Hybridizing Euclidean and Hyperbolic Similarities for Attentively Refining Representations in Semantic Segmentation of Remote Sensing Images

Xin Li, Feng Xu, Fan Liu, Runliang Xia, Yao Tong, Linyang Li, Zhennan Xu, and Xin Lyu

***Abstract*— Attention mechanisms (AMs) have revolutionized the semantic segmentation network in interpreting remote sensing images (RSIs) due to their amazing ability in establishing contextual dependencies. Nevertheless, due to the complex scenes and diverse objects in RSIs, a variety of details and correlations are not available in Euclidean space. Therefore, a similarity-hybrid attention module (SHAM) is devised to attentively learn the hyperbolic and Euclidean attention maps between any two positions, followed by a weighted elementwise summation. The hybrid attention maps posses latent geometric properties of both Euclidean and hyperboloid. Taking commonly used fully convolutional network (FCN) as baseline, hybrid attention-enhanced neural network (HAENet) that embeds SHAM is presented. Experiments on International Society for Photogrammetry and Remote Sensing (ISPRS) Potsdam and DeepGlobe benchmarks reveal its superiority to comparative methods. In addition, the ablation study validates the effectiveness of SHAM compared with other attention modules.**

***Index Terms*— Attention mechanism (AM), hyperbolic geometry, semantic segmentation, similarity-hybrid attention.**

I. Introduction

SEMANTIC segmentation (SS) is essential for accurately interpreting remote sensing images (RSIs). Given an input RSI, SS generates the corresponding pixel-level labels [1]. Therefore, SS plays a vital role in many applications, including water resources management, land cover mapping, and hazard assessment.

Manuscript received 5 September 2022; revised 14 November 2022; accepted 26 November 2022. Date of publication 30 November 2022; date of current version 15 December 2022. This work was supported in part by the Excellent Post-Doctoral Program of Jiangsu Province under Grant 2022ZB166; in part by the National Natural Science Foundation of China under Grant 42104033, Grant 51779100, and Grant 51679103; in part by the Project of Water Science and Technology of Jiangsu Province under Grant 2021063, Grant 2021072, and Grant 2021080; and in part by the Fundamental Research Funds for the Central Universities under Grant B210202080. *(Corresponding author: Feng Xu.)*

Xin Li, Feng Xu, Fan Liu, Zhennan Xu, and Xin Lyu are with the College of Computer and Information and the Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing 211100, China (e-mail: li-xin@hhu.edu.cn; xufeng@hhu.edu.cn; liufan@hhu.edu.cn; zhennanxu@hhu.edu.cn; lvxin@hhu.edu.cn).

Runliang Xia is with the Information Engineering Center, Yellow River Institute of Hydraulic Research, Zhengzhou 450003, China (e-mail: xiarunliang@hky.yrcc.gov.cn).

Yao Tong is with the School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450000, China (e-mail: yaotong@uw.edu).

Linyang Li is with the Surveying and Mapping Institute, PLA Information Engineering University, Zhengzhou 450003, China (e-mail: lilinyang810810@163.com).

Digital Object Identifier 10.1109/LGRS.2022.3225713

Fundamentally, fully convolutional networks (FCNs) exhibit impressive performance in extracting rich features and extend to SS for RSI. However, FCN-based approaches are inherently affected by limited perceptual fields and local context [1].

To capture contextual information from a broader range, the atrous spatial pyramid pooling (ASPP) with multiscale dilation rates was proposed [2]. However, the global context is uncovered due to the inherence of stacking convolutional layers.

Alternatively, the attention mechanism (AM) provides an efficient way of capturing and incorporating global context [3]. An inchoate work, SENet [4], recalibrates channelwise weights to highlight informative feature channels of feature maps. Furthermore, as a mailstone, nonlocal neural network (NLNNet) utilizes a self-AM (SAM), modeling positionwise correlations to refine input features [5]. Afterward, the SAM has been the primary choice in capturing long-range contextual information for SS and the formed Transformer also reached a desired success [6]. Specifically, several RSI-targeted networks, HCANet [7], LANet [8], and HMANet [9], have remarkably promoted segmentation accuracy by designing novel SAM variants with typical deployment. Nonetheless, the existing methods are defined in Euclidean space, in which the features are flattened to fulfill the Euclidean geometry axiom [10].

