# TRANSFORMER-BASED METHOD FOR SEMANTIC SEGMENTATION AND RECONSTRUCTION OF THE MARTIAN SURFACE

Z. Li, B. Wu *, Z. Chen, Y. Ma

Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong – bo.wu@polyu.edu.hk

**KEYWORDS:** Semantic segmentation, reconstruction, deep learning, transformer

## ABSTRACT:

The last decade has witnessed a great advance in deep space exploration, such as the rover missions to Mars. Semantic information on the Martian surface is garnering more attention, for its ability to distinguish the surface landforms for rover traverse planning and facilitating 3D reconstruction. The state-of-the-art studies on semantic segmentation exclusively leveraged transformer-based methods, and the results have been verified to outperform the traditional convolutional neural networks. However, few datasets concerning the Martian surface have been generated, and the publicly available network models were all trained on the common Earth dataset. Constructing a pixel-wise semantic segmentation dataset requires lots of human labor, especially for training a large transformer network. Furthermore, the results of semantic segmentation were typically used for intuitive visualization but seldom exploited in the 3D reconstruction pipeline. To address these problems, this paper presents the following three contributions: (1) introducing an approach to generate a large dataset for Mars in a semi-automatic way; (2) development of a novel variant of transformer designed for multi-view semantic segmentation to improve the accuracy; (3) development of a semantic-aware dense image matching method for improved matching performances assisted with the semantic information. Experimental results using the dataset collected at the Zhurong landing site on Mars have shown superior performances of the proposed methods as compared with traditional methods.

## 1. INTRODUCTION

Semantic reconstruction of the Martian surface is garnering more attention, for its ability to present semantic information in three-dimensional (3D) space, thereby facilitating deep space exploration missions from the aspects of rover traverse planning, risk precautions, and 3D products (Li et al., 2022; Wu et al., 2022). However, the retrieval of the semantic information or semantic segmentation from the 2D image is still an active topic, especially for the Martian surface.

Although many mature semantic segmentation networks such as ViT (Dosovitskiy et al., 2020), and Swin-Transformer (Liu et al., 2021) are publicly available, they have predominantly been trained on conventional Earth datasets, which hinder direct usage without transfer learning. The latest large learning model, namely segment anything model (SAM) (Kirillov et al., 2023), argued it is a zero-shot neural network. While the semantic classes of these segment masks are unavailable, and the utilization of these masks is hence limited to distinguishing characteristic regions. In addition, a conventional neural network designed for segmentation tasks typically takes one image as the input and conducts accuracy evaluation individually. Even if the neural network is capable of yielding proper semantic labels with favorable evaluation metrics, the perspective-invariant trait is not guaranteed or even concerned. Specifically, these inconsistently segmented labels do not conform to the scenario in the real world, which may confuse the downstream visualization or utilization involving multi-view images to achieve semantic reconstruction (Wan et al., 2021).

Since deep-learning is inherently a data-driven method, training with a segmentation dataset constructed with planetary images is also indispensable. But it is still challenging for two-fold reasons: one is constructing a pixel-wise semantic segmentation dataset requires substantial human labor, and the number of planetary images is severely limited (Ma et al., 2023). In 2019, ESA pioneered the public LabelMars project (Schwenzer et al., 2019) to organize a large labeled dataset based on thousands of images from Spirit, Opportunity, and Curiosity. The dataset was then challenged by NASA for its overly specialized categorization and the resulting small volume. The AI4MARS dataset (Swan et al., 2021) was thus proposed based on similar data, but with more intuitive labels, namely, sand, bedrock, soil, and big rock. The associated depth data was also provided, which enhances the versatility of the dataset. Even they claimed that the dataset is comprised of ~35K images, only ~18k images are available, and the detailed distribution of each class is unavailable. The involving tremendous human labor in generating such a dataset makes it hard to be further augmented with more images from the latest rovers. Simulation strategies are hence considered, and Ma et al. (2023) used the OAISYS simulator to add some rocks to the designed surface to generate a large dataset. But the images vary from the real scene on Mars a lot.

Rather than simply visualizing the semantic labels in a 3D reconstruction result, a recent trend in semantic reconstruction is to exploit the semantic cues to achieve semantic-aware algorithms (Zhao et al., 2023; Zheng et al., 2022). Naseer et al. (2017) tested the semantic-aware idea by boosting the feature matching between the images over a long period or with harsh perspective conditions. The superiority of the semantic-aware loop closure detection has also been verified (Zheng et al., 2022), which embedded the semantic labels into the similarity measurement network and led to robust 3D reconstruction results. Zhao et al. (2023) leveraged the segmentation results as guidance and improved the height estimation for single-view UAV images,

* Corresponding author. Email: bo.wu@polyu.edu.hk

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-1643-2023

thereby illustrating the effectiveness of retrieving dense 3D information with the assistance of semantic segmentation. However, multi-view semantic-aware dense image matching is hardly discussed due to the lack of consistent semantic segments.

