


Article

# Enhancing the Ground Truth Disparity by MAP Estimation for Developing a Neural-Net Based Stereoscopic Camera

Hanbit Gil, Sehyun Ryu and Sungmin Woo \* 

Department of Information and Communication Engineering, Korea University of Technology and Education (KOREATECH), Cheonan-si 31253, Republic of Korea; hb3568@koreatech.ac.kr (H.G.); ryurtrt@koreatech.ac.kr (S.R.)

\* Correspondence: innosm@koreatech.ac.kr

**Abstract:** This paper presents a novel method to enhance ground truth disparity maps generated by Semi-Global Matching (SGM) using Maximum a Posteriori (MAP) estimation. SGM, while not producing visually appealing outputs like neural networks, offers high disparity accuracy in valid regions and avoids the generalization issues often encountered with neural network-based disparity estimation. However, SGM struggles with occlusions and textureless areas, leading to invalid disparity values. Our approach, though relatively simple, mitigates these issues by interpolating invalid pixels using surrounding disparity information and Bayesian inference, improving both the visual quality of disparity maps and their usability for training neural network-based commercial depth-sensing devices. Experimental results validate that our enhanced disparity maps preserve SGM's accuracy in valid regions while improving the overall performance of neural networks on both synthetic and real-world datasets. This method provides a robust framework for advanced stereoscopic camera systems, particularly in autonomous applications.

**Keywords:** stereo vision; deep learning; disparity map; MAP estimation; Semi-Global Matching; neural network; interpolation; autonomous driving



**Citation:** Gil, H.; Ryu, S.; Woo, S. Enhancing the Ground Truth Disparity by MAP Estimation for Developing a Neural-Net Based Stereoscopic Camera. *Sensors* **2024**, *24*, 7761. <https://doi.org/10.3390/s24237761>

Academic Editors: Junxing Zheng and Peng Cao

Received: 13 September 2024

Revised: 28 November 2024

Accepted: 2 December 2024

Published: 4 December 2024



**Copyright:** © 2024 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 (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Depth estimation through stereo vision using deep learning is extensively applied in fields like autonomous driving, robotics, and augmented reality [1–3]. The accuracy of this technology hinges on high-quality ground truth data [4]. Although synthetic datasets can produce perfect ground truth depth maps, their use in commercial camera manufacturing is limited due to the differences between synthetic and real-world images, such as variations in lighting conditions, color changes from multiple light sources, and light reflections and transmissions [5,6].

To mitigate these issues, ground truth disparity maps are usually created using traditional stereo-matching methods like Semi-Global Matching (SGM) [7]. SGM is a dependable and precise method that calculates actual distances based on camera calibration parameters [8]. It determines disparity by minimizing the matching cost across multiple scan lines, considering pixel differences in the context of their surrounding regions in both the left and right images. However, SGM encounters difficulties when there is significant occlusion, especially with foreground objects, and in regions with flat surfaces that lack texture, making it hard to accurately compute binocular disparity. In such instances, deeming the disparity values as “invalid” ensures higher confidence in the normal values.

Intel's D-series (Intel Corporation, Santa Clara, CA, USA), one of the popular off-the-shelf stereoscopic cameras, employs 720 p/30 fps real-time circuitry to use the SGM method to generate depth information. The accuracy of distance estimation with this camera is reported to be within two percent of the actual distance, which is accepted for the autonomous driving application of distance measurement. Additionally, Intel's D-series uses infrared (IR) projection to create patterns that help calculate disparity, minimizing ambiguity in textureless

regions [9,10]. Post-processing techniques, such as edge-preserving filtering [11], spatial hole-filling [12], and temporal filtering [13] further improve the disparity map generated by SGM.

However, these post-processing methods are still insufficient in addressing the fundamental challenges of achieving better disparity maps when large invalid areas occur due to severe occlusion or lack of texture in the scenes. Above all, deep learning-based methods exhibit poor generalization when applied to the real-world datasets lacking ground truth. In this study, we propose an interpolation method to enhance SGM disparity maps, enabling their use as ground truth data for training lightweight neural network models designed for disparity estimation in mobile stereoscopic cameras. To validate this, we designed a neural network and its variants specifically tailored for disparity estimation from stereo images and trained the models on a dataset of over 4000 stereo image pairs captured using a custom-designed stereo camera equipped with IR projectors. The experimental results with synthetic and real-world datasets demonstrate that the improved ground truth significantly enhances the network's output, resulting in sharper and more visually accurate disparity maps.

## 2. Related Work

Image interpolation methods primarily include Bilinear Interpolation and Spline Interpolation. Bilinear Interpolation utilizes the values of four neighboring pixels for interpolation, making it easy to implement and computationally efficient. This method is beneficial for real-time processing but struggles to restore fine details. Spline Interpolation, on the other hand, uses a higher-dimensional polynomial spline function to achieve smoother interpolation results, albeit with increased computational costs. It is particularly effective for continuous or smoothly varying data. Partial Differential Equation (PDE)-based inpainting methods restore damaged image regions by propagating surrounding pixel information using mathematical techniques. These approaches preserve edge continuity and important structures through isotropic or anisotropic diffusion, ensuring smooth transitions between damaged and undamaged areas.

Among various specialized interpolation methods for stereo matching, an iterative color-depth Minimum Spanning Tree (MST) cost aggregation algorithm was introduced to improve accuracy in textureless regions [14]. A joint matching cost method was also developed by combining the Sum of Absolute Differences (SAD) and Census transform, enhancing stereo matching accuracy [15]. Additionally, a local stereo matching approach was proposed that integrates an adaptive exponentially weighted moving average (EWMA) filter with Simple Linear Iterative Clustering (SLIC) segmentation, improving both computational efficiency and accuracy [16]. To address occlusion challenges in disparity estimation for dynamic scenes, a local stereo matching method was proposed, combining support weights with motion flow [17].

Several segmentation-based stereo matching techniques [18–20] have also demonstrated effectiveness in handling occluded regions and object boundaries. Additionally, a multi-step disparity refinement framework was introduced to classify outliers and prevent error propagation [21]. An unsupervised stereo matching network has also been proposed, incorporating occlusion handling through a ternary classification system and directed disparity smoothing loss [22]. For textureless regions in stereo images, plane-fitting techniques [23] and non-local cost aggregation methods [24] have been suggested. Geodesic filters were applied to enhance real-time performance and improve disparity accuracy [25,26]. Lastly, belief propagation-based optimization methods were introduced to enhance disparity accuracy, particularly in low-texture areas and occlusion boundaries [27,28].

Disparity refinement methods encompass a range of approaches beyond occlusion handling and interpolation. Techniques using Markov Random Fields (MRF) [29–31] have been widely applied, along with edge-preserving filters [32] box filters [32], fast-weighted median filters [33], and multi-step methods incorporating voting strategies [34–39]. Addi-

tionally, various approaches [40–44] have been proposed to enable real-time stereo matching for FPGA or GPU implementation.

One of the most significant advances in stereo matching has been the adoption of deep learning-based methods. A unified deep learning architecture was proposed to enhance stereo confidence estimation by jointly utilizing the matching cost volume and disparity map, thereby improving stereo matching robustness [45]. For light field (LF) depth estimation, a method was introduced to learn the distribution of subpixel disparities [46]. A CNN was also employed to compute patch similarity, achieving speed and accuracy improvements in stereo matching [47]. Additionally, an unsupervised learning framework for optical flow estimation explicitly modeled occlusions to improve performance in challenging conditions [48]. The use of 3D feature volumes in intermediate layers of deep neural networks has further demonstrated effectiveness in accurate disparity calculation [49–52]. More recently, iterative stereo matching techniques have shown outstanding performance in preserving edges and shapes [53,54].

Deep learning-based interpolation methods enable broad inpainting by learning inferred structural details through neural networks, allowing models to reconstruct cohesive and contextually accurate regions within images. A Generative Multi-column Convolutional Neural Network (GMCNN) was proposed for image inpainting to restore missing regions by propagating context-derived information from surrounding areas [55]. For translation-variant interpolation (TVI) tasks such as image inpainting and super-resolution, a Shepard Convolutional Neural Network (ShCNN) was introduced to propagate information from known to unknown regions through spatially variant pixel weighting [56]. Chen et al. [57] introduced a pluralistic image inpainting approach leveraging latent codes and a bidirectional transformer to effectively restore large masked regions. A mask-aware inpainting framework was also developed to improve feature learning for missing regions [58], while a transformer-based model targeted large-hole inpainting [59]. Recent inpainting approaches have increasingly leveraged diffusion models for enhanced outcomes [60–64].

