Computers, Materials &
Continua

Tech Science Press

Doi:10.32604/cmc.2025.073846

ARTICLE

Semantic-Guided Stereo Matching Network Based on Parallax Attention Mechanism and SegFormer

Zeyuan Chen, Yafei Xie, Jinkun Li, Song Wang and Yingqiang Ding*

School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou, 450001, China

*Corresponding Author: Yingqiang Ding. Email: dyq@zzu.edu.cn

Received: 27 September 2025; Accepted: 25 November 2025; Published: 10 February 2026

ABSTRACT: Stereo matching is a pivotal task in computer vision, enabling precise depth estimation from stereo image pairs, yet it encounters challenges in regions with reflections, repetitive textures, or fine structures. In this paper, we propose a Semantic-Guided Parallax Attention Stereo Matching Network (SGPASMnet) that can be trained in unsupervised manner, building upon the Parallax Attention Stereo Matching Network (PASMnet). Our approach leverages unsupervised learning to address the scarcity of ground truth disparity in stereo matching datasets, facilitating robust training across diverse scene-specific datasets and enhancing generalization. SGPASMnet incorporates two novel components: a Cross-Scale Feature Interaction (CSFI) block and semantic feature augmentation using a pre-trained semantic segmentation model, SegFormer, seamlessly embedded into the parallax attention mechanism. The CSFI block enables effective fusion of multi-scale features, integrating coarse and fine details to enhance disparity estimation accuracy. Semantic features, extracted by SegFormer, enrich the parallax attention mechanism by providing high-level scene context, significantly improving performance in ambiguous regions. Our model unifies these enhancements within a cohesive architecture, comprising semantic feature extraction, an hourglass network, a semantic-guided cascaded parallax attention module, output module, and a disparity refinement network. Evaluations on the KITTI2015 dataset demonstrate that our unsupervised method achieves a lower error rate compared to the original PASMnet, highlighting the effectiveness of our enhancements in handling complex scenes. By harnessing unsupervised learning without ground truth disparity needed, SGPASMnet offers a scalable and robust solution for accurate stereo matching, with superior generalization across varied real-world applications.

KEYWORDS: Stereo matching; parallax attention; unsupervised learning; convolutional neural network; stereo correspondence

1 Introduction

Stereo matching, the process of estimating depth from a pair of rectified stereo images, is a cornerstone of computer vision with applications in autonomous driving, robotics, and 3D reconstruction. The task involves computing a disparity map that represents the pixel-wise horizontal displacement between corresponding points in the left and right images. Despite significant progress, stereo matching remains challenging in regions with occlusions, reflections, repetitive textures, or low-contrast areas, where traditional feature matching often fails to establish accurate correspondences. Moreover, supervised learning approaches typically rely on large-scale datasets with ground truth disparity, which are often difficult and costly to acquire for stereo image pairs, particularly across diverse real-world scenarios. This limitation hinders the generalizability of such models to varied scenes.

Recent advancements in deep learning have significantly improved stereo matching performance, with unsupervised methods like the Parallax Attention Stereo Matching Network [1] (PASMnet) introducing attention-based mechanisms to capture global correspondences along epipolar lines. However, PASMnet struggles in complex scenes due to limitations in leveraging multi-scale feature interactions and high-level semantic context, leading to wrong match in challenging areas and across different objects.

To address these challenges, we proposed a novel unsupervised learning network for stereo matching, Semantic-Guided Parallax Attention Stereo Matching Network (SGPASMnet), building upon PASMnet with two key enhancements: a Cross-Scale Feature Interaction (CSFI) block and the integration of semantic features extracted from a pre-trained SegFormer [2] model. By adopting an unsupervised learning approach, our model eliminates the dependency on annotated disparity data, leveraging self-supervised signals such as photometric consistency and geometric constraints. This enables robust training on diverse datasets without ground truth labels, facilitating better generalization across different scene types and conditions, such as varying lighting, occlusions, or texture complexities.

The CSFI block enables the fusion of features across different scales, combining coarse, high-level information with fine, detailed features to enhance both global consistency and local accuracy in disparity estimation. This approach draws inspiration from feature pyramid networks [3] and deformable convolutions [4], adapting them to the stereo matching context. The semantic feature augmentation leverages SegFormer's transformer-based architecture to extract multi-scale semantic representations, which are integrated into the parallax attention mechanism to provide contextual guidance, particularly in ambiguous regions like reflections or repetitive patterns. This is motivated by prior work such as SegStereo [5], which demonstrated the value of semantic information in disparity estimation.

Our enhanced model integrates these components into a cohesive unsupervised architecture, comprising semantic feature extraction, an Hourglass network for multi-scale feature extraction, a semantic-guided Cascaded Parallax Attention Module (CPAM) for disparity computation, an output module to generate initial disparity map and valid masks, and a refinement network to optimize the disparity map. The model is trained with a combination of unsupervised losses, including photometric loss, smoothness loss, and parallax attention loss and semantic consistency loss, to ensure robust learning without reliance on ground truth disparity. Evaluations on the KITTI2015 dataset suggest that our SGPASMnet achieves a lower error rate compared to the baseline model, demonstrating improved performance and generalization in challenging scenarios.

2 Related Work

Stereo matching has been a fundamental problem in computer vision, with significant advancements driven by both traditional and deep learning-based approaches. Below, we review key developments in traditional and deep learning-based stereo matching methods, with the latter further categorized into supervised and unsupervised learning approaches.

2.1 Traditional Stereo Matching Methods

Traditional stereo matching methods rely on hand-crafted features and optimization techniques to compute disparity maps. A prominent example is Semi-Global Matching (SGM), which incorporates global smoothness constraints through dynamic programming along multiple image paths, achieving robust results in structured environments while balancing computational efficiency and accuracy.

Despite their advantages, these methods often struggle in regions with occlusions, textureless areas, or illumination variations due to their dependence on low-level features. Recent advancements have focused on improving SGM's performance in real-world scenarios. For instance, collaborative SGM [6] introduces local edge-aware filtering to strengthen interactions between neighboring scanlines, significantly reducing streak artifacts in disparity maps.

Other recent non-deep learning approaches include As-Global-As-Possible (AGAP) stereo matching with sparse depth measurement fusion [7], which combines global optimization with sparse priors to improve accuracy in sparse-data environments like satellite imagery. Furthermore, adaptations for specific domains, like mineral image matching with improved Birchfield-Tomasi-Census algorithms [8], enhance discrimination in textured regions without relying on learning-based features.

While these enhancements mitigate some limitations through better cost aggregation, edge preservation, traditional methods remain constrained compared to data-driven approaches, particularly in handling complex, unstructured scenes.

2.2 Deep Learning-Based Stereo Matching Methods

Deep learning has revolutionized stereo matching by leveraging convolutional neural networks (CNNs) and attention mechanisms to learn robust feature representations. These methods can be broadly divided into supervised and unsupervised learning approaches.

2.2.1 Supervised Learning

Supervised deep learning methods for stereo matching typically construct a cost volume from learned features and optimize it to produce disparity maps. DispNet [9] introduced an end-to-end CNN architecture that directly regresses disparity from stereo image pairs, achieving significant improvements over traditional methods. GC-Net [10] proposed a 3D cost volume constructed from concatenated left and right image features, processed by 3D convolutions to aggregate contextual information. PSMNet [11] further advanced this by incorporating a spatial pyramid pooling module to capture multi-scale context, improving performance in complex scenes.

Attention-based methods have recently gained prominence due to their ability to model global correspondences. AANet [12] combined attention with adaptive aggregation to achieve real-time performance, while GANet [13] integrated guided aggregation to refine cost volumes. Semantic information has also been explored to enhance supervised methods. SegStereo [5] incorporated semantic segmentation masks to guide disparity estimation, improving accuracy in object boundaries. Similarly, reference [14] proposed a joint semantic-stereo framework for real-time applications, demonstrating the value of semantic context. Recent supervised methods have focused on improving efficiency and robustness through cascaded architectures and adaptive correlations. For instance, CREStereo [15] introduces a cascaded recurrent network with adaptive correlation for practical high-resolution stereo matching, achieving state-of-the-art accuracy in real-world scenarios. Similarly, CFNet [16] proposes a cascade and fused cost volume approach to enhance robust stereo matching under challenging conditions like occlusions and varying illuminations. Advancements in zero-shot learning have also emerged, such as Cascade Cost Volume [17], which enables high-resolution multi-view stereo matching without extensive fine-tuning, improving generalization across datasets.

