International Journal of Geo-Information

MDPI

Article

Learning Effective Geometry Representation from Videos for Self-Supervised Monocular Depth Estimation

Hailiang Zhao †, Yongyi Kong †, Chonghao Zhang, Haoji Zhang and Jiansen Zhao *

Merchant Marine College, Shanghai Maritime University, Shanghai 200135, China; 202110111138@stu.shmtu.edu.cn (H.Z.); 202210621089@stu.shmtu.edu.cn (Y.K.); 202010413004@stu.shmtu.edu.cn (C.Z.); 202210121093@stu.shmtu.edu.cn (H.Z.)

* Correspondence: jszhao@shmtu.edu.cn

† These authors contributed equally to this work.

Abstract: Recent studies on self-supervised monocular depth estimation have achieved promising results, which are mainly based on the joint optimization of depth and pose estimation via high-level photometric loss. However, how to learn the latent and beneficial task-specific geometry representation from videos is still far from being explored. To tackle this issue, we propose two novel schemes to learn more effective representation from monocular videos: (i) an Inter-task Attention Model (IAM) to learn the geometric correlation representation between the depth and pose learning networks to make structure and motion information mutually beneficial; (ii) a Spatial-Temporal Memory Module (STMM) to exploit long-range geometric context representation among consecutive frames both spatially and temporally. Systematic ablation studies are conducted to demonstrate the effectiveness of each component. Evaluations on KITTI show that our method outperforms current state-of-the-art techniques.

Keywords: self-supervised learning; monocular depth estimation

Citation: Zhao, H.; Kong, Y.; Zhang, C.; Zhang, H.; Zhao, J. Learning Effective Geometry Representation from Videos for Self-Supervised Monocular Depth Estimation. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 193. https://doi.org/10.3390/ijgi13060193

Academic Editors: Wolfgang Kainz, Junxing Zheng and Peng Cao

Received: 2 April 2024
Revised: 4 June 2024
Accepted: 5 June 2024
Published: 11 June 2024

1. Introduction

Understanding the 3D structure of scenes is an essential topic in machine perception, which plays a crucial part in applications such as autonomous driving, robot vision, visual reality and so on [1–4]. For most scenarios, there is vast latent geometric information existing in the input videos. One of the key challenges in this domain is how to acquire effective task-specific geometry representation from videos to help obtain more accurate and reliable depth information.

Recently, there have been some successful attempts [1,2,5] to execute monocular depth estimation and visual odometry prediction together in a self-supervised manner by giving full consideration of the transformation between consecutive frames. In this pipeline, two networks are generally used to predict the depth and camera pose separately, which are then jointly exploited to warp source frames to the target ones, converting the depth estimation problem to a reprojection error minimization process, as shown in Figure 1a.

Despite various extensions of the self-supervised pipeline by adding more penalty items [5–7] or joining with other tasks (optical flow or segmentation) [8,9], these methods only design various high-level loss functions to combine and regularize the network learning, neglecting to leverage valuable geometry representation from videos, e.g., inter-task geometric correlation learning, inter-frame long-range dependency learning, and 3D geometry consistency representation from continuous frames.

Intuitively, modeling the process of perceiving 3D structure from videos can be informed by our human experience. According to the research in biology and neuroscience [10], human brains process motion information during the inference of depth, and conversely, the perceived depth information can bring significant benefits to motion estimation [11]. Inspired by this biological mechanism, we present an Inter-task Attention Module (IAM) to guide the feature-level inter-task geometric correlation learning. It can enhance the interaction

between the depth and pose estimation networks and is effective in making structure and motion information mutually beneficial for improving estimation accuracy.

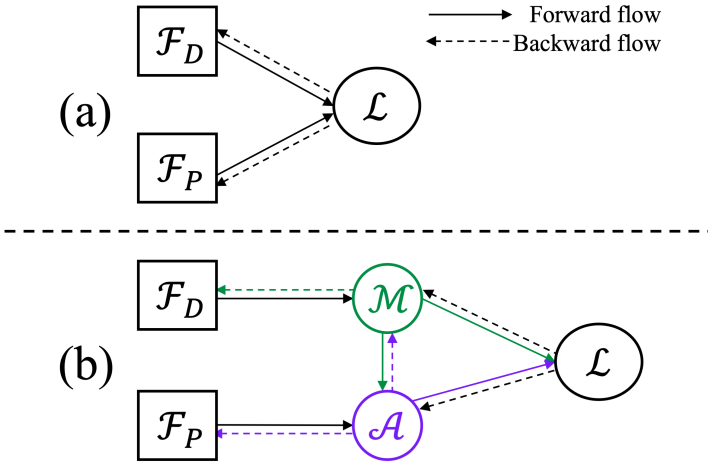


Figure 1. Comparison of the learning process of the general pipeline (**a**) and our method (**b**) for self-supervised monocular depth estimation. Different from the general pipeline that learns the depth feature F_D and the pose feature F_P separately using a 2D photometric loss L, we propose a new scheme for learning better representation from videos. A memory mechanism M is devised to exploit the long-range context from videos for depth feature learning. An inter-task attention mechanism A is devised to leverage depth information for helping pose feature learning, which inversely benefits depth feature learning as well via gradient back-propagation.

Furthermore, many psychologists believe that humans rely on not only immediate sensory feedback but also perception memories from the past for understanding an environment [12,13]. Similarly, it is significant to help networks learn a representation leveraging long-range context and memorizing historical information to disambiguate and realize more precise perception. Therefore, we introduce a Spatial-Temporal Memory Module (STMM) to learn spatial and temporal dependency from video clips and mimic the above perception mechanism of human beings. We embody an STMM based on the Non-local network [14], which is demonstrated to be effective in modeling long-range information, after exploring various attention structures.

