# SemStereo: Semantic-Constrained Stereo Matching Network for Remote Sensing

**Chen Chen[1,2], Liangjin Zhao[1], Yuanchun He[1], Yingxuan Long[1,2], Kaiqiang Chen[1*], Zhirui Wang[1], Yanfeng Hu[1], Xian Sun[1]**

[1]Key Laboratory of Target Cognition and Application Technology, Aerospace Information Research Institute, Chinese Academy of Sciences
[2]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences
{chenchen235, longyingxuan23}@mails.ucas.ac.cn, {zhaolj004896, heyc, chenkq, wangzr, huyf, sunxian}@aircas.ac.cn

## Abstract

Semantic segmentation and 3D reconstruction are two fundamental tasks in remote sensing, typically treated as separate or loosely coupled tasks. Despite attempts to integrate them into a unified network, the constraints between the two heterogeneous tasks are not explicitly modeled, since the pioneering studies either utilize a loosely coupled parallel structure or engage in only implicit interactions, failing to capture the inherent connections. In this work, we explore the connections between the two tasks and propose a new network that imposes semantic constraints on the stereo matching task, both implicitly and explicitly. Implicitly, we transform the traditional parallel structure to a new cascade structure termed Semantic-Guided Cascade structure, where the deep features enriched with semantic information are utilized for the computation of initial disparity maps, enhancing semantic guidance. Explicitly, we propose a Semantic Selective Refinement (SSR) module and a Left-Right Semantic Consistency (LRSC) module. The SSR refines the initial disparity map under the guidance of the semantic map. The LRSC ensures semantic consistency between two views via reducing the semantic divergence after transforming the semantic map from one view to the other using the disparity map. Experiments on the US3D and WHU datasets demonstrate that our method achieves state-of-the-art performance for both semantic segmentation and stereo matching.

## Introduction

Semantic segmentation and stereo matching are two underlying tasks towards semantic urban 3D reconstruction, which requires both semantic and 3d details derived from high-resolution remote sensing images (Kadhim, Mourshed, and Bray 2016; Bosch et al. 2019). They are usually considered as two independent tasks due to the inherent domain gap that characterizes each task, using a semantic segmentation network (Xie et al. 2021; Jing et al. 2021; Kang et al. 2021, 2022) for classification and a stereo matching network for height extraction (Zhang et al. 2019a, 2020; Xu et al. 2022, 2023a,b), respectively. Afterwards, post-processing is conducted to fuse the parallel results (Qin et al. 2019; Kunwar et al. 2020; Sun et al. 2024).

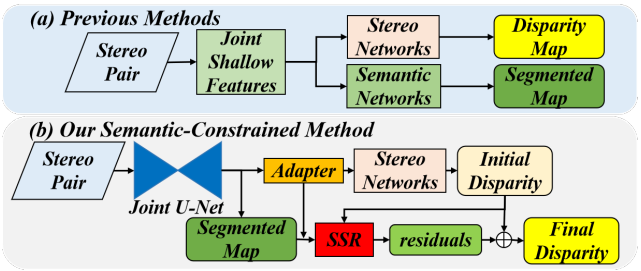


Figure 1: A comparison between our and previous methods.

Further studies (Cheng et al. 2017; Kendall, Gal, and Cipolla 2018; Song et al. 2020; Liao et al. 2023) attempt to fuse the two heterogeneous tasks in a multi-task network to achieve higher accuracy. The typical multi-task learning methods usually adopt a parallel structure with two branches for semantic segmentation and stereo matching respectively, as illustrated in Figure 1 (a). Shallow features are shared to establish implicit and weak connections between the distinct tasks (Zhang et al. 2019b; Dovesi et al. 2020; Liao et al. 2023). Even though it is demonstrated that semantic segmentation improves stereo matching in areas with less texture and occlusions (Yang et al. 2018; Wu et al. 2019; Dovesi et al. 2020), and stereo matching aids in distinguishing confusing categories in semantic segmentation (Liao et al. 2023; Yang et al. 2024), the underlying mechanisms remain unexplored, leading to a failure to capture the inherent connections between the heterogeneous tasks.

In this work, we aim to uncover the connections between semantic categories and disparities in remote sensing and bridge the task domain gap, with the goal of guiding efficient and interpretable network design for improved accuracy. As illustrated in Figure 2, disparities corresponding to the same category are concentrated within a distinct and narrow range in remote sensing images, while this characteristic does not apply to typical images taken from a ground-level perspective. We assume that this might (partially) explain why the heterogeneous tasks can mutually benefit each other. Based on this assumption, we propose a novel network termed *SemStereo* which imposes the semantic constraints on the stereo matching task via implicitly and explicitly modeling their connections.

*Corresponding author.

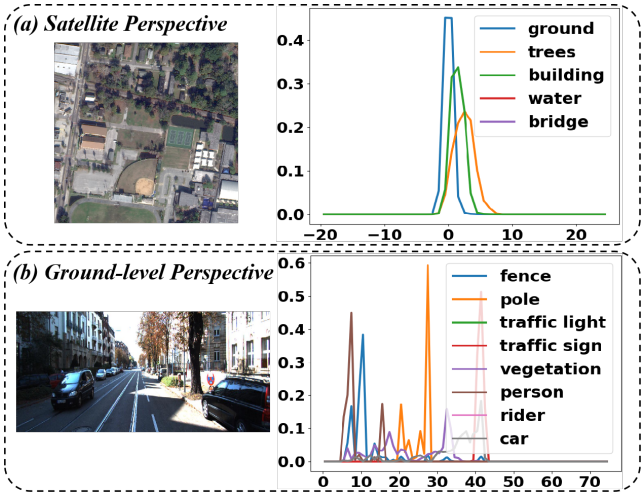


Figure 2: The disparity distribution per semantic category in satellite and ground-level perspectives, respectively. Disparity on the x-axis, probability on the y-axis.

