C-PLES: Contextual Progressive Layer Expansion with Self-attention for Multi-class Landslide Segmentation on Mars using Multimodal Satellite Imagery

Abel A. Reyes$^1$, Sidike Paheding$^1$, A. Rajaneesh$^2$, K.S. Sajinkumar$^1$, Thomas Oommen$^1$

$^1$Michigan Technological University  $^2$University of Kerala

{areyesan, spahedin, skochapp, toomme}@mtu.edu  rajaneesh@keralauniversity.ac.in

Abstract

Landslide segmentation on Earth has been a challenging computer vision task, in which the lack of annotated data or limitation on computational resources has been a major obstacle in the development of accurate and scalable artificial intelligence-based models. However, the accelerated progress in deep learning techniques and the availability of data-sharing initiatives have enabled significant achievements in landslide segmentation on Earth. With the current capabilities in technology and data availability, replicating a similar task on other planets, such as Mars, does not seem an impossible task anymore. In this research, we present C-PLES (Contextual Progressive Layer Expansion with Self-attention), a deep learning architecture for multi-class landslide segmentation in the Valles Marineris (VM) on Mars. Even though the challenges could be different from on-Earth landslide segmentation, due to the nature of the environment and data characteristics, the outcomes of this research lead to a better understanding of the geology and terrain of the planet, in addition, to providing valuable insights regarding the importance of image modality for this task. The proposed architecture combines the merits of the progressive neuron expansion with attention mechanisms in an encoder-decoder-based framework, delivering competitive performance in comparison with state-of-the-art deep learning architectures for landslide segmentation. In addition to the new multi-class segmentation architecture, we introduce a new multi-modal multi-class Martian landslide segmentation dataset for the first time. The dataset will be available at https://github.com/MAIN-Lab/C-PLES

1. Introduction

Landslides are a movement of terrain caused by different factors, including earthquakes, volcanic activity, and heavy rainfalls [17]. Therefore, understanding the cause that triggers their formations help us to unravel the morphological changes in a planet. The study of extraterrestrial terrains has been limited. The current availability of satellite imagery and cutting-edge deep learning (DL) algorithms opens the opportunity to exploit their capabilities for landslide detection outside of Earth. Since Mars has shown evidence that presumes a similar past that the current conditions on Earth, the study of the Martian terrain could have a potential contribution to estimating the transitions related to climate change that could be faced on Earth in a near future [13,31]. The vast presence of landslides on the surface of Mars brings us the possibility to obtain valuable insights from their analysis. With the use of aerial imagery, a typical method to identify landslides on Mars requires the visual interpretation of optical images by geomorphologist analysis [5], a time-consuming procedure that could highly get bias the expertise of the analyst expert. In addition, this procedure based the criteria on similarities with terrestrial landslide [6], which includes shape, size, tone, mottling, texture, pattern of objects, and site topography [3,21]. However, new studies suggested the inclusion of an additional set of criteria to recognize landslides in areas such as the Valles Marineris (VM) region [12], which is considered as a museum of landslides on Mars [38]. Characteristics such as length, width, relative relief, slope statistics, slope of the scarps, and geology are mentioned in addition to an empirical relationship between the volume and the area of the slides. The aforementioned
methods rely upon the use of traditional machine learning algorithms, (i.e. logistic regression), to develop a landslide classifier based on the extracted features. With recent advancements in DL algorithms, several studies have achieved state-of-the-art results in computer vision tasks such as image classification [14,49], object detection [18,33], and semantic segmentation [11,39]. However, there is still a gap between the implementation of cutting-edge techniques and the adoption of those techniques for specific tasks such as landslide segmentation. The use of automated feature extraction, provided by a DL architecture, eliminates any human bias during the feature extraction process and grants an end-to-end framework for landslide analysis, as it is shown in the adoption of Vision Transformers [15] for landslide classification task [27].

In this research work, as shown in Figure 1, we propose an end-to-end DL architecture to facilitate the task of identifying different types of landslides. To this end, the VM region is selected to build a dataset for landslide segmentation and to develop a model to automate this task efficiently. The proposed end-to-end Martian multi-class landslide segmentation model fuses heterogeneous multi-modality imagery data in a contextually aware manner. The main contributions of this research are summarized as follows:

- A new multi-modal multi-class dataset for the Martian landslide segmentation is introduced. The dataset is composed of three different versions, each version with a different image size, which will be publicly available to the research community.