Despite the advantages of deep learning-based stereo matching techniques in achieving natural edge estimation and object description in disparity maps, they are computationally intensive and thus less suited for real-time stereo camera applications. Moreover, generating accurate real-world training data for precise distance estimation remains challenging, as inpainting methods, while visually plausible, often lack accuracy in estimating distances. Furthermore, recent deep learning-based approaches frequently suffer from poor generalization, especially when applied to diverse real-world scenarios. To address these issues, we propose a Maximum a Posteriori (MAP) estimation-based method to enhance disparity maps generated by Semi-Global Matching (SGM), making them more suitable as high-quality training data for lightweight neural networks in real-time stereoscopic camera applications. Our MAP-based interpolation approach not only improves edge consistency and disparity accuracy but also provides a robust foundation for disparity estimation in autonomous and real-time scenarios. The main contributions of our work are as follows:

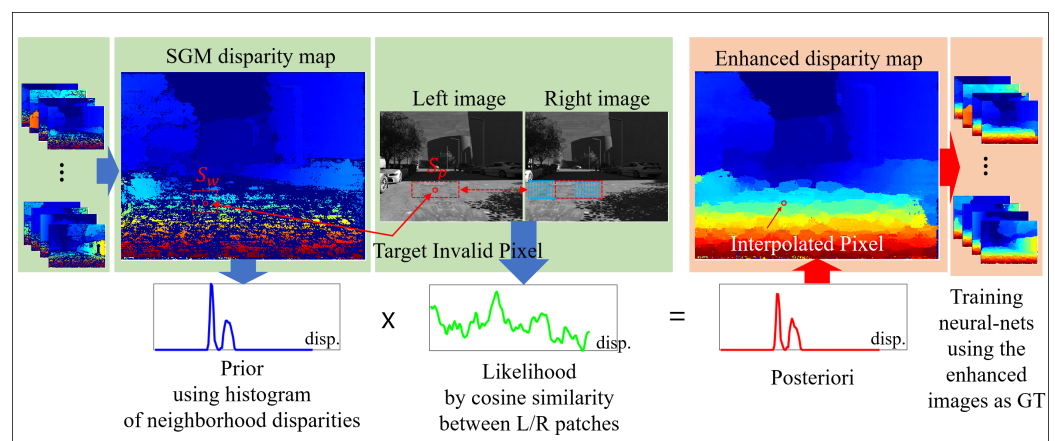
1. **MAP-Based Interpolation with Enhanced Likelihood Calculation for Invalid Disparity Pixels:** Our MAP estimation approach leverages surrounding disparity information as the prior and utilizes cosine similarity as the likelihood, complemented by a preprocessing step that standardizes pixel intensity and applies masking. This posterior probability estimation identifies the most plausible disparity values for invalid regions caused by occlusions, textureless surfaces, and inconsistencies between left and right images in stereo vision.
2. **Practical Validation in Real-World Scenarios Where Ground Truth Is Difficult To Obtain:** The proposed method demonstrates robustness and applicability in both ground truth (GT)-available and GT-unavailable scenarios. In environments without GT, where accurate disparity maps are inaccessible, learning-based methods trained on different datasets often struggle with challenges such as indistinct boundaries and the propagation of incorrect predictions when applied to unseen images.

Our approach overcomes these limitations by employing MAP-based interpolation to generate sharper and more reliable disparity maps. Evaluations conducted on a comprehensive dataset of over 4000 real-world stereo images validate the proposed method's ability to generalize effectively across diverse environments, emphasizing its practical usability in real-world applications where GT is unavailable.

3. **Improved Ground Truth for Lightweight Neural Network Training:** We evaluate the impact of our enhanced ground truth data on various lightweight neural network architectures optimized for mobile applications. Specifically, we compare network performance using original SGM-based ground truth data against ground truth data enhanced by our proposed method. Results indicate that the proposed enhancement significantly improves output quality and generalization of these networks, demonstrating its value for real-time stereoscopic applications with limited computational resources.

### 3. Proposed MAP Estimation

Figure 1 outlines the proposed framework for improving the SGM disparity map by filling in invalid pixels. The process begins with an SGM disparity map that contains missing regions. A prior is calculated using neighboring disparities within a fixed window, while the likelihood is determined through cosine similarity between patches from the left and right images. These prior and likelihood calculations are combined to create a posterior distribution, which estimates the disparities for the invalid pixels. The result is a refined disparity map with interpolated pixels, enhancing the accuracy of the original SGM disparity map.

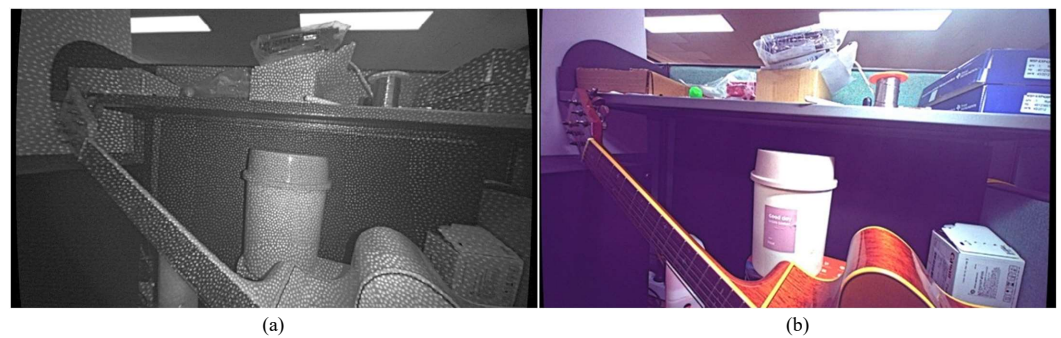


**Figure 1.** The proposed framework for enhancing SGM disparity map.

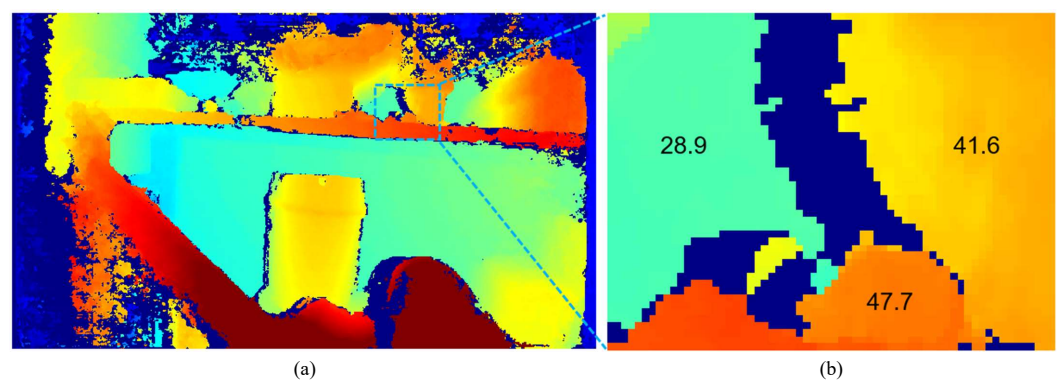
#### 3.1. Invalid Regions of SGM

In SGM, invalid pixels occur when matches cannot be found, often due to occlusions, similar costs in neighboring patches, or disparities between left and right images. Figure 2a,b show the left and RGB images, both captured by a stereo camera system aligned with an IR projector. In Figure 2b, the RGB camera's color filter blocks infrared wavelengths, rendering the IR patterns invisible. However, in Figure 2a, the irregular patterns projected by the IR aid in stereo matching by adding texture to otherwise featureless areas, leading to a denser disparity map. Despite this, some invalid pixels caused by occlusion remain unavoidable. These occlusions are a result of object and camera geometry, where closer objects and greater distance between cameras increase occlusion effects. Figure 3a shows the SGM disparity map created from the left and right images in Figure 2, with a "jet" colormap used to represent closer objects in red and farther objects in blue. Invalid pixels appear as dark blue. Figure 3b provides a magnified view of the red box in Figure 3a, with

numbers representing approximate average disparity values for different regions. In this figure, the disparity map is aligned with the left camera as the reference.



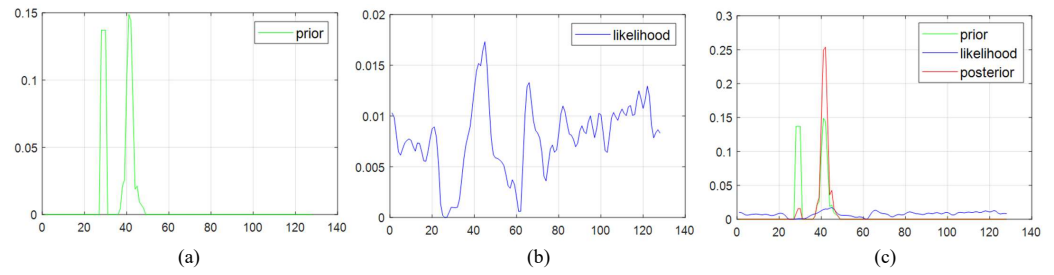
**Figure 2.** Example of left (a) and RGB (b) images.



**Figure 3.** (a) Disparity map generated by SGM for the images in Figure 1. (b) Enlarged view of (a). Dark blue pixels indicate “invalid” regions. The numbers shown represent disparity values for each grouped region.

### 3.2. Prior Probability of Invalid Pixels

To estimate the empty disparities in SGM using MAP, a prior probability  $P(\theta)$  is established.  $\theta$  represents the disparity of an invalid pixel. The key idea in setting the prior probability is that the occluded disparity is limited to the disparities surrounding the target. To do this, a histogram of the valid pixels within a  $S_w$ -sized window centered by a target pixel is constructed in the range of zero to maximum disparity, and then the histogram is normalized so that they can be represented as probabilities. In the resulting normalized histogram, the probabilities tend to be concentrated in distinct disparity spots, leading to zero probability for adjacent high disparity priors, which is unlikely. To address this, a 1D convolution is applied to spread out the initial prior distribution. Figure 4a illustrates the resulting prior distribution of the target pixel from Figure 3b with  $S_w$  set to be  $17 \times 17$ . The horizontal axis represents disparity, and the maximum disparity is set to 128. The convolution does not affect the distribution in the range of 50 to the maximum disparity, indicating that regardless of the likelihood values given later, the posterior will be limited to the non-zero range.



**Figure 4.** (a) Prior probability, (b) Likelihood, and (c) Posterior distribution of an invalid pixel from Figure 3.

### 3.3. Likelihood Probability

To determine the likelihood of invalid pixels  $P(D|\theta)$ , a  $S_p$ -sized patch centered on the target pixel in the left image is compared with corresponding moving cropped regions of the same size in the right image. The similarity between the left and right patches is measured using cosine similarity. In our case, the brightness of the captured image has slight differences between pairs. Thus, the following preprocessing steps are initially performed on the left and right patches to facilitate image comparison:

1. The original captured image is standardized as follows:

$$x_{st} = \frac{x_{ori} - \mu}{\sigma} \quad (1)$$

where  $\mu$  and  $\sigma$  are the average and standard deviation of the patch, respectively.

