SPATIAL-CONTEXT-AWARE DEEP NEURAL NETWORK FOR MULTI-CLASS IMAGE CLASSIFICATION

Jialu ZHANG*†, Qian ZHANG*, Jianfeng REN*, Yitian ZHAO†, Jiang LIU*†‡
* School of Computer Science, University of Nottingham Ningbo China
† Cixi Institute of Biomedical Engineering, Chinese Academy of Sciences
‡ Department of Computer Science and Engineering, Southern University of Science and Technology

ABSTRACT

Multi-label image classification is a fundamental but challenging task in computer vision. Over the past few decades, solutions exploring relationships between semantic labels have made great progress. However, the underlying spatial-contextual information of labels is under-exploited. To tackle this problem, a spatial-context-aware deep neural network is proposed to predict labels taking into account both semantic and spatial information. This proposed framework is evaluated on Microsoft COCO and PASCAL VOC, two widely used benchmark datasets for image multi-labelling. The results show that the proposed approach is superior to the state-of-the-art solutions on dealing with the multi-label image classification problem.

Index Terms— Multi-label, image classification, deep learning, spatial context

1. INTRODUCTION

With the rapid development of technologies, abundant visual information is constantly received. One of the fundamental but challenging problems for image understanding is to label the objects, locations or attributes in the images, possibly with more than one label. Multi-label image classification problem has attracted the attention of researchers in the past few years. However, the rich semantic information and higher-order label co-occurrence are challenging to model [1, 2].

Recently, many deep convolutional neural networks (DCNNs) were developed for single-label image classification problem [3–5], and transforming the multi-label classification problem into multiple binary classification tasks is one of the common strategies [6] to solve the multi-label image classification problem. However, this kind of method ignores the inter-dependencies among labels, which have been proved useful [7, 8]. To tackle this problem, researchers developed various DCNNs [7, 9–14] that can consider all the labels for a given image concurrently. For example, Wang *et al.* designed a sequentially predict model and used a recurrent neural network (RNN) to determine label dependencies [7], while Chen *et al.* and Wang *et al.* employed graph convolutional networks to capture global label dependencies [13, 15]. By exploiting label-correlation information, these approaches made great progress on image multi-labelling. Nevertheless, object spatial information and image context information are not fully exploited in these approaches.

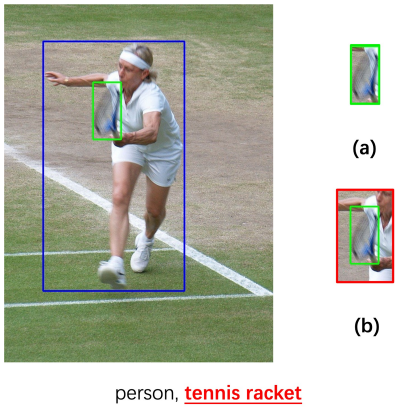

Fig. 1. An illustrative example. Previous methods cannot detect the tennis racket. By exploiting the label dependencies, spatial and context information, the proposed approach could more accurately detect it.

Wei *et al.* [16] and Yang *et al.* [17] addressed this problem by devising a 2-stage pipeline for multi-labelling in which the model generates image patches first and then labels them. However, these methods overemphasize the generated patches, thereby neglecting surrounding context infomation and label dependencies. The idea of object localization is similar to the attention mechanism that has been successfully applied in many vision tasks [10, 11, 18–20]. Fig. 1 illustrates the importance of label dependencies, spatial and context information. Additionally, context has been demonstrated useful in various visual processing tasks, such as recognition and detection [21–23].

To make good use of these dependencies and information, a two-branch spatial-context-aware DCNN is designed, where one branch is designed to extract the spatial information of the objects and the other aims at capturing the image context information. The networks label predictor follows the

This work was supported in part by the National Natural Science Foundation of China under Grant 72071116, and in part by the Ningbo Municipal Bureau of Science and Technology under Grant 2019B10026.

principle of multi-label image classification and utilizes the dependencies among labels. Moreover, with more contextual information, the proposed model performs well in labelling small objects that other models may not be able to capture.

