Supplementary Material of "MTSNet: Joint Feature Adaptation and Enhancement for Text-Guided Multi-view Martian Terrain Segmentation"

Anonymous Authors

Appendix A. Training strategy for image-text pair training data

In our approach, we propose to use terrain-specific text as a segmentation prompt to enhance the efficiency and feasibility of using the model in practical scenarios. When using text as a prompt for Martian terrain segmentation, we need to handle situations where the used text, such as "big rock", does not correspond to the actual content of the Mars surface image being processed, in which case the model's output would be a blank mask. However, if the model is trained solely on image-label (text) pairs that contain valid segmentation regions, it is likely to struggle with such cases. To address this issue, we adopt the following training strategy during the training process: assuming there are M possible Martian terrain categories, for each Mars surface image, even if it does not contain all terrain categories, we generate M sets of corresponding image-text pairs, and set masks to blank for terrain categories that are not present in the image. This strategy significantly improves our model's ability to generate correct masks even if the target terrain is not actually present in the image.

Appendix B. Loss function of model training

To address the potential severe data imbalance issue caused by generating M image-text pairs for each image during training, which can lead to a large number of blank labels, we employ the focal loss for training. The focal loss is defined as follows

$$L_{focal} = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{HW} L_{ij} \tag{1}$$

where

$$L_{ij} = \begin{cases} -\alpha\,(1-p)^{\gamma}\,log(p), & \text{if } y = 1 \\ -(1-\alpha)(p)^{\gamma}log(1-p), & \text{if } y = 0 \end{cases} \tag{2}$$

here, L_{ij} represents the loss associated with pixel j in image i , p denotes the predicted probability of being 1, and y represents the true pixel label. B denotes the batch size, H denotes the height of the mask, and W denotes the width of the mask. The hyperparameters α and γ need to be determined through a validation process.

Appendix C. Additional qualitative analysis

Figure 1 provides supplementary qualitative analysis of Martian terrains segmentation on the AI4Mars [5] dataset. For other models that do not rely on external text prompts (UNet [4] and UNext [6]), the models output segmentation maps for each class. To perform the final class-level mask comparison, we follow this formula:

$$mask_i = \begin{cases} 1, & \text{argmax}\,(O(C)) = i \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

here, i represents the class index, and O represents the output of the segmentation model. Assuming the shape of O is $C \times H \times W$, where C denotes the number of classes, H represents the height of the mask, and W represents the width. It is particularly noteworthy that when a specific Martian terrain is not present in the image, our method can accurately output a blank mask (as shown in the third row of Figure 1), however, other methods give false positive segmentation results.

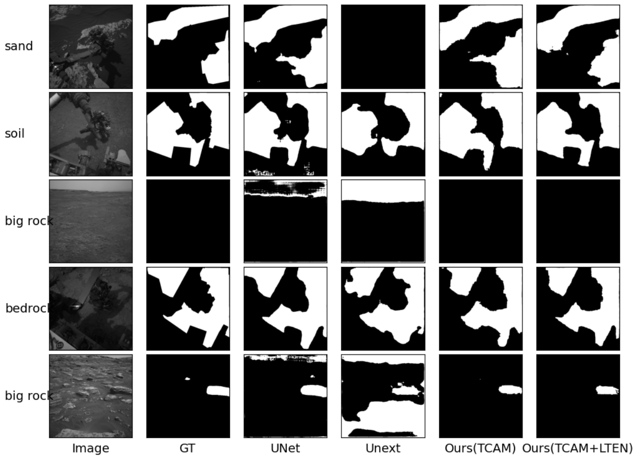

Figure 1: Qualitative Performance on the AI4Mars Dataset (supplementary).

On the ConeQuest [3] dataset, to better observe the performance of our model in different regions, we additionally provide separate

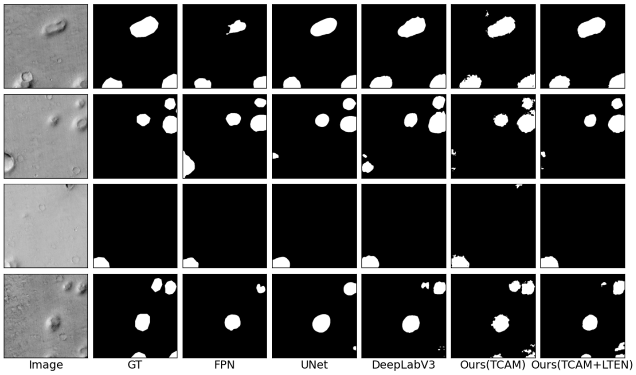

Figure 2: Qualitative Performance on the ConeQuest in the Isidis Planitia (IP) region (supplementary).

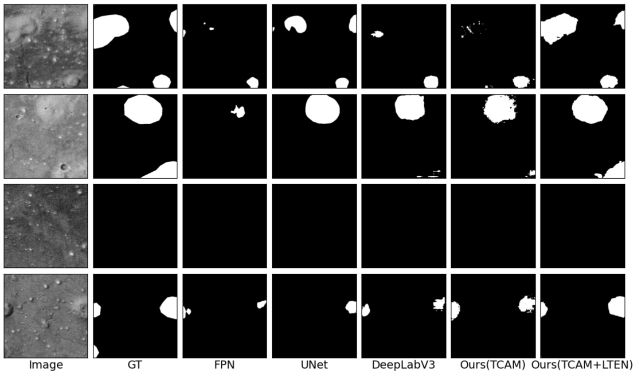

Figure 3: Qualitative Performance on the ConeQuest in the Acidalia Planitia (AP) region (supplementary).

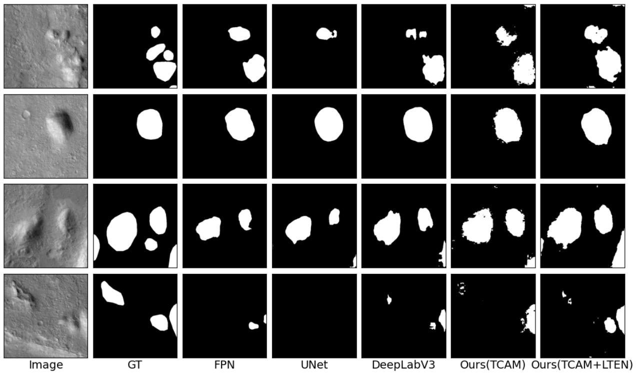

Figure 4: Qualitative Performance on the ConeQuest in the Hypanis (HP) region (supplementary).

visual comparisons for each region, as shown in Figure 2, Figure 3, and Figure 4, respectively. Here, we compare our model with the benchmarked methods including UNet [4], DeepLabV3 [1] and FPN [2].

REFERENCES

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

[2] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[3] Mirali Purohit, Jacob Adler, and Hannah Kerner. 2024. ConeQuest: A Benchmark for Cone Segmentation on Mars. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 6026–6035.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 234–241.

[5] R Michael Swan, Deegan Atha, Henry A Leopold, Matthew Gildner, Stephanie Oij, Cindy Chiu, and Masahiro Ono. 2021. Ai4mars: A dataset for terrain-aware autonomous driving on mars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1982–1991.

[6] Jeya Maria Jose Valanarasu and Vishal M Patel. 2022. Unext: Mlp-based rapid medical image segmentation network. In *International conference on medical image computing and computer-assisted intervention*. Springer, 23–33.