2. The standardized image  $x_{st}$  is then masked out so that the pixels with low intensity are not regarded as distinct features:

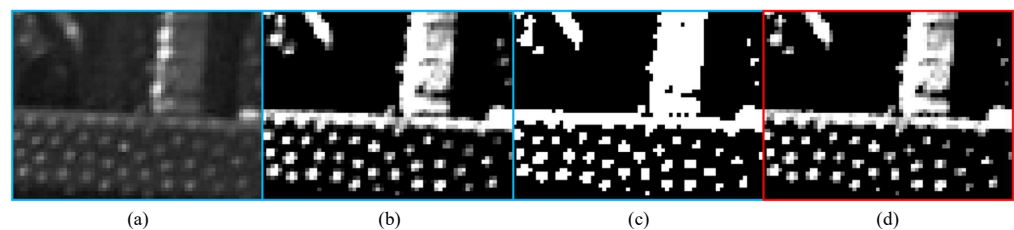
$$\tilde{x} = \begin{cases} x_{st} & \text{if } x_{st} > I_{th}, \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where  $I_{th}$  is the intensity threshold, set to  $-0.7$  throughout the study. Figure 5 illustrates the above preprocessing step, consecutively. Therefore, in the preprocessing step, only the pixels with relatively high intensities are used for comparison.

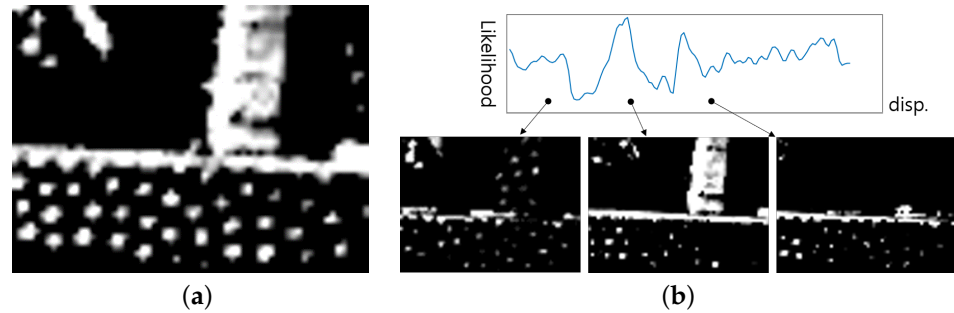
Figure 6 illustrates the left patch and some of the corresponding candidate patches from the right image, used to estimate the invalid pixels. The initial likelihood is calculated using the cosine similarity between the flattened left  $\tilde{x}^L$  and  $i$ -th horizontal displacement image  $\tilde{x}_i^R$  as follows:

$$\cos(\tilde{x}^L, \tilde{x}_i^R) = \frac{\tilde{x}^L \cdot \tilde{x}_i^R}{\|\tilde{x}^L\| \|\tilde{x}_i^R\|} \quad (3)$$

After normalizing the cosine similarities over the range from one to the maximum disparity, the likelihood probability distribution  $P(D|\theta)$  is obtained as illustrated in Figure 4b with  $S_p$  set to a  $48 \times 8$  window. However, cosine similarity is sensitive to noise and geometrical displacement, leading to a probability distribution that is not smooth.



**Figure 5.** Preprocessing steps for the proposed method: (a) Original cropped patch, (b) Standardized patch, (c) Mask, and (d) Masked patch.



**Figure 6.** (a) Left masked cropped patch. (b) Right cropped candidate patches.

### 3.4. Posterior Probability

The posterior probability of the invalid pixels  $P(\theta|D)$  is determined by Bayes' rule and the estimated disparity  $\hat{\theta}_{\text{MAP}}$  by the MAP estimation are determined as follows:

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}} P(D|\theta)P(\theta) \quad (4)$$

Figure 4c illustrates the posterior probability obtained through MAP estimation using the prior from Figure 4a and the likelihood from Figure 4b. Since the posterior probability is the product of the prior probability and the likelihood, any values with a prior probability of zero will also have a posterior probability of zero. Although considering only the likelihood might suggest certain disparity values, combining it with the prior probability, which takes into account the disparity values near the invalid region, results in a more accurate estimation of the optimal disparity value. In the example shown in Figure 3, the disparity of the invalid region is estimated to be 47.7, the value with the highest posterior probability among those with a non-zero prior probability.

## 4. Experimental Result

The performance of the proposed method is evaluated on both synthetic datasets with GT available and real-world datasets where GT is unavailable. Specifically, for the real-world dataset we collected, we demonstrate how the disparity map enhanced by the proposed method improves the output of each lightweight neural network capable of running in real-time on mobile devices.

### 4.1. Performance Evaluation on Synthetic Datasets with Ground Truth

Firstly, the Sceneflow dataset [65], a large-scale synthetic dataset designed for depth estimation, consists of three subsets: FlyingThings3D, Driving, and Monkaa. For this study, we focused on the Driving dataset, which is specifically tailored to autonomous driving scenarios. Figure 7a shows the ground truth disparity maps from the dataset. To generate SGM disparity maps from the left and right images of the Driving dataset, we adjusted the “UniquenessThreshold” parameter,  $U_{th}$  in the SGM function in MATLAB R2023b. Figure 7b presents the resulting disparity maps with  $U_{th} = 10$ , which ensures that the disparities of the valid pixels closely match the ground truth while maximizing the proportion of the valid regions. However, the SGM disparity map still contains some invalid pixels, shown in dark blue, especially on the left sides of objects and in textureless regions. Figure 7c displays the SGM disparity maps with  $U_{th} = 0$ , where no invalid pixels are generated, but a large portion of the disparities is incorrectly estimated. Consequently, the maps in Figure 7b with  $U_{th} = 10$  are considered the most reliable outcomes and are used as input images for comparing the performance of different methods. We compare the proposed method against basic interpolation methods, a PDE-based method, and four neural network-based methods.

The results of linear interpolation and nearest-neighbor interpolation are shown in Figures 7d,e. In these traditional interpolation methods, the pixels marked as invalid were masked and replaced with “NaN” to estimate the disparity. In the inpainting methods,

which aim to identify invalid regions by masking them and restoring the missing areas, the masked regions were replaced with either the median disparity value or “NaN” to minimize errors caused by incorrect disparity estimation. Figure 7f,g,h,i,j show the results produced by the PDE-based inpainting method and the neural network-based methods, ShCNN [56] and GMCNN [55], MADF [58], Chen [57], respectively. ShCNN used a pre-trained model, while GMCNN, MADF, and Chen were directly trained using the ground truth from the Driving dataset. For GMCNN and Chen’s method, the mask is randomly generated according to the original code, while for MADF, the invalid pixels in the SGM output are used as the mask. The proposed method improves upon the SGM disparity map by using it as a prior and interpolating invalid pixels without introducing incorrect estimations, as illustrated in Figure 7k.

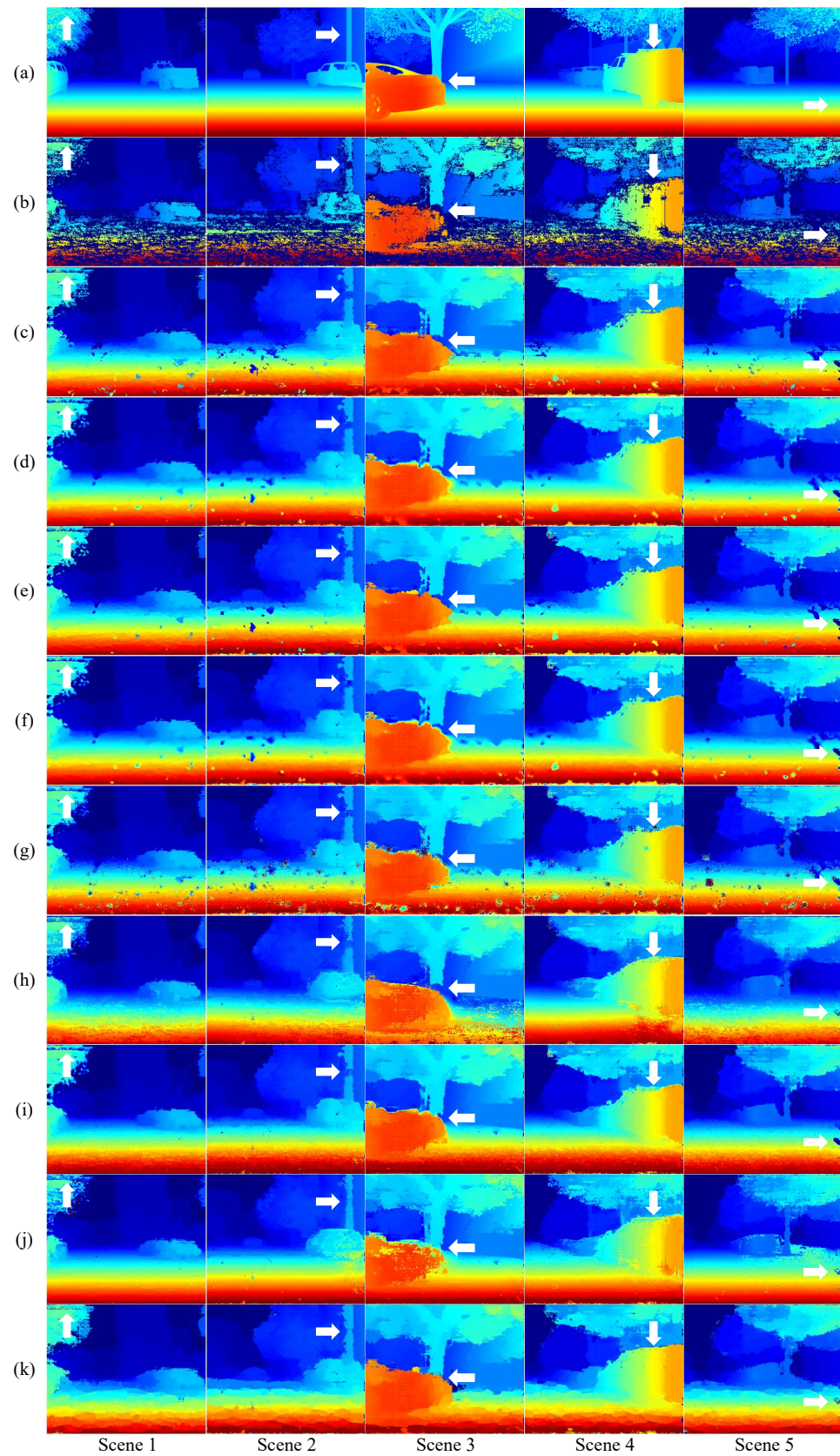
As shown in Figure 7b, a large number of invalid pixels occur in textureless areas, such as the floor, forming a large region. While the proposed method maintains consistent accuracy even in the large invalid regions, the basic interpolation methods and ShCNN tend to estimate incorrect disparity values as the size of the invalid region increases. These methods perform poorly in the lower half of each disparity map in Figure 7, where the proportion of invalid pixels is substantial. Although GMCNN was trained on data with similar scenes and shows sharper details in the upper half of the image, it still struggles with large invalid regions at the bottom, as seen in Figure 7g. MADF fills missing areas with fairly plausible values; however, as observed in the arrow-marked regions in Figure 7h, unexpected empty regions appear, a common issue in deep learning-based methods. Chen’s method, while producing smooth results in regular and repetitive lower regions such as roads, tends to estimate incorrect disparity values in the upper regions, which are rich in structure and non-repetitive patterns, as demonstrated in Figure 7j. In contrast, the proposed method uses the SGM prior to accurately fill invalid regions, preserving the integrity of the original valid data—something that existing methods generally fail to achieve. Notably, the proposed method does not smooth out the valid pixels in the SGM output.

