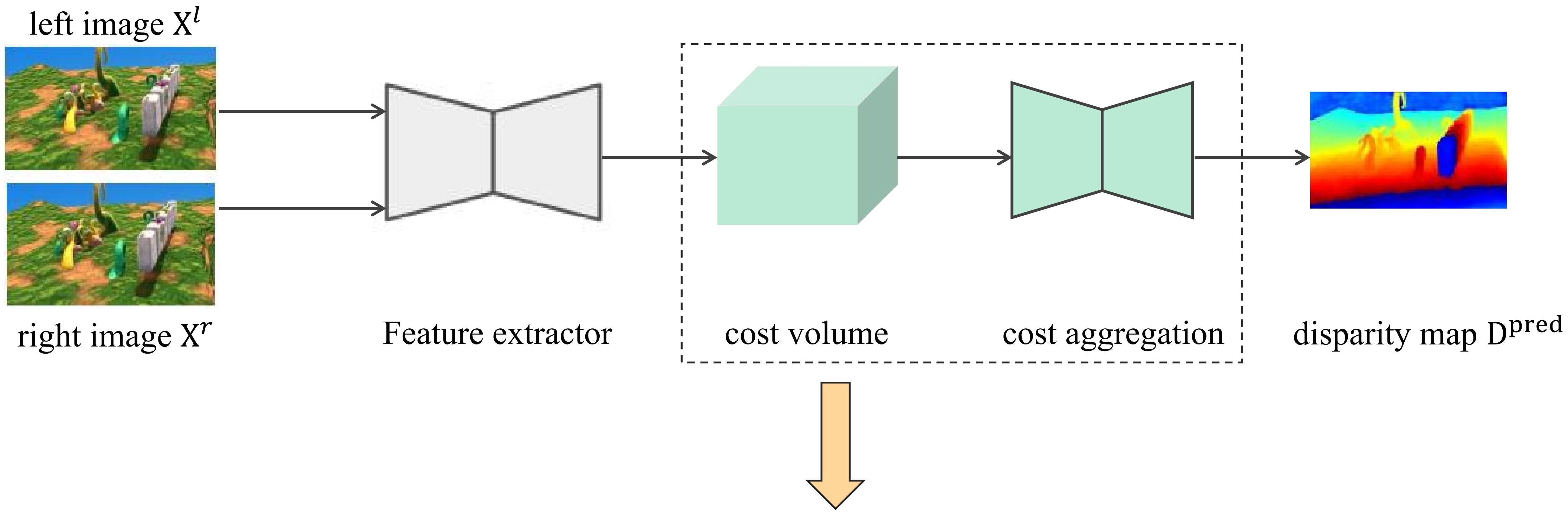
# Domain Generalized Stereo Matching via Hierarchical Visual Transformation

Tianyu Chang [1,3], Xun Yang [1*], Tianzhu Zhang [1], Meng Wang [2]
[1]University of Science and Technology of China [2]Hefei University of Technology
[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
cty8998@mail.ustc.edu.cn {xyang21, tzzhang}@ustc.edu.cn wangmeng@hfut.edu.cn

June 2023

# Stereo Matching (SM)

Goal: Generate a disparity map $D^{pred}$ of the left image : SM Network $F_{\Phi}(X^{l}, X^{r}) \rightarrow D^{pred}$



