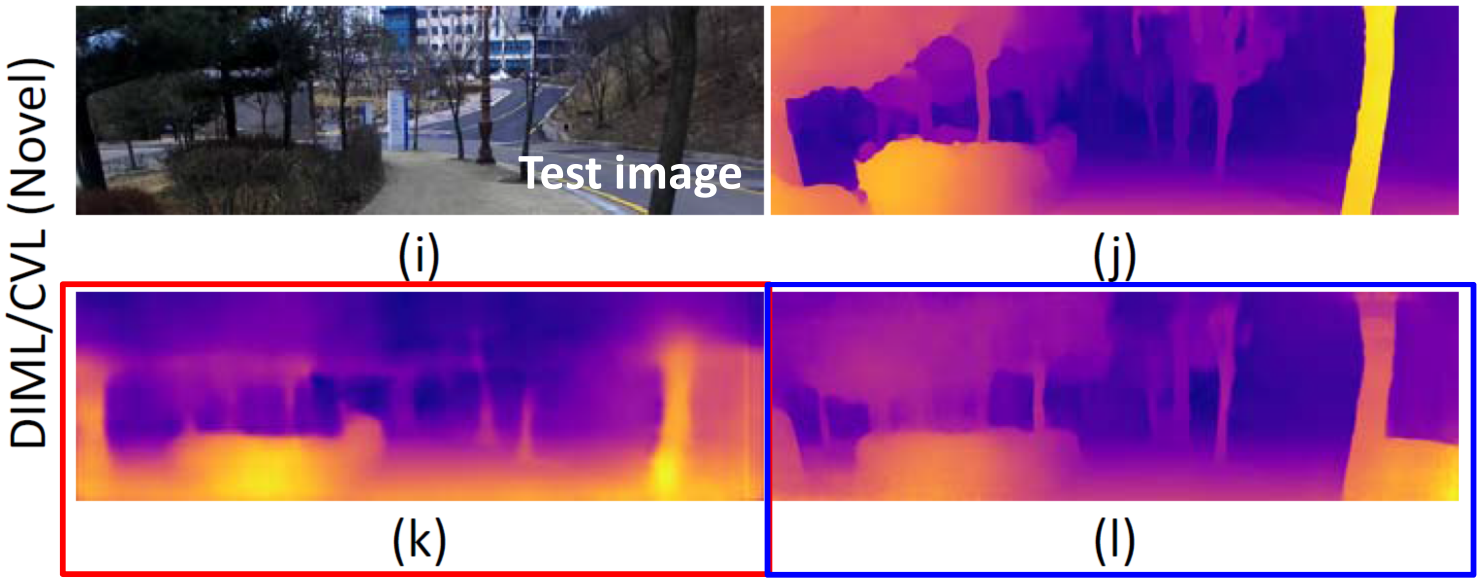
Though both the KITTI and Cityscapes datasets contain driving scenes, a severe domain adaptation problem occurs

[22] Y. Kuznietsov, J. Tsai, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," CVPR, 2017.

(d) and (h) Depth maps of the proposed monocular depth estimation network

Training data: KITTI Stereo + DIML/CVL

Test data: KITTI, Cityscape



(k) Depth maps of state-of-the-art monocular depth estimation network [22]

Training data: KITTI LiDAR

Test data: DIML/CVL

[22] Y. Kuznietsov, J. Tsai, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," CVPR, 2017.

(l) Depth maps of the proposed monocular depth estimation network

Training data: KITTI Stereo + DIML/CVL

Test data: DIML/CVL

Remarks)

1. DIML/CVL dataset is complementary to other datasets.
2. In terms of scene diversity, our strategy to construct massive training data (acquiring stereo images and estimating depth maps) is effective.

For training dataset, K = KITTI, CS = Cityscapes, and Ours = DIML/CVL
Test dataset: Eigen split [17]

[17] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," NIPS, 2014.

Sup.: Supervised approach
Unsup.: Unsupervised approach
Semi-sup.: Semi-supervised approach

1. Our DIML/CVL dataset is complementary to other datasets.
2. In terms of scene diversity, our strategy to construct massive training data (acquiring stereo images and estimating depth maps) is effective.

Method	Training data	Approach	Training Dataset	RMSE(lin)	RMSE(log)	Abs rel	Sqr rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
				Lower is better				Higher is better		
cap 80m										
Eigen *et al.* [17]	Left + LiDAR	*Sup.*	K	7.156	0.270	0.215	1.515	0.692	0.899	0.967
Godard *et al.* [21]	Stereo	*Unsup.*	K	5.927	0.247	0.148	1.344	0.803	0.922	0.964
Godard *et al.* + ***pp*** [21]	Stereo	*UnSup.*	K + CS	4.935	0.206	0.114	0.898	0.861	0.949	0.976
Kuznietsov *et al.* [22]	Left + LiDAR	*Sup.*	K	4.815	0.194	0.122	0.763	0.845	0.957	**0.987**
Kuznietsov *et al.* [22]	Stereo + LiDAR	*Semi-sup*	K	4.621	0.189	0.113	0.741	0.862	0.960	0.986
Luo *et al.* [20]	(Sythetic) Stereo + GT	*Sup.*	K	4.681	0.200	0.102	0.700	0.872	0.954	0.978
Our Method	Left + Pseudo GT	*Semi-sup*	K	4.599	0.183	0.099	0.748	0.880	0.959	0.983
Our Method	Left + Pseudo GT	*Semi-sup*	K + Ours	4.333	0.181	0.098	0.644	0.881	0.963	0.984
Our Method	Left + Pseudo GT	*Semi-sup*	K + CS	4.286	0.177	0.097	0.641	0.882	0.963	0.984
Our Method	Left + Pseudo GT	*Semi-sup*	K + CS + Ours	**4.129**	**0.175**	**0.095**	**0.613**	**0.884**	**0.964**	0.986
cap 50m										
Garg *et al.* [33]	Stereo	*Unsup.*	K	5.104	0.273	0.169	1.080	0.740	0.904	0.962
Godard *et al.* [21]	Stereo	*Unsup.*	K	4.471	0.232	0.140	0.976	0.818	0.931	0.969
Godard *et al.* + ***pp*** [21]	Stereo	*Unsup.*	K + CS	3.729	0.194	0.108	0.657	0.873	0.954	0.979
Kuznietsov *et al.* [22]	Stereo + LiDAR	*Semi-sup*	K	3.518	0.179	0.108	0.595	0.875	0.964	**0.988**
Luo *et al.* [20]	(Sythetic) Stereo + GT	*Sup.*	K	3.503	0.187	0.097	0.539	0.885	0.960	0.981
Our Method	Left + Pseudo GT	*Semi-sup*	K + CS + Ours	**3.162**	**0.162**	**0.091**	**0.505**	**0.901**	**0.969**	0.986

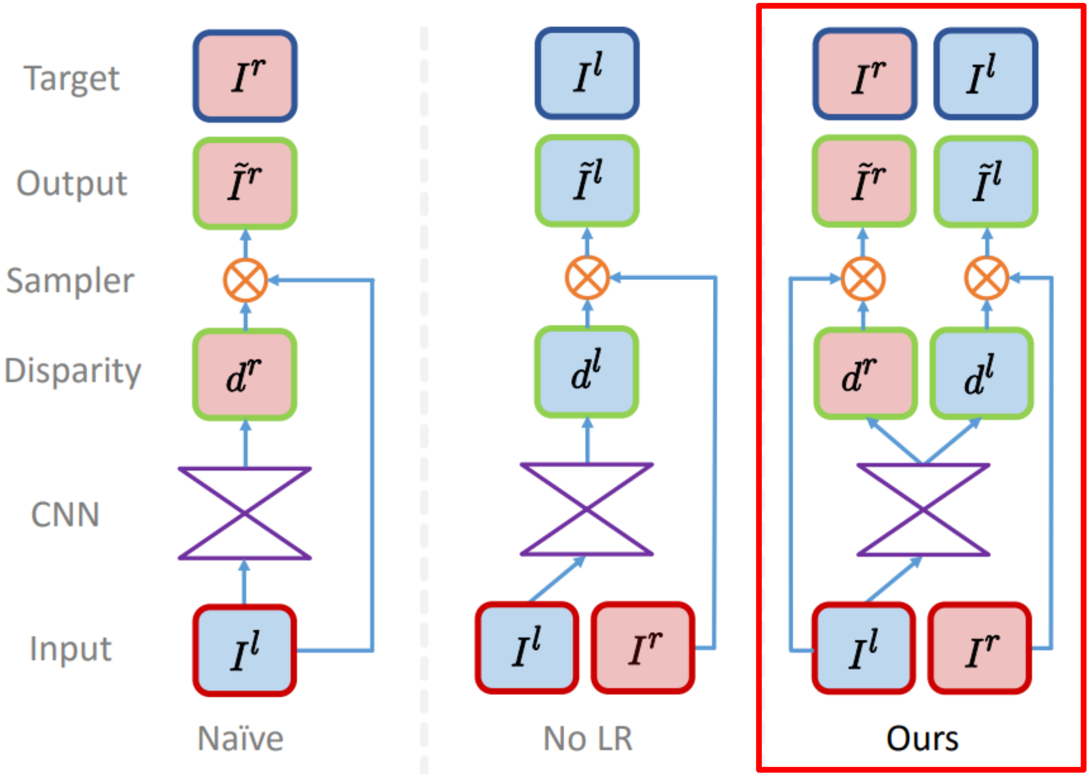
2. Stereo Depth Map vs. Stereo Image

- Unsupervised approach using stereo images [21]
  - Uses stereo images to address the lack of massive training data
  - Proposes an unsupervised reconstruction loss

Left image -> Right image

Left image -> Left image

Left image -> Right image
Right image -> Left image

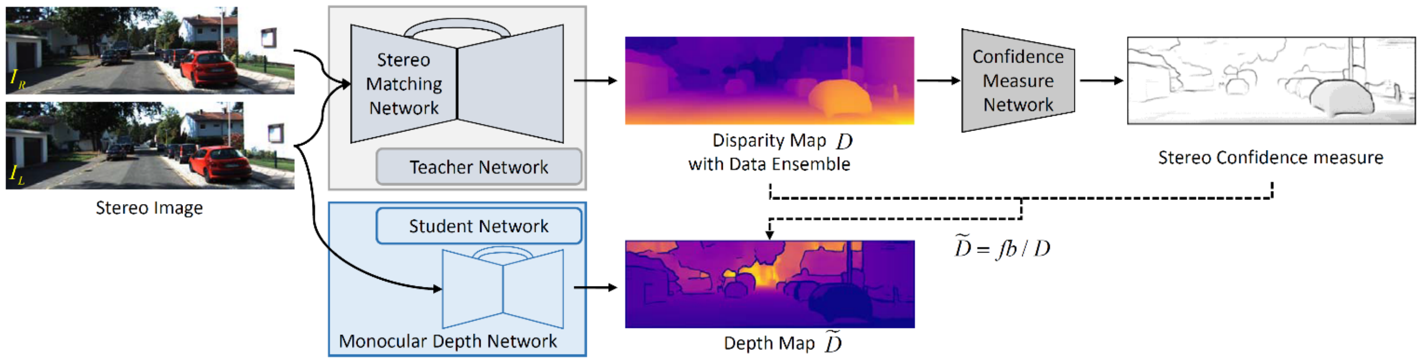


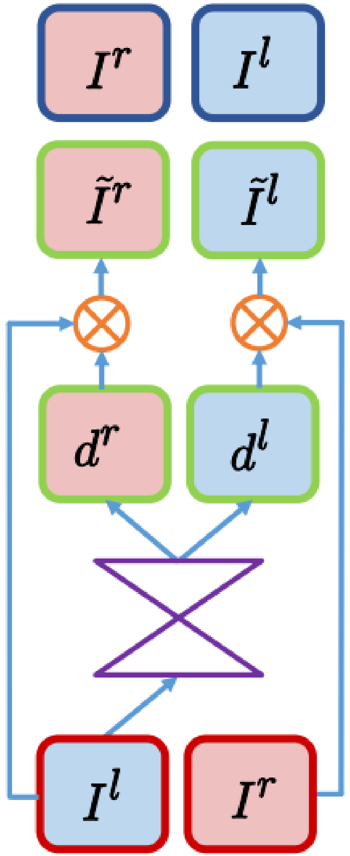
Several ways to use stereo images for training the monocular depth network

[21] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," CVPR, 2016.

2. Stereo Depth Map vs. Stereo Image

Our semi-supervised approach using stereo depth maps **vs.** **Unsupervised approach using stereo images [21]**





Training
Left image -> Right image
Right image -> Left image

Test
Left image -> Left depth map

Teacher network: RGB+D data generation
Training & Test: Left & Right image -> Left depth map & Confidence map

Student network
Training: Left image -> Left depth map (assisted by confidence map)
Test: Left image -> Left depth map

For training dataset, K = KITTI, CS = Cityscapes, and Ours = DIML/CVL
Test dataset: Eigen split [17]

[20] Single view stereo matching, CVPR, 2018.
[21] Unsupervised monocular depth estimation with left-right consistency, CVPR, 2016.
[22] Semi-supervised deep learning for monocular depth map prediction, CVPR, 2017.

Sup.: Supervised approach
Unsup.: Unsupervised approach
Semi-sup.: Semi-supervised approach

1. Our strategy to construct massive training data (acquiring stereo images and estimating depth maps) is effective, when compared to the unsupervised approach.

Method	Training data	Approach	Training Dataset	RMSE(lin)	RMSE(log)	Abs rel	Sqr rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
				Lower is better				Higher is better		
cap 80m										
Eigen *et al.* [17]	Left + LiDAR	*Sup.*	K	7.156	0.270	0.215	1.515	0.692	0.899	0.967
Godard *et al.* [21]	Stereo	*Unsup.*	K	5.927	0.247	0.148	1.344	0.803	0.922	0.964
Godard *et al.* + ***pp*** [21]	Stereo	*UnSup.*	K + CS	4.935	0.206	0.114	0.898	0.861	0.949	0.976
Kuznietsov *et al.* [22]	Left + LiDAR	*Sup.*	K	4.815	0.194	0.122	0.763	0.845	0.957	**0.987**
Kuznietsov *et al.* [22]	Stereo + LiDAR	*Semi-sup*	K	4.621	0.189	0.113	0.741	0.862	0.960	0.986
Luo *et al.* [20]	(Sythetic) Stereo + GT	*Sup.*	K	4.681	0.200	0.102	0.700	0.872	0.954	0.978
Our Method	Left + Pseudo GT	*Semi-sup*	K	4.599	0.183	0.099	0.748	0.880	0.959	0.983
Our Method	Left + Pseudo GT	*Semi-sup*	K + Ours	4.333	0.181	0.098	0.644	0.881	0.963	0.984
Our Method	Left + Pseudo GT	*Semi-sup*	K + CS	4.286	0.177	0.097	0.641	0.882	0.963	0.984
Our Method	Left + Pseudo GT	*Semi-sup*	K + CS + Ours	**4.129**	**0.175**	**0.095**	**0.613**	**0.884**	**0.964**	0.986
cap 50m										
Garg *et al.* [33]	Stereo	*Unsup.*	K	5.104	0.273	0.169	1.080	0.740	0.904	0.962
Godard *et al.* [21]	Stereo	*Unsup.*	K	4.471	0.232	0.140	0.976	0.818	0.931	0.969
Godard *et al.* + ***pp*** [21]	Stereo	*Unsup.*	K + CS	3.729	0.194	0.108	0.657	0.873	0.954	0.979
Kuznietsov *et al.* [22]	Stereo + LiDAR	*Semi-sup*	K	3.518	0.179	0.108	0.595	0.875	0.964	**0.988**
Luo *et al.* [20]	(Sythetic) Stereo + GT	*Sup.*	K	3.503	0.187	0.097	0.539	0.885	0.960	0.981
Our Method	Left + Pseudo GT	*Semi-sup*	K + CS + Ours	**3.162**	**0.162**	**0.091**	**0.505**	**0.901**	**0.969**	0.986

For training dataset, K = KITTI, CS = Cityscapes, and Ours = DIML/CVL
Test dataset: Eigen split [17]

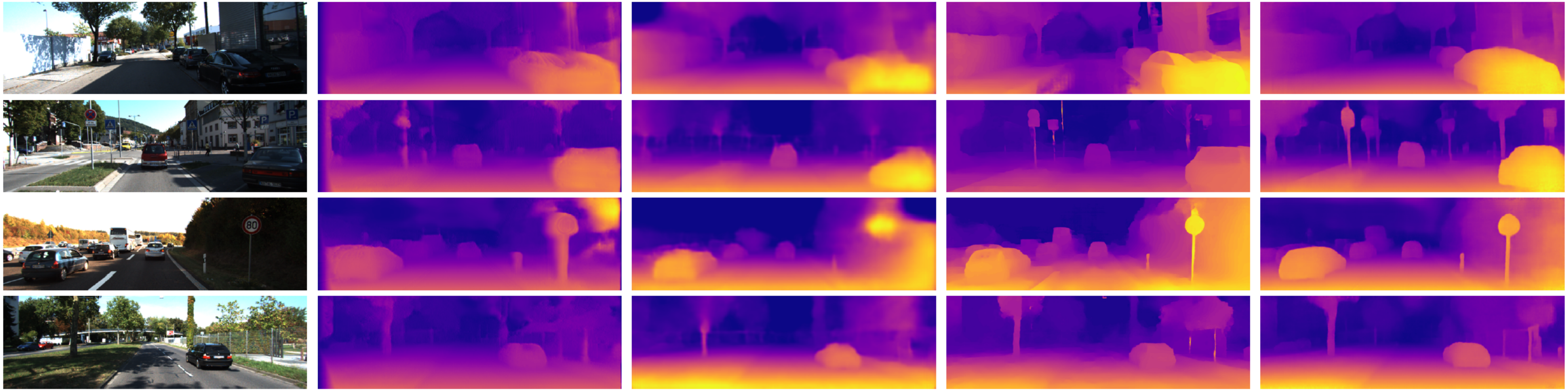
[20] Single view stereo matching, CVPR, 2018.
[21] Unsupervised monocular depth estimation with left-right consistency, CVPR, 2016.
[22] Semi-supervised deep learning for monocular depth map prediction, CVPR, 2017.

