Matching-space Stereo Networks for Cross-domain Generalization

Changjiang Cai

Matteo Poggi

Stefano Mattoccia

Philippos Mordohai

Code: https://github.com/ccj5351/MS-Nets

Motivation

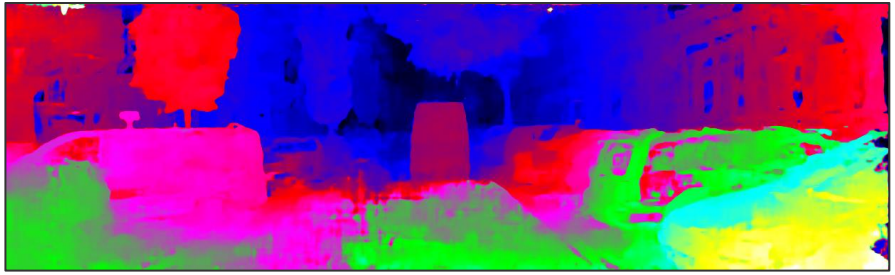
- Annotated data for stereo matching is challenging to collect
  - Expensive LiDAR and Stereo Camera Rig
  - Ground truth depth is **sparse**

- SOTA deep networks generalize poorly to unseen domain
  - Specialize on specific domains when enough data are available for training
  - Less effective at generalization to very different domains or with high variety of image content
- **Domain generalization** is a solution
  - To tackle the domain shift problem, two main strategies are involved
    - synthetic data by graphics engines
    - domain adaptation

Challenges in Synthetic-to-Real

- The large domain gap between synthetic and realistic data still pose difficulties
  - Reflective surfaces, sensor noise and illumination conditions have not been modeled well in the simulators
- Deep stereo networks suffer large accuracy drops moving from synthetic to real scenes
  - ➢ E.g., PSMNet pretrained on Scene Flow produces bad disparity results of KITTI15

Left Image



Disparity Map by PSMNet

Challenges in Domain Adaptation

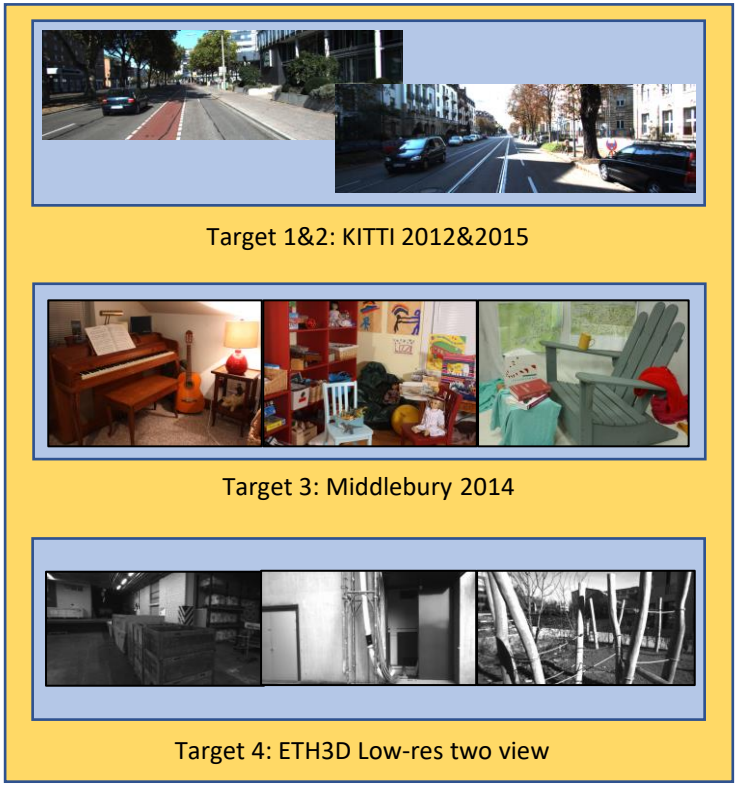
- Domain adaptation requires annotated data from the target domain
- Solutions: unsupervised learning (not in this talk) or methods that generalize well without adaptation
- An advantage of methods that generalize well
  - they can be effective in continuously changing environments, e.g. autonomous driving, without re-training or adaptation
  - being this possibility more appealing for practical applications
- Goal: sacrifice as little accuracy as possible to attain generalization

Domain Generalization

Source Domain: Scene Flow

Generalize

without
finetuning
or adaptation



Target Domains

Domain Generalization

Source Domain: Scene Flow



Monkaa



FlyingThings3D

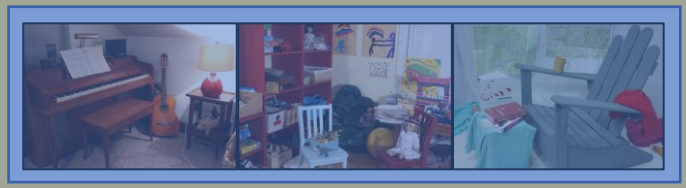


Driving

Generalize

without
finetuning
or adaptation

Target 1&2: KITTI 2012&2015



Target 3: Middlebury 2014

Target 4: ETH3D Low-res two view

Target Domains

Domain Generalization

Source Domain: Scene Flow

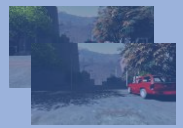


Monkaa



FlyingThings3D

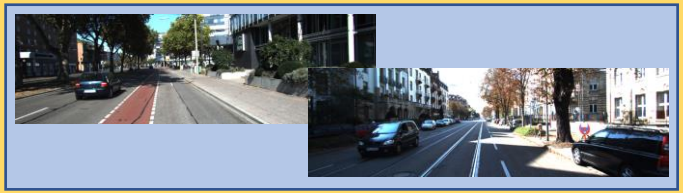




Driving

Generalize

without
finetuning
or adaptation



Target 1&2: KITTI 2012&2015

Target 3: Middlebury 2014

Target 4: ETH3D Low-res two view

Target Domains

Domain Generalization

Source Domain: Scene Flow

Monkaa

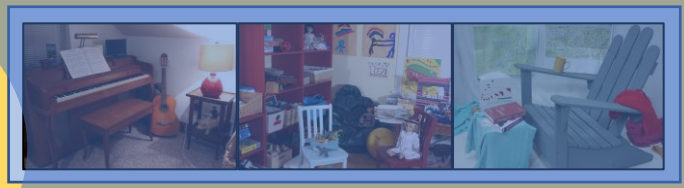
FlyingThings3D

Driving

Generalize

without
finetuning
or adaptation

Target 1&2: KITTI 2012&2015



Target 3: Middlebury 2014

Target 4: ETH3D Low-res two view

Target Domains

Over-specialization to Color Space

- The lack of generalization, or over-specialization, is caused by the learning process being driven by image content
  - Learn how to match pixels by strongly relying on appearance properties
  - Suffer accuracy drops when such content differs from the training data

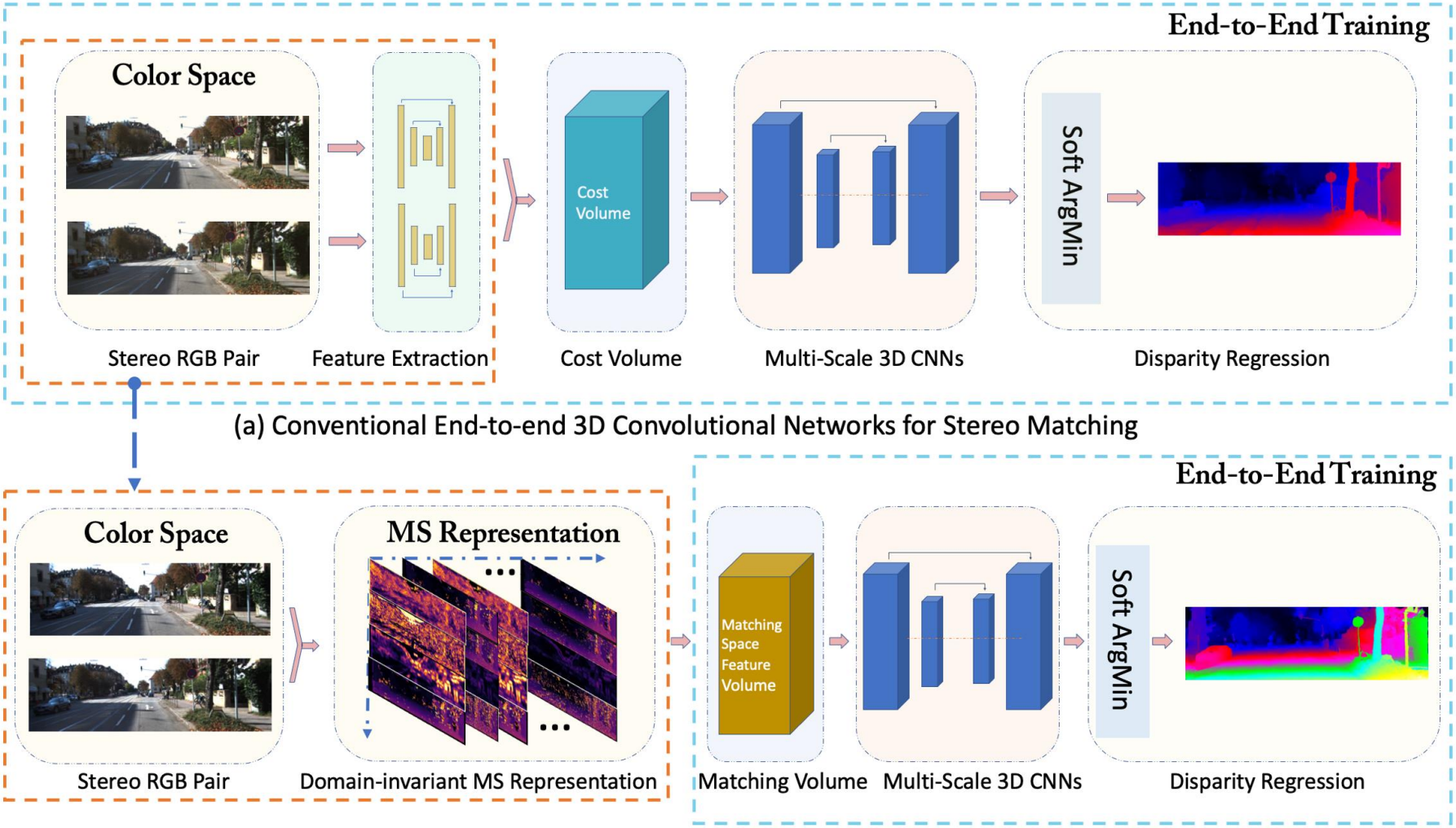
- Better generalization can be achieved by choosing a representation insensitive to common variations of the input images