To quantitatively measure the performance of each method, the Endpoint Error (EPE) is used as follows:

$$\text{EPE} = \frac{1}{|V|} \sum_{i \in V} |\hat{d}_i - d_i^*| \quad (5)$$

where  $V$  and  $|V|$  are the set of valid pixels and the number of valid pixels, respectively.  $\hat{d}_i$  is the estimated disparity value for pixel  $i$ , and  $d_i^*$  is the ground truth disparity value for pixel  $i$ . A summary of the endpoint errors and invalid pixel ratios for the various methods is provided in Table 1. The default SGM method shown in Figure 7b achieved the lowest EPE but marked only about 63 percent of all pixels in the dataset as valid, on average. Reducing  $U_{th}$  to 0 increased the error to 7.77, setting this EPE as the baseline for further enhancement. The proposed method did not achieve the lowest overall EPE in synthetic datasets, primarily due to the regularity of road regions, which favor deep learning-based methods. To analyze this further, we evaluated the error separately for the upper (structure-rich) and lower (road) regions of the images, divided approximately in a 50:50 ratio. While deep learning methods performed better in the lower regions with repetitive patterns, our method showed improved accuracy in the upper regions containing complex structures and objects. This underscores the robustness of our approach in handling irregular and non-repetitive patterns, where learning-based methods often struggle.





**Figure 7.** Disparity map comparisons on the synthetic Driving dataset across different scenes. (a) Ground truth, (b)  $SGM(U_{th} = 10)$ , (c)  $SGM(U_{th} = 0)$ , (d) Linear Interpolation, (e) Nearest Interpolation, (f) PDE, (g) ShCNN [56], (h) GMCNN [55], (i) MADF [58], (j) Chen [57], and (k) The proposed. Invalid regions are shown in darkish blue.

**Table 1.** End-point error by region and invalid pixel ratio comparisons.

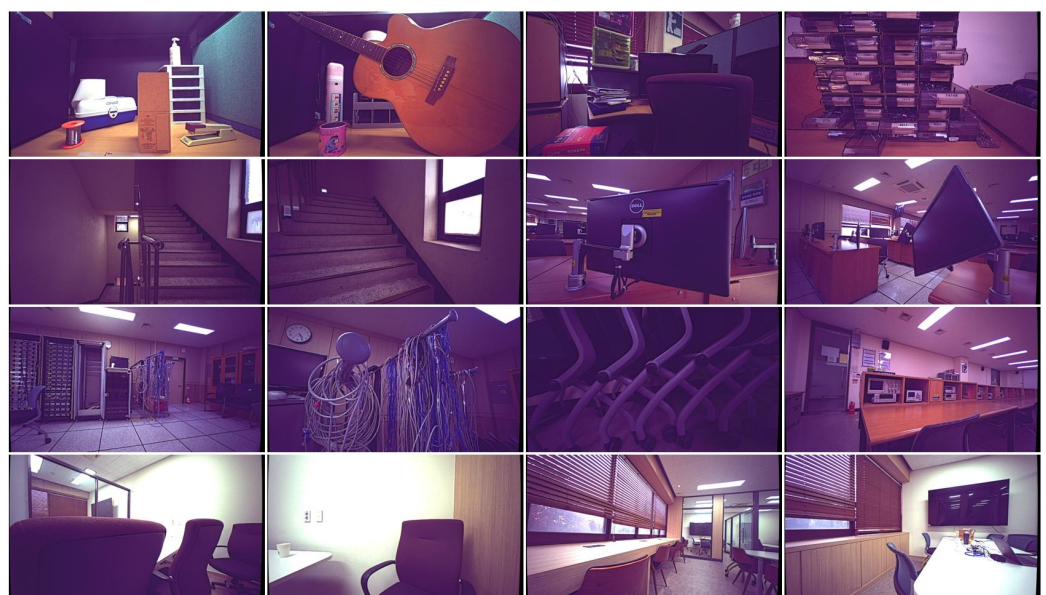
	Metric	SGM( $U_{th} = 10$ )	SGM( $U_{th} = 0$ )	Linear	Nearest	PDE	ShCNN [56]	GMCNN [55]	MADF [58]	Chen [57]	Proposed
EPE	Overall	4.87	7.77	7.30	7.48	7.37	8.87	8.51	6.36	<b>4.08</b>	6.68
	Upper	2.69	3.38	3.31	3.34	3.31	3.53	3.37	3.29	3.38	<b>3.25</b>
	Lower	10.15	12.19	11.33	11.65	11.46	14.06	13.69	9.45	<b>4.79</b>	10.15
	Invalid(%)	37.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15

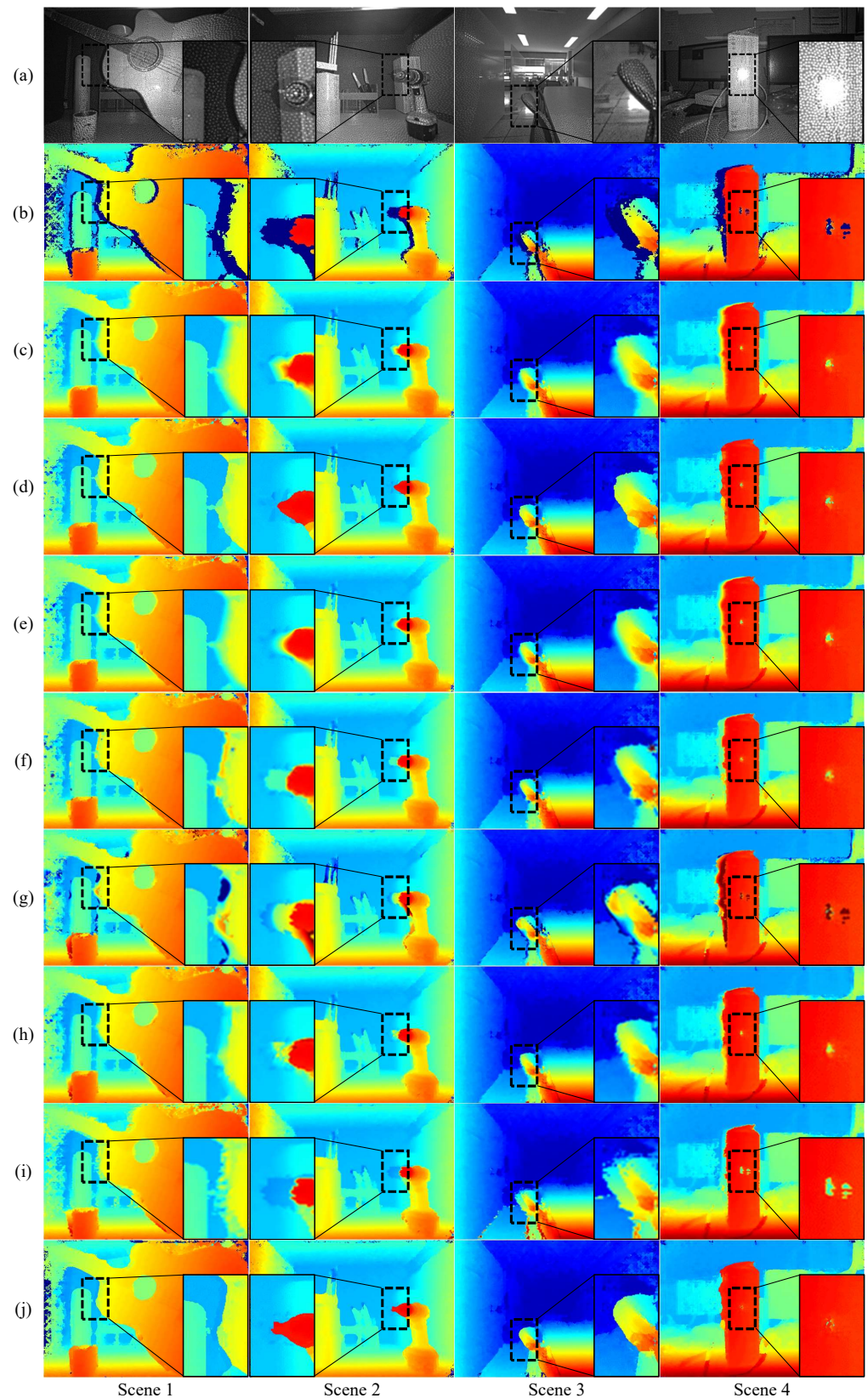
#### 4.2. Self-Generated Real-World Dataset