Specifically, we firstly transform the traditional parallel structure (Figure 1 (a)) into a novel cascade structure termed Semantic-Guided Cascade (SGC) structure (Figure 1 (b)). Instead of only sharing shallow features, we implicitly strengthen the semantic constraints on stereo matching via feeding the deep features right before the segmentation map enriched with semantic information to the stereo network for the initial disparity generation. Explicitly, we propose a Semantic Selective Refinement (SSR) module, which uses semantic segmentation maps to guide the disparity maps refinement via learning the residuals for compensation, as illustrated in Figure 1 (b). Furthermore, we propose a Left-Right Semantic Consistency (LRSC) module, which explicitly imposes the semantic constraint on stereo matching via reducing the divergence of segmentation maps (or semantic feature maps in the absence of semantic supervision) between the two views after converting one view to another based on the disparity map, as illustrated in Figure 3 (c). In summary, we achieve tighter semantic constraints on stereo matching via the implicit deeper semantic feature sharing, the explicit semantic-guided disparity refinement, and the explicit disparity-based cross-view semantic consistency.

Our method achieves state-of-the-art performance on the US3D (Le Saux et al. 2019a,b; Bosch et al. 2019) and WHU (Liu and Ji 2020) datasets for both semantic segmentation and stereo matching tasks, demonstrating the effectiveness of SemStereo. Further ablation studies highlight the significance of modeling the connections between semantic categories and disparities.

## Related Work

**Stereo Matching.** Cost volume construction is crucial in stereo matching, encompassing methods like correlation volume (Mayer et al. 2016; Luo, Schwing, and Urtasun 2016), concatenation volume (Zbontar and LeCun 2015; Kendall et al. 2017) and their combinations (Guo et al. 2019; Shen, Dai, and Rao 2021; Xu et al. 2022). The correlation volume (Mayer et al. 2016; Luo, Schwing, and Urtasun 2016) is derived via computing the inner product between the feature volumes of the two views, producing a single-channel map per disparity level depicting the similarity. Despite its computational efficiency, the collapsed sing-channel map loses the abundant contextual information in the feature volume, reducing accuracy. The concatenation volume (Kendall et al. 2017; Zbontar and LeCun 2015) addresses this by concatenating the feature volumes from both views (Zbontar and LeCun 2015; Kendall et al. 2017), but it lacks explicit modeling of similarity measurements. Further studies (Guo et al. 2019; Shen, Dai, and Rao 2021) aims to combine the strengths of both methods, integrating contextual information and explicitly modeling similarity by concatenating both the correlation volume and concatenation volume. Nevertheless, these methods (Kendall et al. 2017; Zbontar and LeCun 2015; Guo et al. 2019; Shen, Dai, and Rao 2021) suffer from a higher computational cost due to the high dimensions of the cost volumes and the subsequent 3D convolutions. DSMNet (He et al. 2021) simplifies the regularization network via replacing the regular 3D convolutions with factorized 3D convolutions. ACVNet (Xu et al. 2022, 2023b) introduces an attention concatenation volume, which learns attention weights based on the correlation volume to filter the concatenation volume, thereby reducing the need for computationally expensive 3D convolutions. The fast version Fast-ACV (Xu et al. 2022) further enhances efficiency by using the top-k disparity priors from the coarse branch, achieving a better balance between accuracy and speed. In this work, we use the Fast-ACV to derive the initial disparity map, followed by refinement branches, but our focus is on exploring efficient methods and the potential of improving stereo matching via incorporating semantic constraints in remote sensing.

**Stereo Matching Using Semantic Clues.** A pioneering study (Hane et al. 2013) integrates the semantic segmentation results with 3D information through post-processing, based on a category-specific smoothness assumption. It is then expanded to larger, higher-resolution oblique aerial scenes (Blaha et al. 2016) using a hierarchical scheme. These works are non-deep learning methods, resulting in inferior results. Further studies (Qin et al. 2019; Kunwar et al. 2020) improve the performance by separately training two separate networks for distinct tasks. However, these approaches remain limited to post-processing, failing to fully leverage semantic information for stereo matching.

An alternative approach involves integrating pre-acquired semantic labels into stereo matching networks by expanding the input dimensions (Bosch et al. 2019). However, obtaining these semantic labels is highly time-consuming, which limits practical applications. Multi-task learning methods (Zhang et al. 2019b; Dovesi et al. 2020) address this limitation by jointly predicting a semantic segmentation map and a stereo map within an end-to-end trainable network. Experimental analyses reveal that stereo matching benefits from semantic segmentation in textureless (Yang et al. 2018; Wu et al. 2019; Dovesi et al. 2020) and occluded regions, while stereo matching helps to clarify confusing categories

