Paper ID: 2789

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

Introduction —— Multi-label image classification

- Single-label VS Multi-label



Single-label	**airplane**	**buildings**	**cat**
Multi-label	**clouds, airplane, sky**	**buildings, clouds, plants, reflection, sky**	**animal, cat, grass**

Applications

- Image retrieving
- Scene recognition
- Image captioning, etc.

Introduction —— Multi-label image classification

- Existing Solutions

  - Traditional approaches
    a. Hand-crafted features, SIFT, GIST, HOG etc.
    b. SVM, Tree-based approaches, Bayesian etc.

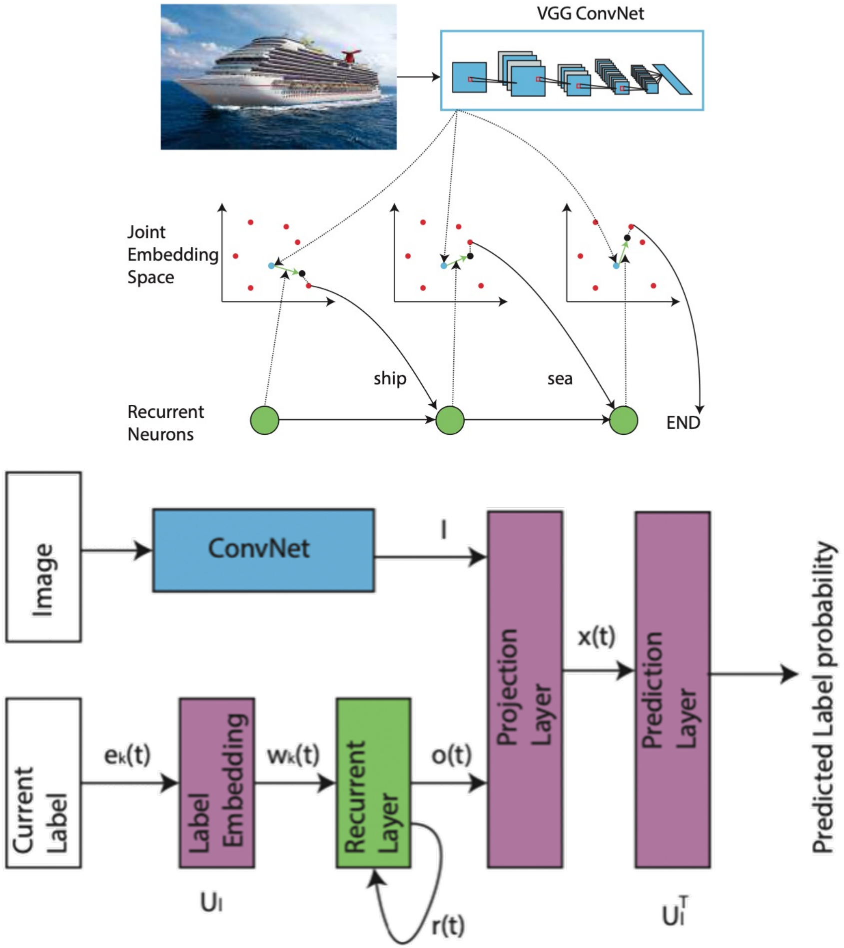
  - Deep learning approaches
    a. Approaches that exploit label inter-dependencies
       a) RNN-based methods
       b) GNN-based methods
       c) Latent space

    b. 2-stage pipeline approaches that utilize the spatial information of objects (region proposal generation & region labeling)

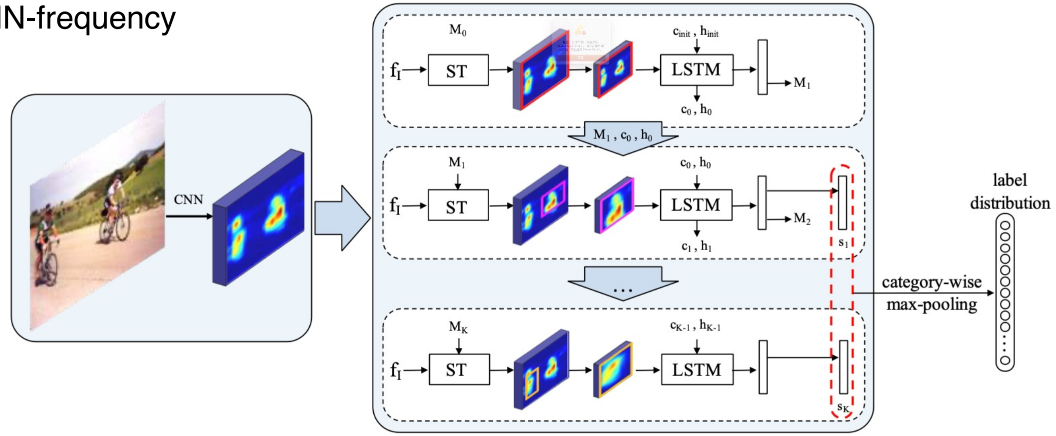
Introduction —— Multi-label image classification

- Approaches that exploit label inter-dependencies —— RNN-based methods

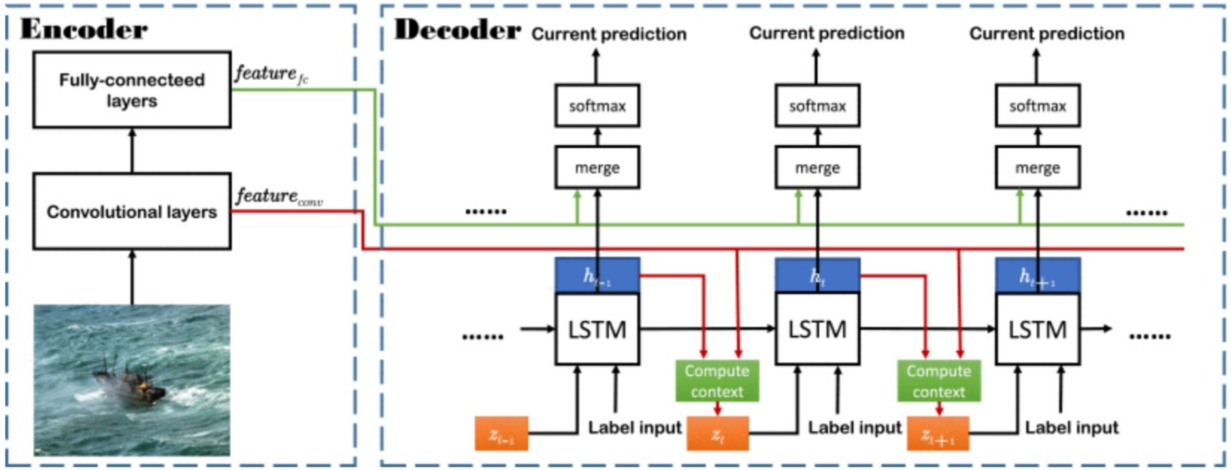
[1] CNN-RNN



[2] RNN-frequency



[3] RNN-attention

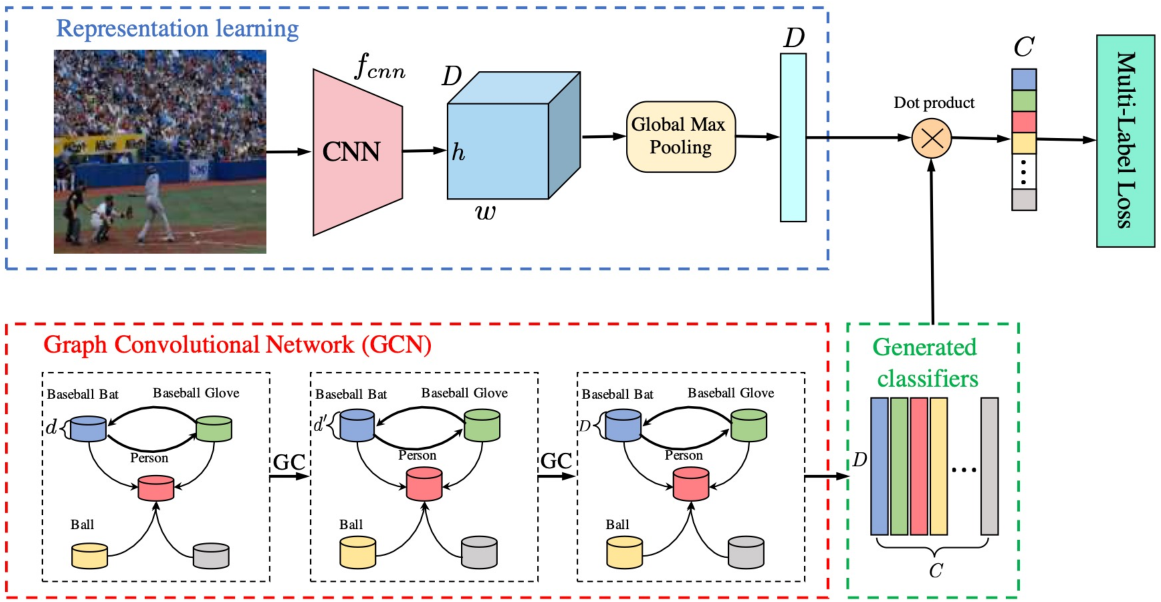


[1] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, W. Xu, CNN-RNN: A unified framework for multi-label image classification, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2285-2294.
[2] F. Lyu, Q. Wu, F. Hu, Q. Wu, M. Tan, Attend and imagine: Multi-label image classification with visual attention and recurrent neural networks, IEEE Trans. Multimedia 21 (2019) 1971-1981
[3] Z. Wang, T. Chen, G. Li, R. Xu, L. Lin, Multi-label image recognition by recurrently discovering attentional regions, in: Proc. IEEE Int. Conf. Comput. Vis., 2017.