Different from existing spatial-context-aware models exploiting the context information between objects [24] or considering the receptive field of the shallow layers as the context information to the pixels on the deep layers of the network [25], the proposed method exploits the spatial context information directly on the feature maps and utilizes the feature maps around the object as the background context. The experimental results on two large benchmark datasets, MS-COCO and PASCAL VOC, demonstrate the effectiveness of the proposed approach compared with state-of-the-art approaches.

The contributions of this paper are summarized as follow: 1) To make use of the spatial and context information to the object, a two-branch spatial-context-aware deep neural network is proposed for multi-label image classification problem. 2) The proposed image-context-aware branch could well exploit both spatial and semantic information of objects. 3) The proposed approach significantly outperforms the state-of-the-art approaches [5, 7, 10, 11, 13–15, 18, 19, 26] on the MS-COCO dataset [27] and PASCAL VOC [28] dataset.

2. METHODOLOGY

2.1. Overall Framework

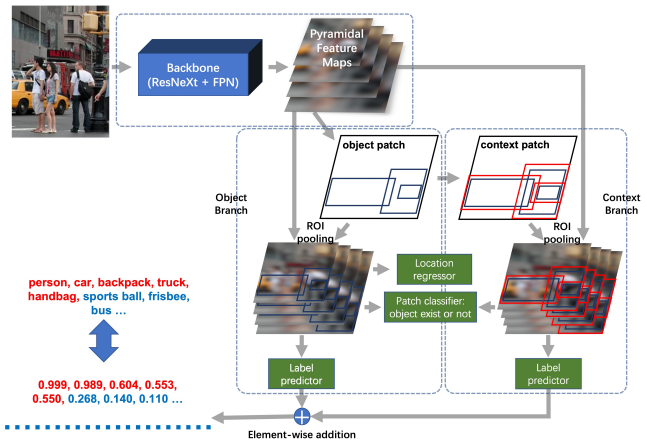

Fig. 2. An overview of the proposed model.

The overall framework of the proposed model is presented in Fig. 2. It consists of three main modules: 1) A feature extractor which integrates ResNeXt-101 [29] with feature pyramid networks (FPN) [30]. 2) Two patch generators. To utilize both spatial and contextual information, the object branch follows the structure in [31] to localize the objects and the context branch considers the image context information. These fragments and feature maps are combined to generate the final visual features. 3) Four patch processors. One regressor for positioning, one classifier for determining the existence of objects in the patch, and two predictors for predicting the label of the patch using a formulation of image multi-labelling.

The patch labelling results from the object branch and context branch are combined by an element-wise classifier fusion to generate the final predicted labels.

2.2. Feature Extractor

ResNeXt-101 [29] followed by FPN [30] is utilized as the feature extractor. The former aggregates a set of transformations to improve the classification capabilities of deep neural networks, and the latter employs the pyramid representations to extract a rich visual semantic abstract. The last pooling and classification layers in ResNeXt-101 are removed and the feature maps from the last convolutional layer are used as the input of FPN. A 4-stage semantic feature pyramid is built in FPN from high to low resolution. Denote the two successive feature extracting networks as $f_R(\mathbf{U};\theta_R)$ and $f_F(\mathbf{Z};\theta_F)$ respectively. Given an input image $\mathbf{I}$, the final generated pyramidal feature $\mathbf{X}$ can be obtained as:

$$\mathbf{X} = f_F\left(f_R\left(\mathbf{I};\theta_R\right);\theta_F\right). \tag{1}$$

2.3. Patch Generators

Based on the generated feature maps, a set of bounding boxes are generated to locate objects, covering both the contextual and spatial information of the objects. However, a tightly cropped bounding box may only contain the information of the object, but ignore the contextual information surrounding the object.

The tightly cropped bounding box in the object branch may neglect the contextual information, so the context branch is utilized to expand it into a larger one with potential object and its surrounding image content. Note that the expanded context patch is only used to determine the label of the object, but not for locating the object. As Fig. 1 shows, confidence in predicting 'tennis racket' increases as information about the surrounding environment, e.g., 'person', is included.

