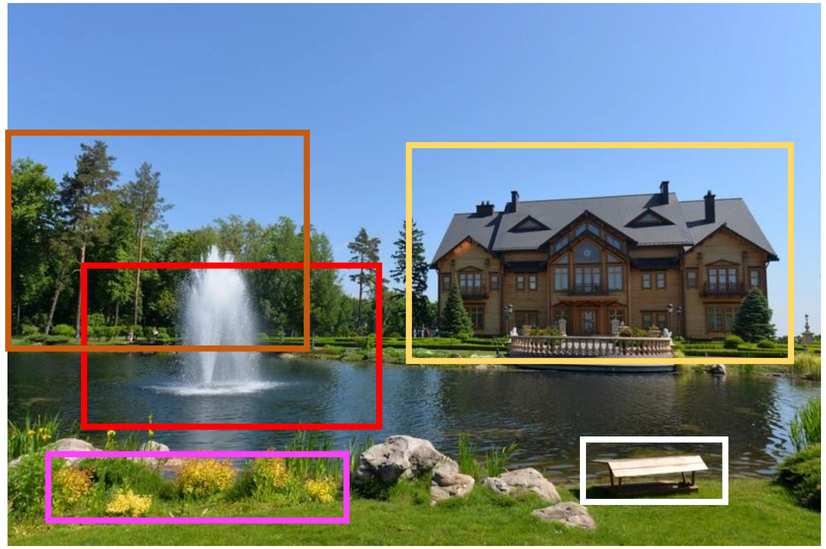
Learning Semantic-Specific Graph Representation for Multi-Label Image Recognition
(ICCV 2019)

ChaeHun Shin

Seoul National University

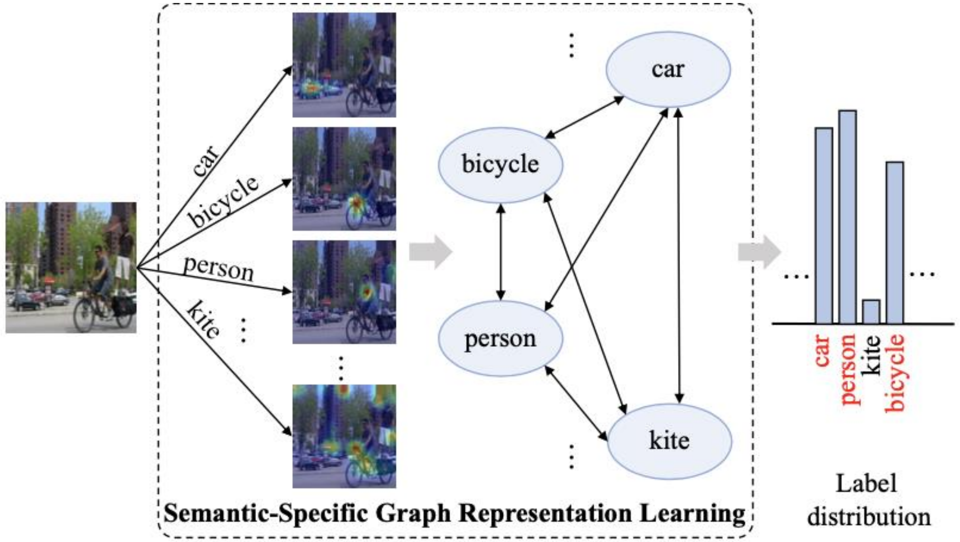
Multi-Label Image Recognition

- Identifies all objects in Image
  - Employing object localization
  - Resort to visual attention networks.



Multi-Label Image classification

- Most of existing algorithms do not consider the dependencies between categories or regions.
- Modeling the dependency between semantics of categories with GNN.
  - Semantic decoupling module
  - Construct a graph based on the statistical label co-occurrence.



Category feature extraction

- Extract feature map of images

$$\mathbf{f}^I = f_{cnn}(I),$$

- Extract semantic embedding vector for each categories

$$\mathbf{x}_c = f_g(w_c),$$

- Semantic guided attention mechanism
    - Guide to focusing more on the semantic-aware regions.
    - Attention weights for each position

$$\tilde{\mathbf{f}}^I_{c,wh} = \mathbf{P}^T \left(\tanh\left((\mathbf{U}^T \mathbf{f}^I_{wh}) \odot (\mathbf{V}^T \mathbf{x}_c)\right)\right) + \mathbf{b},$$

$$\tilde{a}_{c,wh} = f_a(\tilde{\mathbf{f}}^I_{c,wh}).$$

$$a_{c,wh} = \frac{\exp(\tilde{a}_{c,wh})}{\sum_{w',h'} \exp(\tilde{a}_{c,w'h'})}.$$

