# Depth Learning:
# When Depth Estimation Meets Deep Learning

Prof. Liang LIN

SenseTime Research & Sun Yat-sen University

NITRE, June 18, 2018

# Outline

Part 1. Introduction

- Motivation
- Stereo Matching
- Single Image Depth Estimation

Part 2. Depth Learning—Our Progress

- Cascade Residual Learning (CRL)
- Zoom and Learn (ZOLE)
- Single View Stereo Matching (SVS)

Part 3. Conclusion

# Outline

Part 1. Introduction

- Motivation
- Stereo Matching
- Single Image Depth Estimation

Part 2. Depth Learning—Our Progress

- Cascade Residual Learning (CRL)
- Zoom and Learn (ZOLE)
- Single View Stereo Matching (SVS)

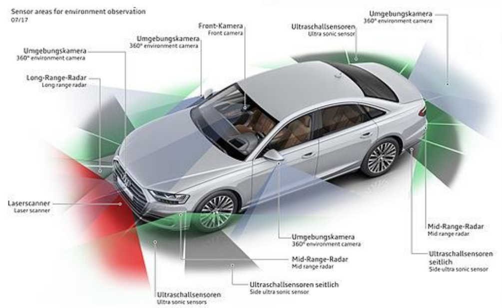
Part 3. Conclusion

# Motivation—Importance of Depth

Depth-of-field Rendering (Bokeh)

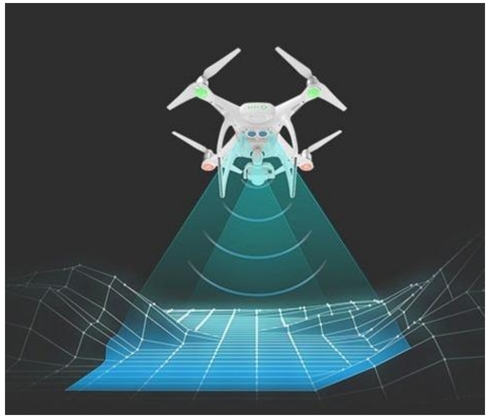

- Face ID


AR / VR



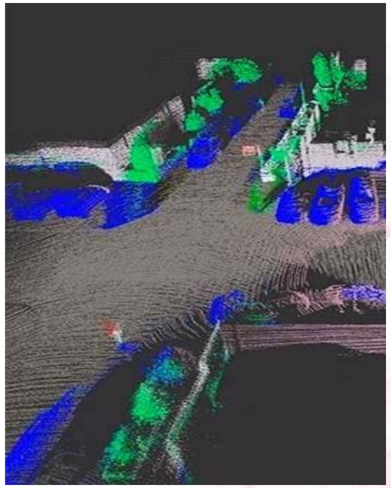

ADAS

Industrial Automation



Obstacle avoidance

Entertainment



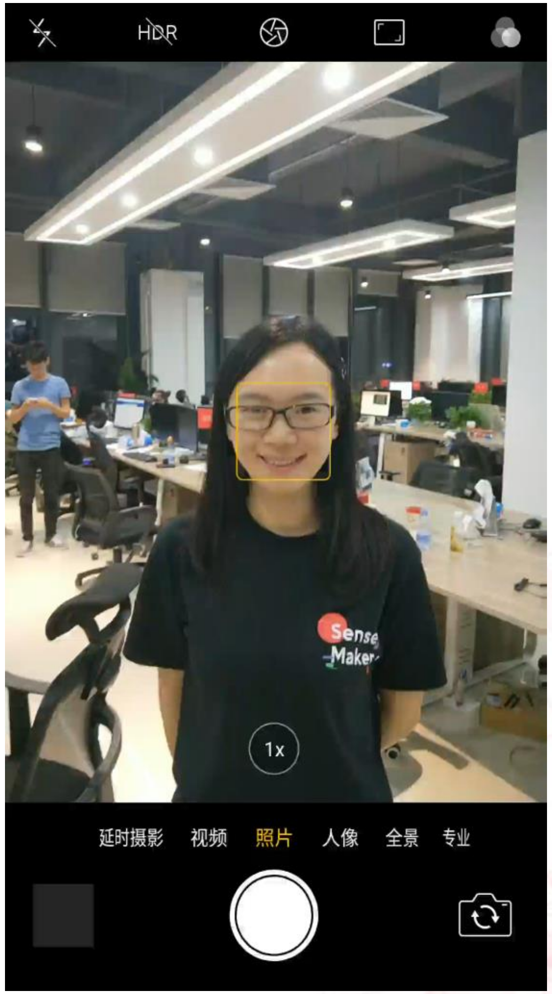
- 3D Reconstruction

# Application: Bokeh in mobile phone

商汤 sensetime


景物虚化效果
Object Bokeh


视频重对焦
Video Refocus

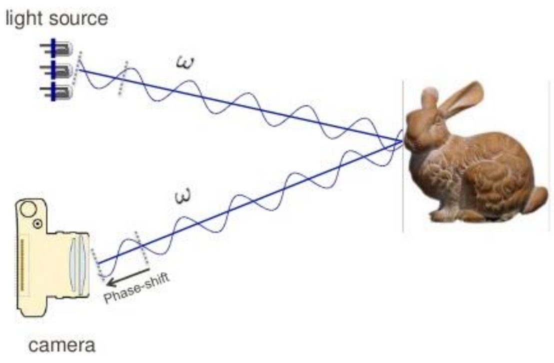



人脸对焦跟随
Face Tracking

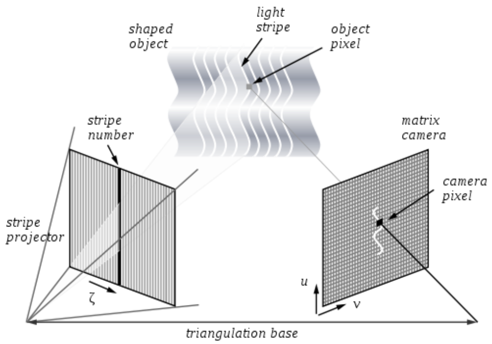
# Applications: AR & Vide Editing

# Motivation—Depth Estimation

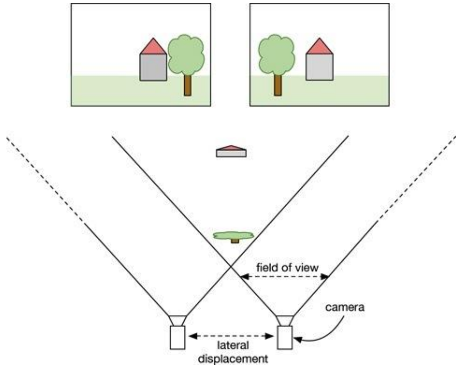
## Multiple ways to estimate depth



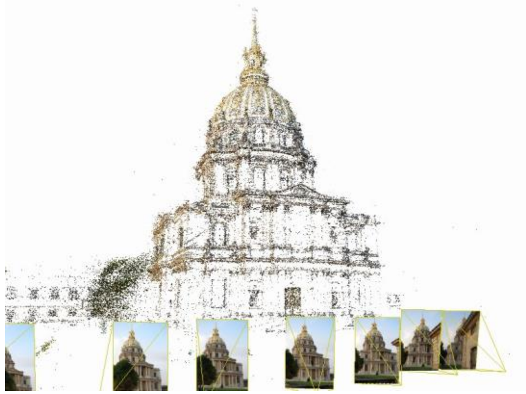

Time of flight



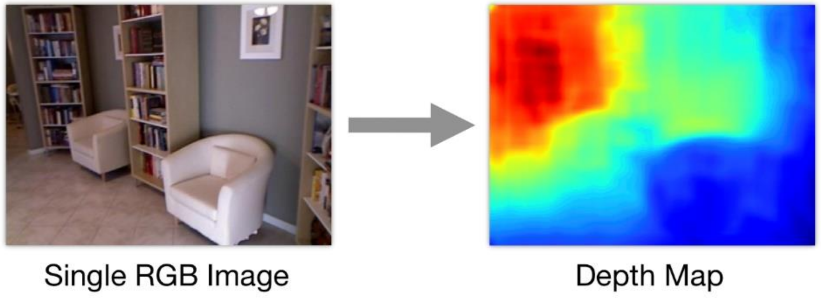

Structure light




Stereo matching



Structure from motion




Single Image Depth Estimation

Most cost-effective approaches

Main focus of this talk

# Outline

Part 1. Introduction

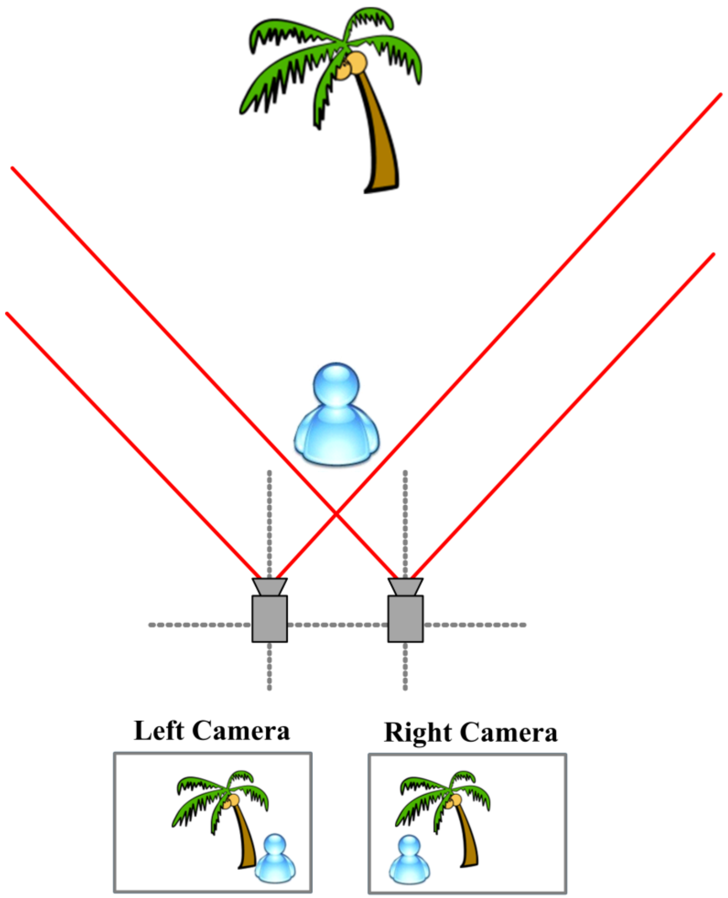
- Motivation
- Stereo Matching
- Single Image Depth Estimation