- To the best of our knowledge, this study is the first attempt to propose an end-to-end DL framework for multi-class Martian landslide segmentation from satellite images. We named our proposed model C-PLES (Contextual Progressive Layer Expansion with Self-attention). Experimental results reveal effectiveness in the performance of C-PLES when compared to state-of-the-art DL architectures. In addition, the ablation study is conducted to support the significance of different components of C-PLES to reach this performance.

- The impact of different image modalities on the contribution of landslide segmentation accuracy is systematically analyzed. Experimental results indicate that imagery modalities, such as the Thermal Emission Imaging System (THEMIS) and Digital Elevation Model (DEM), have the most significant contribution to the segmentation accuracy in terms of Intersection over Union (IoU).

The rest of this paper is organized as follows. Section 2 reviews the landslide detection algorithms. Section 3 describes the proposed C-PLES architecture in detail. Section 4 introduces newly built datasets used in this study. Section 5 provides the experimental results. Section 6 summarizes the component contribution of the model throughout the ablation study. Section 7 concludes our findings in this study along with future research.

2. Related Work

To the best of our knowledge, we have not come across any published research that uses an end-to-end DL framework for multi-class Martian landslide segmentation. As a result, in this section, we discuss the most pertinent machine learning techniques for on-Earth landslides segmentation.

Traditional machine learning algorithms have been widely used in combination with DL and object-based image analysis (OBIA) techniques for landslide segmentation tasks [1,28,44]. For instance, Tavakkoli et al. [44] experimented OBIA with a set of machine learning methods to perform landslide detection, within a stacking machine learning framework. This method utilized random forest, logistic regression, and multilayer perceptron neural network as the set of selected machine learning algorithms (Level 0 of the stacking model), and logistic regression as the meta-learner (Level 1 of the stacking model) to make the final prediction. Overall, they reported the approach resulted in appropriate landslide detection. Keyport et al. [28] reported a comparative analysis for pixel-based landslide detection with the use of very high-resolution (VHR) remote sensing aerial imagery. They performed both a pixel-based and object-oriented analysis (OOA) for landslide mapping, in which the OOA method yielded better results with the presence of less number of false positives. Achariyaviriya et al. [1] utilized a DL approach by combining three versions of the ResNet [22] models, in which each model take a single data modality as an input, such as RGB color image, normalized difference vegetation index (NDVI), and slope factor (SP). Initially, Each ResNet model and the corresponding data modality are trained to generate a set of features per modality. Then, those sets of features are combined to train a final decision tree classifier model. As a result, the use of different modalities, rather than just RGB or gray-scale, helped to improve the performance of their proposed classifier.

The study conducted by Prakash et al. [37] used CNN as a semantic segmentation problem to map landslides. The researchers utilized high-resolution Lidar DEM and cloud-free optical images from Sentinel-2 for mapping. Pixel-based, object-based, and DL methods were used for generating landslide susceptibility maps. The study introduced CNN-based U-Net [40] and ResNet architectures for mapping landslides. The U-Net architecture was used for semantic segmentation, and ResNet was used for feature identification. The authors demonstrated that the U-Net with ResNet strategy outperforms pixel-based and object-based machine learning algorithms on a regional scale for map-
ping larger landslides. However, all three approaches were unsuccessful in detecting minor landslides, and they had difficulty distinguishing individual landslides when they were close together. In summary, there is evidence that supports the use of multimodal inputs to improve the performance of landslide segmentation on Earth. However, it is worth mentioning that the contrast of the terrain and vegetation provides visual characteristics that eventually help to detect a landslide. In addition, another limitation lies in the presence of minor landslides, in which the models tend to fail to separate among individual landslides.

3. Methodology

The proposed C-PLES architecture for multi-class Martian landslides segmentation is illustrated in Figure 2. This architecture follows the pattern of an encoder-decoder structure, in which several components are incorporated to effectively extract pertinent features for multi-class landslide segmentation. Detailed explanations of different components in the C-PLES are described as follows.