The tightly cropped object patch contains the most discriminant information for classify the object, while the expanded image patch containing both the object and additional contextual information. Hence, two branches with different objectives are designed to generate image patches. Specifically, as shown in Fig. 2, the generator in the object branch (left) takes the feature maps as the input and produces a set of rectangular object patches with confidence scores. Only objects within the rectangles are evaluated in this branch, and it is often difficult to detect small objects. Therefore, a patch expansion mechanism is applied to the generator in the context branch. This design allows more contextual information to be captured, and hence improves the accuracy of the patch prediction. Two branches are used together to predict the labels of the input image.

2.4. Multi-task Patch Processors

The generated object patches and context patches are then fed into four dense networks for location regression, patch classification and label prediction, respectively.

The first dense network, location regressor, is designed to accurately locate the objects to explore and utilize the image-level spatial information of different labels. It is guided by the location regression loss L_r. The objective is to maximize the intersection over union (IOU) between the generated bounding boxes and the ground-truth bounding boxes.

$$L_r\left(\hat{t}_i, t_i^*\right) = \sum_{i \in \{\mathrm{x,y,w,h}\}} \phi\left(\hat{t}_i - t_i^*\right), \tag{2}$$

$$\phi(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases}, \tag{3}$$

where $\hat{t}_i$ and t_i^* represent the predicted patches and ground-truth bounding boxes, respectively and x, y, w, h correspond to the centre coordinates, height and width of the proposed patch. $\phi(x)$ is a non-sensitive outlier-removal function for enhancing the training robustness.

The second dense network, patch classifier, determines the confidence whether an object exists in the bounding box. Hence it is a binary classification problem. The cross-entropy loss is used to guide the training process and the loss is defined as follows:

$$L_p\left(\hat{p}, p^*\right) = -p^* \log \hat{p} + \left(1 - p^*\right) \log\left(1 - \hat{p}\right), \tag{4}$$

where $\hat{p}$ and p^* represent the output confidence from the network and ground-truth, respectively. p^* would be assigned to 1 if the maximum IOU between the current patch and the ground-truth one exceeds a certain threshold, and 0 otherwise.

The remaining two dense networks are label predictors to determine which object the bounding box contains. These two dense networks are trained by using the binary cross-entropy (BCE) loss, which can exploit label dependencies in the multi-labelling tasks [7, 8, 10].

Given a one-hot vector of ground-truth labels $\mathbf{y}^* = \left[y^1, y^2, \ldots, y^C\right]^T$, where C represents the number of all possible labels in the dataset. y^i is a binary indicator that $y^l = 1$ if the image contains the label l, and $y^l = 0$ otherwise. The label-prediction loss is denoted as follows:

$$L_l\left(\hat{\mathbf{y}}, \mathbf{y}^*\right) = \sum_{i=1}^{C} y^i \log \sigma\hat{y}^i + \left(1 - y^i\right) \log\left(1 - \sigma\hat{y}^i\right). \tag{5}$$

Similar to $\mathbf{y}^*$ and y^i, $\hat{\mathbf{y}}$ and $\hat{y}^i$ denote the predicted confidence over all possible labels and the i-th category. σ in L_l is a learnable weighting factor, we fix it as 0.5 given by the empirically study.

Hence the total loss function L is represented as follows:

$$L = L_r + \alpha L_p + \beta L_l^O + \gamma L_l^C, \tag{6}$$

where α, β and γ is scale values that utilized to balance the losses and L_l^O and L_l^C represent the loss of the object branch and context branch respectively.

3. EXPERIMENTAL RESULTS

The proposed method is compared with various state-of-the-art models on two benchmark datasets - Microsoft COCO-2017 [27] and PASCAL VOC-2007 [28]. Experimental results indicate that the proposed model significantly and consistently outperforms all the compared solutions.

3.1. Experimental Settings

The proposed model is trained on PyTorch [35]. Stochastic gradient descend strategy is employed for training, with a mini-batch size of 2, a momentum of 0.5 and a decay rate of 0.01. The initial learning rate is 0.002. All images are resized to 416×416 before passing to the model. The proposed model is optimized by using the transfer-learning techniques, i.e., load and freeze the pre-trained weights of the backbone, and then train the remaining modules. Once the training converges, the whole model is jointly optimized.

The MS-COCO dataset [27] and PASCAL VOC [28] are two widely used public benchmark datasets for image labelling. The former contains 80 object categories and the latter covers 20 different categories. All instances in the two datasets are associated with at least one label. The standard evaluation protocol is used as in [11, 13–15, 18, 19, 26].

