Heliyon 9 (2023) e20609

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Aurora retrieval in all-sky images based on hash vision transformer

Hengyue Zhang [a], Hailiang Tang [a,*], Wenxiao Zhang [b]

[a] School of Information Science and Engineering, Qilu Normal University, Jinan, 250200, China
[b] School of Finance and Economics, Shandong University of Engineering and Vocational Technology, Jinan, China

ARTICLE INFO

Dataset link: https://doi.org/10.1029/2018JA025274

Keywords:
Aurora image retrieval
Vision transformer
Deep learning
All-sky images

ABSTRACT

Auroras are bright occurrences when high-energy particles from the magnetosphere and solar wind enter Earth's atmosphere through the magnetic field and collide with atoms in the upper atmosphere. The morphological and temporal characteristics of auroras are essential for studying large-scale magnetospheric processes. While auroras are visible to the naked eye from the ground, scientists use deep learning algorithms to analyze all-sky images to understand this phenomenon better. However, the current algorithms face challenges due to inefficient utilization of global features and neglect the excellent fusion of local and global feature representations extracted from aurora images. Hence, this paper introduces a Hash-Transformer model based on Vision Transformer for aurora retrieval from all-sky images. Experimental results based on real-world data demonstrate that the proposed method effectively improves aurora image retrieval performance. It provides a new avenue to study aurora phenomena and facilitates the development of related fields.

1. Introduction

Auroras are captivating celestial events that occur in coupling between the solar wind and the magnetosphere. These are caused by the collision between charged particles (i.e., electrons and protons) and neutral constituents of the upper atmosphere along magnetic field lines toward Earth [1]. In other words, auroras represent large amounts of energy dissipated by the magnetosphere in the atmosphere [2]. These can be observed by the naked eye from the ground because the magnetospheric serves as a wide screen that contains the projection of magnetosphere maps along magnetic field lines into the upper atmosphere. Since auroras occurred in near-Earth space, the clear morphology of various auroras obtained from the ground is indispensable for researchers understanding of the process of magnetospheric dynamics. Moreover, allowing physicists to study large-scale magnetospheric processes through computer vision techniques from the ground.

With the fast development of imaging technology and imaging equipment, most phenomena in near-Earth space can be captured by anamorphic lenses. Specifically, circular images such as auroras and cloud images are produced by circular fisheye lenses, which are commonly used for observing astronomical phenomena [3]. The circular images are directly converted into rectangular images

* Corresponding author.
E-mail address: 20170333@qilu.edu.cn (H. Tang).

https://doi.org/10.1016/j.heliyon.2023.e20609
Received 14 June 2023; Received in revised form 2 October 2023; Accepted 2 October 2023

Available online 13 October 2023

through transformation functions to store and retrieve them conveniently. However, manual operations introduce artificial errors and labor noise, affecting image quality and research accuracy. Moreover, conventional approaches rely on the human visual system [4], which increases human consumption and is time-consuming. Therefore, it is necessary to design an automatic retrieval algorithm without any human intervention, which provides convenience to physicists in selecting their interested auroras from all-sky images. Several prior machine learning-based methods focus on extracting auroras' intensity, texture, and histogram features to achieve automatic retrieval without any human intervention. For instance, Syrjasuo et al. [5] calculated statistics features, such as brightness distribution and global texture, to retrieve auroras from extensive all-sky image data. To further explore more spatial texture representation (i.e., intensity, shape, and texture), Wang et al. [6] combined the local binary pattern (LBP) [7] operator with a block partition mechanism in their automatic recognition system. Their proposed feature representation approach explored content-based information in the aurora retrieval and classification process. In addition, the reliability and robustness are enhanced by the extensive use of various features. In content-based image retrieval (CBIR), researchers commonly used the scale-invariant feature transform (SIFT) [8] descriptor to generate visual vocabulary, and its discriminative ability can improve retrieval accuracy. Nevertheless, it was unsuitable for aurora images because local representation and contextual information around auroras were limited. To address this problem, Dense-SIFT [9] combined traditional SIFT with a dense rectangular grid that eliminates specific characteristics of aurora images. Although SIFT-based methods had made acceptable progress, they ignored texture extraction around auroras, leading to information loss. To further refined the Dense-SIFT descriptor, Yang et al. proposed [10] a polar embedding structure by leveraging the bag-of-visual words framework to achieve aurora image retrieval in the large-scale database. Experiment results showed that the proposed polar embedding structure could improve aurora image retrieval performance with acceptable calculation consummation. However, machine learning-based methods depend on handcrafted feature selection, which affects the aurora retrieval accuracy and the discriminability of the model because human-designed features lack semantic information and version perception.

