remote sensing

MDPI

Article

Self-Supervised Depth Completion Based on Multi-Modal Spatio-Temporal Consistency

Quan Zhang [1,†], Xiaoyu Chen [1,*,†], Xingguo Wang [1], Jing Han [1], Yi Zhang [1] and Jiang Yue [2]

[1] School of Electronic Engineering and Optoelectronic Technology, Nanjing University of Science and Technology, Nanjing 210000, China
[2] College of Science, Hohai University, Nanjing 210000, China
* Correspondence: 115104000466@njust.edu.cn
† These authors contributed equally to this work.

Abstract: Due to the low cost and easy deployment, self-supervised depth completion has been widely studied in recent years. In this work, a self-supervised depth completion method is designed based on multi-modal spatio-temporal consistency (MSC). The self-supervised depth completion nowadays faces other problems: moving objects, occluded/dark light/low texture parts, long-distance completion, and cross-modal fusion. In the face of these problems, the most critical novelty of this work lies in that the self-supervised mechanism is designed to train the depth completion network by MSC constraint. It not only makes better use of depth-temporal data, but also plays the advantage of photometric-temporal constraint. With the self-supervised mechanism of MSC constraint, the overall system outperforms many other self-supervised networks, even exceeding partially supervised networks.

Keywords: depth completion; lidar data processing; self-supervised; sensor fusion; multi-modal; deep learning

Citation: Zhang, Q.; Chen, X.; Wang, X.; Han, J.; Zhang, Y.; Yue, J. Self-Supervised Depth Completion Based on Multi-Modal Spatio-Temporal Consistency. *Remote Sens.* **2023**, *15*, 135. https://doi.org/10.3390/rs15010135

Academic Editor: Xuan Zhu

Received: 21 November 2022
Revised: 17 December 2022
Accepted: 20 December 2022
Published: 26 December 2022

1. Introduction

In an intelligent traffic system, it has become a fundamental task to obtain the position of objects around. Especially in autonomous driving, the ability to perceive the environment is the basis for the safe and stable operation of autonomous driving [1,2]. The sparse depth provided by LiDAR can support a object segmentation [3] or a simultaneous localization and mapping (SLAM) system [4], but has a poor performance on the scene topology [5]. Moreover, increasing the density of 3D LiDAR measurements means it is cost-prohibitive. Depth completion has always been a research hotspot, especially on low-cost devices. However, the difficulty of obtaining dense annotations makes this technique difficult to implement [6]. Therefore, the self-supervised depth completion have been widely studied in recent years [7].

At present, many supervised depth completions have been proposed. Uhrig et al. [8] proved that convolution still worked on sparse signals, and Ma et al. [6] proved that image information could help reconstruct dense depth images with higher accuracy. Most of depth completion methods have been proposed to generates dense depth map by LiDAR sparse points cloud and RGB images. Due to the sparsity of input depth, convolutional neural networks have difficulty adapting to spatial pixel information [8]. Uhrig et al. [8] proposed a sparse invariant convolution, which enhanced the adaptability of convolution to sparse signals through adaptive sparse weights. This method uses the sparse mask to avoid the undifferentiated calculation of nonexistent points, while it cannot process occluded points or fuse RGB image information.

However, most of the current depth completion networks use RGB images as guidance, because RGB images can provide edge information of objects [9]. The depth map calibrated by internal parameters can be aligned with the RGB images at pixel level, which provides

the theoretical basis for the cross-modal depth completion [7]. Eldesokey et al. [10,11] designed a normalized convolutional layer to introduce RGB image information while completing the depth. Yan and Huang et al. [12,13] proposed a sparse invariant convolution, which can perform feature fusion at each layer.

In terms of fusion methods, it can be divided into signal layer fusion and feature layer fusion [7]. Ma et al. [14] proposed a modular coding network based on ResNet [15] to predict the dense depth map of RGB-D images, and they began to fuse from the signal layer. Jaritz et al. [9,16] chose the strategy of fusion at the feature layer. Another method fuses RGB-D images into multiple-scale feature layers [17–19]. Many introduce physical priors based on signal fusion, such as introducing normal surface [20–22] and semantic segmentation [9,23,24]. Moreover, some methods guide the propagation of sparse depth, according to the correlation of successive pixels on the RGB map [25,26].

However, all of these full-supervised depth completion methods still faces an unavoidable problem on practical application: how to obtain ground truth labels. To face with the lack of ground truth labels, many studies on self-supervised depth completion have been proposed. The self-supervised network is modelled with the intrinsic consistency within the images and the aligned point clouds. Different from the supervised networks which are modelled relying on a large amount of manual labelled ground truths, the proposed self-supervised network automatically construct a relationship between the multi-modal data and the dense depth, which makes it more robust and have better generalization ability.

Self-supervised depth completion method can be divided into stereo vision and monocular vision [7]. Both take the reprojection error as the major constraint. The stereo method [27,28] converts the depth to parallax and calculates the reprojection error with the help of a pre-calibrated baseline. Compared with the latter, the stereo method does not need to consider the influence of external parameters, but it cannot solve the problems caused by occlusion and cost another camera. The monocular method projected adjacent frames to the current frame through pose and depth. Ma et al. [6] first designed a self-supervising framework, with photometric loss as depth supervision. This method requires pose estimation to provide external parameters. To improve the accuracy of the pose estimation, Feng and Choi et al. [29–31] introduced model-based pose estimation module. Among them, Song and Wong et al. [5,32] introduced an odometer to improve their pose network. In a word, existing self-supervised depth completion methods commonly use RGB image reprojection to establish strong constraints for spatial connections. However, different from the dense depth ground truth, the reprojection error is evaluated by photometric value without direct depth supervision. It leads to the difficulty of fusion between image data and the projected sparse depth map [6]. Therefore, Feng et al. [29] generate pseudo dense representations before concatenating the image data and the sparse depth map. Wone et al. [31] pool the sparse depth map while inputting the sparse depth points. It alleviates the problem caused by the photometric evaluation, but the problem of multi-modal information fusion remains.

At present, the self-supervised depth completion have main problems: (1) The reprojection constraint assumes that the scene is static and non-occluded, which reduces the matching success rate of moving objects and occluded regions [6]. Meanwhile, the reprojection error cannot reflect the depth loss between the predicted value and the ground truth, especially in the dark, low-texture parts and distant objects. (2) Depth-temporal information contains a lot of usable information, which should not be discarded in self-supervision. (3) It is not conducive to the fusion of image data and sparse depth map, when the photometric evaluation is used as the depth constraint.

To cope with these challenges, a multi-modal spatio-temporal consistency approach is proposed to help improve the performance of model-based depth completion. We introduce depth-temporal data to reduce the impact of photometric constraints on depth completion, such as dark, low-texture parts and distant objects. Meanwhile, we proposed a depth-temporal consistency constraint to directly supervise depth, so in the whole process of depth completion, RGB image data can significantly improve the effect of depth completion, through effective fusion. However, the sparse depth map are sampled from dynamic scenes,

and the temporal depth after simply stacking cannot be directly used to constrain the depth. Therefore, we designed a photometric reprojection auto-mask to remove the occlusion and displace points. Meanwhile, we introduced the photometric-temporal consistency to provide a global constraint for our depth completion. In addition, multi-modal spatio-temporal consistency is also introduced into the pose estimation to improve the accuracy.

2. Methods

In this section, our self-supervised depth completion framework is shown in Figure 1, which takes a single visible image RGB_t and a LiDAR image D_t as input and generates a dense depth map, *Pred*.