3.1. Encoder-decoder network

An encoder-decoder architecture is a common approach used in semantic segmentation tasks [4,8,50]. This structure exploits multi-scale visual features extracted in the encoder side and recovers the spatial resolution in the decoder side [10]. However, this arrangement by itself tends to lose a certain range of contextual information. The encoder compresses the input into a lower-dimensional representation, which discards some of the spatial information of the original input [29], thus producing segmentation outputs with a lack of details.

3.2. Progressive Neuron Expansion

The concept of Progressive Neuron Expansion (PNE) is adopted from the work presented in [36,42], in which each neuron from the input is progressively expanded following a Maclaurin series expansion of a nonlinear function. Every node, denoted as $S_u$, produced by the expansion is the result of the addition of the $u$ subsequent component in the expansion. $u$ represents a hyperparameter that controls the length (i.e., the number of terms in the Maclaurin series) of the expansion. The series expansion is mathematically expressed as follows:

$$S_u = \sum_{n=1}^{u} c_n x^{p_n} = c_1 x^{p_1} + c_2 x^{p_2} + \ldots + c_u x^{p_u} \quad (1)$$

where $x$ represents the input neuron to be expanded, $c \in \{1, 1/2, 1/3\}$ and $p \in \{1, 2, 3\}$ are the non-trainable coefficients and powers in the Maclaurin series expansion of a nonlinear function, respectively.

3.3. Attention mechanisms

Attention mechanism in DL refers to a family of techniques that allow a neural network to focus on relevant parts of the input data, which is achieved by enabling the model to selectively attend to certain regions of an image, sequence, or another type of input while ignoring the remaining [20,24,34]. In this research, we explore the benefit of two particular attention mechanisms: dot-product attention and multi-head self-attention.

Dot-product attention, also known as Luong-style attention [32], is utilized, in this study, to leverage the production of enriched features from the PNE layer. This attention mechanism is typically used for neural machine transla-
tion [23,41,48]. The Luong-style attention involves a global attention model that considers all the sources in the input when generating the alignment scores and the context vector. To compute the alignment scores, an alignment model $a_i(s)$ is used, expressed as:

$$a_i(s) = align(h_s, S_i)$$  

(2)

where $h_s$ represents the current state to compute the alignment scores, and it is replaced by the $S_i$ term from the PNE expansion. Meanwhile, the $S_i$ term, $i \in [2 : u]$, represents the remaining terms on the expansion to be aligned with the source. The dot-product-based alignment model is mathematically expressed as

$$a_i(h_s, S_i) = h_s^T W_a S_i$$  

(3)

where $W_a$ is a trainable weight matrix. The computed alignment annotations are used to generate a context vector, expressed as

$$c_t = \sum a_i S_i$$  

(4)

And finally, the attention is computed based on a weighted concatenation as follows

$$\tilde{S}_i = tanh(W_c[c_t; S_i])$$  

(5)

where $W_c$ is the trainable concatenation weight matrix.

On the other hand, the multi-head self-attention [46] is characterized by using the scaled dot-product attention mechanism for similarity estimations between the Key, Query, and Value matrices. In our proposed model, they are replaced by the expansion terms $S_1$, $S_2$, and $S_3$, respectively, as described below

$$Attention(S_1, S_2, S_3) = softmax(\frac{S_1 S_2^T}{\sqrt{d_{S_2}}}) S_3$$  

(6)

where $d$ is the dimension of $S_2$ used to properly scale the dot product between $S_1$ and $S_2$. In addition, the multi-head self-attention can be summarized as a generalized version of the aforementioned self-attention mechanism repeated in parallel $h$ times, in which $h$ represent the number of heads with different learnable projections. This is mathematically formulated as follows:

$$MultiHead(S_1, S_2, S_3) = Concat(head_1, ..., head_h) W^O$$  

(7)

where $W^O$ is the learnable weight matrix for the output, and

$$head_i = Attention(S_1 W_1^{S_1}, S_2 W_1^{S_2}, S_3 W_1^{S_3})$$  

(8)

where $W_i^{S_1}$, $W_1^{S_2}$, $W_1^{S_3}$ represent the learned projections for $S_1$, $S_2$, and $S_3$, respectively.

