This CVPR 2021 workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MRSCAtt: A Spatio-Channel Attention-Guided Network for Mars Rover Image Classification

Anirudh S Chakravarthy*
BITS Pilani, India
anirudh.s.chakravarthy@gmail.com

Roshan Roy*
BITS Pilani, India
rroshanroy@gmail.com

Praveen Ravirathinam*
BITS Pilani, India
praveen.ravirathinam@gmail.com

Abstract

As the exploration of human beings pushes deeper into the galaxy, the classification of images from space and other planets is becoming an increasingly critical task. Image classification on these planetary images can be very challenging due to differences in hue, quality, illumination, and clarity when compared to images captured on Earth. In this work, we try to bridge this gap by developing a deep learning network, MRSCAtt (Mars Rover Spatial and Channel Attention), which jointly uses spatial and channel attention to accurately classify images. We use images taken by NASA's Curiosity rover on Mars as a dataset to show the superiority of our approach by achieving state-of-the-art results with 81.53% test set accuracy on the MSL Surface Dataset, outperforming other methods. To necessitate the use of spatial and channel attention, we perform an ablation study to show the effectiveness of each of the components. We further show robustness of our approach by validating with images taken aboard NASA's recently-landed Perseverance rover.

1. Introduction

Mars rover image classification aims to identify important objects in satellite imagery from Mars exploration projects – such as instruments, objects, and surroundings. As ongoing Mars exploration leads to the generation of high volumes of data, it is necessary to categorize important objects from their images across various missions. This would enable further scientific investigation and exploratory missions.

Image classification is a widely-studied domain of high-level Computer Vision, and strong results have been achieved on a variety of tasks. The majority of these studies have been carried out using Earthly images; limited focus has been given to other planetary environments. To advance scientific research on Mars images, Wagstaff *et al*. [25] introduced the MSL Surface dataset. The dataset consists of images captured during the NASA Curiosity mission that contain various classes of objects.

However, the dataset consists of a small number of training images. Usually, for image classification approaches, a large labelled training set is required to train a robust network. To overcome this challenge, we use transfer learning. Similar to other tasks, we begin with ImageNet [5] pre-trained weights, and fine-tune the network to our task. By leveraging the existing representations learned by the base network, we save training computation.

Images in the MSL surface dataset are shot with cameras of different focal lengths, resolutions, and field-of-view. Additionally, the images are captured from different angles and magnifications, having challenging illumination settings. Therefore, it is necessary to obtain a better way to glean semantic information from these challenging image backgrounds.

Wagstaff *et al*. [25] used a pre-trained AlexNet for image classification on the dataset. However, not much additional research has been performed towards perception and recognition on the MSL Surface dataset. In this work, we aim to show the effectiveness of spatial and channel attention in classification on astronomical and other planetary images. The use of spatial attention allows the network to pay attention on important spatial and channel-wise cues that characterize the object categories. Channel attention focuses on the representative features, which enables the learning of semantic context through inter-channel relationships. Using spatial and channel attention enables the network to focus to important information associated with downstream classification and discard the noise associated with the image

*denotes equal contribution.

capturing process.

In summary, we have two main contributions. First, we construct a two-stage deep learning network using spatial and channel attention to leverage spatial and semantic information for mars rover image classification. Second, we experimentally show that our approach achieves state-of-the-art results on the MSL Surface dataset, while using significantly lower parameters. We justify the joint optimization of classification accuracy and model memory footprint for the easier integration of our method with future research works. We believe that with increasing interest in space exploration, classification of images in such other planetary contexts will become important and that this work will help investigate common objects across peaceful exploratory missions in the future.

Our code is available at the following repository: https://github.com/anirudh-chakravarthy/MRSCAtt.

2. Related Work

Image classification. Image classification has been a well explored task in computer vision. LeNet [11] set the foundation for image classification by using convolutional neural networks (CNNs). AlexNet [10] reignited research interest in computer vision tasks by demonstrating the superiority of CNNs for image classification, achieving state-of-the-art results on ImageNet. VGGNet [21] demonstrated the effectiveness of 3×3 convolutional filters, which allowed a deeper network with similar memory usage. GoogLeNet [22] introduced the popular Bottleneck layer in the Inception network, while reducing the computation cost. ResNet [6] introduced residual blocks to solve vanishing gradients for easier optimization for deep CNNs. Xception [4] used depth-wise separable convolutions to reduce the number of parameters in a convolution layer. He *et al*. [7] examines the impact of training and optimization procedures for image classification.