Recently, deep learning-based methods have shown impressive power in Aurora image retrieval. The automatic feature selection strategy (i.e., end-to-end) can avoid human bias and bridge the semantic gap. It can assist physicists in retrieving interested auroras from all-sky images. For instance, Yang et al. [11] proposed a saliency proposal network built on the Mask Region-based Convolutional Neural Network (R-CNN) [12] to reduce the influence of unrelated regions in aurora images. They designed the spherical distortion to replace the rectangular meshing manner, perfectly suitable for aurora images obtained by circular fisheye lenses. Experiment results demonstrated that their method could effectively and accurately retrieve ten categories of auroras from extensive datasets. Another work [13] proposed a hierarchical deep embedding model to enhance the discriminative ability and eliminate mismatches in retrieving aurora images from large-scale data. The well-designed architecture added a polar region pooling layer in the CNN to capture global features, supplementing the local SIFT feature. Local and global features were combined in the hierarchical embedding to improve matching capability. To further suitable circular image retrieval tasks, they introduced a polar meshing scheme, which can reflect the physical information and determine the position of interested regions. CNN-based image retrieval methods gradually replace the human intervention descriptors, which benefit from the pre-training scheme and fine-tuning strategy. To obtain a comprehensive feature representation, hybrid image retrieval methods (i.e., CNN-based methods combined with classic operators) have been proposed and achieved good performance [14]. For instance, Wei et al. [15] proposed a cross-model retrieval architecture that combined off-the-shelf CNN features with handcrafted features of several classic methods, including text, Image-fc6, Image-fc7, Image-BoVW, Image-FT-fc6, and Image-FT-fc7. Specifically, they first pre-trained a traditional CNN on ImageNet, and the fine-tuning phase is performed in the target dataset. And then, a deep semantic matching model was designed to address the cross-modal retrieval. At last, different feature maps of multi-modal data were projected into a semantic space and share the ground truth. Experiments showed that the cross-modal method's performance was superior to pure CNN-based retrieval methods.

Although existing deep learning-based methods have progressed in aurora image retrieval, they focus on utilizing local features but neglect learning contextual information extracted from global features of aurora images. Besides, standard concatenation and pixel-wise addition of feature maps may widen the semantic gap because feature maps from different models or networks always have multiple semantic information. Bearing the above analysis in mind, we propose a Hash-Transformer model by introducing vision transformers, which are a type of deep learning method based on the deep supervised hashing [16] and the Vision Transformer (ViT) [17] for aurora image retrieval. Specifically, both local features and global features of interested regions can be employed in our model. Benefiting from the self-attention mechanism [18] that expands the receptive field and captures long-range dependencies, we can extract the contextual and global representations from the whole aurora images. Furthermore, we design a semantic attention model to merge local and global feature maps, and it can bridge the semantic gap caused by unbefitting feature fusion. In general, the contribution of this work can be summarized as (1) This work aims to assist physicists in their auroras research by automatically retrieving interested images from large-scale all-sky images dataset; (2) The proposed model can simultaneously capture both local receptive field and global receptive field of aurora images, which eliminate human bias and improve retrieval accuracy; (3) We attempt to bridge the semantic gap through a well-designed semantic attention model that contributes to the fusion of different feature representation and interpretability of aurora retrieval model. (4) The experiment result demonstrated that our method achieved superior performance over the state-of-the-art methods.

The remaining section of this manuscript is organized as follows. After introducing related works about aurora image retrieval from all-sky images (Section 1), we presented the details of our proposed deep learning model and the OATH [1] dataset in Section 2. Experiment results of the proposed hash-transformer retrieval model on the OATH dataset are described in Section 3. Finally, the discussion and conclusion of the experiments' outcome are provided in Section 4.

Table 1
Details and explanation of the OATH dataset.

Label	Quantity	Explanation
arc	774	Bounds of aurora across receptive field and have well-defined and sharp edges.
diffuse	1102	The brightness of aurora similar to stars and have blurred edges.
discrete	1400	The brightness of aurora is higher than stars and have well-defined and sharp edges.
cloudy	852	The sky dominated by clouds.
moon	614	The image is consist of the light of the moon.
clear	1082	The image is clear or no aurora.
Total	5824	-

2. Materials and methods

This work introduces a Hash-Transformer aurora retrieval model based on the ViT to assist physicists in selecting their interested aurora images from large-scale all-sky datasets in an automated manner. In particular, the contextual information of interested auroras is extracted from the global feature representation, and it can be captured through long-range dependencies under the self-attention mechanism. Moreover, the proposed semantic attention model merges feature maps from different environments. For example, shapes, textures, and rings of auroras are low-level feature representations, and contextual around auroras in all-sky images is a high-level representation. Furthermore, all experiments are performed on the real-world dataset to guarantee fairness compared to other algorithms. Notably, all algorithms follow an end-to-end strategy, and the dataset is first divided into training and testing sets. In addition, our method can also be performed on other large-scale datasets in a training-testing manner.

2.1. Dataset

Oslo Aurora THEMIS (OATH) dataset collected from Time History of Events and Macroscale Interactions during Substorms (THEMIS) [19] all-sky imaging instruments, which were manually labeled by Clausen et al. [1] and labels including arc, diffuse, discrete, cloudy, moon, and clear. Although the categories overlap, they are only somewhat consistent. Details about the aurora samples of the OATH dataset are provided in Table 1.