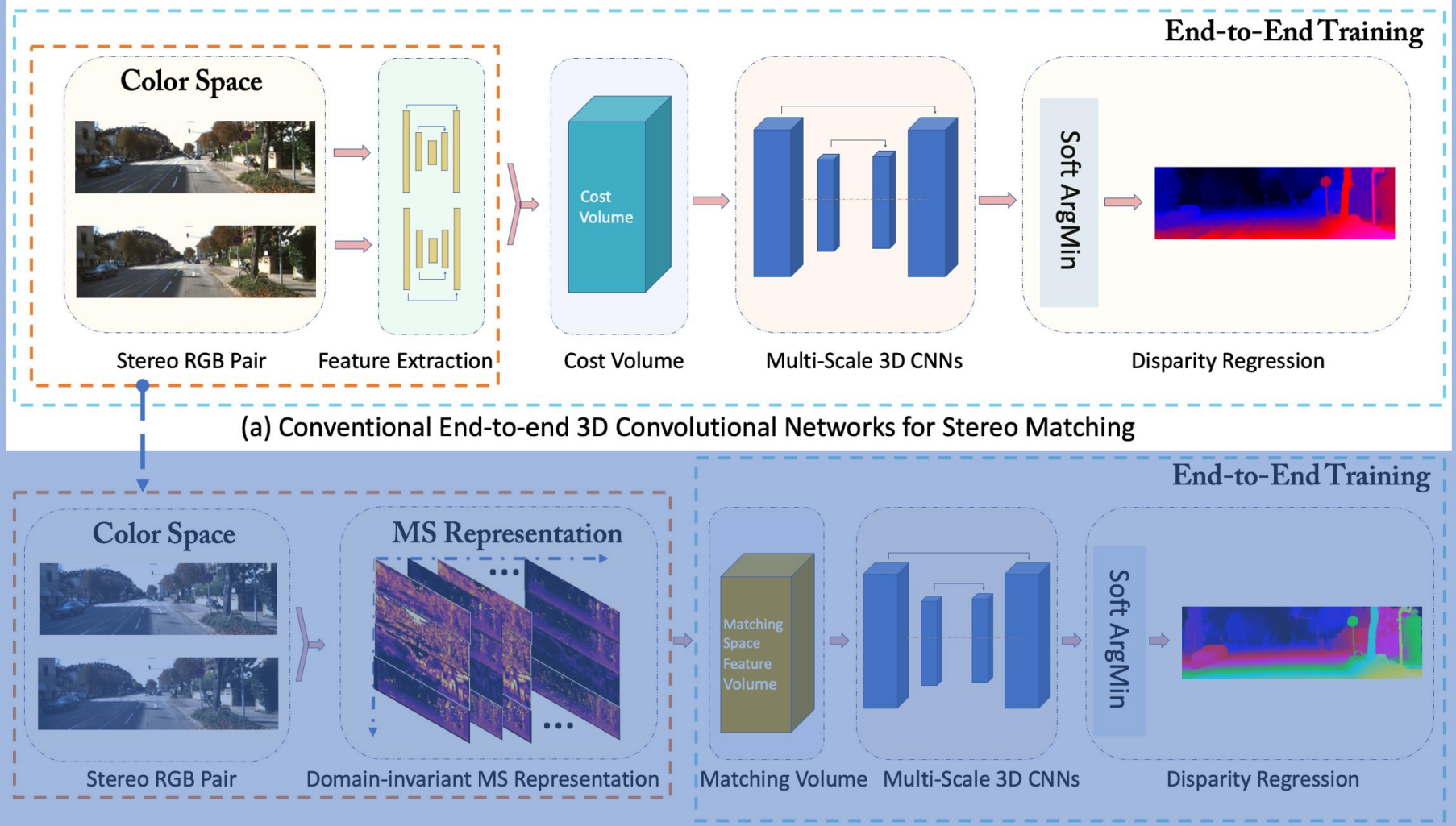
Matching Space Stereo Networks

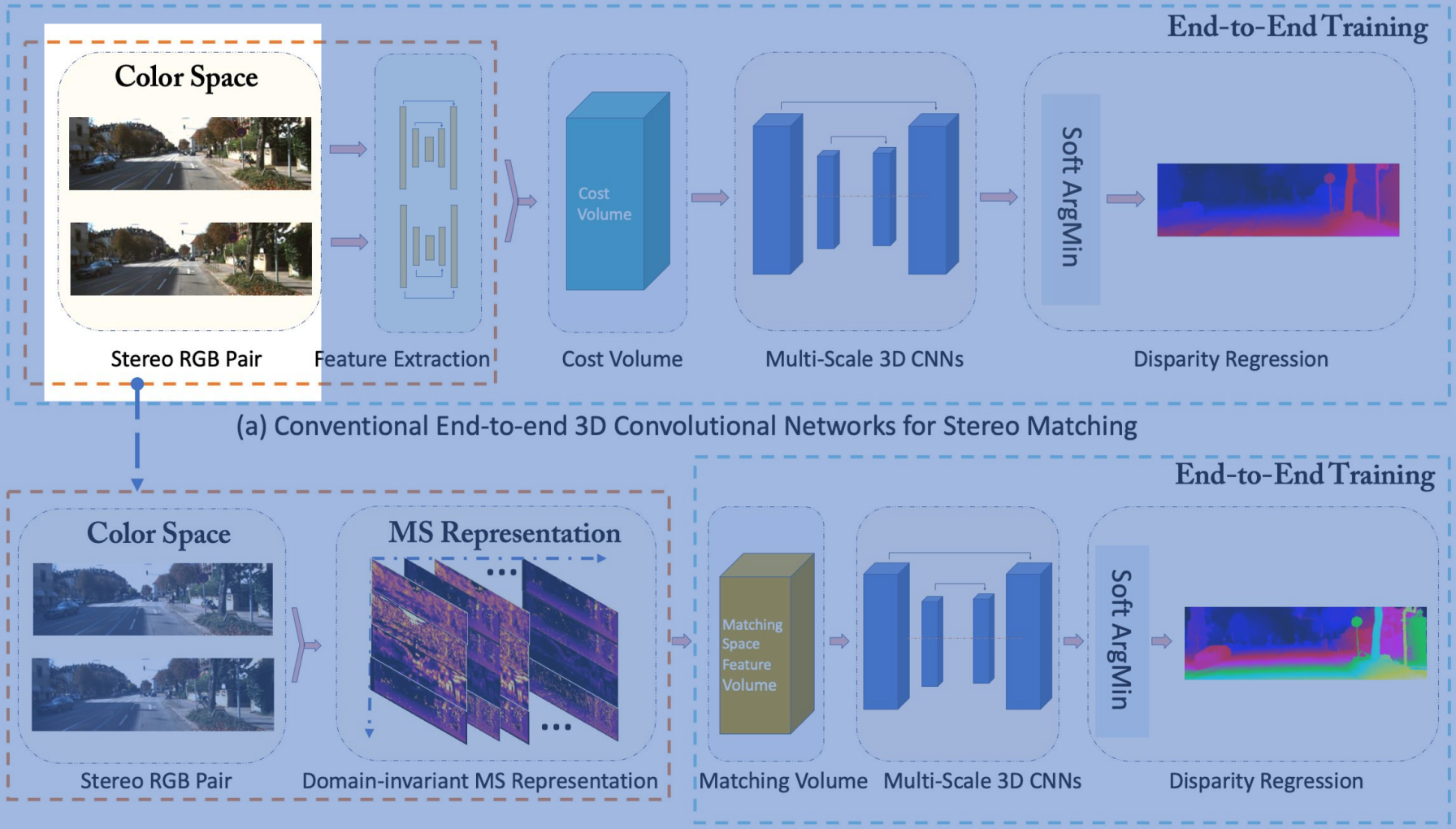
- Replace learning-based feature extraction from RGB with matching functions and confidence measures from conventional wisdom

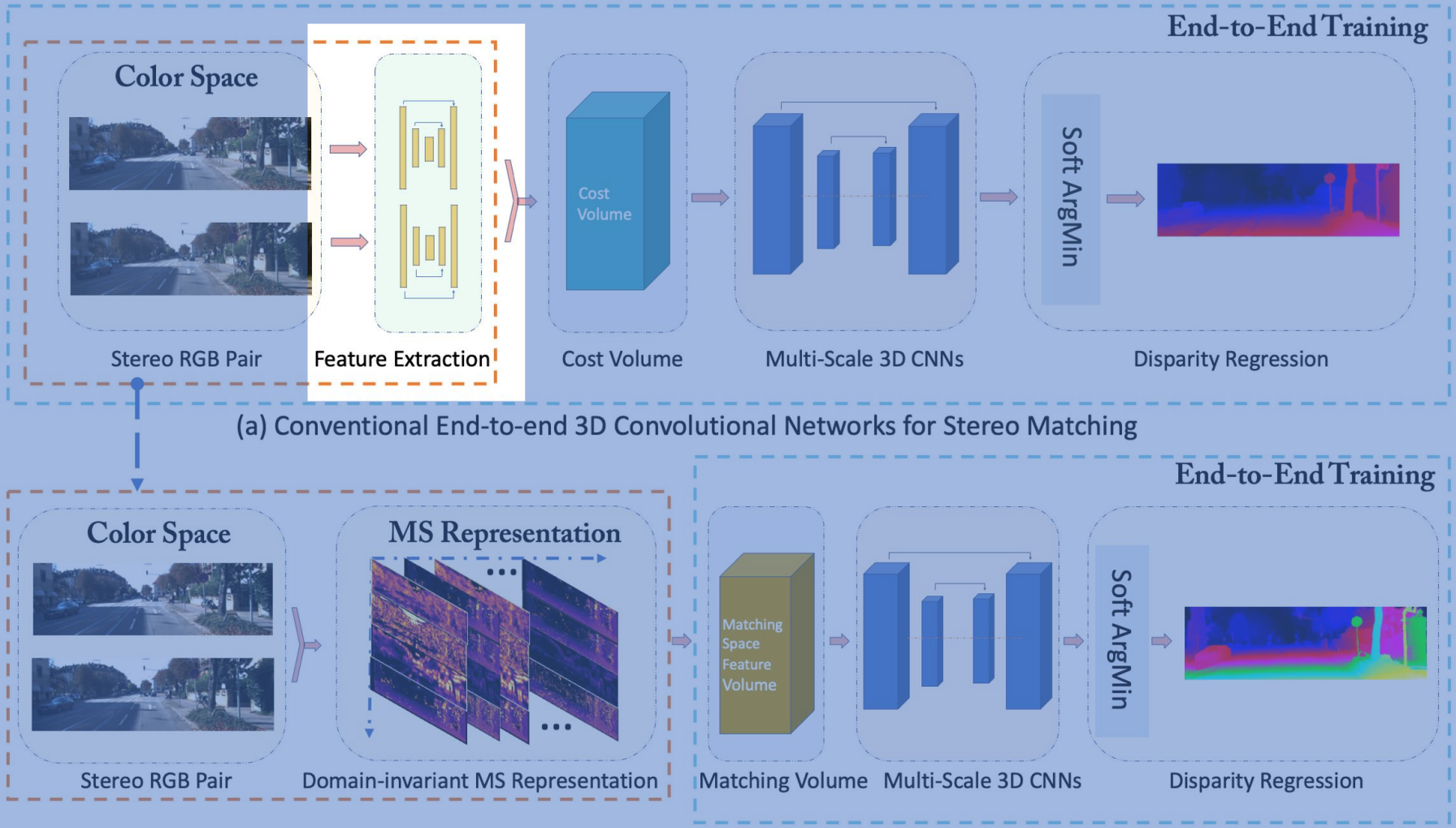
- Move learning process from color space to ***Matching Space (MS)***, avoiding overspecialization to domain specific features

