*Neuro*Explainer: Fine-Grained Attention Decoding to Uncover Cortical Development Patterns of Preterm Infants

Chenyu Xue[1], Fan Wang[2(✉)], Yuanzhuo Zhu[2], Hui Li[3], Deyu Meng[1], Dinggang Shen[4(✉)], and Chunfeng Lian[1(✉)]

[1] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China
chunfeng.lian@xjtu.edu.cn

[2] Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, China
fan.wang@xjtu.edu.cn

[3] Department of Neonatology, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China

[4] School of Biomedical Engineering, ShanghaiTech University, Shanghai, China
dgshen@shanghaitech.edu.cn

Abstract. In addition to model accuracy, current neuroimaging studies require more explainable model outputs to relate brain development, degeneration, or disorders to uncover atypical local alterations. For this purpose, existing approaches typically explicate network outputs in a post-hoc fashion. However, for neuroimaging data with high dimensional and redundant information, end-to-end learning of explanation factors can inversely assure fine-grained explainability while boosting model accuracy. Meanwhile, most methods only deal with gridded data and do not support brain cortical surface-based analysis. In this paper, we propose an *explainable geometric deep network*, the *Neuro*Explainer, with applications to uncover altered infant cortical development patterns associated with preterm birth. Given fundamental cortical attributes as network input, our *Neuro*Explainer adopts a hierarchical attention-decoding framework to learn fine-grained attention and respective discriminative representations in a spherical space to accurately recognize preterm infants from term-born infants at term-equivalent age. *Neuro*Explainer learns the hierarchical attention-decoding modules under subject-level weak supervision coupled with targeted regularizers deduced from domain knowledge regarding brain development. These prior-guided constraints implicitly maximize the explainability metrics (i.e., fidelity, sparsity, and stability) in network training, driving the learned network to output detailed explanations and accurate classifications. Experimental results on the public dHCP benchmark suggest that *Neuro*Explainer led to quantitatively reliable explanation results that are qualitatively consistent with representative neuroimaging studies. The source code will be released on https://github.com/ladderlab-xjtu/NeuroExplainer.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 202–211, 2023.
https://doi.org/10.1007/978-3-031-43895-0_19

1 Introduction

One important task for the neuroscience community is to study atypical alterations in cortices associated with brain development, degeneration, or disorders. For this aim, recent approaches, namely interpretable and explainable deep learning, rely on the training of diagnostic or predictive deep learning models [6,12] with interpretable computations and explainable results. For the aspect of preterm birth, the classification task to differentiate between preterm and term-born infants can help distinguish fine-grained differences on brain cortical surfaces, providing valuable factors for better understanding featured infantile brain development patterns related to different factors.

Although explainable deep learning methods are being actively studied in the machine learning community, they have two challenges when applying to neuroimaging data. First, existing methods typically adopt post-hoc techniques to explain a deep network [13], which is first trained for a specific classification task, and then the underlying (sparse) correlations between its input and output are analyzed offline, e.g., by backpropagating prediction gradients to the shallow layers [8]. Notably, such post-hoc approaches are established upon a common assumption that reliable explanations are the results caused by accurate predictions. This assumption could work in general applications that have large-scale training data, while cannot always hold for neuroimaging and neuroscience research, where available data are typically small-sized and much more complex (e.g., high-resolution cortical surfaces containing noisy, highly redundant, and task-irrelevant information). Second, most of these methods works on gridded data (e.g., images) [2], and does not handle 3D meshes (e.g., brain cortical surfaces) [13]. For these type of data, advanced geometric deep learning methods or mapping original meshes onto a spherical surface [14] suggested promising accuracies in multiple tasks (e.g., parcellation [14], registration [9], and longitudinal prediction [4]), yet the learned models typically lack explainability.