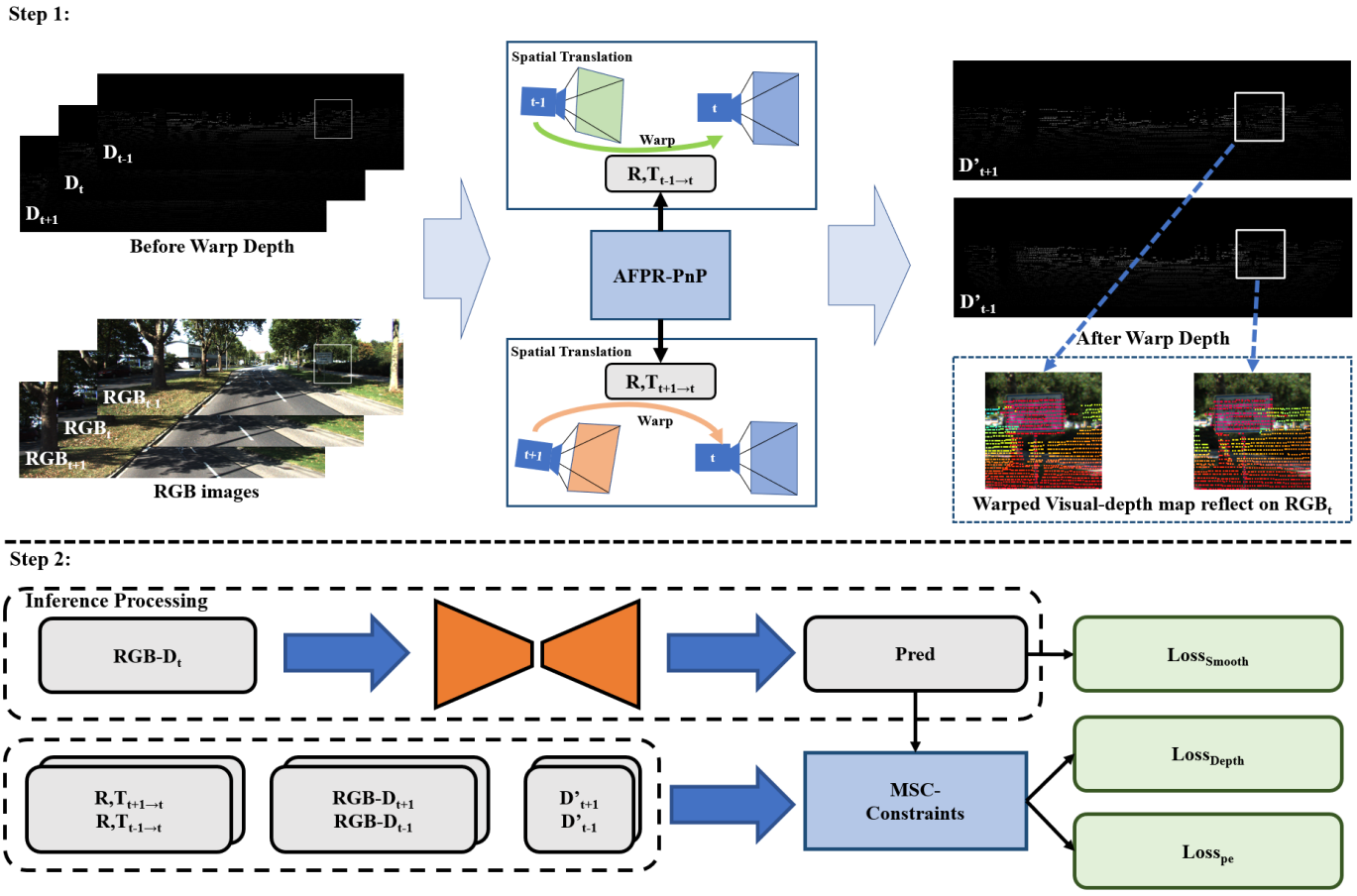

Figure 1. The framework of the proposed self−supervised depth completion network, step 1: spatial translation for preprocessing; step 2, self−supervised training. Gray rectangles are variables, orange is the inference network, blue is computational modules (no parameters to learn), and green is the loss functions.

This methods is divided into two steps: Step 1, spatial translate the depth map of adjacent frames (D_{t-1} and D_{t+1}) into the current camera field, to generate the D'_{t-1} and D'_{t+1}. The pose parameter is provided from AFPR-PnP module. After translation, warped depth points can be reflected on RGB_t. Step 2, self-supervised training procedure based on multi-modal spatio-temporal consistency (MSC) constraint: It requires a sequence of RGB-D images for training. In the inference processing, it needs only a pair of RGB-D images for generating completion depth.

The pose $R, T_{t+i\to t}$ from the current frame $RGB - D_{t+i}$ to the adjacent frame $RGB - D_t$ is calculated by the PnP algorithm, where $i \in \{-1, +1\}$. The depth map of adjacent frames (D_{t-1} and D_{t+1}) can be translated to the current camera field (D'_{t-1} and D'_{t+1}) by $R, T_{t-1\to t}$ and $R, T_{t+1\to t}$. This is so that a multi-modal spatio-temporal consistency constraint can be built for the self-supervising framework. At the same time, a photometric reproject

auto-mask is designed based on the similarity estimation. This automatic mask reduces the displacement errors caused by the depth-temporal point cloud.

The last, an automatic feature points refinement algorithm, is adopted to improve the performance of PnP pose estimation.

This self-supervised framework does not rely on additional sensors, manual tagging efforts, or other learning-based pose estimation algorithms as building blocks. In the inference, only the current frame is needed as an input to generate depth completion.

2.1. Depth-Temporal Consistency Constraint

This paper proposed a self-supervising constraint method based on depth-temporal consistency. The method only needs to obtain synchronous $RGB - D$ image sequences from monocular cameras and LiDAR. The point clouds of the adjacent frames can be transfered to the current frame to constrain depth completion, as shown in Figure 1-step 1.

As is shown in Figure 2, multi-modal spatio-temporal consistency constraint contains two parts, depth-temporal consistency constraint ($Loss_{Depth}$) and photometric similarity evaluation (PSE). The photometric reproject auto-mask generated by PSE was used to assist the depth-temporal consistency constraint on the current frame.

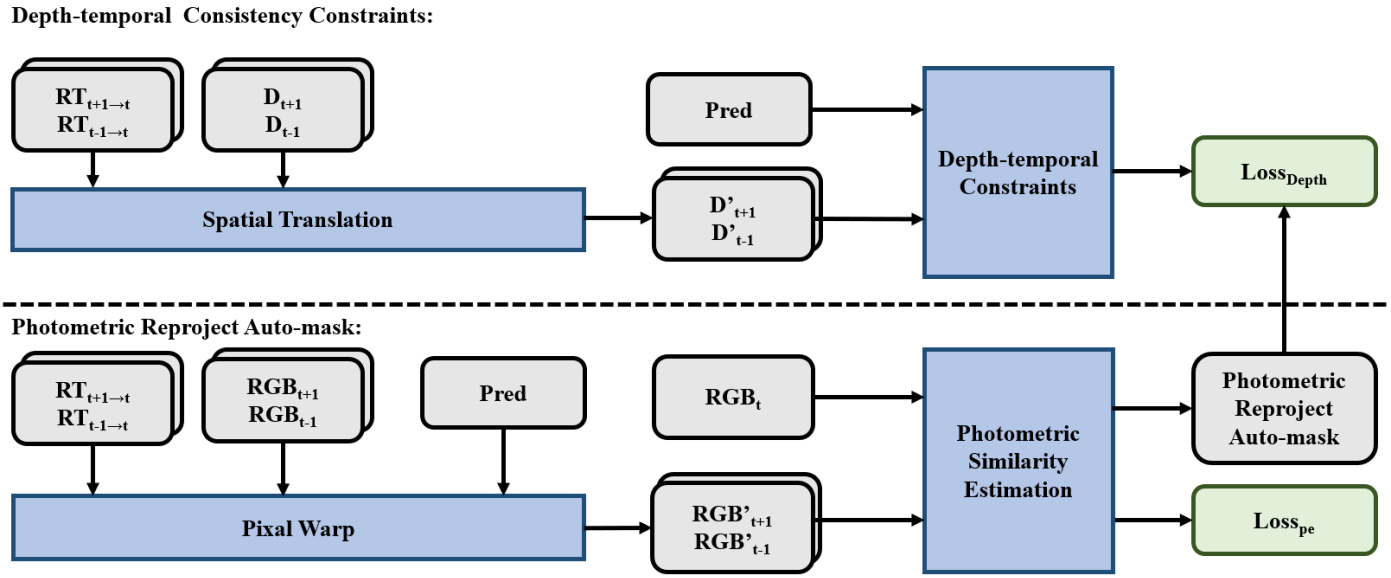

Figure 2. Multi−modal spatio−temporal consistency constraint.