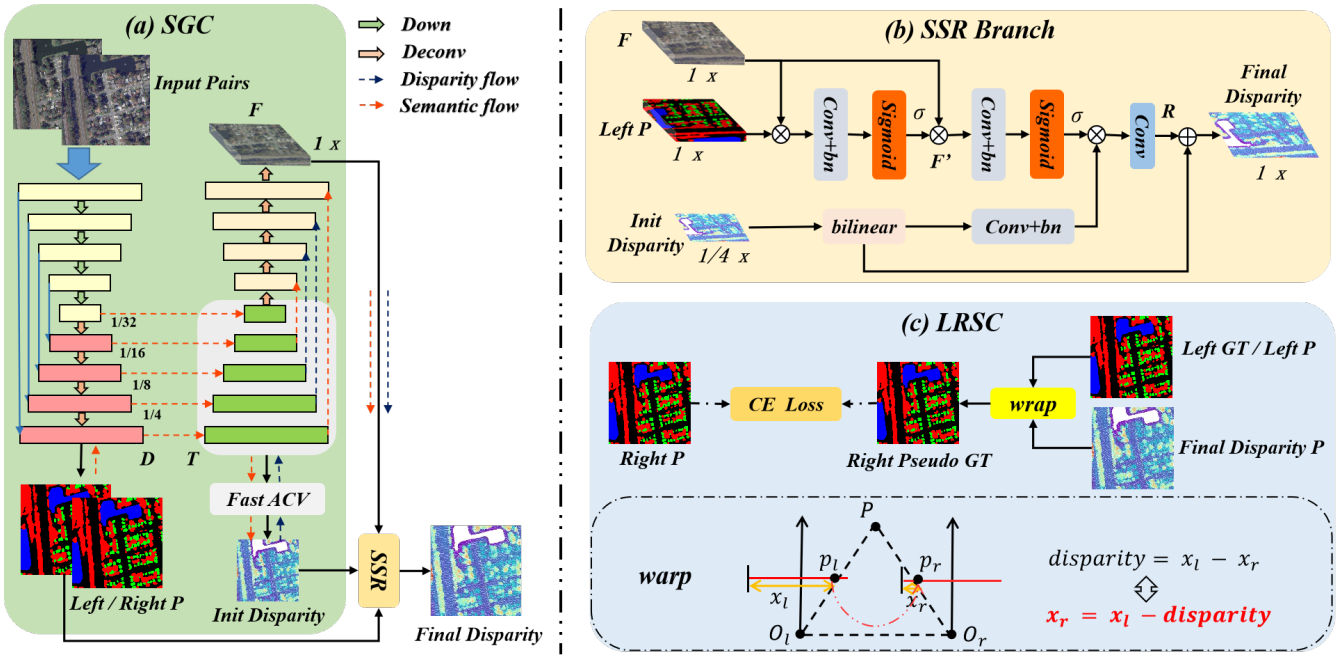


Figure 3: An overview of the SemStereo. It involves (a) a Semantic-Guided Cascade (SGC) structure for generating segmentation and initial disparity maps, (b) a Semantic Selective Refinement (SSR) branch refines the initial disparity under the guidance of semantic information, and (c) a Left-Right Semantic Consistency (LRSC) supervision. P: Prediction, GT: Ground Truth.

for semantic segmentation (Liao et al. 2023). These methods typically employ two parallel branches for the two tasks, sharing a shallow feature extractor, which leads to a weak and implicit coupling between them. A further study (Yang et al. 2024) generates one segmentation map and one disparity map from the same cost volume, strengthening the influence of semantic clues on stereo matching. However, the influence of the semantic clues on the stereo matching remains implicit, and the underlying mechanism is not well understood. In this work, we dive into the connections between semantic categories and stereo matching from the remote sensing perspective and propose a new framework. It both implicitly strengthens the impact of semantic segmentation on stereo matching by sharing deep features (i.e., the Semantic-Guided Cascade structure), and explicitly modeling the intra-class disparity consistency (i.e., the Semantic Selective Refinement branch) and semantic consistency between views (i.e., the Left-Right Semantic Consistency).

## Method

In this section, we provide a comprehensive description of our SemStereo, as illustrated in Figure 3. It involves a Semantic-Guided Cascade (SGC) structure for the generation of an initial disparity, a Semantic Selective Refinement (SSR) branch that refines the disparity under the guidance of semantic segmentation maps, and the Left-Right Semantic Consistency (LRSC) constraints on both views.

### Semantic-Guided Cascade Structure

The Semantic-Guided Cascade structure (Figure 3 (a)) incorporates a U-shaped network for extracting deep shared features for both tasks, along with a Fast-ACV for the disparity computation. This design facilitates deeper feature sharing between the two tasks, thereby strengthening the implicit influence of semantic cues on stereo matching.

**Shared U-shape Feature Extractor.** The U-shaped feature extractor generates shared features for both semantic segmentation and disparity estimation, taking as input an image pair $I^l, I^r \in \mathbb{R}^{3\times H\times W}$. Given the high-resolution input of remote sensing images and the need for efficient extraction of robust features with global receptive fields, we employ MobileViTv2 (Mehta and Rastegari 2021, 2022) as the encoder. This is followed by a decoder with skip connections and transposed convolutions, which progressively derives features at multiple scales $D_i^l, D_i^r \in \mathbb{R}^{C_i'\times \frac{H}{i}\times \frac{W}{i}}$ ($i = 2, 4, 8, 16, 32$).

**Semantic Segmentation.** Attached to the deepest feature volume $D_2^l, D_2^r$ is a simple module for semantic segmentation. It involves a convolution, upsampling, and softmax layer, resulting in the pixel-wise segmentation heat map denoted as $P^l, P^r \in \mathbb{R}^{N\times H\times W}$ ($N$ for the number of semantic classes, the same below).

**Cost Volume for Disparity Computation.** Aiming at strengthening the semantic influence on stereo matching, we adopt a cascaded structure (Figure 1) that utilizes deep shared features $D_i^l$ and $D_i^r$, enriched with semantic information, for initial disparity map generation. This approach contrasts with earlier methods that relied on shallow features, which resulted in weaker connections between tasks (Yang et al. 2018; Zhang et al. 2019b; Dovesi et al. 2020). Specifically, the feature volumes $D_i^l, D_i^r$ are firstly fed into a series of $1\times1$ siamese convolutions to bridge the gap between tasks

and halve the number of channels to ensure an equitable comparison with the baseline (Xu et al. 2022, 2023b), yielding feature volumes $T_i^l, T_i^r \in \mathbb{R}^{C_i'' \times \frac{H}{i} \times \frac{W}{i}}$, where $C_i'' = \frac{C_i'}{2}$. Then we construct the attention concatenation cost volume $V \in \mathbb{R}^{C''' \times \frac{D_{max}}{2} \times \frac{H}{4} \times \frac{W}{4}}$ as Fast-ACV (Xu et al. 2022, 2023b) with a minor adaptation for remote sensing that the disparity spans from negative ($-D_{max}$) to positive values ($D_{max} - 1$). The initial disparity map $d_{init} \in \mathbb{R}^{1 \times \frac{H}{4} \times \frac{W}{4}}$ is derived after regulating the cost volume $V$ as Fast-ACV (Xu et al. 2022, 2023b).

**Feature volume $F$ for SSR.** In addition to the segmentation map and the initial disparity, the third output of SGC is a feature volume containing both semantic and disparity information, as the disparity flow and semantic flow depicted in Figure 3 (a), which is then fed to SSR for further disparity refinement. It generates a comprehensive feature volume $F \in \mathbb{R}^{N \times H \times W}$ by progressively upsampling and concatenating the features $T_i$.

