Learning Stereo Matchability in Disparity Regression Networks

Jingyang Zhang[1], Yao Yao[1], Zixin Luo[1], Shiwei Li[2],
Tianwei Shen[1], Tian Fang[2], Long Quan[1]

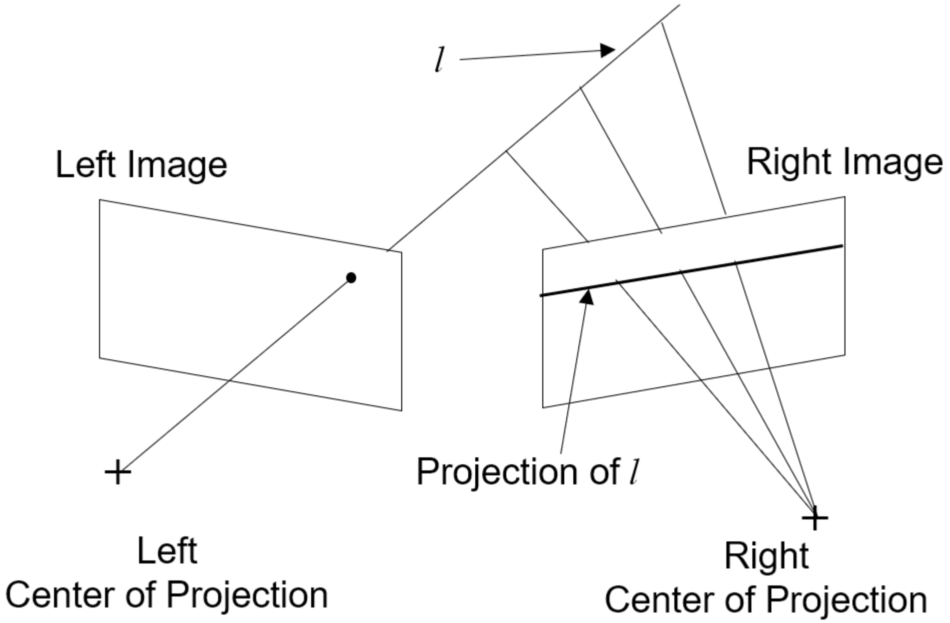
香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

