engineering proceedings

MDPI

Proceeding Paper

Binocular Stereo Vision Disparity Estimation Based on Distilled Internally Normalized Optimized Version 2 with Multi-Scale Attention Fusion †

Chang-Fu Hung, Tzu-Jung Tseng and Jian-Jiun Ding *

Graduate Institute of Communication Engineering, National Taiwan University, Taipei 106319, Taiwan; r11942169@ntu.edu.tw (C.-F.H.); d09942007@ntu.edu.tw (T.-J.T.)

* Correspondence: jjding@ntu.edu.tw; Tel.: +886-2-33669652

† Presented at the 7th Eurasia Conference on IoT, Communication and Engineering 2025 (ECICE 2025), Yunlin, Taiwan, 14–16 November 2025.

Abstract

A stereo vision framework is designed to improve disparity estimation in occluded and boundary regions, targeting autonomous driving scenarios. The proposed architecture combines frozen Distilled Internally Normalized Optimized Version 2 features with a modular three-stage attention fusion strategy, which consists of bottom-up semantic propagation, top-down detail enhancement, and cross-view attention mechanisms. These stages jointly enforce semantic consistency, structural integrity, and accurate correspondence modeling. The fused features are then processed by an Iterative Geometry Encoding and Volumetric regression-based disparity estimation module for multi-stage regression and iterative refinement. A three-phase training pipeline is employed, including pretraining on SceneFlow, fine-tuning on virtual Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) benchmarks, and adaptation to the KITTI and ETH Zurich 3D benchmark dataset. The model achieves an out-of-center, non-occluded pixel error of 7.45% on KITTI2012 and a D1-all error of 4.10% on KITTI2015. Beyond quantitative performance, the proposed method produces visually superior disparity maps. The enhancements of boundary sharpness, occlusion completion, and structural coherence demonstrate the strong potential of the proposed algorithm for real-world deployment in dynamic and complex environments.

Keywords: stereo matching; disparity estimation; DINOv2; attention fusion; multi-level feature integration

1. Introduction

Stereo vision estimates depth by computing disparities between rectified left–right image pairs [1–3]. Compared to active sensors such as Light Detection and Ranging, it offers higher resolution, lower cost, and seamless integration with visual pipelines, making it ideal for applications in autonomous driving and augmented/virtual reality [1]. While recent learning-based approaches such as the Pyramid Stereo Matching Network (PSMNet) [4] and the group-wise architecture [5] have improved robustness in challenging scenarios, accurate disparity estimation remains difficult in occluded and low-texture regions. Vision Transformers (ViTs), particularly Distilled Internally Normalized Optimized Version 2 (DINOv2) [6], have demonstrated strong semantic representation and generalization capabilities through self-supervised learning. However, their single-resolution

Check for updates

Academic Editors: Teen-Hang Meen, Chi-Ting Ho and Cheng-Fu Yang

Published: 31 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

outputs and semantics-oriented design often lack the fine-grained geometric detail required for pixel-level correspondence in stereo matching. To overcome this limitation, we present a hybrid stereo matching framework that integrates frozen DINOv2 features with a multi-scale attention fusion strategy. As illustrated in Figure 1, the developed architecture comprises (1) multi-level feature extraction from selected ViT layers, (2) hierarchical fusion modules that integrate semantic and structural information across scales, and (3) a cost-volume-based disparity refinement module (IGEV). This design enables effective semantic-geometric integration and improves disparity prediction in structurally complex regions. In the developed framework, a multi-scale attention fusion framework was optimized for ViT features, and a frozen foundation model (DINOv2) was integrated into stereo estimation without fine-tuning. The framework improved the visual and geometric quality on the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) benchmarks.

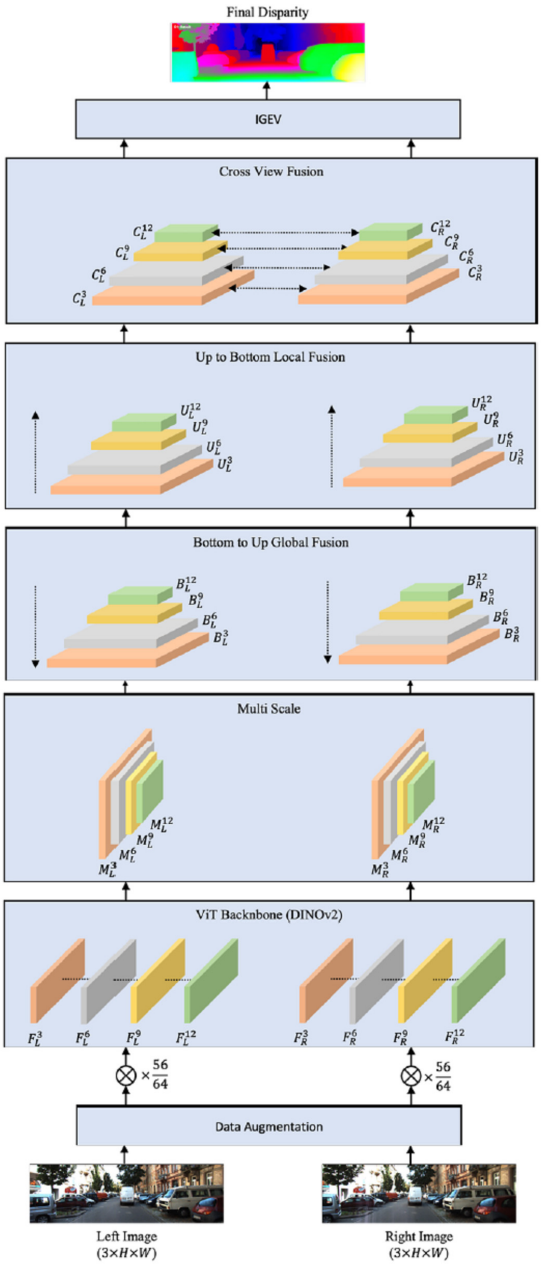


Figure 1. Overall module architecture. Lines indicate the corresponding locations and arrows indicate the relations between inputs and outputs.