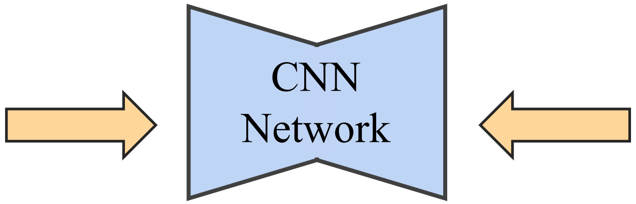

Existing Regular Stereo Matching Methods: Main Improvement Module

Problem: Poor Generalizability of Synthetic-to-realistic Domain

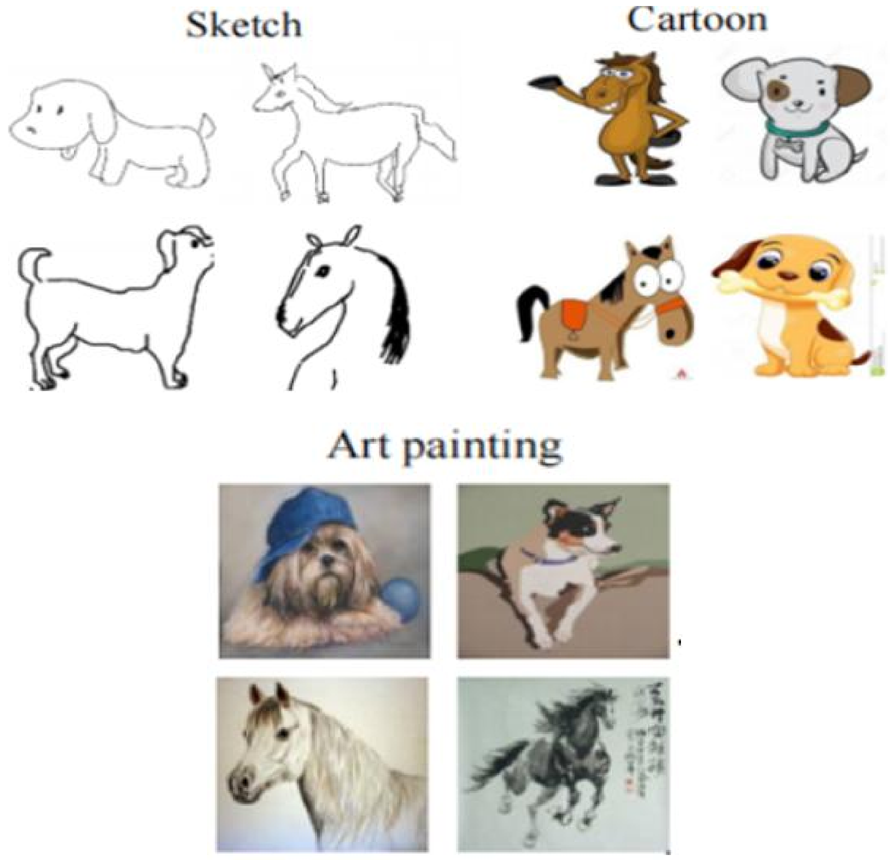
# Domain Generalization

Goal: Train a task model that generalizes well on the unseen target domain data with only source domain training dataset


Source Domain (Training Data)

