electronics

Article

URNet: An UNet-Based Model with Residual Mechanism for Monocular Depth Estimation

Hoang-Thanh Duong, Hsi-Min Chen and Che-Cheng Chang *

Department of Information Engineering and Computer Science, Feng Chia University, Taichung 40724, Taiwan; thanh.duongrsm@gmail.com (H.-T.D.); hmchen@mail.fcu.edu.tw (H.-M.C.)
* Correspondence: checchang@fcu.edu.tw; Tel.: +886-4-24517250 (ext. 3764)

Abstract: Autonomous vehicle systems rely heavily upon depth estimation, which facilitates the improvement of precision and stability in automated decision-making systems. Noteworthily, the technique of monocular depth estimation is critical for one of these feasible implementations. In the area of segmentation of medical images, UNet is a well-known encoder–decoder structure. Moreover, several studies have proven its further potential for monocular depth estimation. Similarly, based on UNet, we aim to propose a novel model of monocular depth estimation, which is constructed from the benefits of classical UNet and residual learning mechanisms and named URNet. Particularly, we employ the KITTI dataset in conjunction with the Eigen split strategy to determine the efficacy of our model. Compared with other studies, our URNet is significantly better, on the basis of higher the precision and lower error rate. Hence, it can deal properly with the depth estimation issue for autonomous driving systems.

Keywords: autonomous vehicle systems; monocular depth estimation; residual mechanism; UNet

check for updates

Citation: Duong, H.-T.; Chen, H.-M.; Chang, C.-C. URNet: An UNet-Based Model with Residual Mechanism for Monocular Depth Estimation. *Electronics* **2023**, *12*, 1450. https://doi.org/10.3390/electronics12061450

Academic Editors: Shiho Kim and Sergio Busquets-Monge

Received: 9 December 2022
Revised: 6 February 2023
Accepted: 15 March 2023
Published: 19 March 2023

1. Introduction

Autonomous vehicle systems rely heavily upon depth estimation, which facilitates the performance of decision-making processes. Currently, active and passive sensing are the two main types of depth estimation techniques. Active sensing, such as LiDAR and infrared sensors, measures distance via the reflection of signals. On the other hand, without emitting signals, passive sensing predicts the depth information via computer vision techniques. For autonomous vehicle systems, the decision-making process needs to detect the type of each object, as well as the relative position between each object and the vehicle to make the optimal decision. Hence, passive sensing is more suitable, since it can realize both at the same time, monocular depth estimation especially. More importantly, monocular depth estimation (MDE) has attracted a large number of researchers, due to its relevance to crucial deep learning topics, including augmented reality (AR) [1], robotics [2], and autonomous vehicle systems [3].

Traditionally, passive depth estimation is reliant on multi-angle geometry, i.e., stereo imaging [4,5]. Nonetheless, its extreme complexity, stringent alignment, and requirements of data input make it tricky for implementation. Over the past decade, convolution neural network (CNN) has undergone significant development. Thus, passive depth estimation has taken a significant turn: monocular depth estimation based on CNN is appealing, due to its utility and ease of implementation. Numerous researchers have proposed various network architectures to generate a depth map from a single image [6–8], where there are two main techniques for implementation, i.e., self-supervised learning and supervised learning based on the utilization of ground truth during the training procedure. Notice that ground truth is the information that is known to be real and provided by measurement. Here are some instances:

- Supervised monocular depth estimation: The ultimate objective of supervised learning is to determine the relationship between the RGB (red–green–blue) image and depth map from the public RGB-D (red–green–blue depth) datasets, such as the KITTI dataset [9] and the NYU V2 dataset [10]. Several researchers have focused on and achieved great success with this technique. Eigen et al. [11] proposed the architectures of coarse-to-fine network and applied the concept of scale-invariant error to measure depth relations to significantly improve depth prediction performance. Fu et al. [12] proposed the spacing-increasing discretization (SID) method to save time and cost. Particualrly, the elimination of the sub-sampling layers and the incorporation of the dilated convolution enable SID to reduce cost and training time and improve accuracy. Teed et al. [13] proposed a DeepV2D model, which estimates monocular degrees on video by combining two classical geometric algorithms in an end-to-end architecture. Hence, DeepV2D needs additional information to produce its depth map.
- Self-supervised monocular depth estimation: Since the depth label is not required for the training procedure, data preparation and reprocessing can be reduced or eliminated. Thus it can save considerable time and energy in the training procedure. In the field of monocular depth estimation, self-supervised learning has gained popularity and made significant advancements. Godard et al. [14] proposed a flexible architecture, which can address the issue of dynamic object of self-supervised learning. They conceptualized, crafted, and enhanced the automatic mask loss and reprojection loss. Next, to improve the accuracy of depth prediction at the boundaries, Wong et al. [15] implemented the concept of residual-based adaptive weight and bilateral cyclic consistency constraint. Ling et al. [16] developed a model of deep learning for completion of monocular depth, which is incorporated with the attention block. To enhance the capabilities of the network, they also designed the notion of multi-warp loss for monocular depth estimation.

UNet is a convolution neural network that was developed and widely applied for the segmentation of medical images. Many researchers have developed their UNet-based models to enhance the performance in medical image segmentation [17–21]. They use a variety of different strategies to alter the conventional model of UNet. As a result, numerous UNet-based networks with improved performance have emerged. Their studies improve the accuracy of automated medical systems. In the depth estimation field, several studies have yielded fruitful results on optimizing the UNet architecture. The authors in [22] came up with the idea for a mixed-scale UNet network called MAPUnet. It had a dense atrous pyramid and was based on the extensively used encoder–decoder structure. The results of a comparison among the proposed network and state-of-the-art methods indicate that the proposed network has superior performance, in terms of error rate and precision. The authors in [23] proposed a new algorithm employing a simple two-tower convolutional neural network, which was designated 2T-UNet. The research eliminated the need for a stereoscopic matching step. Additionally, 2T-UNet not only achieves a high level of accuracy, but is also compatible with mobile devices. Noteworthily, the concepts of encoder, decoder, and skip connection are the primary characteristics of UNet. The encoder is responsible for extracting image features, while the decoder restores the image to its original size and outputs the final result based on the extracted features. The skip connection connects the encoder and decoder, and it is responsible for combining low-level features in the encoder and high-level features in the decoder. Recently, due to the potency of UNet for monocular depth estimation, several researchers have focused on altering and enhancing the internal architecture of encoder [24], decoder [23], and skip connection [22] to improve the efficiency of the depth estimation issue.

