remote sensing

MDPI

Communication

Attention-Based Monocular Depth Estimation Considering Global and Local Information in Remote Sensing Images

Junwei Lv [1,2,3], Yueting Zhang [1,2,3,*], Jiayi Guo [1,2,3], Xin Zhao [1,2,3], Ming Gao [1,2,3] and Bin Lei [1,2]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; lvjunwei19@mails.ucas.ac.cn (J.L.); guojy@aircas.ac.cn (J.G.); zhaoxin195@mails.ucas.ac.cn (X.Z.); gaoming19@mails.ucas.ac.cn (M.G.); leibin@mail.ie.ac.cn (B.L.)
[2] Key Laboratory of Technology in Geo-Spatial Information Processing and Application Systems, Chinese Academy of Sciences, Beijing 100190, China
[3] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China
* Correspondence: zhangyueting06@mails.ucas.ac.cn

Abstract: Monocular depth estimation using a single remote sensing image has emerged as a focal point in both remote sensing and computer vision research, proving crucial in tasks such as 3D reconstruction and target instance segmentation. Monocular depth estimation does not require multiple views as references, leading to significant improvements in both time and efficiency. Due to the complexity, occlusion, and uneven depth distribution of remote sensing images, there are currently few monocular depth estimation methods for remote sensing images. This paper proposes an approach to remote sensing monocular depth estimation that integrates an attention mechanism while considering global and local feature information. Leveraging a single remote sensing image as input, the method outputs end-to-end depth estimation for the corresponding area. In the encoder, the proposed method employs a dense neural network (DenseNet) feature extraction module with efficient channel attention (ECA), enhancing the capture of local information and details in remote sensing images. In the decoder stage, this paper proposes a dense atrous spatial pyramid pooling (DenseASPP) module with channel and spatial attention modules, effectively mitigating information loss and strengthening the relationship between the target's position and the background in the image. Additionally, weighted global guidance plane modules are introduced to fuse comprehensive features from different scales and receptive fields, finally predicting monocular depth for remote sensing images. Extensive experiments on the publicly available WHU-OMVS dataset demonstrate that our method yields better depth results in both qualitative and quantitative metrics.

Keywords: remote sensing; depth estimation; monocular vision; attention

Citation: Lv, J.; Zhang, Y.; Guo, J.; Zhao, X.; Gao, M.; Lei, B. Attention-Based Monocular Depth Estimation Considering Global and Local Information in Remote Sensing Images. *Remote Sens.* **2024**, *16*, 585. https://doi.org/10.3390/rs16030585

Academic Editor: Claudio Piciarelli

Received: 29 December 2023
Revised: 24 January 2024
Accepted: 2 February 2024
Published: 4 February 2024

1. Introduction

With the continuous advancements in drone and satellite technology, there has been a notable enhancement in the quality of remote sensing images. Despite these improvements, a significant limitation persists, given that remote sensing images predominantly exist in a 2D format, lacking essential depth information about the depicted scenes or objects. The integration of depth information is paramount for a comprehensive understanding of the 3D structure inherent in remote sensing scenes. This inclusion facilitates the determination of object distances and the establishment of spatial relationships within the environment. The indispensability of depth information is particularly evident in precise tasks such as 3D reconstruction [1,2], land-cover classification [3–5], and land-change detection [6,7].

Traditional geometry-based methods [8–14] and learning-based methods [12,15–17] can predict depth from images but encounter formidable challenges in implementation. Factors such as high computational costs, the limited availability of remote sensing images for a specific area, and other constraints pose significant obstacles. Multi-view stereo

vision, for example, demands the prediction of camera poses [14] and substantial data from the same scene, which are often scarce in practical remote sensing scenes. Stereo vision methods relying on feature point matching [18] encounter difficulties in low-texture areas and regions with intense specular reflection, leading to compromised depth estimation and 3D reconstruction accuracy. Hence, the exploration of monocular vision depth estimation tasks [19] is of paramount research significance. By deriving depth information from a single view, these approaches offer a promising avenue to surmount the challenges inherent in traditional methods, presenting a more effective solution for remote sensing depth estimation.

Monocular depth estimation stands as a pivotal task in the realm of computer vision, with the objective of deducing depth information for each pixel using a solitary image. To date, several methods have been proposed for monocular depth estimation [20–29].

Nevertheless, it is noteworthy that these methods have predominantly undergone extensive experimentation and validation on public datasets like the KITTI autonomous driving dataset [30] or the NYU Depth V2 indoor scene dataset [31]. Yet, the direct application of these methods to remote sensing scenes is challenging due to their substantial differences. Remote sensing scenes, in contrast to autonomous driving or indoor datasets, exhibit distinctive characteristics. They encompass significantly larger coverage areas than indoor scenes, and the distribution of scenes in autonomous driving datasets tends to be more uniform, featuring similar image structures and styles. In remote sensing scenes, however, the distribution of foreground and background is uncertain. Moreover, these scenes present unique challenges, including occlusion and environmental complexity, which are not easily captured through training on the aforementioned datasets. Additionally, in remote sensing scenarios, there are typically extensive low-texture areas, within which there exist some targets that are not easily distinguishable, among other characteristics. Consequently, adapting existing monocular depth estimation methods for remote sensing necessitates innovative approaches that account for these divergences and challenges.

Building upon the identified challenges and existing research, we present a novel solution to address the complexities associated with remote sensing images. Our proposed approach involves the development of an end-to-end monocular depth estimation method, strategically considering both global and local information. In this framework, we introduce multiple modules that integrate attention mechanisms, thereby showcasing the efficacy of estimating depth from a single image in remote sensing scenes. Our key contributions can be succinctly summarized as follows:

- We propose a monocular depth estimation method for remote sensing scenes, employing an encoder–decoder structure. To enhance depth estimation, we introduce attention mechanisms in both the encoder and decoder, facilitating the improved extraction and fusion of global and local information.
- In the encoder, we introduce a dense neural network with an efficient channel attention mechanism as a feature extractor for remote sensing images, effectively enhancing the efficiency of feature extraction.
- In the decoder, we introduce a dense atrous spatial pyramid pooling module with a convolution block attention module and a global plane guidance module to extract crucial contextual information. Furthermore, a weight allocation mechanism is implemented for each global plane guidance module to enhance the learning of feature and channel importance, contributing to more accurate depth estimation.