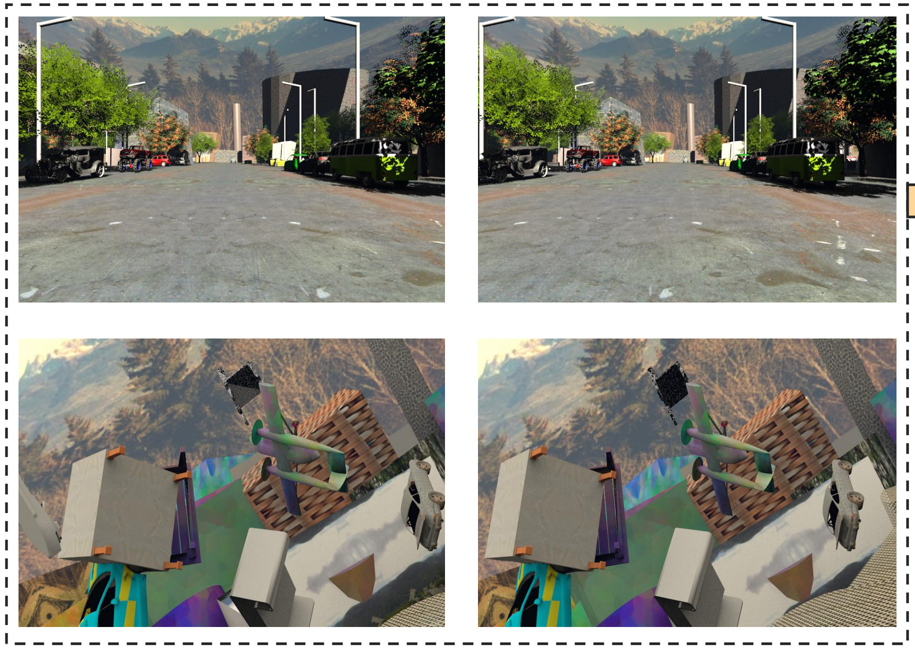



Task Model (Classification)

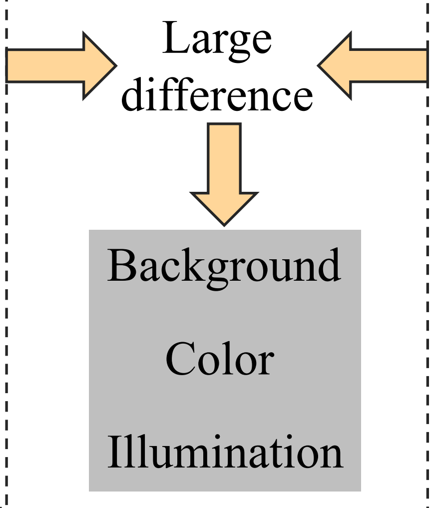



Target Domain (Testing Data)

# Domain Generalized Stereo Matching



Synthetic SM image paris (Training Data)





Realistic SM image paris (Testing Data)

Main Research Problem: how to train an effective SM network on only synthetic data to estimate reliable disparity map on unseen domain.

?

# Motivation

- Key for Domain Generalization

  Learn domain-invariant feature (causal feature)

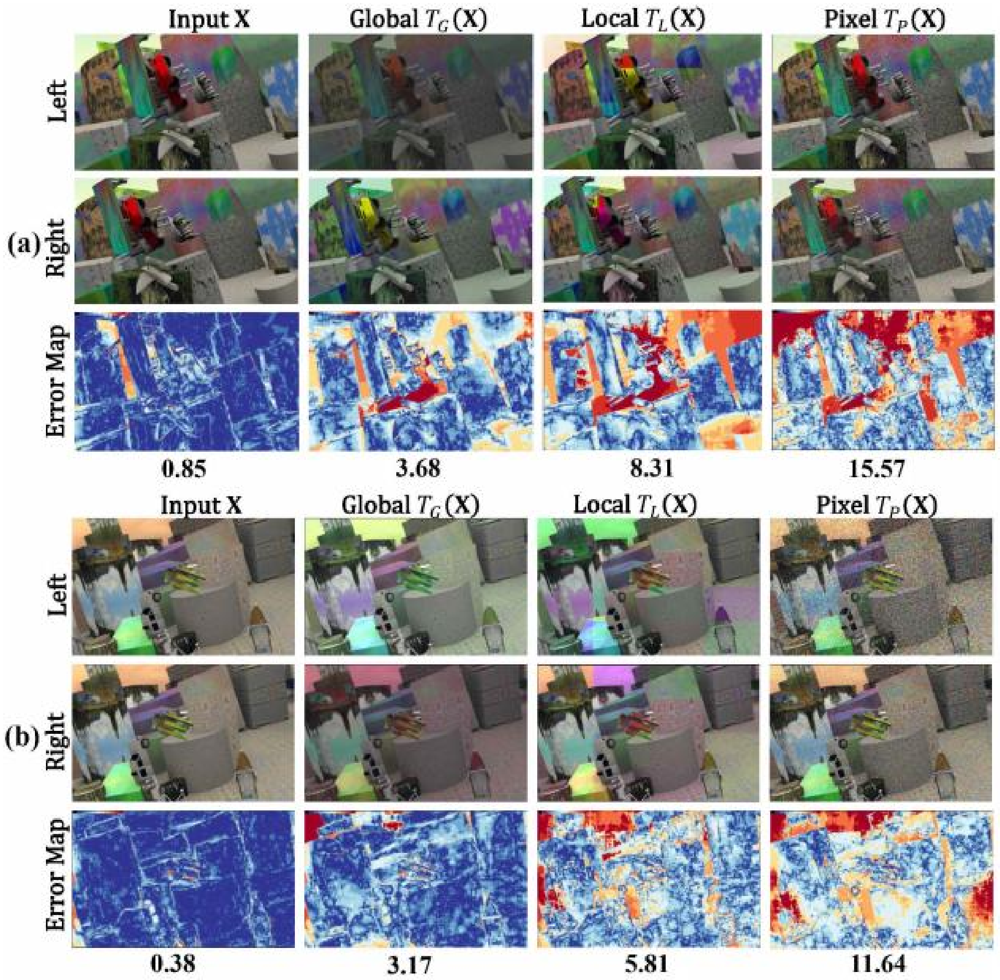
  Causal feature is invariant to certain transformations[1]