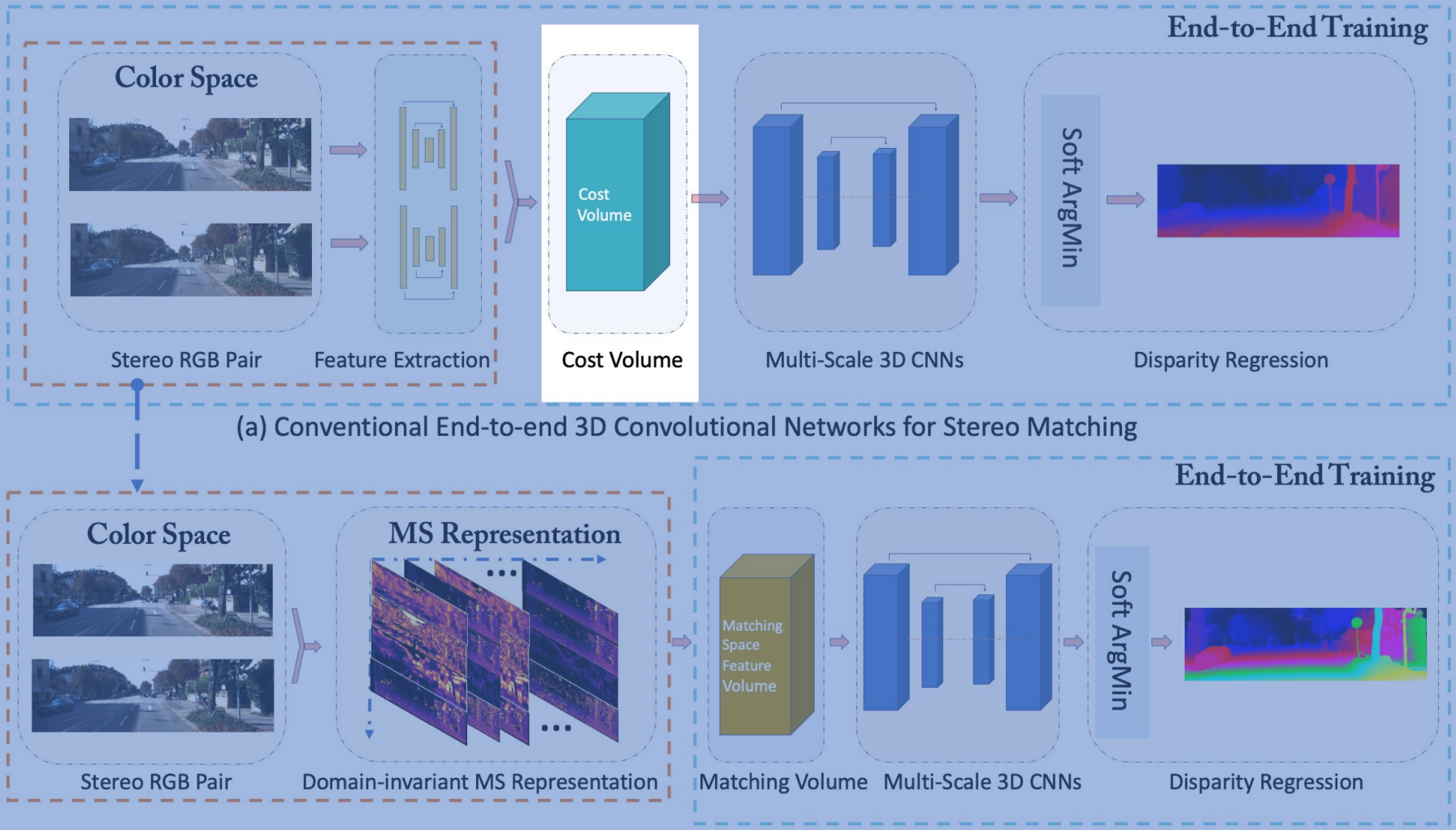
- Modify GCNet and PSMNet architectures to accept MS inputs
  - PSMNet allocates 63.5% of parameters to unary feature extraction
  - GCNet allocates 88.5% of parameters to 3D convolutions

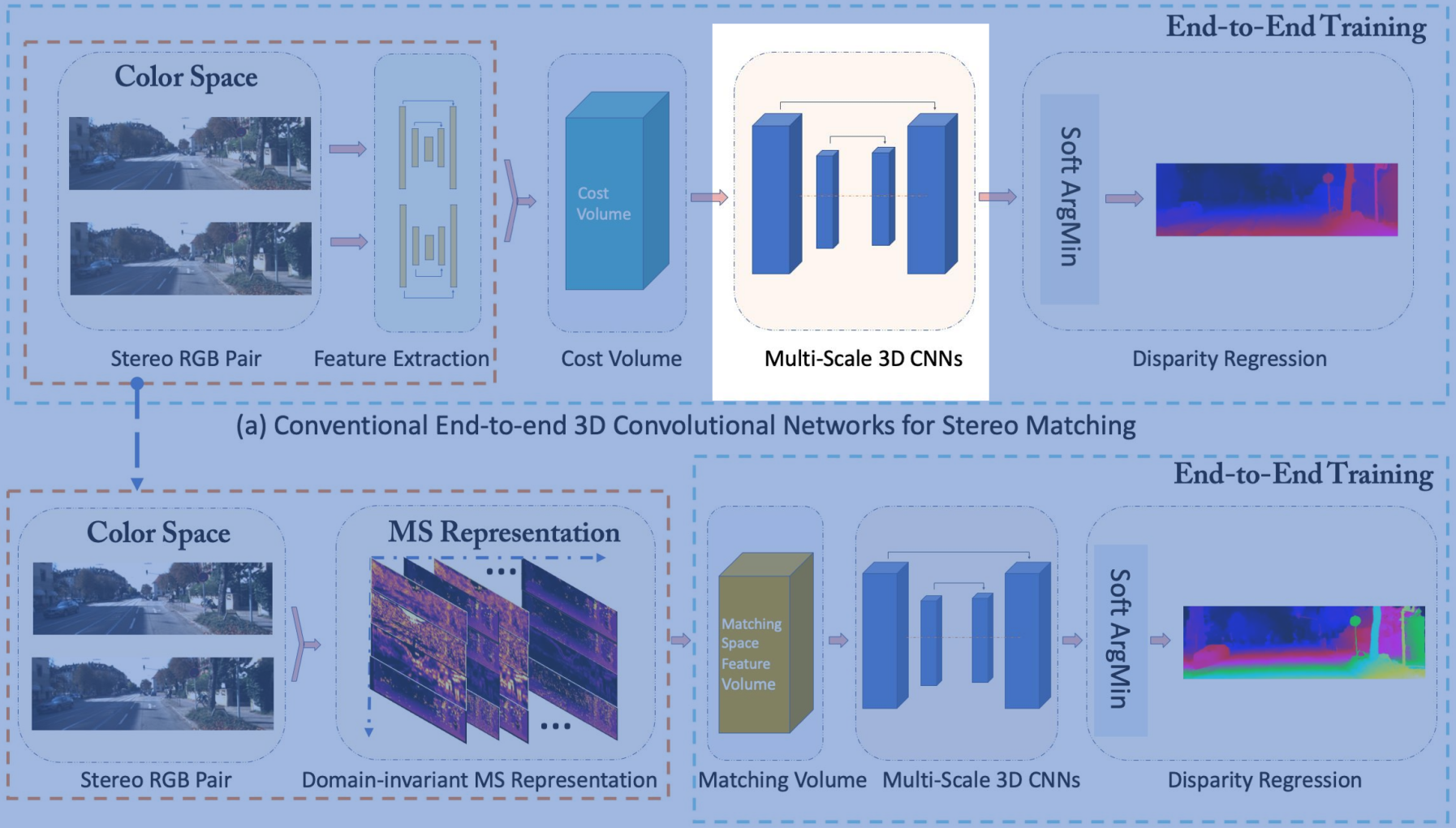

End-to-End Training
Color Space
Cost Volume
Soft ArgMin
Stereo RGB Pair
Feature Extraction
Cost Volume
Multi-Scale 3D CNNs
Disparity Regression
(a) Conventional End-to-end 3D Convolutional Networks for Stereo Matching
End-to-End Training
Color Space
MS Representation
Matching Space Feature Volume
Soft ArgMin
Stereo RGB Pair
Domain-invariant MS Representation
Matching Volume
Multi-Scale 3D CNNs
Disparity Regression
(b) Our Matching Space (MS) End-to-End Stereo Networks for Cross-domain Generalization

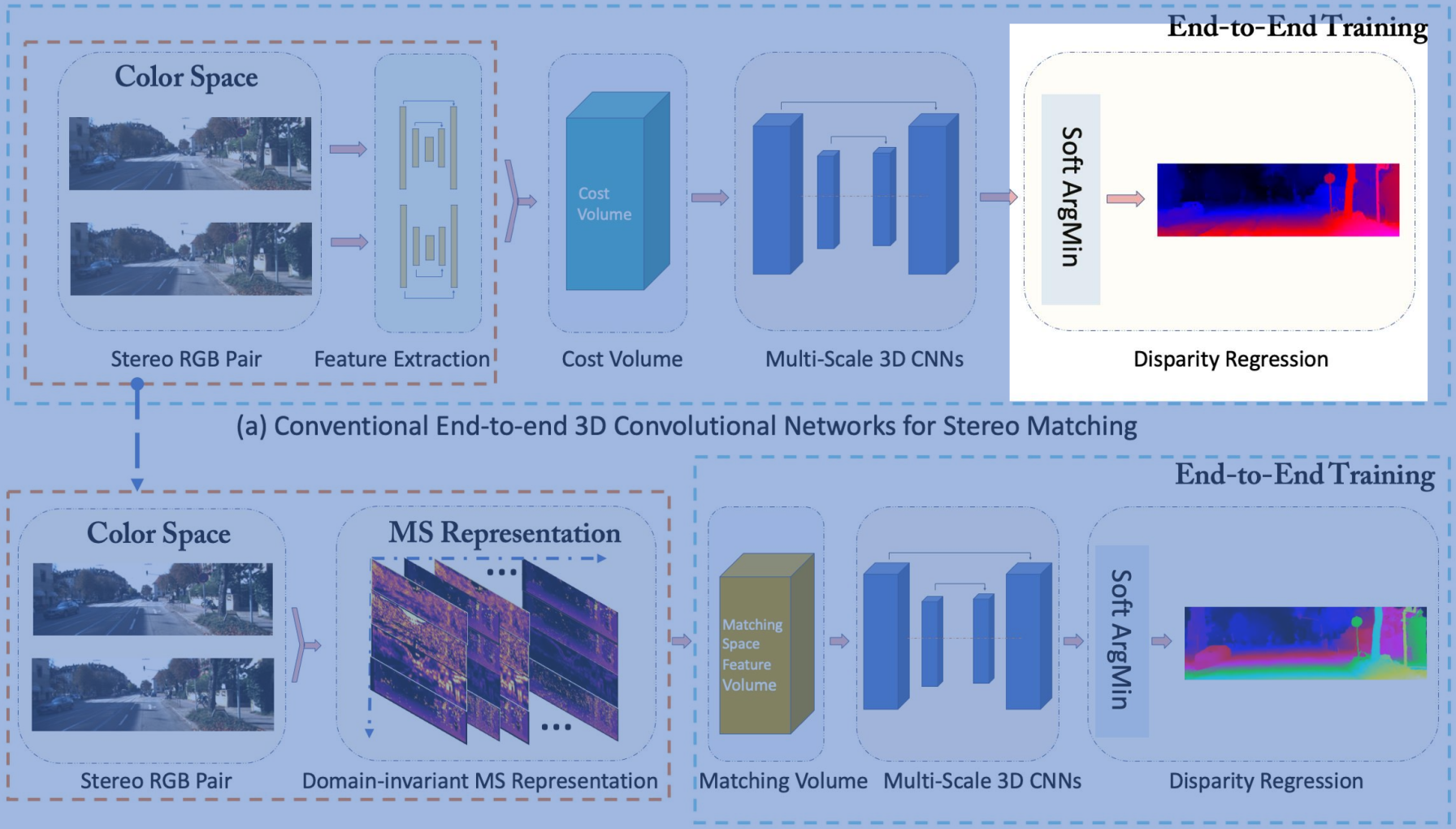

End-to-End Training
Color Space
Cost Volume
Soft ArgMin
Stereo RGB Pair
Feature Extraction
Cost Volume
Multi-Scale 3D CNNs
Disparity Regression
(a) Conventional End-to-end 3D Convolutional Networks for Stereo Matching
End-to-End Training
Color Space
MS Representation
Matching Space Feature Volume
Soft ArgMin
Stereo RGB Pair
Domain-invariant MS Representation
Matching Volume
Multi-Scale 3D CNNs
Disparity Regression
(b) Our Matching Space (MS) End-to-End Stereo Networks for Cross-domain Generalization

