Supporting Information for "Automatic Encoding of Unlabeled Two Dimensional Data Enabling Similarity Searches: Electron Diffusion Regions and Auroral Arcs"

A. W. Smith [1], I. J. Rae [1], J. E. Stawarz [1], W. J. Sun [2], S. Bentley [1], A. Koul [3]

[1]Department of Mathematics, Physics and Electrical Engineering, Northumbria University, Newcastle upon Tyne, UK

[2]Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI, USA

[3]Pinterest Inc, San Francisco, California, USA

Contents of this file

1. Figures S1, S2, S3, S4
2. Tables S1

Introduction

In this supporting information we provide additional details about the data processing and model training.

S1: MMS Data Preparation

There are several steps taken to prepare the MMS velocity electron distributions before they are used to train the model. Figure S1 outlines the process, moving from left to right for two examples (top and bottom). The original electron velocity distributions, within 20° of the "perpendicular" plane, are on the left (Figure S1a, d). We then limit the energy

range to between 2 and 30 keV (Figure S1b, e). Finally, the colour scale is adjusted and the central 100 pixels selected (Figure S1c, f).

S2: Image Augmentations

Here we show a table detailing the random image augmentations selected for the MMS electron distributions (Table S2 column A) and All Sky Images (Table S2 column B), along with the settings of each random augment.

S3: All Sky Image Data and Processing

We use a database of all sky images obtained by the THEMIS cameras in North America (Donovan et al., 2006), the data was then processed by Clausen and Nickisch (2018) to form the OATH dataset. In particular, the all sky images were cropped to remove low elevation angles, and the brightness scaled to between 0 and 1. The OATH dataset represents 126 hours of data, comprising 84 separate intervals and 5824 images. The augmentations applied to the all sky images are detailed in the table in S2, column B. Examples of the augmentations are shown in Figure S2. The format of Figure S2 is similar to that shown in Figure 2 within the main article. Here, in contrast to Figure 2 we see that the images are cropped, to allow the model to focus on smaller structure within the image.

We note here that Clausen and Nickisch (2018) generated a 1001 dimensional embedding of the auroral images prior to classification. In that work the embedding was generated by a pre-trained Inception-v4 network (Szegedy et al., 2016). While similar in some ways to our method we note that the Inception-v4 network was trained with 1.2 million labeled images from 1000 different categories, while our embeddings are generated through bespoke encoders on our specific, but unlabeled datasets.

S4: Model Training

We perform an independent grid search procedure for each of the example data in order to identify the optimum training parameters for the models. The batch size, training epochs and momentum of the optimizer (stochastic gradient descent) are the parameters over which the search is performed. The learning rate (R) was chosen to scale with the batch size, and was defined as:

$$R = 0.05 * \frac{n_{\text{batch}}}{256} \tag{1}$$

where n_{batch} is the batch size, as suggested in Chen and He (2021). For the MMS electron distributions the batch size, training epochs and momentum were found to be 64, 15 and 0.91 respectively. Meanwhile for the auroral images the same parameters were found to be 64, 10 and 0.58. We show the model training curves for the MMS electron distribution data in Figure S3. We can see that the model trains well with the loss plateauing towards the end of the training process. Further, there are no apparent signs of overfitting, i.e. divergence between the validation and training loss.

We also monitor the model to ensure that the output vectors have not collapsed to a constant vector. We define the "collapse" of the model as the average standard deviation of the L2-normalized output vectors, as in Chen and He (2021). In Figure S3 we see that the plateau in the measured collapse occurs at approximately $1/\sqrt{d}$, where d is the dimensionality of the output embedding (512 in this case), indicating that the model has produced output embeddings well distributed on the unit hyper-sphere. A collapse of the output vector to a constant value would manifest as a collapse value of zero (Chen & He, 2021).

S5: All Sky Image Embedding Visualization

Figure S4 shows the distribution of embeddings derived for the auroral all sky images. The visualization shows that similar images are grouped together: for example the moon is present in the images at the top (as well as in the left and lower right), while the central images show auroral arcs near zenith, with broader and even double arcs towards the bottom right of this grouping. As in the examples in the main text, we will not comment quantitatively on these distributions due to the non-convex loss function used within the t-SNE dimension reduction process.