To this end, three contributions are made by this paper to achieve semantic reconstruction on the Martian surface. Firstly, we introduce an approach to generate a large volume dataset in a semi-automatic way, which not only contains the semantic label, but the 3D information (depth and position). Then, a siamese-like Swin-Transformer is proposed specifically designed for multi-view semantic segmentation tasks. Further constrained by the control points calculated through the tie-point matching algorithm, the network could be trained in a semi-self-supervised fashion, considering all kinds of transformations. Finally, a semantic-aware dense image matching is presented, which verifies the correctness of the retrieved semantic information and makes an attempt to incorporate the semantic cues with the RGB information to achieve better disparity images as well.

## 2. TRANSFORMER-BASED METHOD FOR SEMANTIC SEGMENTATION AND RECONSTRUCTION

### 2.1 Overview

As shown in Figure 1, the proposed approach comprises three consecutive phases. In the first step, a training dataset targeting semantic segmentation for planetary surfaces is constructed. The 3D original-textured and semantic-masked mesh models are first generated through a rigorous photogrammetric process facilitated by some manually labeled images. With the virtual cameras defined by interior orientation (IO) and exterior orientation parameters (EO), the original RGB image, the semantic image, the XYZ image, and the depth image could be obtained. Secondly, the siamese swin-transformer is trained pair-wisely on the obtained dataset. With the tie-points calculated between the input images, the contrastive learning could progress in a self-supervised manner hence avoiding labeling issues (i.e., missing small rocks and sand dunes on the far side). The semantic labels are then fed into the dense image matching pipeline, together with the original RGB image, to refine the disparity image by introducing the adapted strategy to each semantic class and the boundary.

### 2.2 Semi-automatic Dataset Construction

Despite directly augmenting the 2D images through perspective transformations (i.e., translation, rotation, scale transform), a more realistic approach to boosting the volume of the dataset is proposed. The core idea of the semi-automatic semantic dataset construction is to fully exploit the 3D reconstruction results of the images, which relies on the premise that the traverse of the rover is typically continued or at least several images share some overlapping regions. Following the *ad hoc* structure from motion (SfM) pipeline (Agarwal et al., 2009; Schonberger and Frahm, 2016), the bundle adjustment could perform based on the tie-points among these images, and the 3D textured mesh model is formed from the dense point clouds calculated from the multi-view stereo (MVS) algorithm (Vu et al., 2012). The semantic-masked 3D model could also be obtained with some manually labeled images.

In the real world, numerous 2D images could be obtained from a 3D world given a virtual camera $\boldsymbol{P}$ defined by both intrinsic orientation parameter $\boldsymbol{K}$ and extrinsic orientation parameters $\boldsymbol{E}$.

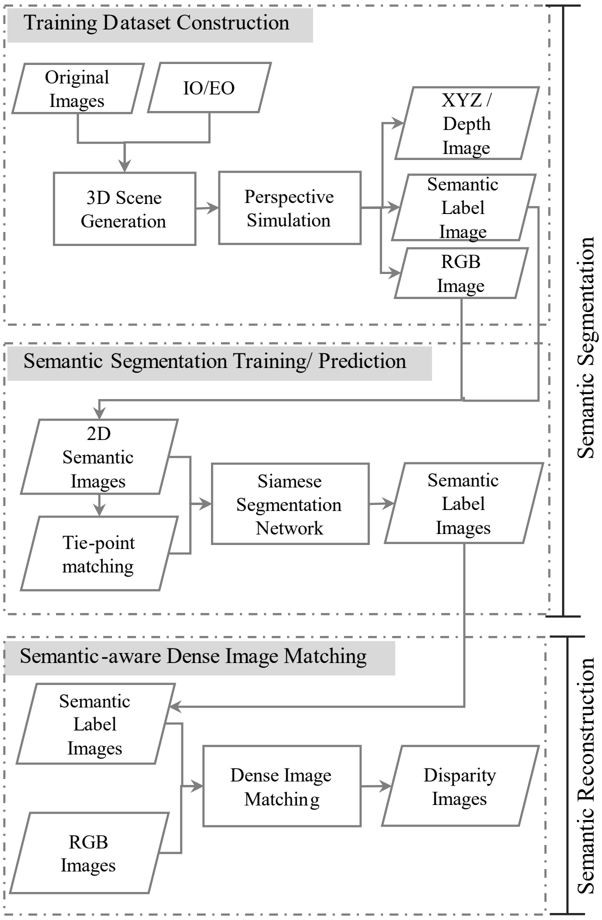

Figure 1. Overview of the proposed workflow.