On the other hand, the residual mechanism was widely adopted for various CNN researches because it overcame the problem of network degradation in the deep learning domain. Consequently, the efficiency of the feature extraction in deeper layers of deep learning networks has been improved via the using of the residual mechanism. He et al. [25] presented a residual learning framework, which has improved precision and convergence

speed of deep learning in deeper networks. Additionally, in comparison to the state of the art, the network has a higher degree of precision. Furthermore, based on [25], Laina et al. [26] proposed a new model with a completely convolutional architecture. The model simultaneously permitted more in-depth training and a reduction in the number of parameters. The results of the comparison indicate that the proposed model has superior performance, in terms of both the increased precision and lower execution time.

In this paper, we introduced URNet, a new network architecture based on the traditional UNet model and residual connections to solve the depth estimation problem. Specifically, we modify the UNet to deal properly with the issue of monocular depth estimation for autonomous vehicles. First, each node from the original UNet is altered based on the residual blocks. Next, most UNet-based models ignore the output feature, where they only concentrate on the last output of a node. However, in our model, additional connections are added between the encoder and decoder, in order to provide a stronger feature map for the decoder. Note that the encoder/decoder is composed of at least one node. The contributions of the paper are summarized as follows:

- We propose URNet, a novel model for monocular depth estimation based on the traditional UNet model and residual blocks. More specifically, we use the attention and ASPP (atrous spatial pyramid pooling) blocks to improve the prediction performance.
- We evaluate the performance of our URNet via the KITTI dataset and compare it with several existing researches.

The remainder of this work is structured as follows: In Section 2, we provide an overview of the topic of monocular depth estimation. In Section 3, we detailedly describe our model, URNet, and the results of the experiments are described in Section 4. Finally, a brief conclusion is presented at the end of this study.

2. Related Work

Early researches regarding monocular depth estimation was realized based on the hand-crafted features. As the first, Torralba and Oliva [27] proposed a source of information for absolute depth estimation that did not rely on specific objects and was derived from the entire scene structure. The authors in [28] estimated the depth information by utilizing the planar layout, which included 3D position and orientation. These were estimated using the Markov random field (MRF), which contains various types of feature values, e.g., edge orientations, chromatic values, and so on. Next, Karsch et al. [29] presented a method that can automatically generate plausible depth maps from videos by making use of non-parametric depth sampling.

Next, with the significant development of convolution neural networks over the past decade, the estimation methods have shifted to the deep learning area. In the literature, several researchers have devoted a great deal of effort to improve the methods for achieving better outcomes, where the KITTI dataset [9], and the NYU V2 dataset [10] are frequently adopted in the experiments. These datasets provided exceptionally precise depth information. The depth information acquired by 3D sensing equipment (such as LiDAR, Kinect, and other similar sensors) is generally employed as the ground truth in supervised learning systems. On the other hand, many researchers use stereo inputs for the monocular depth estimation in unsupervised learning. This research was performed with variety methods and strategies.

The approaches of depth estimation in the field of deep learning are classified according to the type of input images. Numerous studies use unlabeled images as the input data in the training procedure [14–16,30–33]. Wong et al. [15] proposed an adaptive weight design that allows for the regulation of the bilateral cyclic relationship between the left and right disparities. Their model is with high performance on the KITTI dataset. On the other hand, the attention mechanism is extremely well-known for natural language processing (NLP). Ling et al. [16] proposed a new unsupervised architecture for monocular depth estimation by incorporating the concept of attention block. On the KITTI dataset, the architecture demonstrated its superior performance. Noteworthily, the authors confirmed

that the location of attention block is crucial while applying the concept of attention block to a deep learning model. Moreover, utilizing additional features of stereo images is a difficult and crucial task, since the number of features directly influences the performance of a deep learning model. Thus, Ye et al. [34] proposed a novel unsupervised architecture for monocular depth estimation to solve the aforementioned issue. Generally, a network with better depth estimation results is more complex, with a large number of parameters, where both hinder the system performance during the implementation. To address this issue, Liu et al. [35] proposed a real-time and lightweight architecture for unsupervised depth estimation. Due to its lightweight design, the proposed network can operate in real time and small devices, allowing it to be utilized in embedded systems.

Alternatively, numerous researchers employed labeled images as the input data in the training procedure [30,36–40], where the models can achieve superior performance, depending on the quality of input data. More specifically, several researchers proposed their new model for monocular depth estimation using multiple-scale features [30,41,42], and the performance of these models is exceptional. Next, Xu et al. in [40] proposed a pyramid network with the concept of an adaptive fusion block to improve the exploiting capability of the input depth map and the depth estimation performance. DPNet [37] was a comprehensive solution to the problems of inaccurate depth inference and the loss of spatial information, where both the contextual branch and spatial branch were described in detail. In conclusion, they suggested using a refinement module to combine the disparate features obtained from the two different branches, in order to generate a depth map of superior quality. Next, Fu et al. [12] presented the idea of a deep ordinal regression network (DORN) to use the dilated convolution to obtain the high-resolution depth map. Furthermore, Alhashim et al. in [38], proposed a new design for monocular depth recovery by employing transfer learning. This technique acquires the boundaries of objects more accurately than previous researches.

On the other hand, due to the limitation of labeled data, some researchers joined the areas of segmentation and monocular depth estimation. Here, the segmentation results facilitate a more thorough comprehension of the image scenes. Chen et al. [43] proposed SceneNet, which improves location depth estimation by enforcing semantic consistency between stereo pairs. Next, Zhu et al. [44] proposed a monocular self-supervised depth estimation architecture, where the information of segmentation and depth edges was utilized to estimate the boundaries of objects for the purpose of ensuring consistency, as well as optimizing the depth estimation model.

3. Our URNet

In this section, we start to describe our method detailedly, which is named URNet. More specifically, it is based on the concept of UNet and further with the following modifications:

- Modifying the UNet-based nodes with the residual blocks.
- Adding connections between the encoder and decoder.
- Adding the attention blocks to the decoder.
- Utilizing the ASPP block to replace transfer blocks.