The rest of this paper is organized as follows. In Section 2, we introduce the related work. Section 3 introduces the materials and methods in detail. The experimental data and results are displayed in Section 4. Specific discussions are carried out in Section 5, and Section 6 concludes the paper.

2. Related Work

Monocular depth estimation in remote sensing scenes stands as a pivotal research area within the domains of computer vision and remote sensing. While the majority of studies

on monocular depth estimation have been conducted on datasets tailored for outdoor environments, such as the KITTI dataset for autonomous driving, and indoor scenes like the NYU Depth V2 dataset, there remains a scarcity of methods explicitly designed for remote sensing depth estimation. This section aims to offer a succinct overview of traditional geometry-based methods and deep learning-based monocular stereo vision methods, shedding light on techniques dedicated to addressing the unique challenges posed by depth estimation in remote sensing scenes.

2.1. Traditional Geometry-Based Methods

Heiko et al. [9] proposed a semi-global matching method to obtain depth maps of images. But this method is computationally intensive and slow in execution. Saxena et al. [11,12] introduced a technique that employs Markov Random Fields to learn the mapping between input images and output depth. Yet, this approach relies on artificial prior models and is difficult to apply to real-world data. Depth can also be obtained from shadows [13] using geometric relationships, but this method can lead to information loss and significant depth errors in complex scenes. In Structure from Motion (SFM) [14] methods, feature points are identified from multiple images, and then depth information is calculated. This approach is susceptible to interference from moving objects in the scene and can result in substantial depth estimation errors due to correspondences and occlusions.

The aforementioned traditional geometry-based methods, if applied to remote sensing scenes, due to the larger and more complex nature of the scenes, bring about increased computational and temporal costs. Moreover, poor depth results may be yielded due to the complexity of the scenes or the inability to extract feature points in low-texture areas.

2.2. Deep Learning-Based Monocular Methods

Eigen et al. [20] pioneered the utilization of deep convolution neural networks for single-image depth estimation, laying the groundwork for subsequent developments in monocular depth estimation using neural networks. Subsequent refinements by Eigen et al. [21], involving a unified multi-scale network framework with a deeper structure, and Liu et al. [22], introducing a deep convolution neural field to address depth probability optimization, encountered challenges like gradient explosion and vanishing gradients [32]. Laina et al. [23] innovatively applied deep residual networks for depth estimation, inspiring the exploration of DenseNet [33] in the same domain. However, the increased number of parameters in these networks posed significant challenges. Dai et al. [27] introduced the fully convolutional network framework into depth estimation, while Hu et al. [29] proposed a network structure with a multi-scale feature fusion strategy to enhance edge estimation. Cao et al. [24] treated depth estimation as a pixel classification problem, resulting in some information loss during image feature extraction and imprecise outcomes.

Addressing this limitation, Bhat et al. [26] integrated transformers into depth estimation, incorporating attention mechanisms. Chen et al. [28] proposed a structure-aware residual pyramid network and later introduced an improved version. Lee et al. [25] introduced a multi-scale guidance plane method to integrate features at different scales. Shim et al. [34] proposed a monocular depth estimation method employing a self-attention mechanism that showed commendable results on an autonomous driving dataset with a limited depth range. However, this method is not applicable to remote sensing scenes.

For UAV images, Madhuanand et al. [35] proposed a 3D decoder that integrates depth and pose estimation to predict depth. Hermann et al. [36] introduced a self-supervised method for monocular depth prediction in UAV images. Additionally, Chang et al. [37] presented an approach utilizing an encoder–decoder architecture for monocular depth estimation in low-altitude UAV images. To adapt to the characteristics of remote sensing images, Tao et al. [38] introduced a feature pyramid approach for smoke detection in remote sensing scenes. Nonetheless, in the presence of smoke, it becomes challenging to determine the depth of the scene beneath the smoke.

3. Materials and Methods

3.1. Overall Architecture

This paper proposes a monocular depth estimation method for remote sensing images that considers both global and local information. In this method, we introduce attention mechanisms to enhance the learning and fusion of features. The overall architecture is illustrated in Figure 1. The inputs are aerial remote sensing images captured from an oblique perspective. Initially, the images are fed into the encoder, which primarily consists of a dense feature extraction module with attention mechanisms. The dense neural network in the encoder significantly reduces information loss during propagation, promotes feature reuse, and enhances feature representation. The attention mechanism focuses on details and regions of interest in remote sensing images during feature extraction. After feature extraction, the features are input into the decoder, where an attention-enhanced dense atrous spatial pyramid pooling (DenseASPP) module is employed. This module, utilizing atrous convolutions, expands the effective receptive fields of feature maps, obtaining multi-scale information beneficial for complex scenes in remote sensing images. It integrates different-scale feature maps to enhance global scene understanding, ensuring that the network focuses on both global and important local features. Finally, weighted plane guidance modules are introduced to derive different plane guidance from features at different scales. Each plane guidance contributes to the final depth prediction, providing significant constraints. Since features from different scales have varying importance, the network learns the weights for the plane guidance, resulting in monocular depth estimation that closely approximates the true depth in a single-view remote sensing image.

