2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-2923-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISBI52829.2022.9761685

MULTI-TASK LEARNING WITH CONTEXT-ORIENTED SELF-ATTENTION FOR BREAST ULTRASOUND IMAGE CLASSIFICATION AND SEGMENTATION

Meng Xu[1*] *Kuan Huang*[2] *Xiaojun Qi*[1]

[1]Utah State University, Department of Computer Science, Logan, UT
[2]Baylor College of Medicine, Lester and Sue Smith Breast Center, Houston, TX

ABSTRACT

Breast cancer is a great threat to women's health. Automatic analysis of Breast UltraSound (BUS) images can help radiologists make more accurate and efficient diagnoses of breast cancer. We propose a Multi-Task Learning Network with Context-Oriented Self-Attention (MTL-COSA) module to automatically and simultaneously segment tumors and classify them as benign or malignant. The COSA module incorporates prior medical knowledge to guide the network to learn contextual relationships for better feature representations in BUS images. Extensive cross-validation experiments are conducted on two public datasets to evaluate the performance of MTL-COSA and several state-of-the-art methods. MTL-COSA achieves the best classification results and second-best segmentation results compared with deep learning-based methods (5 classification methods and 3 segmentation methods).

Index Terms— context-oriented self-attention, multi-task learning, segmentation, classification, breast ultrasound

1. INTRODUCTION

Breast cancer is the most commonly diagnosed cancer and the fifth leading cause of cancer mortality among women worldwide. It accounts for 1 in 4 cancer cases, and 1 in 6 cancer deaths in women [1]. Early diagnosis and appropriate treatments can increase survival rates. Breast UltraSound (BUS) imaging is a portable, valuable, and widely available diagnosis tool that has been commonly used in the early diagnosis of breast cancer of women in all ages [2]. Automatic analysis of BUS images can help radiologists make more accurate and efficient diagnoses of breast cancer. However, it is still a challenging task due to the poor quality of BUS images and the lack of public training datasets.

BUS image segmentation automatically extracts tumor regions from a BUS image. Recently, many deep learning-based BUS image segmentation methods have been proposed. Amiri *et al.* [3] propose a two-stage method that uses a U-Net to detect tumors and another U-Net to segment the tumor region. Huang *et al.* [4] propose a fully convolutional network for segmentation and use conditional random fields to post-process segmented regions to achieve more accurate results.

BUS image classification automatically classifies breast tumors into benign or malignant categories. Deep Neural Network (DNN) based BUS image classification methods have been recently well studied. Virmani *et al.* [5] show that features extracted by DNNs (e.g., VGG [6] and ResNet [7]) are efficient for BUS image classification. Zhuang *et al.* [8] construct a SEF-Net to extract Region of Interest (ROI) sequence features and construct a GRUC-Net to integrate strong correlation features into the ROI features to classify breast tumors.

BUS image analysis automatically and simultaneously accomplishes segmentation and binary BUS classification. Therefore, it is more appealing and practical than either segmentation or classification. Researchers design Multi-Task Learning (MTL) frameworks to extract breast tumor regions and label segmented tumors as benign or malignant. Zhou *et al.* [9] design an iterative feature-refining training strategy in an MTL framework to guide feature extraction. Singh *et al.* [10] utilize atrous convolution and channel-wise weighting in an MTL network to learn tumor features at different scales and re-balance the relative impact of encoded features.

Recently, researchers integrate either attention mechanisms or prior medical knowledge in DNNs to achieve better BUS segmentation and classification results. Attention mechanisms make networks focus more on the important parts of BUS images and therefore learn better feature representations and achieve better results. Prior medical knowledge provides helpful information to guide either segmentation or classification. For example, Xu *et al.* [11] design a multi-scale self-attention model to extract rich contextual relationships, which leads to better segmentation results. Zhang [12] *et al.* include soft and hard attention in an MTL network to pay more attention to tumor regions and achieve better classification result. Huang *et al.* [4] incorporate prior medical knowledge to correct any conflicting mistakes. The above three systems achieve better segmentation or classification results. However, none involves both attention and prior medical knowledge and explores the feasibility of bringing the output of one task to the network to guide the other task. Furthermore, Huang's system is not an end-to-end model.