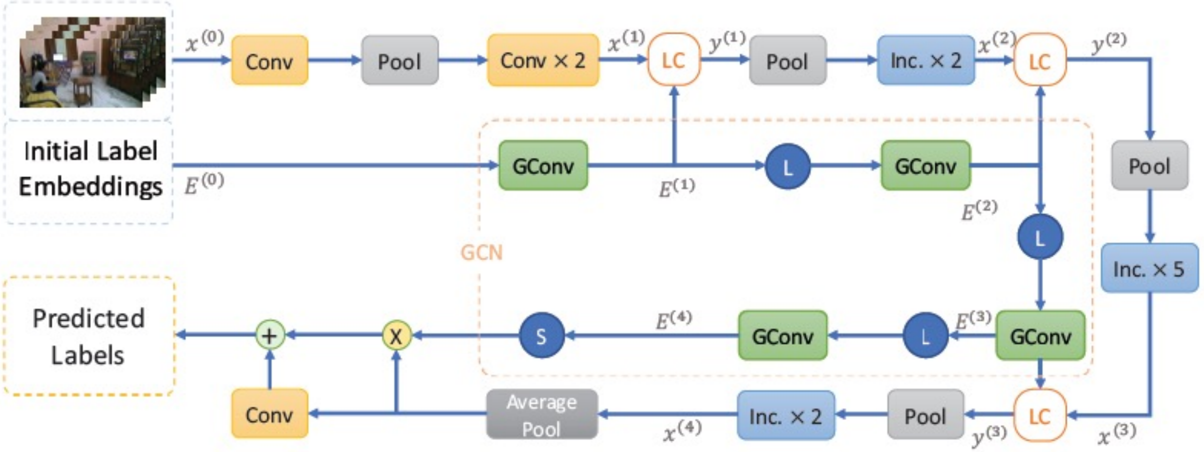
Introduction —— Multi-label image classification

- Approaches that exploit label inter-dependencies —— GNN-based methods

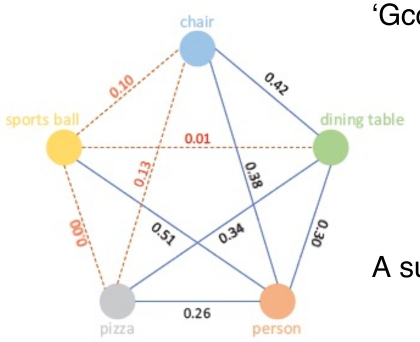
[1] ML-GCN



[2] KSSNet



'Gconv' represents 'Graph Convolution'



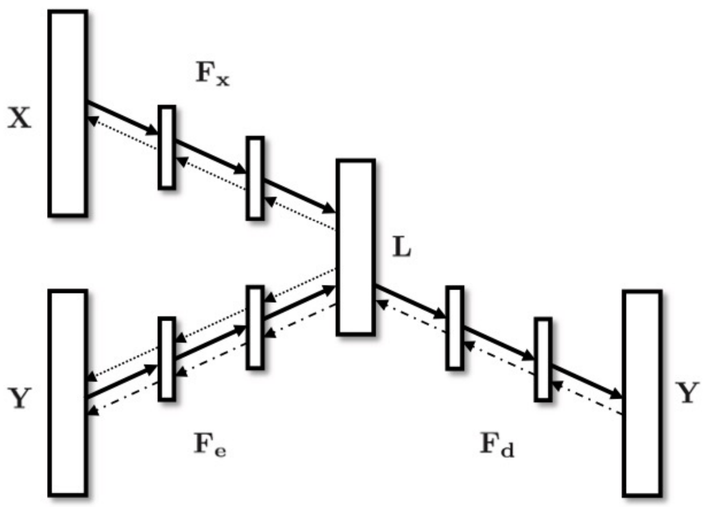
A sub-graph example

[1] Z. Chen, X. Wei, P. Wang, Y. Guo, Multi-label image recognition with graph convolutional networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 5172-5181.
[2] Y. Wang, D. He, F. Li, X. Long, Z. Zhou, J. Ma, S. Wen, Multi-label classification with label graph superimposing, Proc. AAAI Conf. Artif. Intell. 34 (2020) 12265-12272.

Introduction —— Multi-label image classification

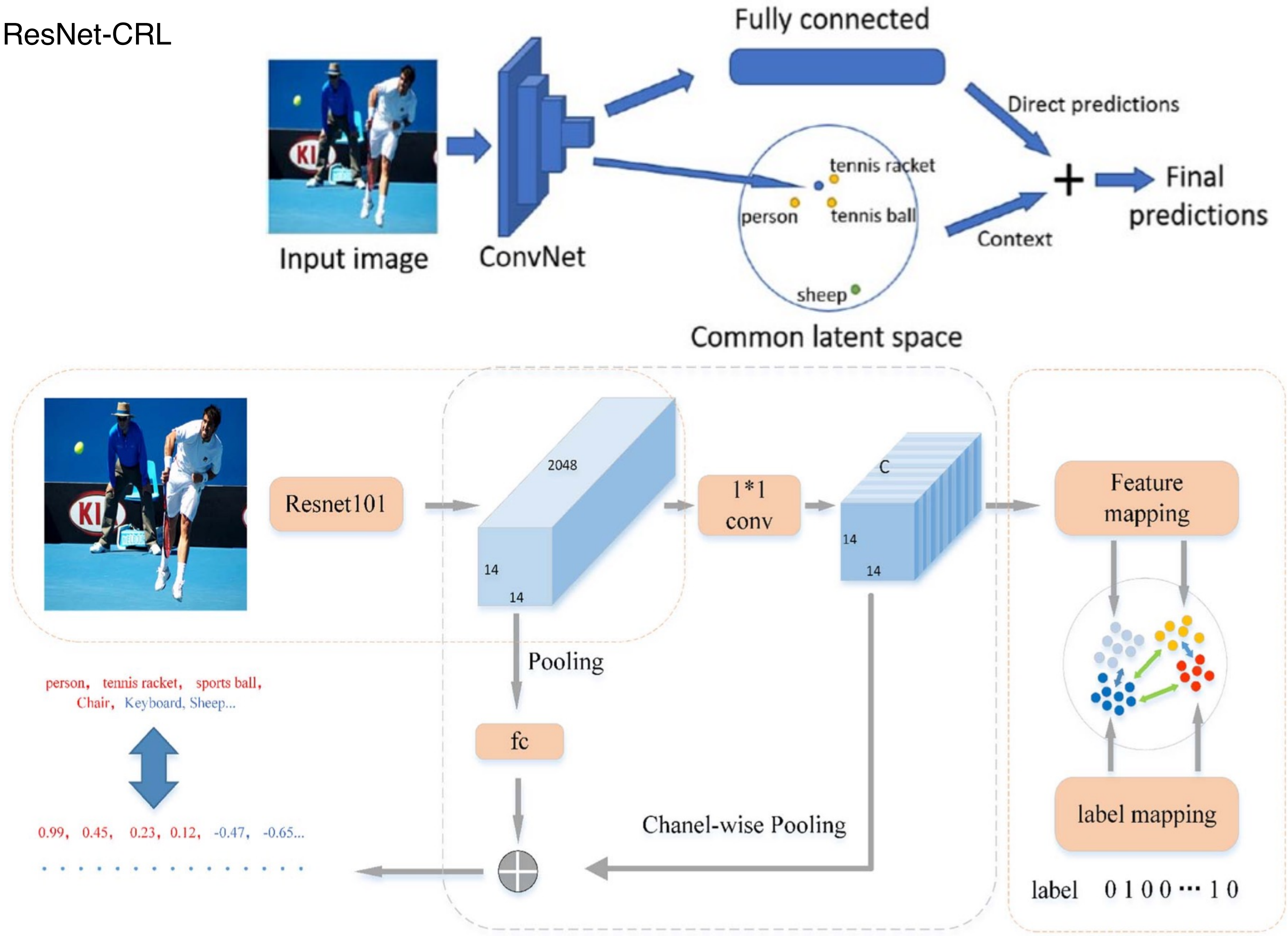
- Approaches that exploit label inter-dependencies —— Latent space

[1] C2AE



'L' represents 'Latent Space'

[2] ResNet-CRL



[1] C.-K. Yeh, W.-C. Wu, W.-J. Ko, Y.-C. F. Wang, Learning deep latent space for multi-label classification, in: Proc. AAAI Conf. Artif. Intell., Vol. 31, 2017.
[2] S. Wen, W. Liu, Y. Yang, P. Zhou, Z. Guo, Z. Yan, Y. Chen, T. Huang, Multilabel image classification via feature/label co-projection, IEEE Trans. Syst. Man Cybern.: Syst.