https://doi.org/10.3390/engproc2026134020

2. Related Work

Stereo vision estimates 3D structure from rectified image pairs by computing disparity, and plays a key role in applications such as autonomous driving, robotics, and AR/VR. The classical four-stage pipeline—feature extraction, cost volume construction, aggregation, and disparity regression, proposed by Scharstein and Szeliski [1]—remains the foundation for continued advances in accuracy and generalization.

2.1. Traditional Methods

Traditional stereo methods used hand-crafted features and local similarity metrics (sum of absolute differences and normalized cross-correlation), but performed poorly in low-texture or occluded regions. Global approaches such as graph cuts [7] and belief propagation [8] improved accuracy through energy minimization but incurred high computational cost. Semi-global matching (SGM) [9] later became a widely adopted trade-off between accuracy and efficiency.

2.2. CNN-Based Methods

Convolutional neural networks (CNNs) have enabled end-to-end stereo matching by learning both feature representations and matching costs directly from data. Geometry and Context Network [3] pioneered the use of 3D convolutions on cost volumes, while PSMNet [4] incorporated spatial pyramid pooling to capture multi-scale context. More recently, Attention Concatenation Volume Network [10] employed occlusion-aware multi-branch cost volumes, achieving further gains in challenging regions.

2.3. Geometry-Guided Methods

Recent methods enhance geometric reasoning by incorporating stereo priors and physical constraints. Guided Aggregation Network [11] employed SGM-inspired guided aggregation, while IGEV [12] constructed a geometry encoding volume with iterative updates. These approaches improve disparity accuracy and consistency, particularly around boundaries and in occluded regions.

2.4. Transformer-Based Methods

Transformers excel at modeling long-range dependencies. Stereo Transformer [13] learned direct correspondences without cost volumes, while Vision Transformer for Attention-based Stereo matching [14] leveraged frozen DINOv2 [6] features with hierarchical attention for multi-scale fusion. Yet, transformer-derived features often lack spatial detail, requiring supplementary modules for geometry-aware tasks.

2.5. Summary and Positioning

Traditional and CNN-based methods focus on local consistency and handcrafted cues, whereas transformer-based models emphasize semantics but lack geometric precision. Geometry-guided methods balance both through learned priors. Leveraging these insights, our framework integrates frozen DINOv2 features with multi-scale attention fusion and employs IGEV for geometry-guided refinement, yielding a unified and robust stereo vision solution.

3. Methodology

While ViTs [6] excel at capturing global semantics, their lack of spatial precision hinders disparity estimation. To overcome this, we adopt a hybrid framework combining frozen DINOv2 features, multi-scale attention fusion, and IGEV-based refinement. As illustrated in Figure 1, the model extracts features from the 3rd, 6th, 9th, and 12th ViT

https://doi.org/10.3390/engproc2026134020

layers, aligns them spatially, fuses semantic and structural cues via bidirectional attention, and performs disparity prediction using the IGEV module [12].

3.1. ViT Backbone

We adopt DINOv2-ViT-Base as the frozen backbone for feature extraction, offering a strong balance between semantic capability and efficiency. Trained via self-supervised learning, it achieves robust semantic alignment without labeled data. As shown in Table 1, ViT-Base provides deeper semantics than ViT-Small while being significantly lighter than ViT-Large, making it well-suited for our stereo framework.

Table 1. DINOv2 model variants and specifications.

Model	Number of Layers	Embedding Dimension	Number of Parameters
ViT-Small	12	384	~22 M
ViT-Base	12	768	~87 M
ViT-Large	24	1024	~304 M

While ViT features provide strong semantic representations, they lack explicit geometric correspondence. To capture hierarchical semantics, we extract features from the 3rd, 6th, 9th, and 12th layers of DINOv2-ViT-Base—denoted as $F^3_{L,R}$, $F^6_{L,R}$, $F^9_{L,R}$, $F^{12}_{L,R}$—each with 768 channels. These layers progressively encode fine textures and edges, local structures, object-level regions, and global scene abstraction. To make them suitable for stereo matching, we apply channel projection, resolution alignment, and directional attention fusion, transforming semantic features into geometry-aware representations.

3.2. Multi-Scale Module

To enable hierarchical fusion across ViT layers, a multi-scale preprocessing module reshapes fixed-resolution features into a spatial pyramid. As shown in Figure 2, features from the 3rd, 6th, 9th, and 12th layers of DINOv2-ViT-Base are projected to 256 channels via 1×1 convolution for dimensional consistency and efficiency. They are then resized according to semantic depth: $\frac{1}{4}H \times \frac{1}{4}W$, $\frac{1}{8}H \times \frac{1}{8}W$, $\frac{1}{16}H \times \frac{1}{16}W$, and $\frac{1}{32}H \times \frac{1}{32}W$, with 3×3 convolution applied to mitigate spatial degradation. Residual connections with batch normalization and ReLU activation preserve semantic cues and gradient flow. The resulting multi-scale outputs serve as inputs to the subsequent attention fusion modules.