Computer vision for Mars. Numerous rovers and satellites being deployed every year, several datasets have been released for computer vision on Mars [25, 20]. Matthies *et al*. [15] provides an outline to the computer vision era on Mars, addressing the challenges, major milestones, and the role of computer vision in Mars exploration. Recently, there has been renewed scientific interest towards vision on Mars. PlaNet [16] used an off-the-shelf RetinaNet [14] object detector for recognizing Transverse Aeolian Ridges (TARs) – the small unusual bedforms on the surface of Mars. SPOC [19] visually identified terrain types and features on Mars using a fully-convolutional neural network, which was successfully deployed on Mars rover missions such as MSL Navcam. Bickel *et al*. [3] used an ensemble of deep learning methods for rockfall distribution and magnitude analysis on the Mars surface. Kerner *et al*. [9] identified surface changes on planetary bodies such as Mars, Earth, and Moon using a convolutional autoencoder and transfer learning with high precision.

Attention. Bahdanau *et al*. [2] introduced the attention mechanism for text summarization. Since Xu *et al*. [27]'s first work on attention in vision, attention modules have demonstrated improved performance in popular image tasks such as classification, detection, and segmentation. Transformer models [23], which have gained prominence lately, also use the attention module. There exist 2 broad categories of visual attention modules – soft and hard attention. Hard attention considers a subset of highly relevant image pixels and thus has significantly lower computation and memory cost. However, it does not have a differentiable cost function. Soft attention considers all input image pixels and has higher resource demand, but can be trained end-to-end.

Soft attention methods are generally used and can be broadly divided into two categories. Channel attention determines the importance of features along the channel dimension. Li *et al*. [13] used global attention on the full image to select specific channel-wise features. [17] fused high-level and low-level features at the channel dimension. Spatial attention extends this idea to enhance feature extraction at the spatial dimension. Jadenberg *et al*. [8] helped understand spatial invariance. The integration of the concepts used in these works helps extract features along both channel and spatial dimensions.

We incorporate the soft attention mechanism in our network– using spatial and channel attention.

3. Proposed Method

The input to our network, MRSCAtt, is a batch of RGB or grayscale images from the MSL Surface Dataset. For each of these images, our method aims to predict a class category for the salient object in the image. In this paper, we propose MRSCAtt (Mars Rover Spatial and Channel Attention), a two-stage network which also jointly uses spatial and channel attention to accurately classify images. Fig 1 shows an illustrative approach to our proposed method.

Attention mechanism has received wide-spread research focus. Recently, Vaswani *et al*. [23] demonstrated the effectiveness of the attention mechanism for computer vision tasks. Attention is known to improve the descriptive ability of features by focusing on important features and ignoring the rest. For the challenging task of image classification on Mars rover images, it is necessary to learn robust representative features. Therefore, we use attention mechanism as a building block for our method.

Given an input image captured from the Mars rover, we extract the image features using the backbone network. This

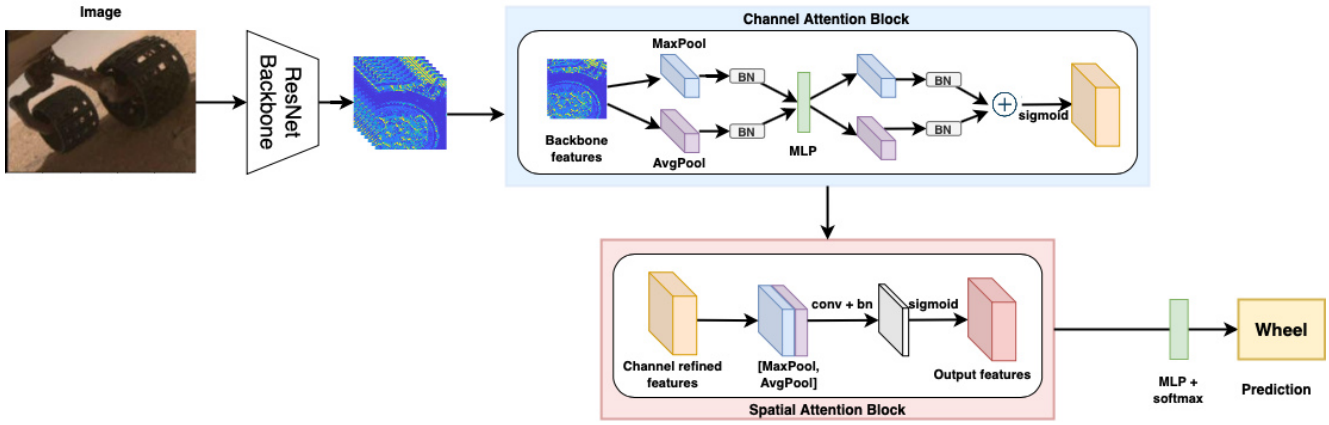


Figure 1: The proposed MRSCAtt architecture.

serves as the first stage of our network, In order to account for challenging variations during image capture (e.g: illumination, focus, angle of capture), we attempt to enrich feature representations using the attention mechanism. To this end, we follow the modules used in [26, 12]. Concretely, the backbone features are first refined using the channel attention block. The channel attention block considers each channel as a feature detector [28], which allows learning the important contextual characteristics from the backbone features. Next, we also use the spatial attention block, which allows the network to focus on the important regions within the backbone features.