This paper presents an *explainable geometric deep network*, called *Neuro*Explainer, with applications to uncover altered infant cortical development patterns associated with preterm birth. *Neuro*Explainer adopts high-resolution cortical attributes as the input to develop a hierarchical attention-decoding architecture working in the sperical space. Distinct to existing post-hoc methods, the *Neuro*Explainer is constructed as an end-to-end framework, where fine-grained explanation factors can be identified in a fully learnable fashion. Our network take advantage of the explainability to boost classification for the high-dimensional neuroimaging data. Specifically, in the framework of weakly supervised discriminative localization, our *Neuro*Explainer is trained by minimizing general classification losses coupled with a set of constraints designed according to prior knowledge regarding brain development. These targeted regularizers drive the network to implicitly optimize the explainability metrics from multiple aspects (i.e., fidelity, sparsity, and stability), thus capturing fine-grained explanation factors to explicitly improve classification accuracies. Experimental results on the public dHCP benchmark suggest that our *Neuro*Explainer led to quantitatively reliable explanation results that are qualitatively consistent with

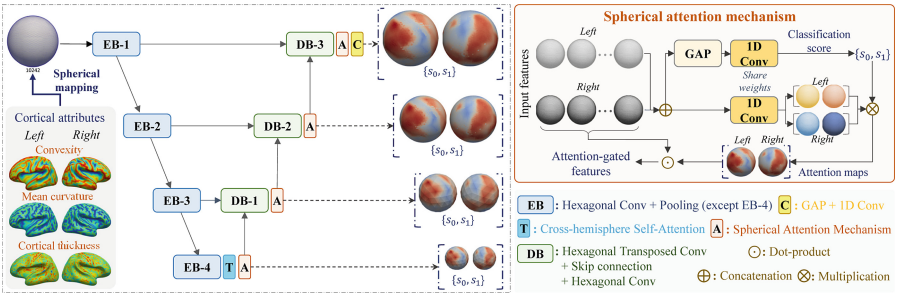

Fig. 1. The schematic diagram of our *Neuro*Explainer architecture and Spherical attention mechanism. Our *Neuro*Explainer learns to capture fine-grained by Spherical attention mechanism explanation factors to boost discriminative representation extraction.

representative neuroimaging studies, implying that it could be a practically useful AI tool for other related cortical surface-based neuroimaging studies.

2 Method

As the schematic diagram shown in Fig. 1, our *Neuro*Explainer works on the high-resolution spherical surfaces of both brain hemispheres (each with $10,242$ vertices). The inputs are fundamental vertex-wise cortical attributes, i.e., thickness, mean curvature, and convexity. The architecture has two main parts, including an encoding branch to produce initial task-related attentions on down-sampled hemispheric surfaces, and a set of attention decoding blocks to hierarchically propagate such vertex-wise attentions onto higher-resolution spheres, finally capturing fine-grained explanation factors on the input high-resolution surfaces to boost the prediction task.

2.1 Spherical Attention Encoding

The starting components of the encoding branch are four spherical convolution blocks (i.e., EB-1 to EB-4 in Fig. 1), with the learnable parameters shared across two hemispheric surfaces. Each EB adopts 1-ring hexagonal convolution [14] followed by batch normalization (BN) and ReLU activation to extract vertex-wise representations, which are then downsampled by hexagonal max pooling [14] (except in EB-4) to serve as the input of the subsequent layer. Based on the outputs from EB, we propose a learnable *spherical attention mechanism* to conduct weakly-supervised discriminative localization.

Specifically, let $\mathbf{F}^l$ and $\mathbf{F}^r \in \mathcal{R}^{162\times M_0}$ be the vertex-wise representations (produced by EB-4) for the left and right hemispheres, respectively. We first concatenate them as a $324 \times M_0$ matrix, on which a self-attention operation [11] is applied to capturing cross-hemisphere long-range dependencies to refine the vertex-wise representations from both hemispheric surfaces, resulting in a

unified feature matrix denoted as $\mathbf{F_0} = [\hat{\mathbf{F}}^l; \hat{\mathbf{F}}^r] \in \mathcal{R}^{324 \times M_0}$. As shown in Fig. 1, $\mathbf{F_0}$ is further global average pooled (GAP) across all vertices to be a holistic feature vector $f_0 \in \mathcal{R}^{1 \times M_0}$ representing the whole cerebral cortex. Both $\mathbf{F_0}$ and f_0 are then *mapped by a same vertex-wise 1D convolution* (i.e., $\mathbf{W}_0 \in \mathcal{R}^{M_0 \times 2}$, without bias) into the categorical space, denoted as $\mathbf{A}_0 = [\mathbf{A}_0^l; \mathbf{A}_0^r] \in \mathcal{R}^{324 \times 2}$ and $\mathbf{s}_0$, respectively. *Notably*, $\mathbf{s}_o$ is supervised by the one-hot code of subject's categorical label, by which $\mathbf{A}_0^l$ and $\mathbf{A}_0^r$ highlight discriminative vertices on the (down-sampled) left and right surfaces, respectively, considering that