However, Bronstein et al. [11] have proved that the images also exhibit a highly non-Euclidean latent anatomy. Besides, it appears in several applications that the dissimilarity measures constructed by experts tend to have non-Euclidean behavior. Therefore, the Euclidean space cannot provide the most powerful or meaningful geometrical representations. It is necessary to exploit hyperbolic representations and take advantage of this property.

Since Ganea et al. [12] derived hyperbolic neural network, projecting feature vectors in Euclidean to hyperboloid endows the representations to capture fundamental data properties, including non-Euclidean visual phenomena and clustering behavior [13]. Especially in RSI, the imaging altitude is always high, bringing the distortions represented in Euclidean space. Therefore, the Euclidean vectorwise similarity is suboptimal in describing the two elements. The latent non-Euclidean similarity is equally essential to measure the elementwise similarity. Apart from designing loss function to fine-tune

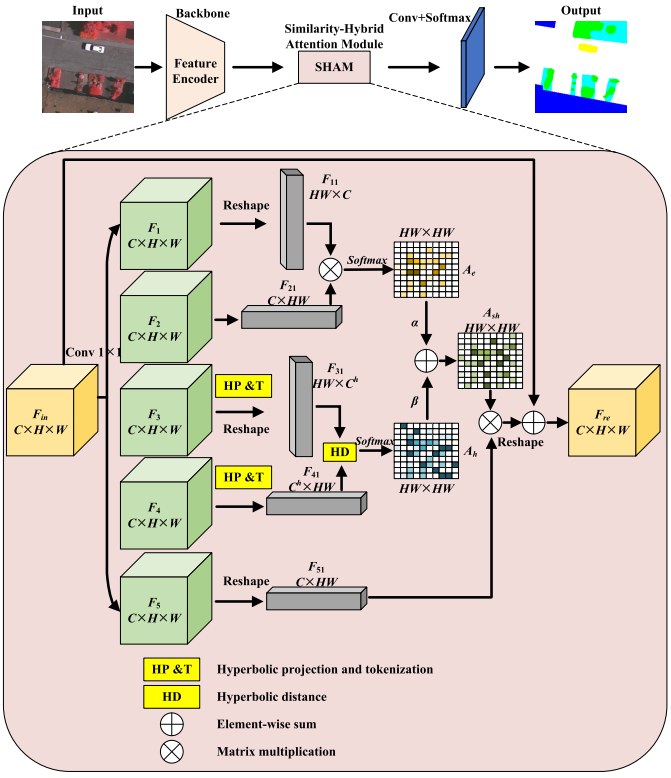

Fig. 1. Framework of HAENet.

the training of a network [14], it is suggested to incorporate hyperbolic representations into network directly.

Motivated by the SAM, a similarity hybrid attention module (SHAM) is proposed to generate and fuse Euclidean and hyperbolic similarities, which are measured with Euclidean and hyperbolic representations. Specifically, we devise a hyperbolic projection and tokenization flow to adapt to the normal attention workflow based on a pseudo-hyperboloid. Therefore, the calculated hyperbolic distance (HD) generates a hyperbolic attention map (HAM). Two contributions are summarized as follows.

1) To comprehensively and effectively exploit Euclidean and hyperbolic geometric representations, we propose an SHAM, which post-fuses the contextual affinities of the two spaces. In addition to performing self-attention in Euclidean space, two parallel branches form a pseudo-hyperboloid to project feature vectors and measure their correlations in hyperbolic space. In this way, the post-fused attention map involves the similarities of Euclidean and hyperbolic representations, making the refined features discriminative. Based on SHAM, the hybrid attention-enhanced neural network (HAENet) is devised to segment RSIs. Experiments on the Potsdam and DeepGlobe benchmarks exhibit competitive performance. Moreover, the ablation study demonstrates the effects of SHAM.

II. Method

A. Overview

As illustrated in Fig. 1, HAENet inherits the encoder–decoder architecture. With an input image, the feature encoder outputs the corresponding representation. Then, the representation is fed into the SHAM. Afterward, the encoded features are refined with global context by a high-fidelity similarity, which derives from hetero-curvature spaces that carry different visual properties. At last, the upsampled SHAM-refined feature is used to classify.

