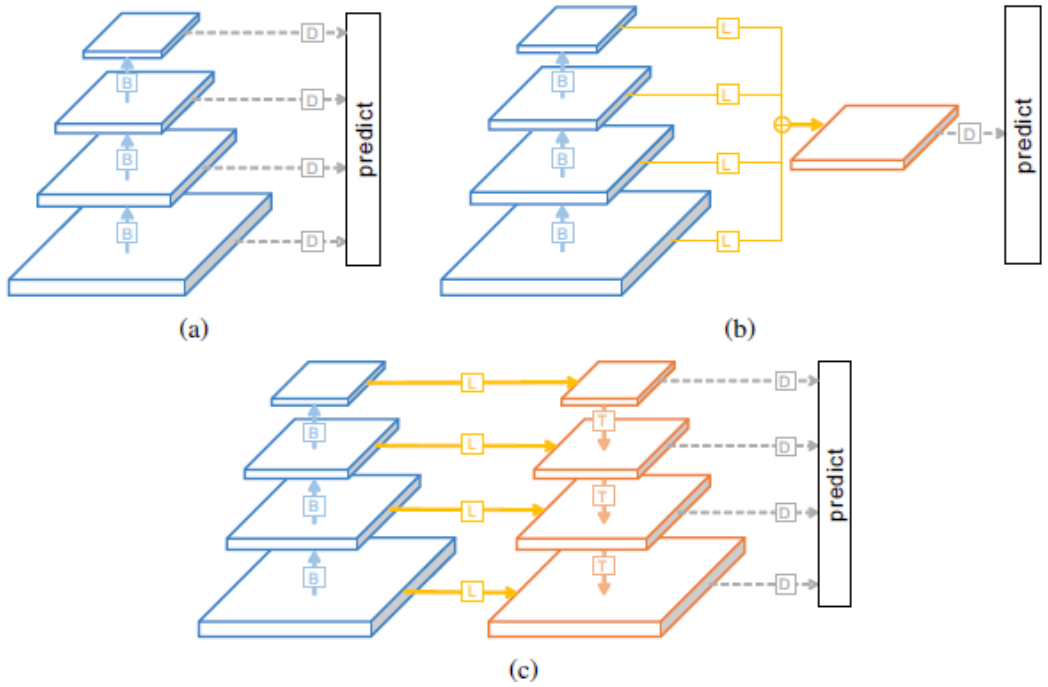
ScarfNet: Multi-scale Features with Deeply Fused and Redistributed Semantics for Enhanced Object Detection

Jin Hyeok Yoo[1], Dongsuk Kum[2], and Jun Won Choi[1]