$$\mathbf{s}_0[i] \propto \left(\mathbf{1}^T \mathbf{F_0}\right) \mathbf{W}_0[:, i] = \mathbf{1}^T \left([\hat{\mathbf{F}}^l; \hat{\mathbf{F}}^r] \mathbf{W}_0[:, i]\right) = \mathbf{1}^T \mathbf{A}_0^l[:, i] + \mathbf{1}^T \mathbf{A}_0^r[:, i], \quad (1)$$

where $\mathbf{s}_o[i]$ ($i = 0$ or 1) in our study denote the prediction scores of preterm and fullterm, respectively, and $\mathbf{1}$ is an unit vector having the same row size with the subsequent matrix. Finally, we define the hemispheric attentions as $\bar{\mathbf{A}}_0^l = \sum_{i=0}^{1} \mathbf{s}_0[i] \mathbf{A}_0^l[:, i]$ and $\bar{\mathbf{A}}_0^r = \sum_{i=0}^{1} \mathbf{s}_0[i] \mathbf{A}_0^r[:, i] \in \mathcal{R}^{324 \times 1}$, respectively, with values spatially varying and depending on the relevance to subject's category.

2.2 Hierarchically Spherical Attention Decoding

The explanation factors captured by the encoding branch are relatively coarse, as the receptive field of a cell on the downsampled surfaces (with 162 vertices after three pooling operations) is no smaller than a hexagonal region of 343 cells on the input surfaces (with 10,242 vertices). To tackle this challenge, we design a spherical attention decoding strategy to hierarchically propagate coarse attentions (from lower-resolution spheres) onto higher-resolution spheres, based on which fine-grained attentions are finally produced to improve classification.

Specifically, *Neuro*Explainer contains three consecutive decoding blocks (i.e., DB-1 to DB-3 in Fig. 1). Each DB adopts both the *attention-gated* discriminative representations from the preceding DB (except DB-1 that uses EB-4 outputs) and the local-detailed representations from the symmetric EB (at the same resolution) as the input. Let the attention-gated representations from the preceding DB be $\mathbf{F}_G^l = \left(\bar{\mathbf{A}}_{in}^l \mathbf{1}_{1 \times M_{in}}\right) \odot \hat{\mathbf{F}}_{in}^l$ and $\mathbf{F}_G^r = \left(\bar{\mathbf{A}}_{in}^r \mathbf{1}_{1 \times M_{in}}\right) \odot \hat{\mathbf{F}}_{in}^r$, respectively, where each row of $\hat{\mathbf{F}}_{in}$ has M_{in} channels, and $\odot$ denotes element-wise dot product. We first upsample $\mathbf{F}_G^l$ and $\mathbf{F}_G^r$ to the spatial resolution of the current DB, by using hexagonal transposed convolutions [14] with learnable weights shared across hemispheres. Then, the upsampled discriminative representations from each hemisphere (say $\tilde{\mathbf{F}}_G^l$ and $\tilde{\mathbf{F}}_G^r$) are channel-wisely concatenated with the local representations from the corresponding EB (say $\mathbf{F}_E^l$ and $\mathbf{F}_E^r$), followed by an 1-ring convolution to produce a unified feature matrix, such as

$$\mathbf{F}_D = [\mathcal{C}_\theta(\tilde{\mathbf{F}}_G^l \oplus \mathbf{F}_E^l); \mathcal{C}_\theta(\tilde{\mathbf{F}}_G^r \oplus \mathbf{F}_E^r)], \quad (2)$$

where $\mathcal{C}_\theta(\cdot)$ denotes 1-ring conv parameterized by θ, and $\oplus$ stands for channel concatenation. In terms of $\mathbf{F}_D$, the attention mechanism described in (1) is further applied to producing refined spherical attentions and classification scores.