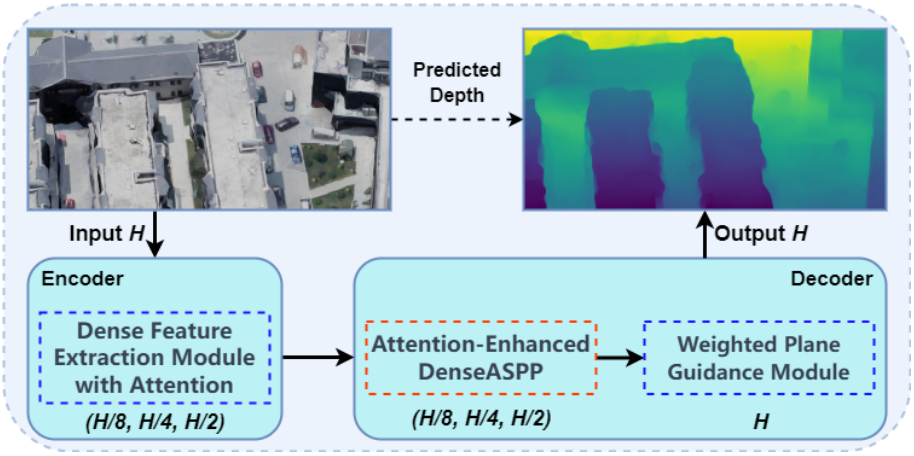

Figure 1. The overall architecture of our work. This architecture comprises an encoder and a decoder, with the input size of the remote sensing image set at 768 × 384. The encoder incorporates a dense feature extraction module with an attention mechanism. In the decoder, the attention-enhanced DenseASPP module and the weighted plane guidance module collaborate to predict the depth for the input remote sensing image.

3.2. Dense Feature Extraction Module with Attention

In the encoder part, we have devised a dense neural network (DenseNet) with efficient channel attention named ECA-DenseNet, as shown on the left side of Figure 2, specifically tailored for feature extraction in remote sensing images.

DenseNet exhibits a densely connected neural network structure, where each layer directly receives feature maps from the preceding layer. This design enhances the reuse of feature maps at different positions, ensuring comprehensive information transfer in each forward pass and the maximal preservation of global features. For remote sensing images, which often contain numerous objects, rich semantic information, and extensive spatial coverage, global features are crucial for the overall effectiveness. Therefore, we chose DenseNet as the foundational network framework.

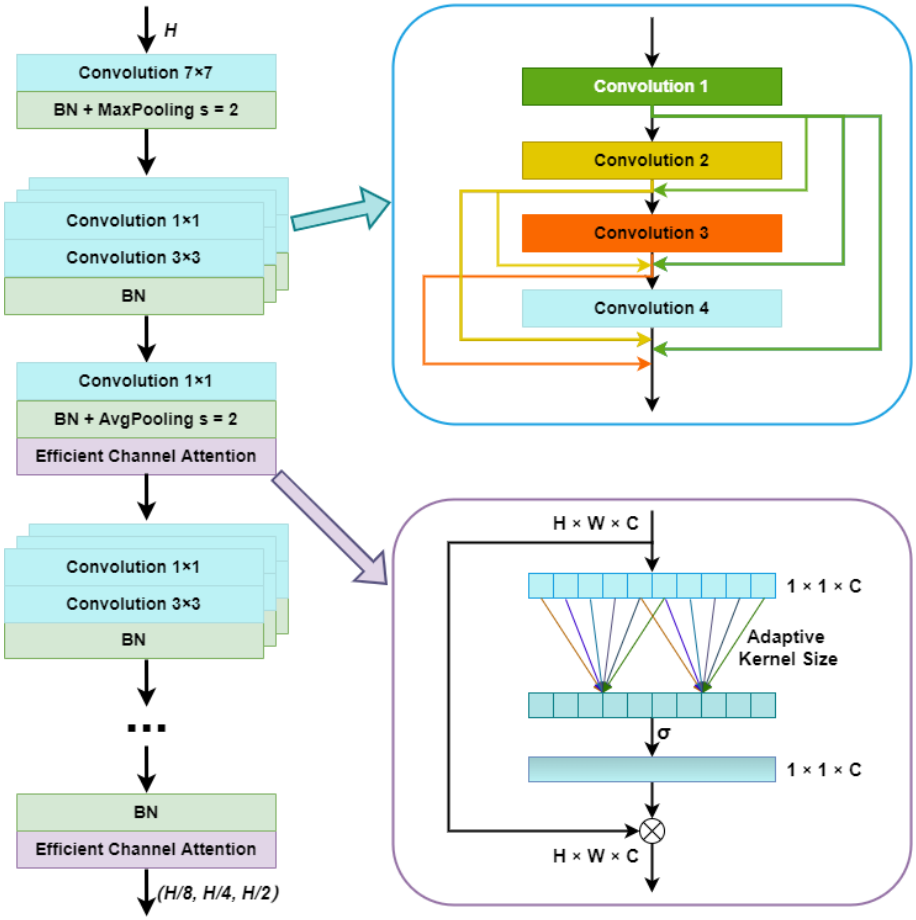

Figure 2. The network architecture of ECA-DenseNet. The left side illustrates the overall structure of the network, the top-right corner provides a schematic diagram outlining the internal connections within each network block, and the bottom-right corner presents a schematic diagram of the ECA model.

While global features are essential, the significance of local information and focused attention to specific details in remote sensing images cannot be understated. To address this, we introduced an attention mechanism.

The efficient channel attention (ECA) mechanism enhances the attention of different channels to features, as depicted in the bottom-right corner of Figure 2. It initially reduces the spatial dimensions to provide each channel with a scalar representation, and the channel weights are adaptively learned through the convolution kernel size and the number of channels. Notably, an attention module has been inserted after each dense block in the network structure, performing the channel attention calculation on the feature maps outputted by each block. This facilitates the rapid and effective identification of noteworthy positional features during each stage of feature extraction.

The design of ECA-DenseNet ensures that, during feature extraction in remote sensing images, it not only emphasizes global features for overall information preservation but also swiftly captures key details. Simultaneously considering both global and local attention information enhances the expressive power for edge features of terrain and complex scenes in remote sensing images. From a training perspective, the dense connection aids in better gradient propagation, significantly alleviating issues such as gradient vanishing and exploding. The ECA, requiring only global average pooling, proves computationally efficient and less prone to overfitting.

3.3. Attention-Enhanced DenseASPP

To better leverage the image features extracted by the feature extractor, we introduce a DenseASPP module that integrates both channel attention and spatial attention, forming a contextual information connection for feature extraction and depth estimation, as shown in Figure 3.