$$\tilde{\boldsymbol{x}} = \boldsymbol{P}\tilde{\boldsymbol{X}} = \boldsymbol{P}\boldsymbol{E}\tilde{\boldsymbol{X}} = \boldsymbol{K}\boldsymbol{R}[\boldsymbol{I}|-\boldsymbol{C}]\tilde{\boldsymbol{X}} \tag{1}$$

where $\boldsymbol{E}$ is composed by the rotation $\boldsymbol{R}$ and the translation $\boldsymbol{C}$. $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{X}}$ denote the homogeneous coordinates of the 2D point $\boldsymbol{x}$ and its corresponding 3D point $\boldsymbol{X}$, respectively. $\boldsymbol{I}$ stands for the identity matrix. The camera matrix $\boldsymbol{K}$ could be further decomposed to describe its relationship with the focal length $f$ and the principal point $(c_x, c_y)$.

$$\boldsymbol{K} = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

To simulate this real imaging process in the computer, the camera matrix is further enriched with the field of view (FoV), near and far viewpoints, to define a view frustum, thus eliminating the outside content. Through this pipeline, a series of virtual images could be rendered given all the above information, as shown in Figure 2. With the knowledge about both the 3D model and the camera, each pixel could be further enriched with the normalized viewing depth of each pixel and its 3D coordinates $(X,\ Y,\ Z)$. By imposing the same camera on the original and the semantic-masked mesh model, the aligned RGB and the semantic images could be acquired simultaneously. The semantic-masked images are then transformed into label images according to the color of the semantic mask. Even if the original labels are still labeled by humans, the automatic simulation algorithm surmounted the problem of a small number of images and significantly reduced labor costs. Furthermore, the dataset is versatile and could assist all kinds of 2D/ 3D vision tasks.

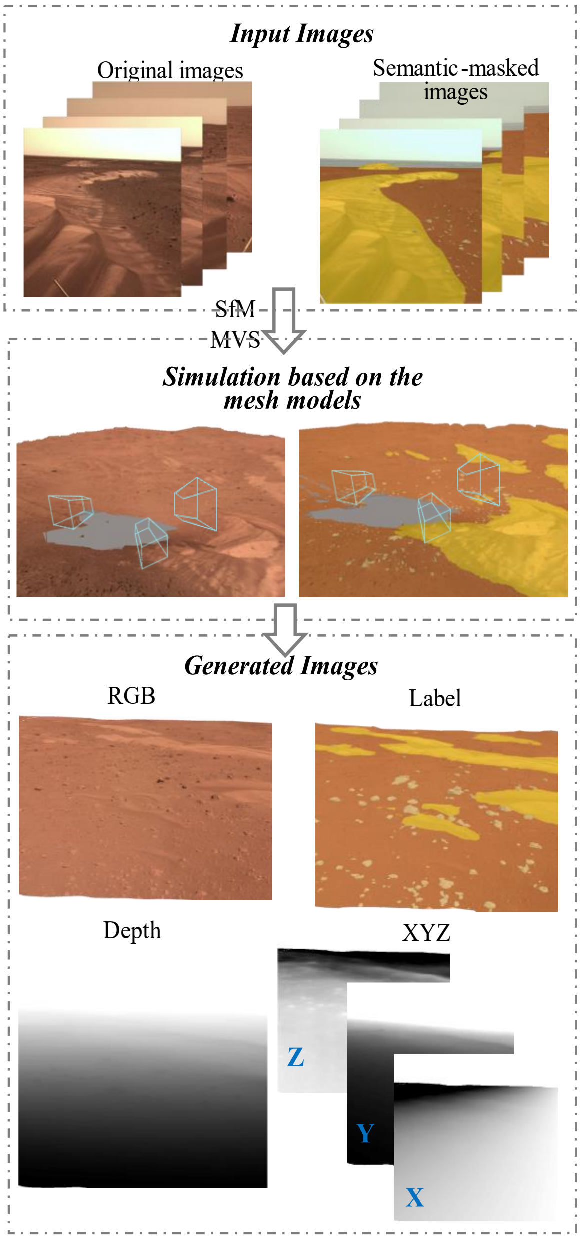

Figure 2. The illustration of the semi-automatic dataset construction.

### 2.3 Siamese Transformer for Semantic Segmentation

Recently, transformers and their variants have been commonly used for semantic segmentation tasks. We propose a new variant of the transformer for multi-view semantic segmentation, as illustrated in Figure 2. The backbone of the network is the state-of-the-art Swin-Transformer structure (Liu et al., 2021), which incorporates the merits of both convolutional-based and transformer-based networks and thus possesses the ability to consider in both local and global fashion.

Given the existence of overlapping regions in the available planetary images, it is natural to consider leveraging this constraint to supervise the segmentation. An intuitive way is to feed the original and transformed images to the neural network and transform the retrieved labels. The contrastive loss could then be established by measuring the similarities between these segmentation results. However, this strategy could not satisfy any two images sharing overlapping observations, whose pixel-wise transformation is hard to obtain. And wrapping an image during the training consumes a large amount of memory. The tie-points are hence introduced to find the corresponding points between the images. Instead of conducting the matching on original images directly, the images are transformed into the semantic masked to guide the tie-points distributed on the certain class and filter some inevitable wrong matches. To further involve more images for constraint in the training stage, the input image pairs are designed. For each image, the overlapped images inside the dataset are randomly chosen to form the counterpart image, as shown in Figure 3.