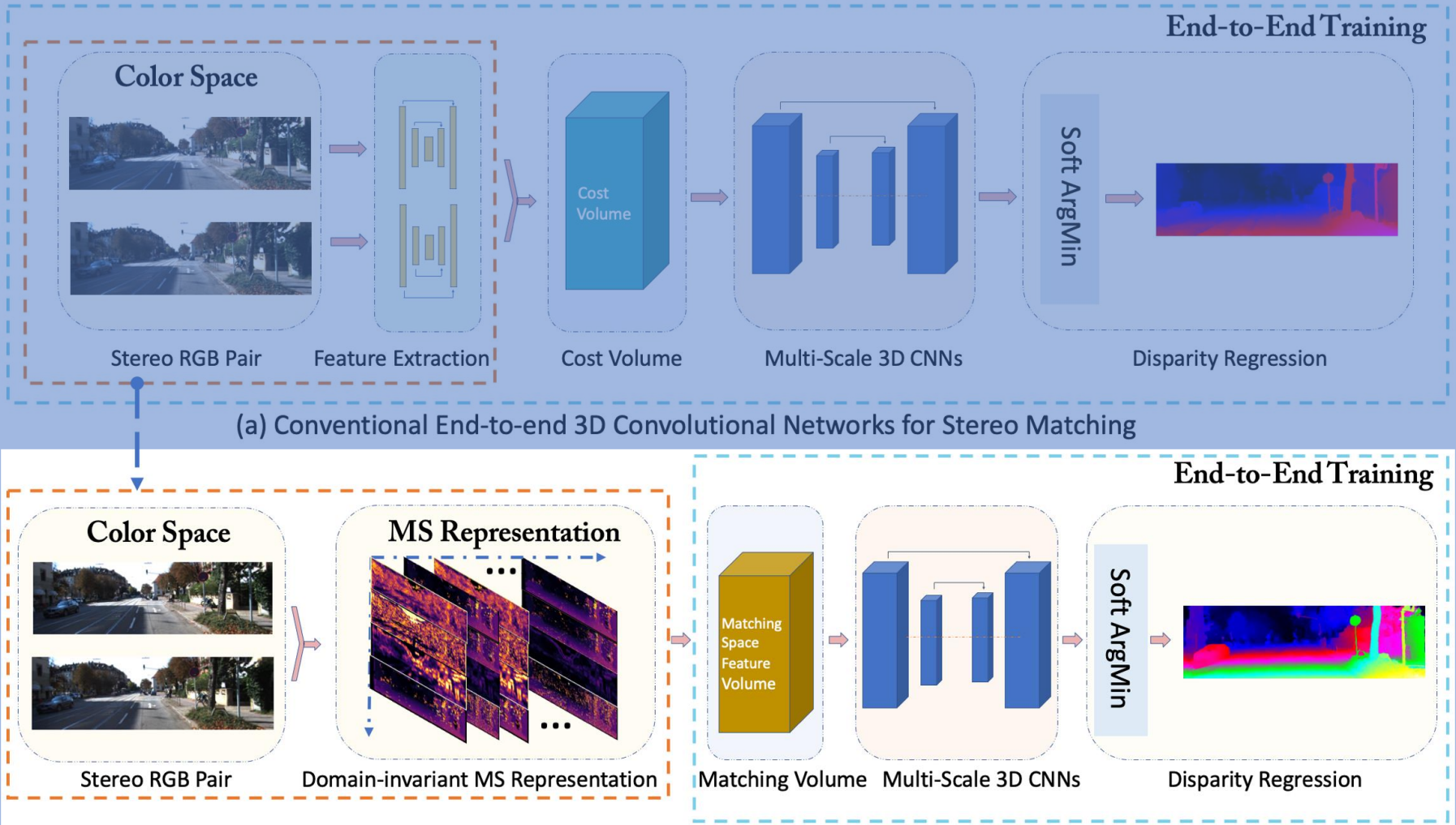

End-to-End Training
Color Space
Stereo RGB Pair
Feature Extraction
Cost Volume
Multi-Scale 3D CNNs
Soft ArgMin
Disparity Regression
(a) Conventional End-to-end 3D Convolutional Networks for Stereo Matching
End-to-End Training
Color Space
MS Representation
Matching Space Feature Volume
Soft ArgMin
Stereo RGB Pair
Domain-invariant MS Representation
Matching Volume
Multi-Scale 3D CNNs
Disparity Regression
(b) Our Matching Space (MS) End-to-End Stereo Networks for Cross-domain Generalization

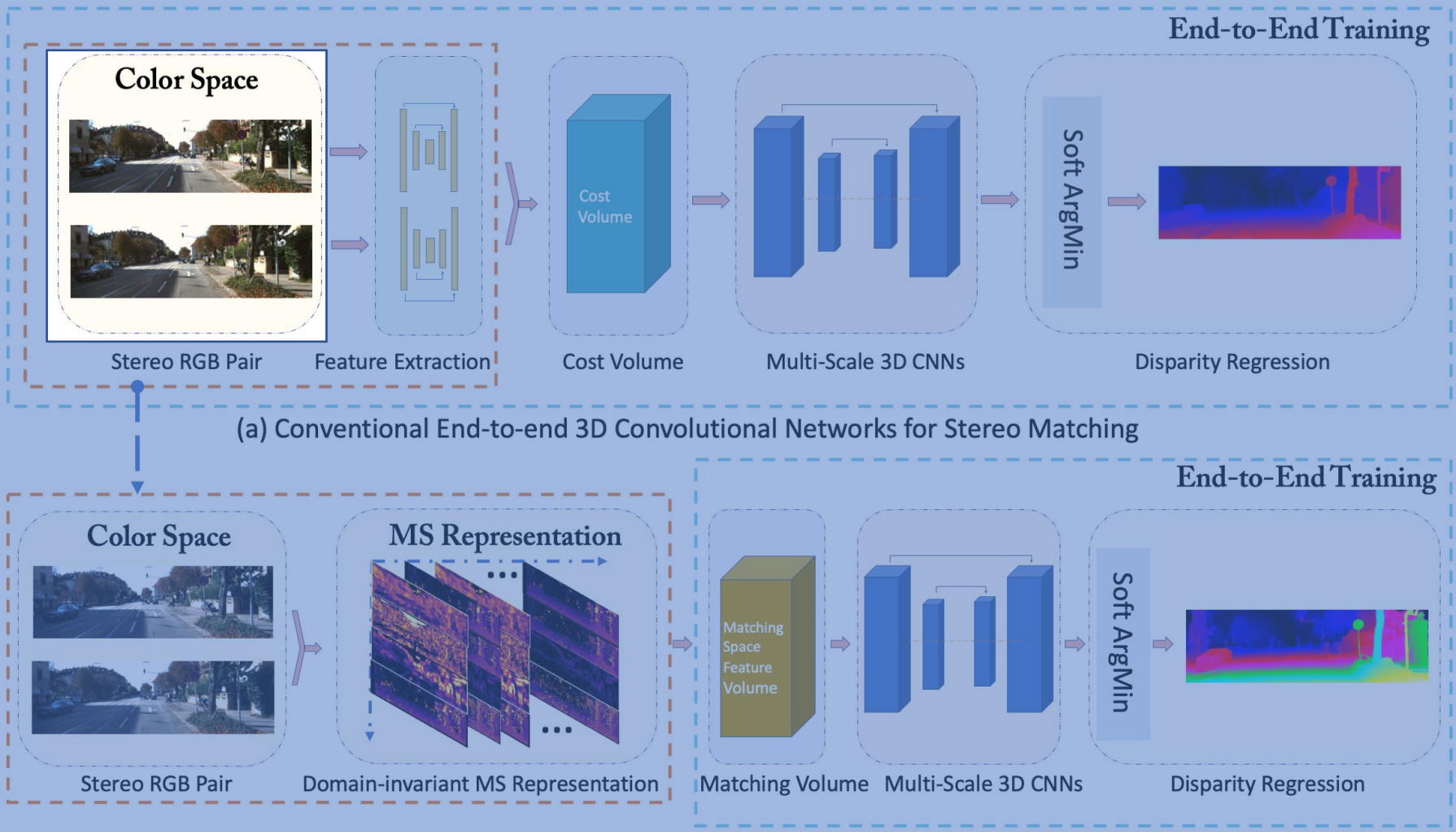

End-to-End Training
Color Space
Cost Volume
Soft ArgMin
Stereo RGB Pair
Feature Extraction
Cost Volume
Multi-Scale 3D CNNs
Disparity Regression
(a) Conventional End-to-end 3D Convolutional Networks for Stereo Matching
End-to-End Training
Color Space
MS Representation
Matching Space Feature Volume
Soft ArgMin
Stereo RGB Pair
Domain-invariant MS Representation
Matching Volume
Multi-Scale 3D CNNs
Disparity Regression
(b) Our Matching Space (MS) End-to-End Stereo Networks for Cross-domain Generalization

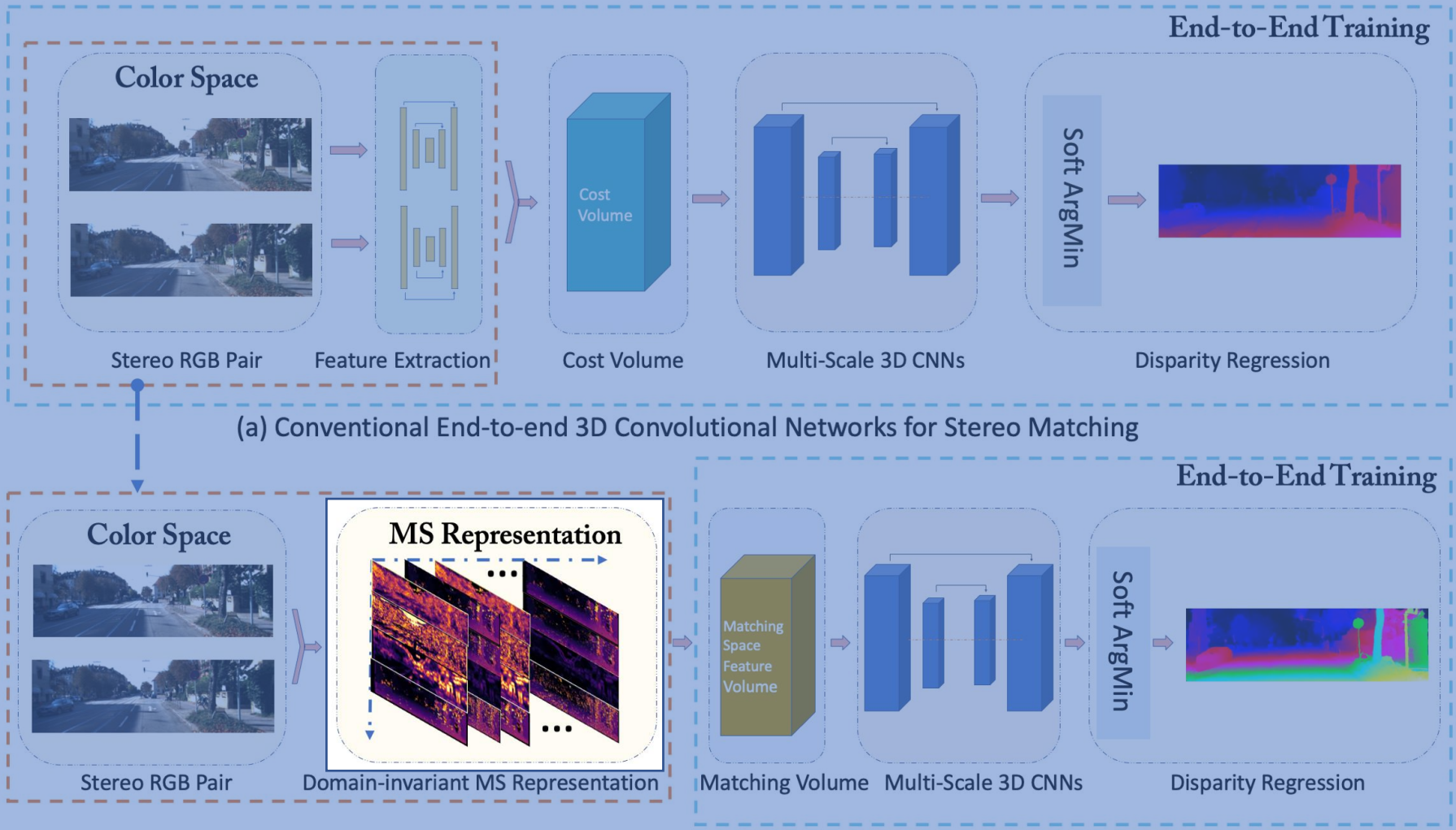

End-to-End Training
Color Space
Cost Volume
Soft ArgMin
Stereo RGB Pair
Feature Extraction
Cost Volume
Multi-Scale 3D CNNs
Disparity Regression
(a) Conventional End-to-end 3D Convolutional Networks for Stereo Matching
End-to-End Training
Color Space
MS Representation
Matching Space Feature Volume
Soft ArgMin
Stereo RGB Pair
Domain-invariant MS Representation
Matching Volume
Multi-Scale 3D CNNs
Disparity Regression
(b) Our Matching Space (MS) End-to-End Stereo Networks for Cross-domain Generalization