Introduction —— Multi-label image classification

- Existing Solutions
  - Traditional approaches
    a. Hand-crafted features, SIFT, GIST, HOG etc.
    b. SVM, Tree-based approaches, Bayesian etc.
  - Deep learning approaches
    a. Approaches that exploit label inter-dependencies
       a) RNN-based methods
       b) GNN-based methods
       c) Latent space

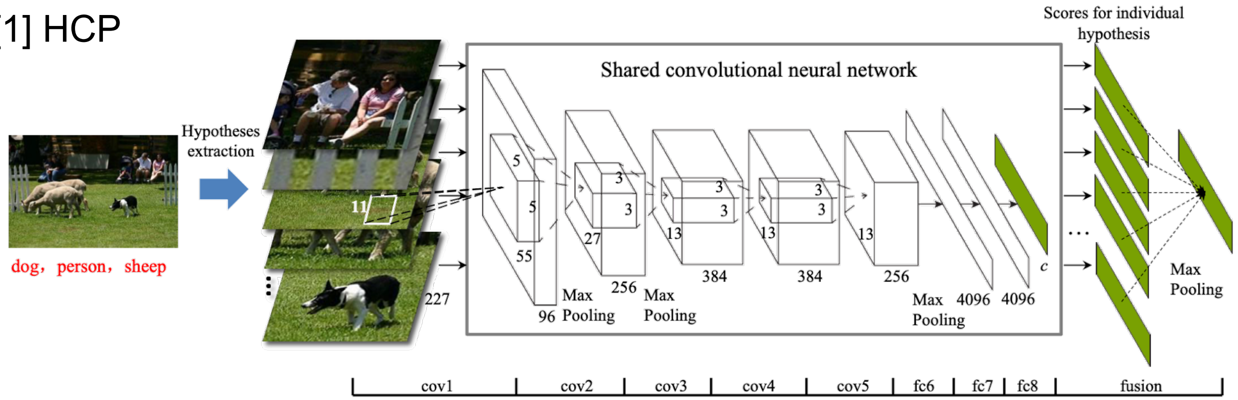
    b. 2-stage pipeline approaches that utilize the spatial information of objects (region proposal generation & region labeling)