- Problem of Existing SM Networks

  Exploiting common artifacts (e.g. consistent local RGB color statistics and overreliance on local chromaticity features) of synthetic stereo images as shortcuts[2].

- Intuitive Idea

  Leverage the visual transformations that do not change the underlying domain-invariant feature to increase the diversity of training domain, thereby enhancing the generalization performance of SM network

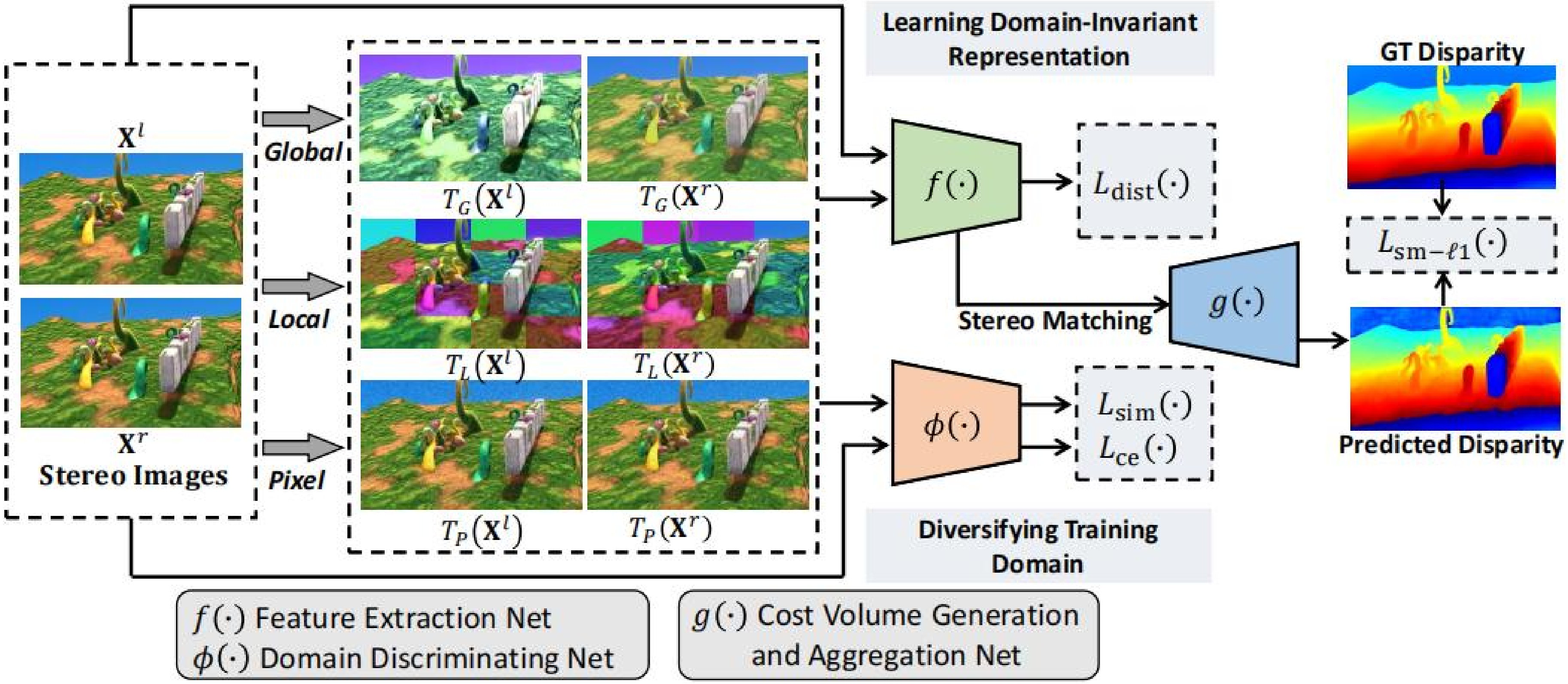



Visualized example of three transformations


[1] Ruoyu Wang. Out-of-distribution generalization with causal invariant transformations. CVPR2022

[2] WeiQin Chuah. Itsa: An information-theoretic approach to automatic shortcut avoidance and domain generalization in stereo matching networks. CVPR2022

# The pipeline of our domain-generalized SM approach




1. **Hierarchical Visual Transformation**: Diversify the distribution of training domain from three complementary perspectives: Global, Local, and Pixel.
2. **Learning Objectives**:
   - Maximizing Cross-Domain Visual Discrepancy: $\min L_{\mathrm{sim}}(\mathbf{X}) = \frac{1}{3}\sum_{J} \mathrm{Cos}\left(\phi(T_J(\mathbf{X})), \phi(\mathbf{X})\right)$, $\min L_{\mathrm{ce}}(\mathbf{X}) = \mathrm{CE}\left(\{\phi(T_J(\mathbf{X})), \phi(\mathbf{X})\}, \mathcal{Y}_d\right)$
   - Minimizing Cross-Domain Feature Inconsistency: $\min L_{\mathrm{dist}}(\mathbf{X}) = \frac{1}{3}\sum_{J} \|f\left(T_J(\mathbf{X})\right) - f\left(\mathbf{X}\right)\|_2$

