ELunarDTMNet: Efficient reconstruction of high-resolution lunar DTM from single-view orbiter images

Hao Chen, *Student Member, IEEE,* Philipp Gläser, Xuanyu Hu, Konrad Willner, Yongjie Zheng, *Student Member, IEEE,* Friedrich Damme, Lorenzo Bruzzone, *Fellow, IEEE* and Jürgen Oberst

***Abstract*—High-resolution Digital Terrain Models (DTMs) are critical for supporting planetary exploration missions and advancing scientific research. Recently, Deep Learning (DL) techniques have been applied to reconstruct high-resolution DTMs from single-view orbiter optical images, particularly for the Moon. However, DL-based methods face challenges in retrieving high-quality multi-scale topographic features, especially in regions with irregular terrains or significant relief. Additionally, their generalization capability across diverse datasets is rarely evaluated. In this paper, we propose an efficient DL-based single-view method with a coarse-resolution DTM as a constraint for high-quality lunar DTM reconstruction, named ELunarDTMNet. This approach introduces a hierarchical transformer-based backbone with a residual-connected mechanism, specifically designed to capture and integrate multi-scale features from single-view lunar images, thereby enhancing prediction accuracy. Meanwhile, given the diverse and complex surface relief, new elevation normalization strategies are proposed to preserve terrain feature contrast while accommodating different elevation distributions. Our method performs well on diverse lunar landscapes with various topographic features and elevation changes. It outperforms existing DL-based methods in accuracy and detail, effectively addressing their encountered challenges. Moreover, the proposed method achieves effective resolutions similar to those of the Shape-From-Shading technique for subtle-scale terrain retrieval, but with enhanced elevation accuracy, illumination robustness, and approximately 850 times faster processing speed. Trained with the Lunar Reconnaissance Orbiter (LRO) Narrow Angle Camera (NAC) images, our model shows superior performance on other high-resolution lunar orbiter images, such as Chang'E-2 imagery.**

***Index Terms*—Deep learning, High-resolution, Lunar DTM reconstruction, Lunar Reconnaissance Orbiter Narrow Angle Camera, Single-view.**

This work was supported by the China Scholarship Council (CSC) under Grant 202006260004. (*Corresponding authors: Hao Chen, Yongjie Zheng*).

Hao Chen, Philipp Gläser, and Friedrich Damme are with the Institute of Geodesy and Geoinformation Science, Technische Universität Berlin, Berlin 10553, Germany (e-mail: hao.chen.2@campus.tu-berlin.de; philipp.glaeser@tu-berlin.de; friedrich.damme@tu-berlin.de).

Xuanyu Hu is with the Institute of Space Technology & Space Applications (LRT 9.1), University of the Bundeswehr Munich, Neubiberg 85577, Germany (e-mail: xuanyuhu@gmail.com).

Konrad Willner is with the Institute of Planetary Research, German Aerospace Center (DLR), Berlin 12489, Germany (e-mail: konrad.willner@dlr.de).

Lorenzo Bruzzone and Yongjie Zheng are with the Department of Information Engineering and Computer Science, University of Trento, Trento 38123, Italy (e-mail: lorenzo.bruzzone@unitn.it; yongjie.zheng@unitn.it).

Jürgen Oberst is with the Institute of Geodesy and Geoinformation Science, Technische Universität Berlin, Berlin 10553, Germany, and also with the Institute of Planetary Research, German Aerospace Center (DLR), Berlin 12489, Germany (e-mail: juergen.oberst@tu-berlin.de).

I. Introduction

HIGH-RESOLUTION Digital Terrain Models (DTMs) derived from optical imaging data of the lunar surface acquired by orbiting spacecraft are critical for interpreting the Moon. They serve as essential tools for scientific analysis of surface morphology, geology, and resource [1]–[3], as well as for supporting engineering applications such as mission planning [4], and high-precision landing site selection [5], [6].

Previously, high-resolution DTMs of the planetary surface have primarily been produced using Stereo-PhotoGrammetry (SPG) [7] and Shape-From-Shading (SFS) [8] techniques. SPG is a well-established method for accurately determining topographic models, which relies on geometric principles involving images taken from stereo or multiple viewpoints [9]–[11]. Despite its reliability, the photogrammetric matching process of stereo pairs may produce artifacts [9], [12], [13]. The Narrow Angle Camera (NAC) onboard the Lunar Reconnaissance Orbiter (LRO) has captured the highest resolution and quality orbital images of the lunar surface to date [14]. While these images offer almost complete coverage of the lunar surface, suitable stereo pairs for SPG terrain modeling are still limited to small areas, as NAC does not have a built-in stereo capability [14]. In contrast, SFS can deduce shape information by estimating slope from a single image, with improved performance when constrained by a coarse-resolution DTM [15]. The technique can eliminate stereo artifacts and achieve DTMs with pixel-level resolution. However, SFS methods can be time-consuming for iterations to converge, and they require a priori knowledge on the reflection parameters of the surface [16]. They also exhibit reduced accuracy in the cross-sun direction [8], [17].

In recent years, the application of Deep Learning (DL) techniques to the reconstruction of high-resolution planetary DTMs from single-view orbiter images has shown substantial progress, as a valuable complement to SPG and SFS [18]–[20]. They utilize existing regional high-quality mapping products, such as SPG-derived high-resolution DTMs and Ortho-Rectified Images (ORIs), to train the model and learn the relationship between images and DTMs [21]. In certain studies, coarse-resolution DTMs are incorporated as model input to optimize model convergence and improve performance [17]. The approaches typically rely on the encoder-decoder framework. The encoder serves as the feature extractor to extract multi-scale representations from fine to coarse

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3501153

resolution, while the decoder is responsible for recovering the spatial resolution and estimating the pixel-level resolution DTM. These methods have demonstrated their effectiveness in automatically reconstructing 3D topography without the requirement of prior knowledge of camera models, orbiter state, or surface properties, offering advantageous processing speed [21], [22].

While various DL-based strategies have been proposed to improve the model performance, the dilemma of achieving high-quality DTM reconstruction persists for all strategies, especially for terrain features with irregular shapes and regions with significant elevation relief [17], [23]. Despite the encoder-decoder framework maintaining the size of the predicted DTM consistent with the input image, DL methods still encounter limitations in capturing subtle and fine-scale details, often resulting in lower effective resolution compared to SFS methods [17], [21]. The medium- and large-scale topographical features exhibit some discrepancies compared to the topographic models derived from SPG methods [17], [24]. Besides, it is noteworthy that some strategies, such as the co-alignment technique used to align predicted DTMs with coarse-resolution reference DTMs for improved elevation accuracy [17], may require additional time. This can significantly exceed the time needed for predicting the DTM itself [17], [25].

In this paper, we present an efficient DL-based approach to lunar DTM reconstruction with a single high-resolution optical image and a coarse-resolution DTM as input (called ELunarDTMNet), demonstrating good performance even in regions with irregular terrain features and significant relief. In contrast to the single-branch encoder module, which processes different inputs within a unified backbone, the dual-branch encoder module, as used in DLunarDTMNet [17], allows for tailored backbones and strategies based on the specific characteristics of images and DTMs. This results in improved accuracy for the reconstructed DTM mosaics, as shown in Fig. 1. Therefore, the proposed method consists of three essential components: a dual-branch encoder module, a fusion module, and a decoder module, trained using a hybrid loss function. Our method achieves high-quality retrieval of multi-scale terrain features while maintaining fast DTM mosaic processing speed, with strong generalization capability across diverse types of high-resolution lunar optical imagery. The contributions of the proposed method can be summarized as follows:

1) ***Multi-scale feature extraction and fusion***: The lunar surface displays various topographic features with varying scales. Unlike previous approaches that typically utilized Convolutional Neural Network (CNN)-based backbones with hierarchical architectures to extract these features from images, we incorporate a hierarchical transformer-based backbone into the image branch of the dual-encoder module to improve the ability to capture features across scales, from local to global. To the best of our knowledge, this represents the first application of a transformer-based backbone to lunar DTM reconstruction. Additionally, a novel residual-connected mechanism is devised to effectively integrate these multi-scale features in the decoder module to improve prediction accuracy.

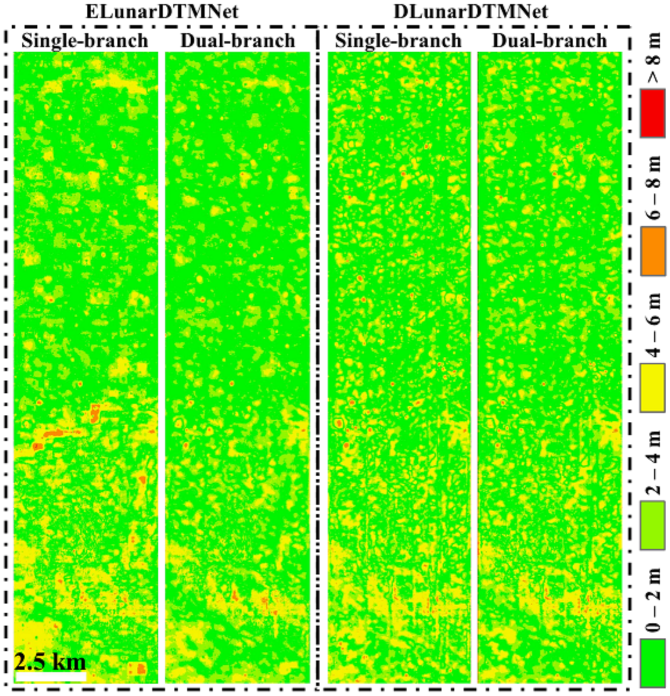


Fig. 1. Reconstruction errors of DTMs using single-branch or dual-branch encoder modules to process different inputs, based on the proposed ELunarDTMNet and DLunarDTMNet, illustrated with an example centered at (334.27°W, 27.96°N).

2) ***New normalization strategies for lunar surface elevations***: Considering the limitations of a narrow normalization range in preserving the contrast of terrain undulations, an elevation-statistics-based DTM normalization strategy is proposed to determine a suitable normalization range, which facilitates the preservation of elevation disparities, thereby contributing to a more faithful and detailed reconstruction result. Besides, the varied lunar topographies lead to diverse elevation distributions. To guide the network in learning useful information unaffected by variations, we propose a mean-normalized loss function to penalize vertical discrepancies based on a uniform mean scale, rather than relying on varying statistics.
3) ***Simplified DTM mosaic process***: DTM mosaicking is an essential step for generating large-scale DTMs. The proposed ELunarDTMNet generates high-resolution DTMs through end-to-end processing, eliminating the need for multiple networks to incrementally produce DTM mosaics from coarse to high resolution. Besides, our method can derive high-quality mosaics without relying on the time-consuming co-alignment technique to enhance elevation accuracy.

II. Related works

Over the past decades, SPG and SFS methods have been widely applied to planetary DTM reconstruction, demonstrating significant advancements [7]–[11], [16], [26]–[28]. This section will specifically focus on the recently developed DL-based single-view methods.

The Martian surface was the initial focal target for applying single-view DL methods to planetary DTM reconstruction. Chen et al. [18], [29] were pioneers in employing CNN architecture for the 3D reconstruction from single Context Camera (CTX) imagery. Their method yielded promising results, particularly in capturing topographic features on flat terrains. Tao et al. [22] adopted a multi-scale architecture (incorporating coarse-, intermediate-, and fine-scale) to estimate DTMs from the High Resolution Stereo Camera (HRSC) images, CTX images, and the High-Resolution Imaging Science Experiment (HiRISE) images. Tao et al. [25] introduced a coarse-to-fine strategy to sequentially reconstruct multi-scale DTMs, utilizing coarser-scale predicted DTMs as constraints to estimate the higher-scale DTMs. Cao et al. [30] evaluated the generalization of their method by using HiRISE images with a higher resolution than those in the training set. Chen et al. [21] applied the single-view DL method for estimating DTMs on the lunar surface. Their approach is distinctive in that it incorporates coarse-resolution DTMs in conjunction with images as input for model training, leading to an enhancement in the elevation accuracy. Nevertheless, the input images and DTMs exhibit significant disparities in terms of resolution and the types of surface information they represent. To alleviate this issue, LDEMGAN implemented a two-step approach, similar to the coarse-to-fine strategy in [25], utilizing two identical CNN networks for DTM prediction [23]. On the other hand, DLunarDTMNet devised a pure CNN-based dual-branch encoder module to independently process input images and DTMs [17]. This approach allows for the customization of a suitable feature extraction architecture for both elements within a single network, thereby eliminating the need for multiple networks. DLunarDTMNet [17] and MADNet [24] were applied to generate large-area, high-resolution DTMs of the lunar surface. Muller et al. [31] applied the method proposed by Tao et al. [25] to generate DTMs for candidate landing sites in the lunar South Pole region. However, their results indicate that further refinement may still be required to achieve high-quality retrieval of features ranging from subtle small-scale details (similar to the SFS method) to large-scale components (comparable to the SPG method).

Training a network requires a well-defined loss function to guide the learning process. Most single-view DTM reconstruction methods employed a hybrid loss function that incorporates various types of loss terms [23]. These terms typically include penalizing the differences between predicted DTMs and their ground truth counterparts in the vertical domain, as well as addressing the discrepancies of high-frequency details in the horizontal domain [17]. The mean absolute error (MAE), mean squared error (MSE), or their variants, are commonly used to quantify and penalize vertical discrepancies [21]–[23]. Nevertheless, the overall elevation of the planetary topography may vary significantly from one area to another. Further, the amplitude and pattern of the topographic variations may also be locally distinct. These variations may pose challenges for MAE or MSE-related loss functions in guiding the training process to effectively learn and adapt to the diverse topographic features.

Besides, single-view image-based DTM reconstruction lacks absolute depth information. SFS methods integrate coarse-resolution DTMs, e.g., derived from SPG and laser altimetry, to provide the absolute depth information [8], [15]. DL methods typically process DTMs by normalizing them to a fixed dimensionless scale, such as [0, 1] using the max-min normalization strategy [19], [22], [23]. This process enhances the effectiveness and fast convergence of the models. However, the unprocessed DTMs often exhibit elevation differences greater than this normalized scale. A narrow normalization range would limit the contrast of the terrain undulations, which is detrimental to the recovery of terrain details, particularly those at subtle small-scale levels. Furthermore, to recover the predicted DTMs from the normalized scale to their original scale with absolute depth information, DLunarDTMNet and MADNet utilized the co-alignment technique, aligning the predicted DTMs with the coarse-resolution reference [17], [24]. While this technique can improve elevation accuracy and yield better results in terrains with small elevational differences, its effectiveness diminishes for seamless DTM mosaic generation in areas with significant relief [17].