Part 2. Depth Learning—Our Progress

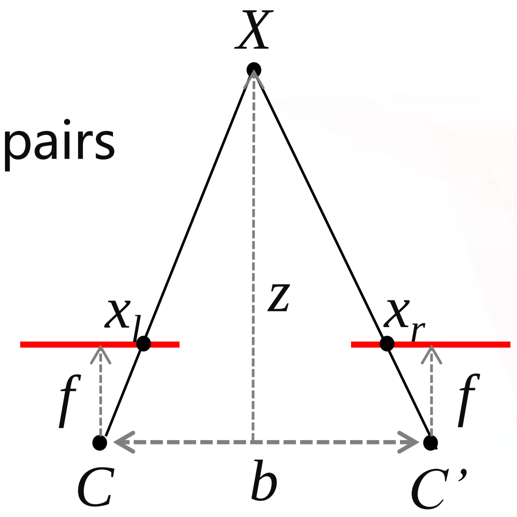
- Cascade Residual Learning (CRL)
- Zoom and Learn (ZOLE)
- Single View Stereo Matching (SVS)

Part 3. Conclusion

# Stereo Matching

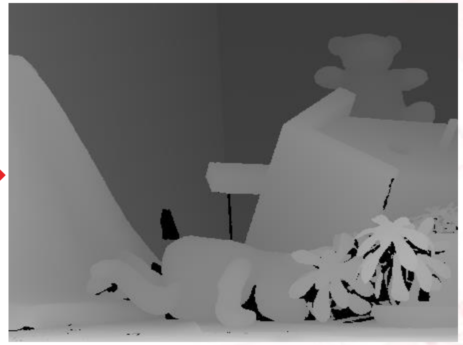



- Estimate pixel correspondence in rectified pairs
- Disparity Map $D(x,y) = x_r - x_l$
- Depth Map $Z(x,y) = \frac{f \cdot b}{D(x,y)}$




Left image

Right image



Disparity Map

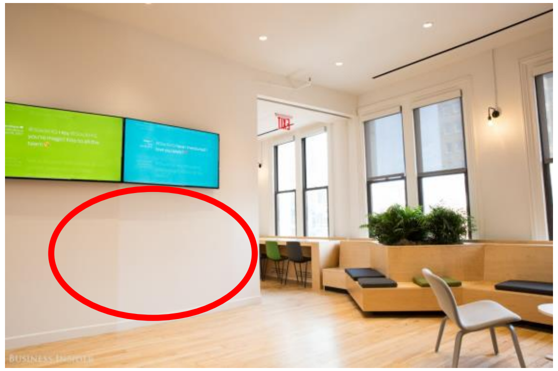
[1] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," . *International Journal of Computer Vision*, 47(1-3):7–42, 2002.

# Stereo Matching—Challenges

Underdetermined (ill-posed)

Photometric variations

Occlusions



Texture-less Areas

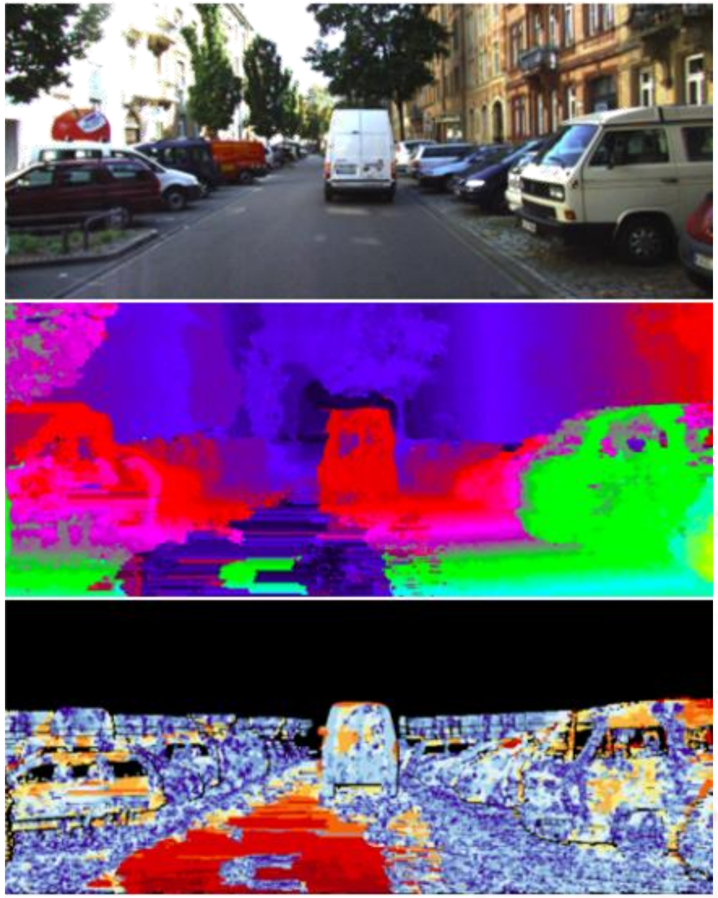
Repetitive patterns

Reflections

# Stereo Matching—Classic Pipeline

- Matching cost computation
  - Compute 3D cost volume along horizontal scanline
- Cost Aggregation
  - Refine matching cost by aggregating information of neighboring pixels
- Disparity Computation
  - Derive the disparity from the matching cost
- Disparity Refinement
  - Post-processing on the disparity map

Typical method: Semi-Global Matching (SGM)



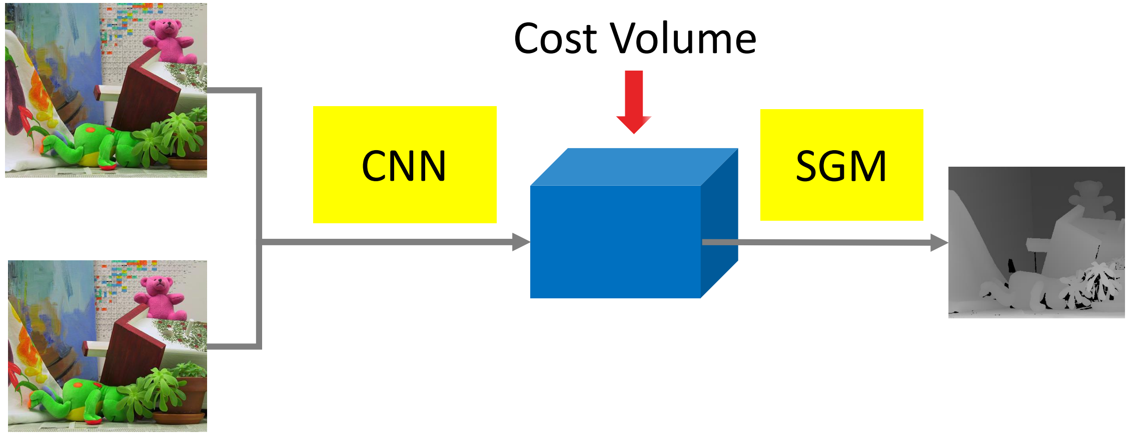

**Erroneous at ill-posed region!**

[2] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information", *IEEE Trans. on pattern analysis and machine intelligence*, 2008.

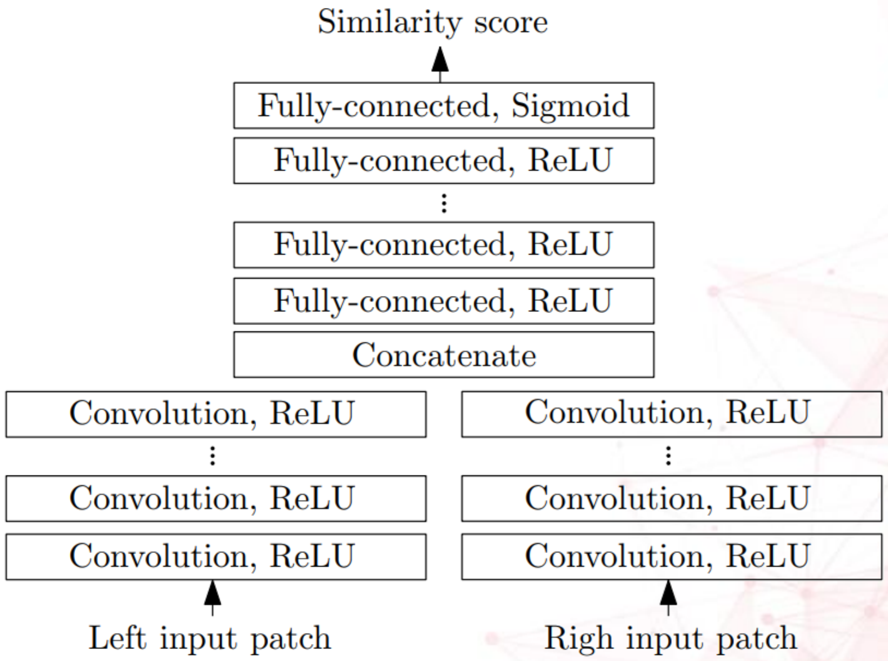
# Stereo Matching Meets Deep Learning (I)

## Matching cost learning

- Train a model to classify patches into two classes (similar and not similar)
- A small set of image pairs with ground-truth disparities generate millions of patches
- Depend on the performance of SGM




Overall pipeline



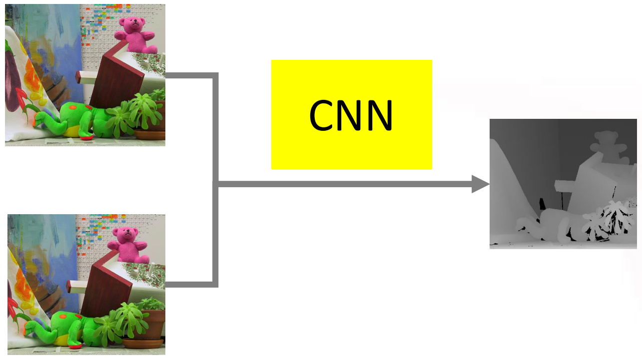

CNN Architecture of [3]

[3] J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 2016.