Finally, as shown in Fig. 1, based on the fine-grained attentions over the input surfaces (each with 10,242 vertices), we use GAP to aggregate the attention-gated representations and apply an 1D conv to output the classification score.

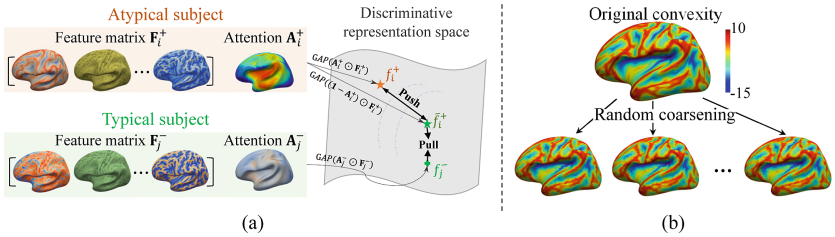

Fig. 2. Brief illustrations of (a) the explanation fidelity-aware contrastive learning strategy, and (b) explanation stability-aware data augmentation strategy.

2.3 Domain Knowledge-Guided Explanation Enhancement

To perform task-oriented learning of explanation factors, we design a set of targeted regularization strategies by considering fundamental domain knowledge regarding infant brain development. Specifically, according to existing studies, we assume that human brains in infancy have generally consistent developments, while the structural/functional discrepancies between different groups (e.g., preterm and term-born) are typically rationalized [1,10]. Accordingly, we require the preterm-altered cortical development patterns captured by our *Neuro*Explainer to be discriminative, spatially sparse, and robust, which suggests the design of the following constraints that concurrently optimize fidelity, sparsity, and stability metrics [13] in deploying an explainable deep network.

Explanation Fidelity-Aware Contrastive Learning. Given the spherical attention block at a specific resolution, we have $\mathbf{A}_i^+$ and $\mathbf{A}_j^- \in \mathcal{R}^{V\times 1}$ as the output attentions for a positive and negative subjects (i.e., preterm and fullterm infants in our study), respectively, and $\mathbf{F}_i^+$ and $\mathbf{F}_j^- \in \mathcal{R}^{V\times M}$ are the corresponding representation matrices. Based on the prior knowledge regarding infant brain development, it is reasonable to assume that $\mathbf{A}_i^+$ highlights atypically-developed cortical regions caused by preterm birth. *In contrast*, the remaining part of the cerebral cortex of a preterm infant (corresponding to $1-\mathbf{A}_i^+$) still growths normally, i.e., looking globally similar to the cortex of a term-born infant.

Accordingly, as the illustration shown in Fig. 2(a), we design a fidelity-aware contrastive penalty to regularize the learning of the attention maps and associated representations to improve their discriminative power. Let $\boldsymbol{f}_i^+ = \mathbf{1}^T\left(\mathbf{A}_i^+\mathbf{1}_{1\times M}\odot\mathbf{F}_i^+\right)$ and $\bar{\boldsymbol{f}}_i^+ = \mathbf{1}^T\left(\{1-\mathbf{A}_i^+\}\mathbf{1}_{1\times M}\odot\mathbf{F}_i^+\right)$ be the holistic feature vector and its inverse for the ith (positive) sample, respectively. Similarly, $\boldsymbol{f}_j^- = \mathbf{1}^T\left(\mathbf{A}_j^-\mathbf{1}_{1\times M}\odot\mathbf{F}_j^-\right)$ denotes the holistic feature vector for the compared jth (negative) sample. By *pushing* $\boldsymbol{f}_i^+$ *away from both* $\bar{\boldsymbol{f}}_i^+$ *and* $\boldsymbol{f}_j^-$, *while pulling* $\bar{\boldsymbol{f}}_i^+$ *close to* $\boldsymbol{f}_j^-$, we define the respective loss as

$$\mathcal{L}_{contra} = \sum_{i\neq j}^{N} ||\bar{\boldsymbol{f}}_i^+ - \boldsymbol{f}_j^-|| + \max(m - ||\bar{\boldsymbol{f}}_i^+ - \boldsymbol{f}_i^+||, 0) + \max(m - ||\boldsymbol{f}_j^- - \boldsymbol{f}_i^+||, 0), \quad (3)$$

where i and j indicate any a pair of positive and negative cases from totally N training samples, and m is a margin setting as 1 in our implementation.