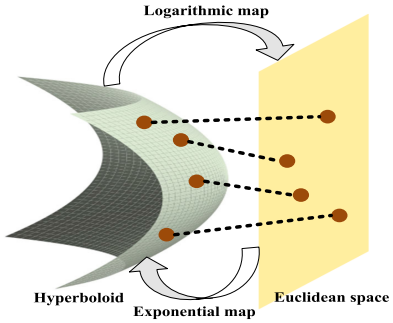

Fig. 2. Space transformation.

B. Hyperbolic Projection and Tokenization

Ascribing to Riemannian manifold and metric, the hyperbolic projection and tokenization module, termed HP&T, enables project feature vector (Euclidean) onto hyperboloid (Poincaré model in this study) with exponential map (see Fig. 2). Formally

$$\exp_v^c(\boldsymbol{x}) = v \oplus_c \left(\tanh\left(\frac{\sqrt{c}\lambda_v^c \|\boldsymbol{x}\|}{2} \right) \frac{\boldsymbol{x}}{\sqrt{c}\|\boldsymbol{x}\|} \right) \tag{1}$$

where $\boldsymbol{x}$ is the feature vector in Euclidean space, $\boldsymbol{v}$ is the anchor, c is a hyperparameter governing curvature and radius of the Poincaré model, $\lambda_v^c = 2(1 - c\|\boldsymbol{x}\|^2)^{-1}$ is a conformal factor, and $\oplus_c$ denotes Möbius addition. The anchor is set to the origin in practice; therefore, (1) turns to

$$\exp_0(\boldsymbol{x}) = \tanh\left(\sqrt{c}\|\boldsymbol{x}\|\right) \frac{\boldsymbol{x}}{\sqrt{c}\|\boldsymbol{x}\|}. \tag{2}$$

After projection, every Euclidean vector has its hyperbolic counterpart. We endow them with sequential positions to achieve tokenization. Thus, HP&T produces position-assisted hyperbolic representations for subsequent operation.

C. Hyperbolic Distance

Analogous to Euclidean space, HD is used to measure the similarity between two arbitrary gyrovectors (see Fig. 3). Inherently, the HD concerns the hyperbolic properties of a specific position in RSI, compensating for the information loss of the Euclidean vector. In this study, the induced distance is given as follows:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \cosh^{-1}\left(1 + 2\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{\left(1 - \|\boldsymbol{x}\|^2\right)\left(1 - \|\boldsymbol{y}\|^2\right)} \right) \tag{3}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are two gyrovectors and belong to one Poincaré model. Therefore, HD is capable of being normalized to provide hyperbolic similarity in accordance with each pair of positions.

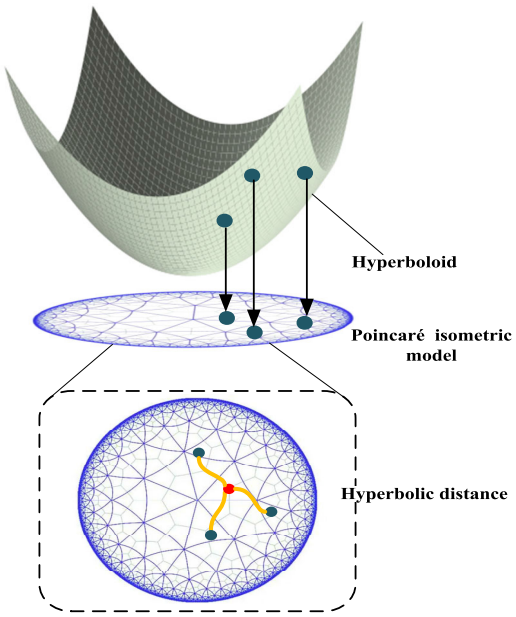

Fig. 3. Illustration of HD with the Poincaré model.

