
Proceedings Paper:
He, S, Lin, Z, Yang, X et al. (13 more authors) (2021) Statistical Dependency Guided Contrastive Learning for Multiple Labeling in Prenatal Ultrasound. In: Lian, C, Cao, X, Rekik, I, Xu, X and Yan, P, (eds.) Machine Learning in Medical Imaging. 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, 27 Sep - 01 Oct 2021, Strasbourg, France. Springer , pp. 190-198. ISBN 978-3-030-87588-6

https://doi.org/10.1007/978-3-030-87589-3_20

Statistical Dependency Guided Contrastive Learning for Multiple Labeling in Prenatal Ultrasound

Shuangchi He[1,2,3⋆], Zehui Lin[1,2,3⋆], Xin Yang[1,2,3], Chaoyu Chen[1,2,3], Jian Wang[1,2,3], Xue Shuang[1,2,3], Ziwei Deng[1,2,3], Qin Liu[1,2,3], Yan Cao[1,2,3], Xiduo Lu[1,2,3], Ruobing Huang[1,2,3], Nishant Ravikumar[4,5], Alejandro Frangi[1,4,5,6], Yuanji Zhang[7], Yi Xiong[7], and Dong Ni[1,2,3(✉)]

[1] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Health Science Center, Shenzhen University, China
nidong@szu.edu.cn
[2] Medical Ultrasound Image Computing (MUSIC) Lab, Shenzhen University, China
[3] Marshall Laboratory of Biomedical Engineering, Shenzhen University, China
[4] Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), University of Leeds, UK
[5] Leeds Institute of Cardiovascular and Metabolic Medicine, University of Leeds, UK
[6] Medical Imaging Research Center (MIRC), KU Leuven, Leuven, Belgium
[7] Department of Ultrasound, Luohu People's Hospital, Shenzhen, China

Abstract. Standard plane recognition plays an important role in prenatal ultrasound (US) screening. Automatically recognizing the standard plane along with the corresponding anatomical structures in US image can not only facilitate US image interpretation but also improve diagnostic efficiency. In this study, we build a novel multi-label learning (MLL) scheme to identify multiple standard planes and corresponding anatomical structures of fetus simultaneously. Our contribution is three-fold. First, we represent the class correlation by word embeddings to capture the fine-grained semantic and latent statistical concurrency. Second, we equip the MLL with a graph convolutional network to explore the inner and outer relationship among categories. Third, we propose a novel cluster relabel-based contrastive learning algorithm to encourage the divergence among ambiguous classes. Extensive validation was performed on our large in-house dataset. Our approach reports the highest accuracy as 90.25% for standard planes labeling, 85.59% for planes and structures labeling and mAP as 94.63%. The proposed MLL scheme provides a novel perspective for standard plane recognition and can be easily extended to other medical image classification tasks.

1 Introduction

Ultrasound (US) is widely used for the evaluation of fetal growth and congenital malformations in routine obstetric examinations [12]. During the scanning, US

⋆ Shuangchi He and Zehui Lin contribute equally to this work.

standard planes (SPs) that contain key anatomical structures (ASs) are selected and subsequent biometric measurements are performed [3]. For example, the abdominal circumference (AC) is measured on the transverse plane of the fetal abdomen with umbilical vein at the level of the portal sinus and stomach bubble visible (Fig. 1). The value of AC is then used to estimate the pre-birth weight of a fetus [12,3]. In clinical practice, the standard plane (SP) selection based on ASs identification is experience-dependent, cumbersome, and suffering from the inter-observer and intra-observer variability [1]. Hence, automatic recognition of SP is desired to improve the examinations.

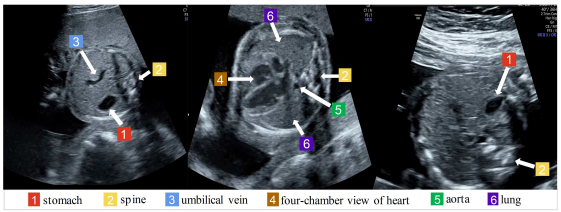

Fig. 1. Left: standard plane of fetal abdomen; Middle: standard plane of four chambers of fetal heart; Right: non-standard plane around fetal abdomen. All images are annotated with multiple anatomical structure labels.

