photonics

MDPI

Article

Self-Supervised Monocular Depth Estimation Based on Channel Attention

Bo Tao [1,2,*], Xinbo Chen [1,2], Xiliang Tong [3], Du Jiang [4] and Baojia Chen [5]

[1] Key Laboratory of Metallurgical Equipment and Control Technology, Ministry of Education, Wuhan University of Science and Technology, Wuhan 430081, China; chenxinbo@wust.edu.cn
[2] Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering, Wuhan University of Science and Technology, Wuhan 430081, China
[3] Precision Manufacturing Institute, Wuhan University of Science and Technology, Wuhan 430081, China; tongxiliang@wust.edu.cn
[4] Research Center for Biomimetic Robot and Intelligent Measurement and Control, Wuhan University of Science and Technology, Wuhan 430081, China; jiangdu@wust.edu.cn
[5] Hubei Key Laboratory of Hydroelectric Machinery Design and Maintenance, China Three Gorges University, CTGU, Yichang 443005, China; cbjia@163.com
* Correspondence: taoboq@wust.edu.cn

Abstract: Scene structure and local details are important factors in producing high-quality depth estimations so as to solve fuzzy artifacts in depth prediction results. We propose a new network structure that combines two channel attention modules in a deep prediction network. The structure perception module (spm) uses a frequency channel attention network. We use frequencies from different perspectives to analyze the channel representation as a compression process. This enhances the perception of the scene structure and obtains more feature information. The detail emphasis module (dem) adopts the global attention mechanism. It improves the performance of deep neural networks by reducing irrelevant information and magnifying global interactive representations. Emphasizing important details effectively fuses features at different scales to achieve more accurate and clearer depth predictions. Experiments show that our network produces clearer depth estimations, and our accuracy rate on the KITTI benchmark has improved from 98.1% to 98.3% in the $\delta < 1.25^3$ metric.

Keywords: monocular depth estimation; deep learning; channel attention; self-supervision

Citation: Tao, B.; Chen, X.; Tong, X.; Jiang, D.; Chen, B. Self-Supervised Monocular Depth Estimation Based on Channel Attention. *Photonics* **2022**, *9*, 434. https://doi.org/10.3390/photonics9060434

Received: 15 May 2022
Accepted: 16 June 2022
Published: 20 June 2022

1. Introduction

Accurately estimating the depth from a single image is a basic task in computer vision, which enables computers to understand real scenes. It has been widely used in robotics navigation [1], autonomous driving [2] and augmented reality [3] for generating high-quality depth from colors instead of using expensive LIDAR sensors. Although monocular cameras are cheap and lightweight, the task of depth estimation is still challenging to the traditional SfM (Structure from motion) algorithms. Recently, supervised methods [4–8] have achieved some success. Nevertheless, they rely on a large amount of ground truth, which can only be obtained sparingly by expensive LiDAR sensors [9]. Self-supervised methods use geometric constraints as the only source of supervision for monocular videos [10] or simultaneous stereo image pairs [11].

In recent years, due to its powerful ability to extract data features and express complex relationships, deep learning has attracted more attention in solving traditional monocular depth estimation problems. Deep learning [12–19] is committed to mining the hidden rules of data from large datasets and then using the learned rules to predict the results. It is expected that the models obtained through learning will have a good generalization ability. Convolutional neural networks are used to extract semantic features that contain structural information of the scene.

The local detail feature is another feature that emphasizes the boundaries of objects and tries to produce clear depth maps. U-Net [20] is the framework used by most depth estimation networks, in which the decoder fuses features based on concatenation and basic convolution. We find that these operations fail to retain sufficient detail and accurately obtain spatial information, resulting in the inefficient integration of features at different levels and blurred artifacts in regions of depth discontinuity.

We propose a new monocular depth estimation network based on the attention mechanism, which has different channel attention modules above the depth network to enhance the performance of features by capturing more contextual information about the scene geometry. In this paper, a deep convolutional neural network model for monocular depth estimation is proposed, and the overall architecture of the network has the form of an encoder–decoder. The encoder completes the feature extraction process, and the decoder completes the feature output process. The main contributions of our work are as follows:

(1) A new network architecture is proposed, which combines two channel attention modules in the depth prediction network to capture more contextual information of the scene and emphasize detailed features.
(2) The spm is based on frequency channel attention to enhance the perception of the scene structure and obtain more feature information. The dem is based on the channel attention mechanism to efficiently fuse features at different scales and emphasize important details to obtain clearer depth estimates.
(3) The superior performance of the proposed method is validated on the KITTI benchmark and the Make 3D dataset.

2. Related Work

2.1. Supervised Depth Estimation

Deep learning has been continuously developed, and it has shown a relatively high performance in image processing, such as image classification [21], target detection [22], semantic segmentation [23–29], etc. Estimating depth is an inherently uncertain problem because there may be multiple seemingly reasonable depths for the pixels in an image. Recently, supervised methods have shown their ability to predict models. They can correctly estimate the depth of color images. Various supervised methods based on deep learning are continuously being explored. However, they require high-quality real depth of the ground truth, which can be expensive.