The network architecture is shown in Figure 1, and the detailed settings of the network architecture are presented in Table 1. In Figure 1, the encoder and decoder are the two primary building constituents of the proposed scheme, and each consists of four nodes that are inspired by the residual mechanism of ResNet architecture [25]. In addition, the ASPP block plays a critical role as the unique transfer node of the model, which is first placed at the base of the network. After the decoding procedure is complete, the ASPP block is used again to improve the quality of the output. Notably, our method connects the encoder and decoder blocks via two connections, as opposed to the single connection of conventional UNet. The relationship of one pair of nodes between the encoder and decoder of our method is shown in Figure 2.

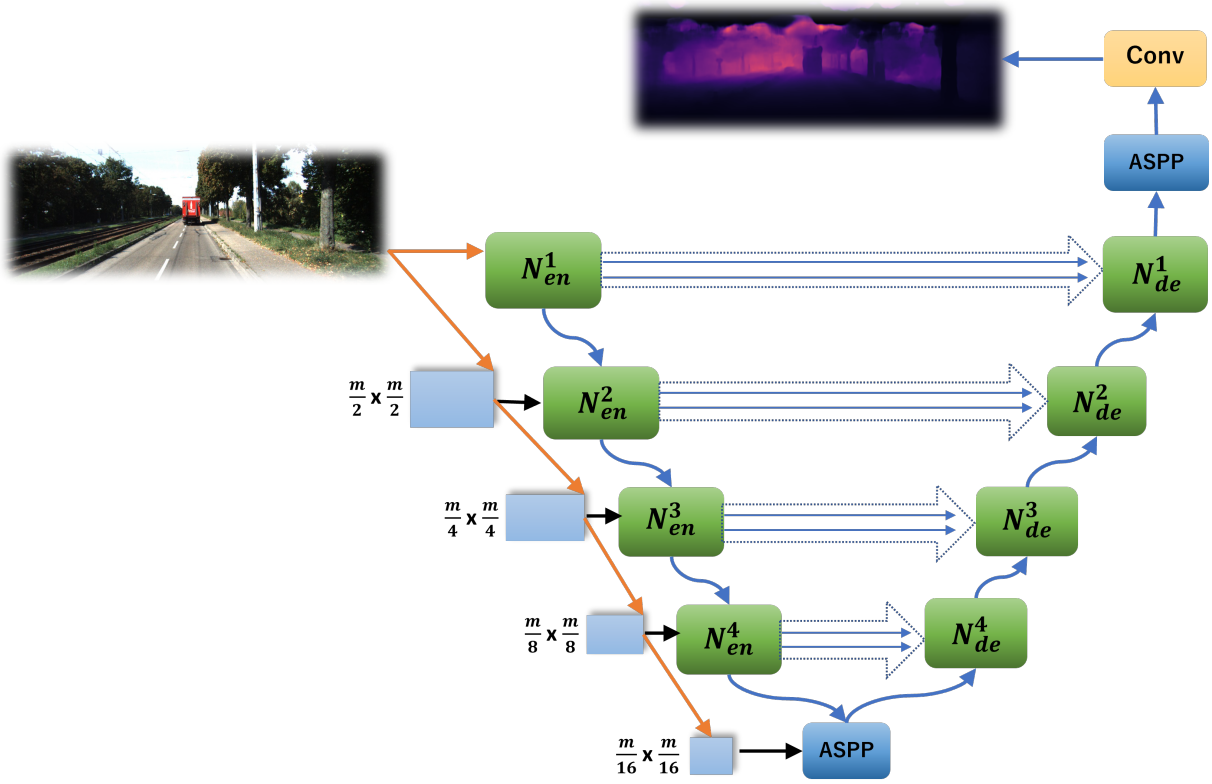

Figure 1. The overall design of our network architecture.

Table 1 provides the outline of the structure of our URNet model. The model makes use of 3 × 3 convolutions. Moreover, the dropout rate is set to 0.3 of ASPP blocks, and the attention blocks are incorporated into four nodes of the decoder to improve the efficiency of feature extraction. Finally, the additional connections are added between the encoder and decoder to gain more features than the conventional UNet.

Table 1. The architecture of URNet.

Node	Encoder	Decoder
1	$[64 \times (BN + ReLU + Conv)^2] + Max\ pooling + AD + SE$	$AT + Up\text{-}sample + BN + ReLU + Conv + AD$
2	$[128 \times (BN + ReLU + Conv)^2] + Max\ pooling + AD + SE$	$AT + Up\text{-}sample + BN + ReLU + Conv + AD$
3	$[256 \times (BN + ReLU + Conv)^2] + Max\ pooling + AD + SE$	$AT + Up\text{-}sample + BN + ReLU + Conv + AD$
4	$[512 \times (BN + ReLU + Conv)^2] + Max\ pooling + AD + SE$	$AT + Up\text{-}sample + BN + ReLU + Conv + AD$
Tr	$ASPP$	

Note that the transition, squeeze excite, addition, and attention block are denoted by the abbreviations Tr, SE, AD, and AT, respectively.

Next, two important components, attention and ASPP blocks, are introduced, and the detailed network architectures are shown in Figure 3.

- Attention blocks: Natural language processing is the main area where the attention mechanism has been utilized [45–47]. Recently, it has also been used in depth estimation tasks, such as pixel-wise prediction [48–50]. The attention blocks are responsible for focusing and extracting additional features. Therefore, the implementation of attention blocks facilitates the enhancement of performance of our model. Because of the effectiveness of the attention blocks, in both the topics of natural language processing and computer vision, we decided to include the design of attention blocks in the decoder part of our architecture. This helps us to zero in on the most crucial aspects of the feature maps. More specifically, our attention blocks are composed of three smaller blocks, shown in Figure 3a. The first block receives the input from

the encoder, and the second block receives the input from the decoder. Then, the outputs of the above two blocks will be concatenated as the input for the third block. Finally, the result will be the multiplication of the outputs of the third block and the second block.

- Atrous spatial pyramid pooling (ASPP): Chen et al. [51] came up with the idea of ASPP. Because of its benefits, ASPP has been used more in the topics of segmentation and depth prediction in the past few years [33,42,52]. The authors in [33] suggested a novel algorithm, called DenseASPP, for densely connected atrous spatial pyramid pooling. The blocks accept scales of various sizes, including larger and smaller sizes. In our model, the ASPP is used to extract multiple rates of network endpoint and transfer node characteristics. More specifically, our ASPP block is composed of three smaller blocks. Each block is comprised of the 3×3 convolution, ReLU activation, and Batch normalization (Figure 3b), where the dilation parameters in the convolution layers of three blocks are set to 1, 2, and 3, respectively. Finally, the outputs of three blocks are concatenated and then subjected to the 3×3 convolution.