The loss function $\mathcal{L}_{all}$ is thus comprised of two parts. While the first part $\mathcal{L}_{Label}$ examines the cross-entropy loss between the images and the supervised labels, the second part $\mathcal{L}_{corre}$ punishes the inconsistent segmentation between the input images. After several warm-up epochs, the weight of the first part is expected to decrease and make images supervise each other to mitigate the not thoroughly labeled issue due to the complexity of the landforms. It is worth noting that as some unlabelled landforms may be retrieved during the training, the tie-points should be calculated based on the union of the prediction and the labeled masks.

$$\mathcal{L}_{all} = \mathcal{L}_{Label} + \mathcal{L}_{corre} \tag{3}$$

With respect to the prediction, the network could function in either a single- or multi-image version, as the two inputs share the same segmentation network.

### 2.4 Semantic-aware Dense Image Matching

Before deep learning based method, texture-aware dense image matching was attempted by utilizing the boundaries extracted automatically by the Sobel or Canny operator (Hu et al., 2016; Rothermel et al., 2012). Still, this algorithm fails to retrieve reasonable boundaries when it comes to the textureless planetary images, and severely suffers from dashed edges and noisy problems, as shown in Figure 4. While the boundary of the segments is not only continuous but also meaningful, the canny edges lead to unreasonable noises. Moreover, the texture-aware algorithm defines the texture by the gradient and the standard deviation of the intensities and adjusts the parameters accordingly. Due to the substantial differences among the camera sensors, the defined texture may not be ubiquitous enough. Fortunately, these two issues could be tackled properly by gauging the acquired consistent semantic labels.

The previous texture-aware algorithm is first extended to a semantic version. Specifically, instead of adapting the involving parameters according to the metrics of the intensities, they are now fine-tuned by the semantic labels. Furthermore, the feature descriptor of each pixel could be established considering the trait of each semantic class and the distance to the semantic boundaries. Semantic label similarity is measured and aggregated with the RGB's difference to serve as the joint descriptor. As for the perception, which is also decisive for the descriptor (Wang et al., 2022), the minimum region size is predefined for each semantic class. With the knowledge that insufficient texture requires a larger perception to construct a distinctive descriptor and small ones are preferred by the discontinuities, the minimum size is grown adaptively. The improved features and parameters are then injected into the conventional pipeline to calculate the pixel-wise disparity image.

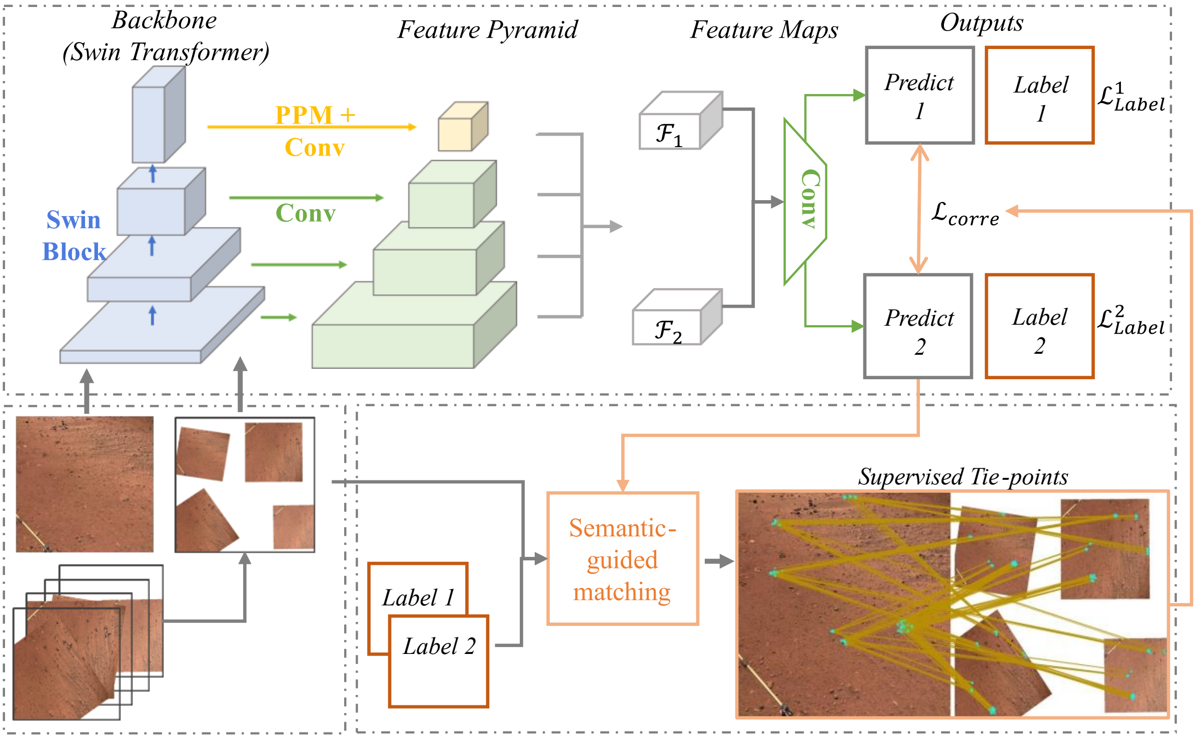

Figure 3. Illustration of the architecture of the Siamese Swin-transformer.