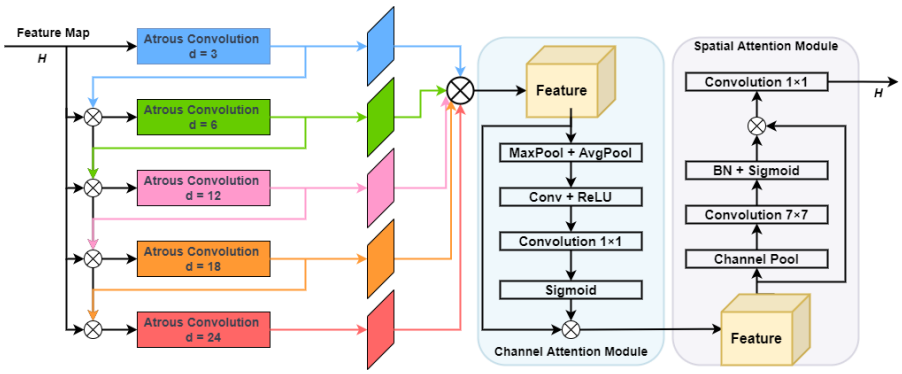

Figure 3. The architecture of attention-enhanced DenseASPP.

To account for global structures, we employ atrous convolution modules, which expand the effective receptive field of the convolution kernel. This expansion enables each kernel to capture information from a larger distance, ranging from a small to a large scale. This approach proves beneficial for addressing long-range dependencies in the image, particularly suited for remote sensing images. Simultaneously, the atrous convolution maintains the resolution of the feature map, aiding subsequent tasks. By adjusting the dilation rate, dilated convolution obtains feature representations with different receptive field sizes, capturing multi-scale features that comprehensively represent the image and provide rich information for the subsequent network. Importantly, dilated convolution does not introduce additional parameters when enlarging the receptive field, avoiding an increase in the training burden or difficulty compared to other architectures.

Similar to the feature extractor, we implement dense connections in this section through DenseASPP. Dense connections facilitate efficient feature transfer and reuse by directly transmitting features from the previous layers to subsequent layers. Additionally, dense connections contribute to a more compact and efficient network structure. In contrast to regular network propagation, DenseASPP enhances learning capabilities during training, effectively addressing gradient-vanishing issues and enhancing the model's expressive and generalization capabilities.

Our approach involves a comprehensive consideration of both global and local structures, enhancing the learning of local features through the incorporation of channel and spatial attention modules. In the channel attention module, we perform channel dimension pooling within the feature block, followed by convolution and sigmoid layers to generate weights for each channel. Simultaneously, employing spatial attention addresses the high spatial complexity of remote sensing images, concentrating on significant spatial positions by deriving weights for each position through channel pooling and sigmoid activation. This strategic combination reinforces the representation of crucial positions in the image.

In summary, our proposed attention-enhanced DenseASPP effectively interprets and infers complex scenes and targets in remote sensing images. It accomplishes this by accurately expanding the global receptive field while intensifying attention toward key image locations.

3.4. Weighted Plane Guidance Module

To enhance the accuracy of predictions in the final depth output, we introduce a weighted plane guidance module to constrain the depth predictions before the ultimate prediction, as illustrated in Figure 4. Each plane guidance module outputs features with the same dimension H as the input image. Under the combined influence of the previous

plane guidance module and the features of different scales, guidance planes with features of varying scales are obtained. Prior to the final depth prediction, multiple guidance planes are weighted and aggregated, contributing to the ultimate depth estimation, thereby enhancing the accuracy of the results.

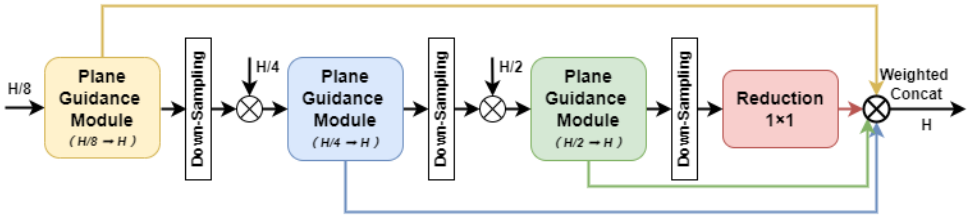

Figure 4. The architecture of the weighted plane guidance module.

The aforementioned module outputs a 4D vector (n_1, n_2, n_3, n_4). Utilizing ray-plane intersection [39], we can construct the local depth corresponding to the 4D vector according to Equation (1). Different scales correspond to distinct guiding planes, each encapsulating information from various scales. Depth estimation under global information emphasizes coarser scales, while depth estimation under local information primarily focuses on learning at finer scales.

$$\tilde{d}_i = \frac{n_4\sqrt{u_i^2 + v_i^2 + 1}}{n_1 u_i + n_2 v_i + n_3}, \tag{1}$$

where $\tilde{d}_i$ is the local depth, and (u_i, v_i) are the normalized coordinates of the pixels.

To accommodate both global and local information within the features of remote sensing images, we assign weights to the guidance planes from different scales. As the network undergoes training, it learns the significance of various guidance planes for depth estimation. Ultimately, the weighted guidance planes are concatenated and contribute to the final depth prediction layer.

The incorporation of the weighted plane guidance module serves to effectively establish the relationship between internal features and the final depth estimation, providing a robust constraint to the prediction process.

3.5. Loss Function

The scale-invariant error (SIE) is the proportional error between the depth estimation error and the true depth value. Its definition is as follows:

$$\mathbf{SIE} = \frac{1}{N}\sum_i (\log \hat{D}_i - \log D_i)^2 - \frac{\lambda}{N^2}\left(\sum_i \log \hat{D}_i - \log D_i\right)^2, \tag{2}$$

where N represents the number of the images, D_i is the ground truth of the depth, $\hat{D}_i$ is the predicted depth, and $\lambda = 0.5$. The **SIE** is often used as a loss function to address scale uncertainty issues in depth estimation tasks. Because this function is independent of the absolute scale of the depth values, even if the depth values are scaled proportionally, the scale-invariant error remains consistent. This is particularly useful for handling diverse scenes in remote sensing images and different datasets.