In recent years, deep learning-based methods have witnessed significant growth in automated SP recognition. Chen et al. [3] proposed a composite neural network framework for the automatic recognition of three SPs. Burgos-Artizzu et al. [1] evaluated a large set of state-of-the-art convolutional neural networks for the classification of more than 6 maternal / fetal US planes. Cai et al. [2] presented a convolutional neural network (CNN) framework SonoEyeNet for the detection of SPs. They found that the eye movement tends to focus on the existence of ASs. These methods could distinguish the SPs from the non-standard ones directly with the plane-level labels. However, they did not explicitly incorporate the clues of key ASs, which limited the clinical interpretability and possible guidance for novice sonographers. Lin et al. [9] focused on the detection of key ASs, providing fine-grained information of SPs. However, as shown in Fig. 1, the presence of anatomical structure (AS) alone does not guarantee an accurate identification of the SP, as the SP is also defined by the global image appearance and subtle details [5]. Furthermore, the extensive annotations of each structure with bounding boxes are also labor-intensive and are difficult to obtain. Therefore, new frameworks and methods need to be devised to recognize SP and provide additional information on key ASs simultaneously.

In this paper, we build a novel multi-label learning (MLL) scheme to recognize multiple SPs and corresponding key ASs at the same time. Our contribution is three-fold. (*i*) Inspired by natural language processing techniques, the word embedding [7] is introduced to model the latent concurrency and statistical dependency among different classes, including SPs and ASs. These kinds of cues

prove to be strong guidance for MLL prediction. (*ii*) To further capture the topological structures in the label space of the word embeddings, graph convolutional network (GCN) [4] is explored to propagate information between multiple classes to capture the inner and outter relationship among ASs and SPs. (*iii*) To tackle the high intra-class variation and low inter-class variation of different SPs and ASs (Fig. 1), we further devise a cluster relabel-based contrastive learning (CRC) to align the similarity and increase discrimination across different classes. We conduct extensive experiments on a large dataset which contains 9742 US images from 920 fetuses and 39 object classes (including 10 SPs and 29 ASs). Experiments prove that, the proposed MLL method can achieve promising results in classifying multiple SPs and identifying associated key ASs.

2 Methodology

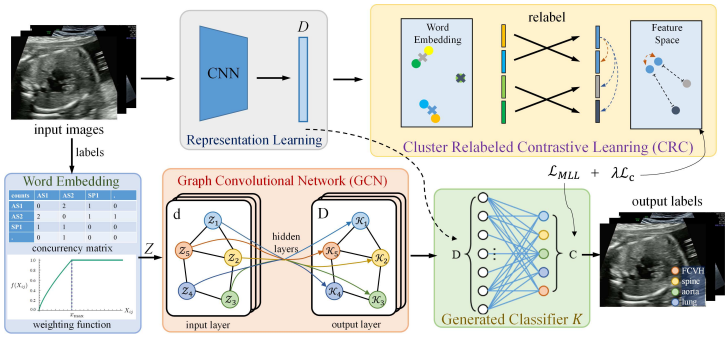

Fig. 2. Overall framework of the proposed MLL for US image recognition. The word embeddings $Z \in \mathbb{R}^{C \times d}$ are generated based on the concurrency matrix and weighting function. Stacked GCNs are learned over the graph to map these word embeddings into an inter-dependent object classifier, i.e., $\boldsymbol{K} \in \mathbb{R}^{C \times D}$. CRC is used to improve the discrimination of the classifier. The classifier is then applied to the image representation from the input image via a CNN for MLL image recognition.

Fig. 2 is the schematic view of our proposed method. We propose a MLL framework to recognize the multiple SPs and ASs simultaneously. To exploit the statistical dependency among classes, we firstly generate statistical word embeddings from label annotations. Then, we utilize GCN to model the hierarchical relationship among the classes. Further more, we propose CRC to align the high-level representation among samples of the same category. The MLL recognition output is obtained through representation learning and generated classifier.