## Semantic Selective Refinement Branch

To explicitly leverage intra-class disparity consistency in remote sensing scenes (Figure 2), we introduce the Semantic-Selective Refinement (SSR) branch. This branch uses a channel attention mechanism to selectively learn disparity residual errors from the semantic prediction, which are then applied to refine the initial disparity map.

As shown in Figure 3 (b), we first compute the inner production of the feature volume $F$ and the predicted semantic map $P^l \in \mathbb{R}^{N \times H \times W}$. Notably, we preserve the multi-channel outputs from the semantic segmentation, where each channel represents the probability map for a specific class, thereby maintaining the confidence level of each pixel belonging to a certain class. Next, we compute their correlation score using a $1 \times 1$ point convolution followed by Batch Normalization and a Sigmoid activation function to generate a weight map $\sigma(F \cdot P^l)$. This weight map is then applied to filter the feature volume $F$, yielding a new feature volume $F' \in \mathbb{R}^{N \times H \times W}$ as follows:

$$F' = \sigma(F \cdot P^l) \cdot F, \tag{1}$$

where $\sigma$ denotes $1 \times 1$ CNN followed by Batch Normalization and Sigmoid.

Furthermore, we apply the new feature volume $F'$ to generate a new weight map $\sigma(F')$ to filter the initial disparity map. Specifically, we perform bilinear upsampling on $d_{init}$ to obtain $d'_{init} \in \mathbb{R}^{1 \times H \times W}$. Subsequently, we normalize $d'_{init}$ and use a $3 \times 3$ convolution to expand the channels to the number of classes, denoted as $d''_{init} \in \mathbb{R}^{N \times H \times W}$. We multiply $d''_{init}$ with the new weight map $\sigma(F')$. Finally, we utilize a $1 \times 1$ CNN to obtain single channel residuals $R$ and refine the final disparity map $d_{final}$ by adding with $d''_{init}$:

$$R = Conv(\sigma(F') \cdot d''_{init}), \tag{2}$$

$$d_{final} = R + d'_{init}. \tag{3}$$

## Left-Right Semantic Consistency Supervision

We further impose a semantic consistency constraint between both views after warping the semantic segmentation map (or feature map in the absence of pixel-wise annotations) of the left view to the right view based on the refined disparity map.

Disparity measures the horizontal pixel difference between corresponding points in left ($p_l$) and right ($p_r$) images:

$$disparity = x_l - x_r, \tag{4}$$

where $x_l$ and $x_r$ are the horizontal coordinates of $p_l$ and $p_r$. Therefore, we leverage this relationship by warping the semantic labels of the left view $GT^l$ to the right using the corresponding disparity, thus obtaining pseudo-semantic labels $GT^r$ for the right semantic supervision (See Figure 3 (c)). Additionally, given the high cost of semantic annotations, we account for scenarios where such annotations may be unavailable. In these cases, simply replacing $GT^l$ with $P^l$ enables our method to operate in a self-supervised manner.

To achieve that goal, we minimize the discrepancy between the semantic maps after warping using the Cross-Entropy (CE) loss as an auxiliary loss $\mathcal{L}_{LRSC}$:

$$R^{gt} = \begin{cases} \text{warp}(GT^l, d_{final}), & \text{if } GT^l \text{ is available} \\ \text{warp}(P^l, d_{final}), & \text{if } GT^l \text{ is not available} \end{cases} \tag{5}$$

$$\mathcal{L}_{LRSC} = \mathcal{L}_{CE}(P^r, GT^r). \tag{6}$$

## Loss Function

For the semantic segmentation task, we combine Dice loss and CE loss to leverage the strengths of both:

$$\mathcal{L}_{Seg} = \mathcal{L}_{CE}(L^p, L^{gt}) + \mathcal{L}_{Dice}(L^p, L^{gt}). \tag{7}$$

For the stereo matching task, Smooth $L_1$ loss is used:

$$\mathcal{L}_{Disp} = \sum_{i}^{n} \lambda_i \mathcal{S}mooth_{L1}(d_i - d^{gt}), \tag{8}$$

where $d_i$ are the predicted disparity map of different stages, and $d^{gt}$ is the disparity ground truth, $\lambda_i$ are hyperparameters that control the relative weights of the different stages.

For the entire multi-task model, our joint loss is given by,

$$\mathcal{L} = \mathcal{L}_{Disp} + \alpha \mathcal{L}_{Seg} + \beta \mathcal{L}_{LRSC}, \tag{9}$$

where $\alpha$ and $\beta$ are hyperparameters to control the relative weights of losses.

# Experiments

## Datasets

**US3D** (Bosch et al. 2019) contains 2,139 pairs of satellite stereo images from Jacksonville and 2,153 from Omaha, each with corresponding semantic labels. We randomly select 1,500 pairs from Jacksonville for training, 139 for validation, and 500 for testing, and use Omaha for generalization verification.

**WHU** (Liu and Ji 2020) is an aerial dataset from 8,316 real aerial images in the training set and 2,618 in the test set, covering an area of $6.7 \times 2.2\ km^2$ over Meitan County, China, with a ground resolution of approximately 0.1 meters. However, it lacks corresponding semantic labels.

| | Model | SGC | SSR | LRSC | Stereo Matching | | Semantic | |
|---|---|---|---|---|---|---|---|---|
| | | | | | EPE(Pixel)↓ | D1(%)↓ | mIOU(%)↑ | PA(%)↑ |
| 1 | Baseline | | | | 1.2087 | 7.28 | 75.84 | 93.65 |
| 2 | SGC-Net | ✓ | | | 0.9995 | 4.98 | 75.74 | 93.70 |
| 3 | SGC-SSR-Net | ✓ | ✓ | | 0.9702 | 4.76 | 76.85 | 93.83 |
| 4 | SemStereo | ✓ | ✓ | ✓ | **0.9582** | **4.58** | **77.02** | **94.13** |
| 5 | Baseline* | | | | 1.2260 | 7.47 | - | - |
| 6 | SGC-Net* | ✓ | | | 1.0499 | 5.61 | - | - |
| 7 | SGC-SSR-Net* | ✓ | ✓ | | 1.0164 | 5.34 | - | - |
| 8 | SemStereo* | ✓ | ✓ | ✓ | **0.9956** | **5.00** | - | - |

