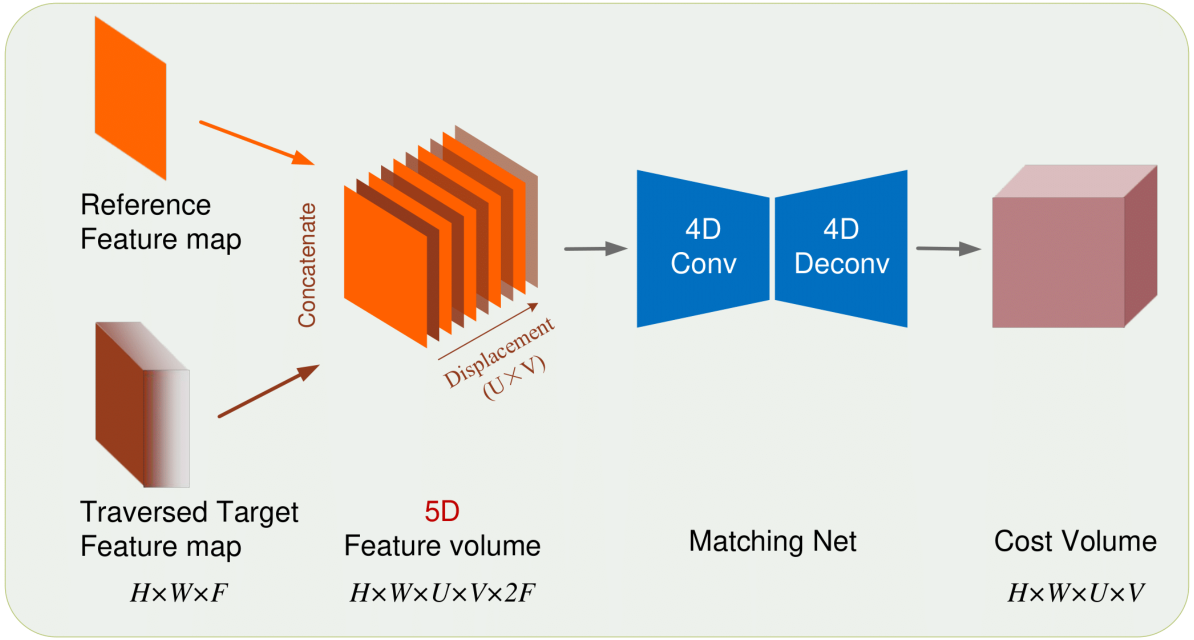
Displacement-Invariant Matching Cost Learning for Accurate Optical Flow Estimation

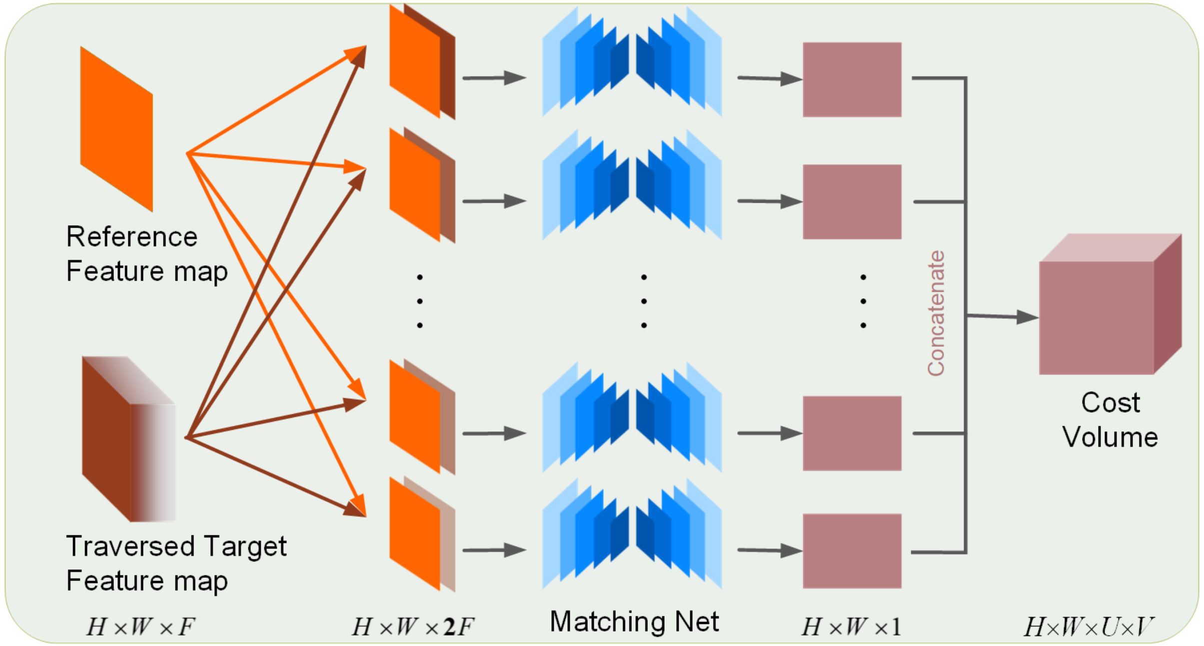
***Jianyuan Wang[1], *Yiran Zhong[1,5], Yuchao Dai[2], Kaihao Zhang[1,4], Pan Ji[3], Hongdong Li[1,5]**

[1]Australian National University, [2]Northwestern Polytechnical University, [3]NEC Labs America, [4]Tencent AI Lab, [5]ACRV

Displacement-Invariant Cost Learning



Volumetric Approach



Our Method

Table 1. Per Layer Analysis of Processing a 5D feature Volume $(K \times U \times V \times \lambda H \times \lambda W)$

Methods	Kernel	Params	Ratio	Theoretical Inference Memory	ratio
4D convolutions	$(K,K,3,3,3,3)$	$81K^2$	$9K$	$K \times U \times V \times \lambda H \times \lambda W$	$U \times V$
Ours	$(K,3,3)$	$9K$	1	$K \times \lambda H \times \lambda W$	1

NEURAL INFORMATION PROCESSING SYSTEMS

Displacement-Invariant Cost Learning

Table 2. Ablation study on different cost computation metrics.