- Final feature map of images for each class

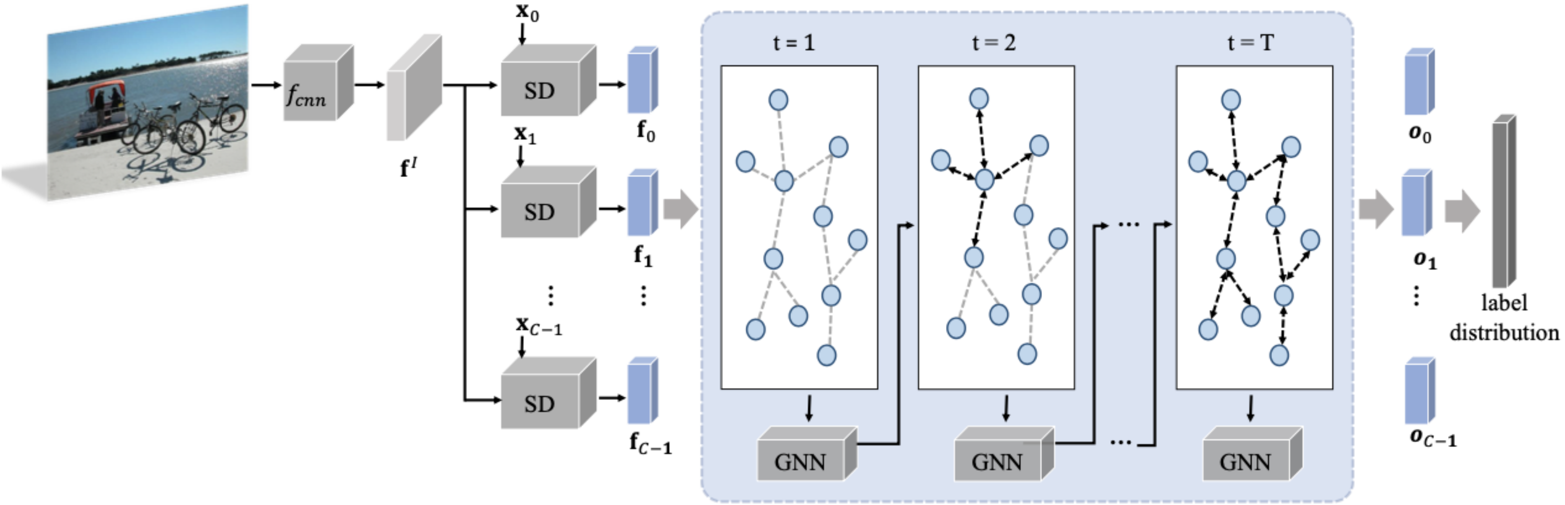
$$\mathbf{f}_c = \sum_{w,h} a_{c,wh} \mathbf{f}_{c,wh}$$

$$\{\mathbf{f}_0, \mathbf{f}_1, \ldots, \mathbf{f}_{C-1}\}$$

Graph construction

- Construct the graph $\mathcal{G} = \{V, A\}$ based on the statistical label co-occurrence.
  - Node set $V = \{v_0, v_1, \dots v_{c-1}\}$ represents each category.
  - Edge set $A = \{a_{00}, a_{01}, \dots a_{(c-1)(c-1)}\}$, where $a_{cc'}$ means the probability of the existence of object belonging to category c' in the presence of object belonging to category c.

- Each hidden features of nodes are initialized with the feature vectors. $\mathbf{h}_c^0 = \mathbf{f}_c.$



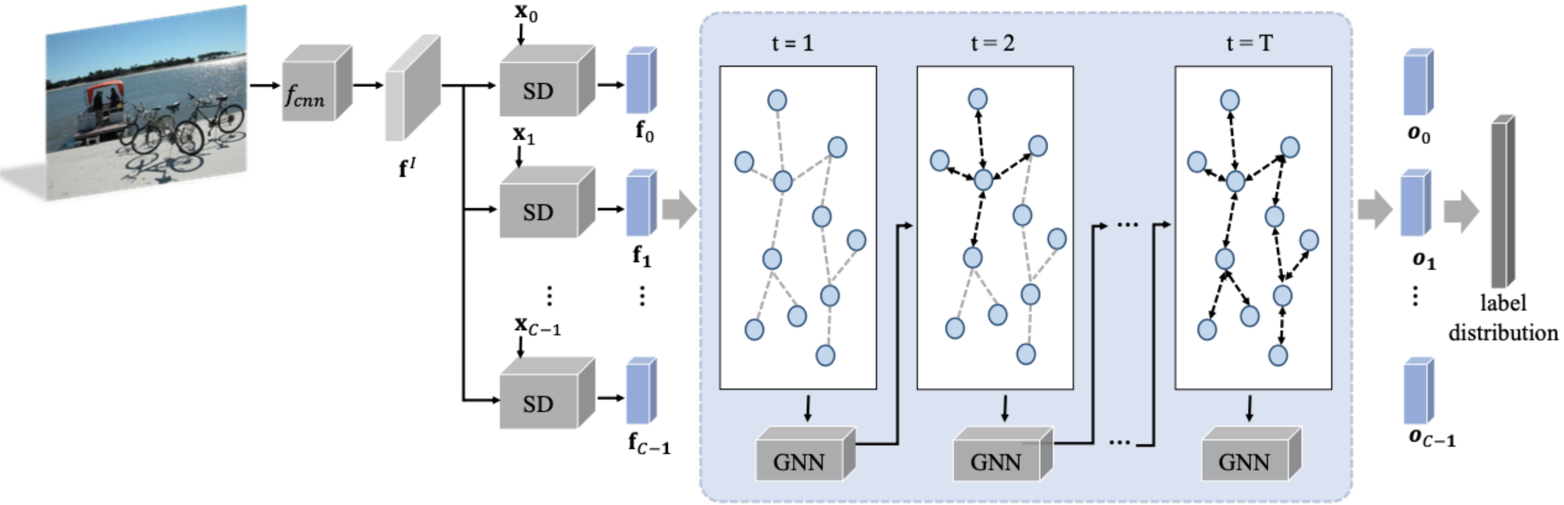
Graph propagation

- Propagate as aggregating message from its neighbor nodes.
    - Nodes interact with each other under the guidance of the statistical label co-occurrence.