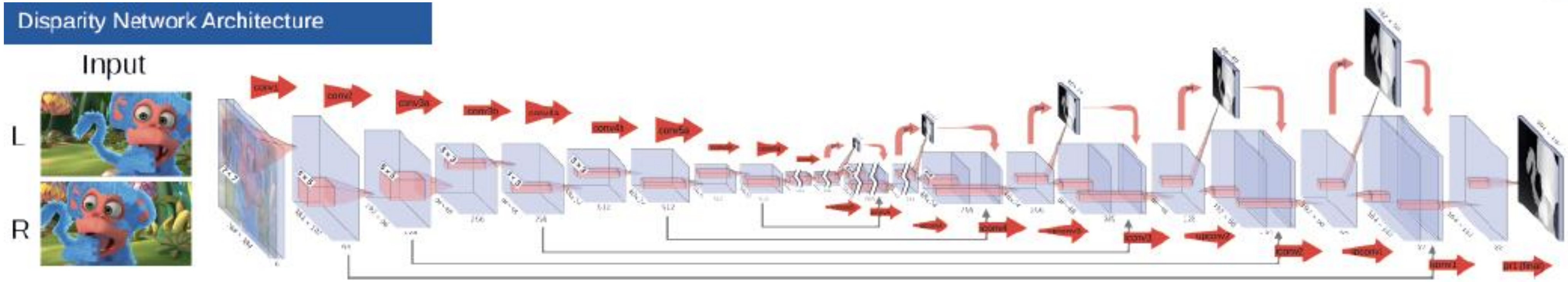
# Stereo Matching Meets Deep Learning (II)

## End-to-end Learning

- Train an end-to-end model to regress disparity
- A large-scale dataset is needed to train a good model
- Usually via hourglass structure
- Variations:
  - Correlation layer to compute cost volume, e.g., DispNetC [4]
  - Unsupervised learning with left-right check, e.g., [5]




Overall pipeline




[4] N. Mayer, et al,. "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation." In *Proc. IEEE CVPR*, 2016.

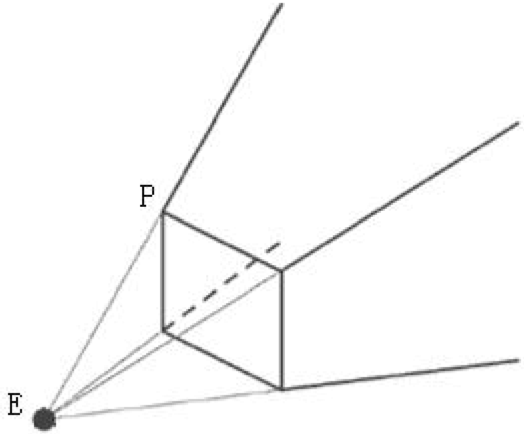
[5] C. Zhou, H. Zhang, X. Shen, and J. Jia. "Unsupervised Learning of Stereo Matching." In *Proc. IEEE CVPR*, 2017.

# Outline

# Single Image Depth Estimation

## Estimate depth from a single image

- Monocular depth estimation: $D = F(I)$
- Highly-ill posed, infinite configurations
- Deep learning to rescue!

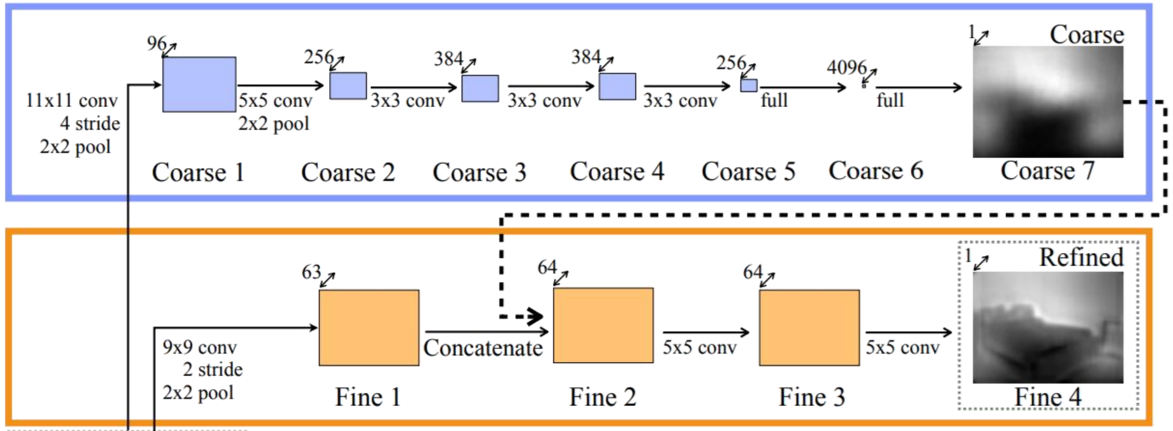



Needs semantic info to understand the scene

# Single Image Depth Estimation Meets DL (I)

## Direct regression with CNN

- Train a model end-to-end to regress scene depth directly, e.g., [6]
- Issue: generalization, requires lots of data





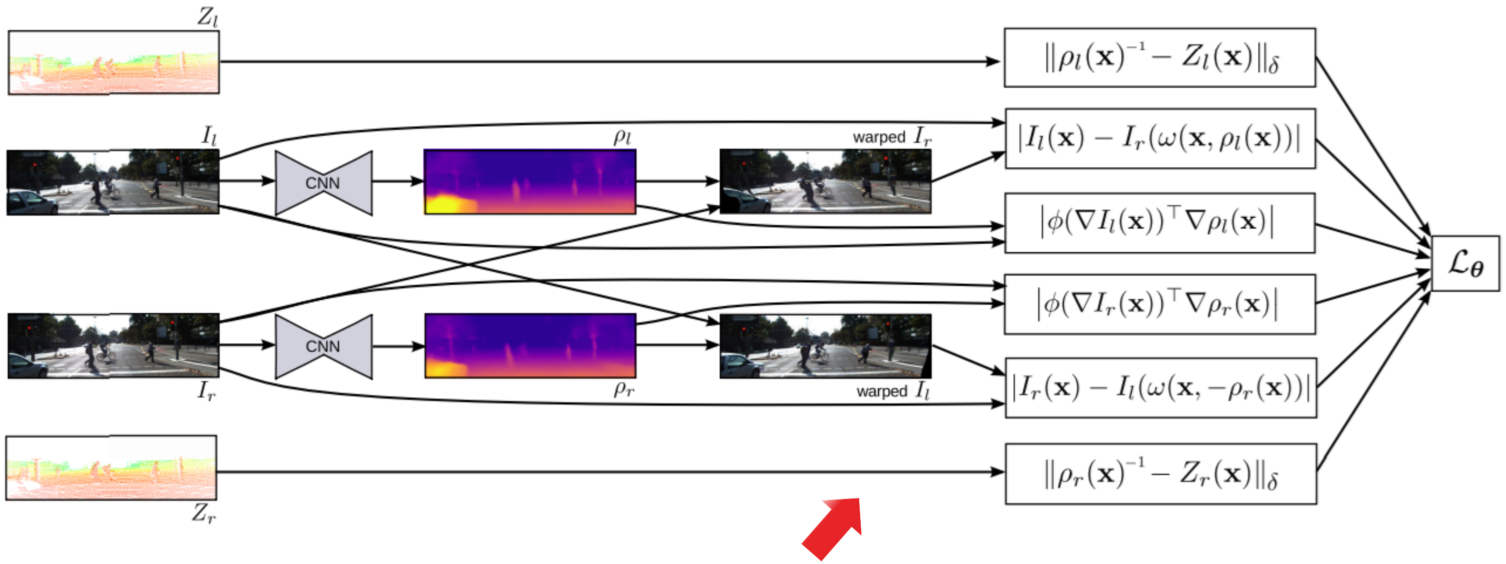
| | | Coarse | | | | | Fine |
|---|---|---|---|---|---|---|---|
| Layer | input | 1 | 2,3,4 | 5 | 6 | 7 | 1,2,3,4 |
| Size (NYUDepth) | 304x228 | 37x27 | 18x13 | 8x6 | 1x1 | 74x55 | 74x55 |
| Size (KITTI) | 576x172 | 71x20 | 35x9 | 17x4 | 1x1 | 142x27 | 142x27 |
| Ratio to input | /1 | /8 | /16 | /32 | – | /4 | /4 |

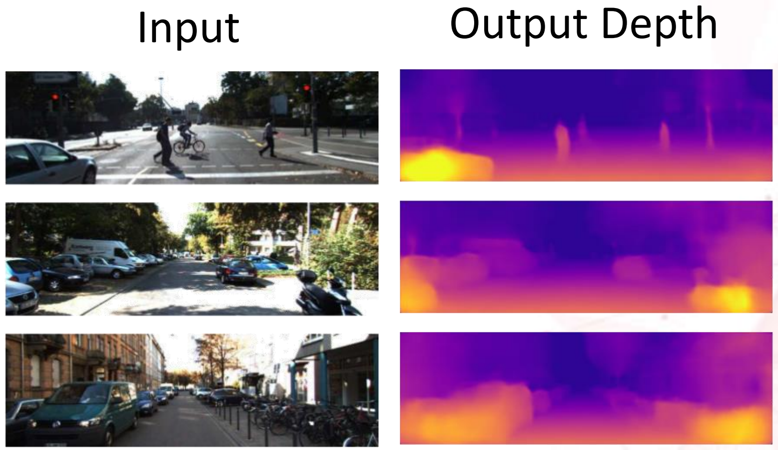

[6] D. Eigen, C. Puhrsch, R. Fergus "Depth map prediction from a single image using a multi-scale deep network." In *Proc. NIPS*, pp. 2366-2374, 2016.

# Single Image Depth Estimation Meets DL (II)

## Semi-supervised/Unsupervised schemes

- Use stereo pairs for training
- Based on left-right consistency and smoothness of depth




[8] defines a series of loss functions based on left-right consistency and depth smoothness

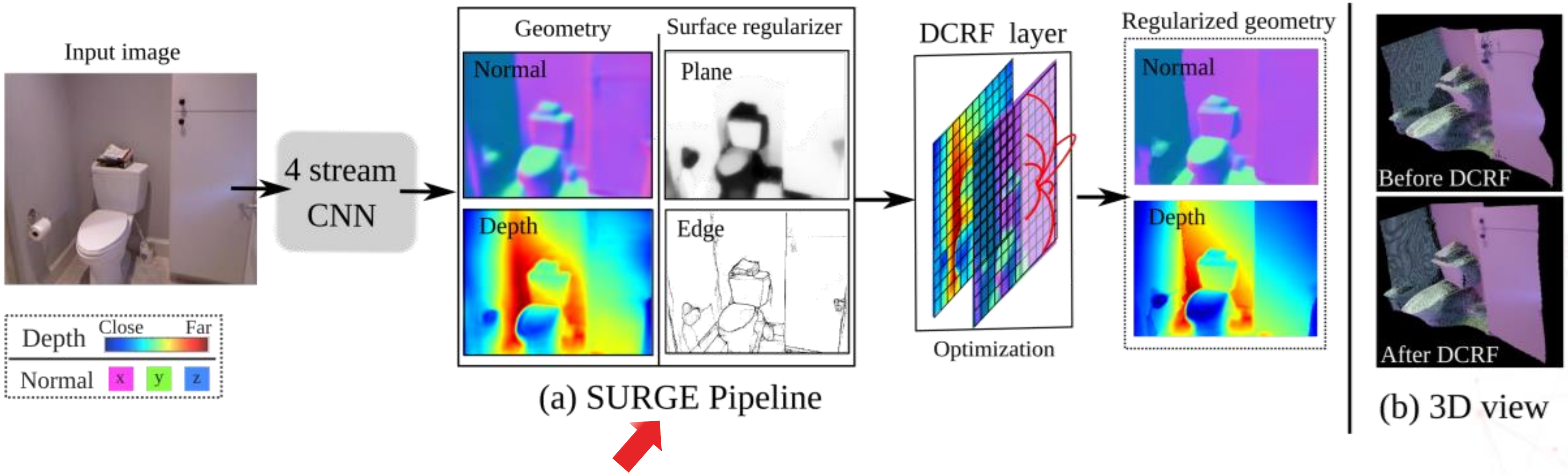

[7] C. Godard, O. Mac Aodha, G.J. Brostow"Unsupervised monocular depth estimation with left-right consistency," In *Proc. CVPR*, 2017.