Explanation Sparsity-Aware Regularization. According to the specified prior knowledge regarding infant brain development, the attention maps produced by our *Neuro*Explainer should have two featured properties in terms of sparsity. That is, the attention map for a preterm infant (e.g., $\mathbf{A}_i^+$) should be sparse, considering that altered cortical developments are assumed to be localized. In contrast, the attention map for a healthy term-born infant (e.g., $\mathbf{A}_j^-$) should not be spatially informative, as all brain regions growth typically without abnormality. To this end, we design a straightforward entropy-based regularization to enhance results' explainability, such as

$$\mathcal{L}_{entropy} = \sum_{i \neq j}^{N} \mathbf{1}^T \left\{ \mathbf{A}_i^+ \odot \log(\mathbf{A}_i^+) - \mathbf{A}_j^- \odot \log(\mathbf{A}_j^-) \right\}, \tag{4}$$

where i and j indicate a positive and a negative cases from totally N training samples, respectively, and $\mathbf{1}$ is an unit vector to sum up the values of all vertices.

Explanation Stability-Aware Regularization. We enhance the explanation stability of our *Neuro*Explainer from two aspects. *First*, we require the spherical attention mechanisms to *robustly* decode from complex cortical-surface data fine-grained explanation factors to produce accurate predictions. To this end, we randomize the surface coarsening step by quantifying a vertex's cortical attributes (on the downsampled surface) as the average of a random subset of the vertices from the respective hexagonal region of the highest-resolution surface, such as the examples summarized in Fig. 2(b). Considering that the network is trained to produce consistently accurate predictions for all these variants with perturbations, it inversely enhances the stability of learned explanation factors.

Second, as described in Sect. 2.2, we design a cross-scale consistency regularization to refine the decoding branch. Specifically, let $\mathbf{A}_i^l$ and $\mathbf{A}_i^h$ be the spherical attentions from two different DB blocks. We simply minimize

$$\mathcal{L}_{consistent} = \sum_{i=1}^{N} \left(\mathbf{A}_i^l - \mathbf{A}_i^h \right)^2, \tag{5}$$

which encourages spherical attentions at different resolutions to be consistent.

Implementation Details. In our implementation, the feature representations produced by EB-1 to EB-4 in Fig. 1 have 32, 64, 128, and 256 channels, respectively. Correspondingly, DB-1 to DB-3, and the final classification layer have 256, 128, 64, and 32 channels, respectively. The network was trained end-to-end by minimizing the cross-entropy classification losses defined at three different spatial resolutions (overall denoted as $\mathcal{L}_{CE}$), coupled with the regularization terms introduced in Sec. 2.3, such as

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{contrast} + \lambda_2 \mathcal{L}_{entropy} + \lambda_3 \mathcal{L}_{consistent}, \tag{6}$$

where the tuning parameters were empirically set as $\lambda_1 = 0.2$, $\lambda_3 = 0.5$, and $\lambda_3 = 0.1$. The network parameters were updated by using Adam optimizer for 500 epochs, with the initial learning rate setting as 0.001 and bath size as 20.

3 Experiments

Dataset and Experimental Setup. We conducted experiments on the dHCP benchmark [5]. The structural MRIs of 700 infants scanned at term-equivalent ages (35–44 weeks postmenstrual age) were studied, including 143 preterm and 557 term-born infants. These subjects were randomly split as a training set of 500 infants (89 preterm and 411 fullterm), and a test set of the remaining 200 infants (54 preterm and 146 fullterm), where test and training sets were from different subjects. Using the data-augmentation strategy described in Sect. 2.3, the training set was augmented to have roughly 1,250 subjects from each category for balanced network training. The input spherical surfaces contain 10,242 vertices, and each of them has three morphological attributes, i.e., cortical thickness, mean curvature, and convexity.

Table 1. Classification results obtained by the competing geometric deep networks and different variants of our *Neuro*Explainer.