Sup.: Supervised approach
Unsup.: Unsupervised approach
Semi-sup.: Semi-supervised approach

1. Our approach outperforms state-of-the-arts.

Method	Training data	Approach	Training Dataset	RMSE(lin)	RMSE(log)	Abs rel	Sqr rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
				Lower is better				Higher is better		
cap 80m										
Eigen *et al.* [17]	Left + LiDAR	*Sup.*	K	7.156	0.270	0.215	1.515	0.692	0.899	0.967
Godard *et al.* [21]	Stereo	*Unsup.*	K	5.927	0.247	0.148	1.344	0.803	0.922	0.964
Godard *et al.* + ***pp*** [21]	Stereo	*UnSup.*	K + CS	4.935	0.206	0.114	0.898	0.861	0.949	0.976
Kuznietsov *et al.* [22]	Left + LiDAR	*Sup.*	K	4.815	0.194	0.122	0.763	0.845	0.957	**0.987**
Kuznietsov *et al.* [22]	Stereo + LiDAR	*Semi-sup*	K	4.621	0.189	0.113	0.741	0.862	0.960	0.986
Luo *et al.* [20]	(Sythetic) Stereo + GT	*Sup.*	K	4.681	0.200	0.102	0.700	0.872	0.954	0.978
Our Method	Left + Pseudo GT	*Semi-sup*	K	4.599	0.183	0.099	0.748	0.880	0.959	0.983
Our Method	Left + Pseudo GT	*Semi-sup*	K + Ours	4.333	0.181	0.098	0.644	0.881	0.963	0.984
Our Method	Left + Pseudo GT	*Semi-sup*	K + CS	4.286	0.177	0.097	0.641	0.882	0.963	0.984
Our Method	Left + Pseudo GT	*Semi-sup*	K + CS + Ours	**4.129**	**0.175**	**0.095**	**0.613**	**0.884**	**0.964**	0.986
cap 50m										
Garg *et al.* [33]	Stereo	*Unsup.*	K	5.104	0.273	0.169	1.080	0.740	0.904	0.962
Godard *et al.* [21]	Stereo	*Unsup.*	K	4.471	0.232	0.140	0.976	0.818	0.931	0.969
Godard *et al.* + ***pp*** [21]	Stereo	*Unsup.*	K + CS	3.729	0.194	0.108	0.657	0.873	0.954	0.979
Kuznietsov *et al.* [22]	Stereo + LiDAR	*Semi-sup*	K	3.518	0.179	0.108	0.595	0.875	0.964	**0.988**
Luo *et al.* [20]	(Sythetic) Stereo + GT	*Sup.*	K	3.503	0.187	0.097	0.539	0.885	0.960	0.981
Our Method	Left + Pseudo GT	*Semi-sup*	K + CS + Ours	**3.162**	**0.162**	**0.091**	**0.505**	**0.901**	**0.969**	0.986

Visual Comparison



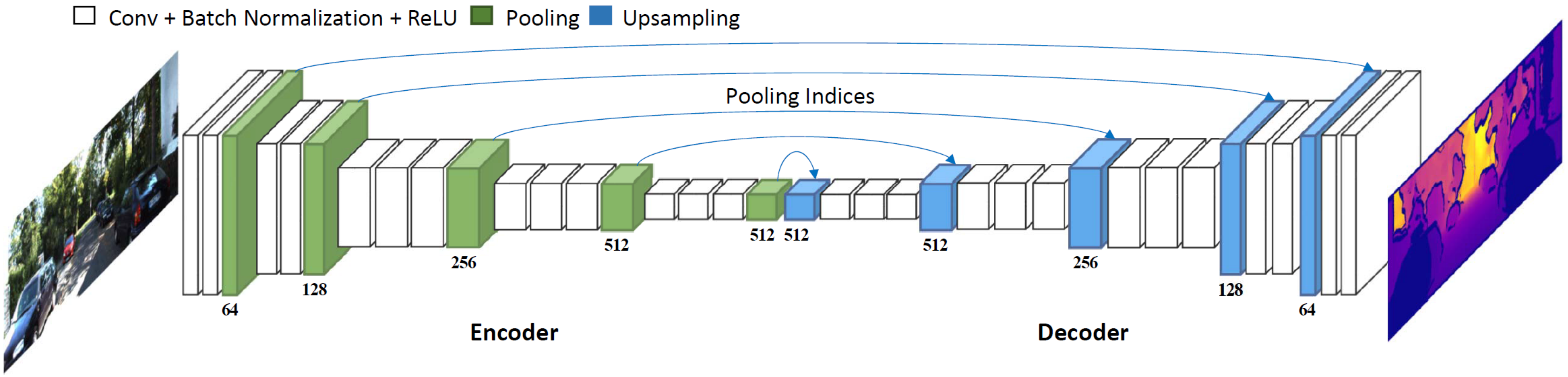
(a) (b) (c) (d) (e)

(b) Unsupervised approach [21] trained with stereo image pairs of the KITTI + Cityscapes
(c) Kuznietsov et al. [22] trained with stereo image pairs and ground truth depth map of KITTI
(d) Luo et al. [20] trained with left image and ground truth depth map of Flying Things synthetic dataset [9]
(e) the proposed method trained with KITTI + Cityscapes + DIML/CVL dataset.

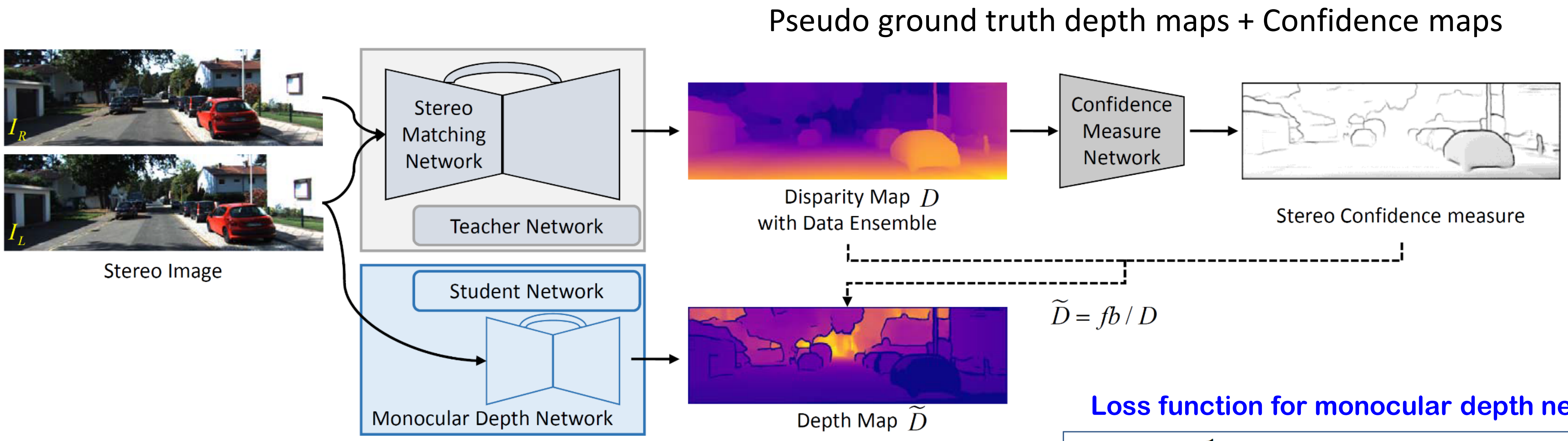
[20] Single view stereo matching, CVPR, 2018.
[21] Unsupervised monocular depth estimation with left-right consistency, CVPR, 2016.
[22] Semi-supervised deep learning for monocular depth map prediction, CVPR, 2017.

Recap: Our Monocular Depth Network is Just a Simple Encoder-decoder!

- Even with such a simple baseline architecture using an encoder-decoder, we achieve outstanding performance.
- It is expected that using more sophisticate networks produces more accurate depth maps.



3. Effect of Confidence Map



Loss function for monocular depth network

$$\mathcal{L}_c = \frac{1}{\sum_p M_p} \sum_p M_p \cdot \left| \hat{D}(p) - \tilde{D}(p) \right|_1,$$

$$M_p = \begin{cases} 1, & \text{if } C(p) \geq \tau \\ 0, & \text{if } C(p) < \tau \end{cases}.$$

$C(p)$: confidence at pixel p

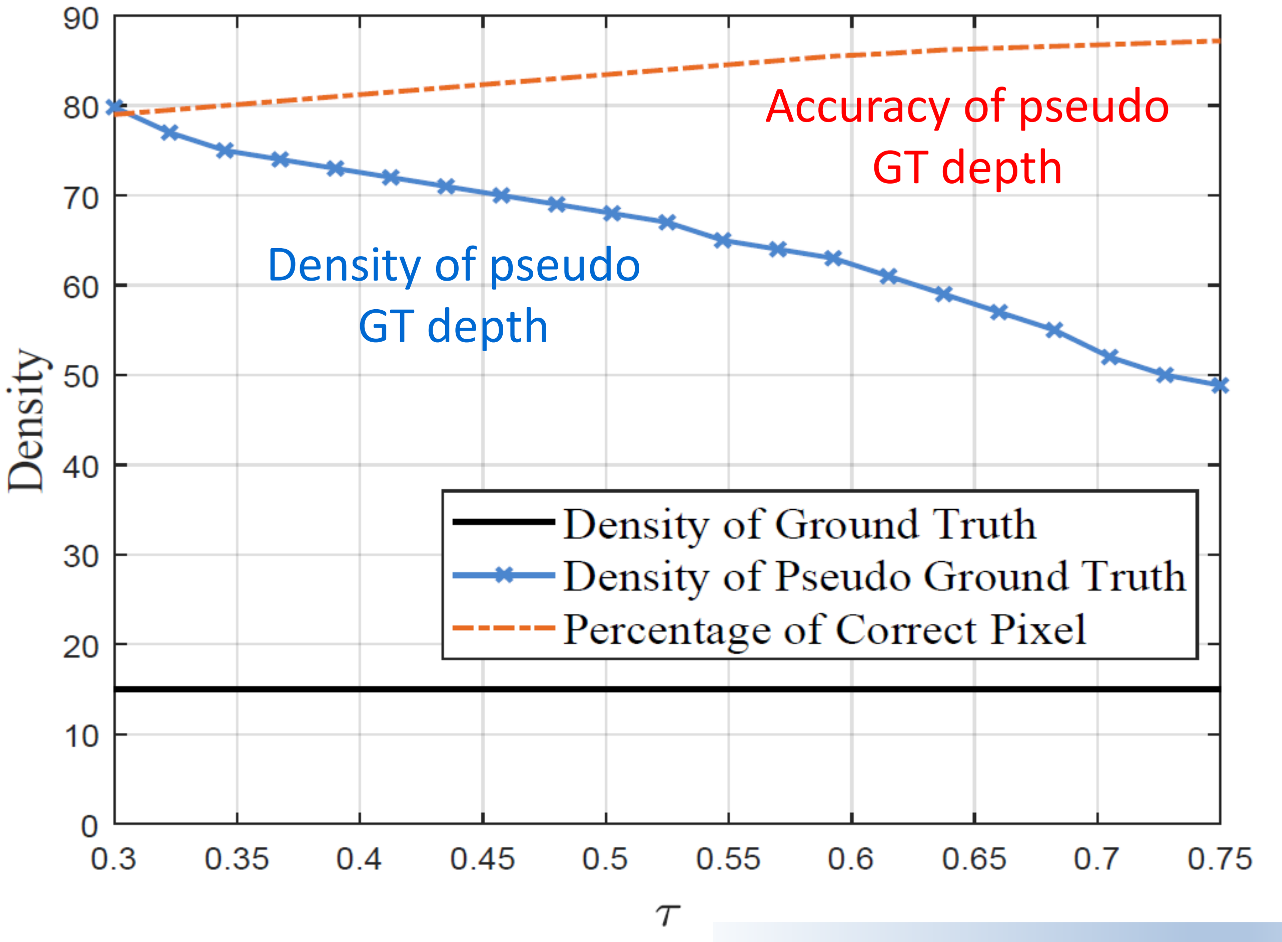
τ: Confidence threshold ($0 \leq \tau \leq 1$)

Question

1. How much does the confidence map C(p) have on the final performance?
2. Only confident depth values are used. What is the best way to set the confidence threshold?

이화여자대학교
EWHA WOMANS UNIVERSITY

3. Effect of Confidence Map



Trade-off between Density vs. Accuracy in pseudo GT depth maps

Loss function for monocular depth network

$$\mathcal{L}_c = \frac{1}{\sum_p M_p} \sum_p M_p \cdot \left| \hat{D}(p) - \tilde{D}(p) \right|_1,$$

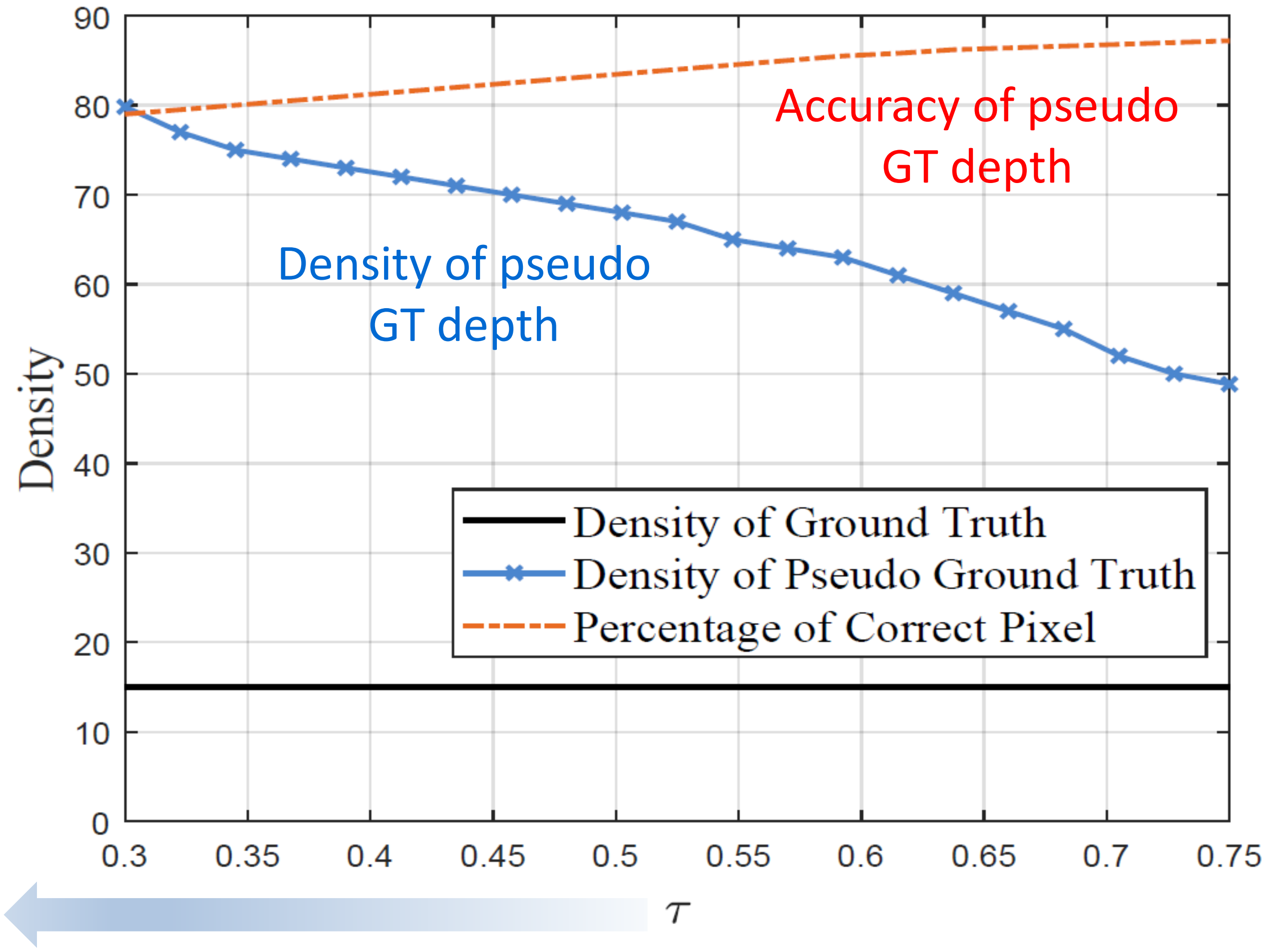
$$M_p = \begin{cases} 1, & \text{if } C(p) \geq \tau \\ 0, & \text{if } C(p) < \tau \end{cases}.$$

$C(p)$: confidence at pixel p

τ: Confidence threshold ($0 \leq \tau \leq 1$)

The higher τ, the better the accuracy of pseudo ground truth depth maps. However, this reduces the density of the depth maps.