$$\mathbf{a}_c^t = \left[\sum_{c'}(a_{cc'})\mathbf{h}_c^{t-1}, \sum_{c'}(a_{c'c})\mathbf{h}_c^{t-1}\right].$$

- Updates the hidden states of nodes via a gated mechanism.

$$\begin{aligned}
\mathbf{z}_c^t &= \sigma(\mathbf{W}^z\mathbf{a}_c^t + \mathbf{U}^z\mathbf{h}_c^{t-1}) \\
\mathbf{r}_c^t &= \sigma(\mathbf{W}^r\mathbf{a}_c^t + \mathbf{U}^r\mathbf{h}_c^{t-1}) \\
\widetilde{\mathbf{h}_c^t} &= \tanh\left(\mathbf{W}\mathbf{a}_c^t + \mathbf{U}(\mathbf{r}_c^t \odot \mathbf{h}_c^{t-1})\right) \\
\mathbf{h}_c^t &= (1 - \mathbf{z}_c^t) \odot \mathbf{h}_c^{t-1} + \mathbf{z}_c^t \odot \widetilde{\mathbf{h}_c^t}
\end{aligned}$$



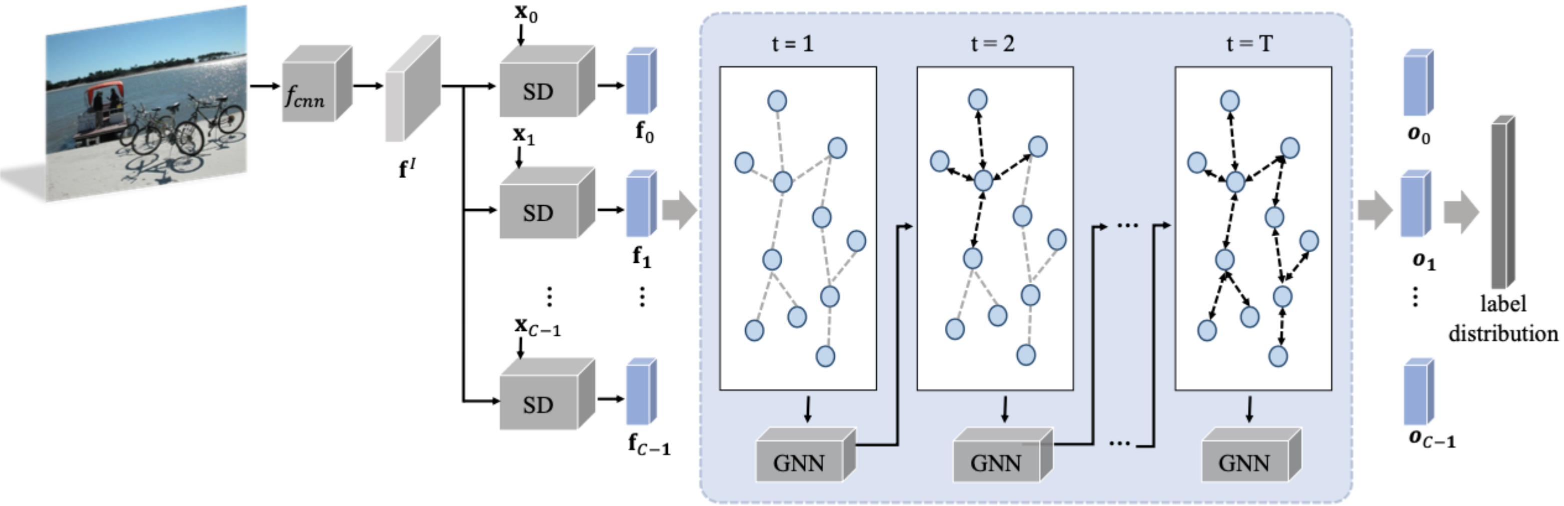
Optimization

- Predict presence of category based on the final hidden states and initial hidden states

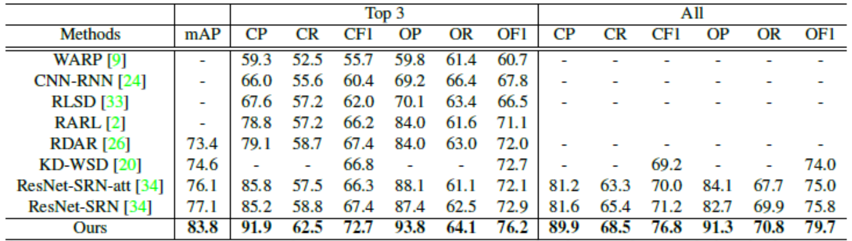
$$\begin{aligned} \mathbf{o}_c &= f_o(\mathbf{h}_c^T, \mathbf{h}_c^0) \\ s_c &= f_c(\mathbf{o}_c) \end{aligned}$$

$$p_{ic} = \sigma(s_{ic}).$$

$$\mathcal{L} = \sum_{i=0}^{N-1} \sum_{c=0}^{C-1} \left(y_{ic} \log p_{ic} + (1 - y_{ic}) \log(1 - p_{ic}) \right).$$

