Iran University of Science and Technology
Deep Learning (Spring 2019)

Weakly Supervised Instance Segmentation using Image Level Annotations

Arman Ali Mohammadi
Department of Computer Engineering
Iran University of Science
and Technology
ali_ar@comp.iust.ac.ir

Abstract

In this paper, we propose an approach to boost current state-of-the-art architecture of weakly supervised instance segmentation with the use of an special attention mechanism which simulates the notion of biological "top-down" "bottom-up" attention systems. Weakly supervised instance segmentation reduce labeling cost with loosing a little accuracy in segmentation. But we show that attention mechanisms can improve these results. In this work Residual Attention Network is used in combination with Peak Response Map method and achieved better results compared to original PRM.

1 Introduction

Instance segmentation has been an area of research for decades, trying to assign a distinct object of interest to an image's pixel [1]. The main approaches to this problem need pixel level annotated ground-truth which is very costly or even impossible in some cases (for example brain tumor which ground-truth labels provided by the high quality specialists are about 75 percent precise). Weakly supervised instance segmentation works with coarser labels and thus it is much more cheaper thus more desirable in real world application, but current state-of-the art approaches have limited functionalities that causes accuracy drop for around 5-10 percent depending on dataset.

Instance segmentation is challenging itself, because it needs to distinguish individual object instances in a scene and this individual object or the labeling convention of an specific dataset can widely differ from one another. For example one dataset labels hammer as a handle and a piece of metal separately while others label the hammer as a whole. Another challenge just like almost all computer vision tasks is occlusion, which occurs when objects are hiding partially behind another object [1]. Weakly supervised instance segmentation have an extra challenge of poor labeling. In other words we expect the system to learn and produce an output that we have never taught to it directly, thus we need to add more complicated learning mechanisms so that the system be able to extract useful information by itself.

Having only image level class labels, state-of-the-art PRM method leverages natural class aware feature maps extracted from top layers of a convolutional neural network. These feature maps are like heat maps specifying region of presence of an important feature, which must be an area containing discriminative features of object network trained to classify. Since these feature maps are not intended to specify the region of presence of an object of interest, instance segmentation using these features is not promising. Therefore some tricks such as peak stimulation are used to force

(LateX template borrowed from NeurIPS 2019)

the network to learn discriminative feature maps but still PRM lakes a meaningful method to learn reliable instance aware feature maps. In this work we are going explore these weaknesses and try to find solutions for such problems.

2 Related work / Background

PRM Method

Using image level supervision for solving instance segmentation problem is becoming an area of interest in computer vision research. Some approaches has been taken for dealing with this problem [2, 3, 4, 5, 6] among these we found PRM method a more reasonable and reliable method which utilizes natural capabilities of CNN networks, also more similar to biological alternative of the system (human visual system). Here we try to take an overall glimpse of this method.

For CNN we know that in each level of network a convolution filter can represent and extract visual cues from image. Also it has been discovered that the result of performing these filters gives feature maps that object activations are apparent in them. Main idea of PRM method [3] is to leverage these feature maps named Class Response Maps (CRM) extracted from last layer of a convolutional neural network. in this paper a simple CNN classifier such as VGG or ResNet is trained with image level class labels showing an objects presence in the image. their method consists of three main parts.

1. Fully Convolutional Architecture

   To prevent loss of spatial information, dense classifier at the end of the network is converted to a FCN architecture [7] that naturally preserve spatial information throughout the forwarding. therefore training this network for classification can capture visual cues and represent instance aware feature maps.

2. Peak Stimulation

   To stimulate peaks to emerge from class response maps a method similar to non-maxima suppression is used. This method like a pooling layer can be implemented as a layer in conventional deep-learning frameworks. Having a single CRM (feature map of top convolutional layer) peak stimulation is performed by sliding a window over the feature map. Peaks of a response map are defined to be the local maximums within this window and the location of peaks are denoted as (j, j). Also during the forward pass a sampling kernel is performed to compute classification confidence scores for each class response map corresponding to an object class. this kernel is as follows:

   $$G^c_{x,y} = \sum_{k=1}^{N^c} f(x - i_k, y - j_k)$$

   where (j_k, j_k) is the coordinate of the k-th peak, and f is a sampling function (in this setting f is Dirac delta function that aggregates features from peaks only). So the network uses peaks only to make decision by averaging valid peaks of each class response map computing confidence scores. it can be seen that the output of this layer will be confidence scores for each class.

