Pseudo-Stereo Inputs: A Solution to the Occlusion Challenge in Self-Supervised Stereo Matching

Ruizhi Yang[1,2,3,4], Xingqiang Li[1,2,4⋆], Jiajun Bai[1,2,4], and Jinsong Du[1,2,4]

[1] Shenyang Institute of Automation, Chinese Academy of Sciences
[2] Liaoning Liaohe Laboratory
[3] University of Chinese Academy of Sciences
[4] Liaoning Key Laboratory of Intelligent Detection and Equipment Technology
{yangruizhi, lixingqiang, baijiajun}@sia.cn, jsdu_sia@163.com

Abstract. Self-supervised stereo matching holds great promise by eliminating the reliance on expensive ground-truth data. Its dominant paradigm, based on photometric consistency, is however fundamentally hindered by the occlusion challenge—an issue that persists regardless of network architecture. The essential insight is that for any occluders, valid feedback signals can only be derived from the unoccluded areas on one side of the occluder. Existing methods attempt to address this by focusing on the erroneous feedback from the other side, either by identifying and removing it, or by introducing additional regularities for correction on that basis. Nevertheless, these approaches have failed to provide a complete solution. This work proposes a more fundamental solution. The core idea is to transform the fixed state of one-sided valid and one-sided erroneous signals into a probabilistic acquisition of valid feedback from both sides of an occluder. This is achieved through a complete framework, centered on a pseudo-stereo inputs strategy that decouples the input and feedback, without introducing any additional constraints. Qualitative results visually demonstrate that the occlusion problem is resolved, manifested by fully symmetrical and identical performance on both flanks of occluding objects. Quantitative experiments thoroughly validate the significant performance improvements resulting from solving the occlusion challenge. The code is available at https://github.com/qrzyang/Pseudo-Stereo.

1 Introduction

Stereo matching, a cost-effective method for obtaining reliable 3D information, has been a longstanding focus of research in computer vision [26]. In recent years, learning-based stereo matching methods have emerged, demonstrating superior performance in terms of accuracy and efficiency [1,18,20]. Nevertheless, the dominant paradigm remains fully reliant on ground-truth annotations, fundamentally

⋆ Corresponding authors

limiting its potential. Self-supervised stereo matching [16, 31, 32, 37] was introduced to resolve this annotation issue and initially progressed in tandem with supervised methods, but its development soon reached a bottleneck, hindering further advancement.

The guiding principle of self-supervised stereo matching is the photometric consistency between different views [6]. This principle, however, inherently fails in regions where the consistency assumption is violated, resulting in erroneous feedback signals. Occlusions are the most critical challenge, as they produce large-scale, persistent erroneous feedback that severely degrades network performance. Consequently, a significant body of work has focused on addressing this problem [3, 6, 16, 17, 35], including related domains facing the same challenge [10, 12, 22, 33]. Their approaches can be broadly summarized into a two-pronged strategy: one is to identify occluded regions to prevent the backpropagation of erroneous gradients, and the other is to compensate for the missing information in these areas by introducing additional sources of feedback.

Information restoration for occluded areas remains a major challenge, and existing methods have yet to achieve satisfactory results. Mainstream approaches rely on providing auxiliary feedback through mechanisms like an edge-aware smoothness loss [6, 12, 31], while others attempt to exploit local disparity regularities [17, 35]. Fundamentally, these methods depend on information from the immediate neighborhood, which presumes a high degree of regularity and correlation between the occluded region and its surroundings—an assumption that is often unreliable. Some works [4, 6, 16] advocate for using a flip data augmentation to source information from symmetric regions. However, as demonstrated in Fig. 1, this strategy is ineffective. Due to the coupled nature and fixed directionality of the input and feedback signal, occlusions consistently appear on a fixed side of the occluding object in the primary view, and the flip operation fails to alter this geometric reality. In contrast, methods [29, 36] that introduce a third viewpoint—using a central primary view and selecting the minimum loss from the left and right views—have proven effective at correctly resolving occlusions. Nevertheless, these solutions are impractical for standard settings. They either necessitate specialized, perfectly symmetric trinocular camera hardware [36] or require a full 3D scene reconstruction via methods like NeRF to synthesize the novel view [29]. Consequently, such techniques cannot be directly applied for self-supervised training on conventional binocular image data.

In this work, we propose a novel occlusion handling method that acquires reliable feedback, instead of relying on simple inference from local neighbors, and also without introducing any additional constraints. The core idea, illustrated in Fig. 1, is to decouple the network input and its photometric feedback signal by using a pseudo third-view input. This principle effectively reframes the problem from one of static, one-sided occlusions to one of dynamic, two-sided occlusions with equal probability. The execution of this idea is non-trivial, and we identified and solved three core challenges. First, how to generate the pseudo third-view input with minimal cost and using it to achieve the aforementioned transformation to symmetric occlusions. Second, given that symmetric occlusions imply

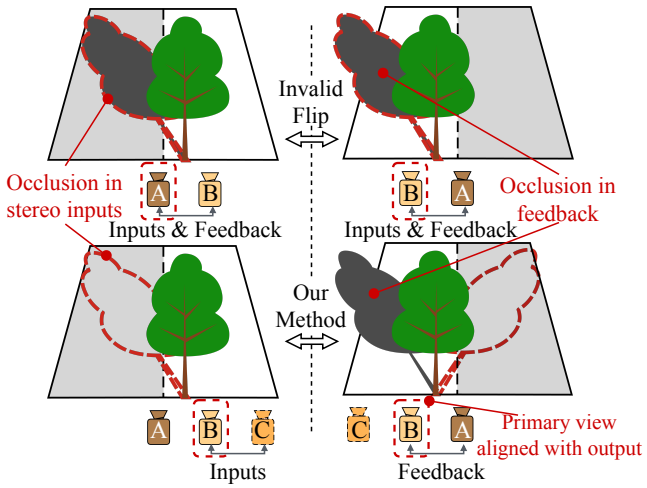

Fig. 1: An illustration of our motivation. (Top) The fixed-side occlusion problem where standard flipping is ineffective, as the occlusion side remains constant relative to the inputs. (Bottom) Our method decouples inputs and feedback using a pseudo-view (C), breaking the persistent directional error signal.

that each side of an occluder receives correct and incorrect feedback with equal probability, how to guide the network to converge to the correct solution. Third, how to ensure proper network convergence and prevent overfitting, given the distinguishable nature of the pseudo-input compared to the real input.