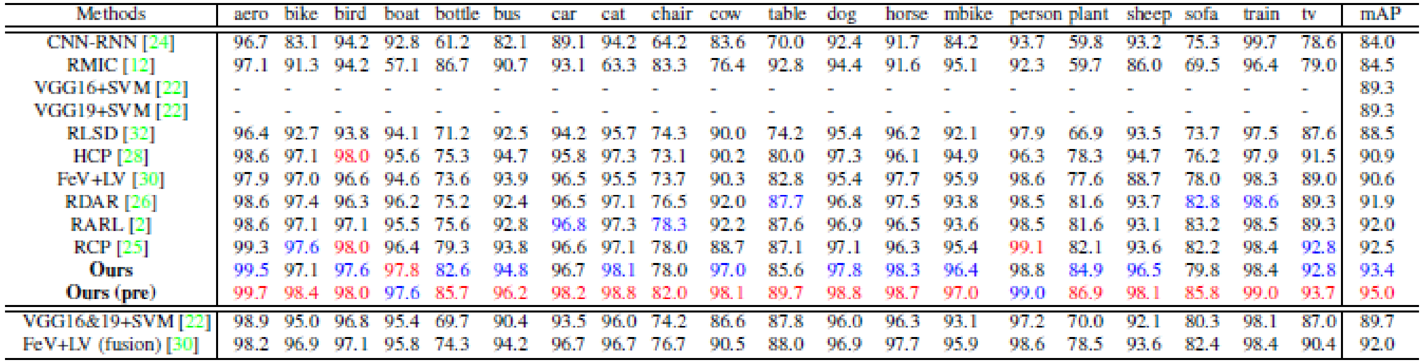


Experiments



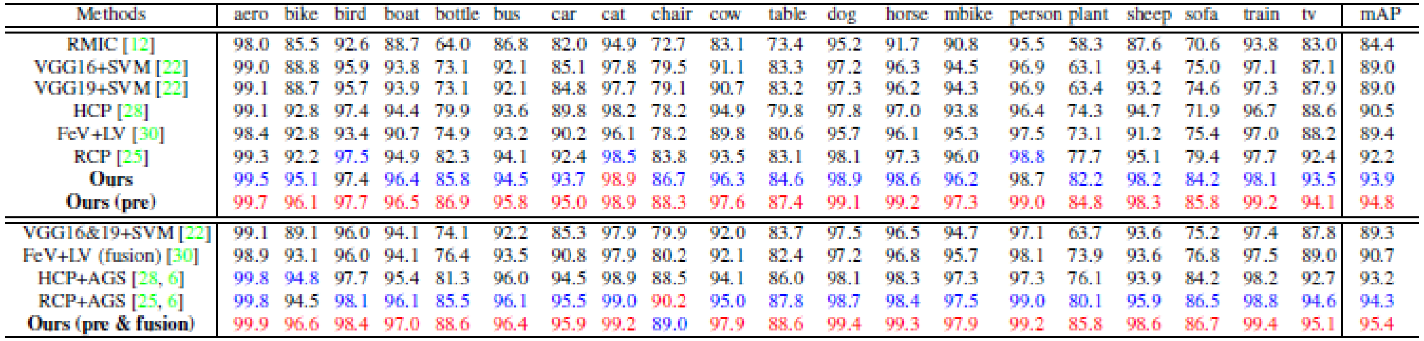
Methods	mAP	Top 3						All					
		CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
WARP [9]	-	59.3	52.5	55.7	59.8	61.4	60.7	-	-	-	-	-	-
CNN-RNN [24]	-	66.0	55.6	60.4	69.2	66.4	67.8	-	-	-	-	-	-
RLSD [33]	-	67.6	57.2	62.0	70.1	63.4	66.5	-	-	-	-	-	-
RARL [2]	-	78.8	57.2	66.2	84.0	61.6	71.1						
RDAR [26]	73.4	79.1	58.7	67.4	84.0	63.0	72.0	-	-	-	-	-	-
KD-WSD [20]	74.6	-	-	66.8	-	-	72.7	-	-	69.2	-	-	74.0
ResNet-SRN-att [34]	76.1	85.8	57.5	66.3	88.1	61.1	72.1	81.2	63.3	70.0	84.1	67.7	75.0
ResNet-SRN [34]	77.1	85.2	58.8	67.4	87.4	62.5	72.9	81.6	65.4	71.2	82.7	69.9	75.8
Ours	**83.8**	**91.9**	**62.5**	**72.7**	**93.8**	**64.1**	**76.2**	**89.9**	**68.5**	**76.8**	**91.3**	**70.8**	**79.7**

Table 1. Comparison of mAP, CP, CR, CF1 and OP, OR, OF1 (in %) of our framework and state-of-the-art methods under the settings of all and top-3 labels on the Microsoft COCO dataset. "-" denotes the corresponding result is not provided.



Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
CNN-RNN [24]	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	99.7	78.6	84.0
RMIC [12]	97.1	91.3	94.2	57.1	86.7	90.7	93.1	63.3	83.3	76.4	92.8	94.4	91.6	95.1	92.3	59.7	86.0	69.5	96.4	79.0	84.5
VGG16+SVM [22]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	89.3
VGG19+SVM [22]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	89.3
RLSD [32]	96.4	92.7	93.8	94.1	71.2	92.5	94.2	95.7	74.3	90.0	74.2	95.4	96.2	92.1	97.9	66.9	93.5	73.7	97.5	87.6	88.5
HCP [28]	98.6	97.1	98.0	95.6	75.3	94.7	95.8	97.3	73.1	90.2	80.0	97.3	96.1	94.9	96.3	78.3	94.7	76.2	97.9	91.5	90.9
FeV+LV [30]	97.9	97.0	96.6	94.6	73.6	93.9	96.5	95.5	73.7	90.3	82.8	95.4	97.7	95.9	98.6	77.6	88.7	78.0	98.3	89.0	90.6
RDAR [26]	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92.0	87.7	96.8	97.5	93.8	98.5	81.6	93.7	82.8	98.6	89.3	91.9
RARL [2]	98.6	97.1	97.1	95.5	75.6	92.8	96.8	97.3	78.3	92.2	87.6	96.9	96.5	93.6	98.5	81.6	93.1	83.2	98.5	89.3	92.0
RCP [25]	99.3	97.6	98.0	96.4	79.3	93.8	96.6	97.1	78.0	88.7	87.1	97.1	96.3	95.4	99.1	82.1	93.6	82.2	98.4	92.8	92.5
Ours	99.5	97.1	97.6	97.8	82.6	94.8	96.7	98.1	78.0	97.0	85.6	97.8	98.3	96.4	98.8	84.9	96.5	79.8	98.4	92.8	93.4
Ours (pre)	99.7	98.4	98.0	97.6	85.7	96.2	98.2	98.8	82.0	98.1	89.7	98.8	98.7	97.0	99.0	86.9	98.1	85.8	99.0	93.7	95.0
VGG16&19+SVM [22]	98.9	95.0	96.8	95.4	69.7	90.4	93.5	96.0	74.2	86.6	87.8	96.0	96.3	93.1	97.2	70.0	92.1	80.3	98.1	87.0	89.7
FeV+LV (fusion) [30]	98.2	96.9	97.1	95.8	74.3	94.2	96.7	96.7	76.7	90.5	88.0	96.9	97.7	95.9	98.6	78.5	93.6	82.4	98.4	90.4	92.0

Table 2. Comparison of AP and mAP in % of our framework and state-of-the-art methods on the PASCAL VOC 2007 dataset. Upper part presents the results of single model and lower part presents those that aggregate multiple models. "Ours" and "Ours (pre)" denote our framework without and with pre-training on the COCO dataset. The best and second best results are highlighted in red and blue, respectively. "-" denotes the corresponding result is not provided. Best viewed in color.



Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
RMIC [12]	98.0	85.5	92.6	88.7	64.0	86.8	82.0	94.9	72.7	83.1	73.4	95.2	91.7	90.8	95.5	58.3	87.6	70.6	93.8	83.0	84.4
VGG16+SVM [22]	99.0	88.8	95.9	93.8	73.1	92.1	85.1	97.8	79.5	91.1	83.3	97.2	96.3	94.5	96.9	63.1	93.4	75.0	97.1	87.1	89.0
VGG19+SVM [22]	99.1	88.7	95.7	93.9	73.1	92.1	84.8	97.7	79.1	90.7	83.2	97.3	96.2	94.3	96.9	63.4	93.2	74.6	97.3	87.9	89.0
HCP [28]	99.1	92.8	97.4	94.4	79.9	93.6	89.8	98.2	78.2	94.9	79.8	97.8	97.0	93.8	96.4	74.3	94.7	71.9	96.7	88.6	90.5
FeV+LV [30]	98.4	92.8	93.4	90.7	74.9	93.2	90.2	96.1	78.2	89.8	80.6	95.7	96.1	95.3	97.5	73.1	91.2	75.4	97.0	88.2	89.4
RCP [25]	99.3	92.2	97.5	94.9	82.3	94.1	92.4	98.5	83.8	93.5	83.1	98.1	97.3	96.0	98.8	77.7	95.1	79.4	97.7	92.4	92.2
Ours	99.5	95.1	97.4	96.4	85.8	94.5	93.7	98.9	86.7	96.3	84.6	98.9	98.6	96.2	98.7	82.2	98.2	84.2	98.1	93.5	93.9
Ours (pre)	99.7	96.1	97.7	96.5	86.9	95.8	95.0	98.9	88.3	97.6	87.4	99.1	99.2	97.3	99.0	84.8	98.3	85.8	99.2	94.1	94.8
VGG16&19+SVM [22]	99.1	89.1	96.0	94.1	74.1	92.2	85.3	97.9	79.9	92.0	83.7	97.5	96.5	94.7	97.1	63.7	93.6	75.2	97.4	87.8	89.3
FeV+LV (fusion) [30]	98.9	93.1	96.0	94.1	76.4	93.5	90.8	97.9	80.2	92.1	82.4	97.2	96.8	95.7	98.1	73.9	93.6	76.8	97.5	89.0	90.7
HCP+AGS [28, 6]	99.8	94.8	97.7	95.4	81.3	96.0	94.5	98.9	88.5	94.1	86.0	98.1	98.3	97.3	97.3	76.1	93.9	84.2	98.2	92.7	93.2
RCP+AGS [25, 6]	99.8	94.5	98.1	96.1	85.5	96.1	95.5	99.0	90.2	95.0	87.8	98.7	98.4	97.5	99.0	80.1	95.9	86.5	98.8	94.6	94.3
Ours (pre & fusion)	99.9	96.6	98.4	97.0	88.6	96.4	95.9	99.2	89.0	97.9	88.6	99.4	99.3	97.9	99.2	85.8	98.6	86.7	99.4	95.1	95.4