D. Similarity-Hybrid Attention Module

SHAM unitedly models and aggregates self-attentive dependencies in Euclidean and hyperboloid spaces in parallel. As depicted in Fig. 1 (pink box), five sub-branches are devised. Initially, the input feature is convolved with a kernel size of 1×1, generating five representations for self-attention paradigm. With the first and second branches (from the top), the vanilla self-attention that captures long-range dependencies in Euclidean space is implemented. Formally

$$A_e = \text{Softmax}\left(\frac{F_{11} \times F_{21}}{\sqrt{C}}\right) \tag{4}$$

where $A_e \in \mathbb{R}^{\text{HW}\times\text{HW}}$ is the attention map that implies Euclidean similarity, $F_{11} \in \mathbb{R}^{\text{HW}\times C}$ and $F_{21} \in \mathbb{R}^{C\times\text{HW}}$ are feature maps (Euclidean), and C is the channel number.

To compute hyperbolic similarity matrix, the convolved F_3 and F_4 are fed into HP&T initially. Then, one of the tokenized gyrovectors associated with the sequential positions is reshaped with a size of HW $\times$ C^h, while the other one is with $C^h \times$ HW, where C^h is the dimension of gyrovector. Referring to (3), the pairwise distances are obtained followed by a Softmax layer:

$$A_h = \text{Softmax}\left(\frac{F_{31} \times F_{41}}{\sqrt{C^h}}\right) \tag{5}$$

where $A_h \in \mathbb{R}^{\text{HW}\times\text{HW}}$ is the attention map that implies hyperbolic similarity, $F_{31} \in \mathbb{R}^{\text{HW}\times C^h}$ and $F_{41} \in \mathbb{R}^{C^h\times\text{HW}}$ are gyrovectors (hyperbolic), and C^h is the dimension of gyrovectors.

Attempting to aggregate the similarity in hetero-curvature spaces, a weighted summation is applied to A_e and A_h

$$A_{sh} = \alpha \cdot A_e + \beta \cdot A_h \tag{6}$$

where $A_{\text{sh}} \in \mathbb{R}^{\text{HW}\times\text{HW}}$ is the similarity-hybrid attention map, and α and β are two learnable coefficients (initially set as 0.5 and 0.5).

Eventually, the refined feature maps F_{re} are given as follows:

$$F_{re} = F_{\text{in}} \oplus (F_{51} \times A_{sh}) \tag{7}$$

where $F_{51} \in \mathbb{R}^{C\times\text{HW}}$, and $\oplus$ is the elementwise summation.

In summary, vanilla self-attention still suffers from uncertainty in segmenting RSIs, especially for easy-confused and edge-surrounding pixels, because the similarity in Euclidean space cannot provide sufficient discriminability. SHAM allows the network learn the hybrid similarity in a single module with acceptable computations. Consequently, the refined features could supply more comprehensive contextual cues for inference.

III. Experiments

A. Datasets

Two benchmarks, namely, International Society for Photogrammetry and Remote Sensing (ISPRS) Potsdam and DeepGlobe, are examined. Thirty-eight aerial imagery tiles (20/4/14 tiles for training/validation/test) are collected and annotated for the Potsdam benchmark with six categories (clutter is ignored in evaluation). Every tile has a size of 6000 $\times$ 6000 with a spatial resolution of 5 cm. The DeepGlobe benchmark is acquired from a satellite platform with a spatial resolution of 50 cm. Specifically, 1146 images of size 2448 $\times$ 2448 pixels are available (803/171/172 images for training/validation/test).

B. Evaluation Metrics

To evaluate the performance, we calculate the pixelwise intersection over union (IoU) with formula

$$\text{IoU}_j = \frac{\sum_{i=1}^{n} TP_{ij}}{\sum_{i=1}^{n} TP_{ij} + \sum_{i=1}^{n} FP_{ij} + \sum_{i=1}^{n} FN_{ij}} \tag{8}$$

where TPs, FPs, and FNs are the number of true positives, false positives, and false negatives in image i with class j, respectively. Moreover, the mIoU over k classes is given as $\text{mIoU} = (1/k)\sum_{j=1}^{k} \text{IoU}_j$. F_1 score and OA are as follows:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{9}$$

$$OA = \frac{TP + TN}{TP + FP + FN + TN} \tag{10}$$

where precision = (TP/TP + FP), and recall = (FP/TP + FN).