Eigen et al. [4] used multiscale neural networks for dense pixel depth estimation, with one estimating a coarse global depth prediction and the other estimating a local fine prediction generated by the first network. Eigen et al. [5] improved the framework of prediction depth, surface normal vectors and semantic tags and deepened the network structure. Laina et al. [6] proposed a full convolution network which uses a residual network to extract global information and designed residual upsampling blocks to improve the quality of depth estimation. Many methods have turned depth estimation into a classification problem. Cao et al. [7] divided the continuous depth values into several boxes and classified them by CNN. Fu et al. [8] proposed a discretization strategy of increasing distance to discretize the depth and reformulated deep network learning as an ordered regression problem to prevent the loss of over-reinforcement. The methods of supervision require annotated datasets, and the cost of obtaining these datasets is very high.

2.2. Self-Supervised Monocular Depth Estimation

Due to the limitations of supervised methods, the self-supervised approach was developed. Godard [11] extended the reconstruction constraint by the loss of parallax smoothness and the loss of left–right depth consistency. Zhou et al. [10] proposed a method for learning depth and self-motion from monocular videos by training a depth estimation network and an individual attitude network. Tosi et al. [30] introduced a traditional depth estimation method, which uses the inverse Huber loss and image reconstruction loss. Wong et al. [31] proposed a new objective function to elucidate the bilateral cyclic relationship

between the left and right parallax and introduced an adaptive regularization scheme to deal with co-visible and occluded regions in stereo pairs. The full CNN framework proposed in [32] uses monocular images and the corresponding optical flow to estimate the exact depth map. In [33], to improve the consistency of the estimated depth and the self-motion between consecutive frames, 3D constraints were proposed. Guizilini et al. [34] learned to compress and unpack in order to decompress by symmetric packing, thus preserving detailed features. The state-of-the-art framework proposed by Godard et al. [35] is Monodepth2, which introduces minimal reprojection loss to handle occlusion and automatic masking schemes to robustly remove invalid pixels.

2.3. Self-Attention Mechanism

Self-attention mechanisms are used in monocular depth estimation. Wang et al. [36] modeled spatiotemporal correlations between video sequences and images by aggregating global contexts specific to each location of the query. Zhang et al. [37] learned a better image generator by incorporating the self-attention mechanism into the GAN framework. Fu et al. [38] designed two types of attention modules to enhance the feature representation ability of scene segmentation. Johnston et al. [39] captured the background of similar disparity values in discontinuous areas by exploring the similarity of features in spatial dimensions.

Compared to previous work, we demonstrate that more relative depth information is obtained from more distant regions due to different channels. We can obtain a better depth estimation performance by capturing globally relevant information along the channel dimension and distinguishing different features.

3. Self-Supervised Depth Estimation and Network Models

3.1. Network Model

3.1.1. Attention U-Net Architecture

In this paper, the network architecture combining a depth network with a pose network is adopted. In this case, the depth network uses an encoder and decoder architecture integrated with U-Net [20] and ResNet18 [40]. The input is a single frame of an image at a certain moment. The pose network uses the same encoder as the depth network, and three neighboring frames are used as inputs. The pose network adopts PoseNet without mask prediction [10]. The output is the predicted depth of each pixel, and the overall structure of the network model is shown in Figure 1.

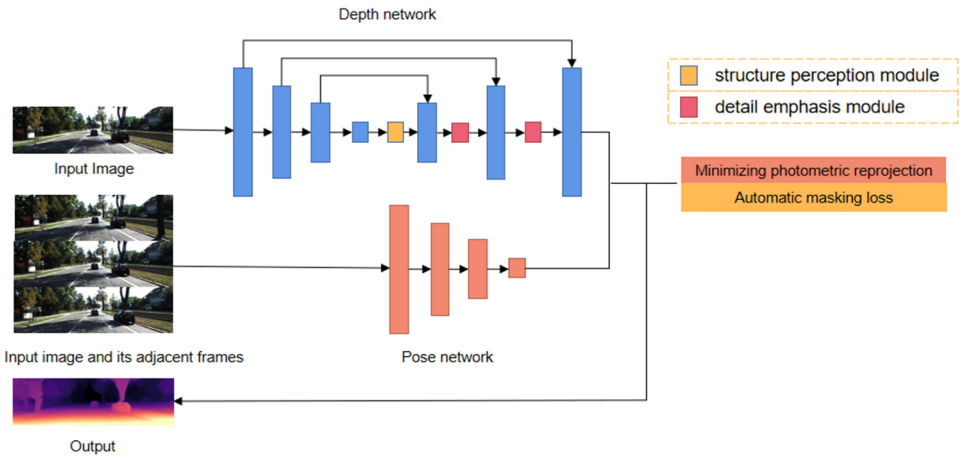

Figure 1. Overall network architecture.

As shown in Figure 2, the U-Net [20] network architecture is a lightweight all-convolution neural network which includes two parts: the encoder and decoder. We use a residual network. The last layer of the encoder connects the spm, and skip connections are used to facilitate the gradient and information flow throughout the model. The

encoder connects the dem through skip connections to obtain a clear depth map. Finally, we use the nearest neighbor interpolation to successively sample the predicted depth map at multiple scales until the resolution of the original input is reached. In addition, the training loss is calculated using the higher input resolution.