References

Chen, X., & He, K. (2021). Exploring simple Siamese representation learning. In *Proceedings of the ieee computer society conference on computer vision and pattern recognition* (pp. 15745–15753). doi: 10.1109/CVPR46437.2021.01549

Clausen, L. B. N., & Nickisch, H. (2018, jun). Automatic classification of auroral images from the Oslo Auroral THEMIS (OATH) dataset using machine learning. *Journal of Geophysical Research: Space Physics*. Retrieved from `http://doi.wiley.com/10.1029/2018JA025274` doi: 10.1029/2018JA025274

Donovan, E., Mende, S., Jackel, B., Frey, H., Syrjäsuo, M., Voronkov, I., ... Connors, M. (2006, sep). The THEMIS all-sky imaging array—system design and initial results from the prototype imager. *Journal of Atmospheric and Solar-Terrestrial Physics*, *68*(13), 1472–1487. doi: 10.1016/J.JASTP.2005.03.027

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2016, feb). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 4278–4284. Retrieved from `https://arxiv`

.org/abs/1602.07261v2 doi: 10.1609/aaai.v31i1.11231

Torbert, R. B., Burch, J. L., Argall, M. R., Alm, L., Farrugia, C. J., Forbes, T. G., ... Khotyaintsev, Y. (2017, dec). Structure and Dissipation Characteristics of an Electron Diffusion Region Observed by MMS During a Rapid, Normal-Incidence Magnetopause Crossing. *Journal of Geophysical Research: Space Physics*, *122*(12), 11901–11916. Retrieved from http://doi.wiley.com/10.1002/2017JA024579 doi: 10.1002/2017JA024579

Zhou, M., Deng, X. H., Zhong, Z. H., Pang, Y., Tang, R. X., El-Alaoui, M., ... Lindqvist, P.-A. (2019, jan). Observations of an Electron Diffusion Region in Symmetric Reconnection with Weak Guide Field. *The Astrophysical Journal*, *870*(1), 34. Retrieved from https://iopscience.iop.org/article/10.3847/1538-4357/aaf16fhttps://iopscience.iop.org/article/10.3847/1538-4357/aaf16f/meta doi: 10.3847/1538-4357/AAF16F