Introduction —— Multi-label image classification

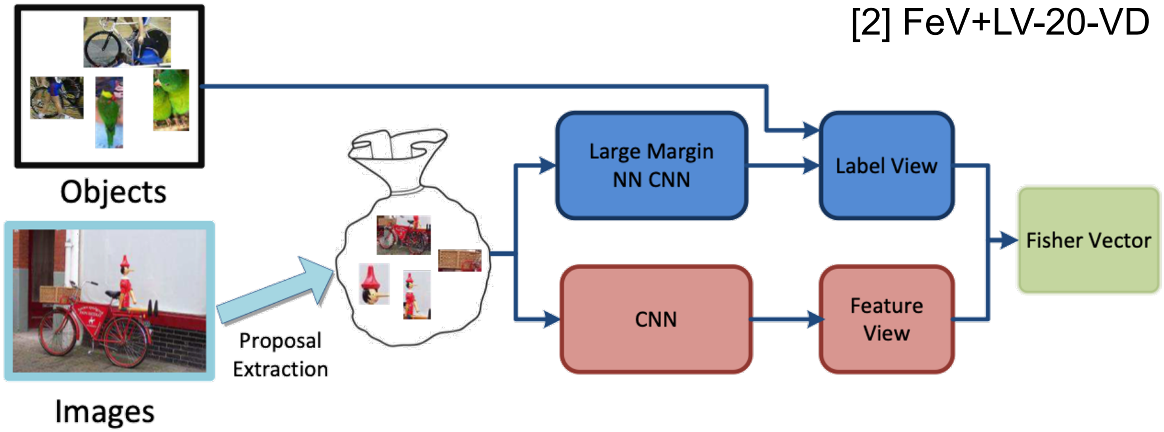
- 2-stage pipeline approaches

Region proposal generation & Region labeling

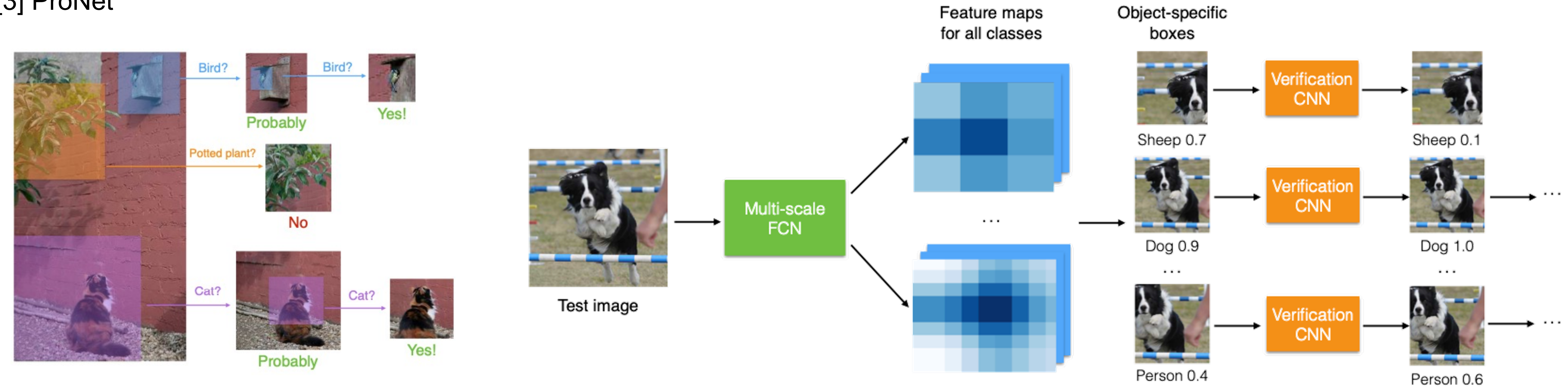
[1] HCP



[2] FeV+LV-20-VD



[3] ProNet

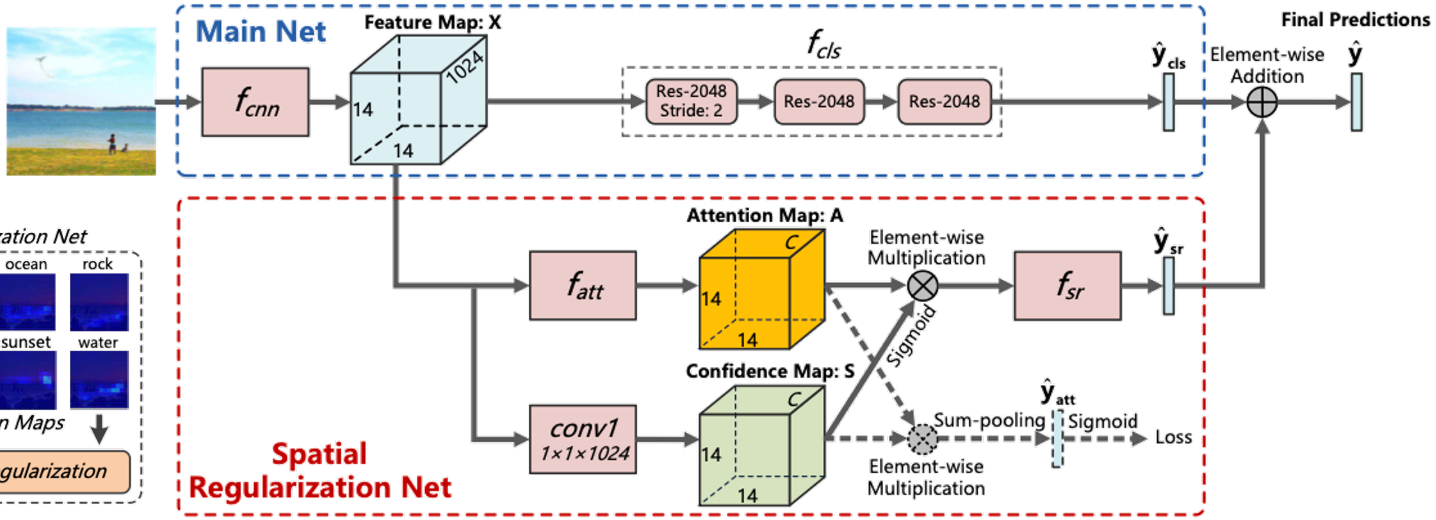


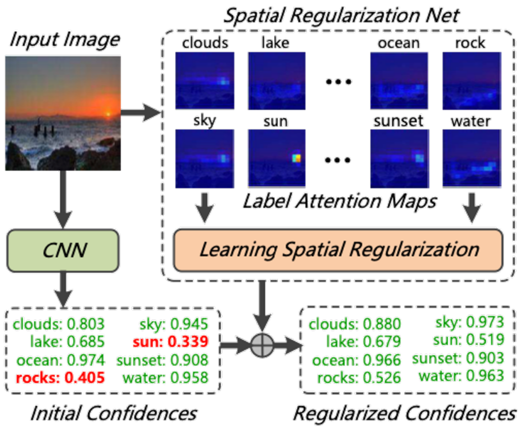
[1] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, S Yan, CNN: Single-label to multi-label, arXiv preprint arXiv:1406.5726.
[2] H. Yang, J. Tianyi Zhou, Y. Zhang, B.-B. Gao, J. Wu, J. Cai, Exploit bounding box annotations for multi-label object recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016.
[3] C. Sun, M. Paluri, R. Collobert, R. Nevatia, L. Bourdev, ProNet: Learning to propose object-specific boxes for cascaded neural networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,

Introduction —— Multi-label image classification

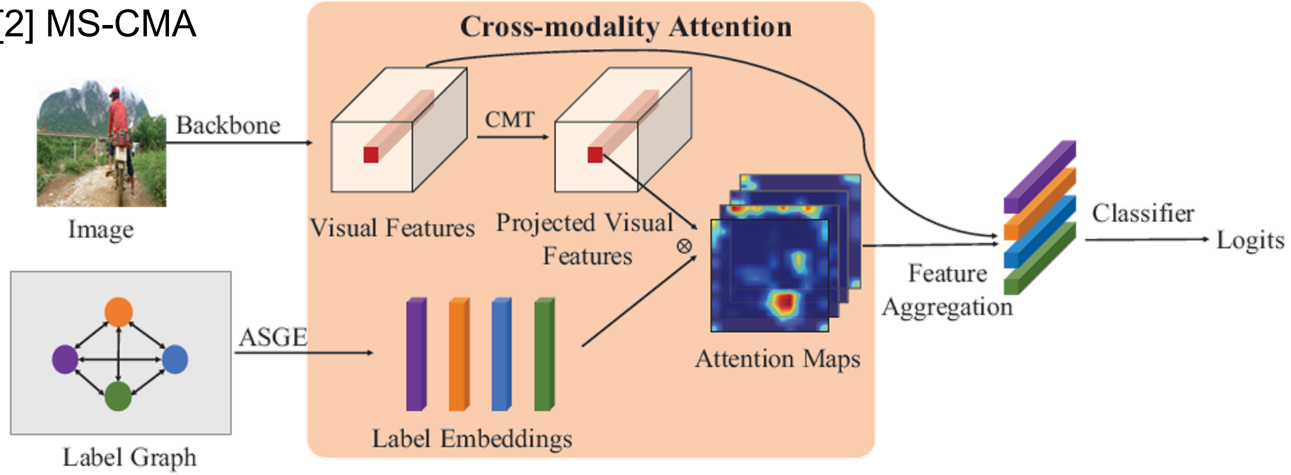
- 2-stage pipeline approaches – ‘attention map’ techniques

[1] ResNet-SRN





[2] MS-CMA

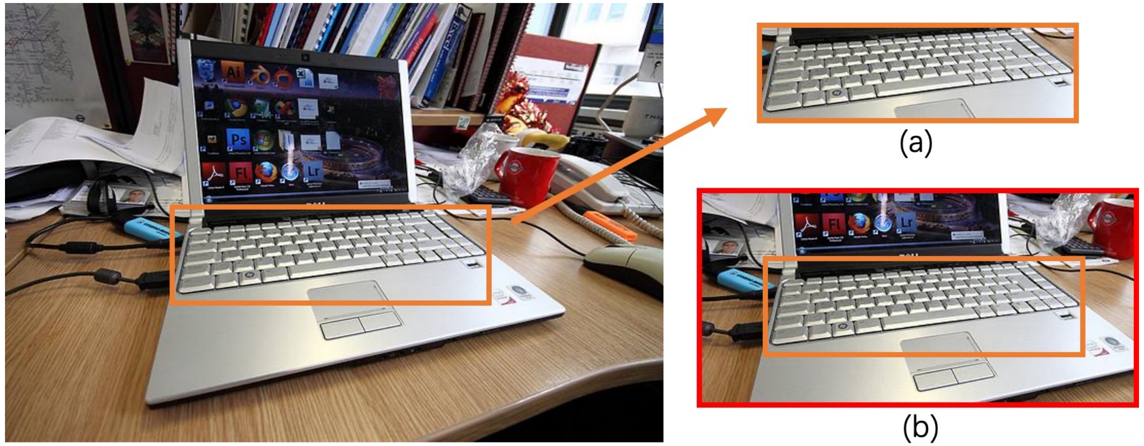


[1] F. Zhu, H. Li, W. Ouyang, N. Yu, X. Wang, Learning spatial regularization with image-level supervisions for multi-label image classification, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5513-5522.
[2] R. You, Z. Guo, L. Cui, X. Long, Y. Bao, S. Wen, Cross-modality attention with semantic graph embedding for multi-label classification, Proc. AAAI Conf. Artif. Intell. 34 (2020) 12709-12716.

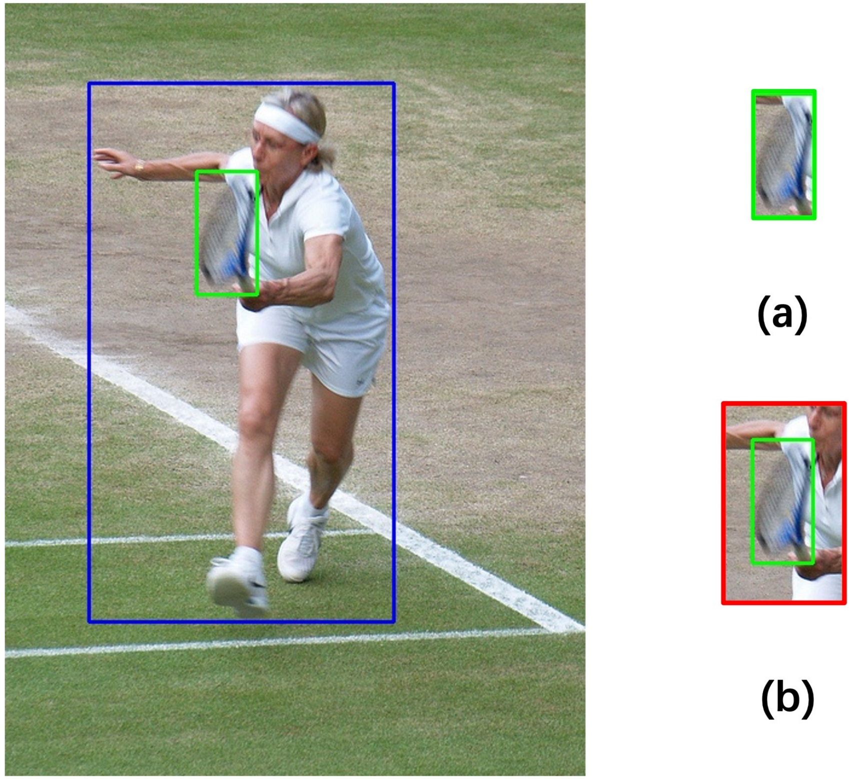
Introduction– multi-label image classification

Existing problem

The background context is considered harmful to the object detection due to the increase of the intra-class variations, and hence totally removed.