3. Effect of Confidence Map



Trade-off between Density vs. Accuracy in pseudo GT depth maps

Loss function for monocular depth network

$$\mathcal{L}_c = \frac{1}{\sum_p M_p} \sum_p M_p \cdot \left| \hat{D}(p) - \tilde{D}(p) \right|_1,$$

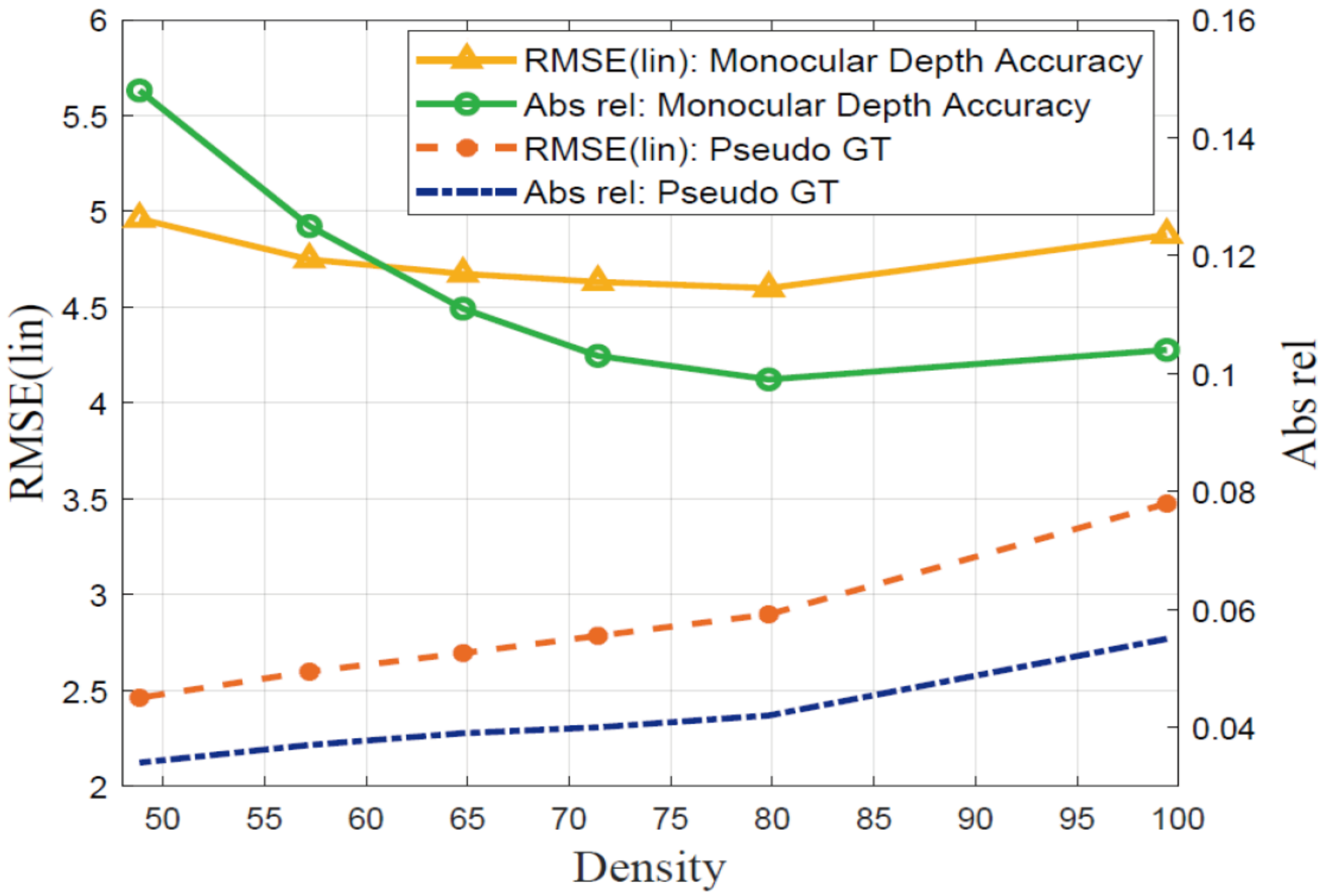
$$M_p = \begin{cases} 1, & \text{if } C(p) \geq \tau \\ 0, & \text{if } C(p) < \tau \end{cases}.$$

$C(p)$: confidence at pixel p

τ: Confidence threshold ($0 \leq \tau \leq 1$)

The lower τ, the higher the density of pseudo ground truth depth maps.
However, this decreases the accuracy of the depth maps.

3. Effect of Confidence Map

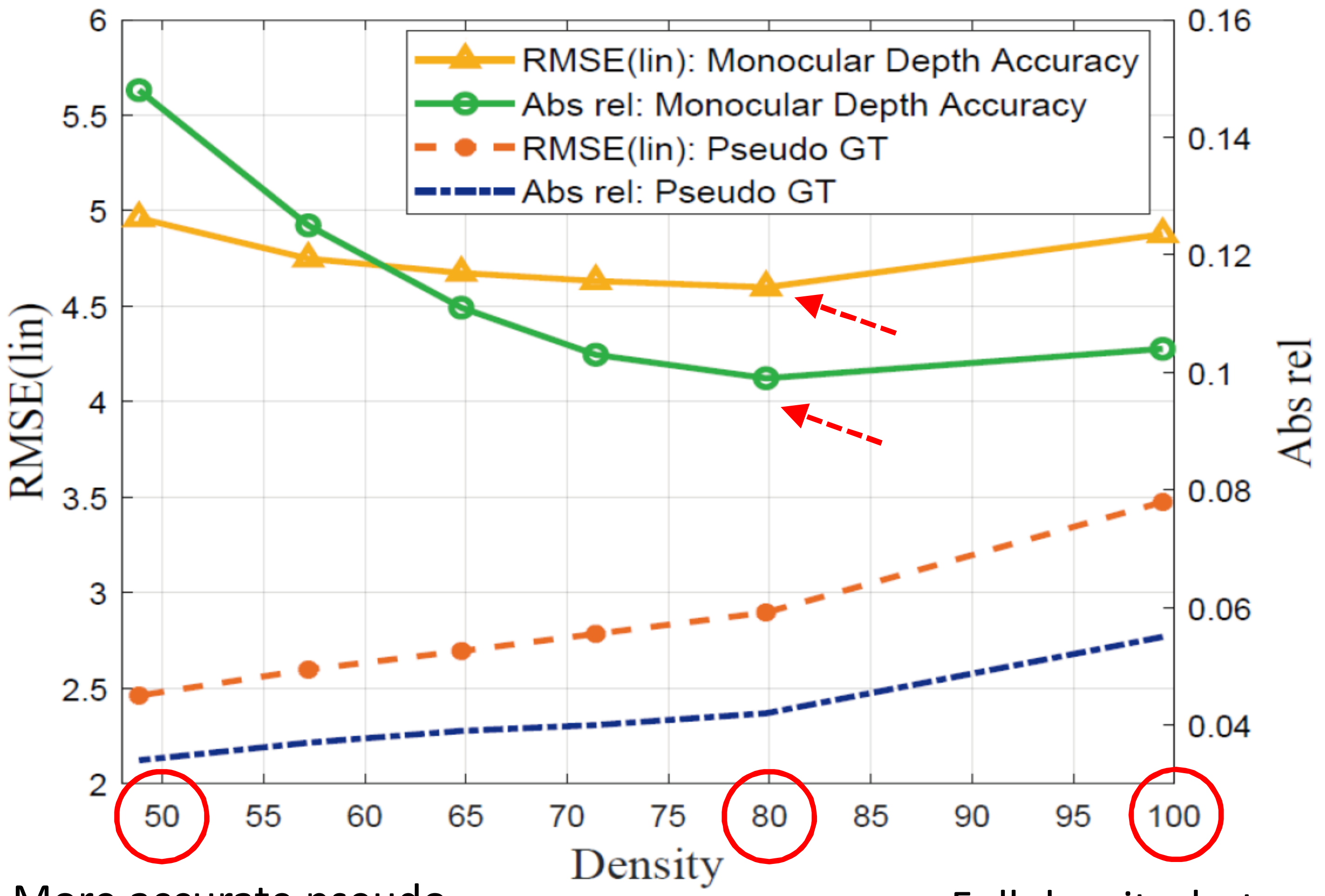


Accuracy of the monocular depth network w.r.t. the density of the pseudo GT depth maps

Accuracy of the monocular depth network

Accuracy of Pseudo Ground Truth (GT) depth maps

3. Effect of Confidence Map



More accurate pseudo GT depth maps yet with too low density

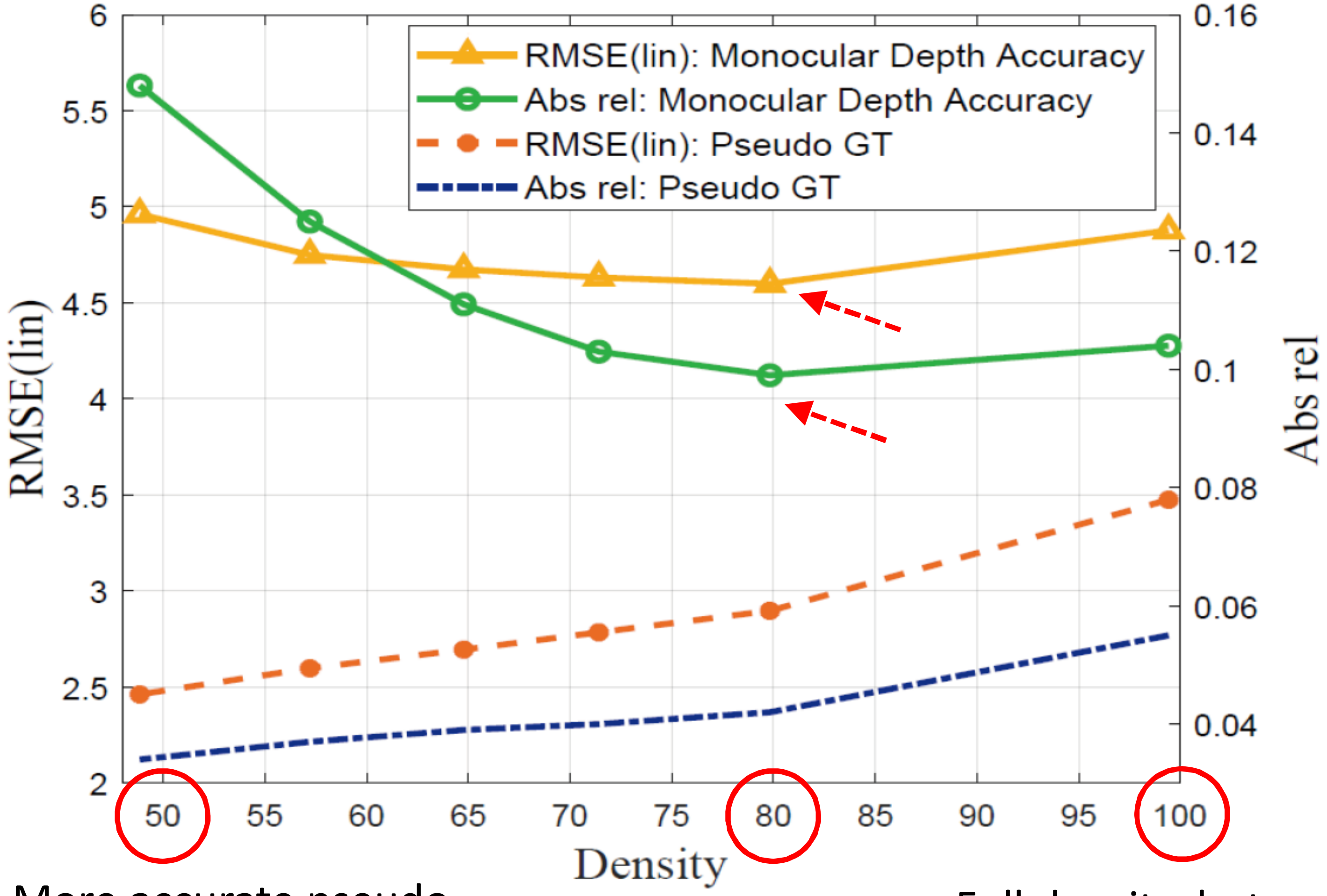
Full density, but pseudo GT depth maps become worse. (Not using confidence map)

Accuracy of the monocular depth network w.r.t. the density of the pseudo GT depth maps

Accuracy of the monocular depth network

Accuracy of Pseudo Ground Truth (GT) depth maps

3. Effect of Confidence Map



More accurate pseudo GT depth maps yet with too low density

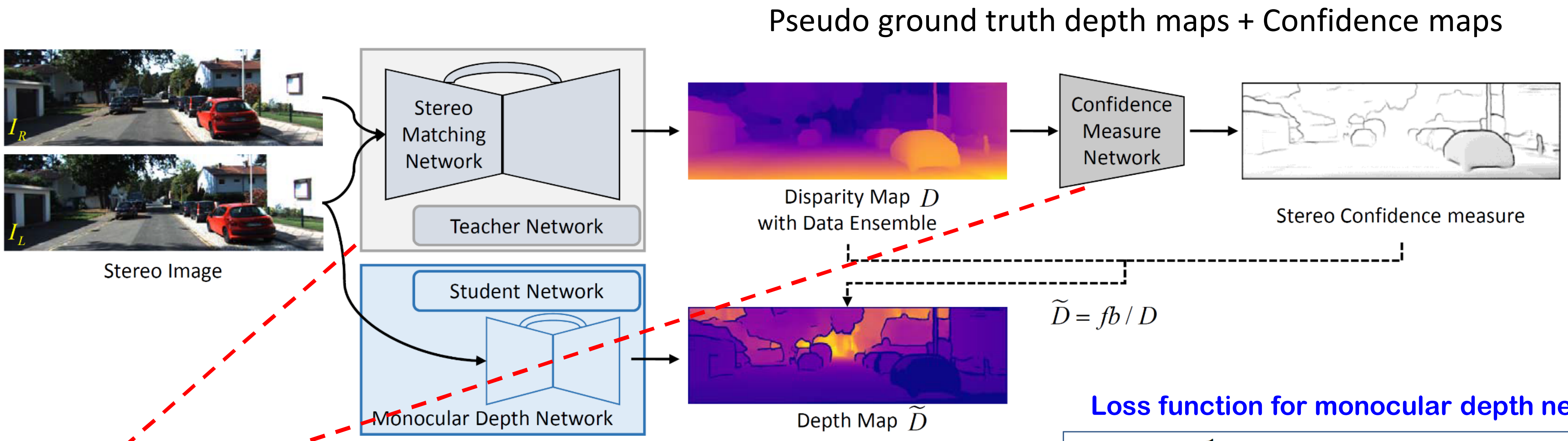
Full density, but pseudo GT depth maps become worse. (Not using confidence map)

Accuracy of the monocular depth network w.r.t. the density of the pseudo GT depth maps

Conclusion

1. The monocular depth network achieves the best accuracy when the density is about 80% (τ =0.3)
2. More accurate pseudo GT depth maps do NOT necessarily lead to better training for monocular depth network. Density also matters.

4. Why do we choose a semi-supervised approach?



Loss function for monocular depth network

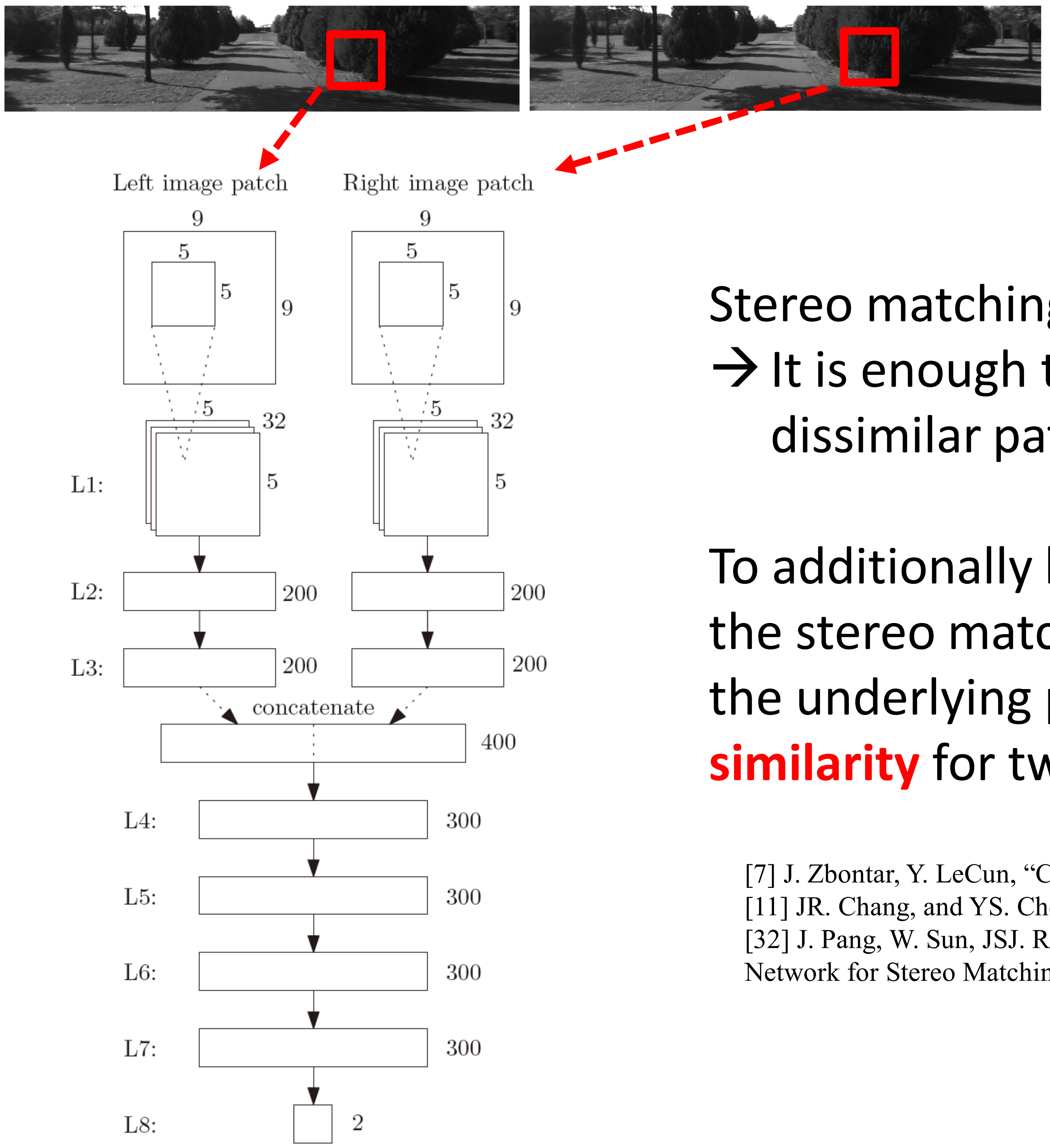
$$\mathcal{L}_c = \frac{1}{\sum_p M_p} \sum_p M_p \cdot \left| \hat{D}(p) - \tilde{D}(p) \right|_1,$$

$$M_p = \begin{cases} 1, & \text{if } C(p) \geq \tau \\ 0, & \text{if } C(p) < \tau \end{cases}.$$

Stereo matching network & confidence estimation network are trained using training data in a supervised manner.