According to previous work [1], we also follow their pre-processing operations. Specifically, 5824 all-sky images are randomly selected from the OATH dataset. The raw aurora images are cropped by 15% to eliminate irrelevant pixels. Subsequently, the intensity of each raw aurora image is normalized to the interval [0, 1] to enhance the contrast of aurora regions. Fig. 1 illustrates some samples from the OATH dataset. Notably, the categories of all-sky images are not exclusive. For example, the top three rows (in red) represent the all-sky image with the aurora, and the bottom three (in blue) represent the all-sky image without the aurora.

2.2. Overview of proposed model

As illustrated in Fig. 2(a)-(e), the pipeline of our model mainly includes three phases, pre-training on ImageNet under classification task, fine-tuning on aurora images, and online search.

Pre-training on ImageNet under classification task. In our study, we perform a ViT-based hash model for aurora image retrieval from all-sky images. The weight of pre-training in our model follows the standard vision transformer [17], which trains the final output of the model on the ImageNet dataset under the image classification task. In particular, input images are first split into 16 or 32 patches and linearly embedding them. After that, the well-designed linear embeddings combined with unique position embeddings are fed into a standard transformer encoder. Finally, the classification head is attached to the last layer of the transformer encoder. To reduce calculated consumption, we split input images into patch sizes 16. The pre-training model is publicly available on GitHub.[1]

Fine-tuning on aurora images. Since images in ImageNet are ordinary images captured from daily life, we provided fine-tuning on aurora images to adapt higher resolution aurora images. Specifically, the MLP head for image classification in the ViT model is replaced by a hash block to achieve the aurora image retrieval task. The feature map after the transformer encoder is reshaped and through the dropout layer. During the nonlinear activation operation and a linear projection, the feature map is embedded into a size of 1024. At last, another linear projection layer is performed to generate hash codes.

Online search. When given a query aurora image, the same operations in the vision transformer model are performed to finish the feature extraction phase. Subsequently, the hash representation is obtained via a well-designed hash block, and it generates the hash code. Then, the hash code of the query image is compared with the hash code index of fine-tuning phase to calculate the

[1] https://github.com/jeonsworld/ViT-pytorch.

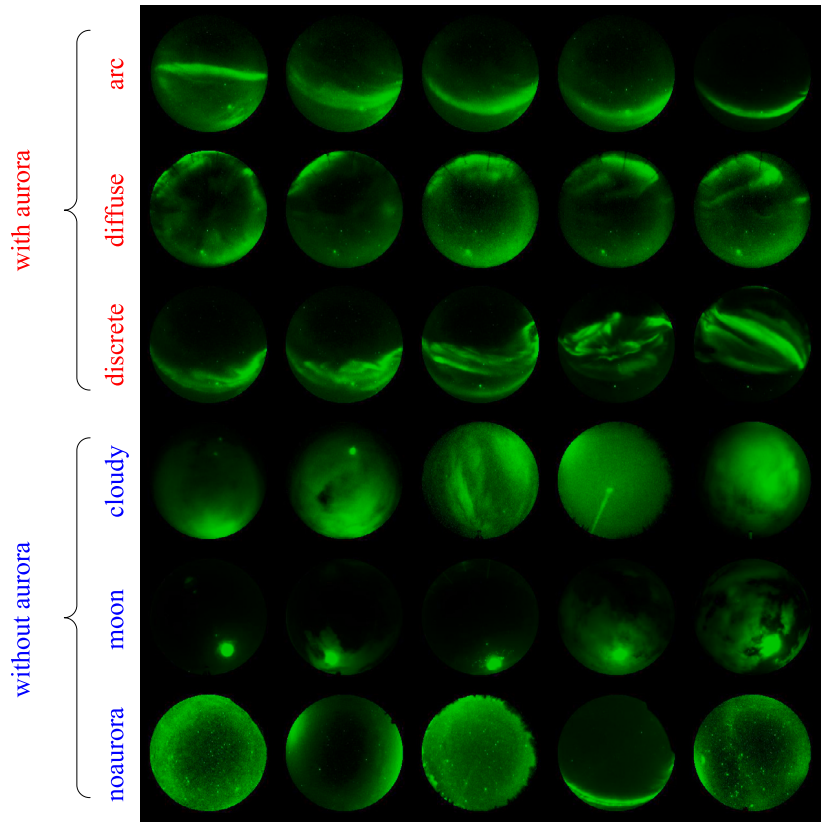

Fig. 1. Samples of all-sky images in OATH dataset. The top three rows (in red) represent the all-sky image with the aurora, and the bottom three (in blue) represent the all-sky image without the aurora.

similarity score. Finally, the aurora image with the highest similarity score is retrieved from the large-scale dataset (i.e., the OATH dataset).

2.3. Hash-Transformer model

An overview of the proposed Hash-Transformer model is depicted in Fig. 3. Motivated by the Transformer-based Hamming Hashing model [20–22] and the Vision Transformer Hashing model [23], our aurora retrieval model includes four stages, patch embedding, position embedding, transformer encoder, and hash model.

