Spacecraft Depth Completion Based on the Gray Image and the Sparse Depth Map

XIANG LIU
HONGYUAN WANG
ZHIQIANG YAN
YU CHEN
Harbin Institute of Technology, Harbin, China
XINLONG CHEN
WEICHUN CHEN
China Academy of Space Technology, Beijing, China

Perceiving the 3D structure of the spacecraft is a prerequisite for successfully executing many on-orbit space missions, and it can provide critical input for many downstream vision algorithms. In this paper, we propose to sense the 3D structure of spacecraft using light detection and ranging sensor (LIDAR) and a monocular camera. To this end, Spacecraft Depth Completion Network (SDCNet) is proposed to recover the dense depth map based on gray image and sparse depth map. Specifically, SDCNet decomposes the spacecraft depth completion task into foreground segmentation subtask and foreground depth completion subtask, which segments the spacecraft region first and then performs depth completion on the segmented foreground area. In this way, the background interference to foreground spacecraft depth completion is effectively avoided. Moreover, an attention-based feature fusion module is also proposed to aggregate the complementary information between different inputs, which deduces the correlation between different features along the channel and the spatial dimension sequentially. Besides, four metrics are also proposed to evaluate object-level depth completion performance. Empirical experiments on the dataset demonstrate the effectiveness of the proposed SDCNet, which achieves 0.225 m mean absolute error of interest and 0.778 m mean absolute truncation error, surpassing state-of-the-art methods by a large margin. The pose estimation experiment is also conducted based on the depth completion results, and the experimental results indicate that the predicted dense depth map could meet the needs of downstream vision tasks.

I. INTRODUCTION

With the rapid development of aerospace technology, numerous on-orbit missions oriented to noncooperative spacecraft [1], [2] have emerged. Among them, perceiving the 3-D structure of spacecraft and providing it to downstream vision algorithms (such as pose estimation [3], [4], component detection [5], 3-D reconstruction [6], etc.) are vital to ensure the successful execution of these tasks.

At present, stereo vision systems [7] and active time-of-flight (TOF) cameras [8] are the primary options for perceiving the 3-D structure of noncooperative spacecraft. Unfortunately, limited by the installation baseline length and power consumption, both of them work at close distances (generally less than 20 m), which brings great challenges to space on-orbit tasks. Moreover, stereo vision systems generally work poorly on objects with smooth surfaces or repetitive textures due to their reliance on the quality of the extracted feature points. Inspired by autonomous driving technology, this article attempts to sense the 3-D structure of spacecraft at a long distance (maximum to 250 m) using a light detection and ranging sensor (LIDAR) and monocular camera. To this end, we propose a depth completion algorithm to recover the 3-D structure of spacecraft using a gray image and sparse depth map.

The depth completion task has been attracting considerable research due to its importance in various fields, and numerous depth completion methods for ground scenes have been proposed. Nevertheless, due to the distinct working scenarios and data characteristics, the current depth completion methods oriented to ground scenarios cannot be directly applied to spacecraft depth completion. Fig. 1 shows the difference between autonomous driving scene data and space on-orbit mission data. The differences in the data collected from different scenarios can be embodied in the following aspects. 1) The data from ground scenarios contains rich background information. On the contrary, the on-orbit data mainly comprises the spacecraft and the irrelevant background. Considering the spacecraft is the sole object to recovering dense depth (unaffected by background changes), the spacecraft depth completion task can be regarded as object-level depth completion to some extent. 2) Due to the different working distances, the depth map of spacecraft obtained by LIDAR is more sparse than the depth map of ground scenarios, bringing significant challenges to spacecraft depth completion. 3) Compared to the ground scene, the on-orbit lighting conditions are more complex, and some areas of the spacecraft are invisible due to lighting shadows.

To alleviate the problem, we propose a Spacecraft Depth Completion Network (SDCNet) for the 3-D structure recovery of spacecraft using a gray image and sparse depth map.
II. RELATED WORKS

Over the last decades, numerous depth completion algorithms with the guidance of optical images have been proposed due to the importance of depth completion in various applications, which can be roughly divided into two categories: the early fusion model and the late fusion model.

The early fusion model generally takes the concatenation of RGB-D as input and predicts dense depth through a U-Net-alike structure. For instance, Sparse-to-dense [9] concatenates the RGB images and sparse depth and adopts the encoder–decoder structure to regress depth for each pixel. Several methods [10], [11], [12] also extract low-level features from the optical image and depth map separately and concatenate them at the first layer of the encoder-decoder network. Moreover, the coarse-to-fine strategy is frequently adopted to generate more accurate depth estimation results [13], [14], [15]. S2DNet [13] utilizes the sparse-to-dense [9] to predict the coarse-level depth map first. The estimated coarse depth map is then concatenated with the input image and fed into the fine network for fine-level depth map estimation. Many methods [15], [16], [17], [18], [19], [20]; also adopt the spatial diffusion process to refine the coarse depth map and achieve promising results. These methods utilize the encoder-decoder structure to simultaneously predict a blur depth map and the affinity matrix. Then the spatial diffusion process is used to generate a final refined depth map according to the predicted affinity matrix.