In the channel attention block, in addition to max-pooling and average pooling layers, we also introduce batch normalization to reduce the covariance shift. We also include random dropout to prevent overfitting. Finally, the category prediction is obtained using a linear layer from the obtained output features.

3.1. Channel Attention

The channel attention block outputs a refined feature map given the backbone features. The channel attention block aims to learn inter-channel dependencies and dynamically decide which channels to give importance to for downstream classification. Concretely, channel attention guides the network about 'what' to look for in the backbone features. Given input feature $F \in R^{C\times H\times W}$, (where C denotes the number of channels, and H, W denote the height and width of the feature maps), the channel attention block learns a filter $W_C \in R^C$.

The filter W_C learns to prioritize important semantic information across the channel dimension. This filter W_C is then applied to the input feature F, to output channel-refined features R,

$$R = W_C(F) \circ F \tag{1}$$

where $\circ$ denotes the element-wise product. The output features R, therefore, contain regions corresponding to the important semantic context within the image, with the other regions suppressed.

To learn the filter W_C (Eq. 2), the features F are aggregated along spatial dimensions. Global average pooling and global max pooling are used to generate feature descriptors. The multi-layer perceptron (MLP) is then used to learn relationships within the descriptors. The resultant feature vectors from the MLP are then combined using the element-wise addition operation.

$$W_C(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{2}$$

Here, σ denotes the sigmoid activation function. To reduce the co-variate shift in the feature descriptors, we use batch normalization after each layer in the MLP. This normalizes the activations for each layer, thereby stabilizing training and improving generalization performance.