cup, laptop, mouse, **keyboard**, book



person, **tennis racket**

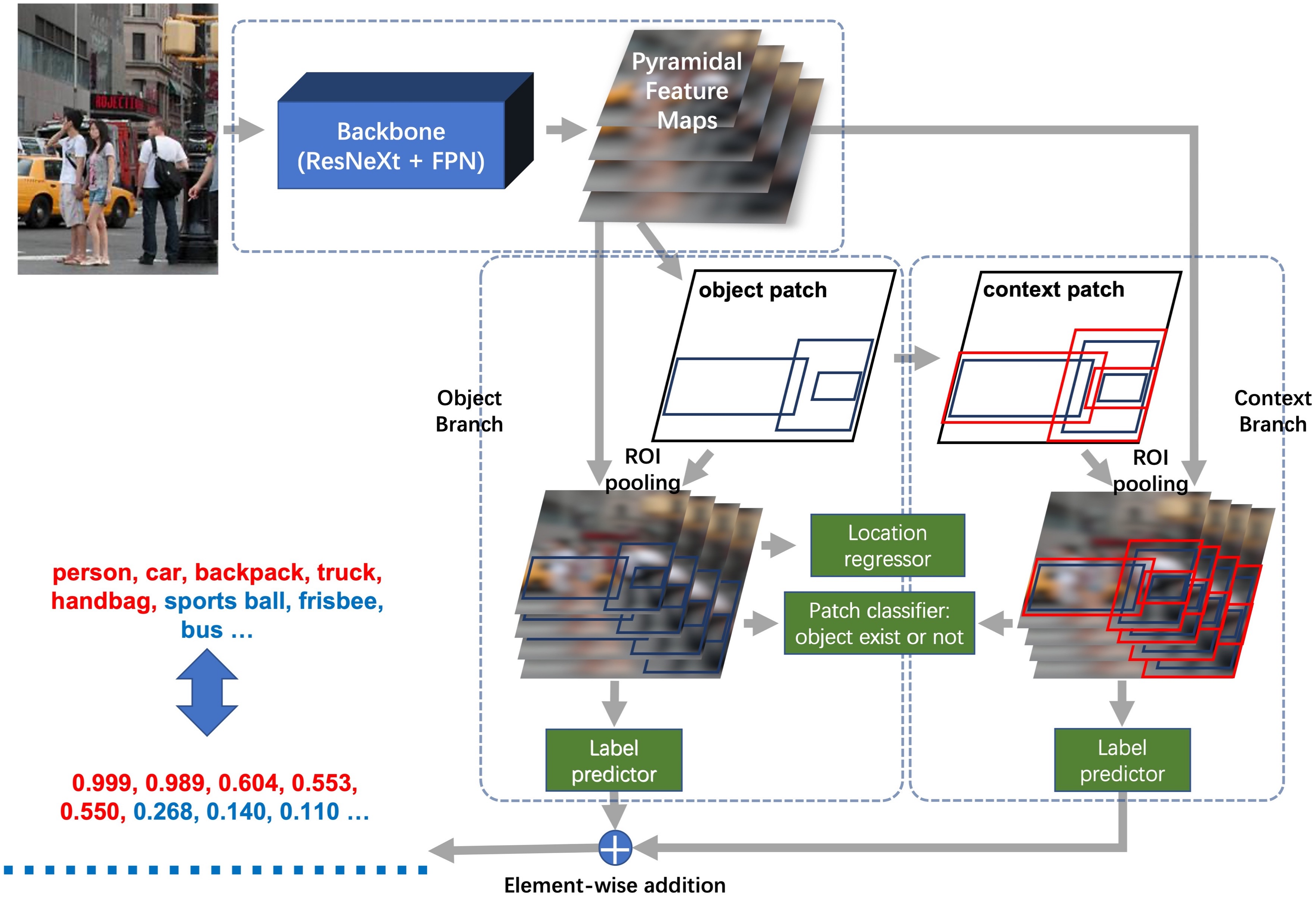
Contributions

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.

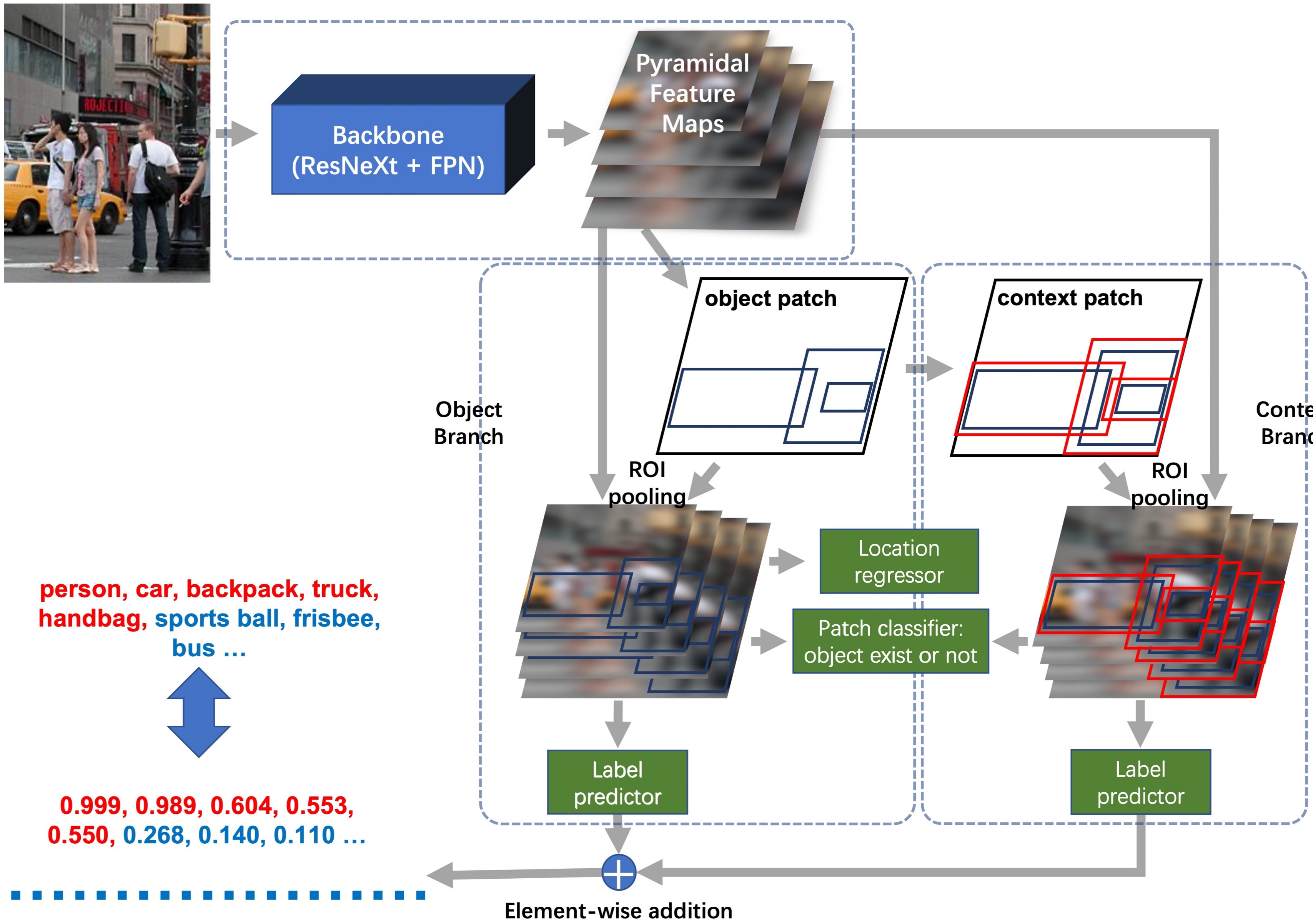
2. The proposed image-context-aware branch could well exploit both spatial and semantic information of objects.

3. The proposed approach significantly outperforms the state-of-the-art approaches on the MS-COCO dataset and PASCAL VOC dataset.