III. Method

Fig. 2 illustrates the framework of our proposed ELunarDTMNet for pixel-level resolution DTM reconstruction of the lunar surface. Our network utilizes a high-resolution optical image captured by an orbiter and a coarse-resolution DTM as input, with a photogrammetric-derived high-resolution DTM serving as the ground truth for training and validation. The process of DTM normalization and image pre-processing is detailed in Section III-A. The ELunarDTMNet consists of three main components: a dual-branch encoder module based on hierarchical transformer blocks for images and Convolutional blocks (Conv blocks) for DTMs, a fusion module, and a residual-connected decoder module, as described in Section III-B. In Section III-C, we introduce a hybrid loss function that incorporates terms targeting the errors in both the vertical and horizontal domains. Besides, the steps for deriving the large-sized DTM mosaics from the small-sized predicted DTM tiles are presented in Section III-D.

A. DTM normalization and image pre-processing

1) Elevation-statistics-based DTM normalization: The max-min normalization strategy might oversimplify the variations by stretching all values into a narrow range (Fig. 3a), potentially obscuring subtle, but crucial, topographic details with less pronounced elevation differences. Given the wide range of elevation differences among the samples in the training set, all exceeding 1 m (Fig. 3b), we propose an elevation-statistics-based normalization strategy to determine the normalization range. The mean value of the elevation differences within the training dataset provides a measure of the average tendency of topographic relief. Considering the substantial coverage of flat areas compared to regions with significant elevation differences, we derive the mean value based on the inverse of the elevation difference to determine

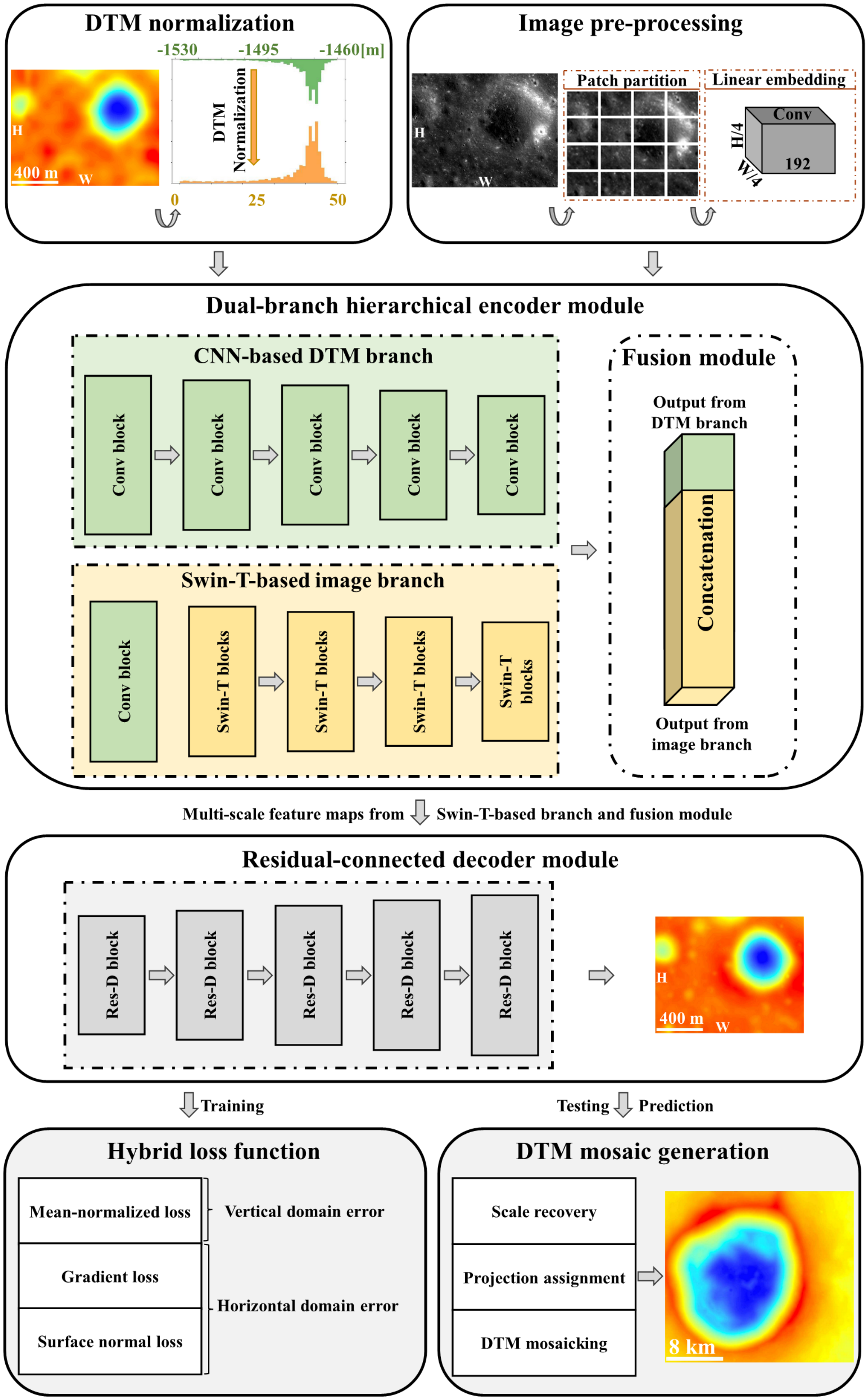


Fig. 2. The framework of our proposed ELunarDTMNet for pixel-level resolution DTM reconstruction. The CNN-based DTM branch is designed to process input DTMs, while the Swin-T-based image branch is designed to process input images. Swin-T: Swin Transformer, Res-D: Residual-connected Decoder, H: image height, W: image width.

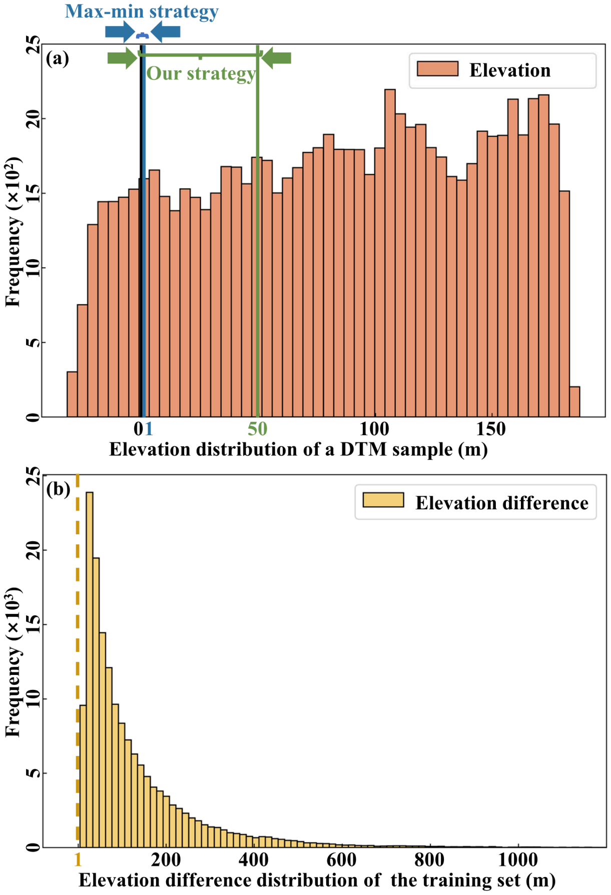


Fig. 3. DTM normalization strategy. (a) Elevation distribution of a DTM sample, with normalized range determined by the max-min strategy [0, 1] and our strategy [0, 50]. (b) Distribution of elevation differences among samples in the training set, all of which exceed 1 m.

the normalization range, which is equivalent to increasing the weight of flat areas in determining this value:

$$NR = \frac{n}{\left(\sum_{p=1}^{n} \frac{1}{\Delta Z_p}\right)}, \tag{1}$$

$$Z_{\text{norm},p} = NR \cdot \left(\frac{Z_p - \min(Z_p)}{\Delta Z_p}\right), \tag{2}$$

where n denotes the number of DTM samples in the dataset. Z_p refers to the original elevation value of sample p, ΔZ_p represents the elevation difference in sample p. NR is the normalized range. $Z_{\text{norm},p}$ represents the normalized value of sample p. By using the actual elevation difference to determine the normalization range, our approach can increase the sensitivity and representations of subtle terrain features.

2) Image pre-processing: To pre-process input images, we apply the max-min normalization strategy to scale the gray values to the range [0, 1], as described by DLunarDTMNet [17]. As presented in Section III-B1, one of the encoder branches in the network, designed specifically for image feature extraction, is primarily based on the Swin-T architecture. Instead of directly processing images as input, the Swin-T architecture employs a patch partition operation to split the input images into non-overlapping 4 × 4 patches [32]. These patches are subsequently concatenated into one volume. Following this, the linear embedding step utilizes a convolution operation to project this volume into a higher channel dimension [32]. In this study, we considered 192 channels; the resulting higher-dimension volume (H/4 × W/4 × 192) is then fed into the Swin-T blocks to extract multi-scale features, as shown in Fig. 4.

B. Network architecture

1) Dual-encoder module for input image and DTM: Given the different data types of high-resolution optical orbital images, which display fine textures of the surface variations by image intensity, and coarse-resolution DTMs, which provide direct high-precision (albeit relatively sparse) elevation measurements, this study adopts two separate encoder branches with different design model architectures as feature extractors. The images showcase a wide range of topographic features, from regular to irregular, at various scales, such as craters with diameters ranging from several meters to kilometers. Previous approaches to planetary DTM estimation relied on CNN-based backbones, such as ResNet [21], [33], and DenseNet [22], [34], for multi-scale feature extraction. The advent of transformer architectures in computer vision offers a promising alternative to overcome CNN limitations in capturing global information from the input image [35], [36]. Unlike CNNs, which process data locally and sequentially, the attention mechanism of transformers makes the network simultaneously attend to all positions when processing an image [37]. An improved hierarchical transformer, the so-called “Swin Transformer (Swin-T)”, demonstrates a notable proficiency in capturing features from a local to global scale [32], [38], aligning well with the multi-scale characteristic of lunar surface features.

We incorporate the Swin-T backbone into the image encoder branch to extract multi-scale terrain features. Swin-T considers four distinct multi-scale feature map sizes, similar to the five multi-scale feature map sizes in ResNet utilized by DLunarDTMNet [17]. In contrast to DLunarDTMNet, this study replaces the max-pooling operation and the subsequent ResNet blocks with Swin-T blocks to construct the encoder architecture, as illustrated in Fig. 4. Unlike the max-pooling (Fig. 4c-1) and convolution (Fig. 4c-2) operations used in ResNet-based backbone to reduce spatial dimension, Swin-T uses patch merging technique to select elements of the feature maps at positions with a spacing of 2 in the row and column directions, forming new patches. Then, all the patches are concatenated together, increasing the number of channels to four times, while reducing the width and height resolutions by half (Fig. 4c-3) [32]. The architecture of stage 1 in ResNet is preserved to extract local features from the input image. Its output is directly passed to the decoder module, enhancing the model’s capability to recover local features, as described in Section III-B2.

The image-only architecture performs well in capturing 2D shapes of terrain features but lacks accuracy in elevation for the DTM estimation task. Inspired by the SFS [15] and DL [17], [23], [29] methods, which leverage coarse-resolution

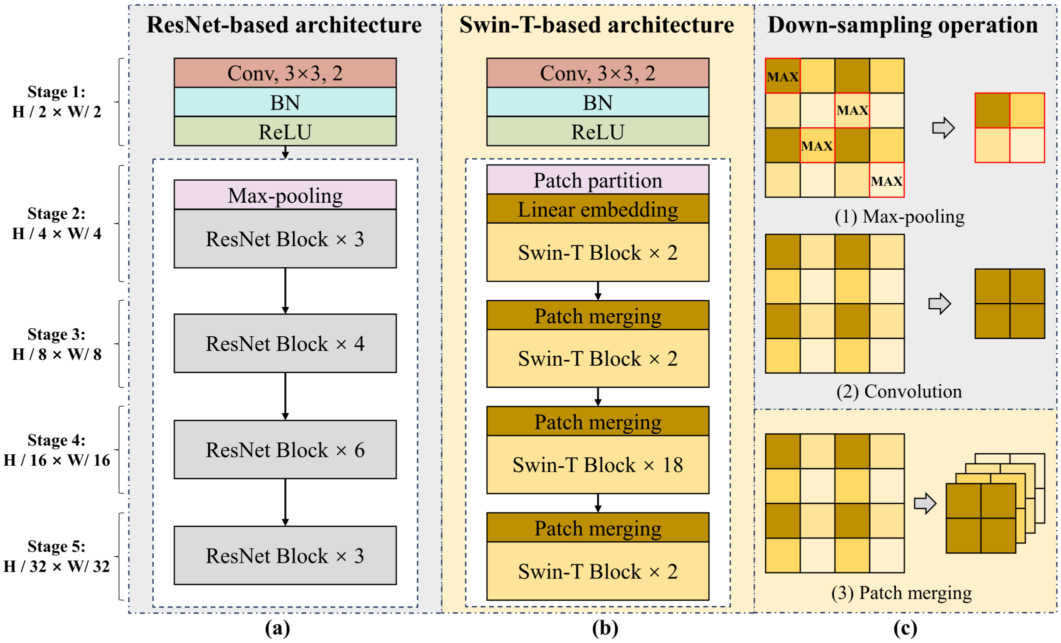


Fig. 4. The architecture of the encoder branch for extracting multi-scale features from high-resolution optical images captured from an orbiter. (a) ResNet-based architecture from DLunarDTMNet, (b) Swin-T-based architecture introduced in this paper, and (c) spatial dimension reduction methods of ResNet using max-pooling (1) and convolution with a stride of 2 (2) operations, while Swin-T using patch merging (3). The values depicted on the left side indicate the output size of each stage. Conv: convolution operation, BN: batch normalization, ReLU: an activation function.

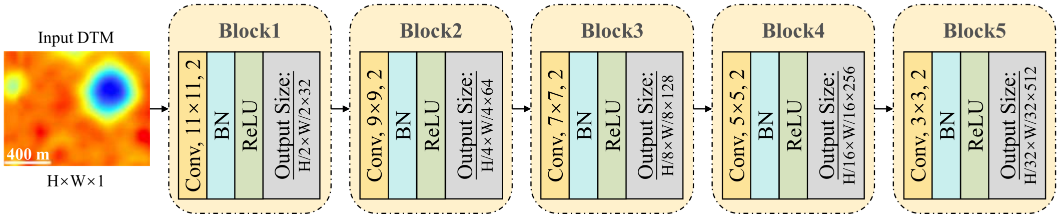


Fig. 5. The architecture of the encoder branch for extracting features from coarse-resolution DTMs.