   [my contribution to this part is using a dynamic method to select valid class threshold and peak threshold]

3. Peak Back-propagation

   To generate fine detailed and instance aware representations a probability back-propagation process is proposed. this process can be interpreted as a procedure that a walker starts from the peak and walks randomly to the bottom layer. The top-down relevance of each location in the bottom layer is formulated its probability of being visited by the walker. with the probability back-propagation we can localize most relevant spacial locations for each class peak response to generate fine-detailed instance-aware visual cues referred to as Peak Response Maps (PRM).

Attention

'Attention' mechanisms have become a significant area of research and attracted many researchers attention in recent years [8, 9, 10, 11]. Among these we found RAN a more reasonable method and more similar to biological alternative of the notion (human visual system). In the following sections

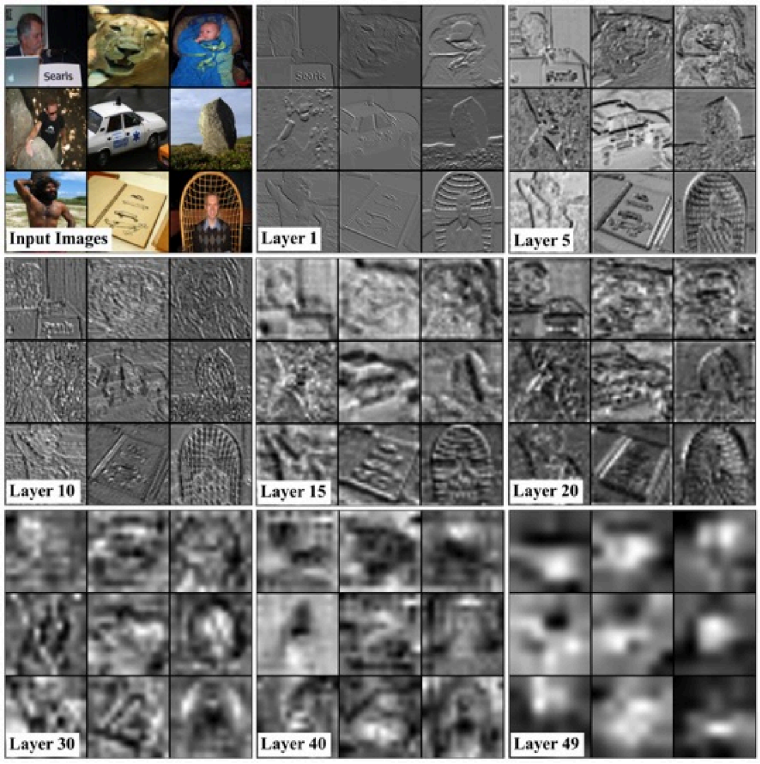

Figure 1: Visualization of example features of layers 1, 10, 20, 30, 40, and 49 of a deep convolutional neural network [15].

we firstly try to grasp some understanding of the notion then demonstrate selected approach and bring justifications for our selection.

Top-Down Bottom-up Attention

In human visual system, attention can be focused volitionally by "top-down" signals derived from task demands and automatically by "bottom-up" signals from salient stimuli. Volitional shifts of attention are taught to depend on "top-down" signals derived from knowledge about current task (e.g. finding your lost keys) whereas automatic "bottom-up" capture of attention is derived by properties inherent in stimuli that is by salience (e.g. a moving object) [12, 13].