2.1 Multi-label Learning with Word Embeddings

CNN is known for its ability in representation learning. As shown in Fig. 2, our MLL learning scheme is built upon a CNN to learn the feature of an image. In

specific, we use ResNet [6] as the backbone model. Given an input image $\boldsymbol{I}$ with a size of 448×448 pixels, we can obtain an image-level feature $\boldsymbol{x}$:

$$\boldsymbol{x} = f_{cnn}(\boldsymbol{I}, \theta_{cnn}) \in \mathbb{R}^D, \tag{1}$$

where θ_{cnn} indicates model parameters and $D = 512$.

Inspired by the natural language processing techniques which aim to model the statistical dependency among words, phrases and sentences, we try to capture the fine-grained semantic dependency that exists among the SPs and ASs in the label space following the spirit of word embedding [7]. Since it is intractable to model the relationship among the SP and AS labels in prenatal US directly using the word embeddings pre-trained on natural languages, we build a corpus based on the labels from the trainng US dataset (An image sample is considered as a sentence, and the SP category and AS labels of the sample are considered as words.), and use the GloVe [11] to train the word embeddings. According to the label-based sentences, we construct a concurrency matrix X and use it as GloVe input. X_{ij} represents the number of times class i and class j appear together on the same sample in the dataset. Then, the relationship between word embeddings and the co-occurrence matrix is formulated as:

$$w_i^T \tilde{w_j} + b_i + \tilde{b_j} = log(X_{ij}), \tag{2}$$

where $w_i \in \mathbb{R}^d$ is word embedding and $\tilde{w_j} \in \mathbb{R}^d$ is separate context word embedding which reduces overfitting. b_i and $\tilde{b_j}$ are corresponding bias terms.

We can obtain the final word embedding output $z = w_i^T + \tilde{w_j}$ by optimizing the following loss function:

$$J = \sum_{i,j=1}^{C} f(X_{ij})(w_i^T \tilde{w_j} + b_i + \tilde{b_j} - log(X_{ij})), \tag{3}$$

where C is the size of the vocabulary (i.e. our class number, 39), f is the weighting function [11] that adjusts the frequency of concurrency in the corpus. Word embeddings matrix $Z = \{z_i\}_{i=1}^C \in \mathbb{R}^{C \times d}$ hence encodes the statistical dependency and distribution relationships among different labels and can be further explored in the following sections.

2.2 GCN for Class Dependency Learning

It is important to capture the internal relationships between ASs and SPs and leverage this relationship to improve the classification performance in multi-label US image recognition. In this paper, inspired by [4], we explore the GCN to model the class dependency in prenatal US images, which is an effective and flexible way to capture the topological structures in the word embeddings label space represented by Z. Specially, GCN is built to directly map the nodes (i.e. word embeddings Z) of the graph into an inter-dependent classifier (Fig. 2). The GCN based mapping function is defined as:

$$\boldsymbol{G}^{l+1} = h(\hat{\boldsymbol{B}}\boldsymbol{G}^l\boldsymbol{W}^l), \tag{4}$$

where $\boldsymbol{G}^l \in \mathbb{R}^{C\times d}$ are feature descriptions (C denotes the number of nodes and d indicates the dimensionality of node feature) and $\hat{\boldsymbol{B}} \in \mathbb{R}^{C\times C}$ is the normalized version of correlation matrix, and $h(\cdot)$ denotes a non-linear operation. In every back-propagation, the transformation matrix $\boldsymbol{W}^l \in \mathbb{R}^{d\times d^{'}}$ will be updated.