[8] Y Kuznietsov, J Stückler and L. Bastian, "Semi-supervised deep learning for monocular depth map prediction," In *Proc. CVPR*, 2017.

# Single Image Depth Estimation Meets DL (III)

## CNN with expressive modeling

- Integrate geometric models (e.g., [9]) or image models (e.g., CRF, [10]) into CNN




[9] classifies pixels into planes and edges, enforce two points on the same plane have same normal

[9] P. Wang, et al., "SURGE: surface regularized geometry estimation from a single image," In *Proc. NIPS*, pp. 172-180, 2016.

[10] F. Liu, et al., "Deep convolutional neural fields for depth estimation from a single imag,." In *Proc. CVPR*, pp. 5162-5170, 2015.

# Part 2. Depth Learning—Our Progress

# Outline

Part 1. Introduction

- Motivation
- Stereo Matching
- Single Image Depth Estimation

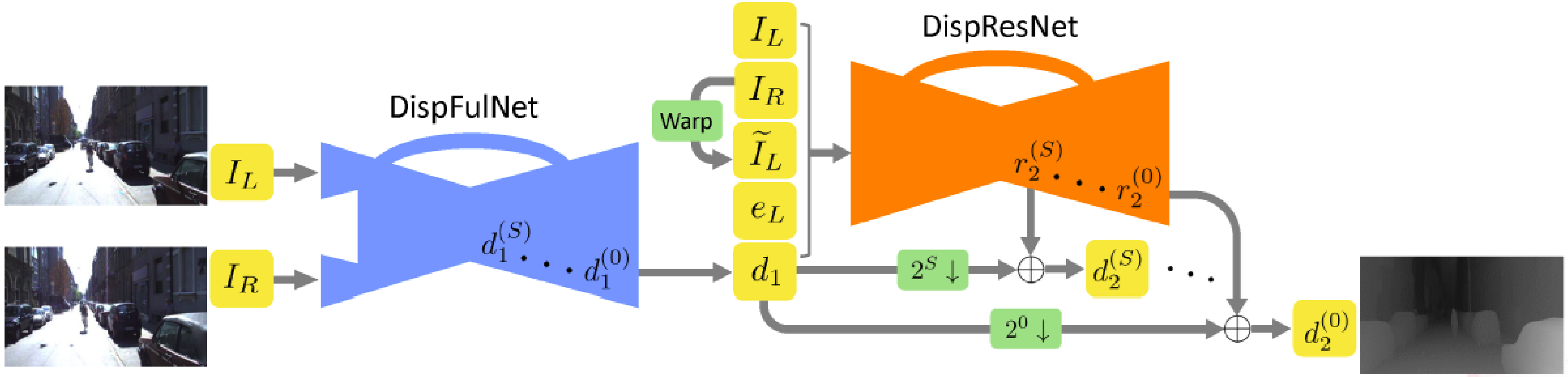
Part 2. Depth Learning—Our Progress

- Cascade Residual Learning (CRL)
- Zoom and Learn (ZOLE)
- Single View Stereo Matching (SVS)

Part 3. Conclusion

# Cascade Residual Learning: A Two-stage Convolutional Neural Network for Stereo Matching

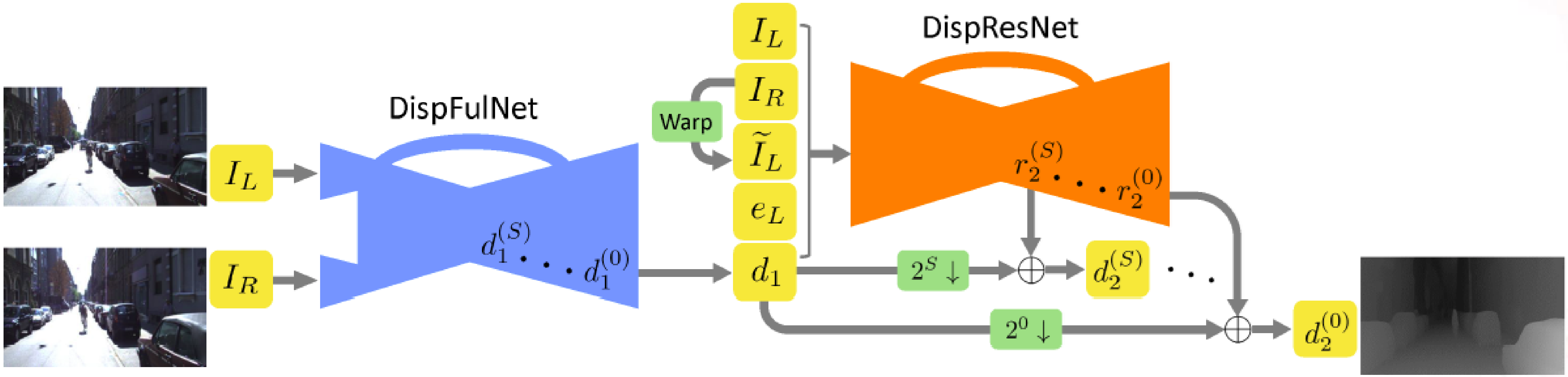
Jiahao Pang, Wenxiu Sun, Jimmy Ren, Chengxi Yang and Yan Qiong,
SenseTime Research




[11] J. Pang, et al., "Cascade residual learning: A two-stage convolutional neural network for stereo matching," In *Proc. ICCV Workshop*, 2017.

# Cascade Residual Learning (I)

- Cascade residual networks for accurate disparity estimation




Two-stage disparity estimation:

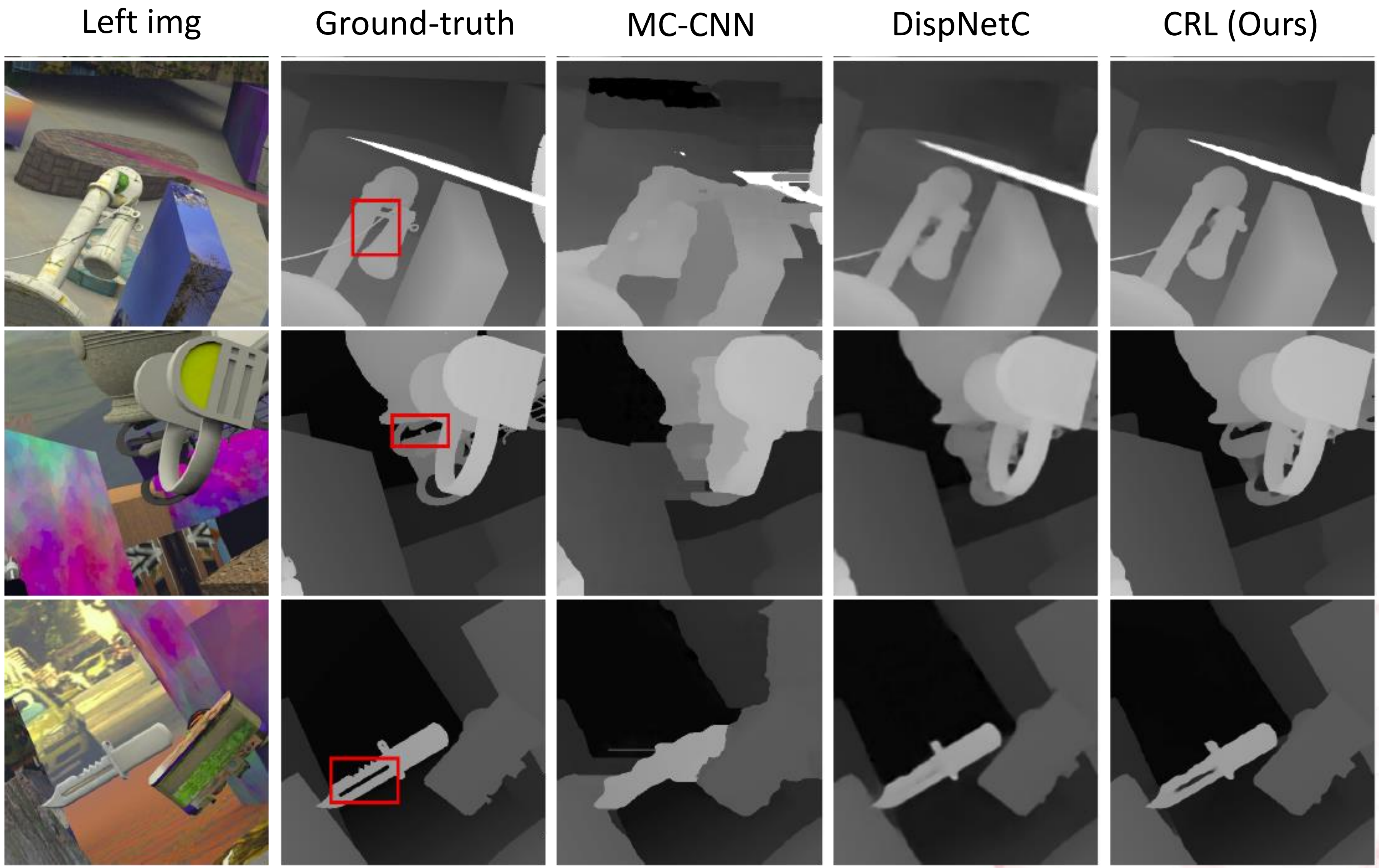
- **DispFulNet**: Resembles DispNet, use extra deconvolution modules to obtain full resolution output
- **DispResNet**: Inspired by ResNet, the second stage learns the residual signals across multiple scales