Table 1: Ablation quantitative evaluation of Semantic Guided-Cascade (SGC) framework, Semantic Selective Refinement (SSR) branch, and Left-Right Semantic Consistency (LRSC) supervision on US3D Jacksonville test set. * : Without explicit semantic label supervision. **Bold**: Best.

| Method | US3D | | WHU | |
|---|---|---|---|---|
| | EPE(Pixel)↓ | D1(%)↓ | EPE(Pixel)↓ | D1(%)↓ |
| StereoNet (Khamis et al. 2018) | 1.6053 | 12.13 | 0.3881 | 1.413 |
| $S^2$Net†(Liao et al. 2023) | 1.439 | 10.05 | - | - |
| $S^3$Net†(Yang et al. 2024) | 1.403 | 9.58 | - | - |
| DSMNet†(He et al. 2021) | 1.2776 | 7.94 | - | - |
| HMSMNet (He et al. 2022) | 1.2338 | 7.91 | 0.2745 | 0.904 |
| GwcNet (Guo et al. 2019) | 1.2120 | 6.99 | 0.2549 | 0.862 |
| PSMNet (Chang and Chen 2018) | 1.1770 | 6.87 | 0.2432 | 0.814 |
| IGEV-Stereo (Xu et al. 2023a) | 1.2051 | 7.32 | - | - |
| ACVNet (Xu et al. 2022, 2023b) | 1.2836 | 7.73 | 0.2422 | 0.737 |
| Fast-ACVNet (Xu et al. 2022, 2023b) | 1.1706 | 7.06 | 0.2257 | 0.740 |
| SemStereo* (ours) | 0.9956 | 5.00 | **0.2236** | **0.731** |
| SemStereo (ours) | **0.9582** | **4.58** | - | - |

Table 2: Quantitative comparison of stereo matching between our SemStereo and state-of-the-art models on US3D and WHU test set. †: The results are obtained from the official declaration. * : Without explicit semantic label supervision. **Bold**: Best.

| Method | Zero-shot | | 50 Pairs Fine-tuning | | 500 Pairs Fine-tuning | |
|---|---|---|---|---|---|---|
| | EPE(Pixel)↓ | D1(%)↓ | EPE(Pixel)↓ | D1(%)↓ | EPE(Pixel)↓ | D1(%)↓ |
| StereoNet (Khamis et al. 2018) | 1.6719 | 11.76 | 1.6599 | 11.53 | 1.4890 | 9.58 |
| PSMNet (Chang and Chen 2018) | 1.5163 | **9.27** | 1.3746 | 7.33 | 1.2135 | 5.53 |
| GwcNet (Guo et al. 2019) | 1.5342 | 9.32 | 1.3494 | 7.15 | 1.2467 | 5.85 |
| HMSMNet (He et al. 2022) | 1.5271 | 9.41 | 1.3293 | 6.95 | 1.1279 | 5.04 |
| IGEV-Stereo (Xu et al. 2023a) | 1.5120 | 9.91 | 1.3659 | 8.71 | 1.2004 | 5.71 |
| ACVNet (Xu et al. 2022, 2023b) | 1.5647 | 9.36 | 1.4070 | 7.55 | 1.3500 | 6.77 |
| Fast-ACVNet (Xu et al. 2022, 2023b) | 1.6132 | 11.13 | 1.4134 | 8.37 | 1.1753 | 5.78 |
| SemStereo (ours) | **1.4996** | 9.70 | **1.3206** | **6.79** | **1.1002** | **4.54** |

Table 3: Quantitative comparison of generalization performance across cities on the US3D test set. **Bold**: Best.

## Implementation Details

We implement SemStereo using PyTorch and conduct our experiments on two NVIDIA A40 GPUs. The hyperparameters for the loss function are set as follows: $\lambda_0 = 1$, $\lambda_1 = 0.6$, $\lambda_2 = 0.5$, $\lambda_3 = 0.3$, and $\alpha, \beta = 1$. We compare our approach with a range of state-of-the-art methods from both the computer vision and remote sensing communities. For fairness, we standardize the configuration parameters: the optimizer is set to Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, the batch size is 4, and we use the original resolution without any augmentation techniques. We train each stage for 48 epochs, starting with an initial learning rate ($lr_0$) of 0.001, which decays by half after epochs 12, 22, 30, 38, and 44. The disparity range varies by dataset: US3D is set to $[-64, 64)$ and WHU is set to $[0, 128)$, following the settings used in previous work (He et al. 2021).