To address the above issues, we propose a novel end-to-

*Corresponding author: Meng Xu. Email: meng.xu0201@gmail.com

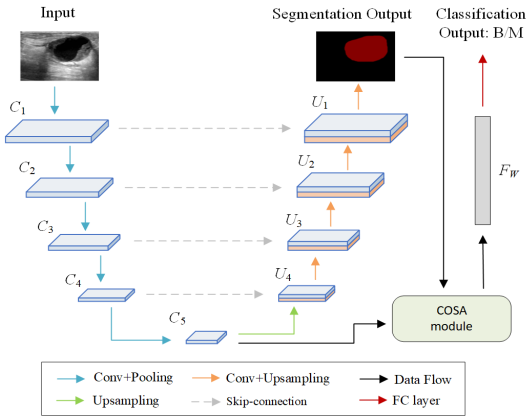

Fig. 1. An overview of the proposed MTL-COSA.

end MTL framework named MTL-COSA (COSA stands for Context-Oriented Self-Attention) to incorporate the COSA module in an MTL DNN to achieve better segmentation and classification results. Our major contributions are: (1) Proposing an MTL DNN for simultaneous breast tumor classification and segmentation. (2) Adopting the self-attention model [13] to focus more on each of three regions (background, tumor, and margin) to learn contextual relationships within each region for better feature representations. (3) Designing a COSA module that incorporates prior medical knowledge to achieve better segmentation and classification. (4) Conducting extensive experiments on two public datasets by comparing the proposed MTL-COSA with 5 state-of-the-art deep learning-based classification methods and 3 state-of-the-art deep learning-based segmentation methods.

2. THE PROPOSED MTL-COSA

The proposed MTL-COSA utilizes segmentation outputs to automatically estimate the tumor boundary, which is treated as prior medical knowledge, to be integrated into the COSA module to guide the MTL DNN to learn contextual relationships for better feature representations to improve both segmentation and classification accuracy. In this section, we first present the architecture of MTL-COSA and then explain the COSA module in detail.

2.1. Architecture Overview

The architecture of the proposed MTL-COSA is illustrated in Fig. 1. It consists of three branches: backbone feature extraction, segmentation, and classification. Segmentation and classification branches use the same feature map extracted by the backbone to produce segmentation and classification results, respectively. The backbone network is ResNet-101 [7], which has five convolutional blocks. We use C_i to denote the output of one of the five blocks, where integer i corresponds to a block number ranging from 1 to 5. Each C_i is at different scales in different depths, where scales decrease and depths increase with increasing i.

The backbone feature extraction branch performs downsampling operations, and the segmentation branch performs upsampling operations. These two branches pair together to form a U-shape structure. We use U_i to denote a upsampled feature map in the segmentation branch, where i ranges from 1 to 4. Starting with C_5, the feature map extracted by the backbone, we concatenate its upsampled feature map with the feature map C_4 to obtain U_4, which integrates local and semantic information from both blocks 5 and 4. We repeat the similar operation to get U_3, U_2, and U_1. The feature map U_{i-1} contains local spatial and high-level semantic information from i^{th} and $i-1^{th}$ blocks. A 3×3 convolution is then applied to U_1, followed by bilinear interpolation and softmax layer to generate the final segmentation result.

The feature map C_5 extracted by the backbone feature extraction branch together with the segmentation output is fed into the proposed COSA module to compute rich contextual relationships in BUS images and generate a weighted feature vector F_W, which is passed to a fully connected layer to generate the classification result. The COSA module will be explained in Subsection 2.2.

The overall loss of MTL-COSA is the weighted sum of the loss of the segmentation branch $\mathcal{L}_{seg}$ and the loss of the classification branch $\mathcal{L}_{cls}$.