Table 3. Comparison of AP and mAP in % of our model and state-of-the-art methods on the PASCAL VOC 2012 dataset. Upper part presents the results of single model and lower part presents those that aggregate multiple models. "Ours" and "Ours (pre)" denote our framework without and with pre-training on the COCO dataset. "Ours (pre & fusion)" denotes fusing our two scale results. The best and second best results are highlighted in red and blue, respectively. Best viewed in color.

Experiments

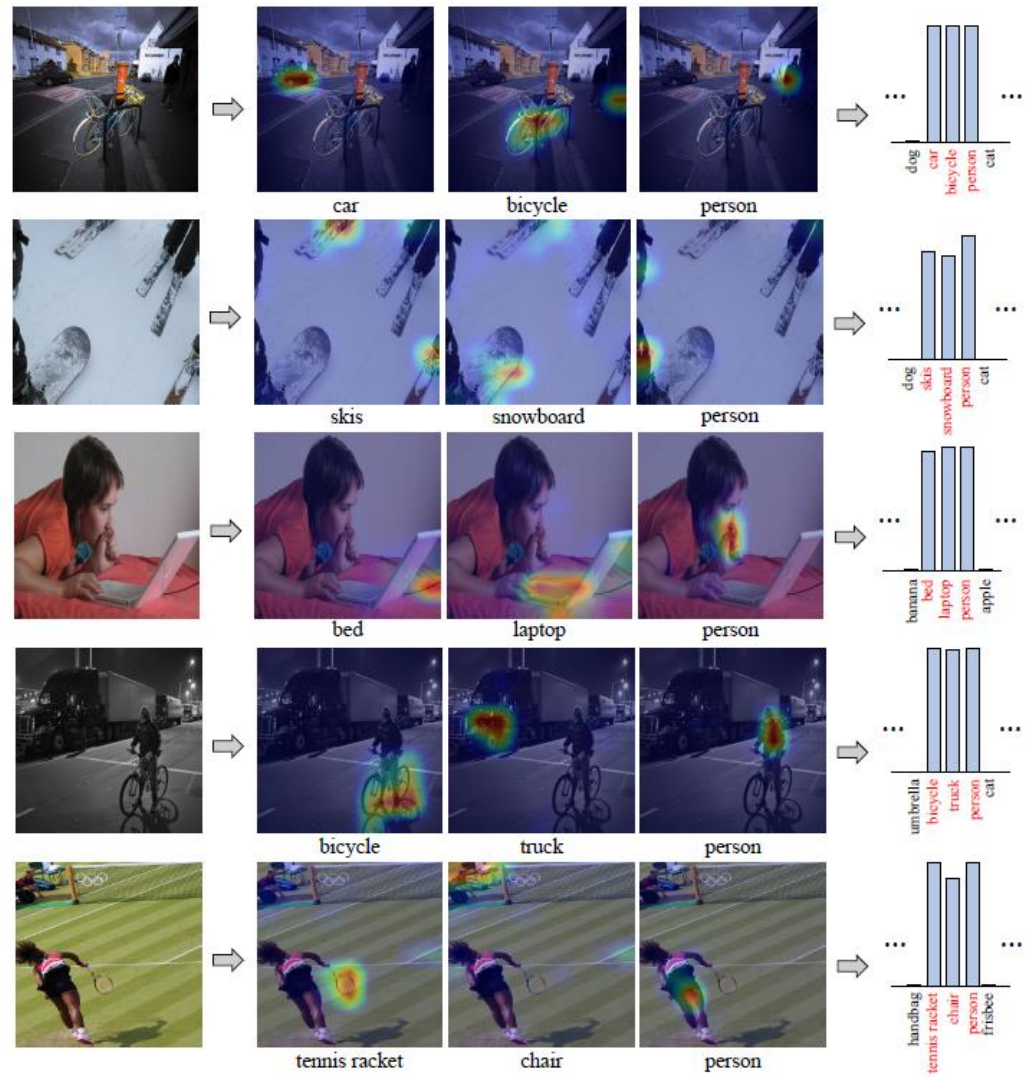
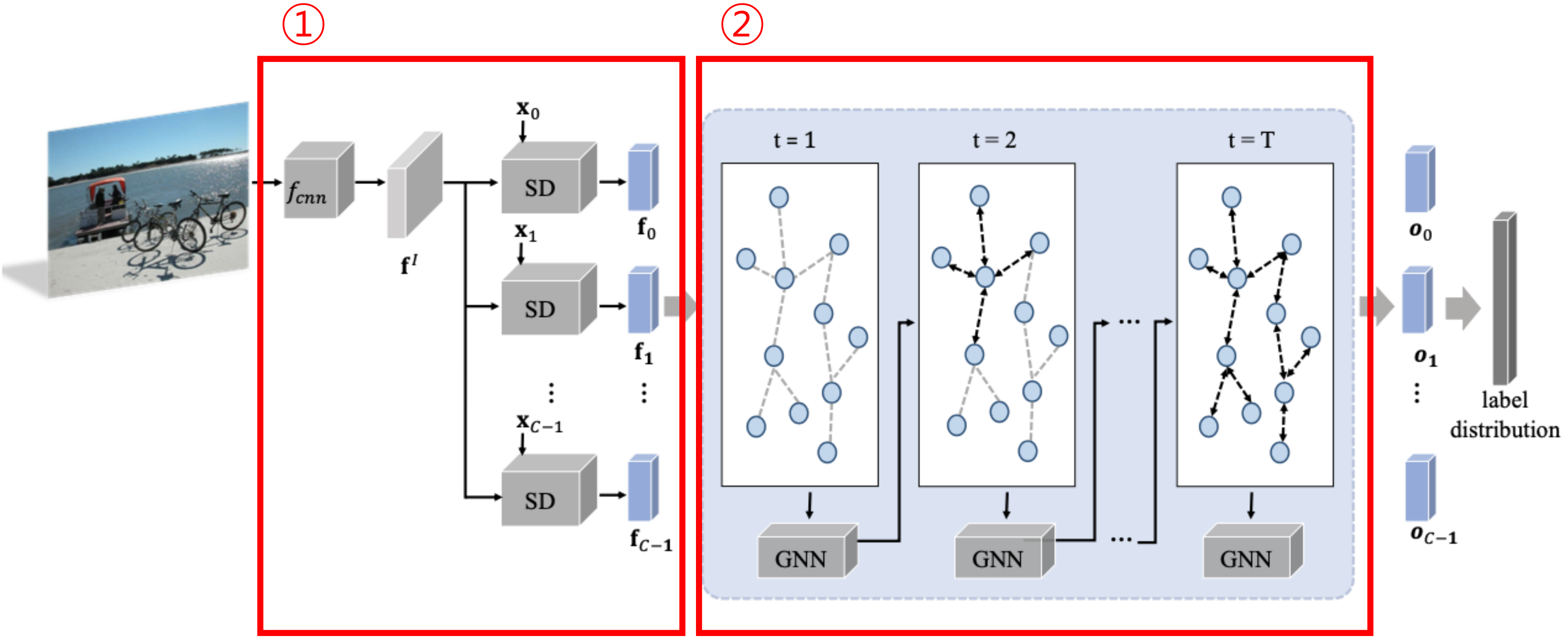