The late fusion model usually utilizes two branches to extract features from the optical image and depth map, respectively, and then perform feature fusion at intermediate layers. Inspired by guided image filtering [21], GuideNet [22] introduces the guided convolution module, which generates data-driven spatially-variant kernels from the RGB image features to integrate the edge information in RGB into depth map features. FCFRNet [23] utilizes the channel shuffle and the energy-based fusion operation to mix two features extracted from different inputs. On the basis of GuideNet [22], RigNet [24] adopts the repetitive hourglass network to generate more clear image...
Fig. 2. Overall architecture of Spacecraft Depth Completion Network (SDCNet), which comprises a foreground segmentation subnet (FSNet) and foreground depth completion subnet (FDCNet).

Fig. 3. Detailed structure parameter of the FSNet.

guidance. Moreover, the repetitive guidance module is also proposed to progressively generate structure-detailed depth map features. FuseNet [25] replaces the 2-D convolution in the depth map branch with 3-D continuous convolution to extract 3-D features of the depth map. The 3-D features are then projected into image space to form a sparse depth feature map. The sparse depth feature map is finally fused with the RGB feature map for the dense depth prediction.

III. SPACECRAFT DEPTH COMPLETION METHOD

The overall network architecture of our proposed SDCNet is illustrated in Fig. 2. The SDCNet comprises a foreground segmentation subnet (FSNet) and foreground depth completion subnet (FDCNet). The FSNet predicts the probability that each pixel belongs to the spacecraft and segments the foreground area for subsequent spacecraft depth completion. Then the FDCNet regresses the depth of the segmented foreground area by fusing gray image and sparse depth image information.

A. Foreground Segmentation Subnet

Considering the on-orbit data are generally composed of spacecraft and irrelevant background, the spacecraft depth completion task can be equivalent to the object-level depth completion task. At this time, if regressing the depth of all pixels directly, the imbalance of the number of foreground and background samples will inevitably degrade network performance. Moreover, the network will also tend to learn the weight minimizing the overall pixel depth error (including the foreground and background) instead of minimizing the spacecraft depth error. To avoid these problems mentioned above, we design a simple FSNet to segment the foreground region, which is performed on depth regression subsequently. The detailed structure parameter of the FSNet is shown in Fig. 3.

Since the foreground segmentation subtask belongs to the pixel’s binary classification problem, FSNet adopts the encoder-decoder network structure to predict the probability that each pixel belongs to the foreground. Moreover, to fully utilize the complementary information of the gray image and sparse depth map, FSNet takes their concatenation as input. Finally, the designed FSNet only contains 8 convolutional layers and 0.005 M learnable parameters, which can filter out background interference on the subsequent spacecraft depth regression while keeping the negligible overhead.

B. Foreground Depth Completion Subnet

After segmenting the foreground region, the FDCNet is designed to regress the pixel depth in the segmented foreground region. As shown in Fig. 2, FDCNet is composed of the gray image feature extraction branch and the sparse depth map completion branch, both of which adopt the encoder-decoder structure. The gray image feature extraction branch aims to extract the geometric structure and context information from the gray image, providing critical cues for the subsequent spacecraft depth prediction. At the same time, the sparse depth map completion branch predicts the pixel depth utilizing the features extracted from the gray image and the depth map. Considering the adverse effect of sparse data on convolution operations [26], we adopt simple morphological operations to preprocess the sparse depth map, generating coarse pseudodense depth maps and feeding the result into the foreground depth completion.
As shown in Fig. 5, the max-pooling and average-pooling operations are used to extract the global feature for each region. Moreover, the intraregion pixel adaptive weighting operation is also used to characterize the detail features. Finally, the extracted region features under the same channel are spliced to form the channel feature vector. Supposing the original feature map size is $C \times H \times W$, the dimension of the final embedded feature vector is $C \times d_\ell$, where $d_\ell = 3HW/S^2$.

After embedding the feature maps into feature vectors along the channel dimension, the multithead coattention mechanism is introduced to deduce the correlation between different features along the channel dimension. Assuming the feature maps of the gray image and depth map extracted by the convolutional neural network are $f_g \in \mathbb{R}^{C \times H \times W}$ and $f_s \in \mathbb{R}^{C \times H \times W}$, respectively, $f_g^v \in \mathbb{R}^{C \times d_\ell}$ and $f_s^v \in \mathbb{R}^{C \times d_\ell}$ are the embedded feature vector of $f_g$ and $f_s$, respectively. The query of $f_g^v$ and the key of $f_s^v$ can be calculated as

$$
Q_i = f_g^v W_g^Q, i \in [1, n] \quad \text{and} \quad K_i = f_s^v W_s^K, i \in [1, n],
$$

(1)

where $W_g^Q \in \mathbb{R}^{d_\ell \times (d_\ell/n)}$ and $W_s^K \in \mathbb{R}^{d_\ell \times (d_\ell/n)}$ are the parameter matrices, $Q \in \mathbb{R}^{C \times (d_\ell/n)}$ and $K \in \mathbb{R}^{C \times (d_\ell/n)}$ are the query and the key of $f_g^v$ and $f_s^v$, respectively, $n$ is the number of attention heads, and $i$ is the index of attention heads. Then the cross-channel attention map $\omega \in \mathbb{R}^{C \times C}$ between different features can be calculated as

$$
\omega_i = \text{softmax} \left( \frac{Q K_i^T}{\sqrt{d_\ell}} \right).
$$

(2)