In computer vision we use similar terminology and refer to attention mechanisms derived by task specific context (non-visual) as "top-down", and purely visual feed-forward attention mechanisms as "bottom-up" [14]. The "bottom-up" visual attention is triggered by stimulus, where a saliency is captured as the distinction of image locations, regions or objects in terms of low level cues such as color, intensity, orientation, shape, T-conjunctions, X-conjunctions, etc [15]. As we know CNNs are able to extract visual features from an image in different layers (Figure 1). This feature extraction mechanism is task independent and entirely from visual cues. a simple feed forward process of CNN network can mimic "bottom-up" attention of human visual system. The "top-down" visual attention uses prior knowledge, expectations or rewards as high level visual factors to identify the target of interest [15, 11]. A Fully Convolutional Network trained for a specific classification task must be able to attend to most informative regions of an image for that specific task. Without loss of spacial information in FCN architecture, network feedback (by back-propagation) can somehow simulate human visual cortex mechanism in which is enhanced by "top-down" attention stimuli feedback so that non-relevant neurons will be suppressed in feedback loops when searching for objects that leads to selectivity in neuron activations [16].

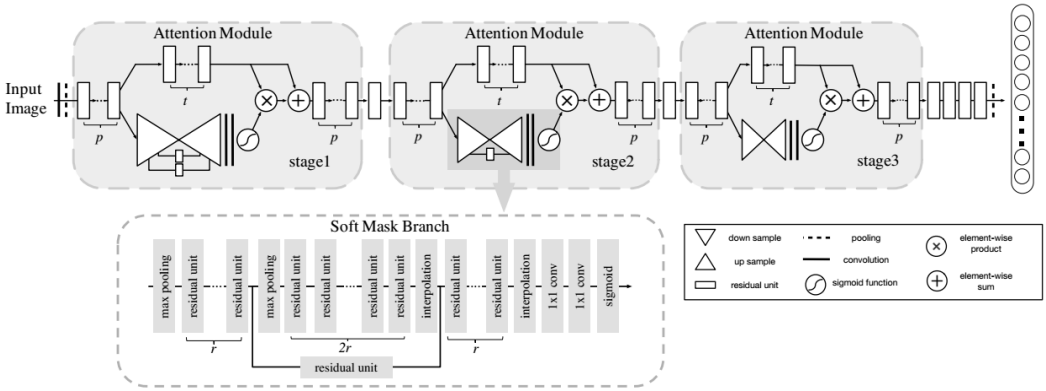

Figure 2: An example architecture of Residual Attention Network [17].

Residual Attention Network

Residual Attention Network (RAN) is a convolutional neural network using attention mechanism built by stacking attention modules which generate attention aware features. Inside each attention module, "bottom-up" "top-down" feedforward structure is used to unfold the feedforward and feedback attention process into a single feedforward process. In the RAN method "bottom-up" "top-down" feedforward structure is used as part of attention module to add soft weights to features of a residual network block. this structure can mimic "bottom-up" feedforward process and "top-down" attention feedback in a single feedforward process which allows development of an end-to-end trainable network with "top-down" attention [17].

RAN is constructed by stacking multiple attention modules (Figure 2). Each attention module is divided into two branches: mask branch and trunk branch. trunk branch performs feature processing and can be adapted to any state-of-the-art network structure. In this work pre-activation Residual Unit is used in which in the main branch of residual block Bach-Normalization and ReLU activation are performed before Convolution layers and ReLU activation after skip-connection at the end of the block is removed. Given trunk branch output $T(x)$ with input x, the mask branch uses "bottom-up" "top-down" structure to learn same size mask $M(x)$ that is used to softly weight output features of trunk branch $T(x)$. Having this output weighting procedure, the output masks act like control gates for neurons of trunk branch (Figure 3). It can be seen that this procedure acts like how we described the process and goal of "top-down" attention mechanism.

In Attention Modules, the attention mask can not only serve as a feature selector during forward inference, but also as a gradient update filter during back-propagation. This property makes Attention Module robust to noisy labels. Mask branches can prevent wrong gradients from noisy labels to update trunk parameters. In Attention Module, each trunk branch has its own mask branch to learn attention that is specialized for its features.