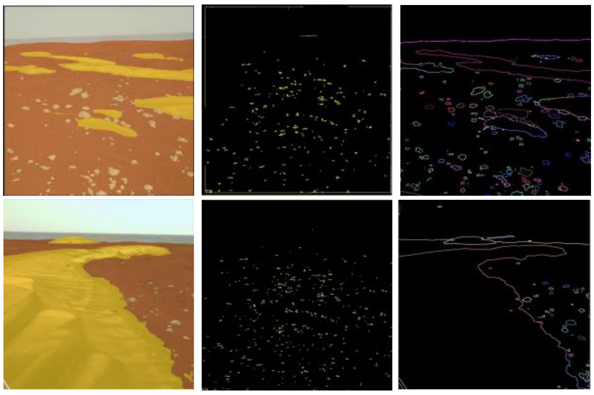

Figure 4. Illustration of the superiority of the semantic edge. The first column shows the original semantic-masked image, and the second and the third column represent the edges retrieved by Sobel/Canny operator and the semantic segmentation, respectively.

## 3. EXPERIMENTAL EVALUATION

### 3.1 Dataset Description

In this paper, the image dataset collected at the Zhurong landing site (Wu et al., 2021; Wu et al., 2022) on Mars is leveraged to evaluate the proposed approach. The images were collected by the Navigation and Terrain Camera (NaTeCam) onboard the Zhurong rover, comprising 2048 × 2048 pixels. Label files, recording the shooting time, position to the lander, and IO/EO parameters of each image, were also provided. The images within the same rover station are 360° panorama observations of the surroundings, which possess a nearly 30° perspective angle difference from the neighbor images.

Nine representative classes are defined, namely, soil, rock, sand, crater, shadow, wheel track, rover, far side, and other mechanism material (Rothrock et al., 2016), as shown in Figure 5. ~500 original images were carefully labeled by human labor, dated from 18th May 2021 to 14th March 2021.

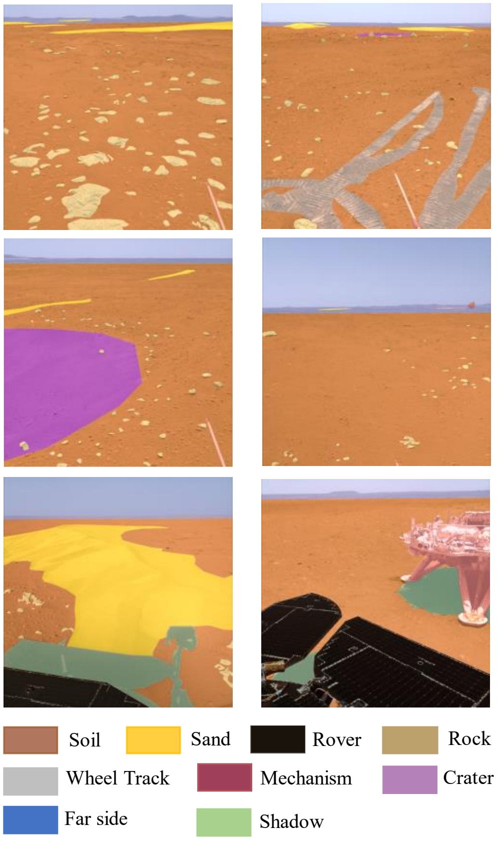

Figure 5. Illustration of the labeled semantic classes based on six representative images.

## 3.2 Experiment Results

### 3.2.1 Dataset Construction Results

The 3D mesh models are first generated through SfM and MVS pipeline. However, holes and over-interpolation may occur due to the inevitable occlusion problem resulting from the perspective of the rover. The examples of these situations are visualized in Figure 6. The virtual cameras are thus defined on the basis of the original cameras to avoid these defects. Empirically, the distance between the position of the virtual and the original camera should be within 8 meters in the direction away from the rover. The rotation and the IO parameters could be more flexible with a favorable position. Even the algorithm is limited by the quality of the 3D mesh, the amount of the images is enriched 20 times to ~10 K in a 3D manner. It is worth noting that the number of virtual cameras for one 3D model is not a constant number. The semantic labels are also considered to balance the sample of each class, and the amount is hence adapted automatically. Considering the memory of the GPU, the images are then cropped into patches of 512 × 512 pixels through three levels of scale pyramids.