# Cascade Residual Learning (II)

- Ranked 1st on KITTI Stereo 2015 from Mar. 2017 – Sept. 2017

| | Method | Setting | Code | D1-bg | D1-fg | D1-all | Density | Runtime | Environment |
|---|---|---|---|---|---|---|---|---|---|
| 1 | CRL | | | 2.48 % | 3.59 % | 2.67 % | 100.00 % | 0.47 s | Nvidia GTX 1080 |
| J. Pang, W. Sun, J. Ren, C. Yang and Y. Qiong: Cascade Residual Learning: A Two-stage Convolutional Neural Network for Stereo Matching. arXiv preprint arXiv:1708.09204 2017. | | | | | | | | | |
| 2 | GC-NET | | | 2.21 % | 6.16 % | 2.87 % | 100.00 % | 0.9 s | Nvidia GTX Titan X |
| A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach and A. Bry: End-to-End Learning of Geometry and Context for Deep Stereo Regression. arXiv preprint arxiv:170 | | | | | | | | | |
| 3 | DRR | | | 2.58 % | 6.04 % | 3.16 % | 100.00 % | 0.4 s | Nvidia GTX Titan X |
| S. Gidaris and N. Komodakis: Detect, Replace, Refine: Deep Structured Prediction For Pixel Wise Labeling. arXiv preprint arXiv:1612.04770 2016. | | | | | | | | | |
| 4 | L-ResMatch | | code | 2.72 % | 6.95 % | 3.42 % | 100.00 % | 48 s | 1 core @ 2.5 Ghz (C/C++) |
| A. Shaked and L. Wolf: Improved Stereo Matching with Constant Highway Networks and Reflective Loss. arXiv preprint arxiv:1701.00165 2016. | | | | | | | | | |
| 5 | Displets v2 | | code | 3.00 % | 5.56 % | 3.43 % | 100.00 % | 265 s | >8 cores @ 3.0 Ghz (Matlab + C/C++) |
| F. Guney and A. Geiger: Displets: Resolving Stereo Ambiguities using Object Knowledge. Conference on Computer Vision and Pattern Recognition (CVPR) 2015. | | | | | | | | | |
| 6 | D3DNet | | | 2.88 % | 6.60 % | 3.50 % | 100.00 % | 0.35 s | Nvidia GTX Titan X |

# Cascade Residual Learning (III)



# Outline

Part 1. Introduction

- Motivation
- Stereo Matching
- Single Image Depth Estimation

Part 2. Depth Learning—Our Progress

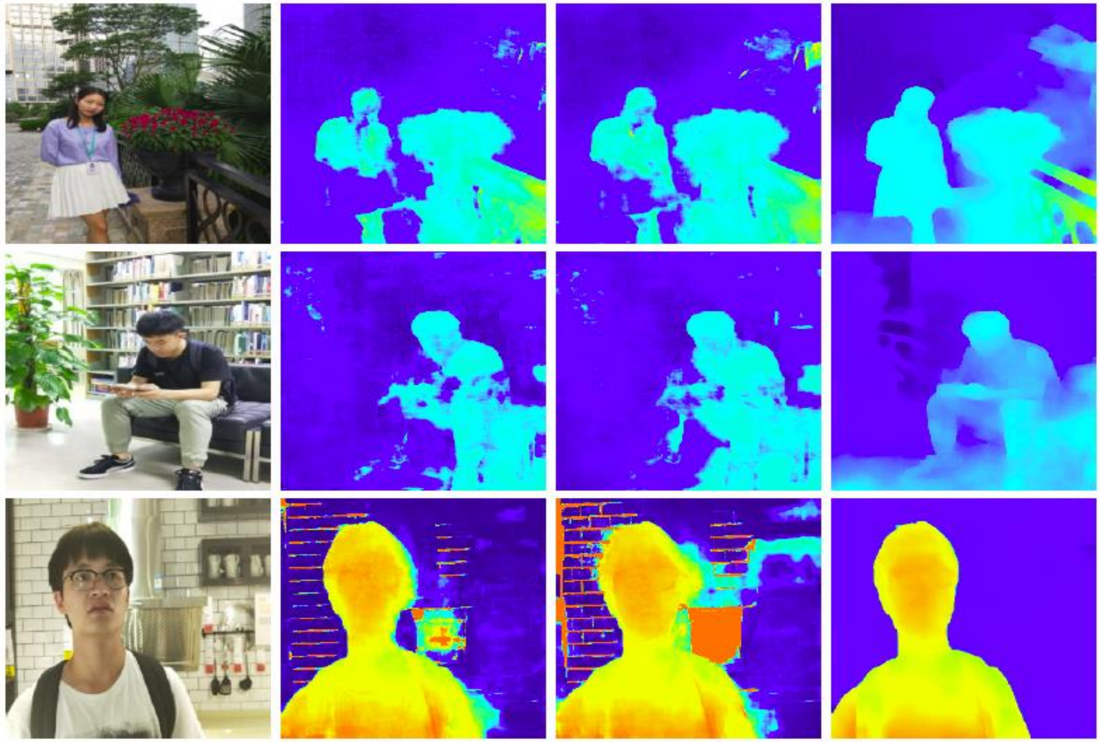
- Cascade Residual Learning (CRL)
- Zoom and Learn (ZOLE)
- Single View Stereo Matching (SVS)

Part 3. Conclusion

# Zoom and Learn: Generalizing Deep Stereo Matching to Novel Domains

Jiahao Pang[1], Wenxiu Sun[1], Chengxi Yang[1], Jimmy Ren[1], Ruichao Xiao[1], Jin Zeng[1], Liang Lin[1,2]

[1] SenseTime Research, [2] Sun Yat-sen University



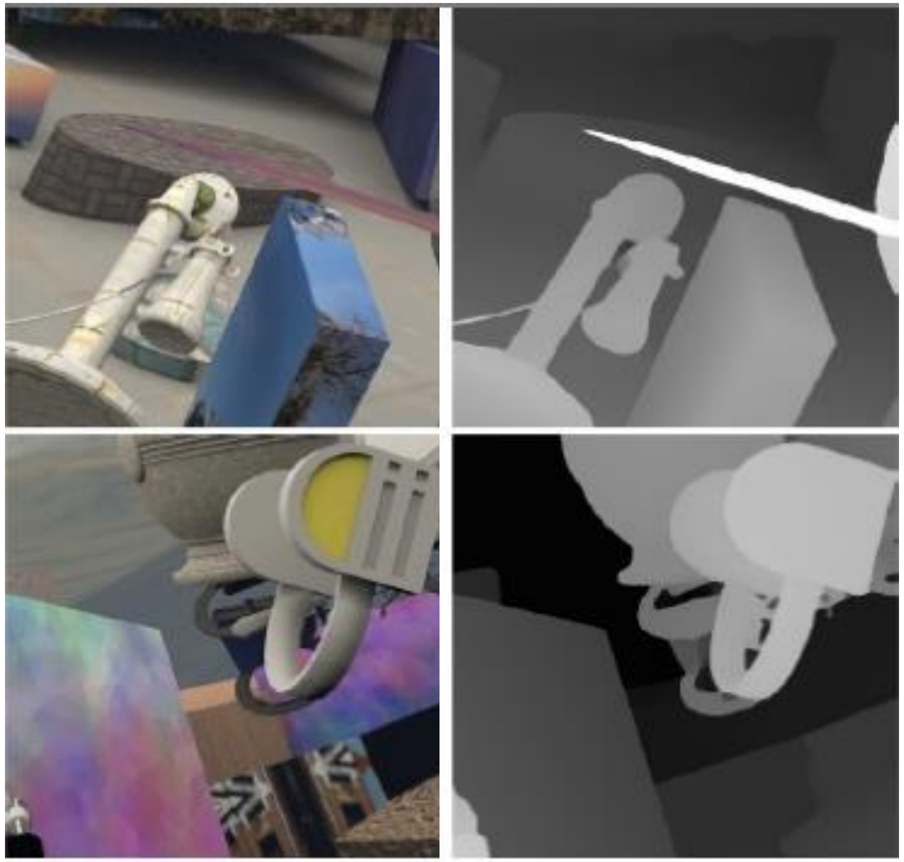
Input | Competing Methods | ZOLE (Ours)

[12] J. Pang, et al., "Zoom and Learn: Generalizing Deep Stereo Matching to Novel Domains," CVPR 2018.

# Zoom and Learn—Motivation

- Great results on pre-trained synthetic dataset

Source domain



However

- Directly apply model to a new domain brings horrible results

Target domain

# Zoom and Learn—Method (I)

- Finetuning in the target domain?
  No, hard to collect ground-truth depth in new domain, *e.g*. bokeh
- Our solution: a self-adaptation approach

| Synthetic data, with ground-truth | + | Stereo pairs in new domain, **no** ground-truth |
|---|---|---|

- Key observation: up-sampled stereo pair leads to disparity map with extra details
- Stereo matching CNN $S$ parameterized by $\boldsymbol{\Theta}$, stereo pair $P$ , consider two schemes

Scheme A

$$D = S\left(P; \boldsymbol{\Theta}\right)$$

Scheme B

$$D' = \frac{1}{r} \cdot \downarrow_r \left(S\left(\uparrow_r(P); \boldsymbol{\Theta}\right)\right)$$

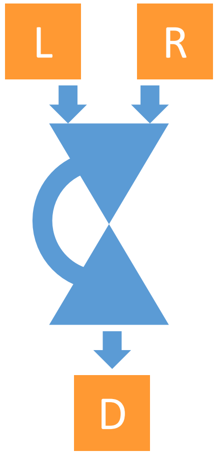
# Zoom and Learn—Method (II)

$$D' = \frac{1}{r} \cdot \downarrow_r \left( S \left( \uparrow_r (P); \mathbf{\Theta} \right) \right)$$