Our experiments validate that the proposed full framework effectively addresses the occlusion challenge, manifested by the symmetric and accurate estimates on both sides of occluders shown in Fig. 6, and the significant performance gains reported in Tab. 1. Furthermore, subsequent experiments (Fig. 7 and Tab. 2) confirm that occlusion is a core bottleneck in self-supervised stereo matching, independent of network architecture or dataset. We demonstrate that our proposed method successfully overcomes this fundamental bottleneck across diverse network architectures and datasets.

In summary, the main contributions of this work are as follows.

1. We propose a novel occlusion handling strategy that uses pseudo-stereo inputs to decouple network input and feedback. This fundamentally reframes the learning problem in occluded regions.
2. We introduce a complete and robust framework to realize this idea, ensuring convergence under feedback conflicts and preventing overfitting to synthetic artifacts.
3. Extensive experiments show our method mitigates the occlusion bottleneck to achieve significant performance improvements, generalizing effectively across various network backbones and datasets.

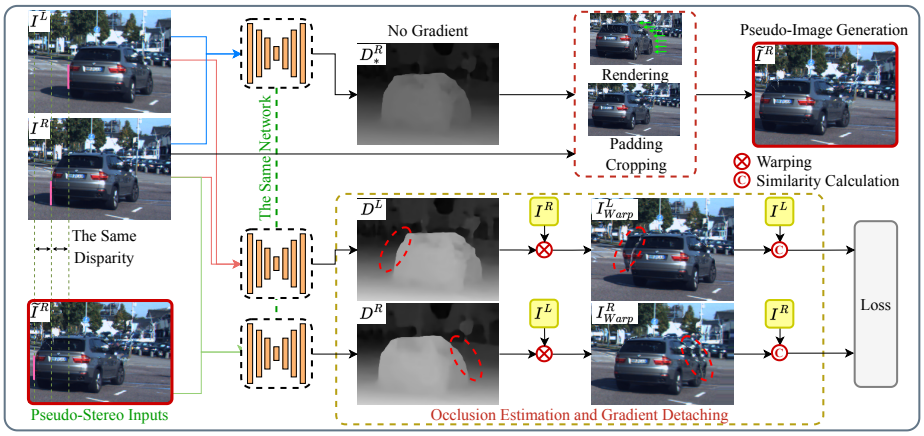

Fig. 2: An overview of our proposed pipeline. We leverage a pseudo-input to decouple network inputs and feedback. This enables a symmetric training paradigm where occlusions are probabilistically relocated, providing valid feedback for both sides of occluders. An occlusion estimation and gradient detaching mechanism is employed to resolve the inherent feedback conflicts.

2 Related Work

2.1 Learning-Based Stereo Matching Backbone

The rapid advancement of deep learning has given rise to powerful stereo matching architectures [8, 28]. Representative works like PSMNet [1] introduced 3D convolutions for cost aggregation, while more recent methods like GMStereo [34] incorporate transformers [30] for more capable feature matching. Recently, approaches such as MonSter [2] have shown success by integrating features from pre-trained monocular depth estimators. Despite their increasingly powerful feature representation and matching capabilities, these models are fundamentally designed for a supervised paradigm and require accurate ground-truth disparities. When applied to self-supervised learning, where the supervision signal can be ambiguous or erroneous, their advanced feature induction capabilities, while still beneficial, do not address the core bottleneck. Although some works [3,16,31] have proposed architectural modifications specifically for self-supervised occlusion handling, their visual results reveal that the underlying problem is not significantly mitigated. This suggests that the solution to the occlusion problem lies primarily within the loss formulation rather than the network architecture.

2.2 Self-Supervised Stereo and Occlusion Handling

Self-supervised stereo methods are built upon the principle of photometric consistency between views [4, 31, 32, 35, 37]. Consequently, a large body of work has focused on strategies to address this principle's primary weakness: the erroneous feedback generated by occlusions. A common consensus is to identify occluded

regions via a mask and exclude them from the photometric loss, typically through pixel-wise multiplication [12, 22, 31–33]. Method [12] compared various occlusion inference methods [11, 22, 33], demonstrating that all benefit performance, and crucially established that the gradient stop for occlusion mask is an important strategy.

With occlusions masked, however, an information void is created. Various methods have been proposed to compensate for this. Without introducing external constraints, this compensation is typically provided by a standard smoothness loss [12,22,31–33]. [12] demonstrated that different smoothness edge-weights can significantly impact performance. Nevertheless, since smoothness is not always a reliable or correct convergence target, this approach remains unsatisfactory. Some methods [4,6,16] employ flip strategy and regard left-right consistency directly as a loss term, attempting to use it as a key solution. However, this does not fundamentally mitigate the occlusion problem, as an incorrect disparity can still satisfy the consistency. A more targeted line of work [17, 35] have designed particular losses to extract more reliable information from local neighborhoods. Among these, [17] focuses on interpolating from background neighbors rather than foreground ones, while [35] encourages occluded regions and their neighborhoods to conform to a 3D geometry inferred from monocular cues. These methods are more tailored for occlusion handling and have achieved performance gains, but their solutions are not robust as they rely on the strong assumption that neighboring information is sufficiently correlated with the occluded content.