[1] Hanyang University, [2] Korea Advanced Institute of Science and Technology

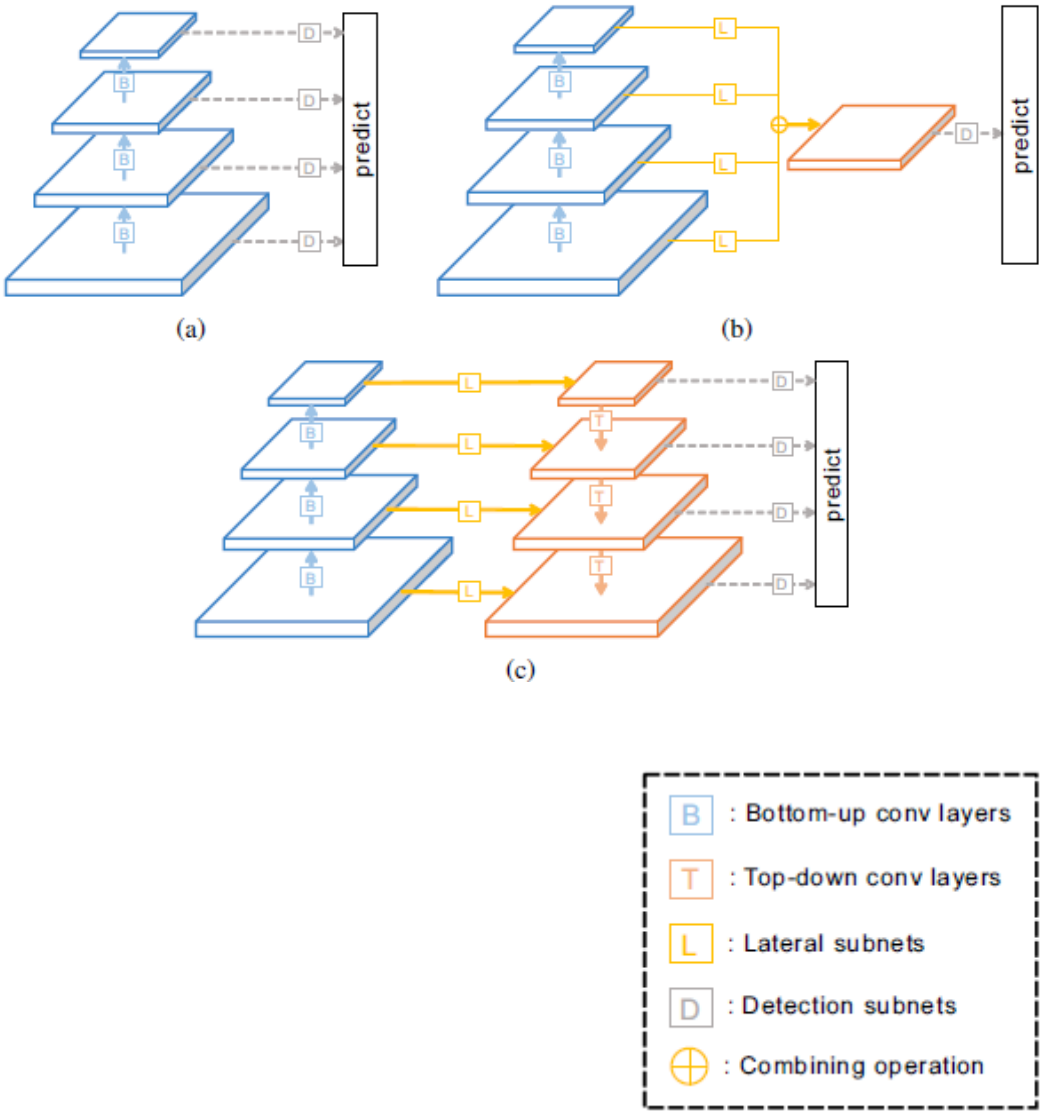
Introduction

- Pyramidal features in object detection
  - Feature Pyramid Method
    - To detect objects of various sizes, object detectors often exploit the ***multiscale feature maps called feature pyramids as shown in (a)*** , which are obtained by the backbone network.
      - ➢ However, the bottom-level features are ***not deep enough to exhibit high-level semantics*** underlying in the objects and their surroundings.
  - Top-down-based Method
    - To provide the contextual information to the bottom-level features, ***top-down-based method using lateral connections*** are proposed as illustrated in Fig. 1 (c).
    - These methods include DSSD [1], FPN [2], StairNet [3], and HR-FPN [4].