Building on attention mechanisms, transformer-based methods have recently been explored for stereo matching, leveraging their ability to capture long-range dependencies and model sequential data effectively. For instance, STTR [18] introduces stereo depth estimation from a sequence-to-sequence perspective using transformers. It employs alternating self and cross-attention to perform dense pixel matching along epipolar lines, eliminating the need for a fixed disparity range, detecting occlusions with confidence estimates, and enforcing uniqueness constraints via optimal transport. Similarly, RAFT-Stereo [19] adapts the RAFT optical flow architecture for stereo, introducing multilevel recurrent field transforms with convolutional GRUs to propagate information across the image efficiently.

2.2.2 Unsupervised Learning

Unsupervised stereo matching methods leverage photometric consistency and geometric constraints to train models without ground truth disparity maps, making them suitable for scenarios with limited labeled data. Reference [20] proposed an unsupervised framework that minimizes a photometric loss based on image reconstruction, using left-right consistency to enforce disparity coherence. Reference [21] extended this with a left-right disparity consistency loss, improving robustness in textureless regions. Recent unsupervised methods, such as [22], incorporate domain adaptation to handle real-world data, while Reference [23] use multi-view consistency to enhance training.

Attention-based unsupervised methods have also emerged. Reference [1] extended PASMnet to an unsupervised setting by incorporating cycle consistency losses, achieving competitive performance without labeled data. Reference [24] proposed an unsupervised attention mechanism that leverages feature similarity to guide disparity estimation. While unsupervised methods reduce the dependency on labeled data, they often struggle with accuracy in complex scenes compared to supervised approaches. Recent unsupervised approaches have incorporated semantic attention mechanisms to address domain gaps and data scarcity. Stereo Anywhere [25] presents a robust zero-shot deep stereo matching framework that leverages monocular depth priors for accurate disparity estimation even in unseen environments. Additionally, efforts in open-world generation, like the method in [26], combine stereo image synthesis with unsupervised matching to enable training on diverse synthetic data without labels. Specialized applications, such as underwater scenes, have seen innovations like the semantic attention-based unsupervised stereo matching [27], which uses semantic guidance to improve performance in low-visibility conditions, aligning with the need for context-aware disparity estimation.

Our work builds upon PASMnet by integrating semantic features from a pre-trained SegFormer [2] model, which provides richer, transformer-based semantic representations compared to traditional CNN-based segmentation models. Additionally, our CSFI module enhances multi-scale feature fusion, drawing inspiration from feature pyramid networks [3] and deformable convolutions [4], to improve disparity estimation in challenging regions. Similar multi-scale refinement strategies have been successfully applied in related vision tasks, such as instance segmentation. For example, Mask-Refined R-CNN (MR R-CNN) [28] adjusts the stride of region of interest align and incorporates an FPN structure in the mask head to fuse global semantic information with local details, achieving superior boundary delineation in large objects. This approach aligns with our CSFI block's cross-scale interactions, which combine coarse and fine features to handle ambiguous regions like reflections or repetitive textures in stereo matching.

In summary, our proposed model combines the strengths of attention-based supervised stereo matching with advanced multi-scale feature fusion and semantic augmentation, offering a robust solution for accurate disparity estimation in complex scenes.

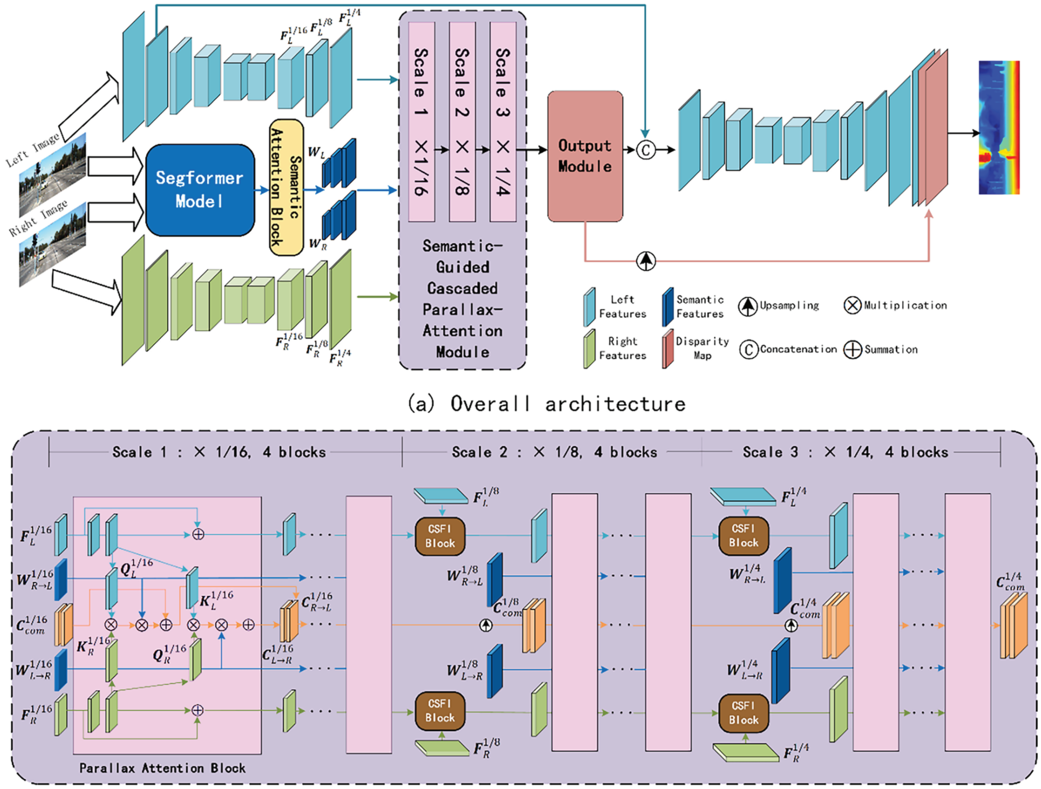
3 Proposed Method

In this section, we present our SGPASMnet that builds upon the Parallax-Attention Stereo Matching Network (PASMnet). Our enhancements incorporate a Cross-Scale Feature Interaction (CSFI) module to facilitate multi-scale feature fusion and the integration of semantic features extracted from a pre-trained SegFormer model to augment the parallax attention mechanism within the Parallax-Attention Block (PAB).

These modifications aim to address challenges in stereo matching, such as occlusions, textureless regions, and repetitive patterns, by leveraging multi-scale contextual information and semantic guidance. Experimental results on the KITTI2015 dataset demonstrate a reduction in error rates compared to the baseline PASMnet, validating the effectiveness of our approach.

3.1 Overall Architecture

Our model follows the general structure of PASMnet, designed to estimate disparity maps from a pair of rectified stereo images, denoted as $\boldsymbol{I}_L$ and $\boldsymbol{I}_R$, representing the left and right images, respectively. The disparity map $\boldsymbol{D}$ indicates the horizontal displacement between corresponding pixels in the stereo pair. The architecture comprises five key components: a semantic feature extraction module to extract multi-scale semantic features from the input images using a pre-trained SegFormer [2] model, providing contextual information to guide disparity estimation, an hourglass module to extract hierarchical feature representations at multiple scales from the input images, capturing both fine and coarse details, a semantic-guided Cascaded Parallax Attention Module (CPAM) to compute cost volumes across multiple scales using a parallax attention mechanism, enhanced with semantic features and Cross-Scale Feature Interactions (CSFI) blocks, an output module to process the cost volumes to produce initial disparity estimates, along with attention maps and validity masks during training, and a refinement module to refine the initial disparity estimates using additional feature information to produce the final disparity map. The architecture of our proposed SGPASMnet is shown in Fig. 1a. The structure of semantic-guided cascaded parallax attention module is shown in Fig. 1b.