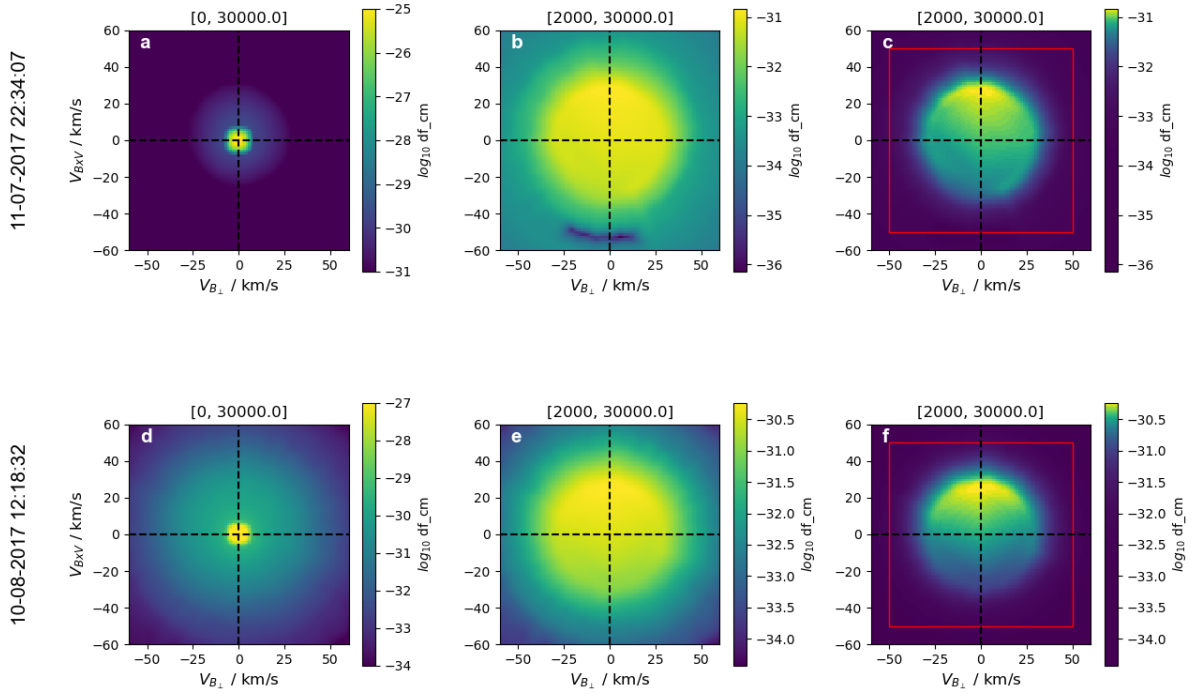


Figure S1. Data preparation for the MMS electron velocity distributions for two example intervals (top: Torbert et al. (2017), bottom: Zhou et al. (2019)). Left (a, d): the original electron velocity distributions, with the x axis defined by the V_{bulk} projected onto the plane normal the magnetic field, and the y axis defined by $B \times V_{bulk}$. Center (b, e): the electron velocity distributions with the energy range limited to between 2 and 30 keV. Right (c, f): the same as in the centre but with an adjusted color scale and the central 100 pixels highlighted with a red square.

Augmentations	**A: Electron Distributions**	**B: All Sky Images**
Mirror	Horizontal: 0.5	Vertical: 0.7; Horizontal: 0.7
Blur [Kernel Size]	0.4 [0.025]	0.7 [0.025]
Crop	n/a	0.3 - 1.
Hue	n/a	n/a
Saturation [Degree]	0.4 [0.3]	n/a
Contrast [Degree]	0.4 [0.6]	n/a
Brightness [Degree]	0.4 [0.3]	n/a

Table S2. Table showing possible augmentations, their probability and properties (in square brackets).