To evaluate the performance of the proposed method in enhancing ground truth SGM images used for training a neural network in disparity estimation for autonomous driving, we developed a comprehensive real-world dataset using a stereoscopic camera. In real-world scenarios, unlike the proposed method, deep neural networks often struggle to obtain ground truth. Therefore, their performance is evaluated by comparing how well a neural network trained on a specific dataset performs on unseen datasets. The self-generated dataset consists of synchronized left, right, and RGB images along with IR patterns to enhance the matching algorithm. A total of 4777 image pairs with the size of  $848 \times 480$  were captured, covering distances ranging from 20 cm to infinity in various autonomous robotic environments.

##### 4.2.1. Ground Truth Disparity Map Comparison for Training Neural-Nets

Figure 8 presents examples of the RGB images from the captured scenes. We first generated disparity maps using SGM and then applied the proposed interpolation method to these maps. Figure 9a shows the images captured by the left camera. As illustrated in Figure 9b, the initial SGM disparity maps contain invalid pixels, indicated by a dark blue color, particularly on the left side of objects. The presence of these invalid pixels is reduced compared to the synthetic dataset due to the IR patterns that create comparable features on flat surfaces. Unlike existing methods, the enhanced disparity maps produced by the proposed method clearly define object boundaries, as highlighted in the marked areas of the first to third columns in Figure 9. In the fourth column, it is evident that SGM produces incorrect disparities in the center of the marked rectangle along with invalid pixels. This error propagates through the interpolation methods, resulting in a significant amount of incorrect disparity, as shown in the comparison methods.

**Figure 8.** Example captured images of real-world indoor scenes.



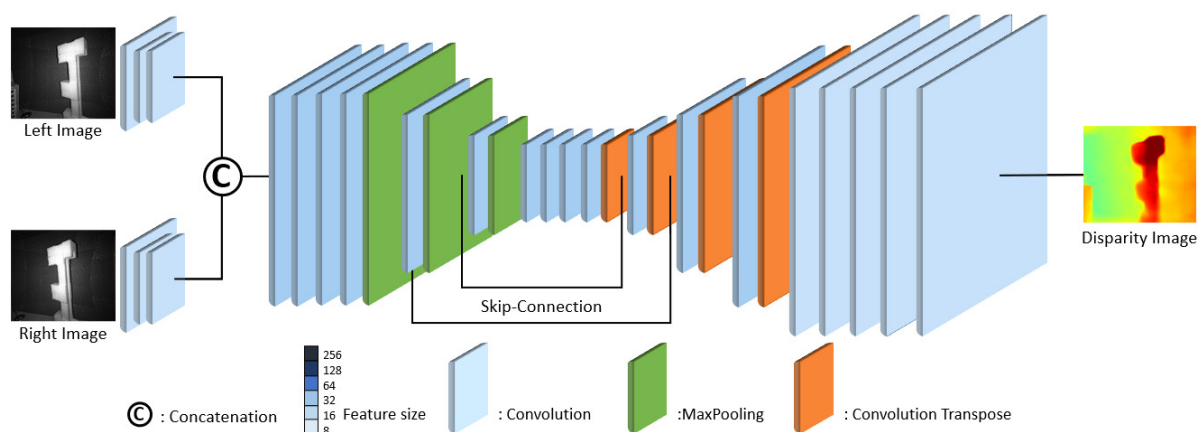
**Figure 9.** Disparity map comparisons across different real-world scenes. (a) Input left images, (b)  $SGM(U_{th} = 10)$ , (c) Linear Interpolation, (d) Nearest Interpolation, (e) PDE, (f) Shepard inpainting [56], (g) GMCNN [55], (h) MADF [58], (i) Chen [57], and (j) The proposed. The insets highlight areas with significant differences, particularly in challenging regions with occlusions and textureless surfaces. Invalid regions are shown in darkish blue.

While deep learning-based comparison methods showed numerically superior performance on the synthetic dataset, they performed significantly worse than the proposed method in terms of subjective quality on unseen real-world data, which differs from the synthetic data. In real-world scenarios, where obtaining GT is challenging, neural network-based inpainting techniques struggle to achieve reliable interpolation performance. In contrast, the proposed method demonstrates comparable performance on real-world data to that observed on synthetic datasets, highlighting its ability to generalize effectively across diverse datasets, even under challenging real-world conditions.

#### 4.2.2. Neural Network Output Map Comparison

To evaluate how the enhanced disparity maps influence the training of neural networks when used as GT, we conducted extensive experiments with neural network models under two different scenarios using a self-generated real-world dataset. In the first scenario, the models were trained on the original disparity maps produced by SGM, serving as a baseline for performance comparison. These maps contain invalid pixels, particularly on the left side of objects. In the second scenario, the models were trained on the enhanced maps generated by the proposed interpolation method, designed to improve the quality and accuracy of the disparity maps. This dual-training approach enabled a thorough evaluation of the proposed interpolation method's impact on the overall performance of the neural network models, particularly their ability to produce visually appealing and accurate results in the interpolated regions.

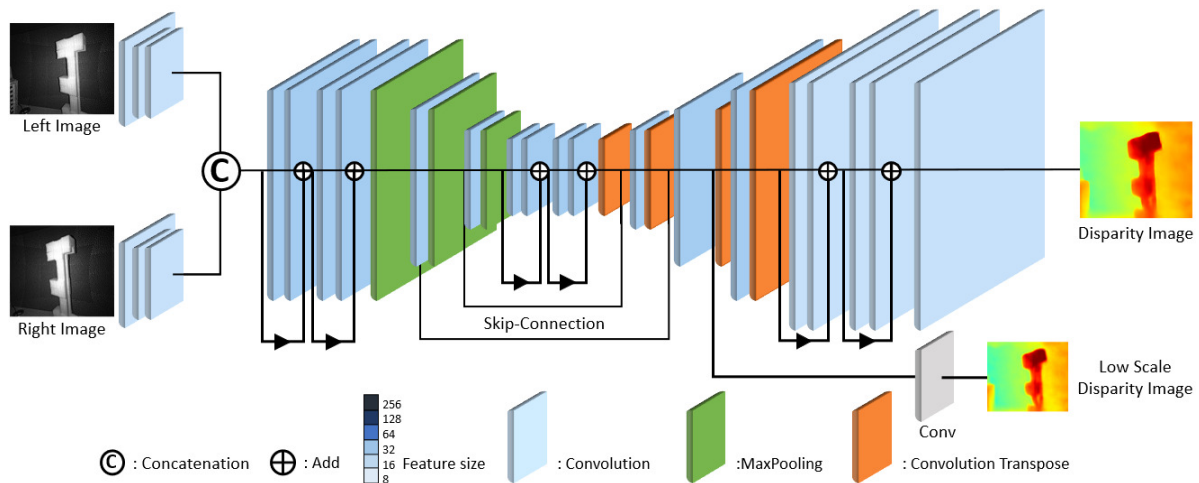
To this end, we compared three lightweight neural network models, each optimized for on-device AI, with PSMNet [50], a state-of-the-art disparity regression model. The first model in Figure 10 features a simple Encoder–Decoder architecture featuring Convolutional Layers with 16–32 features, Pooling Layers, and Convolutional Transpose Layers for upsampling. The Encoder reduces the input image dimensions through Convolutional and Pooling Layers, extracting key features. The Decoder then upsamples these features to restore the image to its original size, ultimately generating a pixel-wise disparity map. This basic neural network model includes approximately 220.93 K parameters and requires 21.8 GFLOPs to achieve real-time 30fps operation, which is more efficient compared to the 28.6 GFLOPs needed by YoloV8s. All FLOPs are measured based on an image size of  $640 \times 480$ .