In summary, the learning process of our method is shown in Figure 1b, and our main contributions are as follows:

- We devise an Inter-task Attention Module to exploit the inter-task geometric correlation between depth and pose estimation networks. It learns attention maps from depth information as guidance to help the pose network identify key regions to be targeted. To the best of our knowledge, this is the first attempt to propose this idea for exploiting the inter-task geometric correlation in self-supervised monocular depth estimation.
- We introduce a Spatial-Temporal Memory Module in a depth estimation network to leverage the spatial and temporal geometric context among consecutive frames, which is effective for utilizing historical information and improving estimation results.
- We conduct comprehensive empirical studies on the KITTI dataset, and the single-frame inference result of our method outperforms state-of-the-art methods by a relative gain of 6.6% based on the major evaluation metric.

2. Related Work

2.1. Inter-Task Monocular Video Learning

Ref. [1] proposed a fully unsupervised end-to-end network for training with monocular videos that can jointly predict the depth and pose transformation between consecutive frames. The core technique is a spatial transformer network [15] to synthesize target frames from source frames, which converts the depth estimation problem to a reprojection error

minimization process. This pipeline was then extended by plenty of researchers. Ref. [6] added a feature-based warping loss upon the original photometric loss and trained the networks with stereo image pairs to resolve the scale ambiguity. Ref. [2] further proposed an auto-masking strategy to handle situations where the camera is static or objects move at the same speed as the camera, yielding more accurate results. Moreover, some works combined depth and pose estimation with other tasks, e.g., normal, segmentation, and optical flow estimation. Ref. [7] implemented the estimation of normal in scenes and incorporated an edge-aware depth-normal consistency constraint. Ref. [16] used the Mask R-CNN [17] model to extract semantic information and obtain pre-computed object masks to filter out moving objects. Ref. [8] defined a cascaded network to jointly learn depth, pose, and optical flow for handling rigid motions and moving objects separately, using a forward-backward coherence loss. Ref. [18] also presented an architecture to simultaneously learn depth, ego-motion, and optical flow and focused on enforcing cross-task consensus between depth and optical flow. JPerceiver [19] jointly learn depth estimation, visual odometry, and Bird's-Eye-View segmentation. Despite the progress made by these methods, almost all of them follow the pipeline in [1] that uses separate networks to learn depth, pose, and other tasks without any interactions before being combined into the final loss. By contrast, we propose an IAM to learn geometry information from the depth network as guidance to help the pose network learn more valuable representation for pose estimation. Notably, the two tasks are joint-optimized via high-level photometric error, which enables an interaction between two networks via gradient back-propagation, meaning that depth tasks can also benefit from the IAM and learn more useful representation to improve the estimation results.

2.2. Long-Range Representation Learning

Taking videos instead of single images as input is extremely important for many applications, such as autonomous driving, robotic vision, and drones. However, the rich long-range dependency, including spatial and temporal correlations, is still far from being fully utilized to eliminate ambiguity and obtain more consistent estimation. UnDeepVO [20] was proposed as the first end-to-end visual odometry by combining CNNs and two stacking LSTMs to achieve simultaneous representation learning and sequential modelling of the monocular VO. Kumar et al. [21] proposed a convolutional LSTM-based network architecture for depth to capture inter-frame dependencies and variations. Wang et al. [22] also adopted ConvLSTM architecture, multi-view reprojection, and forward-backward consistency constraints to utilize the temporal information effectively. These research efforts demonstrated that utilizing long-range information from videos is helpful in learning more effective representation for both depth and pose networks and improving the estimation accuracy. However, convolutional and recurrent operations both process a local neighborhood, either in space or time [14]. And, all these RNN methods focused only on temporal dependency learning without long-range spatial context, which is not that useful for single image inference. Recently, Transformer-based methods [23–27] are attracting researcher's attention to use the stronger backbone to extract better visual representation. MonoViT [26] introduces a Vision Transformer-based encoder for self-supervised monocular depth estimation, leveraging both local and global reasoning capabilities to achieve state-of-the-art performance on the KITTI dataset. PixelFormer [25] was proposed as a novel pixel query refinement approach for monocular depth prediction, using a Skip Attention Module to effectively fuse global and local features. MonoFormer [27] was introduced as a deep analysis of self-supervised monocular depth estimation models, and the authors proposed methodological enhancements to improve their generalization across various environments. These Transformer-based methods achieve impressive performance but require larger network structures and consume more computational resources. In this paper, aligning with our lightweight network design, we introduce an STMM module to learn long-range geometric relationships both spatially and temporally among pixels in consecutive frames. We embody STMM based on the Non-local network [14] after exploring various attention structures. STMM is demonstrated to be beneficial for both multi-frame training and single-frame inference.

3. Method

3.1. Problem Definition

A typical self-supervised monocular depth estimation pipeline is mainly built upon the perspective projection among consecutive frames. Taking $<I_t, I_{t+1}, \ldots, I_{t+n}>$ as a training video within time window N, once the depth D_{t+n} and camera transformation $T_{t+n\to t}$ are obtained, we can warp the source frame I_{t+n} to reconstruct the target frame $\hat{I}_{t+n\to t}$ using the differentiable bilinear sampling approach [15], which can be formulated as

$$D_t^{ij} I_t^{ij} = KRK^{-1} D_{t+n}^{ij} I_{t+n}^{ij} + Kt, \tag{1}$$

where I_{t+n}^{ij} is the homogeneous coordinate given image I_{t+n} and D_{t+n}^{ij} denotes the depth value of the view I_{t+n}. Given a rotation matrix R, translation vector t, and camera intrinsics K, the transformed homogeneous coordinate I_t^{ij} and depth D_t^{ij} can be obtained. Thus, the reprojected image coordinate $\hat{I}_{t+n\to t}$ can be acquired by dehomogenization of $D_t^{ij} I_t^{ij}$.

Then, the self-supervised learning is conducted based on the difference between the synthetic view $\hat{I}_{t+n\to t}$ and the original view I_t:

$$L_e = L_r(I_t, \hat{I}_{t+n\to t}). \tag{2}$$

Here, L_r denotes a consistency measurement loss.