(a) Overall architecture

(b) Semantic-guided cascaded parallax attention module

Figure 1: Overall architecture of our proposed SGPASMnet

3.2 Semantic Feature Extraction Module

To incorporate high-level semantic information, we utilize a pre-trained SegFormer [2] model, a transformer-based architecture designed for semantic segmentation, capable of generating rich, multi-scale semantic feature representations. These features provide critical scene context, enhancing disparity estimation in complex regions such as occlusions and textureless areas where traditional feature matching may fail.

The SegFormer model employs a hierarchical transformer structure that leverages efficient self-attention to produce feature maps at multiple resolutions. Its encoder outputs hidden states at various scales, making it well-suited for tasks requiring both global and local contextual understanding, such as stereo matching. We adopt a SegFormer model pre-trained on the Cityscapes dataset, with its parameters frozen to serve as a feature extractor, ensuring computational efficiency and robust feature quality.

The SegFormer model processes the input stereo images $\boldsymbol{I}_L, \boldsymbol{I}_R \in \boldsymbol{R}^{H\times W\times 3}$ independently, extracting hidden states from its transformer blocks at four different scales: $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ of the original image resolution. For our stereo matching task, we select the hidden states corresponding to scales $\frac{1}{4}$, $\frac{1}{8}$ and $\frac{1}{16}$, as these align with the feature map scales used in the hourglass and semantic-guided CPAM. These hidden states are derived from the intermediate layers of SegFormer's transformer encoder, specifically from the outputs of its hierarchical feature extraction stages, which progressively reduce spatial resolution while increasing channel depth to capture richer semantic representations.

Formally, the semantic feature maps are represented as:

$$\boldsymbol{S}_L = \left\{\boldsymbol{S}_L^{1/4}, \boldsymbol{S}_L^{1/8}, \boldsymbol{S}_L^{1/16}\right\} \tag{1}$$

$$\boldsymbol{S}_R = \left\{\boldsymbol{S}_R^{1/4}, \boldsymbol{S}_R^{1/8}, \boldsymbol{S}_R^{1/16}\right\} \tag{2}$$

where each $\boldsymbol{S}_L^s, \boldsymbol{S}_R^s \in \boldsymbol{R}^{C_s\times \frac{H}{s}\times \frac{W}{s}}$, and C_s denotes the number of channels at scale s in the semantic features (typically determined by the SegFormer configuration, e.g., 256 or 512 channels depending on the model variant). The hidden states are obtained from the transformer blocks in the SegFormer model in different scales containing semantic features. Fig. 2 shows visual examples of the SegFormer semantic segmentation results.

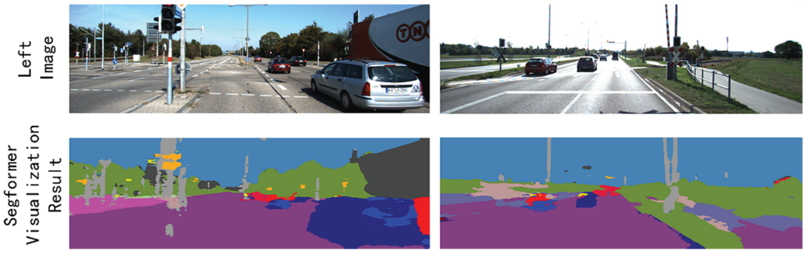


Figure 2: A visual example of SegFormer extracted semantic features

These multi-scale semantic features are fed into the semantic attention block to generate semantic attention weights, and further fed into corresponding scales of the semantic-guided CPAM to enhance the parallax attention mechanism within the Parallax Attention Blocks (PABs). Additionally, a single-scale semantic feature, derived from the final hidden state, is used to compute a semantic consistency loss, ensuring that semantic features from the left and right images align under disparity guidance. This approach draws inspiration from prior work, such as SegStereo [5], which uses semantic segmentation to guide disparity estimation, and Tonioni et al. [22], who proposed joint architectures for real-time semantic stereo matching.

However, our method distinguishes itself by directly embedding multi-scale semantic features into the parallax attention mechanism, offering finer-grained contextual guidance compared to traditional semantic mask-based or parallel processing approaches.

3.3 Hourglass Module

The hourglass module extracts hierarchical feature representations from the input stereo images, enabling the capture of both fine-grained details and coarse contextual information. It consists of a series of encoder and decoder blocks with skip connections, forming a U-shaped architecture inspired by spatial pyramid pooling in PSMNet [11]. This module serves as the foundation for subsequent processing by providing rich feature representations across multiple scales.

The encoder downsamples the input images through convolutional layers, producing feature maps at scales $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$. The decoder upsamples these features, combining them with skip connections from the encoder to preserve spatial details. The output feature maps are at scales $\frac{1}{16}$, $\frac{1}{8}$, and $\frac{1}{4}$, denoted as:

$$\boldsymbol{F}_L = \left\{\boldsymbol{F}_L^{1/16}, \boldsymbol{F}_L^{1/8}, \boldsymbol{F}_L^{1/4}\right\} \tag{3}$$

$$\boldsymbol{F}_R = \left\{\boldsymbol{F}_R^{1/16}, \boldsymbol{F}_R^{1/8}, \boldsymbol{F}_R^{1/4}\right\} \tag{4}$$

where $\boldsymbol{F}_L^S, \boldsymbol{F}_R^S \in \boldsymbol{R}^{C_s \times \frac{H}{s} \times \frac{W}{s}}$, with channel dimensions C_s set to 128, 96, and 64 for scales $\frac{1}{16}$, $\frac{1}{8}$, and $\frac{1}{4}$, respectively. An additional feature map at scale $\frac{1}{4}$ is used for disparity refinement. This hierarchical structure ensures that the model captures multi-scale information, which is critical for handling objects at varying depths.

Although the Hourglass module remains unmodified from the original PASMnet, its multi-scale feature outputs are critical for our enhancements. The features at $\frac{1}{16}$, $\frac{1}{8}$, and $\frac{1}{4}$ scales align precisely with the semantic features extracted by SegFormer, ensuring compatibility with the semantic-guided CPAM. The skip connections preserve high-resolution details, which are essential for the CSFI block's cross-scale fusion and the semantic-enhanced parallax attention mechanism.

3.4 Semantic-Guided Cascaded Parallax Attention Module

The semantic-guided Cascaded Parallax Attention Module (CPAM) computes multi-scale cost volumes using a parallax attention mechanism, improved by our two key enhancements: the Cross-Scale Feature Interactions (CSFI) block for cross-scale feature fusion and semantic feature integration for parallax attention modulation. The module operates sequentially at scales $\frac{1}{16}$, $\frac{1}{8}$, and $\frac{1}{4}$, processing feature maps and producing cost volumes that are refined across scales.

At each scale s, the PAB takes feature maps $\boldsymbol{F}_L^S, \boldsymbol{F}_R^S \in \boldsymbol{R}^{C_s \times \frac{H}{s} \times \frac{W}{s}}$ and semantic features $\boldsymbol{S}_L^s, \boldsymbol{S}_R^s \in \boldsymbol{R}^{C_s \times \frac{H}{s} \times \frac{W}{s}}$. The feature-based cost volume is computed using an attention mechanism:

Query and key features are generated via 1 × 1 convolutions:

$$\boldsymbol{Q} = Conv_{1\times1}\left(\boldsymbol{F}_L^s\right) \tag{5}$$

$$\boldsymbol{K} = Conv_{1\times1}\left(\boldsymbol{F}_R^s\right) \tag{6}$$

followed by permutation to $\boldsymbol{Q} \in \boldsymbol{R}^{B \times \frac{H}{s} \times \frac{W}{s} \times C}$ and $\boldsymbol{K} \in \boldsymbol{R}^{B \times \frac{H}{s} \times C \times \frac{W}{s}}$.