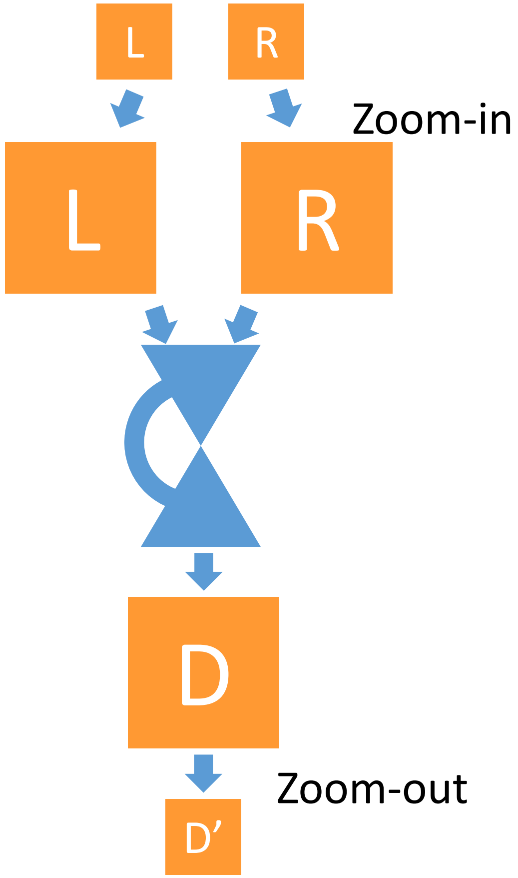
Scheme B brings more details

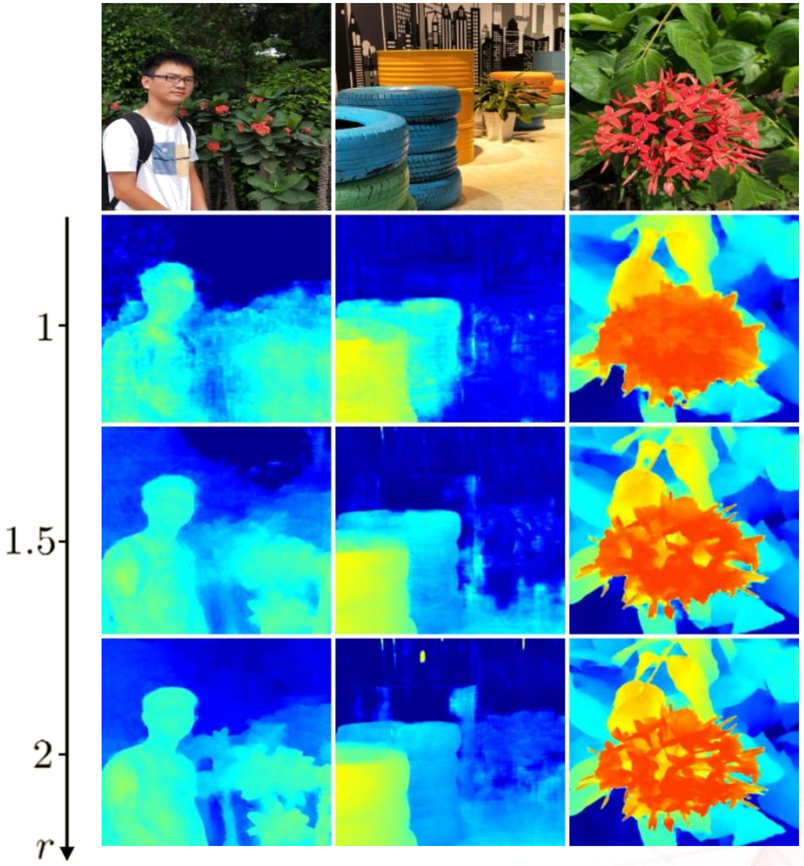
$$D = S \left( P; \mathbf{\Theta} \right)$$

Scheme A




Scheme B



# Zoom and Learn—Method (III)

- However
  - A bigger *r* does not necessarily mean better results
  - Performance first improves then deteriorates

Results of DispNetC on KITTI 2015

| Network | Resolution | | | | |
|---|---|---|---|---|---|
| | 896 | 1280 | 1664 | 2048 | 2432 |
| DispNetC | 14.26% | 9.97% | **8.81%** | 9.17% | 10.53% |
| DispNetS | 18.95% | 11.61% | 9.18% | **8.64%** | 9.08% |

- Analysis
  - Up-sampling means perform stereo matching at subpixel accuracy
  - Larger input translates to smaller receptive field

Strategy
- Let the CNN **learn its own** higher-res details
- Graph Laplacian regularization to keep desired edges while smoothing out artifacts

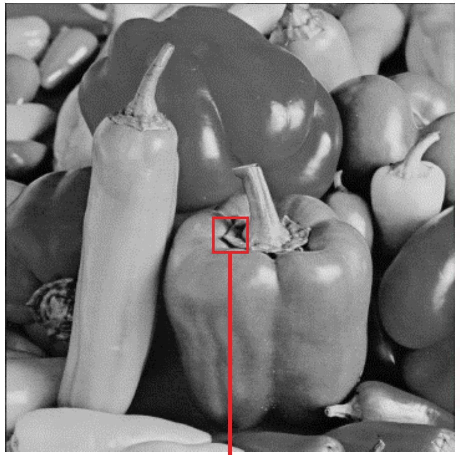
# Zoom and Learn—Method (IV)

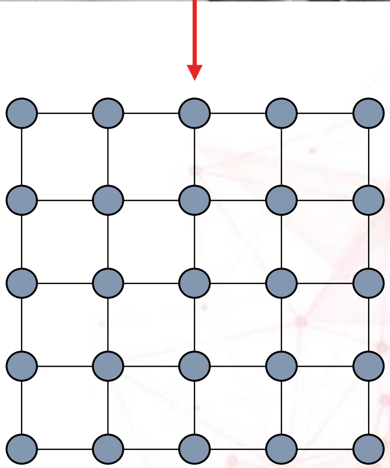
- Minimize graph Laplacian regularizer: widely used for image restoration

$$\mathbf{s}^{\mathrm{T}}\mathbf{L}\mathbf{s} = \sum_{(i,j)\in\mathcal{E}} w_{ij}(\mathbf{s}(i) - \mathbf{s}(j))^2$$

$\mathbf{L}$ : Graph Laplacian matrix

- Bigger weights $w_{ij}$ , higher similarity of pixels $i$ and $j$, force $\mathbf{s}(i)$ close to $\mathbf{s}(j)$
- Graph Laplacian regularization—simple & perform well on many restoration problems
  - Denoising
  - Super-resolution
  - Deblurring
  - Dequantization of JPEG images
  - Bit-depth enhancement
- We integrate it as a loss function in our work





Graph of a $5 \times 5$ patch,
(may not be a grid graph)



[13] J. Pang and G. Cheung, “Graph Laplacian regularization for image denoising: Analysis in the continuous domain,” *IEEE Trans. Image Process*, 2017.

# Zoom and Learn—Method (V)

- Settings

$S(\cdot;\boldsymbol{\Theta}^{(0)})$ : init stereo network pre-trained with synthetic data, $\boldsymbol{\Theta}^{(0)}$ : model parameter

$N$ stereo pairs $P_i = (L_i, R_i), 1 \le i \le N$ for training

The first $N_{\text{dom}}$ pairs are real stereo pairs of the target domain, no ground-truth

The rest $N_{\text{syn}} = N - N_{\text{dom}}$ pairs are synthetic data, they have ground truth disparities $D_i$

- We solve for a new set of model parameter $\boldsymbol{\Theta}^{(k+1)}$ at iter. $k$
  - First create a set of pseudo ``ground-truths'' for the $N_{\text{dom}}$ real stereo pairs by zooming (up-sampling)

$$D_i = \frac{1}{r} \cdot \downarrow_r \left(S\left(\uparrow_r(P_i);\boldsymbol{\Theta}^{(k)}\right)\right), 1 \le i \le N_{\text{dom}}$$

  - $D_i \in \mathbb{R}^{Mm}$ , we divide a disparity map $D_i$ into $M$ square patches
  - Matrix extracting the $j$ -th patch from $D_i$ is denoted as $\mathbf{R}_j$, $\mathbf{R}_j \in \mathbb{R}^{m \times Mm}$

# Zoom and Learn—Method (VI)

- Formulation

Data term: drives $\mathbf{s}_{ij}$ to be similar to $\mathbf{d}_{ij}$

Smoothness term: graph Laplacian regularization

$$\boldsymbol{\Theta}^{(k+1)} = \arg\min_{\boldsymbol{\Theta}} \sum_{i=1}^{N_{\text{dom}}} \sum_{j=1}^{M} \|\mathbf{s}_{ij} - \mathbf{d}_{ij}\|_1 + \lambda \cdot \mathbf{s}_{ij}^{\mathrm{T}} \mathbf{L}_{ij}^{(k)} \mathbf{s}_{ij} + \tau \cdot \sum_{i=N_{\text{dom}}+1}^{N} \|S(P_i; \boldsymbol{\Theta}) - D_i\|_1,$$

$$\text{s.t.} \ \ \mathbf{s}_{ij} = \mathbf{R}_j \cdot \text{vec}\left(S(P_i; \boldsymbol{\Theta})\right), \ \ \mathbf{d}_{ij} = \mathbf{R}_j \cdot \text{vec}\left(D_i\right)$$

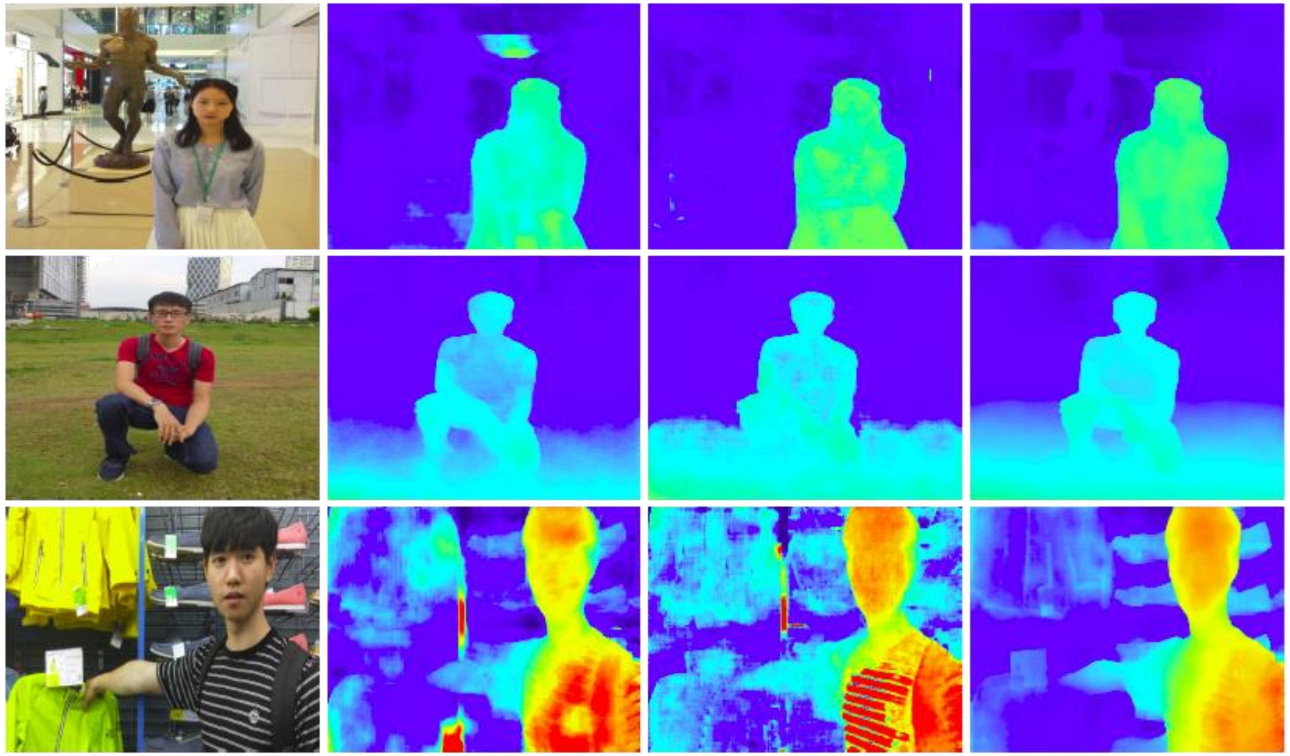
Feasibility: a stereo model works well for the target domain should perform reasonably on the synthetic data