Patch embedding and Position embedding. According to the standard Transformer, the received input is a one-dimensional sequence of token embeddings. To handle aurora images (two-dimensional sequence), we split the aurora image $I \in R^{H\times W\times C}$ into several two-dimensional patches $I_P \in R^{N\times(P^2\cdot C)}$, where (H, W) denote the height and width of the aurora image, C denotes the number of channels, P denotes the length of each two-dimensional patch, and $N = HW/P^2$ is the number of patches. Subsequently, the patches are flattened and projected to a D-dimensional linear embedding, a trained linear projection. After linear projection, the position embeddings are added to the patch embeddings to preserve spatial and Position information of aurora images. The resulting embedding serves as input to the transformer encoder stage.

Transformer encoder. Fig. 3(a) shows that we use the transformer encoder to extract feature representations. Specifically, the transformer encoder includes the stack of L transformer blocks, each with a consistent structure. Notably, we replace the multi-head self-attention (MSA) [18] of the standard transformer encoder with the Semantic Attention Model (SAM) to merge local and global feature representations, as depicted in Fig. 4. After the standard transformer encoder update, it consists of SAM and Multi-layer perceptrons (MLP) blocks. Before every block, the first layer performed is layer normalization (LN). Subsequently, the output of the LN serves as the input of SAM which has three projections for Query, Key, and Value tensors. In addition, the residual mechanism and skip connections are applied to the updated structure. The final output includes the MLP block with dropout layers and a GELU [24] activation function.

Hash model. As shown in Fig. 3(b), to facilitate aurora image retrieval and obtain the hash code, we design a hash block behind the output layer of the transformer encoder to explore hash features. In particular, the hash block comprises a group of operations that includes double convolutions, batch normalization, and ReLU activation. Furthermore, a dropout and a linear projection layer are performed to change the dimension of feature maps behind the ReLU activation function. Finally, the hash feature extracted from the hash block is applied to generate the hash code. The hash model consists of the hash block and hash code.

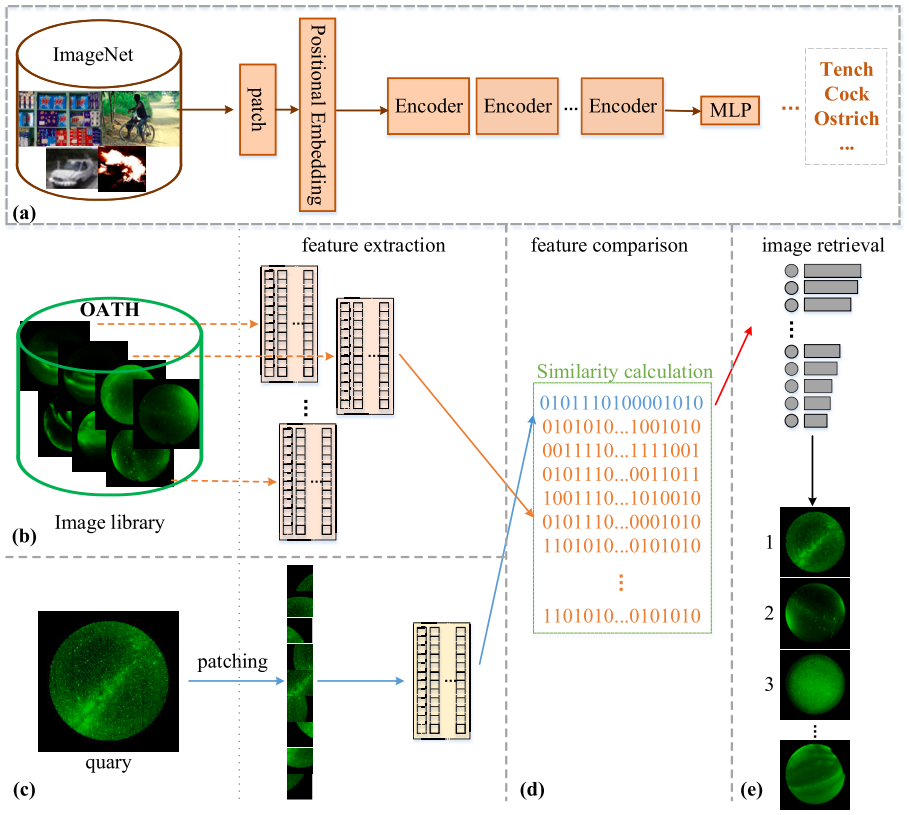

Fig. 2. The pipeline of the proposed method for aurora image retrieval. (a) Pre-training; (b) Fine-tuning; (c) Online search; (d) Feature comparison; (e) image retrieval.