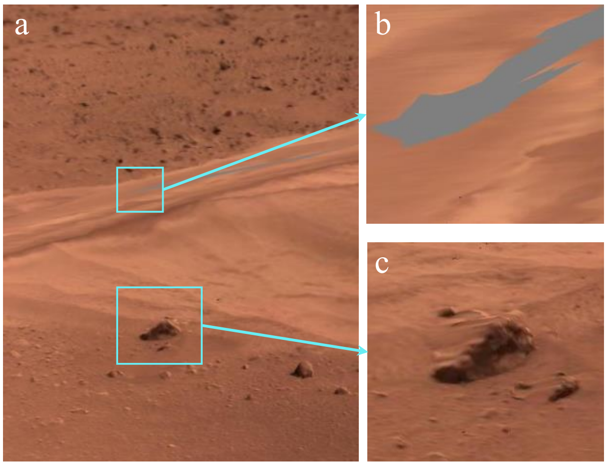

Figure 6. Illustration of the defects of the 3D mesh model.

### 3.2.2 Semantic Segmentation Results

The training was implemented with the PyTorch framework (Paszke et al., 2019) on a single NVIDIA GTX 3090 GPU. AdamW optimizer (Loshchilov and Hutter, 2017) was used for faster convergence. Beginning with the publicly available pretrained tiny version of the Swin-transformer (Swin-T) (Liu et al., 2021), the training was processed 80 epochs, with an initial $3e^{-4}$ learning rate. With respect to the tie-points, the state-of-the-art SuperGlue algorithm (Sarlin et al., 2020) is utilized for two-fold reasons. Firstly, it shows the superior capability to retrieve abundant tie-points even for textureless image pairs suffering from large perspective variations. And it is a GPU-based algorithm whose efficiency is guaranteed.

The representative results are shown in Figure 7, comparing our results with the manual labels and the results yielded by the retrained Swin-T based on our dataset. Intuitively, the transfer learned segmentation results are aligned well with the manually labeled ones, and some small rocks missed by humans are also retrieved. The large regions are all segmented to the correct semantic class, even for the wheel track and crater class suffering from the insufficient issue. This is mainly attributed to the semi-automatic dataset construction approach, which renders more simulated images for these seldomly observed landforms. It is worth noting that the crater here is not exactly the same as the ones defined in the satellite images (Wang and Wu, 2019), which is just a depressed area in the terrain. However, these results still suffer from wrong and incomplete segmentations. Benefiting from the self-supervision strategy, the prediction results generated from our approach are typically better than the traditional Swin-T model in the aspect of the accuracy and the details of the segments. While the incorrect segmentations are all solved, more rocks could be segmented as well. The complete detection of the rock leads to a comprehensive understanding of the terrain, which not only facilitates the rover to decide the path intelligently, but also provides more reliable data support for the following scientific analysis. The results also indicate that the over-fitting issue of the neural network could be effectively avoided by the cross-check between the overlapping images.

To quantitively analyze the results, the three commonly-used indexes, namely, mean intersection over union (MIoU), mean pixel accuracy (MPA), and frequency weighted accuracy (FWAcc), are analyzed. As suggested by Table 1, both networks possess favorable accuracy, while our approach is a little bit better in terms of the mIoU. However, these metrics are not that rigorous due to the aforementioned label issues.

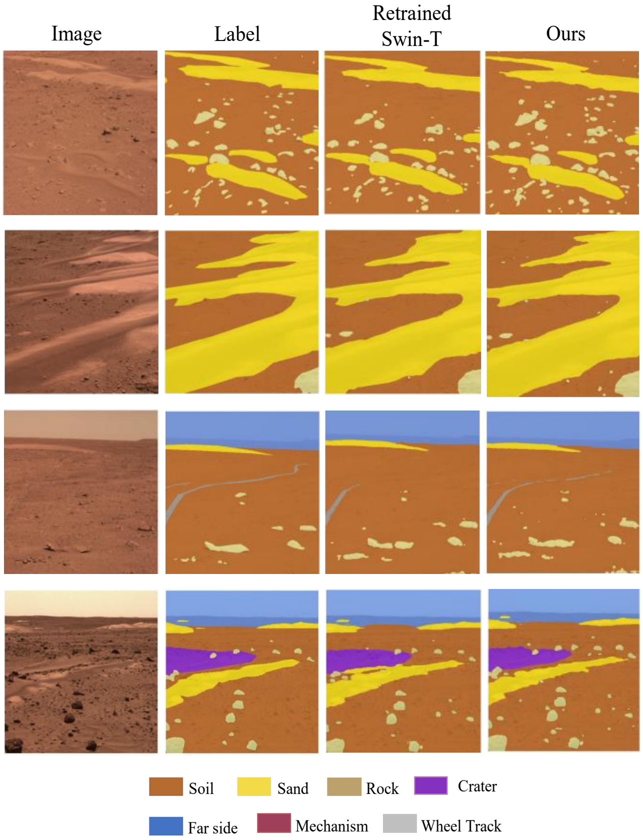

Figure 7. A representative semantic segmentation result, indicating the superiority of the proposed Siamese-like Swin-T. The segmented results are overlaid on the original images. The first two columns present the original images and the manually labeled images. The last two columns are the results generated from the Swin-T after transfer learning and our approach.

| Experiment | mIoU | MPA | FWAcc |
|---|---|---|---|
| Swin-T | 86.08 | 94.33 | 96.03 |
| Ours | 88.25 | 95.78 | 96.91 |

Table 1. Quantitative analysis of the semantic segmentation results.