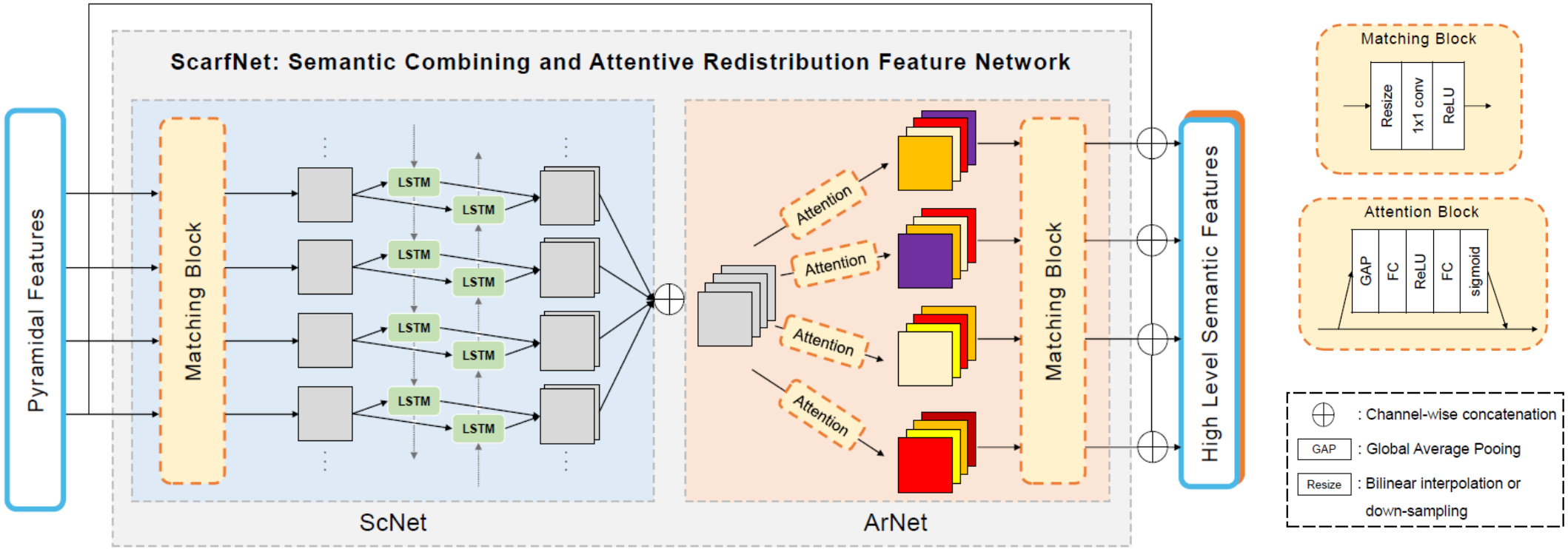
Introduction

- Pyramidal features in object detection
  - Motivation of Proposed ScarfNet
    1. The capacity of the current architectures for generating top-down features might not be large enough to generate strong semantics for all scales.
    2. The biLSTM is used to combine the multiscale features to incorporate strong semantics for feature pyramids.



Proposed Network : ScarfNet

- Overall Architecture
  1. ScNet (Semantic Combining Network)
     - Combining the scattered semantic information using biLSTM
  2. ArNet (Attentive Redistribution Network)
     - Redistributing the fused semantics back to each pyramid level using the channel-wise attention model.