The features of the gray image are then integrated into the depth map features according to the cross-channel attention, which can be expressed as

$$
\tilde{f}_i = \text{reshape} \left( F_g^v + \omega F_g^v \right)
$$

(3)

where $\text{reshape}(\cdot)$ denotes vector reshape operation, $F_g^v \in \mathbb{R}^{C \times (H \times W)}$ and $F_s^v \in \mathbb{R}^{C \times (H \times W)}$ are vectorized versions of $f_g$ and $f_s$, respectively. $\tilde{f}_i \in \mathbb{R}^{C \times H \times W}$ represents the single-head fused feature, aggregating the grayscale and depth image information. Similar to Transformer [28], the multithead attention mechanism is adopted to generate advanced fused features, which perform the single-head function in parallel. The interhead fused features adaptive weighting is then adopted to enhance the representation ability of the fused features. The multithead attention mechanism can be expressed as

$$
\tilde{f}_s = \text{Conv} \left( \left[ \tilde{f}_1; \tilde{f}_2; \ldots; \tilde{f}_n \right] \right)
$$

(4)

where $n$ denotes the number of attention heads, and $[\cdot; \cdot]$ means concatenate operation. In this work, we set $n$ to 4.

The cross-channel attention layer can fully mine the correlation between heterologous feature maps along channel dimensions. However, it ignores the different information richness of the gray image features and the depth map features at different spatial locations. For instance, affected by lighting conditions, the shadow region in the gray image may fail to perceive spacecraft, while the LIDAR can still

![Fig. 4. Detailed structure of the feature fusion module, which is mainly composed of cross-channel attention layer and spatial attention layer. The cross-channel attention layer deduces the correlation between different features along the channel dimension, while the spatial fusion layer infers which feature is more informative along the spatial dimension.](image)

LIU ET AL.: SPACECRAFT DEPTH COMPLETION BASED ON THE GRAY IMAGE
provide accurate ranging information in this region. On the other hand, in the regions with satisfactory lighting conditions, due to the sparseness of LIDAR ranging information, LIDAR loses the specific edge information. At this time, the gray image can provide clear edge features to guide spacecraft structure recovery. Therefore, the spatial attention layer is designed to retain more informative features at different spatial locations, which is complementary to the cross-channel attention module.

As shown in Fig. 4, the spatial attention layer takes the feature \( f_s \subset \mathbb{R}^{C \times H \times W} \) output from the cross-channel attention layer and the gray image feature \( f_g \subset \mathbb{R}^{C \times H \times W} \) as input and outputs the final fused feature \( f_{out} \subset \mathbb{R}^{C \times H \times W} \) for subsequent depth prediction. Like CBAM [29], we first apply max-pooling and average-pooling operations along the channel axis on \( f_s \) and \( f_g \) and concatenate them to generate an efficient descriptor. Then a convolution layer and the sigmoid function are applied to generate the spatial attention map \( W_{\text{spatial}} \subset \mathbb{R}^{1 \times H \times W} \), which encodes the richness of information at each location in the depth map. The final fused feature is the spatial-variant weighting of \( f_s \) and \( f_g \), which can be expressed as

\[
 f_{out} = W_{\text{spatial}} \odot f_s + (1 - W_{\text{spatial}}) \odot f_g 
\]

where \( \odot \) denotes elementwise multiplication. During multiplication, the spatial attention map \( W_{\text{spatial}} \) is broadcasted along the channel dimension.

C. Loss Functions

The loss functions of SDCNet consist of two parts: the foreground prediction loss for FSNet and the foreground depth completion loss for FDCNet. For the foreground prediction loss, the binary cross-entropy loss is adopted to supervise the foreground region segmentation, which can be expressed as

\[
 \mathcal{L}_{FSNe t} = \frac{1}{N} \sum_{i} \left( -y_i \log (\hat{y}_i) + (1 - y_i) \log (1 - \hat{y}_i) \right) 
\]

where \( y_i \) and \( \hat{y}_i \) are the foreground ground truth label and predicted foreground probability, respectively. If the pixel belongs to the foreground region, its ground truth label is set to 1. Otherwise, it is set to 0.

The foreground depth completion loss aims to minimize the error between the network predicted depth and the ground truth depth in the predicted foreground region, which can be calculated as

\[
 \mathcal{L}_{FDCNet} = \frac{1}{N(\hat{A})} \sum_{x_i \in \hat{A}} |\hat{d}_i - d_i| 
\]

where \( \hat{A} \) represents the set of pixels predicted as foreground pixels in FSNet. \( N(\hat{A}) \) denotes the number of elements in set \( \hat{A} \), \( \hat{d}_i \) and \( d_i \) denote the predicted depth and ground truth depth for pixel \( x_i \), respectively.

D. Evaluation Metrics

For the foreground segmentation subtask, we use intersection over union (IOU) and intersection over interest (IOI) to evaluate the performance of FSNet, which can be calculated as

\[
 IOU = \frac{N(A \cap \hat{A})}{N(A \cup \hat{A})} 
\]

\[
 IOI = \frac{N(A \cap \hat{A})}{N(A)} 
\]

where \( \hat{A} \) denotes the set of pixels predicted as foreground pixels in FSNet, \( A \) denotes the set of pixels belonging to the foreground in the ground truth label. \( N(\cdot) \) is the function that counts the number of elements in the set. The IOI is mainly used to evaluate the retention rate of the ground truth foreground. The larger the IOI, the fewer foreground pixels are misclassified as background, and the more complete the retained foreground information is.