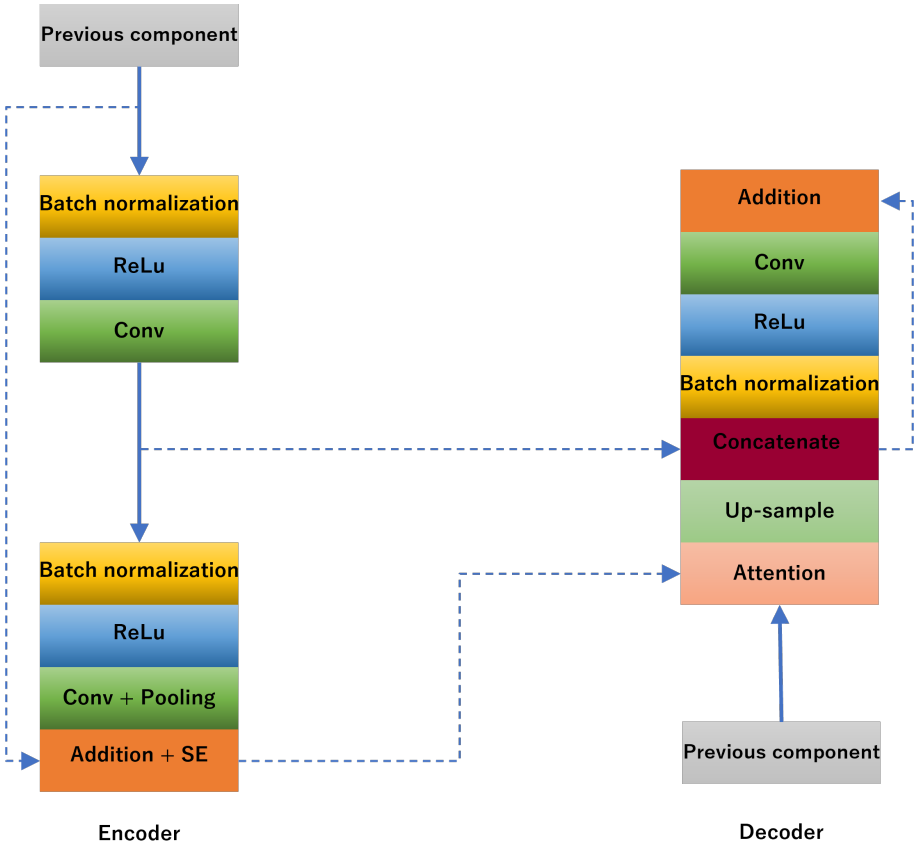

Figure 2. The relationship between the encoder and decoder.

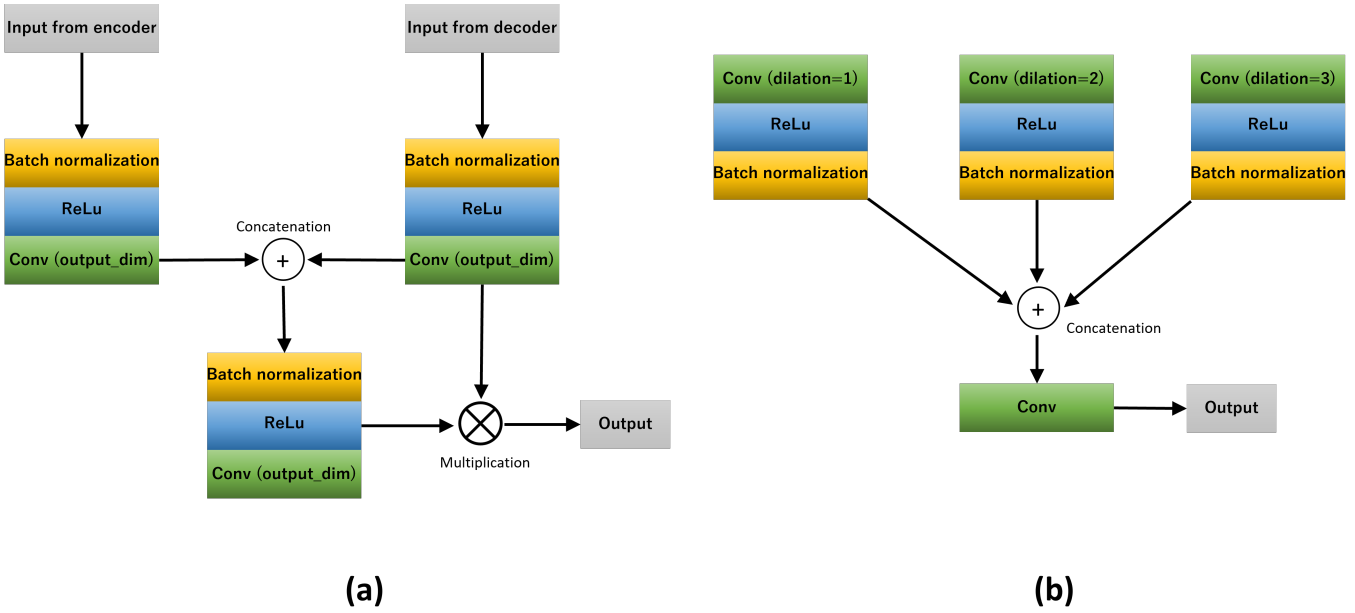

Figure 3. (**a**) the structure of attention mechanism and (**b**) the structure of ASPP block.

4. Experiments and Results

4.1. Dataset

The KITTI and NYU datasets have distinct depth collections, since KITTI employs Lidar, while NYU uses Kinect to sense environments. The KITTI data collection is renowned for its coverage of outdoor environments. Hence, we chose the KITTI dataset [9] for model training and evaluation, due to the fact that URNet was designed specifically for autonomous driving systems. The dataset consists of numerous road configurations from different driving situations, where the resolution of an image is 1242 $\times$ 375 pixels. Noteworthily, the Eigen split strategy [30] for performance comparison among different models is adopted in this paper. Namely, the testing set contains 697 images selected from 29 scenarios, and the training set includes 23,488 images chosen from 32 scenes.