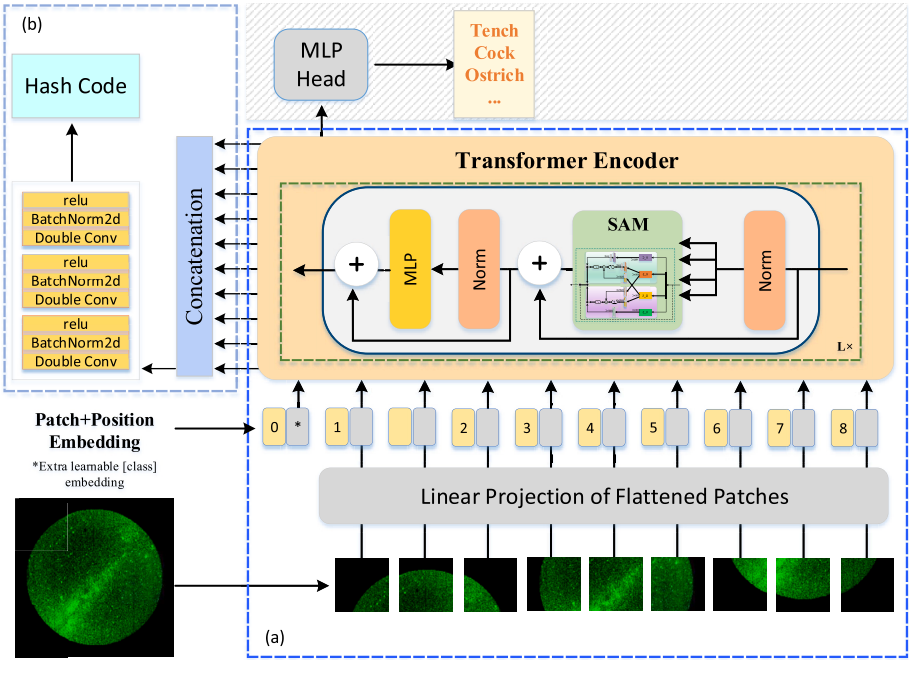

Fig. 3. The overview of the proposed Hash-Transformer model. (a) Embedding and transformer encoder; (b) Hash block.

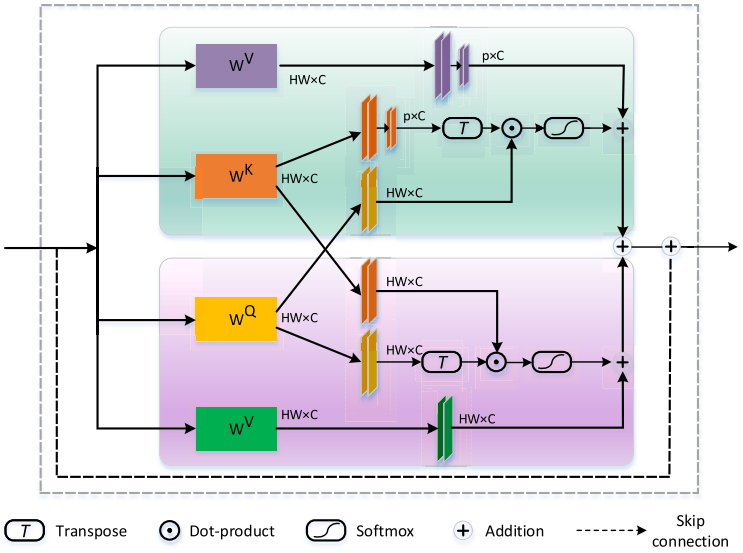

Fig. 4. The structure of the semantic attention model consists of local-attention (**top**) and global-attention (**bottom**) branches.

2.4. Semantic attention model

To bridge the semantic gap during the fusion of different feature representations, we designed a semantic attention model to enhance the interaction between local and global information, which shares the parameter in calculating Query and Key tensors. Furthermore, motivated by the efficient paired-attention mechanism [25], we replace the multi-head self-attention model of the standard transformer encoder with the semantic attention model, which is more suitable for aurora image retrieval because spatial and semantic features can be used simultaneously. The semantic attention model is shown in Fig. 4. The feature maps of local ($\hat{\boldsymbol{X}}_l$) and global ($\hat{\boldsymbol{X}}_g$) attention are formulated as follows:

$$\hat{\boldsymbol{X}}_l = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{shared}}\,\boldsymbol{K}^{\top}_{\mathrm{projected}}}{\sqrt{d}}\right)\cdot \tilde{\boldsymbol{V}}_{\mathrm{local}}\,, \tag{1}$$

$$\hat{\boldsymbol{X}}_g = \tilde{\boldsymbol{V}}_{\mathrm{global}}\cdot \mathrm{Softmax}\left(\frac{\boldsymbol{K}_{\mathrm{shared}}\,\boldsymbol{Q}^{\top}_{\mathrm{shared}}}{\sqrt{d}}\right), \tag{2}$$

$$\hat{\boldsymbol{X}} = Conv_{1\times1}(Conv_{3\times3}(\hat{\boldsymbol{X}}_l + \hat{\boldsymbol{X}}_g)), \tag{3}$$

where the output of the semantic attention model is $\hat{\boldsymbol{X}}$. $\boldsymbol{Q}_{\mathrm{shared}}$ and $\boldsymbol{K}_{\mathrm{shared}}$ denote shared Queries and Keys. $\boldsymbol{K}^{\top}_{\mathrm{projected}}$ and $\tilde{\boldsymbol{V}}_{\mathrm{local}}$ denote projected shared Keys and projected local value, respectively.