However,

- Smaller training data is needed
- Less sensitive to the domain adaption problem



Our observation:
The stereo matching network is relatively free from the domain adaption problem.

Stereo matching aims to find **similar patches**.
→ It is enough to train the network with similar patches and dissimilar patches [7].

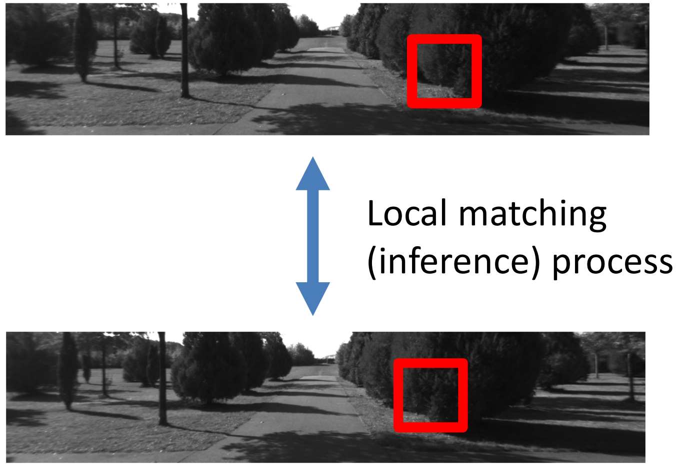
To additionally leverage a global context, some methods train the stereo matching network using two images at once [11, 32], the underlying principle is to locally explore the **patch-level similarity** for two-view matching.

[7] J. Zbontar, Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," CVPR, 2015.
[11] JR. Chang, and YS. Chen, "Pyramid stereo matching network," CVPR, 2018.
[32] J. Pang, W. Sun, JSJ. Ren, C. Yang, and Q. Yan, "Cascade Residual Learning: A Two-Stage Convolutional Neural Network for Stereo Matching," CVPR, 2017.

The stereo matching network is relatively free from the domain adaption problem, while the monocular depth estimation network often suffers from it.

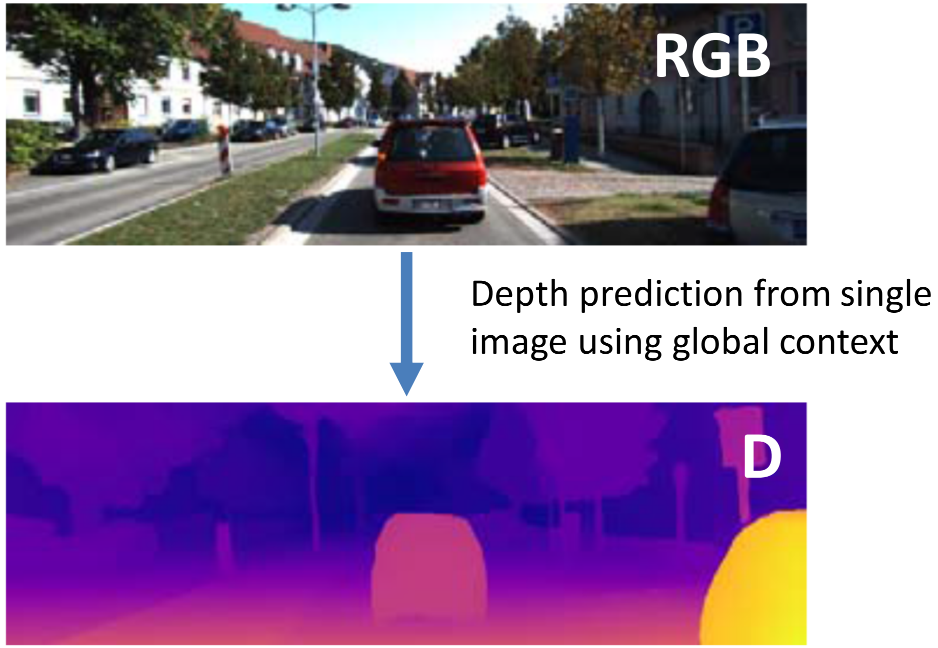
Stereo matching

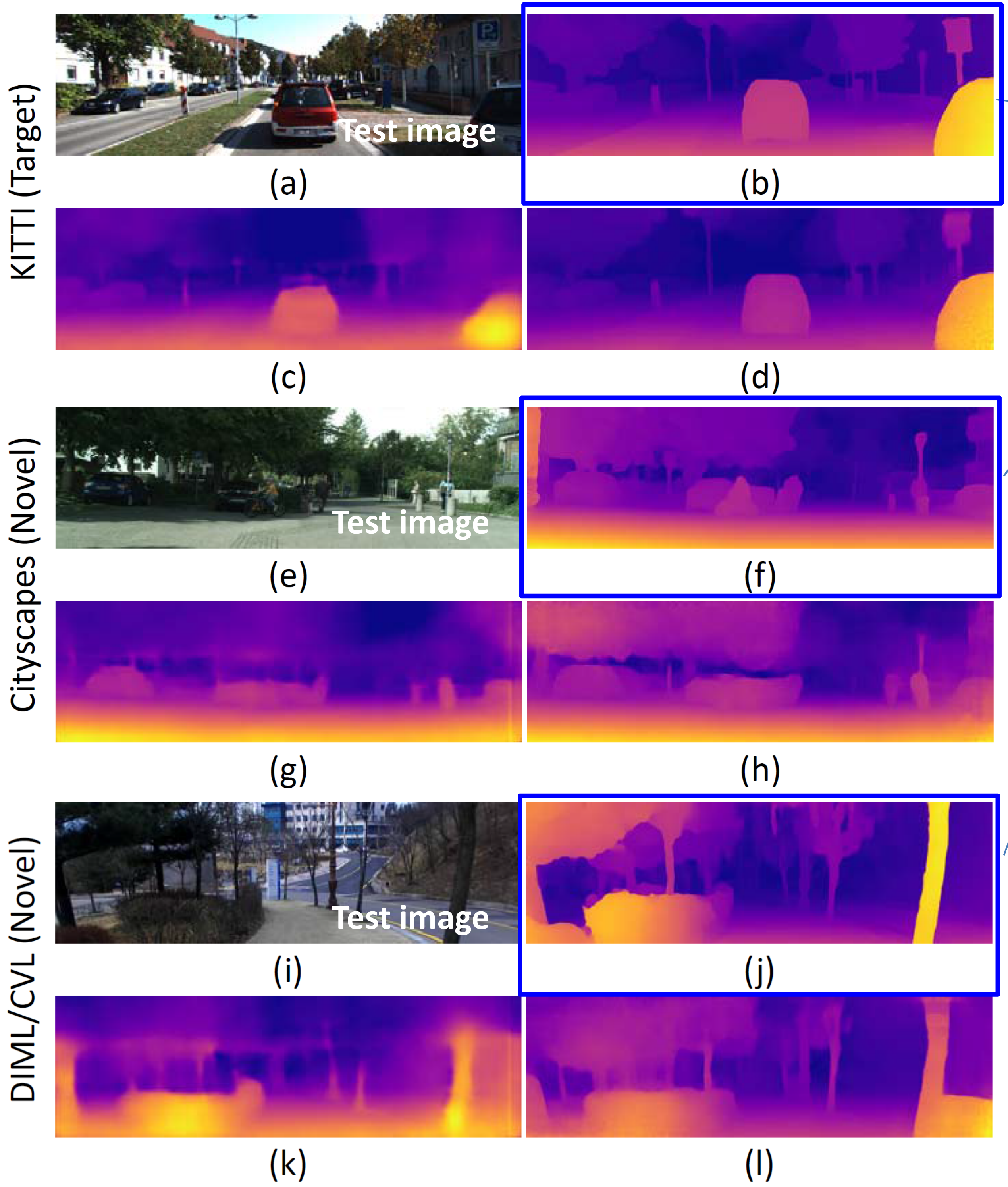
- Finding a pair of similar patches is a *local* inference process.



Monocular depth estimation

- Estimating an overall 3D layout requires seeing an entire image.
- *Global* context does matters.





(b) (f) (j) Depth maps of deep stereo matching network [32],
Training data: KITTI LiDAR
Test data: KITTI, Cityscape, DIML/CVL

The stereo matching network is relatively free from the domain adaption problem.

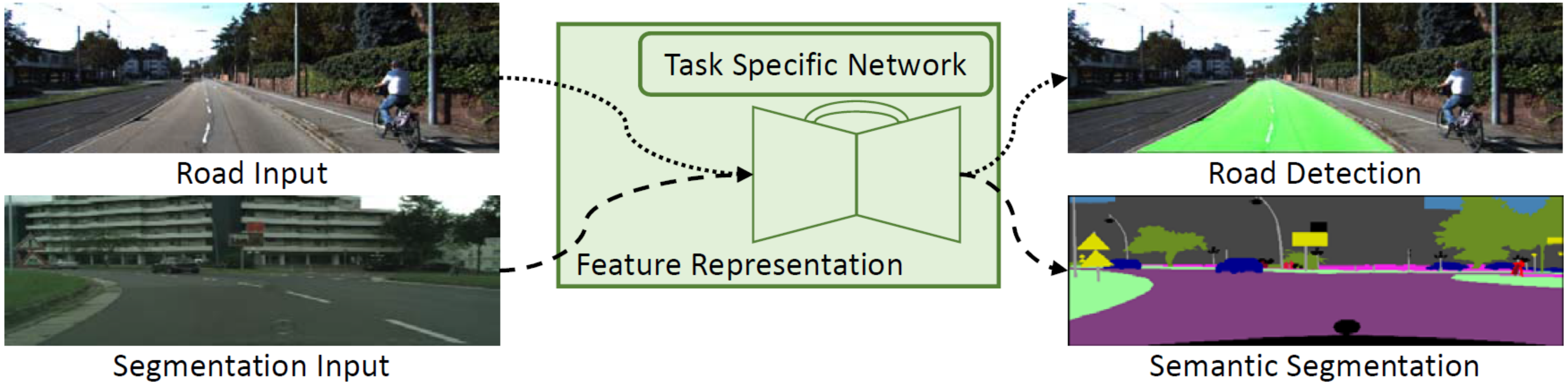
[32] J. Pang, W. Sun, JSJ. Ren, C. Yang, and Q. Yan, "Cascade Residual Learning: A Two-Stage Convolutional Neural Network for Stereo Matching," CVPR, 2017.

Similarly, the confidence measure network is less sensitive to the domain adaptation problem,
as it is trained with a pair of patches.

Transfer Learning using Monocular Depth Network

- **Pre-trained model for road detection and semantic segmentation**

Our pre-trained monocular depth network can be transferred as a pretext task for training road detection and semantic segmentation



Transfer Learning using Monocular Depth Network

- Semantic segmentation

Semantic Segmentation		
Initialization	Pretext	mean IoU
Scratch	-	52.27
ImageNet pre-trained model [47]	Classification	66.27
K	Depth	62.82
K + Ours	Depth	64.54
K + CS	Depth	65.02
K + CS + Ours	Depth	65.47

Random initialization of training weights

1. Starting with ImageNet pre-trained model
2. Finetuning the network with small amount of semantic segmentation training data

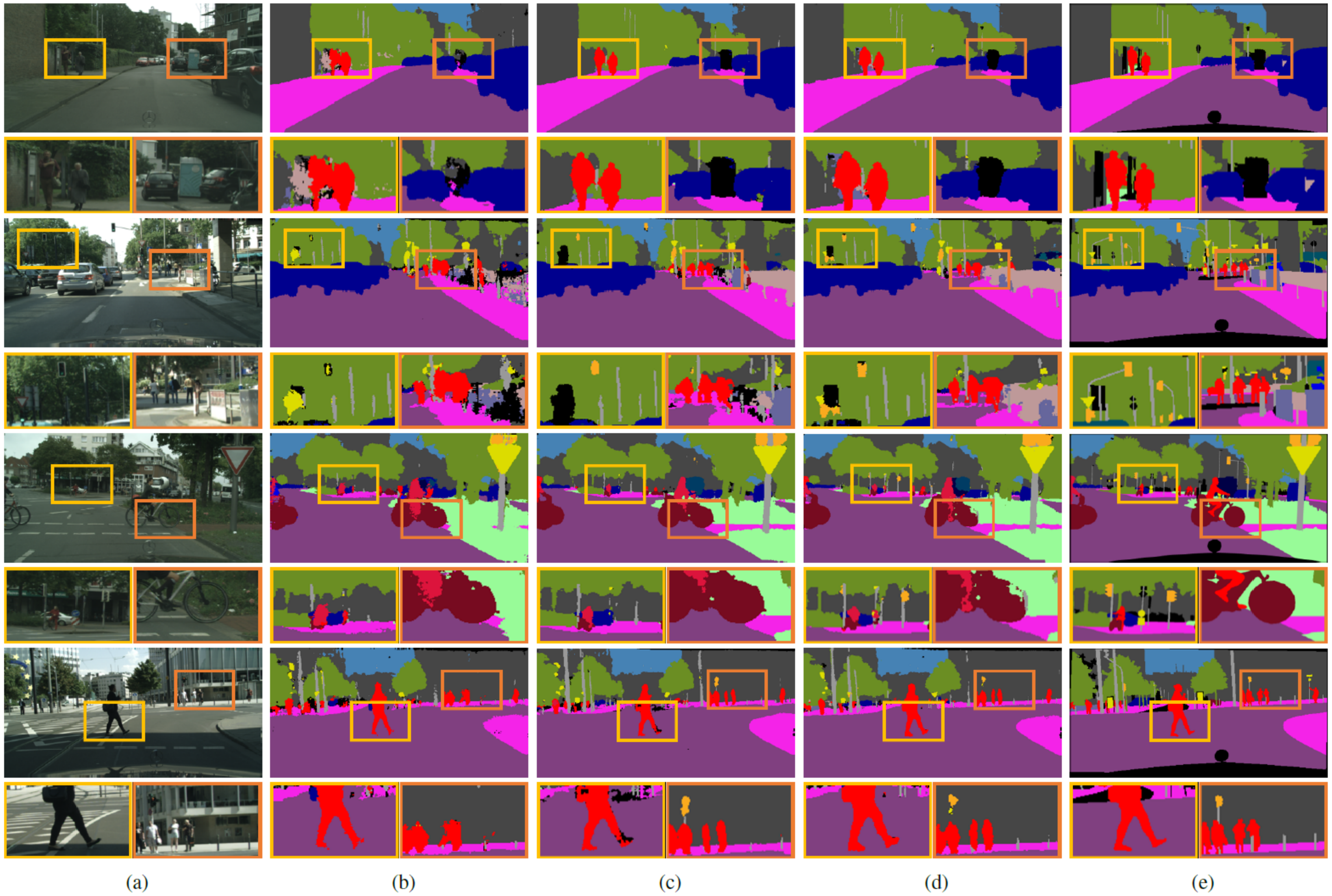
1. Starting with our pre-trained model
2. Finetuning the network with small amount of semantic segmentation training data

Training data for finetuning

Semantic segmentation: Cityscapes dataset
(a small amount of manually annotated training data)

Remarks)

1. Our dataset (DIML/CVL) is complementary to other dataset.
2. Our pre-trained model is comparable to the ImageNet pre-trained model.