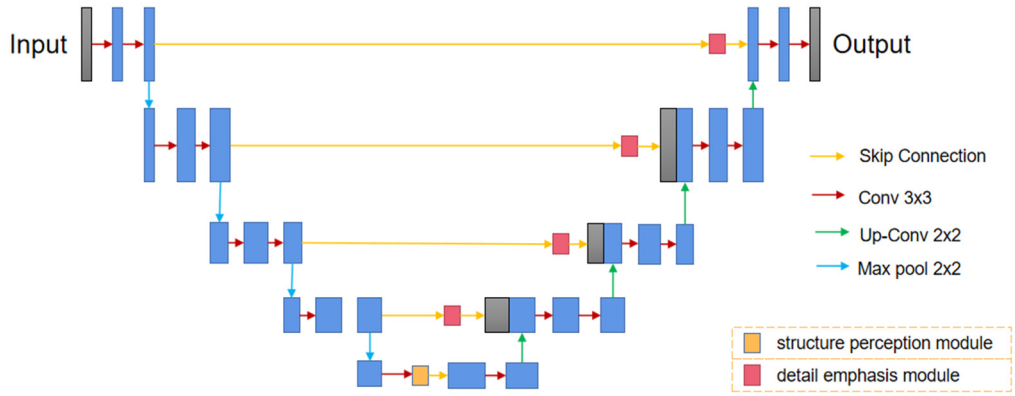

Figure 2. The structure of the depth network with channel attention added.

3.1.2. Depth Network

Tables 1 and 2 show the encoder and decoder structure of the depth network. Every two convolutional blocks form a building block with the following structure: Conv (3×3)-BN-ReLU-Conv (3×3)-BN. The BN (Batch Normalization) layer is the batch normalization layer. In the whole model, each convolutional kernel has a convolutional step of 2. The output part of the convolutional kernel uses the ReLU activation function uniformly. Corresponding to the encoder, the decoder part of the depth network uses a convolutional kernel with a step size of 1 and a size of 3×3. Here are the symbols in the tables: k, kernel size; s, step size; c, number of output channels; activation, activation function. The encoder block is represented by econv.

Table 1. The structure of the encoder network.

Layers	k	s	c	Activation
convl	7	2	64	ReLU
maxpool	3	2	64	-
econv1	3	1	64	ReLU
econv2	3	2	128	ReLU
econv3	3	2	256	ReLU
econv4	3	2	512	ReLU
spm	3	-	512	-

Table 2. The structure of the decoder network.

Layers	k	s	c	Activation
upconv6	3	1	512	ELU
dem	3	1	512	ReLU
iconv6	3	1	512	ELU
upconv5	3	1	256	ELU
dem	3	1	256	ReLU
iconv5	3	1	256	ELU
disp5	3	1	1	Sigmoid
upconv4	3	1	128	ELU
dem	3	1	128	ReLU
iconv4	3	1	128	ELU
disp4	3	1	1	Sigmoid

Table 2. *Cont.*

Layers	k	s	c	Activation
upconv3	3	1	64	ELU
dem	3	1	64	ReLU
iconv3	3	1	64	ELU
disp3	3	1	1	Sigmoid
upconv2	3	1	32	ELU
dem	3	1	32	ReLU
iconv2	3	1	32	ELU
disp2	3	1	1	Sigmoid
upconv1	3	1	16	ELU
dem	3	1	16	ReLU
iconv1	3	1	16	ELU
disp1	3	1	1	Sigmoid

3.2. Structure Perception Module

GAP (global average pooling) is an average operation which can be regarded as the simplest frequency spectrum of the input, that is, the component with a frequency of 0. However, it is not good to only use separated gap information in the channel attention. Since the separate spectrum is not good, several more frequency components can be considered, which introduce a multispectral channel attention mechanism.

Based on theoretical analysis, GAP is a special form of *2DDCT* (two-dimensional discrete cosine transform), which is proportional to the lowest frequency component in *2DDCT*. We can see that using GAP in the channel attention mechanism means that only the information from the lowest frequency is kept. All the components from the other frequencies are discarded, which also encode useful information patterns representing the channel and should not be lost.

In order to use more information in the input characteristic graph X, we can use multiple frequency components, including GAP. First, we take the input characteristic graph X and divide it into multiple groups according to the channel: $[X_0, X_1, \ldots, X_{n-1}]$. The original number of channels is C. After the division, the number of channels in each group is C', and $C' = C/n$, C should be a multiple of n. For each group, a specific frequency component of *2DDCT* is assigned, and this specific frequency component needs to be chosen in advance.

Thus, the *2DDCT* can be used as a preprocessing of channel attention. Inspired by [41], we apply the channel attention module to weight different local features. As shown in Figure 3, the channel attention module tries to learn and model the correlations between different channel mappings. We put this attention module at the beginning of the decoder part to integrate global information into local features and improve the representation of local features.

$$\begin{aligned} &Freq^i = 2DDCT^{u.v}(X^I), \\ &= \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} X^i_{:,\,h,w} B^{u,v}_{h,w} \\ &s.t.i \in \{0, 1, \ldots, n-1\}, \end{aligned} \tag{1}$$

Here, $[u, v]$ is the *2D* index of the frequency component, $Freq^i$ is a C'-dimensional vector. The overall preprocessing vector is the one that splices all:

$$Freq = cat([Freq^0, Freq^1, \ldots, Freq^{n-1}]), \tag{2}$$

Here, $Freq$ is the obtained multispectral vector, and the channel attention of this multispectral can be written as:

$$ms\_att = sigmoid(fc(Freq)). \tag{3}$$

It can be seen that different combinations of $[u, v]$ may be used for different groups, that is, different frequency components can be used for each group. In this way, the single spectrum of the GAP is expanded into multiple spectra.