DTMs as a constraint to enhance performance, we use a separate encoder branch dedicated to capturing elevation information from coarse-resolution DTMs. Given that the input DTM provides limited information compared to images with detailed textures, the architecture of the DTM encoder branch consists of only five simple Conv blocks. Each block is constructed with ‘a convolution layer $\rightarrow$ a batch normalization layer $\rightarrow$ a Rectified Linear Unit (ReLU) activation function’, as depicted in Fig. 5. Their convolution kernels are 11 × 11, 9 × 9, 7 × 7, 5 × 5, and 3 × 3 pixels in size, respectively. The stride of the convolution operation in each block is set to 2 to down-sample the spatial dimensions of the feature maps, creating a hierarchical structure. This ensures that the output dimensions from each block match those of the image encoder branch.

2) Fusion module: The dual-encoder module yields two sets of feature maps: one from the image encoder branch sized at H/32 × W/32 × 1536, denoted as E_i^5, and another from the DTM branch sized at H/32 × W/32 × 512, denoted as E_d^5. The fusion model is adopted to concatenate the outputs from the two branches into one volume. Firstly, E_i^5 and E_d^5 are combined into a single H/32 × W/32 × 2048 volume. Then, a 3 × 3 kernel convolutional layer and a batch normalization layer are applied to reduce the channel dimensions, resulting in a new volume (E^5) sized at H/32 × W/32 × 1536:

$$E^5 = \text{BN}\left(\text{Conv}_{3\times3}\left(\text{Concate}(E_i^5, E_d^5)\right)\right), \tag{3}$$

where Concate represents the concatenation operation.

The output volume from the fusion module will serve as the initial input to the decoder module. However, this volume, with coarse resolution, is primarily suited for capturing abstract or global features. In contrast to the inherently coarse-resolution information of the input DTM, the multi-scale features extracted from the Swin-T-based encoder branch, particularly those from the earlier stages, are enriched with fine-detailed features. To facilitate DTM predictions that encompass information across a wider range of scales, with a particular emphasis on fine details, a strategic integration operation is introduced based on our dual-encoder module. This integration involves establishing direct connections between the output features from the initial four stages of the image encoder branch (E_i^4, E_i^3, E_i^2, E_i^1) and their corresponding counterparts within the decoder module (D^2, D^3, D^4, D^5):

$$\text{Concate}\left(E_i^u, D^{6-u}\right), \tag{4}$$

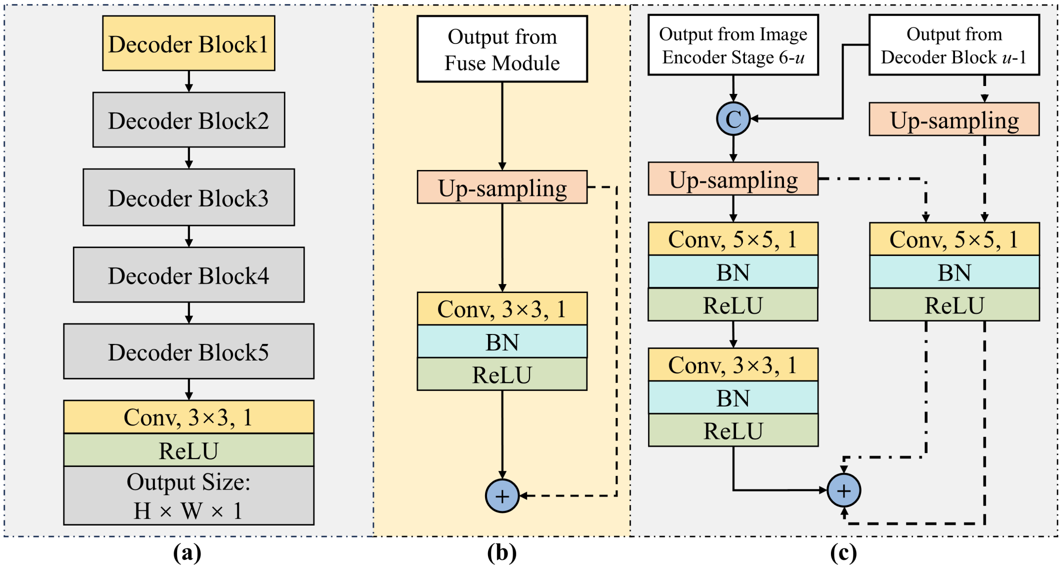


Fig. 6. The architecture of the decoder module. (a) Illustration of the overall structure, (b) details of decoder block1, and (c) details of subsequent decoder blocks, where u ranges from 2 to 5. The dash-dot line in (c) represents the residual connection fashion of UPB, while the dashed line represents the residual connection fashion proposed in this study. 'C' represents the concatenation operation. '+' represents the addition operation.

where u represents the image encoder stage number, ranging from 4 to 1.

3) Residual-connected decoder module: The decoder module utilizes multi-scale feature maps obtained from the image encoder branch and the fusion module as input. It gradually enhances feature resolution while simultaneously reducing the feature channel dimension through five consecutive blocks followed by a 3 × 3 kernel convolutional layer and a ReLU activation function (Fig. 6a). This produces a single-channel output with pixel-level resolution, matching the size of the input image, and serving as the predicted DTM. Regarding the decoder blocks, we incorporate residual connections to establish short-circuit connections, depicted as dashed lines in Fig. 6b and Fig. 6c. In contrast to the Up-Projection Blocks (UPBs) [17], [25], [39], which establish a residual connection with two convolution operation branches after the concatenation and up-sampling operations (shown by the dash-dot line in Fig. 6c), we establish a distinct residual connection for outputs originating from the preceding decoder block (indicated by the dashed line in Fig. 6c). Specifically, only one branch in UPBs is preserved to pass the information concatenated from the previous decoder block and the counterpart from the image encoder branch. Unlike the Swin-T blocks in the encoder module, the decoder blocks are positioned closer to the model output, allowing them to retain a richer array of information from earlier stages in the network. Therefore, a separate branch consisting of an up-sampling operation followed by a 5 × 5 kernel Conv block is designed to more effectively pass features from the previous decoder block in the network. These two separate branches form a new type of residual connection of the decoder module, which can empower the network to effectively integrate features, ultimately enhancing the accuracy of our DTM predictions.

C. Hybrid loss function

The datasets, constructed from high-resolution optical images captured by an orbiter, encompass a wide range of scenarios, demonstrating significant variations in the shapes and scales of both regularly and irregularly shaped topographical features. These variations pose challenges when training DL models for effective DTM estimation. To address these challenges and achieve accurate vertical predictions while capturing well-defined and properly shaped terrain features in the horizontal domain, we define a hybrid loss function that guides the model in minimizing the residuals in both domains:

$$l(z, z^*) = \phi l_{mn-elev}(z, z^*) + \gamma l_{grad}(z, z^*) + \lambda l_{norm}(z, z^*), \tag{5}$$

where z and z^* represent the ground truth and the predicted DTMs, respectively. $l_{mn-elev}$, l_{grad}, and l_{norm} are three loss terms. ϕ, γ, and λ are the related weight parameters.

While the DTM is normalized to a fixed range during the pre-processing stage, its elevation distribution deviates significantly from a uniform distribution pattern. This disparity is illustrated in Fig. 7a, where the mean values of different DTMs show considerable differences. The mean absolute error ($l1$ loss) directly measured from these DTMs is greatly influenced by the process of finding the average scale of the scenes [40]. This constitutes a substantial portion of the total error and impacts the network performance [40]. Motivated by this, we introduce a mean-normalized loss function ($l_{mn-elev}$), measuring the errors on the predicted DTM normalized by the mean elevation values. The mean-normalized DTMs are with uniform mean values as displayed in Fig. 7b.

The $l_{mn-elev}$ loss is defined by applying the $l1$ loss over the mean-normalized values, i.e.,

$$\bar{z} = \sum_{q=1}^{m} z_q/m; \quad \bar{z}^* = \sum_{q=1}^{m} z_q^*/m, \tag{6}$$

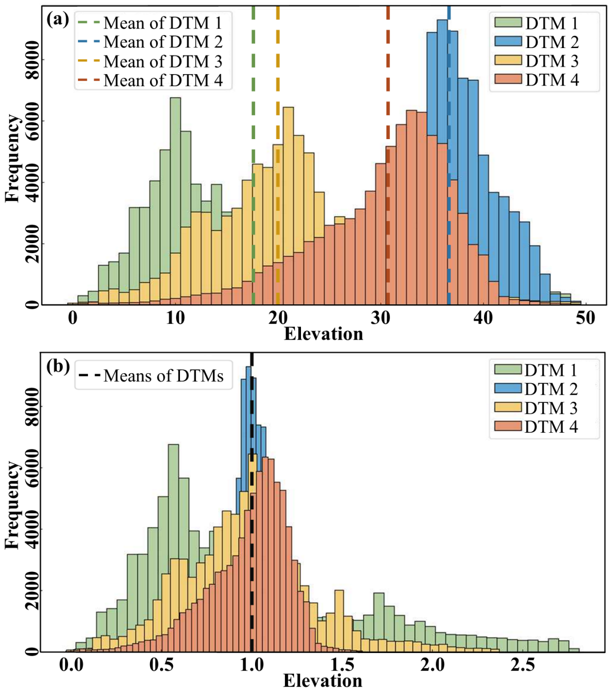


Fig. 7. Examples of the elevation distribution of DTMs. (a) DTMs normalized to a fixed range during the pre-processing stage. (b) DTMs further normalized based on mean elevation values.

$$l_{mn-elev}(z, z^*) = \sum_{q=1}^{m} |z_q/\bar{z} - z_q^*/\bar{z}^*|/m, \tag{7}$$

where z_q represents the pixel elevation value in the ground truth, z_q^* is the corresponding one in the predicted DTM, and m is the number of pixels in the DTM. A small offset can be added when the dataset contains samples with $\bar{z} = 0$.

The last two terms in Eq. (5) serve to penalize errors in the horizontal domain. As presented in DLunarDTMNet, l_{grad}, is the $l1$ loss defined over the gradient of the predicted DTM and the ground truth, aimed at improving the accuracy of terrain detail prediction. The third loss term, l_{norm}, measures the angle between the surface normals of the predicted DTM and ground truth, enhancing sensitivity to small surface undulations [17].

D. DTM mosaic generation

To construct large-sized DTM mosaics from small-sized predicted tiles, three essential steps are required: scale recovery, projection assignment, and DTM mosaicking [19], [21], [22], [25]. Due to the improved performance in multi-scale terrain feature reconstruction and elevation recovery, our approach achieves seamless and high-quality DTM mosaicking with direct scale recovery. This is accomplished by a straightforward application of the inverse transformation of DTM normalization (as presented in Eq. (2); Section III-A1):

$$Z_{sr} = Z_{pred} \cdot (\max(Z) - \min(Z))\, / NR + \min(Z), \tag{8}$$

where Z_{sr} denotes the scale recovered DTM, and Z_{pred} refers to the predicted DTM. Z serves as the reference DTM, supplying both real scale information and projection details for projection assignment. After recovering both the real scale and the projection information, the small-sized DTM tiles are integrated into a contiguous DTM mosaic for the entire large-sized orbiter imagery using the dem_mosaic tool provided in the Ames Stereo Pipeline [26].

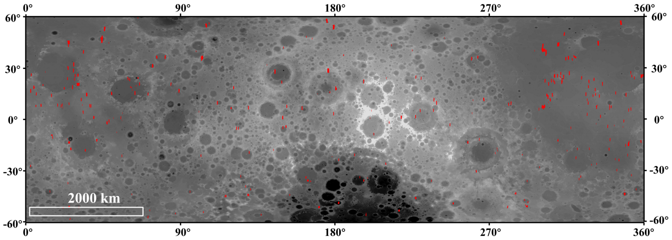


Fig. 8. The distribution of SPG-derived NAC products released on LROC PDS within +/- 60° longitude. The sparse red pixels indicate the coverage area of these products, while the gray background represents the SLDEM.

IV. Datasets and implementation details

A. Training and validation Datasets

The NAC onboard LRO is a highly advanced system designed to perform high-resolution (0.5 – 2 m) imaging of the lunar surface. While the availability of stereo NAC pairs suitable for terrain modeling is spatially limited across the entire lunar surface, high-quality DTMs and ORIs have been generated for some local regions using the SPG method [7]. These products, mostly with a spatial resolution of 2 – 5 m, are accessible through the LROC Planetary Data System (PDS) node. Fig. 8 illustrates the distribution of these products across lunar highlands and maria within a longitude range of +/- 60°. These datasets exhibit diverse terrain characteristics and geomorphological elements, with elevation ranging from -9.1 km to 10.8 km. The NAC images used to derive these products were captured at solar elevation angles ranging from 18.55° to 63.65°. The varied and heterogeneous attributes make them an optimal data source for constructing training sets for high-resolution single-view lunar DTM estimation. Meanwhile, the 512 pixel per degree resolution photogrammetric models from the Selenological and Engineering Explorer (SELENE) have been co-registered with Lunar Orbiter Laser Altimeter (LOLA) profiles to generate an improved, semi-global (within +/- 60° longitude) DTM, the so-called "SLDEM" [41]. It offers high elevation accuracy and maintains consistent horizontal positioning with NAC-derived products. In this study, we utilize the NAC ORIs along with the corresponding SLDEM of the respective regions as input sources, while the NAC DTMs serve as the ground truth for training and validating the model.

To unify the resolution of inputs for the dual-encoder module in the proposed ELunarDTMNet, SLDEM is interpolated to match the resolution of the corresponding NAC ORIs by spline interpolation [42]. The NAC operates in push-broom mode as a line-scanner camera, offering images with dimensions of up to 5, 000 pixels × 50, 000 pixels. Consequently, the SPG-derived NAC DTMs and ORIs are of large size. Following DLunarDTMNet, we split the data into small-sized tiles with 256 pixels × 320 pixels without any overlap. These

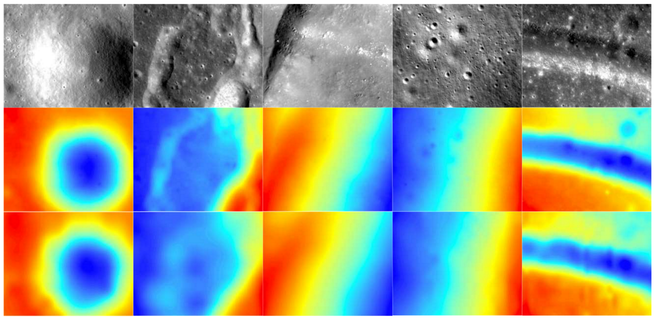


Fig. 9. Examples of the small-sized samples: top panel is the NAC ORIs, middle panel is the SPG NAC DTMs, bottom panel is the corresponding SLDEM.

tiles are then used to construct training and validation sets, with each sample containing the NAC DTM and ORI, along with the corresponding SLDEM covering the same small area.