Spatial-context-aware deep neural network for multi-label image classification



Spatial-context-aware deep neural network for multi-label image classification

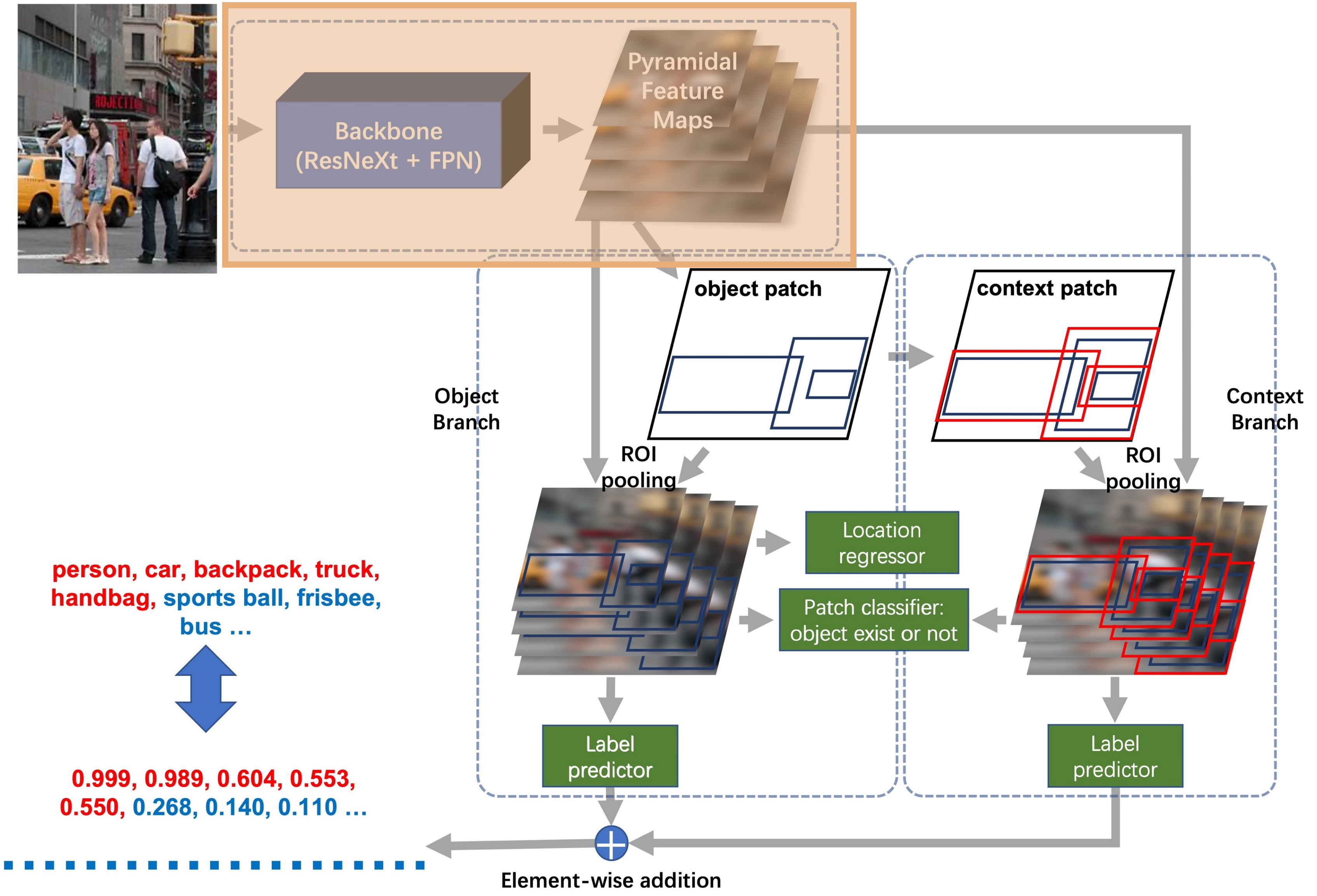


person, car, backpack, truck, handbag, sports ball, frisbee, bus …

0.999, 0.989, 0.604, 0.553, 0.550, 0.268, 0.140, 0.110 …

- A feature extractor
- Two patch generators
- Four patch processors

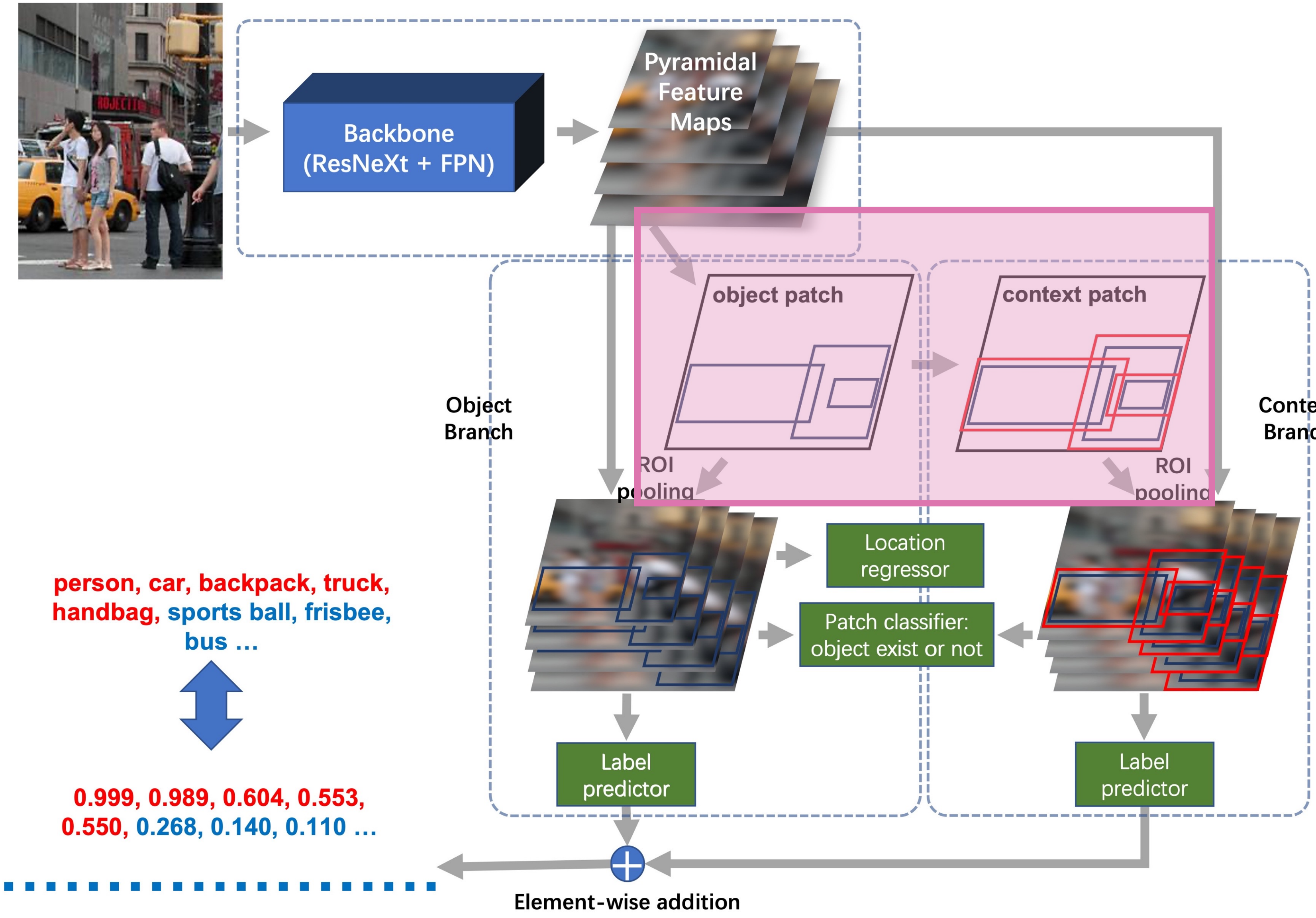
Spatial-context-aware deep neural network for multi-label image classification



ResNeXt-101 followed by FPN 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.