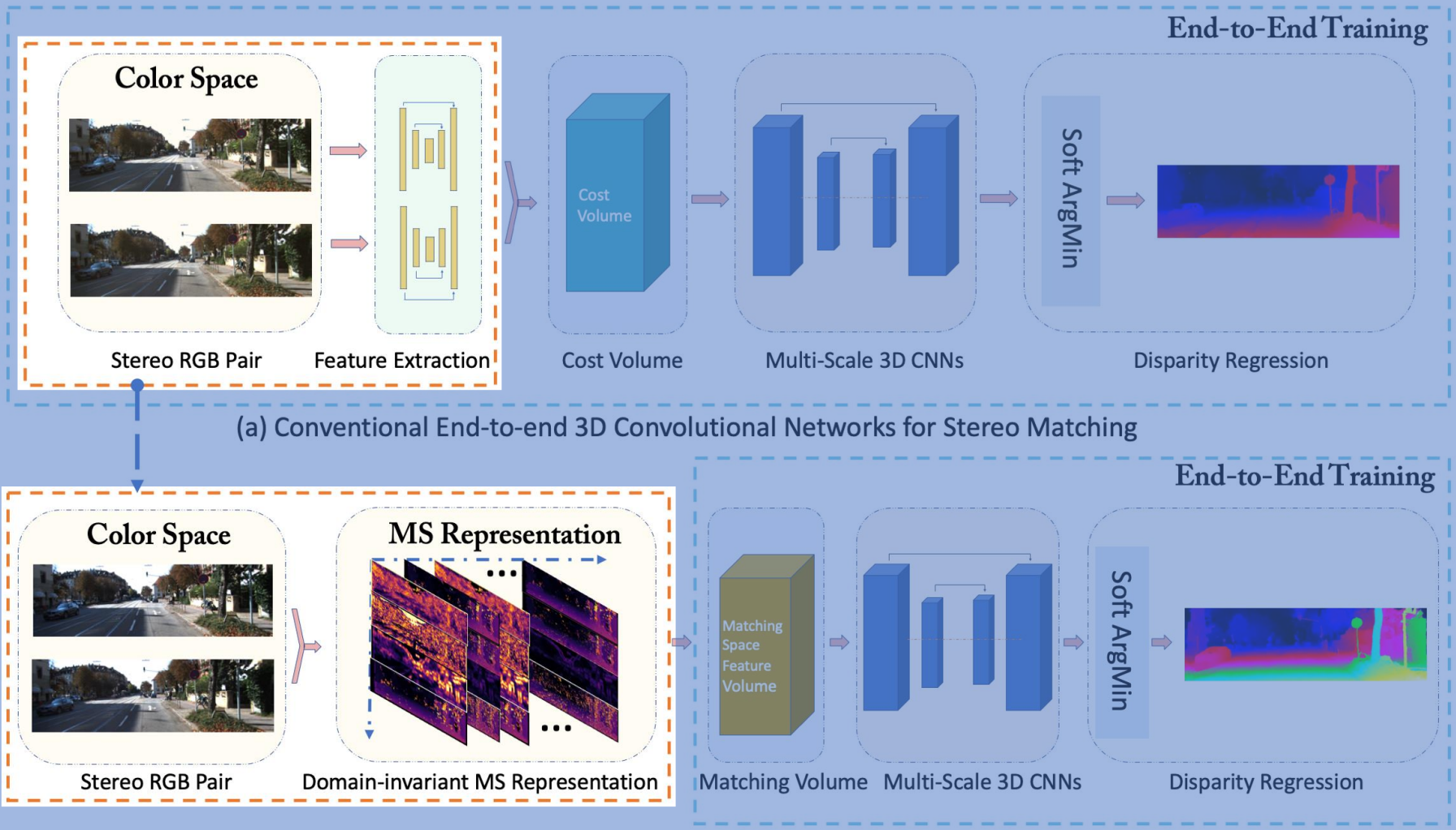

End-to-End Training
Color Space
Cost Volume
Soft ArgMin
Stereo RGB Pair
Feature Extraction
Cost Volume
Multi-Scale 3D CNNs
Disparity Regression
(a) Conventional End-to-end 3D Convolutional Networks for Stereo Matching
End-to-End Training
Color Space
MS Representation
Matching Space Feature Volume
Soft ArgMin
Stereo RGB Pair
Domain-invariant MS Representation
Matching Volume
Multi-Scale 3D CNNs
Disparity Regression
(b) Our Matching Space (MS) End-to-End Stereo Networks for Cross-domain Generalization

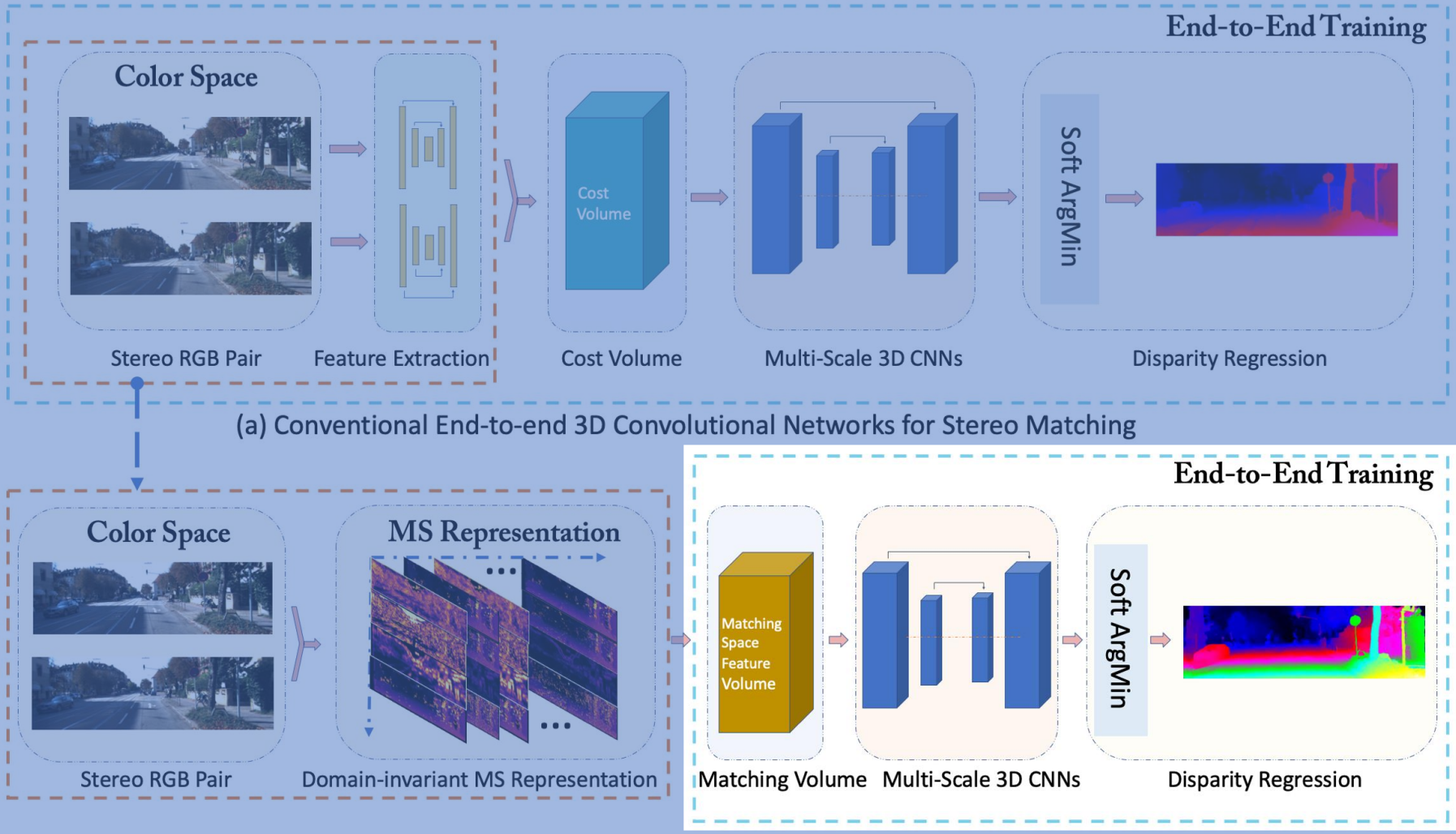

End-to-End Training
Color Space
Cost Volume
Soft ArgMin
Stereo RGB Pair
Feature Extraction
Cost Volume
Multi-Scale 3D CNNs
Disparity Regression
(a) Conventional End-to-end 3D Convolutional Networks for Stereo Matching
End-to-End Training
Color Space
MS Representation
Matching Space Feature Volume
Soft ArgMin
Stereo RGB Pair
Domain-invariant MS Representation
Matching Volume
Multi-Scale 3D CNNs
Disparity Regression
(b) Our Matching Space (MS) End-to-End Stereo Networks for Cross-domain Generalization


End-to-End Training
Color Space
Cost Volume
Soft ArgMin
Stereo RGB Pair
Feature Extraction
Cost Volume
Multi-Scale 3D CNNs
Disparity Regression
(a) Conventional End-to-end 3D Convolutional Networks for Stereo Matching
End-to-End Training
Color Space
MS Representation
Matching Space Feature Volume
Soft ArgMin
Stereo RGB Pair
Domain-invariant MS Representation
Matching Volume
Multi-Scale 3D CNNs
Disparity Regression
(b) Our Matching Space (MS) End-to-End Stereo Networks for Cross-domain Generalization


End-to-End Training
Color Space
Cost Volume
Soft ArgMin
Stereo RGB Pair
Feature Extraction
Cost Volume
Multi-Scale 3D CNNs
Disparity Regression
(a) Conventional End-to-end 3D Convolutional Networks for Stereo Matching
End-to-End Training
Color Space
MS Representation
Matching Space Feature Volume
Soft ArgMin
Stereo RGB Pair
Domain-invariant MS Representation
Matching Volume
Multi-Scale 3D CNNs
Disparity Regression
(b) Our Matching Space (MS) End-to-End Stereo Networks for Cross-domain Generalization


End-to-End Training
Color Space
Cost Volume
Soft ArgMin
Stereo RGB Pair
Feature Extraction
Cost Volume
Multi-Scale 3D CNNs
Disparity Regression
(a) Conventional End-to-end 3D Convolutional Networks for Stereo Matching
End-to-End Training
Color Space
MS Representation
Matching Space Feature Volume
Soft ArgMin
Stereo RGB Pair
Domain-invariant MS Representation
Matching Volume
Multi-Scale 3D CNNs
Disparity Regression
(b) Our Matching Space (MS) End-to-End Stereo Networks for Cross-domain Generalization


End-to-End Training
Color Space
Stereo RGB Pair
Feature Extraction
Cost Volume
Cost Volume
Multi-Scale 3D CNNs
Soft ArgMin
Disparity Regression
(a) Conventional End-to-end 3D Convolutional Networks for Stereo Matching
End-to-End Training
Color Space
MS Representation
Stereo RGB Pair
Domain-invariant MS Representation
Matching Space Feature Volume
Matching Volume
Multi-Scale 3D CNNs
Soft ArgMin
Disparity Regression
(b) Our Matching Space (MS) End-to-End Stereo Networks for Cross-domain Generalization


End-to-End Training
Color Space
Cost Volume
Soft ArgMin
Stereo RGB Pair
Feature Extraction
Cost Volume
Multi-Scale 3D CNNs
Disparity Regression
(a) Conventional End-to-end 3D Convolutional Networks for Stereo Matching
End-to-End Training
Color Space
MS Representation
Matching Space Feature Volume
Soft ArgMin
Stereo RGB Pair
Domain-invariant MS Representation
Matching Volume
Multi-Scale 3D CNNs
Disparity Regression
(b) Our Matching Space (MS) End-to-End Stereo Networks for Cross-domain Generalization