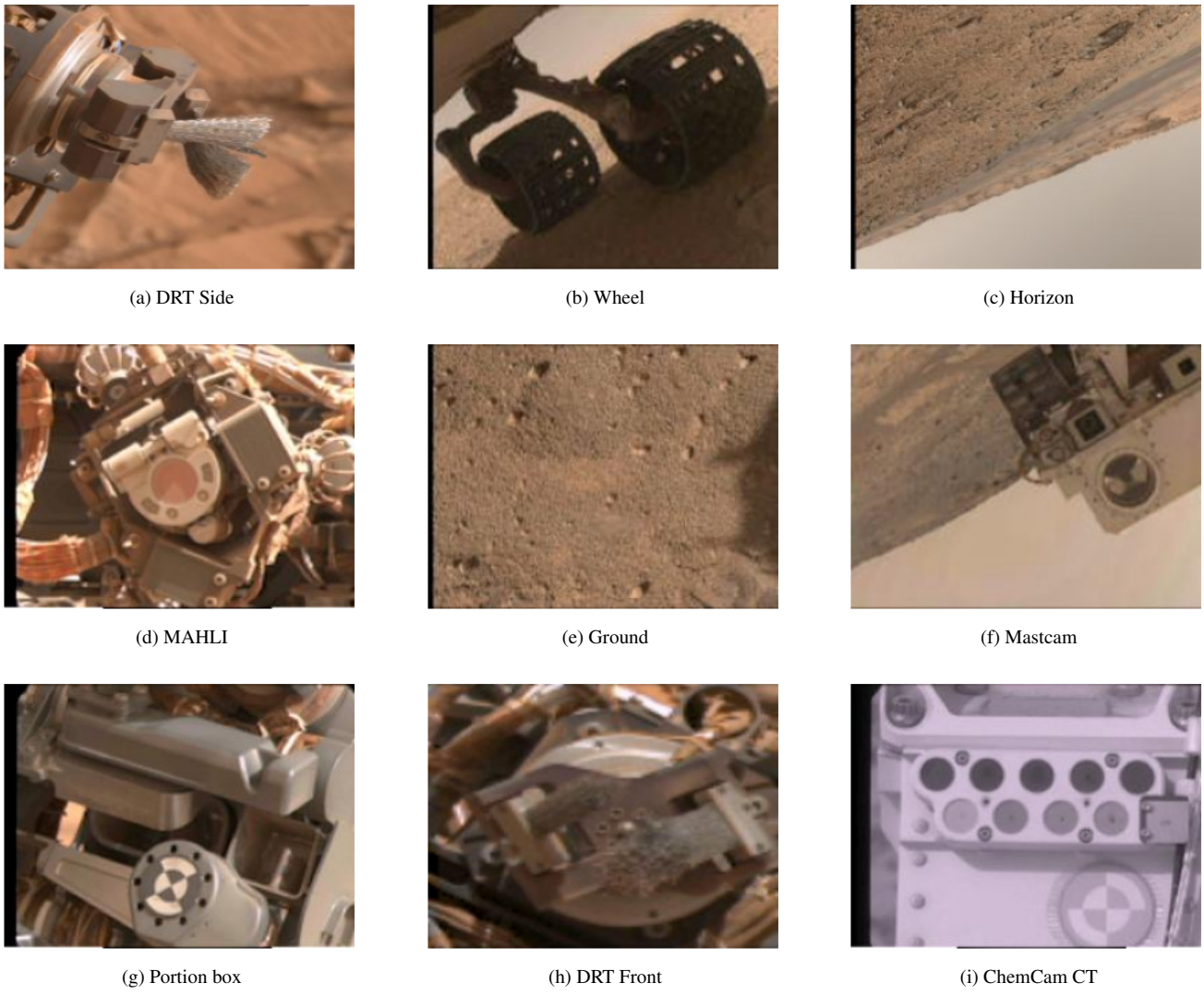
3.2. Spatial Attention

The Mars surface images have significant spatial variations in terms of background and angle of capture. Therefore, it is essential to capture robust spatial information to achieve better context understanding. To this end, we utilize the spatial attention module. For the channel-refined features $R \in R^{C\times H\times W}$, the spatial attention block learns a spatial filter $W_S \in R^{H\times W}$, which is then applied to the R,

$$O = W_S(R) \otimes R \tag{3}$$

where $\otimes$ denotes tensor multiplication. The filter W_S allows the network to learn 'where' to look for important information.

The filter W_S (Eq. 4) is learnt from two feature descriptors generated for each spatial dimension using global aver-



(a) DRT Side (b) Wheel (c) Horizon

(d) MAHLI (e) Ground (f) Mastcam

(g) Portion box (h) DRT Front (i) ChemCam CT

Figure 2: Some images from the MSL Surface Dataset with corresponding category labels

age pooling and global max pooling. These feature descriptors are concatenated and fed to a convolution layer with a 7×7 filter and sigmoid activation (denoted by σ). We also perform batch normalization post the convolution layer.

$$W_S(R) = \sigma(conv^{7\times7}([AvgPool(R), MaxPool(R)])) \tag{4}$$

4. Experiments

In this section, we evaluate the performance of our network, MRSCAtt, on the MSL Surface Dataset.

We approach this task as a multi-class classification challenge. [18] demonstrated that using extracted features from models pre-trained on object classification datasets such as ImageNet are beneficial for aerial and remote sensing images. Thus, we incorporate transfer learning.

Mars surface exploration is a relatively unexplored domain. With growing research, we anticipate the introduction of multiple data sets with addition of more classes in future years. Redeploying models for more classes requires parameter fine-tuning with newer data. In contrast to previous work, we propose an architecture whose training requires minimal memory footprint — to minimize the cost of future fine-tuning. Thus, apart from classification accuracy, we also prioritize the need for fewer trainable parameters.

4.1. Dataset

The MSL Surface Dataset [25] consists of 6691 images of the Mars surface environment that were collected

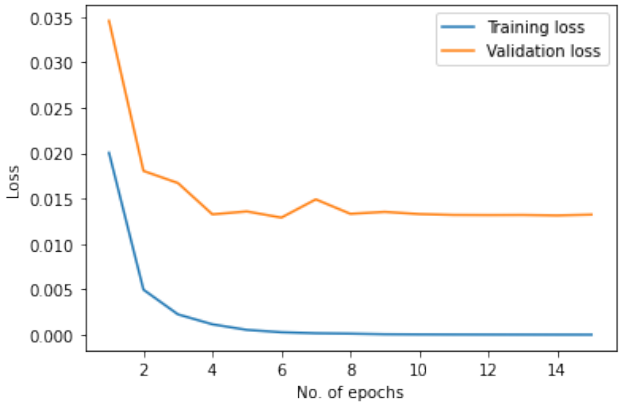


Figure 3: Loss curves for MRSCAtt on the MSL Surface Dataset.

by three instruments on the MSL (Curiosity) rover: Mastcam Right eye, Mastcam Left eye, and MAHLI (Mars Hand Lens Imager). The average size of images is 193×256, with slight variations. The dataset consists of 24 classes as identified by a Mars mission scientist, and contains RGB images, grayscale images, and images with some other instrument filter-induced colours. Images are captured across various settings of illumination, angles, and magnifications, making classification challenging. Some images with their corresponding category labels are illustrated in Fig 2.