$$\mathbf{X} = f_F(f_R(\mathbf{I}; \theta_R); \theta_F)$$

Spatial-context-aware deep neural network for multi-label image classification

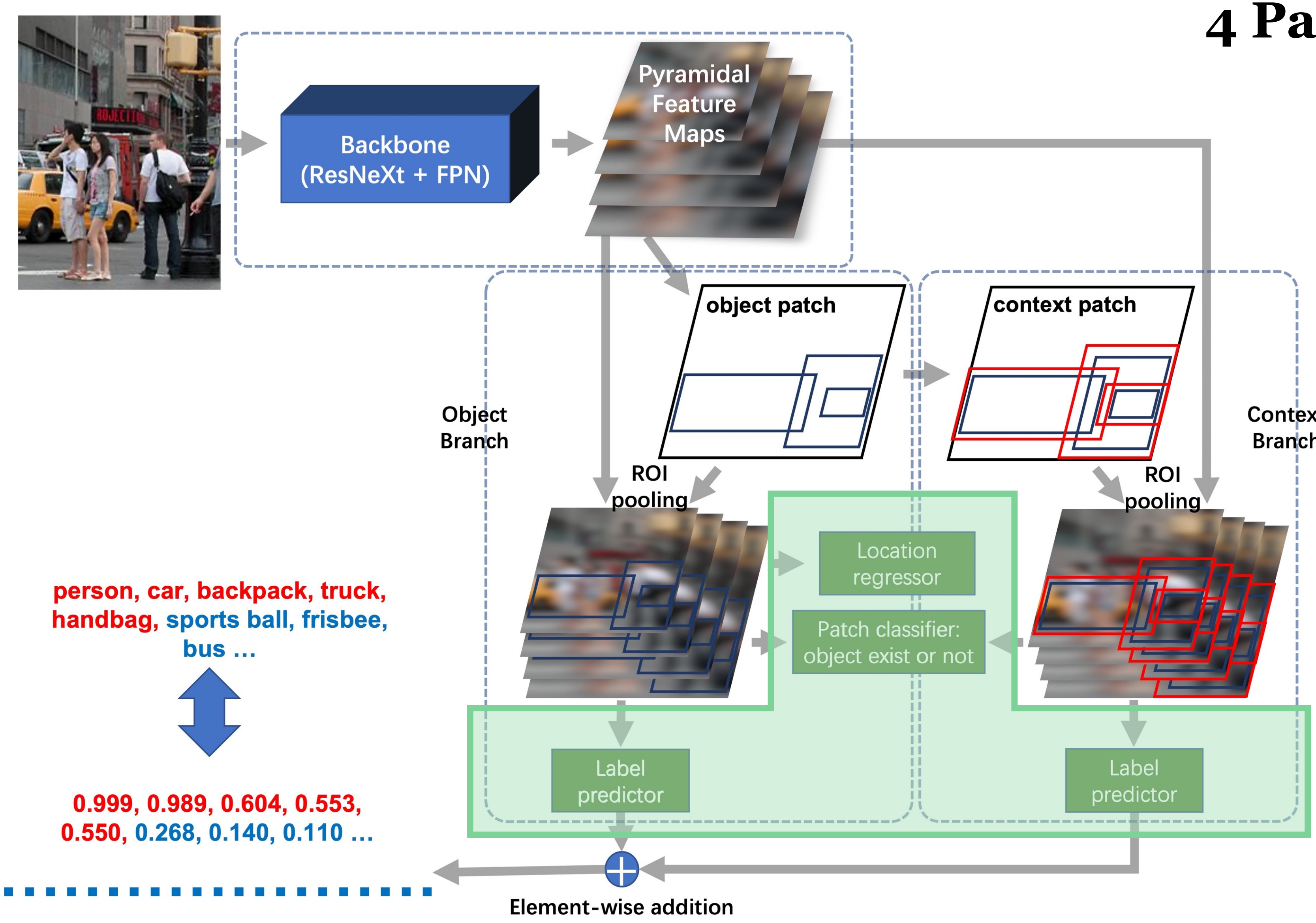


2 Patch Generators:

1. One is used to generate tightly cropped bounding boxes that contain the most discriminant information for classifying objects in the object branch

2. One is responsible for generating expanded image patches that containing both the object and additional contextual information in the context branch

Spatial-context-aware deep neural network for multi-label image classification



4 Patch Processors:

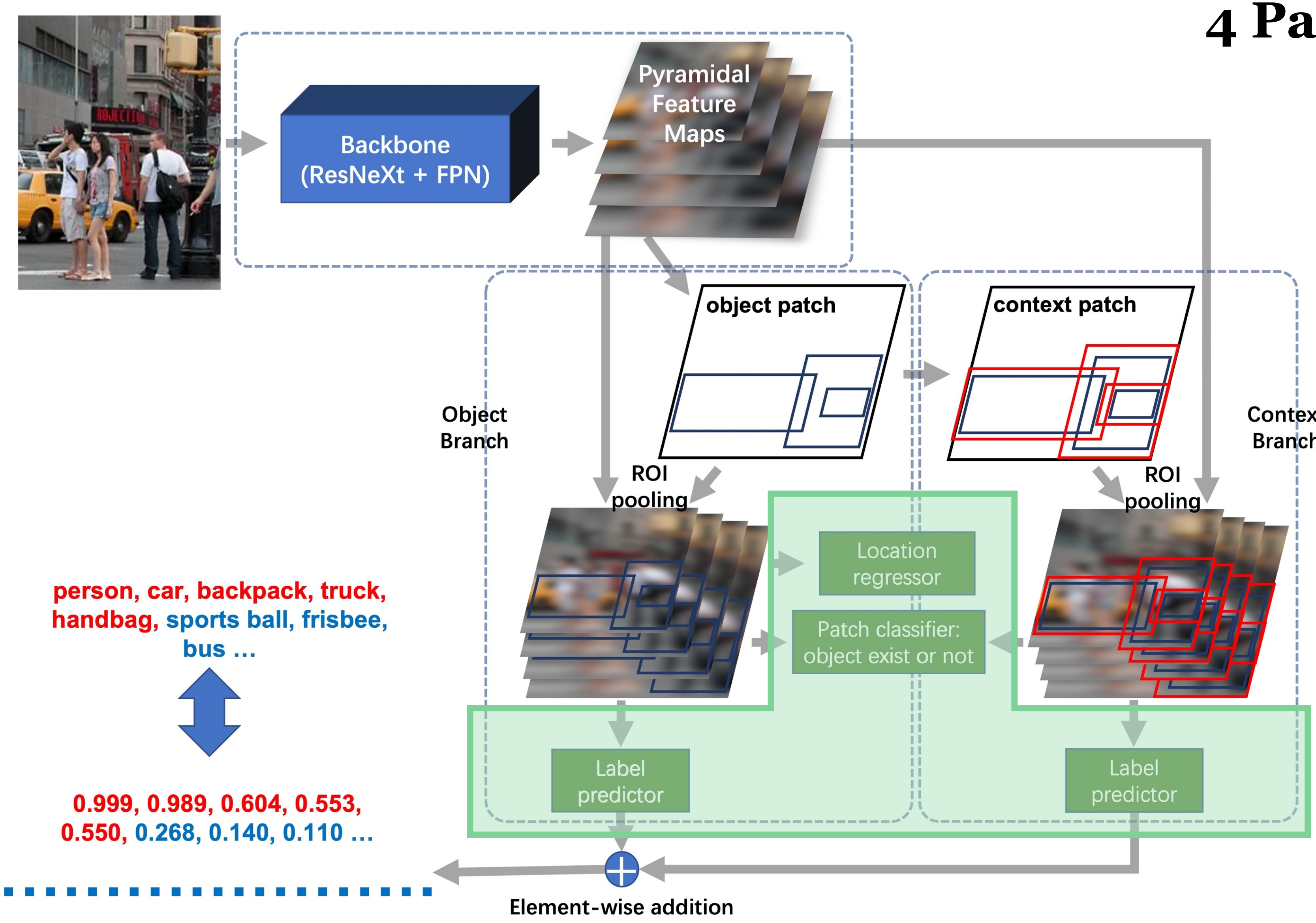
1. **Location regressor**, designed to accurately <u>locate the objects</u> to explore and utilize the image-level spatial information of different labels.

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.