### 3.4. C-PLES: The Context Progressive Layer Expansion

The proposed DL architecture for multi-class landslide segmentation introduces a context enrichment Block, named C-PLES, that leverages the merits of the extracted features through the use of the PNE layers and the combination of global attention and local context, given by the dot attention mechanism and the dilated convolution, respectively. By the use of the C-PLES, we aimed to address the issue of context loss inherent in the encoder-decoder architectures. The C-PLES block is detailed and illustrated in Figure 2, and it is composed of an initial 2D dilated convolution. A dilated convolution increases the receptive field of the visual feature extractor and provides more contextual information from the encoder to map it with the decoder [47]. Then, a PNE layer takes the output of the dilated convolution to produce the expansion of each hidden neuron (i.e., $S_1$, $S_2$, and $S_3$ as we used three terms in Maclaurin series expansion). This expansion enriches the visual feature representation, and the different numerical characteristics comprised by each term from the expansion generate a suitable scenario for the usage of attention mechanisms. Accordingly, the three nodes from the expansion are fed as the inputs for the attention layer (either dot-product attention or MHA). This attention layer helps the architecture to focus on a relevant part of the input. Considering computational efficiency, we select dot-product attention in our C-PLES block. In the end, the attention layer produced an attention mask that is multiplied by the original input of the C-PLES block. The output is then sent to the respective decoder. Mathematically, the C-PLES block can be represented as

<table>
<thead>
<tr>
<th>Table 1. Detailed structure of the proposed C-PLES architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong></td>
</tr>
<tr>
<td><strong>Encoding</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Bridge</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Decoding</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Output</strong></td>
</tr>
</tbody>
</table>
The C-PLES architecture combines the merits of an encoder-decoder architecture with the addition of a C-PLES block which serves to join each level of the encoder side with the corresponding level in the dilated 2D convolution operation. In addition, a dropout layer is also added to each set of convolutions to alleviate overfitting issues and stabilize the learning process during training.

In order to maximize the performance of the proposed C-PLES architecture, extensive experiments are performed to determine the best suitable loss function for the multi-class segmentation task. Since the dataset used in this study contains a significantly larger ratio of pixels belonging to the background, which makes the dataset heavily unbalanced, and thus we propose to fuse the focal loss and dice coefficient loss to effectively address this issue.

The focal loss [30] helps the model to effectively penalize easy and hard examples from imbalanced classes during the training process in a task such as image segmentation. Focal loss applies a modulating term \((1 - p_i)^\gamma\) to the regular cross-entropy loss. This modulating factor reduces the loss contribution from easy examples and focuses the learning on hard-miss classified examples. Mathematically, the focal loss function is expressed as follows:

\[
L_{\text{FocalLoss}} = - \sum_{i=1}^{n} (1 - p_i)^\gamma \log_b(p_i)
\]

where \(n\) is the number of samples and \(p\) is predicted probability. In order to address the class imbalance problem of positive and negative examples, a weighted parameter \((\alpha)\) is added as the inverse class frequency, extending the mathematical representation of the focal loss as:

\[
L_{\text{FocalLoss}} = - \sum_{i=1}^{n} \alpha_i (1 - p_i)^\gamma \log_b(p_i)
\]

In addition, dice coefficient loss (DCL) [26,43] was also explored, considering that the dice coefficient is a popular metric to calculate similarities between images. The DCL is mathematically expressed as:

\[
L_{\text{DCL}} = 1 - \frac{2 \sum_{i}^{N} p_i g_i + 1}{\sum_{i}^{N} p_i^2 + \sum_{i}^{N} g_i^2 + 1}
\]

where 1 is added to the numerator and denominator to avoid undefined cases for scenarios in which \(p = g = 0\).