As shown in Fig. 2, for the first layer of the stacked GCNs, the input is the word embeddings matrix Z. The output of the last GCN layer is $\boldsymbol{K} \in \mathbb{R}^{C\times D}$, which matches the dimensionality of the image representations extracted by the CNN. $\boldsymbol{K}$ contains the class dependency and hence regularizes the CNN prediction $\boldsymbol{x}$ as the final classifier. The multi-label prediction scores can be computed by applying the learned classifier to the image representation as follows:

$$\hat{\boldsymbol{y}} = \boldsymbol{K}\boldsymbol{x}, \tag{5}$$

where the ground truth labels of an image is represented as $\boldsymbol{y}$ with $y^i = \{0, 1\}$ denoting whether label i appears in the image or not. The training of the whole network uses the traditional MLL classification loss as follows:

$$\mathcal{L}_{MLL} = \sum_{c=1}^{C} y^c log(\sigma(\hat{y}^c)) + (1 - y^c) log(1 - \sigma(\hat{y}^c)). \tag{6}$$

2.3 Cluster Relabeled Contrastive Learning

Borrowing the idea of supervised contrastive learning [8], we propose to use contrastive learning (CL) to further increase the discriminative ability of learning. In CL, the samples belonging to the same class are encouraged to be similar to each other, while that of the different classes are encouraged to be different in high dimensional feature space. However, this principle can not be directly applied to our multi-label circumstance. One sample may have labels overlapped with the other samples, thus it is difficult to define the positive and negative sample pairs. On the other hand, semantically related concepts in the word embeddings space are found to be naturally close to each other [10]. Therefore, we propose to assign every sample a new single label based on the cluster of the word embeddings and perform supervised contrastive learning.

Specially, as shown in Fig. 2, we perform the k-means clustering algorithm in the word embeddings label space $Z \in \mathbb{R}^{C\times d}$. We use C as the sample size and d as the dimensionality to generate N clusters. Each sample with original multi-label $\boldsymbol{y}$ is represented as a vector $\overline{z_p}$. It is calculated through the mean value of the $\{z_i \in \mathbb{R}^d | y_i = 1, i = 1, \cdots, C\}$. The new single label $\boldsymbol{y}^*$ with $y_i^* = \{0, 1\}$, $i \in [1, N]$ is assigned to the sample according to the nearest distance among these N cluster centroids. For our multi-label task, N is set to 10.

The contrastive loss to drive the learning of relabeled samples is defined as:

$$\mathcal{L}_c = \sum_i \sum_{j,k\neq i} \alpha(1 - sim(\boldsymbol{x}_i, \boldsymbol{x}_j)) + \beta(1 + sim(\boldsymbol{x}_i, \boldsymbol{x}_k)), \tag{7}$$

where $\boldsymbol{x}$ is the image representation, i and j is the positive sample pair with same $\boldsymbol{y}^*$, i and k is the negative sample pairs with different $\boldsymbol{y}^*$. $sim(\boldsymbol{a}, \boldsymbol{b}) =$

$\frac{\boldsymbol{a}^T\boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|}$ is the cosine similarity between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. α and β are the hyperparameter to weight the similarity. Since there are fewer pairs of positive samples than negative samples, we empirically set α to 0.75 and β to 0.25 to balance the loss weights. The total loss of the proposed method is defined as the summation of MLL loss $\mathcal{L}_{MLL}$ and contrastive loss $\mathcal{L}_c$

$$\mathcal{L} = \mathcal{L}_{MLL} + \lambda\mathcal{L}_c, \tag{8}$$

where λ is the hyperparameter to weight the contrastive loss. λ is set to 0.1 based on the validation results.

3 Experimental Results

Implementation Details. Our dataset contains 9742 prenatal US images from 920 fetuses, including 10 types of SP and 29 types of AS. The gestational age ranges from 18 to 28 week. The image size was set to 448×448. An experienced sonographer provided the ground truth labels. The dataset was randomly split into 4331, 2643 and 2768 images in fetus level for training, validation and testing. There was no overlap of fetus among datasets. Adequate data augmentation were performed. The model was implemented in PyTorch with an RTX 2080Ti GPU. We used Adam optimizer (learning rate 0.001) to train GloVe for 256 epochs to obtain 512-dimensional word embeddings. SGD optimizer (learning rate 0.01) is used to train the model for 100 epochs to obtain the MLL classifier.
Quantitative and Qualitative Analysis. We evaluated the classification in terms of the average overall precision (OP), recall (OR), F1 (OF1) and the average per-class precision (CP), recall (CR), F1 (CF1). The mean average precision (mAP), Hamming loss (HL), the accuracy of the standard plane classification (SP_ACC) and the multi-label classification accuracy that exactly matches the categories of all targets on the image (MLL_ACC) were also taken into consideration. Table 1 illustrates the detailed evaluation results.