  - graph Laplacian $\mathbf{L}_{ij}^{(k)}$ are pre-computed, based on the left image and the predictions

- In practice, we resolve the optimization with stochastic gradient descent

# Zoom and Learn—Results (I)

- Refer to our paper for the detailed setting



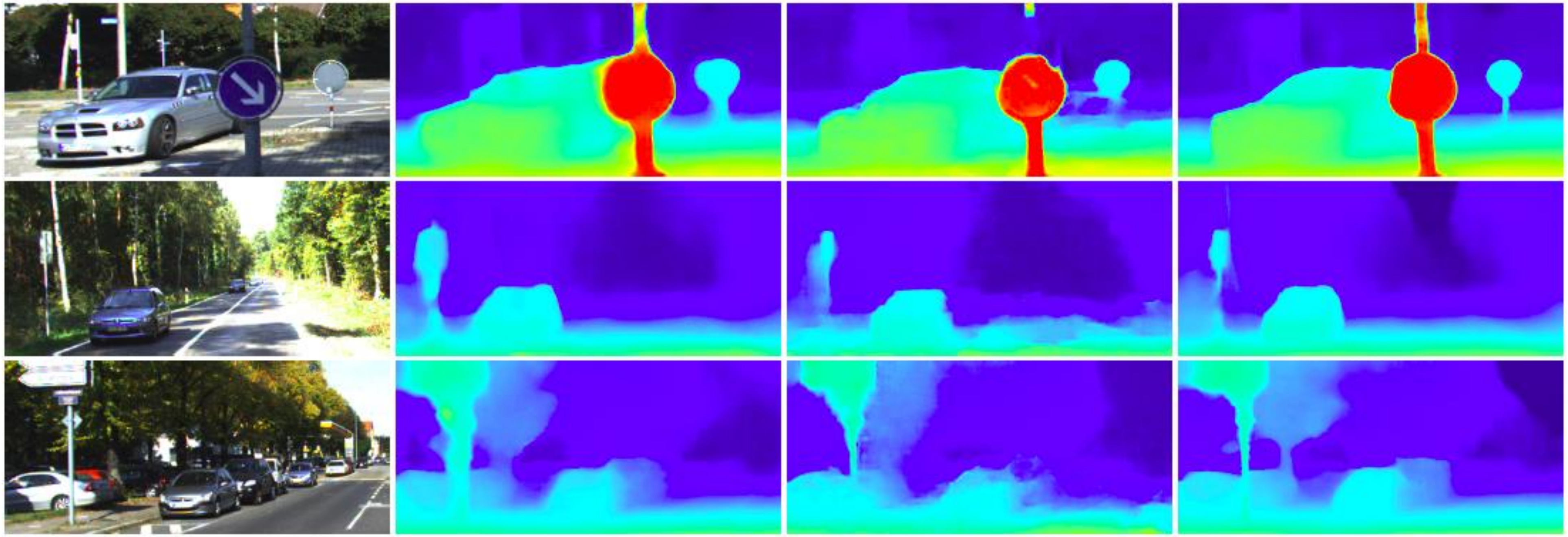
Left img  Tonioni *et al.* [14]  DispNetC  ZOLE (Ours)

| Dataset | Metric | | Model | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Tonioni | | DispNetC | | DispNetC-80 | | ZOLE | |
| Smartphone | PSNR | SSIM | 22.92 | 0.845 | 21.99 | 0.790 | 22.39 | 0.817 | **23.12** | **0.855** |
| FlyingThings3D-80 | EPE | 3ER | 1.08 | 6.79% | 1.03 | 5.63% | **0.93** | **5.11%** | 1.11 | 6.54% |

[14] A. Tonioni, M. Poggi, S. Mattoccia, and L. Di Stefano. “Unsupervised adaptation for deep stereo,” In *IEEE ICCV*, Oct 2017.

# Zoom and Learn—Results (II)

- Refer to our paper for detailed setting



Input | Tonioni *et al.* | DispNetC | ZOLE (Ours)

| Metric | Model | | |
|---|---|---|---|
| | Tonioni | DispNetC | ZOLE |
| EPE | 1.27 | 1.64 | **1.25** |
| 3ER | 7.06% | 11.41% | **6.76%** |

# Outline

Part 1. Introduction

- Motivation
- Stereo Matching
- Single Image Depth Estimation

Part 2. Depth Learning—Our Progress

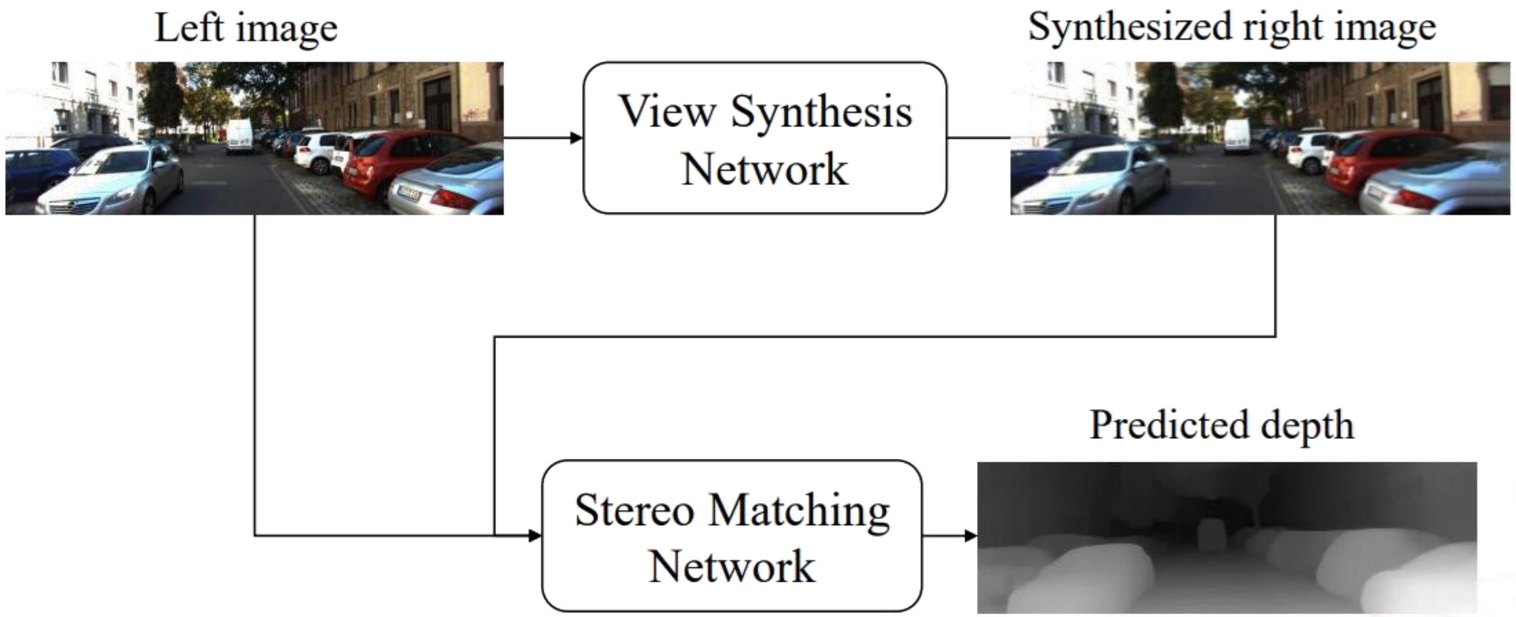
- Cascade Residual Learning (CRL)
- Zoom and Learn (ZOLE)
- Single View Stereo Matching (SVS)

Part 3. Conclusion

# Single View Stereo Matching

Yue Luo[1], Jimmy Ren[1], Mude Lin[1,2], Jiahao Pang[1], Wenxiu Sun[1], Hongsheng Li[3] , Liang Lin[1,2]