4.2. Detailed Settings

The PyTorch framework [53] is utilized to construct our design, and the detailed settings are listed below:

- Epochs: 50.
- Batch size: 4.
- Optimizer: AdamW [54].
- Values of power and momentum: 0.90 and 0.999.
- Values of weight decay: 0.0005 (encode) and 0 (decode).
- Values of learning rate: 10^{-4} (initially) and 10^{-5} (finally).

Moreover, the training images from the KITTI dataset are randomly cropped to 704 $\times$ 352 pixels. Then, the probability of these images of being horizontally flipped is set to 50%, and the values of brightness, color, and gamma of these images are also modified randomly during the interval $[0.9, 1.1]$. More importantly, the two extra settings of the training images will increase the level of confidence, while realizing our design in real-world environment, since we also consider the tolerance of computer vision.

On the other hand, the loss function is one of the important aspects during the training procedure. In order to train URNet, we implemented the loss function that Eigen et al. [11] suggested, based on the basis of a scale-invariant error. This allowed us to train the network

more effectively. The loss function was utilized for the training process in a great number of UNet-based models [22,42]. The specification of the loss function is as follows:

$$D_{(g)} = \frac{1}{T}\sum g_i^2 - (\frac{1}{T}\sum g_i)^2 + (1-\lambda)(\frac{1}{T}\sum g_i)^2, \quad (1)$$

and it can be simplified as:

$$D_{(g)} = \frac{1}{T}\sum g_i^2 - \frac{\lambda}{T^2}(\sum g_i)^2, \quad (2)$$

where $g_i = log(\tilde{\gamma}_i) - log(\gamma_i)$, γ_i is the ground truth depth value of pixel i, $\tilde{\gamma}_i$ the predicted depth value of pixel i, T representing the number of pixels with valid ground truth values, and $\lambda = 0.85$. Next, since scaling the range of the loss function properly will improve convergence and the final result, the final version is set to:

$$L = \delta\sqrt{D_{(g)}}, \quad (3)$$

where δ is s a constant and set to 10.

4.3. Evaluation Metrics

Several metrics are often used to evaluate the efficiency and robustness of the network. We refer to part of these measure methods introduced in [30] for the evaluation in this paper, i.e., root mean squared error (RMSE) (Equation (3)), Root Mean Squared Logarithmic Error (RMLSE) (Equation (4)), Mean Relative Error (MRE) (Equation (5)), and accuracy (Equation (6)).

$$\text{RMSE} = \sqrt{\frac{1}{|T|}\sum_{\gamma\in T} \| \gamma - \gamma^* \|^2}, \quad (4)$$

$$\text{RMLSE} = \sqrt{\frac{1}{|T|}\sum_{\gamma\in T} \| \log(\gamma) - \log(\gamma^*) \|^2}, \quad (5)$$

$$\text{MRE} = \frac{1}{|T|}\sum_{\gamma\in T} \frac{\| \gamma - \gamma^* \|}{\gamma^*}; \quad (6)$$

$$Accuracy = \%\, of\ \gamma_i\ s.t.\ max(\frac{\gamma}{\gamma^*}, \frac{\gamma^*}{\gamma}) = \sigma < th, th \in (1.25, 1.25^2, 1.25^3), \quad (7)$$

where γ represents the predicted depth value of pixel i, γ^* the ground truth depth value of pixel i, and T is the number of valid ground truth pixels.

4.4. Depth Estimation Performance

First, we use the preliminary comparisons between the image results of Alhashim's method [38] and ours to show some interesting viewpoints. Particularly, in Figure 4, the original RGB images (column (a)), the results of our URNet (column (b)), the results of Alhashim's method (column (c)), and ground-truth from LiDAR (column (d)) are provided. The images indicated that our method produced more accurate estimates of the boundaries of distant signs, people, and vehicles. To clearly highlight the performance of the proposed method, we have marked those with red boxes. However, the test images contain gaps in either the sky or the top sections. We believe that this is due to the extremely sparse ground truth depth data. Due to the lack of proper depth values in certain regions of the images throughout the whole dataset, it is incredibly difficult to train the network appropriately for these regions.