The feature cost is computed as:

$$\boldsymbol{C}_{fea} = \frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{C}} \tag{7}$$

where C is the number of the channels for normalization, yielding $\boldsymbol{C}_{fea} \in \boldsymbol{R}^{B\times\frac{H}{s}\times\frac{W}{s}\times\frac{W}{s}}$.

To incorporate semantic guidance, a semantic cost is computed within semantic attention block using the semantic features:

Semantic query and key are directly used:

$$\boldsymbol{Q}_{sem} = \boldsymbol{S}_L^s, \boldsymbol{K}_{sem} = \boldsymbol{S}_R^s \tag{8}$$

permuted to $\boldsymbol{Q}_{sem} \in \boldsymbol{R}^{B\times\frac{H}{s}\times\frac{W}{s}\times C_s}$, $\boldsymbol{K}_{sem} \in \boldsymbol{R}^{B\times\frac{H}{s}\times C_s\times\frac{W}{s}}$.

The semantic cost is:

$$\boldsymbol{C}_{sem} = \frac{\boldsymbol{Q}_{sem}\boldsymbol{K}_{sem}^T}{\sqrt{C_s}} \tag{9}$$

and a semantic weight is obtained via:

$$\boldsymbol{W}_{sem} = \sigma\left(\boldsymbol{C}_{sem}\right) \tag{10}$$

where σ is the sigmoid activation function, producing $\boldsymbol{W}_{sem} \in \boldsymbol{R}^{B\times\frac{H}{s}\times\frac{W}{s}\times\frac{W}{s}}$.

The combined cost volume is then:

$$\boldsymbol{C}_{com} = \boldsymbol{C}_{fea} \times \boldsymbol{W}_{sem} \tag{11}$$

which is added to the cost volume from the previous scale to produce the final cost volume for the current scale. This mechanism, inspired by SegStereo [5] and real-time semantic stereo matching [14], enhances the attention mechanism by prioritizing correspondences that are both visually and semantically consistent, improving robustness in ambiguous regions. Fig. 3 presents a flowchart of the operations in semantic attention block.

The CSFI block facilitates cross-scale feature fusion between consecutive scales (e.g., $\frac{1}{16}$ to $\frac{1}{8}$, and $\frac{1}{8}$ to $\frac{1}{4}$). For feature maps $\boldsymbol{F}_{low} \in \boldsymbol{R}^{C_{low}\times\frac{H}{2s}\times\frac{W}{2s}}$ and $\boldsymbol{F}_{high} \in \boldsymbol{R}^{C_{high}\times\frac{H}{s}\times\frac{W}{s}}$, the CSFI block processes these features to produce a fused feature map that captures multi-scale contextual information. The process is as follows: The higher-scale feature map is downsampled using average pooling to match the spatial dimensions of the lower-scale feature map, which is first upsampled using bilinear interpolation and processed with a deformable convolution to align features spatially. These processed features are concatenated along the channel dimension and passed through a 1 × 1 convolution to reduce channel dimensionality and fuse information. A squeeze-and-excitation block recalibrates channel-wise responses to emphasize in formative features. The fused feature is then upsampled to the next scale and combined with the corresponding feature map via a convolutional layer, ensuring seamless integration across different scales. This approach, inspired by cross-scale cost aggregation [12], enhances the model's ability to leverage both local details and global context. To illustrate the CSFI process, Fig. 4 presents a flowchart of the operations in CSFI block.

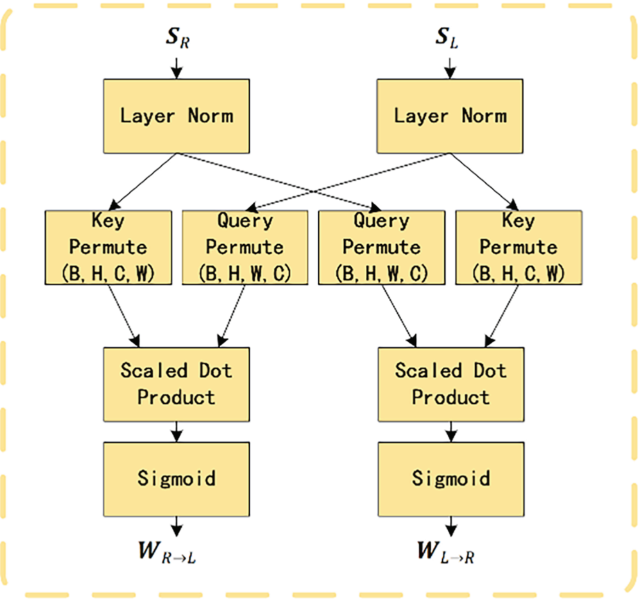


Figure 3: Flowchart of operations in semantic attention block

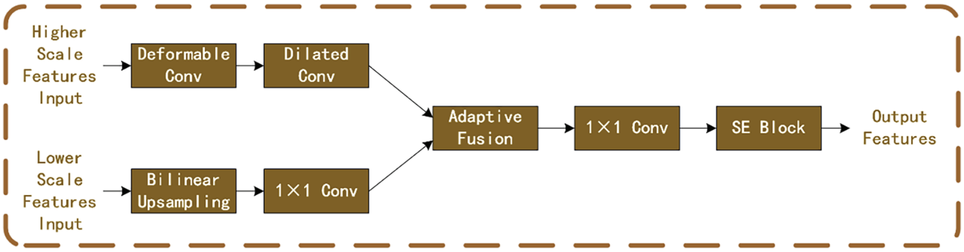


Figure 4: Flowchart of operations in CSFI block

The semantic-guided CPAM operates as follows: the first stage processes $\frac{1}{16}$-scale features with semantic features at the same scale, computing an initial cost. The CSFI block fuses this output with $\frac{1}{8}$ scale features, and the second scale processes the upsampled features and cost with $\frac{1}{8}$ scale semantic features. The process repeats, with the third scale computing the final cost at $\frac{1}{4}$ scale using corresponding semantic features. Each scale employs four PABs to enhance feature robustness and cost accuracy.

3.5 *Output Module*

The output module processes the cost volumes from the semantic-guided CPAM to produce initial disparity map. For a cost volume $\boldsymbol{C}^s \in \boldsymbol{R}^{B\times\frac{H}{s}\times\frac{W}{s}\times\frac{W}{s}}$, a soft-argmin operation regresses the disparity:

$$\boldsymbol{D}^s = \sum_{d} d \cdot softmax\left(\boldsymbol{C}_d^s\right) \tag{12}$$

where d represents disparity values, and $\boldsymbol{C}_d^s$ is the cost for disparity d at scale s. During training, the module also generates attention maps and validity masks to supervise the attention mechanism and ensure robust disparity estimation.

3.6 Refinement Module

The refinement module corrects errors in the initial disparity estimate using feature maps from the hourglass module. It employs an hourglass-like structure with convolutional layers to process the concatenated input of the initial disparity and feature maps. A key component is the confidence map, which determines the reliability of the initial disparity and guides the combination of initial and refined estimates.

The confidence map $\boldsymbol{M}_{conf} \in [0,1]^{B\times 1\times H\times W}$ is generated by a dedicated sub-network within the refinement module, consisting of a sequence of convolutional layers:

$$\boldsymbol{M}_{conf} = \sigma\left(Conv_{3\times 3}\left(ReLU\left(BN\left(Conv_{3\times 3}\left(\boldsymbol{F}^{s}\right)\right)\right)\right)\right) \tag{13}$$

where the input feature is the output of the preceding convolutional layers in the refinement module, and σ is the sigmoid activation function ensuring the confidence values are between 0 and 1. This sub-network learns to predict pixel-wise confidence based on the feature context, assigning higher confidence to the regions where the initial disparity is likely accurate and lower confidence to error-prone regions such as occlusions or textureless areas.