C. Implement Details

Three bands, R, G and B, are used as the input channels. In practice, we split the raw image into subpatches with a spatial size of 256 $\times$ 256 for training. With the Tesla V100-32 GB GPU, the comparative methods are reproduced. The batch size is 16, and the learning policy is poly decay with an initial learning rate of 0.0001 and the momentum of 0.9. Several data augmentations are deployed, including 90°, 180°, and 270° rotation, horizontally and vertically flip. Commonly, the SGD optimizer and the cross-entropy loss are used in this study.

TABLE I

RESULTS OF THE ISPRS POTSDAM DATASET. CATEGORYWISE F_1 SCORE, MEAN F_1 (OVER ALL CLASSES), AND MIOU ARE LISTED, WHERE BOLD INDICATES THE BEST

Methods	Impervious surfaces	Building	Low vegetation	Tree	Car	mean F_1	mIoU	OA
FCN-8s [17]	81.46	85.17	75.22	81.15	75.08	79.62	65.93	79.15
U-Net [18]	83.61	87.30	79.41	83.84	81.38	83.11	69.11	82.23
DeepLabV3+ [2]	89.15	91.07	83.02	86.69	82.58	86.50	71.35	83.58
LANet [8]	93.05	97.19	87.30	88.04	91.19	91.35	80.22	89.31
HCANet [7]	**94.11**	96.82	85.96	88.17	92.25	91.46	81.01	89.37
HMANet [9]	93.85	96.56	87.65	**89.12**	92.34	91.90	82.15	89.93
FCN+SEB [4]	90.47	96.57	86.21	87.51	81.07	88.37	72.29	86.79
FCN+CBAM [15]	91.37	96.49	86.00	87.40	83.22	88.90	73.33	86.62
FCN+DAB [16]	91.55	96.54	86.17	87.42	84.68	89.27	75.91	87.11
FCN+NLB [5]	91.37	96.49	86.00	87.40	83.22	88.90	73.33	88.01
HAENet (ours)	93.91	**97.41**	**88.11**	**89.12**	**95.29**	**92.77**	**84.28**	**90.12**

The ResNet50 with eight times downsampling is adopted as the backbone. Besides, the max epoch is set to 500. We produce the hyperbolic embeddings referred to geoopt.[1] The feature embeddings are first transformed from Euclidean space to hyperbolic space and then mapped onto the Poincaré model for distance (similarity) calculation. The procedure is position-related; thus, the pseudo-hyperbolic space and its related parameters are unnecessary to be trained and optimized.

All comparative methods are trained from scratch without bells and whistles. As the first group of experiments, we compare proposed HAENet to state-of-the-art (SOTA) methods, including LANet, HCANet, and HMANet. Specifically, several fundamental networks are compared, such as FCN-8s, U-Net, and DeepLab V3+. Second, we compared several attention models based on FCN, including squeeze and excitation block (SEB) in SENet [4] (termed FCN + SEB), convolution block attention module (CBAM) [15] (termed FCN + CBAM), and dual attention block (DAB) in DANet [16] (termed FCN + DAB). Specifically, FCN + nonlocal block (NLB) in [5]) is evaluated as the ablative models. This model is identical to removing hyperbolic-related branches of SHAM.

D. Results of Potsdam Dataset

As presented in Table I, the categorywise F_1 score, mean F_1, and mIoU are collected on test set. In general, our HAENet outperforms to others on the Potsdam benchmark. The mean F_1 score and mIoU of HAENet are the best compared with the recent-proposed SOTA methods, such as LANet, HCANet, and HAMNet. An increase of 1%/2% of mean F_1 score/mIoU is obtained compared with HMANet. Some typical baselines are susceptible to imbalanced distribution, intraclass variations, and interclass similarities of RSIs, resulting in low accuracy, below 87%/72% of mean F_1 score/mIoU. Although RSI-targeted networks have made extensive efforts, the high-dimensional non-Euclidean properties are ignored by them. As for buildings suffering from occlusion, hyperbolic representations can stretch this part and inject it into the gyrovector for similarity measurement. HAENet reaches a peak of over 97% for the classification F_1 scores of buildings. At the same time, the hybrid similarity aggregation allows for more distinguishable representations of cars, rising by about 3% compared with the second-order network. Two random samples of Potsdam test set are predicted by HAENet in Fig. 4, where the vast majority of pixels are correctly classified.