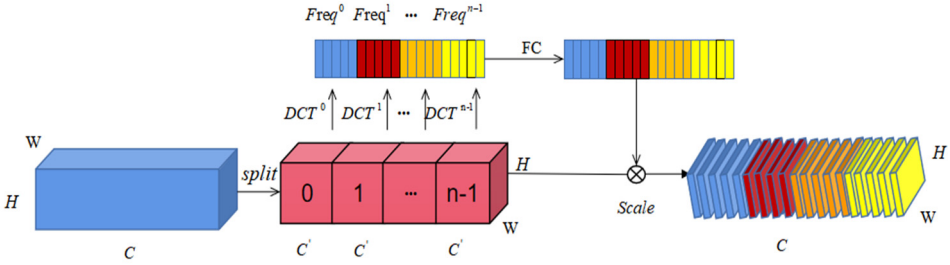

Figure 3. Structure perception module.

3.3. Detail Emphasis Module

The jump connects the low-level information of the encoder, while the high-level information contains richer spatial details. Further processing of local details is crucial and cannot be simply fused, which contains semantic differences between features of different levels. It leads to blurred artifacts in the predicted depth map. We predict some sharp edges by processing local details. The network can easily recover accurate depth predictions. Therefore, by using a channel focus mechanism, the network is enabled to focus on specific channel features. Inspired by [42], the detail emphasis module allows us to highlight some important details and effectively integrate features at different scales. 3D alignment is used to retain 3D information. Then, two-layer MLP (Multilayer Perceptron) is used to magnify the spatial correlation of the cross-dimensional channel. (MLP is an encoder–decoder structure with a compression ratio of r, which is the same as for BAM.) As shown in Figure 4, given an input feature map $F_1 \in R^{CxHxW}$ and an output, F_2 is defined as:

$$F_2 = M_C(F_1) \otimes F_1 \tag{4}$$

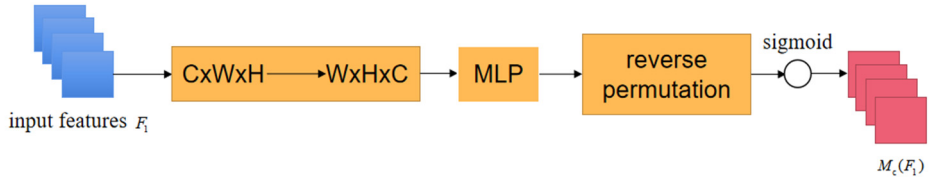

Figure 4. Detail emphasis module.

3.4. Loss Function

The minimized photometric reprojection is used to deal with the object occlusion problem, while the automatic masking loss is used to deal with the artifacts caused by motion. The per-pixel minimum photometric reprojection error L_p is experssed as:

$$L_p = \min_{t'} pe(I_t, I_{t'\to t}) \tag{5}$$

$$I_{t'\to t} = I_{t'} \langle proj(D_t, T_{t'\to t}, K) \rangle \tag{6}$$

where I_t is the target image, $I_{t'}$ is the source image, $T_{t'\to t}$ represents the camera pose of each source image with respect to the target image and pe is the photometric error, consisting of $SSIM$ (Structural Similarity index) and $L1$ loss. The expression is:

$$pe(I_t, I_{t'\to t}) = \partial \frac{1 - SSIM(I_t, I_{t'\to t})}{2} + (1 - \partial)||I_t - I_{t'\to t}||_1 \tag{7}$$

The *SSIM* is used to measure the similarity of the target image to the estimated image. The *L*1 loss is used to make the difference between the pixel values in the target image and those in the estimated image and take the absolute value. The value of ∂ is set to 0.85.

In this paper, an automatic masking method is used to solve the artifact problem of processing images. The pixel mask loss μ is applied to the masking loss by selectively weighting the pixels, $\mu \in \{0,1\}$ is calculated automatically during the forward pass of the network. The function expression is:

$$\mu = [\min pe(I_t, I_{t' \to t}) < \min pe(I_t, I_{t'})] \tag{8}$$

where [] is the Iverson bracket.

Furthermore, an edge-aware smoothness regularization term L_s is used in order to regularize the differences in the texture-free regions.

$$L_s = |\partial_x d_t^*| e^{-|\partial_x I_t|} + |\partial_y d_t^*| e^{-|\partial_y I_t|} \tag{9}$$

where d_t^* is the average normalized inverse depth to stop the estimated depth from shrinking.

The final loss function L is a weight sum of photometric loss L_p and smoothness loss L_s:

$$L = \mu L_p + \lambda L_s \tag{10}$$

where λ is a constant with a value of 0.001. The value of λ is referred to Godard's setting, which is the most commonly used optimal parameter setting, and L_p is the combination of *SSIM* and *L*1.

4. Experiments and Analysis

Our approach is compared with previous related works on the KITTI dataset [9] to demonstrate that our proposed approach can improve the accuracy of model prediction and reduce the error in model estimation. An ablation study of the attention mechanism is also studied based on monodepth2.

4.1. Implementation Details

Our network model is implemented via PyTorch, and during training, 20 epochs are trained. The batchsize is set to 12, and the resolution of the input and output is 640 × 192. We use ResNet-18 pre-trained on ImageNet [43] as the encoder. For the first 15 epochs, we use the Adam [44] optimizer, with a learning rate of 10^{-4}. The learning rate for the last five epochs is 10^{-5}. The GPU is NVIDIA GeForce GTX 1080Ti, and the training duration is 12 h.