We perform data filtering to exclude data with horizontal offset discrepancies between NAC DTMs and ORIs, as well as samples containing null values and image artifacts, such as evident intensity errors. Additionally, due to the presence of artifacts in both the NAC DTM and SLDEM, the elevation data in certain samples may display considerable inconsistencies, which need to be excluded. Specifically, we compare the elevation difference between NAC DTM and SLDEM for each sample. In cases where the disparity exceeds three times the standard deviation (STD), we discard the samples,

$$\Delta Z_D = \Delta Z_N - \Delta Z_S, \tag{9}$$

$$\begin{aligned} &\text{If } \mu - 3\sigma < \Delta Z_{Dp} < \mu + 3\sigma, \text{ retain;} \\ &\qquad\qquad\qquad\qquad\qquad \text{else, discard,} \end{aligned} \tag{10}$$

where ΔZ_N represents the elevation difference of the NAC DTM, while ΔZ_S represents the elevation difference of the SLDEM. μ and σ are the mean and STD of ΔZ_D. ΔZ_{Dp} is the value of ΔZ_D at sample p. Examples of the samples are shown in Fig. 9.

B. Testing dataset

To assess the effectiveness of our proposed ELunarDTMNet compared to the state-of-the-art single-view DL-based approaches to lunar DTM reconstruction, including DLunarDTMNet, LDEMGAN, and MADNet [17], [23], [24], we perform a comparative analysis reported in Section V-A. We select four typical areas (Apollo 11, Imbrium, Lichtenberg, and Highland Ponds (Hponds)) as used in LDEMGAN, depicted in Fig. 10 (a to d), to evaluate the performance of DLunarDTMNet, LDEMGAN, and the proposed ELunarDTMNet. Considering the focus of MADNet on the Von Kármán Crater, the landing area of the Chang'E-4 mission, we select two additional areas, the Chang'E-4 landing site and the MonsTai, as depicted in Fig. 10e and Fig. 10f, to assess the performance of their method alongside our ELunarDTMNet. Table I provides the basic information for each area. These areas comprise a variety of terrain types, ranging from relatively flat areas with small elevation differences (e.g., Apollo 11 and Chang'E-4 landing site) to regions with significant variations in elevation (e.g., Lichtenberg and Hponds). Alongside the ubiquitous multi-scale craters, these areas also showcase distinct features such as mountains, central peaks, impact melts, etc.

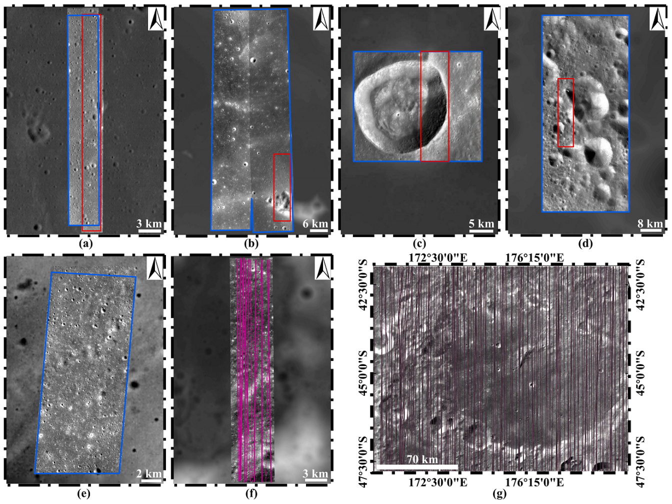


Fig. 10. The NAC images of the test areas with the SLDEM as background: (a) Apollo 11 (M150361817LE, M150361817RE, M150368601RE); (b) Imbrium (M1106095239LE, M1106095239RE, M183697099LE, M183697099RE); (c) Lichtenberg (M191052644LE, M191052644RE, M191059794LE, M191059794RE); (d) Hponds (M182432168LE, M182432168RE, M182439316LE, M182439316RE, M182425021LE); (e) Chang'E-4 landing site (M1303619844, M1303619844); (f) MonsTai (M123417906, M123417906). The red boxes in (a), (b), (c), and (d) indicate the areas considered for comparing the performance of our method with the SFS method. (g) shows the Change'E-2 imagery for the Von Kármán Crater area. The blue polygons mark the regions of SPG DTMs for accuracy evaluation. The purple tracks in (f) and (g) indicate the LOLA tracks used for accuracy assessment due to the absence of publicly released or reliable SPG DTMs for reference.

For a comparative analysis between the single-view SFS method [8] and our proposed method, as described in Section V-B1, we specifically use four sub-regions indicated by red boxes in Fig. 10 to test the results. To evaluate the robustness of both our method and SFS approach under varying illumination conditions, we select two different NAC images that collectively cover the landing site of Chang'E-3 for validation, as elaborated in Section V-B2. Additionally, to evaluate the generalization capability of our model (trained on the NAC-based dataset) on other datasets, we employ imagery captured by the Chang'E-2 orbiter, which provides the second-highest resolution lunar global imaging, as input images. This is detailed in Section V-C. The experimental area is the Von Kármán Crater, covered by 7 m resolution Chang'E-2 imagery spanning 175 km × 218.6 km, as indicated in Fig. 10g.

C. Implementation details

We implement our proposed network using the Pytorch framework and train it on a single NVIDIA TITAN RTX with 24G memory. We perform the network training with a batch size of 4 and an initial learning rate of 0.0001. The cosine annealing learning rate strategy is employed for the learning rate decay [43]. To assess the effectiveness of our proposed improvements (the ablation study in Section VI), we randomly select 40k samples for the training set and 5k samples for the validation set. The network is trained with 15k iterations. As

TABLE I
THE BASIC INFORMATION OF THE NAC IMAGE-BASED TESTING AREAS.

Area name	MinLon, MinLat (°, °)	MaxLon, MaxLat (°, °)	Typical features	Area size (km^2)	NOI	Resolution (m)	ER (km)
Apollo 11	23.37, 0.29	23.52, 1.24	Crater	117	3	1.5	0.26
Imbrium	333.37, 27.42	334.42, 29.90	Mountain	1949	4	1.5	1.56
Lichtenberg	291.89, 31.47	292.93, 32.23	Large crater	607	4	1.5	2.83
Hponds	166.92, 41.22	168.36, 43.64	Impact melt	2401	5	1.7	3.49
Chang'E-4	177.32, -45.99	177.80, -45.31	Crater	173	2	1	0.13
MonsTai	176.32, -44.33	176.58, -43.33	Central peak	171	2	1	1.14

"MinLon", "MinLat", "MaxLon", "MaxLat" are abbreviations for minimum longitude, minimum latitude, maximum longitude, and maximum latitude, respectively. ”NOI” and ”ER” are abbreviations for number of images and elevation range, respectively.

for the best-performing network determined by the ablation study, we employ 162k samples for the final training process with 60k iterations, and 18k samples for the validation. For the initial 5k iterations, the network is trained only with the loss function $l_{mn-elev}$, followed by the addition of l_{grad} and l_{norm} for subsequent training. The weights (ϕ of $l_{mn-elev}$, γ of l_{grad}, and λ of l_{norm}) in Eq. (5), are empirically set as 10, 1, and 1, respectively, based on experiments. The NAC DTM and SLDEM in each sample are normalized to [0, 50], meaning that NR is set to 50 according to our normalization strategy.

For the ablation study, we quantitatively compare the improvements made by the proposed method using seven metrics on the validation set, based on the normalized scale. These metrics are defined as follows:

Average Relative Error (REL):

$$\left(\sum_{q=1}^{m}|z_q - z_q^*|/z_q\right)/m, \tag{11}$$

Root Mean Squared Error (RMSE):

$$\sqrt{\sum_{q=1}^{m}(z_q - z_q^*)^2/m}, \tag{12}$$

Average Log10 error:

$$\sum_{q=1}^{m}|\log_{10}(z_q) - \log_{10}(z_q^*)|/m, \tag{13}$$

Threshold accuracy (a_j), including three metrics:

$$\% \text{ of } z_q \text{ s.t. } \max\left(\frac{z_q}{z_q^*}, \frac{z_q^*}{z_q}\right) = a_j < 1.25^j \text{ for } j = 1, 2, 3, \tag{14}$$

Peak Signal to Noise Ratio (PSNR):

$$10 \cdot \log_{10}\left(\max(z_q^*)^2/\text{mse}(z_q, z_q^*)\right), \tag{15}$$

where z_q represents the pixel elevation value in the ground truth, z_q^* is the corresponding one in the predicted DTM, and m is the number of pixels in the DTM. The REL, RMSE, Log10, and a_j are the metrics used to evaluate elevation accuracy. In general, lower values of REL, RMSE, and Log10, while higher values of a_j, indicate better quality. PSNR is the metric that measures image similarity and quality, where higher values indicate better results.

During the training process, we validate the trained model with the validation set every 5k iterations. We utilize the well-trained model of the proposed ELunarDTMNet, determined at the iteration with the lowest RMSE values on the validation set, to predict DTMs on the testing dataset. For the generated DTM mosaics with recovered scales, we adopt the five metrics proposed by LDEMGAN to assess the reconstruction errors and compare them with other single-view DL methods (DLunarDTMNet, LDEMGAN, and MADNet) and the SFS method [8]. These error metrics include the percentages of errors (less than 2 m, 4 m, and 10 m) between the reconstructed DTM and the ground truth, called Reconstruction Error (RE; Eq. (16)), in addition to MAE and RMSE.

$$\% \text{ of } z_q \text{ s.t. } |z_q - z_q^*| = RE < t \text{ for } t = 2, 4, 10, \tag{16}$$

$$\text{MAE} = \sum_{q=1}^{m}|z_q - z_q^*|/m, \tag{17}$$

where z_q represents the pixel elevation value in the ground truth, which is derived from the SPG method [7]. In areas where reliable SPG DTMs are unavailable, we utilize LOLA tracks as the ground truth. Here, all tracks are first co-registered to the reconstructed DTM by minimizing the height residuals between each track and the DTM [44], [45]. This is achieved by shifting the LOLA tracks over the DTM and calculating the STD at each location. The smallest STD between the reconstructed DTM and each shifted LOLA track represents the correct location of the track and is employed to evaluate the accuracy:

$$\text{STD} = \sqrt{\sum_{w=1}^{r}(z_{laser,w} - z_{DTM,w}^*)^2/(r-1)}, \tag{18}$$

where r represents the number of LOLA spots, $z_{laser,w}$ denotes the elevation value of one LOLA spot, and $z_{DTM,w}^*$ represents the corresponding one in the reconstructed DTM.

V. EXPERIMENTAL RESULTS

A. Results of the comparison with state-of-the-art DL methods

In this section, we conduct a comparative analysis between the proposed ELunarDTMNet and the state-of-the-art single-view DL methods for lunar DTM reconstruction. The reconstruction errors using SPG-derived DTMs as reference [7] are provided in Table II. Overall, the performance of our ELunarDTMNet is excellent compared to LDEMGAN across all four areas on Apollo 11, Imbrium, Lichtenberg, and Hponds. Particularly, in the case of Hponds where there are large elevation differences, we observe a sharp decrease of 72% in MAE and 68% in RMSE. Both DLunarDTMNet and

TABLE II
COMPARISON OF RECONSTRUCTION ERRORS OF THE PROPOSED ELUNARDTMNET WITH STATE-OF-THE-ART DL METHODS IN THE TESTING AREAS.

Area name	Method	RE < 2m (%)	RE < 4m (%)	RE < 10m (%)	MAE (m)	RMSE (m)
Apollo 11	LDEMGAN	48.32	73.06	91.54	2.76	3.02
	DLunarDTMNet	58.68	88.83	99.71	2.02	2.62
	ELunarDTMNet	**59.11**	**89.16**	**99.79**	**1.98**	**2.56**
Imbrium	LDEMGAN	42.27	70.47	85.67	3.03	3.64
	DLunarDTMNet	**52.41**	**83.17**	99.38	**2.36**	3.09
	ELunarDTMNet	51.76	81.69	**99.64**	2.37	**3.06**
Lichtenberg	LDEMGAN	**27.52**	45.11	68.89	9.26	10.52
	DLunarDTMNet	21.27	41.47	80.12	6.47	8.71
	ELunarDTMNet	23.31	**45.37**	**83.20**	**5.89**	**7.83**
Hponds	LDEMGAN	25.52	45.95	62.51	9.43	10.65
	DLunarDTMNet	43.20	71.96	96.56	3.21	5.14
	ELunarDTMNet	**47.27**	**77.72**	**99.15**	**2.66**	**3.44**
Chang'E-4	MADNet	33.72	63.13	95.00	3.83	5.27
	ELunarDTMNet	**55.42**	**85.86**	**99.53**	**2.19**	**2.86**

The best results are in **bold**. The numbers reported by LDEMGAN are from their original paper, whereas DLunarDTMNet were obtained by using their code repository. The metrics reported by MADNet are from their released DTM, which has been co-aligned to the SPG DTM in this study.

the proposed ELunarDTMNet maintain promising accuracy. It is worth noting that DLunarDTMNet required the co-alignment technique to ensure the predicted DTMs are highly consistent with the reference data and further maintain self-consistency among adjacent DTM tiles. However, the metrics of DLunarDTMNet in Apollo 11, Lichtenberg, and Hponds areas are still inferior to those of our method, particularly in regions with significant elevation variations, such as Lichtenberg and Hponds. For the Imbrium region, where the reconstruction errors of DLunarDTMNet, specifically those less than 2 m and 4 m, and MAE, exhibit better performance compared to our method, the degradation is only marginal. Besides, our proposed ELunarDTMNet also exhibits a significantly higher level of elevation accuracy compared to that of MADNet. For instance, at the Chang'E-4 landing site, utilizing SPG DTM as the ground truth, our approach surpasses MADNet across all five metrics. In the case of MonsTai, our STD with LOLA tracks (the distributions are shown in Fig. 10f) is 1.36 m, significantly lower than that of MADNet with 10.86 m.

Fig. 11 presents the reconstructed DTMs for the Lichtenberg and Hponds regions using the proposed ELunarDTMNet. It also includes the corresponding RE maps and local area Hill-Shaded (HS) maps from the ELunarDTMNet and DLunarDTMNet reconstructions for performance comparison. As shown in the RE maps, DLunarDTMNet displays more areas with large errors in steep terrains, such as the wall of craters on the Lichtenberg and Hponds regions. The HS maps in local area 1 show DLunarDTMNet displays striped artifacts. The results from the SPG DTM exhibit matching artifacts and incongruities with the NAC image. In contrast, our method captures the intricate texture and resembles the NAC image. Local area 2 highlights a region characterized by irregular impact melts. The proposed ELunarDTMNet manages to reconstruct the linear meandering features. In contrast, the SPG method and DLunarDTMNet fall short of recovering these terrain features. Also, the SPG method fails to reconstruct some craters with diameters smaller than tens of meters, as marked by the yellow arrows. In area 3, the shapes of craters with diameters of several hundred meters, such as those highlighted by yellow circles, derived from DLunarDTMNet, deviate from the NAC images. In contrast, both the SPG and the proposed methods yield more accurate results. Besides, while our method uses a coarse-resolution DTM (e.g., SLDEM) as input to provide elevation constraints, it is not affected by the lack of topographic details in SLDEM and effectively recovers the fine topographic features. Moreover, unlike DLunarDTMNet, the proposed ELunarDTMNet eliminates the need for the co-alignment technique during the DTM mosaic generation stage. This key difference significantly streamlines our DTM reconstruction process. To compare the computing time of both methods, we select a specific area within the Imbrium region with 3, 400 pixels × 14, 920 pixels (114 km^2), indicated by the red box in Fig. 10b. We can efficiently generate the DTM mosaic in just 12 minutes. In contrast, DLunarDTMNet requires 742 minutes to accomplish the same area DTM, which is about 60 times longer than the time of the proposed technique.