(a) input images
(b) From scratch
(c) Using ImageNet pre-trained model
(d) Using our pre-trained model
(e) Ground truth annotations

Transfer Learning using Monocular Depth Network

- Road detection

Fmax: F1-measurement
AP: average precision

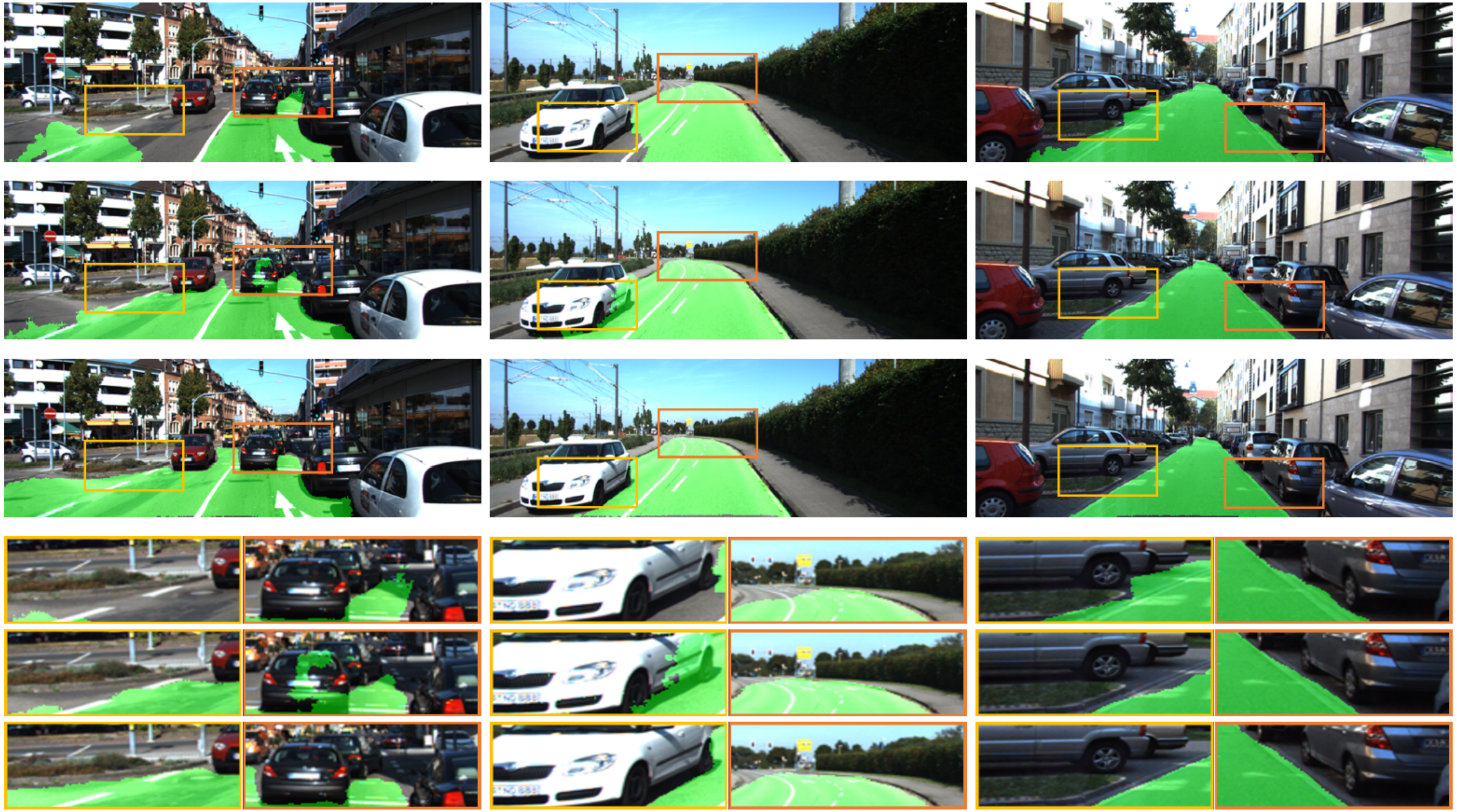
Road Detection			
Initialization	Pretext	Fmax	AP
Scratch	-	93.82	90.87
ImageNet pre-trained model [47]	Classification	94.28	92.25
K	Depth	94.41	92.04
K + Ours	Depth	94.92	92.28
K + CS	Depth	95.12	93.09
K + CS + Ours	Depth	95.65	94.46

Random initialization of training weights

1. Starting with ImageNet pre-trained model
2. Finetuning the network with small amount of road detection training data

1. Starting with our pre-trained model
2. Finetuning the network with small amount of road detection training data

<u>Training data for finetuning</u>
Road detection: KITTI road benchmark
(a small amount of manually annotated training data)



(a) UM (b) UMM (c) UU

Results learned from scratch

Using ImageNet pre-trained model

Our pre-trained model

UM: single lane road with markings
UU: single lane road without markings
UMM: multi-lane road with markings

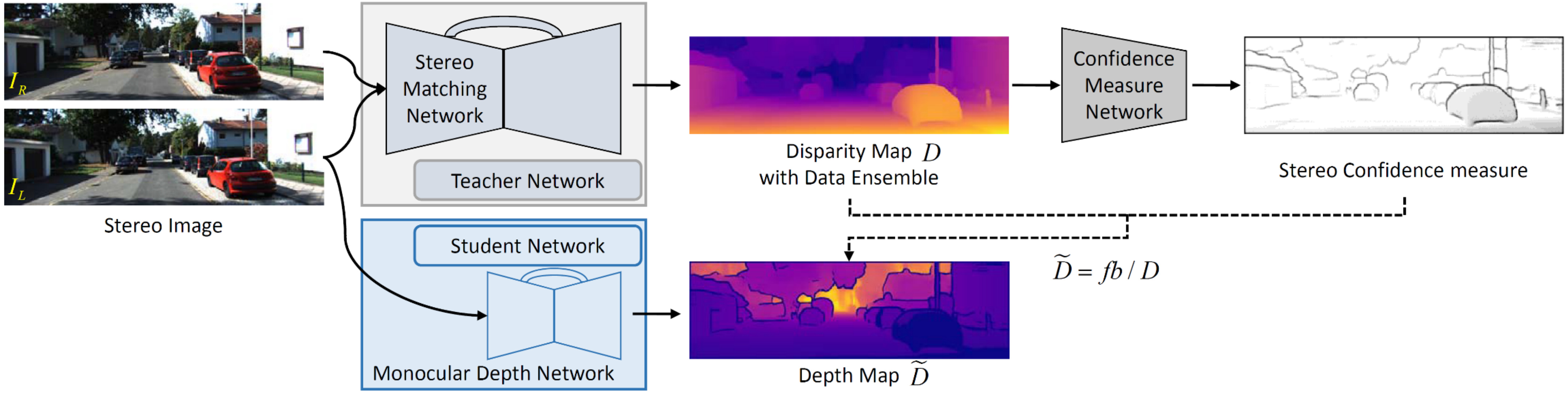
Conclusion

- DIML/CVL RGB+D dataset
  - 1 million outdoor scenes
  - Consisting of left and right color images, disparity maps, depth maps, confidence maps

- Semi-supervised learning approach for monocular depth estimation

Training
1. Left & Right image -> Left depth map & Confidence map
2. Left image -> Left depth map (assisted confidence map)

Test
Left image -> Left depth map



Conclusion

- Remarks on the proposed semi-supervised method
  - Our DIML/CVL dataset is complementary to other datasets.
  - Our strategy to construct massive training data (acquiring stereo images and estimating depth maps) is effective.
  - Our approach outperforms state-of-the-arts.
  - Confidence map is effective in addressing estimation errors of pseudo ground truth depth maps

Stereo Confidence Estimation

1. Deep learning based approach for confidence estimation

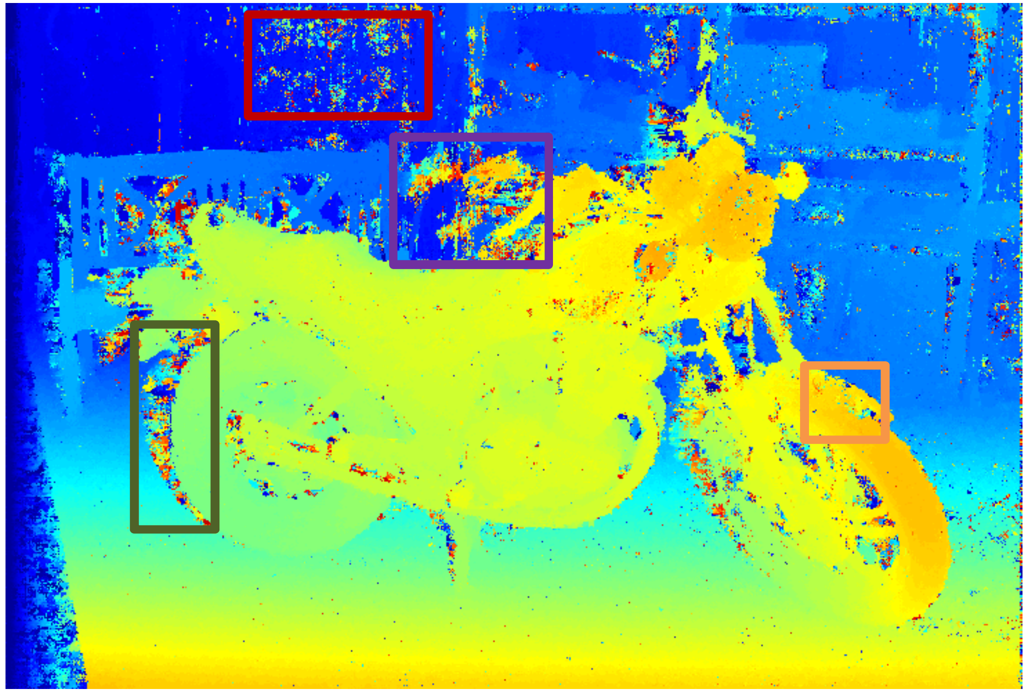
LAF-Net: Locally Adaptive Fusion Networks for Stereo Confidence Estimation, IEEE CVPR 2019 (oral presentation)

Stereo Confidence Estimation

- **Challenges on Stereo Matching**

Left image

Right image



Disparity

: textureless regions

: reflection regions

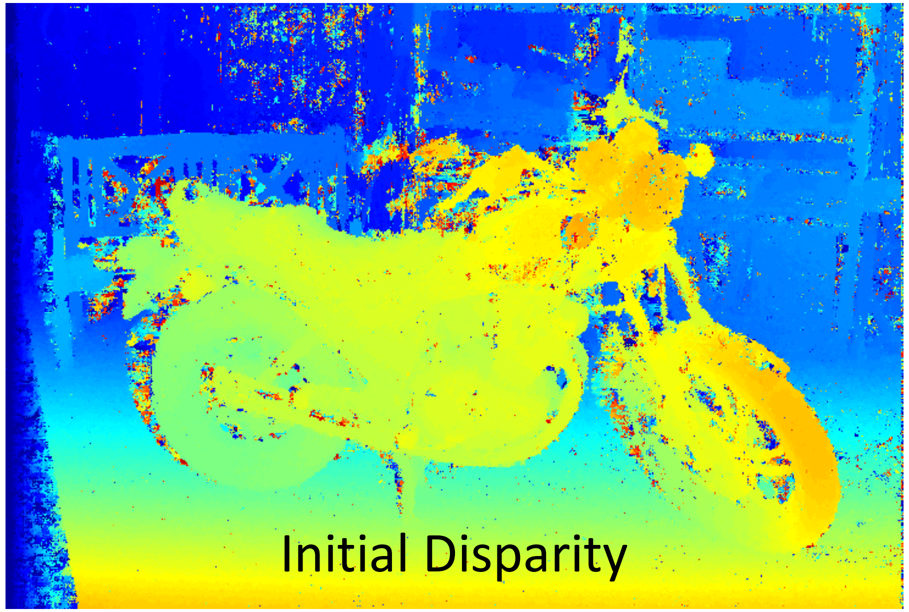
: occlusion regions

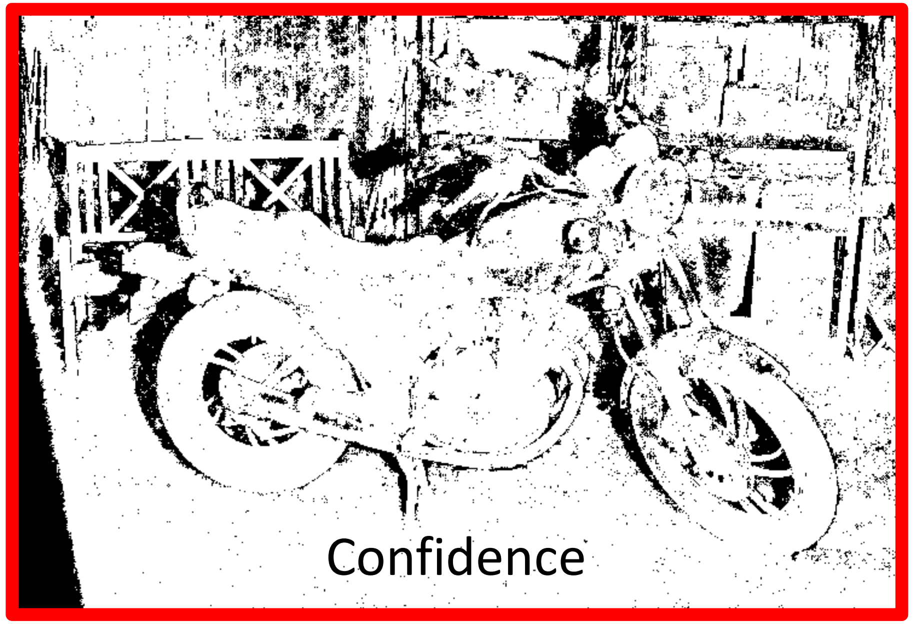
: illumination variations

  – Stereo Matching remains still <u>an unsolved problem</u> due to its inherent challenging elements, e.g., <u>textureless, reflection, occlusion regions, and illumination variations</u>

Stereo Confidence Estimation

- **Confidence estimation**
  - Confidence map indicates whether an estimated depth is reliable or not

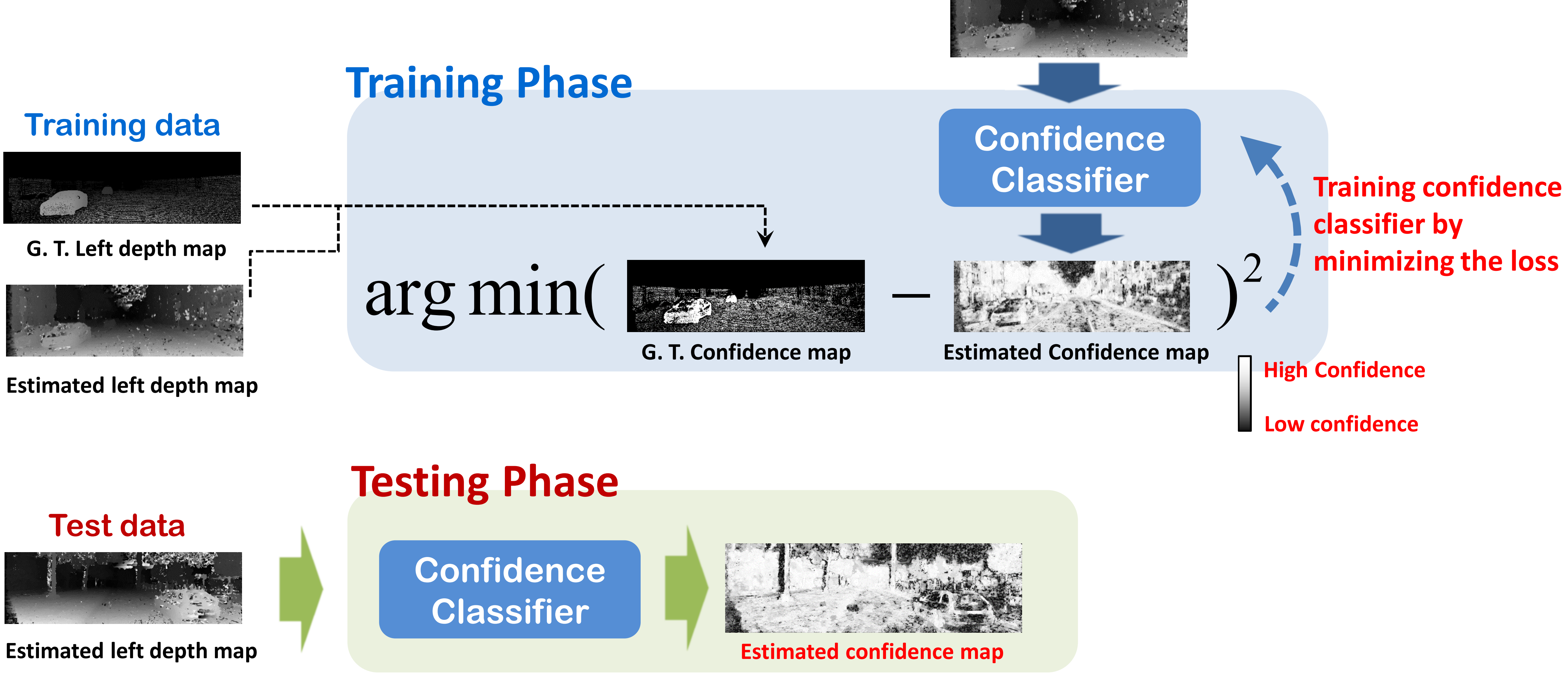




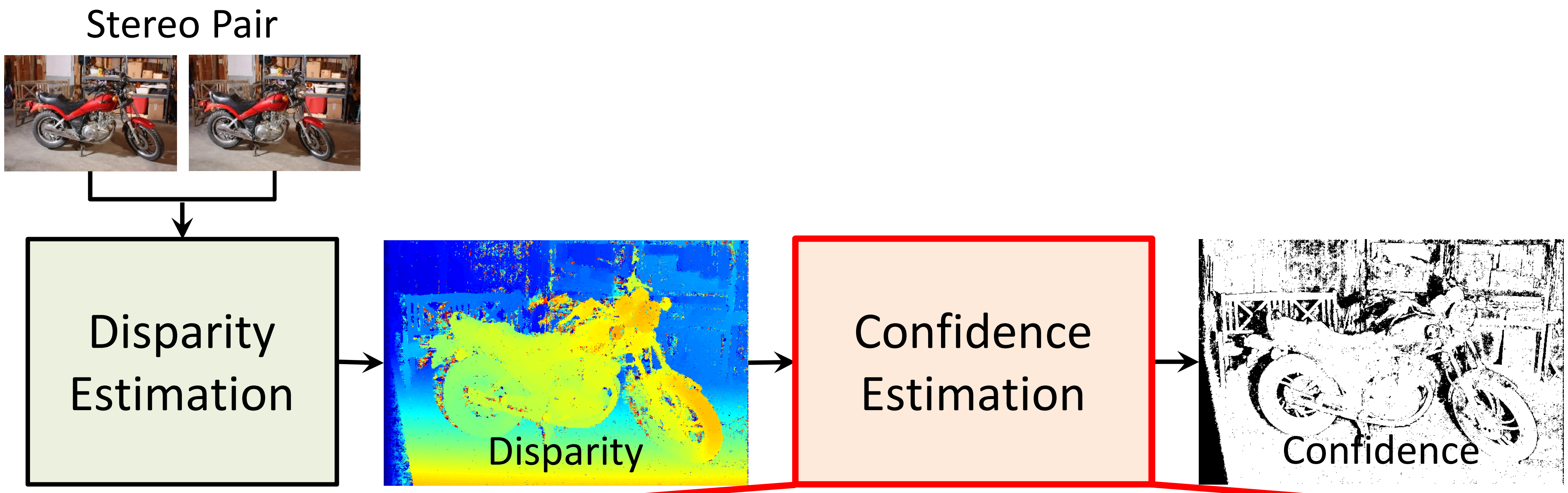
(0: unreliable <-> 1: reliable)

Related Work: Learning Approach based Confidence Estimation

Goal: designing a **confidence classifier**



Related Work: Learning Approach based Confidence Estimation



1) Confidence Feature Design:

→ How to extract robust and discriminative confidence features?