4.2. KITTI Results

We use the KITTI dataset for all experiments, using the data segmentation of Eigen et al. [4] to increase the distance to 80 m. Before training, the same preprocessing as that performed by Zhou et al. [10] is performed to remove some static frames. In the end, 39,810 frames are used for training, 4424 frames are used for verification and 697 frames are used for testing. When training these networks, the same camera inherent focal length and average focal length are used for all frames. As shown in Figure 5, for road signs and utility poles, our model yields clearer results. As shown in Table 3, our network outperforms other self-supervised methods in most metrics.

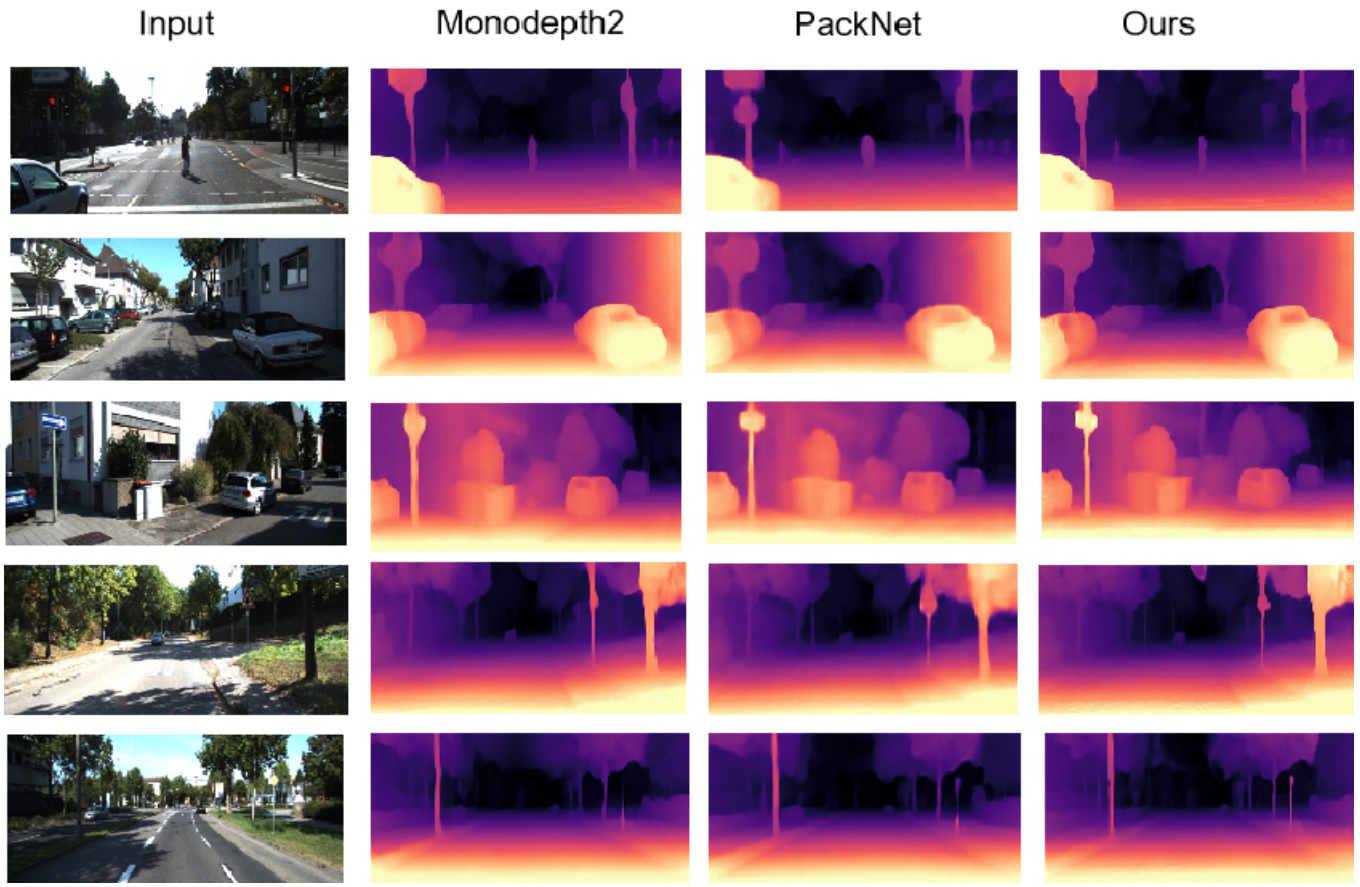

Figure 5. Qualitative results of KITTI feature splitting. In terms of boundary and object details, such as trees, pedestrians and signs, our model is better.

Table 3. Quantitative results. Comparison of our method with existing methods for the intrinsic splitting of methods used in KITTI 2015. The best effect categories for each method are shown in bold. Lower values are better for Abs Rel, Sq Rel, RMSE and RMSE log, and higher values are better for $\delta < 1.25$, $\delta < 1.25^2$ and $\delta < 1.25^3$. In the table, M: Self-supervised mono supervision.