Fig. 12 displays the reconstructed DTMs derived from the proposed ELunarDTMNet and performance comparison with the MADNet in the Chang'E-4 landing site and MonsTai areas. MADNet exhibits larger reconstruction errors in terrain feature reconstruction, especially in the MonsTai region, where elevation differences reach up to 1.14 km. The HS maps in local area 1 show MADNet encounters challenges in accurately reconstructing the depth information of craters with a diameter of several hundred meters. This limitation results in subtle changes in shadow intensity, as indicated by the four craters highlighted with yellow circles. Conversely, the results of our technique show changes with more contrast in shadow intensity for craters of this scale, demonstrating consistency with the NAC image. In local area 2, MADNet shows deviations in the shape characteristics of some craters, with diameters of several tens of meters, compared to the NAC image and our results, as indicated by the yellow arrows. Additionally, the irregular pimple-shaped feature, highlighted by the red arrow, is missing in MADNet. In contrast, our proposed ELunarDTMNet accurately captures the shape of this feature.

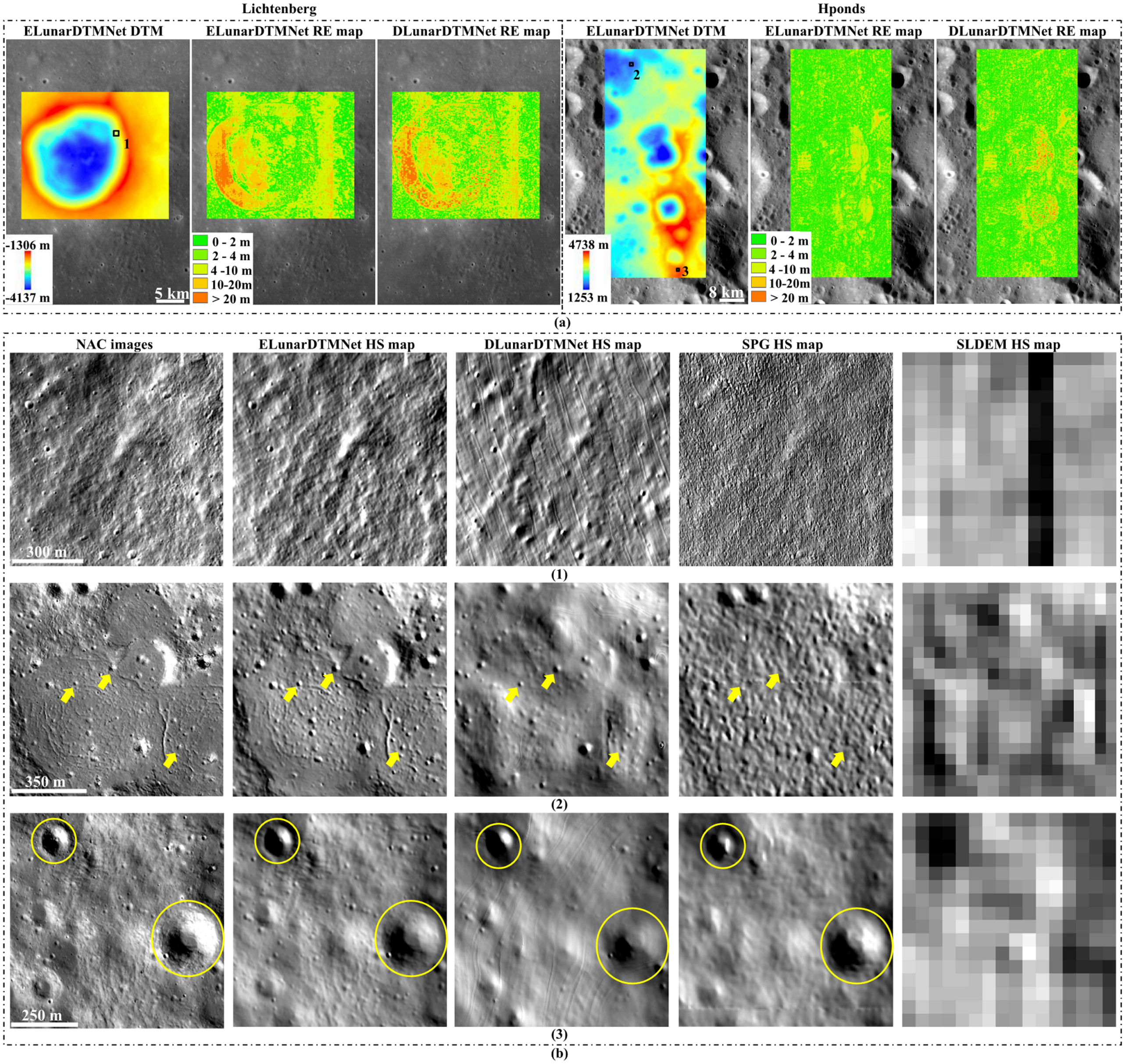


Fig. 11. The reconstructed DTMs using the proposed ELunarDTMNet and their performance comparison with the DLunarDTMNet in the Lichtenberg and Hponds regions. (a) The DTMs generated by ELunarDTMNet, and the RE maps generated by ELunarDTMNet and DLunarDTMNet, (b) the NAC images, as well as the HS maps derived from ELunarDTMNet, DLunarDTMNet, SPG method [7], and SLDEM on the local areas. The black boxes in (a) indicate the locations of the local areas. The yellow arrows point to small-scale craters that the SPG method fails to capture, whereas the yellow circles highlight craters that DLunarDTMNet struggles to accurately predict.

B. *Results of the comparison with the single-view SFS method*

1) Comparison of accuracy, resolution, and efficiency: In this section, we compare the performance of the proposed ELunarDTMNet with that of the SFS method [8], [26] using the same NAC images and SLDEM as input. To ensure a fair evaluation, we adopt the same tiling process as our method to produce SFS DTMs, where the input data is split into smaller tiles (256 pixels × 320 pixels) with padding (100 pixels). The tiles are then blended to create full SFS DTM mosaics. This assessment is conducted across four different areas, as indicated by the red boxes in Fig. 10. Table III shows the reconstruction errors obtained by comparing the generated DTMs with the SPG DTMs [7]. While three metrics of the SFS method perform better than our method in the Apollo 11 region, the difference in advantage is marginal. However, our method consistently outperforms the SFS method in the Imbrium, Lichtenberg, and Hponds regions.

Fig. 13 shows the reconstructed DTMs of the Imbrium and Lichtenberg regions, as well as the HS maps and RE maps for local areas. Both the SFS method and our proposed ELunarDTMNet depict terrain features consistently with the SPG DTMs, exhibiting similar elevation trends. The SFS method is renowned for its capability to reconstruct DTMs with detailed terrain information. Among previous DL methods, DLunarDTMNet shows competitive elevation accuracy, as shown in Table II, but it fails to preserve some terrain

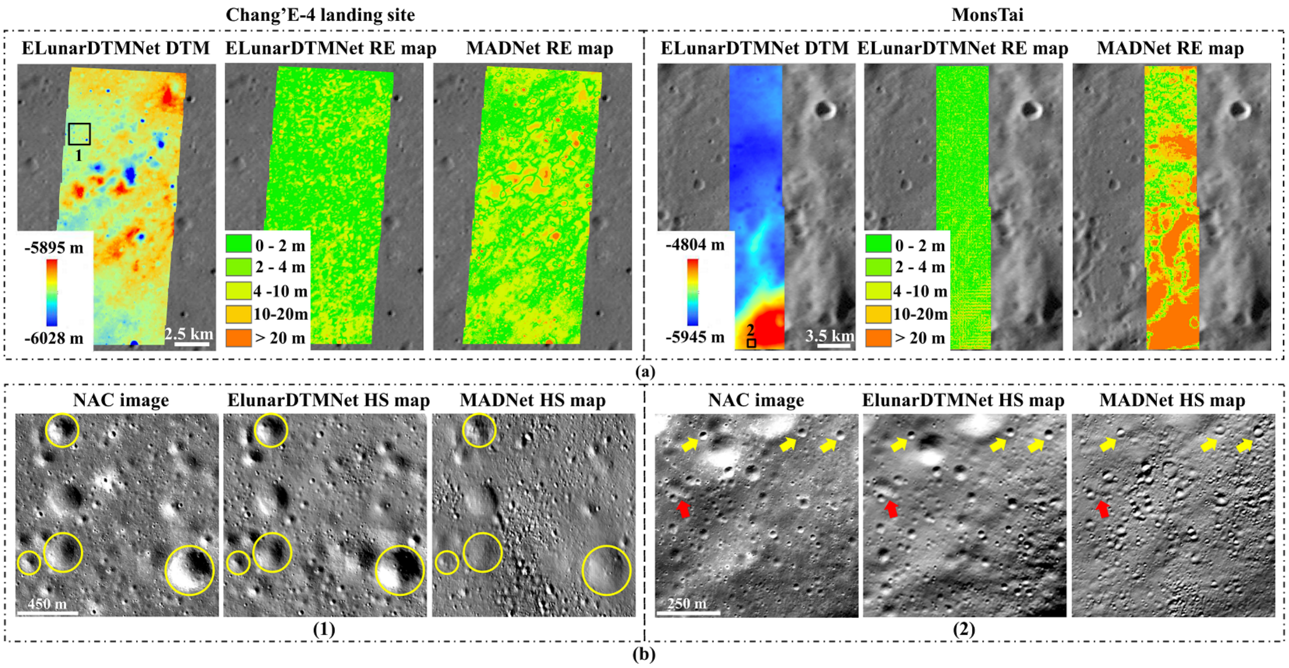


Fig. 12. The reconstructed DTMs using the proposed ELunarDTMNet and their performance comparison with the MADNet in the Chang'E-4 landing site and MonsTai areas. (a) The DTMs generated by ELunarDTMNet, and the RE maps generated by ELunarDTMNet and MADNet, (b) the NAC images, and the HS maps derived from ELunarDTMNet and MADNet on the local areas. The SPG DTM [7] in the Chang'E-4 landing site area and the SLDEM [41] in the MonsTai area are used as the reference DTMs to determine the RE maps. The results from MADNet have been co-aligned to the SPG DTM (SLDEM) in this study. The black boxes in (a) indicate the locations of the local areas. The yellow circles and arrows indicate some craters that MADNet struggles to accurately predict. The red arrows point to an irregularly shaped feature that MADNet fails to reconstruct.

TABLE III
COMPARISON OF DTM RECONSTRUCTION ERRORS OBTAINED BY THE SFS METHOD [8] AND THE PROPOSED ELUNARDTMNET.

Metric	Method	Apollo 11	Imbrium	Lichtenberg	Hponds
RE<2m (%)	SFS	56.14	52.39	28.00	49.82
	ELunarDTMNet	**57.06**	**67.79**	**30.30**	**59.50**
RE<4m (%)	SFS	**87.78**	84.04	55.81	80.13
	ELunarDTMNet	87.69	**92.57**	**61.01**	**87.39**
RE<10m (%)	SFS	**99.67**	99.20	96.99	99.34
	ELunarDTMNet	99.64	**99.88**	**98.64**	**99.86**
MAE (m)	SFS	2.11	2.35	3.98	2.51
	ELunarDTMNet	**2.09**	**1.69**	**3.59**	**2.03**
RMSE (m)	SFS	**2.71**	3.08	4.84	3.25
	ELunarDTMNet	2.72	**2.24**	**4.34**	**2.67**

The best results are in **bold**.

details. The HS maps of the local areas and further localized areas illustrate that the proposed ELunarDTMNet outperforms DLunarDTMNet in capturing small-scale and subtle terrain features, demonstrating similar effective resolution as the SFS method. In Fig. 13b, the SFS method demonstrates artifacts in the dark region marked by the yellow circle. The RE maps on these local areas show that the proposed method achieves better accuracy compared to both the SFS method and DLunarDTMNet. Furthermore, we evaluate the processing time of the SFS method in generating the DTM mosaic on the Imbrium region. This area is also the same one used in Section V-A to evaluate the computation time of both DLunarDTMNet and the proposed ELunarDTMNet. In contrast to our approach which takes about 12 minutes, the SFS method requires a significantly longer computational time of 10, 243 minutes to derive the DTM mosaic. This sharp difference of more than 850 times highlights the superior efficiency of our approach.

2) Influence of varying illumination conditions: The NAC images reveal variations in surface appearance due to differing illumination conditions. To investigate the robustness of the proposed ELunarDTMNet and the SFS method in handling such varying illuminations, we analyze two NAC images (M1154358210RE and M1164944600RE) of the same area in the Chang'E-3 landing site, taken under different solar azimuth and elevation angles. As shown in Fig. 14, we conduct cross-illumination verification and generate HS maps on a local area to assess the performance. Despite the DTMs being derived from two different NAC images, the HS maps generated from our DTMs demonstrate a high level of consistency. For example, the wrinkle ridges exhibit identical shapes between our results and remain consistent with the NAC images. In contrast, when the illumination condition differs from that of the NAC image used to generate SFS DTMs, the HS maps from the SFS method show obvious discrepancies. Specifically, the map derived from the illumination condition of M1154358210RE, based on M1164944600RE-derived DTM (the fifth column of Fig. 14a), fails to properly model the shape of wrinkle ridges.