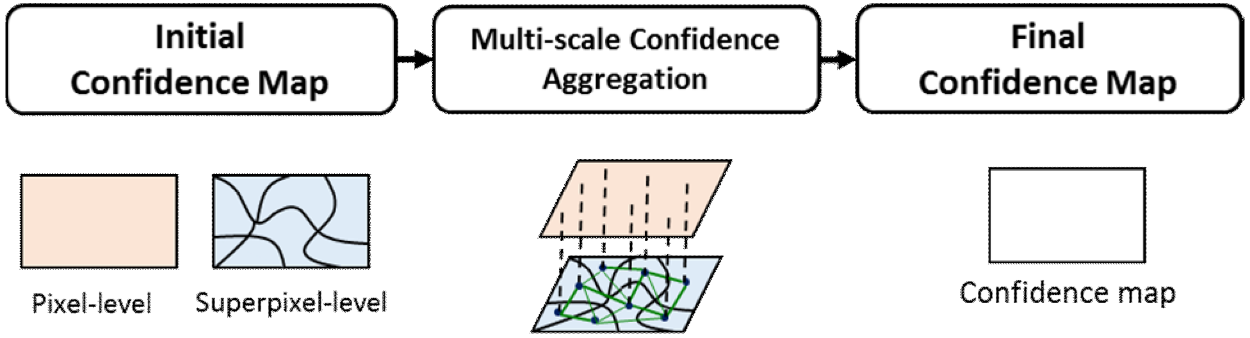
2) Confidence Estimator Learning:

→ How to learn confidence features effectively?

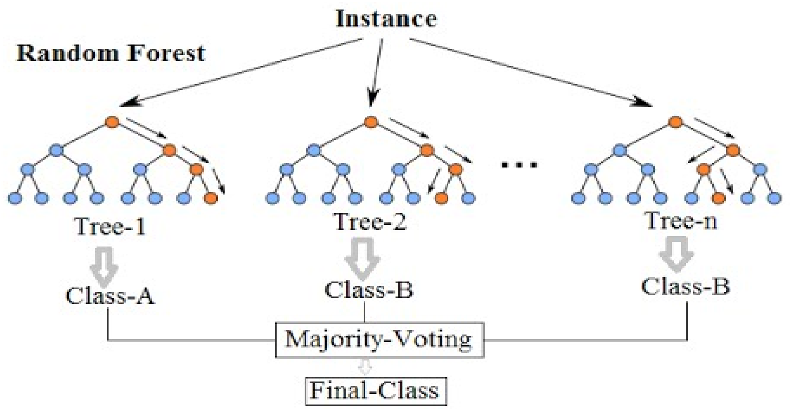
Related Work: Learning Approach based Confidence Estimation

Handcrafted approach for learning confidence classifier [20, 27]

Confidence Feature Design



Confidence Estimator Learning



[20] S. Kim, D. Min, S. Kim, and K. Sohn, “Feature augmentation for learning confidence measure in stereo matching,” IEEE Trans. Image Processing, 2017.
[27] M. Park and K. Yoon, “Leveraging stereo matching with learning-based confidence measures,” CVPR 2015

Related Work: Learning Approach based Confidence Estimation

Handcrafted Confidence Features [12]

- Entire cost curve / Local properties of the cost curve
- Local minima of the cost curve

} From Matching Cost

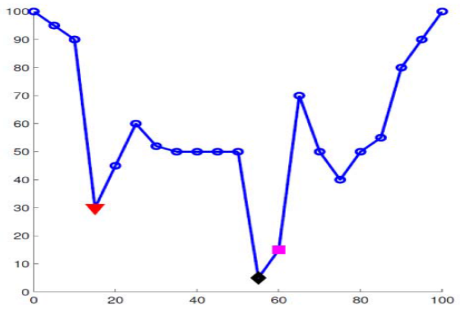
- The consistency between the left and right disparity maps
- Median deviations of disparity values

} From Disparity

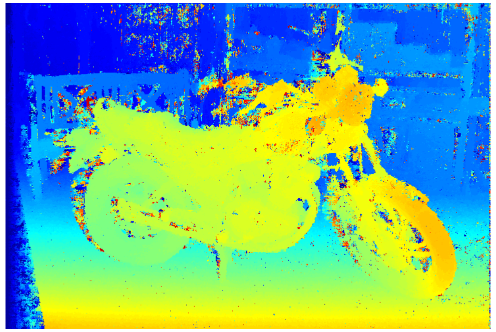
- Image gradients
- Zero mean sum of absolute differences

} From Color Image

- Etc.



Matching cost



Disparity

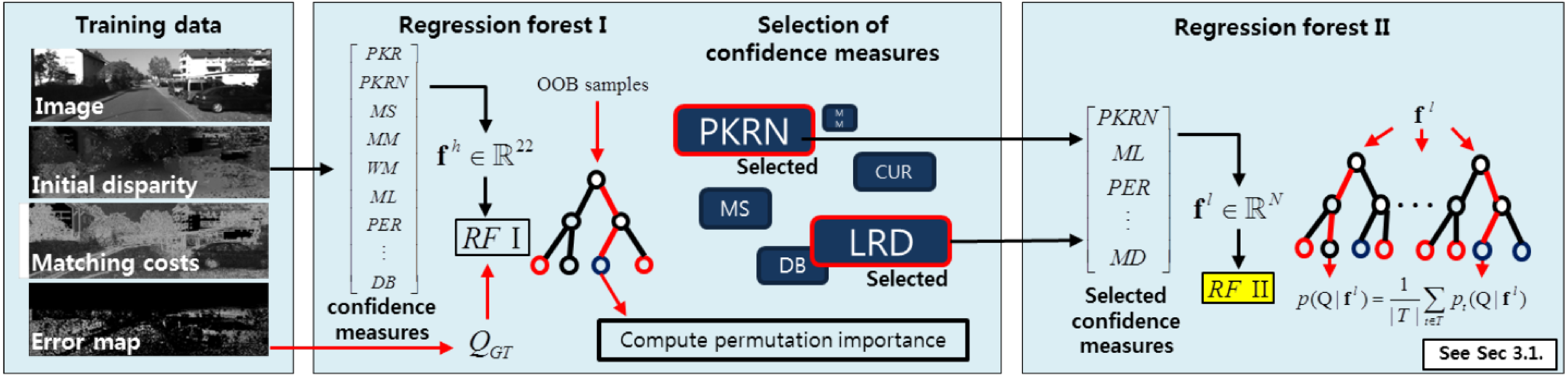
Color image

[12] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. IEEE Trans. Pattern Anal. Mach. Intell.,, 2012.

Related Work: Learning Approach based Confidence Estimation

Combination of Handcrafted Confidence Features

[Haeusler et al. CVPR’13, Spyropoulos et al. CVPR’14, Park et al. CVPR’15, Poggi et al. 3DV’16]



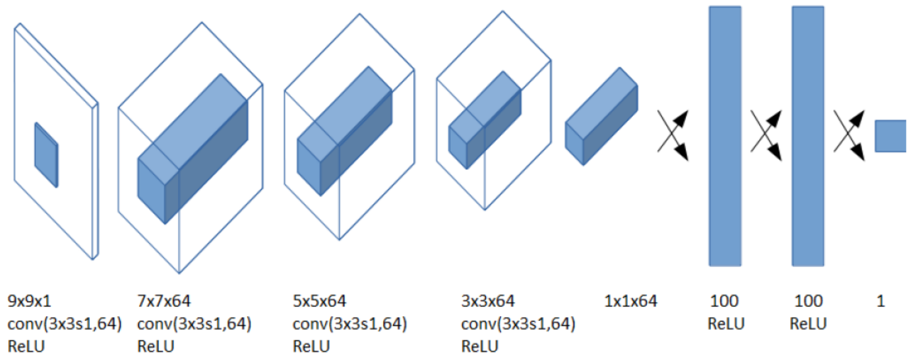
이화여자대학교
EWHA WOMANS UNIVERSITY

Related Work: Learning Approach based Confidence Estimation

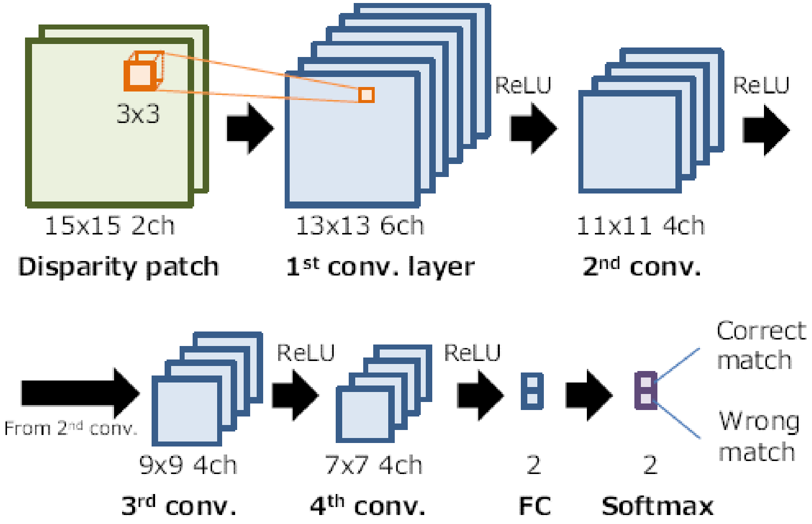
Handcrafted confidence features → **NOT** optimal!

→ **Convolutional Neural Networks (CNNs)-based Approaches**

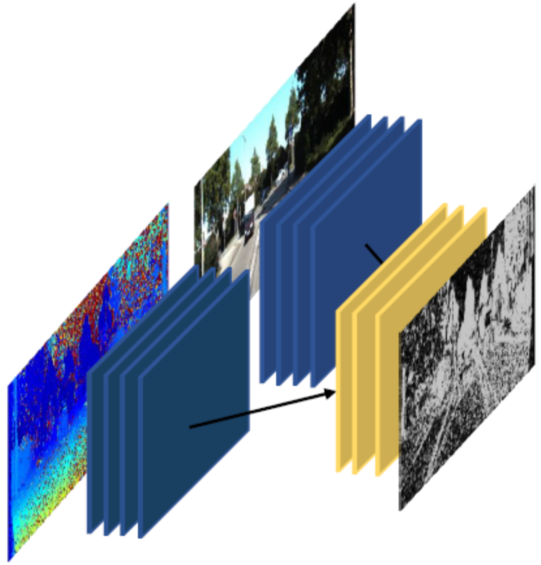
⇨ *Learning confidence features from* ***disparity*** *and/or* ***color image***



CCNN [Poggi et al., BMVC’16]



PBCP [Seki et al., BMVC’16]



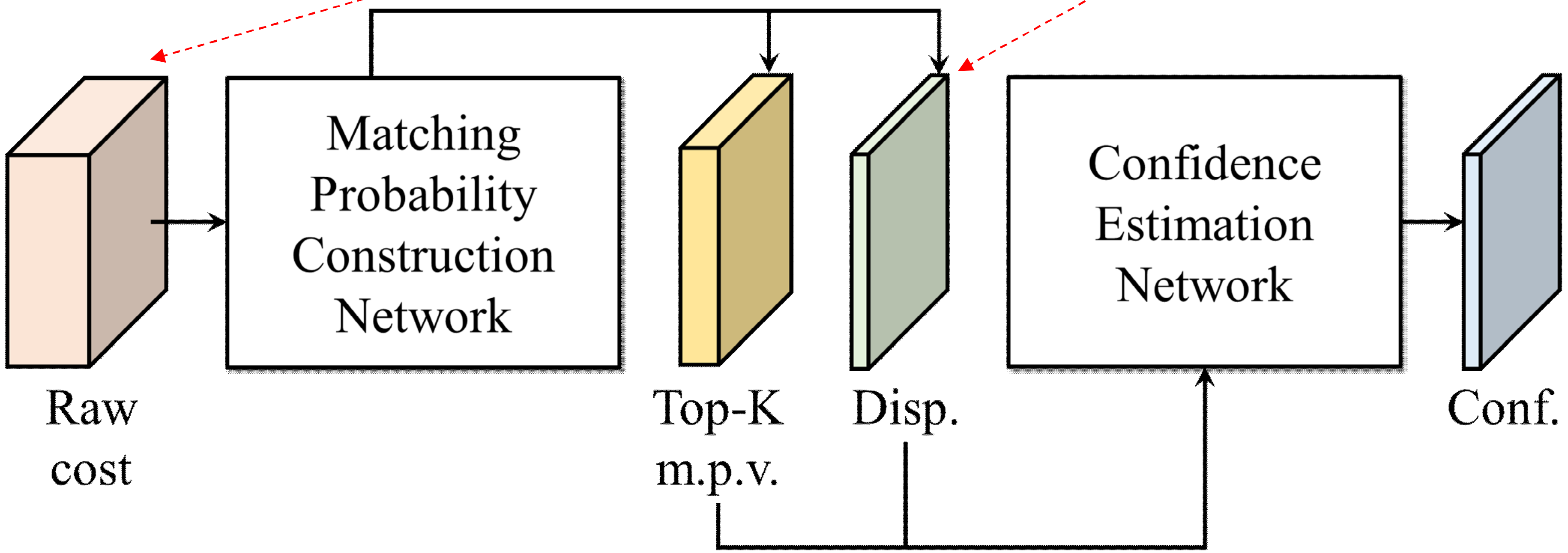
LFN [Fu et al., WACV’18]

Related Work: Learning Approach based Confidence Estimation

Handcrafted confidence features → **NOT** optimal!

→ **Convolutional Neural Networks (CNNs)-based Approaches**

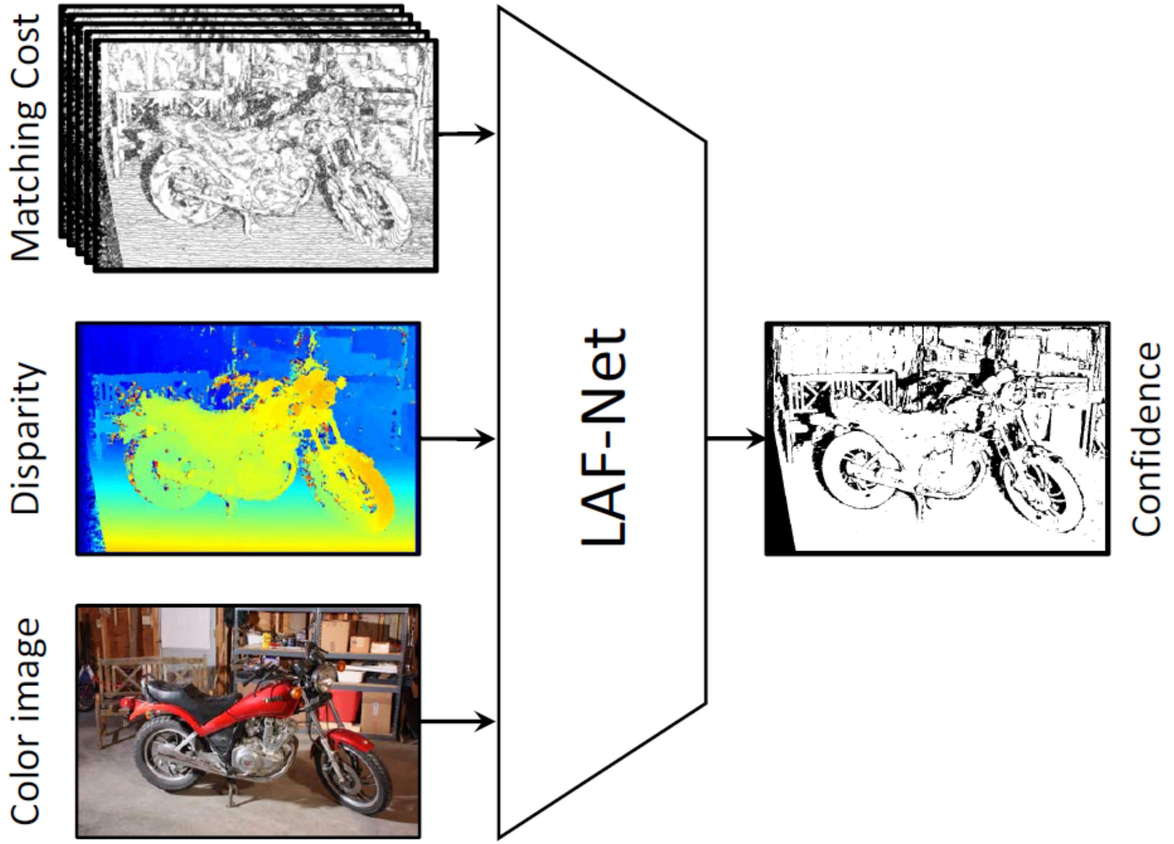
⇨ *Learning confidence features from **matching cost** and **disparity***



[21] S. Kim, D. Min, S. Kim, and K. Sohn, "Unified confidence estimation networks for robust stereo matching," IEEE Trans. Image Processing, 2019.