# Performance comparison with SOTA domain generalized SM networks

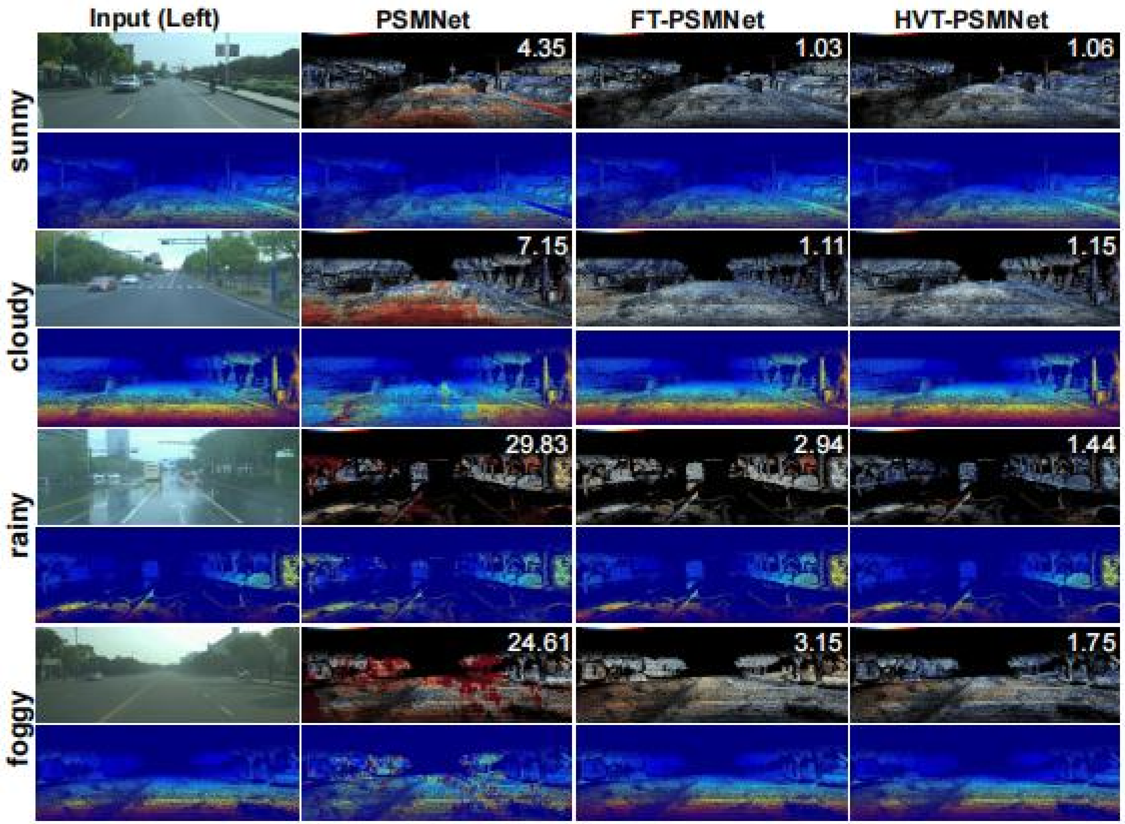
| Baselines | Methods | KITTI 2015 | | KITTI 2012 | | Middlebury(H) | | ETH3D | | References |
|---|---|---|---|---|---|---|---|---|---|---|
| | | EPE | D1(3px) | EPE | D1(3px) | EPE | D1(2px) | EPE | D1(1px) | |
| - | GANet [42] | 2.31 | 11.7 | 1.93 | 10.1 | 5.41 | 20.3 | 1.33 | 14.1 | CVPR 2019 |
| | CasStereo [9] | 2.42 | 11.9 | 2.12 | 11.8 | 3.71 | 17.2 | 0.87 | 7.8 | CVPR 2020 |
| | DSMNet [43] | 1.46 | 6.5 | 1.26 | 6.2 | 2.62 | 13.8 | 0.69 | 6.2 | ECCV 2020 |
| PSMNet [3] | PSMNet [3] | 3.17 | 16.3 | 2.69 | 15.1 | 7.65 | 34.2 | 2.33 | 23.8 | CVPR 2018 |
| | MS-PSMNet [2] | 1.64* | 7.8 | 2.33* | 14.0 | 4.72* | 19.8 | 1.42* | 16.8 | 3DV 2020 |
| | FC-PSMNet [46] | 1.58* | 7.5 | 1.42* | 7.0 | 4.14* | 18.3 | 1.25* | 12.8 | CVPR 2022 |
| | ITSA-PSMNet [5] | 1.39* | 5.8 | 1.09* | 5.2 | 3.25* | 12.7 | 0.94* | 9.8 | CVPR 2022 |
| | Graft-PSMNet [17] | 1.32 | 5.3 | 1.09 | 5.0 | 2.34 | 10.9 | 1.16 | 10.7 | CVPR 2022 |