C. Generalization to other high-resolution orbiter imagery

We employ the well-trained model based on the NAC dataset to directly predict DTMs using the 7 m resolution Chang'E-2 imagery (Fig. 10g), in conjunction with SLDEM, for the Von Kármán Crater area. The results are compared

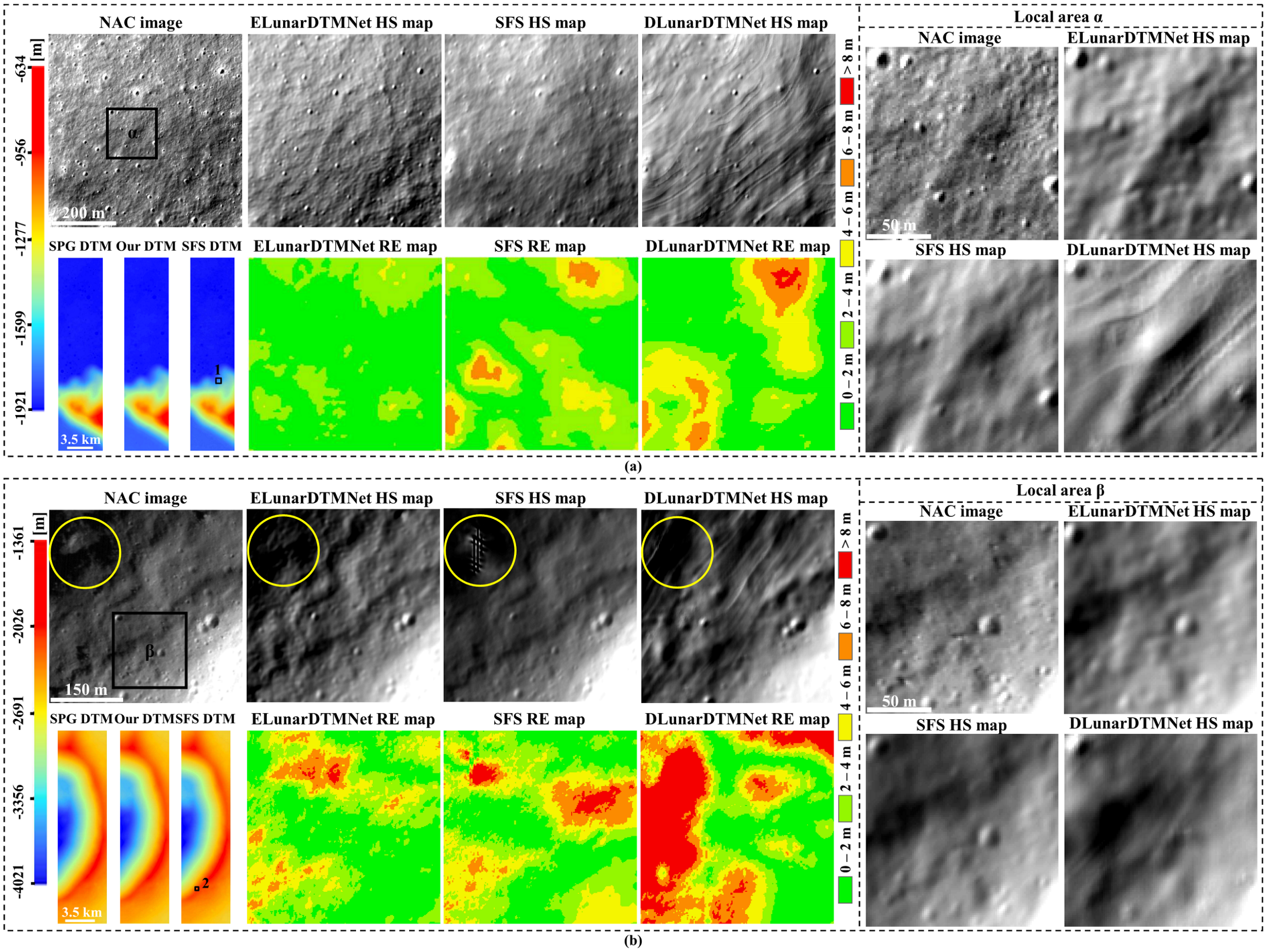
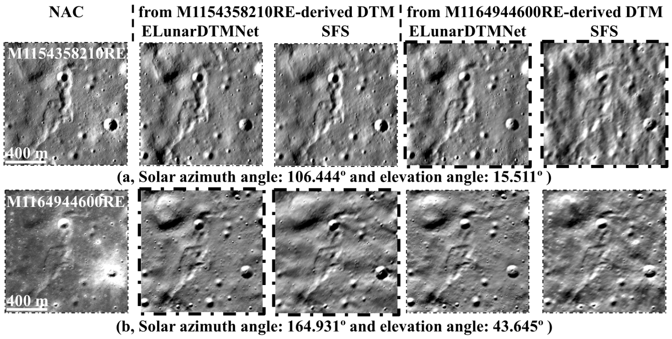


Fig. 13. The reconstructed DTMs by the proposed ELunarDTMNet and the SFS method [8], with comparisons of local areas using HS maps and RE maps. (a) Imbrium region, and (b) Lichtenberg region. The black boxes 1 and 2 on the SFS DTMs mark the locations of the local areas shown in the left panels. The locations of further localized areas in the right panels are indicated by the black boxes labeled α and β.



(a, Solar azimuth angle: 106.444° and elevation angle: 15.511°)

(b, Solar azimuth angle: 164.931° and elevation angle: 43.645°)

Fig. 14. Cross-illumination verification for the proposed ELunarDTMNet and the SFS method [8] centered at (19.60°W, 44.43°N). (a) M1154358210RE and (b) M1164944600RE. From left to right, the figures show: the NAC images, the proposed ELunarDTMNet (second and fourth columns) and the SFS (third and fifth columns) results from M1154358210RE-derived and M1164944600RE-derived DTMs. The HS maps in columns two through five are generated using the same illumination conditions as the left NAC images. Bolded borders on the maps indicate that the corresponding DTMs are generated from another NAC image.

TABLE IV
THE STD BETWEEN CHANG'E-2 IMAGERY-DERIVED DTMS AND LOLA TRACKS.

Methods	ELunarDTMNet	MADNet	SPG [46]
STD (m)	**3.726**	17.892	4.133

The best results are in **bold**.

with DTMs derived from Chang'E-2 imagery using MADNet and the SPG method [46]. Table IV presents a comparison of STDs between the Chang'E-2 imagery-derived DTMs and the LOLA tracks. Our method demonstrates the highest level of accuracy, with MADNet exhibiting a STD approximately five times larger than ours.

Fig. 15a shows the features on the MADNet-derived DTM appear blurred, whereas those on our and SPG DTMs are distinctly defined. When focusing on the enlarged pink box area, it becomes evident that the elevation trends remain consistent between the SPG method and our proposed ELunarDTMNet. On the contrary, the DTM from MADNet displays noticeable discrepancies, such as the three craters indicated by the yellow arrows. Fig. 15b shows a comparison of the HS maps for three Chang'E-2 DTMs across two local areas. Some terrain details

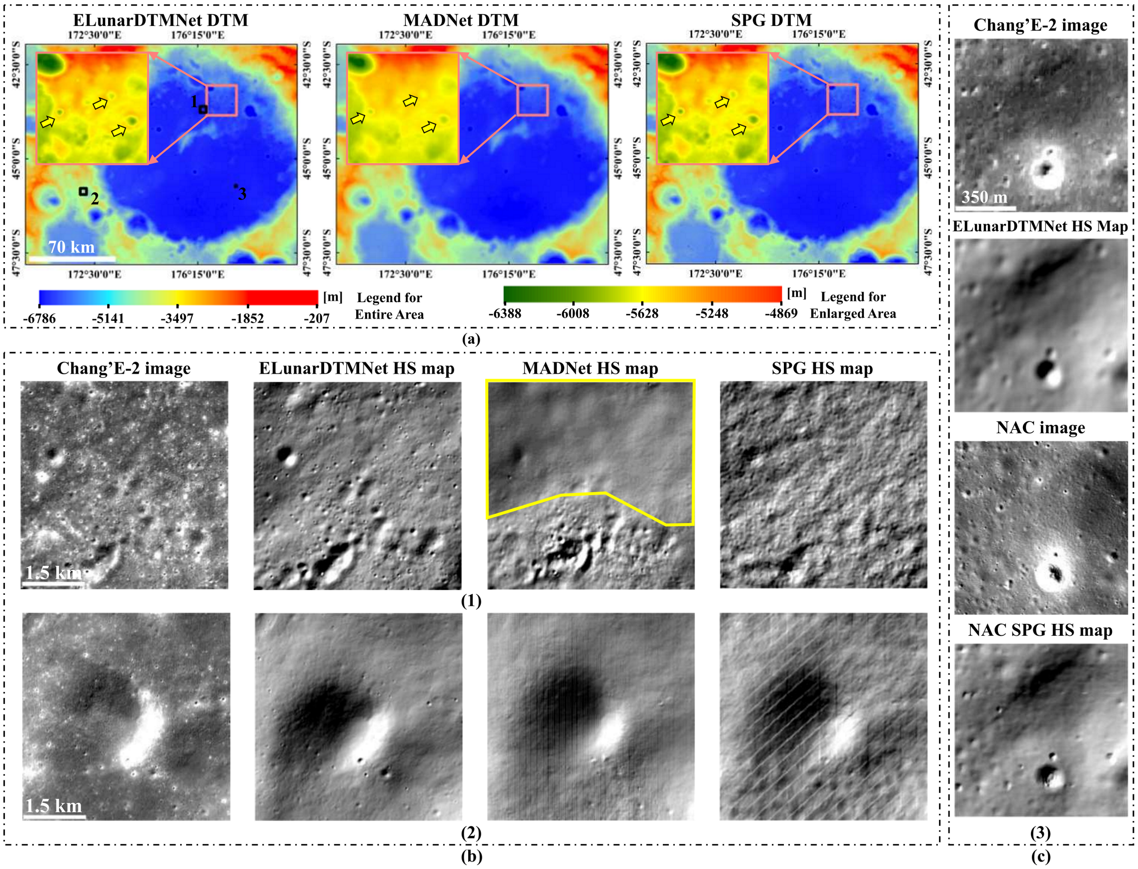


Fig. 15. The reconstructed DTMs of the Von Kármán Crater area using the Chang'E-2 imagery. (a) The DTMs reconstructed by the proposed ELunarDTMNet with a resolution of 7 m, MADNet with a resolution of 14 m, and SPG method [46] with a resolution of 20 m, (b) the Chang'E-2 imagery, and the HS maps derived from ELunarDTMNet, MADNet, and SPG method in local areas 1 and 2. The yellow arrows point to some craters where MADNet exhibits discrepancies compared to both SPG and our methods. The irregular yellow polygon indicates the areas identified by MADNet as over-smoothed. The black boxes in (a) indicate the locations of local areas. (c) presents the Chang'E-2 imagery and the corresponding HS map generated by the proposed ELunarDTMNet, along with the NAC image and the HS map derived from the SPG method based on NAC images in local area 3.

are not recovered by MADNet, such as the region, enclosed by the irregular yellow polygon, showing over-smoothing. The SPG-derived DTM also fails to accurately represent terrain features, showing noticeable artifacts. In contrast, our proposed ELunarDTMNet efficiently recovers the topographic details on the depicted areas. Besides, Fig. 15c compares the HS map generated by the proposed ELunarDTMNet using Chang'E-2 imagery with the HS map from the SPG method based on NAC images. While the Chang'E-2 imagery has a lower resolution than the NAC images, the effective resolution of our derived Chang'E-2 DTM is close to that of the SPG DTM derived from the NAC images.

VI. Discussion

A. Ablation study during training and validation stages

To validate the effectiveness of our proposed ELunarDTMNet, we perform five comprehensive ablation studies. Each study aims to test specific enhancements, including the DTM normalization strategies, the coarse-resolution DTM constraints, the image encoder branch backbones, the residual-connected mechanisms in the decoder module, and the loss functions. The training and validation logs for each enhancement and variant, based on our proposed DTM normalization strategy, are presented in Fig. 16. As the number of iterations increases, the training losses gradually decrease and converge (Fig. 16a). Concurrently, the overall accuracy demonstrates an increase on the validation set, with a_1 increasing (Fig. 16b) and RMSE decreasing (Fig. 16c). The significant increase in the loss curve at 5k iterations is due to the incorporation of the l_{grad} and l_{norm} terms during training. From the perspective of validation accuracy, their inclusion effectively guides the network to better performance. The quantitative comparison of well-trained models for each variation on the validation set is presented in Table V. It shows that our enhancements in each variation effectively improve prediction accuracy, with our proposed ELunarDTMNet achieving the best performance.

B. Ablation study during testing stage

Table VI presents the reconstructed errors of the DTM mosaics derived from the well-trained models for each variation, in three testing areas (Apollo 11, Imbrium, and Lichtenberg), as marked by the red boxes in Fig. 10. Using SLDEM as input significantly improves accuracy, which is essential for achieving high-performance results. In the Imbrium region, our metrics for the reconstruction errors less than 2 m and 10 m do not achieve the highest accuracy but still rank second

TABLE V
QUANTITATIVE COMPARISON OF WELL-TRAINED MODELS FOR EACH VARIATION ON THE VALIDATION SET BASED ON OUR PROPOSED DTM NORMALIZATION STRATEGY.

Architecture variant		a_1	a_2	a_3	REL	RMSE	LOG10	PSNR
Encoder module	No SLDEM	0.7276	0.8976	0.9540	0.2408	4.9484	0.083	21.8864
	ResNet-based	0.8995	0.9744	0.9909	0.1048	2.3309	0.0402	29.3058
Decoder module	UPB	0.9092	0.9779	0.9917	0.1024	2.1776	0.0381	29.6611
Loss function	l_1 loss	0.8929	0.9718	0.9894	0.1120	2.3978	0.0417	28.9389
ELunarDTMNet		**0.9109**	**0.9787**	**0.9924**	**0.0964**	**2.1654**	**0.0375**	**29.9208**

The best results are in **bold**.