Distinct from the aforementioned methods, some works [29, 36] acquire reliable, non-local feedback by leveraging additional views. Specifically, [36] uses a fixed, symmetric multi-view camera hardware where a region occluded from one direction is visible from another. [29] achieves the same goal by rendering a third-view image using Neural Radiance Fields [24] (NeRF). These two approaches provide a conceptually correct pathway for occlusion handling: obtaining unoccluded information from an auxiliary perspective. However, both methods either depend on specialized hardware or require additional data for scene reconstruction, imposing external constraints that prevent their direct application to conventional binocular datasets. This is inconsistent with the goal of self-supervised learning, which aims for direct training on any target scene.

In this paper, we propose a self-supervised method that can acquire the same reliable feedback as these multi-view approaches without depending on any such external constraints.

3 Method

The motivation and the overall framework of our method are illustrated in Fig. 1 and Fig. 2, respectively. Section 3.1 details the core idea in our approach: a pseudo-stereo inputs strategy that decouples input images from feedback images. Section 3.2 further discusses the practical challenges of this design and proposes solutions, including addressing feedback conflicts and overfitting problems.

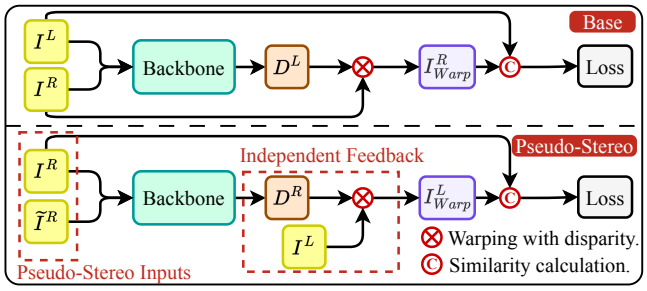

Fig. 3: A comparison between the existing self-supervised pipeline and our pseudo-stereo inputs strategy. In contrast to existing methods, our method decouples the input stereo images from the feedback stereo images. This allows for flexible feedback using image pairs with different occlusion positions.

3.1 Pseudo-Stereo Inputs Strategy

This section elucidates the principles and implementation of our pseudo-stereo inputs strategy, which addresses the occlusion challenge by decoupling the input image pairs from the feedback image pairs in self-supervised training. The core method for this decoupling is to utilize image pairs containing a generated pseudo-image as inputs instead of original image pairs. In contrast, the feedback calculation relies solely on the original images to ensure the reliability and accuracy of the feedback. This decoupling removes the fixed relative positional constraints on the input and feedback images, allowing them to flexibly select positions where occlusions occur, thereby addressing the issue of continuous information loss due to fixed occlusions. An intuitive illustration of the principle is shown in Fig. 1.

Pseudo-Stereo Inputs Figure 3 visually illustrates the distinction between this strategy and existing self-supervised paradigms. In this strategy, the left view remains a real image from the dataset, while the right view is replaced with a generated pseudo-image. Specifically, existing methods use a stereo image pair I^L,I^R from the dataset as network inputs and output a disparity map D^L aligned with I^L. Where I and D denote color images and disparity maps, respectively. Superscripts L and R indicate correspondence to the original left and right viewpoints. In contrast, when using pseudo-inputs, an estimated disparity D^R and I^R are used to generate a pseudo-image, denoted as $\widetilde{I}^R$. At this point, the network inputs become I^R as the left view and $\widetilde{I}^R$ as the right view, producing a dense disparity map D^R aligned with I^R. The images used for feedback should be the original ones to ensure reliability. Therefore, the original image I^L is warped using D^R to obtain I^L_{warp}, which is aligned with I^R. I^R and I^L_{Warp} are then employed for loss calculation. When I^L is as the left input, the occluded areas appear to the left of the occluders; conversely, when I^R is as the left input, the occluded areas are on the right side of the occluders. This strategy ensures that the network can consistently have a probability to capture information from

both sides of the occluders, thus resolving the problem where previous methods invariably lacked information from one side of the occluders.

Method for Generating Pseudo-Images While numerous methods for novel view synthesis exist [13, 14, 24], they typically require additional data to train the models themselves. Incorporating such methods would introduce extraneous data and priors, which contradicts the fundamental principle of self-supervised learning. Therefore, our approach must exclusively utilize information available within the original training pipeline. In the standard stereo estimation pipeline, the commonly used warping operation is a form of image synthesis; this involves using D^L, which is aligned with I^L, to warp I^R into a novel image that aligns with I^L. However, this approach cannot be used for generating pseudo-images. The reason is that we do not have a disparity map aligned with the pseudo-image prior to its generation. Consequently, our approach employs a pseudo-image generation method that more closely resembles a simplified rendering process as illustrated in Fig. 4. For I^R, each pixel can be relocated backward using D^R to generate $\widetilde{I}^R$. If multiple pixels map to the same location in $\widetilde{I}^R$, only the pixel with the maximum disparity is kept. This corresponds to the occlusion effect during rendering. As shown in Fig. 5, this process yields a pseudo-image containing missing pixels. To make it as similar to the real image as possible, it is necessary to pad these missing pixels. The padding method and how closely the resulting image resembles a real one directly influence the severity of the overfitting problem. This issue is discussed in detail in Sec. 3.2.

Discussion on Using Pseudo Images as Inputs As described earlier, our method does not use a previously trained network to generate novel views. Instead, the generation of pseudo-images relies on the network that is currently undergoing training. In other words, images generated in the early stages of training, which severely lack realism, are also used as inputs during the training process. This raises concerns about the network's convergence. However, in fact, forcing the network to exploit monocular cues during stereo training is a common and effective practice. As an example, in the supervised stereo matching method [27], random patches are used to occlude the right view to accomplish this objective.