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{seg} + \beta \cdot \mathcal{L}_{cls} \tag{1}$$

where α and β are weights of losses from segmentation and classification branches, respectively. $\alpha + \beta = 1$. Cross-entropy is employed to compute both $\mathcal{L}_{seg}$ and $\mathcal{L}_{cls}$.

2.2. COSA Module

The self-attention mechanism and prior medical knowledge are commonly used in BUS image segmentation [11, 4]. However, they have not been well studied in the field of BUS image classification and segmentation. To the best of our knowledge, we are the first to incorporate the segmentation results into self-attention [13] to simultaneously segment tumors and classify them as benign or malignant. The segmentation results contain the shape of extracted tumors, which can be used as the estimated prior medical knowledge, to guide the proposed MLT-COSA to learn contextual relationships in BUS to better represent features and therefore achieve better classification and segmentation results.

Fig. 2 illustrates the proposed COSA module. It takes the feature map C_5 and segmentation output as inputs and then outputs a weighted feature vector F_W. Following the findings in [14] that posterior acoustic shadowing in the background region, tumor shape, and tumor margin are three characteristics to differentiate benign and malignant tumors, we split the segmentation output into three regions (background, tumor, and margin) to capture the three vital characteristics. To this

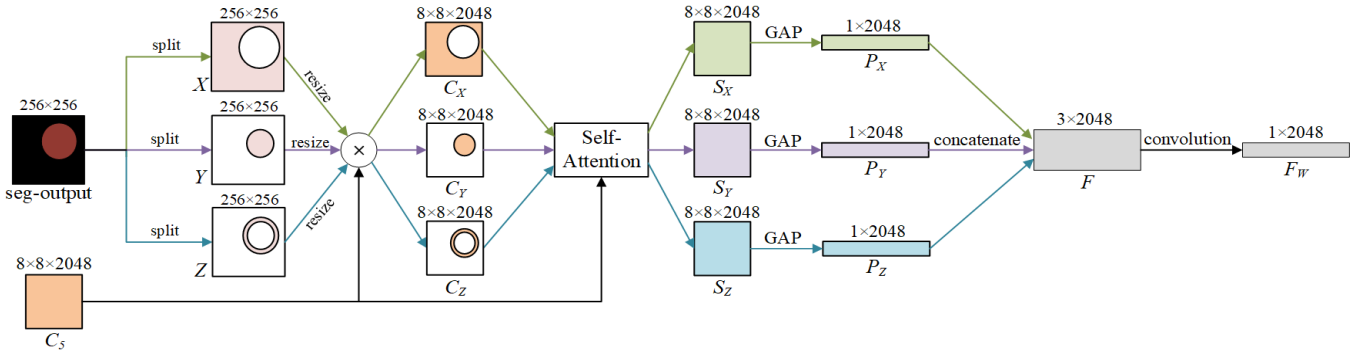

Fig. 2. Illustration of the proposed COSA module.

end, we apply the Sobel edge detector on the segmentation result to find the tumor contour. We then use a 5×5 square structuring element to perform a dilation operation on the tumor contour to find the inner and outer boundaries. The tumor region falls within the inner boundary. The margin region falls between the inner and the outer boundaries. The background region falls outside the outer boundary. The COSA module employs self-attention to focus on learning contextual relationships to better represent features in each region without the interference of other regions. To facilitate the description of the COSA, we label the dimension of the data at each step in Fig. 2. For the segmentation output of size 256×256, three non-overlapping binary maps X, Y, and Z respectively capture background, tumor, and margin regions, where the pink region contains values of 1's and the white region contains values of 0's. The union of three pink regions is a binary map of size 256×256 with all 1's. Region maps X, Y, and Z are then resized to 8×8 and individually multiplied with C_5 to generate three regional feature maps C_X, C_Y, and C_Z, which respectively contain features of background, tumor, and margin regions. C_X, C_Y, and C_Z together with C_5 is individually fed into the self-attention module [13] to compute contextual relationships in the background, tumor, and margin regions, respectively. This self-attention module builds upon the module in [11] to take two inputs and then output an attentive feature map. Attentive feature maps S_X, S_Y, and S_Z produced by the self-attention module are individually fed into a Global Averaging Pooling (GAP) layer to generate their corresponding feature vectors P_X, P_Y, and P_Z, which are concatenated to construct a new feature vector F of size 3×2048. F is resized to $1 \times 2048 \times 3$ and a 1×1 convolution filter is then applied to F to generate F_W of size $1 \times 2048 \times 1$. Lastly, the final weighted feature F_W is resized to 1×2048.