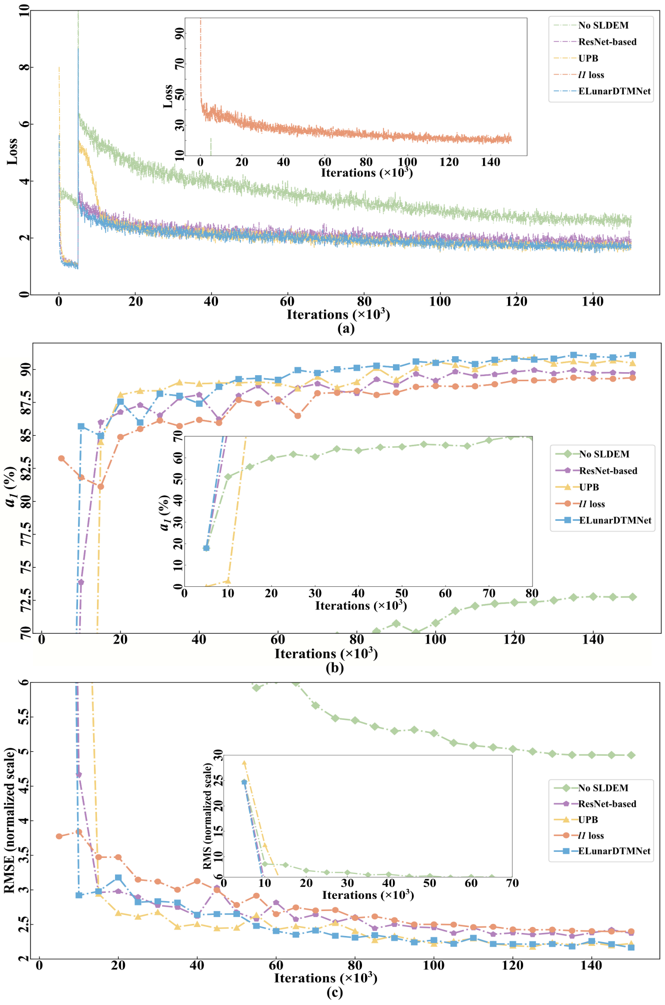


Fig. 16. The training and validation logs of the ablation study based on our proposed DTM normalization strategy. (a) Training curves, (b) and (c) metrics a_1 and RMSE on the validation set every 5k iterations, respectively. 'No SLDEM' indicates training the network without using SLDEM as input. 'ResNet-based' indicates using ResNet-based instead of Swin-T-based backbone to extract features from images. 'UPB' indicates using UPB instead of our proposed residual-connected mechanism to construct the decoder module. '$l1$ loss' indicates using $l1$ loss instead of our proposed mean-normalized loss function to train the network. 'ELunarDTMNet' represents using our proposed method to train the network. The max-min DTM normalization strategy is not compared in this context as it uses a different normalization range, resulting in inconsistent metric scales.

in terms of accuracy. While some variations, such as the UPB mechanism on the decoder module, show good performance, our proposed ELunarDTMNet outperforms them all with the lowest MAE and RMSE across all three areas.

In addition, we provide a detailed comparison in local areas to evaluate the effectiveness of our proposed DTM normalization strategy, the Swin-T-based backbone for the image encoder branch, and the mean-normalized loss function in fine topography retrieval. Fig. 17 compares the max-min DTM normalization strategy with our proposed elevation-statistics-based strategy. Fig. 17a and Fig. 17b present the areas featuring multi-scale craters. The HS maps show that the application of the max-min strategy does not effectively recover some small-scale craters. Notably, the crater with a diameter of 300 m, indicated by the yellow circles, displays artifacts rather than accurate details under the max-min normalization strategy. Moreover, this strategy also fails to effectively capture the edge of the flat impact melt at the center of Fig. 17c, as well as the subtle textures surrounding it. Conversely, our proposed strategy performs well in capturing these details with improved accuracy as shown in the RE maps.

Fig. 18 illustrates a comparison between the ResNet-based backbone used in DLunarDTMNet and our Swin-T-based backbone. While the ResNet-based backbone demonstrates good performance in reconstructing the shapes of most craters in Fig. 18a, it displays some artifacts, as marked by the yellow arrows. In contrast, the Swin-T-based backbone maintains consistency with the NAC images and does not exhibit these artifacts. In the left ellipse of Fig. 18b, the result using the ResNet-based backbone displays evident indications of over-smoothing. The right ellipse in Fig. 18b highlights the region where the reconstructed shapes by the ResNet-based backbone show discrepancies with the NAC image. The region, indicated by the yellow ellipse in Fig. 18c, displays albedo variances as topographic relief. Overall, the results derived from the Swin-T-based backbone align more consistently with the NAC images. This is further evidenced by the RE maps. Our method demonstrates better accuracy than that obtained by the ResNet-based backbone.

Fig. 19 depicts the comparison between $l1$ loss and our proposed mean-normalized loss function in penalizing differences between predicted and true values. The HS map obtained with the $l1$ loss training model in Fig. 19a exhibits discrepancies from the NAC image for craters with diameters of several hundred meters. In contrast, our proposed loss function effectively guides the model in predicting the shapes of these craters. Fig. 19b and Fig. 19c highlight that while

TABLE VI
COMPARISON OF RECONSTRUCTION ERRORS FROM WELL-TRAINED MODELS FOR EACH VARIATION IN THREE TESTING AREAS.

Area name	Method	RE < 2m (%)	RE < 4m (%)	RE < 10m (%)	MAE (m)	RMSE (m)
Apollo 11	Max-min strategy	55.52	85.73	99.62	2.18	2.84
	No SLDEM	42.59	72.84	98.59	2.94	3.80
	ResNet-based	54.32	85.63	99.59	2.21	2.87
	UPB	55.50	85.93	**99.65**	2.16	2.80
	$l1$ loss	55.83	86.36	99.63	2.17	2.82
	ELunarDTMNet	**57.33**	**87.16**	**99.65**	**2.10**	**2.73**
Imbrium	Max-min strategy	65.66	90.80	**99.86**	1.80	2.38
	No SLDEM	47.53	73.02	92.94	3.53	5.70
	ResNet-based	63.90	88.76	99.66	1.92	2.61
	UPB	**67.47**	91.45	99.79	**1.74**	2.34
	$l1$ loss	64.50	89.89	99.80	1.85	2.47
	ELunarDTMNet	67.03	**91.72**	99.82	**1.74**	**2.32**
Lichtenberg	Max-min strategy	26.01	53.76	97.11	4.08	4.91
	No SLDEM	15.70	31.19	66.17	11.31	19.65
	ResNet-based	20.93	45.31	95.13	4.67	5.52
	UPB	25.15	53.11	97.61	4.08	4.86
	$l1$ loss	23.04	49.06	95.93	4.41	5.27
	ELunarDTMNet	**27.06**	**55.70**	**97.97**	**3.91**	**4.69**

The best results are in **bold**. The second-best results are underlined.

the $l1$ loss-trained model can predict local albedo variances as craters, it struggles to capture subtle surface features, for example, the elephant hide features. Conversely, the proposed approach can capture these features, aligning closely with the NAC images. The RE maps indicate enhanced accuracy in our results, illustrating the effectiveness of our proposed loss function in optimizing the network performance.

C. Accuracy analysis for the large bright slopes

In this study, the coarse-resolution SLDEM is used as the input DTM, which differs from the high-resolution SPG DTM, as shown in Fig. 20. Overall, our predicted DTM reduces the discrepancy to the SPG DTM; however, the bright slopes (sun-facing sides of the terrain) exhibit larger errors compared to the shaded slopes. Further investigations will focus on understanding and reducing this effect, thus improving accuracy within areas with extended bright slopes.

VII. CONCLUSION

In this paper, we have proposed an efficient single-view DL-based method for high-quality DTM reconstruction of the lunar surface from a high-resolution optical image captured by an orbiter constrained by a coarse-resolution DTM (ELunarDTMNet). It first incorporates the Swin-T architecture into the image encoder branch to improve the model's ability to capture multi-scale features. In the decoder module, we introduce a new residual-connected mechanism to improve prediction accuracy. Moreover, we implement an elevation-statistics-based DTM normalization strategy to preserve terrain feature contrast, and a mean-normalized loss function to accommodate the complex elevation distribution of the lunar surface. Extensive comparative experiments demonstrate that our proposed enhancements lead to an effective improvement in network performance, yielding the following results:

1) ***High-quality terrain feature retrieval***: The proposed ELunarDTMNet demonstrates superior sensitivity in the retrieval of multi-scale features, including irregular terrains and regions with significant relief, outperforming state-of-the-art single-view DL methods. It can generate fine terrain details comparable to the SFS method [8], with improved elevation accuracy and robustness under different illuminations. The medium- and large-scale terrains in our results demonstrate a higher level of consistency with the SPG method [7], while effectively mitigating artifacts that may be present in the SPG DTMs.
2) ***Optimized processing speed and generalization capability***: While DLunarDTMNet demonstrated a processing speed advantage in reconstructing DTM mosaics compared to the SFS method [8], [26], the proposed ELunarDTMNet further significantly reduces processing speed by simplifying the DTM mosaic generation process. Our method is evaluated on diverse datasets and demonstrates strong generalization across different types of lunar optical imagery, such as NAC and Chang'E-2 orbiter images.

In the future, we will explore the potential of more DL models, such as the Diffusion Model [47], [48], for lunar DTM reconstruction. We will also focus on high-quality DTM reconstruction for challenging areas, such as the lunar South Pole region, and apply the proposed method to relevant scientific analyses and engineering tasks. Additionally, we will explore the feasibility of applying our method to shape modeling of other celestial bodies.

ACKNOWLEDGMENTS

The authors would like to thank the LRO Mission Operation Center, LROC Science Operations Center, and Chang'E-2 mission team for their great work in producing and maintaining the data used in this study.

DATA AVAILABILITY

The NAC images and SPG-derived products are available at https://www.lroc.asu.edu/. SLDEM and LOLA profiles can be accessed from the https://ode.rsl.wustl.edu/moon/index.aspx. The Chang'E-2 SPG-derived DTMs and ORIs are available at https://moon.bao.ac.cn/. The MADNet-derived DTMs

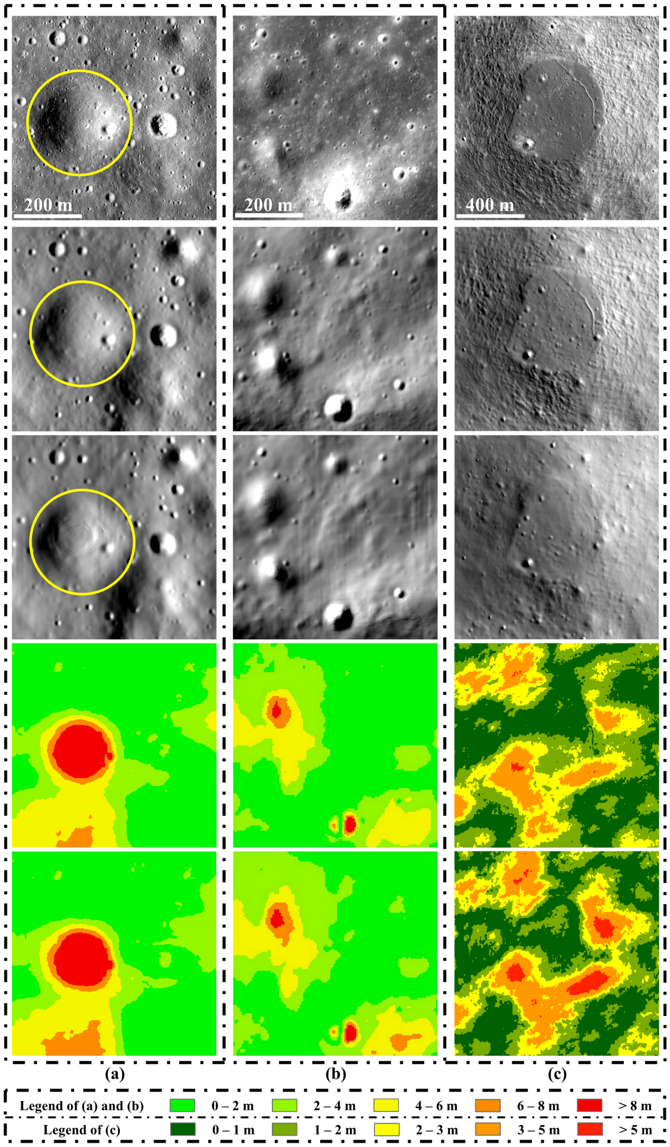


Fig. 17. Performance comparison using different DTM normalization strategies in three local areas. Areas (a) centered at (23.45°, 0.70°), (b) centered at (334.25°, 27.86°), and (c) centered at (167.28°, 42.50°). From top to bottom, the images show: the NAC images, the HS maps using the proposed and the max-min normalization strategies, and the RE maps using the proposed and the max-min normalization strategies. The yellow circles indicate the crater where artifacts appear with the max-min strategy.

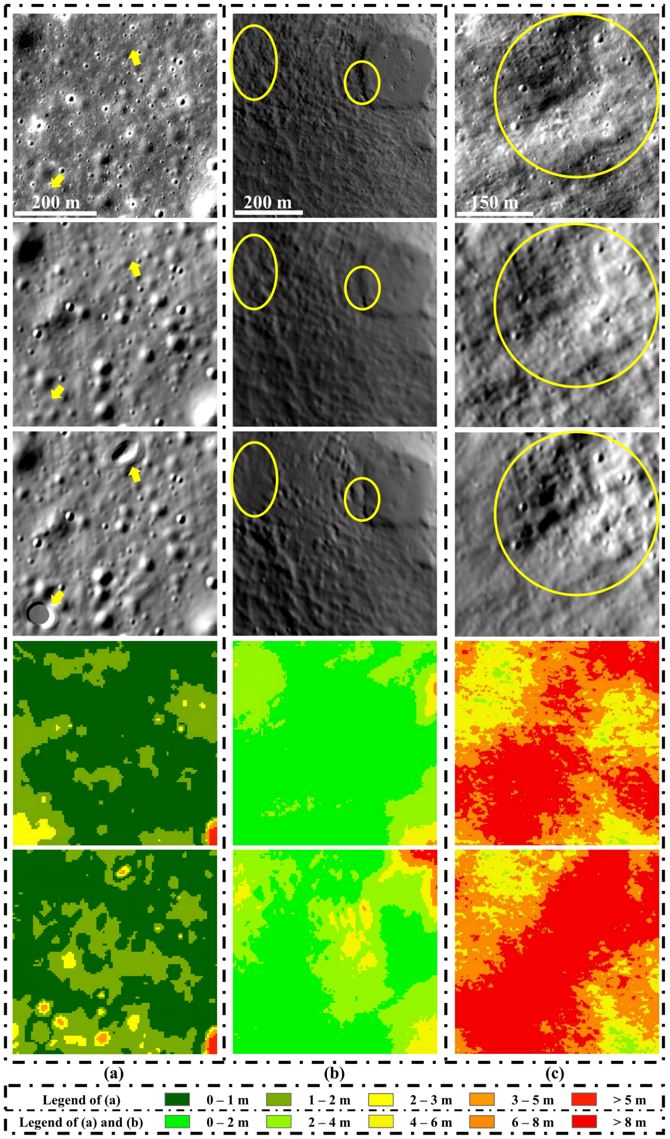


Fig. 18. Performance comparison using different image feature extraction backbones in three local areas. Areas (a) centered at (334.34°, 27.97°), (b) centered at (167.33°, 42.42°), and (c) centered at (292.46°, 31.92°). From top to bottom, the figures show: the NAC images, the HS maps using our Swin-T-based backbone and the ResNet-based backbone as employed in DLunarDTMNet, and the RE maps using our Swin-T-based backbone and the ResNet-based backbone. The yellow arrows indicate artifacts reconstructed by the ResNet-based backbone. The yellow ellipses highlight areas that exhibit over-smoothing or inconsistencies with the NAC images using the ResNet-based backbone.

are available at https://www.cosmos.esa.int/web/psa/ucl-mssl_moon_von_karman_v1.0.