Standard evaluation metrics such as mean average precision (mAP), F1-score, precision and recall are used in experiments. The prefix 'macro-' indicates the average over all categories, while 'micro-' indicates the average over all samples. Therefore, 'macro-' is susceptible to the rare categories and 'micro-' is easily dominated by the major classes [36]. Denote 'macro-' as '-C' and 'micro-' as '-O'.

The proposed method is compared against state-of-the-art models including CNN-RNN [7], ResNet101 [5], RNN-Attention [32], ResNet101-SRN [10], RNN-frequency [33], DELTA [34], ResNet101-ACfs [11], DecoupleNet [12], ML-GCN [13], ResNet101-CRL [18], KSSNet [15], MS-CMA [19], WSL-GCN [26] and C-Tran [14], among which CNN-RNN [7], DecoupleNet [12], ML-GCN [13], KSSNet [15] and WSL-GCN [26] exploit semantic relations by capturing global label dependencies. ResNet101 [5] is initially designed for single-label image classification while also performs well in multi-labelling by using appropriate loss functions. RNN-Attention [32], ResNet101-SRN [10], RNN-frequency [33], DELTA [34], ResNet101-ACfs [11] and ResNet101-ACfs [11] exploit spatial information by extending attention ideas, while ResNet101-CRL constructs explicit context relations by feature-label co-projection. MS-CMA [19] adapts the

Method	mAP	F1-C	P-C	R-C	F1-O	P-O	R-O
CNN-RNN (CVPR, 2016, [7])	61.2	60.4	66.0	55.6	67.8	69.2	66.4
ResNet101 (CVPR, 2016, [5])	75.2	69.5	80.8	63.4	74.4	82.2	68.0
RNN-Attention (ICCV, 2017, [32])	-	67.4	79.1	58.7	72.0	84.0	63.0
ResNet101-SRN (CVPR, 2017, [10])	77.1	71.2	81.6	65.4	75.8	82.7	69.9
RNN-frequency (TMM, 2019, [33])	64.7	-	-	-	-	-	-
DELTA (PR, 2019, [34])	71.3	-	-	-	-	-	-
ResNet101-ACfs (CVPR, 2019, [11])	77.5	72.2	77.4	68.3	76.3	79.8	73.1
DecoupleNet (ICASSP, 2019, [12])	82.2	76.3	83.1	71.6	79.5	84.7	74.8
ML-GCN (CVPR, 2019, [13])	83.0	78.0	85.1	72.0	80.3	85.8	75.4
ResNet101-CRL (TSMC-S, 2020, [18])	81.1	75.8	81.2	70.8	78.1	83.6	73.3
KSSNet (AAAI, 2020, [15])	83.7	77.2	84.6	73.2	81.5	**87.8**	76.2
MS-CMA (AAAI, 2020, [19])	83.8	78.4	82.9	74.4	81.0	84.4	77.9
WSL-GCN (PR, 2021, [26])	84.8	-	-	-	-	-	-
C-Tran (CVPR, 2021, [14])	85.1	79.9	**86.3**	74.3	81.7	87.7	76.5
The Proposed	**86.0**	**80.3**	84.0	**77.5**	**83.2**	85.9	**80.6**

Table 1. Comparison results on the MS-COCO dataset. The best results are shown in bold.