3 Proposed method

In this work we use Residual Attention module for instance aware class response map extraction used by PRM method as this method has a meaningful approach to learn reliable instance aware feature maps by simulating "top-down" and "bottom-up" attention mechanism of human visual system.

To do this we first convert RAN architecture to a fully convolutional architecture such that spacial information can be preserved during training process. We also use Peak Stimulation technique to amplify networks attendance to objects of interest. After training for a multi label image classification task, feature maps of top layer of network can be used to create PRMs.

In main approach of PRM method, some constant thresholds are used for valid class response map and then valid peaks selection. These thresholds are chosen experimentally for a specific task. We tried to solve this issue by exploring distribution of values of class response map aggregation

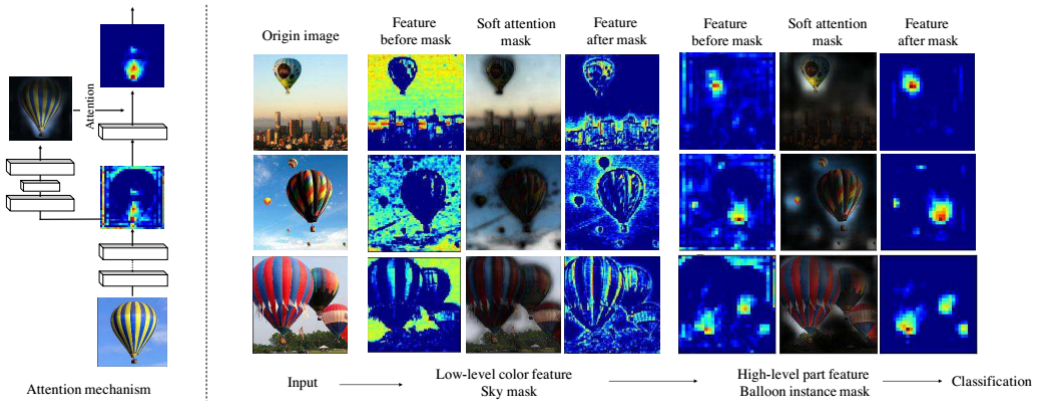

Figure 3: An example of how attention control gates can lead to clearer and better object aware feature maps [17].

and peaks of an individual map and found third quartile a meaningful criterion for choosing valid class response maps in which CRMs with higher than 75% show confidence higher than 75% about presence of an object belonging to a specific class. For peak selection also third quartile is useful since most peaks extracted by local maximum extraction window are not valid peaks and peaks take place in the upper quartile. But we achieved same results using mean of peak values for original PRM architecture so we found mean value a good threshold for our purpose. We use same technique in our method.

We used probability back-propagation method proposed in [3] to generate Peak response maps. Each peak is back-propagated through network and gradients reached to the bottom layer of network are PRM of that peak. We examined two gradient back-propagation paths, first we used entire network, then we removed soft mask branches from gradient back-propagation graph. No significant difference was found (although 0.2 of dataset was used for training RA network).

4 Results

Dataset used for this work is PASCAL VOC 2012 with twenty classes for classification task. We used first 20 percent of data, 1 percent of it left off for validation during training classifier.

Since PRM gives dense peak response maps not segmentations exactly and to get segmentations a combination of these PRMs with an off-the-shelf object proposal method is needed therefore we skip this extra part and do a qualitative comparison [3, 18].

All image inputs resized to 448*448 and normalized. For training phase, random horizontal flip with probability of 0.5 was performed as data augmentation. Model hyper parameter configuration for training was as follows. SGD with 0.01 learning rate, 1.0e-4 weight decay and 0.9 momentum is used. Multi Label Soft Margin Loss is used for training loss calculation [19, 20]. We used 0.2 of dataset was used for training with 0.1 of this left off for validation (it was not possible for us to use whole dataset also default hyper parameters was used since we couldn't explore hyper parameter space for parameter selection). We showed these data to our model with batches of size 8 for 4 epochs. The results comparing to original PRM model (having original PRM, used pre-trained ResNet on image net and fine-tuned whole model on entire VOC dataset) are shown in Figure 4. As we can see has improved result Peak Response maps in which boundaries of the object are extracted. Despite not optimized and not fully trained model due to computational and time limit, our approach performs better in some cases (Figure 5) compared to original PRM method.