Furthermore, the evaluation of the transformed images is also performed to test the transform-invariant strength of our approach. Three experiments involving all the translation, rotation, scale, and real-world transformations are exhibited in Figure 8. Two highlighted regions are marked by the white and blue ellipse, respectively. Despite the incomplete or incorrect retrieval problem, the semantic labels in the ellipses calculated by the Swin-T are not strictly aligned across the experiments. While the Siamese Swin-T tends to maintain a similar pattern even with more segmented rocks.

### 3.2.3 Dense Matching Results

Figure 9 shows the dense image matching results based on the semantic segmentation results. The proposed semantic-aware dense image matching is compared with the original and the texture-aware ones. Generally, the original dense image matching could give reasonable disparity results. But the speckle effect in the sand region is obvious, and the boundary region between the soil and sand region is facing severe discontinue and incorrect issues. Even the texture-aware algorithm slightly improves the results with narrowed no data region, the incorrect disparity still exists, especially for the second experiment. By introducing the correct retrieved semantic information, the disparities are strictly aligned with the semantic boundary with further closed gaps, and the speckle effects in the sand region are mitigated.

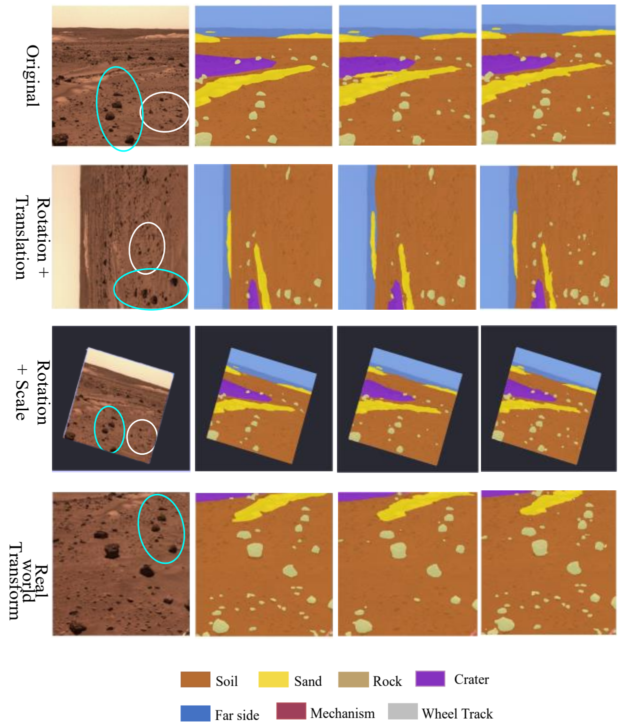

Figure 8. Illustration of the invariant ability of the Siamese Swin-T. The segmented results are overlaid on the original images. The first two columns present the original images and the manually labeled images. The last two columns are the results generated from the Swin-T after transfer learning and our approach.

## 4. CONCLUSIONS

In this paper, we present an effective approach to generate a large and versatile training dataset semi-automatically by introducing the original 3D information. A new variant of the Swin transformer is proposed targeting multi-view semantic segmentation, which fully exploits the overlapping information to supervise the segmentation in a self-supervised manner. Semantic-aware dense image matching is hence performed, incorporating the semantic segments to guide the adaptively matching. The performance of the proposed approach is validated with the real Martian dataset at the Zhurong landing site, indicating that abundant training data could be generated and further guarantee the accuracy of the multi-view semantic segmentation and reconstruction.

Our future efforts will focus on incorporating the 3D information in the dataset to improve the semantic results, and exploring more elegant approaches to achieve semantic 3D reconstruction.