3.3. Feature Fusion Mechanisms

While ViTs capture strong global semantics, their features often lack geometric precision, particularly near occlusions and edges. To address this, a three-stage fusion strategy is employed: Bottom-to-Up Global Fusion transfers high-level semantics, Up-to-Bottom Local Fusion adds fine geometric details, and Cross-View Attention Fusion strengthens stereo correspondence. This pipeline integrates semantic, geometric, and cross-view cues for improved disparity estimation.

1. Bottom-to-Up Global Fusion: This module facilitates top-down semantic enhancement by propagating abstract context from deeper to shallower layers, thereby improving feature consistency across scales. Given input features $M^l \in \mathbb{R}^{C \times H_l \times W_l}$ and deeper features $M^{l+3} \in \mathbb{R}^{C \times H_{l+3} \times W_{l+3}}$, the fused output $B^l \in \mathbb{R}^{C \times H_l \times W_l}$ is computed through the following steps:
   (a) Query embedding: M^l is divided into non-overlapping $p \times p$ patches, each flattened into $m^q_i \in \mathbb{R}^{C \cdot p^2}$, and projected into queries.

$$q_i = W_q m^q_i + b_q,\ W_q \in \mathbb{R}^{(Cp^2) \times d},\ q_i \in \mathbb{R}^d \tag{1}$$

All queries are obtained from $Q \in \mathbb{R}^{N_q \times d}$, where $N_q = \frac{H_l}{p} \cdot \frac{W_l}{p}$, W_q is the weight, m_i are the values within the patch, b_q is the offset, C is the scaling factor, q is the output, d is the depth of the output, and $H_l \times W_l$ is the size of each layer.

(b) Key/value embedding: the deeper feature M^{l+3} is flattened into $\mathrm{m}_j^{kv} \in \mathbb{R}^{\mathrm{C}}$, and linearly projected into the following.

$$k_j = W_k m_j^{kv} + b_k,\ v_j = W_v m_j^{kv} \tag{2}$$

yielding key and value matrices $K, V \in \mathbb{R}^{N_{kv} \times d}$.

(c) Multi-head attention: Each head performs the following.

$$Attn_h(Q_h, K_h, V_h) = softmax\left(\frac{Q_h K_h^{\mathrm{T}}}{\sqrt{d_h}}\right) V_h \tag{3}$$

Concatenated results form $Z \in \mathbb{R}^{N_q \times d}$.

(d) Gated attention: To control semantic flow, we compute:

$$g_i = \sigma\left(W_2 \cdot ReLU\left(W_1\left[q_i; \hat{k}_i\right]\right)\right),\ \ g_i \in [0,1]^d \tag{4}$$

where $\hat{k}_i \in \mathbb{R}^{\mathrm{C}}$ is pooled from M^{l+3}, and W_1, W_2 are MLP weights.

(e) Feature reconstruction: The gated output is projected back as follows.

$$\hat{m}_i^q = W_f \hat{z}_i,\ W_f \in \mathbb{R}^{d \times (Cp^2)} \tag{5}$$

The final output is:

$$B^l = DropPath\left(\hat{M}^l\right) + \mathrm{Y} \cdot M^l \tag{6}$$

This mechanism strengthens shallow features with deep semantic cues while maintaining spatial resolution. The overall design is shown in Figure 3. The resulting fused features $B_{L,R}^3$, $B_{L,R}^6$, B_{LR}^9, $B_{L,R}^{12}$ are passed to the Up-to-Bottom fusion stage.

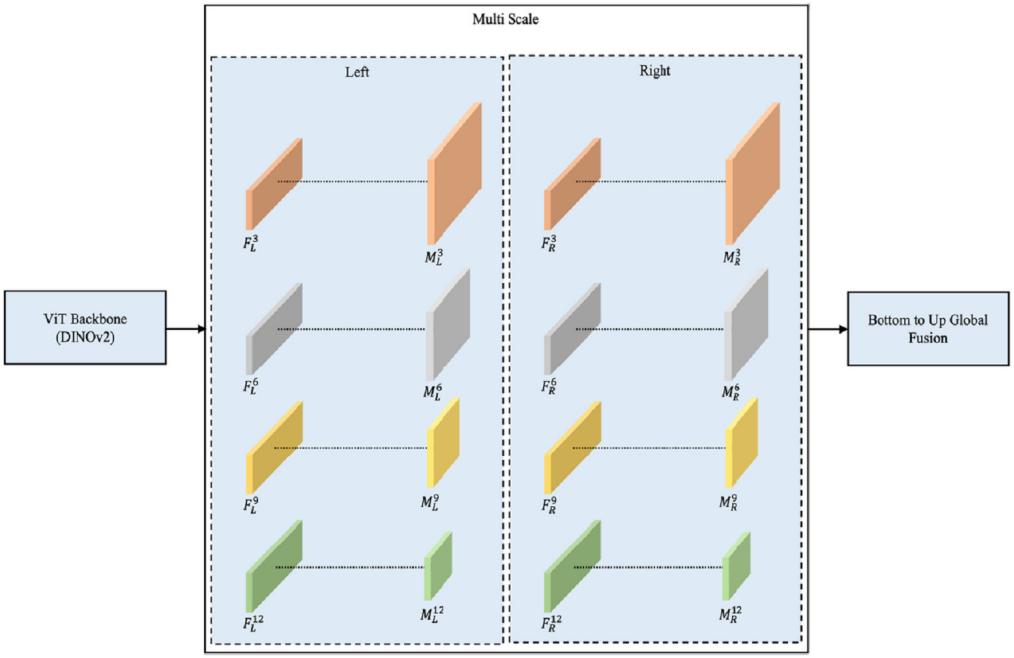


Figure 2. Multi-scale module architecture. Lines indicate the corresponding locations and arrows indicate the relations between inputs and outputs.