2.1.1. Spatial Translation

We take $RGB - D_t$ at time t, and the adjacent frame as $RGB - D_{t+i}$, where $i \in \{-1, 1\}$. Pose parameters are estimated with $RGB - D_t$ and $RGB - D_{t+i}$, containing two groups of the rotation matrix and translation $(R, T_{t+i \to t})$. The detail of pose estimation is introduced in Section 2.3.

The external parameter matrix can be expressed as:

$$T_{t+i \to t} = \begin{bmatrix} R_{t+i} & T_{t+i} \\ 0^3 & 1 \end{bmatrix}, \; T_{t \to t+i} = \begin{bmatrix} R_{t+i}^{-1} & -T_{t+i} \\ 0^3 & 1 \end{bmatrix}. \tag{1}$$

The information of the LiDAR points at time $t + i$ can be transferred to the camera coordinate system at time t through external parameters, and the process can be expressed as:

$$\begin{bmatrix} x'_{t+i} & y'_{t+i} & z'_{t+i} & 1 \end{bmatrix}^T = T_{t+i \to t} \cdot \begin{bmatrix} x_{t+i} & y_{t+1} & z_{t+1} & 1 \end{bmatrix}^T, \tag{2}$$

where $D'_{t+i} = z'_{t+i}$. The translated depth map can be calculated by internal parameters:

$$D'_{t+i}\begin{bmatrix} u & v & 1 \end{bmatrix}^T = K \cdot \begin{bmatrix} x'_{t+i} & y'_{t+i} & z'_{t+i} & 1 \end{bmatrix}^T. \tag{3}$$

We can obtain the sparse depth map D'_{t+i} under the camera coordinate system at the time of t.

2.1.2. Depth-Temporal Consistency Module

Sequential frames contain a lot of spatial structure information. For example, adjacent RGB images have most of the same scenes as the current frame, as do depth images. Different from the point cloud of the current frame (D_t), it provides a lot of depth information in the blank block. However, it cannot be used directly until it is transferred to the current camera field.

The depth-temporal consistency loss function ($Loss_{Depth}$) is built with D'_{t+i} and $Pred$. To ground the predictions to a metric scale, we minimize the L^2 difference between predictions $pred$ and the sparse warped depth map D'_{t-1} and D'_{t+1} over its domain (Ω_{t-1} and Ω_{t+1}):

$$\begin{aligned} Loss_{Depth} = & \frac{1}{|\Omega_{t-1}|} \sum_{x \in \Omega_{t-1}} ||D'_{t-1}(x) - pred(x)||^2 \\ & + \frac{1}{|\Omega_{t+1}|} \sum_{x \in \Omega_{t+1}} ||D'_{t+1}(x) - pred(x)||^2. \end{aligned} \tag{4}$$

2.2. Photometric-Temporal Consistency Constraint

Because of the object movement and occlusion, D'_{t+i} contains many displaced depth points affecting the accuracy of the constraint. Inspired by Godard et al. [33], we select the most similar part of the pixel structure to retain the depth points of this part. The similarity of warped images (RGB'_{t+i} and RGB_t) can be used to select the minimum value points in $E_{ph}|_{t+i}$ as the automatic photometric reprojection mask ($mask_{t+i}$).

We multiply $mask_{t+i}$ times D'_{t+i}, and combine their results as a new depth map D'_t.

2.2.1. Pixal Warp

Similarly with Section 2.1.1, RGB images RGB_{t+i} will be mapped to the camera coordinate system at t through the pose and predicted depth, as shown in Figure 3. Take the pixel point of the predicted depth image $Pred$ at (u, v), then the pixel point coordinates after the predicted depth can be expressed as:

$$\begin{bmatrix} u' & v' & 1 \end{bmatrix}^T = Pred(u, v) K^{-1} \begin{bmatrix} u & v & 1 \end{bmatrix}^T. \tag{5}$$

The sampling function of the warped RGB image RGB'_{t+i} at (u, v) can be expressed as:

$$RGB'_{t+i}(u, v) = (RGB_{t+i}(KT_{t \to t+i}) \begin{bmatrix} u' & v' & 1 \end{bmatrix})^T. \tag{6}$$

Finally, we can obtain the RGB image RGB'_{t+i} under the camera coordinate system at the time of t.

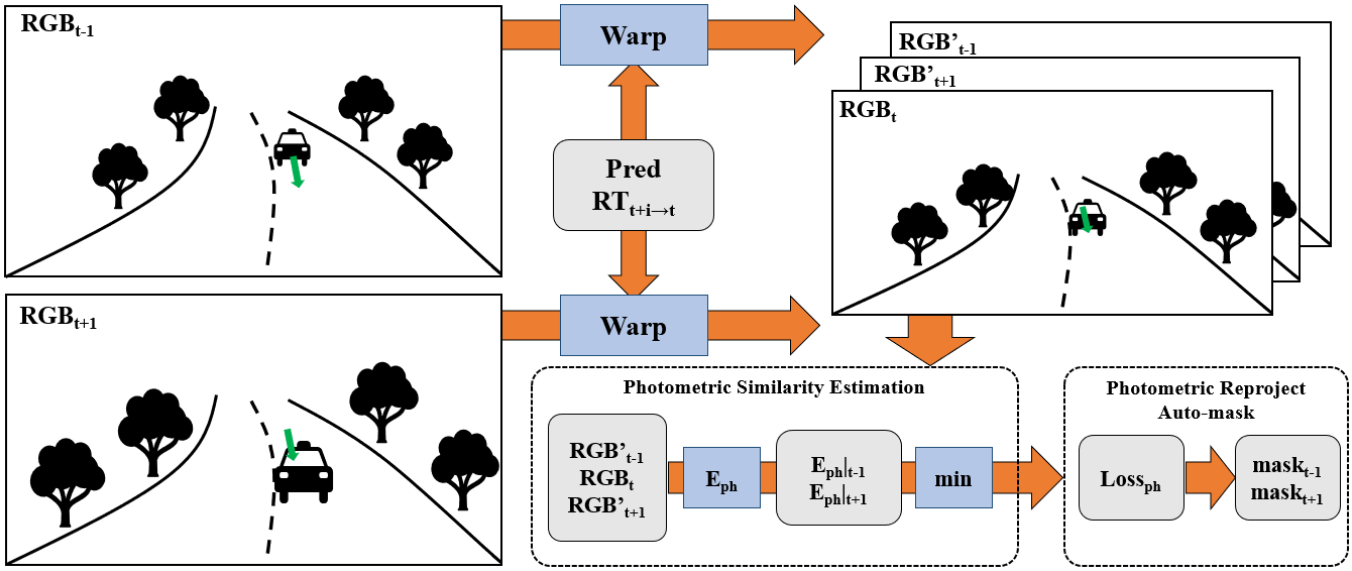

Figure 3. Warp RGB_{t+i} to RGB'_{t+i} with pose $(R, T_{t+i\to t})$ and completion depth (*Pred*), then generate the automatic photometric reprojection mask.