To evaluate our network, we report the overall accuracy (OA) and class-wise accuracy on the classification task.

4.2. Implementation Details

Following the dataset split proposed in [25], The training set consists of 3746 images, the validation set consists of 1640 images and the test set consists of 1305 images. The training, validation, and tests sets have been pre-split based on the day of image acquisition. Since the size of the training set is small, we use varying data augmentation techniques to boost the size of the training set. We perform augmentation with random rotations (clockwise, anticlockwise, 72°) and random flips (horizontal, 180°). Apart from the regularizing effect of augmentation, we also use dropout with probability 0.2 to prevent overfitting.

We implement MRSCAtt using the PyTorch framework. We use the ResNet-50 backbone [6] pre-trained on ImageNet for feature extraction. We replace the final fully connected layer to map to the 24 classes in the MSL Surface dataset. To maintain our goal of using fewer trainable parameters for training, we freeze the first four residual blocks and keep the last block as trainable. We train our network for 15 epochs, with a learning rate of 0.0001 and a batch size of 64. We use Adam as the optimizer and the cross-entropy loss function to train our network. The training and validation loss curves are plotted in Fig 3.

Method	Train	Valid	Test
Random	4.2%	4.2%	4.2%
Most common	62.5%	5.3%	19.5%
AlexNet [25]	98.7%	72.8%	66.7%
ResNet-50 [6]	**100%**	80.37%	76.16%
MRSCAtt (ours)	99.60%	**82.74%**	**81.53%**

Table 1: Classification accuracy on the MSL Surface dataset.

Class	R-50	MRSCAtt
APXS	100%	100%
APXS CT	100%	100%
ChemCam CT	**95.23%**	90.48%
Chemin inlet open	55.95%	**95.24%**
Drill	55.0%	55.0%
DRT front	**3.33%**	0%
DRT side	34.66%	**64.67%**
Ground	93.7%	**96.85%**
Horizon	**87.5%**	73.61%
Inlet	0%	**12.5%**
MAHLI	58.33%	**91.67%**
MAHLI CT	70.17%	**85.96%**
Mastcam	71.87%	**93.75%**
Mastcam CT	83.33%	83.33%
Observation tray	**100%**	91.67%
Portion box	89.58%	**91.67%**
Portion tube	77.77%	77.77%
Portion tube opening	100%	100%
REMS UV sensor	**69.44%**	63.89%
Rover rear deck	92.85%	**100%**
Scoop	**100%**	80%
Wheel	–	–

Table 2: Class-wise performance on the test set of the MSL Surface Dataset.

4.3. Results

We compare the performance of MRSCAtt with other networks in Table 1. For reference, we also added the performance of a random classifier and a classifier which always votes for the most frequent object category. [25] proposed a transfer learning method with AlexNet [10] backbone with complete pre-training (no frozen layers) that achieved a test set accuracy of 66.7%. We trained a ResNet-50 network following the same protocol for MRSCAtt training, and achieved an accuracy of 76.16% on the test set. However, the introduction of spatial and channel attention blocks significantly improve the performance, demonstrating performance an improvement of 5.37% on the test set. Using spatial and channel attention, MRSCAtt achieves state-of-the-art results on the MSL Surface dataset. We also