Method	areoplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	**mAP**
CNN-RNN (CVPR, 2016, [7])	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	**99.7**	78.6	84.0
ResNet101 (CVPR, 2016, [5])	99.5	97.7	97.8	96.4	65.7	91.8	96.1	97.6	74.2	80.9	85.0	98.4	96.5	95.9	98.4	70.1	88.3	80.2	98.9	89.2	89.9
RNN-Attention (ICCV, 2017, [32])	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92.0	87.7	96.8	97.5	93.8	98.5	81.6	93.7	82.8	98.6	89.3	91.9
RNN-frequency (TMM, 2019, [33])	97.0	92.5	93.8	93.3	59.3	82.6	90.6	92.0	73.4	82.4	76.6	92.4	94.2	91.4	95.3	67.9	88.6	70.1	96.8	81.5	85.6
DELTA (PR, 2019, [34])	98.2	95.1	95.8	95.7	71.6	91.2	94.5	95.9	79.4	92.5	85.6	96.7	96.8	93.7	97.8	77.7	95.0	81.9	99.0	87.9	91.1
ML-GCN (CVPR, 2019, [13])	99.5	98.5	98.6	98.1	80.8	94.6	97.2	98.2	82.3	95.7	86.4	98.2	98.4	96.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0
ResNet101-CRL (TSMC-S, 2020, [18])	**99.9**	98.4	97.8	**98.8**	81.2	93.7	97.1	98.4	**82.7**	94.6	87.1	98.1	97.6	96.2	98.8	83.2	96.2	84.7	99.1	93.5	93.8
WSL-GCN (PR, 2021, [26])	99.7	98.5	**99.0**	97.8	86.2	96.2	98.3	**99.3**	81.1	95.9	88.0	**99.2**	**98.6**	97.1	**99.4**	85.0	**97.5**	84.3	99.0	94.0	94.7
The Proposed	99.4	**98.8**	98.0	98.6	**90.5**	**98.3**	**98.6**	98.4	81.3	**96.2**	**88.6**	96.7	**98.6**	**99.0**	99.3	**87.0**	**97.5**	**87.3**	98.6	**95.7**	**95.3**

Table 2. Comparison results on the PASCAL VOC dataset. The best results are shown in bold.

attention mechanism to a cross-modality version with graph embedding. C-Tran [14] exploits the dependencies among visual features and labels to tackle the image multi-labeling.

3.2. Comparisons to State-of-the-art Approaches

The comparison results to the state-of-the-art approaches on the MS-COCO dataset are summarized in Table 1. It is clear to see that the proposed model significantly outperforms all the state-of-the-art models in terms of the key evaluation metrics such as mAP and F1-score.

As shown in bold in Table 1, the proposed method achieves an overall mAP of 86.3%, a per-class F1 score of 80.6% and an overall F1-score of 83.1%, which greatly surpasses the previous best solution, C-Tran [14] by 1.2%, 0.7% and 1.4%, respectively. It shows that the proposed approach performs well on exploiting the semantic information, spatial and object context information and label dependencies.

Since the results of the state-of-the-art models are collected from the corresponding origional paper, only eight previous solutions are selected for the performance comparison on PASCAL VOC [28] dataset, and only comparing on the key evaluation metrics, mAP. The comparison results are summarized in Table 2. The proposed model achieves an mAP of 95.3 %, surpassing all the state-of-the-art models.

To visually demonstrate the effectiveness of the proposed model, several sample detection results for are shown in Fig. 3. The marked labels and bounding boxes prove that the proposed method can effectively utilize the context information. Objects in red are the ones that can not be detected by previous approaches but now can be detected by the proposed model by utilizing the spatial and contextual information. From Fig. 3, it can be seen that the proposed model plays a key role in the annotation of the easily neglected small objects such as the knife on the table and the person sailing the boat.

Fig. 3. Sample detection results. Small or partially occluded objects could be better detected with the background context.

4. CONCLUSION

In this paper, to make use of the spatial information of objects and the contextual information around the object, a spatial-context-aware deep neural network is designed for multi-class image classification problem. The object localization and the patch expansion enable the model to leverage both the semantic and spatial information of objects. The comparisons to the state-of-the-art solutions on the MS-COCO dataset and PASCAL VOC dataset demonstrate that the proposed network significantly and consistently outperforms all compared models.

References

[1] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[2] C. Silla and A. Frexitas, "A survey of hierarchical classification across different application domains," *Data Mining Knowl. Discovery*, vol. 22, no. 1-2, pp. 31–72, 2011.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[6] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[7] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *CVPR*, 2016, pp. 2285–2294.

[8] C. Yeh, W. Wu, W. Ko, and Y. Wang, "Learning deep latent space for multi-label classification," *AAAI*, vol. 31, no. 1, pp. 28382844, 2017.

[9] C. Tsai and H. Lee, "Adversarial learning of label dependency: A novel framework for multi-class classification," in *ICASSP*, 2019, pp. 3847–3851.

[10] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *CVPR*, 2017, pp. 5513–5522.