3. EXPERIMENTS

In this section, we describe BUS datasets, explain evaluation metrics, and compare experimental results of the proposed MTL-COSA, 3 DNN-based segmentation methods, and 5 DNN-based classification methods.

Datasets. We use two publicly available BUS datasets, Dataset B [15] and Dataset BUSI [16], in our experiments. Dataset B contains 163 BUS images with an average size of 760×570 pixels, where 110 images have benign tumors and 53 images have malignant tumors. Dataset BUSI contains 830 BUS images with an average size of 500×500 pixels, where 487 images have benign tumors, 210 images have malignant tumors, and 133 images have no tumors. We perform image segmentation and binary classification on 860 BUS images with benign and malignant tumors in both datasets.

Evaluation Metrics. We conduct 5-fold cross-validation on the above two datasets to evaluate the segmentation and classification performance of each compared method. We compute the average value of each metric after performing 5 runs on the test set. In each fold, images with benign and malignant tumors are split into training and testing sets according to their ratio in each dataset. Segmentation metrics are True Positive Ratio (TPR), False Positive Ratio (FPR), mean Intersection over Union (mIoU), Dice's Coefficient (DSC), and Area Error Ratio (AER). Classification metrics are TPR, FPR, Classification Accuracy (ACC), Precision (PRE), F_1-score, and Receiver Operating Characteristic (ROC) curve.

Results. We compare the proposed MTL-COSA with several state-of-the-art methods in terms of segmentation and classification accuracy. The compared segmentation methods are U-shape ResNet-101 [7] (UResNet), Multi-Task Learning (MTL) with a classification branch added to UResNet, MTL-SA with conventional self attention [13] added to MTL, and the proposed MTL-COSA with COSA added to MTL. MTL feeds C_5 into a fully connected layer for classification while passing C_5 into U_4 to U_1 for segmentation. MTL-SA applies self attention [13] to C_5 and the attentive C_5 is used to generate classification and segmentation results in the same way as MTL. The compared classification methods are VGG-16 [17], LeNet [18], ResNet-101, MTL, MTL-SA, and MLT-COSA.

Table 1 summarizes segmentation results of all compared methods on each dataset. The Single Task (ST) segmentation network, UResNet, achieves better TPR, mIoU, and DSC values than all MTL methods on both datasets. It is reasonable because adding a classification branch to UResNet

Table 1. Summary of tumor segmentation results (%)

Datasets	Methods		TPR	FPR	mIoU	DSC	AER
Dataset B [15]	ST	UResNet	84.39	34.03	86.11	82.25	49.64
	MT	MTL	**82.18**	23.03	85.00	79.92	40.85
		MTL-SA	78.54	**15.50**	84.44	78.90	36.96
		MTL-COSA	82.17	17.14	**85.87**	**81.82**	**34.97**
Dataset BUSI [16]	ST	UResNet	77.09	33.34	82.27	77.63	56.25
	MT	MTL	77.52	39.11	81.48	76.38	61.59
		MTL-SA	75.28	37.05	81.09	75.61	61.77
		MTL-COSA	**77.85**	**37.01**	**82.13**	**77.48**	**59.15**