The cornerstone of our method's stable convergence is that the left input and the feedback image are always real and reliable. The effective feedback compels the network to leverage monocular cues in the initial stages of training to enhance prediction accuracy. As accuracy increases, the precision of pseudo-image generation also improves, leading to better binocular information for the network. This results in a stable and progressively enhancing iterative process. The experimental results in Section 4.3 also support this point. In these experiments, even when no real images were used for the right input (i.e., using I^L, $\widetilde{I}^L$ or I^R, $\widetilde{I}^R$ as network inputs), the network still converges successfully.

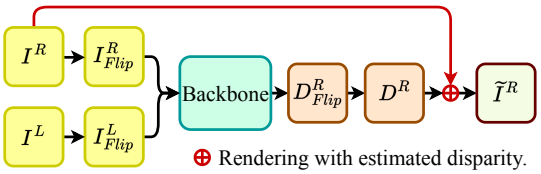

Fig. 4: Schematic diagram of the pseudo-images generation using the rendering method, where Backbone is the model currently under training rather than a previously well-trained one. The subscript "Flip" denotes horizontal flipping.

3.2 Feedback Conflicts and Overfitting Problems

Simply applying the pseudo-stereo inputs strategy with no further modifications already enables a significant performance improvement over the standard paradigm, which can be evident from the ablation study in Sec. 4.3. However, the feedback conflicts and overfitting issues hinder further performance enhancements and prevent stable convergence.

Feedback Conflicts The feedback conflicts issue manifests as unstable disparity estimates in occluded regions. This phenomenon can be readily explained. Despite our strategy in self-supervised feedback ensuring a 50% probability of obtaining correct information on the occluded side, there remains an equal 50% probability of receiving incorrect information. Under the influence of this conflicting feedback, the network will attempt to minimize the loss of both correct and incorrect information simultaneously, finally resulting in a compromised outcome in the occluded region. More importantly, the gradient of the photometric loss is negatively correlated with image similarity. As training progresses and overall image similarity increases, the global average gradient diminishes to a low value. However, occluded regions, which lack correct correspondences, produce a large color discrepancy when warped even with the correct disparity.

Occlusion estimation and detaching the propagation of loss from occluded pixels offers a promising solution to this challenge. While consistency check [9] is an established method for occlusion estimation, it suffers from two notable drawbacks. First, it incurs additional computational cost by requiring the inference of a disparity map for the opposing view. Second, it is prone to errors where incorrect disparities coincidentally satisfy the consistency check, a problem particularly prevalent at object boundaries. We implement a solution that leverages the inherent properties of the current view rather than relying on information from another view. Specifically, we identify invalid pixels during the rendering process—those falling outside the image boundaries or occluded by other pixels—and classify them as occlusion pixels. This method effectively minimizes the risk of misclassifying pixels and prevents erroneous feedback from propagating. After mitigating erroneous feedback stemming from occluded pixels using this approach, we observe a notable improvement in the network's performance. Detailed results can be found in the ablation study presented in

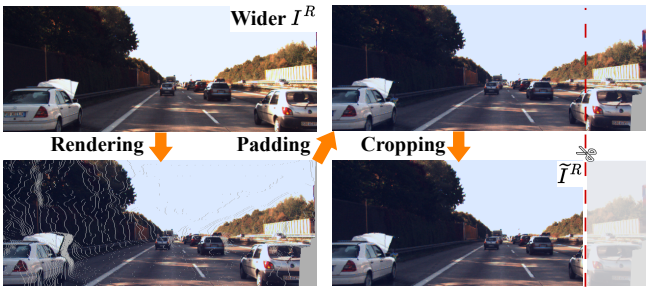

Fig. 5: Visual demonstration of the pseudo-image generation process. Issues with small-scale pixel loss and large-scale pixel missing at edges are mitigated using padding and cropping.

Sec. 4.3. Subsequently, the photometric loss L_p can be expressed as:

$$L_p = \frac{1}{N}\sum_{ij}\left(p \cdot pe(I^L_{ij}, I^R_{ij}) + (1-p)\cdot pe(I^R_{ij}, \widetilde{I}^R_{ij})\right)\cdot \widetilde{O}_{ij} \tag{1}$$

$$\begin{aligned} \text{where } pe(I^A, I^B) = &\frac{\alpha}{2}(1 - SSIM(I^A, I^B)) \\ &+ (1-\alpha)\|I^A - I^B\|_1 \end{aligned} \tag{2}$$

Here, p and $1-p$ represent the probabilities of the respective items, both being constant at 0.5. $\widetilde{O}_{ij}$ denotes the occlusion estimation value, where it equals 1 for non-occluded pixels and 0 for occluded ones. $\alpha = 0.85$ as in [7].

Overfitting Problems Incorporating occlusion estimation into the pseudo-stereo inputs strategy further enhances performance. Nevertheless, extended long-term training stability tests have revealed the overfitting problem, manifested as a gradual decline in accuracy after reaching a peak. Critically, self-supervised training, by its nature, operates without labeled data and thus lacks a validation set. Consequently, early stopping cannot be relied upon to prevent overfitting; instead, the network must be guaranteed to converge stably.

We identify the root cause of this overfitting as the network's ability to distinguish between real and pseudo images in the later stages of training. Specifically, in current strategies, the use of real or pseudo image pairs directly determines the location of occlusions. This enables the network to classify inputs by detecting whether they are pseudo images, thereby converging different results for each input type. This renders the network no longer "consistently have a probability to capture information from both sides of the occluders" as expected.

These observations suggest that the generated pseudo-images contain sufficient features for the network to distinguish them. Based on this inference, we initially attempted to generate more indistinguishable images to prevent the network from classifying the inputs. When generating images using the rendering method described in Sec. 3.1, notable features include small-scale voids at

Table 1: Comparison with existing methods on the KITTI online benchmark.