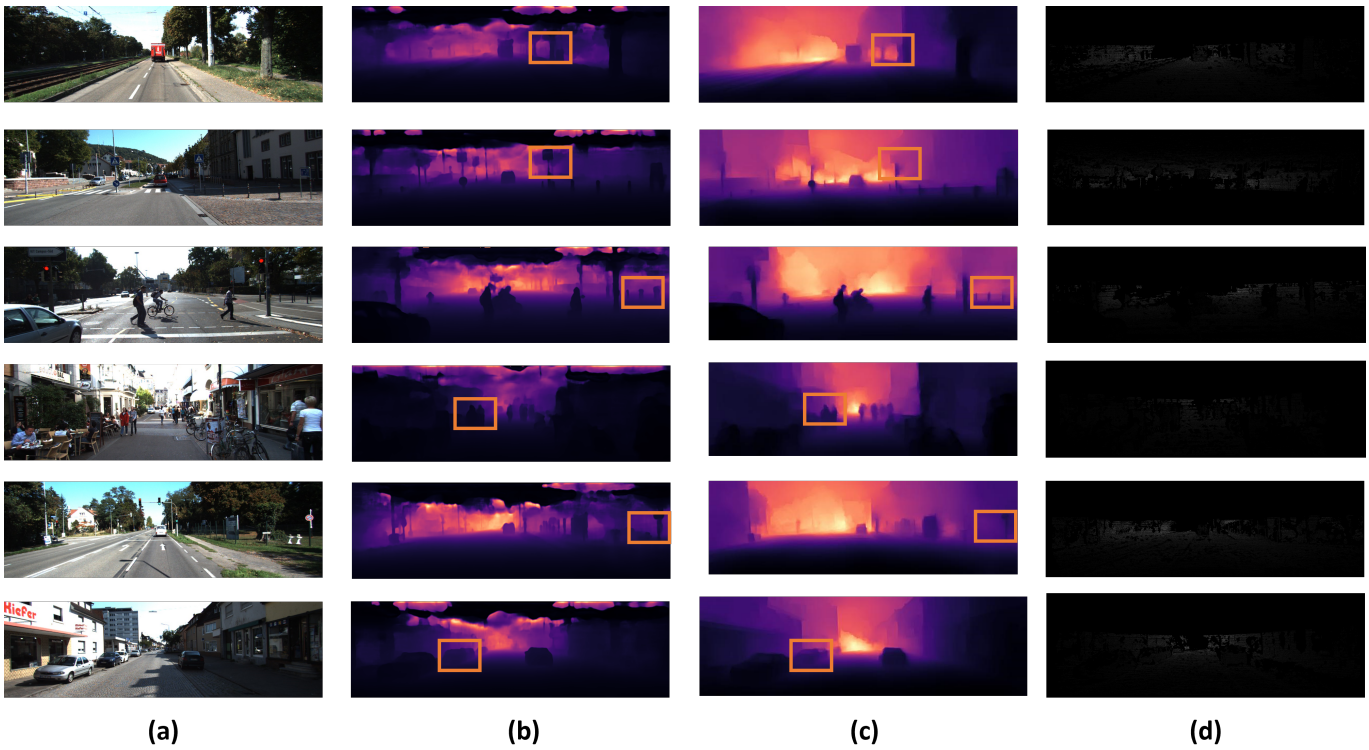

Figure 4. The KITTI qualitative results of different methods for six images (Eigen split): (**a**) RGB images, (**b**) our results, (**c**) results of Alhashim [38], and (**d**) ground-truth from LiDAR.

Next, to demonstrate the effectiveness of URNet in comparison to [38] further, we computed the error of depth estimation of each pixel, and it is the difference between the prediction of algorithm and the ground truth from KITTI dataset. The statistics presented in Figures 5 and 6 have shown some more detailed information. In Figure 5, which includes the mean error of an image of the 652 images in KITTI, we can observe that our algorithm possesses lower mean error of depth estimation and obviously more stabler. In the Figure 6, the entire range of depth estimation is separated into 10 intervals to observe more detailed comparisons among different sensed range. According to the Figure 6, we also possess better performance, especially for the intervals of shorter distance, [0, 9.4) and [9.4, 18.8), that are more important for collision avoidance of smart vehicles. Hence, on the basis of above two statistics, the level of confidence of the improvement of our algorithm can be guaranteed.

On the other hand, Table 2 displays the comprehensive results and comparisons among various depth recovery algorithms and ours, where parts of the materials are from [40]. In this table, the type "Stereo" indicates self-supervised learning by utilizing stereo supervision, and the type "Depth" indicates the supervised learning methods by utilizing depth supervision. Our method is compared to these existing works [14–16,30–32,34–38,40,41]. More specifically, consider the first three comparisons, RMSE, MRE, and RMLSE, where the lower the better. Our method has the best performance, with values of 3.249, 0.088, and 0.131, respectively. Then, in the last comparison, accuracy, where the higher the better. Again, our strategy produced the best results for different thresholds, $\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$, with respective values of 0.912, 0.981, and 0.995. The experimental results indicate that our method is superior to those listed in Table 2. The results also confirmed that applying residual blocks, attention blocks, and ASPP blocks to conventional UNet is effective.

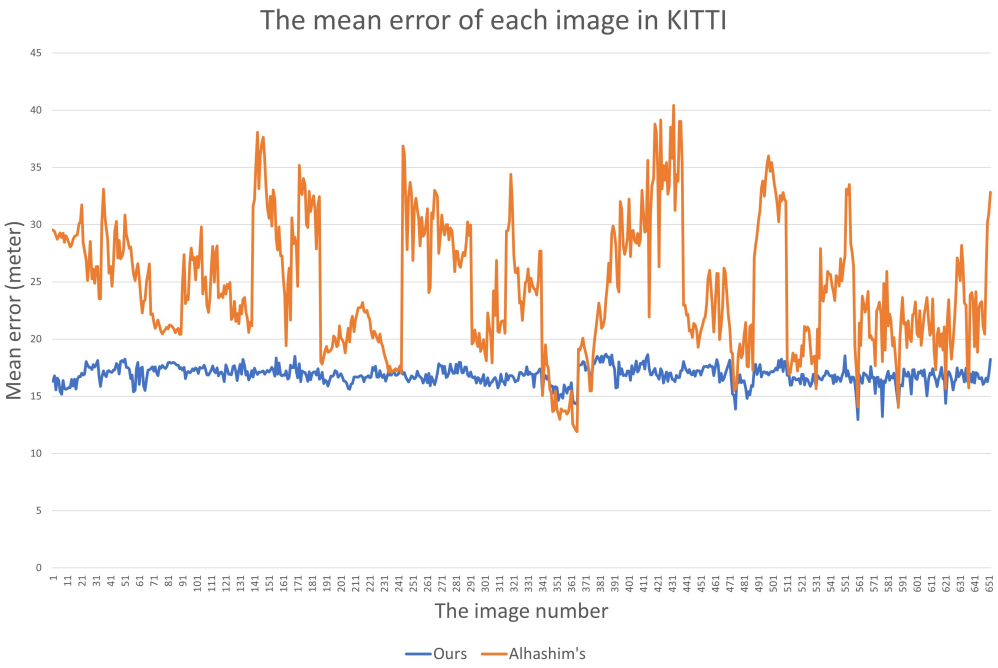

Figure 5. The mean error of each image in the set of test images.

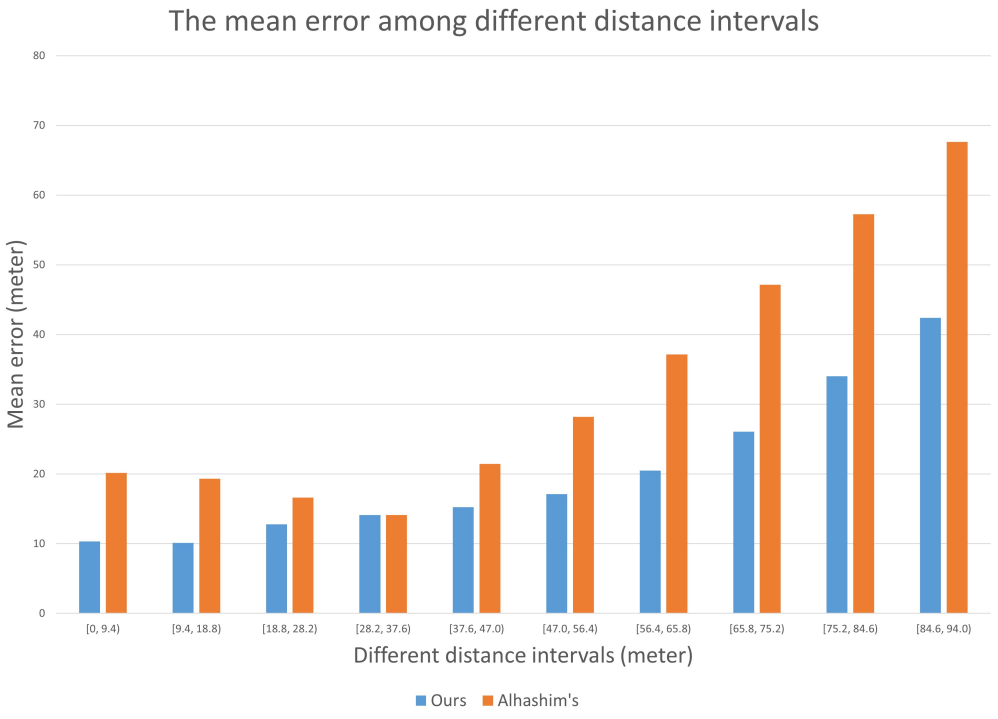

Figure 6. The mean error for various distance intervals.