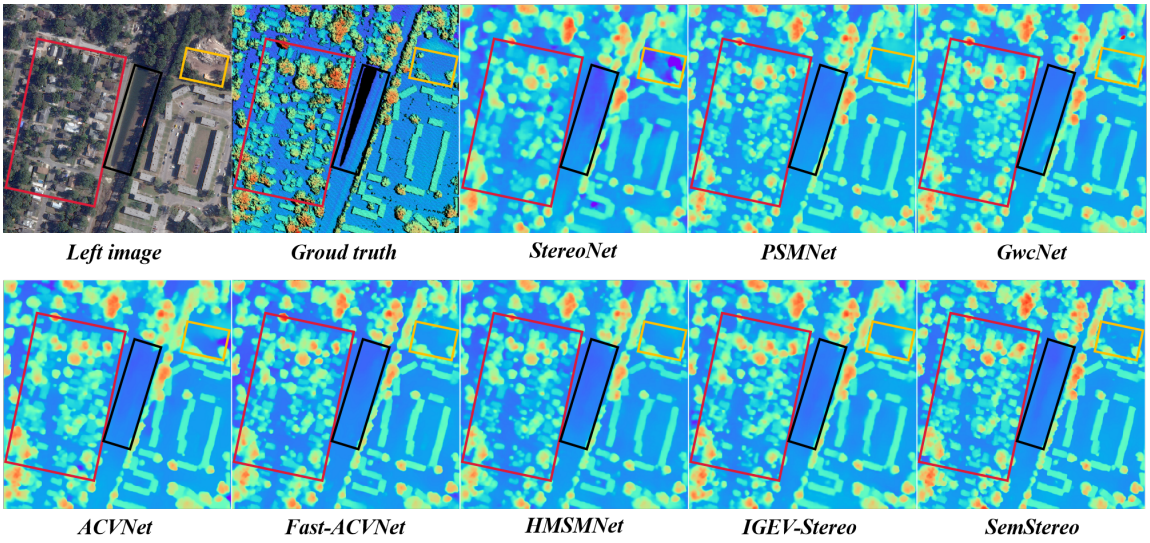


Figure 4: Qualitative comparison of results with other state-of-the-art models on the US3D test set. Red box area: Our SemStereo achieves clearer boundaries in dense buildings; Black box area: Our model has clearer boundaries even for areas without disparity labels; Yellow box area: The prediction of our model preserves more details.

| Method | PA(%)↑ | mIOU(%)↑ | IOU Per Class(%)↑ | | | | |
|---|---|---|---|---|---|---|---|
| | | | Ground | Trees | Building Roof | Water | Bridge/ Elevated Road |
| FCN-8s (Long, Shelhamer, and Darrell 2015) | 88.32 | 58.89 | 86.87 | 57.62 | 75.65 | 46.59 | 27.73 |
| UNet (Ronneberger, Fischer, and Brox 2015) | 90.76 | 65.98 | 86.11 | 64.04 | 81.93 | 57.53 | 40.27 |
| DeepLabV3 (Chen et al. 2017) | 92.00 | 66.53 | 87.62 | 67.89 | 84.75 | 50.20 | 42.19 |
| PSPNet (Zhao et al. 2017) | 91.33 | 67.06 | 86.74 | 65.12 | 82.83 | 52.80 | 47.83 |
| SegFormer (Xie et al. 2021) | 90.45 | 63.60 | 85.57 | 63.37 | 81.08 | 49.35 | 38.65 |
| $S^2$Net†(Liao et al. 2023) | - | 69.10 | 83.67 | 66.82 | 79.92 | 80.38 | 34.68 |
| $S^3$Net†(Yang et al. 2024) | - | 67.39 | 81.94 | 66.39 | 73.45 | 79.23 | 35.96 |
| SemStereo* (ours) | 92.99 | 67.57 | 89.08 | 70.60 | 87.42 | 48.37 | 42.38 |
| SemStereo (ours) | **94.13** | **77.02** | **90.84** | **74.63** | **88.30** | **68.94** | **62.37** |

Table 4: Quantitative comparison of semantic accuracy between our SemStereo and state-of-the-art models on the US3D test set. †: The results are obtained from the official declaration. * : Without stereo matching supervision. **Bold**: Best.

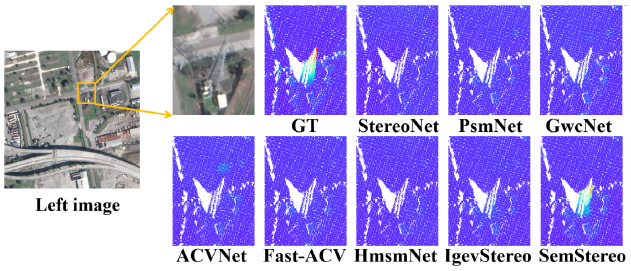


Figure 5: Qualitative comparison of results for local details with other state-of-the-art models on the US3D test set. Only our SemStereo can effectively estimate the disparities of the signal tower and its surrounding small-scale object.

## Ablation Study

We conduct ablation experiments to assess the effectiveness of our proposed Semantic-Guided Cascade (SGC) structure by comparing it with standard parallelized frameworks (Zhang et al. 2019b; Dovesi et al. 2020; Liao et al. 2023) (Figure 1), the Semantic Selective Refinement (SSR) by replacing it with a common bilinear upsampling method (Chang and Chen 2018; Guo et al. 2019), and the Left-Right Semantic Consistency (LRSC) by removing it.

As shown in Table 1, the SGC structure improves the baseline method by 31.6% in the D1 metric and 17.3% in the EPE metric (Line 2 vs. Line 1), demonstrating its effectiveness. Comparing Line 2 with Line 6, we observe that incorporating semantic supervision results in an 11.2% improvement in D1 and a 4.8% improvement in EPE, underscoring the critical role of semantic information in enhancing stereo matching. The introduction of SSR further enhances the D1 metric by 4.4% and the EPE by 2.9% (Line 2 vs. Line 3).

The introduction of LRSC improves the D1 metric by 3.8% and the EPE metric by 1.2% (Line 3 vs. Line 4). In the absence of explicit semantic annotations, LRSC still en-

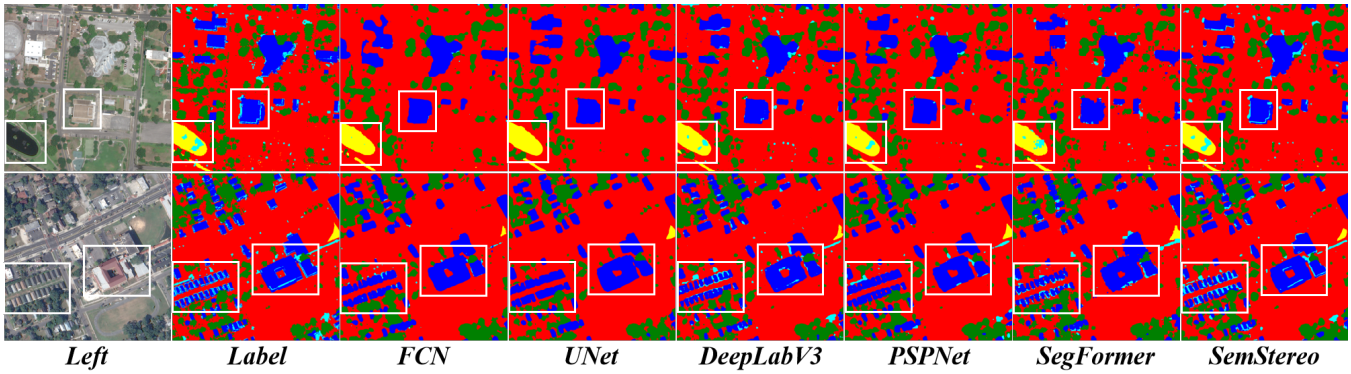


Figure 6: Qualitative comparison of semantic results with other classic state-of-the-art models on the US3D test set. Our SemStereo delivers clearer boundaries both in dense building clusters and on large objects, while also preserving more details.