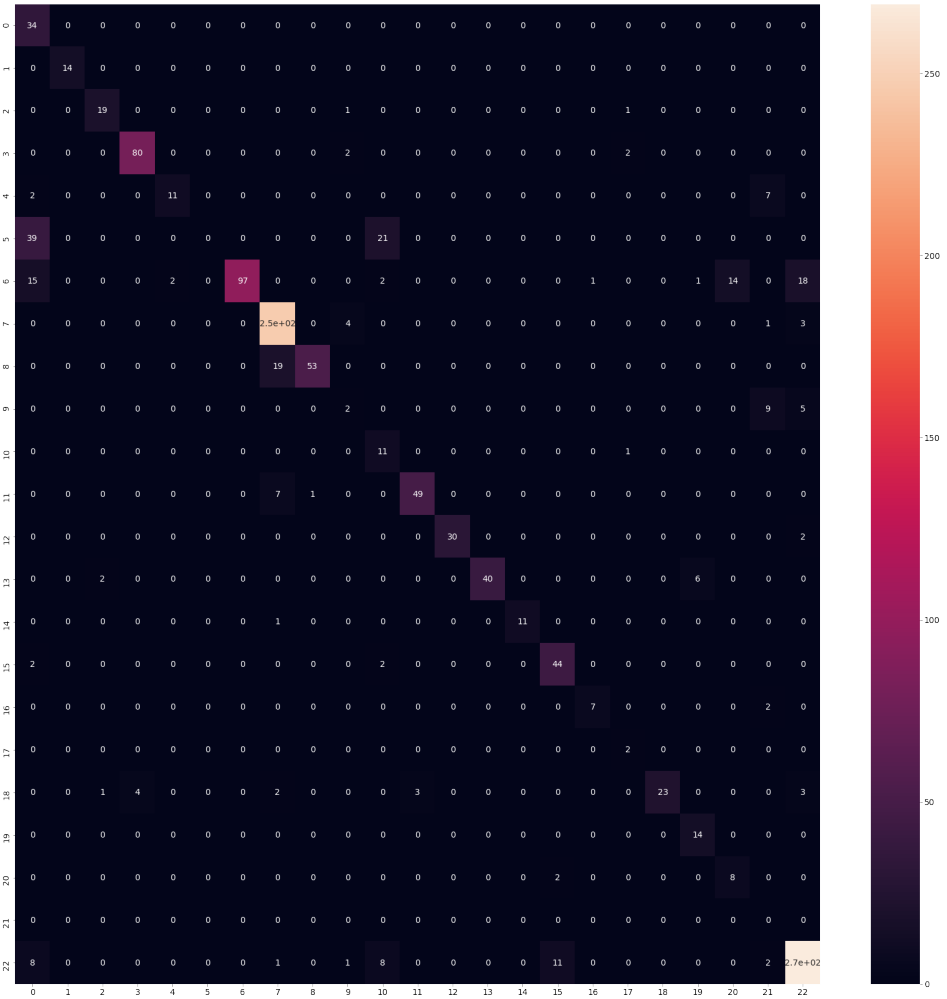


Figure 4: The confusion matrix for MRSCAtt on the MSL Surface test set.

note that in [25], the entire network is retrained on the MSL Surface dataset. Since we only unfreeze the final residual block, we use much fewer parameters. Therefore, not only do we achieve state-of-the-art performance, we also utilize significantly fewer parameters.

We also examine the performance of our network using a confusion matrix in Fig 4. We note that the test set labels miss 2 categories, explaining the size of the confusion matrix. Despite large class imbalance in the test set, MRSCAtt is able to reliably classify most of the categories accurately. However, on closer inspection of the confusion matrix, we observe a failure case of our method. Our network never predicts the *DRT Front* class and often misclassifies it as *APXS* or *MAHLI CT*.

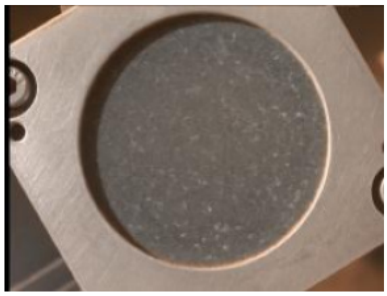
To analyse this in greater detail, we also compare class-wise performances in Table 2. As demonstrated in Table 1, the baseline model is prone to overfitting. Although classes such as *Scoop* have 100% baseline accuracy, they have extremely low precision. In these classes, our MRSCAtt model has better precision and thus generalizes better to provide drastically improved accuracy for classes such as *Chemin inlet open*, *DRT side* and *MAHLI*. In classes such as *APXS*, *APXS CT*, and *Portion Tube opening*, MRSCAtt retains 100% classification performance because the precision for these classes is sufficiently high. However, as we found previously, *DRT Front* and *Inlet* are frequently misclassified, perhaps due to dataset imbalance. We also observe that these classes often appear visually very similar, which may explain the model's confusion for these classes.

We also visualize the performance of MRSCAtt on the test set of the MSL Surface Dataset. We illustrate the correctly classified images in Fig 5.

4.4. Ablation Study

To understand where the performance improvement stems from, we perform an ablation study on our compo-