In the depth completion task, the mean absolute error (MAE) and the root-mean-squared error (RMSE) are often used to evaluate the performance of depth completion. However, since MAE and RMSE are the statistics of all pixel depth prediction errors, the results are easily unstable in the object-level depth completion task due to the change in the ratio of the foreground and background. Therefore, four new metrics to evaluate the quality of object-level depth completion results are proposed, including the MAE of interest (MAEI), the mean absolute truncation error (MATE), the RMSE of interest (RMSEI), and the RMS truncation error (RSTE).
error (RMSTE), which can be calculated as

$$\text{MAEI} = \frac{1}{N(A \cap \hat{A})} \sum_{x_i \in A \cap \hat{A}} |\hat{d}_i - d_i|$$  \hspace{1cm} (10)

$$\text{RMSEI} = \sqrt{\frac{1}{N(A \cup \hat{A})} \sum_{x_i \in A \cup \hat{A}} |\hat{d}_i - d_i|^2}$$  \hspace{1cm} (11)

$$\text{MATE}(\alpha) = \frac{1}{N(A \cup \hat{A})} \min_{x_i \in A \cup \hat{A}} (|\hat{d}_i - d_i|, \alpha)$$  \hspace{1cm} (12)

$$\text{RMSTE}(\alpha) = \sqrt{\frac{1}{N(A \cup \hat{A})} \sum_{x_i \in A \cup \hat{A}} \min_{x_i \in A \cup \hat{A}} (|\hat{d}_i - d_i|^2, \alpha^2)}$$  \hspace{1cm} (13)

where $\hat{A}$ and $A$ denote the set of pixels predicted as foreground pixels and the set of pixels belonging to the foreground in the ground truth, respectively. $\hat{d}_i$ and $d_i$ denote the predicted depth and ground truth depth for pixel $x_i$. $N(\cdot)$ is the function that counts the number of elements in the set, and the parameter $\alpha$ is the truncation threshold.

MAEI, RMSEI, MATE, and RMSTE essentially calculate MAE and RMSE in different image areas. MAEI and RMSEI calculate MAE and RMSE in the intersection area of the predicted foreground and the ground truth foreground, which can more intuitively reflect the depth completion accuracy of the spacecraft itself. On the other hand, MATE and RMSTE calculate MAE and RMSE in the union area of the predicted foreground and the ground truth foreground. Considering the error of the misclassified pixels is large and depends on the object’s distance, threshold truncation is performed to prevent it from causing excessive influence on MAE and RMSE. In our work, we set the truncation threshold $\alpha$ to 10 m.

### IV. SATELLITE DEPTH COMPLETION DATASET CONSTRUCTION

Training a deep network for spacecraft depth completion requires an extensive collection of satellite images and LIDAR data. To date, there is no public dataset for spacecraft depth completion tasks. This article constructs a large-scale satellite depth completion dataset for training and testing spacecraft depth completion algorithms.

We performed the imaging simulation of spacecraft using blender software based on 126 satellite CAD models. The specific camera and LIDAR parameters are shown in Tables I and II. Specifically, the LIDAR parameters are derived from the JAGUAR product of Innovusion company.

During the simulation, a sphere with the earth’s texture plays the role of the earth. The angle between the relative direction of the earth and the observation platform and the observation direction is randomly sampled in $[-5^\circ, 5^\circ]$ to capture the earth in the optical image frequently. Moreover, we randomly set the satellite solar plane size to 3–8 m and the body size to 1–3 m. The satellite’s distance to the observation platform is set in the range of 50–250 m, and the satellite’s attitude relative to the observation platform is randomly sampled from pose space. The maximum angle between the illumination and observation directions is set to 70°.

We simulated 64 sets of data under different observation conditions for each satellite model, resulting in 8064 sets of gray images and LIDAR depth maps. Fig. 6 shows some simulation results of different satellite models. To ensure the generalization performance of the algorithm to different satellite models, we divided the simulation data of 126 satellite models into three subsets: training set (simulation results of 99 models), validation set (simulation results of nine models), and test set (simulation results of 18 models), resulting in 6336, 576, and 1152 sets of data for training, validation, and testing. In this way, we can ensure that the data used for validation and testing is invisible to the depth completion network, putting forward high requirements for the applicability of the algorithm to unknown satellite models.

### V. EXPERIMENTS

#### A. Experiment Setup

The proposed SDCNet is implemented in python using the Paddle library and trained on an Nvidia RTX 3090 GPU. During training, we first train the FSNet using the standard binary cross-entropy loss, then freeze the weights in FSNet and train the whole network. We train the FSNet and the whole model using the Adam [30] optimizer for 50 and 100 epochs, with an initial learning rate of 0.001 and a weight decay of 0.001. In addition, data augmentation techniques, including random flip and image jitter, are adopted.

#### B. Experiment Results

We compare our method with state-of-the-art depth-completion methods on the constructed satellite depth-completion dataset, including Sparse-to-dense [9], CSPN [17], GuideNet [22], FCFRNet [23], RigNet [24], PENet [15], and DySPN [20]. All the methods are trained on the same training set and evaluated on the same test

| Parameter          | Value   |
|--------------------|---------|
| Focal length/mm    | 50      |
| Field of view/°    | 7.38×7.38 |
| Image resolution/pixel | 512×512 |
| Sensor size/mm     | 6.449×6.449 |

| Parameter          | Value   |
|--------------------|---------|
| Maximum range/m    | 280     |
| Range error/cm     | <3      |
| Vertical angular resolution/° | 0.13 |
| Horizontal angular resolution/° | 0.09 |
Fig. 6. Examples of simulation results of different satellite models. From top to bottom: The gray image, the sparse depth map which is dilated for better visualization, and the ground truth dense depth map. (a) Optical camera imaging simulation results. (b) LIDAR sparse ranging simulation results. (c) Ground truth dense depth map.