| | **HVT-PSMNet** | **1.14±0.02** | **4.9±0.12** | **0.93±0.02** | **4.3±0.06** | **1.46±0.13** | **10.2±0.16** | **0.47±0.03** | **6.9±0.23** | Ours |
| GwcNet [10] | GwcNet [10] | 3.43 | 22.7 | 2.77 | 20.2 | 7.23 | 37.9 | 2.78 | 54.2 | CVPR 2019 |
| | FC-GwcNet [46] | 1.72* | 8.0 | 1.45* | 7.4 | 5.14* | 21.1 | 1.13* | 11.7 | CVPR 2022 |
| | ITSA-GwcNet [5] | 1.33* | 5.4 | 1.02* | 4.9 | 2.73* | 11.4 | 0.62* | 7.1 | CVPR 2022 |
| | **HVT-GwcNet** | **1.15±0.02** | **5.0±0.11** | **0.88±0.02** | **3.9±0.13** | **1.29±0.13** | **10.3±0.21** | **0.46±0.08** | **5.9±0.26** | Ours |
| CFNet [25] | CFNet [25] | 1.71 | 6.0 | 1.04 | 5.2 | 3.24 | 15.4 | 0.48 | 5.72 | CVPR 2021 |
| | ITSA-CFNet [5] | **1.09** | **4.7** | 0.87 | 4.2 | 1.87 | 10.4 | 0.45 | 5.1 | CVPR 2022 |
| | **HVT-CFNet** | 1.10±0.04 | 4.9±0.16 | **0.85±0.02** | **4.0±0.14** | **1.79±0.22** | **10.2±0.16** | **0.39±0.02** | **4.5±0.24** | Ours |
| RAFT [16] | RAFT [16] | 1.26 | 5.7 | 1.01 | 5.1 | 1.92 | 12.6 | 0.36 | 3.3 | 3DV 2021 |
| | **HVT-RAFT** | **1.12±0.02** | **5.2±0.09** | **0.87±0.02** | **3.7±0.08** | **1.37±0.11** | **10.4±0.14** | **0.29±0.01** | **3.0±0.09** | Ours |