Matching Functions and Confidence Measures

- We adopt four matching functions and associated confidence scores
- Four matchers include
  - normalized cross correlation (NCC)
  - zero-mean sum of absolute differences (ZSAD)
  - census transform (CENSUS)
  - absolute differences of the horizontal Sobel operator (SOBEL)
- Four confidence scores
  - each matcher's likelihood, a confidence measure of each disparity for a given pixel
  - obtained by converting the cost curve to a probability density function for each disparity under consideration

Datasets: Sim2Real

- **Scene Flow (SF)**
  - synthetic dataset of 39k stereo pairs with dense ground truth
  - 3 subsets: Driving, Monkaa, and FlyingThings3D
- **KITTI 2015 (KT15) & KITTI 2012 (KT12)**
  - a real dataset of street views, with sparse ground truth captured by LiDAR
  - around 200 training and 200 testing stereo pairs
- **Middlebury 2014 (MB)**
  - Indoor scenes with high variability, with dense ground truth captured by structured light
  - 15 training, 15 testing and 12 extra stereo pairs
- **ETH3D Low-res two view (ETH)**
  - 27 training and 20 testing stereo pairs
  - Quasi-dense ground truth

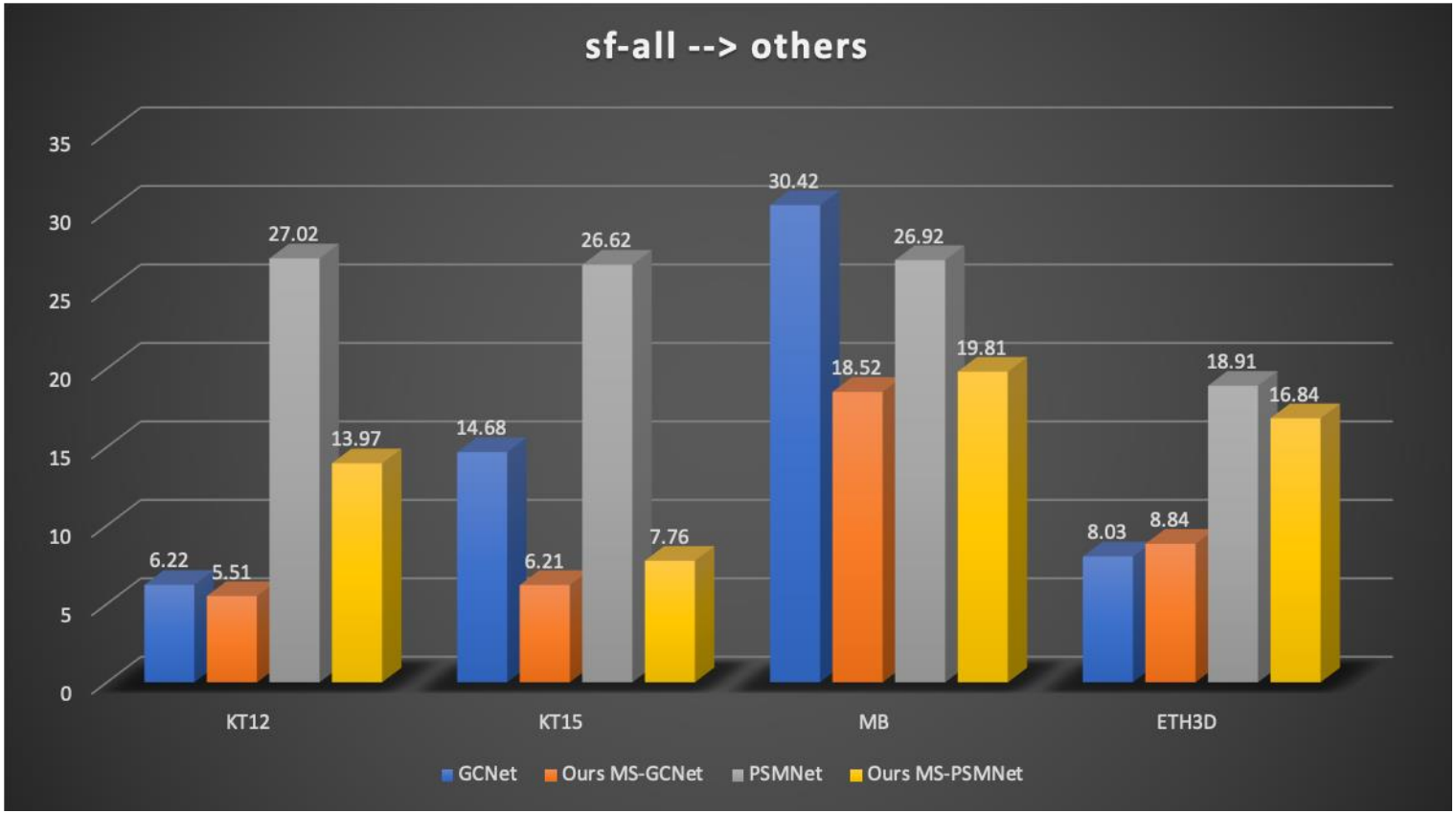
Scene Flow (SF)

Target 1&2: KITTI 2012&2015

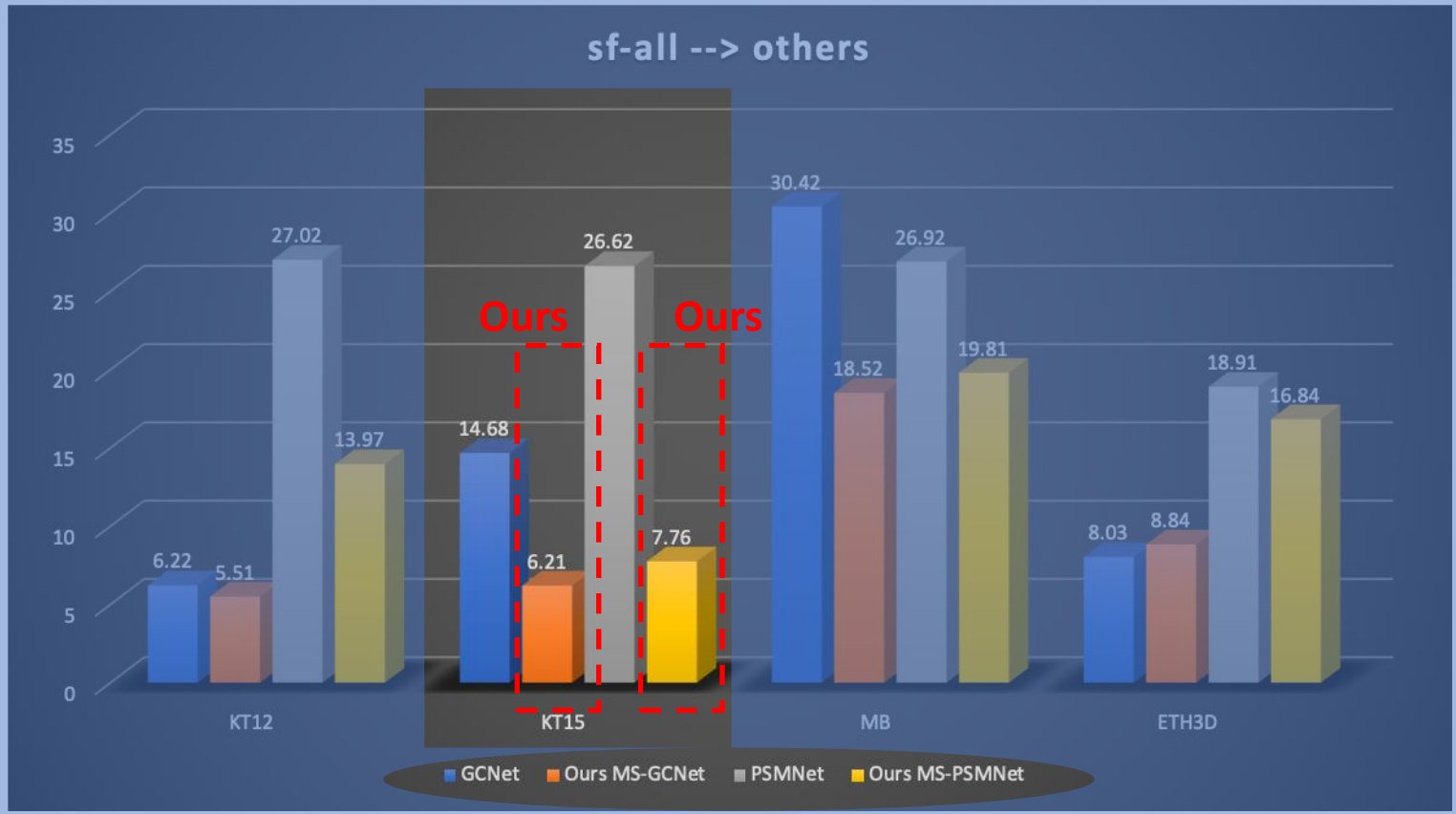
Target 3: Middlebury 2014

Target 4: ETH3D Low-res two view

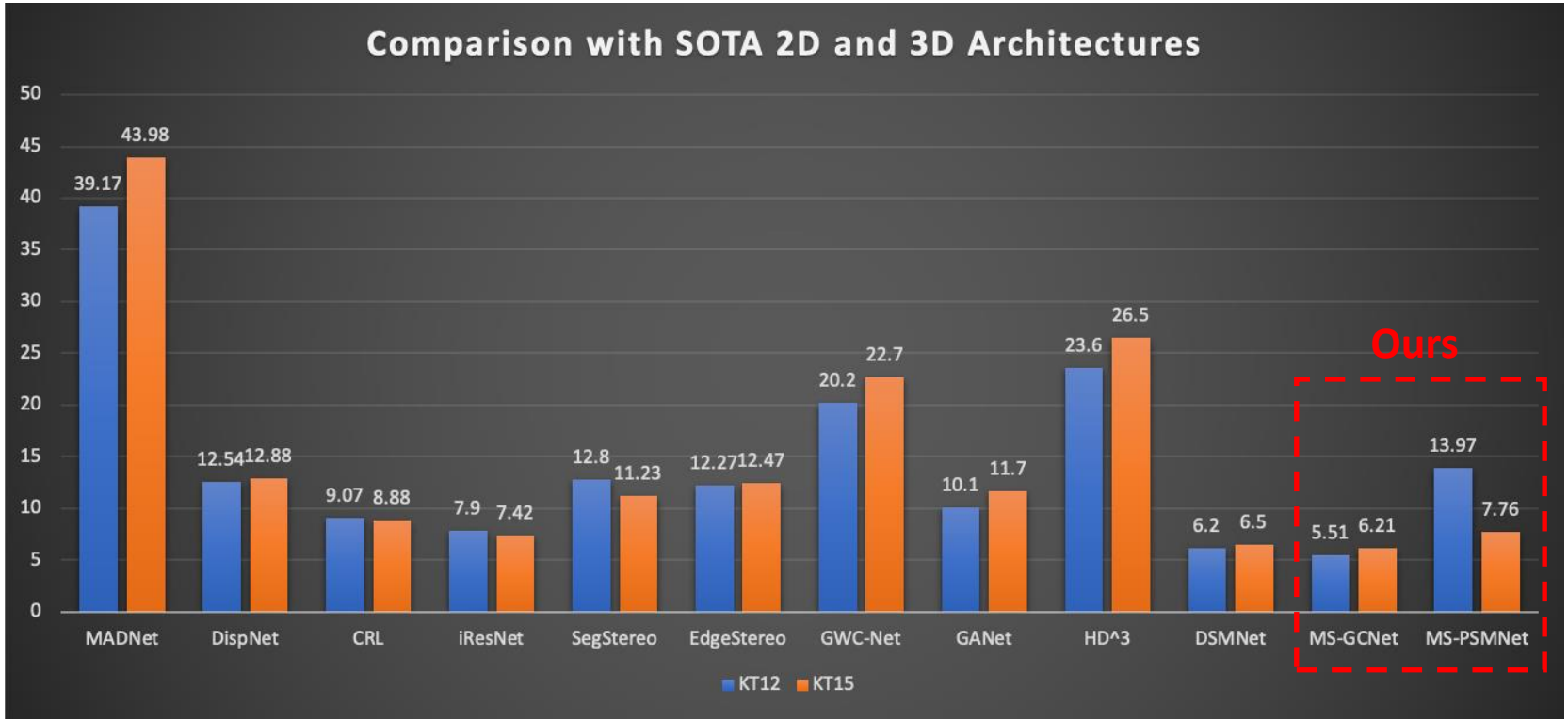
MSNets: Sim2Real (sf-all -> real)