[1]The Hong Kong University of Science and Technology

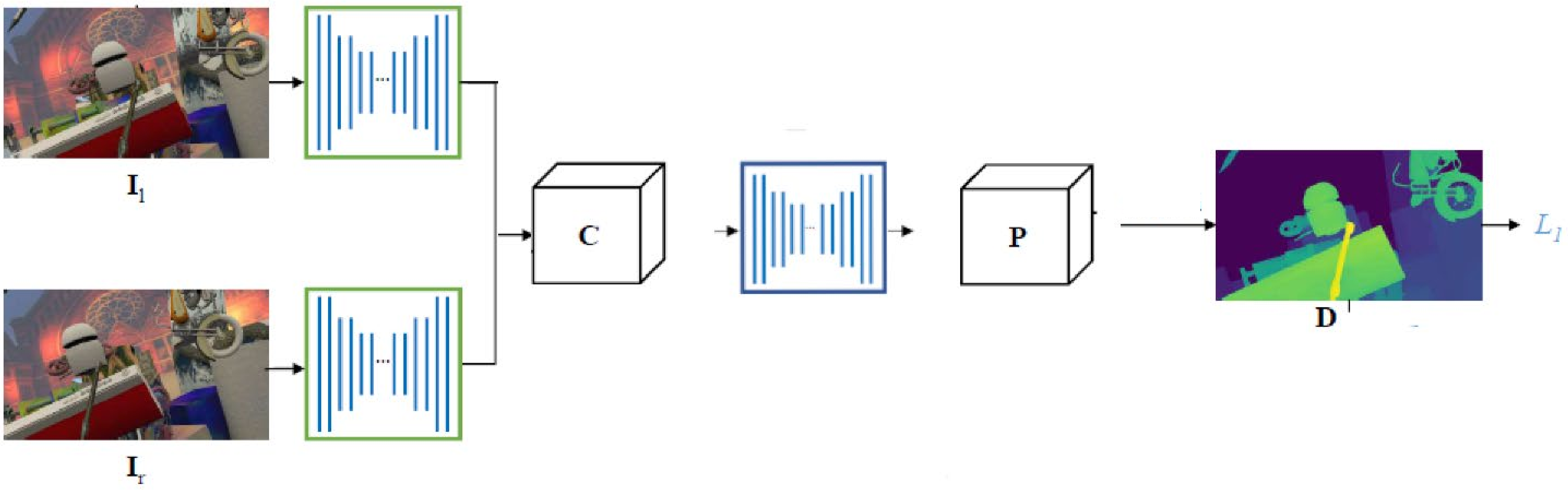
[2]Everest Innovation Technology

Introduction of Stereo Matching

- Input: Left and right rectified images
- Output: Per pixel disparity map aligned with the left image
- Key idea: The corresponding pixels should be photo-consistent



Learning-based Stereo



Feature Extraction

Feature Aggregation

Volume Regularization

Soft Argmax

L1 Loss

Matchability issue

- No or multiple hypothesis with high photo-consistency
  - Occlusion: e.g. object boundary
  - Non-Lambertian: e.g. specular surface
  - Textureless: e.g. large plain with single color

Solution in Previous Methods

- Traditional methods
  - Check the uni-modality of the probability over all the hypothesis
- Learning-based methods
  - Directly estimate the confidence/uncertainty from input image

Matchability to Uncertainty

- Define Matchability as the entropy of the estimated probability distribution

$$M(x,y) = \sum_{d=0}^{D-1} p(x,y,d) \log p(x,y,d)$$

- Transform the matchability to uncertainty by a simple 2D CNN

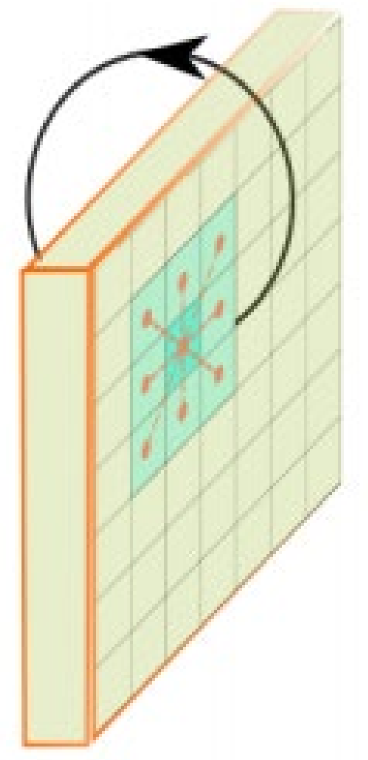
Joint Estimation of Disparity and Uncertainty

- Training
  - Assume a Laplacian distribution $p(x|\mu, b) = \frac{1}{2b}\exp(-\frac{|x-\mu|}{b})$
  - Model disparity as the location parameter μ, uncertainty as the scale parameter b
  - Let d be the estimated disparity, u be the estimated uncertainty, d_{gt} be the ground truth depth
  - Minimize the negative log likelihood

$$L = \frac{1}{u}\left|d_{gt} - d\right| + \log u$$

Recover the Unmatchable pixels

- Recover the the unmatchable pixels by the neighboring values
  - Use convolutional spatial propagation network
  - Can be viewed as anisotropic diffusion
  - Diffusion kernel is different for each pixel
  - Diffusion kernel is estimated from image, disparity and matchability