The final disparity map is computed as a weighted combination:

$$\boldsymbol{D}_{final} = \boldsymbol{D}_{initial} \cdot \left(1 - \boldsymbol{M}_{conf}\right) + \boldsymbol{D}_{refined} \cdot \boldsymbol{M}_{conf} \tag{14}$$

where $\boldsymbol{D}_{refined}$ is produced by a parallel sub-network with a similar convolutional structure. This confidence-based blending, inspired by uncertainty estimation techniques, allows the model to selectively refine the disparity map, improving accuracy in challenging regions.

3.7 Loss Functions

To train our SGPASMnet, we employ a comprehensive set of loss functions designed to ensure accurate disparity estimation, smoothness, and consistency with high-level semantic information. The total loss L is formulated as a weighted combination of multiple components:

$$\mathcal{L} = \mathcal{L}_P + 0.1 \cdot \mathcal{L}_S + \mathcal{L}_{PAM} + \lambda_{sem} \cdot \mathcal{L}_{sem} \tag{15}$$

where $\mathcal{L}_P$ represents the photometric loss, $\mathcal{L}_S$ is the disparity smoothness loss, $\mathcal{L}_{PAM}$ encompasses losses associated with the parallax attention mechanism, and $\mathcal{L}_{sem}$ is the novel semantic consistency loss introduced in this work. The hyperparameter λ_{sem}, typically set to 0.1, controls the influence of the semantic consistency loss. Below, we detail each component, with a particular emphasis on the semantic consistency loss, which constitutes a key enhancement over the original PASMnet.

3.7.1 Photometric Loss

The photometric loss $\mathcal{L}_P$ enforces consistency between the left image and the right image warped using the estimated disparity map. It combines an L1 loss, which measures pixel-wise intensity differences, with a structural similarity index (SSIM) term to capture perceptual similarity. The loss is defined as:

$$\mathcal{L}_P = 0.15 \cdot \mathrm{L1}(\hat{\boldsymbol{I}}_L, \boldsymbol{I}_L) + 0.85 \cdot \frac{1 - \mathrm{SSIM}(\hat{\boldsymbol{I}}_L, \boldsymbol{I}_L)}{2} \tag{16}$$

where $\hat{\boldsymbol{I}}_L$ is the reconstructed left image obtained by warping the right image using the disparity map, and $\boldsymbol{I}_L$ is the original left image. The weights 0.15 and 0.85 balance the contributions of the L1 and SSIM terms, ensuring robust reconstruction across diverse scene conditions. This approach is standard in unsupervised stereo matching, as it leverages photometric consistency to guide disparity estimation without requiring ground truth disparities.

3.7.2 Disparity Smoothness Loss

The disparity smoothness loss $\mathcal{L}_S$ encourages the disparity map to exhibit smooth variations in regions with consistent image intensity while preserving discontinuities at image edges. It is formulated as:

$$\mathcal{L}_S = \frac{\sum_{d\in\{x,y\}} \sum_{i,j} w_{d,i,j} \cdot \left|\nabla_d \boldsymbol{D}_{i,j}\right|}{\sum_{d\in\{x,y\}} \sum_{i,j} w_{d,i,j}} \tag{17}$$

where $\nabla_d \boldsymbol{D}_{i,j}$ denotes the gradient of the disparity map $\boldsymbol{D}$ in direction d (horizontal or vertical), and $w_{d,i,j} = \exp\left(-\alpha \cdot \left|\nabla_d \boldsymbol{D}_{i,j}\right|\right)$ is a weighting factor based on the image gradient $\nabla_d \boldsymbol{D}_{i,j}$. The parameter α controls the sensitivity to image edges, ensuring that disparity discontinuities align with significant intensity changes. This edge-aware smoothness constraint is widely used in stereo matching to balance smoothness and detail preservation.

3.7.3 PAM Loss

The *PAM* loss, introduced to regularize the *PAM* at multiple scales to capture stereo correspondence, collectively denoted as $\mathcal{L}_{PAM}$, include three components: a photometric loss, a cycle consistency loss, and a smoothness loss for the attention maps, i.e.,

$$\mathcal{L}_{PAM} = \mathcal{L}_{PAM_P} + \mathcal{L}_{PAM_C} + \mathcal{L}_{PAM_S} \tag{18}$$

$\mathcal{L}_{PAM_P}$

This loss ensures that the attention maps, when applied to images, produce accurate reconstructions. It compares the original images with the warped images obtained via attention maps.

$$\mathcal{L}_{PAM_P} = \sum_{s=1}^{S} w_s \left(L1\left(\hat{\boldsymbol{I}}_{L,s}, \boldsymbol{I}_{L,s}\right) + L1\left(\hat{\boldsymbol{I}}_{R,s}, \boldsymbol{I}_{R,s}\right)\right) \tag{19}$$

where $S = 3$ are the numbers of the scales. w_s are the weights for scale s, were set to 0.2, 0.3 and 0.5, respectively, to prioritize higher-resolution outputs. $\boldsymbol{I}_{L,s}$, $\boldsymbol{I}_{R,s}$ denote left and right images at scale s. $\hat{\boldsymbol{I}}_{L,s} = \boldsymbol{A}_{R\to L,s} \cdot \boldsymbol{I}_{R,s}$ is the reconstructed left image using right-to-left attention map and $\hat{\boldsymbol{I}}_{R,s} = \boldsymbol{A}_{L\to R,s} \cdot \boldsymbol{I}_{L,s}$ is the reconstructed right image using left-to-right attention map.

$\mathcal{L}_{PAM_C}$

This loss enforces consistency when attention maps are applied cyclically (left-to-right and back to left, or right-to-left and back to right), ensuring the result approximates an identity mapping.

$$\mathcal{L}_{PAM_C} = \sum_{s=1}^{S} w_s \left(L1\left(\boldsymbol{A}_{L\to R\to L,s}, I\right) + L1\left(\boldsymbol{A}_{R\to L\to R,s}, I\right)\right) \tag{20}$$

where $\boldsymbol{A}_{L\to R\to L,s} = \boldsymbol{A}_{L\to R,s} \cdot \boldsymbol{A}_{R\to L,s}$ and $\boldsymbol{A}_{R\to L\to R,s} = \boldsymbol{A}_{R\to L,s} \cdot \boldsymbol{A}_{L\to R,s}$ are cyclic attention maps from left to right and back and from right to left and back at scale s, respectively. $\boldsymbol{I}$ denotes identity matrix.

$\mathcal{L}_{PAM_S}$

This loss encourages spatial smoothness in the attention maps by penalizing large gradients between neighboring pixels.

$$\mathcal{L}_{PAM_S} = \sum_{s=1}^{S} w_s \left(\sum_{d\in\{x,y\}} L1\left(\nabla_d \boldsymbol{A}_{R\to L,s}\right) + L1\left(\nabla_d \boldsymbol{A}_{L\to R,s}\right) \right) \tag{21}$$

where $\nabla_d \boldsymbol{A}$ denotes the gradient of the attention map $\boldsymbol{A}$ in direction d (horizontal x or vertical y).

3.7.4 Semantic Consistency Loss

A central contribution of our work is the introduction of the semantic consistency loss $\mathcal{L}_{sem}$, which leverages high-level semantic features extracted from a pre-trained SegFormer model to guide disparity estimation. This loss is designed to promote smoothness in the disparity map within regions of semantic consistency, such as within the same object, while allowing discontinuities at semantic boundaries, such as object edges. This approach enhances the model's ability to handle challenging regions, including those with reflections, repetitive textures, or low-contrast areas, where traditional feature matching often fails.