Method	KITTI 2012 Test				KITTI 2015 Test			
	3-noc (%)	3-all (%)	EPE-noc	EPE-all	D1-bg (%)	D1-fg (%)	D1-all (%)	D1-noc (%)
OASM [16]	6.39	8.60	1.3	2.0	6.89	19.42	8.98	7.39
PASMnet [31]	-	-	-	-	5.41	16.36	7.23	6.69
OASM-DDS [17]	4.81	5.69	1.1	1.2	4.46	15.76	6.51	6.02
UHP [35]	6.05	7.09	1.2	1.3	5.00	13.7	6.45	5.93
Flow2Stereo [19]	4.58	5.11	1.0	1.1	5.01	14.62	6.61	6.29
CRD-Fusion [3]	4.38	5.40	0.9	1.1	4.59	13.68	6.11	5.69
Pseudo-Stereo-PSMNet (ours)	3.46	4.08	0.8	0.9	3.11	12.52	4.68	4.40
Pseudo-Stereo-MonSter (ours)	**3.08**	**3.62**	**0.7**	**0.8**	**2.93**	**11.67**	**4.39**	**4.22**

occluded areas and large-scale missing areas along the right edge, as shown in Fig. 5. For small-scale voids, we fill them using the average value of adjacent non-void pixels. For the missing parts at the right edge, we adopt a wider generation strategy to produce images that are wider than the inputs. This ensures that the resulting cropped images no longer suffer from large-scale missing regions at their edges. It is evident that the generated images now possess a higher degree of realism. Under this strategy, experiments demonstrate an improvement in overfitting issues; however, they have not been completely eradicated. This implies that the network is still able to extract subtle features for differentiation.

We further implement the fully pseudo-stereo inputs strategy. This strategy no longer relies on real image pairs but instead uses fully pseudo-stereo images as input. To be specific, the network's inputs are changed to image pairs I^L,$\widetilde{I}^L$ and I^R, $\widetilde{I}^R$. At this point, since all the inputs are pseudo-stereo inputs, the network naturally cannot differentiate them further. As shown by the experiments in Sec. 4.3, the network achieves stable convergence to high performance and avoids overfitting under this strategy. At this point, (1) becomes,

$$L_p = \frac{1}{N}\sum_{ij}\left(p \cdot pe(I_{ij}^L, \widetilde{I}_{ij}^L) + (1-p)\cdot pe(I_{ij}^R, \widetilde{I}_{ij}^R)\right)\cdot \widetilde{O}_{ij} \tag{3}$$

In addition to L_p, the same edge-aware smoothness L_s as in [7] is adopted, and the disparity values are also mean-normalized like them.

$$L_s = |\partial_x D^{L^*}|e^{-|\partial_x I^L|} + |\partial_y D^{L^*}|e^{-|\partial_y I^L|}, \tag{4}$$

where $D^{L^*} = D^L/\overline{D^L}$. In our experiments, it is observed that an overly large smoothing loss at the early stages hinders the model's exploration capability and has a slight negative impact on its final performance. Therefore, we modify the smoothing term to a parameter that gradually reaches the set value, reducing its influence in the early stages. The final loss can be expressed as:

$$L = L_p + \lambda L_s. \tag{5}$$

λ is the smoothing term coefficient, starting from 0.001 and gradually increasing to 0.5 over 10k iterations.

4 Experiments

In this section, we present experiments to validate that our method effectively addresses the occlusion challenge. The logic of our experimental validation is as follows. First, in Sec. 4.1, we conduct a direct comparison against existing methods on the mainstream benchmark datasets they established. Then, we perform additional experiments across various datasets and backbone networks to verify the generalizability of our method in Sec. 4.2. Finally, Sec. 4.3 presents ablation studies to validate the contribution of each proposed component. We emphasize the visualization of occluded regions to confirm the resolution of the occlusion problem, while also showcasing the performance improvements achieved by solving it.

Datasets and Backbones. For evaluation, we use the KITTI 2012 [5] and KITTI 2015 [23] datasets, which consist of real-world driving scenes and are the established benchmarks for comparing with previous methods [3,16,19,31,35]. To validate the robustness of our method, we also use two additional datasets with vastly different styles: SceneFlow [20] and Spring [21], with the latter being used at half resolution owing to our limitations in computational resources. For the network architectures, we select three representative models with highly distinct designs—PSMNet [1], GMStereo [34], and MonSter [2]—to verify the general applicability of our method across different backbones.

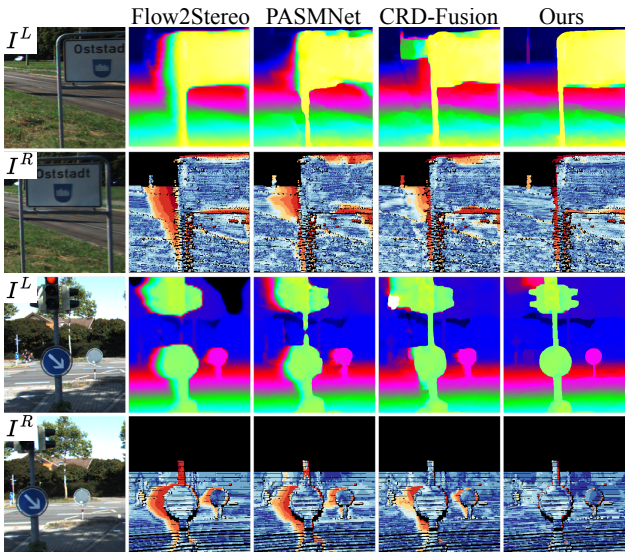

Fig. 6: Qualitative comparisons. All disparity and error maps sourced from the KITTI 2015 online benchmark. Only our method produces accurate results on both sides of occluders, demonstrating that the occlusion problem is resolved.