Method	Chairs	KITTI-15 train		Sintel-train (EPE)	
	EPE	EPE	Fl-all	Clean	Final
Dot Product	1.86	10.39	31.1	2.57	4.06
Cosine Similarity	1.84	10.45	30.2	2.55	4.03
3-Layer MLP	1.76	9.83	28.9	2.45	3.98
DICL	1.33	8.78	23.8	2.11	3.85

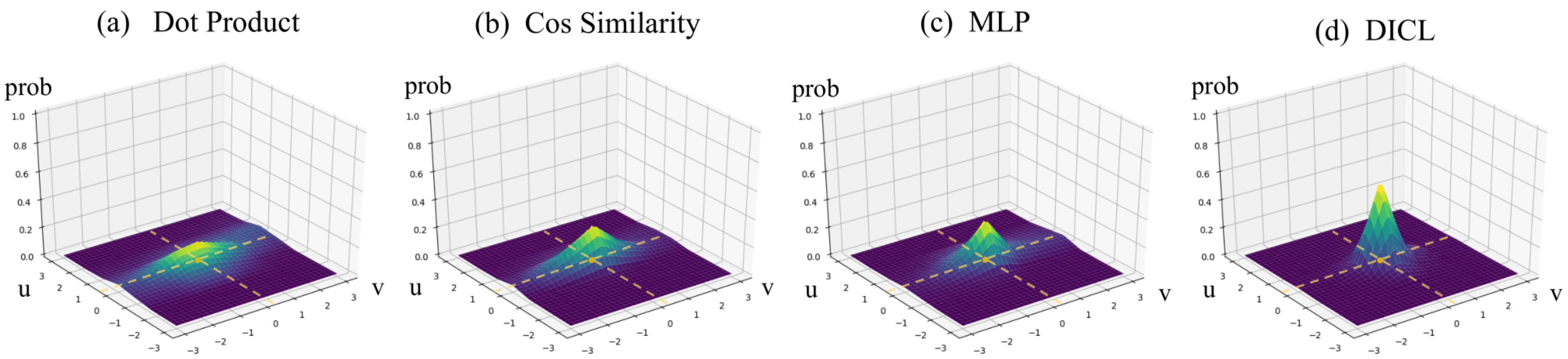


Figure 1. Qualitative Example of the Displacement Probability Distribution with Different Kinds of Matching Costs. The intersection of two yellow lines shows the ground truth location.

Displacement Aware Projection

Learned Reweighting Parameters

Reweighted Matching Costs

$$C'_{\mathbf{u}} = \sum_{\mathbf{v}\in\mathcal{U}} w_{(\mathbf{u},\mathbf{v})} C_{\mathbf{v}},$$

Costs Computed by DICL

All Candidate Displacements

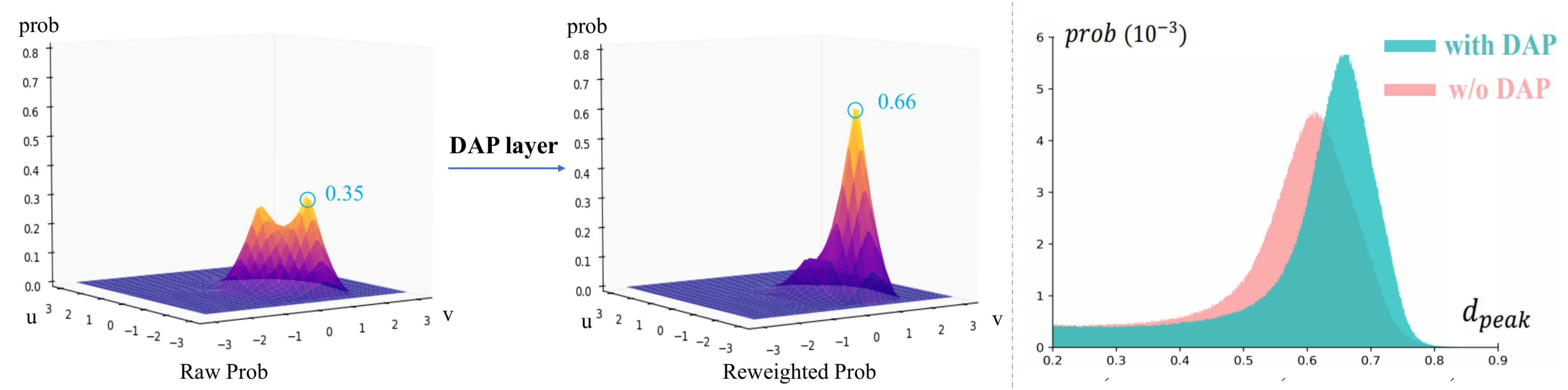


Figure 2. The left column compares an example pixel's displacement probability, before and after using DAP layer. The right column shows the histogram of the d_{peak} distribution with and without the DAP layer. d_{peak} represents the difference value between the highest and the second probability among the displacements.