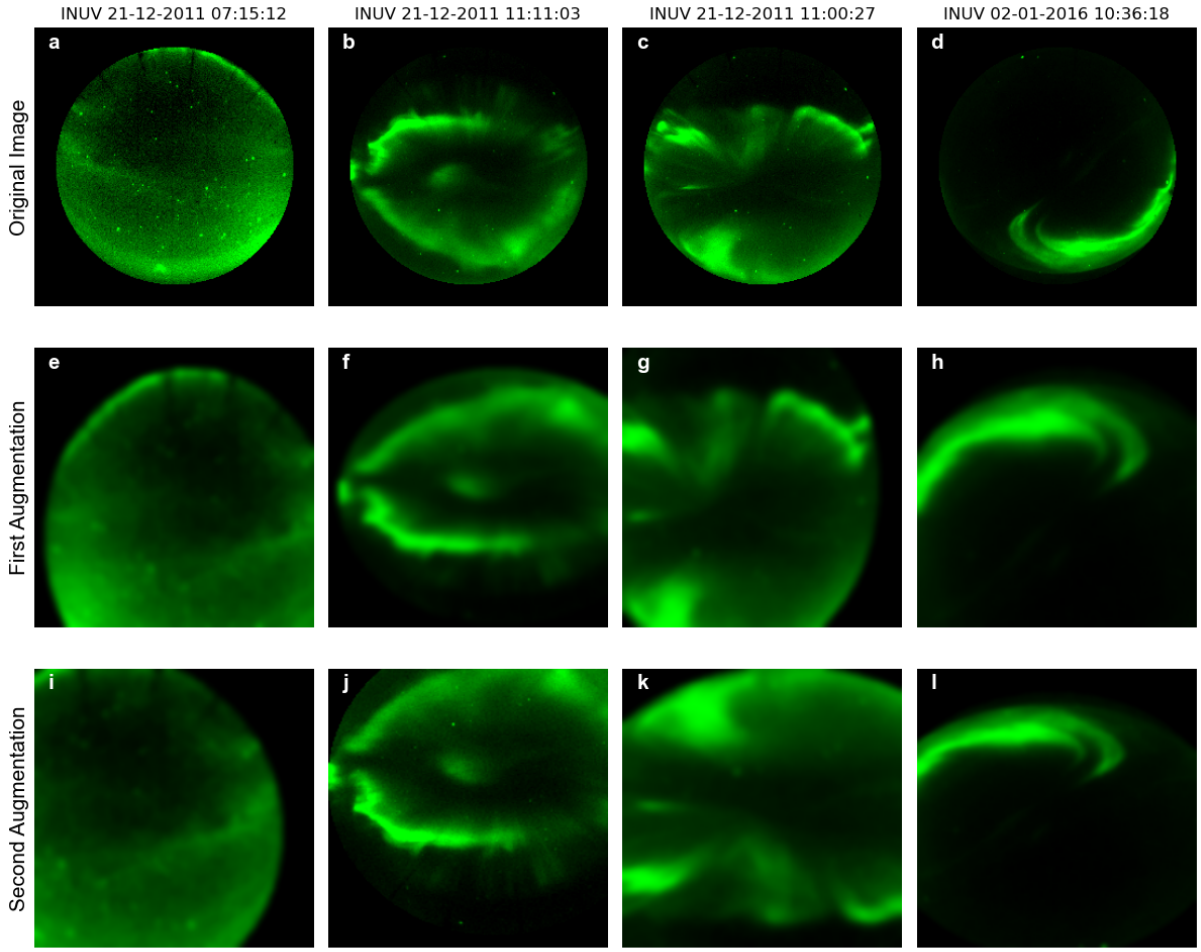


Figure S2. Four example images with augmented versions. Top (a - d): The original all sky images (Clausen & Nickisch, 2018). Middle (e - h): the first augmentation with a random assignment from Table S2, column B. Bottom (i - l): the second random augmentation, again with a random assignment augmentations.

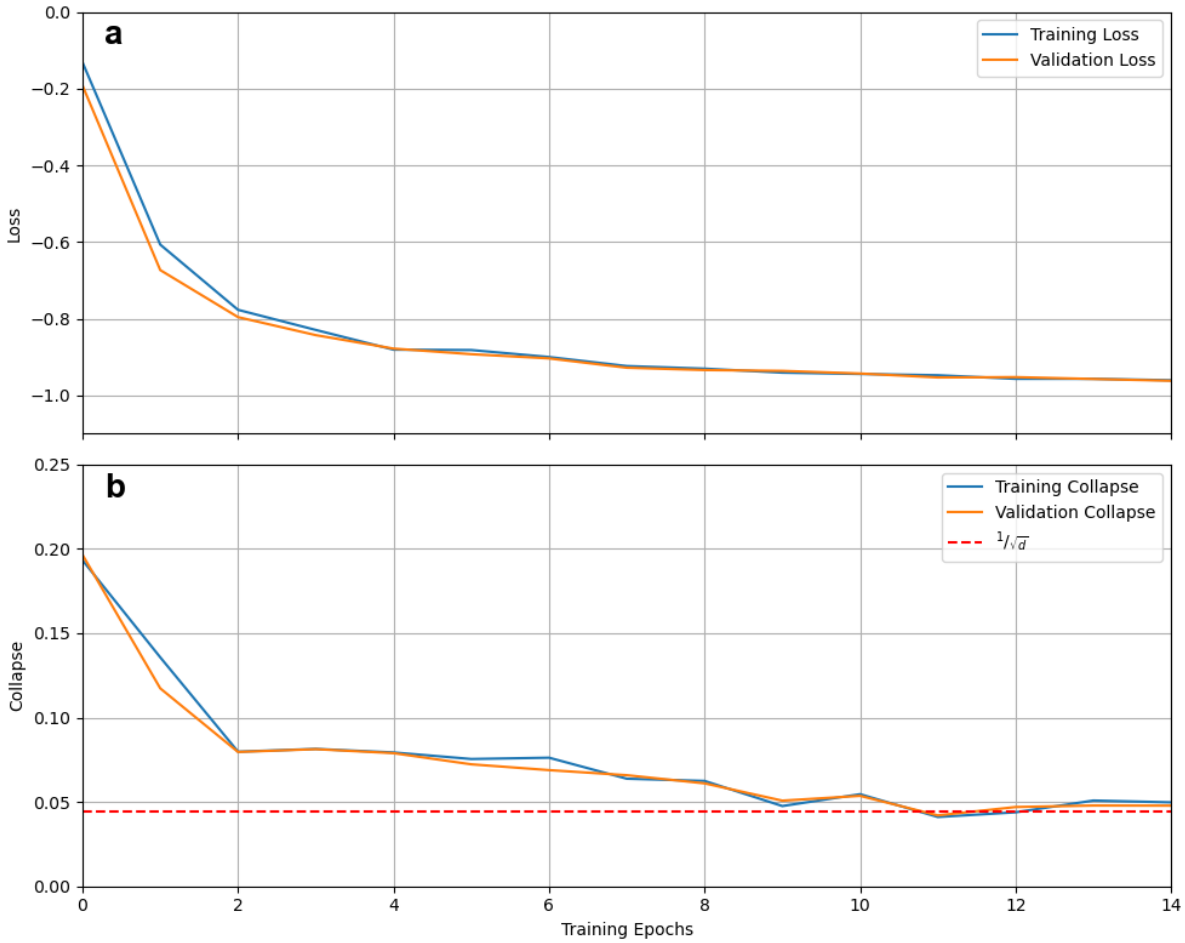


Figure S3. Training curves for the MMS electron distribution training set. Top (a): the negative cosine similarity loss as a function of training epoch. Bottom (b): the model collapse, as defined by the (average) per-channel standard deviation of the L2-normalised output (Chen & He, 2021). We mark $^1/_{\sqrt{d}}$ here, where $d = 512$ (the dimensions of the output embedding).