Ablation study was conducted to compare different methods, including MLL without GCN and CL (Single-MLL), MLL with contrastive learning (MLL-CL, the non-relabeled version of CRC), MLL with CRC (MLL-CRC), MLL with GCN (MLL-GCN) [4], MLL with GCN and vallina contrastive learning (MLL-GCN-CL) and the full model (MLL-GCN-CRC). We also compared with state-of-the-art methods, including CNN-RNN [13] and SRN [14]. All the above methods were pre-trained with ImageNet. The ResNet34 served as the network backbone for Single-MLL, MLL-CL, MLL-CRC, MLL-GCN, MLL-GCN-CL and MLL-GCN-CRC. We can draw the following conclusions from the Table 1:

(a) GCN significantly improves the model performance (4% in MLL_ACC) under both CL and CRC conditions (i.e., MLL-GCN-CL vs. MLL-CL and MLL-GCN-CRC vs. MLL-CRC). It is attributed to the informative class dependency extracted from the statistical word embeddings by the GCN. A similar conclusion can be deduced through the comparison between MLL-GCN and Single-MLL.

(b) Comparing the MLL-GCN, MLL-GCN-CL and MLL-GCN-CRC, we can draw the conclusion that, CL can increase the discriminative ability of our

Table 1. Quantitative evaluation of multi-label classification methods (in %).

Method	SP_ACC	MLL_ACC	mAP	HL	OP	OR	OF1	CP	CR	CF1
CNN-RNN	80.07	76.45	83.15	3.83	-	-	-	-	-	-
SRN	86.95	66.17	91.74	2.15	90.13	89.63	89.88	86.81	88.40	87.60
Single-MLL	88.26	81.04	93.75	1.64	92.00	92.80	92.40	88.84	89.61	89.22
MLL-CL	88.37	81.37	93.67	1.65	92.02	92.65	92.33	88.84	89.43	89.13
MLL-CRC	88.37	81.37	93.73	1.63	92.22	92.63	92.42	89.09	89.57	89.33
MLL-GCN	89.27	84.83	94.30	1.51	92.64	93.31	92.98	89.64	90.29	89.97
MLL-GCN-CL	90.07	85.52	94.62	1.45	92.43	94.16	93.28	89.67	91.74	90.69
MLL-GCN-CRC	**90.25**	**85.59**	**94.63**	**1.40**	**92.68**	**94.42**	**93.54**	**89.87**	**92.14**	**90.99**

method by about 0.7% in MLL_ACC. Besides, we can observe that the CRC methods consistently give better perfomances than the CL methods. The relabeled operation in CRC incorporating the fine-grained semantic in word embeddings space further boosts the similarity alignment of CL.

(c) Among all the state-of-the-art methods (CNN-RNN lacks some result due to its design), the proposed full model MLL-GCN-CRC achieves the best results regarding both the SPs classification and ASs identification. The statistical knowledge via graph manner and similarity alignment in MLL contributes to the capture of class dependency.