Overall Architecture

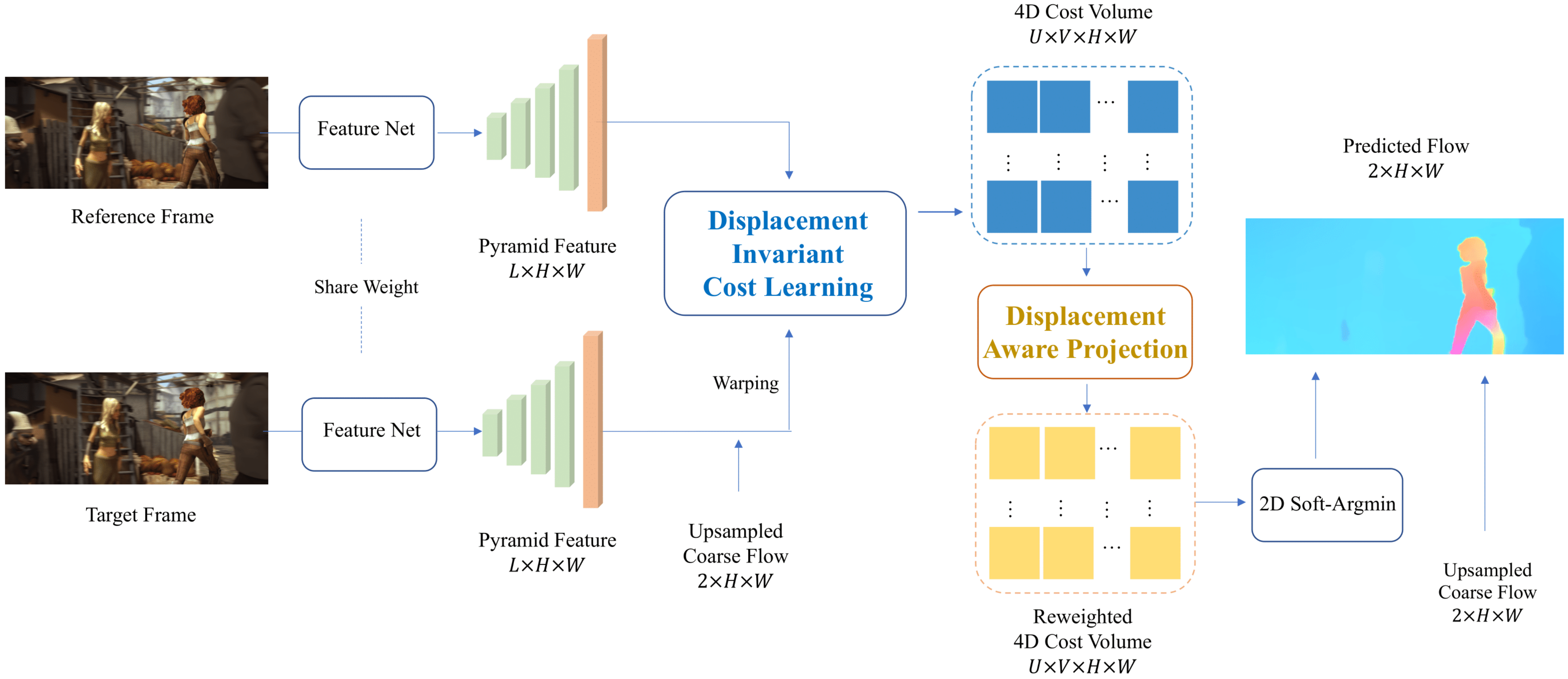


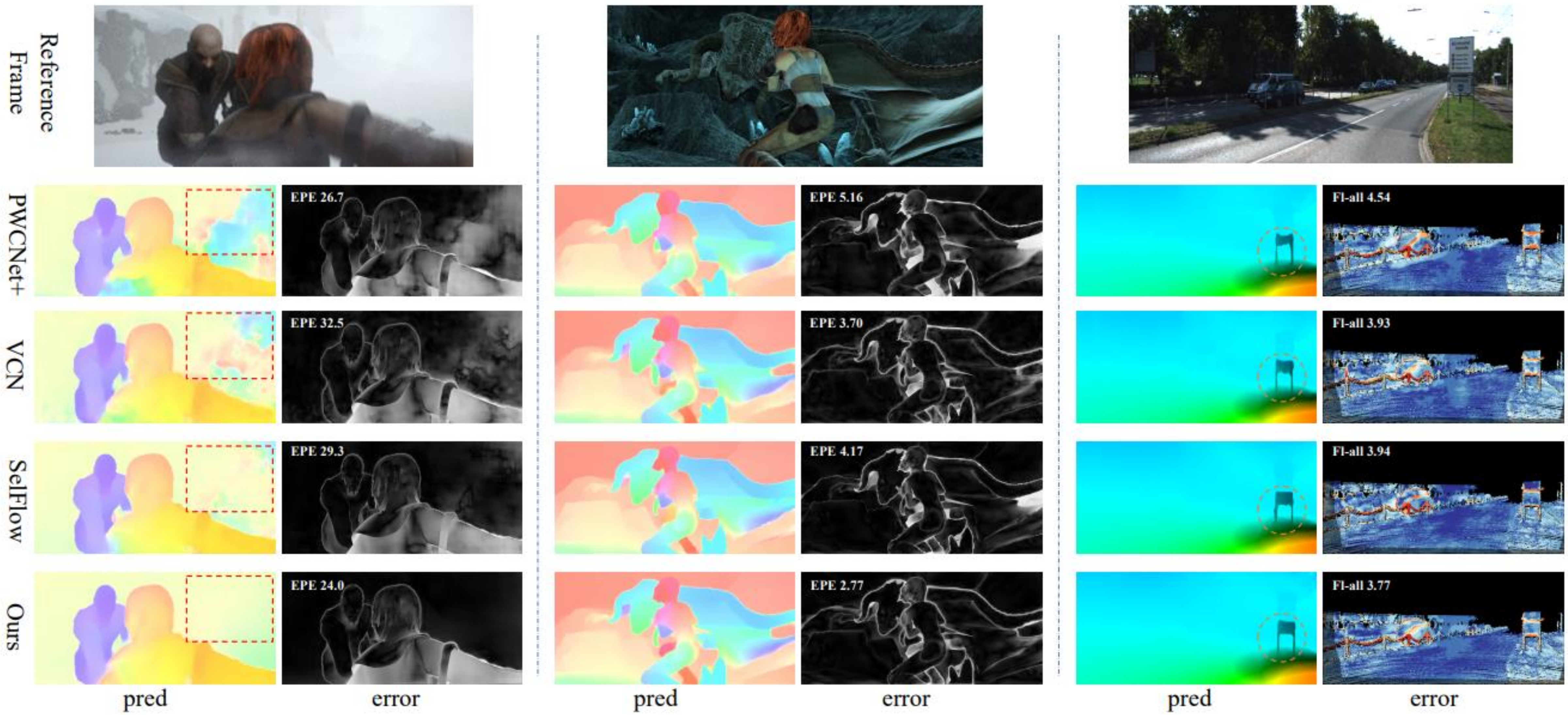
Figure 3. The feature net outputs features at five pyramid levels. For each level, our displacement-invariant cost learning module compares the reference feature map and the target feature map at each displacement and builds a 4D cost volume. Our displacement-aware projection layer reweights the learned cost volume to make it unimodal. A 2D soft-argmin layer projects the cost volume to optical flow

Benchmark Results

Table 3. Quantitative Results on KITTI 2015 and Sintel Datasets. The symbol ‘C+T’ indicates a model pre-trained on the Chair and Things datasets while ‘+K/S’ means further fine-tuned on the KITTI or Sintel dataset. Parentheses means the results are reported on its training dataset.