System Overview

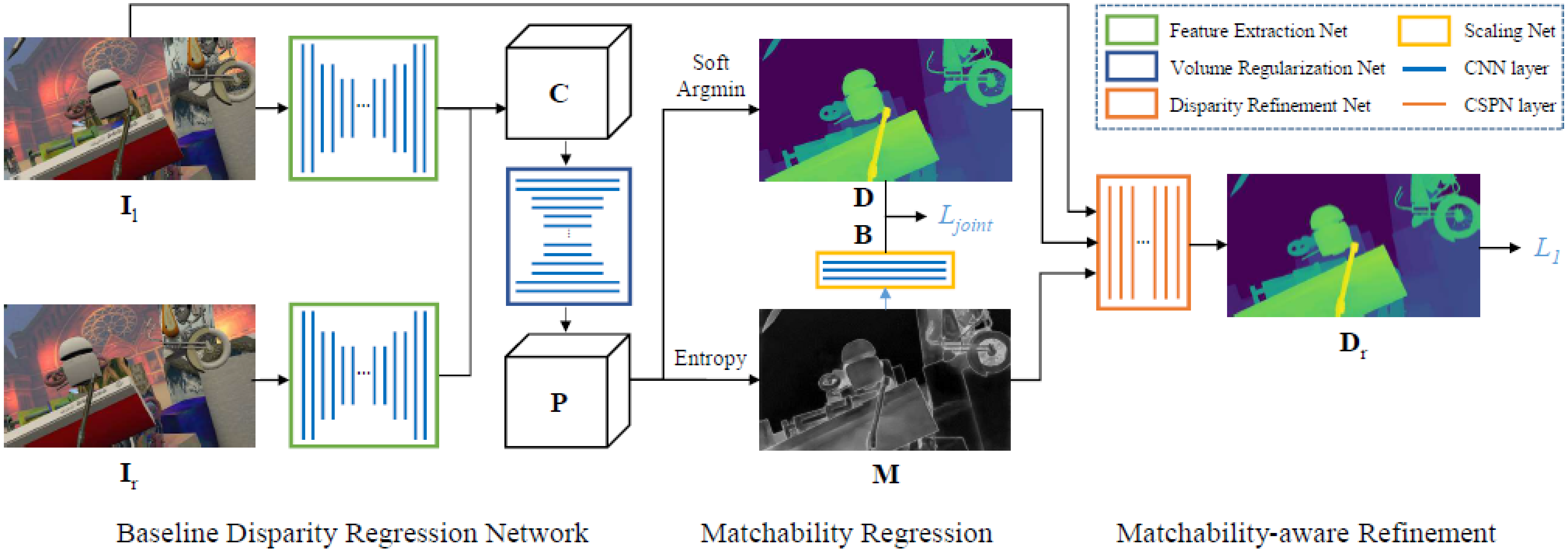
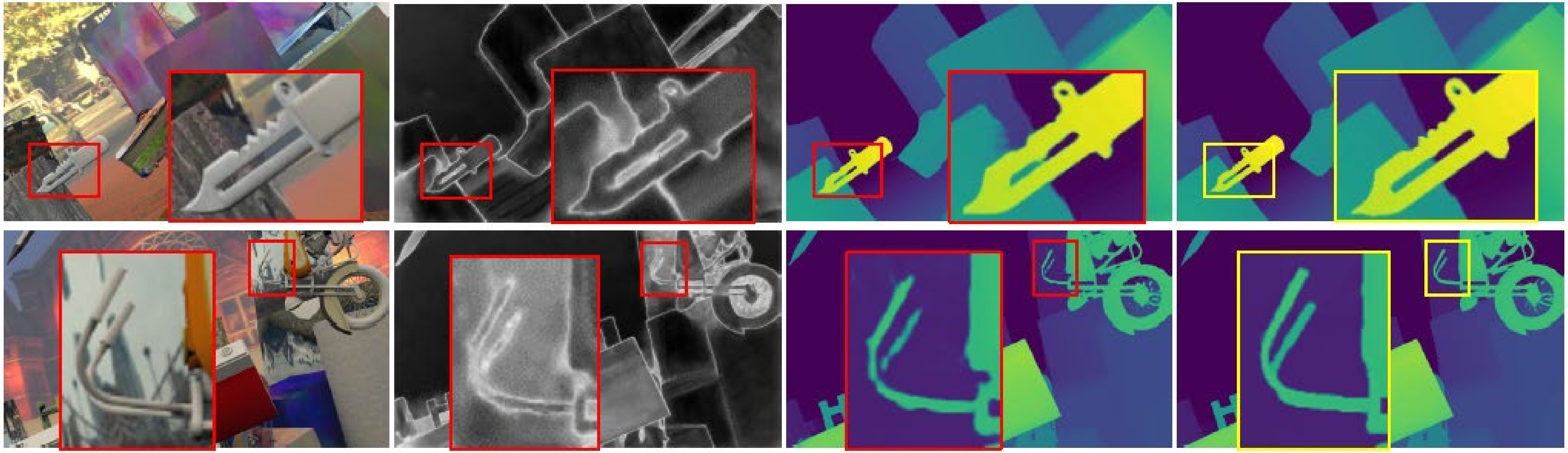


Fig. 1. The proposed framework. Our network contains a baseline disparity regression network, where the image features are extracted though a 2D UNet and the cost volume are regularized via a 3D UNet. The matchability and the initial disparity maps are respectively regressed from the probability volume using the *soft-argmin* and the *entropy* operations. Finally we use the matchability information and input image semantics to refine the disparity output.