This error equation can also be reformulated as follows:

$$\mathbf{SIE} = \frac{1}{N}\sum_i (\log \hat{D}_i - \log D_i)^2 - \frac{1}{N^2}\left(\sum_i \log \hat{D}_i - \log D_i\right)^2 + (1-\lambda)\left(\frac{1}{N}(\sum_i \log \hat{D}_i - \log D_i)\right)^2 \tag{3}$$

This equation offers a more transparent breakdown of the error components, encompassing the weighted square mean of the variance and the error in logarithmic space. The

parameter λ provides flexibility in adjusting the emphasis on minimizing variance in the error. In summary, our loss is expressed as follows:

$$Loss_{SIE} = \text{Min}(c\sqrt{\mathbf{SIE}}), \tag{4}$$

where λ is set to 0.85, and we use $c = 10$ as a constant.

4. Experiments and Results

4.1. Datasets

The WHU-OMVS dataset [40,41], developed by Wuhan University, serves as a dataset for oblique aerial images tailored for city-scale 3D scene reconstruction tasks and monocular depth estimation in remote sensing imagery. Captured at an approximate flight altitude of 220 m above the ground, the dataset includes extensive urban remote sensing images, along with the corresponding depths. The dataset comprises a total of 32,175 images obtained by slicing large-scale remote sensing images, and each image has dimensions of 768 $\times$ 384 pixels and is subject to a certain amount of actual capture jitter. We divided these images into training and testing sets with a ratio of 40:1, following the partitioning scheme used in the KITTI and NYU datasets. We ensured that the scenes in the testing set did not appear in the training set to maintain independent scenario settings.

The LEVIR-NVS dataset [42] is a newly proposed dataset for remote sensing images. The dataset comprises 16 scenes, including mountains, cities, schools, stadiums, colleges, and more, with each scene containing 21 multi-view images and corresponding depth maps. The LEVIR-NVS dataset is used as the test dataset for our method.

4.2. Implementation Details

Our model underwent training and testing on an NVIDIA RTX 3090 GPU, utilizing the PyTorch framework. Throughout the training process, the Adam optimizer was employed with a learning rate set to 10^{-4}, accompanied by a weight decay rate of 10^{-2}, and a batch size set to 8. Furthermore, epsilon $\epsilon = 10^{-3}$ was incorporated into the optimizer to mitigate potential division by zero in the denominator, thereby enhancing stability in computations. Both training and testing images were of the size 768 $\times$ 384. For testing, we employed several quality assessment metrics, as outlined below.

4.3. Quality Assessment Metrics

Similar to various monocular depth estimation methods, we adopted metrics from prior work [20] as our evaluation criteria for convenient comparison with other studies.

The Average Relative Error (Abs Rel) represents the average depth error. The Squared Relative Error (Sq Rel) is utilized to measure the relative error of the depth estimation model concerning the depth values. The Root-Mean-Squared Error (RMSE) gauges the difference between predicted values and actual values. The RMSE accounts for the variance of prediction errors and is more sensitive to large errors. The Root-Mean-Squared Error in log space (RMSE log) is employed to assess the model's performance by calculating the square root in logarithmic space. The RMSE log is more sensitive to depth errors than the RMSE.

Smaller values for these metrics indicate a more effective depth estimation by the model. The metrics are specifically defined as follows:

$$\text{Abs Rel} = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i - \hat{D}_i|}{D_i} \tag{5}$$

$$\text{Sq Rel} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i - \hat{D}_i}{D_i}\right)^2 \tag{6}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(D_i - \hat{D}_i)^2} \tag{7}$$

$$\text{RMSE log} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(D_i) - \log(\hat{D}_i))^2} \tag{8}$$

The Accuracy Threshold measures the accuracy for a certain threshold range and is defined as follows:

$$\text{Accuracy Threshold} = \%\text{ of } D_i \text{ s.t. } \text{Max}\left(\frac{D_i}{\hat{D}_i}, \frac{\hat{D}_i}{D_i}\right) < thr, \text{ where } thr = 1.25,\ 1.25^2,\ 1.25^3 \tag{9}$$

In the equations above, N represents the number of images, D_i is the ground truth of depth, and $\hat{D}_i$ is the predicted depth. In Equation (9), thr represents the threshold of the measuring range.

4.4. Results

4.4.1. Quantitative and Qualitative Evaluation Compared with Other Methods

We assessed our method from qualitative and quantitative perspectives against other monocular depth estimation methods, including Monodepth [17], BTS [25], and Adabins [26].

In Table 1 above, we present the results of experiments conducted with these methods on the WHU-OMVS dataset. We calculated depth estimation outcomes using the aforementioned metrics for quantitative analysis and comparison. Table 1 clearly indicates that our method demonstrated superior quantitative performance.

Table 1. Quality assessment metrics comparison on WHU-OMVS dataset.

Model Name	Abs Rel ↓	Sq Rel ↓	RMSE ↓	$\text{RMSE}_{\log}$ ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Monodepth	0.214	8.142	20.191	0.259	0.761	0.925	0.967
BTS	0.127	2.859	15.028	0.218	0.855	0.961	0.983
Adabins	0.130	2.719	13.881	0.184	0.867	0.965	0.985
Ours	**0.085**	**1.427**	**9.605**	**0.134**	**0.928**	**0.981**	**0.991**

The best results are highlighted in bold in the table.

Furthermore, we conducted qualitative visual comparisons among various methods using the WHU-OMVS dataset, employing reference depth for evaluation.

Upon an overall assessment of the depth, each method exhibits satisfactory performance. However, upon closer scrutiny, our results demonstrate superior estimates in terms of details and edges as shown in Figure 5. For example, it is evident that our depth results provide clearer and more refined delineation of object contours, as well as depth level differentiation on surfaces of buildings or other targets. Notably, in the top row within the red box, our method accurately detects and estimates the depth of the outermost trees along the edges. In the middle-position boxes, our method successfully estimates the position and depth of ground vehicles, a capability not matched by other methods. Additionally, in the red box in the fourth row, a region with a hole in the remote sensing image results in a deeper depth value, a nuanced detail captured only by our method. These qualitative results affirm that our method not only effectively estimates the global depth of remote sensing images but also excels in capturing the depth of endpoints and details within the images.