References

[1] M. A. Kreslavsky, J. W. Head, G. A. Neumann, M. A. Rosenburg, O. Aharonson, D. E. Smith, and M. T. Zuber, "Lunar topographic roughness maps from lunar orbiter laser altimeter (lola) data: Scale dependence and correlation with geologic features and units," *Icarus*, vol. 226, no. 1, pp. 52–66, 2013.

[2] E. Speyerer, S. Lawrence, J. Stopar, P. Gläser, M. Robinson, and B. Jolliff, "Optimized traverse planning for future polar prospectors based on lunar topography," *Icarus*, vol. 273, pp. 337–345, 2016.

[3] W. Cao, Z. Xiao, F. Luo, Y. Ma, H. Ouyang, and R. Xu, "Emplacement mechanism of ponded light plains on the moon: Insight from topography roughness," *Icarus*, vol. 415, p. 116071, 2024.

[4] M. S. Menon, A. Kothandhapani, N. S. Sundaram, V. Raghavan, and S. Nagaraj, "Terrain-based analysis as a design and planning tool for operations of a lunar exploration rover for the teamindus lunar mission," in *2018 SpaceOps Conference*, 2018, p. 2494.

[5] D. De Rosa, B. Bussey, J. T. Cahill, T. Lutz, I. A. Crawford, T. Hackwill, S. van Gasselt, G. Neukum, L. Witte, A. McGovern *et al.*, "Characterisation of potential landing sites for the european space agency's lunar lander project," *Planetary and space science*, vol. 74, no. 1, pp. 224–246, 2012.

[6] P. Gläser, J. Oberst, G. Neumann, E. Mazarico, E. Speyerer, and M. Robinson, "Illumination conditions at the lunar poles: Implications for future exploration," *Planetary and Space Science*, vol. 162, pp. 170–178, 2018.

[7] M. Henriksen, M. Manheim, K. Burns, P. Seymour, E. Speyerer, A. Deran, A. Boyd, E. Howington-Kraus, M. R. Rosiek, B. A. Archinal *et al.*, "Extracting accurate and precise topography from lroc narrow

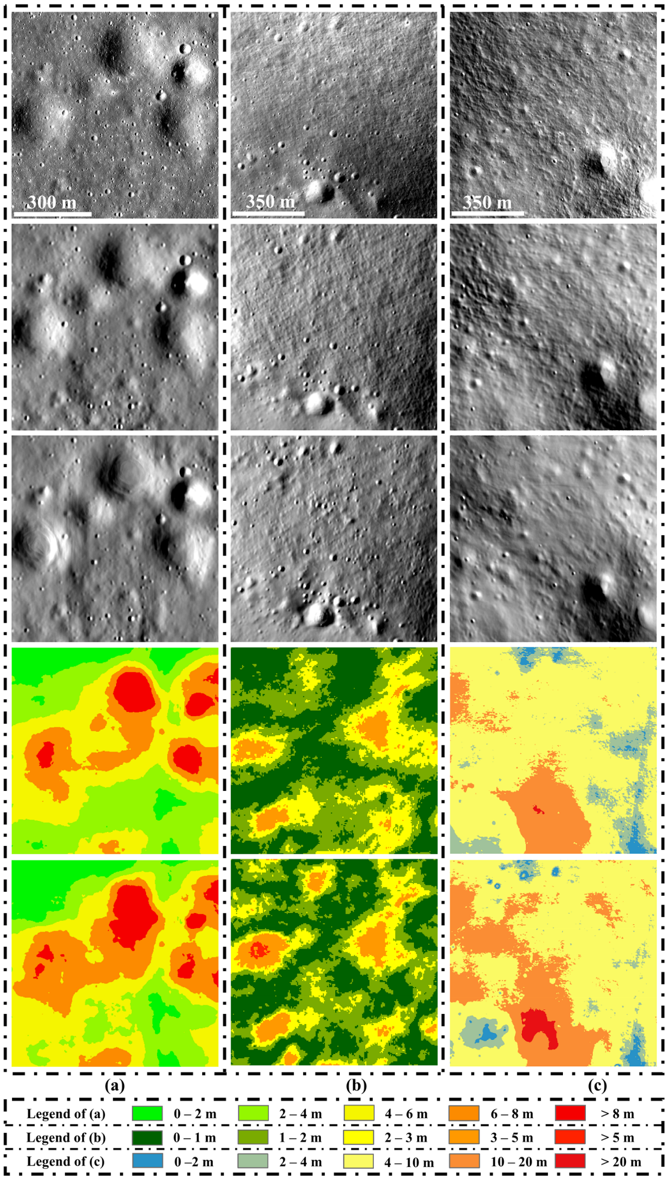

300 m
350 m
350 m
(a)
(b)
(c)
Legend of (a) 0 – 2 m 2 – 4 m 4 – 6 m 6 – 8 m > 8 m
Legend of (b) 0 – 1 m 1 – 2 m 2 – 3 m 3 – 5 m > 5 m
Legend of (c) 0 –2 m 2 – 4 m 4 – 10 m 10 – 20 m > 20 m

Fig. 19. Performance comparison using our proposed mean-normalized loss and $l1$ loss in three local areas. Areas (a) centered at (23.46°, 0.66°), (b) centered at (292.48°, 32.04°), and (c) centered at (167.24°, 42.13°). From top to bottom, the images show: the NAC images, the HS maps using the mean-normalized and the $l1$ losses, and the RE maps using the mean-normalized and the $l1$ losses.

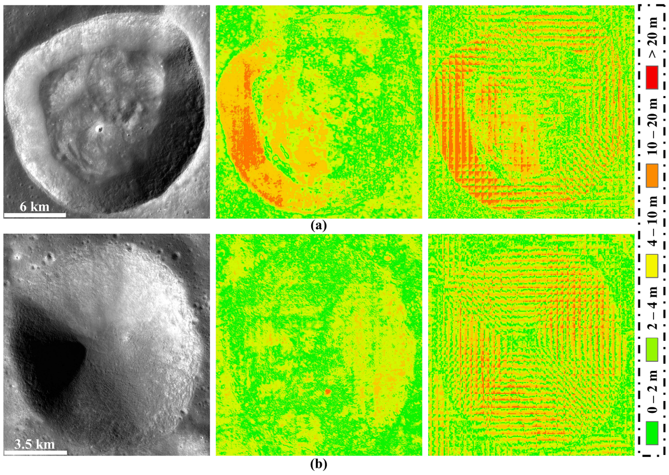


Fig. 20. Accuracy analysis for the large bright slopes in the Lichtenberg (a) and Hponds (b) regions. From left to right, the images show: the NAC image, the RE map between our DTM and SPG DTM, and the RE map between SLDEM and SPG DTM.

angle camera stereo observations," *Icarus*, vol. 283, pp. 122–137, 2017.

[8] O. Alexandrov and R. A. Beyer, "Multiview shape-from-shading for planetary images," *Earth and Space Science*, vol. 5, no. 10, pp. 652–666, 2018.

[9] R. L. Kirk, E. Howington-Kraus, M. R. Rosiek, J. A. Anderson, B. A. Archinal, K. J. Becker, D. Cook, D. M. Galuszka, P. E. Geissler, T. M. Hare *et al.*, "Ultrahigh resolution topographic mapping of mars with mro hirise stereo images: Meter-scale slopes of candidate phoenix landing sites," *Journal of Geophysical Research: Planets*, vol. 113, no. E3, 2008.

[10] F. Preusker, J. Oberst, J. W. Head, T. R. Watters, M. S. Robinson, M. T. Zuber, and S. C. Solomon, "Stereo topographic models of mercury after three messenger flybys," *Planetary and Space Science*, vol. 59, no. 15, pp. 1910–1917, 2011.

[11] F. Scholten, J. Oberst, K.-D. Matz, T. Roatsch, M. Wählisch, E. Speyerer, and M. Robinson, "Gld100: The near-global lunar 100 m raster dtm from lroc wac stereo image data," *Journal of Geophysical Research: Planets*, vol. 117, no. E12, 2012.

[12] A. N. Stein, A. Huertas, and L. Matthies, "Attenuating stereo pixel-locking via affine window adaptation," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.* IEEE, 2006, pp. 914–921.

[13] Q. Yang and N. Ahuja, "Stereo matching using epipolar distance transform," *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4410–4419, 2012.

[14] M. S. Robinson, S. Brylow, M. e. Tschimmel, D. Humm, S. Lawrence, P. Thomas, B. W. Denevi, E. Bowman-Cisneros, J. Zerr, M. Ravine *et al.*, "Lunar reconnaissance orbiter camera (lroc) instrument overview," *Space science reviews*, vol. 150, pp. 81–124, 2010.

[15] B. Wu, W. C. Liu, A. Grumpe, and C. Wöhler, "Construction of pixel-level resolution dems from monocular images by shape and albedo from shading constrained with low-resolution dem," *ISPRS journal of photogrammetry and remote sensing*, vol. 140, pp. 3–19, 2018.

[16] A. Grumpe, F. Belkhir, and C. Wöhler, "Construction of lunar dems based on reflectance modelling," *Advances in Space Research*, vol. 53, no. 12, pp. 1735–1767, 2014.

[17] H. Chen, X. Hu, P. Gläser, H. Xiao, Z. Ye, H. Zhang, X. Tong, and J. Oberst, "Cnn-based large area pixel-resolution topography retrieval from single-view lroc nac images constrained with sldem," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 9398–9416, 2022.

[18] Z. Chen, B. Wu, and W. C. Liu, "Mars3dnet: Cnn-based high-resolution 3d reconstruction of the martian surface from single images," *Remote Sensing*, vol. 13, no. 5, p. 839, 2021.

[19] R. La Grassa, I. Gallo, C. Re, G. Cremonese, N. Landro, C. Pernechele, E. Simioni, and M. Gatti, "An adversarial generative network designed for high-resolution monocular depth estimation from 2d hirise images of mars," *Remote Sensing*, vol. 14, no. 18, p. 4619, 2022.

[20] L. Yang, Z. Zhu, L. Sun, and D. Zhang, "Global attention-based dem: A planet surface digital elevation model-generation method combined with a global attention mechanism," *Aerospace*, vol. 11, no. 7, p. 529, 2024.

[21] H. Chen, X. Hu, and J. Oberst, "Pixel-resolution dtm generation for the lunar surface based on a combined deep learning and shape-from-shading (sfs) approach," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 3, pp. 511–516, 2022.

[22] Y. Tao, S. Xiong, S. J. Conway, J.-P. Muller, A. Guimpier, P. Fawdon, N. Thomas, and G. Cremonese, "Rapid single image-based dtm estimation from exomars tgo cassis images using generative adversarial u-nets," *Remote Sensing*, vol. 13, no. 15, p. 2877, 2021.

[23] Y. Liu, Y. Wang, K. Di, M. Peng, W. Wan, and Z. Liu, "A generative adversarial network for pixel-scale lunar dem generation from high-resolution monocular imagery and low-resolution dem," *Remote Sensing*, vol. 14, no. 21, p. 5420, 2022.

[24] Y. Tao, J.-P. Muller, S. J. Conway, S. Xiong, S. H. Walter, and B. Liu, "Large area high-resolution 3d mapping of the von kármán crater: Landing site for the chang'e-4 lander and yutu-2 rover," *Remote Sensing*, vol. 15, no. 10, p. 2643, 2023.

[25] Y. Tao, J.-P. Muller, S. Xiong, and S. J. Conway, "Madnet 2.0: Pixel-scale topography retrieval from single-view orbital imagery of mars using deep learning," *Remote Sensing*, vol. 13, no. 21, p. 4220, 2021.

[26] R. A. Beyer, O. Alexandrov, and S. McMichael, "The ames stereo pipeline: Nasa's open source software for deriving and processing terrain data," *Earth and Space Science*, vol. 5, no. 9, pp. 537–548, 2018.

[27] B. D. Boatwright and J. W. Head, "Shape-from-shading refinement of lola and lroc nac digital elevation models: Applications to upcoming human and robotic exploration of the moon," *The Planetary Science Journal*, vol. 5, no. 5, p. 124, 2024.

[28] R. Jia, B. Wu, W. C. Liu, Y. Peng, S. Krasilnikov, L. Sheng, and S. Peng, "Shadow-constrained shape-from-shading for pixel-wise 3d surface reconstruction at the lunar south pole," *Geo-spatial Information Science*, pp. 1–19, 2024.

[29] Z. Chen, B. Wu, and W. C. Liu, "Deep learning for 3d reconstruction of the martian surface using monocular images: A first glance," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 1111–1116, 2020.

[30] J. Cao, R. Huang, Z. Ye, Y. Xu, and X. Tong, "Generative 3d reconstruction of martian surfaces using monocular images," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 48, pp. 51–56, 2024.

[31] J.-P. Muller, Y. Tao, and S. H. Walter, "Digital terrain models of the nasa artemis sites from single lroc-nac images using the ucl madnet retrieval system," in *European Planetary Science Congress*. Berlin, Germany, 2024, pp. EPSC2024–1170.

[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[35] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[36] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[38] C. Zhang, L. Wang, S. Cheng, and Y. Li, "Swinsunet: Pure transformer network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[39] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 239–248.

[40] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[41] M. Barker, E. Mazarico, G. Neumann, M. Zuber, J. Haruyama, and D. Smith, "A new lunar digital elevation model from the lunar orbiter laser altimeter and selene terrain camera," *Icarus*, vol. 273, pp. 346–355, 2016.

[42] R. Franke, "Smooth interpolation of scattered data by local thin plate splines," *Computers & mathematics with applications*, vol. 8, no. 4, pp. 273–281, 1982.

[43] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[44] P. Gläser, I. Haase, J. Oberst, and G. Neumann, "Co-registration of laser altimeter tracks with digital terrain models and applications in planetary science," *Planetary and Space Science*, vol. 89, pp. 111–117, 2013.

[45] M. K. Barker, E. Mazarico, G. A. Neumann, D. E. Smith, M. T. Zuber, J. W. Head, and X. Sun, "A new view of the lunar south pole from the lunar orbiter laser altimeter (lola)," *The Planetary Science Journal*, vol. 4, no. 9, p. 183, 2023.

[46] X. Ren, J. Liu, F. Wang, W. Wang, L. Mu, and H. Li, "A new lunar global topographic map products from chang'e-2 stereo camera image data," in *European Planetary Science Congress*, vol. 9. Centro de Congressos do Estoril Cascais, Portugal, 2014, pp. EPSC2014–344.

[47] D. Zheng, X.-M. Wu, Z. Liu, J. Meng, and W.-s. Zheng, "Diffuvolume: Diffusion model for volume based stereo matching," *arXiv preprint arXiv:2308.15989*, 2023.

[48] Y. Duan, X. Guo, and Z. Zhu, "Diffusiondepth: Diffusion denoising approach for monocular depth estimation," *arXiv preprint arXiv:2303.05021*, 2023.