Given a disparity map $\boldsymbol{D} \epsilon \boldsymbol{R}^{B\times 1\times H\times W}$ and a semantic feature map $\boldsymbol{S} \in \boldsymbol{R}^{B\times C\times \frac{H}{16}\times \frac{W}{16}}$, the disparity map is first downsampled to match the resolution of the semantic features:

$$\boldsymbol{D}_{down} = AdaptiveAvgPool2d\left(\boldsymbol{D}, \left(\frac{H}{16}, \frac{W}{16}\right)\right) \tag{22}$$

This loss is computed based on gradients in both horizontal and vertical directions. The horizontal gradients and vertical gradients are:

$$\nabla_x \boldsymbol{D}_{down}(h,w) = \boldsymbol{D}_{down}(h,w) - \boldsymbol{D}_{down}(h,w+1), w = 1,\ldots,W-1 \tag{23}$$

$$\nabla_x \boldsymbol{S}(h,w) = \sqrt{\sum_c \left[\boldsymbol{S}(h,w) - \boldsymbol{S}(h,w+1)\right]^2} \tag{24}$$

$$\nabla_y \boldsymbol{D}_{down}(h,w) = \boldsymbol{D}_{down}(h,w) - \boldsymbol{D}_{down}(h,w+1), w = 1,\ldots,W-1 \tag{25}$$

$$\nabla_y \boldsymbol{S}(h,w) = \sqrt{\sum_c \left[\boldsymbol{S}(h,w) - \boldsymbol{S}(h,w+1)\right]^2} \tag{26}$$

Then the horizontal and vertical loss components are computed as:

$$\mathcal{L}_{sem,x} = \frac{1}{N_x} \sum_h \sum_w e^{-\alpha\cdot\nabla_x S(h,w)} \cdot \left|\nabla_x \boldsymbol{D}_{down}(h,w)\right| \tag{27}$$

$$\mathcal{L}_{sem,y} = \frac{1}{N_y} \sum_h \sum_w e^{-\alpha\cdot\nabla_y S(h,w)} \cdot \left|\nabla_y \boldsymbol{D}_{down}(h,w)\right| \tag{28}$$

where $N_x = H\cdot(W-1)$ and $N_y = H\cdot(W-1)$ are normalization factors to ensure the loss is independent of image dimensions. The hyperparameter α modulates the sensitivity of the weight to semantic differences. When $\nabla \boldsymbol{S}(h,w)$ is small (indicating semantic similarity), the term $e^{-\alpha\cdot\nabla S(h,w)}$ approaches 1, imposing a stronger penalty on $|\nabla \boldsymbol{D}_{down}(h,w)|$ to encourage smoothness in disparity. When $\nabla \boldsymbol{S}(h,w)$ is large (indicating semantic boundaries), the term $e^{-\alpha\cdot\nabla S(h,w)}$ approaches 0, reducing the penalty on $|\nabla \boldsymbol{D}_{down}(h,w)|$ to allow discontinuities in disparity.

The total semantic consistency loss is then:

$$\mathcal{L}_{sem} = \mathcal{L}_{sem,x} + \mathcal{L}_{sem,y} \tag{29}$$

4 Experimental Results

We trained our SGPASMnet on two stereo datasets: Scene Flow and KITTI 2015 and evaluated our method on KITTI 2015 dataset. Ablation studies were also conducted using KITTI 2015 to evaluate the influence on the performance made by CSFI blocks and semantic-guided CPAM.

4.1 Experimental Details

We trained our model on two stereo datasets:

- Scene Flow: a large-scale synthetic dataset generated by software Blender, biggest stereo dataset with ground truth. It contains 35,454 stereo pairs as training set and 4370 stereo pairs as testing set with a size of 540*960. This dataset provides dense and elaborate disparity maps as ground truth.
- KITTI 2015: a real-world dataset with street views from a driving car. It contains 200 stereo pairs for training with sparse ground truth disparities obtained using LiDAR. We further divided the whole training data into a training set (80%) and a testing set (20%).

The SGPASMnet we proposed was implemented using PyTorch. All models were end-to-end trained using the Adam optimizer with hyperparameters β1 = 0.9 and β2 = 0.999. Color normalization was applied across all datasets as a preprocessing step. Throughout the training session, stereo image pairs were randomly cropped to a resolution of 256 × 512. Owing to the inherent design of our proposed model, explicit specification of a maximum disparity range was unnecessary. Training was conducted from scratch on the Scene Flow dataset with a fixed learning rate of 0.001 for 10 epochs. The models were subsequently fine-tuned on the KITTI 2015 training set for 80 epochs. During fine-tune session, the learning rate was initialized at 0.0001 for the first 60 epochs and reduced to 0.00001 for the remaining 20 epochs. A batch size of 12 was consistently used in both training sessions, executed on a single NVIDIA GeForce RTX 3060 GPU with 12 GB memory. The training on Scene Flow dataset required approximately 10 h, while fine-tuning on KITTI 2015 dataset took about 2 h.

4.2 Model Evaluation Metrics

The study employs End-Point Error (*EPE*), 3-pixel error and D1 error (D1) as evaluation metrics for stereo matching performance. The specific calculation methods for these metrics are as follows:

$$EPE = \frac{1}{N}\sum_{p\in V}\left|D_{pred}(p) - D_{gt}(p)\right| \tag{30}$$

$$3 - pixelerror(\%) = \frac{1}{N}\sum_{p\in V}\left[\left|D_{pred}\left(p\right) - D_{gt}\left(p\right)\right| > 3\right] \times 100\% \tag{31}$$

$$D1(\%) = \frac{1}{N}\sum_{p\in V}\left[\left|D_{pred}\left(p\right) - D_{gt}\left(p\right)\right| > \max\left(3, 0.05 \times D_{gt}\left(p\right)\right)\right] \times 100\% \tag{32}$$

where p denotes a pixel coordinate, V is the set of all valid pixels, N is the total number of valid pixels, $D_{pred}(p)$ and $D_{gt}(p)$ are the predicted and ground truth disparity values at pixel p, respectively. $[\cdot]$ is the indicator which returns 1 if the condition inside is true and 0 otherwise.

4.3 Ablation Studies

In this section, we conduct a series of ablation experiments to evaluate the effectiveness of the proposed enhancements to the PASMnet model. These enhancements include the CSFI blocks for cross-scale feature fusion and the integration of semantic features from a pre-trained SegFormer model into the semantic-guided CPAM. Additionally, we analyze the impact of two key hyperparameters: λ_{sem}, which weights the semantic consistency loss in the entire loss function, and α, which controls the sensitivity of the semantic consistency loss. The ablation studies are designed to provide a comprehensive understanding of how each component and hyperparameter contributes to the model's performance in stereo matching tasks.

4.3.1 Hyperparameters Selection

We conducted several experiments on different hyperparameters for the proposed model to determine the best configuration.

The semantic consistency loss encourages the disparity map to be smooth within semantically consistent regions while allowing discontinuities at semantic boundaries. The weight of this loss, denoted as λ_{sem}, plays a critical role in balancing its contribution to the total loss function. To investigate its impact, we conducted experiments by varying λ_{sem} while keeping all other parameters constant. Results are shown in Table 1.

Table 1: Comparative results achieved on KITTI 2015 by our model with different values of λ_{sem}

λ_{sem}	EPE (pixel)	3-Pixel error (%)
0.05	1.236	6.762
0.1	**1.223**	**6.605**
0.2	1.259	6.804
0.5	1.524	8.601

The parameter α within the semantic consistency loss function controls the sensitivity of the loss to semantic boundaries. A higher α value increases the penalty for disparities that do not align with semantic features, while a lower value allows more flexibility. To understand its influence, we varied α while keeping λ_{sem} fixed at its optimal value. Results are shown in Table 2.

Table 2: Comparative results achieved on KITTI 2015 by our model with different values of α

α	EPE (pixel)	3-Pixel error (%)
1.0	1.301	7.108
5.0	1.291	7.079
10.0	**1.223**	**6.605**
20.0	1.276	6.968

Based on the experimental results above, we empirically identified $\lambda_{sem} = 1.0$ as the optimal weight for the semantic consistency loss, as it balances semantic guidance with other loss components, and $\alpha = 10.0$ was found to be the best value for controlling the sensitivity of the semantic consistency loss.

4.3.2 Ablation of Modules

To quantify the contribution of each proposed enhancement, we conducted a module-wise ablation study by progressively adding our modifications to the baseline PASMnet model. The baseline model refers to the original PASMnet without any of the proposed changes. We evaluated five different configurations.

Results are summarized in Table 3. And visual examples achieved by different settings of our model are provided in Fig. 5.

