南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

The 2nd LID Challenge (Weakly Supervised Object Localization)

Zhendong Wang, Zhenyuan Chen, Chen Gong
Nanjing University Of Science and Technology
LEAP Group@PCA Lab

Outline

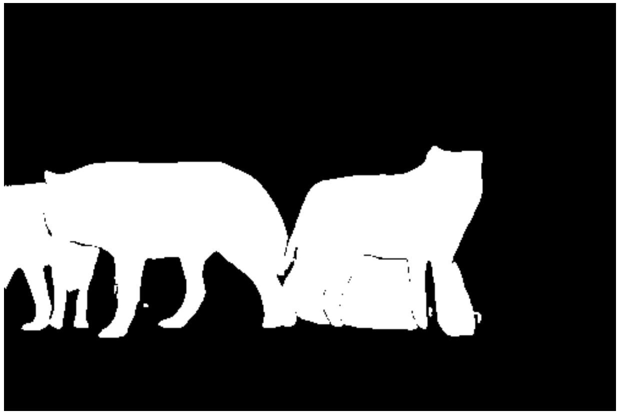
1 Problem Analysis

2 Model Description

3 Experimental Results

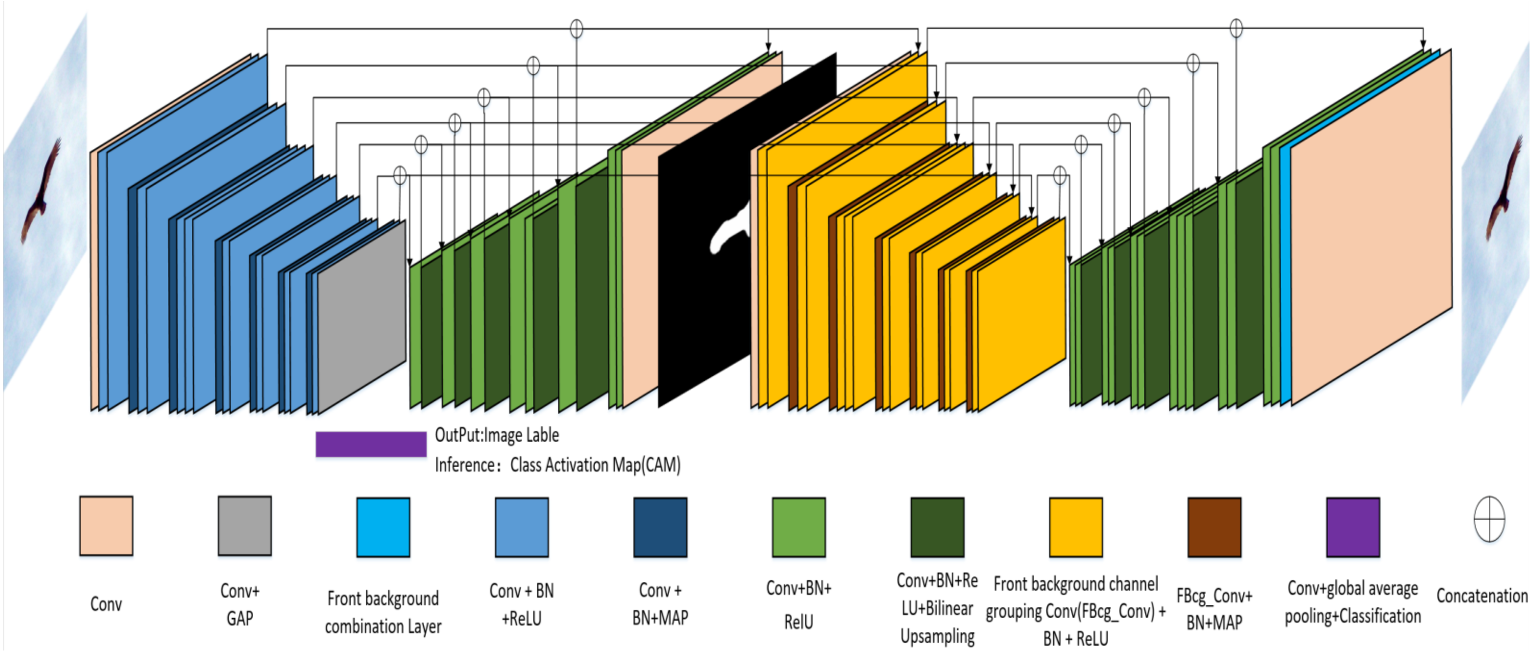
4 Future Improvements

Problem Analysis



- Weakly-Supervised Object Localization: localize the objects in an image only with image-level annotations.
- Only using image-level labels is too weak, so we try to utilize the information of raw image to promote the edge and other details.
- We designed a network composed of two auto-encoder parts.