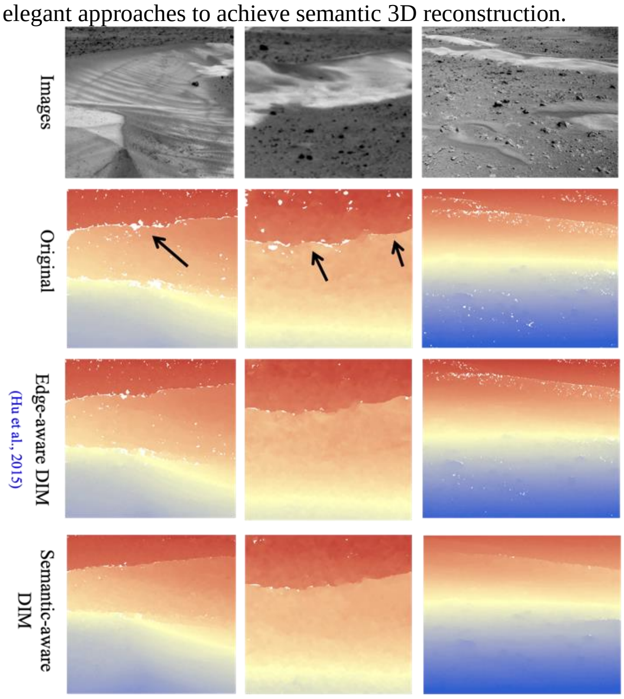

Figure 9. The evaluation of the proposed semantic-aware dense image matching algorithm. The three columns are corresponding to the three representative regions.

## ACKNOWLEDGMENTS

This work was supported by grants from the Research Grants Council of Hong Kong (RIF Project No: R5043-19, Project No: 15215822, Project No: PolyU 15210520, Project No: PolyU 15219821).

## REFERENCES

Agarwal, S., Snavely, N., Simon, I., Seitz, S.M., Szeliski, R., 2009. Building Rome in a Day. Proceedings of the IEEE International Conference on Computer Vision, pp.72-79.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Hu, H., Chen, C., Wu, B., Yang, X., Zhu, Q., Ding, Y., 2016. Texture-aware dense image matching using ternary census transform. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences. III-3, 59-66.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., 2023. Segment anything. arXiv preprint arXiv:2304.02643.

Li, Z., Wu, B., Liu, W.C., Chen, Z., 2022. Integrated photogrammetric and photoclinometric processing of multiple HRSC images for pixelwise 3-D mapping on Mars. IEEE Transactions on Geoscience and Remote Sensing 60, 1-13.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 10012-10022.

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Ma, C., Li, Y., Xiao, Z., Zhang, W., Mo, L., Li, A., 2023. SimMars6K. Zenodo.

Naseer, T., Oliveira, G.L., Brox, T., Burgard, W., 2017. Semantics-aware visual localization under challenging perceptual conditions. 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 2614-2620.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32.

Rothermel, M., Wenzel, K., Fritsch, D., Haala, N., 2012. SURE: Photogrammetric surface reconstruction from imagery, Proceedings LowCost 3D Workshop, Berlin.

Rothrock, B., Kennedy, R., Cunningham, C., Papon, J., Heverly, M., Ono, M., 2016. SPOC: Deep learning-based terrain classification for mars rover missions. American Institute of Aeronautics and Astronautics (AIAA) SPACE Forum 2016, p. 5539.

Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. SuperGlue: Learning feature matching with graph neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4937-4946.

Schonberger, J.L., Frahm, J.M., 2016. Structure-from-Motion revisited. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.4104-4113.

Schwenzer, S., Woods, M., Karachalios, S., Phan, N., Joudrier, L., 2019. Labelmars: Creating an extremely large martian image dataset through machine learning. 50th Annual Lunar and Planetary Science Conference, pp. 1970.

Swan, R.M., Atha, D., Leopold, H.A., Gildner, M., Oij, S., Chiu, C., Ono, M., 2021. AI4MARS: A dataset for terrain-aware autonomous driving on mars. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1982-1991.

Vu, H.H., Labatut, P., Pons, J.P., Keriven, R., 2012. High accuracy and visibility-consistent dense multiview stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 889-901.

Wan, Y., Li, Y., You, Y., Guo, C., Lijin, 2021. Semantic Dense Reconstruction with Consistent Scene Segments. arXiv pre-print server.

Wang, Y., Wu, B., 2019. Active machine learning approach for crater detection from planetary imagery and digital elevation models. IEEE Transactions on Geoscience and Remote Sensing 57, 5777-5789.

Wang, Y., Gu, M., Zhu, Y., Chen, G., Xu, Z., Guo, Y., 2022. Improvement of AD-Census algorithm based on stereo vision. Sensors 22, pp. 9050.

Wu, B., Dong, J., Wang, Y., Rao, W., Sun, Z., Li, Z., Tan, Z., Chen, Z., Wang, C., Liu, W.C., Chen, L., Zhu, J., Li, H., 2022. Landing site selection and characterization of Tianwen-1 (Zhurong Rover) on Mars. Journal of Geophysical Research: Planets 127, e2021JE007137.

Zhao, W.F., Persello, C., Stein, A., 2023. Semantic-aware unsupervised domain adaptation for height estimation from single-view aerial images. ISPRS Journal of Photogrammetry and Remote Sensing 196, 372-385.

Zheng, T., Zhang, G., Han, L., Xu, L., Fang, L., 2022. BuildingFusion: Semantic-aware structural building-scale 3D reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 2328-2345.