**Figure 10.** Basic CNN-based model for disparity estimation.

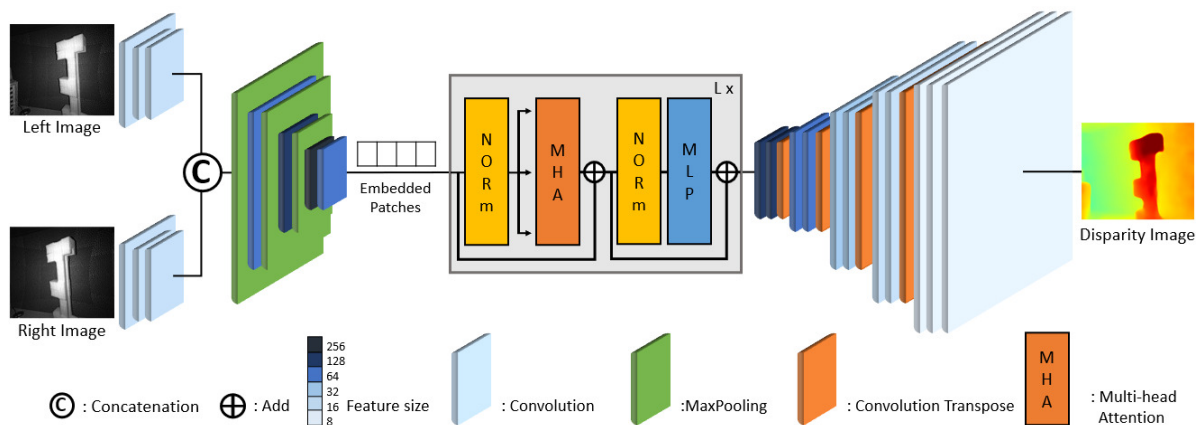
The second model in Figure 11 incorporates a residual learning algorithm [66] to the Convolutional Layers to enhance the edges in invalid areas of the ground truth. There are no scores assigned to the invalid pixels in the training of the neural network, which makes the left-sided region of the object unclear. Rather than assigning pixel-by-pixel scores at the original image size, the second model employs low-scale intermediate features to capture structural information of disparity of the original size [65]. Therefore, low-

scale disparity feature maps in the intermediate layer are extracted, convolved to reduce one channel disparity map, and additional loss at this scale is attached to the main loss. The model is designed with approximately 202.87 K parameters and requires 21.7 GFLOPs for computation.



**Figure 11.** ResNet-based model for disparity estimation. Note that the additional low-scale intermediate feature maps are used to capture structural information of disparity at the original size.

The third model replaces the Encoder of the baseline model with a Vision Transformer (ViT) [67] and modifies the Decoder accordingly. Figure 12 illustrates the structure of this third model. This neural network model is designed with approximately 1.79 M parameters and requires 14.9 GFLOPs for computation.



**Figure 12.** Vision Transformer-based model for disparity estimation. This model replaces the Encoder of the baseline model with a Vision Transformer (ViT) and modifies the Decoder from Figure 10 accordingly.

For comparison, we also evaluate these models against PSMNet. While PSMNet, with its 5.22 million parameters and 893.8 GFLOPs, is not feasible for onboard processing, it provides a useful benchmark for assessing the performance of the three models. Table 2 outlines the computational requirements for each neural network architecture.

**Table 2.** Computation comparison of neural networks tested, with FLOPs calculated for an image size of  $640 \times 480$ .

	CNN-Based	ResNet-Based	ViT-Based	PSMNet [50]
# parameters	220.93 K	202.87 K	1.79 M	5.22 M
FLOPs	21.8 G	21.7 G	14.9 G	893.8 G

A total of 4299 images were used for training, while an additional 478 images were set aside for testing. All models employed the L1 loss function as specified in Equation (6), with a batch size of 64 and a learning rate of 0.001, optimized using the Adam optimizer with exponential decay. During training, images were randomly cropped to  $256 \times 256$  pixels, and intensity levels were scaled within the range from 0.5 to 1.5. To exclude invalid regions, a mask, as described in Equation (6), was applied during loss calculation. Additionally, the maximum disparity  $d_{max}$  was set to 64 to cover the effective range of the camera.

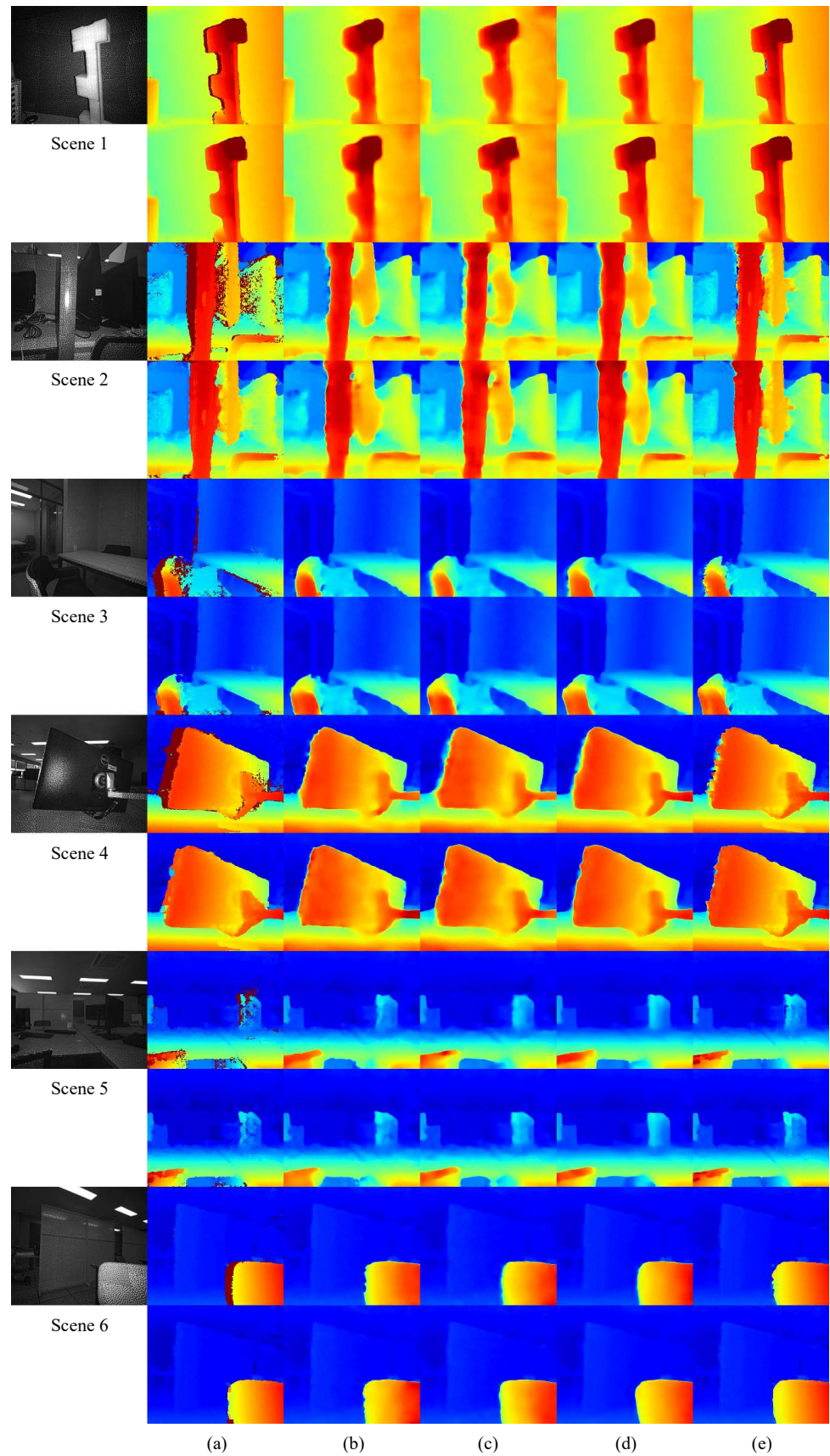
$$Loss = \sum_{i \in V} \left| \hat{d}_i - \min(d_{max}, d_i^*) \right| \quad (6)$$

The disparity maps presented in Figure 13 illustrate the outcomes from different neural network architectures. The first and second rows of each scene in Figure 13 represent models trained with SGM GT and the proposed GT, respectively. Across all architectures, the original SGM GT produces disparity maps with unclear edges, particularly along the left boundaries of objects where the ground truth was uncertain due to occlusions. In contrast, the proposed GT generates outputs with improved clarity, displaying slightly thicker but sharp boundaries as seen in the second row of the neural network models. PSMNet does produce more accurate disparities compared to the more concise models, the presence of invalid pixels results in a less organized overall disparity map, faithfully reproducing the unstructured details of SGM. The proposed method, by filling in the holes, creates cleaner and more organized boundaries.

Table 3 presents a comparison of endpoint errors across different neural network architectures. PSMNet achieved the lowest endpoint errors, followed by the transformer-based model, both when trained with SGM GT and the proposed GT. In the ResNet-based architecture, the proposed GT not only delivered visually improved disparity map but also a numerically superior disparity map compared to the original SGM GT.

**Table 3.** End-point error comparisons.

	CNN-Based	ResNet-Based	ViT-Based	PSMNet [50]
SGM GT	1.83	2.13	1.74	1.26
Proposed GT	1.91	2.03	1.87	1.35



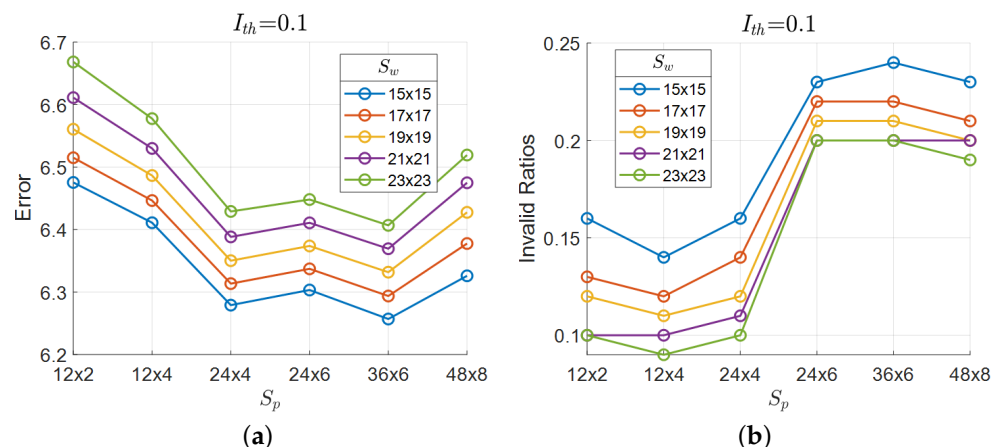
**Figure 13.** Disparity map comparisons across various scenes using different models. (a) GT, (b) CNN-based, (c) ResNet-based, (d) ViT-based, (e) PSMNet [50]. The first row of each scene is trained with the original SGM GT, and the second row with the proposed GT. Invalid regions are shown in darkish red.