3. Experiment and results

3.1. Implementation details and evaluation metrics

Our experiments are performed on a computer with Intel Xeon® Silver 4314R CPU, NVIDIA GeForce GTX 3080 GPU, and 64 GB memory. The computations are implemented on CentOS7 64-bit platform with PyTorch. In the training phase, we utilized the Adam optimizer with a learning rate $1e^{-4}$ and weight decay of 0.8. The entire model was trained for 500 epochs with a batch size 32. Furthermore, the patch size is set to 16. Thus, the number of patches is $N = 196$. And the number of Transformer blocks $L = 12$.

To evaluate the performance of our proposed method and compare it with state-of-the-art algorithms, we compute the Average Precision (AP) [26], a standard evaluation metric for image retrieval tasks. The Mean Average Precision (mAP) is the proportion of accurately retrieved images to the total retrieved images.

$$AP = (1/R) * \sum_{k=1}^{R} P(k)\cdot \mathrm{rel}_k, \tag{4}$$

in this formula, R represents the total number of retrieval results, $P(k)$ is the precision at the $k-th$ retrieved item, and rel_k indicates the relevance of the $k-th$ retrieved item.

$$mAP = (1/N) * \sum_{i=1}^{N} AP_i, \tag{5}$$

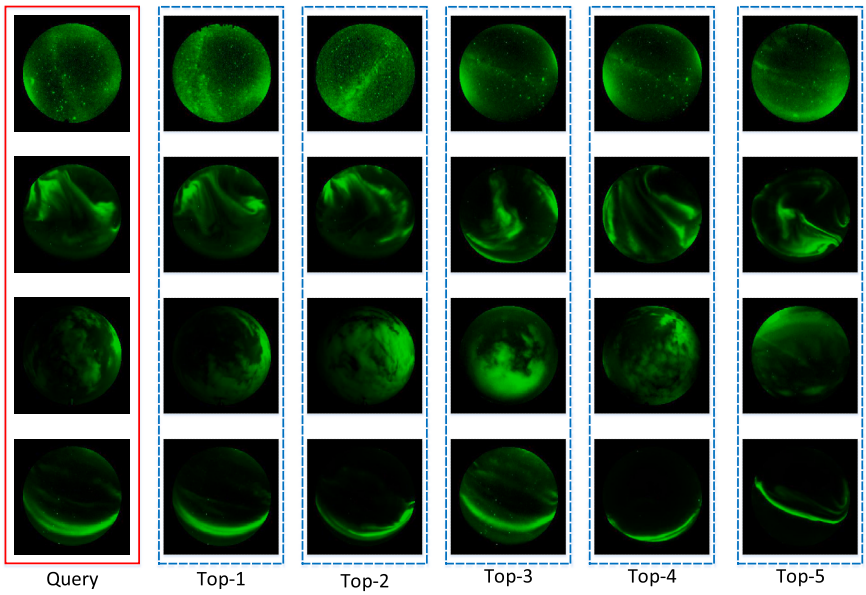

Fig. 5. The structure of the semantic attention model consists of local-attention (**top**) and global-attention (**bottom**) branches.

Table 2
The aurora retrieval results comparison in terms of mAP, Recall, and Precision with state-of-the-art methods on the OATH dataset.

Methods	mAP	Recall	Precision
SIFT [8]	0.73	0.80	0.74
R-CNN [27]	0.85	0.82	0.89
Mask-CNN [12]	0.88	0.84	0.86
HashNet [28]	0.91	0.83	0.90
TransHash [20]	0.93	0.86	0.95
Our	0.96	0.92	0.97

where N denotes the number of query images, AP_i denotes the average precision for each query image. Besides, precision and recall are also used to judge the performance of aurora image retrieval. The precision and recall can be formulated as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{6}$$

$$Recall(Sensitivity) = \frac{TP}{TP + FN}, \tag{7}$$

where TP, FP, and FN denote true positive, false positive, and false negative, respectively. In summary, Average Precision (AP) represents the average accuracy of retrieval results at different Recall levels. It is calculated by multiplying the Precision of each retrieval result by its relevance label, summing them up, and dividing by the total number.

3.2. Experimental results

Fig. 5 illustrates the retrieval performance of aurora images on the OATH dataset. We present some examples of top-ranking images by our model. The proposed method can accurately retrieve similar aurora images related to the queries. To further evaluate the effectiveness of the proposed Hash-Transformer retrieval model, we compare it with the state-of-the-art retrieval models by using the same dataset. We divide baseline models into two categories, i.e., CNN-based models and Hash-based models, including the Scale-invariant feature transform (SIFT), Region-convolutional neural network (R-CNN), Mask-convolutional neural network (Mask-CNN), HashNet, and TransHash. As demonstrated in Table 2, our methods achieve satisfactory results and superior performance compared with the existing models.

3.3. Ablation study

To further verify the effectiveness of the hyper-parameters of our model, we evaluate different batch size settings in the OATH dataset. As shown in Fig. 6, our experiments' best batch size value is 32. We also conduct mAP, Recall, and Precision experiments

Fig. 6. Visualization of the batch size with different settings.