	Method	Time	K-15 train		K-15 test	S-train (EPE)		S-test (EPE)	
		(s)	EPE	Fl-all	Fl-all	Clean	Final	Clean	Final
	EpicFlow [27]	15.00	-	-	26.29	-	-	4.12	6.29
	DCFlow [35]	8.60	-	15.1	14.86	-	-	3.54	5.12
C+T	FlowNet2 [12]	0.12	10.08	30.0	-	2.02	**3.54**	3.96	6.02
	PWCNet [29]	0.03	10.35	33.7	-	2.55	3.93	-	-
	LiteFlowNet [9]	0.09	10.39	28.5	-	2.48	4.04	-	-
	LiteFlowNet2 [10]	0.04	8.97	25.9	-	2.24	3.78	-	-
	HD^{3}F [39]	0.08	13.17	24.0	-	3.84	8.77	-	-
	VCN [37]	0.18	**8.36**	25.1	-	2.21	3.62	-	-
	Ours-w/o DAP	0.08	8.78	23.8	-	2.11	3.85	-	-
	Ours	0.08	8.70	**23.6**	-	**1.94**	3.77	-	-
+K/S	FlowNet2 [12]	0.12	(2.30)	(8.6)	11.48	(1.45)	(2.01)	4.16	5.74
	PWCNet+ [30]	0.03	(1.50)	(5.3)	7.72	(1.71)	(2.34)	3.45	4.60
	LiteFlowNet [9]	0.09	(1.62)	(5.6)	9.38	(1.35)	(1.78)	4.54	5.38
	LiteFlowNet2 [10]	0.04	(1.47)	(4.8)	7.74	(1.30)	(1.62)	3.45	4.90
	IRR-PWC [11]	0.21	(1.63)	(5.3)	7.65	(1.92)	(2.51)	3.84	4.58
	HD^{3}F [39]	0.08	(1.31)	(4.1)	6.55	(1.87)	(**1.17**)	4.79	4.67
	SelFlow [20]	0.09	(1.18)	-	8.42	(1.68)	(1.77)	3.74	4.26
	VCN [37]	0.18	(1.16)	(4.1)	**6.30**	(1.66)	(2.24)	2.81	4.40
	Ours-w/o DAP	0.08	(1.09)	(3.8)	-	(1.30)	(1.72)	-	-
	Ours	0.08	(**1.02**)	(**3.6**)	6.31	(**1.11**)	(1.60)	**2.12**	**3.44**

Benchmark Results



Adversarial Attack

Table 4. Performance Against Adversarial Attacks. The patch size used by the adversarial attack is indicated by pixels, e.g., 25 × 25. The column 'Diff' denotes the relative EPE difference after attacks. The results are reported on the KITTI 2015 training set

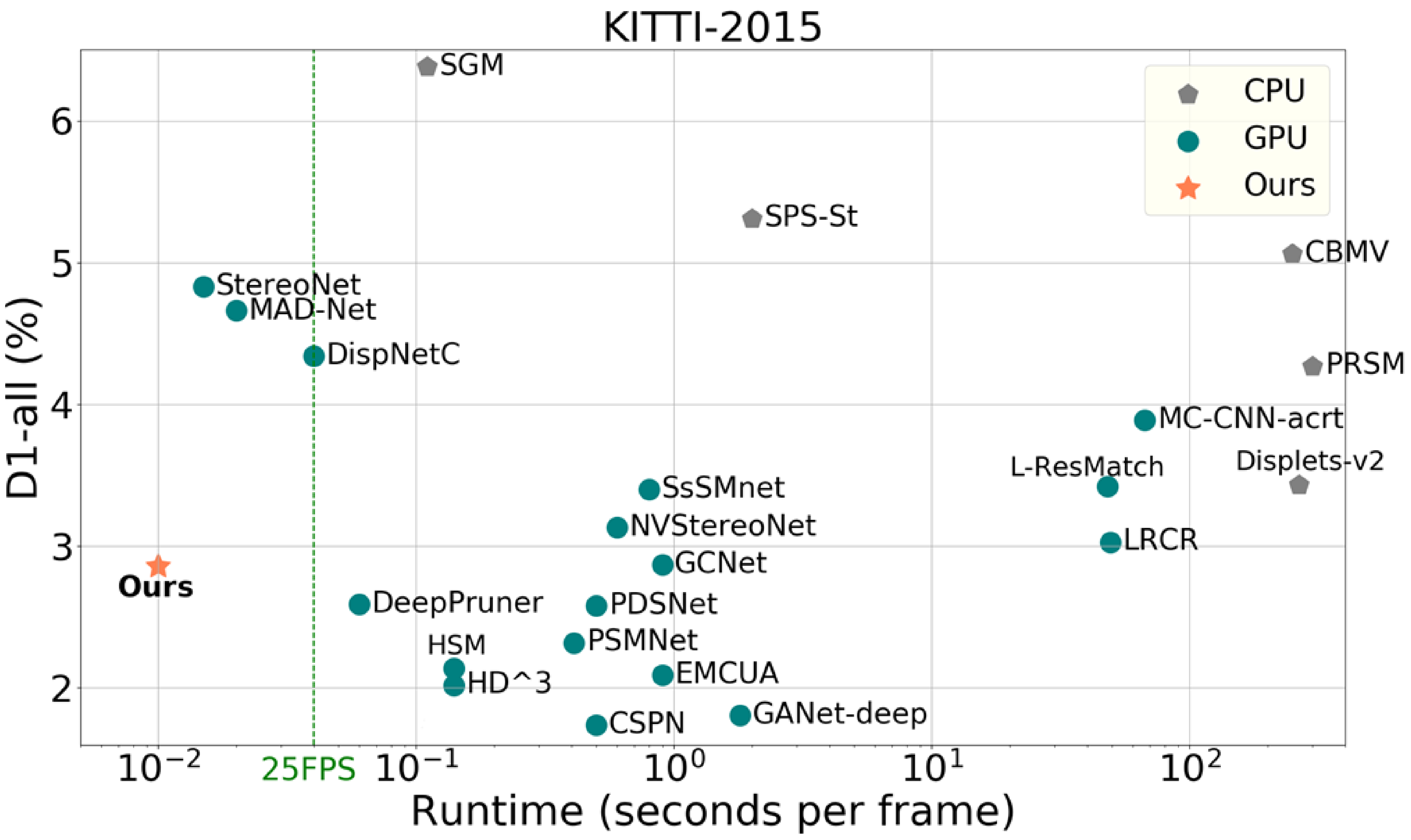
	Unattacked	25x25		51x51		102x102		153x153	
Network	EPE	EPE	Diff	EPE	Diff	EPE	Diff	EPE	Diff
FlowNetC [3]	14.56	29.07	+14.51	40.27	+25.51	82.41	+67.85	95.32	+80.76
FlowNet2 [4]	11.90	17.04	+5.14	24.42	+12.52	38.57	+26.67	59.58	+47.68
SpyNet [8]	20.26	20.59	+0.33	21.00	+0.74	21.22	+0.96	21.00	+0.74
PWC-Net [10]	11.03	11.37	+0.34	11.50	+0.47	11.86	+0.83	12.52	+1.49
Back2Future [5]	17.49	18.04	+0.55	18.24	+0.75	18.73	+1.24	18.43	+0.94
Ours	8.98	9.17	+0.19	9.30	+0.32	9.52	+0.54	9.61	+0.63