[1]https://github.com/geoopt/geoopt

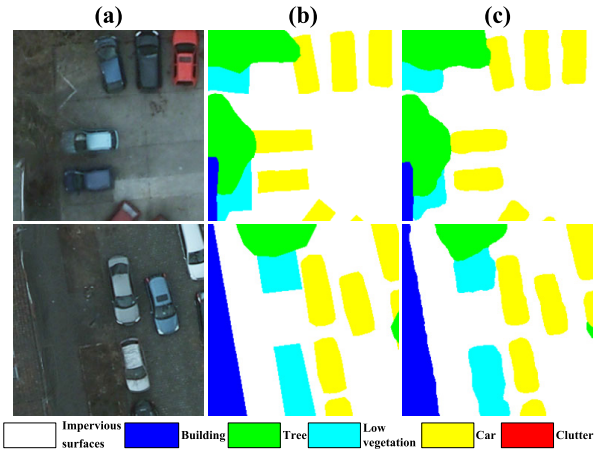

Fig. 4. Visualization of random samples. (a) RGB image. (b) Ground truth. (c) HAENet prediction.

E. Results of DeepGlobe Dataset

The DeepGlobe benchmark has a lower spatial resolution and covers a broader range than aerial images, where fine-grained visual features are difficult to be learned. As shown in Table II, all methods have experienced degradation. However, the proposed HAENet reaches the highest accuracy, with 82.93%/67.78% of mean F_1 score/mIoU. Specifically, HAENet leads in all categories except for rangeland, of which a 0.02% $F1$ score is dropped than HMANet. More than 95% of the F_1 score for water areas is calculated. Concerning imaging conditions, satellite RSIs are orthographic and insensitive to light. Less than a 1% increase of mean F_1 score is observed with the suboptimal method. Two random samples of DeepGlobe test set are predicted in Fig. 5. In summary, satellite RSIs have an indistinctive hyperbolic property, though a slight improvement is reached than other SOTA methods.

F. Ablation Study

With the same setup and network baseline of FCN-8s, we embedded SEB, CBAM, and DAB at the end of the encoder. The results are listed in Tables I and II. Overall, the proposed SHAM enables the best refinement of encoded

TABLE II

RESULTS OF THE DEEPGLOBE DATASET. CATEGORYWISE F_1 SCORE, MEAN F_1 (OVER ALL CLASSES), AND MIOU ARE LISTED, WHERE BOLD INDICATES THE BEST

Methods	Urban land	Agriculture land	Rangeland	Forest land	Water	Barren land	Unknown	mean F_1	mIoU	OA
FCN-8s [17]	70.43	81.37	69.26	67.95	83.73	61.28	58.21	70.32	55.31	73.13
U-Net [18]	76.73	85.87	75.46	74.04	85.37	59.37	58.21	73.58	59.44	75.15
DeepLabV3+ [2]	77.42	86.40	77.81	74.67	87.36	65.45	61.26	75.77	59.59	78.16
LANet [8]	81.15	90.01	83.33	78.00	90.11	71.12	66.02	79.96	62.81	84.38
HCANet [7]	82.12	91.19	84.42	78.89	92.01	72.23	66.77	81.09	63.78	85.02
HMANet [9]	86.50	92.88	**85.51**	77.99	93.07	72.25	66.92	82.16	64.62	85.19
FCN+SEB [4]	78.82	87.19	79.97	76.63	89.12	65.08	61.45	76.89	60.48	81.14
FCN+CBAM [15]	80.08	88.23	80.95	75.52	87.91	68.67	65.56	78.13	61.45	81.92
FCN+DAB [16]	79.68	87.20	79.22	75.98	88.35	67.18	63.33	77.28	60.79	80.13
FCN+NLB [5]	80.19	89.01	81.38	75.59	89.02	69.23	64.78	78.46	61.71	82.79
HAENet (ours)	**86.69**	**94.11**	84.49	**80.92**	**95.02**	**72.29**	**66.98**	**82.93**	**67.78**	**86.92**

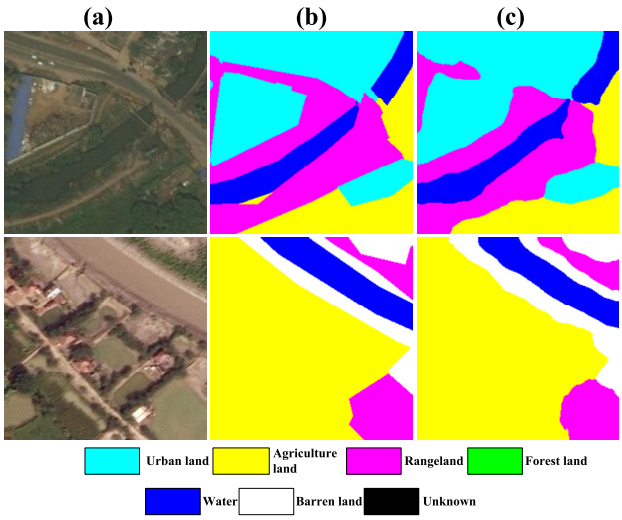

Fig. 5. Visualization of random samples. (a) RGB image. (b) Ground truth. (c) HAENet prediction.

representations. Compared with SEB, the mIoU on the Potsdam test set rises from 72.29% to 84.28%. When testing DeepGlobe, the increase slightly drops to about 7% of mIoU. CBAM and DAB have similar effects on two datasets with the cascaded and parallel post-fusion manners of two attention modules. As described in III-C, FCN + NLB is the ablative model by removing hyperbolic-related branches. Numerically, NLB refines the learned feature maps with respect to position. However, the latent non-Euclidean similarity is not introduced. With the fusion of hetero-spaces' attention maps, massive invisible cues are exploited to accurately measure the similarity of different objects, producing more fidelity similarity by incorporating hyperbolic geometry.

IV. CONCLUSION

This letter proposes a novel SHAM, which involves the latent non-Euclidean visual properties by attentively fusing position-associated attention maps in Euclidean and hyperboloid spaces, respectively. The experiments conducted on the ISPRS Potsdam and DeepGlobe benchmarks validate its efficacy and superiority to several methods. Moreover, the ablation study examined the effects of SHAM. This study opens a new direction for the interpretation of RSIs in a non-Euclidean view.

REFERENCES

[1] X. Li et al., "Dual attention deep fusion semantic segmentation networks of large-scale satellite remote-sensing images," *Int. J. Remote Sens.*, vol. 42, no. 9, pp. 3583–3610, May 2021.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[3] G. Brauwers and F. Frasincar, "A general survey on attention mechanisms in deep learning," *IEEE Trans. Knowl. Data Eng.*, early access, Nov. 9, 2021, doi: 10.1109/TKDE.2021.3126456.

[4] H. Jie, S. Li, S. Gang, and S. Albanie, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Apr. 2017.

[5] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[6] L. Ding et al., "Looking outside the window: Wide-context transformer for the semantic segmentation of high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[7] X. Li, F. Xu, R. Xia, X. Lyu, H. Gao, and Y. Tong, "Hybridizing cross-level contextual and attentive representations for remote sensing imagery semantic segmentation," *Remote Sens.*, vol. 13, no. 15, p. 2986, Jul. 2021.

[8] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2021.

[9] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.

[10] W. Peng, T. Varanka, A. Mostafa, H. Shi, and G. Zhao, "Hyperbolic deep neural networks: A survey," 2021, *arXiv:2101.04562*.

[11] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond Euclidean data," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, Jul. 2017.

[12] O. Ganea, G. Bécigneul, and T. Hofmann, "Hyperbolic neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[13] S. Liu, J. Chen, L. Pan, C.-W. Ngo, T.-S. Chua, and Y.-G. Jiang, "Hyperbolic visual embedding learning for zero-shot recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9270–9278.

[14] B. Chen, W. Peng, X. Cao, and J. Rning, "Hyperbolic uncertainty aware semantic segmentation," 2022.

[15] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[16] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.

[17] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, Oct. 2015, pp. 234–241.