Qualitative Results



(a) Left Image (b) Matchability Map (c) Initial Disparity Map (d) Refined Disparity Map

Fig. 2. Illustrations on intermediate results of the proposed network. From left to right: (a) the left input image; (b) the regressed matchability map; (c) the initial disparity map; (d) the refined disparity map. These two samples clearly shows the effectiveness of the matchability-aware disparity refinement.

Quantitative Results

TABLE I

QUANTITATIVE RESULTS ON KITTI 2012 & 2015 STEREO BENCHMARKS OVER NON-OCCLUDED REGIONS (NOC) AND ALL PIXELS (ALL). THE D1 ERROR IS THE PERCENTAGE OF PIXELS WITH DISPARITY ERROR LARGER THAN 3 PX AND 5% OF THE GROUND TRUTH.

Methods	KITTI 2015						KITTI 2012					
	Noc			All			Noc			All		
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	**D1-all**	>2px	>3px	EPE	>2px	>3px	EPE
DispNetC [7]	4.11 %	3.72 %	4.05 %	4.32 %	4.41 %	4.34 %	7.38 %	4.11 %	0.9 px	8.11 %	4.65 %	1.0 px
MC-CNN [13]	2.48 %	7.64 %	3.33 %	2.89 %	8.88 %	3.89 %	3.90 %	2.43%	0.7 px	5.45 %	3.63 %	0.9 px
GC-Net [1]	2.02 %	5.58 %	2.61 %	2.21 %	6.16 %	2.87 %	2.71 %	1.77 %	0.6 px	3.46 %	2.30 %	0.7 px
PSMNet [2]	1.71 %	4.31 %	2.14 %	1.86 %	4.62 %	2.32 %	2.44 %	1.49 %	0.5 px	3.01 %	1.89 %	0.6 px
DSM (Ours)	1.66 %	4.16 %	2.07 %	1.83 %	4.56 %	2.28 %	2.25 %	1.39 %	0.5 px	2.83 %	1.79 %	0.5 px
SegStereo [20]	1.76 %	3.70 %	2.08 %	1.88 %	4.07 %	2.25 %	2.66 %	1.68 %	0.5 px	3.19 %	2.03 %	0.6 px
GwcNet [19]	1.61 %	3.49 %	1.92 %	1.74 %	3.93 %	2.11 %	2.16 %	1.32 %	0.5 px	2.71 %	1.70 %	0.5 px
EdgeStereo [21]	1.69 %	2.94 %	1.89 %	1.84 %	3.30 %	2.08 %	2.32 %	1.46 %	0.4 px	2.93 %	1.83 %	0.5 px
GANet [4]	1.40 %	3.37 %	1.73 %	1.55 %	3.82 %	1.93 %	2.18 %	1.36 %	0.5 px	2.79 %	1.80 %	0.5 px
CSPN [3]	1.40 %	2.67 %	1.61 %	1.52 %	2.88 %	1.74 %	1.79 %	1.19 %	-*	2.27 %	1.53 %	-*

*Not reported by the paper or the benchmark

Light-weight Model

- Reduce the expensive 3D CNN to save time
- 20fps at 320x576 input

TABLE V
COMPARISON OF QUALITY AND RUNNING TIME BETWEEN THE LIGHTWEIGHT MODEL AND OTHER METHODS ON SCENEFLOW TEST SET.

Settings	EPE (px)	>1px (%)	>3px (%)	Time (s)
Baseline	0.875	9.07	4.30	0.32
DSM	0.761	8.31	4.07	0.34
Baseline (lightweight)	0.952	9.66	4.56	0.15
DSM (lightweight)	0.806	8.75	4.08	0.17

Thank you

Code available at https://github.com/jzhangbs/DSM

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY