Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 11, No. 3, 2025

Research on the Fast-GwcNet 3D Reconstruction Method for Crime Scenes

Yao Zhou, Lixin Zhao and Fanliang Bu *

China People's Public Security University of China Beijing, China
*** Corresponding author:** Fanliang Bu (Email: bufanliang@sina.com)

Abstract: The 3D reconstruction of the case scene is a key technology for judicial evidence collection and physical evidence analysis, which restores the details of the 3D scene of the x scene through high-precision stereo matching, which significantly surpasses the clue mining ability of 2D images. However, traditional algorithms are prone to matching ambiguity in complex scene environments (such as weakly textured physical evidence, dynamic occlusion, and non-Lambert surfaces), and at the same time, to meet the needs of accuracy and efficiency of 3D reconstruction of case scenes, this paper designs a lightweight stereo matching algorithm based on the GwcNet basic model, in which the S2A attention mechanism is introduced to improve the feature extraction, the MPM module and a new 3D convolution fusion multi-level convolution fusion with cost filtering cost are used. At the same time, the SAFM module is used to fuse the multi-level disparity map of parallax prediction, so that the matching accuracy and running time of the experimental results of the KITTI2012 and KITTI2015 datasets can be improved, and the running time is reduced, which has significant judicial application value.

Keywords: Crime Scenes; Fast-GwcNet 3D Reconstruction Method; S2A Attention Mechanism; MPM Module.

1. Introduction

With the rapid development of three-dimensional vision technology and deep learning, stereo matching, as a core component of 3D reconstruction, has demonstrated significant value in fields such as autonomous driving [1-2], virtual reality [3-4], intelligent security[5], and more. Especially in the scene of judicial evidence collection and case scene restoration, high-precision 3D reconstruction technology is not only a key tool to restore the truth, but also a technical guarantee to maintain social fairness and justice. The traditional stereo matching method relies on artificially designed features (such as SIFT and Census transform) and full semi-global and local optimization strategies, which is easy to produce matching ambiguity under weak texture, occlusion and complex lighting conditions, and is difficult to meet the needs of accuracy and robustness in the case scene. In recent years, breakthroughs in deep learning technology have brought innovative methods to three-dimensional matching, and a series of end-to-end networks have significantly improved the matching accuracy and efficiency in complex scenarios through multi-level feature learning and cost optimization.

Early stereo matching methods based on deep learning (such as DispNet[6]) created a precedent for end-to-end parallax prediction, in which the encoder-decoder architecture directly regressed the parallax map through multi-scale feature fusion, avoiding the complex post-processing process of traditional methods, and making a breakthrough in real-time. However, due to the lack of cost volume modeling, such methods are difficult to meet the challenges of occlusion and repetitive textures. To this end, GCNet[7] introduces 4D cost volume (spatial × parallax dimension) and 3D convolution aggregation strategy for the first time to capture more comprehensive information in spatial and parallax dimensions, combines geometric constraints with semantic contexts, and improves the understanding and matching ability of many complex scenes and geometric structures. After that, PSMNet[8] fuses multi-scale global features through the Spatial Pyramid Pooling (SPP) module to reduce the matching error problem of regions with more or fewer textures. GwcNet[9] further proposes a group-related strategy to enhance feature diversity with multiple sets of parallel cost volumes, and achieve a balance between accuracy and efficiency. At the same time, lightweight design has become an important research direction, and StereoNet[10] realizes real-time high-precision reconstruction on the mobile terminal through low-resolution cost matching and edge-aware upsampling. GANet[11] innovatively designs the Guided Aggregation Layer (LGA/RGA) and explicitly models the geometry a priori to sharpen the parallax edges, providing a new idea for complex texture scenes.

Despite significant progress made by existing methods, high-precision reconstruction of crime scenes still faces multiple challenges. Firstly, the scene environment often suffers from interference such as motion blur and non-Lambertian reflections (e.g., bloodstains, glass), which reduces the reliability of feature matching. Secondly, current algorithms' utilization of multi-scale features mostly relies on fixed-weight fusion, making it difficult to adaptively distinguish between critical areas (such as fingerprints, footprints) and background redundant information.

In response to the aforementioned issues, this paper proposes a 3D reconstruction method tailored for crime scenes, with the following core contributions:

1.Multi-scale Feature Enhancement Module: By incorporating the S2Attention mechanism and dynamic weight allocation, this module dynamically enhances feature representation and matching saliency in critical regions (such as fingerprints, tool marks) through a dual-pathway of spatial-semantic enhancement at multiple scales, while suppressing background interference.

2.Multi-level Cost Optimization Architecture Network: This network employs a Multi-level Adaptive Feature Matching (MAPM) module to replace traditional 3D

convolutions, combined with a multi-scale cost filtering convolutional fusion structure, to achieve redundant information compression, enhanced matching costs, and mitigation of the negative impact of motion blur, while reducing the number of parameters.

3.Edge Consistency Fusion Strategy: By introducing a Spatial Adaptive Feature Fusion (SAFM) module, this strategy guides the fusion of multi-level disparity predictions with edge information, jointly optimizing the geometric consistency of disparity edges, and enhancing the detail reconstruction capability for trace evidence.

The method proposed in this paper has demonstrated its superiority on public datasets (KITTI 2012, KITTI 2015, SceneFlow) and real crime scene datasets, providing reliable 3D reconstruction technology support for forensic evidence collection compared to mainstream algorithms (such as GwcNet, GANet).

2. Introduction to the GwcNet Algorithm

The GwcNet stereo matching algorithm process includes feature extraction, construction of group-wise correlation volumes and concatenation, 3D aggregation, disparity regression, and optimization. To improve the accuracy of stereo matching, learn correlation measurements, and reduce runtime, the model employs group-wise correlation volumes for cost aggregation, which reduces parameters while retaining rich information. Additionally, it optimizes the disparity map through an improved 3D aggregation module. The network structure of GwcNet is illustrated in Figure 1. GwcNet adopts the following steps:

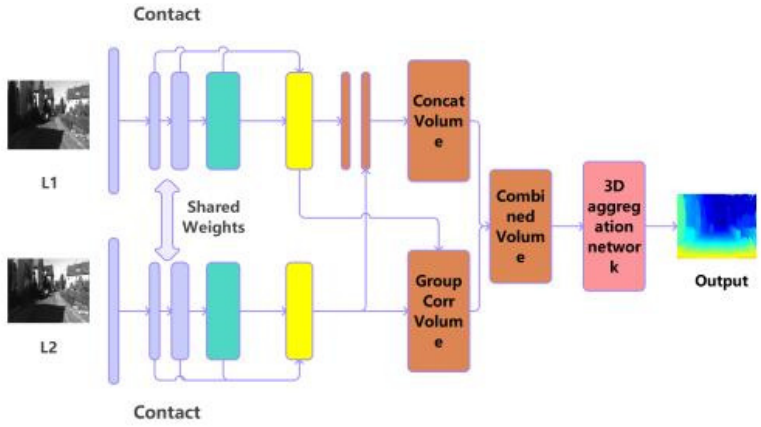

Fig 1. GwcNet flowchart

2.1. Feature Extraction

During the feature extraction phase, the model processes the left and right images using a network similar to ResNet, as seen in the efficient and concise PSMNet algorithm, to obtain feature maps at three different scales. To address the issue of degradation in large correlation regions, these three feature maps of varying scales are subsequently concatenated into a single feature map with 320 channels, which serves as the input for the 3D cost volume construction.

2.2. Build Group Correlations and Connectomes

GwcNet introduces a Group-wise Correlation mechanism that differs from traditional methods which directly concatenate left and right features. Instead, the feature maps fused from the left and right images during the feature extraction stage are divided into multiple groups along the channel dimension. Each group has a 40-channel group-wise correlation volume and a 24-channel concatenation volume, which are fused into a 64-channel cost volume. For each group of features, the similarity between corresponding positions in the left and right images is calculated separately, generating multiple groups of correlation maps. These are used to construct a 4D cost volume containing multiple groups of correlation maps, which serves as the input for 3D aggregation (H×W×D×G, where D is the disparity range and G is the number of groups).

2.3. 3D Aggregation Network

The 3D convolutional aggregation network adopts a stacked 3D hourglass network structure. This design consists of two convolutional layers and three stacked 3D hourglass modules. Each module internally contains four normalized 3D convolutional layers and ReLU activations, specifically designed for extracting and aggregating feature information. Additionally, each module outputs a disparity map, as shown in Figure 2. Furthermore, an auxiliary module is added after the first two convolutional layers to generate disparity map 0 for learning low-level features. The 3D hourglass network removes some unnecessary connections and retains two skip connections to enhance speed and improve the fusion of key and edge information features. Notably, all convolutional layers are 3D convolutions and 3D deconvolutions, enabling the network to simultaneously process information in both spatial and disparity dimensions, thereby reducing computation time.

2.4. Disparity Regression and Optimization

After the 4D cost volume aggregation is completed, the disparity maps output by the four modules are first processed through two 3D convolutional layers to produce a 1-channel 4D volume. This volume is then upsampled to restore it to the size of the original image. Along the disparity dimension, a softmax function is applied to convert it into a probability volume (as shown in Equation 1). After obtaining the disparity values from the four modules, the final loss function value is calculated by weighting (as shown in Equation 2). Smooth L1 loss (as shown in Equation 3) is used to compute the loss for each output.

$$\tilde{d} = \sum_{k=0}^{D_{max}-1} k \cdot p_k \quad (1)$$

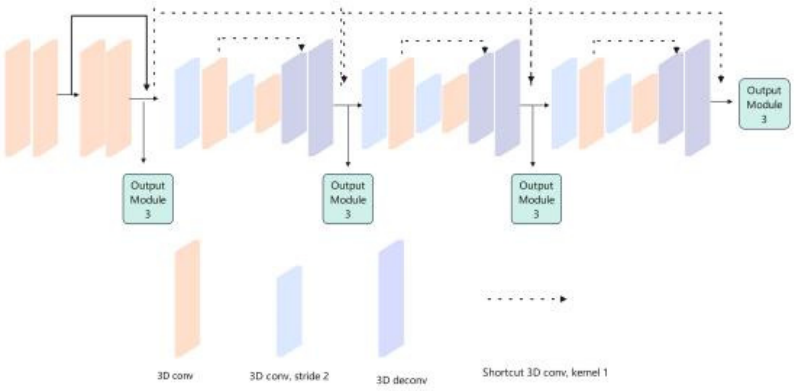

Fig 2. 3D hourglass network structure

$$L = \sum_{i=0}^{i=3} \lambda_i \cdot \mathrm{Smooth}_{L_1}(\tilde{d}_i - d^*) \quad (2)$$

$$\mathrm{Smooth}_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if} |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \quad (3)$$

3. Optimize the GwcNet Algorithm

Due to the original GwcNet model's lack of using an MLP spatial pyramid pooling module in the feature extraction stage, leading to inadequate extraction of spatial local information, this paper proposes the S2Attention mechanism to enhance spatial displacement and increase the richness of local information, as well as a divided attention mechanism to fuse multi-scale features. In the construction of cost volumes for stereo matching, concatenation and group-wise correlation are used to build 3D or 4D cost volumes. To create a cost volume that ensures accuracy and efficient runtime while addressing the issue of failure in areas with little or excessive texture, which can produce identical disparity values, this paper introduces the Multi-level Adaptive Patch Matching (MAPM) method. MAPM extracts multi-level feature maps to capture information at different scales and enhances similarity measurement through self-learned weight patch matching of local information. Additionally, to optimize and fuse the four disparity maps generated by the 3D aggregation network, the SAFM (Scale-Aware Feature Modulation) module is proposed. This module dynamically selects representative features using a multi-scale feature modulation mechanism to increase the model's ability to capture features at different scales. Furthermore, it supplements local context information through a convolutional channel mixer to compensate for missing information in the global context.

3.1. S2Attention Mechanisms

The S2Attention mechanism, as illustrated in Figure 3, is a key mechanism based on the S²-MLPv2 visual backbone network that fuses feature maps after undergoing different spatial displacement operations using a Multi-Layer Perceptron (MLP). First, the input feature map X passes through an MLP spatial pyramid pooling module layer, which expands the number of channels to three times the original, resulting in a new feature map. Subsequently, the expanded feature map is split into three parts along the channel dimension, with each part having the same size as the original feature map. This step aims to divide the feature map into different groups for applying different spatial displacement operations to them.

Next, the three feature maps undergo related operations: the first one undergoes a spatial displacement operation, which involves cyclic shifts in the up, down, left, and right directions; the second one undergoes asymmetric opposite spatial displacement operations to complement the first feature map; and the third one remains unchanged. Finally, the three feature maps after spatial displacement operations are sent to a divided attention module.

In this module, each feature map first undergoes average pooling to reduce its dimensions and extract key information. Then, after passing through a series of MLP layers, these feature maps are further transformed and generate final attention weights. These weights are used to weight the original feature maps, thereby generating the final feature map.

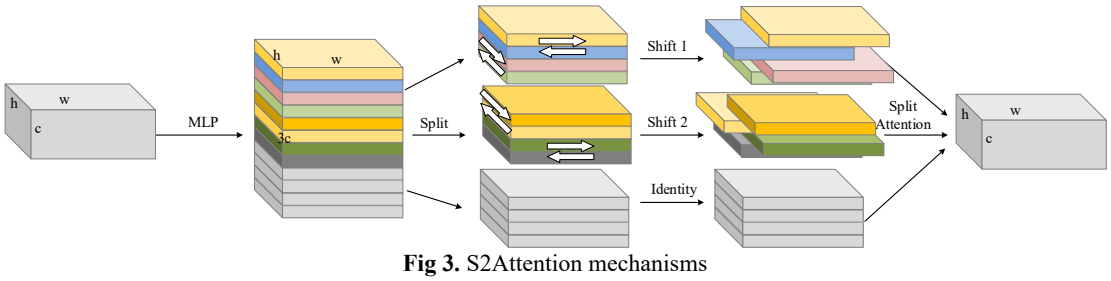

Fig 3. S2Attention mechanisms

3.2. SAFM Module

The Spatial Adaptive Feature Modulation (SAFM) module, as shown in Figure 4, focuses on achieving precise 3D reconstruction through a network architecture that dynamically adjusts features at the pixel level. First, the input features are equally divided into four sub-feature layers along the channel dimension, with each sub-layer maintaining the

same spatial size as the original feature map but with the number of channels reduced to one-fourth of the original.

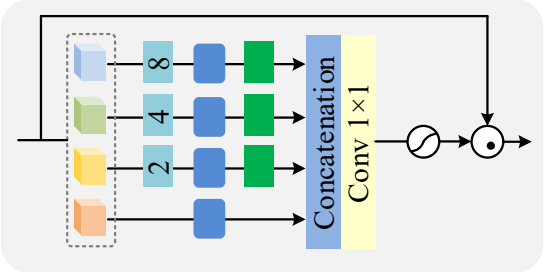

Fig 4. SAFM module

Next, a multi-scale feature pyramid is generated through downsampling operations: the lowest layer retains the original resolution, while the other three upper layers undergo 2x, 4x, and 8x downsampling, respectively, to form different levels of dimensionality-reduced representations. For computational convenience, these three downsampling branches all use 3×3 depthwise convolutions to extract neighborhood information, followed by bilinear interpolation upsampling to restore the original dimensions and spatial size.

Finally, three feature maps that have undergone spatial displacement are fed into the segmentation attention module. In this module, each feature map is averaged pooled to reduce the dimension and extract key information. Then, these feature maps are further processed through a series of MLP layers to generate attention weights. These weights are used to weight the original feature map to produce an optimized feature map.

3.3. MAPM Module and Filtered Attention Weights

3.3.1. MAPM Module

The Multi-level Feature Aggregation Module (MFAM) is a module that generates highly accurate and discriminative matching cost objects, as shown in Figure 5. Firstly, the multi-scale feature map obtained from the model feature extraction stage and the number of channels is 64, 128 and 128 to calculate the matching cost of pixels, and the connection and grouping operations are carried out at the same time. To deal with the matching problem of different disparity and texture regions, for each pixel, the dilated convolution kernel that selects the expansion rate related to the feature map hierarchy is used to calculate the matching cost, to ensure that the correct number of pixels is used when calculating the matching cost of the multi-scale hierarchical feature map. At the same time, the weights in the convolutional kernel have adaptive learning, and the matching cost of some pixels is calculated by weighting. Finally, the matching costs of the three levels are spliced into a multi-level local matching volume, such as Equations 4 and 5.

$$C^{l_k}_{patch}(g,d,x,y) = \frac{1}{N_f/N_g}\sum_{(i,j)\in\Omega^k}\omega^k_{ij}\cdot C^g_{ij}(d,x,y)$$
$$C^g_{ij}(d,x,y) = \langle f^g_l(x-i,y-j), f^g_r(x-i-d,y-j)\rangle \quad (4)$$

$$C_{patch} = \mathrm{Concat}\left\{C^{l_1}_{patch}, C^{l_2}_{patch}, C^{l_3}_{patch}\right\} \quad (5)$$

3.3.2. Filtered Attention Weights

In order to further optimize the processing of these matching costs, 3D convolution, 3D hourglass network is used for regularization, and the number of channels is compressed by another convolutional layer, and finally the attention weight A is generated. Subsequently, the weight A is used to eliminate the redundant information in the initial splicing volume, so as to obtain the attention stitching volume with stronger representation ability. where the attention splicing volume is as shown in Equation 6.

$$C_{ACV}(i) = A \odot C_{concat}(i) \quad (6)$$

4. Experimental Analyses

4.1. Quantitative Analysis

4.1.1. Ablation Experiments on SceneFlow Datasets

Table 1. Results of ablation experiments at SceneFlow in the test set

method	S2Attention	SAFM	MAPM	>1px/%	>2px/%	>3px/%	D1/%	EPE (px)
GwcNet				8.03	4.47	3.30	2.71	0.76
Base-S2A	✓			7.75	4.33	3.12	2.52	0.73
Base-S2A-SAFM	✓	✓		7.66	4.21	3.09	2.47	0.70
Base-S2A-SAFM-MAPM	✓	✓	✓	7.09	3.96	2.85	2.30	0.65

Our proposed method (Base-S2A-SAFM-MAPM) demonstrates significant superiority in stereo matching tasks on the SceneFlow dataset by integrating three modules: S2Attention, SAFM, and MAPM. As in Table 1 Compared to the baseline model GwcNet, our method achieves notable reductions in the mismatch rate (>1px/%, D1/%) and the average end-point error (EPE (px)), with improvements of 0.94% and 0.11, respectively. S2Attention enhances the features of critical regions, SAFM improves boundary clarity, and MAPM reduces mismatches and optimizes computational efficiency. The synergistic effect of these three modules significantly enhances matching accuracy and robustness, providing a more stable solution for applications in the public security sector.

4.1.2. KITTI 2012 Experiments

In Table 2, our proposed method (Ours) demonstrates significant superiority in the stereo matching task. Compared to other state-of-the-art methods, Ours reduces error rates across multiple metrics, including > 2px / %, > 3px/%, and >5px/%. Specifically, in the Noc and All scenarios, Ours achieves optimized performance with error rates of 2.11% and 2.59%, and 1.25% and 1.64%, respectively, showcasing higher matching accuracy. Meanwhile, the Mean Error (px) remains below 0.5 pixels, on par with the current best methods. Furthermore, our method requires only 0.30 seconds for execution, highlighting its efficiency advantage. Overall, Ours excels in terms of accuracy, robustness, and efficiency, providing a superior solution for stereo matching tasks.

From the data in Table 3, it can be seen that the error index of the proposed method in the KITTI2015 dataset under All pixels and Noc pixels reaches the lowest value, which is 1.98% for D1-all (All pixels) and 1.81% for D1-all (Noc pixels). At the same time, the performance in the background (D1-bg) and foreground (D1-fg) large areas is also optimized to 1.50% and 3.77%, respectively, which proves that it matches more

textures with high accuracy. In addition, the method in this paper can be calculated in only 0.30 seconds on the running time.

Table 2. Experimental results of each method in the KITTI 2012 test set

method	>2px/%		>3px/%		>5px/%		Mean Error(px)		Times/s
	Noc	All	Noc	All	Noc	All	Noc	All	
DispNetC[12]	7.38	8.11	4.11	4.65	2.05	2.39	0.9	1.0	0.06
MC-CNN-act[13]	3.90	5.45	2.43	3.63	1.64	2.39	0.7	0.9	67
GC-Net[7]	2.71	3.46	1.77	2.30	1.12	1.46	0.6	0.7	0.9
iResNet-i2[14]	2.69	3.34	1.71	2.16	1.06	1.32	0.5	0.6	0.12
SegStereo[15]	2.66	3.19	1.68	2.03	1.00	1.21	0.5	0.6	0.6
PSMNet[8]	2.44	3.01	1.49	1.89	0.90	1.15	0.5	0.6	0.41
GANet[11]	2.18	2.79	1.36	1.80	0.83	1.10	0.5	0.5	0.36
GwcNet[9]	2.16	2.71	1.32	1.70	0.80	1.03	0.5	0.5	0.32
Our	2.11	2.59	1.25	1.64	0.75	0.96	0.5	0.5	0.30

4.1.3. KITTI 2012 Experiments

Table 3. Experimental results of each method in the KITTI 2012 test set

method	All pixels/%			Noc pixels/%			Times/s
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all	
DispNetC[12]	4.32	4.41	4.34	4.11	3.72	4.05	0.06
GC-Net[7]	2.21	6.16	2.87	2.02	5.58	2.61	0.9
iResNet-i2e2[14]	2.14	3.45	2.36	1.94	3.20	2.15	0.22
PSMNet[8]	1.86	4.62	2.32	1.71	4.31	2.14	0.41
SegStereo[15]	1.88	4.07	2.25	1.76	3.70	2.08	0.6
GANet-15[11]	1.55	3.82	1.93	1.40	3.37	1.73	0.36
GwcNet[9]	1.74	3.93	2.11	1.61	3.49	1.92	0.32
Our	1.50	3.77	1.98	1.48	3.42	1.81	0.30

4.2. Qualitative Analysis

4.2.1. KITTI 2012 Experiments

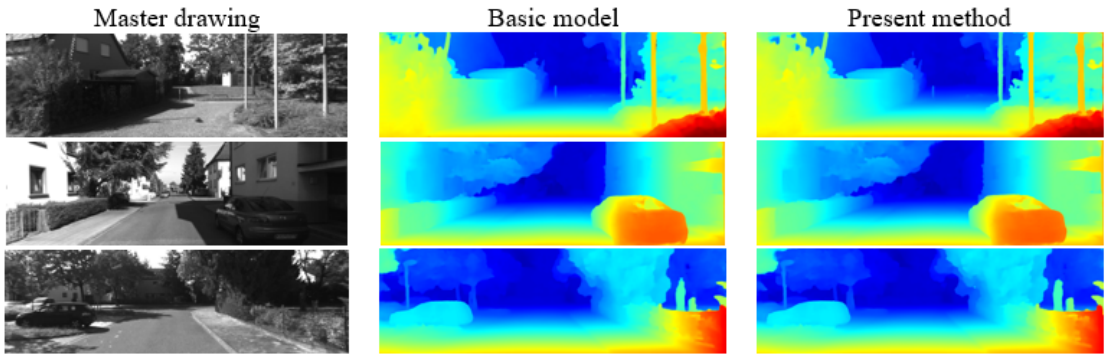

Fig 5. Comparison of the effects of the original model and our method on the KITTI2012 dataset

It can be clearly observed from Figure 5 that the proposed method has a reduction in the error value, which is manifested in the reduction of the degree of deviation of the color distribution. In the first image, the glass area of the house appears as a large white block in the original model, lacking detailed information. The method in this paper can more accurately restore the shape of the window and the specific size and contour of the three poles, showing a stronger ability to depict details. In the second image, the original model can only show a rough overall outline of the shape of the trees in the dark area, but the proposed method can capture the more detailed concave and convex structure and contour features, making the shape of the trees more realistic and natural, and the length and thickness of the dividing line on the upper edge of the fence are more clear and fine, and the clarity of the boundary area is significantly improved. In the third image, the shape of the car in the image is closer to the real curvilinear shape, and the ghosting of the trees is also effectively alleviated, and the white boundary line on the side of the highway becomes more obvious under the effect of this method, and the continuity and clarity of the lines are improved.

4.2.2. KITTI 2015 Experiments

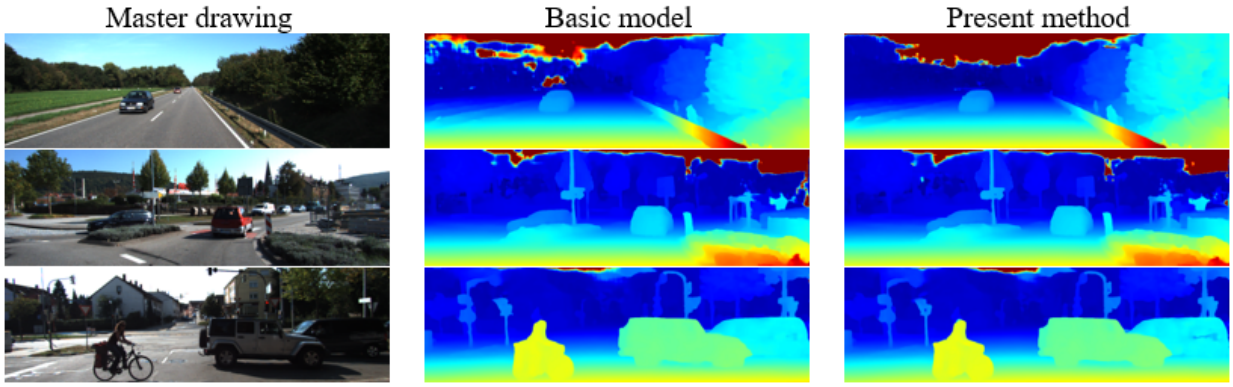

Fig 6. Comparison of the effects of the original model and our method on the KITTI2015 dataset

As can be seen from Figure 6, the experimental results using the KITTI2015 dataset demonstrate the superiority of the proposed method in the stereo matching effect in dynamic large scenes. Specifically, in the first image, the scene contains a vast blue sky and white clouds, and the proposed method significantly reduces the mismatch rate and successfully avoids misconfusing the parallax of the woods below with the sky region, so as to more accurately preserve the true parallax distribution of different regions. In the second image, small trees are shown side by side, and the proposed method can demarcate the outline of each tree more clearly and accurately, effectively reducing the problem of blurring the boundaries between trees, and further improving the ability to recognize details. The third image shows a pair of warning signs with a triangular red border and a white shape inside, which can better deal with the shape segmentation problem of the same color area in a large area, ensure that the shape and color characteristics of the warning signs are accurately restored, and avoid the problem of color overflow or parallax error.

5. Summary

In this paper, a lightweight three-dimensional matching network for 3D reconstruction of case scenes is proposed, which innovatively integrates S2Attention attention mechanism, spatial adaptive feature fusion module (SAFM) and multi-level adaptive patch matching module (MAPM). By dynamically enhancing feature extraction in key areas to express rich information, optimizing edge fusion strategies and adaptive matching local computing, the matching accuracy in more complex scenes (weak texture/occlusion) is significantly improved. This method provides reliable technical support for the 3D reconstruction of traces and physical evidence in judicial forensics with efficient and accurate matching characteristics, and has both academic innovation and engineering practical value.

References

[1] Moritz Menze, Andreas Geiger. Object Scene Flow for Autonomous Vehicles. [J], Computer Vision and Pattern Recognition, 2015: 3061-3070.

[2] Tianyuan Y, Mao Y, Jiawei Y, Yicheng L, Yue W, Hang Z, et al. PreSight: Enhancing Autonomous Vehicle Perception with City-Scale NeRF Priors[J], ECCV 2024, 2024.

[3] Jonas K, Songyou P, Zuzana K, Marc P, Torsten S, et al. Wild Gaussians: 3D Gaussian Splatting in the Wild[J], NeurIPS 2024, 2024.

[4] Thomas M, Alex E, Christoph S, Alexander K, et al. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. [J], ACM Transactions on Graphics, 2022, 41(4): 1-15.

[5] Hao W, Jing H, Huili C, Haozhe L, Yu-Kun L, Lu F, Kun L, et al. Crowd3D: Towards Hundreds of People Reconstruction from a Single Image.[J], Computing Research Repository, 2023: 8937-8946.

[6] Nikolaus M, Eddy I, Philip H, Philipp F, Daniel C, Alexey D, Thomas B, et al. A Large Dataset To Train Convolutional Networks For Disparity, Optical Flow, And Scene Flow Estimation[J], Computing Research Repository, 2016, abs/1512. 02134 (1): 4040-4048.

[7] Alex K, Hayk M, Saumitro D, Peter H, Ryan K, Abraham B, Adam B, et al. End-to-End Learning of Geometry and Context for Deep Stereo Regression.[C], IEEE International Conference on Computer Vision, 2017.

[8] Jia-Ren Chang, Yong-Sheng Chen. Pyramid Stereo Matching Network[J], CoRR, 2018, abs/1803.08669: 5410-5418.

[9] Xiaoyang G, Kai Y, Wukui Y, Xiaogang W, Hongsheng L, et al. Group-Wise Correlation Stereo Network[J], Computer Vision and Pattern Recognition, 2019: 3268-3277.

[10] Sameh K, Sean F, Christoph R, Adarsh K, Julien V, Shahram I, et al. StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction.[J], Computer Vision – ECCV 2018 Lecture Notes in Computer Science, 2018: 596-613.

[11] Feihu Z, Victor A P, Ruigang Y, Philip H S T, et al. GA-Net: Guided Aggregation Net for End-To-End Stereo Matching[J], 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, abs/1904.06587: 185-194.

[12] Nikolaus M, Eddy I, Philip H, Philipp F, Daniel C, Alexey D, Thomas B, et al. A Large Dataset To Train Convolutional Networks For Disparity, Optical Flow, And Scene Flow Estimation[J], Computing Research Repository, 2016, abs/1512. 02134(1): 4040-4048.

[13] Jure Zbontar, Yann LeCun. Computing the Stereo Matching Cost with a Convolutional Neural Network[C], Computer Vision and Pattern Recognition, 2015: 1592-1599.

[14] Zhengfa L, Yiliu F, Yulan G, Hengzhu L, Wei C, Linbo Q, Li Z, Jianfeng Z, et al. Learning for Disparity Estimation Through Feature Constancy[C], Computer Vision and Pattern Recognition, 2017.

[15] Guorun Y, Hengshuang Z, Jianping S, Zhidong D, Jiaya J, et al. SegStereo: Exploiting Semantic Information for Disparity Estimation. [J], arXiv (Cornell University), 2018, abs/1807. 11699: 660-676.