We finally fuse both loss functions, denoted as \(L_{\text{Total}}\), in which a weight coefficient factor \(\beta\) is added to the focal loss to force the model to adjust the penalization for class imbalance cases. The \(L_{\text{Total}}\) can be formulated as

\[
L_{\text{Total}} = (1 - \beta)L_{\text{DCL}} + \beta L_{\text{FocalLoss}}
\]
4. Dataset

The multi-class Mars landslide segmentation dataset is an imagery inventory for the Valles Marineris region, a vast system canyon that runs along the Martian equator surface. The dataset comprises a total of 682 manually annotated landslides, distributed as follows: 133 debris flow (DF), 275 rock avalanches (RA), and 274 slumps (S).

The image data used are the RGB, greyscale, DEM, slope, and thermal data. RGB data used is the Mars Viking Colorized Global Mosaic, which has a spatial resolution of approximately 232 meters. It is a derivative of NASA’s Viking Mission to Mars. The grey scale data is Context Camera (CTX) imagery of the Mars Reconnaissance Orbiter (MRO) mission, having a spatial resolution of 5 m and bandwidth of 600-800 nm. The elevation data of the study area, which is the DEM, is Mars MGS MOLA-MEX HRSC Blended DEM with a spatial resolution of 200 m. The slope information is derived from the DEM using ESRI ArcGIS software and has a spatial resolution of 200 m. Thermal data is the thermal inertia of the Martian surface, generated from the Thermal Emission Imaging System (THEMIS) with a spatial resolution of 100 m.

Landslides annotation.

The landslides were classified based on visual interpretation using satellite imagery in a Geographic Information System (GIS) environment, which was based on typical morphological features like flow direction, the shape of scarp [38, 45], area of depletion, and accumulation [19], irregular and jumbled terrain, hummocks, and lateral levees. These landslides were digitized as polygons.

Patch Sample landslides. The multi-modal composited image is available as a TIF file, with a dimension of 10,245 by 4,098 pixels with a total of 7 channels. In addition, the dataset is also available as a set of patches in several resolutions (without overlapping), distributed as follows: 122, 240, and 955 patches of 256 × 256, 128 × 128, and 64 × 64 respectively with the corresponding segmentation mask. Figure 5 illustrates a set of 256 × 256 samples from the proposed multi-class Martian landslide multimodal dataset with the respective manually annotated segmentation mask.

5. Results and discussion

5.1. Experimental setup

We empirically demonstrate the effectiveness of the proposed C-PLES architecture using the multimodal Martian multi-class landslide segmentation dataset and compare its performance with several state-of-the-art deep network architectures, including U-Net [40], Attention U-Net [35], TransUNet [9], R2UNet [2], UNet3+ [25], UNet++ [51], and SwinUNet [7]. The performance of all models is evaluated in terms of the four standard assessment metrics: a) mean intersection over union (mIoU), b) precision, c) recall, and d) F1-score, all of which are typically used for landslide detection studies. All the experiments, including competing models, are run for a total of 150 epochs, with a batch size of 4, 8, and 16 samples for an input patch size of 256 × 256, 128 × 128, and 64 × 64, respectively. The L\textsubscript{total} is used as a loss function in all the experiments as well, with \( \beta \) set as 0.7 after a grid search. The training process is optimized with the Adam algorithm using a scheduled learning rate, which is initially set to 0.001 and reduced by a factor of 0.01 in every epoch. As part of the pipeline in the training process, a data augmentation stage is added with a rotation range of 90 degrees, horizontal and vertical flip activated with zoom and reflect fill mode. The number of steps per batch size is computed as the ratio between the number of samples by the corresponding batch size. In addition, the validation loss was monitored to save checkpoints of the models with the best performance during the training process. In our exper-

1https://nssdc.gsfc.nasa.gov/planetary/viking.html
5.2. Segmentation output performance comparison

For each image in the testing dataset, our model produced a segmentation mask that delineates the regions of interest of each type of landslide. The quantitative result to evaluate the performance of our C-PLES architecture and compare it with the aforementioned state-of-the-art is shown in Table 2. As can be observed, the proposed C-PLES outperforms all the competing models in terms of mIoU and F1 scores for all three different patch sizes. The best mIoU (0.5764) is achieved by our proposed C-PLES method. In addition, the highest F1 score is also achieved by the C-PLES (0.6948). This is a promising indicator that our C-PLES model has the ability to effectively balance precision and recall across all classes, thereby achieving an overall superior landslide segmentation performance, identifying pixels that belong to a particular landslide class while minimizing the presence of misclassification.