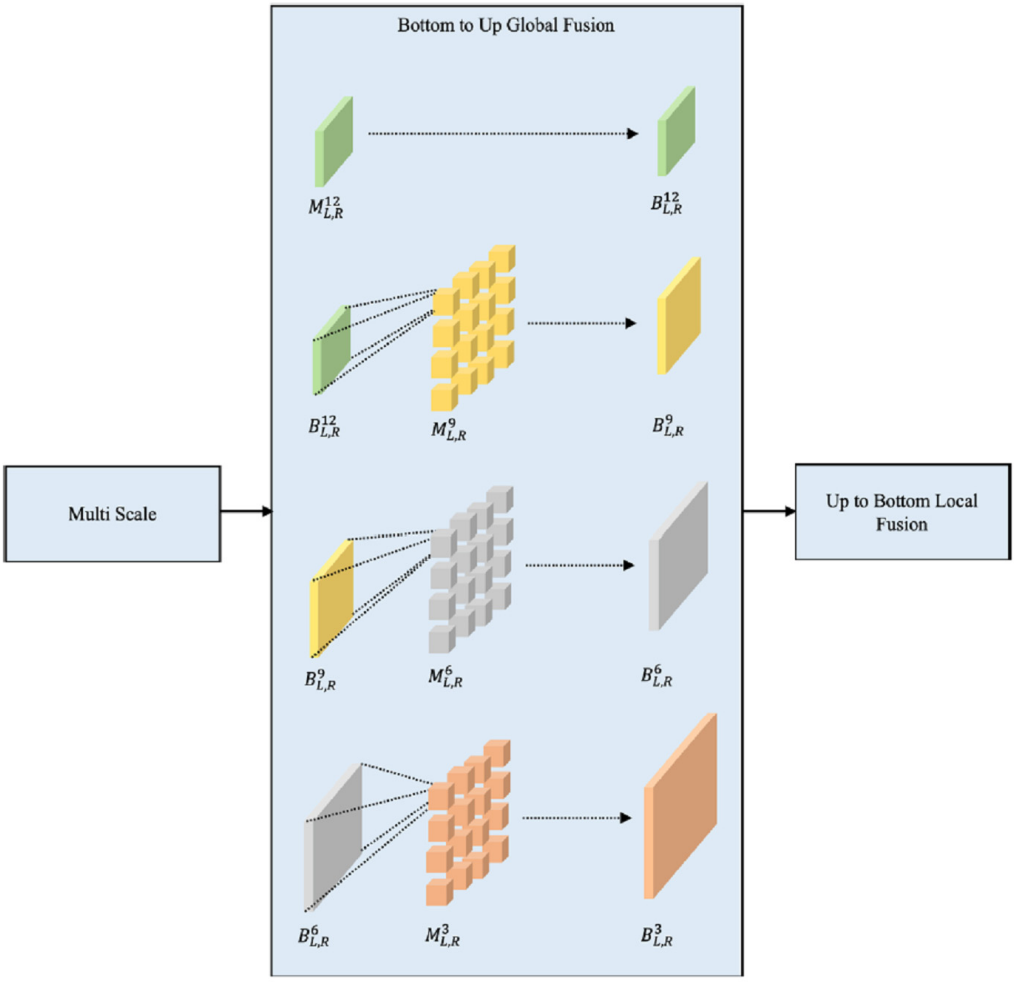


Figure 3. Bottom-to-up global fusion module. Lines indicate the corresponding locations and arrows indicate the relations between inputs and outputs.

2. Up-to-bottom local fusion: To refine coarse semantic features with spatial precision, the Up-to-Bottom Local Fusion module injects geometric details from shallower layers using localized multi-head attention. Given the deeper-layer feature $B_q^l \in \mathbb{R}^{C\times H_l\times W_l}$ and the shallower-layer feature $B^{l-3} \in \mathbb{R}^{C\times H_{l-3}\times W_{l-3}}$, the fused output is $U^l \in \mathbb{R}^{C\times H_l\times W_l}$. The overall architecture is illustrated in Figure 4.

   (a) Query embedding: The input is partitioned into non-overlapping $p \times p$ patches. Each patch $b_i^q \in \mathbb{R}^{C\cdot p^2}$ s projected to the following.

   $$q_i = W_q b_i^q + b_q,\ \ q_i \in \mathbb{R}^d \tag{7}$$

   All queries form $Q \in \mathbb{R}^{N_q\times d}$, where $N_q = \frac{H_l}{p}\cdot\frac{W_l}{p}$.

   (b) Key/value embedding: Shallower-layer tokens are projected as follows.

   $$k_j = W_k b_j^{kv} + b_k,\ v_j = W_v b_j^{kv} + b_v \tag{8}$$

   This yields $K, V \in \mathbb{R}^{N_{kv}\times d}$.