Attacked
Reference Frame

Unattacked Flow

Attacked Flow

Stereo Matching Extension



Stereo Matching Extension

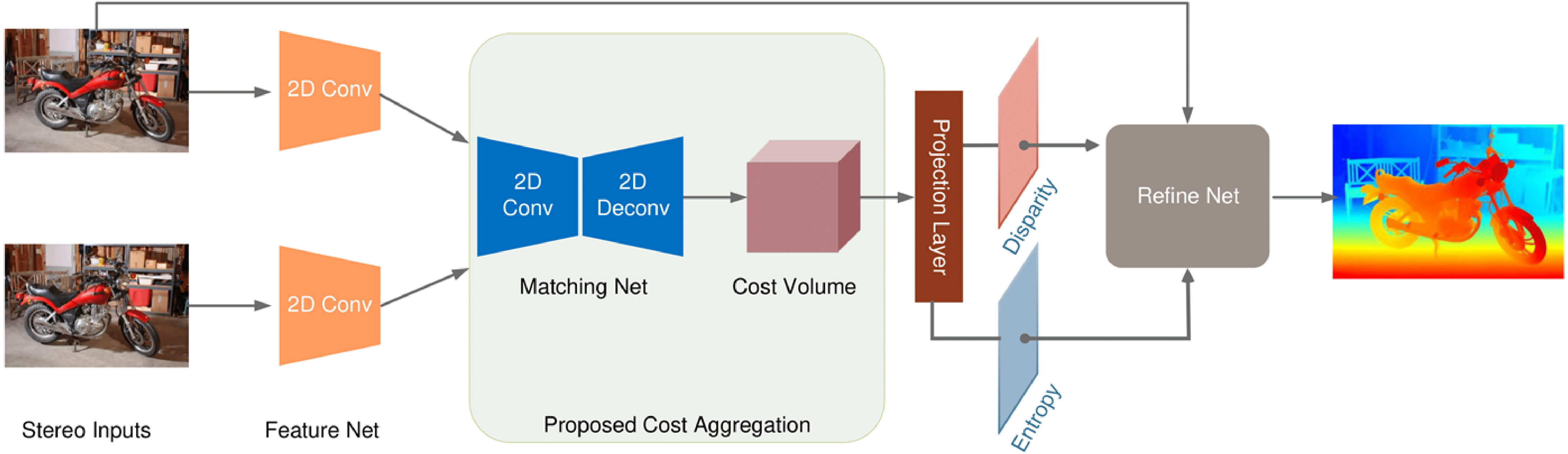


Figure 4. Overall Architecture of Our Stereo Matching Extension

Feature Net: 8 layers with spatial pyramid pooling.
Matching Net: 17 layers with skip connected U-net.
Projection Layer:
    Project cost volume to disparity map
    Compute entropy map from cost volume

Refine Net: take left image, entropy map and disparity map as input.
Loss Functions: smooth l_1 loss on $\boldsymbol{d}\_coarse$ and $\boldsymbol{d}\_refine$

Stereo Matching Extension

Table 5. Benchmark Quantitative Results.

Results on KITTI 2015 *test* set. Bold indicates the best, while underline indicates the second best.

	Method	Runtime(s)	Non-occ (%) bg	Non-occ (%) fg	Non-occ (%) all	All (%) bg	All (%) fg	All (%) all
Below 10 FPS	MC-CNN [29]	67.00	2.48	7.64	3.33	2.89	8.88	3.89
	SGMnet [23]	67.00	2.23	7.44	3.09	2.66	8.64	3.66
	PDSnet [28]	0.50	2.09	3.68	2.36	2.29	4.05	2.58
	CRL [18]	0.47	2.32	3.68	2.36	2.48	3.59	2.67
	SDRNet [1]	0.23	1.57	4.58	2.06	1.72	4.95	2.26
	PSMnet [2]	0.41	1.71	4.31	2.14	1.86	4.62	2.32
	GC-Net [11]	0.90	2.02	3.12	2.45	2.21	6.16	2.87
	M2S_CSPN [4]	0.50	1.40	**2.67**	**1.61**	**1.51**	**2.88**	**1.74**
	HSM [31]	**0.15**	1.63	3.40	1.92	1.80	3.85	2.14
	EMCUA [17]	0.90	1.50	3.88	1.90	1.66	4.27	2.09
	GA-Net-15 [34]	0.36	1.40	3.37	1.73	1.55	3.82	1.93
	DPruner_Best [24]	0.18	1.71	3.18	1.95	1.87	3.56	2.15
Above 10 FPS	StereoNet [12]	0.02	-	-	-	4.30	7.45	4.83
	DN-CSS [9]	0.07	2.23	4.96	2.68	**2.39**	5.71	2.94
	MAD-Net [27]	0.02	3.45	8.41	4.27	3.75	9.2	4.66
	DispNetC [15]	0.04	4.11	**3.72**	4.05	4.32	**4.41**	4.34
	Our2D	**0.01**	**2.12**	3.88	**2.42**	2.51	4.62	**2.86**

Results on ETH3D *test* dataset. Bold indicates the best, while underline indicates the second best.

Methods	time(s)	EPE	rmse	bad-4.0	bad-2.0	bad-1.0	A99
HSM [31]	**0.14**	0.29	0.67	0.68	1.48	4.25	3.25
SDRNet [1]	0.15	0.34	0.71	0.50	1.66	6.02	3.07
iResNet [1]	0.20	0.25	0.59	**0.34**	1.20	4.04	2.70
DPruner [24]	0.16	**0.28**	**0.58**	**0.34**	**1.04**	**3.82**	**2.61**
PSMnet [2]	0.41	0.36	0.75	0.54	1.31	5.41	3.38
DN-CSS [9]	0.07	**0.24**	**0.56**	**0.38**	**0.96**	**3.00**	2.89
Our2D	**0.01**	0.32	0.63	0.53	1.25	4.82	**2.79**