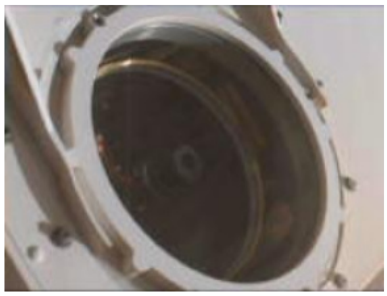
(a) MAHLI CT



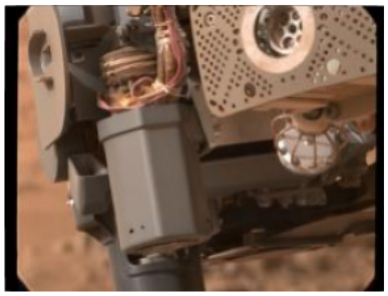
(b) APXS CT

(c) Ground

(d) Horizon



(e) Chemin inlet open



(f) APXS

(g) MAHLI

(h) Portion Box

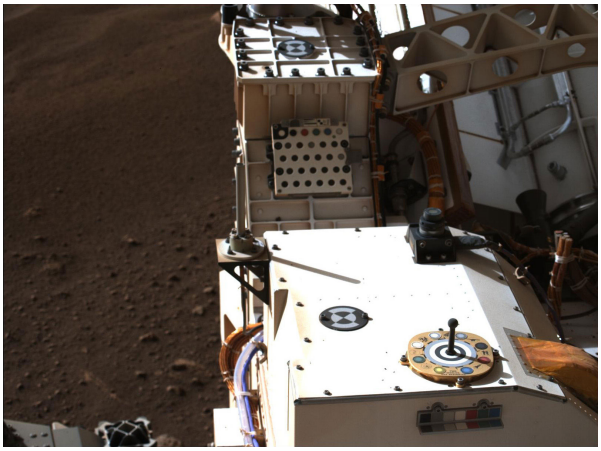
(i) ChemCam CT

Figure 5: Some predictions of MRSCAtt on the MSL Surface Dataset.

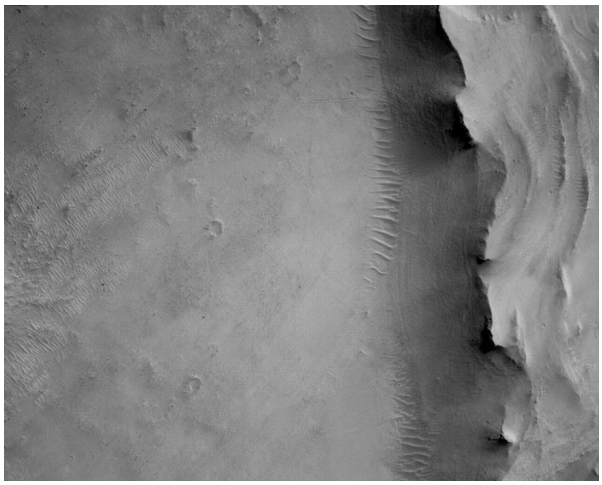
Backbone	Spatial	Channel	Test accuracy	No. of Parameters
ResNet-50 (res-5 frozen)	×	×	68.89%	51.2K
ResNet-50	×	×	76.16%	15.015M
ResNet-50	✓	×	76.86%	15.016M
ResNet-50	×	✓	81.07%	17.126M
ResNet-50 (ours)	✓	✓	81.53%	17.127M

Table 3: Ablation study of the components in MRSCAtt.

nents, as shown in Table 3. The baseline accuracy using ResNet-50 with the final residual block unfrozen during training, is 76.16% on the test set. Using spatial attention, we achieve a performance of 76.86%. This is surprising, because the MSL Surface dataset consists of complex variations and heterogeneity. However, we hypothesize that the ResNet backbone is sufficiently able to capture the spatial variations across the dataset, thereby explaining this small increment. To examine this claim, we train a MRSCAtt with the entire backbone network frozen. In this setting, we observe a test set accuracy of 68.89%, which is significantly lower and seems to support our understanding.



(a) Observation Tray



(b) Ground

Figure 6: Category predictions of our network on images captured by the Perseverance Rover

Channel attention contributes the majority of the performance improvement, despite requiring a few thousand additional trainable parameters. This follows from [26], which suggests that learning inter-channel dependencies enables better understanding since each channel is a feature detector. Finally, the combination of both spatial and channel attention in MRSCAtt enables the learning of both spatial and inter-channel dependencies and achieves 81.53% test set accuracy, while adding only 2 million additional trainable parameters.

4.5. Perseverance Rover

NASA's Mars 2020 mission is a follow-up to the 2016 Curiosity Mission. Recently in 2021, NASA's Perseverance Rover landed on Mars. It is part of the Mars exploration program that aims to gather evidence for signs of habitability on Mars. The cameras aboard the Perseverance rover, like the Curiosity rover, are also designed to capture high-resolution images of the planet's surface and other rover tools.

In order to validate the effectiveness of our method and prove its generalization capacity for further Mars rover missions, we used our network to classify images collected from the Perseverance mission's image gallery[1] in Fig 6. Since there are no ground truths available for these images at this time, we cannot report metrics for this experiment. However, we show empirical results obtained using inference on MSRCAtt. Fig 6a is predicted as an *Observation Tray*, while Fig 6b is predicted as *Ground* category. This empirical verification shows us that MSRCAtt along with the channel and spatial attention modules can also be deployed across several NASA missions for image classification on Mars.