TABLE III

| Methods                   | MAEI/m | MATE/m | RMSEI/m | RMSTE/m | Average time/ms |
|---------------------------|--------|--------|---------|---------|-----------------|
| Sparse-to-dense/           | 3.871/3.680 | 5.892/5.845 | 4.742/4.616 | 7.053/6.839 | 8.6             |
| Sparse-to-dense*           | 2.732/3.383  | 3.863/4.660  | 3.549/4.262  | 5.024/5.747  | 93.6            |
| CSPN/CSPN*                 | 2.241/0.593  | 2.688/1.048  | 3.623/1.691  | 3.994/2.558  | 65.5            |
| GuideNet/GuideNet*         | 2.484/1.670  | 2.886/2.031  | 3.832/3.051  | 4.058/3.345  | 112.4           |
| FCFRNet/FCFRNet*           | 1.397/0.451  | 2.700/1.699  | 1.971/1.515  | 4.195/3.662  | 111.0           |
| RigNet/RigNet*             | 0.720/0.567  | 1.232/1.307  | 1.405/1.326  | 2.644/3.007  | 196.9           |
| DySPN/DySPN*               | 1.406/0.383  | 2.060/0.927  | 2.227/1.042  | 3.370/2.418  | 93.8            |
| SDCNet                    | **0.225**   | **0.778**    | **0.691**   | **2.304**    | 64.9            |

* Denotes that the extra morphological preprocessing operation is adopted on the LIDAR depth map. The average inference time is tested on an NVIDIA RTX 3090 GPU, including the time consumed by the morphological preprocess (about 5ms). Bold values represent the optimal values of different evaluation criteria among all methods.

set. Considering that the spacecraft LIDAR depth maps are more sparse compared to ground scenes, to ensure a fair comparison, we optionally adopt the extra morphological preprocessing operation on the sparse depth map to generate the pseudodense depth map and feed it into the existing depth completion network. Table III lists the quantitative results of different methods. It can be seen from Table III that, compared to taking the sparse depth map as input directly, the depth completion accuracy of most methods improved dramatically with adopting extra
morphological operation on the sparse depth map. This phenomenon demonstrates that too sparse data inhibits convolution operations from extracting instructive features, which inspires us to adopt the morphological preprocessing before the foreground depth completion subnet. In addition, thanks to the lightweight FSNet and the efficient feature fusion module, the proposed SDCNet outperforms all existing methods by a large margin in terms of accuracy. Regarding computing efficiency, sparse-to-dense maintains the fastest inference speed due to its simple structure while performing the poorest in accuracy. Excluding the worst-performed sparse-to-dense and comparable GuideNet, SDCNet is also significantly faster than other existing methods.

Fig. 7 shows some qualitative depth completion examples of SDCNet. The first and second rows in Fig. 7 are the gray images and the sparse depth maps, respectively. They are the inputs of the depth completion network. The third row in Fig. 7 shows the foreground segmentation results predicted by FSNet. It can be seen that, thanks to the complementary information of gray images and depth maps, the designed lightweight foreground segmentation can accurately segment the foreground region even if the spacecraft overlaps with the earth. The fourth and fifth rows
are the predicted dense depth map and ground truth also Effects of Different Inputs to FSNet: We choose $V$ with adopting extra morphological preprocessing as also shows the change of depth Effectiveness of Different Component: $V$ conducts to verify the effectiveness of each component including the FSNet, the cross-channel and spatial attention layers to aggregate proposed in our method, including the FSNet, the cross-channel attention layer (CCA), and the spatial attention layer (SA).

As shown in Table IV, when the FSNet solely takes the gray image as input, this version of FSNet performs worst, and the IOU and IOI of the prediction results are only 68.37% and 81.85%. This is mainly because the optical camera can simultaneously observe the spacecraft and the earth. When the spacecraft overlaps with the earth, it is challenging to segment the spacecraft with a lightweight network accurately. Moreover, the optical camera is sensitive to the on-orbit lighting condition, and the regions in the shadow are generally invisible to the gray images. When the FSNet solely takes the sparse depth map as input, the IOU and IOI increase to 86.82% and 93.27%, respectively. This is thanks to the fact that LIDAR data does not contain the background and is robust to illumination. Nonetheless, considering that the LIDAR cannot provide exact edge structure information due to its sparse-ranging results, the accuracy of foreground prediction solely using the sparse depth map is still unsatisfactory. The final version takes the gray image and the sparse depth map jointly into the FSNet, and prediction results are significantly improved, with the IOU and IOI reaching 94.37% and 97.30%, respectively. On one hand, the depth image can provide a rough contour of the foreground, even if part of its feature is annihilated in the gray image due to poor illumination. On the other hand, the gray image can provide clear edge structure information in satisfactory lighting conditions. In this way, FSNet can take full advantage of the complementary information of different sensor data and accurately classify whether a pixel belongs to the foreground under different lighting conditions.