Proposed Method

- First confidence estimation approach that makes full use of tri-modal input (**matching cost, disparity, and color image**)



Key issue

How to fuse such heterogeneous inputs well (matching cost, disparity, and color image)

⇨ **Attention Network**
Scale Inference N

tri-modal input
→ tri-modal features

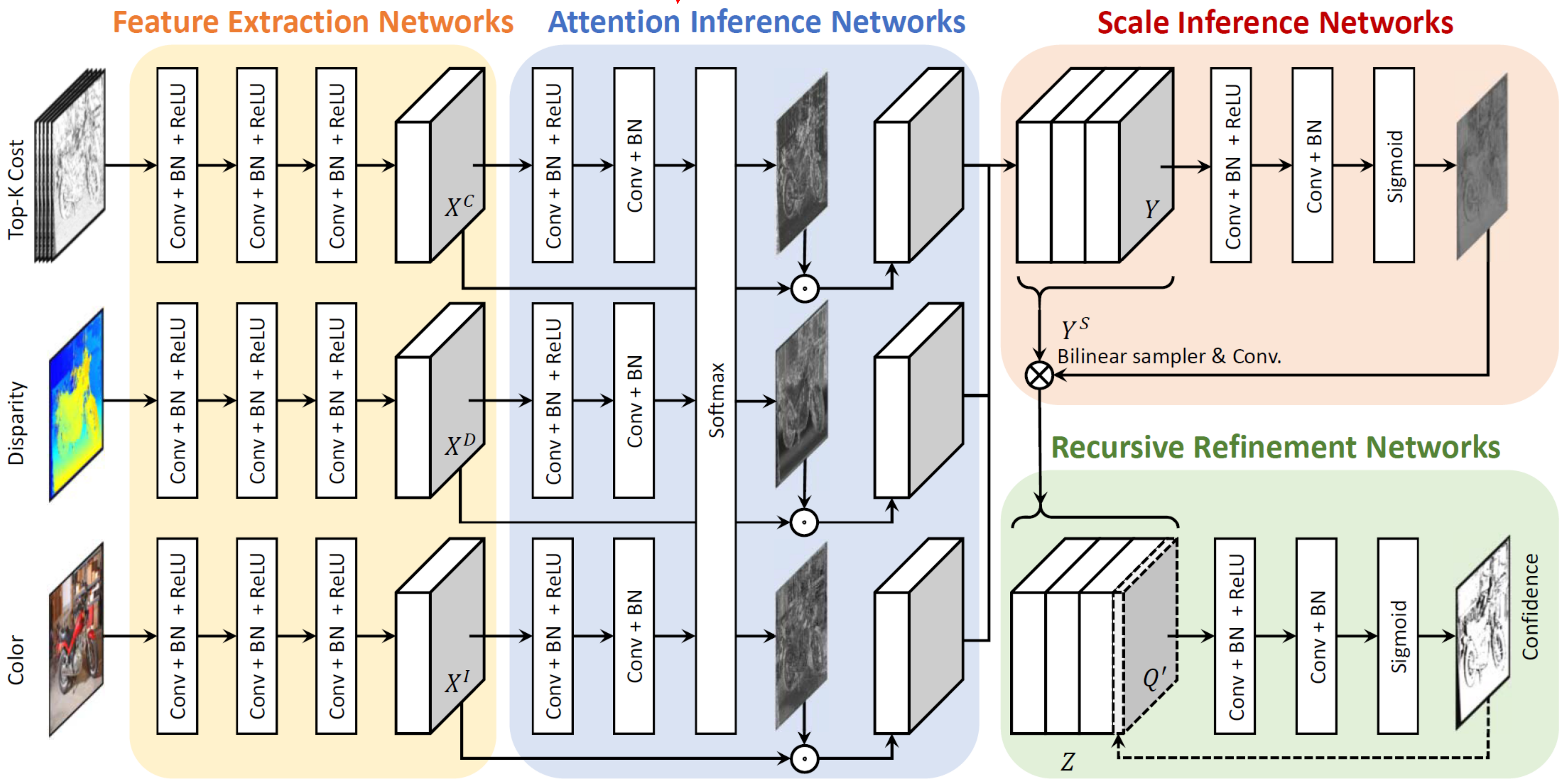
Extract features from tri-modal input
(Top-K matching cost, disparity, color image)

tri-modal features
→ confidence features

Infer locally-varying attention weights of the tri-modal features

confidence features
→ refined confidence features

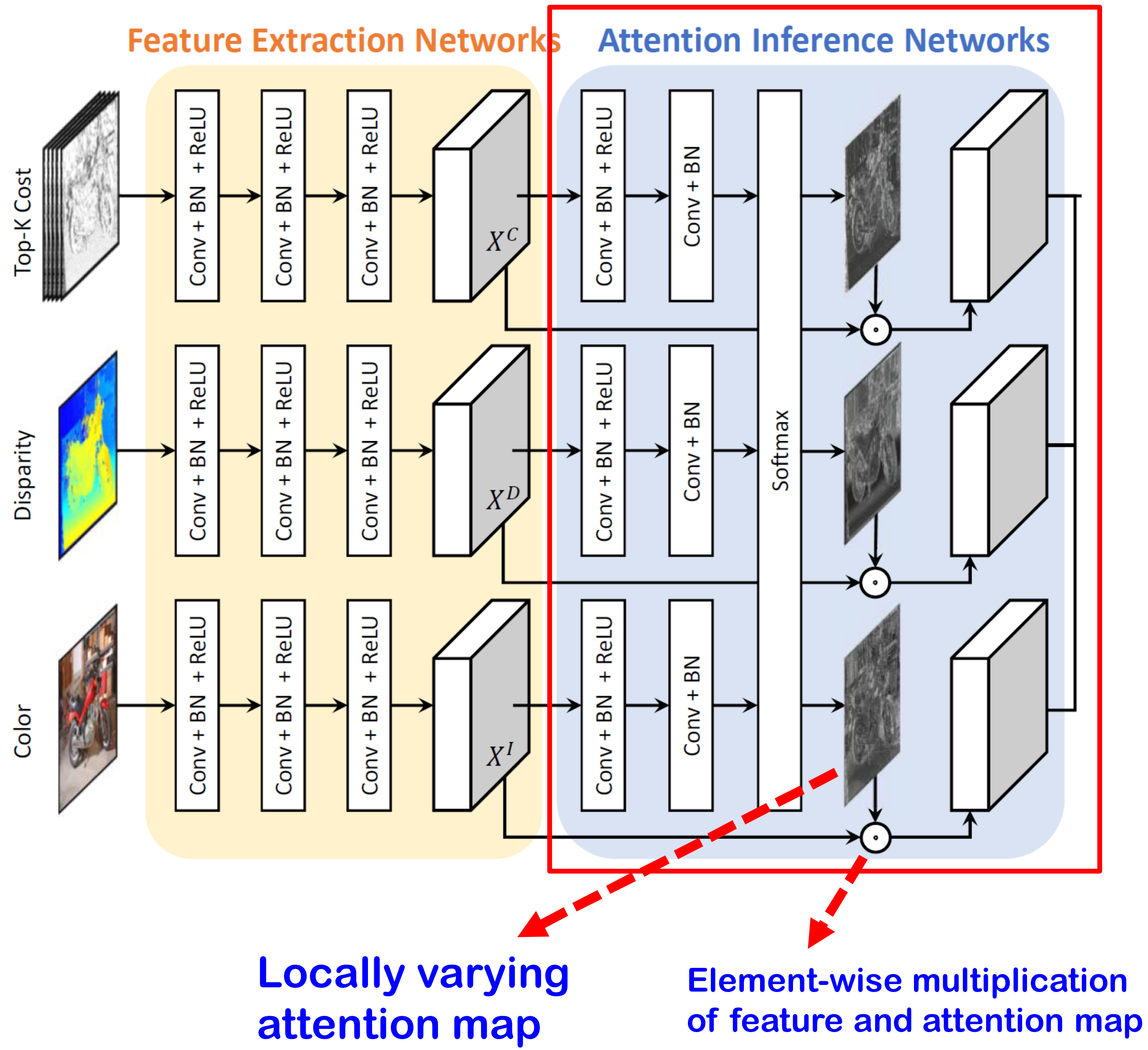
Determine optimal receptive fields for confidence features



refined confidence features
+ current confidence map
→ confidence map

Estimate the confidence map in a recursive manner

Attention Inference Networks



- **Several methods to fuse tri-modal input**
  1. Direct concatenation of tri-modal input

     This yields a poor performance due to their heterogeneous attributes of tri-modal input.

  2. Concatenation of tri-modal features

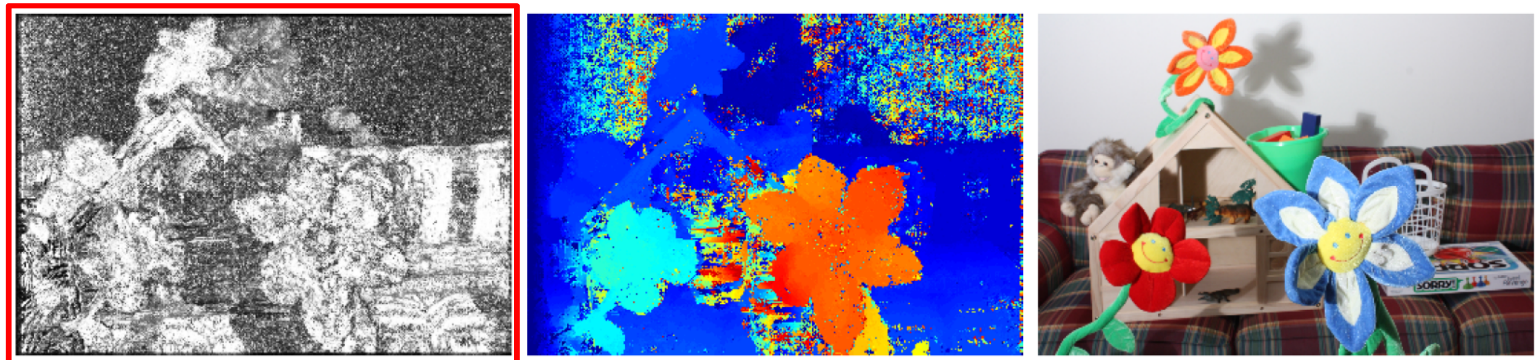
     Fusion weights are always fixed so it does not fuse them optimally.

  3. **The proposed method**

     Attention inference networks:
     - Infer locally-varying attention map of the tri-modal feature.
     - Attention map is determined dynamically conditioned on input tri-modal features.

Attention Inference Networks

Attention maps for different input modalities



Top-1 matching cost | Disparity | Color image

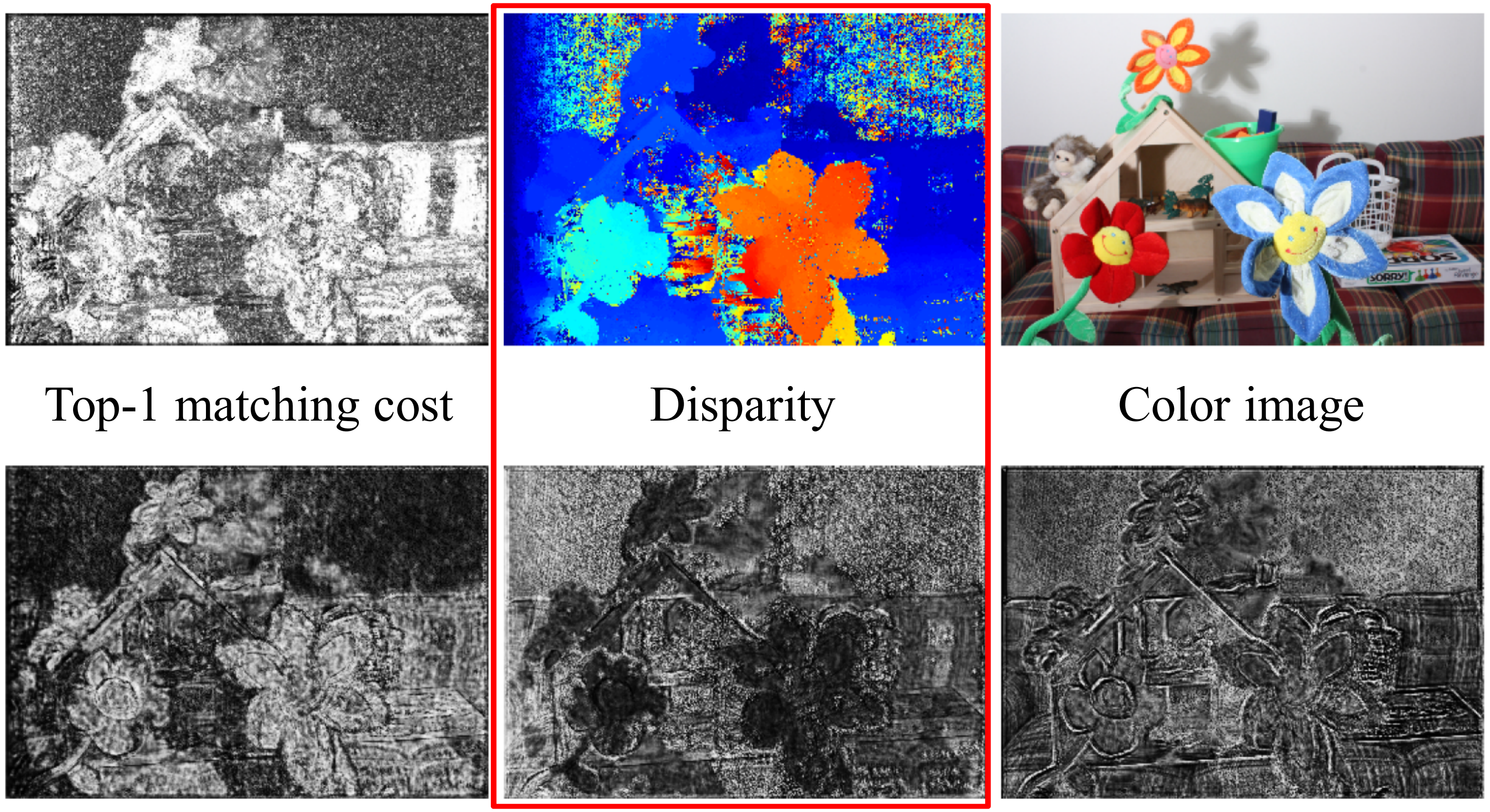
Attention map of matching cost | Attention map of disparity | Attention map of color image

The attention of top-K matching cost is high for pixels with high matching probability.

[8] R. Haeusler, R. Nair, and D. Kondermann, "Ensemble learning for confidence measures in stereo vision," CVPR 2013

Attention Inference Networks

Attention maps for different input modalities



Top-1 matching cost | Disparity | Color image

Attention map of matching cost | Attention map of disparity | Attention map of color image

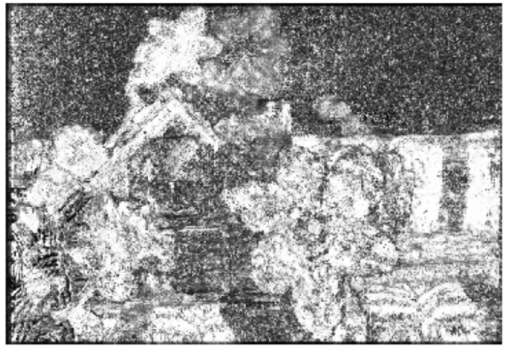
The attention of top-K matching cost is high for pixels with high matching probability.

The attention of disparity is high in noisy region, indicating informative features can be extracted from the different disparity assignments.
(similar to VAR or MDD [8] in handcrafted features)

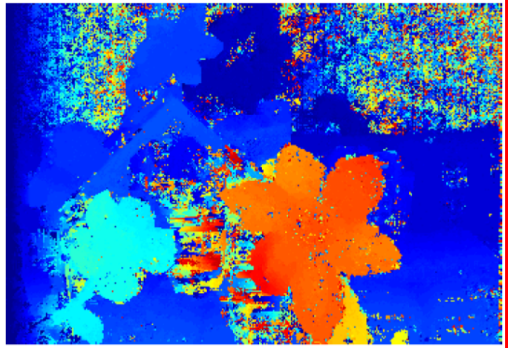
[8] R. Haeusler, R. Nair, and D. Kondermann, "Ensemble learning for confidence measures in stereo vision," CVPR 2013

Attention Inference Networks

Attention maps for different input modalities

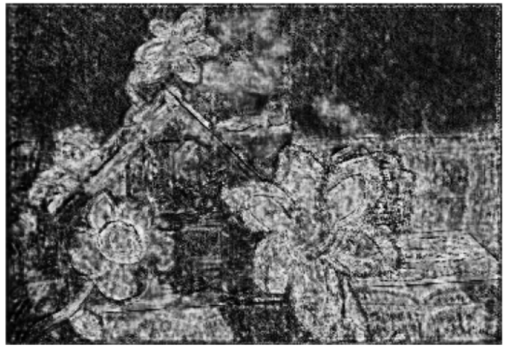


Top-1 matching cost

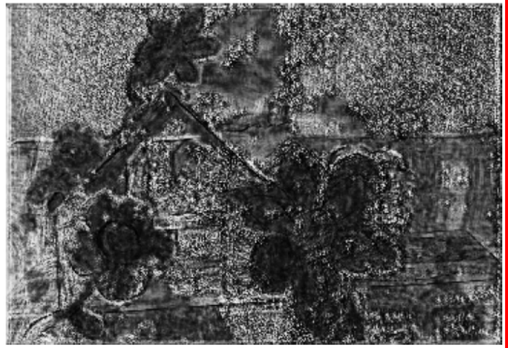


Disparity

Color image



Attention map of matching cost



Attention map of disparity

Attention map of color image

The attention of top-K matching cost is high for pixels with high matching probability.

