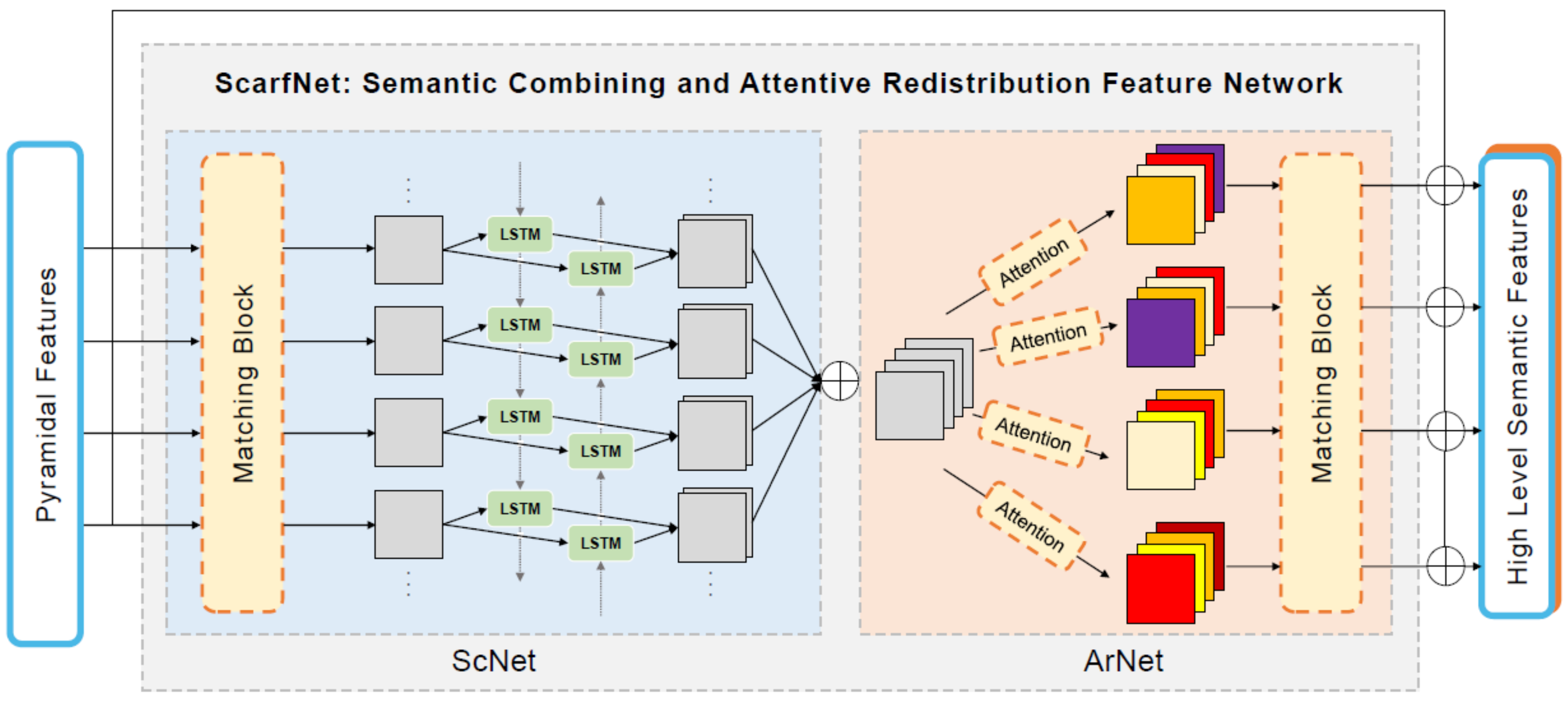
SCARFNET: MULTI-SCALE FEATURES WITH DEEPLY FUSED AND REDISTRIBUTED SEMANTICS FOR ENHANCED OBJECT DETECTION

Jin Hyeok Yoo, Dongsuk Kum, and Jun Won Choi
Signal Processing & Artificial-intelligence Laboratory
Hanyang University, Seoul Korea

Summary

- We introduce a new deep architecture for closing the semantic gaps between the multiscale feature maps.
- The proposed ScarfNet generates new multiscale feature maps with deeply fused and redistributed semantics by using the combination of biLSTM and the channel-wise attention model.
- For the first time in the literature, the biLSTM is used to combine the multiscale features to incorporate strong semantics for feature pyramids.
- The biLSTM model can produce deeply fused semantic information using the recurrent connection over different pyramid scales.
- The evaluation shows that our method offers significant improvement over the baseline detectors as well as other competitive detectors.