[1] https://mars.nasa.gov/mars2020/multimedia/images/

5. Conclusion

In this work, we present a deep learning method , MRSCAtt, that jointly incorporates channel and spatial attention mechanism to classify images of the MSL Surface Dataset. Not only do we achieve state-of-the-art classification results, we do so with significantly fewer trainable parameters compared to existing work. This allows the model to be relatively inexpensive during training and accurate at deployment. This is significant because of the potential need to fine-tune the model to accommodate the dynamic nature of Mars research – an increasing number of classification categories.

Our ablation studies demonstrate the effectiveness of our approach. By adding only 2 million additional trainable parameters to the backbone network, we significantly boost the classification performance. After validating our model's performance on NASA's Perseverance rover images, we believe that this is a feasible step towards the Mars rover image classification challenge. With new advances with rovers and satellites and thus new datasets in the upcoming years [24, 1], scientific investigation will continue. Therefore, the generalization of MRSCAtt is an extremely important property for the upcoming decades. In the future, we aim to evaluate our method for classification on other planetary images. We also plan to use transformer networks for learning attention mechanisms.

References

[1] Abigail R. Azari, John B. Biersteker, Ryan M. Dewey, Gary Doran, Emily J. Forsberg, Camilla D. K. Harris, Hannah R. Kerner, Katherine A. Skinner, Andy W. Smith, Rashied Amini, Saverio Cambioni, Victoria Da Poian, Tadhg M. Garton, Michael D. Himes, Sarah Millholland, and Suranga Ruhunusiri. Integrating machine learning for planetary science: Perspectives for the next decade, 2020. 8

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2

[3] V. T. Bickel, S. J. Conway, P. A. Tesson, A. Manconi, S. Loew, and U. Mall. Deep learning-driven detection and mapping of rockfalls on mars. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:2831–2841, 2020. 2

[4] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017. 2

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2009. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 5

[7] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li. Bag of tricks for image classification with convolutional neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019. 2

[8] Max Jaderberg, K. Simonyan, Andrew Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *NIPS*, 2015. 2

[9] H. R. Kerner, K. L. Wagstaff, B. D. Bue, P. C. Gray, J. F. Bell, and H. Ben Amor. Toward generalized change detection on planetary surfaces with convolutional autoencoders and transfer learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(10):3900–3918, 2019. 2

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, page 1097–1105, 2012. 2, 5

[11] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[12] Haifeng Li, Kaijian Qiu, Li Chen, Xiaoming Mei, Liang Hong, and Chao Tao. Scattnet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 2020. 3

[13] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *CoRR*, abs/1805.10180, 2018. 2

[14] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 2

[15] Larry Matthies, Mark Maimone, Andrew Johnson, Yang Cheng, Reg Willson, Carlos Villalpando, Steve Goldberg, Andres Huertas, Andrew Stein, and Anelia Angelova. Computer vision on mars. *International Journal of Computer Vision*, 75(1):67 – 92, October 2007. 2

[16] Timothy Nagle-McNaughton, Timothy McClanahan, and Louis Scuderi. Planet: A neural network for detecting transverse aeolian ridges on mars. *Remote Sensing*, 12:21–25, 2020. 2

[17] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *CoRR*, abs/1804.03999, 2018. 2

[18] Otavio Penatti, Keiller Nogueira, and Jefersson dos Santos. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains. In *Proceedings of the IEEE conference on computer vision and pattern recognition Workshops*, pages 44–51, 06 2015. 4

[19] Brandon Rothrock, Ryan Kennedy, Chris Cunningham, Jeremie Papon, Matthew Heverly, and Masahiro Ono. Spoc: Deep learning-based terrain classification for mars rover missions. *AIAA SPACE 2016*, page 5539, 2016. 2

[20] S. P. Schwenzer, M. Woods, S. Karachalios, N. Phan, and L. Joudrier. Labelmars: Creating an extremely large martian image dataset through machine learning. In *50th Lunar and Planetary Science Conference*, 2019. 2

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR*, 2015. 2

[22] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 2

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NIPS*, 2017. 2

[24] Kiri Wagstaff, Steven Lu, Emily Dunkel, Kevin Grimes, Brandon Zhao, Jesse Cai, Shoshanna B Cole, Gary Doran, Raymond Francis, Jake Lee, and Lukas Mandrake. Mars image content classification: Three years of nasa deployment and recent advances. *arXiv preprint arXiv:2102.05011*, 2021. 8

[25] Kiri Wagstaff, You Lu, Alice Stanboli, Kevin Grimes, Thamme Gowda, and Jordan Padams. Deep mars: Cnn classification of mars imagery for the pds imaging atlas. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1, 2, 4, 5, 6

[26] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3, 8

[27] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 2048–2057. JMLR.org, 2015. 2

[28] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 3