Training Details. We maintain consistent parameters across different datasets to demonstrate the robustness of our method. All training is conducted using two RTX 4090 GPUs in PyTorch [25]. We employ the Adam optimizer [15] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and the cosine annealing schedule. Following the settings in [1], we crop images to 512×256 during training, with a batch size of 8. Data augmentation is applied to the input images but not to the feedback images. This augmentation consists of random cropping, brightness enhancement, and contrast adjustment. All training processes start from scratch without any pre-training, ground-truth labels, or post-processing involved.

Computational Overhead. The pseudo-image generation requires only a single gradient-free forward pass of the network, introducing no additional GPU memory consumption. The total training time increases by approximately 25%. During inference, the pseudo-image generation is not used, so the computational cost is identical to the baseline network.

4.1 Comparisons with State-of-the-art

We first present our submitted results on the KITTI online leaderboard in Tab. 1. The evaluation metrics are identical to those in the online benchmarks. Specifically, "noc" and "all" denote non-occluded pixels and all pixels respectively. "n-noc/all" refers to the percentage of pixels exceeding a difference of "n" from ground truth. "D1" indicates the proportion of pixels where the end-point error is more than 3px and 5%. "EPE" represents the average difference between estimated values and ground truth.

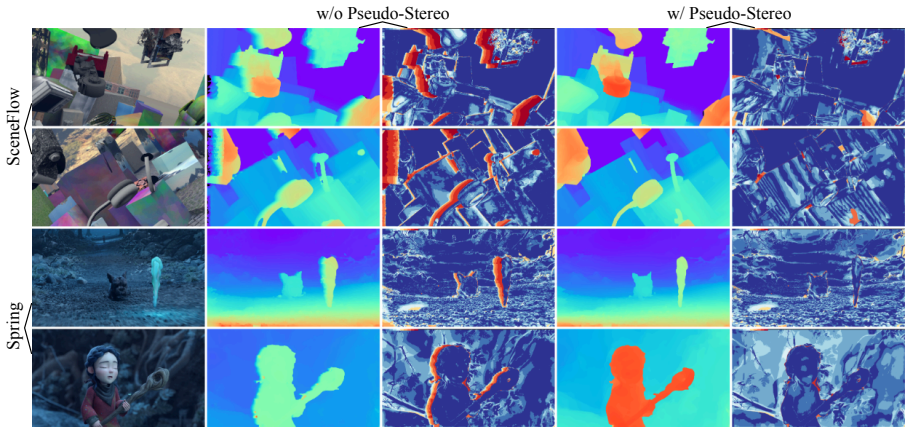

Fig. 7: Qualitative results of our proposed Pseudo-Stereo method on the SceneFlow [20] and Spring [21] datasets, demonstrating its ability to resolve the occlusion challenge across various scenarios.

The results in Tab. 1 demonstrate that our method significantly outperforms other methods across the board. We present results using two distinct back-

Table 2: Evaluation on various backbones and datasets.

Backbone	w/o Pseudo-Stereo		w/ Pseudo-Stereo	
	EPE (px) ↓	3-px (%) ↓	EPE (px) ↓	3-px (%) ↓
SceneFlow [20] Dataset				
PSMNet [1]	5.41	16.85	**3.10**	**9.80**
GMStereo [34]	5.09	15.01	**3.11**	**8.49**
MonSter [2]	5.28	16.15	**2.32**	**7.86**
Spring [21] Dataset				
PSMNet [1]	2.41	6.58	**2.30**	**5.12**
GMStereo [34]	2.33	6.31	**2.18**	**4.78**
MonSter [2]	2.37	6.56	**1.77**	**4.65**

bones: PSMNet [1] and MonSter [2]. The former is chosen to ensure that the performance gains are attributable to our method rather than to advancements in network architecture, thereby validating the credibility of our contribution. The latter validates the generalizability and applicability of our method to the latest architectures. These performance improvements are attributed to our successful resolution of the occlusion problem, which is visually demonstrated in Fig. 6. The figure presents representative occlusion scenarios alongside the disparity and error maps from different methods. It is clearly visible that other methods exhibit varying degrees of error on the left side of the occluder, which indicates that the occlusion problem remained unresolved in these prior works. In contrast, only our method produces results that are symmetric and accurate on both sides of the occluder. This validates that the primary objective of this work—to mitigate the occlusion challenge—has been largely accomplished.

4.2 Evaluation on Various Backbones and Datasets

This section aims to demonstrate the robustness and effectiveness of our approach across various backbone architectures and datasets. As shown in Fig. 7, the occlusion problem is prevalent across various datasets, manifesting as erroneous disparity inference on the left side of foreground occluders (indicated by red pixels in the error maps). The proposed Pseudo-Stereo strategy demonstrates clear effectiveness in these occluded regions, yielding accurate results on both sides of the occluder. This qualitative improvement translates to quantitative gains, as shown in Tab. 2, where applying our method to different backbones on various datasets all leads to substantial accuracy enhancements attributable to solving the occlusion problem.

4.3 Ablation Study

In this section, the roles of the various components proposed in Sec. 3 within the overall framework are studied and discussed. All evaluations are performed on the KITTI 2015 training set. The components involved include the pseudo-stereo

Table 3: Ablation study of our proposed components. PS: Pseudo-stereo inputs strategy. Occ: Occlusion estimation and gradient detaching. RG: Refined pseudo-image generation. FPS: Fully pseudo-stereo inputs strategy.

Config.	PS	Occ	RG	FPS	D1 (%)	EPE	Remarks
(a)					5.67	1.17	/
(b)		✓			5.73	1.24	Over-smoothing in occlusions
(c)	✓				4.99	1.07	Overfitting
(d)	✓	✓			4.08	0.99	Overfitting
(e)	✓			✓	5.17	1.11	/
(f)	✓	✓		✓	4.31	1.04	/
(g)	✓	✓	✓	✓	4.06	1.01	/

Table 4: Ablation study of the naive flip baseline and occlusion-region metrics. Occlusion regions are identified using the rendering-based method described in Sec. 3.2.