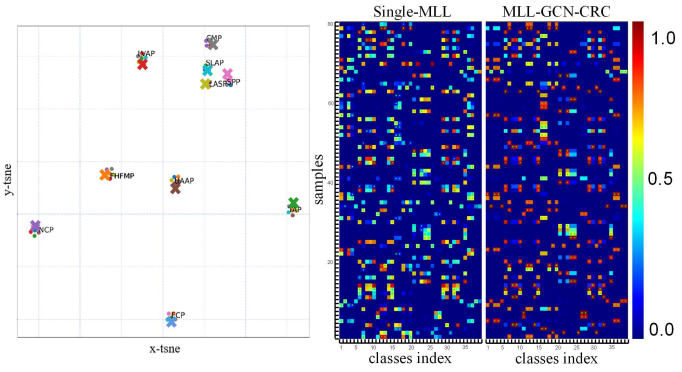

Fig. 3. Left: the t-SNE visualization of the word embeddings space of 39 classes. The dots of different colors represent categories. The cross represents the center of cluster. Right: two score matrices of Single-MLL and our proposed MLL-GCN-CRC. Each row in the matrix represents a sample class prediction. The little green points indicate the true class and the color of matrix element indicate the predicted class scores.

In Fig. 3, we can observe that the related ASs and SPs embedding clustered together naturally, which builds a more semantic-reasonable label space. On the other hand, this result supports the feasibility of our CRC. The score matrices

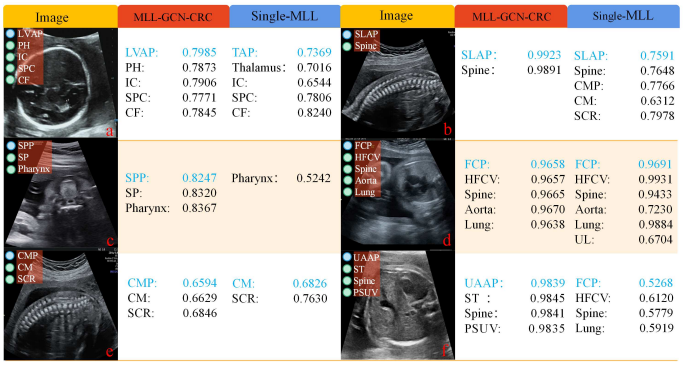

Fig. 4. Typical results of multi-label recognition on fetal US. Red box for ground truth. Blue circle for SPs and green circle for ASs.

in Fig. 3 further illustrates the MLL prediction of some examples. More sample score matrices can be found in Fig. 5 in the Appendix. It can be observed that the proposed method MLL-GCN-CRC obtains more matched cases (i.e. green point locates in the region with high score) than the Single-MLL does. This phenomenon reflects that the statistical knowledge encoded by GCN and the discriminative power enhanced by CRC are beneficial in promoting the class prediction and reducing the false positives.

Fig. 4 shows the prediction comparisons of six samples. Among the six SPs of fetal LVAP, SPP, CMP, SLAP, FCP, and UAAP (see the detailed name list of SP and AS in the Table 2 of Appendix) , our MLL-GCN-CRC obtaines high scores and correct predictions for most of the SPs and ASs (Fig. 4(a)(c)). More comparison results can be found in the Fig. 6 of Appendix. On the contrary, the Single-MLL presents mis-classifications and false positives (Fig. 4(b)(f)).

4 Conclusion

In this paper, we propose a novel multi-label learning scheme (MLL-GCN-CRC) for multiple standard planes and corresponding anatomical structures recognition in prenatal ultrasound. Following the spirit of word embedding, the statistical concurrency knowledge is explored to capture the latent class dependency between standard planes and anatomical structures. A GCN is designed to further encode the dependency among the word embeddings. By performing relabeling based on the clusters in word embeddings space, the contrastive learning boosts the classification performance. Experiments on large dataset show that the proposed method obtains promising performances. Our proposed design is general and may inspire the community for multi-task labeling.

Acknowledgment This work was supported by the SZU Top Ranking Project (No. 86000000210).