The attention of disparity is high in noisy region, indicating informative features can be extracted from the different disparity assignments.
(similar to VAR or MDD [8] in handcrafted features)

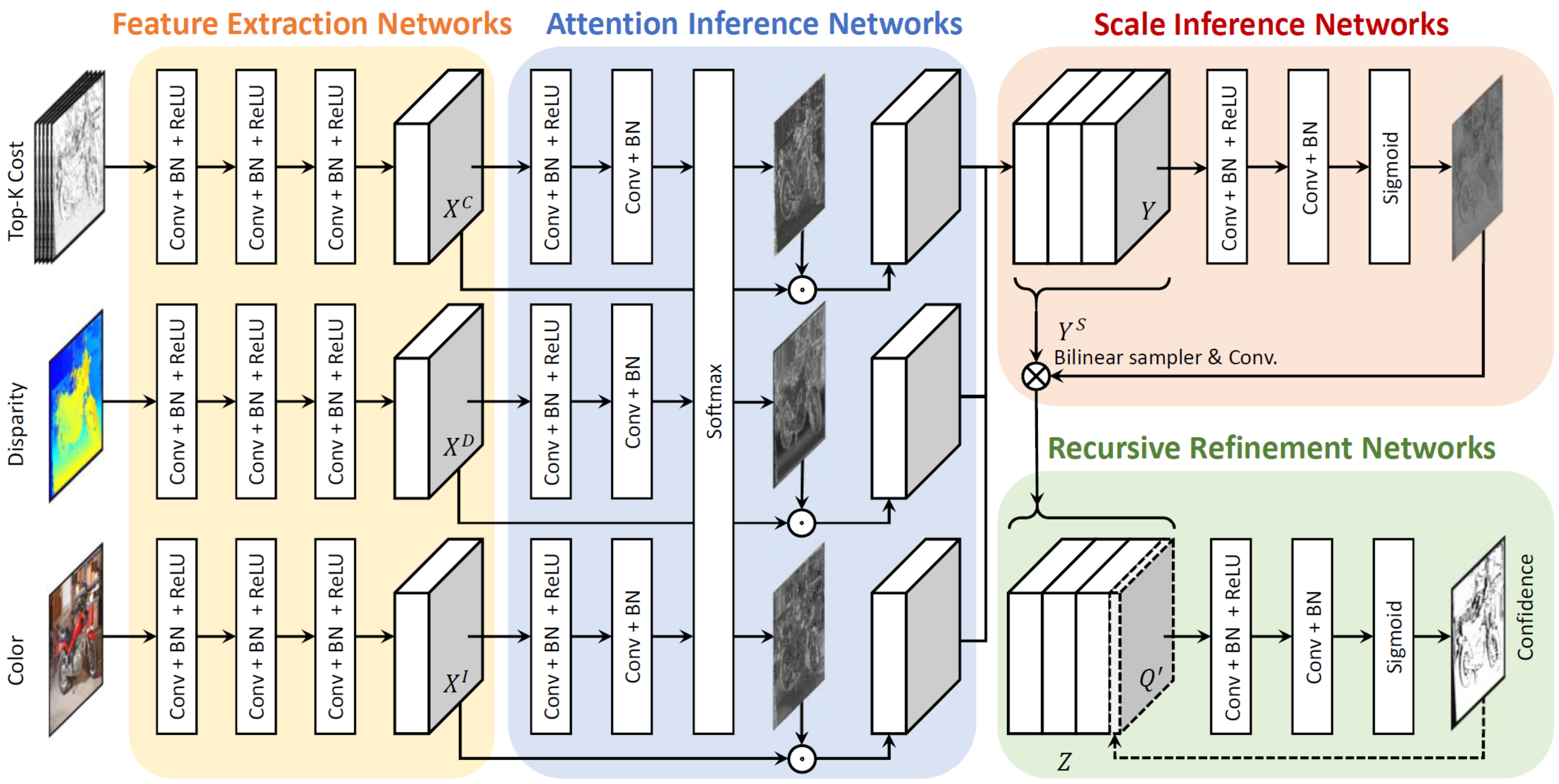
The attention of color image is high near image boundary, indicating that a image texture gives a useful cue.

[8] R. Haeusler, R. Nair, and D. Kondermann, "Ensemble learning for confidence measures in stereo vision," CVPR 2013

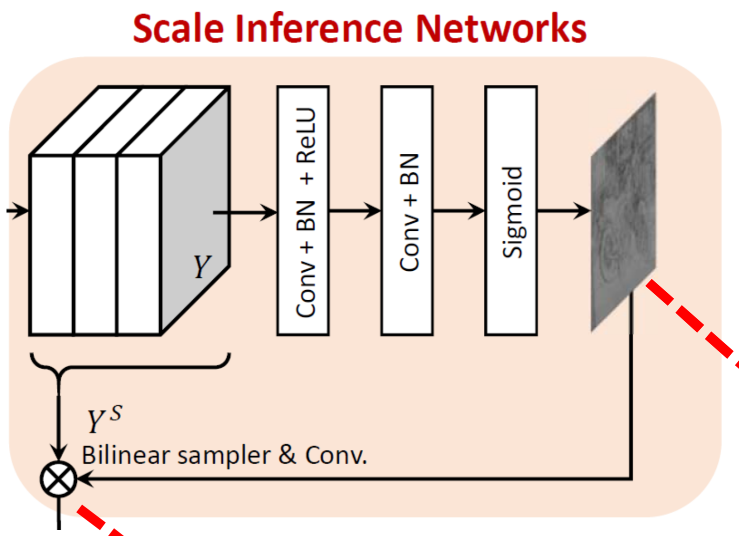
Scale Inference Network

confidence features
→ refined confidence features

Determine optimal receptive fields for confidence features



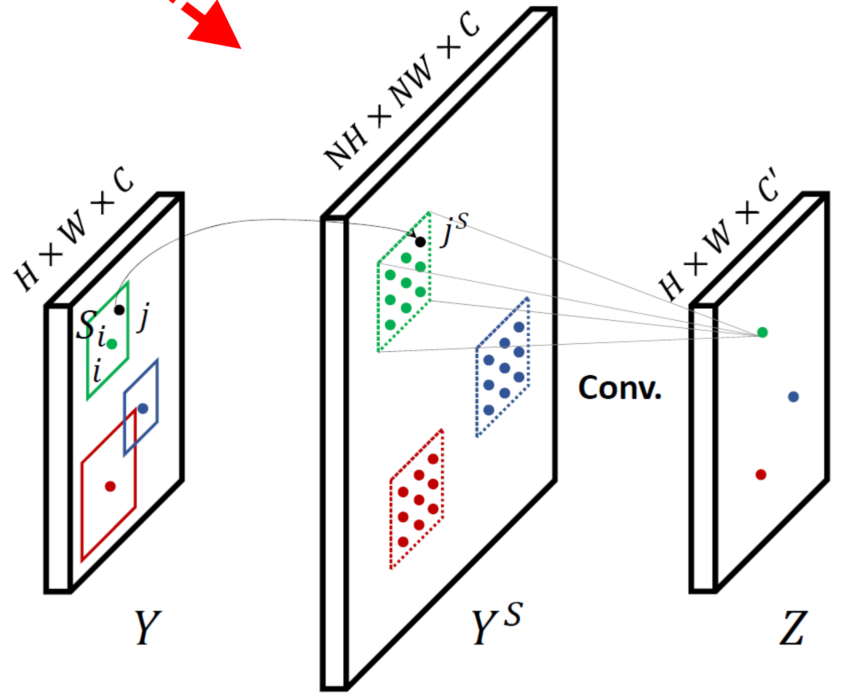
Scale Inference Network



The optimal receptive fields for confidence features vary at each pixel.

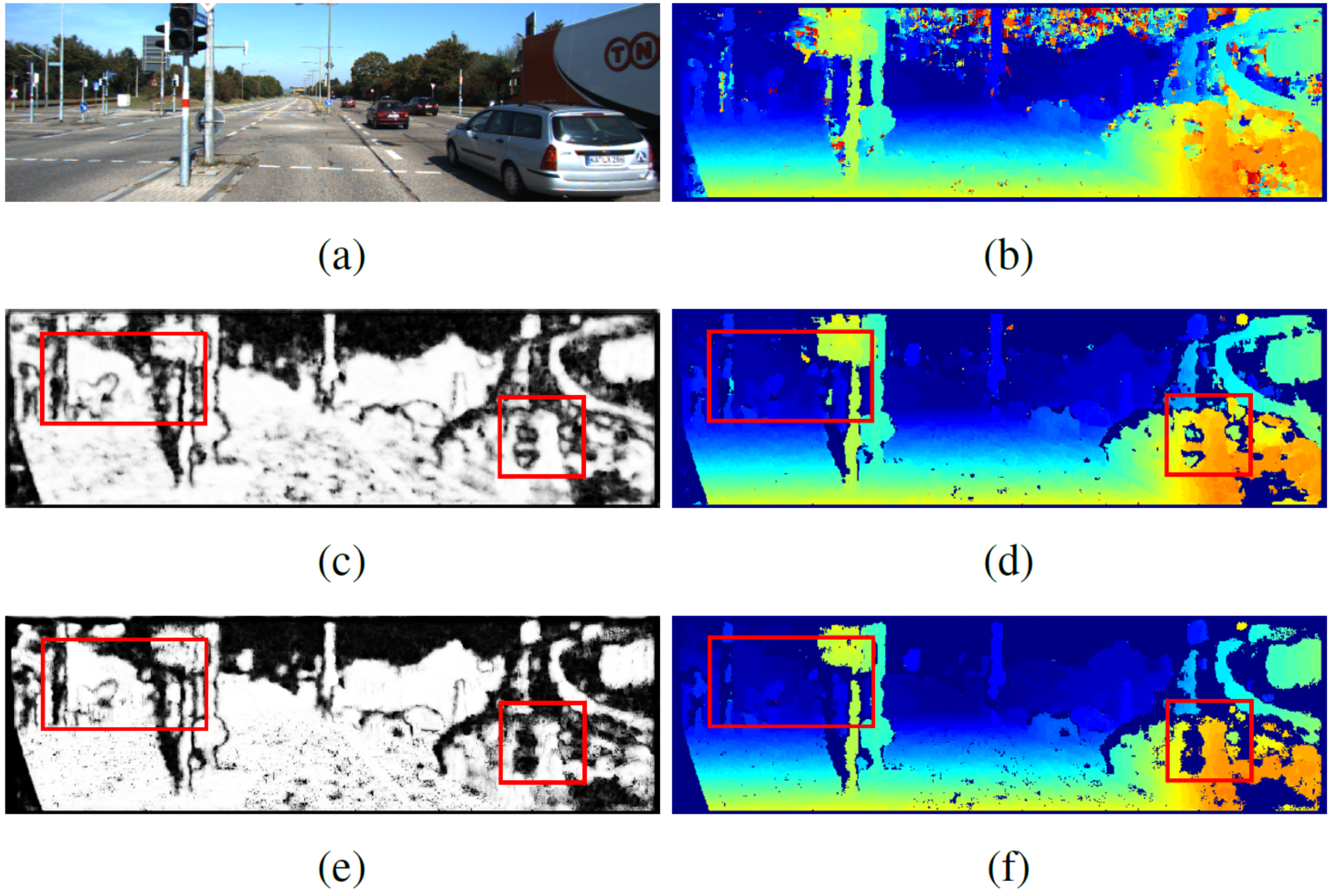
-> Scale inference networks are used to determine optimal receptive fields for confidence features.

1. **Optimal scale is inferred for each pixel**



2. **(Y → Y^s) Using locally-varying sampling grid, the convolution activations Y are resampled into Y^s.**
3. **(Y^s → Z) Convolution is applied with a stride of N.** (N=3, in this example)

Recursive Refinement Networks



(a) Left color image
(b) Initial disparity
(c) Estimated confidence without recursive module
(d) Thresholded disparity with (c)
(e) Estimated confidence with recursive module
(f) Thresholded disparity with (e).

Mismatched pixels in the red boxes are reliably detected with the proposed recursive confidence refinement networks.

Experimental Setup

Implementation Details

– **Raw matching cost:** Census-SGM [Hirschmuller, TPAMI'08],
MC-CNN [Zbontar et al., CVPR'15]

Datasets

– Training: MPI Sintel dataset and KITTI 2012 dataset
– Test: Middlebury 2006 (MID 2006), Middlebury 2014 (MID 2014), and KITTI 2015 dataset

Comparison with other methods

– **Handcrafted approaches:** Haeusler et al. [8], Spyropoulos et al. [38], Park and Yoon [27], Poggi and Mattoccia [29], Kim et al. [20]
– **CNN-based approaches:** CCNN [30], PBCP [36], Kim et al. [21], LFN [7], ConfNet [39], LGC-Net [39]

For references, refer to "LAF-Net: Locally Adaptive Fusion Networks for Stereo Confidence Estimation", CVPR 2019

Ablation study

Ablation study of input tri-modal data

Area Under Curve (AUC): The lower, the better

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

Area Under Curve (AUC)

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 three inputs and three sub-networks leads to a substantial performance gain.

Evaluation metric: AUC?

Sparsification curve: draws a bad pixel rate while successively removing pixels in descending order of confidence values in the disparity map
Area under curve (AUC): area of the sparsification curve

이화여자대학교
EWHA WOMANS UNIVERSITY

Comparison with state-of-the-arts

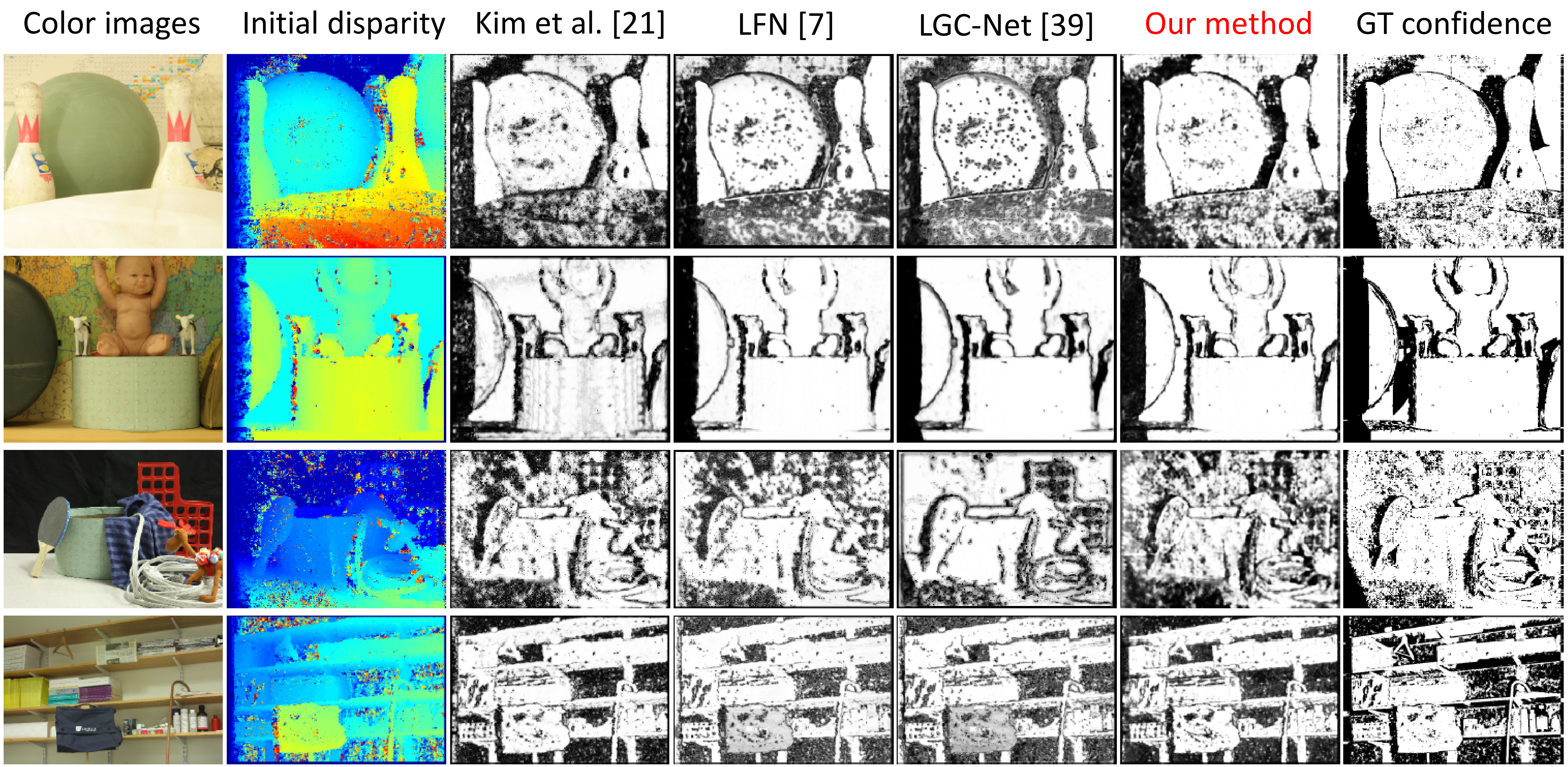
Average AUC

Matching cost: census-based SGM and MC-CNN

Test data: Middlebury 2006 (MID 2006), Middlebury 2014 (MID 2014), and KITTI 2015 datasets

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

Comparison with state-of-the-arts



Conclusion

- 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 the heterogeneous tri-modal input
- Recursive refinement networks improves the accuracy.
- Further study
  - How confidence estimation networks could be learned in an unsupervised manner

Experimental Setup

- The sparsification curve draws a bad pixel rate while successively removing pixels in descending order of confidence values in the disparity map