leads to a weight reduction of the segmentation branch in the loss function. This reduction is determined by α in Eq. (1), with smaller α leading to more weight reduction in segmentation and therefore leading to worse segmentation results. Among three MTL methods, MTL-COSA achieves the highest mIoU of 85.87%, the highest DSC of 81.82%, and the lowest AER of 34.97% on Dataset B. Among three MTL methods, MTL-COSA outperforms other two methods in all five metrics. Overall, MTL-COSA achieves the best segmentation performance on both datasets compared to other MTL methods. The segmentation performance is dropped for all MTL methods when comparing with the ST method since less weight is employed in training to reduce segmentation error. However, MTL-COSA maintains the smallest drop in segmentation performance due to its integration of both attention mechanism and prior medical knowledge.

Table 2. Summary of tumor classification results (%)

Datasets	Methods		TPR	FPR	ACC	PRE	F_1
Dataset B [15]	ST	VGG-16	85.33	33.27	79.16	84.10	84.60
		LeNet	**92.73**	51.82	77.99	78.26	84.74
		ResNet	90.91	**29.82**	**84.09**	**86.23**	**88.38**
	MT	MTL	**92.73**	31.64	84.69	86.09	88.88
		MTL-SA	90.82	25.64	85.34	88.24	89.22
		MTL-COSA	91.73	**22.36**	**87.08**	**89.36**	**90.41**
Dataset BUSI [16]	ST	VGG-16	89.22	20.95	85.92	89.83	89.51
		LeNet	90.84	43.33	79.74	81.44	85.81
		ResNet	**93.56**	**17.14**	**90.10**	**91.90**	**92.68**
	MT	MTL	93.57	14.76	90.86	92.98	93.22
		MTL-SA	**94.27**	16.19	90.87	92.39	93.30
		MTL-COSA	92.20	**10.00**	**91.48**	**95.05**	**93.59**

Table 2 summarizes classification results of all compared methods on each dataset. Among the ST classification methods, ResNet achieves the best overall performance. Among the MTL classification methods, MTL-COSA achieves the best performance in terms of FPR, ACC, PRE, and F_1 scores. All MTL classification methods achieve better performance than all ST classification methods in terms of ACC, PRE, and F_1 scores. At least one MTL classification method achieve better performance than all ST classification methods in terms of TPR and FPR scores. It is clear that classification results are significantly improved for MTL methods. Adding a segmentation branch, which has enough training samples, makes the network learn better feature representations and therefore achieve better classification results. This improvement surpasses the performance drop caused by weight reduction of the classification branch in the loss function.

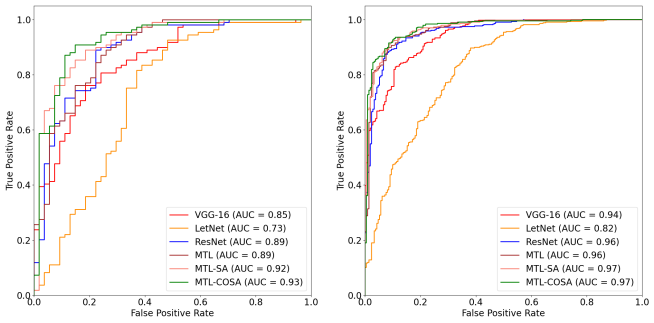

Fig. 3. ROC curves of six compared classification methods on Dataset B(left) and Dataset BUSI (right).

Fig. 3 shows ROC curves with values of Area Under the ROC Curve (AUC) for each method listed in Table 2. Among three ST classification methods, ResNet yields the highest AUC of 0.89 on Dataset B and 0.96 on Dataset BUSI. Among three MTL methods, the proposed MTL-COSA achieves the highest AUC of 0.93 on Dataset B and 0.97 on Dataset BUSI. The MTL method without attention and prior medical knowledge achieves the worst classification results on both datasets, which are comparable with the classification results obtained by the best ST method. It is clear from Table 2 and Figure 3 that the COSA module guides the MLT network to utilize the estimated prior medical knowledge in the attention mechanism to learn better feature representations and achieve better classification results and comparable segmentation results than ST methods.