Config	D1-all	EPE-all	D1-occ	EPE-occ	D1-noc	EPE-noc
base	5.67	1.17	20.82	3.24	4.37	1.01
naive flip	5.61	1.14	20.77	3.20	4.34	1.00
ours	4.06	1.01	9.05	1.36	3.68	0.97

inputs strategy, the refined pseudo-image generation method, the occlusion estimation, and the fully pseudo-stereo inputs strategy. The experimental results are shown in Tab. 3.

The results for settings (a-d) indicate that both the pseudo-stereo inputs and occlusion estimation are essential for performance enhancement, but they concurrently induce overfitting. Conversely, the results from (e-g), where overfitting is absent, prove that our fully pseudo-stereo input strategy is the key to resolving the overfitting problem. In Fig. 8, we plot the training curves to visually illustrate this phenomenon. The comparison between (f) and (g) reveals that when the generated pseudo-image suffers from significant artifacts, adopting the fully pseudo-stereo strategy can lead to a degradation in accuracy. This issue is effectively resolved when our refined pseudo-image generation strategy is used. Overall, each component plays an indispensable role within the entire framework. The combination of all components enables our model to achieve both high accuracy and stable training.

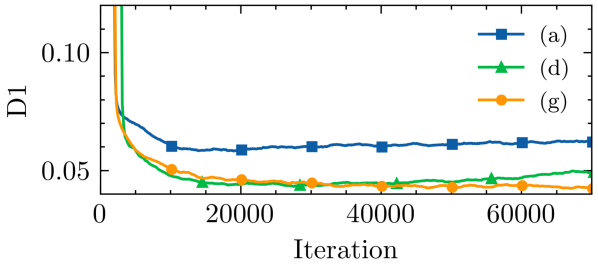

Fig. 8: Visualization of the overfitting problem. By adopting the fully pseudo-stereo input strategy in (f), the increasing D1 error trend observed in (c) is eliminated.

Table 4 provides further experiments to quantitatively demonstrate that the naive flip in Fig. 1 is ineffective for resolving occluded regions. The naive flip configuration yields nearly identical results to the baseline (5.61 vs. 5.67 D1-all), confirming that simple flipping cannot resolve the fixed-side occlusion problem. In contrast, our method substantially reduces occlusion errors: D1-occ from 20.82% to 9.05% and EPE-occ from 3.24 to 1.36 px.

5 Conclusion

In this work, we propose a self-supervised stereo matching paradigm with the core idea of a pseudo-stereo inputs strategy. This method provides a solution to the occlusion issue, which has consistently hindered the performance of self-supervised stereo matching. Extensive experiments validate our method's success in handling occlusions and demonstrate its robustness and generality across diverse datasets and network backbones.

Despite our significant progress on occlusions, this work does not propose explicit solutions for other challenges, such as textureless and reflective regions. In fact, unlike occlusions, providing accurate feedback for these ill-posed regions is inherently intractable with binocular cues alone. A promising solution is the multi-frame self-supervised paradigm, as explored in [32]. This creates a powerful synergy: temporal consistency across frames is key to resolving ambiguities in areas like textureless regions, while the accuracy of the single-frame disparity estimation is core to resolving depth ambiguities of moving objects. The substantially enhanced reliability of disparity estimation achieved by our work provides a much stronger foundation for this paradigm, creating significant potential to better leverage multi-frame information and tackle the remaining challenges in self-supervised 3D perception.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62101547, the Liaoning Provincial Science and Technology Program Project under Grant 2025JH2/101330119, and the Fundamental Research Project of SIA under Grant 2025JC3K01.

References

1. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5410–5418 (2018)
2. Cheng, J., Liu, L., Xu, G., Wang, X., Zhang, Z., Deng, Y., Zang, J., Chen, Y., Cai, Z., Yang, X.: MonSter: Marry Monodepth to Stereo Unleashes Power. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6273–6282 (Jun 2025)

3. Fan, X., Jeon, S., Fidan, B.: Occlusion-Aware Self-Supervised Stereo Matching with Confidence Guided Raw Disparity Fusion. In: 2022 19th Conference on Robots and Vision (CRV). pp. 132–139 (May 2022). https://doi.org/10.1109/CRV55824.2022.00025
4. Fang, I., Wen, H.C., Hsu, C.L., Jen, P.C., Chen, P.Y., Chen, Y.S., et al.: ES3Net: Accurate and efficient edge-based self-supervised stereo matching network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4472–4481 (2023)
5. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361. IEEE, Providence, RI (Jun 2012). https://doi.org/10.1109/CVPR.2012.6248074
6. Godard, C., Aodha, O.M., Brostow, G.J.: Unsupervised Monocular Depth Estimation with Left-Right Consistency. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6602–6611. IEEE, Honolulu, HI (Jul 2017). https://doi.org/10.1109/CVPR.2017.699
7. Godard, C., Aodha, O.M., Firman, M., Brostow, G.: Digging Into Self-Supervised Monocular Depth Estimation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3827–3837. IEEE, Seoul, Korea (South) (Oct 2019). https://doi.org/10.1109/ICCV.2019.00393
8. Hamid, M.S., Abd Manap, N., Hamzah, R.A., Kadmin, A.F.: Stereo matching algorithm based on deep learning: A survey. Journal of King Saud University-Computer and Information Sciences **34**(5), 1663–1673 (May 2022). https://doi.org/10.1016/j.jksuci.2020.08.011
9. Hirschmuller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 2, pp. 807–814 vol. 2 (Jun 2005). https://doi.org/10.1109/CVPR.2005.56
10. Hur, J., Roth, S.: Self-Supervised Multi-Frame Monocular Scene Flow. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2683–2693. IEEE, Nashville, TN, USA (Jun 2021). https://doi.org/10.1109/CVPR46437.2021.00271
11. Janai, J., Guney, F., Ranjan, A., Black, M., Geiger, A.: Unsupervised learning of multi-frame optical flow with occlusions. In: Computer Vision — ECCV 2018. LNCS, vol. 11213, pp. 713–731. Springer (2018). https://doi.org/10.1007/978-3-030-01270-0_42
12. Jonschkowski, R., Stone, A., Barron, J.T., Gordon, A., Konolige, K., Angelova, A.: What matters in unsupervised optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 557–572. Springer (2020)
13. Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics **42**(4), 1–14 (Aug 2023). https://doi.org/10.1145/3592433
14. Khan, N., Xiao, L., Lanman, D.: Tiled Multiplane Images for Practical 3D Photography. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10420–10430. IEEE, Paris, France (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.00959
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)