Competing Mehtods	ACC	AUC	SEN	SPE
SphericalCNN [14]	0.93	0.92	0.76	**0.98**
SphericalMoNet [9]	0.85	0.93	0.65	0.92
SubdivNet [3]	0.79	0.67	0.74	0.80
*Neuro*Explainer (**ours**)	**0.95**	**0.97**	**0.94**	0.95
w/o $\mathcal{L}_{contrast}$ (3)	0.88	0.89	0.80	0.91
w/o $\mathcal{L}_{entropy}$ (4)	0.91	0.96	0.74	0.97
w/o $\mathcal{L}_{consistent}$ (5)	0.89	0.95	0.89	0.88

Table 2. Quantitative explanation results obtained by the competing post-hoc approaches and our end-to-end *Neuro*Explainer.

Competing Methods		Fidelity	Sparsity	Stability
CAM [15] +	SphericalCNN	0.24	0.91	0.77
	SphericalMoNet	0.55	0.93	0.58
	SubdivNet	0.06	0.97	0.53
Grad-CAM [7] +	SphericalCNN	0.22	0.99	0.77
	SphericalMoNet	0.42	0.98	0.58
	SubdivNet	0.16	0.96	0.53
*Neuro*Explainer (**ours**)		**0.56**	0.73	**0.96**

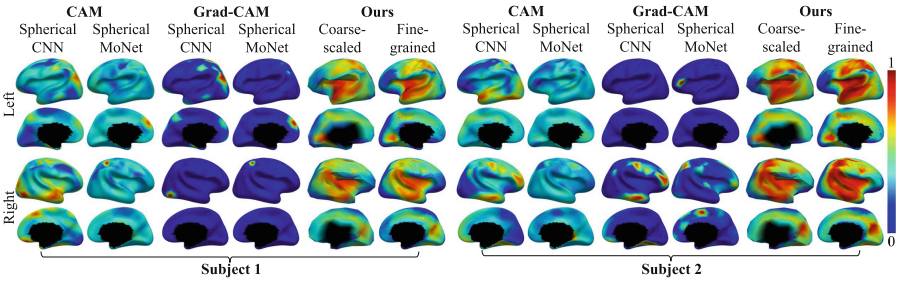

Fig. 3. Typical examples of the explanation factors captured by different methods. Higher values indicate larger links to preterm birth.

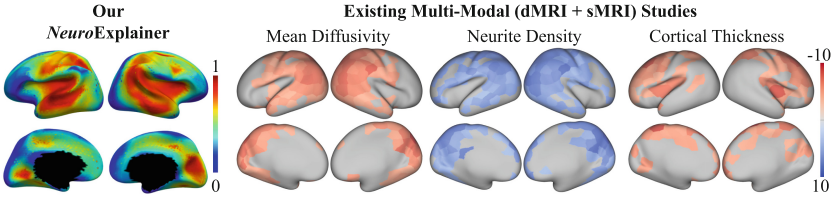

Fig. 4. Comparison of the *individualized* preterm-altered developments uncovered by *Neuro*Explainer with the *group-wise* multi-modal studies [1].

For classification, our *Neuro*Explainer was compared with three representative geometric networks, including a spherical network based on 1-ring convolution (**SphericalCNN**) [14], a MoNet reimplementation working on spherical surfaces (**SphericalMoNet**) [9], and **SubdivNet** [3] working on original meshes. The classification performance was quantified in terms of accuracy (**ACC**), area under the ROC curve (**AUC**), sensitivity (**SEN**), and specificity (**SPE**).

On the other hand, the explanation performance of our *Neuro*Explainer was compared with two representative feature-based explanation approaches, i.e., **CAM** [15] and **Grad-CAM** [7]. The explanation performance was quantitatively evaluated in terms of three metrics [13], i.e., **Fidelity**, **Sparsity**, and **Stability**. Please refer to [13] for more details regarding these metrics.

Classification Results. The classification results obtained by different competing methods are summarized in Table 1, from which we can have at least *two observations*. **1)** Our *Neuro*Explainer consistently led to better classification accuracies in terms of all metrics (especially SEN and AUC), suggesting that it can reliably identify featured development patterns associated with preterm birth to make accurate predictions in such an imbalanced learning task. **2)** These results imply that our idea to capture fine-grained explanation factors in an end-to-end fashion to boost discriminative representation extraction is beneficial for deploying an accurate classification model. **3)** To check the efficacy of the prior-

induced regularization strategies, we orderly removed them from the loss function (6) to quantify the respective influences. From Table 1, we can see that all the three regularizations demonstrated *significant but different* improvements on classification, implying their complementary roles in boosting explainable representation learning.