Table 2. KITTI dataset-based comparisons.

Methods	Type	RMSE	MRE	RMLSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Godard et al. [14]	Stereo	4.863	0.115	0.193	0.877	0.959	0.981
Watson et al. [31]	Stereo	4.695	0.106	0.193	0.875	0.958	0.980
Wong et al. [15]	Stereo	4.172	0.126	0.217	0.840	0.941	0.973
Tosi et al. [32]	Stereo	4.714	0.111	0.199	0.864	0.954	0.979
Ling et al. [16]	Stereo	5.206	0.121	0.214	0.843	0.944	0.975
Ye et al. [37]	Stereo	4.810	0.105	0.196	0.861	0.947	0.978
Eigen et al. [30]	Depth	7.156	0.190	0.246	0.692	0.899	0.967
Liu et al. [35]	Depth	4.977	0.127	NR	0.838	0.948	0.980
Fang et al. [36]	Depth	4.075	0.098	0.174	0.889	0.963	0.985
Ye et al. [34]	Depth	4.978	0.112	0.210	0.842	0.947	0.973
Pei et al. [41]	Depth	4.054	0.098	NR	0.893	0.968	0.987
Alhashim et al. [38]	Depth	4.170	0.093	NR	0.886	0.963	0.986
Chen et al. [49]	Depth	3.597	0.0955	0.159	0.893	0.970	0.989
Gan et al. [39]	Depth	3.933	0.098	0.173	0.890	0.964	0.985
Xu et al. [40]	Depth	3.842	0.092	0.185	0.895	0.974	0.990
Our results (URNet)	Depth	3.249	0.088	0.131	0.912	0.981	0.995

5. Conclusions

In this paper, our URNet is presented, which is a new model for monocular depth estimation based on the enhancement of UNet with the residual learning mechanism, the additional linkages between the encoder and decoder, and the ASPP and attention blocks to improve its performance. Particularly, we use the KITTI dataset to realize our experiments. In the comparisons, four different evaluation metrics with various parameters are adopted. The results show that our system is the best among those algorithms both error rate and precision. Hence, our proposed model can deal properly with depth estimation issue and improve the performance of decision-making processes of autonomous driving systems.

For the future work, we intend to design some novel methods for different data of depth estimation for the purpose of getting better performace. Additionally, the fusion of distinct data is an interesting topic for us. However, if excessive data are used, it is obvious that the model would make a demand for a lot of system memory and may not acquire the better performance. This is the tricky part of trade-off as well. Not all devices are compatible due to the huge number of URNet parameters. The forthcoming edition will have a lightweight architecture as one of its most important characteristics.

Author Contributions: Conceptualization, H.-T.D. and H.-M.C.; methodology, H.-T.D. and C.-C.C.; software, H.-T.D.; validation, H.-T.D., H.-M.C. and C.-C.C.; formal analysis, H.-T.D., H.-M.C. and C.-C.C.; writing–original draft preparation, H.-T.D. and C.-C.C.; writing–review and editing, H.-M.C. and C.-C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by National Science and Technology Council, Taiwan, R.O.C. under grant 109-2221-E-035-055-MY3 and 109-2221-E-035-067-MY3.

Data Availability Statement: Not applicable.

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

References

1. Bertels, M.; Jutzi, B.; Ulrich, M. Automatic Real-Time Pose Estimation of Machinery from Images. *Sensors* **2022**, *22*, 2627. [CrossRef] [PubMed]
2. Avinash, A.; Abdelaal, A.E.; Salcudean, S.E. Evaluation of increasing camera baseline on depth perception in surgical robotics. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 5509–5515.
3. Chuah, W.; Tennakoon, R.; Hoseinnezhad, R.; Bab-Hadiashar, A. Deep learning-based incorporation of planar constraints for robust stereo depth estimation in autonomous vehicle applications. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 6654–6665. [CrossRef]