2.2.2. Photometric Reproject Auto-Mask

We map the adjacent RGB images to the current RGB image for similarity evaluation by referring to the current predicted depth. The evaluation function E_{ph} contains L1 loss and pixel structure loss.

The photometric error of RGB_t and RGB'_{t+i} can be represented as:

$$E_{ph}|_{t+i} = \frac{\omega}{2}(1 - SSIM(RGB'_{t+i}, RGB_t)) + (1-\omega)\|RGB'_{t+i}, RGB_t\|^1, \tag{7}$$

where ω is the weight between 0 and 1 , and the t is time.

We select the adjacent depth point based on the minimum photometric error. Specifically, we took the minimum value of E_{ph} as photometric loss $Loss_{ph}$. This photometric loss can overcome the effects of partial occlusion and displacement [33].

It can be expressed as:

$$Loss_{ph} = Min\{E_{ph}|_{t-1}, E_{ph}|_{t+1}\}, \tag{8}$$

and the photometric loss function is:

$$Loss_{pe} = \frac{1}{|\Omega|}\sum_{(x,y)\in\Omega} Loss_{ph}(x,y). \tag{9}$$

Among them, $Loss_{ph}$ generated by selecting the minimum value from $E_{ph}(RGB'_{t+1}, RGB_t)$ and $E_{ph}(RGB'_{t-1}, RGB_t)$, which means the selected region is the unobstructed part.

The photometric reproject auto-mask according to this characteristics, which represents the most similar part of the warped adjacent RGB images, as shown in Figure 3.

The auto-mask can be represented as:

$$mask_{t+i}(x,y) = \begin{cases} 1 & \text{if } Loss_{ph}(x,y) = E_{ph}|_{t+i}(x,y) \\ 0 & otherwise. \end{cases} \tag{10}$$

The auto-mask also represents the most similar part of the depth when it overlaps with the depth image. For the occluded part in a adjacent depth, the auto-mask will have a

higher response in another adjacent depth image, to complementing the occluded depth. In the selection of adjacent depths, the auto-mask can provide an available reference. The stitched depth still works on moving objects, unlike photometric errors.

The stitched depth D'_t is shown in Figure 4, represented as:

$$D'_t = \sum_{i}^{-1,1} (D'_{t+i} * mask_{t+i}). \tag{11}$$

The sequence depth loss function can be represented as:

$$Loss_{Depth} = \frac{1}{|\Omega|} \sum_{x \in \Omega} ||D'_t(x) - pred(x)||^2, \tag{12}$$

where Ω is the domain of depth map D'_t.

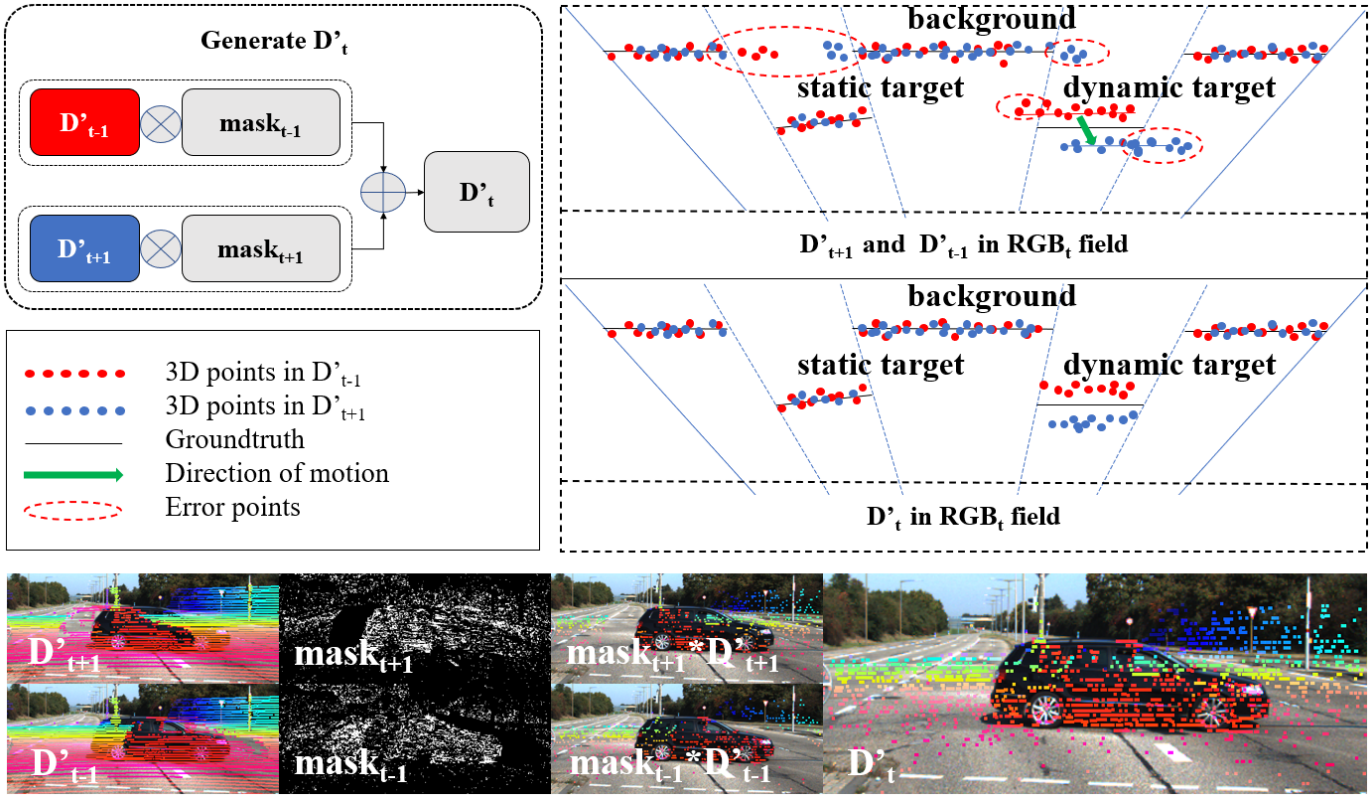

Figure 4. D'_t is generated by splicing high similarity points in D'_{t+i} (visual process in the bottom). D'_t participates in the loss function $Loss_{Depth}$.

Our multi-modal spatio-temporal consistency constraints include depth-temporal consistency constraint (DC) and photometric-temporal consistency constraint (PC). We implement both constraints with $Loss_{Depth}$ and $Loss_{pe}$.

2.3. Automatic Feature Points Refinement

In this method, the stitched depth comes from the translated adjacent depth. The pose estimation network gives constantly adjusted poses, which will bring a heavy calculation to the preprocessing. Therefore, we made the preprocessing dataset with the PnP algorithm, which including the sparse depth after obtaining the fixed pose and translated depth.

Since there are many moving objects in the traffic scene, the feature points from fast-moving objects will affect the estimation in previous PnP algorithms. The automatic feature points refinement proposed as shown in Figure 5, on the left side, the blue point is the

matched pair of feature points, the gray point is the camera coordinate system position of feature points (each pair of feature points maps to the same gray point), and the red point is the fast displacement point in the two imaging. On the right side, we line the matched feature points, the green line is the point pair successfully selected, and the red line is the point pair filtered by AFPR-PnP algorithm (the left column is the PnP algorithm designed by Ma, and the right column is ours).