- The synthetic-to-realistic generalization performances of all the baselines are consistently improved by our HVT in all settings.
- The improvements of generalization performance brought by HVT on the Middlebury and ETH3D datasets seem to be much larger that those on the KITTI 2012 and 2015 datasets.
- Our HVT-enhanced methods almost outperform all the SOTA methods except ITSA-CFNet on KITTI 2015.

# Robustness to Complex Realistic Scenarios

| Methods | Sunny | Cloudy | Rainy | Foggy | Avg. |
|---|---|---|---|---|---|
| PSMNet [3] | 62.5 | 60.1 | 60.5 | 68.6 | 63.9 |
| FT-PSMNet [5] | **4.0** | **2.9** | 11.5 | 6.5 | 6.3 |
| FC-PSMNet [46] | 4.9 | 4.3 | **7.2** | 6.2 | 5.7 |
| ITSA-PSMNet [5] | 4.8 | 3.2 | 9.4 | 6.3 | 5.9 |
| **HVT-PSMNet** | 4.2 | 3.1 | 8.7 | **5.6** | **5.4** |
| GwcNet [10] | 18.1 | 24.7 | 28.2 | 28.3 | 24.8 |
| FT-GwcNet [5] | **3.1** | **2.5** | 12.3 | 6.0 | 6.0 |
| ITSA-GwcNet [5] | 4.4 | 3.3 | 9.8 | 5.9 | 5.9 |
| **HVT-GwcNet** | 3.4 | 3.5 | **8.6** | **5.6** | **5.3** |

Robustness comparison of different methods on the DrivingStereo [3] dataset collected from complex realistic scenarios: Sunny, Cloudy, Rainy, and Foggy.