Proposed Network


ScarfNet: Semantic Combining and Attentive Redistribution Feature Network
Pyramidal Features
Matching Block
LSTM
Attention
Matching Block
High Level Semantic Features
ScNet
ArNet

Overall Architecture

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

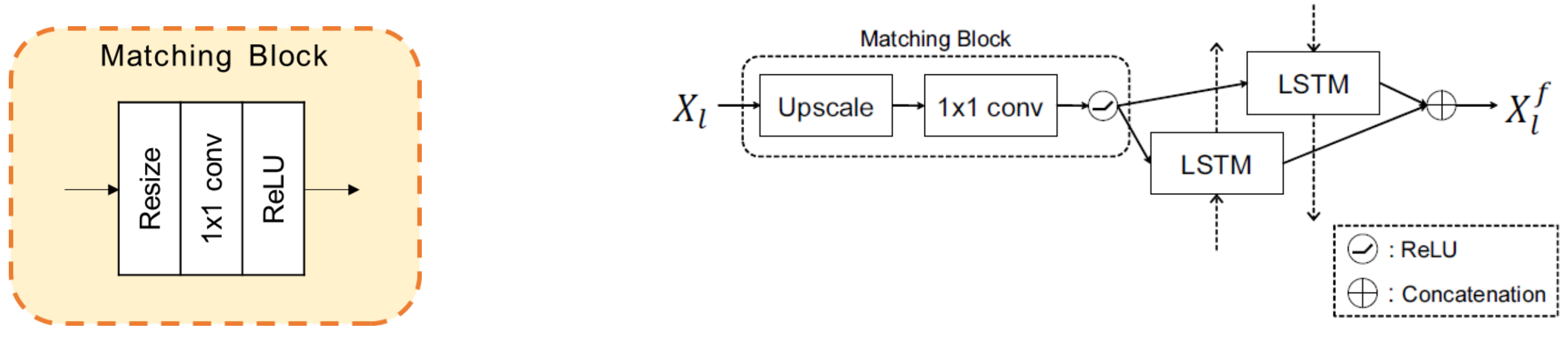
➢ Detailed procedures:

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.
$$X'_l = X_l \oplus ArNet(X^f_{n-k+1:n})$$

- The overall procedures can be expressed as
$$X'_l = ScarfNet(X_{n-k+1:n}) = X_l \oplus ArNet_l(ScNet(X_{n-k+1:n})),$$

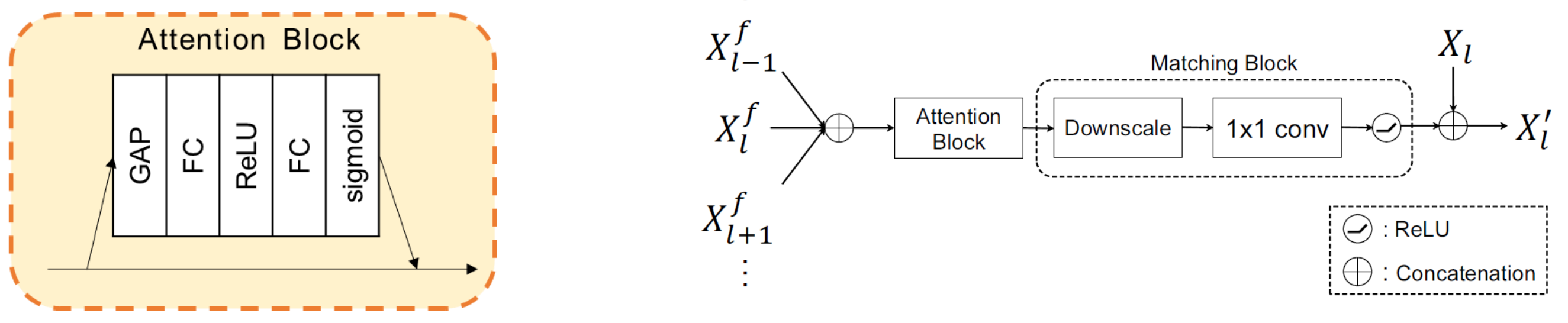
ScNet (Semantic Combining Network)

- ***Objective : combine the scattered semantic information***
- 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.
- biLSTM : The biLSTM model can selectively fuse the contextual information in multiscale features through the gating function.



ArNet (Attentive Redistribution Network)

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



Experimental Results

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	~ 833 × 500	15.3	37.6	59.1	40.7	19.2	41.8	52.3
		ResNeXt-101	FPN	~ 1333 × 800	10.3	41.9	63.9	45.9	25.0	45.3	52.3
	Scarf Faster R-CNN (ours)	ResNeXt-101	SCARF	~ 833 × 500	13.8	38.5	59.9	41.5	19.1	42.9	**54.1**
		ResNeXt-101	SCARF	~ 1333 × 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	~ 833 × 500	15.4	34.4	53.1	36.8	14.7	38.5	49.1
		ResNeXt-101	FPN	~ 1333 × 800	9.3	40.8	61.1	44.1	24.1	44.2	51.2
	Scarf RetinaNet (ours)	ResNet-101	SCARF	~ 833 × 500	13.6	35.1	53.8	37.7	15.8	38.7	49.0
		ResNeXt-101	SCARF	~ 1333 × 800	8.4	**41.6**	**62.0**	**44.6**	**24.5**	**45.5**	**52.3**

- Table provides the detection accuracy of the algorithms tested on the MS COCO
- The experiment was conducted on various baseline detectors and feature pyramid modules.
- The proposed Scarf SSD513 and Scarf RetinaNet achieve the significant performance gain over the baselines.

Ablation Studies

	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

- Table shows how the performance of our method improves as we add bi-LSTM and channel-wise attention to the baseline one by one.
- The biLSTM offers the 1.6% AP gain over the baseline and combination of biLSTM and channel-wise attention adds 1.9% AP gain.

Conclusions

- 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.