3.2. Network Architecture

As shown in Figure 2a, our network is composed of two main networks for depth estimation and pose estimation, respectively. Meanwhile, the pose network is split into two branches for the estimation of rotation and translation. The proposed IAM is used to address the importance of geometric correlation representation between the depth and pose tasks, while the proposed STMM is used to exploit the long-range geometric relevance among continuous frames. The details of IAM and STMM are presented later.

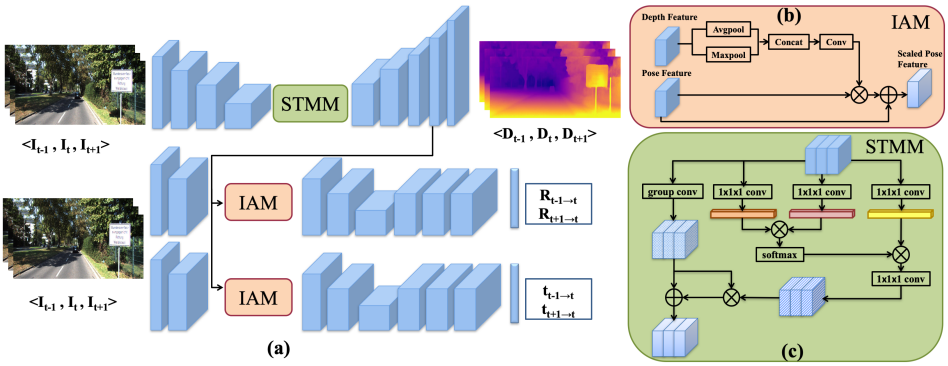


Figure 2. Illustration of our network framework (**a**) and the architecture of the IAM (**b**) and the STMM (**c**). The network takes three consecutive frames as input to learn the long-range geometric correlation representation by introducing STMM after the encoder. The pose network is split into two branches to predict rotation R and translation t separately. The IAM is applied after the second convolution layer of both R and t branches, learning valuable geometry information to assist R and t branches in leveraging inter-task correlation representation.

The depth network adopts an encoder–decoder architecture in a U-shape with skip connections similar to DispNet [28]. The encoder is a Resnet18 [29] network pre-trained on ImageNet [30]. Our depth decoder is similar to that of [2], using sigmoid activation functions in multi-scale side outputs and ELU nonlinear functions [31] otherwise. Most importantly, we take a three-frame snippet as the sequential input and stack the encoded features as the input of the STMM to learn the temporal and spatial geometric correlations

among video sequences during training. The outputs are then decoded into a three-frame depth sequence.

The pose network takes two consecutive frames as input at each time and outputs the corresponding pose transformation based on an encoder–decoder structure as well. To generate more accurate estimations, the network is divided into two branches to calculate R and t, respectively. In the feature encoding phase, the IAM is employed to produce the attention from depth features as guidance for the R and t branches.

3.2.1. Inter-Task Attention Module

The IAM aims to leverage the latent geometric correlation between depth and pose estimation tasks during learning. To exploit geometry information, features from the penultimate layer of the depth decoder are first stacked in the same order as the input sequence of the IAM. In the IAM, the features are first processed by an average pooling layer and a max pooling layer along the channel axis and then concatenated together as a compact representation, as previous studies [32,33] show that pooling layers can help highlight features. Furthermore, a subsequent convolution layer is used to obtain the attention maps.

The varying weights regarding different pixels in the learned attention maps guide the R and t branches in deciding what feature should be the focus and prioritized. Therefore, we use the attention maps to obtain scaled pose features by element-wise multiplication, which are then added to the original pose features as a residual item. A schematic diagram of the IAM is provided in Figure 2b, which can be formulated as

$$F'_p = F_p(W_c[AVP(F_{mn}); MAP(F_{mn})]) + F_p. \tag{3}$$

Here, F_{mn} represents the stacked depth features encoded from continuous frames I_m and I_n by the depth network, while F_p and F'_p denote the original pose feature and the attended one, respectively. The average and max pooling layers are represented by AVP and MAP, respectively, while W_c denotes the learnable weight of the convolution layer.

Intuitively, the geometric patterns of learned attention maps for the two branches should be different or even opposite, as nearby regions tend to matter more to translation, while distant pixels may play a more important role in deciding the rotation. The ablation study and visualization results demonstrate that the IAM does learn different attention patterns for the R and t branches from geometry information. They are utilized to guide the pose network to learn more valuable and effective representation, improving estimation accuracy for both tasks via joint optimization.

3.2.2. Spatial-Temporal Memory Module

Instead of learning representation from each frame individually [1,2,5] in the depth network, we introduce an STMM to leverage and aggregate long-range geometric correlation from both a spatial context and a temporal context to obtain a more representative feature embedding for depth estimation. To this end, various attention structures can be leveraged, including SE [34], CBAM [32], and Non-local attention [14]. After a comparison study, we found that Non-local attention is more effective at capturing long-range context. Thereby, we chose it to embody STMM in this work. First, the encoded depth features from several consecutive frames are concatenated together and used as the input of the Non-local block. Then, the attention map for each frame is obtained, which is multiplied and added to the original depth features after a group convolution layer. The group number is the same as the number of input frames. As shown in Figure 2c, in STMM, the aggregated depth feature F'_D is calculated as follows:

$$F'_D = NL(F_D)W_\delta F_D + W_\delta F_D, \tag{4}$$

where F_D is the input depth feature and NL means the operation of Non-local block. W_δ is the learnable weight of the group convolution layer.

4. Experiments

4.1. Depth Estimation Result

Although our method is trained with three-frame snippet input, it can infer single image depth during inference by stacking the same encoded features three times before being fed into the STMM. Following the common evaluation protocol [1], we report the single-frame inference results in the following experiments, although better results can be achieved by leveraging multiple frames. The extensive experimental results on KITTI are presented in Table 1, from which it is clear that our method outperforms all prior works trained with monocular and even stereo videos in a self-supervised manner. The visual results shown in Figure 3 demonstrate our method can generate more accurate and sharper depth maps, especially for challenging situations, such as moving objects, distant objects, and fine structures. More experiment results for both depth and pose estimation can be found in Supplementary Materials.