   (c) Local attention: For each q_i, attention is restricted to a local window $\mathbb{N}_i$.

$$Attn_i^h\left(q_i^h, k_j^h, v_j^h\right) = \sum_{j\in\mathbb{N}_i} softmax\left(\frac{q_i^h\left(k_j^h\right)^{\mathrm{T}}}{\sqrt{d_h}}\right)v_j^h. \quad (9)$$

Its outputs are concatenated and projected into $Z \in \mathbb{R}^{N_q\times d}$.

(d) Dynamic gating: Using pooled context $\hat{k}_i$, a gate vector $g_i \in [0,1]^d$ is computed by the following.

$$g_i = \sigma\left(W_2 \cdot ReLU\left(W_1\left[q_i; \hat{k}_i\right]\right)\right),\ \hat{z}_i = g_i \odot z_i \quad (10)$$

(e) Reconstruction: Each $\hat{z}_i$ is mapped back to patch space.

$$\hat{b}_i^q = W_f\hat{z}_i,\ \hat{B}^l \to U^l = DropPath\left(\hat{B}^l\right) + \Upsilon \cdot B^l. \quad (11)$$

This fusion aligns deep semantics with localized structure, improving fine-grained accuracy. The outputs $U_{L,R}^3$, $U_{L,R}^6$, U_{LR}^9, $U_{L,R}^{12}$ are used in the final cross-view fusion stage.

3. Cross-view attention fusion: To enhance semantic consistency and geometric alignment between stereo views, a Transformer-based Cross-View Attention Fusion module is adopted as illustrated in Figure 5. At each scale l, the fused left and right features U_L^l, $U_R^l \in \mathbb{R}^{C\times H_l\times W_l}$ are flattened into sequences.

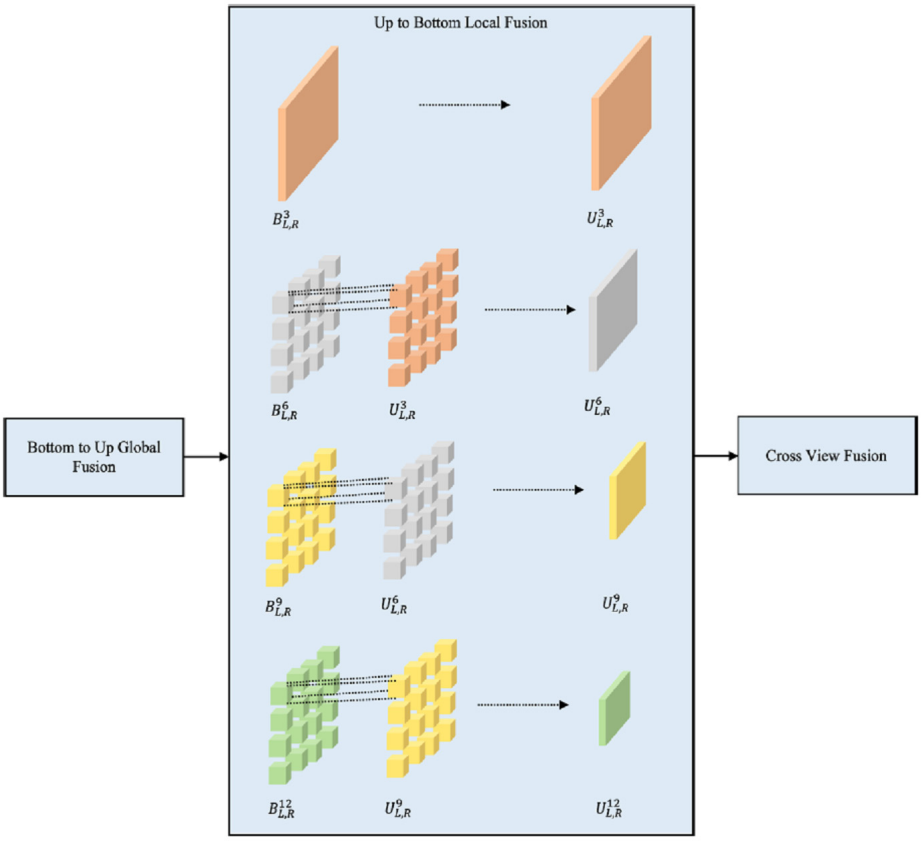


Figure 4. Up-to-bottom local fusion module. Lines indicate the corresponding locations and arrows indicate the relations between inputs and outputs.