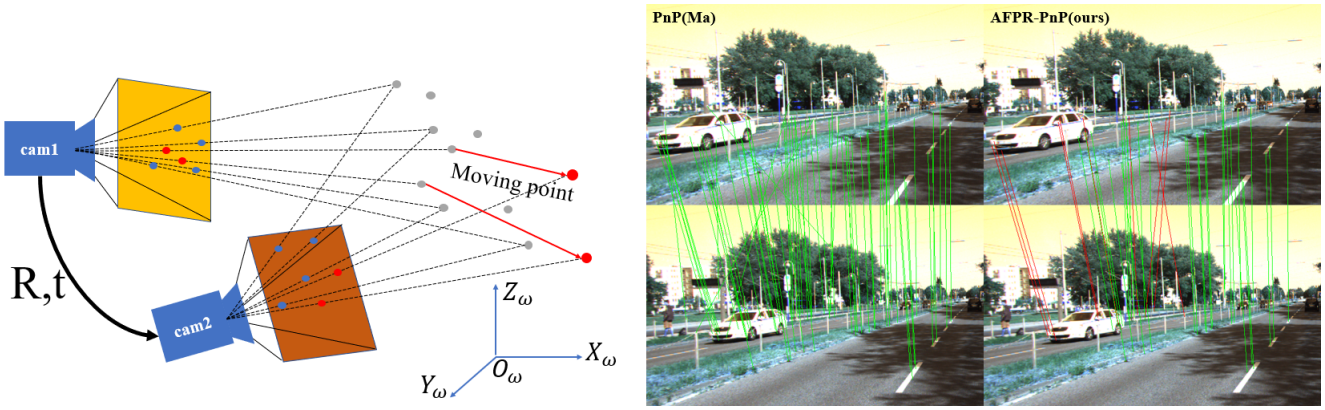

Figure 5. Automatic feature points' refinement (AFPR).

Our method adopt the evaluation of moving distance in 3d space to remove interference points. For mismatched points, some of these pairs can be filtered in 2D space, while others are indistinguishable in 2D space, as shown in Figure 5. And such problems can be improved by converting to 3D space coordinate. Assuming that the coordinate of point p_t on the visible image is (u, v), and the depth at here is d, the 3D space position can be expressed as ps_t:

$$ps_t = K^{-1} \cdot d \cdot \begin{bmatrix} u & v & 1 \end{bmatrix}^T. \tag{13}$$

Transform the feature points into 3D space position coordinates, and calculate their pixel distance D_{pixel} and space distance D_{space}.

$$D_{pixel} = ||p_t - p_{t+i}||^2, \tag{14}$$

$$D_{space} = ||ps_t - ps_{t+i}||^2. \tag{15}$$

We removed the point pairs whose pixel distance D_{pixel} exceeded th_{d2} and space coordinate distance D_{space} exceeded th_{d3}.

3. Experimental Evaluation

3.1. Datasets and Setup

This chapter briefly describes the preprocessing of datasets and design experiments to verify the validity of model.

Datasets: We used KITTI https://www.cvlibs.net/datasets/kitti/raw_data.php (accessed on 21 November 2022) as our data set. In the data pre-processing, we deleted the data of the static scene and the camera on the right. It reduces duplicate scenes and detection points. There are 85,342 training sets. For the test, we selected 1000 test set images for comparison.

Before training, we preprocess the data set. We calculate the Pose parameter through the PnP algorithm, translate the point cloud of the adjacent frame to the current camera coordinate system by parameter, and then splice the translated point cloud according to Formula (11) to generate D'_t. Since the PnP algorithm we use is unchanged for each frame, the input results can be reused after saving, thus reducing a lot of training pressure.

Our preprocessing is given in the following Algorithm 1 flow chart:

Algorithm 1 Data set preprocessing

INPUT: RGB images and Lidar data set I_{train}
For *each episode*$(RGB_t, D_t, RGB_{t+i}, D_{t+i}) \in I_{train}$,where $i \in \{-1, 1\}$ do:
1. Calculate the $Rt_{t+i \to t}$ from $RGB - D_{t+i}$ to $RGB - D_t$, through the AFPR-PnP.
2. We transferred the D_{t+i} to the same coordinate system as D_t through $Rt_{t+i \to t}$ and internal parameters, then obtained D'_{t+i} from Equation (3).
3. The generated D'_{t+i} and Rt_{t+i} are saved as the pre-preprocessing set $D_{train}|_i$.

Setup: We used PyTorch to train the model and cropped the input image to 352*1216. In the pre-processing stage, color correction and noise superposition are performed on the visible light data. We have a learning rate of 10^{-5}. Set the batch size to 4, learn 30 epochs, and train the model with an NVIDIA TITAN RTX. We tested the validity of proposed module in three test directions. Our depth completion network adopts the network structure of $sparse - to - dense$, and the ResNet18 network experiment verifies the superiority of $Loss_{Depth}$ and PnP algorithms. ResNet34 is used to verify the validity of the deep information mining module. Finally, in comparison with other methods, we also use resnet34.

The final loss function:

$$Loss = \omega_0 Loss_{Depth} + \omega_1 Loss_{Smooth} + \omega_2 Loss_{pe} \tag{16}$$

where ω_0, ω_1 and ω_2 are respectively weights of $Loss_{Depth}$ (Function (12)), $Loss_{Smooth}$, and $Loss_{pe}$ (Function (9)).

During the training, the semi-dense depth labels provided by KITTI were not involved in the calculation. We use the KITTI2012 prediction set https://www.cvlibs.net/datasets/kitti/raw_data.php (accessed on 21 November 2022) and our output results to calculate RMSE error as a standard to evaluate the depth completion accuracy because it is the most representative indicator. RMSE can be expressed as:

$$RMSE = (\frac{1}{|\Omega|} \sum_{x \in \Omega} |\hat{z}(x) - z_{gt}(x)|^2)^{\frac{1}{2}}, \tag{17}$$

$\hat{z}$ denotes the completed depth; z_{gt} denotes the ground-truth depth.
Our training process is given in the following Algorithm 2 flow chart:

Algorithm 2 Flow chart of self-supervised sequence depth constraint training algorithm.

INPUT: RGB images and Lidar data set I_{train}, the pre-preprocessing set $D_{train}|_i$
For *each episode*$(RGB_t, D_t, RGB_{t+i}, D_{t+i}) \in I_{train}$ and $(D'_{t+i}, Rt_{t+i}) \in D_{train}|_i$ do:
1. Input RGB_t, D_t into network, predict output$Pred$.
2. By outputting depth $Pred$, the PnP algorithm estimates pose Rt_{t+i} and camera internal parameters and maps RGB_{t+i} to the camera coordinate system of RGB_t to calculate photometric loss $E_{ph}(RGB'_{t+i})$. The formula is given by (7).
3. he photometric loss $E_{ph}(RGB'_{t+i})$ is used to select and splice D_{t+i} to obtain D'_t, and the relation is given by (11).
4. Calculate the total loss Function (16).
For *each episode*$(RGB_t, D_t, Gt_t) \in I_{test}$ do:
1. Input RGB_t, D_t into network, predict output$Pred$.
2. Calculate the RMSE errors of A with B by the error Formula (17).

3.2. Results of The Ablation Experiments

Our multi-modal spatio-temporal consistency (MSC) constraint is embodied in using the consistency of sequence multi-modal information to constrain.

In this section, we compare three self-supervised methods: depth-temporal consistency (DC) constraint, photometric-temporal consistency(PC) constraint, and multi-modal spatio-temporal consistency (MSC) constraint.

As shown in Table 1, the depth completion network can converge smoothly with stitched depth D'_t participation in the constraint. As seen in Figure 4, mask comes from the part of the minimum photometric loss, which is multiplied by D'_{t+i} to filter out most of the wrong displacement points of D'_t. This further demonstrates the effectiveness of DC constraints.