References

1. Burgos-Artizzu, X.P., Coronado-Gutiérrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F., Gratacós, E.: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Scientific Reports **10**(1), 1–12 (2020)
2. Cai, Y., Sharma, H., Chatelain, P., Noble, J.A.: Sonoeyenet: Standardized fetal ultrasound plane detection informed by eye tracking. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 1475–1478. IEEE (2018)
3. Chen, H., Wu, L., Dou, Q., Qin, J., Li, S., Cheng, J.Z., Ni, D., Heng, P.A.: Ultrasound standard plane detection using a composite neural network framework. IEEE transactions on cybernetics **47**(6), 1576–1586 (2017)
4. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5177–5186 (2019)
5. Dong, J., Liu, S., Liao, Y., Wen, H., Lei, B., Li, S., Wang, T.: A generic quality control framework for fetal ultrasound cardiac four-chamber planes. IEEE journal of biomedical and health informatics **24**(4), 931–942 (2019)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Hinton, G.E., et al.: Learning distributed representations of concepts. In: Proceedings of the eighth annual conference of the cognitive science society. vol. 1, p. 12. Amherst, MA (1986)
8. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. arXiv preprint arXiv:2004.11362 (2020)
9. Lin, Z., Li, S., Ni, D., Liao, Y., Wen, H., Du, J., Chen, S., Wang, T., Lei, B.: Multi-task learning for quality assessment of fetal head ultrasound images. Medical image analysis **58**, 101548 (2019)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
11. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
12. Salomon, L., Alfirevic, Z., Berghella, V., Bilardo, C., Hernandez-Andrade, E., Johnsen, S., Kalache, K., Leung, K.Y., Malinger, G., Munoz, H., et al.: Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound in Obstetrics & Gynecology **37**(1), 116–126 (2011)
13. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: Cnn-rnn: A unified framework for multi-label image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2285–2294 (2016)
14. Zhu, F., Li, H., Ouyang, W., Yu, N., Wang, X.: Learning spatial regularization with image-level supervisions for multi-label image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5513–5522 (2017)

Appendix

Table 2. Abbreviation and full name of structures and standard plane. In the Abbreviation column, blue represents the standard plane, black represents the anatomical structure, and the horizontal line represents no abbreviation.

Abbreviation	Full name
SLAP	long axis plane of the spine
CMP	plane of the conus medullary position
TAP	the axial plane at the level of the thalamus
LVAP	the axial plane at the level of the lateral ventricle
NCP	coronal plane of the nasolabial
FHFMP	midsagittal plane of the fetal head and face
SPP	soft palate plane
FCP	four-chamber view plane
UAAP	upper abdominal axial plane
FLAP	long axis plane of the femur
CF	cerebral falx
PH	posterior horn
SPC	cavity of septum pellucidum
CM	conus medullaris
SCR	sacro-coccyx region
-	thalamus
IC	intact cranium
NA	apex of nose
NB	nasal bone
-	palate
-	mandible
SP	soft palate
-	pharynx
FCVH	four-chamber view of heart
-	aorta
-	lung
ST	stomach
PSUV	umbilical vein at the level of the portal sinus
FD	femur diaphysis
-	spine
UL	upper lip
LL	lower lip
-	chin
-	nostril