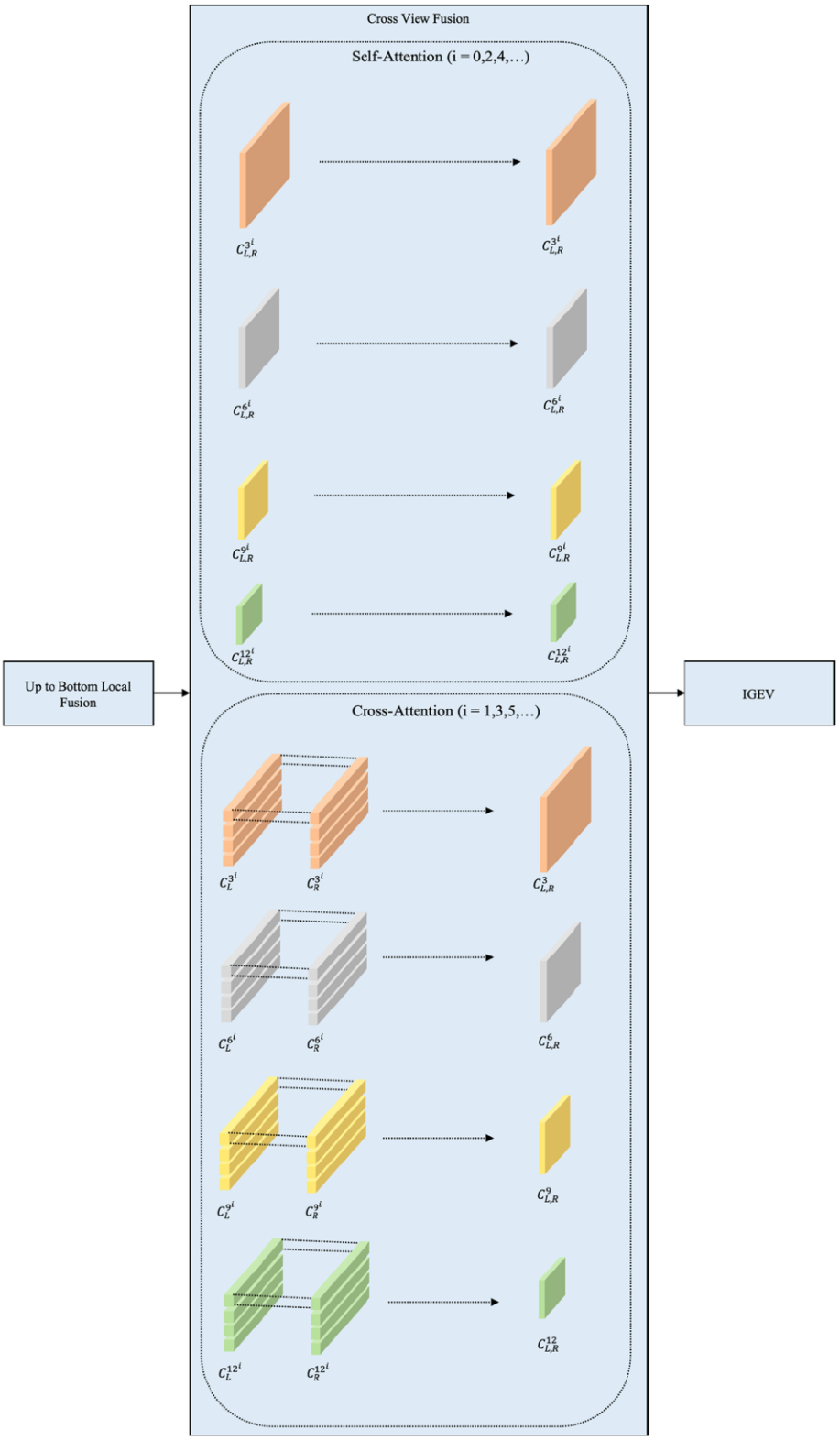


Figure 5. Cross-view fusion module. Lines indicate the corresponding locations and arrows indicate the relations between inputs and outputs.

$$X_{L,R}^{l} = Flatten\left(U_{L,R}^{l}\right) \in \mathbb{R}^{C \times N_l},\ N_l = H_l \times W_l. \tag{12}$$

The module consists of T alternating layers of attention and feedforward blocks. Even-numbered layers apply local self-attention within spatial windows, while odd-numbered

layers perform cross-view attention along the epipolar axis. For even layers, attention is computed within a local window $\omega_x(i)$.

$$\hat{x}^t_{L,i} = \sum_{j\in\omega_x(i)} softmax\left(\frac{q_i(k_j)^{\mathrm{T}}}{\sqrt{d}}\right)v_j \in \mathbb{R}^d, \tilde{x}^t_{L,i} = W_o\hat{x}^t_{L,i} + x^t_{L,i},\ W_o \in \mathbb{R}^{C\times d}. \quad (13)$$

For odd layers, cross-attention uses left-view queries and right-view keys/values.

$$\hat{x}^t_{L,i} = \sum_{j\in\omega_x(i)} softmax\left(\frac{q_i(k_j)^{\mathrm{T}}}{\sqrt{d}}\right)v_j \in \mathbb{R}^d, \tilde{x}^t_{L,i} = W_o\hat{x}^t_{L,i} + x^t_{L,i},\ W_o \in \mathbb{R}^{C\times d}. \quad (14)$$

Each attention output passes through a two-layer feedforward network with a residual connection.

$$z_i = W_2 \cdot RELU\left(W_1\tilde{x}^t_i\right) + \tilde{x}^t_i \in \mathbb{R}^C \quad (15)$$

After T layers, the outputs are reshaped back to spatial form.

$$C^l_{L,R} = Reshape\left(z_{(L,R),1},\ z_{(L,R),2}, \ldots, z_{(L,R),N}\right). \quad (16)$$

The resulting $C^l_{L,R}$ features are used for disparity prediction and refinement via the IGEV module.

3.4. Iterative Geometry Encoding Volume (IGEV)

We adopt the IGEV framework 12 as the disparity estimation backbone for its robustness and accuracy on stereo benchmarks. IGEV constructs a 4D cost volume by warping right-view features with disparity hypotheses and fusing them with left-view features. This improves prediction in occluded and low-texture regions. Without modifying IGEV, we feed in multi-scale features (1/4 to 1/32 resolution, 256 channels) extracted by our front-end. The original geometry encoding, hourglass aggregation, and recurrent refinement modules are retained to ensure consistent and effective disparity estimation.