16. Li, A., Yuan, Z.: Occlusion aware stereo matching via cooperative unsupervised learning. In: Computer Vision — ACCV 2018. LNCS, vol. 11361, pp. 197–213. Springer (2018)
17. Li, A., Yuan, Z., Ling, Y., Chi, W., Zhang, S., Zhang, C.: Unsupervised occlusion-aware stereo matching with directed disparity smoothing. IEEE Transactions on Intelligent Transportation Systems **23**(7), 7457–7468 (Jul 2022). https://doi.org/10.1109/TITS.2021.3070403
18. Li, K., Wang, L., Zhang, Y., Xue, K., Zhou, S., Guo, Y.: LoS: Local structure-guided stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19746–19756 (2024)
19. Liu, P., King, I., Lyu, M.R., Xu, J.: Flow2Stereo: Effective Self-Supervised Learning of Optical Flow and Stereo Matching. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6647–6656 (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.00668
20. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4040–4048. IEEE, Las Vegas, NV, USA (Jun 2016). https://doi.org/10.1109/CVPR.2016.438
21. Mehl, L., Schmalfuss, J., Jahedi, A., Nalivayko, Y., Bruhn, A.: Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4981–4991. IEEE, Vancouver, BC, Canada (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.00482
22. Meister, S., Hur, J., Roth, S.: UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7251–7259 (2018). https://doi.org/10.1609/AAAI.V32I1.12276
23. Menze, M., Geiger, A.: Object Scene Flow for Autonomous Vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3061–3070 (2015)
24. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision (ECCV). pp. 405–421. Springer (2020)
25. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 32, pp. 8024–8035 (2019)
26. Poggi, M., Tosi, F., Batsos, K., Mordohai, P., Mattoccia, S.: On the Synergies between Machine Learning and Binocular Stereo for Depth Estimation from Images: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 5241–5262 (2022). https://doi.org/10.1109/TPAMI.2021.3070917
27. Tankovich, V., Hane, C., Zhang, Y., Kowdle, A., Fanello, S., Bouaziz, S.: HITNet: Hierarchical Iterative Tile Refinement Network for Real-time Stereo Matching. In: 2021 Ieee/Cvf Conference on Computer Vision and Pattern Recognition, Cvpr 2021. pp. 14357–14367 (2021). https://doi.org/10.1109/CVPR46437.2021.01413
28. Tosi, F., Bartolomei, L., Poggi, M.: A Survey on Deep Stereo Matching in the Twenties. International Journal of Computer Vision **133**(7), 4245–4276 (2025). https://doi.org/10.1007/s11263-024-02331-0
29. Tosi, F., Tonioni, A., De Gregorio, D., Poggi, M.: NeRF-Supervised Deep Stereo. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition

(CVPR). pp. 855–866. IEEE, Vancouver, BC, Canada (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.00089
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
31. Wang, L., Guo, Y., Wang, Y., Liang, Z., Lin, Z., Yang, J., An, W.: Parallax attention for unsupervised stereo correspondence learning. IEEE transactions on pattern analysis and machine intelligence **44**(4), 2108–2125 (2020)
32. Wang, Y., Wang, P., Yang, Z., Luo, C., Yang, Y., Xu, W.: UnOS: Unified Unsupervised Optical-Flow and Stereo-Depth Estimation by Watching Videos. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8063–8073. IEEE, Long Beach, CA, USA (Jun 2019). https://doi.org/10.1109/CVPR.2019.00826
33. Wang, Y., Yang, Y., Yang, Z., Zhao, L., Wang, P., Xu, W.: Occlusion Aware Unsupervised Learning of Optical Flow. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4884–4893. IEEE, Salt Lake City, UT (Jun 2018). https://doi.org/10.1109/CVPR.2018.00513
34. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Yu, F., Tao, D., Geiger, A.: Unifying flow, stereo and depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(11), 13941–13958 (2023). https://doi.org/10.1109/TPAMI.2023.3298645
35. Yang, R., Li, X., Cong, R., Du, J.: Unsupervised Hierarchical Iterative Tile Refinement Network With 3D Planar Segmentation Loss. IEEE Robotics and Automation Letters **9**(3), 2678–2685 (Mar 2024). https://doi.org/10.1109/LRA.2024.3359545
36. Yuan, W., Zhang, Y., Wu, B., Zhu, S., Tan, P., Wang, M.Y., Chen, Q.: Stereo matching by self-supervision of multiscopic vision. arXiv preprint arXiv:2104.04170 (2021). https://doi.org/10.48550/ARXIV.2104.04170
37. Zhou, C., Zhang, H., Shen, X., Jia, J.: Unsupervised Learning of Stereo Matching. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 1576–1584. IEEE, Venice (Oct 2017). https://doi.org/10.1109/ICCV.2017.174