Although our proposed C-PLES achieved competitive results in terms of recall among all the models evaluated, it is important to note that the precision metric is slightly lower compared to some state-of-the-art. This indicates a higher rate of false positives in the segmentation, as the model trade-off the performance of recall over precision to reduce missing important features in landslide segmentation. Future work may focus on improving the precision of C-PLES while maintaining a high recall rate, to achieve more accurate and complete segmentation of landslide areas. Figure 6 shows a visual comparison of the segmentation outputs of the proposed C-PLES and competing methods.

Figure 7 provides the segmentation output of C-PLES for the Martian Valles Marineris region. Four random samples are zoomed in and compared to the predicted segmentation with the corresponding ground truth. In the comparison, it is noticeable that the highest IoU (each class IoU) is for the Slump landslide class, while the lowest one is for the Rock Avalanches landslide class. Another visible aspect in Figure 7, is the lack of confidence in the model to identify the boundaries of the landslide, especially when different classes of landslides are very close among them. In some cases, the model tends to generalize the prediction of all the landslides within a patch input towards the class of the biggest landslide. This limitation may be linked to the quantitative results in terms of precision.

6. Ablation study

Contribution of image modality. Several image modalities were merged in the Martian multi-class landslide segmentation dataset. As mentioned before, the dataset contains a total of seven bands: RGB (3 bands), Gray-scale imagery, DEM, THEMIS, and Slope. In order to analyze the contribution of each modality to the performance of the proposed C-PLES, extensive experiments are conducted. The experimental results are detailed in Table 3, which is the result of testing on unseen 256 × 256 patches data samples from the dataset. The training hyperparameters are set as similar to the ones reported in Section 5, however, only mIoU and IoU (IoU_{DF} : Debris Flow, IoU_{S} : Slumps, and IoU_{RA} : Rock Avalanches) are considered as evaluation
Table 3. Importance of different image modalities to the prediction accuracy.

<table>
<thead>
<tr>
<th>Image modality</th>
<th>mIoU</th>
<th>IoU_{DP}</th>
<th>IoU_{DP}</th>
<th>IoU_{RA}</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGB</td>
<td>0.5935</td>
<td>0.0001</td>
<td>0.5779</td>
<td>0.0001</td>
</tr>
<tr>
<td>RGB + DEM</td>
<td>0.6427</td>
<td>0.1427</td>
<td>0.5973</td>
<td>0.1627</td>
</tr>
<tr>
<td>RGB + THEMIS</td>
<td>0.4043</td>
<td>0.0091</td>
<td>0.5206</td>
<td>0.0917</td>
</tr>
<tr>
<td>RGB + Slope</td>
<td>0.4322</td>
<td>0.0099</td>
<td>0.5616</td>
<td>0.1587</td>
</tr>
<tr>
<td>RGB + GS</td>
<td>0.6894</td>
<td>0.0023</td>
<td>0.5619</td>
<td>0.0743</td>
</tr>
<tr>
<td>RGB + DEM + THEMIS</td>
<td>0.5215</td>
<td>0.3454</td>
<td>0.5854</td>
<td>0.1557</td>
</tr>
<tr>
<td>RGB + DEM + Slope</td>
<td>0.5335</td>
<td>0.4419</td>
<td>0.4827</td>
<td>0.2202</td>
</tr>
<tr>
<td>RGB + DEM + GS</td>
<td>0.4233</td>
<td>0.0064</td>
<td>0.6143</td>
<td>0.0730</td>
</tr>
<tr>
<td>RGB + DEM + THEMIS + Slope</td>
<td>0.5698</td>
<td>0.4840</td>
<td>0.5485</td>
<td>0.2485</td>
</tr>
<tr>
<td>RGB + DEM + THEMIS + GS</td>
<td>0.3937</td>
<td>0.0025</td>
<td>0.5773</td>
<td>0.0002</td>
</tr>
<tr>
<td>RGB + DEM + THEMIS + Slope + GS</td>
<td>0.5764</td>
<td>0.4953</td>
<td>0.6153</td>
<td>0.2052</td>
</tr>
</tbody>
</table>