## 5. Discussion

### Parameter Selection for Proposed Method

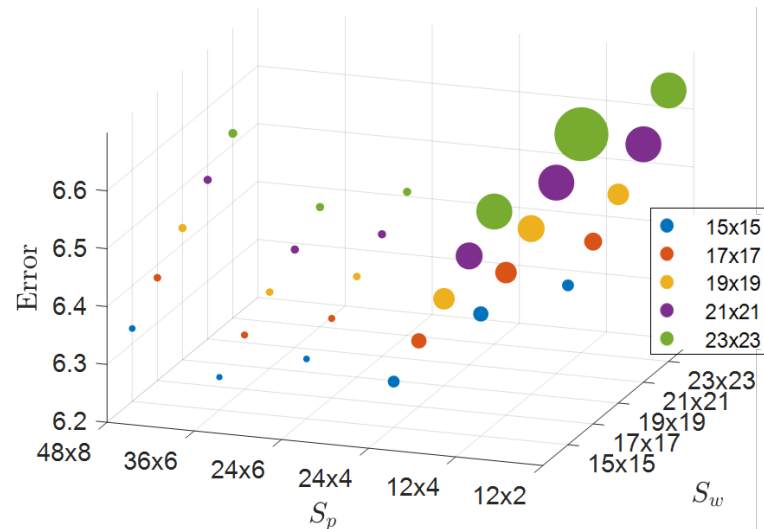
In the proposed MAP-based disparity estimation, there are controllable parameters: the window size  $S_w$  for the prior, the patch size  $S_p$  for the likelihood, and the intensity threshold  $I_{th}$  for masking. To determine the optimal combination of these parameters for minimizing error while also reducing invalid regions, we randomly selected 400 images from the Driving dataset. Given the numerous possible combinations, we prioritized  $S_w$  and  $S_p$  in the initial analysis. Disparity maps were generated using the proposed method, varying  $S_w$  and  $S_p$  while keeping  $I_{th}$  fixed.

Figure 14a,b show results for  $I_{th} = 0.1$ . From Figure 14a, the combination of a  $15 \times 15$  window for  $S_w$  and a  $36 \times 6$  patch for  $S_p$  yielded the lowest error. However, as shown in Figure 14b, smaller  $S_w$  and larger  $S_p$  tended to produce more invalid regions. Therefore, selecting parameters requires balancing the goals of minimizing error and reducing invalid regions. After evaluating these factors, we chose  $17 \times 17$  for  $S_w$  and  $24 \times 4$  for  $S_p$ . Similar results were observed for other values of  $I_{th}$ . Figure 15 illustrates the relationship between these parameters and error in 3D space, with the intensity threshold held constant. In the plot, point size represents the reciprocal of the invalid pixel ratios, with larger points indicating smaller invalid pixel ratios. Generally, a larger  $S_p$  is preferable, while a smaller  $S_w$  yields better results. Consequently, we selected a  $17 \times 17$  window for  $S_w$  and a  $24 \times 4$  patch for  $S_p$  to balance error reduction and invalid pixel ratio minimization. Finally, with  $S_w$  and  $S_p$  fixed, we tested intensity thresholds among 10 different values, including “None,” and determined  $I_{th} = -0.7$  as optimal.



**Figure 14.** Relationship between patch size  $S_p$  and both error and invalid pixel ratios for various prior window sizes  $S_w$  at a fixed intensity threshold  $I_{th} = 0.1$ . (a) shows how smaller values of  $S_w$  and larger values of  $S_p$  tend to minimize error, with an optimal configuration observed around  $S_w = 17 \times 17$  and  $S_p = 24 \times 4$ . (b) demonstrates that smaller values of  $S_w$  generally lead to higher invalid pixel ratios, while smaller values of  $S_p$  help in reducing invalid pixel ratios. This indicates a trade-off in parameter selection between minimizing error and reducing invalid pixel ratios.





**Figure 15.** 3D visualization of error-based on prior window size  $S_w$  and patch size  $S_p$ , with point size indicating the reciprocal of invalid pixel ratios.

## 6. Conclusions and Future Work

In this paper, we introduced a novel interpolation method to address invalid regions in SGM disparity maps using MAP estimation. By leveraging SGM outputs as prior probabilities and incorporating cosine similarity from surrounding regions, our approach effectively enhances the visual quality of disparity maps, particularly in challenging areas with large invalid regions, while suppressing irrelevant outliers. Although the interpolation may not always achieve exact numerical accuracy, the visually coherent disparity maps produced by our method enhance the performance of neural network models by providing more realistic and consistent training data. Moreover, the experimental results demonstrate the robustness of the proposed method in enhancing disparity maps, making it highly applicable to real-world environments where ground truth is unavailable. For future work, we aim to expand our ground truth enhancement methods to accommodate a broader range of real-world scenarios and challenges. Additionally, we plan to explore more sophisticated neural network architectures and optimization strategies to further enhance the accuracy and efficiency of disparity estimation for autonomous and real-time applications.

**Author Contributions:** Conceptualization, S.W.; Software, H.G. and S.R.; Validation, H.G.; Writing—original draft, H.G.; Writing—review & editing, S.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2023-00248854).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

## References

1. Peng, R.; Wang, R.; Wang, Z.; Lai, Y.; Wang, R. Rethinking Depth Estimation for Multi-View Stereo: A Unified Representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8645–8654.
2. Fan, R.; Wang, L.; Bocus, M.J.; Pitas, I. Computer stereo vision for autonomous driving. *arXiv* **2020**, arXiv:2012.03194.
3. Cui, Y.; Chen, R.; Chu, W.; Chen, L.; Tian, D.; Li, Y.; Cao, D. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 722–739. [[CrossRef](#)]

4. Rajpal, A.; Cheema, N.; Illgner-Fehns, K.; Slusallek, P.; Jaiswal, S. High-Resolution Synthetic rgb-d Datasets for Monocular Depth Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1188–1198.
5. Toschi, M.; De Matteo, R.; Spezialetti, R.; De Gregorio, D.; Di Stefano, L.; Salti, S. Relight my Nerf: A Dataset for Novel View Synthesis and Relighting of Real World Objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 20762–20772.
6. Ummenhofer, B.; Agrawal, S.; Sepulveda, R.; Lao, Y.; Zhang, K.; Cheng, T.; Richter, S.; Wang, S.; Ros, G. Objects With Lighting: A Real-World Dataset for Evaluating Reconstruction and Rendering for Object Relighting. In Proceedings of the 2024 International Conference on 3D Vision (3DV), Davos, Switzerland, 18–21 March 2024; pp. 137–147.
7. Kallwies, J.; Engler, T.; Forkel, B.; Wuensche, H.J. Triple-SGM: Stereo Processing Using Semi-Global Matching with Cost Fusion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 192–200.
8. Hirschmuller, H. Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 2, pp. 807–814.
9. Jeong, J.C.; Shin, H.; Chang, J.; Lim, E.G.; Choi, S.M.; Yoon, K.J.; Cho, J.i. High-quality stereo depth map generation using infrared pattern projection. *ETRI J.* **2013**, *35*, 1011–1020. [[CrossRef](#)]
10. Xu, Y.; Yang, X.; Yu, Y.; Jia, W.; Chu, Z.; Guo, Y. Depth Estimation by Combining Binocular Stereo and Monocular Structured-Light. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1746–1755.
11. Li, H.; Chan, T.N.; Qi, X.; Xie, W. Detail-preserving multi-exposure fusion with edge-preserving structural patch decomposition. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 4293–4304. [[CrossRef](#)]
12. Ma, F.; Cavalheiro, G.V.; Karaman, S. Self-Supervised Sparse-to-Dense: Self-Supervised Depth Completion from Lidar and Monocular Camera. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 3288–3295.
13. Zhao, Y.; Krähenbühl, P. Real-Time Online Video Detection with Temporal Smoothing Transformers. In *Computer Vision—ECCV 2022, Proceedings of the 17th European Conference, Tel Aviv, Israel, 23–27 October 2022*; Springer: Cham, Switzerland, 2022; pp. 485–502.
14. Yao, P.; Zhang, H.; Xue, Y.; Zhou, M.; Xu, G.; Gao, Z. Iterative Color-Depth MST Cost Aggregation for Stereo Matching. In Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, USA, 11–15 July 2016; pp. 1–6.
15. Chai, Y.; Cao, X. Stereo Matching Algorithm Based on Joint Matching Cost and Adaptive Window. In Proceedings of the 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 12–14 October 2018; pp. 442–446.
16. Yang, S.; Lei, X.; Liu, Z.; Sui, G. An efficient local stereo matching method based on an adaptive exponentially weighted moving average filter in SLIC space. *IET Image Process.* **2021**, *15*, 1722–1732. [[CrossRef](#)]
17. Yang, J.; Wang, H.; Ding, Z.; Lv, Z.; Wei, W.; Song, H. Local stereo matching based on support weight with motion flow for dynamic scene. *IEEE Access* **2016**, *4*, 4840–4847. [[CrossRef](#)]
18. Bleyer, M.; Rother, C.; Kohli, P. Surface Stereo with Soft Segmentation. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1570–1577.
19. Yamaguchi, K.; McAllester, D.; Urtasun, R. Efficient Joint Segmentation, Occlusion Labeling, Stereo and Flow Estimation. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13; Springer: Cham, Switzerland, 2014; pp. 756–771.
20. Ulusoy, A.O.; Black, M.J.; Geiger, A. Semantic Multi-View Stereo: Jointly Estimating Objects and Voxels. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4531–4540.
21. Ye, X.; Gu, Y.; Chen, L.; Li, J.; Wang, H.; Zhang, X. Order-based disparity refinement including occlusion handling for stereo matching. *IEEE Signal Process. Lett.* **2017**, *24*, 1483–1487. [[CrossRef](#)]
22. Li, A.; Yuan, Z.; Ling, Y.; Chi, W.; Zhang, S.; Zhang, C. Unsupervised occlusion-aware stereo matching with directed disparity smoothing. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 7457–7468. [[CrossRef](#)]
23. Xie, Y.; Zeng, S.; Chen, L. A Novel Disparity Refinement Method Based on Semi-Global Matching Algorithm. In Proceedings of the 2014 IEEE International Conference on Data Mining Workshop, Shenzhen, China, 14 December 2014; pp. 1135–1142.
24. Yang, Q. A Non-Local Cost Aggregation Method for Stereo Matching. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1402–1409.
25. Sun, X.; Mei, X.; Jiao, S.; Zhou, M.; Liu, Z.; Wang, H. Real-time local stereo via edge-aware disparity propagation. *Pattern Recognit. Lett.* **2014**, *49*, 201–206. [[CrossRef](#)]
26. Hosni, A.; Bleyer, M.; Gelautz, M.; Rhemann, C. Local Stereo Matching Using Geodesic Support Weights. In Proceedings of the 2009 16th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, 7–10 November 2009; pp. 2093–2096.
27. Yang, Q.; Wang, L.; Yang, R.; Stewénius, H.; Nistér, D. Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 492–504. [[CrossRef](#)]