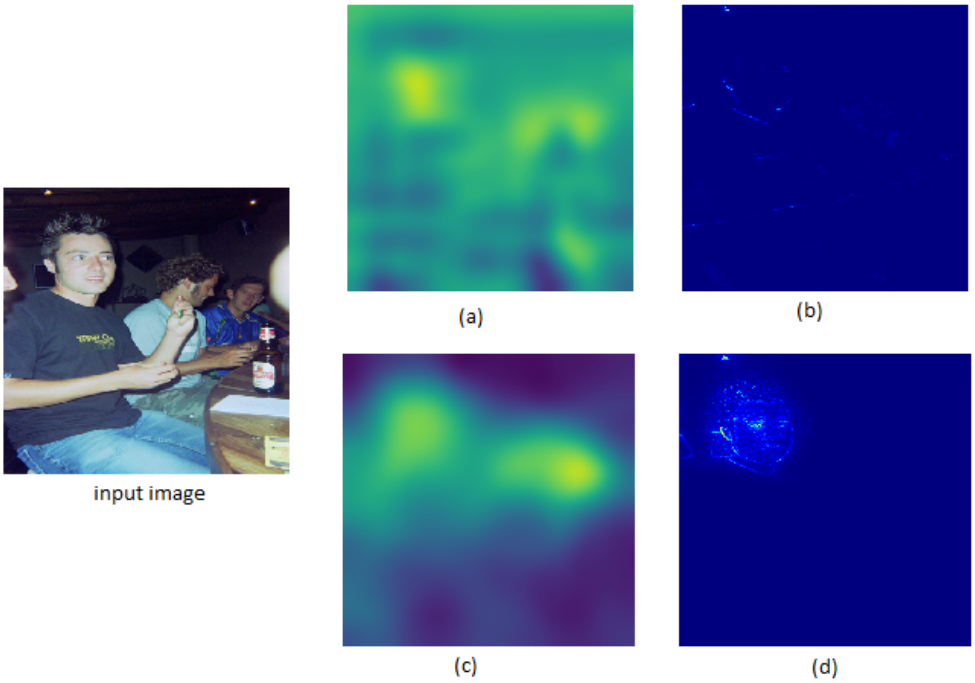

Figure 4: a) Class Response map for attention augmented PRM. b) Peak Response map for attention augmented PRM. c) Class response map for original PRM. d) Peak Response map for original PRM.