Figure 4. Several examples of input images (left), semantic feature maps corresponding to categories with top 3 highest confidences (middle), and predicted label distribution (right). The ground truth labels are highlighted in red.

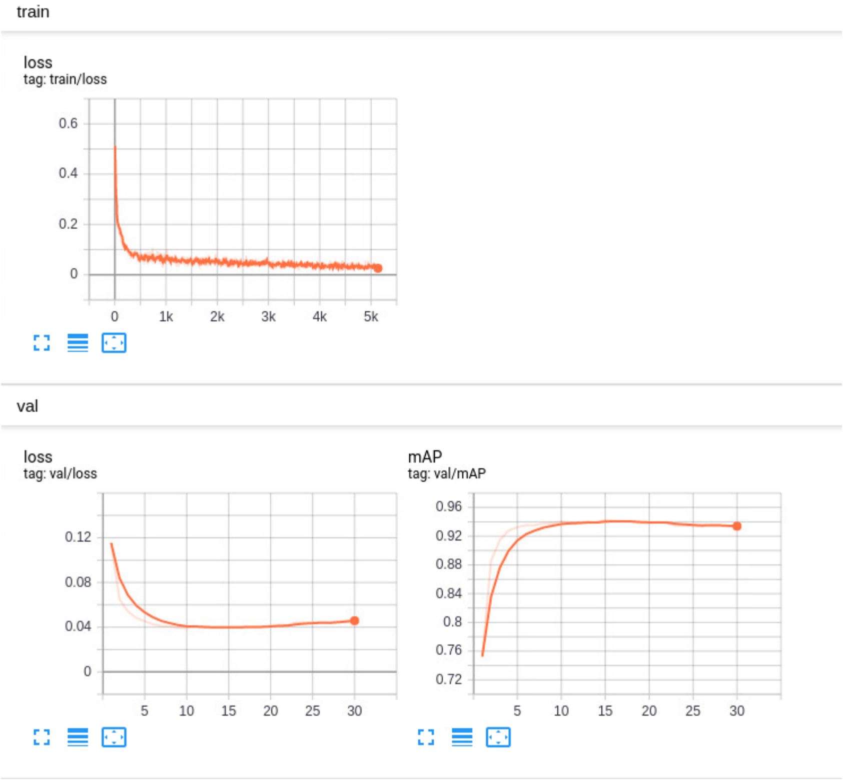
Conclusion

- Modeling the dependency between semantics of categories with GNN.
  - Semantic-guided attention feature map
  - Graph propagation based on the statistical label co-occurence.