hances D1 by 6.4% and EPE by 2.0% (Line 7 vs. Line 8). Additionally, incorporating semantic supervision results in an 8.4% improvement in D1 and a 3.7% improvement in EPE (Line 4 vs. Line 8). This demonstrates that LRSC effectively enhances performance by enforcing semantic consistency across views, both with and without explicit semantic supervision, and achieves even better results when semantic annotations are provided.

## Comparisons with State-of-the-art

**Stereo Matching** We compare our SemStereo with the state-of-the-art stereo methods on US3D (Le Saux et al. 2019a,b; Bosch et al. 2019) and WHU (Liu and Ji 2020), as presented in Table 2.

When semantic labels are unavailable, our downgraded model, SemStereo*, still surpasses state-of-the-art methods on both US3D and WHU datasets. With the introduction of semantic labels during training, our full model, SemStereo, achieves an additional 14.6% improvement in the D1 metric, representing a more substantial enhancement compared to previous methods. SemStereo outperforms IGEV-Stereo (Xu et al. 2023a) by 37.7%, Fast-ACVNet (Xu et al. 2022, 2023b) by 35.4%, HMSMNet (He et al. 2022) by 42.35%, GwcNet (Guo et al. 2019) by 34.76%, and PSMNet (Chang and Chen 2018) by 33.6% on the D1 metric.

To evaluate generalization across cities, we assess zero-shot capabilities and conduct transfer learning with limited data from Omaha. We randomly select 50 and 500 pairs from Omaha for fine-tuning over 12 and 48 epochs, respectively, using 1500 pairs for validation. As shown in Table 3, our model achieves state-of-the-art performance with an EPE of 1.4996 in zero-shot scenarios. As the amount of fine-tuning data increases, our model's performance improves. When fine-tuning is performed with 500 pairs, our model's performance on Omaha closely matches the results from the original city (Jacksonville), with D1 scores of 4.54% vs. 4.58%. Additionally, our model demonstrates a clear advantage over other models, improving D1 by 9.9% and EPE by 2.5%.

A qualitative comparison of stereo matching results on the US3D test set is shown in Figure 4 and Figure 5. SemStereo exhibits clearer boundaries and more detailed features, owing to enhanced feature interactions between tasks and the incorporation of explicit inter-task constraints.

**Semantic Segmentation** As shown in Table 4, we compare our model with established semantic segmentation networks, including FCN (Long, Shelhamer, and Darrell 2015), UNet (Ronneberger, Fischer, and Brox 2015), DeepLabV3 (Chen et al. 2017), PSPNet (Zhao et al. 2017), and SegFormer (Xie et al. 2021). SemStereo achieves state-of-the-art results, outperforming SegFormer by 14.45% in mIoU and PSPNet by 9.96% in mIoU. Additionally, our model shows significant advantages over other semantic stereo methods, such as $S^2$Net and $S^3$Net (Liao et al. 2023; Yang et al. 2024). The inclusion of stereo matching supervision enhances semantic accuracy by 9.45% mIoU in the full SemStereo. Qualitative results in Figure 6 demonstrate that SemStereo delivers clearer boundaries on dense building clusters and large objects while preserving more details, attributable to the enhanced disparity label supervision.

## Conclusion

In this work, we advocate for a deeper exploration of the complementary relationship between semantic segmentation and stereo matching in remote sensing scenes. We introduce an implicit method, SGC, which enhances feature sharing between these tasks. Additionally, we propose two explicit constraints: Semantic-Selective Refinement (SSR) to leverage disparity consistency within the same semantic category, and Left-Right Semantic Consistency (LRSC) to ensure semantic consistency across views. Our experiments on the US3D and WHU datasets demonstrate the state-of-the-art performance of SemStereo. Ablation studies validate the effectiveness of the SGC, SSR, and LRSC modules, highlighting the mutual benefits of semantic and disparity information. We also find that semantic instances exhibit a closer relationship with disparities compared to semantic categories, and this relationship can be extended to more general scenarios. Future research will focus on explicitly modeling these.

## Acknowledgments

This work was supported by the National Nature Science Foundation of China under Grant 62331027, and supported by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDA0360303.

## References

Blaha, M.; Vogel, C.; Richard, A.; Wegner, J. D.; Pock, T.; and Schindler, K. 2016. Large-scale semantic 3d reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3176–3184.

Bosch, M.; Foster, K.; Christie, G.; Wang, S.; Hager, G. D.; and Brown, M. 2019. Semantic Stereo for Incidental Satellite Images. In *The IEEE Winter Conference on Applications of Computer Vision*.

Chang, J.-R.; and Chen, Y.-S. 2018. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5410–5418.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Cheng, J.; Tsai, Y.-H.; Wang, S.; and Yang, M.-H. 2017. Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*, 686–695.

Dovesi, P. L.; Poggi, M.; Andraghetti, L.; Martí, M.; Kjellström, H.; Pieropan, A.; and Mattoccia, S. 2020. Real-time semantic stereo matching. In *2020 IEEE international conference on robotics and automation (ICRA)*, 10780–10787. IEEE.

Guo, X.; Yang, K.; Yang, W.; Wang, X.; and Li, H. 2019. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3273–3282.