Method	Train	Abs Rel	Sq Rel	RMSE	RMSE Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
DDVO [36]	M	0.151	1.257	5.583	0.228	0.810	0.936	0.974
DF-Net [45]	M	0.150	1.124	5.507	0.223	0.806	0.933	0.973
Ranjan [46]	M	0.148	1.149	5.464	0.226	0.815	0.935	0.973
EPC++ [47]	M	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2depth [48]	M	0.141	1.026	5.291	0.215	0.816	0.945	0.979
Monodepth [11]	M	0.124	1.388	6.125	0.217	0.818	0.929	0.966
SGDdepth [49]	M	0.117	0.907	4.844	0.196	0.875	0.954	0.979
Monodepth2 [35]	M	0.115	0.903	4.863	0.193	0.877	0.959	0.981
PackNet-SfM [34]	M	0.111	0.785	4.601	0.189	0.878	0.960	0.982
HR-Depth [50]	M	0.109	0.792	4.632	0.185	0.884	0.962	**0.983**
Johnston [39]	M	**0.106**	0.861	4.699	0.185	0.889	0.962	0.982
Ours	M	0.107	**0.765**	**4.532**	**0.184**	**0.893**	**0.963**	**0.983**

4.3. Make 3D Results

Make3D [51] consists of RGB monocular images and their depth maps, but stereo images are not available. Therefore, it is not possible to train an unsupervised depth estimation model on this dataset. Our model is equally valid when tested on other datasets. We tested Make3D using the model trained at KITTI 2015, employing the same camera parameters as those provided by the KITTI dataset. We cropped the input images according

to the aspect ratio requirements of the images in the model. As shown in Table 4, our method outperforms other self-supervised methods.

Table 4. The error measure results on the Make3D dataset. In the table, M: Self-supervised mono supervision, S: Self-supervised stereo supervision.

Method	Train	Abs Rel	Sq Rel	RMSE	RMSE Log
Monodepth [11]	S	0.544	10.94	11.760	0.193
Zhou [10]	M	0.383	5.321	10.470	0.478
DDVO [36]	M	0.387	4.720	8.090	0.204
Monodepth2 [35]	M	0.322	3.589	7.417	0.163
Ours	M	**0.314**	**3.112**	**7.048**	**0.159**

4.4. Ablation Study

Tables 5 and 6 show the ablation study based on Monodepth2, and the backbones are ResNet 18 and ResNet 50, respectively. It shows that all of our contributions achieve a steady improvement in almost all evaluation metrics and obtain a consistent performance gain on different backbones. A visualization of the improvement of different components is shown in Figure 6.

Table 5. Comparison of prediction accuracy before and after adding the structure perception module and the detail emphasis module, respectively, on the KITTI dataset using ResNet 18 (R18).

Method	Backbone	Abs Rel	Sq Rel	RMSE	RMSE Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Baseline (monodepth2)	R18	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Baseline + spm	R18	0.111	0.833	4.768	0.191	0.881	0.961	0.982
Baseline + dem	R18	**0.110**	0.812	4.733	**0.190**	**0.882**	0.961	0.982
Ours	R18	**0.110**	**0.810**	**4.678**	**0.190**	**0.882**	**0.962**	**0.983**

Table 6. Comparison of prediction accuracy before and after adding the structure perception module and the detail emphasis module, respectively, to the KITTI dataset using ResNet 50 (R50).

Method	Backbone	Abs Rel	Sq Rel	RMSE	RMSE Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Baseline (monodepth2)	R50	0.110	0.831	4.642	0.187	0.883	0.962	0.982
Baseline + spm	R50	0.109	0.768	4.554	**0.183**	0.885	**0.963**	**0.983**
Baseline + dem	R50	0.109	0.772	4.593	0.185	0.886	0.962	0.982
Ours	R50	**0.107**	**0.765**	**4.532**	0.184	**0.893**	**0.963**	**0.983**

4.5. Discussion

Based on monodepth2, our approach adds spm and dem. The experimental results show that incorporating the attention mechanism is effective. We use a pre-trained residual network as the backbone to extract semantic features, which are then fed into spm and generate new features to explicitly enhance the perception of the scene structure. In the decoding phase, we gradually recover the spatial resolution, use skip connections to facilitate the gradient and information flow throughout the model and use dem to produce fine details. While our addition of the attention mechanism improves the performance of the monocular depth estimation network, there are some shortcomings: the network becomes more complex and takes longer of a cycle.