2) Effectiveness of Different Component: We choose the state-of-the-art depth completion network GuideNet [22] with adopting extra morphological preprocessing as the baseline, which has the same structure as FDCNet except employs the guided convolution module to fuse the features extracted from the optical image and the depth map. On this basis, the FSNet is first introduced to enable the network to focus on recovering the foreground region depth (the second version in Table V). It can be seen that the introduction of the FSNet significantly improves depth completion accuracy. This phenomenon reflects that the background pixels will interfere with the depth recovery of the object itself and degrade the network performance. At the same time, it verifies the effectiveness of decomposing the object-level depth completion task into the foreground segmentation subtask and foreground depth completion subtask. Further replacing the guided convolution module with the proposed cross-channel attention layer (the third version in Table V), the depth prediction error MAEI and MATE are further reduced by 5.6 and 5.3 cm, respectively. Similarly, replacing the guided convolution module with the spatial attention layer can also decrease the depth completion error (the fourth version in Table V). The final version (the fifth version in Table V), which successively uses the cross-channel and spatial attention layers to aggregate features of different inputs, achieves the best performance, verifying the effectiveness of the multisource feature fusion module.

C. Ablation Studies

In this section, we first conduct additional experiments to explore how different input choices affect the performance of the FSNet. Moreover, ablation studies are also conducted to verify the effectiveness of each component proposed in our method, including the FSNet, the cross-channel attention layer (CCA), and the spatial attention layer (SA).

1) Effects of Different Inputs to FSNet: In order to analyze the influence of different inputs on the performance of the foreground prediction network, we train the FSNet with different inputs, and the quantitative experimental results are listed in Table IV.
TABLE V
Quantitative Results of SDCNet With Different Components

| Baseline | FSNet | CCA | SA | MAEI/m | MATE/m | RMSEI/m | RMSTE/m |
|----------|-------|-----|----|--------|--------|---------|---------|
| ✓        | ✓     | ✓   | ✓  | 0.593  | 1.048  | 1.691   | 2.558   |
| ✓        | ✓     | ✓   | ✓  | 0.324  | 0.863  | 1.024   | 2.383   |
| ✓        | ✓     | ✓   | ✓  | 0.268  | 0.810  | 0.840   | 2.331   |
| ✓        | ✓     | ✓   | ✓  | 0.290  | 0.835  | 0.846   | 2.345   |
| ✓        | ✓     | ✓   | ✓  | 0.225  | 0.778  | 0.691   | 2.304   |

Bold values represent the optimal values of different evaluation criteria among all versions.

Fig. 9. Some qualitative examples of spacecraft pose estimation. The green, blue, and red point clouds denote the source, target, and transformed source point cloud, respectively.

D. Application for Pose Estimation

To evaluate the predicted depth quality and explore the feasibility of utilizing the predicted dense depth map in downstream vision tasks, we conduct the spacecraft pose estimation experiment based on the results of SDCNet. Considering that the noncooperative/model-free spacecraft pose estimation task aims to estimate the spacecraft’s pose changes relative to the observation platform, which can be equivalent to the interframe point cloud registration problem. Therefore, the previous frame point cloud (referred to as the source point cloud) and the current frame point cloud (referred to as the target point cloud) are needed in this experiment. To this end, we simulated a sequence of data containing 36 frames for each satellite in the test set, and the Euler angle and position along each axis are randomly increased within \([5°, 10°]\) and \([1 \text{ m}, 2 \text{ m}]\) for each adjacent frame. The simulated data is then fed into SDCNet to obtain the predicted dense map, which is used as the input for the pose estimation experiments.

In this experiment, we choose the famous two-stage point cloud registration method (sample consensus initial alignment (SAC-IA) [31] + iterative closest point (ICP) [32]) as our pose estimation method. Since SAC-IA+ICP requires the point clouds of spacecraft as input, we convert the predicted depth map into the point cloud according to the camera parameters. Moreover, the statistical outlier removal operation is also adopted to improve point cloud quality. Fig. 9 shows several pose estimation visualization examples on the testing point cloud. It can be seen that the transformed source point cloud (the point clouds in red) aligns well with the target point cloud (the point clouds in blue), which implies the pose estimation results are with high accuracy. According to statistics, the average three-axis rotation and translation errors are \(1.4°\) and \(1.121 \text{ m}\), respectively. The sparse point clouds obtained by LIDAR are also fed into the pose estimation method for comparison. The average three-axis rotation and translation errors are \(15.9°\) and \(18.262 \text{ m}\), respectively. It can be seen that the predicted dense depth can boost the pose estimation accuracy dramatically, which reflects the necessity of depth completion and verifies that the proposed depth completion method can be applied in downstream vision tasks.