Table 1. Comparison of self-supervised methods for different sequence modality constraints. The bold numbers are the best.

Method	*Network*	$Auto-Mask$	$Loss_{Depth}$	$Loss_{Smooth}$	$Loss_{pe}$	*Layers*	*RMSE*
$s2d-DC$	*s2dNet*	√	0.010	0.1		18	4885.72
$s2d-PC$	*s2dNet*			0.1	1.0	18	2299.49
$s2d18-MSC$	*s2dNet*	x	0.010	0.1	1.0	18	1379.70
$s2d18-MSC-auto$	*s2dNet*	√	0.010	0.1	1.0	18	1316.93
$s2d34-MSC$	*s2dNet*	x	0.010	0.1	1.0	34	1241.58
$s2d34-MSC-auto$	*s2dNet*	√	0.010	0.1	1.0	34	**1212.69**
$kb-MSC$	*kbNet*	x	0.010	0.1	1.0		1527.37
$kb-MSC-auto$	*kbNet*	√	0.010	0.1	1.0		1289.67
$s2d18-MSC$ 1	*s2dNet*	√	0.001	0.1	1.0	18	1503.27
$s2d18-MSC$ 2	*s2dNet*	√	0.010	0.1	0.1	18	1577.12

Since RGB and depth data values in different range, and we only regularized RGB images, and the values of $Loss_{Depth}$ and $Loss_{Smooth}$ are different from $Loss_{pe}$ in scale. We tested different values of ω_0 and ω_2. There is much space for optimization in the adjusting parameters.

MSC constraint: As can be observed from Table 1, the MSC method has obvious advantages over single-modal constraint in stable convergence: the lack of any modal constraint ($Loss_{Depth}$ or $Loss_{pe}$) will crimp the network performance.

(1) Depth-temporal consistency constraint (DC): Many works involve the input depth map in training so that the network will not lose too much depth information. However, due to the displacement between the RGB camera and LiDAR, some of the background depth and the foreground depth are mixed in the same occluded area. Retaining input depth means retaining these erroneous background depth. Our method improved this practice and the stitched sequence depth to replace the current depth.

(2) Photometric-temporal consistency constraint (PC): The sequence depth constraint is too sparse to provide the global constraint. Thus, we introduce the sequence photometric constraint. As shown in the Table 1, the sequence photometric constraint is also indispensable, and the superiority of the sequence multi-modal constraint is also proved. In the DC constraint, this temporal depth data cannot constrain depth completion directly; it contains displaces depth points. This points can affect the performance of depth completion, which was also shown in subsequent experiments. Therefore, we introduce photometric reproject auto-mask to remove these error points, and the experiment proves that this auto-mask is useful.

(3) Multi-modal spatio-temporal consistency constraint (MSC): The MSC constraint contains DC and PC, containing three loss functions. We tested the effect of their weights on the two network structure, S2D [6] and KBNet [31]. As can be observed in the Table 1, the network has the best performance when $\omega_0 = 0.01$, $\omega_1 = 0.1$, and $\omega_2 = 1$. We tested several weight ratios of $Loss_{Depth}$ and $Loss_{pe}$ and finally took $\omega_0 = 0.01$, $\omega_1 = 0.1$, and $\omega_2 = 1$ as the value of the following experiment.

In addition, we change the input for the experiment, which further verifies that the network can not only mine spatial information but also learn the depth completion from RGB information. As shown in the Table 2, input RGB or Gray images can assist depth completion to generate more accurate depth than input depth only.

Table 2. Experiment and result comparison of spatial information matching mask. The bold numbers are the best.

Method	Input	$Loss_{Depth}$	$Loss_{Smooth}$	$Loss_{pe}$	*Layers*	*RMSE*
1	$RGB + Depth$	0.01	0.1	1	34	**1212.69**
2	$Depth$	0.01	0.1	1	34	1264.01
3	$Gray + Depth$	0.01	0.1	1	34	1255.51

AFPR-PnP: In this experiment, we demonstrate the reliability of the PnP algorithm based on automatic feature point refinement. Since S2D [6] have the same pose estimation method (PnP algorithm) with ours. We preprocess the same dataset with the PnP algorithm proposed by Ma et al. , train the network in the same way, and compare it with our method. As shown in Figure 5, our PnP algorithm filtered error points, including points on moving cars and different objects with repeating textures.

Specifically, we set $th_{d2} = 100(pixel)$, $th_{d2} = 3(meter)$.

As shown in Table 3, we performed ablation experiments with S2D [6] and KBNet [31], the performance for depth completion is enhanced with these methods.

Table 3. Experiment and result comparison of AFPR-PnP algorithms. The bold numbers are the best.

Method	Network	*Depth Constraint*	$Auto - Mask$	*Pose*	*RMSE*
$s2d$	s2dNet	D_t	x	PnP(Ma)	1476.76
$s2d + DC$	s2dNet	$D'_{t-1} + D'_{t+1}$	x	PnP(Ma)	1379.70
$s2d + DC + PC$	s2dNet	D'_t	$\surd$	PnP(Ma)	1322.37
$s2d + MSC + AFPR$	s2dNet	D'_t	$\surd$	PnP(AFPR)	**1316.93**
$kbnet$	kbNet	D_t	x	PnP(Ma)	1495.51
$kbnet + DC$	kbNet	$D'_{t-1} + D'_{t+1}$	x	PnP(Ma)	1361.9
$kbnet + DC + PC$	kbNet	D'_t	$\surd$	PnP(Ma)	1311.12
$kbnet + MSC + AFPR$	kbNet	D'_t	$\surd$	PnP(AFPR)	**1289.67**

Discussion: In Figure 6, we compare the performance of model-based depth completion with and without MSC constraint (S2D+MSC(ours) and S2D(Ma)) in detail:

(1) Experiments show that the performance of dark, low-texture parts and distant object is significantly improved(① ② ③ ④).
(2) Sequence depth replaces the input depth as a constraint, so the network does not fully preserve the input depth with its error points. In addition, our photometric similarity estimation module filters out the occlusion points in adjacent depth. Compared with the scheme without MSC constraints, our method reduces the influence of displace points of sparse depth input on depth completion (⑤ ⑥).
(3) For dynamic targets (④ ⑦), our method has better adaptability.
(4) Our method enhances the fusion ability of the network for multi-modal information, so there will not be a large number of residual sparse points in the image(①②⑤⑥⑦).

3.3. Results of The Comparative Experiments

As shown in Table 4, we compared the sequence multi-modal constrain self-supervised algorithm with other algorithms on the validation dataset and test set of KITTI2012, and the lower the indicator, the better. We selected DepthComp [5], VOICED [32], SelfDeco [30] and ddp [27] network pose estimation methods to compare with Ma and other PnP pose estimation methods. At the same time, we compare different SLAM acquisition methods. It can be observed that our PnP position estimation method even exceeds the performance of the pose estimation network methods. The stereo images self-supervised method [28] avoids the effect of pose translation with the cost of one more camera, which exceeds the self-supervised scheme of the partial pose estimation network. But our method still has an advantage over it with only a single camera.

PnP has the advantages of scene generalization and no training but at the cost of reducing accuracy [5]. However, we still obtain good results with AFPR-PnP algorithm. Additionally, we classify existing constraint methods into multi-modal spatio-temporal consistency constraint (MSC) and photometric-temporal constraint (PC). It can be observed that our MSC constraint method contained DC and PC is more advantageous than only PC.

Discussion: From the comparison (Figure 7) of the output results, our depth completion is better for the interior of the object, while for the edge of the object, our depth completion has fewer wrong completions.

Table 4. The experimental results compared with other methods. The bold numbers are the best.