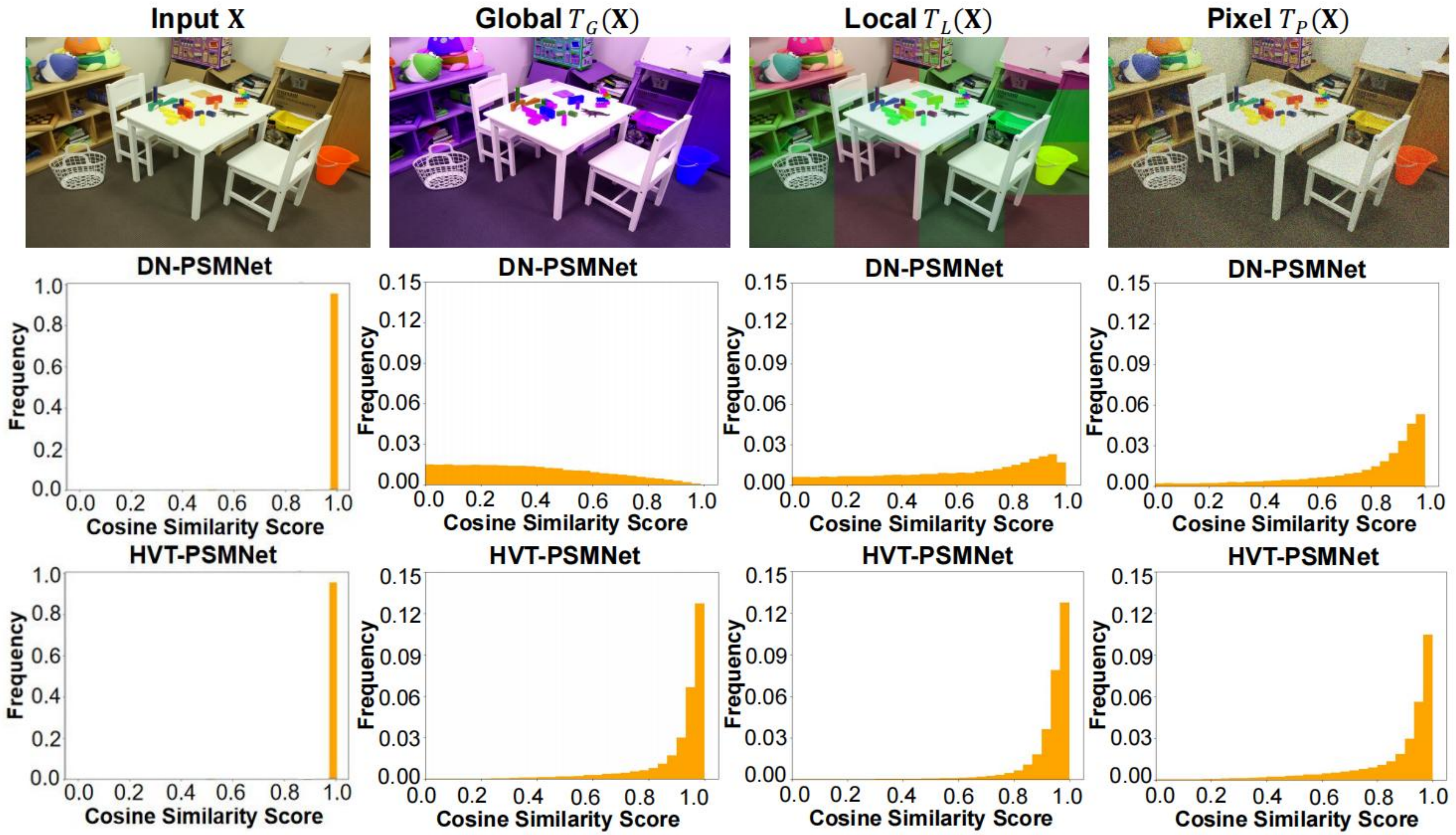



Qualitative results on the DrivingStereo [3] dataset.

Our methods obtain the best overall performance (5.4% and 5.3%) w.r.t. the average D1 error rate over the four groups of weather conditions, which demonstrates the efficacy of HVT and the strong robustness of HVT-based methods.

[3] Guorun Yang. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. CVPR2019

# Learning Domain-Invariant features




Histograms of feature cosine similarity scores respectively on DN-PSMNet model (see second row) and HVT-PSMNet model (see third row) between original feature and original, global transformed, local transformed and pixel transformed features.

Thank you for your careful listening