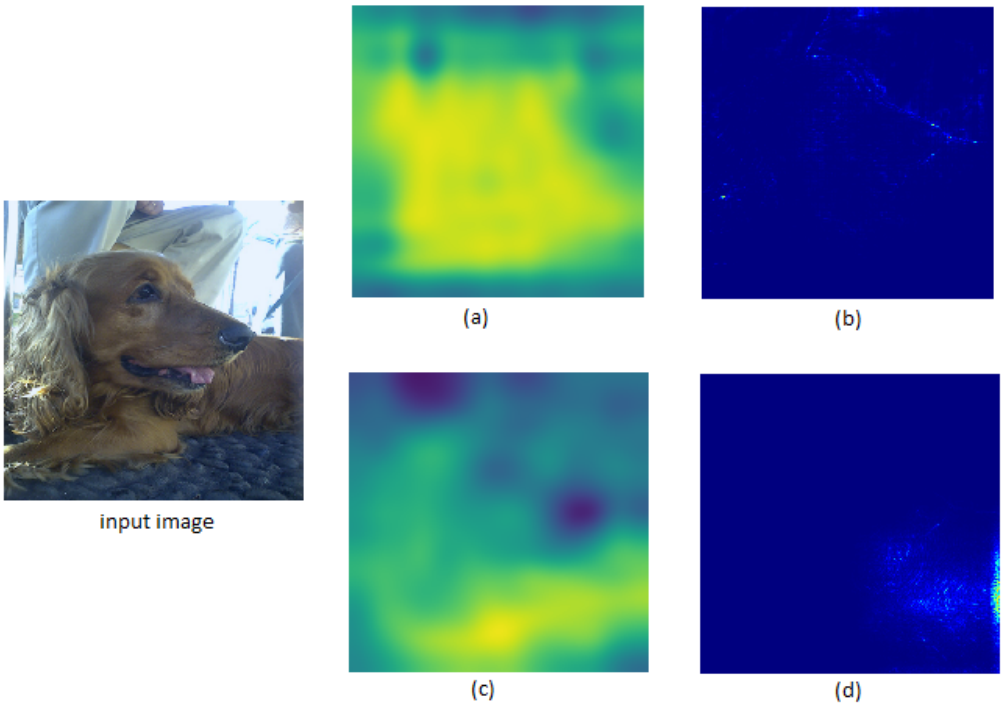

Figure 5: a) Class Response map for attention augmented PRM. b) Peak Response map for attention augmented PRM. c) Class response map for original PRM. d) Peak Response map for original PRM.

5 Discussion

In this work we explored weakly instance segmentation with novel method. Our approach have better result in the same learning time. This method is capable of extract fine detailed object boundaries from an image unlike the original method which just highlights an area of the object itself. We achieve this by using "bottom-up" "top-down" attention module which focus on the important areas (like borders and object properties) using visual only beside task-specific information. We achieved qualitatively better results compares to original PRM.

References

[1] M. Ren and R. S. Zemel, "End-to-end instance segmentation with recurrent attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6656–6664, 2017.

[2] H. Cholakkal, G. Sun, F. S. Khan, and L. Shao, "Object counting and instance segmentation with image-level supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12397–12405, 2019.

[3] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao, "Weakly supervised instance segmentation using class peak response," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3791–3800, 2018.

[4] S. Liao, Y. Sun, C. Gao, P. S. KP, S. Mu, J. Shimamura, and A. Sagata, "Weakly supervised instance segmentation using hybrid networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1917–1921, IEEE, 2019.

[5] J. Ahn, S. Cho, and S. Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2209–2218, 2019.

[6] W. Zhang, S. Zeng, D. Wang, and X. Xue, "Weakly supervised semantic segmentation for social images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2718–2726, 2015.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.

[8] P. Rodríguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, and X. Roca, "A painless attention mechanism for convolutional neural networks," 2018.

[9] X. Liu, T. Xia, J. Wang, Y. Yang, F. Zhou, and Y. Lin, "Fully convolutional attention networks for fine-grained recognition," *arXiv preprint arXiv:1603.06765*, 2016.

[10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[11] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018.

[12] T. J. Buschman and E. K. Miller, "Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices," *science*, vol. 315, no. 5820, pp. 1860–1862, 2007.

[13] F. Katsuki and C. Constantinidis, "Bottom-up and top-down attention: different processes and overlapping neural systems," *The Neuroscientist*, vol. 20, no. 5, pp. 509–521, 2014.

[14] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6077–6086, 2018.

[15] A. Mahdi, J. Qin, and G. Crosby, "Deepfeat: a bottom-up and top-down saliency model based on deep features of convolutional neural nets," *IEEE Transactions on Cognitive and Developmental Systems*, 2019.

[16] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, *et al.*, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2956–2964, 2015.

[17] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164, 2017.

[18] Y. Zhu, Y. Zhou, H. Xu, Q. Ye, D. Doermann, and J. Jiao, "Learning instance activation maps for weakly supervised instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3116–3125, 2019.

[19] M. Lapin, M. Hein, and B. Schiele, "Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 7, pp. 1533–1554, 2017.

[20] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "Cnn: Single-label to multi-label," *arXiv preprint arXiv:1406.5726*, 2014.