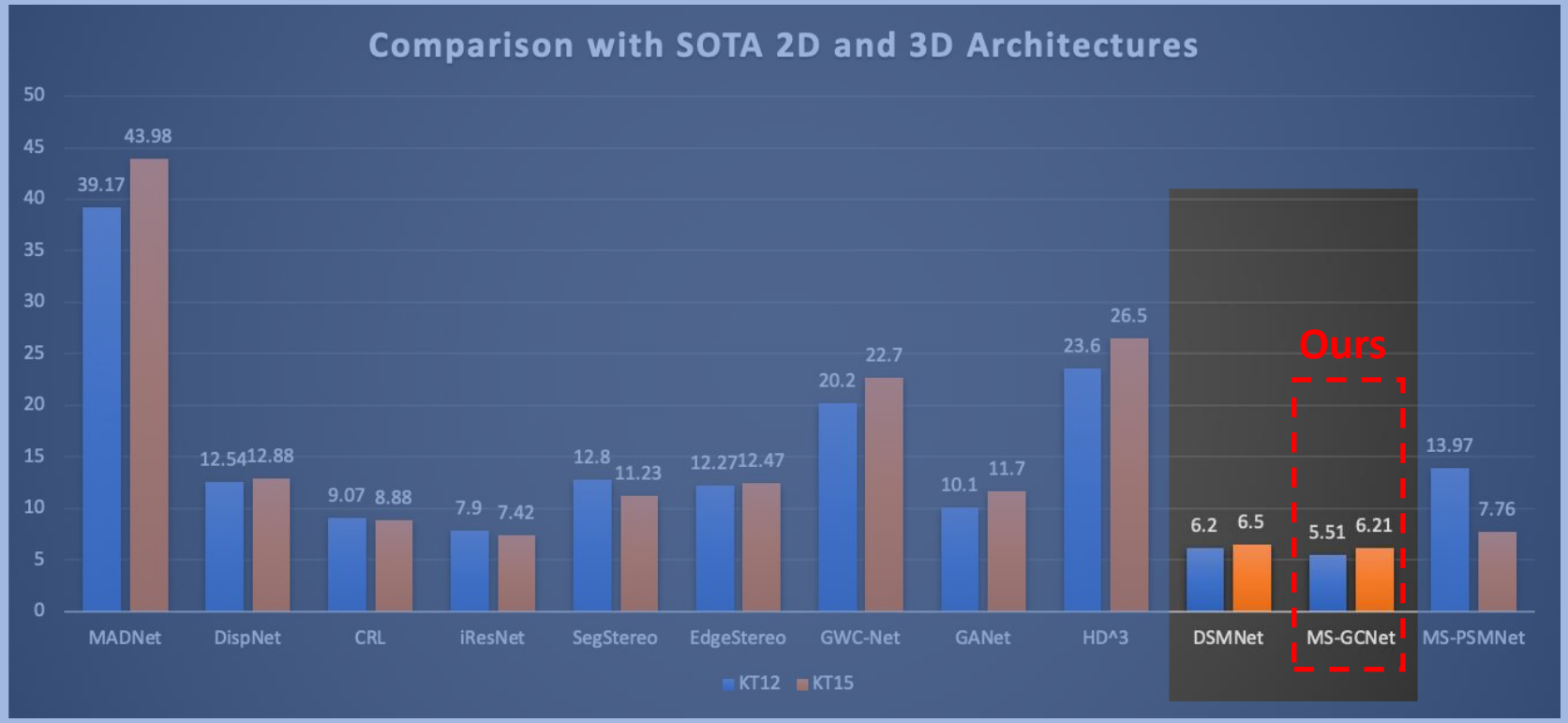
MSNets: Sim2Real (sf-all -> real)



MSNets: Sim2Real (VS SOTA Networks)



MSNets: Sim2Real (VS SOTA Networks)

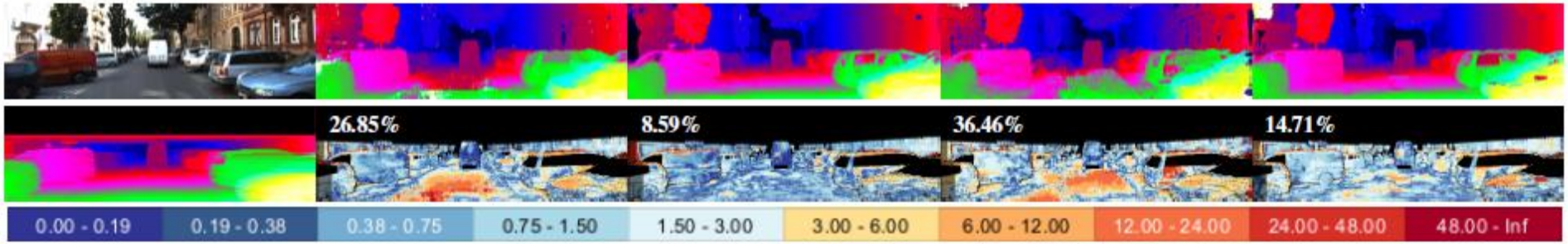


MSNets: Evaluation on Real Benchmark KITTI 2015

Models	All-D1 %			Noc-D1 %		
	bg	fb	all	bg	fg	all
MS-GCNet (Ours)	2.58	6.83	3.29	2.19	5.59	2.75
GC-Net	2.21	6.16	2.87	2.02	5.58	2.61
MS-PSMNet (Ours)	2.15	5.01	2.63	1.99	4.52	2.41
PSM-Net	**1.86**	**4.62**	**2.32**	**1.71**	**4.31**	**2.14**

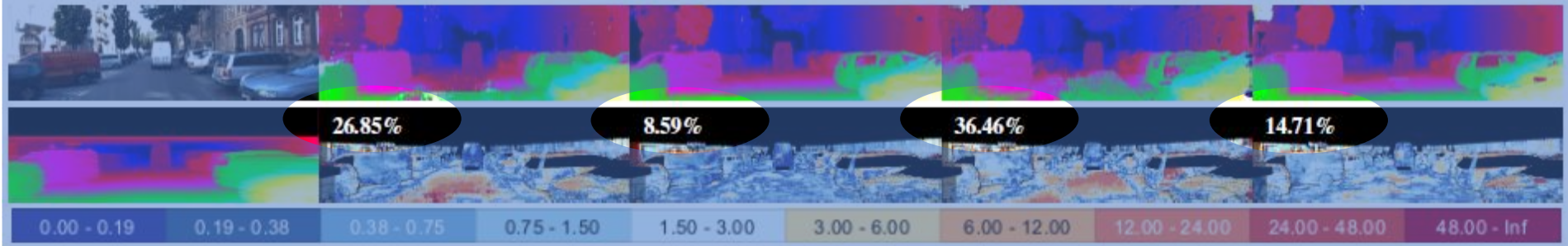
Test results on KITTI 2015 Benchmark

MSNets: Qualitative Results on KT15



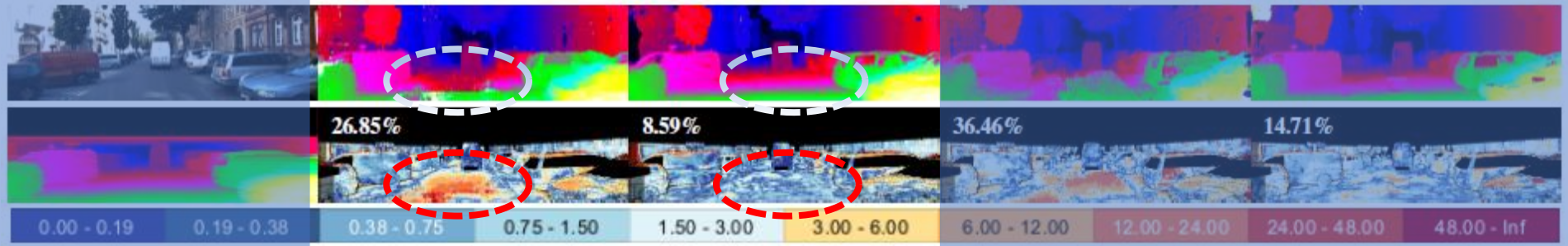
Input | GCNet | MS-GCNet (Ours) | PSMNet | MS-PSMNet (Ours)

MSNets: Qualitative Results on KT15



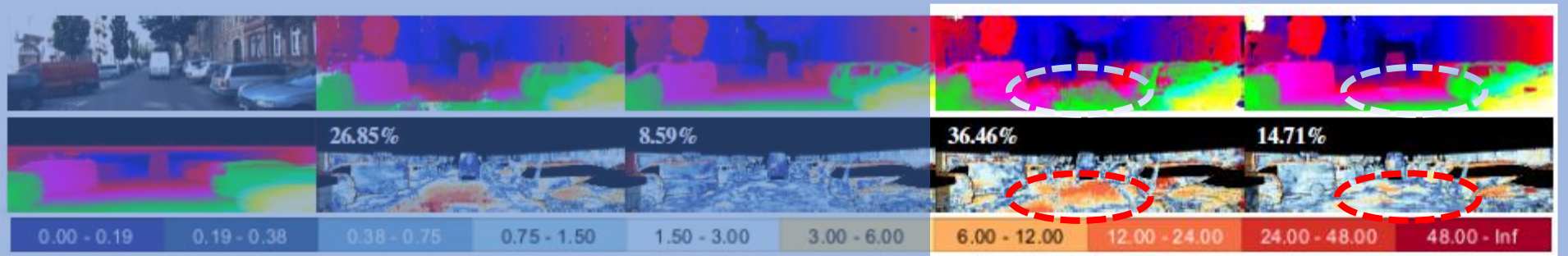
Input | GCNet | MS-GCNet (Ours) | PSMNet | MS-PSMNet (Ours)

MSNets: Qualitative Results on KT15



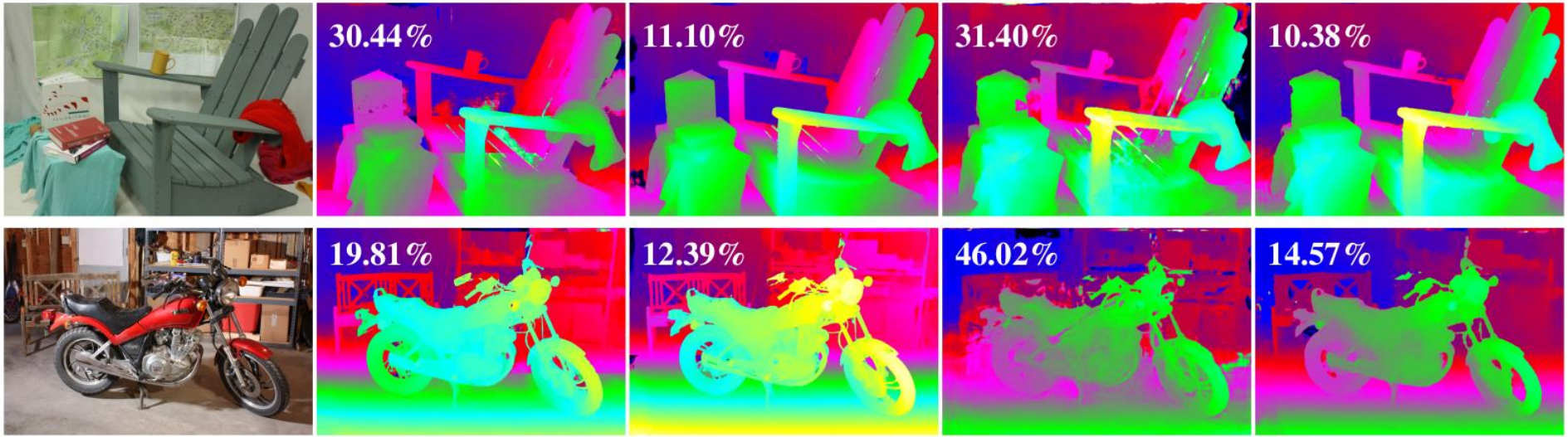
Input | GCNet | MS-GCNet (Ours) | PSMNet | MS-PSMNet (Ours)