Additionally, we conducted tests of our model on LEVIR-NVS remote sensing dataset. Table 2 presents the quantitative metric results, and Figure 6 illustrates the qualitative visual effects. Due to the poor generalization of other methods in remote sensing images, we cannot obtain depth maps that can distinguish scenes. Therefore, we only present our results compared with the ground truth. Based on the testing results obtained from the

LEVIR-NVS dataset, it can be observed that our metrics surpass those of other methods, indicating the effectiveness of our approach for improving remote sensing scenes. Visually, our testing yields depth estimates for some of the important regional targets within the scene, demonstrating that our attention mechanism is operational, hence validating the efficacy and superiority of our method. However, it is also apparent that the generalization of our approach requires further enhancement.

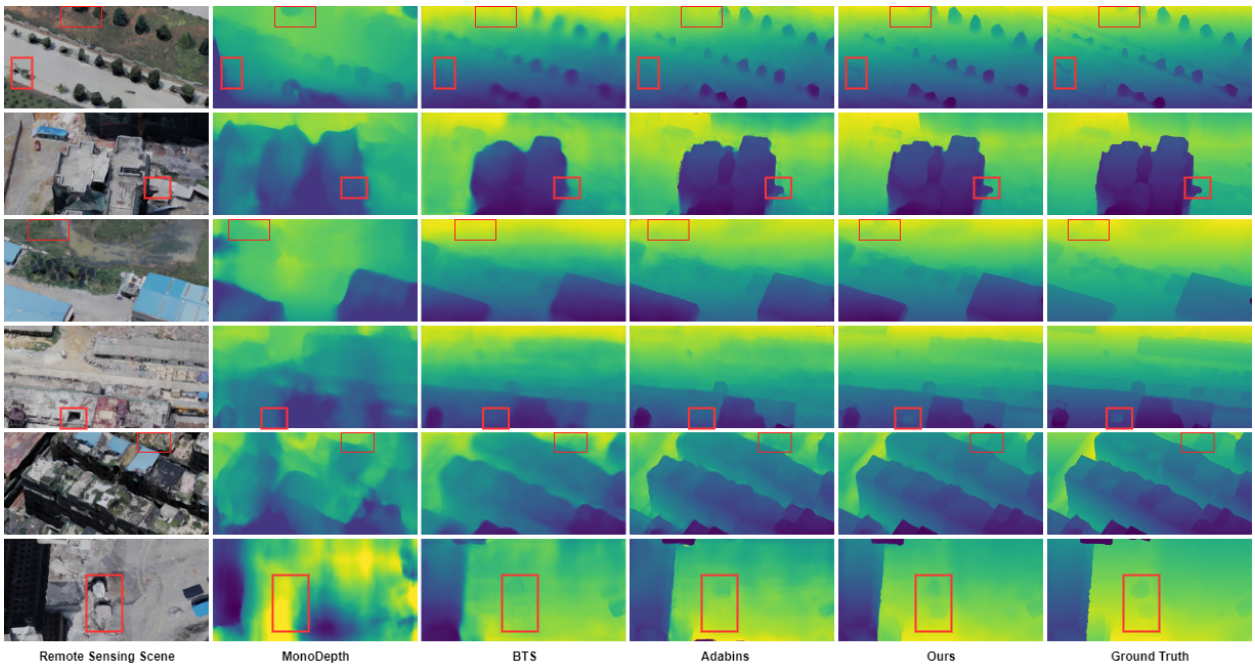

Figure 5. Qualitative comparison on WHU-OMVS dataset.

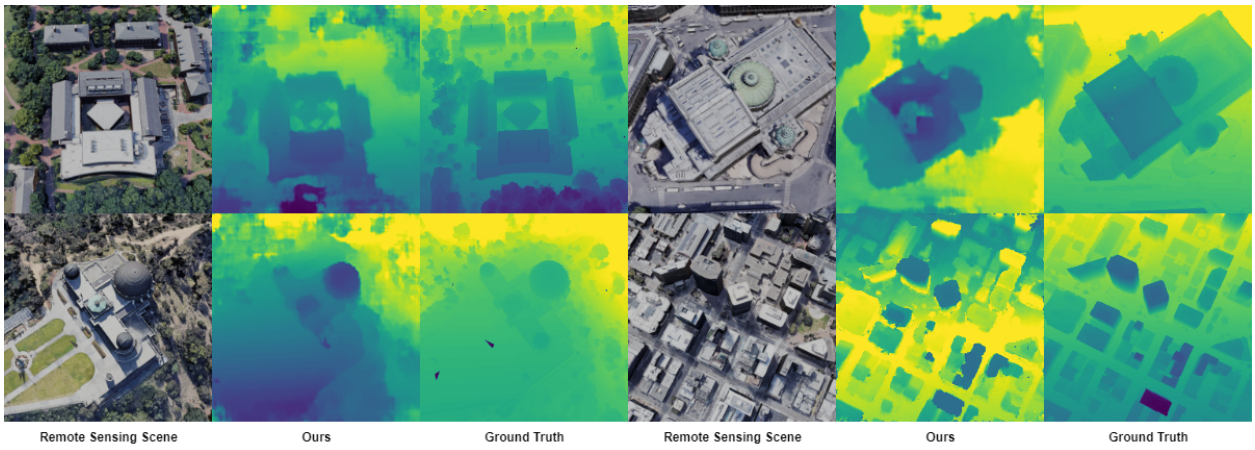

Figure 6. Qualitative comparison with models trained on WHU-OMVS dataset and tested on LEVIR-NVS dataset.

Table 2. Quality assessment metrics comparison with models trained on WHU-OMVS dataset and tested on LEVIR-NVS dataset.

Model Name	Abs Rel ↓	Sq Rel ↓	RMSE ↓	$RMSE_{log}$ ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Monodepth	0.710	0.695	10.978	0.351	0.539	0.831	0.924
BTS	0.649	0.630	10.357	0.282	0.588	0.874	0.936
Adabins	0.597	0.613	10.286	0.311	0.601	0.887	0.939
Ours	**0.526**	**0.567**	**10.009**	**0.260**	**0.703**	**0.921**	**0.959**

The best results are highlighted in bold in the table.

In summary, the results establish that our method performs exceptionally well on complex remote sensing images, validating the superiority and effectiveness of our approach.