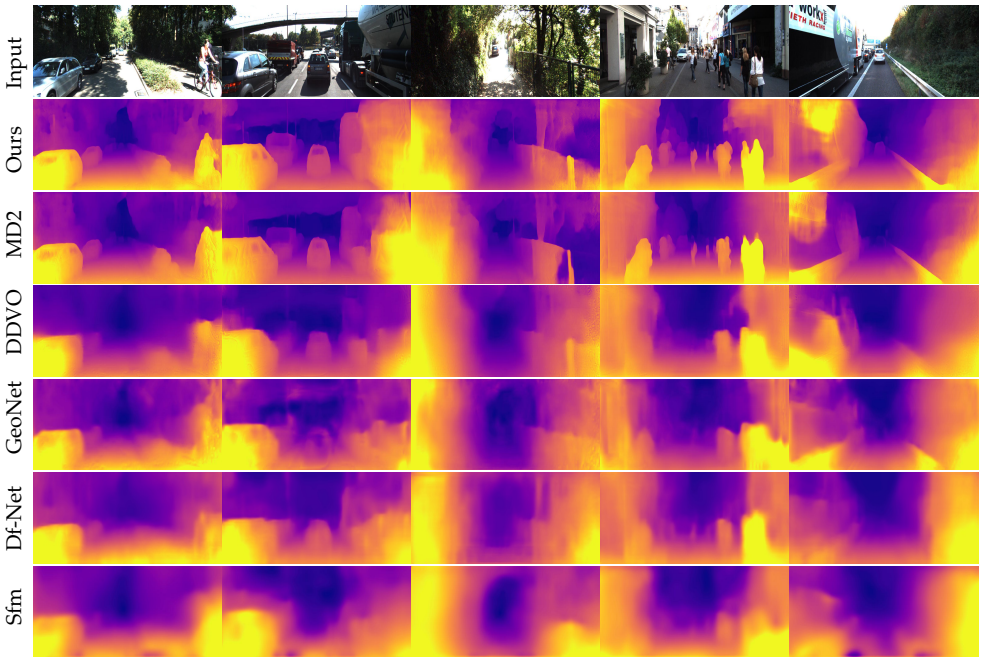


Figure 3. Qualitative results on KITTI test set. Our method produces more accurate depth maps with low-texture regions, moving vehicles, delicate structures, and object boundaries.

Table 1. Quantitative performance of single depth estimation over KITTI test set [35]. For a fair comparison, all the results are evaluated, taking 80 m as the maximum depth threshold. The "S" and "M" in the train column mean stereo and monocular inputs for training, while "R18" and "R50" denote the used Resnet [29] version. "†" means updated result after publication. We train our models using only KITTI without any post-processing. The best results are illustrated with bold text.

Methods	Train	Error Metric ↓				Accuracy Metric ↑		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
[36] †	S	0.152	1.226	5.849	0.246	0.784	0.921	0.967
MD (R50) † [37]	S	0.133	1.142	5.533	0.230	0.830	0.936	0.970
SuperDepth [38]	S	0.112	0.875	4.958	0.207	0.852	0.947	0.977
MD2 [2]	S	0.107	0.849	4.764	0.201	0.874	0.953	0.977
[1] †	M	0.183	1.595	6.709	0.270	0.734	0.902	0.959
[5]	M	0.163	1.240	6.220	0.250	0.762	0.916	0.968
GeoNet † [8]	M	0.149	1.060	5.567	0.226	0.796	0.935	0.975
DDVO [39]	M	0.151	1.257	5.583	0.228	0.810	0.936	0.974
DF-Net [18]	M	0.150	1.124	5.507	0.223	0.806	0.933	0.973

Table 1. *Cont.*

Methods	Train	Error Metric ↓				Accuracy Metric ↑		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Struct2depth [16]	M	0.141	1.026	5.142	0.210	0.845	0.845	0.948
CC [9]	M	0.140	1.070	5.326	0.217	0.826	0.941	0.975
SC-SFMLearner [40]	M	0.137	1.089	5.439	0.217	0.830	0.942	0.975
HR [41]	M	0.121	0.873	4.945	0.197	0.853	0.955	0.982
MD2(R18) [2]	M	0.115	0.882	4.701	0.190	0.879	0.961	0.982
DeFeat [42]	M	0.126	0.925	5.035	0.200	0.862	0.954	0.980
[43]	M	0.113	**0.704**	4.581	0.184	0.871	0.961	**0.984**
Ours (R18)	M	**0.106**	0.761	**4.545**	**0.182**	**0.890**	**0.965**	0.983
Ours (R50)	M	**0.105**	0.731	**4.412**	**0.181**	**0.891**	**0.965**	0.983

4.2. Evaluation of Generalization Ability

Though our models were only trained on KITTI [44], competitive results can be achieved on unseen datasets without any fine-tuning. We evaluated our method on two outdoor datasets: Make3D [45] and Cityscapes [46]. In Table 2, our model outperforms other self-supervised methods on the Make3D test protocol, showing good domain adaptation ability. The qualitative comparison in Figure 4 on Cityscapes provides additional intuitive evidence on the generalization ability. More test results can be found in Supplementary Materials.

Table 2. Quantitative results on the Make3D dataset. The best results are illustrated with bold text.