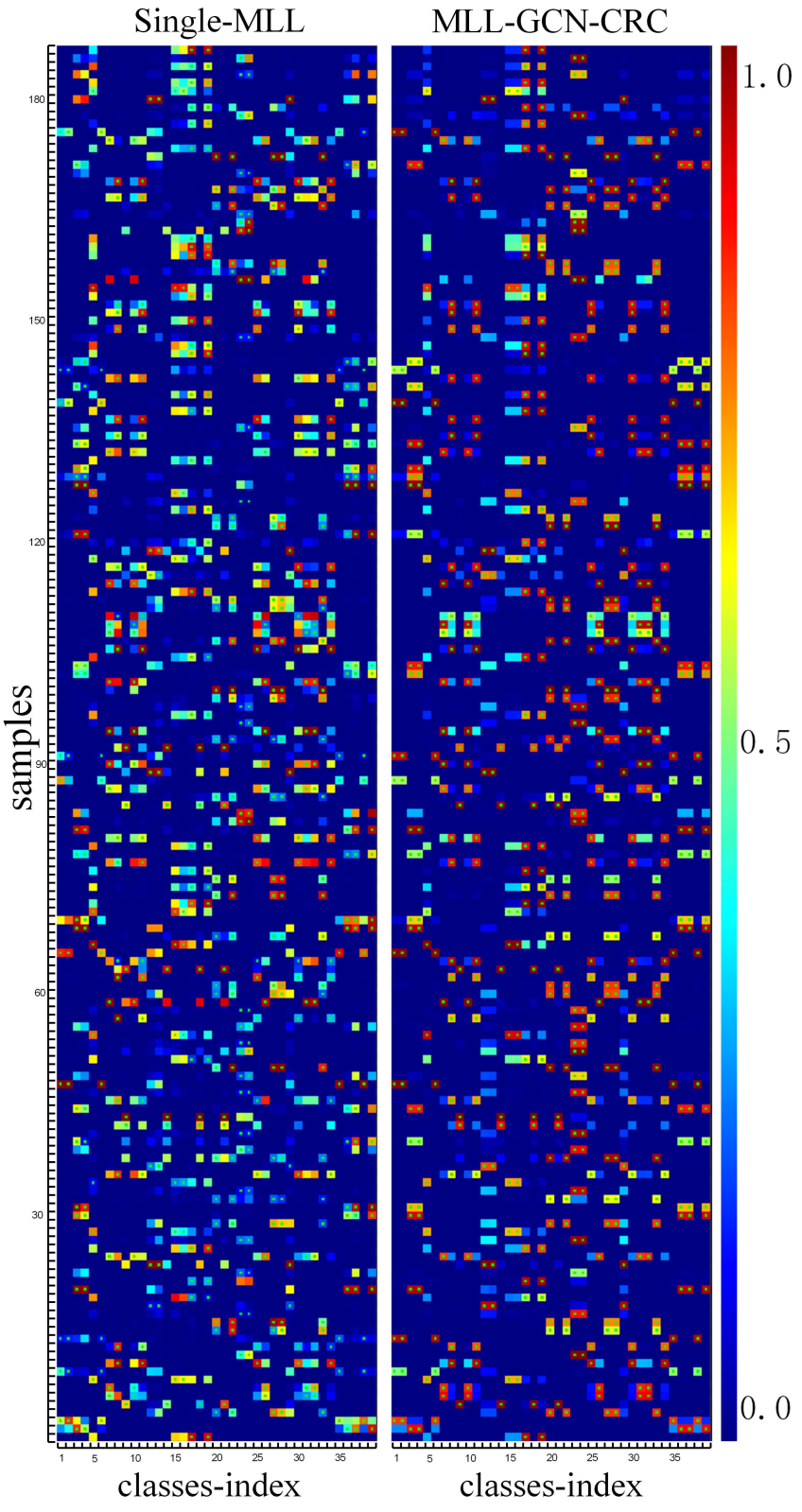

Fig. 5. Two score matrices of Single-MLL and MLL-GCN-CRC. Each row in the matrix represents an example class prediction. The little green points indicate the true label and the color of matrix element indicate the model output score.

Image	Queue	MLL_GCN_CRC	Single_MLL	Image	Queue	MLL_GCN_CRC	Single_MLL
	**FCP**	✓	✓		**FLAP**	✓	×
	FCVH	✓	✓		**FD**	✓	✓
	Spine	✓	✓				
	Aorta	✓	×				
	Lung	✓	✓				
	UAAP		✓				
	ST		✓				
	Spine		✓				
	PSUV		✓				
	**LVAP**	✓	✓				
	LVPH	✓	✓		**FCP**	✓	✓
	IC	✓	✓		**FCVH**	✓	✓
	SPC	✓	✓		**Spine**	✓	×
	CF	✓	✓		**Aorta**	✓	×
	SPC		✓		**Lung**	✓	✓

Fig. 6. The result comparison between Single-MLL and MLL-GCN-CRC. In the Queue column, the bold font represents the category in the ground true, the red bold represents the standard plane category, and the gray non-bold represents the category that does not exist in the ground true. In the MLL-GCN-CRC and Single-MLL columns, a check indicates that the category is predicted, and a cross indicates that the category is not predicted.