4.4.2. Ablation Study

In the previously described methodology, we formulated a feature extractor using DenseNet with an attention mechanism, a contextual connector employing DenseASPP that integrates channel attention and spatial attention, and a depth estimation constraint through a weighted plane guidance module. In this section, we present various ablation experiments conducted to validate the efficacy of each module in depth estimation and assess the corresponding improvements in metrics. Our approach was compared with the baseline BTS, which lacks an attention mechanism. In this baseline configuration, we utilized DenseNet161 without attention, standard ASPP, and unweighted plane guidance.

Table 3 encompasses various configurations involving the three modules within the network. Through the examination of depth estimation metrics, it is evident that each module makes a distinct contribution, resulting in improvements in the outcomes. This observation underscores the effectiveness of our method in monocular depth estimation for remote sensing images.

Table 3. Quality assessment metrics from ablation study of our method on WHU-OMVS dataset.

Model Name	Abs Rel ↓	Sq Rel ↓	RMSE ↓	$\text{RMSE}_{\log}$ ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Baseline	0.127	2.859	15.028	0.218	0.855	0.961	0.983
Model with ED	0.107	1.937	11.925	0.157	0.899	0.973	0.989
Model with AD	0.108	2.035	11.958	0.161	0.897	0.972	0.987
Model with WPG	0.107	2.004	11.939	0.156	0.900	0.974	0.989
Model without ED	0.099	1.654	11.004	0.148	0.909	0.978	0.991
Model without AD	0.095	1.580	10.701	0.144	0.913	0.978	0.991
Model without WPG	0.097	1.622	10.962	0.147	0.911	0.978	0.991
Complete model	**0.085**	**1.427**	**9.605**	**0.134**	**0.928**	**0.981**	**0.991**

ED: ECA-DenseNet; AD: attention-enhanced DenseASPP; WPG: weighted plane guidance. The results of the complete model are highlighted in bold in the table.

5. Discussion

This paper proposes a monocular depth estimation method for remote sensing images. In order to achieve more accurate depth estimation results, we introduce several novel modules: the ECA-DenseNet feature extraction module, the attention-based DenseASPP contextual information fusion module, and the weighted plane guidance module. The introduction of these modules follows the principle of simultaneously considering global and local feature information.

Extensive experiments were conducted on the WHU-OMVS dataset, and quantitative metrics and qualitative results were compared with those of the monocular depth estimation methods MonoDepth [17], BTS [25], and Adabins [26]. As shown in the results, MonoDepth produces decent results in some simple scenes but fails to distinguish the position and depth in complex scenes. Although BTS performs well globally, it struggles with depth discrimination in detailed locations, and the contours in the depth are not clear. Adabins exhibits good performance in depth contours and certain objects, but it suffers from depth loss, such as exceeding the applicable depth range, and fails to identify and predict depth for specific small targets.

In contrast, our results are closest to the ground truth. Our method exhibits the clearest depth hierarchy in the global structure, discerns depth changes at the edges, and demonstrates excellent depth estimation in detail.

As observed from the visual results in Figure 5, our method is capable of feature extraction and depth estimation even at the edges of remote sensing scenes, exhibiting satisfactory global visual effects. We note that in remote sensing scenarios, such as rooftops or large areas of uniform buildings, extensive low-texture regions are common. Within

these regions, small targets that are difficult to detect can be found, as in the scenes from the third and fifth rows. Our approach accurately identifies and estimates depths in low-texture areas, showing notable improvements for remote sensing imagery compared to other methods. This success can be attributed to the contributions of our method, which continuously enhances global information throughout the process while also integrating attention mechanisms into each module to give extra focus to important targets. Comparative and ablation experiments indicate the effectiveness of the proposed modules, showcasing their applicability to remote sensing images.

Furthermore, we have also employed the LEVIR-NVS dataset for testing. Overall, our test results are able to generate corresponding depths at crucial locations, reflecting the role of the attention mechanism. However, there is still substantial room for improvement in the overall visual effects, and hence, we are considering how to enhance the generalization performance of our model in future work.

In summary, our method provides a valuable solution for monocular depth estimation in remote sensing images, offering potential applications in 3D reconstruction and aiding tasks such as remote sensing target detection and segmentation.

In future work, we will consider improving remote sensing depth estimation in binocular or multi-view vision to enhance the accuracy in 3D reconstruction tasks. Additionally, we will continue to delve into monocular remote sensing depth estimation to further improve the method's generalization performance.

6. Conclusions

This paper introduces an attention-based monocular depth estimation method for remote sensing images, taking into account both global and local information. The input image undergoes feature extraction through a dense neural network equipped with an attention mechanism, effectively capturing both global and local details. The connector of the dense atrous spatial pyramid pooling module efficiently conveys the global structural features of remote sensing images. By incorporating channel attention and spatial attention, it enhances the focus on critical positions within complex remote sensing scenes. The weighted plane guidance module integrates guidance planes from different scales, providing depth constraints from various features and ultimately producing the depth corresponding to a single remote sensing image. Comprehensive ablation experiments on the WHU-OMVS public dataset validate the effectiveness of the proposed methodological framework. The results compared with other methods also prove the superiority and effectiveness of our method and its potential benefits for remote sensing 3D reconstruction and target detection and segmentation tasks.

Author Contributions: Conceptualization, J.L., J.G., Y.Z. and X.Z.; methodology, J.L.; validation, J.L.; investigation, J.L.; writing—original draft preparation, J.L.; writing—review and editing, J.L., J.G., Y.Z., X.Z. and M.G.; supervision, J.G., Y.Z. and B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The National Natural Science Foundation of China, grant numbers 61991421 and 61991420, and the Key Research and Development Program of Aerospace Information Research Institute Chinese Academy of Sciences, grant number E1Z208010F.

Data Availability Statement: The data presented in this study are available in article.