Methods	Train	Abs Rel	Sq Rel	RMSE	log10
[47]	D	0.475	6.562	10.05	0.165
[37]	S	0.544	10.94	11.760	0.193
[1]	M	0.383	5.321	10.470	0.478
DDVO [39]	M	0.387	4.720	8.090	0.204
MD2 [2]	M	0.322	3.589	7.417	0.163
Ours	M	**0.316**	**3.200**	**7.095**	**0.158**

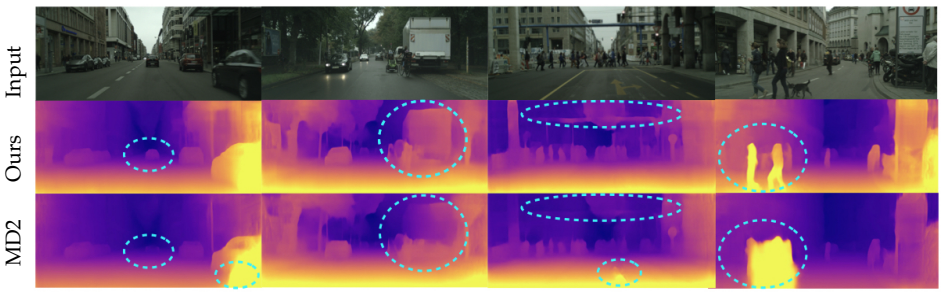


Figure 4. Visual results evaluated on the Cityscapes dataset. The evaluation uses models trained on KITTI without any refinement. Compared with the methods in [2], our method generates higher-quality depth maps and captures moving and slim objects better. The difference is highlighted with the dashed circles.

4.3. Ablation Study

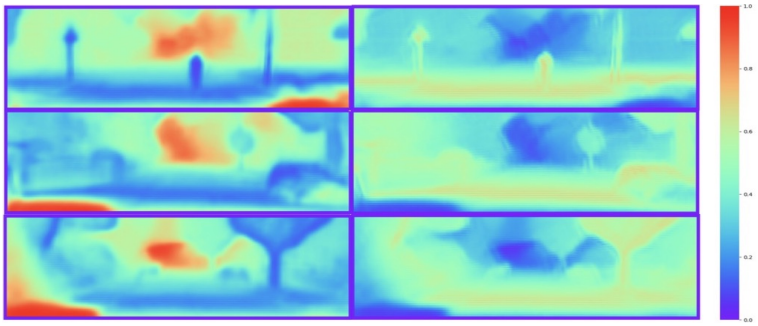
The ablation study was conducted on KITTI to highlight the effects of individual components of our model. Table 3 shows the detailed results by removing specific component(s) from our model. The in-depth analysis of each part is given in corresponding sections. Moreover, we conducted systematic experiments to test the performance under various training conditions, listed in Table 3, including input resolution (1024×320 vs. 640×192), backbone networks (Resnet18 vs. Resnet50), and with/without pretraining.

Table 3. Ablation results on KITTI with each individual component removed and using backbone networks (Resnet18 or Resnet50) and different resolutions of input videos during training. The term "plain" means removing all components, while "pre" means pretraining on ImageNet. The best results are illustrated with bold text.

	Resolution	Net	IAM	STMM	Abs Rel	Sq Rel	RMSE	RMSE Log
Ours (full)	1024 × 320	R50	√	√	**0.105**	**0.731**	**4.412**	**0.181**
Ours (full)	1024 × 320	R18	√	√	**0.106**	**0.761**	**4.545**	**0.182**
Ours w/o IAM	1024 × 320	R18		√	0.112	0.844	4.815	0.190
Ours w/o STMM	1024 × 320	R18	√		0.111	0.829	4.799	0.190
Ours w/o STMM	1024 × 320	R18	√	CBAM	0.109	0.778	4.591	0.186
Ours w/o STMM	1024 × 320	R18	√	SE	0.111	0.799	4.704	0.187
Ours (plain)	1024 × 320	R18			0.120	0.915	4.972	0.196
Ours (full) smaller	640 × 192	R18	√	√	0.110	0.809	4.616	0.185
Ours (full) w/o pre	640 × 192	R18	√	√	0.124	0.847	4.713	0.196

Effect of Inter-Task Attention Module

As mentioned before, when predicting the pose from two consecutive frames, we believe that different geometry information has a different impact on the estimation of rotation and translation. Thus, we introduce valuable attention guidance learned by the IAM into the prediction of R and t. The attention maps learned for the two branches during training are visualized in Figure 5, with color variation denoting different weight values. The attention maps indicate that the two branches did learn different geometric priorities from the depth information to help with their own estimation and conversely improve the depth estimation result, as shown in Table 3. The learned geometric patterns demonstrate that the estimation of R attaches more importance to farther regions and corner places, while the t branch values closer areas more. Our IAM adopts an attention mechanism and works in a generalized representation learning manner to utilize the geometric correlation between depth and pose, which can also be useful in other similar tasks to improve estimation quality.



R branch *t* branch

Figure 5. The visualization of learned attention maps in the IAM. It indicates the IAM places distinct emphasis on different regions for two branches to improve their estimation.

Effects of STMM on distant objects. The motivation of the STMM is to leverage the rich temporal and spatial geometric dependency among continuous frames. By exploiting the temporal information of depth features from three consecutive frames as input, The STMM is helpful for utilizing historical knowledge within the time window and enhancing the estimation of distant objects. During inference, the input of networks is a single image, and our STMM can exploit only spatial correlations from the pixels within the single image. The ablation study results shown in Table 3 demonstrate the benefit of STMM

compared with the models replacing STMM with other attention structures (CBAM and SE). To better evaluate our model's performance on estimating distant objects, we segmented each scene into two groups of pixels according to a distance of 20 m, following [41] to ensure fairness. We conducted an ablation test for the estimation of distant objects, and the results are listed in Table 4. The results show that removing the STMM severely decreases the performance of our model for distant objects, which demonstrates the effectiveness of STMM in distant scenes.

Table 4. Ablation study and comparison for distant objects.