28. Sun, J.; Li, Y.; Kang, S.B.; Shum, H.Y. Symmetric Stereo Matching for Occlusion Handling. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 2, pp. 399–406.
29. Mozerov, M.G.; Van De Weijer, J. Accurate stereo matching by two-step energy minimization. *IEEE Trans. Image Process.* **2015**, *24*, 1153–1163. [[CrossRef](#)] [[PubMed](#)]
30. Heitz, F.; Bouthemy, P. Multimodal estimation of discontinuous optical flow using Markov random fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **1993**, *15*, 1217–1232. [[CrossRef](#)]
31. Yamaguchi, K.; Hazan, T.; McAllester, D.; Urtasun, R. Continuous Markov Random Fields for Robust Stereo Estimation. In Proceedings of the Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Proceedings, Part V 12; Springer: Cham, Switzerland, 2012; pp. 45–58.
32. Hosni, A.; Rhemann, C.; Bleyer, M.; Rother, C.; Gelautz, M. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 504–511. [[CrossRef](#)] [[PubMed](#)]
33. Wu, W.; Li, L.; Jin, W. Disparity Refinement Based on Segment-Tree and Fast Weighted Median Filter. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3449–3453.
34. Sun, X.; Mei, X.; Jiao, S.; Zhou, M.; Wang, H. Stereo Matching with Reliable Disparity Propagation. In Proceedings of the 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, Hangzhou, China, 16–19 May 2011; pp. 132–139.
35. Zhan, Y.; Gu, Y.; Huang, K.; Zhang, C.; Hu, K. Accurate image-guided stereo matching with efficient matching cost and disparity refinement. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *26*, 1632–1645. [[CrossRef](#)]
36. Mei, X.; Sun, X.; Zhou, M.; Jiao, S.; Wang, H.; Zhang, X. On Building an Accurate Stereo Matching System on Graphics Hardware. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 467–474.
37. Jiao, J.; Wang, R.; Wang, W.; Dong, S.; Wang, Z.; Gao, W. Local stereo matching with improved matching cost and disparity refinement. *IEEE Multimed.* **2014**, *21*, 16–27. [[CrossRef](#)]
38. Zhang, K.; Lu, J.; Yang, Q.; Lafruit, G.; Lauwereins, R.; Van Gool, L. Real-time and accurate stereo: A scalable approach with bitwise fast voting on CUDA. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 867–878. [[CrossRef](#)]
39. Stentoumis, C.; Grammatikopoulos, L.; Kalisperakis, I.; Karras, G. On accurate dense stereo-matching using a local adaptive multi-cost approach. *ISPRS J. Photogramm. Remote Sens.* **2014**, *91*, 29–49. [[CrossRef](#)]
40. Miclea, V.C.; Nedeveschi, S. Real-time semantic segmentation-based stereo reconstruction. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 1514–1524. [[CrossRef](#)]
41. Chen, Z.; Dong, P.; Li, Z.; Yao, R.; Ma, Y.; Fang, X.; Deng, H.; Zhang, W.; Chen, L.; An, F. Real-Time FPGA-Based Binocular Stereo Vision System with Semi-Global Matching Algorithm. In Proceedings of the 2021 IEEE 34th International System-on-Chip Conference (SOCC), Las Vegas, NV, USA, 14–17 September 2021; pp. 158–163.
42. Jin, S.; Cho, J.; Dai Pham, X.; Lee, K.M.; Park, S.K.; Kim, M.; Jeon, J.W. FPGA design and implementation of a real-time stereo vision system. *IEEE Trans. Circuits Syst. Video Technol.* **2009**, *20*, 15–26.
43. Cambuim, L.F.; Oliveira, L.A., Jr.; Barros, E.N.; Ferreira, A.P. An FPGA-based real-time occlusion robust stereo vision system using semi-global matching. *J. Real-Time Image Process.* **2020**, *17*, 1447–1468. [[CrossRef](#)]
44. Zhang, Q.; Xu, L.; Jia, J. 100+ Times Faster Weighted Median Filter (WMF). In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2830–2837.
45. Kim, S.; Min, D.; Kim, S.; Sohn, K. Unified confidence estimation networks for robust stereo matching. *IEEE Trans. Image Process.* **2018**, *28*, 1299–1313. [[CrossRef](#)] [[PubMed](#)]
46. Chao, W.; Wang, X.; Wang, Y.; Wang, G.; Duan, F. Learning sub-pixel disparity distribution for light field depth estimation. *IEEE Trans. Comput. Imaging* **2023**, *9*, 1126–1138. [[CrossRef](#)]
47. Zbontar, J.; LeCun, Y. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.* **2016**, *17*, 1–32.
48. Wang, Y.; Yang, Y.; Yang, Z.; Zhao, L.; Wang, P.; Xu, W. Occlusion Aware Unsupervised Learning of Optical Flow. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4884–4893.
49. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 3273–3282.
50. Chang, J.R.; Chen, Y.S. Pyramid Stereo Matching Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5410–5418.
51. Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; Tan, P. Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2495–2504.
52. Bangunharcana, A.; Cho, J.W.; Lee, S.; Kweon, I.S.; Kim, K.S.; Kim, S. Correlate-and-Excite: Real-Time Stereo Matching via Guided Cost Volume Excitation. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 3542–3548.

53. Wang, F.; Galliani, S.; Vogel, C.; Pollefeys, M. Itermv: Iterative Probability Estimation for Efficient Multi-View Stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8606–8615.
54. 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, Vancouver, BC, Canada, 17–24 June 2023; pp. 21919–21928.
55. Wang, Y.; Tao, X.; Qi, X.; Shen, X.; Jia, J. Image Inpainting via Generative Multi-Column Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 31 (NeurIPS 2018), Montréal, QC, Canada, 3–8 December 2018.
56. Ren, J.S.; Xu, L.; Yan, Q.; Sun, W. Shepard Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015), Montréal, QC, Canada, 7–12 December 2015.
57. Chen, H.; Zhao, Y. Don't Look into the Dark: Latent Codes for Pluralistic Image Inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 7591–7600.
58. Zhu, M.; He, D.; Li, X.; Li, C.; Li, F.; Liu, X.; Ding, E.; Zhang, Z. Image inpainting by end-to-end cascaded refinement with mask awareness. *IEEE Trans. Image Process.* **2021**, *30*, 4855–4866. [[CrossRef](#)]
59. Li, W.; Lin, Z.; Zhou, K.; Qi, L.; Wang, Y.; Jia, J. Mat: Mask-Aware Transformer for Large Hole Image Inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10758–10768.
60. Zhang, Z.; Wu, B.; Wang, X.; Luo, Y.; Zhang, L.; Zhao, Y.; Vajda, P.; Metaxas, D.; Yu, L. AVID: Any-Length Video Inpainting with Diffusion Model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 7162–7172.
61. Liu, H.; Wang, Y.; Qian, B.; Wang, M.; Rui, Y. Structure Matters: Tackling the Semantic Discrepancy in Diffusion Models for Image Inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 8038–8047.
62. Wei, C.; Mangalam, K.; Huang, P.Y.; Li, Y.; Fan, H.; Xu, H.; Wang, H.; Xie, C.; Yuille, A.; Feichtenhofer, C. Diffusion Models as Masked Autoencoders. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 16284–16294.
63. Li, X.; Guo, Q.; Abdelfattah, R.; Lin, D.; Feng, W.; Tsang, I.; Wang, S. Leveraging Inpainting for Single-Image Shadow Removal. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 13055–13064.
64. Sargsyan, A.; Navasardyan, S.; Xu, X.; Shi, H. Mi-gan: A Simple Baseline for Image Inpainting on Mobile Devices. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 7335–7345.
65. 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 Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4040–4048.
66. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
67. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth  $16 \times 16$  words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

**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.