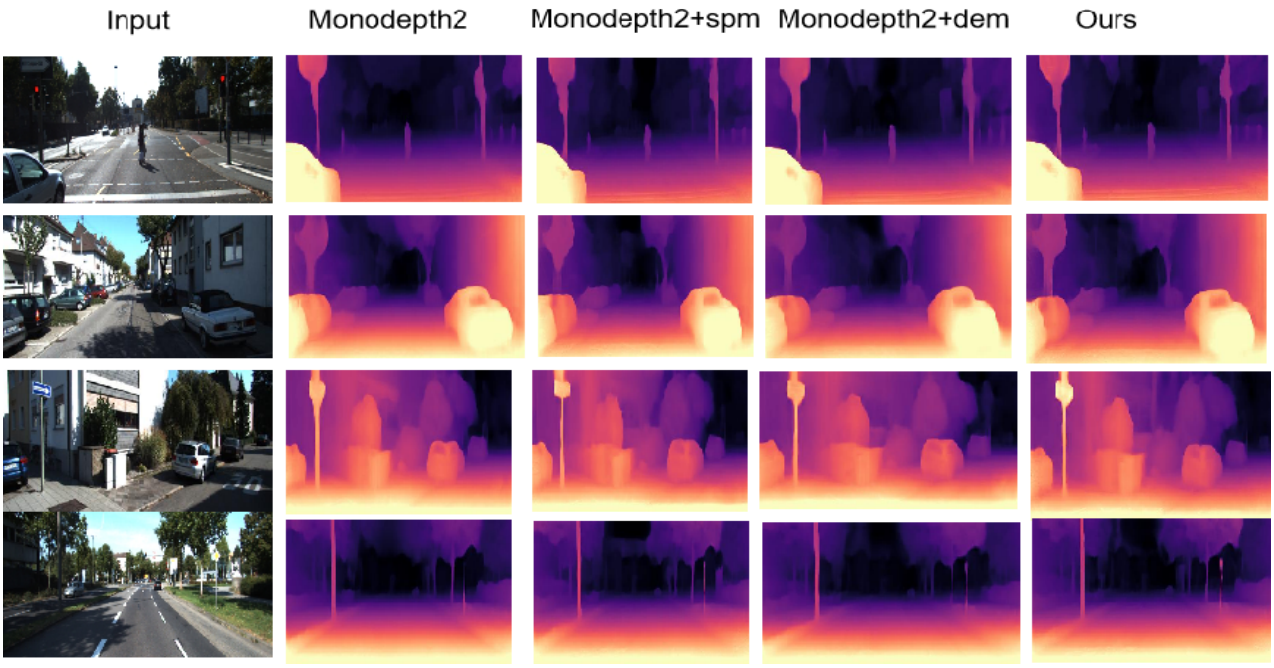

Figure 6. Visualization of the ablation study. From left to right: input image, result of Monodepth2, result of Monodepth2 with the spm, result of Monodepth2 with the dem and result of combining the two.

5. Conclusions

In this paper, a new network structure is proposed based on a self-supervised monocular depth estimation model. Two different channel attention modules are added. The spm uses a frequency channel attention network to enhance the perception of the scene structure and obtain more feature information. The dem employs a channel attention mechanism. It emphasizes some important details and can effectively fuse features at different scales to achieve more accurate and clearer depth prediction. Our network has achieved advanced results on the KITTI dataset.

Author Contributions: Conceptualization, B.T.; methodology, X.C.; software, X.C.; validation, B.T., X.C. and X.T.; formal analysis, X.C.; investigation, X.C.; resources, D.J.; data curation, B.T.; writing—original draft preparation, X.C.; writing—review and editing, B.T.; visualization, B.C.; supervision, B.T.; project administration, B.T.; funding acquisition, B.T. 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 number (51505349, 51575407); the Hubei Provincial Department of Education, grant number D20201106; and the Open Fund of the Hubei Key Laboratory of Hydroelectric Machinery Design & Maintenance in China, Three Gorges University (2021KJX13).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: http://make3d.cs.cornell.edu/index.html (Make3D dataset). http://www.cvlibs.net/datasets/kitti/eval_object.php?obj_benchmark (KITTI dataset) accessed on 1 May 2022.

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

References

1. DeSouza, G.N.; Kak, A.C. Vision for mobile robot navigation: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 237–267. [CrossRef]
2. Menze, M.; Geiger, A. Object scene flow for autonomous vehicles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3061–3070.