Methods	Dist	Abs Rel	Sq Rel	RMSE	RMSE Log
HR [41]	≤20	0.102	0.391	1.959	0.146
Ours	≤20	**0.076**	**0.189**	**1.410**	**0.123**
HR [41]	>20	0.187	**2.430**	9.695	0.305
Ours	>20	**0.152**	2.556	**9.226**	**0.211**
Ours w/o STMM	>20	0.179	2.803	9.667	0.291
Ours w/o IAM	>20	0.158	2.643	9.346	0.234

4.4. Evaluation with Improved Ground Truth

The main evaluation method proposed by Eigen [35] uses the reprojected raw LIDAR points as ground truth, which brings severe effects on the estimation of tricky cases, such as occlusion, object motion, and so on. To conduct a fair comparison with [2], we also adopted the annotated depth map from the official KITTI website as ground truth to evaluate methods. These annotated depth maps introduced by [48] tackle the above-mentioned tough cases to improve ground truth using stereo pair. We compared our models with other self-supervised methods, as shown in Table 5. The results demonstrate that our method outperforms all previous methods, including both monocular and stereo training approaches.

Table 5. Quantitative performance of a single depth estimation using an annotated depth map [48] as ground truth. For a fair comparison, all the results are evaluated, taking 80 m as the maximum depth threshold. The resolution column indicates the size of input images during training. We trained our network using only KITTI without any post-processing. The best results are illustrated with bold text. "†" means updated result after publication.

Methods	Train	Resolution	Error Metric ↓				Accuracy Metric ↑		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
MD R50 † [37]	S	512 × 256	0.109	0.811	4.568	0.166	0.877	0.967	0.988
3Net (VGG) [49]	S	512 × 256	0.119	0.920	4.824	0.182	0.856	0.957	0.985
3Net (R50) [49]	S	512 × 256	0.102	0.675	4.293	0.159	0.881	0.969	0.991
SuperDepth [38]	S	1024 × 382	0.090	0.542	3.967	0.144	0.901	0.976	0.993
MD2 [2]	S	1024 × 320	0.085	0.537	3.868	0.139	0.912	0.979	0.993
Zhou et al. † [1]	M	416 × 128	0.176	1.532	6.129	0.244	0.758	0.921	0.971
Mahjourian et al. [5]	M	416 × 128	0.134	0.983	5.501	0.203	0.827	0.944	0.981
GeoNet [8]	M	416 × 128	0.132	0.994	5.240	0.193	0.833	0.953	0.985
DDVO [39]	M	416 × 128	0.126	0.866	4.932	0.185	0.851	0.958	0.986
CC [9]	M	832 × 256	0.123	0.881	4.834	0.181	0.860	0.959	0.985
MD2 (R18) [2]	M	1024 × 320	0.090	0.545	3.942	0.137	0.914	0.983	0.995
Ours (R18)	M	1024 × 320	**0.083**	**0.447**	**3.667**	**0.126**	**0.924**	**0.986**	**0.997**

4.5. Single-Scale Evaluation

Monocular training methods usually need a scaling step during evaluation because monocular solutions do not have a certain metric scale during training. For evaluation, ref. [1] calculated the median of each predicted depth map and the ground truth as the scaling factor. However, using a distinct scaling factor for every frame may cause an unfair

advantage in contrast to stereo methods, which use a certain scale for all images, according to [2].

In [2], the authors changed this evaluation protocol by taking the median of all the scaling ratios of the depth maps on the test set as a constant scale for all test images. To conduct a fair comparison, we adopted this modified protocol to validate our methods. The quantitative comparison can be found in Table 6, in which our method still outperforms all previous approaches. The standard deviation σ_{scale} of our method is also lower than other methods, which indicates our approach can generate more consistent depth map scales.

Table 6. Quantitative performance using single scale on KITTI Eigen test set [35]. "†" means updated result after publication.

Methods	σ_{scale}	Error Metric ↓				Accuracy Metric ↑		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou et al. † [1]	0.210	0.258	2.338	7.040	0.309	0.601	0.853	0.940
Mahjourian et al. [5]	0.189	0.221	1.663	6.220	0.265	0.665	0.892	0.962
GeoNet [8]	0.172	0.202	1.521	5.829	0.244	0.707	0.913	0.970
DDVO [39]	0.108	0.147	1.014	5.183	0.204	0.808	0.946	0.983
CC [9]	0.162	0.188	1.298	5.467	0.232	0.724	0.927	0.974
MD2 (R18) [2]	0.093	0.109	0.623	4.136	0.154	0.873	0.977	0.994
Ours (R18)	**0.082**	**0.096**	**0.507**	**3.828**	**0.139**	**0.898**	**0.983**	**0.996**

4.6. Results with Post-Processing

To finish the comprehensive comparison with the previous state-of-the-art work [2], we also evaluated our method with post-processing. This technique was proposed by [37] to improve stereo-based methods, but it has proved effective for monocular training methods as well. As shown in Table 7, this post-processing step did improve the result of our methods. In addition, the performance of our models exceeds the post-processed results of [2] even without post-processing.

Table 7. Evaluation results with post-processing compared with MD2 [2].

Methods	Resolution	Error Metric ↓				Accuracy Metric ↑		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
MD2 [2]	1024 × 320	0.115	0.882	4.701	0.190	0.879	0.961	0.982
MD2 (pp) [2]	1024 × 320	0.112	0.838	4.607	0.187	0.883	0.962	0.982
Ours (R18)	1024 × 320	**0.106**	**0.761**	**4.545**	**0.182**	**0.890**	**0.965**	**0.983**
Ours (R18 + pp)	1024 × 320	**0.104**	**0.726**	**4.457**	**0.180**	**0.893**	**0.965**	**0.984**
Ours (R50)	1024 × 320	**0.105**	**0.731**	**4.412**	**0.181**	**0.891**	**0.965**	**0.983**
Ours (R50 + pp)	1024 × 320	**0.104**	**0.709**	**4.352**	**0.179**	**0.894**	**0.966**	**0.984**

4.7. Inference Speed

The depth inference task usually plays an important role in autonomous driving and robotic vision. In these fields, there generally are strict requirements for calculation speed. To test the practicability of models, we calculated the inference speed of our models under the condition with a GPU or CPU device. In Table 8, we list the average time cost for testing 697 frames of Eigen's test set [35].

Table 8. Inference speed on Eigen's test set [35]. "Time" means the total time required for the inference for 697 frames.