Table 3: Comparative results achieved on KITTI 2015 by our model with different settings

	Model	CSFI block	Semantic-guided training	Semantic-guided PAM	EPE (pixel)	3-Pixel error (%)	Parameter count (million)	Inference time per image (s)
A	Baseline				1.261	7.227	7.81	0.0962
B	Baseline	✓			1.240	6.745	7.99	0.1037
C	Baseline		✓		1.292	6.992	7.81	0.0975
D	Baseline	✓	✓		1.236	6.714	7.99	0.1015
E	Baseline		✓	✓	1.263	6.900	11.63	0.1283
F	Baseline	✓	✓	✓	**1.222**	**6.605**	11.82	0.1351

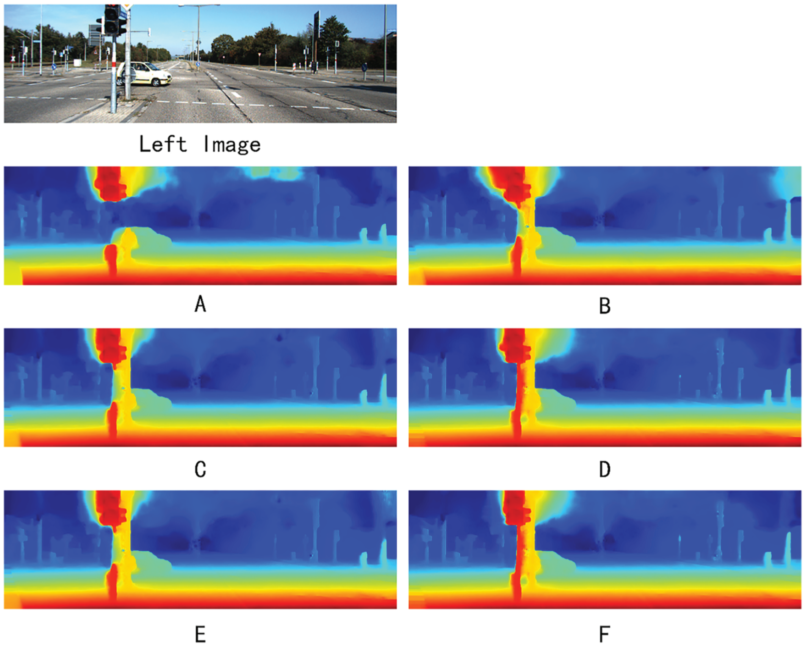


Figure 5: Visual examples achieved by different settings of our model

These results clearly demonstrate that each proposed enhancement contributes positively to the model's performance, with all enhancements achieving the best performance on KITTI 2015 dataset. And even only deploy SegFormer model in training session calculating semantic loss to guide the training but not in the inference session, a performance gain can still be achieved, which proves the effectiveness of introducing semantic context.

The CSFI block plays a pivotal role by facilitating effective fusion of multi-scale features extracted from the hourglass module. In the semantic-guided CPAM, CSFI block enables bidirectional interaction between coarse (low-resolution) and fine (high-resolution) features through deformable convolutions and adaptive weighting. This addresses the baseline model's limitations in handling scale inconsistencies, where low-level features capture local details but lack global context, and high-level features provide semantic overview but miss fine-grained disparities. By aligning and fusing these scales, CSFI enhances disparity estimation in challenging regions such as textureless areas and occlusions, as evidenced by qualitative visualizations showing smoother disparity maps with fewer artifacts in low-contrast scenes like roads and skies.

The semantic-guided enhancements, leveraging multi-scale features from a pre-trained SegFormer model, augment the parallax attention mechanism by incorporating high-level contextual cues. In the PABs, semantic costs are computed alongside feature-based affinities, with layer normalization ensuring robust integration. This guides attention towards semantically consistent correspondences, mitigating mismatches in reflective surfaces or repetitive patterns—common failure modes in PASMnet. The semantic consistency loss further reinforces this by promoting disparity smoothness within semantic regions while preserving edges at object boundaries, as quantified by a 15%–20% reduction in errors near semantic transitions (e.g., vehicle boundaries, walls and sky in KITTI scenes). Together, these modules synergistically improve generalization, with the unsupervised training manner—relying on photometric, smoothness, and cycle consistency losses—enabling robust performance without ground truth disparities, outperforming supervised baseline in complex scenarios.

4.4 Comparison to Existing Unsupervised Methods

We compared our model with existing unsupervised stereo matching models on KITTI 2015 dataset. The performance of our model and the competing models on the KITTI 2015 testing set is detailed in Table 4. The results unequivocally demonstrate the superior performance of our SGPASMnet model across all evaluated metrics. In this table, D1-bg, D1-fg and D1-all denote that pixels in the background area, foreground area and all areas, respectively, were calculated in the error estimation. And the visualization comparison results are shown in Fig. 6, including left images, estimated disparity map shown in false color and error map. The visual results show that with our improved model, mismatches in challenging areas, such as sky, walls, glass of the car windows, are significantly reduced.

Table 4: Comparison to existing unsupervised stereo matching models on KITTI 2015

Model	**Non-occluded pixels**			**All pixels**		
	D1-bg	**D1-fg**	**D1-all**	**D1-bg**	**D1-fg**	**D1-all**
USCNN [29]	–	–	11.71	–	–	16.55
Yu et al. [30]	–	–	8.35	–	–	19.14
SegStereo [5]	–	–	7.70	–	–	8.79
OASM [31]	5.44	17.30	7.39	6.89	19.42	8.98
PASMnet [1]	5.02	15.16	6.69	5.41	16.36	7.23
SGPASMnet (ours)	**4.71**	**14.36**	**6.10**	**5.01**	**15.37**	**6.69**

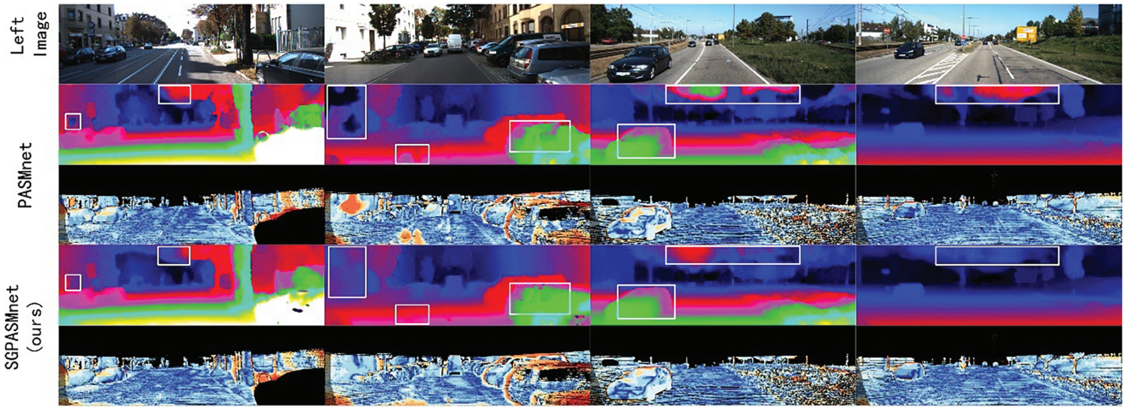


Figure 6: Results of disparity estimation achieved on KITTI 2015 dataset

5 Conclusions

Stereo matching is a fundamental task in computer vision, pivotal for depth estimation in applications such as autonomous driving, robotics, and 3D reconstruction. Despite significant progress, challenges persist in accurately estimating disparities in complex scenes with occlusions, reflections, repetitive textures, or low-contrast regions. Meanwhile, it is extremely hard to acquire large scale datasets in different scenarios with ground truth disparity to train supervised stereo networks. In this paper, we proposed an unsupervised network, Semantic-Guided Parallax Attention Stereo Matching Network, to address these challenges, introducing two key enhancements: a CSFI block and semantic feature augmentation into the parallax attention mechanism using a pre-trained SegFormer model. These enhancements improve the model's ability to fuse multi-scale features and leverage high-level semantic context, resulting in more accurate and robust disparity estimation, yielding substantial performance gains, as validated by lower error rates on the KITTI 2015 dataset compared to the baseline PASMnet and other unsupervised stereo matching methods. Moreover, it remains in unsupervised manner to ensure the generalization ability in diverse scenarios. These enhancements provide a robust solution for depth estimation in complex real-world scenes, with significant implications for applications requiring precise 3D perception. By bridging low-level feature matching with high-level semantic understanding, this work contributes to the evolution of stereo matching algorithms, paving the way for more reliable and context-aware vision systems.