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

References

1. Geiger, A.; Ziegler, J.; Stiller, C. Stereoscan: Dense 3d reconstruction in real-time. In Proceedings of the 2011 IEEE Intelligent Vehicles Symposium (IV), Baden-Baden, Germany, 5–9 June 2011; pp. 963–968.
2. Remondino, F. Heritage recording and 3D modeling with photogrammetry and 3D scanning. *Remote Sens.* **2011**, *3*, 1104–1138. [CrossRef]

3. Lv, Z.; Zhang, P.; Sun, W.; Benediktsson, J.A.; Li, J.; Wang, W. Novel Adaptive Region Spectral–Spatial Features for Land Cover Classification with High Spatial Resolution Remotely Sensed Imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5609412. [CrossRef]
4. Immitzer, M.; Vuolo, F.; Atzberger, C. First experience with Sentinel-2 data for crop and tree species classifications in central Europe. *Remote Sens.* **2016**, *8*, 166. [CrossRef]
5. Hu, F.; Xia, G.S.; Hu, J.; Zhang, L. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]
6. Lv, Z.; Huang, H.; Sun, W.; Jia, M.; Benediktsson, J.A.; Chen, F. Iterative Training Sample Augmentation for Enhancing Land Cover Change Detection Performance With Deep Learning Neural Network. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 1–14. [CrossRef] [PubMed]
7. Lv, Z.; Huang, H.; Li, X.; Zhao, M.; Benediktsson, J.A.; Sun, W.; Falco, N. Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective. *Proc. IEEE* **2022**, *110*, 1976–1991. [CrossRef]
8. Sun, C. Fast stereo matching using rectangular subregioning and 3D maximum-surface techniques. *Int. J. Comput. Vis.* **2002**, *47*, 99–117. [CrossRef]
9. Hirschmuller, H. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 328–341. [CrossRef]
10. Zhang, S. *High-Resolution, Real-Time 3-D Shape Measurement*; Stony Brook University: Stony Brook, NY, USA, 2005.
11. Saxena, A.; Chung, S.H.; Ng, A.Y. Learning Depth from Single Monocular Images. In Proceedings of the 18th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 5–8 December 2005.
12. Saxena, A.; Schulte, J.; Ng, A.Y. Depth estimation using monocular and stereo cues. In Proceedings of the 20th International Joint Conference on Artifical Intelligence, Hyderabad, India, 6–12 January 2007.
13. Zhang, R.; Tsai, P.S.; Cryer, J.E.; Shah, M. Shape from Shading: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 690–706. [CrossRef]
14. Schönberger, J.L.; Frahm, J.M. Structure-from-Motion Revisited. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4104–4113. [CrossRef]
15. Yao, Y.; Luo, Z.; Li, S.; Fang, T.; Quan, L. Mvsnet: Depth inference for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 767–783.
16. Dai, J.; He, K.; Sun, J. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1635–1643.
17. Godard, C.; Mac Aodha, O.; Brostow, G.J. Unsupervised monocular depth estimation with left-right consistency. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 270–279.
18. Tareen, S.A.K.; Saleem, Z. A comparative analysis of sift, surf, kaze, akaze, orb, and brisk. In Proceedings of the 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 3–4 March 2018; pp. 1–10.
19. Bhoi, A. Monocular depth estimation: A survey. *arXiv* **2019**, arXiv:1901.09402.
20. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
21. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2650–2658.
22. Liu, F.; Shen, C.; Lin, G.; Reid, I. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 2024–2039. [CrossRef]
23. Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; Navab, N. Deeper depth prediction with fully convolutional residual networks. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 239–248.
24. Cao, Y.; Wu, Z.; Shen, C. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 3174–3182. [CrossRef]
25. Lee, J.H.; Han, M.K.; Ko, D.W.; Suh, I.H. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv* **2019**, arXiv:1907.10326.
26. Bhat, S.F.; Alhashim, I.; Wonka, P. Adabins: Depth estimation using adaptive bins. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 4009–4018.
27. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 379–387.
28. Chen, X.; Chen, X.; Zha, Z.J. Structure-aware residual pyramid network for monocular depth estimation. *arXiv* **2019**, arXiv:1907.06023.
29. Hu, J.; Ozay, M.; Zhang, Y.; Okatani, T. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1043–1051.
30. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]

31. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. 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: Berlin/Heidelberg, Germany, 2012; pp. 746–760.
32. Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; JMLR Workshop and Conference Proceedings; pp. 249–256.
33. Iandola, F.; Moskewicz, M.; Karayev, S.; Girshick, R.; Darrell, T.; Keutzer, K. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv* **2014**, arXiv:1404.1869.
34. Shim, K.; Kim, J.; Lee, G.; Shim, B. Depth-Relative Self Attention for Monocular Depth Estimation. *arXiv* **2023**, arXiv:2304.12849.
35. Madhuanand, L.; Nex, F.; Yang, M.Y. Self-supervised monocular depth estimation from oblique UAV videos. *ISPRS J. Photogramm. Remote Sens.* **2021**, *176*, 1–14. [CrossRef]
36. Hermann, M.; Ruf, B.; Weinmann, M.; Hinz, S. Self-supervised learning for monocular depth estimation from aerial imagery. *arXiv* **2020**, arXiv:2008.07246.
37. Chang, R.; Yu, K.; Yang, Y. Self-Supervised Monocular Depth Estimation Using Global and Local Mixed Multi-Scale Feature Enhancement Network for Low-Altitude UAV Remote Sensing. *Remote Sens.* **2023**, *15*, 3275. [CrossRef]
38. Tao, H. Smoke Recognition in Satellite Imagery via an Attention Pyramid Network With Bidirectional Multi-Level Multi-Granularity Feature Aggregation and Gated Fusion. *IEEE Internet Things J.* **2023**. [CrossRef]
39. Haines, E. Fast ray-convex polyhedron intersection. *Graph. Gems II* **1991**, 247–250. . [CrossRef]
40. Liu, J.; Ji, S. A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6050–6059.
41. Liu, J.; Gao, J.; Ji, S.; Zeng, C.; Zhang, S.; Gong, J. Deep learning based multi-view stereo matching and 3D scene reconstruction from oblique aerial images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *204*, 42–60. [CrossRef]
42. Wu, Y.; Zou, Z.; Shi, Z. Remote Sensing Novel View Synthesis with Implicit Multiplane Representations. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5627613. [CrossRef]