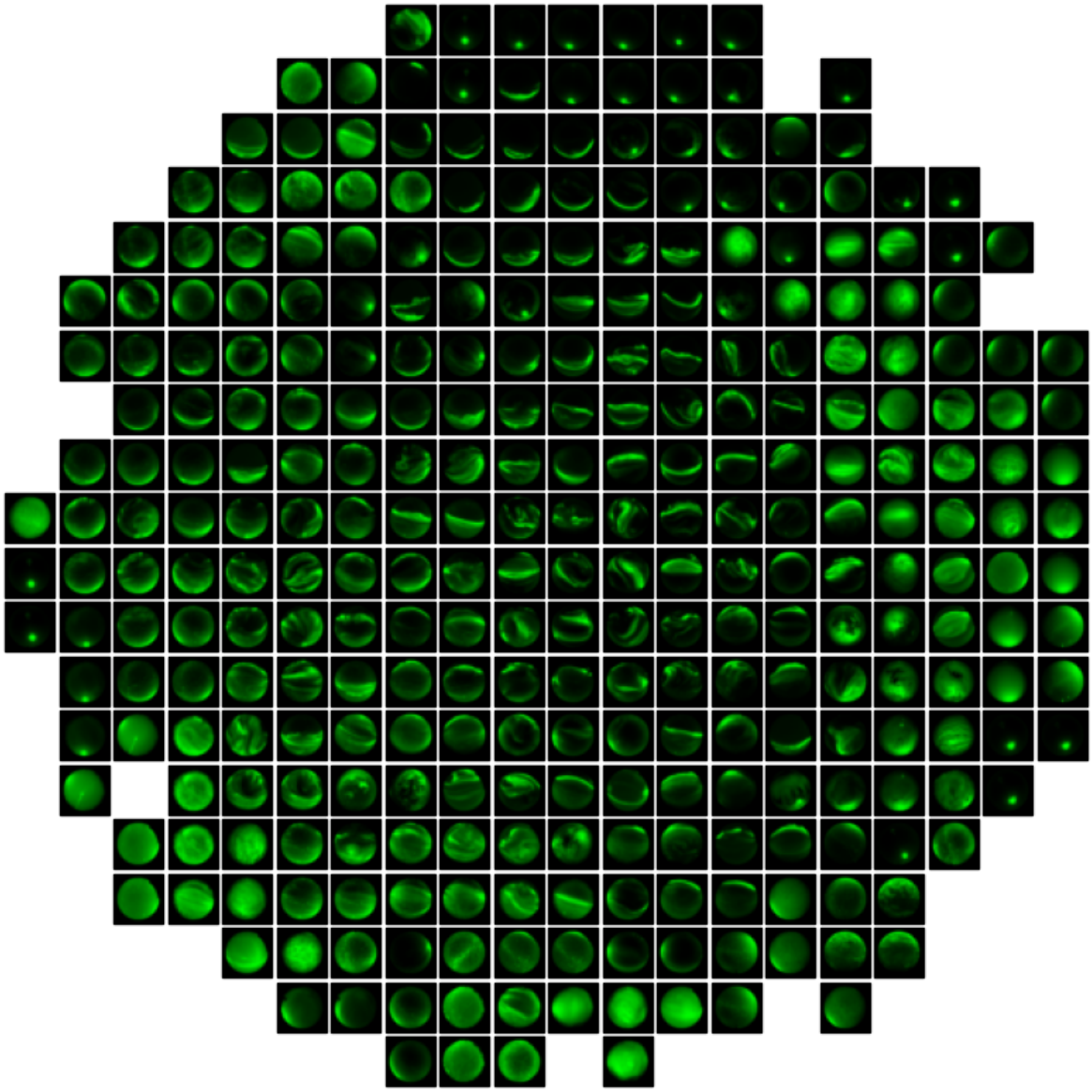


Figure S4. Two dimensional t-SNE derived distribution of image embeddings for auroral all-sky images.