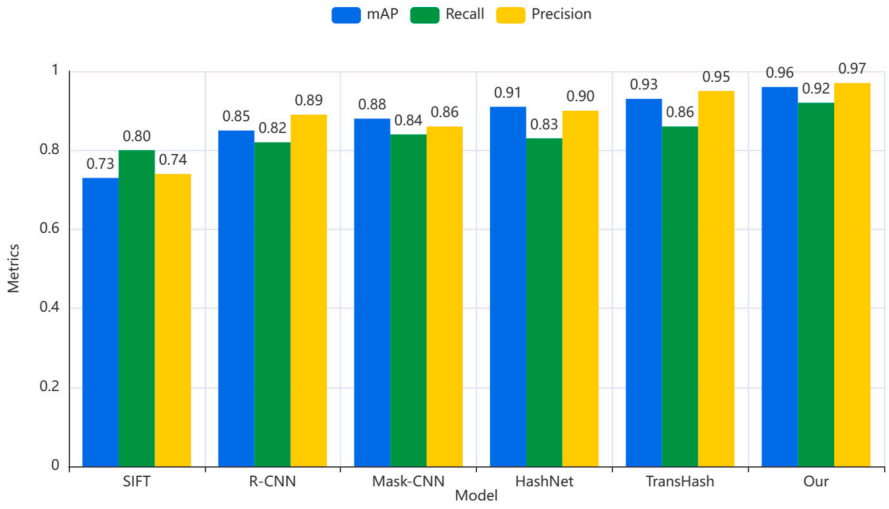

Fig. 7. Comparison with state-of-the-art models on the OATH dataset.

in Table 2. Our proposed SAGS-Net outperformed these models and achieved an mAP of 96%, Recall of 92%, and Precision of 97%. Furthermore, we validate our model on the OATH dataset to confirm the robustness and generalization. Various state-of-the-art models based on CNN-based and Hash-based are compared. As illustrated in Fig. 7, our model outperforms existing models on the real-world dataset. In summary, the ablation results of our model on the real-world dataset demonstrated a notable performance improvement. It can be attributed to as follows. On the one hand, our model increases the performance of global-local feature utilization to eliminate human bias and improve retrieval accuracy. On the other hand, we bridge the semantic gap through the semantic attention model that contributes to the fusion of different feature representations simultaneously.

4. Conclusions

In this work, we proposed a Hash-Transformer approach for retrieving aurora images from the large-scale dataset, which provides convenience to physicists in selecting their interested auroras from all-sky images. Unlike the existing aurora retrieval methods that only focus on the utilization of spatial features in aurora images, we attempt to explore the contextual information around the whole image via the vision transformer with long-range dependencies. We further introduce a semantic attention model to capture local and global feature representations during the feature extraction phase. Accordingly, local and global receptive fields from aurora areas can be obtained to improve the performance of auroras retrieval in all-sky images. In addition, the well-designed semantic attention model is embedded into the standard transformer encoder to bridge the semantic gap. In experiments, our approach is performed on the OATH dataset with six categories (including arc, diffuse, discrete, cloudy, moon, and clear) in a training-testing manner. We calculate the mAP, recall, and precision scores to evaluate the performance of our proposed method and compare it with state-of-

the-art algorithms. The experiment results demonstrated that our method achieved superior performance over the state-of-the-art methods.

Although our method has progressed in aurora image retrieval, this study has disadvantages. Limited by the number of labeled images, only one real-world dataset is used in the experiments. Thus a private aurora dataset should be used in future work. In addition, a few auroras struggle to retrieve because the shape of auroras is variable and irregular, affecting our retrieval model's performance.

In the future, the Hash-Transformer retrieval method will perform another real-world dataset containing more aurora images and more types of auroras. Furthermore, the prompt mechanism will be embedded into the process of retrieval.

Institutional review

Not applicable.

Abbreviations

The following abbreviations are used in this manuscript:

MSA	Multi-head self-attention
SAM	Semantic attention model
MLP	Multi-layer perceptrons
LN	Layer normalization
AP	Average precision
mAP	Mean average precision
TP	True positive
FP	False positive
FN	False negative

CRediT authorship contribution statement

Hengyue Zhang: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Hailiang Tang:** Conceptualization, Data curation, Formal analysis, Methodology, Software, Supervision, Writing – original draft, Writing – review & editing. **Wenxiao Zhang:** Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in this research is publicly available in the OATH dataset at https://doi.org/10.1029/2018JA025274, reference number [1].

References

[1] L.B.N. Clausen, H. Nickisch, Automatic classification of auroral images from the Oslo auroral THEMIS (OATH) data set using machine learning, J. Geophys. Res. Space Phys. 123 (7) (2018) 5640–5647, https://doi.org/10.1029/2018JA025274, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018JA025274, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JA025274.

[2] J.E. Borovsky, J.A. Valdivia, The Earth's magnetosphere: a systems science overview and assessment, Surv. Geophys. 39 (5) (2018) 817–859, https://doi.org/10.1007/s10712-018-9487-x.