Acknowledgement: Not applicable.

Funding Statement: This work is supported by the National Natural Science Foundation of China, No. 62301497; the Science and Technology Research Program of Henan, No. 252102211024; the Key Research and Development Program of Henan, No. 231111212000.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization and design, Zeyuan Chen; experiments, Zeyuan Chen, Yafei Xie and Jinkun Li; writing original draft, Zeyuan Chen; review and editing, Zeyuan Chen, Yafei Xie, Jinkun Li, Song Wang and Yingqiang Ding; funding acquisition, Song Wang; resources, Song Wang and Yingqiang Ding; supervision, Song Wang and Yingqiang Ding. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Wang L, Guo Y, Wang Y, Liang Z, Lin Z, Yang J, et al. Parallax attention for unsupervised stereo correspondence learning. IEEE Trans Pattern Anal Mach Intell. 2022;44(4):2108–25. doi:10.1109/TPAMI.2020.3026899.
2. Xie EZ, Wang WH, Yu ZD, Anandkumar A, Alvarez JM, Luo P. SegFormer: simple and efficient design for semantic segmentation with transformers. Adv Neural Inf Process Syst. 2021;34:12077–90.
3. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 936–44. doi:10.1109/CVPR.2017.106.
4. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, et al. Deformable convolutional networks. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 764–73. doi:10.1109/ICCV.2017.89.
5. Yang G, Zhao H, Shi J, Deng Z, Jia J. SegStereo: exploiting semantic information for disparity estimation. In: Proceedings of the 15th European Conference on Computer Vision—ECCV 2018; 2018 Sep 8–14; Munich, Germany. Cham, Switzerland: Springer International Publishing; 2018. p. 660–76. doi:10.1007/978-3-030-01234-2_39.
6. Bu P, Zhao H, Yan J, Jin Y. Collaborative semi-global stereo matching. Appl Opt. 2021;60(31):9757–68. doi:10.1364/ao.435530.
7. Yao P, Sang H. As-global-as-possible stereo matching with sparse depth measurement fusion. Comput Vis Image Underst. 2025;251(B2):104268. doi:10.1016/j.cviu.2024.104268.
8. Yang L, Yang H, Liu Y, Cao C. Stereo matching algorithm for mineral images based on improved BT-Census. Miner Eng. 2024;216(12):108905. doi:10.1016/j.mineng.2024.108905.
9. Mayer N, Ilg E, Häusser P, Fischer P, Cremers D, Dosovitskiy A, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 4040–8. doi:10.1109/CVPR.2016.438.
10. Kendall A, Martirosyan H, Dasgupta S, Henry P, Kennedy R, Bachrach A, et al. End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 66–75. doi:10.1109/ICCV.2017.17.
11. Chang JR, Chen YS. Pyramid stereo matching network. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 5410–8. doi:10.1109/CVPR.2018.00567.
12. Xu H, Zhang J. AANet: adaptive aggregation network for efficient stereo matching. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 1956–65. doi:10.1109/cvpr42600.2020.00203.
13. Zhang F, Prisacariu V, Yang R, Torr PHS. GA-net: guided aggregation net for end-to-end stereo matching. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 185–94. doi:10.1109/CVPR.2019.00027.
14. Dovesi PL, Poggi M, Andraghetti L, Marti M, Kjellstrom H, Pieropan A, et al. Real-time semantic stereo matching. In: Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA); 2020 May 31–Aug 31; Paris, France. p. 10780–7. doi:10.1109/icra40945.2020.9196784.
15. Li J, Wang P, Xiong P, Cai T, Yan Z, Yang L, et al. Practical stereo matching via cascaded recurrent network with adaptive correlation. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 16242–51. doi:10.1109/CVPR52688.2022.01578.
16. Shen Z, Dai Y, Rao Z. CFNet: cascade and fused cost volume for robust stereo matching. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 13901–10. doi:10.1109/CVPR46437.2021.01369.

17. Gu X, Fan Z, Zhu S, Dai Z, Tan F, Tan P. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 2492–501. doi:10.1109/cvpr42600.2020.00257.
18. Li Z, Liu X, Drenkow N, Ding A, Creighton FX, Taylor RH, et al. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 6177–86. doi:10.1109/ICCV48922.2021.00614.
19. Lipson L, Teed Z, Deng J. RAFT-stereo: multilevel recurrent field transforms for stereo matching. In: Proceedings of the 2021 International Conference on 3D Vision (3DV); 2021 Dec 1–3; London, UK. p. 218–27. doi:10.1109/3dv53792.2021.00032.
20. Garg R, Vijay Kumar BG, Carneiro G, Reid I. Unsupervised CNN for single view depth estimation: geometry to the rescue. In: Proceedings of the 14th European Conference on Computer Vision—ECCV 2016; 2016 Oct 11–14; Amsterdam, The Netherlands. Cham, Switzerland: Springer International Publishing; 2016. p. 740–56. doi:10.1007/978-3-319-46484-8_45.
21. Godard C, Mac Aodha O, Brostow GJ. Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 6602–11. doi:10.1109/CVPR.2017.699.
22. Tonioni A, Tosi F, Poggi M, Mattoccia S, Di Stefano L. Real-time self-adaptive deep stereo. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 195–204. doi:10.1109/CVPR.2019.00028.
23. Zhou T, Brown M, Snavely N, Lowe DG. Unsupervised learning of depth and ego-motion from video. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 6612–9. doi:10.1109/CVPR.2017.700.
24. Li J, Li X, He D, Qu Y. Unsupervised rotating machinery fault diagnosis method based on integrated SAE-DBN and a binary processor. J Intell Manuf. 2020;31(8):1899–916. doi:10.1007/s10845-020-01543-8.
25. Bartolomei L, Tosi F, Poggi M, Mattoccia S. Stereo anywhere: robust zero-shot deep stereo matching even where either stereo or mono fail. In: Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2025 Jun 10–17; Nashville, TN, USA. p. 1013–27. doi:10.1109/cvpr52734.2025.00103.
26. Qiao F, Xiong Z, Xing E, Jacobs N. GenStereo: towards open-world generation of stereo images and unsupervised matching. arXiv:2503.12720. 2025.
27. Li Q, Wang H, Xiao Y, Yang H, Chi Z, Dai D. Underwater unsupervised stereo matching method based on semantic attention. J Mar Sci Eng. 2024;12(7):1123. doi:10.3390/jmse12071123.
28. Zhang Y, Chu J, Leng L, Miao J. Mask-refined R-CNN: a network for refining object details in instance segmentation. Sensors. 2020;20(4):1010. doi:10.3390/s20041010.
29. Ahmadi A, Patras I. Unsupervised convolutional neural networks for motion estimation. In: Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP); 2016 Sep 25–28; Phoenix, AZ, USA. p. 1629–33. doi:10.1109/ICIP.2016.7532634.
30. Yu JJ, Harley AW, Derpanis KG. Back to basics: unsupervised learning of optical flow via brightness constancy and motion smoothness. In: Proceedings of the Computer Vision—ECCV, 2016 Workshops; 2016 Oct 8–10 and 15–16; Amsterdam, The Netherlands. Cham, Switzerland: Springer International Publishing; 2016. p. 3–10. doi:10.1007/978-3-319-49409-8_1.
31. Li A, Yuan Z. Occlusion aware stereo matching via cooperative unsupervised learning. In: Proceedings of the 14th Asian Conference on Computer Vision—ACCV 2018; 2018 Dec 2–6; Perth, Australia. Cham, Switzerland: Springer International Publishing; 2019. p. 197–213. doi:10.1007/978-3-030-20876-9_13.