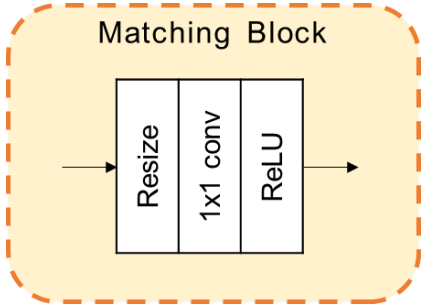


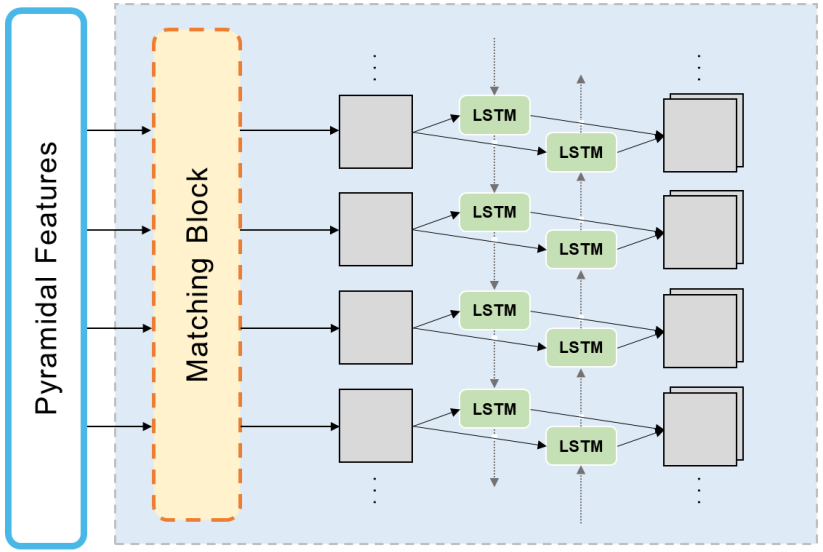
Proposed Network : ScarfNet

- Overall Architecture
  - Procedures
    1. Backbone network generate k pyramidal features
       - $X_{n-k+1:n} = [X_{n-k+1}, X_{n-k+2}, \dots, X_n]$
    2. ScNet produces the feature maps $X^f_{n-k+1:n}$
       - $X^f_{n-k+1:n} = ScNet(X_{n-k+1:n})$
    3. Concatenate the output features of ScNet $X^f_{n-k+1:n}$.
    4. ArNet produce the high-level semantic feature map and concatenated with the original feature to produce the final output feature X'_l.
       - $X'_l = X_l \oplus ArNet\left(X^f_{n-k+1:n}\right)$
  - The overall procedures can be expressed as
    - $X'_l = ScarfNet(X_{n-k+1:n}) = X_l \oplus ArNet_l\left(ScNet(X_{n-k+1:n})\right)$,

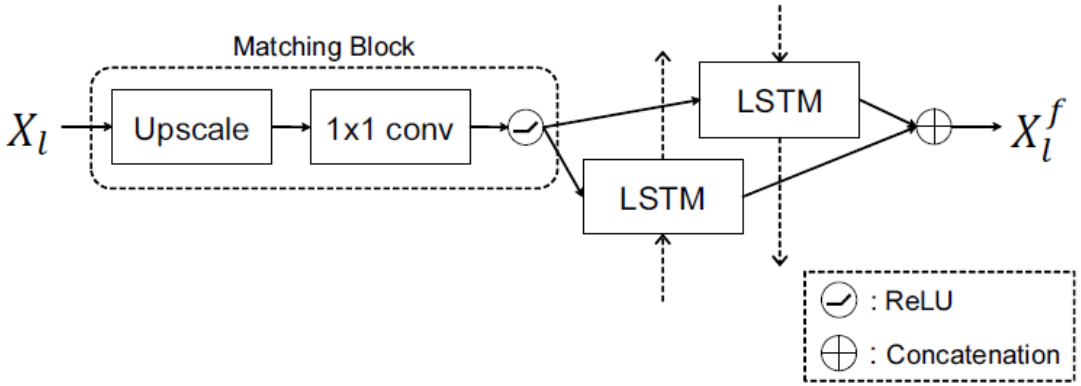
Proposed Network : ScarfNet

- ScNet (Semantic Combining Network)
  - *Objective : combine the scattered semantic information*
    1. Matching block
       - Resizes the pyramidal features such that they have the same size. Then, it adjusts the channel dimension of the input using the 1x1 convolutional layer.
    2. biLSTM
       - The biLSTM model can selectively fuse the contextual information in multiscale features through the gating function.