[3] X. Yang, N. Wang, B. Song, X. Gao, BoSR: a CNN-based aurora image retrieval method, Neural Netw. 116 (2019) 188–197, https://doi.org/10.1016/j.neunet.2019.04.012, https://www.sciencedirect.com/science/article/pii/S0893608019301108.

[4] J. Lian, T. Liu, Y. Zhou, Aurora classification in all-sky images via CNN–transformer, Universe 9 (5) (2023), https://doi.org/10.3390/universe9050230, https://www.mdpi.com/2218-1997/9/5/230.

[5] M.T. Syrjäsuo, E.F. Donovan, Diurnal auroral occurrence statistics obtained via machine vision, Ann. Geophys. 22 (4) (2004) 1103–1113, https://doi.org/10.5194/angeo-22-1103-2004, https://angeo.copernicus.org/articles/22/1103/2004/.

[6] Q. Wang, J. Liang, Z.-J. Hu, H.-H. Hu, H. Zhao, H.-Q. Hu, X. Gao, H. Yang, Spatial texture based automatic classification of dayside aurora in all-sky images, J. Atmos. Sol.-Terr. Phys. 72 (5) (2010) 498–508, https://doi.org/10.1016/j.jastp.2010.01.011, https://www.sciencedirect.com/science/article/pii/S1364682610000441.

[7] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987, https://doi.org/10.1109/TPAMI.2002.1017623.

[8] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110, https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[9] A. Bosch, A. Zisserman, X. Munoz, Image classification using random forests and ferns, in: 2007 IEEE 11th International Conference on Computer Vision, 2007, pp. 1–8.

[10] X. Yang, X. Gao, Q. Tian, Polar embedding for aurora image retrieval, IEEE Trans. Image Process. 24 (11) (2015) 3332–3344, https://doi.org/10.1109/TIP.2015.2442913.
[11] X. Yang, N. Wang, B. Song, X. Gao, Aurora image search with saliency deep features, IEEE Access 7 (2019) 65996–66006, https://doi.org/10.1109/ACCESS.2019.2917723.
[12] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.
[13] X. Yang, X. Gao, B. Song, B. Han, Hierarchical deep embedding for aurora image retrieval, IEEE Trans. Cybern. 51 (12) (2021) 5773–5785, https://doi.org/10.1109/TCYB.2019.2959261.
[14] L. Zheng, Y. Yang, Q. Tian, Sift meets CNN: a decade survey of instance retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 40 (5) (2018) 1224–1244, https://doi.org/10.1109/TPAMI.2017.2709749.
[15] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, S. Yan, Cross-modal retrieval with CNN visual features: a new baseline, IEEE Trans. Cybern. 47 (2) (2017) 449–460, https://doi.org/10.1109/TCYB.2016.2519449.
[16] H. Liu, R. Wang, S. Shan, X. Chen, Deep supervised hashing for fast image retrieval, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2064–2072.
[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021, https://openreview.net/forum?id=YicbFdNTTy.
[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 6000–6010.
[19] E. Donovan, S. Mende, B. Jackel, H. Frey, M. Syrjäsuo, I. Voronkov, T. Trondsen, L. Peticolas, V. Angelopoulos, S. Harris, M. Greffen, M. Connors, The THEMIS all-sky imaging array—system design and initial results from the prototype imager, J. Atmos. Sol.-Terr. Phys. 68 (13) (2006) 1472–1487, https://doi.org/10.1016/j.jastp.2005.03.027, passive Optics Aeronomy, https://www.sciencedirect.com/science/article/pii/S1364682606001118.
[20] Y. Chen, S. Zhang, F. Liu, Z. Chang, M. Ye, Z. Qi, TransHash: transformer-based hamming hashing for efficient image retrieval, in: Proceedings of the 2022 International Conference on Multimedia Retrieval, ICMR '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 127–136.
[21] A. El-Nouby, N. Neverova, I. Laptev, H. Jégou, Training vision transformers for image retrieval, arXiv:2102.05644, 2021.
[22] C.H. Song, J. Yoon, S. Choi, Y. Avrithis, Boosting vision transformers for image retrieval, arXiv:2210.11909, 2022.
[23] S.R. Dubey, S.K. Singh, W.-T. Chu, Vision transformer hashing for image retrieval, in: 2022 IEEE International Conference on Multimedia and Expo (ICME), 2022, pp. 1–6.
[24] D. Hendrycks, K. Gimpel, Gaussian error linear units (GELUs), arXiv:1606.08415, 2020.
[25] A. Shaker, M. Maaz, H. Rasheed, S. Khan, M.-H. Yang, F.S. Khan, UNETR++: delving into efficient and accurate 3D medical image segmentation, arXiv:2212.04497, 2023.
[26] S.R. Dubey, A decade survey of content based image retrieval using deep learning, IEEE Trans. Circuits Syst. Video Technol. 32 (5) (2022) 2687–2704, https://doi.org/10.1109/TCSVT.2021.3080920.
[27] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '14, IEEE Computer Society, USA, 2014, pp. 580–587.
[28] Z. Cao, M. Long, J. Wang, P.S. Yu, HashNet: deep learning to hash by continuation, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5609–5618.