Model Description


OutPut:Image Lable
Inference：Class Activation Map(CAM)
Conv
Conv+ GAP
Front background combination Layer
Conv + BN +ReLU
Conv + BN+MAP
Conv+BN+ RelU
Conv+BN+ReLU+Bilinear Upsampling
Front background channel grouping Conv(FBcg_Conv) + BN + ReLU
FBcg_Conv+ BN+MAP
Conv+global average pooling+Classification
Concatenation

- We designed a network with two auto-encoders.
- In the first part, we train a classifier with global average pooling under the supervision of image-level annotations. Then we use the binary images generated by CAMs as pseudo pixel-level annotations.
- In the second part, we expect to recover the raw image from binary image in order to get the refined binary image for the next iteration.

Model Description

Our loss function:

$$L = \sum_{k=1}^{K} a_k l^k + \frac{1}{2}\sum_{i=1}^{c}\left(\hat{y}_i - y\right)^2 + l_{out}$$

$$l_{out} = -\sum_{(r,c)}\left[I(r,c)\log\left(P(r,c)\right) + \left(1 - I(r,c)\right)\log\left(1 - P(r,c)\right)\right]$$

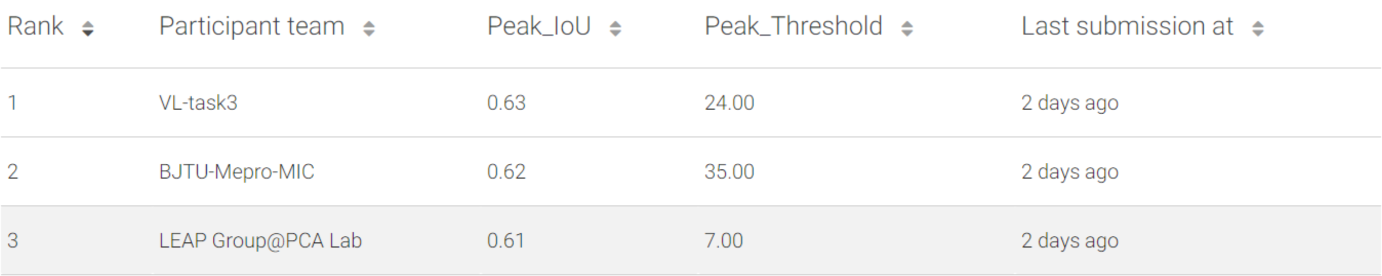
$$l^k = l_{bce}^k + l_{ssim}^k$$

$$l_{bce} = -\sum_{(r,c)}\left[G(r,c)\log\left(S(r,c)\right) + \left(1 - G(r,c)\right)\log\left(1 - S(r,c)\right)\right]$$

$$l_{ssim} = 1 - \frac{(2u_x u_y + 0.01^2)\left(2\sigma_{xy} + 0.03^2\right)}{\left(u_x^2 + u_x^2 + 0.01^2\right)\left(\sigma_x^2 + \sigma_y^2 + 0.03^2\right)}$$

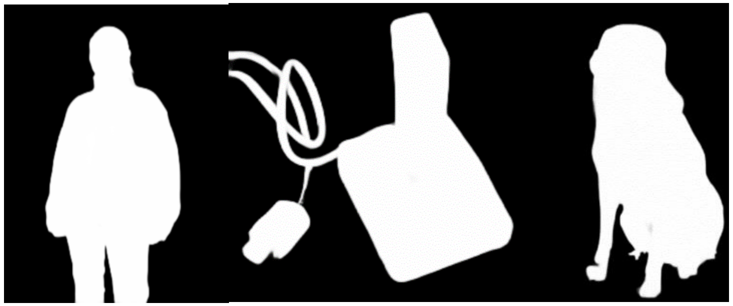
1. The loss l^k between binary image prediction and pseudo pixel-level annotations.
2. Mean square loss of class prediction.
3. Cross entropy loss l_{out} between the final output image and the raw image.

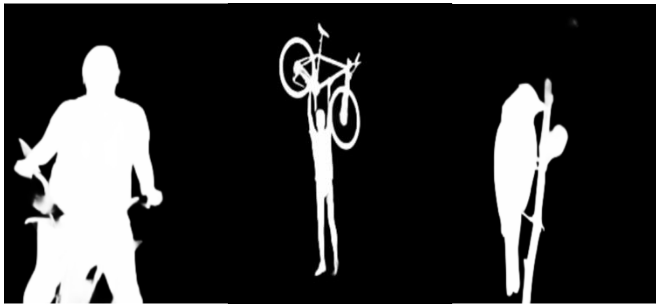
Experimental Results



Rank ⬍	Participant team ⬍	Peak_IoU ⬍	Peak_Threshold ⬍	Last submission at ⬍
1	VL-task3	0.63	24.00	2 days ago
2	BJTU-Mepro-MIC	0.62	35.00	2 days ago
3	LEAP Group@PCA Lab	0.61	7.00	2 days ago