4. Experiments

To enhance semantic understanding, geometric reasoning, and generalization in stereo matching, a three-stage training pipeline is employed: pretraining, intermediate finetuning, and either target-specific finetuning or direct evaluation. The experimental setup utilizes four benchmark datasets—FlyingThings3D [15], Virtual KITTI (vKITTI) [16], KITTI 2012 [17], and KITTI 2015 [18]—covering synthetic, semi-synthetic, and real-world domains. Their respective roles are summarized in Table 2.

Table 2. Summary of the adopted datasets.

Datasets	Type	Resolution	Role
SceneFlow	Synthetic	960 × 540	Pretraining
vKITTI	Semi-synthetic	1242 × 375	Intermediate Finetuning
KITTI2012	Real	1240 × 376	Target Finetuning & Testing
KITTI2015	Real	1240 × 376	Target Finetuning & Testing

4.1. Fine-Tuning Strategy

The model is trained in three stages: SceneFlow pretraining, vKITTI finetuning, and KITTI adaptation. A selective freezing strategy is applied to shift learning focus from semantic representation to disparity refinement.

- SceneFlow: Freeze DINOv2, train all other modules using FlyingThings3D;
- vKITTI: Freeze DINOv2 and the multi-scale module, train fusion modules and IGEV;

- KITTI: Freeze all modules except IGEV, which is finetuned on KITTI 2012 or 2015.

This progressive scheme improves training stability and enhances domain adaptation.

4.2. Comparison with Other Methods

We compare the performance of the proposed method with that of PSNNet [4], GWCNet [5], IGEV [12], and PWCNet [19]. Visual comparisons in Figures 6–8 show that the proposed method surpasses state-of-the-art approaches, offering sharper object boundaries, improved distant background estimation, and better geometric consistency in occluded areas.

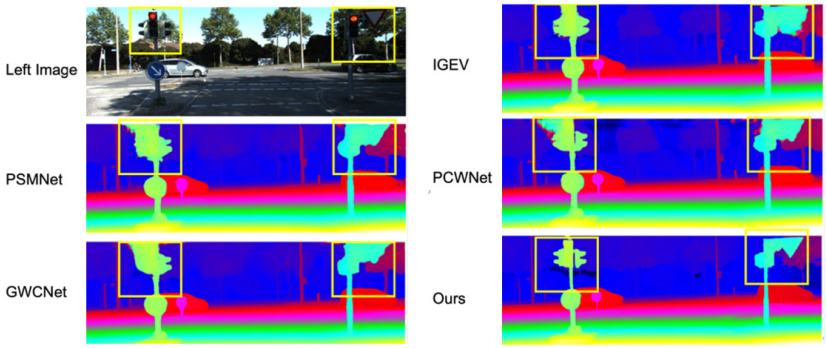


Figure 6. KITTI2015 disparity results comparison. We use different colors to represent different depths. A lighter color means a shorter distance. There are traffic lights on both the left and right sides of the image. Our method preserves the complete structure of these traffic lights more effectively than other approaches, avoiding the common issue of blending with the sky background.

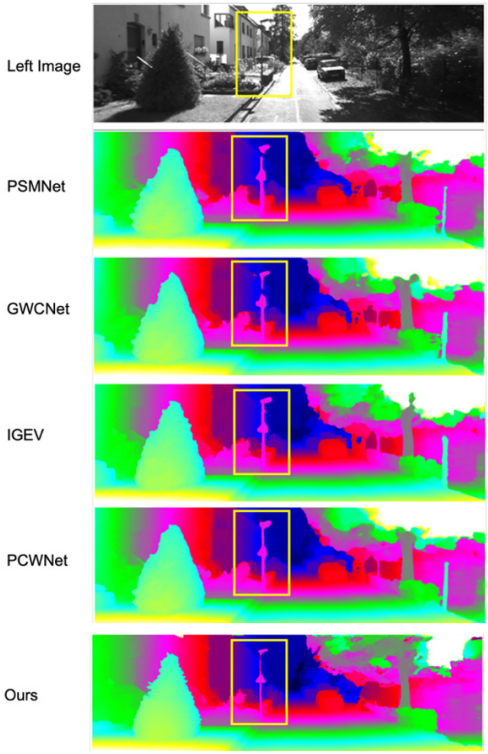


Figure 7. KITTI2012 disparity results comparison. We use different colors to represent different depths. A lighter color means a shorter distance. This example highlights the superior boundary refinement achieved by our method compared to other approaches. Our results show finer detail across multiple objects, such as utility poles, where the predicted shapes more closely match the original dimensions and maintain clearer edges.