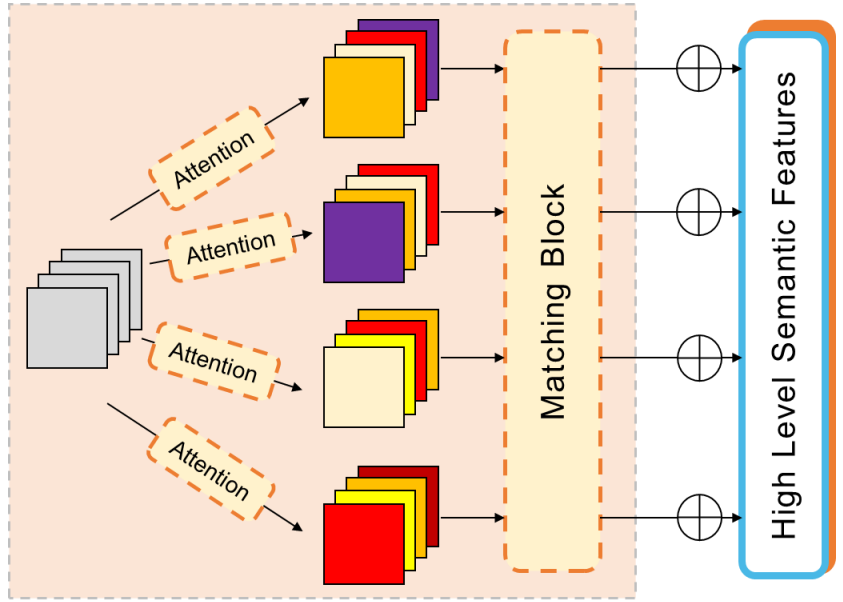


ScNet

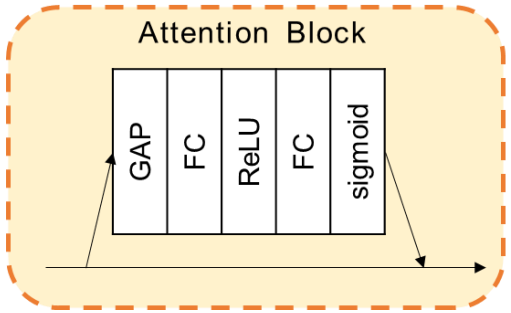


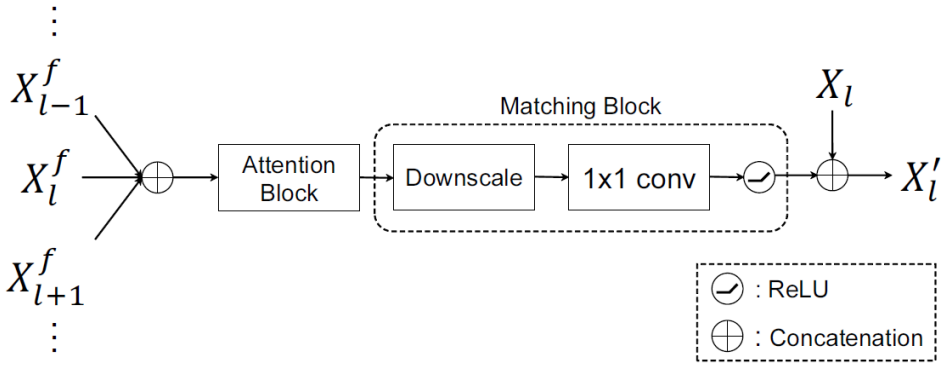
Proposed Network : ScarfNet

- ArNet (Attentive Redistribution Network)
  - *Objective : produce the high-level semantic feature map*
    1. Attention Block
       - After channel-wise concatenation of the outputs of ScNet, apply the channel-wise attention to them.
    2. Matching Block
       - The matching block down-samples the attentive feature maps to the original size of the pyramidal features
    3. High-Level Semantic features
       - Finally, the output of the matching block is concatenated with the original feature to produce the highly semantic feature.



ArNet





Experiments

- Experimental Setup
  - Baseline Detectors
    - Faster R-CNN [5]
    - SSD [6]
    - RetinaNet [7]
  - Datasets
    - PASCAL VOC 2007 [8]
    - PASCAL VOC 2012 [8]
    - MS COCO [9]

Experiments on PASCAL VOC

- Results on PASCAL VOC 2007, 2012

Method	Backbone	Input size	mAP (%)	
			VOC 2007	VOC 2012
StairNet [14]	VGG-16	300 × 300	78.8	76.4
Faster R-CNN [5]	VGG-16	∼ 1000 × 600	73.2	70.4
ION [23]	VGG-16	∼ 1000 × 600	76.5	76.4
SSD300* [7]	VGG-16	300 × 300	77.5	75.8
Scarf SSD300 (ours)	VGG-16	300 × 300	79.4	77.2
SSD512* [7]	VGG-16	512 × 512	79.8	78.5
Scarf SSD512 (ours)	VGG-16	512 × 512	**81.6**	**79.8**
SSD321 [12]	ResNet-101	321 × 321	77.1	75.4
SSD513 [12]	ResNet-101	513 × 513	80.6	79.4
DSSD321 [12]	ResNet-101	321 × 321	78.6	76.3
DSSD513 [12]	ResNet-101	513 × 513	81.5	80.0
R-FCN [27]	ResNet-101	∼ 1000 × 600	80.5	77.6
Faster R-CNN† [5]	ResNet-101	∼ 833 × 500	81.1	-
Scarf Faster R-CNN (ours)	ResNet-101	∼ 833 × 500	82.3	-
RetinaNet500† [10]	ResNet-101	∼ 833 × 500	83.0	-
Scarf RetinaNet500 (ours)	ResNet-101	∼ 833 × 500	**83.5**	-