Explanation Results. The quantitative explanation results are summarized in Table 2. Notably, the three metrics should be analyzed concurrently in evaluating a network's explainability [13], as the isolated quantification of a single metric could be biased. From Table 2, we can observe that our *Neuro*Explainer led to significantly better Fidelity and Stability, under reasonable Sparsity, suggesting that it can robustly identify rationalized preterm-altered cortical patterns from high-dimensional inputs for preterm infant recognition. Also, we visually compared the attention maps produced by different competing methods, with two typical examples presented in Fig. 3. From Fig. 3, we can see that, compared with post-hoc explanation methods, our end-to-end *Neuro*Explainer stably produced more reasonable attentions. For example, our *Neuro*Explainer led to group-wisely more consistent explanations across subjects. Also, it produced more consistent results across hemispheres, without using any related training constraints.

Finally, we compared the *individualized* preterm-altered cortical development patterns uncovered by our *Neuro*Explainer with representative *group-wise* multimodal (dMRI and sMRI) quantitative analyses presented in [1]. As shown in Fig. 4, we can see that our observations in this paper are consistent with [1]. The discriminative cortical regions captured by our *Neuro*Explainer (using solely morphological features) are largely overlapped with the group-wise significantly different regions identified by [1] in terms of the mean diffusivity, neurite density, and cortical thickness, respectively. For example, they both highlighted some specific regions in the inferior parietal, medial occipital, and superior temporal lobe, and posterior insula, which is worth deeper evaluations in the future.

4 Conclusion

In the paper, we have proposed an geometric deep network, i.e., *Neuro*Explainer, to learn fine-grained explanation factors from complex cortical-surface data to boost discriminative representation extraction and accurate classification model construction. On the benchmark dHCP database, our *Neuro*Explainer achieved better performance than existing post-hoc approaches in terms of both explainability and prediction accuracy, in uncovering preterm-altered infant cortical development patterns. The proposed method could be a promising AI tool applied to other similar cortical surface-based neuroimage and neuroscience studies.

Funding. This work was supported in part by NSFC Grants (Nos. 62101431 & 62101430), and STI 2030-Major Projects (No. 2022ZD0209000).

References

1. Dimitrova, R., et al.: Preterm birth alters the development of cortical microstructure and morphology at term-equivalent age. Neuroimage **243**, 118488 (2021)
2. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. Commun. ACM **63**(1), 68–77 (2019)
3. Hu, S.M., et al.: Subdivision-based mesh convolution networks. ACM Trans. Graph. (TOG) **41**(3), 1–16 (2022)
4. Liu, P., Wu, Z., Li, G., Yap, P.-T., Shen, D.: Deep modeling of growth trajectories for longitudinal prediction of missing infant cortical surfaces. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 277–288. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_21
5. Makropoulos, A., et al.: The developing human connectome project: a minimal processing pipeline for neonatal cortical surface reconstruction. Neuroimage **173**, 88–112 (2018)
6. Ouyang, J., Zhao, Q., Adeli, E., Zaharchuk, G., Pohl, K.M.: Self-supervised learning of neighborhood embedding for longitudinal MRI. Med. Image Anal. **82**, 102571 (2022)
7. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 618–626 (2017)
8. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
9. Suliman, M.A., Williams, L.Z., Fawaz, A., Robinson, E.C.: A deep-discrete learning framework for spherical surface registration. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13436, pp. 119–129. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_12
10. Thompson, D.K., et al.: Tracking regional brain growth up to age 13 in children born term and very preterm. Nat. Commun. **11**(1), 1–11 (2020)
11. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
12. Yang, Z., et al.: A deep learning framework identifies dimensional representations of Alzheimers disease from brain structure. Nat. Commun. **12**(1), 1–15 (2021)
13. Yuan, H., Yu, H., Gui, S., Ji, S.: Explainability in graph neural networks: a taxonomic survey. IEEE Trans. Pattern Anal. Mach. Intell. **45**, 5782–5799 (2022)
14. Zhao, F., et al.: Spherical U-net on cortical surfaces: methods and applications. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 855–866. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_67
15. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)