Table 4. Model performance by varying components of the proposed C-PLES architecture. key: [Best, Second Best]

<table>
<thead>
<tr>
<th>Model</th>
<th>mIoU</th>
<th>F1 score</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-PLES w/ dot self-attention</td>
<td>0.5597</td>
<td>0.6894</td>
<td>0.6754</td>
<td>0.7025</td>
</tr>
<tr>
<td>C-PLES w/ MHA</td>
<td>0.5951</td>
<td>0.7159</td>
<td>0.7153</td>
<td>0.7175</td>
</tr>
<tr>
<td>C-PLES w/o Attention</td>
<td>0.5479</td>
<td>0.6670</td>
<td>0.6590</td>
<td>0.6859</td>
</tr>
<tr>
<td>C-PLES w/o PNE</td>
<td>0.5513</td>
<td>0.6703</td>
<td>0.6657</td>
<td>0.6822</td>
</tr>
</tbody>
</table>

Figure 7. Segmentation map of the landslides in the Martian - Valles Marineris region. Four random regions are zoomed in to show in detail a comparison between the ground truth and the predicted segmentation. The IoU scores for each class are shown on the bottom right side of the figure.

metrics. On every round of training, a different set of modalities is used with the RGB modality as a baseline. According to the quantitative results, the best performance is reached by the full set of modalities. However, the best performance to segment the Rock Avalanches (IoU_{RA} of 0.2485) is given under the absence of the gray-scale image (RGB+DEM+THEMIS+Slope). For segmenting Debris Flow landslides, the absence of slopes reduces the ability of the network to discriminate this landslide from the others considerably. It is worth mentioning that, on the other hand, the presence of the gray-scale imagery improves the performance to identify Slumps landslides. Overall, our results suggest that the combination of the full set of modalities is beneficial to achieve better performance in general.

Importance of model components. To understand the inner working of the proposed C-PLES architecture, we evaluate the performance of the architecture with a different set of components that are present or absent. For instance, a major emphasis is on the overall impact of the inclusion or absence of the C-PLES block within the architecture. Table 4 summarizes the results obtained in this ablation study. Due to computational constraints, this experiment is performed with a 64 × 64 patch size dataset, allowing us to efficiently compare the performance of the proposed architecture but varying the attention mechanism. The best performance is reached with C-PLES architecture using the multi-head self-attention (MHA) mechanism. In this approach, the C-PLES block internally takes the output from the PNE layer as the query, key, and value matrices for the multi-head self-attention. The computation of multiple heads allows the model to capture a larger range of trained weights (a set of weights by each head) and thus capture different types of relationships between the different parts of the input sequence in comparison to a regular attention mechanism. In addition, as we hypothesized, the C-PLES architecture using the dot product attention mechanism, which is the version used in the majority of the experiments reported in this paper, provides better performance when it is compared with the version of the C-PLES without attention mechanism and without the PNE. Additionally, the C-PLES block under the absence of the PNE performed slightly better than its counterpart without an attention mechanism. This indicates the importance of the attention mechanism to selectively focus on relevant parts of the image and capture spatial relationships between objects and their context. Finally, the combination of attention with enriched feature representation by PNE provides a harmonically improved performance for Martian landslide segmentation.

7. Conclusions

In this research work, we proposed the C-PLES architecture for Martian multi-class landslide segmentation. The presented research opens the door not only to the analysis of the Martian landslides but also to the same task on Earth. The C-PLES block is introduced as a plug-and-play module that leverages the merits of two main components: the progressive expansion neurons and the attention mechanism. In addition, a Martian multimodal multi-class landslide dataset is introduced and used to evaluate the performance of the proposed method. Our experiments showed that the proposed C-PLES achieved state-of-the-art results for the Martian landslide segmentation task, which proves the importance of the C-PLES block that consists of progressive neuron expansion and attention mechanism to capture pertinent visual features from the different image modalities. The aforementioned dataset will be available to the research community.
References


[25] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei


[48] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun,