VI. CONCLUSION

Aiming at the limited work distance of the existing stereo vision system and active time-of-flight (TOF) camera, this article proposes to sense the 3-D structure of spacecraft at a long distance (maximum to 250 m) using LIDAR and a monocular camera. To this end, a novel SDCNet is proposed to recover the dense depth map using a gray image and sparse depth map. Considering that the irrelevant background inevitably interferes with the spacecraft depth...
recovery, the object-level spacecraft depth completion task is decomposed into the foreground segmentation subtask and the foreground depth completion subtask. Specifically, a lightweight FSNet is designed for foreground region segmentation first, and the pixel’s depth in the segmented region is regressed using the FDCNet. Moreover, we design an attention-based feature fusion module to deduce the correlation between different features along the channel and the spatial dimension sequentially, integrating the geometric features and context the gray image provides into the depth map feature. Four new metrics for the object-level depth completion task are also proposed to evaluate depth completion results, including MAEI, MATE, RMSEI, and RMSTE. Besides, we construct a large-scale satellite depth completion dataset based on 126 satellite CAD models, containing 6336, 576, and 1152 sets of data for training, validation, and testing the spacecraft depth completion algorithms. The construction of the satellite depth completion dataset solves the lack of satellite data for training and testing depth completion methods. Empirical experiments on the dataset demonstrate that our method achieves state-of-the-art depth completion performance, which achieves 0.225 m mean absolute error of interest and 0.778 m mean absolute truncation error. Finally, the spacecraft pose estimation experiment is also conducted based on the depth completion results, which achieves 1.4° rotation error and 1.121 m translate error, verifying that the predicted depth map could meet the needs of downstream vision tasks. The proposed spacecraft depth completion method has the potential to be integrated into space on-orbit service systems, which can perceive the fine 3-D structure of spacecraft at a long distance using LIDAR and optical camera.

REFERENCES

[1] Y. Gao and S. Chien, “Review on space robotics: Toward top-level science through space exploration,” Sci. Robot., vol. 2, no. 7, Jun. 2017, Art. no. eaan5074, doi: 10.1126/scirobotics.aan5074.
[2] L. Chen, P. Huang, and J. Cai, “Extracting and matching lines of low-textured region in close-range navigation for tethered space robot,” IEEE Trans. Ind. Electron., vol. 66, no. 9, pp. 7131–7140, Sep. 2019, doi: 10.1109/TIE.2018.2879286.
[3] J. Ventura, A. Fleischner, and U. Walter, “Pose tracking of a non-cooperative spacecraft during docking maneuvers using a time-of-flight sensor,” in Proc. AIAA Guid. Nav. Control Conf., 2016, Art. no. 0875, doi: 10.2514/6.2016-0875.
[4] X. Liu, H. Wang, X. Chen, W. Chen, and Z. Xie, “Position awareness network for non-cooperative spacecraft pose estimation based on point cloud,” IEEE Trans. Aerosp. Electron. Syst., vol. 59, no. 1, pp. 507–518, Feb. 2023, doi: 10.1109/TAES.2022.3182307.
[5] Q. Wei, Z. Jiang, and H. Zhang, “Robust spacecraft component detection in point clouds,” Sensors, vol. 18, no. 4, Mar. 2018, Art. no. 933, doi: 10.3390/s18040933.
[6] R. Volpe, M. Sabatini, and G. B. Palmerini, “Reconstruction of the shape of a tumbling target from a chaser in close orbit,” in Proc. Aerosp. Conf., 2020, pp. 1–11, doi: 10.1109/AERO4725.2020.9172529.
[7] W. Xu, Q. Xue, H. Liu, X. Du, and B. Liang, “A pose measurement method of a non-cooperative GEO spacecraft based on stereo vision,” in Proc. Int. Conf. Control Autom. Robot. Vis., 2012, pp. 966–971, doi: 10.1109/ICARCV.2012.6485288.
[8] H. G. Martínez, G. Giorgi, and B. Eissfeller, “Pose estimation and tracking of non-cooperative rocket bodies using time-of-flight cameras,” Acta Astronautica, vol. 139, pp. 165–175, Oct. 2017, doi: 10.1016/j.actaastro.2017.07.002.
[9] F. Ma and S. Karaman, “Sparse-to-dense: Depth prediction from sparse depth samples and a single image,” in Proc. Int. Conf. Robot. Autom., 2018, pp. 4796–4803, doi: 10.1109/ICRA.2018.8460184.
[10] S. Imran, Y. Long, X. Liu, and D. Morris, “Depth coefficients for depth completion,” in Proc. Conf. Comput. Vis. Pattern Recognit., 2019, pp. 12438–12447, doi: 10.1109/CVPR.2019.01273.
[11] F. Ma, G. V. Cavalheiro, and S. Karaman, “Self-supervised sparse-to-dense: Self-supervised depth completion from LIDAR and monocular camera,” in Proc. Int. Conf. Robot. Autom., 2019, pp. 3288–3295, doi: 10.1109/ICRA.2019.8793637.
[12] B.-U. Lee, K. Lee, and I. S. Kweon, “Depth completion using plane-residual representation,” in Proc. Comput. Vis. Pattern Recognit. Conf., 2021, pp. 13911–13920, doi: 10.1109/CVPR46437.2021.01370.
[13] P. Pambharde and S. Murala, “S2DNet: Depth estimation from single image and sparse samples,” IEEE Trans. Comput. Imag., vol. 6, pp. 806–817, 2020, doi: 10.1109/TCI.2020.2981761.
[14] L. Liu, Y. Liao, Y. Wang, A. Geiger, and Y. Liu, “Learning steering kernels for guided depth completion,” IEEE Trans. Image Process., vol. 30, pp. 2850–2861, 2021, doi: 10.1109/TIP.2021.3055629.
[15] M. Hu, S. Wang, B. Li, S. Ning, L. Fan, and X. Gong, “PENet: Towards precise and efficient image guided depth completion,” in Proc. Int. Conf. Robot. Autom., 2021, pp. 13656–13662, doi: 10.1109/ICRA48506.2021.9561035.
[16] S. Liu, S. D. Mello, J. Gu, G. Zhong, M.-H. Yang, and J. Kautz, “Learning affinity via spatial propagation networks,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 1519–1529, doi: 10.48550/arXiv.1710.01020.
[17] X. Cheng, P. Wang, and R. Yang, “Learning depth with convolutional spatial propagation network,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 10, pp. 2361–2379, Oct. 2020, doi: 10.1109/TPAMI.2019.2947374.
[18] X. Cheng, P. Wang, C. Guan, and R. Yang, “CSPN++: Learning context and resource aware convolutional spatial propagation networks for depth completion,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 10615–10622, doi: 10.1609/aaai.v34i07.6635.
[19] J. Park, K. Joo, Z. Hu, C.-K. Liu, and I. S. Kweon, “Non-local spatial propagation network for depth completion,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 120–136, doi: 10.1007/978-3-030-58601-0_8.
[20] Y. Lin, T. Cheng, Q. Zhong, W. Zhou, and H. Yang, “Dynamic spatial propagation network for depth completion,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 1636–1646, doi: 10.1609/aaai.2021.20055.
[21] K. He, J. Sun, and X. Tang, “Guided image filtering,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 6, pp. 1397–1409, Jun. 2013, doi: 10.1109/TPAMI.2012.213.
[22] J. Tang, F.-F. Tian, W. Feng, J. Li, and P. Tan, “Learning guided convolutional network for depth completion,” IEEE Trans. Image Process., vol. 30, pp. 1116–1129, Dec. 2020, doi: 10.1109/TIP.2020.3040528.
[23] L. Liu et al., “FCFR-Net: Feature fusion based coarse-to-fine residual learning for depth completion,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 2136–2144, doi: 10.1609/aaai.2021.08270.
[24] Z. Yan, K. Wang, X. Li, Z. Zhang, J. Li, and J. Yang, “RigNet: Repetitive image guided network for depth completion,” in Proc. Eur. Conf. Comput. Vis., 2022, pp. 214–230, doi: 10.1007/978-3-031-19812-0_13.
[25] Y. Chen, B. Yang, M. Liang, and R. Urtasun, “Learning joint 2D-3D representations for depth completion,” in Proc. Int. Conf. Comput. Vis., 2019, pp. 10023–10032, doi: 10.1109/ICCV.2019.01012.
[26] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, “Sparsity invariant CNNs,” in Proc. Int. Conf. 3D Vis., 2017, pp. 11–20, doi: 10.1109/3DV.2017.00012.
[27] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proc. Int. Conf. Comput. Vis., 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.
[28] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, arXiv:1706.03762.
[29] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 3–19, doi: 10.48550/arXiv.1807.06521.
[30] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Representation, 2015, pp. 1–15.
[31] R. B. Rusu, N. Blodow, and M. Beetz, “Fast point feature histograms (FPFH) for 3D registration,” in Proc. Int. Conf. Robot. Autom., 2009, pp. 3212–3217, doi: 10.1109/ROBOT.2009.5152473.
[32] P. J. Besl and N. D. McKay, “A method for registration of 3-D shapes,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 14, no. 2, pp. 239–256, Feb. 1992, doi: 10.1117/12.57955.