https://doi.org/10.3390/engproc2026134020

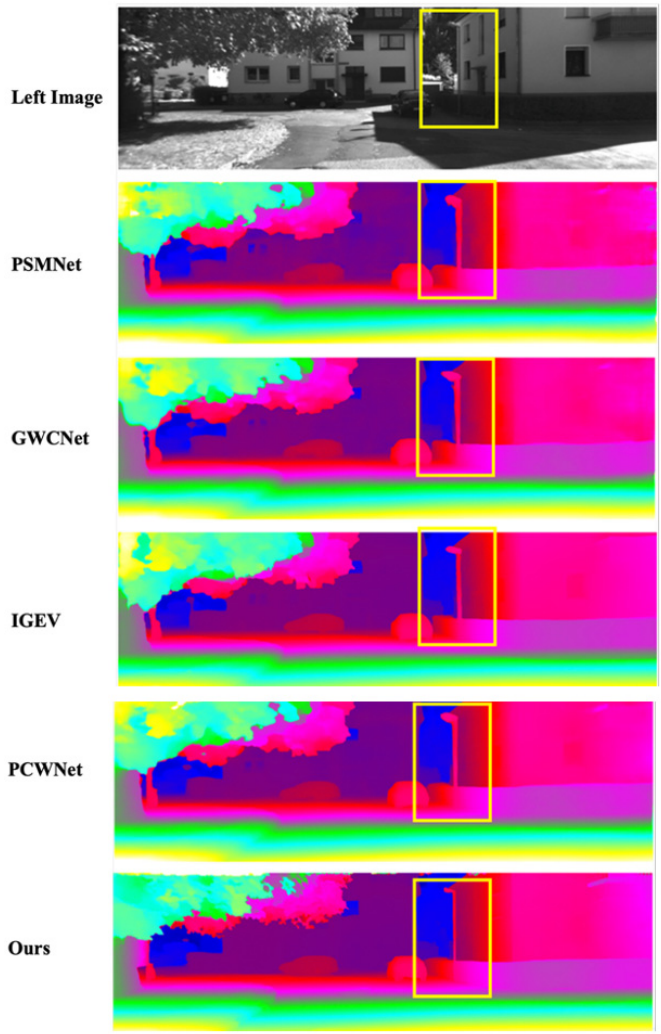


Figure 8. KITTI2012 disparity results comparison. We use different colors to represent different depths. A lighter color means a shorter distance. The reconstructed streetlights in our disparity map also exhibit better alignment with the ground truth in both size and structural detail.

5. Conclusions

We developed a stereo disparity estimation framework that adapts high-level semantic features from a pretrained ViT into geometry-aware representations using a three-stage fusion strategy: bottom-to-up global fusion, up-to-bottom local fusion, and cross-view attention. Combined with a multi-stage training scheme on synthetic and real-world datasets, the method improved accuracy and structural consistency on KITTI benchmarks, particularly in occluded and boundary regions. Despite these gains, performance degradation on the Middlebury and ETH Zurich 3D benchmark dataset highlights the need for enhanced cross-domain generalization. The results confirm the feasibility of integrating frozen ViT features into stereo vision through semantic-to-geometric fusion, offering a promising direction for future research.

Author Contributions: Conceptualization, C.-F.H. and J.-J.D.; methodology, C.-F.H.; software, C.-F.H.; validation, C.-F.H.; formal analysis, C.-F.H.; investigation, C.-F.H.; resources, C.-F.H.; data curation, C.-F.H.; writing—original draft preparation, C.-F.H.; writing—review and editing, T.-J.T. and J.-J.D.; visualization, C.-F.H. and T.-J.T.; supervision, J.-J.D.; project administration, J.-J.D.; and funding acquisition, J.-J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science and Technology Council, Taiwan, under the contract of NSTC 114-2221-E-002-122-MY2.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Scharstein, D.; Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* **2002**, *47*, 7–42. [CrossRef]
2. Geiger, A.; Roser, M.; Urtasun, R. Efficient large-scale stereo matching. In *Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 25–38.
3. Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; Bry, A. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2017; pp. 66–75.
4. Chang, J.R.; Chen, Y.S. Pyramid stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 5410–5418.
5. Guo, X.; Yang, K.; Yang, W.; Wang, X.; Li, H. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2019; pp. 3273–3282.
6. Ayzenberg, L.; Raja, G.; Hayit, G. DINOv2 based self supervised learning for few shot medical image segmentation. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*; IEEE: New York, NY, USA, 2024; pp. 1–5.
7. Kolmogorov, V.; Zabih, R. Computing visual correspondence with occlusions using graph cuts. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2001; Volume 2, pp. 508–515.
8. Yang, Q.; Wang, L.; Yang, R.; Wang, S.; Liao, M.; Nister, D. Real-time global stereo matching using hierarchical belief propagation. In *Proceedings of the British Machine Vision Conference*; BMVA Press: Durham, UK, 2006; Volume 6, pp. 989–998.
9. Hirschmuller, H. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2005; Volume 2, pp. 807–814.
10. Xu, G.; Cheng, J.; Guo, P.; Yang, X. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2022; pp. 12981–12990.
11. Zhang, F.; Prisacariu, V.; Yang, R.; Torr, P.H. Ga-Net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2019; pp. 185–194.
12. Xu, G.; Wang, X.; Ding, X.; Yang, X. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2023; pp. 21919–21928.
13. Li, Z.; Liu, X.; Drenkow, N.; Ding, A.; Creighton, F.X.; Taylor, R.H.; Unberath, M. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2021; pp. 6197–6206.
14. Liu, C.W.; Chen, Q.; Fan, R. Playing to vision foundation model's strengths in stereo matching. *IEEE Trans. Intell. Veh.* 2024. *Early Access*. [CrossRef]
15. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 4040–4048.
16. Gaidon, A.; Wang, Q.; Cabon, Y.; Vig, E. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 4340–4349.
17. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2012; pp. 3354–3361.
18. Menze, M.; Geiger, A. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2015; pp. 3061–3070.
19. Shen, Z.; Dai, Y.; Song, X.; Rao, Z.; Zhou, D.; Zhang, L. PCW-Net: Pyramid combination and warping cost volume for stereo matching. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 280–297.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.