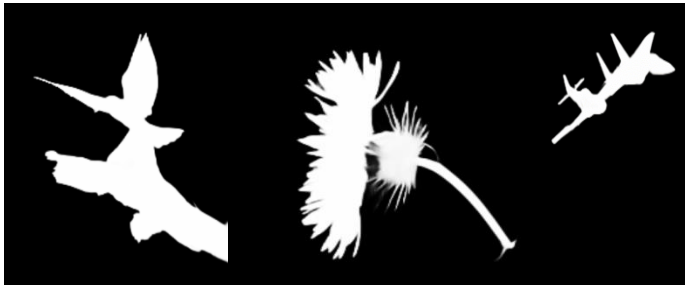
- Our model achieved 61% Peak_IoU in test dataset.
- Because of the wrong choice of one param of output image function, the Peak_Threshold is only 7.
- We corrected this mistake and improved the Peak_Threshold to 30.

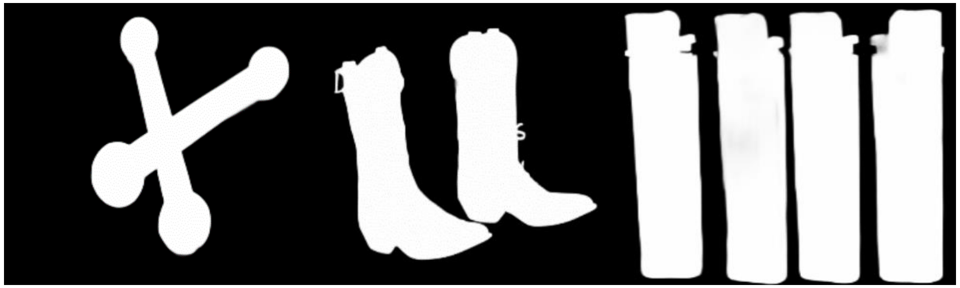




Above output images show that our method works well on images with a single object or overlapping objects.

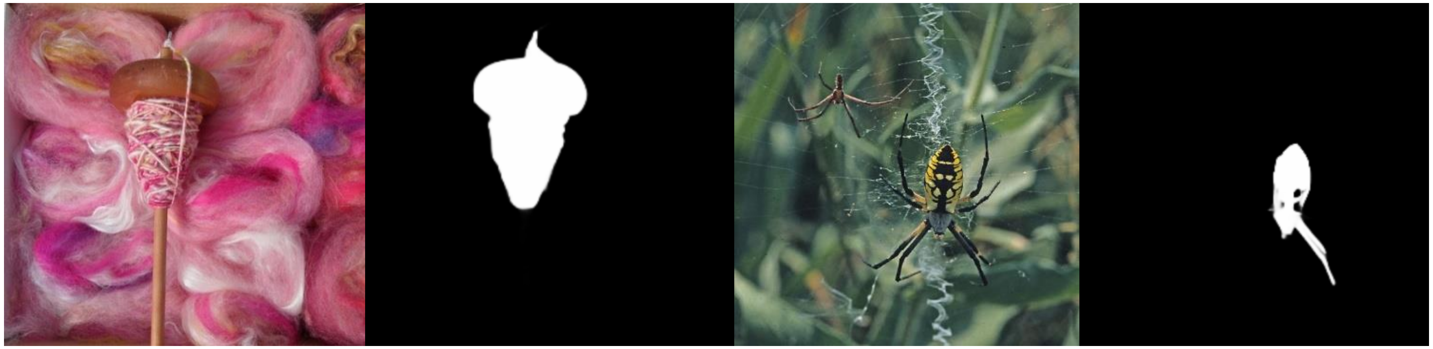
Experiment Results





- The three images in left panel demonstrate that our model can localize the objects with small local complex edge structure clearly.
- The three images in right panel show that our method is also applicable to the images containing multiple instances belonging to one category.