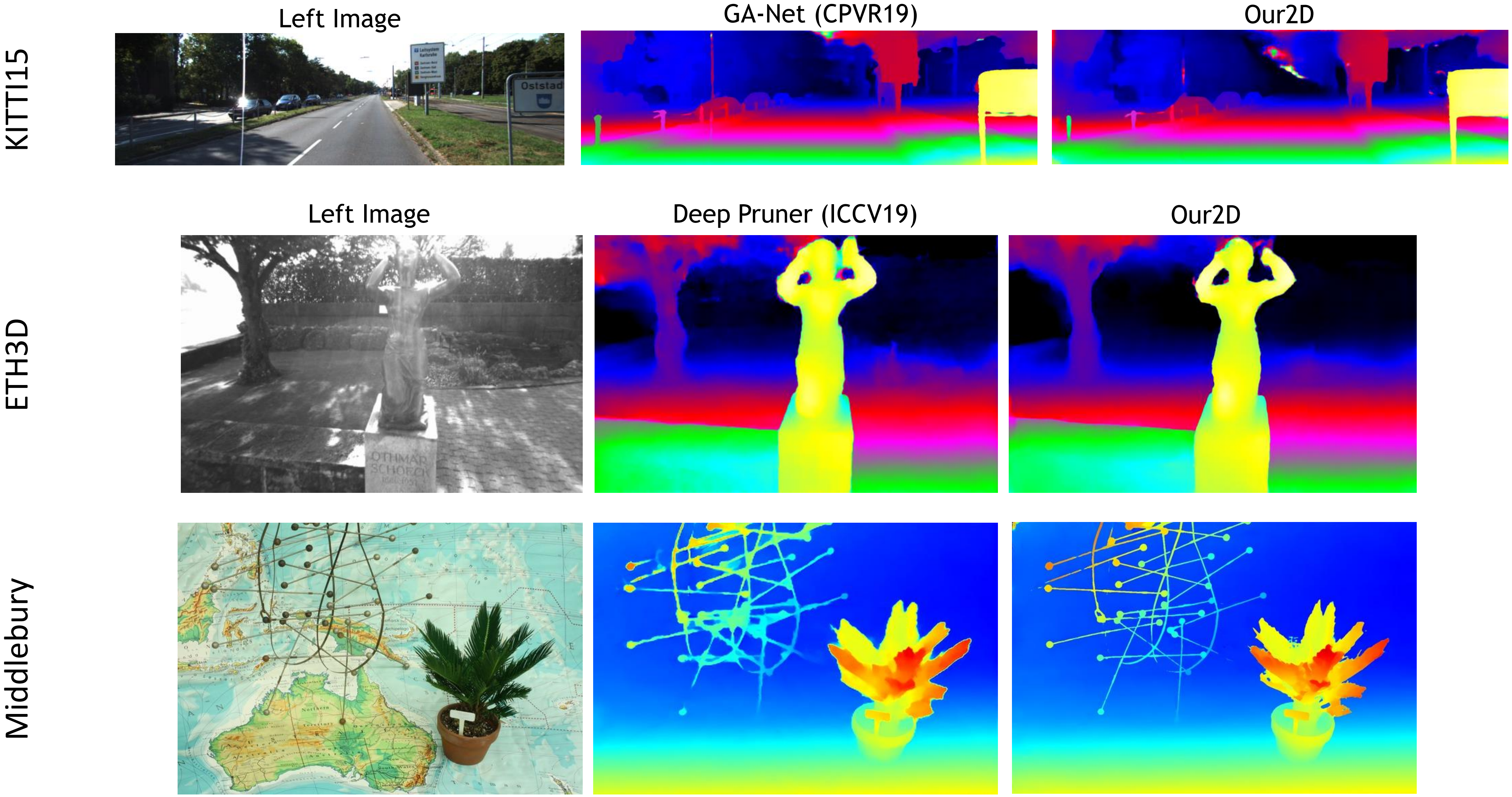
Results on Middlebury 2014 *test* dataset. Bold indicates the best, while underline indicates the second best.

Methods	time(s)	EPE	rmse	bad-4.0	bad-2.0	bad-1.0	A99
SGM [7]	0.32	5.32	20.0	12.2	18.4	31.1	109
HSM [31]	0.51	**2.07**	**10.3**	**4.83**	**10.2**	**24.6**	**39.2**
iResNet [1]	0.34	3.31	11.3	12.6	22.9	38.8	48.6
DPruner [24]	0.13	4.80	14.7	15.9	30.1	52.3	67.7
PSMNet [2]	0.64	6.68	19.4	23.5	42.1	63.9	84.5
DN-CSS [9]	0.66	4.04	13.9	14.7	22.8	36.0	58.8
Our2D	**0.04**	3.12	13.8	7.22	15.4	35.1	55.6

NEURAL INFORMATION PROCESSING SYSTEMS

Stereo Matching Extension

Figure 5. Benchmark Qualitative Results.



NEURAL INFORMATION PROCESSING SYSTEMS

Thanks

Code is available at:

AUSTRALIAN CENTRE FOR ROBOTIC VISION

Hierarchical Neural Architecture Search for Deep Stereo Matching

*Xuelian Cheng[1,5], *Yiran Zhong[2], Mehrtash Harandi[1,7], Yuchao Dai[3], Xiaojun Chang[1], Hongdong Li[2,6], Tom Drummond[1], Zongyuan Ge[1,4,5]

[1]Faculty of Engineering, Monash University, [2]Australian National University, [3]Northwestern Polytechnical University, [4]eResearch Centre, Monash University, [5]Airdoc Research Australia, [6]ACRV, [7]DATA61 CSIRO