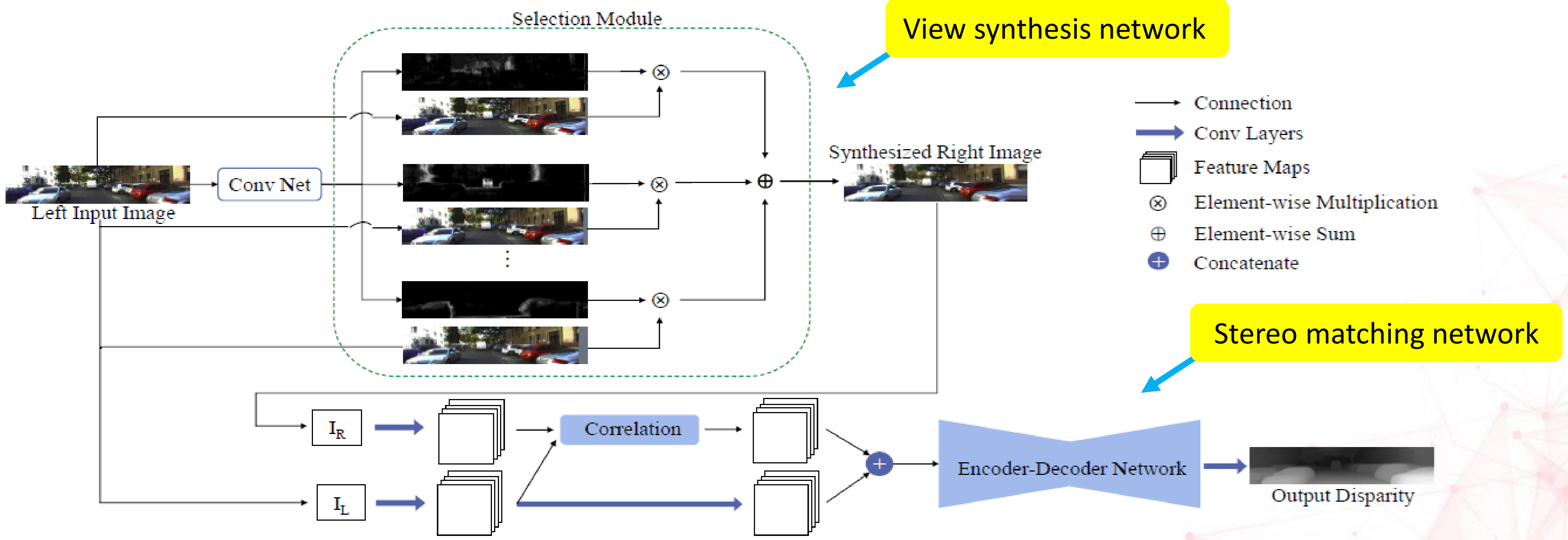
[1]SenseTime Research, [2]Sun Yat-sen University, [3]CUHK




[15] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, L. Lin, “Single view stereo matching, ” CVPR 2018. (Spotlight)

# Single View Stereo Matching—Method (I)

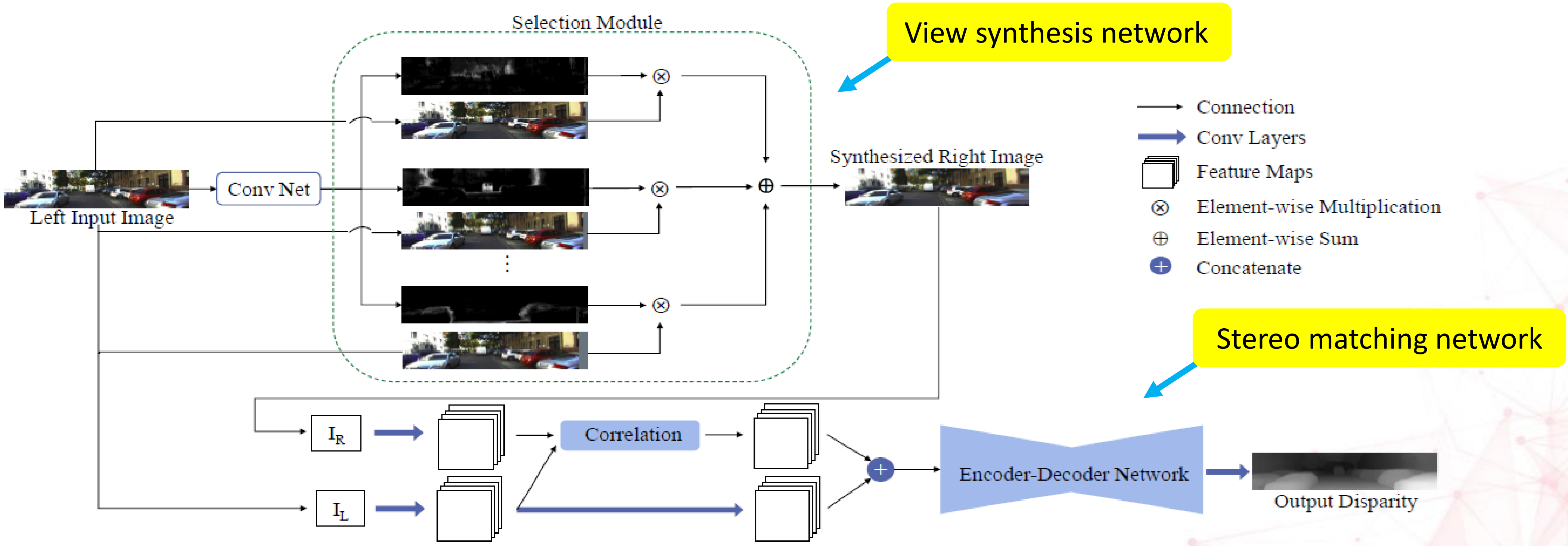
- Decomposed into two stages: view synthesis & stereo matching
- View synthesis network: based on Deep3D [16]
  - Formulate the left-to-right transformation with differentiable selection module
- Stereo matching network: based on DispFulNet of our CRL



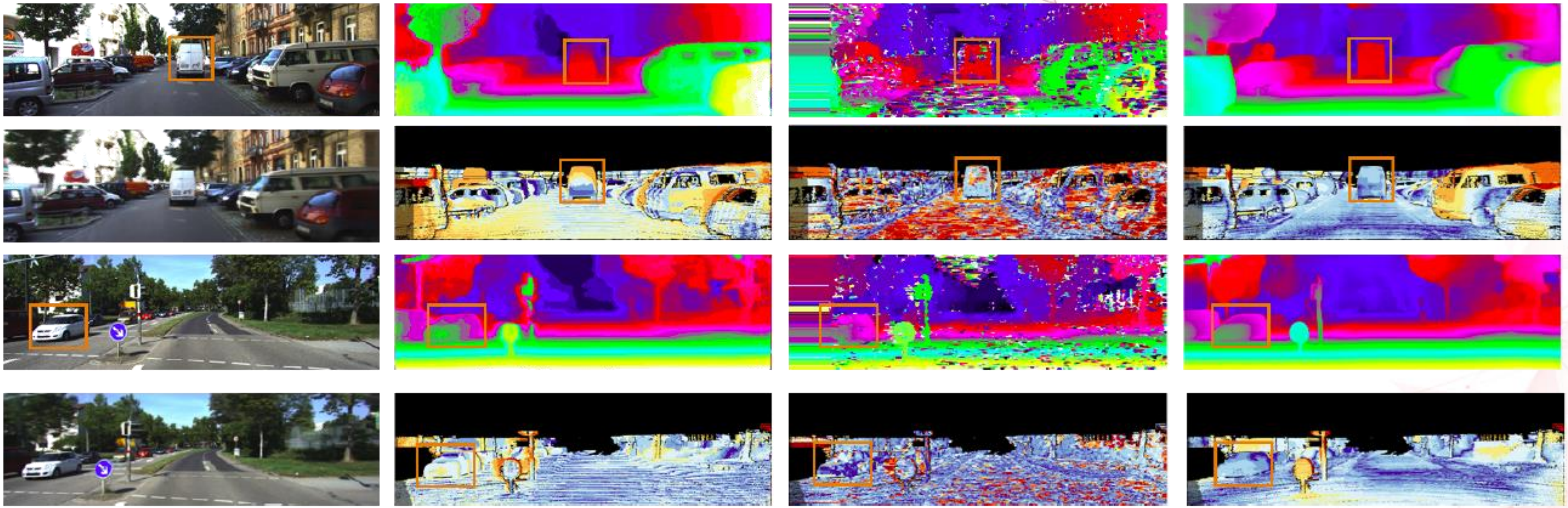

[16] J Xie, R Girshick, A Farhadi , "Deep3D: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks," ECCV 2016.

# Single View Stereo Matching—Method (II)

- Benefits
  - Explicitly encode the geometric transformation within individual networks to better solve the problem
  - Demand on labelled depth data is greatly alleviated



# Single View Stereo Matching—Results (I)



Input | Gordard *et al.* [7] | OCV-BM (Stereo) | SVS (Ours)

| Method | D1-bg | D1-fg | D1-all |
|---|---|---|---|
| Godard *et al.* | 27.00 | 28.24 | 27.21 |
| OCV-BM | **24.29** | 30.13 | 25.27 |
| Ours | 25.18 | **20.77** | **24.44** |

- Ours produce sharp edges
- Outperforms stereo block matching method!

# Single View Stereo Matching—Results (I)

Qualitative results on KITTI Eigen test set

Qualitative results on Make3D dataset (top two rows) and Cityscapes dataset (bottom two rows) using the model trained on KITTI

# Single View Stereo Matching—Results (II)

# Conclusion

Part 1. Introduction

- Motivation
- Stereo Matching
- Single Image Depth Estimation

Part 2. Depth Learning—Our Progress

- Cascade Residual Learning (CRL)
- Zoom and Learn (ZOLE)
- Single View Stereo Matching (SVS)

Part 3. Conclusion

# Part 3. Conclusion

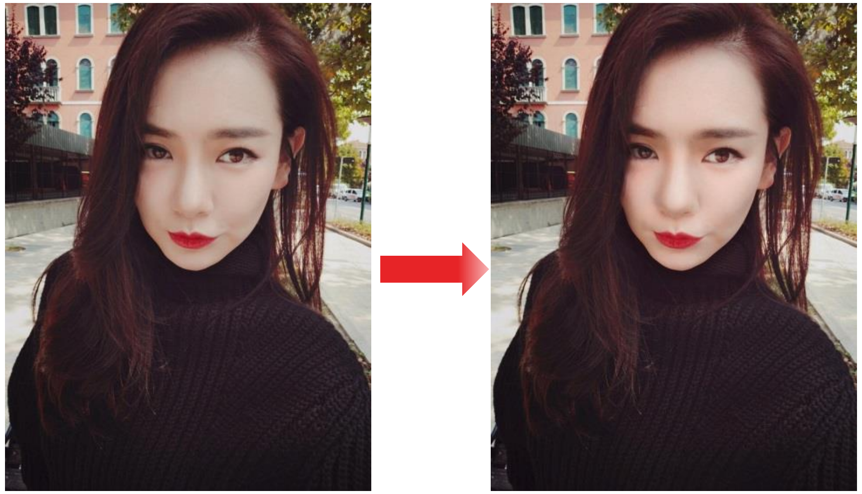
# Conclusion

- Depth estimation is important in many areas
- Deep learning brings a new era for depth estimation
- Pure data-driven approach can be fragile (e.g., single image depth estimation)
- Combine data-driven approaches and model-based approaches is a new trend
- Welcome to SenseTime to MAKE  IT HAPPEN!

# About Our Team: Image Restoration and Enhancement

Depth estimation, computational photography, image processing on smartphones


Our bokeh solution on recently launched VIVO V9



3D relighting on smartphones

**Get involve!** Internship and full-time employee are welcome

Location: Hong Kong, Shenzhen, Beijing, Shanghai, etc

# Thank You!

# Q & A