Device	Time (s)	Speed (f/s)
GPU	60.9	11.5
CPU	8721.1	0.08

The inference speeds of our model on the GPU and CPU devices are significantly different. Frames of the KITTI [44] dataset were collected at 10 Hz, and our inference speed on the GPU device is over 10 fps, which indicates the practicability of our method on the GPU device. However, the speed on the CPU device is much lower, which will be improved in our future work.

5. Visual Odometry

Trajectory estimation is also very important for environment perception [1,50]. Since the accuracies of the pose and depth estimations are correlated, our proposed method not only produces a high-quality depth estimation but also improves the accuracy of the pose estimation. We used the 00-08 sequences of the KITTI odometry split for training and the 09-10 sequences for the evaluation, as in [1]. We compared the absolute trajectory errors calculated in overlapping five-frame and three-frame snippets with various methods. In Table 9, our results are clearly superior to the latest self-supervised monocular training methods and also close to the traditional ORB-SLAM method with the loop closure step. The estimated trajectories by our models and other methods are shown in Figure 6 for comparison purposes.

Table 9. Results of the visual odometry on the KITTI Odometry dataset. "Frame" means the number of frames used when calculating absolute trajectory error. "†" means updated result after publication. The best results are illustrated with bold text.

Methods	Sequence09	Sequence10	Frame
ORB-SLAM	0.014 ± 0.008	0.012 ± 0.011	
ORB-SLAM (short)	0.064 ± 0.141	0.064 ± 0.130	
[1]	0.021 ± 0.017	0.020 ± 0.015	5
Zhou et al. †	0.016 ± 0.009	0.013 ± 0.009	5
DDVO [39]	0.045 ± 0.108	0.033 ± 0.074	5
DF-Net [18]	0.017 ± 0.007	0.015 ± 0.009	5
[2]	0.017 ± 0.008	0.015 ± 0.010	5
Ours	$\mathbf{0.015 \pm 0.007}$	$\mathbf{0.015 \pm 0.009}$	5
[5] (no ICP)	0.014 ± 0.010	0.013 ± 0.011	3
[5] (with ICP)	0.013 ± 0.010	0.012 ± 0.011	3
[5]	0.013 ± 0.010	0.012 ± 0.011	3
[2]	0.013 ± 0.007	0.011 ± 0.008	3
Ours	$\mathbf{0.009 \pm 0.005}$	$\mathbf{0.010 \pm 0.007}$	3

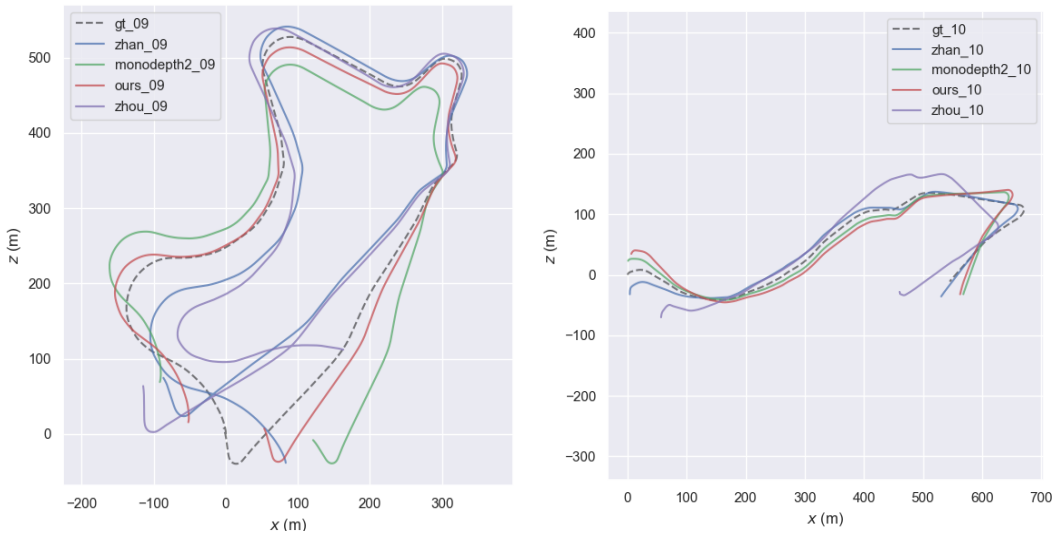


Figure 6. Visual comparison of the visual odometry trajectories. Full trajectories are plotted using the Evo visualization tool [51].

6. Conclusions

Our work is dedicated to the self-supervised monocular depth estimation problem with a focus on learning more effective task-specific representation during learning. In our method, the IAM can actively explore the geometric correlation between depth- and pose-estimation tasks by learning attentive representation from depth to guide the pose network to highlight and leverage more valuable geometry information, which improves the estimation quality of depth and pose. We also introduce an STMM to learn the spatial and temporal geometric dependencies among sequential frames, which are helpful for utilizing long-range historical knowledge within the time window to perceive distant objects. Experimental results demonstrated that our method is superior to existing state-of-the-art approaches and can generate higher-quality depth maps. In our future work, we will explore more powerful network architectures, such as Transformers and their corresponding attention mechanisms.

Author Contributions: Conceptualization, Hailiang Zhao and Jiansen Zhao; data curation, Hailiang Zhao; experiments, Hailiang Zhao and Yongyi Kong; funding acquisition, Hailiang Zhao and Jiansen Zhao; investigation, Chonghao Zhang and Haoji Zhang. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Key Research and Development Program, China (Grant no. 2021YFC2801004), National Natural Science Foundation of China (Grant no. 51709167), National Natural Science Foundation of China (Grant No. 52331012).

Data Availability Statement: The data presented in this study are available in the KITTI dataset at https://www.cvlibs.net/datasets/kitti/, reference number [44].