Experiments on MS COCO

- Results on MS COCO

Method	Network	Backbone	Module	Input size	fps	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
two-stage	Faster R-CNN* [5]	ResNeXt-101	FPN	$\sim 833 \times 500$	15.3	37.6	59.1	40.7	19.2	41.8	52.3
		ResNeXt-101	FPN	$\sim 1333 \times 800$	10.3	41.9	63.9	45.9	25.0	45.3	52.3
	Scarf Faster R-CNN (ours)	ResNeXt-101	SCARF	$\sim 833 \times 500$	13.8	38.5	59.9	41.5	19.1	42.9	**54.1**
		ResNeXt-101	SCARF	$\sim 1333 \times 800$	8.9	**42.8**	**64.3**	**47.1**	**26.0**	**45.7**	52.9
one-stage	SSD513 [12]	ResNet-101	-	513×513	12.5	31.2	50.4	33.3	10.2	34.5	49.8
	DSSD513 [12]	ResNet-101	DSSD	513×513	10.0	33.2	53.3	35.2	13.0	35.4	51.1
	Scarf SSD513 (ours)	ResNet-101	SCARF	513×513	11.5	**34.5**	**54.1**	**36.3**	**15.1**	**36.1**	**51.6**
	RetinaNet [10]	ResNet-101	FPN	$\sim 833 \times 500$	15.4	34.4	53.1	36.8	14.7	38.5	49.1
		ResNeXt-101	FPN	$\sim 1333 \times 800$	9.3	40.8	61.1	44.1	24.1	44.2	51.2
	Scarf RetinaNet (ours)	ResNet-101	SCARF	$\sim 833 \times 500$	13.6	35.1	53.8	37.7	15.8	38.7	49.0
		ResNeXt-101	SCARF	$\sim 1333 \times 800$	8.4	**41.6**	**62.0**	**44.6**	**24.5**	**45.5**	**52.3**

Experiments (Ablation study)

- Results

	Method	mAP
Ablation study	Basedline (SSD)	77.5
	biLSTM	79.1
	biLSTM + channel-wise attention	**79.4**
Other fusion strategy (used with channel-wise attention)	1x1 conv.-based fusion	78.9
	uniLSTM	78.7
	Top-down structure with lateral connections	78.6

Conclusion

- Conclusion
  - In this study, we developed a deep architecture that generates multiscale features with strong semantics to reliably detect the objects in various sizes.
  - Our ScarfNet method transforms the pyramidal features produced by the baseline detector into evenly abstract features. ScarfNet fuses the pyramidal features using biLSTM and distributes the semantics back to each multiscale feature.
  - We verified through experiments conducted with PASCAL VOC and MS COCO datasets that the proposed ScarfNet method significantly increases the detection performance over the baseline detectors.
  - Our object detector achieves the state-of-the-art performance on the PASCAL VOC and COCO benchmarks.

References

- References

[1] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, “Dssd: Deconvolutional single shot detector,” *arXiv preprint arXiv:1701.06659,* 2017.

[2] T.-Y. Lin, P. Doll´ar, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, “Feature pyramid networks for object detection,” *IEEE conference on Computer Vision and Pattern Recognition (CVPR),* 2017.

[3] S. Woo, S. Hwang, and I. S. Kweon, “Stairnet: Top-down semantic aggregation for accurate one shot detection*,” IEEE Winter Conference on Applications of Computer Vision (WACV),* pp. 1093–1102, 2018.

[4] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, “High-resolution representations for labeling pixels and regions,” *arXiv preprint arXiv:1904.04514,* 2019.

[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *Advances in Neural Information Processing Systems (NIPS),* pp. 91–99, 2015.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” *European Conference on Computer Vision (ECCV),* pp. 21–37, 2016.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” *IEEE International Conference on Computer Vision (ICCV),* pp. 2999–3007, 2017.

[8] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *IEEE International Conference on Computer Vision (ICCV),* vol. 88, no. 2, pp. 303–338, 2010.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, “Microsoft coco: Common objects in context,” *European Conference on Computer Vision (ECCV),* pp. 740–755, 2014.