MSNets: Qualitative Results on KT15



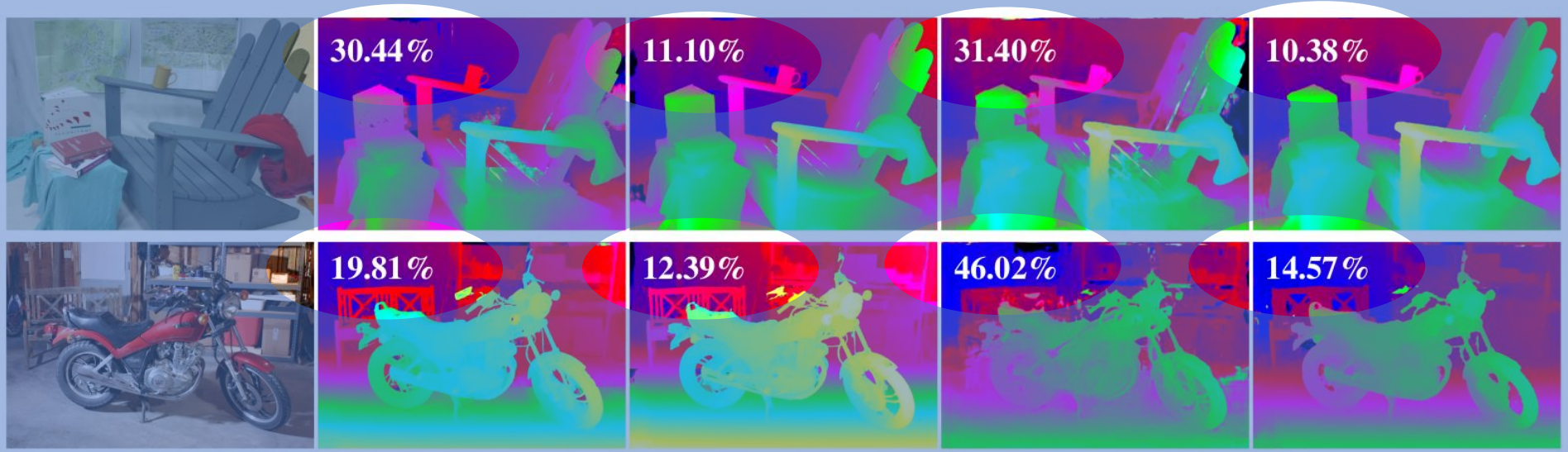
Input | GCNet | MS-GCNet (Ours) | PSMNet | MS-PSMNet (Ours)

MSNets: Qualitative Results on Middlebury



Input | GCNet | MS-GCNet (Ours) | PSMNet | MS-PSMNet (Ours)

MSNets: Qualitative Results on Middlebury



Input

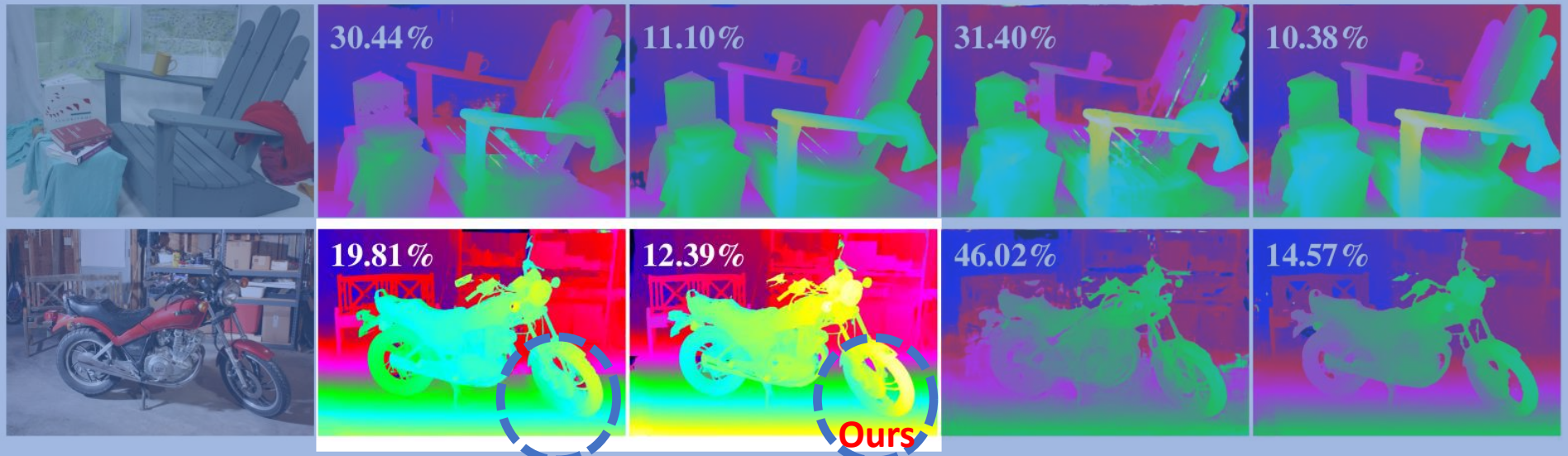
GCNet

MS-GCNet (Ours)

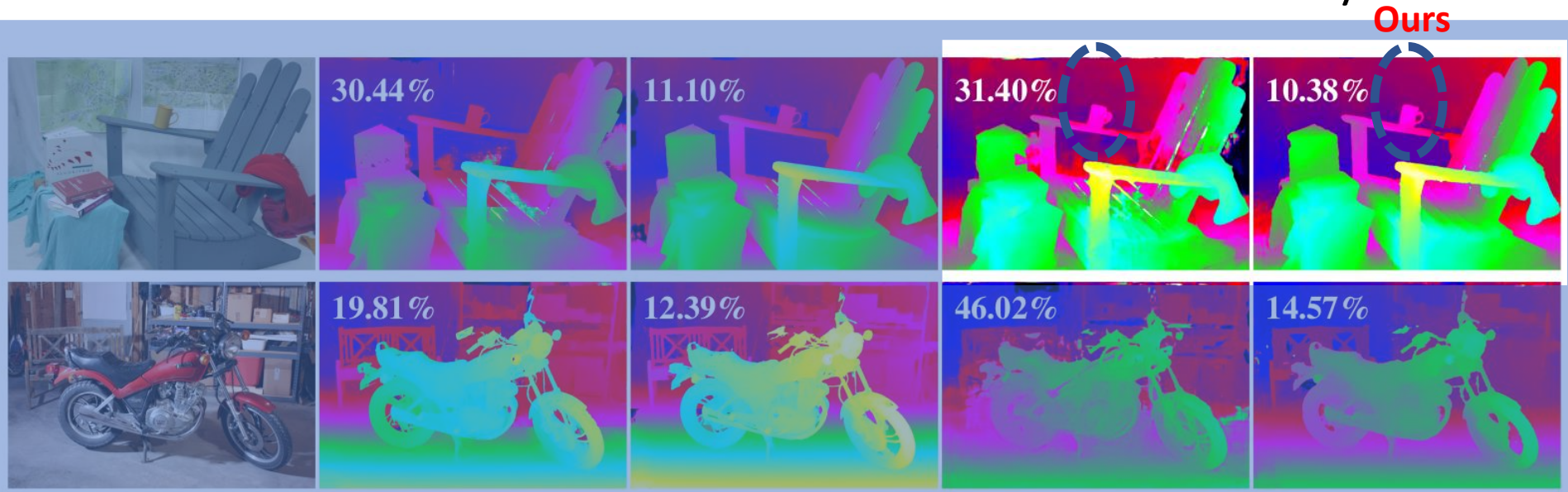
PSMNet

MS-PSMNet (Ours)

MSNets: Qualitative Results on Middlebury



MSNets: Qualitative Results on Middlebury

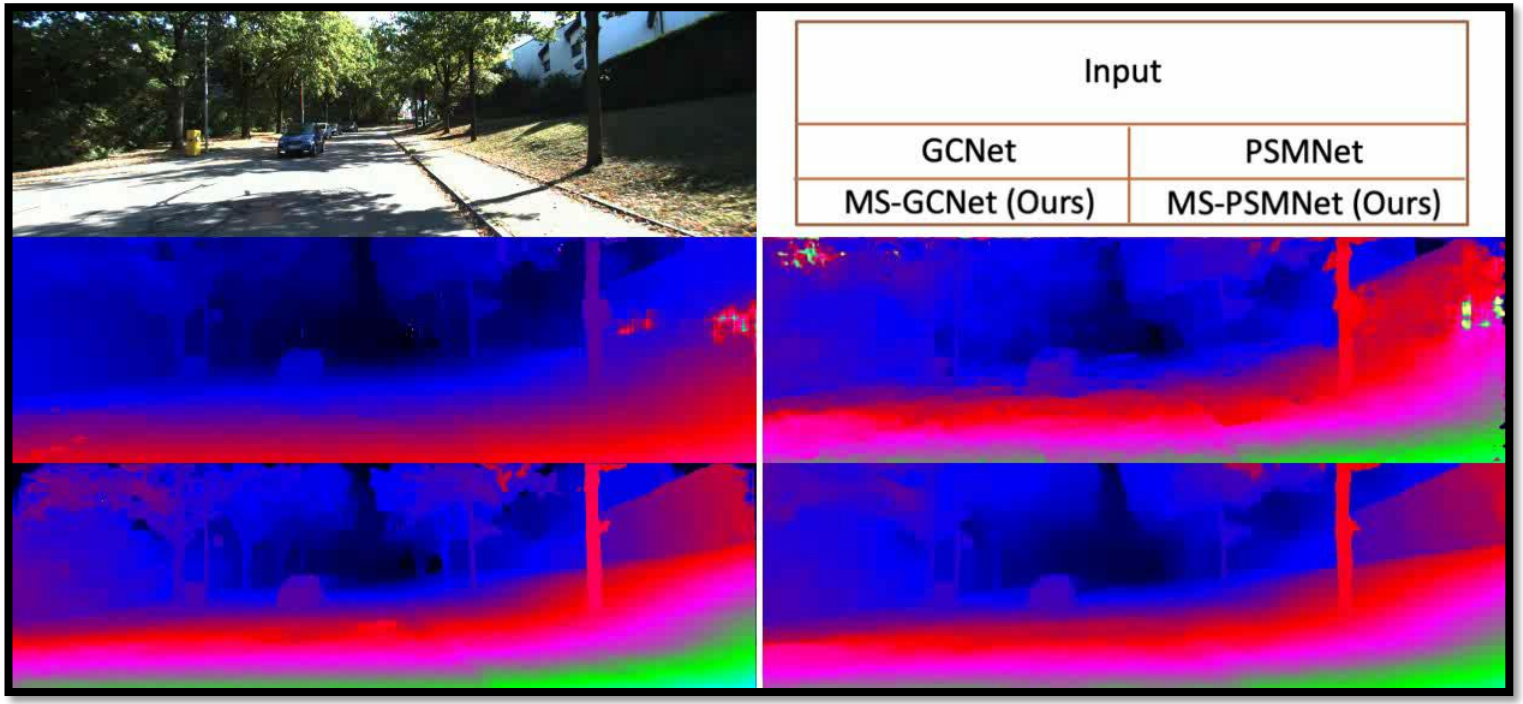


MSNets: Sim2Real (sf-all -> KT Raw)

Input

Baselines

Ours



Summary

- We show that not exposing CNNs directly to image appearance leads to better generalization properties
- A novel family of architectures, MS-Nets, and one of its possible implementations built on conventional wisdom and popular 3D networks
- An extensive set of experiments highlighting the behavior of both 3D and MS-Nets under domain shift
- Code is available at https://github.com/ccj5351/MS-Nets