The Proposed Pipeline

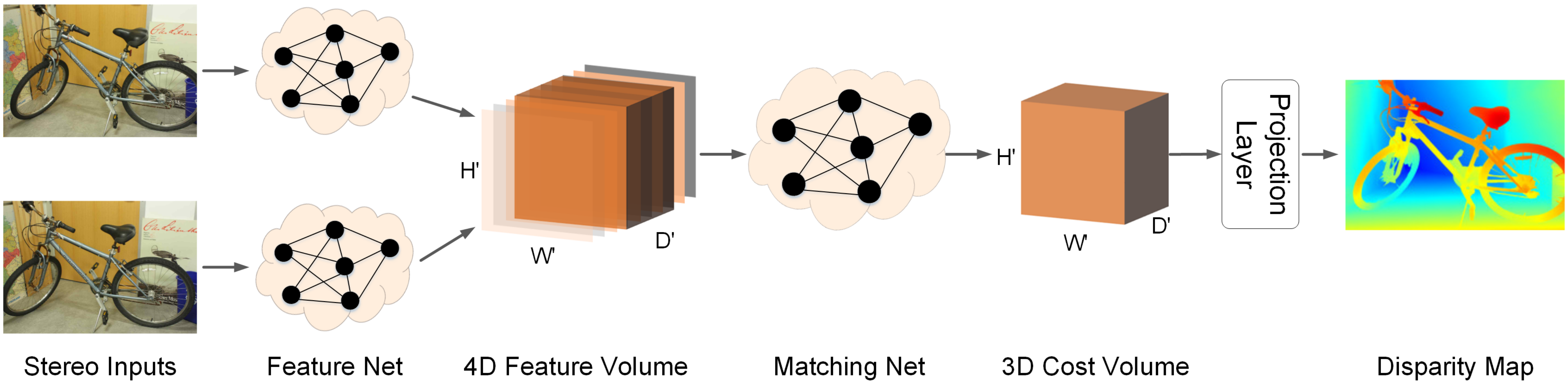


Figure 1. The pipeline of our proposed stereo matching network. Given a pair of stereo images, the Feature Net produces feature maps that are processed by the Matching Net to generate a 3D cost volume. The disparity map can be projected from the cost volume with soft-argmin operation. Feature Net and Matching Net are the only two modules that contain trainable parameters, we utilize the NAS technique to select the optimal structures.

Refined Searching Space

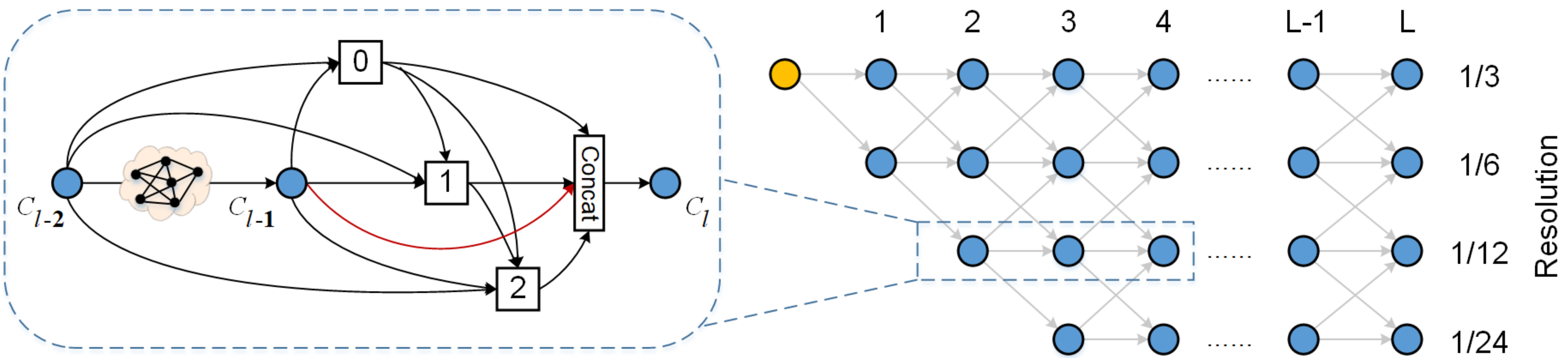
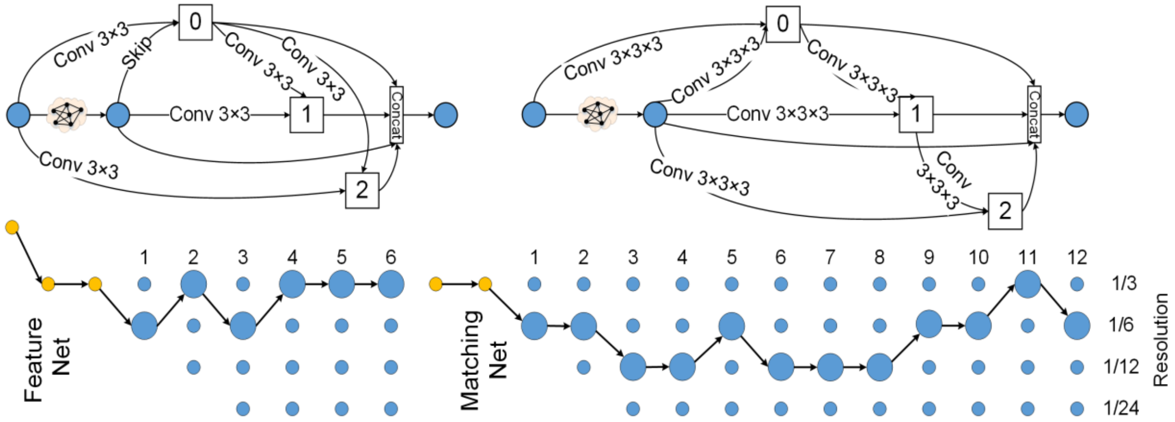


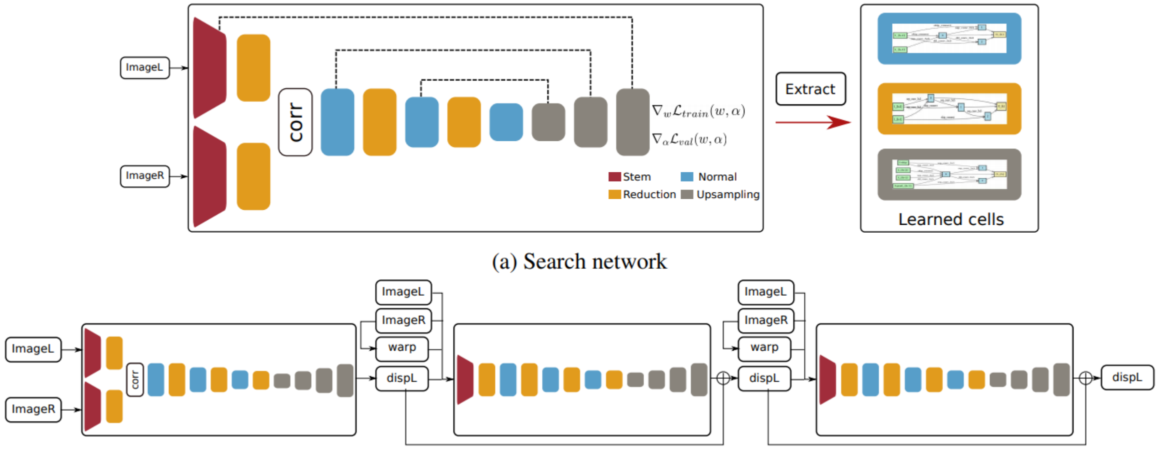
Figure 2. Our Refined Search Space. Left: cell level search space; Right: our network level search space. The red arrow on the left represents the proposed residual connections. We set $\mathrm{L}^{\mathrm{F}} = 6$ for Feature Net and $\mathrm{L}^{\mathrm{M}} = 12$ for Matching Net.