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. 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.
2. Godard, C.; MacAodha, O.; Firman, M.; Brostow, G.J. Digging into self-supervised monocular depth estimation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3828–3838.
3. Zhao, H.; Zhang, Q.; Zhao, S.; Chen, Z.; Zhang, J.; Tao, D. Simdistill: Simulated multi-modal distillation for bev 3d object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 18–22 November 2024.
4. Shu, Y.; Hu, A.; Zheng, Y.; Gan, L.; Xiao, G.; Zhou, C.; Song, L. Evaluation of ship emission intensity and the inaccuracy of exhaust emission estimation model. *Ocean. Eng.* **2023**, *287*, 115723. [CrossRef]
5. 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, 18–22 June 2018; pp. 5667–5675.
6. Zhan, H.; Garg, R.; Weerasekera, C.S.; Li, K.; Agarwal, H.; Reid, I. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
7. Yang, Z.; Wang, P.; Xu, W.; Zhao, L.; Nevatia, R. Unsupervised learning of geometry with edge-aware depth-normal consistency. *arXiv* **2017**, arXiv:1711.03665.
8. Yin, Z.; Shi, J. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1983–1992.
9. 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 Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12240–12249.
10. Kim, H.R.; Angelaki, D.E.; DeAngelis, G.C. The neural basis of depth perception from motion parallax. *Philos. Trans. R. Soc. B Biol. Sci.* **2016**, *371*, 20150256. [CrossRef] [PubMed]
11. Colby, C. Perception of Extrapersonal Space: Psychological and Neural Aspects. *Int. Encycl. Soc. Behav. Sci.* **2001**, 11205–11209. [CrossRef]
12. Gieseking, J.J.; Mangold, W.; Katz, C.; Low, S.; Saegert, S. *The People, Place, and Space Reader*; Routledge: London, UK, 2014.
13. Gregory, R.L. Knowledge in perception and illusion. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* **1997**, *352*, 1121–1127. [CrossRef] [PubMed]

14. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
15. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Glasgow, UK, 2015.
16. 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.
17. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
18. 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, Munich, Germany, 8–14 September 2018; pp. 36–53.
19. Zhao, H.; Zhang, J.; Zhang, S.; Tao, D. Jperceiver: Joint perception network for depth, pose and layout estimation in driving scenes. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022.
20. Li, R.; Wang, S.; Long, Z.; Gu, D. Undeepvo: Monocular visual odometry through unsupervised deep learning. In Proceedings of the IEEE International Conference on Robotics and Automation, Brisbane, Australia, 21–25 May 2018; pp. 7286–7291.
21. CS Kumar, A.; Bhandarkar, S.M.; Prasad, M. Depthnet: A recurrent neural network architecture for monocular depth prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 283–291.
22. Wang, R.; Pizer, S.M.; Frahm, J.-M. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5555–5564.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
24. Zhao, H.; Zhang, J.; Chen, Z.; Yuan, B.; Tao, D. On Robust Cross-view Consistency in Self-supervised Monocular Depth Estimation. *Mach. Intell. Res.* **2024**, *21*, 495–513. [CrossRef]
25. Agarwal, A.; Arora, C. Attention attention everywhere: Monocular depth prediction with skip attention. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023.
26. Zhao, C.; Zhang, Y.; Poggi, M.; Tosi, F.; Guo, X.; Zhu, Z.; Huang, G.; Tang, Y.; Mattoccia, S. Monovit: Self-supervised monocular depth estimation with a vision transformer. In Proceedings of the International Conference on 3D Vision, Prague, Czechia, 12–15 September 2022.
27. Bae, J.; Moon, S.; Im, S. Deep digging into the generalization of self-supervised monocular depth estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023.
28. Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; Brox, T. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
29. 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.
30. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
31. Clevert, D.-A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units(elus). *arXiv* **2015**, arXiv:1511.07289.
32. Woo, S.; Park, J.; Lee, J.-Y.; So Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
33. Zagoruyko, S.; Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv* **2016**, arXiv:1612.03928.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
35. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. In Proceedings of the NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2366–2374.
36. Garg, R.; BG, V.K.; Carneiro, G.; Reid, I. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 740–756.
37. Godard, C.; MacAodha, 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.
38. Pillai, S.; Ambruş, R.; Gaidon, A. Superdepth: Self-supervised, super-resolved monocular depth estimation. In Proceedings of the IEEE International Conference on Robotics and Automation, Montreal, QC, Canada, 20–24 May 2019; pp. 9250–9256.
39. 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, 18–22 June 2018; pp. 2022–2030.

40. Bian, J.-W.; Li, Z.; Wang, N.; Zhan, H.; Shen, C.; Cheng, M.-M.; Reid, I. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Glasgow, UK, 2019.
41. Zhou, J.; Wang, Y.; Qin, K.; Zeng, W. Unsupervised high-resolution depth learning from videos with dual networks. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
42. Spencer, J.; Bowden, R.; Hadfield, S. DeFeat-Net: General Monocular Depth via Simultaneous Unsupervised Representation Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 14402–14413.
43. Zhao, W.; Liu, S.; Shu, Y.; Liu, Y.-J. Towards better generalization: Joint depth-pose learning without posenet. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
44. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012.
45. 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]
46. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understandin. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
47. 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]
48. Uhrig, J.; Schneider, N.; Schneider, L.; Franke, U.; Brox, T.; Geiger, A. Sparsity invariant cnns. In Proceedings of the 2017 International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017.
49. Poggi, M.; Tosi, F.; Mattoccia, S. Learning monocular depth estimation with unsupervised trinocular assumptions. In Proceedings of the International Conference on 3D Vision, Verona, Italy, 5–8 September 2018.
50. Zhao, J.; Yan, Z.; Zhou, Z.; Chen, X.; Wu, B.; Wang, S. A ship trajectory prediction method based on GAT and LSTM. *Ocean. Eng.* **2023**, *289*, 116159. [CrossRef]
51. Grupp, M. evo: Python Package for the Evaluation of Odometry and SLAM. 2017. Available online: https://github.com/MichaelGrupp/evo (accessed on 5 May 2023).