Method	***SLAM***	***PC/MSC***	***RMSE***
Kitti2012 Depth Completion Validation Dataset			
S2D [6]	PnP	*PC*	1342.33
DepthComp [5]	PnP	*PC*	1330.88
DepthComp	PoseNet	*PC*	1282.81
SelfDeco [30]	PoseNet	*PC*	1212.89
KBNet(withPnP)	PnP	*PC*	1289.67
our	PnP	*MSC*	**1212.69**
Kitti2012 Depth Completion Test Dataset			
S2D	PnP	*PC*	1299.85
IP-Basic [34]	PnP	*PC*	1288.46
KBNet(with PnP)	PnP	*PC*	1223.59
DFuseNet [28]	Stereo	/	1206.66
DDP [27]	PoseNet	*PC*	1263.19
DepthComp	PoseNet	*PC*	1216.26
VOICED(VGG8) [32]	PoseNet	*PC*	1164.58
VOICED(VGG11)	PoseNet	*PC*	1169.97
ours	PnP	*MSC*	**1156.78**

By synthesizing all the completion results, it can be observed that the self-supervised completion still has the completion ability for the areas not detected by LiDAR. Compared with the supervised depth completion, it does not rely on manual labelled ground truths and has a better application prospect.

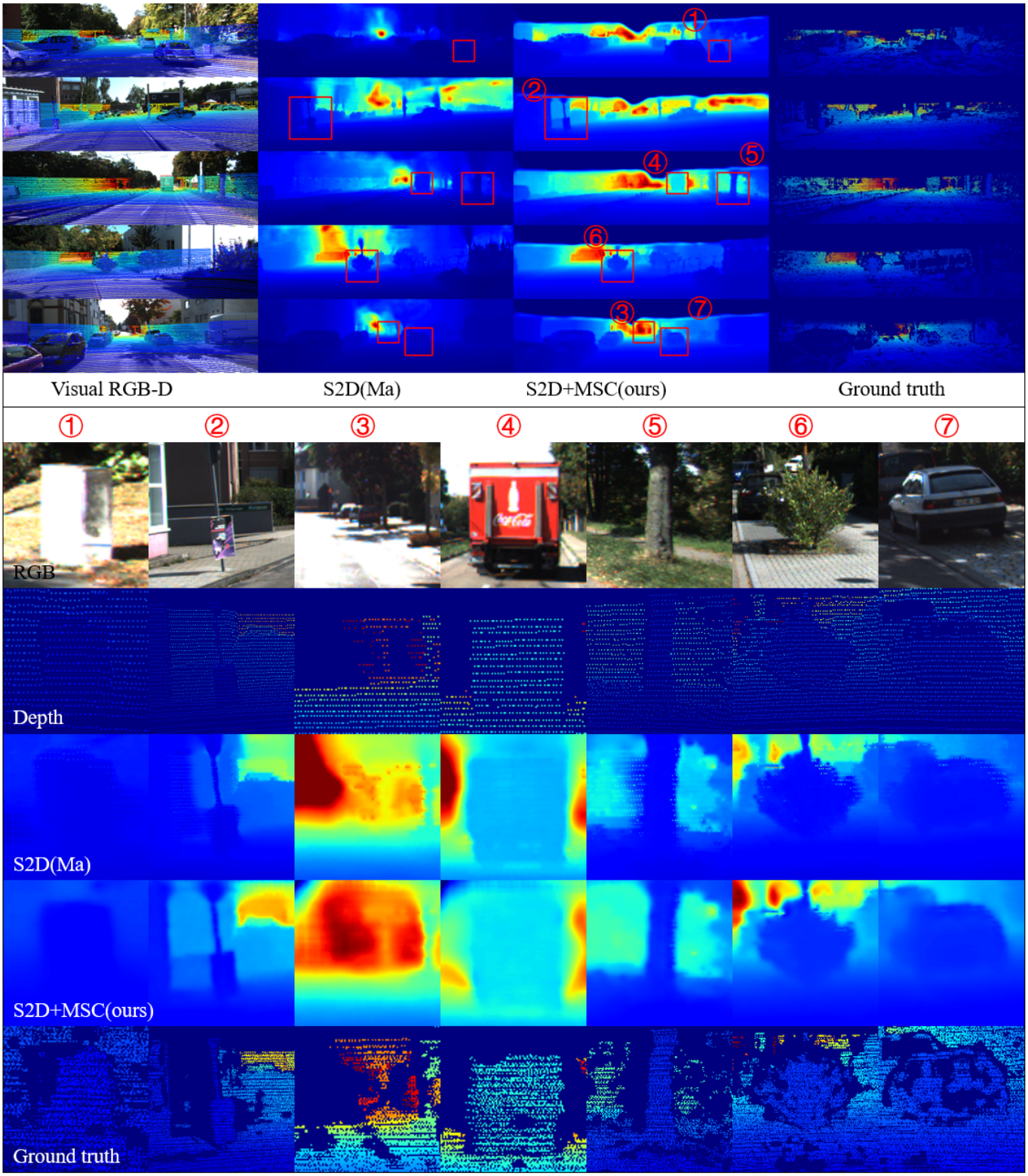

Figure 6. Results of our self-supervised depth completion on the prediction dataset.

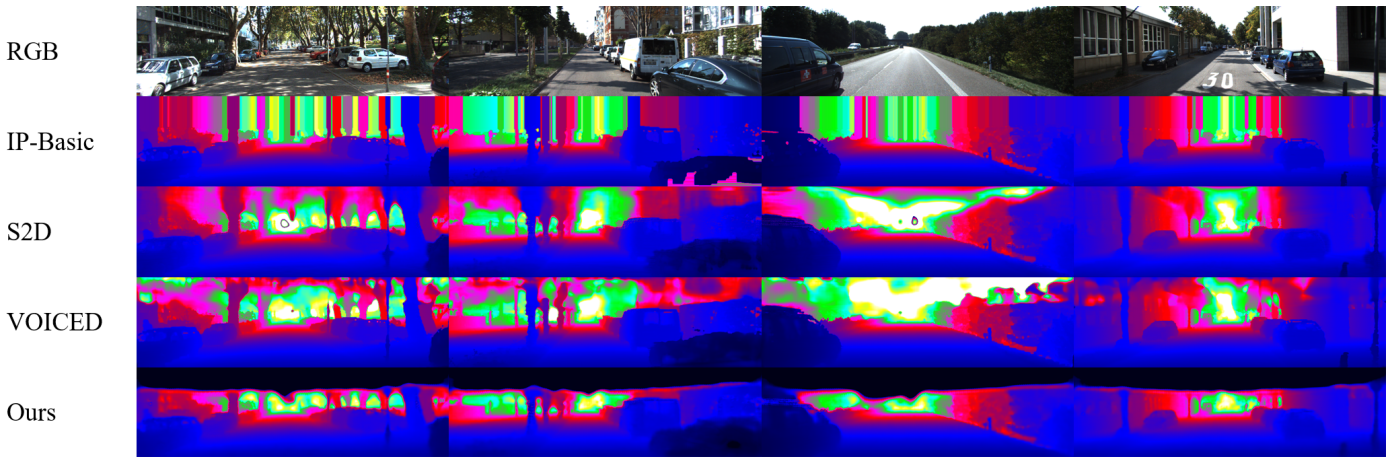

Figure 7. The results of our self-supervised depth completion compared with others.

4. Conclusions

To sum up, this paper completed the design of self-supervised depth completion without the ground truth. Among the self-supervised depth completion algorithms, our method is the only one that exploits the MSC constraint. The unique advantage of this method is that it can improve the performance of depth completion on moving objects and occluded/dark light/low texture parts, making the use of the multi-modal spatio-temporal information to the greatest extent. Our experiment demonstrates the effectiveness of our method, even surpassing many supervised methods. However, our method still has much room for improvement. In future work, we will put forward more improvement schemes for the point cloud denoising of sequence LiDAR. Besides, a pose estimation for real-time spatial translation needs to be proposed, and we will focus on improving its estimation accuracy. On this basis, we can still observe the great potential for this kind of self-supervision framework.