Ours vs AutoDispNet

Our Searched Architecture



AutoDispNet Architecture



(a) Search network

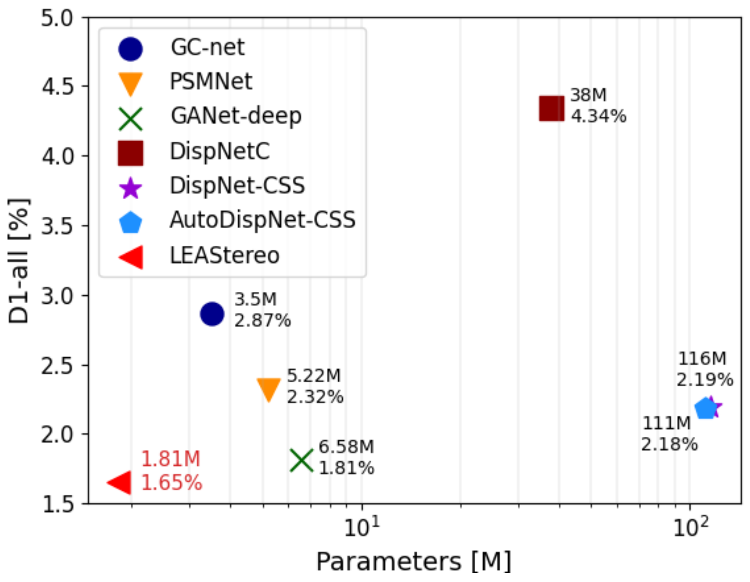
(b) AutoDispNet-CSS

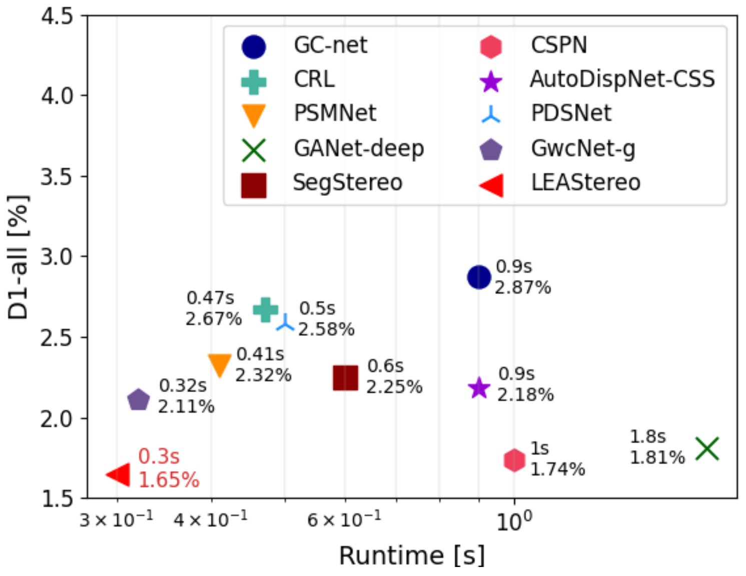
Table 1. Comparing with AutoDispNet, our method boosts the performance of **32.12%** in accuracy and **66.67%** in inference speed with only **1.7%** of the parameters.

	Search Level	Params	KITTI 2012	KITTI 2015	Runtime
AutoDispNet	Cell	111M	1.70%	2.18%	0.9s
Ours	Full Network	1.8M	1.13%	1.65%	0.3s

Benchmark Results

KITTI 2015





Middlebury 2014

Set: test dense | test sparse | training dense | training sparse

Metric: bad 0.5 | bad 1.0 | bad 2.0 | bad 4.0 | avgerr | rms | A50 | A90 | A95 | A99 | time | time/MP | time/GD

Mask: nonocc | all

☐ plot selected ☐ show invalid Reset sort Reference list

Date	Name (bad 4.0 (%))	Res	Avg (Weight)	Austr MP: 5.6 nd: 290 im0 im1 GT nonocc	AustrP MP: 5.6 nd: 290 im0 im1 GT nonocc	Bicyc2 MP: 5.6 nd: 250 im0 im1 GT nonocc	Class MP: 5.7 nd: 610 im0 im1 GT nonocc	ClassE MP: 5.7 nd: 610 im0 im1 GT nonocc	Compu MP: 1.5 nd: 256 im0 im1 GT nonocc	Crusa MP: 5.5 nd: 800 im0 im1 GT nonocc	CrusaP MP: 5.5 nd: 800 im0 im1 GT nonocc	Djemb MP: 5.7 nd: 320 im0 im1 GT nonocc	DjembL MP: 5.7 nd: 320 im0 im1 GT nonocc	Hoops MP: 5.7 nd: 410 im0 im1 GT nonocc	Livgrm MP: 5.9 nd: 320 im0 im1 GT nonocc	Nkuba MP: 5.5 nd: 570 im0 im1 GT nonocc	Plants MP: 5.6 nd: 320 im0 im1 GT nonocc	Stairs MP: 5.2 nd: 450 im0 im1 GT nonocc
05/28/20	☑ LEAStereo	H	**2.75** 1	3.57 14	2.46 6	1.68 3	1.60 6	3.47 4	1.87 5	**1.57** 1	1.59 5	1.44 3	4.02 3	**4.11** 1	5.01 3	**5.13** 1	**2.89** 1	**3.22** 1
05/26/18	☐ NOSS_ROB	H	3.46 2	2.69 4	2.37 4	1.99 7	1.19 2	**3.11** 1	1.75 3	1.63 5	**1.26** 1	1.87 15	5.60 5	4.63 3	5.66 4	10.4 19	3.60 9	7.02 22
03/09/19	☐ 3DMST-CM	H	3.57 3	3.07 10	2.77 11	1.80 4	1.79 10	5.43 16	2.58 10	1.71 7	2.01 15	1.48 4	6.86 9	5.14 5	**4.41** 1	8.85 8	3.84 12	6.29 12

Thanks

NEURAL INFORMATION PROCESSING SYSTEMS

Airdoc

AUSTRALIAN CENTRE FOR ROBOTIC VISION

DATA 61

CSIRO