[11] H. Guo, K. Zheng, X. Fan, H. Yu, and S. Wang, "Visual attention consistency under image transforms for multi-label image classification," in *CVPR*, 2019, pp. 729–739.

[12] L. Liu, S. Guo, W. Huang, and M. Scott, "Decoupling category-wise independence and relevance with self-attention for multi-label image classification," in *ICASSP*, 2019, pp. 1682–1686.

[13] Z. Chen, X. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *CVPR*, 2019, pp. 5172–5181.

[14] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," in *CVPR*, 2021, pp. 16478–16488.

[15] Y. Wang, D. He, F. Li, X. Long, Z. Zhou, J. Ma, and S. Wen, "Multi-label classification with label graph superimposing," *AAAI*, vol. 34, no. 07, pp. 12265–12272, 2020.

[16] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "CNN: Single-label to Multi-label," *arXiv e-prints*, p. arXiv:1406.5726, 2014.

[17] H. Yang, J. Zhou, Y. Zhang, B. Gao, J. Wu, and J. Cai, "Exploit bounding box annotations for multi-label object recognition," in *CVPR*, 2016, pp. 280–288.

[18] S. Wen, W. Liu, Y. Yang, P. Zhou, Z. Guo, Z. Yan, Y. Chen, and T. Huang, "Multilabel image classification via feature/label co-projection," *IEEE Trans. Syst., Man, Cybern., Syst.*, pp. 1–10, 2020.

[19] R. You, Z. Guo, L. Cui, X. Long, Y. Bao, and S. Wen, "Cross-modality attention with semantic graph embedding for multi-label classification," *AAAI*, vol. 34, no. 07, pp. 12709–12716, 2020.

[20] C. He, S. Lai, and K. Lam, "Improving object detection with relation graph inference," in *ICASSP*, 2019, pp. 2537–2541.

[21] F. Yang, J. Ren, Z. Lu, J. Zhang, and Q. Zhang, "Rain-component-aware capsule-GAN for single image de-raining," *Pattern Recognit.*, vol. 123, pp. 108377, 2022.

[22] J. Ren, X.D. Jiang, and J. Yuan, "Learning LBP structure by maximizing the conditional mutual information," *Pattern Recognit.*, vol. 48, no. 10, pp. 3180–3190, 2015.

[23] M. Sina, N. Mehrdad, B. Ali, G. Sina, and H. Mohammad, "Cagnet: Content-aware guidance for salient object detection," *Pattern Recognit.*, vol. 103, pp. 107303, 2020.

[24] K. Bardool, T. Tuytelaars, and J. Oramas, "A systematic analysis of a context aware deep learning architecture for object detection.," in *BNAIC/BENELEARN*, 2019, vol. 2491.

[25] Y. Kong, M. Feng, X Li, H. Lu, X. Liu, and B. Yin, "Spatial context-aware network for salient object detection," *Pattern Recognit.*, vol. 114, pp. 107867, 2021.

[26] Y. Liu, W. Chen, H. Qu, S. Mahmud, and K. Miao, "Weakly supervised image classification and pointwise localization with graph convolutional networks," *Pattern Recognit.*, vol. 109, pp. 107596, 2021.

[27] T. Lin, M. Maire, S. Belongie, J. Hays, et al., "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[28] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vision*, vol. 88, pp. 303–338, 2010.

[29] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017, pp. 1492–1500.

[30] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017, pp. 2117–2125.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2016.

[32] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *ICCV*, 2017, pp. 464–472.

[33] F. Lyu, Q. Wu, F. Hu, Q. Wu, and M. Tan, "Attend and imagine: Multi-label image classification with visual attention and recurrent neural networks," *IEEE Trans. Multimedia*, vol. 21, pp. 1971–1981, 2019.

[34] W. Yu, Z. Chen, X. Luo, W. Liu, and X. Xu, "Delta: A deep dual-stream network for multi-label image classification," *Pattern Recognit.*, vol. 91, pp. 322–331, 2019.

[35] A. Paszke, S. Gross, F. Massa, et al., "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019, pp. 8026–8037.

[36] L. Tang, S. Rajan, and V. Narayanan, "Large scale multi-label classification via metalabeler," in *Proc. 18th Int. Conf. WWW*. ACM, 2009, pp. 211–220.