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

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

Spatial-context-aware deep neural network for multi-label image classification



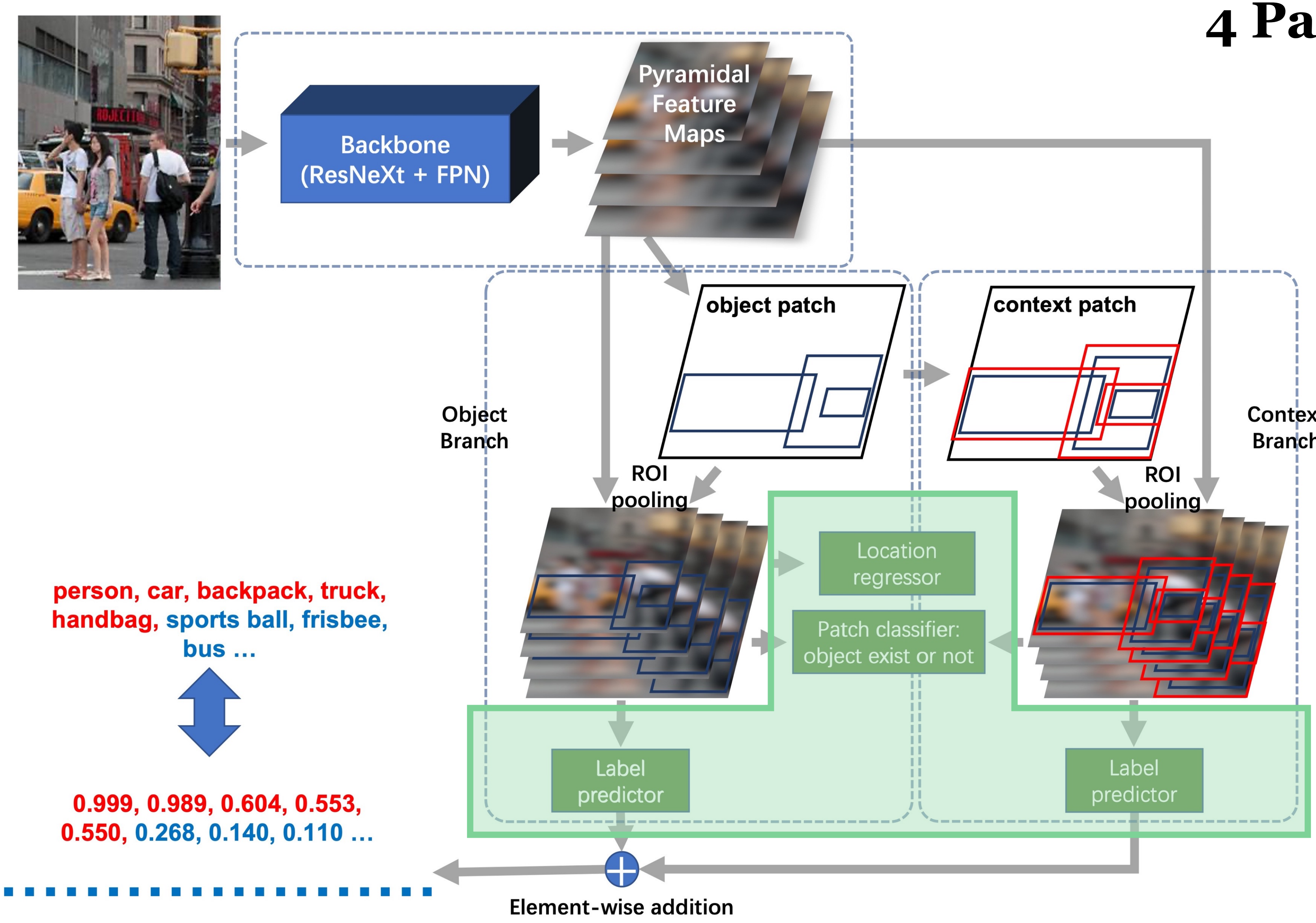
4 Patch Processors:

2. 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

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

- $\hat{p}$ - predicted confidence score
- p^* - ground-truth label

Spatial-context-aware deep neural network for multi-label image classification



4 Patch Processors:

1. **2 Label predictors**, designed 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

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

- $\hat{\mathbf{y}}$ - predicted score over all possible labels
- $\hat{y}^i$ - predicted score of the i-th category
- σ – weighting factor

Spatial-context-aware deep neural network for multi-label image classification —— Experiments

DATASET

- **Microsoft COCO 2017**
  - Train – 82,783
  - Test - 40,504
  - 80 categories
  - ≈ 2.9 labels/image
- **PASCAL VOC 2007**
  - Train/Val – 5011
  - Test – 4952
  - 20 categories
  - ≈ 1.4 labels/image

Evaluation Metrics

- Mean Average Precision (mAP)
- Macro Precision (P-C)
- Macro Recall (R-C)
- Macro F1 (F1-C)
- Micro Precision (P-O)
- Micro Recall (R-O)
- Micro F1 (F1-O)

Spatial-context-aware deep neural network for multi-label image classification —— Experimental Results on COCO

Method	mAP	F1-C	P-C	R-C	F1-O	P-O	R-O
CNN-RNN (CVPR, 2016)	61.2	60.4	66.0	55.6	67.8	69.2	66.4
ResNet101 (CVPR, 2016)	75.2	69.5	80.8	63.4	74.4	82.2	68.0
RNN-Attention (ICCV, 2017)	-	67.4	79.1	58.7	72.0	84.0	63.0
ResNet101-SRN (CVPR, 2017)	77.1	71.2	81.6	65.4	75.8	82.7	69.9
RNN-frequency (TMM, 2019)	64.7	-	-	-	-	-	-
DELTA (PR, 2019)	71.3	-	-	-	-	-	-
ResNet101-ACfs (CVPR, 2019)	77.5	72.2	77.4	68.3	76.3	79.8	73.1
DecoupleNet (ICASSP, 2019)	82.2	76.3	83.1	71.6	79.5	84.7	74.8
ML-GCN (CVPR, 2019)	83.0	78.0	85.1	72.0	80.3	85.8	75.4
ResNet101-CRL (TSMC-S, 2020)	81.1	75.8	81.2	70.8	78.1	83.6	73.3
KSSNet (AAAI, 2020)	83.7	77.2	84.6	73.2	81.5	87.8	76.2
MS-CMA (AAAI, 2020)	83.8	78.4	82.9	74.4	81.0	84.4	77.9
WSL-GCN (PR, 2021)	84.8	-	-	-	-	-	-
C-Tran (CVPR, 2021)	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

Spatial-context-aware deep neural network for multi-label image classification —— Experimental Results on PASCAL

Method	areoplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable
CNN-RNN (CVPR, 2016)	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0
ResNet101 (CVPR, 2016)	99.5	97.7	97.8	96.4	65.7	91.8	96.1	97.6	74.2	80.9	85.0
RNN-Attention (ICCV, 2017)	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92.0	87.7
RNN-frequency (TMM, 2019)	97.0	92.5	93.8	93.3	59.3	82.6	90.6	92.0	73.4	82.4	76.6
DELTA (PR, 2019)	98.2	95.1	95.8	95.7	71.6	91.2	94.5	95.9	79.4	92.5	85.6
ML-GCN (CVPR, 2019)	99.5	98.5	98.6	98.1	80.8	94.6	97.2	98.2	82.3	95.7	86.4
ResNet-CRL (TSMC-S, 2020)	**99.9**	98.4	97.8	**98.8**	81.2	93.7	97.1	98.4	**82.7**	94.6	87.1
WSL-GCN (PR, 2021)	99.7	98.5	**99.0**	97.8	86.2	96.2	98.3	**99.3**	81.1	95.9	88.0
The Proposed	99.4	**98.8**	98.0	98.6	**90.5**	**98.3**	**98.6**	98.4	81.3	**96.2**	**88.6**

	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
CNN-RNN (CVPR, 2016)	92.4	91.7	84.2	93.7	59.8	93.2	75.3	**99.7**	78.6	84.0
ResNet101 (CVPR, 2016)	98.4	96.5	95.9	98.4	70.1	88.3	80.2	98.9	89.2	89.9
RNN-Attention (ICCV, 2017)	96.8	97.5	93.8	98.5	81.6	93.7	82.8	98.6	89.3	91.9
RNN-frequency (TMM, 2019)	92.4	94.2	91.4	95.3	67.9	88.6	70.1	96.8	81.5	85.6
DELTA (PR, 2019)	96.7	96.8	93.7	97.8	77.7	95.0	81.9	99.0	87.9	91.1
ML-GCN (CVPR, 2019)	98.2	98.4	96.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0
ResNet-CRL (TSMC-S, 2020)	98.1	97.6	96.2	98.8	83.2	96.2	84.7	99.1	93.5	93.8
WSL-GCN (PR, 2021)	**99.2**	**98.6**	97.1	**99.4**	85.0	**97.5**	84.3	99.0	94.0	94.7
The Proposed	96.7	**98.6**	**99.0**	99.3	**87.0**	**97.5**	**87.3**	98.6	**95.7**	**95.3**

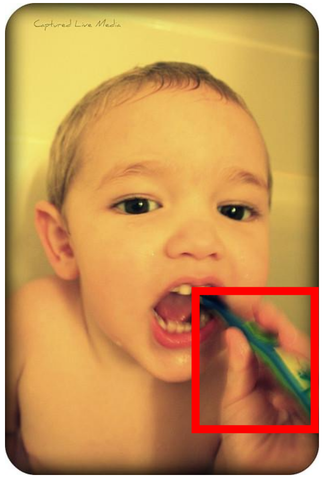
Spatial-context-aware deep neural network for multi-label image classification —— Qualitative Results

person, **bottle**

bottle, dining table, chair, **potted plant**

person, parking meter, umbrella, **dining**

person, boat, dog



person, **toothbrush**

bottle, **knife**, spoon, sandwich, dining table

person, **handbag**, **teddy bear**

bottle, dining table, person

Spatial-context-aware deep neural network for multi-label image classification —— reflection and further study

1. The trade-off between adding contextual information and increasing intra-class variations need to be balanced.

2. Context information is important for multi-labeling, however, only the object-to-background context is utilized in this work. Relations between objects, i.e., the object-to-object information, is not fully exploited.

University of Nottingham
UK | CHINA | MALAYSIA

THANK YOU