Implementation Details. All experiments are conducted on Ubuntu 18.04 system, Intel(R) Core(TM) CPU i5-11600K 3.9 GHz, and 2 NVIDIA GeForce 1080Ti graphics cards. To train all networks, Adam optimizer is used with learning rate of 0.0001, momentum β_1 of 0.9, momentum β_2 of 0.99, and weight decay of 0.0005. Batch size is 12 and the number of training epochs is 100. All BUS images are resized to 256×256 as the input. Weights of two branches α and β are empirically set to be 0.8 and 0.2, respectively. Other values significantly reduce the mIoU value of segmentation results.

4. CONCLUSION

We propose a novel MTL-COSA network for simultaneous BUS image segmentation and binary classification. The COSA module utilizes the segmentation output to gain estimated prior medical knowledge and use it to learn contextual relationships for better feature representations in BUS images. MTL-COSA achieves significant classification improvement and comparable segmentation performance on two datasets when comparing with other state-of-the-art deep learning-based methods.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by [15, 16]. Ethical approval was not required as confirmed by the license attached with the open access data.

6. ACKNOWLEDGMENTS

No funding was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

7. REFERENCES

[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, Isabelle S., A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] O. Ginsburg, C. Yip, A. Brooks, A. Cabanes, M. Caleffi, J. A. Dunstan Yataco, B. Gyawali, et al., "Breast cancer early detection: A phased approach to implementation," *Cancer*, vol. 126, pp. 2379–2393, 2020.

[3] M. Amiri, R. Brooks, B. Behboodi, and H. Rivaz, "Two-stage ultrasound image segmentation using u-net and test time augmentation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 6, pp. 981–988, 2020.

[4] K. Huang, Y. Zhang, H. Cheng, P. Xing, and B. Zhang, "Semantic segmentation of breast ultrasound image with fuzzy deep learning network and breast anatomy constraints," *Neurocomputing*, vol. 450, pp. 319–335, 2021.

[5] J. Virmani, R. Agarwal, et al., "Deep feature extraction and classification of breast ultrasound images," *Multimedia Tools and Applications*, vol. 79, no. 37, pp. 27257–27292, 2020.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[8] Z. Zhuang, W. Ding, S. Zhuang, A.N.J. Raj, et al., "Tumor classification in automated breast ultrasound (abus) based on a modified extracting feature network," *Computerized Medical Imaging and Graphics*, vol. 90, pp. 101925, 2021.

[9] Y. Zhou, H. Chen, Y. Li, Q. Liu, X. Xu, et al., "Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images," *Medical Image Analysis*, vol. 70, pp. 101918, 2021.

[10] V.K. Singh, H.A. Rashwan, M. Abdel-Nasser, Md Sarker, M. Kamal, et al., "An efficient solution for breast tumor segmentation and classification in ultrasound images using deep adversarial learning," *arXiv preprint arXiv:1907.00887*, 2019.

[11] M. Xu, K. Huang, Q. Chen, and X. Qi, "Mssa-net: Multi-scale self-attention network for breast ultrasound image segmentation," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 827–831.

[12] G. Zhang, K. Zhao, Y. Hong, X. Qiu, K. Zhang, et al., "Sha-mtl: soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–7, 2021.

[13] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7354–7363.

[14] S. Gokhale, "Ultrasound characterization of breast masses," *The Indian Journal of Radiology & Imaging*, vol. 19, no. 3, pp. 242, 2009.

[15] M.H. Yap, G. Pons, J. Martí, S. Ganau, M. Sentís, et al., "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 4, pp. 1218–1226, 2017.

[16] W. Al-Dhabyani, M. Gomaa, and A. Khaled, H.and Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, pp. 104863, 2020.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] Y. LeCun, Lé. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.