3. Newcombe, R.A.; Lovegrove, S.J.; Davison, A.J. Dtam: Dense tracking and mapping in real-time. In Proceedings of the 2011 International Conference on Computer Vision, Washington, DC, USA, 6–13 November 2011; pp. 2320–2327.
4. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *arXiv* **2014**, arXiv:1406.2283.
5. Eigen, D.; Fergus, R. Predicting depth, surface normal and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2650–2658.
6. 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.
7. Cao, Y.; Wu, Z.; Shen, C. Estimating depth from monocular images as classifification using deep fully convolutional residual networks. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 3174–3182. [CrossRef]
8. Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Tao, D. Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2002–2011.
9. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
10. Zhou, T.; Brown, M.; Snavely, N.; Lowe, D.G. Unsupervised learning of depth and ego-motion from video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1851–1858.
11. Godard, C.; Mac Aodha, O.; Brostow, G.J. Unsupervised monocular depth estimation with left-right consistency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 270–2792.
12. Tao, B.; Liu, Y.; Huang, L.; Chen, G.; Chen, B. 3D reconstruction based on photoelastic fringes. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e6481. [CrossRef]
13. Tao, B.; Wang, Y.; Qian, X.; Tong, X.; He, F.; Yao, W.; Chen, B.; Chen, B. Photoelastic Stress Field Recovery Using Deep Convolutional Neural Network. *Front. Bioeng. Biotechnol.* **2022**, *10*, 818112. [CrossRef] [PubMed]
14. Hao, Z.; Wang, Z.; Bai, D.; Tao, B.; Tong, X.; Chen, B. Intelligent detection of steel defects based on improved split attention networks. *Front. Bioeng. Biotechnol.* **2022**, *9*, 810876. [CrossRef]
15. Jiang, D.; Li, G.; Sun, Y.; Hu, J.; Yun, J.; Liu, Y. Manipulator grabbing position detection with information fusion of color image and depth image using deep learning. *J. Ambient Intell. Humaniz. Comput.* **2021**, *12*, 10809–10822. [CrossRef]
16. Tao, B.; Huang, L.; Zhao, H.; Li, G.; Tong, X. A time sequence images matching method based on the siamese network. *Sensors* **2021**, *21*, 5900. [CrossRef]
17. Jiang, D.; Li, G.; Tan, C.; Huang, L.; Sun, Y.; Kong, J. Semantic segmentation for multiscale target based on object recognition using the improved Faster-RCNN model. *Future Gener. Comput. Syst.* **2021**, *123*, 94–104. [CrossRef]
18. Wang, H.M.; Lin, H.Y.; Chang, C.C. Object Detection and Depth Estimation Approach Based on Deep Convolutional Neural Networks. *Sensors* **2021**, *21*, 4755. [CrossRef]
19. Ming, Y.; Meng, X.; Fan, C.; Yu, H. Deep learning for monocular depth estimation: A review. *Neurocomputing* **2021**, *438*, 14–33. [CrossRef]
20. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
21. Zhang, F.; Zhu, X.; Ye, M. Fast Human Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
22. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards Balanced Learning for Object Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
23. Lyu, H.; Fu, H.; Hu, X.; Liu, L. Esnet: Edge-based segmentation network for real-time semantic segmentation in traffic scenes. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1855–1859.
24. Sun, Y.; Huang, P.; Cao, Y.; Jiang, G.; Yuan, Z.; Dongxu, B.; Liu, X. Multi-objective optimization design of ladle refractory lining based on genetic algorithm. *Front. Bioeng. Biotechnol.* **2022**, *10*, 900655. [CrossRef]
25. Liu, Y.; Jiang, D.; Tao, B.; Qi, J.; Jiang, G.; Yun, J.; Huang, L.; Tong, X.; Chen, B.; Li, G. Grasping Posture of Humanoid Manipulator Based on Target Shape Analysis and Force Closure. *Alex. Eng. J.* **2022**, *61*, 3959–3969. [CrossRef]
26. Bai, D.; Sun, Y.; Tao, B.; Tong, X.; Xu, M.; Jiang, G.; Chen, B.; Cao, Y.; Sun, N.; Li, Z. Improved single shot multibox detector target detection method based on deep feature fusion. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e6614. [CrossRef]
27. Liu, Y.; Xu, M.; Jiang, G.; Tong, X.; Yun, J.; Liu, Y.; Chen, B.; Cao, Y.; Sun, N.; Li, Z. Target localization in local dense mapping using RGBD SLAM and object detection. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e6655. [CrossRef]
28. Liu, Y.; Li, C.; Jiang, D.; Chen, B.; Sun, N.; Cao, Y.; Tao, B.; Li, G. Wrist angle prediction under different loads based on GAELM neural network and sEMG. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e6574. [CrossRef]
29. Yang, Z.; Jiang, D.; Sun, Y.; Tao, B.; Tong, X.; Jiang, G.; Xu, M.; Yun, J.; Liu, Y.; Chen, B.; et al. Dynamic Gesture recognition using surface EMG signals based on multi-stream residual network. *Front. Bioeng. Biotechnol.* **2021**, *9*, 779353. [CrossRef]

30. Tosi, F.; Aleotti, F.; Poggi, M.; Mattoccia, S. Learning monocular depth estimation infusing traditional stereo knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9799–9809.
31. Wong, A.; Soatto, S. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5644–5653.
32. Mancini, M.; Costante, G.; Valigi, P.; Ciarfuglia, T.A. Fast robust monocular depth estimation for obstacle detection with fully convolutional networks. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, 9–14 October 2016; pp. 4296–4303.
33. Mahjourian, R.; Wicke, M.; Angelova, A. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5667–5675.
34. Guizilini, V.; Ambrus, R.; Pillai, S.; Raventos, A.; Gaidon, A. 3d packing for self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2485–2494.
35. Godard, C.; Mac Aodha, O.; Firman, M.; Brostow, G.J. Digging into self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 3828–3838.
36. Wang, C.; Buenaposada, J.M.; Zhu, R.; Lucey, S. Learning depth from monocular videos using direct methods. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2022–2030.
37. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 7354–7363.
38. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 10–15 June 2019; pp. 3146–3154.
39. Johnston, A.; Carneiro, G. Self-supervised monocular trained depthestimation using self-attention and discrete disparity volume. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4756–4765.
40. 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.
41. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 783–792.
42. Liu, Y.; Shao, Z.; Hoffmann, N. Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions. *arXiv* **2021**, arXiv:2112.05561.
43. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
44. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
45. Zou, Y.; Luo, Z.; Huang, J.B. DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency. In Proceedings of the European Conference on Computer Vision (ECCV) 2018, Munich, Germany, 8–14 September 2018.
46. Ranjan, A.; Jampani, V.; Balles, L.; Kim, K.; Sun, D.; Wulff, J.; Black, M.J. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, Long Beach, CA, USA, 16–17 June 2019.
47. Luo, C.; Yang, Z.; Wang, P.; Wang, Y.; Xu, W.; Nevatia, R.; Yuille, A. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2624–2641. [CrossRef]
48. Casser, V.; Pirk, S.; Mahjourian, R.; Angelova, A. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8001–8008.
49. Klingner, M.; Termöhlen, J.A.; Mikolajczyk, J.; Fingscheidt, T. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 582–600.
50. Lyu, X.; Liu, L.; Wang, M.; Kong, X.; Liu, L.; Liu, Y.; Chen, X.; Yuan, Y. Hr-depth: High resolution self-supervised monocular depth estimation. *arXiv* **2020**, arXiv:2012.07356.
51. Saxena, A.; Sun, M.; Ng, A.Y. Make3d:Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 824–840. [CrossRef] [PubMed]