Author Contributions: Conceptualization, Q.Z.; Methodology, Q.Z.; Software, Q.Z.; Validation, Q.Z.; Investigation, Q.Z.; Resources, X.C., J.H. and Y.Z.; Writing—original draft, Q.Z.; Writing—review & editing, Q.Z., X.W. and J.Y.; Supervision, X.C., J.H. and Y.Z.; Project administration, X.C.; Funding acquisition, J.H. All authors have read and agreed to the pub-lished version of the manuscript.

Funding: This work was supported by National Natural Science Foundation of China (62101256) and China Postdoctoral Science Foundation (2021M691591).

Data Availability Statement: Publicly available datasets were analyzed in this study. This data canbe found here: https://www.cvlibs.net/datasets/kitti/eval_depth.php, accessed on 21 November 2022.

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

References

1. Newcombe, R.A.; Izadi, S.; Hilliges, O.; Molyneaux, D.; Kim, D.; Davison, A.J.; Kohi, P.; Shotton, J.; Hodges, S.; Fitzgibbon, A. Kinectfusion: Real-time dense surface mapping and tracking. In Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality, Basel, Switzerland, 26–29 October 2011; pp. 127–136.
2. Zhang, J.; Singh, S. LOAM: Lidar odometry and mapping in real-time. In Proceedings of the Robotics: Science and Systems, Berkeley, CA, USA, 12–16 July 2014; Volume 2, pp. 1–9.
3. Li, Y.; Le Bihan, C.; Pourtau, T.; Ristorcelli, T.; Ibanez-Guzman, J. Coarse-to-fine segmentation on lidar point clouds in spherical coordinate and beyond. *IEEE Trans. Veh. Technol.* **2020**, *69*, 14588–14601. [CrossRef]
4. Zhou, H.; Zou, D.; Pei, L.; Ying, R.; Liu, P.; Yu, W. StructSLAM: Visual SLAM with building structure lines. *IEEE Trans. Veh. Technol.* **2015**, *64*, 1364–1375. [CrossRef]
5. Song, Z.; Lu, J.; Yao, Y.; Zhang, J. Self-Supervised Depth Completion From Direct Visual-LiDAR Odometry in Autonomous Driving. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 11654–11665. [CrossRef]
6. Ma, F.; Cavalheiro, G.V.; Karaman, S. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 3288–3295.

7. Cui, Y.; Chen, R.; Chu, W.; Chen, L.; Tian, D.; Li, Y.; Cao, D. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 722–739. [CrossRef]
8. 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; pp. 11–20.
9. Jaritz, M.; De Charette, R.; Wirbel, E.; Perrotton, X.; Nashashibi, F. Sparse and dense data with cnns: Depth completion and semantic segmentation. In Proceedings of the 2018 International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018; pp. 52–60.
10. Eldesokey, A.; Felsberg, M.; Khan, F.S. Propagating confidences through cnns for sparse data regression. *arXiv* **2018**, arXiv:1805.11913.
11. Eldesokey, A.; Felsberg, M.; Khan, F.S. Confidence propagation through cnns for guided sparse depth regression. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2423–2436. [CrossRef] [PubMed]
12. Yan, L.; Liu, K.; Belyaev, E. Revisiting sparsity invariant convolution: A network for image guided depth completion. *IEEE Access* **2020**, *8*, 126323–126332. [CrossRef]
13. Huang, Z.; Fan, J.; Cheng, S.; Yi, S.; Wang, X.; Li, H. Hms-net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion. *IEEE Trans. Image Process.* **2019**, *29*, 3429–3441. [CrossRef] [PubMed]
14. Ma, F.; Karaman, S. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; pp. 4796–4803.
15. 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.
16. Wei, M.; Zhu, M.; Zhang, Y.; Sun, J.; Wang, J. An Efficient Information-Reinforced Lidar Deep Completion Network without RGB Guided. *Remote. Sens.* **2022**, *14*, 4689. [CrossRef]
17. Hu, M.; Wang, S.; Li, B.; Ning, S.; Fan, L.; Gong, X. Penet: Towards precise and efficient image guided depth completion. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 13656–13662.
18. Li, A.; Yuan, Z.; Ling, Y.; Chi, W.; Zhang, C.; et al. A multi-scale guided cascade hourglass network for depth completion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020; pp. 32–40.
19. Liu, L.; Song, X.; Lyu, X.; Diao, J.; Wang, M.; Liu, Y.; Zhang, L. FCFR-Net: Feature fusion based coarse-to-fine residual learning for depth completion. *arXiv* **2020**, arXiv:2012.08270.
20. Zhang, Y.; Funkhouser, T. Deep depth completion of a single rgb-d image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 175–185.
21. Qiu, J.; Cui, Z.; Zhang, Y.; Zhang, X.; Liu, S.; Zeng, B.; Pollefeys, M. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3313–3322.
22. Xu, Y.; Zhu, X.; Shi, J.; Zhang, G.; Bao, H.; Li, H. Depth completion from sparse lidar data with depth-normal constraints. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2811–2820.
23. Nazir, D.; Liwicki, M.; Stricker, D.; Afzal, M.Z. SemAttNet: Towards Attention-based Semantic Aware Guided Depth Completion. *arXiv* **2022**, arXiv:2204.13635.
24. Yue, J.; Wen, W.; Han, J.; Hsu, L.T. 3D Point Clouds Data Super Resolution-Aided LiDAR Odometry for Vehicular Positioning in Urban Canyons. *IEEE Trans. Veh. Technol.* **2021**, *70*, 4098–4112. [CrossRef]
25. Cheng, X.; Wang, P.; Yang, R. Learning depth with convolutional spatial propagation network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2361–2379. [CrossRef] [PubMed]
26. Cheng, X.; Wang, P.; Guan, C.; Yang, R. Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10615–10622.
27. Yang, Y.; Wong, A.; Soatto, S. Dense depth posterior (ddp) from single image and sparse range. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3353–3362.
28. Shivakumar, S.S.; Nguyen, T.; Miller, I.D.; Chen, S.W.; Kumar, V.; Taylor, C.J. Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Long Beach, CA, USA, 15–20 June 2019; pp. 13–20.
29. Feng, Z.; Jing, L.; Yin, P.; Tian, Y.; Li, B. Advancing self-supervised monocular depth learning with sparse liDAR. In Proceedings of the Conference on Robot Learning, Auckland, New Zealand, 14–18 December 2022; pp. 685–694.
30. Choi, J.; Jung, D.; Lee, Y.; Kim, D.; Manocha, D.; Lee, D. Selfdeco: Self-supervised monocular depth completion in challenging indoor environments. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 467–474.
31. Wong, A.; Soatto, S. Unsupervised depth completion with calibrated backprojection layers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12747–12756.

32. Wong, A.; Fei, X.; Tsuei, S.; Soatto, S. Unsupervised depth completion from visual inertial odometry. *IEEE Robot. Autom. Lett.* **2020**, *5*, 1899–1906. [CrossRef]
33. Godard, C.; Mac Aodha, O.; Firman, M.; Brostow, G.J. Digging into self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3828–3838.
34. Ku, J.; Harakeh, A.; Waslander, S.L. In defense of classical image processing: Fast depth completion on the cpu. In Proceedings of the 2018 15th Conference on Computer and Robot Vision (CRV), Toronto, ON, Canada, 8–10 May 2018; pp. 16–22.