Failure cases:

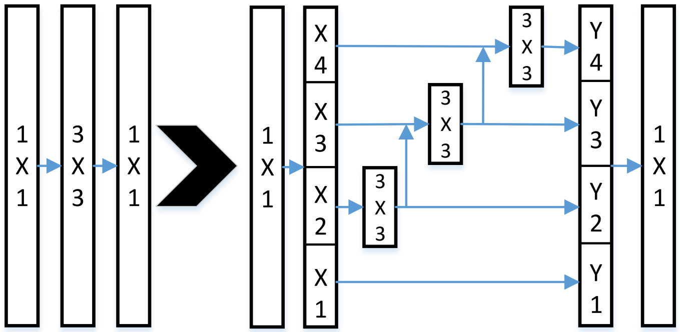


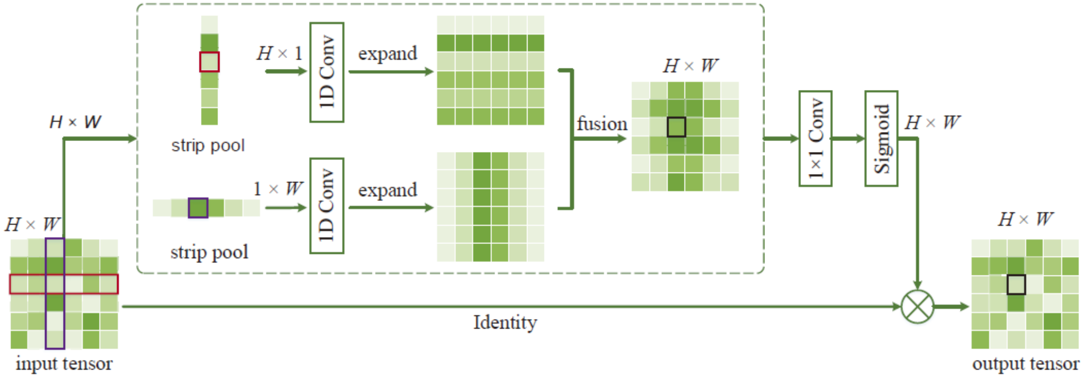
We found our method does not work well on situations as shown above.
One is with special strips, and the other is with some interfering information.

Future Improvements

Our plan:

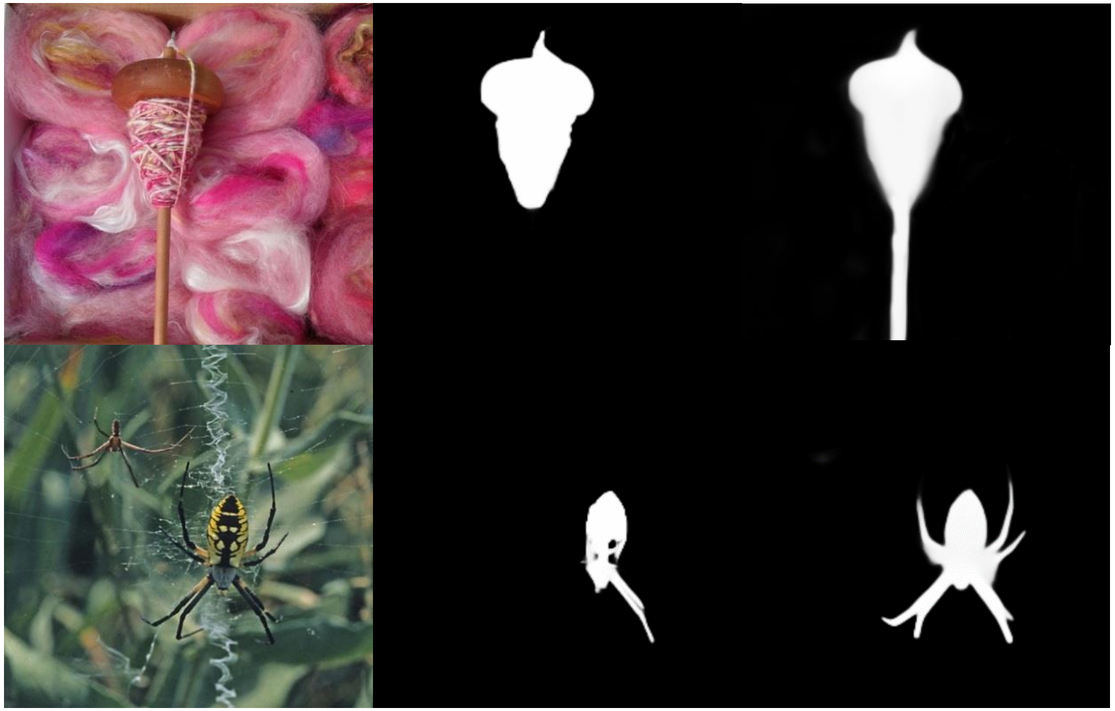
1. Bringing Res2Net structure to downsampling layers.
2. Providing information of rich scales by Integrating the channels.





3. Add Strip Pooling in our encoder-decoder model, as Strip Pooling can help the network better exploit long-range dependencies.

Future Improvements



We carried out some preliminary experiments. The results show that our improvements promote the effects on some challenging images.

THANKS FOR LISTENING