4. Scharstein, D.; Pal, C. Learning conditional random fields for stereo. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
5. Scharstein, D.; Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* **2002**, *47*, 7–42. [CrossRef]
6. Xu, D.; Wang, W.; Tang, H.; Liu, H.; Sebe, N.; Ricci, E. Structured attention guided convolutional neural fields for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3917–3925.
7. Kuznietsov, Y.; Stuckler, J.; Leibe, B. Semi-supervised deep learning for monocular depth map prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6647–6655.
8. Godard, C.; Mac Aodha, O.; Brostow, G. Unsupervised monocular depth estimation with left-right consistency. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6602–6611.
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. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. In Proceedings of the European Conference on Computer Vision, Berlin, Germany, 7–13 October 2012; pp. 746–760.
11. 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**, *3*, 2366–2374.
12. 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, 18–23 June 2018; pp. 2002–2011.
13. Teed, Z.; Deng, J. Deepv2d: Video to depth with differentiable structure from motion. *arXiv* **2018**, arXiv:1812.04605.
14. Godard, C.; Aodha, O.M.; Firman, M.; Brostow, G. Digging into self-supervised monocular depth estimation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 Octtober–2 November 2019; Volume 2019, pp. 3827–3837. [CrossRef]
15. Wong, A.; Soatto, S. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 5637–5646. [CrossRef]
16. Ling, C.; Zhang, X.; Chen, H. Unsupervised Monocular Depth Estimation Using Attention and Multi-Warp Reconstruction. *IEEE Trans. Multimed.* **2022**, *24*, 2938–2949. [CrossRef]
17. Tran, S.T.; Cheng, C.H.; Nguyen, T.T.; Le, M.H.; Liu, D.G. TMD-Unet: Triple-Unet with multi-scale input features and dense skip connection for medical image segmentation. *Healthcare* **2021**, *9*, 54. [CrossRef]
18. Tran, S.T.; Cheng, C.H.; Liu, D.G. A multiple layer U-Net, U n-Net, for liver and liver tumor segmentation in CT. *IEEE Access* **2020**, *9*, 3752–3764. [CrossRef]
19. Tran, S.T.; Nguyen, T.T.; Le, M.H.; Cheng, C.H.; Liu, D.G. TDC-Unet: Triple Unet with Dilated Convolution for Medical Image Segmentation. *Int. J. Pharma Med. Biol. Sci.* **2022**, *11*, 1–7. [CrossRef]
20. Isensee, F.; Petersen, J.; Klein, A.; Zimmerer, D.; Jaeger, P.F.; Kohl, S.; Wasserthal, J.; Koehler, G.; Norajitra, T.; Wirkert, S.; et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv Preprint* **2018**, arXiv:1809.10486.
21. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv Preprint* **2021**, arXiv:2105.05537.
22. Yang, Y.; Wang, Y.; Zhu, C.; Zhu, M.; Sun, H.; Yan, T. Mixed-Scale Unet Based on Dense Atrous Pyramid for Monocular Depth Estimation. *IEEE Access* **2021**, *9*, 114070–114084. [CrossRef]
23. Choudhary, R.; Sharma, M.; Anil, R. 2T-UNET: A Two-Tower UNet with Depth Clues for Robust Stereo Depth Estimation. *arXiv Preprint* **2022**, arXiv:2210.15374.
24. Zhao, T.; Pan, S.; He, X. ResUnet++ for Sparse Samples-based Depth Prediction. In Proceedings of the 2021 IEEE 15th International Conference on Electronic Measurement & Instruments (ICEMI), Harbin, China, 9–11 August 2021; pp. 242–246.
25. 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, 26 June–1 July 2016; pp. 770–778.
26. 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.
27. Torralba, A.; Oliva, A. Depth estimation from image structure. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 1226–1238. [CrossRef]
28. 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]
29. Karsch, K.; Liu, C.; Kang, S.B. Depth extraction from video using non-parametric sampling. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 775–788.

30. Eigen, D.; Fergus, R. Predicting depth, surface normals 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. [CrossRef]
31. Watson, J.; Firman, M.; Brostow, G.J.; Turmukhambetov, D. Self-Supervised Monocular Depth Hints. *arXiv Preprint* **2019**, arXiv.1909.09051.
32. Tosi, F.; Aleotti, F.; Poggi, M.; Mattoccia, S. Learning Monocular Depth Estimation Infusing Traditional Stereo Knowledge. *arXiv* **2019**, arXiv:1904.04144.
33. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3684–3692.
34. Ye, X.; Fan, X.; Zhang, M.; Xu, R.; Zhong, W. Unsupervised Monocular Depth Estimation via Recursive Stereo Distillation. *IEEE Trans. Image Process.* **2021**, *30*, 4492–4504. [CrossRef]
35. Liu, J.; Li, Q.; Cao, R.; Tang, W.; Qiu, G. MiniNet: An extremely lightweight convolutional neural network for real-time unsupervised monocular depth estimation. *Isprs J. Photogramm. Remote. Sens.* **2020**, *166*, 255–267. [CrossRef]
36. Fang, Z.; Chen, X.; Chen, Y.; Gool, L.V. Towards good practice for cnn-based monocular depth estimation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1091–1100.
37. Ye, X.; Chen, S.; Xu, R. DPNet: Detail-preserving network for high quality monocular depth estimation. *Pattern Recognit.* **2021**, *109*, 107578. [CrossRef]
38. Alhashim, I.; Wonka, P. High quality monocular depth estimation via transfer learning. *arXiv* **2018**, arXiv:1812.11941.
39. Gan, Y.; Xu, X.; Sun, W.; Lin, L. *Monocular Depth Estimation with Affinity, Vertical Pooling, and Label Enhancement*; Springer: Berlin/Heidelberg, Germany, 2018.
40. Xu, H.; Li, F. Multilevel Pyramid Network for Monocular Depth Estimation Based on Feature Refinement and Adaptive Fusion. *Electronics* **2022**, *11*, 2615. [CrossRef]
41. Pei, M. MSFNet:Multi-scale features network for monocular depth estimation. *arXiv* **2021**, arXiv:2107.06445.
42. Lee, J.H.; Han, M.K.; Ko, D.W.; Suh, H. From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation. *arXiv* **2019**, arXiv:1907.10326.
43. Chen, P.Y.; Liu, A.H.; Liu, Y.C.; Wang, Y.C.F. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2619–2627. [CrossRef]
44. Zhu, S.; Brazil, G.; Liu, X. The edge of depth: Explicit constraints between segmentation and depth. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13116–13125.
45. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008
46. Hu, D. An introductory survey on attention mechanisms in NLP problems. In Proceedings of the SAI Intelligent Systems Conference, London, UK, 5–6 September 2019; pp. 432–448.
47. Lei, S.; Yi, W.; Ying, C.; Ruibin, W. Review of attention mechanism in natural language processing. *Data Anal. Knowl. Discov.* **2020**, *4*, 1–14.
48. Liu, P.; Zhang, Z.; Meng, Z.; Gao, N. Monocular depth estimation with joint attention feature distillation and wavelet-based loss function. *Sensors* **2021**, *21*, 54. [CrossRef]
49. Chen, S.; Fan, X.; Pu, Z.; Ouyang, J.; Zou, B. Single image depth estimation based on sculpture strategy. *Knowl.-Based Syst.* **2022**, *250*, 109067. [CrossRef]
50. Makarov, I.; Bakhanova, M.; Nikolenko, S.; Gerasimova, O. Self-supervised recurrent depth estimation with attention mechanisms. *Peerj Comput. Sci.* **2022**, *8*, e865. [CrossRef]
51. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
52. Song, M.; Lim, S.; Kim, W. Monocular depth estimation using laplacian pyramid-based depth residuals. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 4381–4393. [CrossRef]
53. Imambi, S.; Prakash, K.B.; Kanagachidambaresan, G.R. Pytorch. In *Programming with TensorFlow: Solution for Edge Computing Applications*; Springer: Cham, Switzerland, 2021; pp. 87–104.
54. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.