# Matching-space Stereo Networks for Cross-domain Generalization

Changjiang Cai[1], Matteo Poggi[2], Stefano Mattoccia[2], Philippos Mordohai[1]
[1]Stevens Institute of Technology, [2]University of Bologna

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

## 1. Motivation

- Annotated data for stereo matching is challenging to collect
  - ➢ Expensive LiDAR, Stereo Camera Rig
  - ➢ Ground truth depth is **sparse**
- SOTA deep networks generalize poorly to unseen domain
  - ➢ E.g., PSMNet suffers large accuracy drops moving from synthetic (pretrained on Scene Flow) to real scenes (KT15)

Left Image  Disparity Map by PSMNet

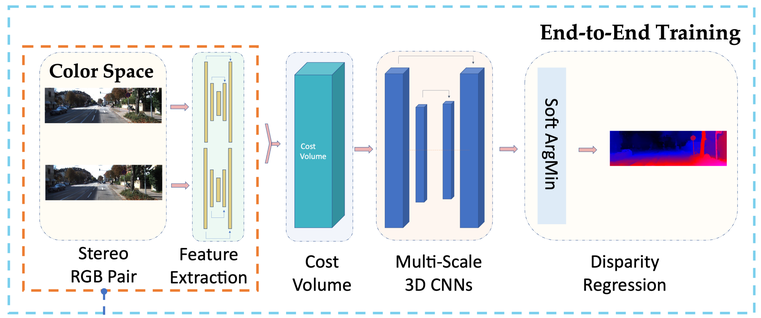
- **Domain generalization**
  - ➢ A method that generalizes well without adaptation is a solution
  - ➢ Effective in continuously changing environments, e.g. autonomous driving, without re-training or adaptation
- **Goal**
  - ➢ Sacrifice as little accuracy as possible to attain generalization

## 2. Over-specialization to Color Space

- Learning process is driven by image content
- Better generalization can be achieved by choosing a representation **insensitive** to common variations of RGBs

## 3. Matching Space Stereo Networks

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




## 4. Matching Functions and Confidences

- Four matching functions and the associated confidence scores
- Matchers include
  - ➢ normalized cross correlation (NCC)
  - ➢ zero-mean sum of absolute differences (ZSAD)
  - ➢ census transform (CENSUS)
  - ➢ absolute differences of the horizontal Sobel operator (SOBEL)
- 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

$$L_z(x_L, y, d) = \frac{\exp\left(-\frac{(C_z(x_L,y,d)-C_{z,min})^2}{2\sigma_z^2}\right)}{\sum_i \exp\left(-\frac{(C_z(x_L,y,d_i)-C_{z,min})^2}{2\sigma_z^2}\right)}$$

## 5. Experimental Results

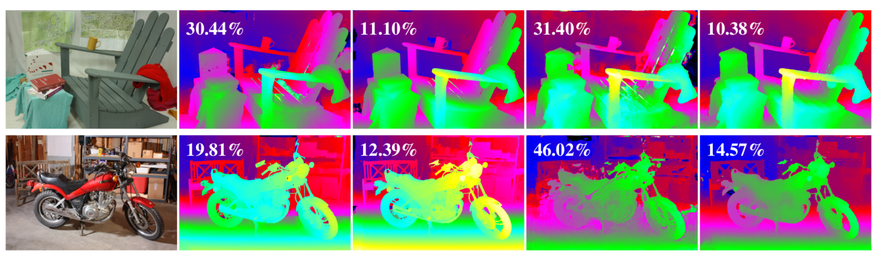
- **Domain Generalization Training and Evaluation**
  - ➢ Networks are trained in source domain Scene Flow
  - ➢ Evaluated in target domains (KITTI 2012&2015, Middlebury 2014 and ETH3D Low-res two view datasets) without finetuning or adaptation
- **Sim2Real (sf-all → real)**

| Target domains | Models (Sim2Real) | | | |
|---|---|---|---|---|
| | GCNet | MS-GCNet(Ours) | PSMNet | MS-PSMNet(Ours) |
| KT12 | 6.22 | **5.51** | 27.02 | **13.97** |
| KT15 | 14.68 | **6.21** | 26.62 | **7.76** |
| MB | 30.42 | **18.52** | 26.92 | **19.81** |
| ETH3D | **8.03** | 8.84 | 18.91 | **16.84** |

- **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** |

- **Qualitative Results on Middlebury**




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

- **Comparison with SOTA 2D and 3D architectures**

| Target Domains | MAD-Net | Disp-Net | CRL | iRes-Net | Seg-Stereo | Edge-Stereo | GWC-Net | GA-Net | HD3 | DSM-Net | MS-GCNet | MS-PSMNet |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| KT12 | 39.17 | 12.54 | 9.07 | 7.90 | 12.80 | 12.27 | 20.20 | 10.10 | 23.60 | 6.20 | **5.51** | 13.97 |
| KT15 | 43.98 | 12.88 | 8.88 | 7.42 | 11.23 | 12.47 | 22.70 | 11.70 | 26.50 | 6.50 | **6.21** | 7.76 |