Hane, C.; Zach, C.; Cohen, A.; Angst, R.; and Pollefeys, M. 2013. Joint 3D scene reconstruction and class segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 97–104.

He, S.; Li, S.; Jiang, S.; and Jiang, W. 2022. HMSM-Net: Hierarchical multi-scale matching network for disparity estimation of high-resolution satellite stereo images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188: 314–330.

He, S.; Zhou, R.; Li, S.; Jiang, S.; and Jiang, W. 2021. Disparity estimation of high-resolution remote sensing images with dual-scale matching network. *Remote Sensing*, 13(24): 5050.

Jing, H.; Wang, Z.; Sun, X.; Xiao, D.; and Fu, K. 2021. PSRN: Polarimetric space reconstruction network for PolSAR image semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 10716–10732.

Kadhim, N.; Mourshed, M.; and Bray, M. 2016. Advances in remote sensing applications for urban sustainability. *Euro-Mediterranean Journal for Environmental Integration*, 1(1): 7.

Kang, J.; Wang, Z.; Zhu, R.; Sun, X.; Fernandez-Beltran, R.; and Plaza, A. 2021. PiCoCo: Pixelwise contrast and consistency learning for semisupervised building footprint segmentation. *IEEE journal of selected topics in applied earth observations and remote sensing*, 14: 10548–10559.

Kang, J.; Wang, Z.; Zhu, R.; Xia, J.; Sun, X.; Fernandez-Beltran, R.; and Plaza, A. 2022. DisOptNet: Distilling semantic knowledge from optical images for weather-independent building segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7482–7491.

Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; and Bry, A. 2017. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, 66–75.

Khamis, S.; Fanello, S.; Rhemann, C.; Kowdle, A.; Valentin, J.; and Izadi, S. 2018. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 573–590.

Kunwar, S.; Chen, H.; Lin, M.; Zhang, H.; D'Angelo, P.; Cerra, D.; Azimi, S. M.; Brown, M.; Hager, G.; Yokoya, N.; et al. 2020. Large-scale semantic 3-D reconstruction: Outcome of the 2019 IEEE GRSS data fusion contest—Part A. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 922–935.

Le Saux, B.; Yokoya, N.; Hänsch, R.; and Brown, M. 2019a. 2019 ieee grss data fusion contest: large-scale semantic 3d reconstruction. *IEEE Geoscience and Remote Sensing Magazine (GRSM)*, 7(4): 33–36.

Le Saux, B.; Yokoya, N.; Hansch, R.; Brown, M.; and Hager, G. 2019b. 2019 data fusion contest [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 7(1): 103–105.

Liao, P.; Zhang, X.; Chen, G.; Wang, T.; Li, X.; Yang, H.; Zhou, W.; He, C.; and Wang, Q. 2023. S 2 Net: A Multitask Learning Network for Semantic Stereo of Satellite Image Pairs. *IEEE Transactions on Geoscience and Remote Sensing*.

Liu, J.; and Ji, S. 2020. A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6050–6059.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Luo, W.; Schwing, A. G.; and Urtasun, R. 2016. Efficient deep learning for stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5695–5703.

Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; and Brox, T. 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4040–4048.

Mehta, S.; and Rastegari, M. 2021. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*.

Mehta, S.; and Rastegari, M. 2022. Separable self-attention for mobile vision transformers. *arXiv preprint arXiv:2206.02680*.

Qin, R.; Huang, X.; Liu, W.; and Xiao, C. 2019. Pairwise stereo image disparity and semantics estimation with the combination of u-net and pyramid stereo matching network. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, 4971–4974. IEEE.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Shen, Z.; Dai, Y.; and Rao, Z. 2021. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13906–13915.

Song, X.; Zhao, X.; Fang, L.; Hu, H.; and Yu, Y. 2020. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. *International Journal of Computer Vision*, 128: 910–930.

Sun, X.; Huang, X.; Mao, Y.; Sheng, T.; Li, J.; Wang, Z.; Lu, X.; Ma, X.; Tang, D.; and Chen, K. 2024. GABLE: A first fine-grained 3D building model of China on a national scale from very high resolution satellite imagery. *Remote Sensing of Environment*, 305: 114057.

Wu, Z.; Wu, X.; Zhang, X.; Wang, S.; and Ju, L. 2019. Semantic stereo matching with pyramid cost volumes. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7484–7493.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Xu, G.; Cheng, J.; Guo, P.; and Yang, X. 2022. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12981–12990.

Xu, G.; Wang, X.; Ding, X.; and Yang, X. 2023a. Iterative Geometry Encoding Volume for Stereo Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21919–21928.

Xu, G.; Wang, Y.; Cheng, J.; Tang, J.; and Yang, X. 2023b. Accurate and efficient stereo matching via attention concatenation volume. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, G.; Zhao, H.; Shi, J.; Deng, Z.; and Jia, J. 2018. Segstereo: Exploiting semantic information for disparity estimation. In *Proceedings of the European conference on computer vision (ECCV)*, 636–651.

Yang, Q.; Chen, G.; Tan, X.; Wang, T.; Wang, J.; and Zhang, X. 2024. S3Net: Innovating Stereo Matching and Semantic Segmentation with a Single-Branch Semantic Stereo Network in Satellite Epipolar Imagery. *arXiv preprint arXiv:2401.01643*.

Zbontar, J.; and LeCun, Y. 2015. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1592–1599.

Zhang, F.; Prisacariu, V.; Yang, R.; and Torr, P. H. 2019a. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 185–194.

Zhang, J.; Skinner, K. A.; Vasudevan, R.; and Johnson-Roberson, M. 2019b. Dispsegnet: Leveraging semantics for end-to-end learning of disparity estimation from stereo imagery. *IEEE Robotics and Automation Letters*, 4(2): 1162–1169.

Zhang, Y.; Chen, Y.; Bai, X.; Yu, S.; Yu, K.; Li, Z.; and Yang, K. 2020. Adaptive unimodal cost volume filtering for deep stereo matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12926–12934.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.