Xiang Liu received the B.Sc. degree in optoelectronic information science and engineering from the Harbin Institute of Technology, Weihai, China, in 2017, and the M.S. degree in optical engineering from the Harbin Institute of Technology, Harbin, Heilongjiang, China, in 2019. He is currently working toward the Ph.D. degree in optical engineering with the Space Optical Engineering Center, Harbin Institute of Technology, Harbin, China. His research interests include the satellite component recognition, pose estimation, and behavior understanding of spacecraft based on computer vision.

Hongyuan Wang received the B.S. degree in thermal power engineering and the Ph.D. degree in optical engineering from the Harbin Institute of Technology, Harbin, China, in 2000 and 2008, respectively. He is currently a Professor with the Harbin Institute of Technology. His major research interests include the application of deep learning method in aerospace, photoelectric imaging system modeling and simulation, guidance, and navigation of spacecraft based on computer vision.

Zhiqiang Yan received the B.S. and Ph.D. degrees in optical engineering from the Harbin Institute of Technology, Harbin, China, in 2017 and 2023, respectively. He is currently a Postdoctoral in optical engineering with the Harbin Institute of Technology. His research interests include imaging simulation, visual SLAM, and pose estimation of non-cooperative targets.

Yu Chen received the B.S. degree in optoelectronics information science and engineering in 2021 from the Harbin Institute of Technology, Harbin, China, where she is currently working toward the M.S. degree in optical engineering. Her research interests include multimodal data fusion and satellite pose estimation.

Xinlong Chen received the Ph.D. degree in aircraft design from the Harbin Institute of Technology, Harbin, China, in 2008. In 2008, he joined the China Academy of Space Technology, Beijing, China, where he is currently a Professor with the Qian Xuesen Laboratory of Space Technology, and also the Director of Research Center of Space Debris Removal. His research interests include the overall design and research of spacecraft in space security and in-orbit services.

Weichun Chen received the M.S. degree in thermal power engineering from the Dalian University of Technology, Dalian, China, in 2002. In 2002, he joined the China Academy of Space Technology, Beijing, China, where he is currently a Senior Engineer with the Qian Xuesen Laboratory of Space Technology. His main research interests include debris removal, navigation, control and dynamics for spacecraft, and novel technology development for space security.