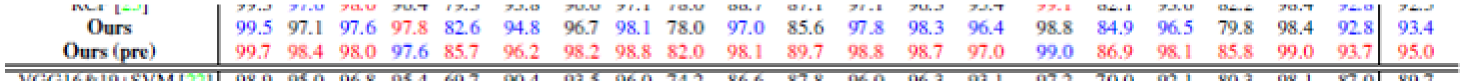
Implementation

- My implementations in https://github.com/chaehunshin/snu_gcn_project
  (only train the PASCAL VOC 2012 dataset)

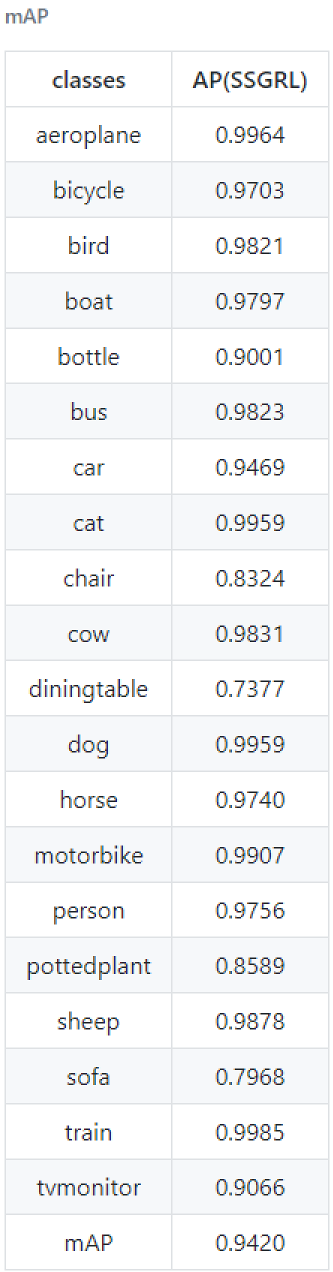


Implementation

- mAP results
  - Better than the original performance reported in the paper. (without pretrained with the COCO dataset)



Ours	99.5	97.1	97.6	97.8	82.6	94.8	96.7	98.1	78.0	97.0	85.6	97.8	98.3	96.4	98.8	84.9	96.5	79.8	98.4	92.8	93.4
Ours (pre)	99.7	98.4	98.0	97.6	85.7	96.2	98.2	98.8	82.0	98.1	89.7	98.8	98.7	97.0	99.0	86.9	98.1	85.8	99.0	93.7	95.0

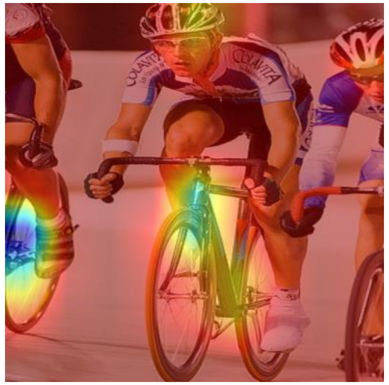


mAP

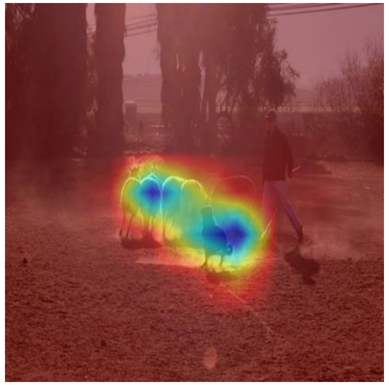
classes	AP(SSGRL)
aeroplane	0.9964
bicycle	0.9703
bird	0.9821
boat	0.9797
bottle	0.9001
bus	0.9823
car	0.9469
cat	0.9959
chair	0.8324
cow	0.9831
diningtable	0.7377
dog	0.9959
horse	0.9740
motorbike	0.9907
person	0.9756
pottedplant	0.8589
sheep	0.9878
sofa	0.7968
train	0.9985
tvmonitor	0.9066
mAP	0.9420

Implementation

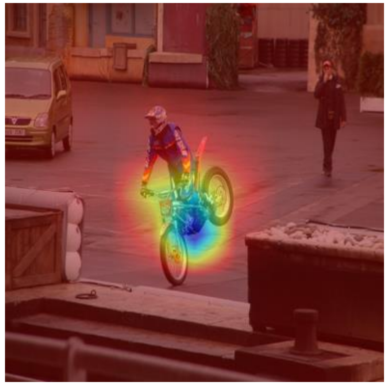
- Attention map results



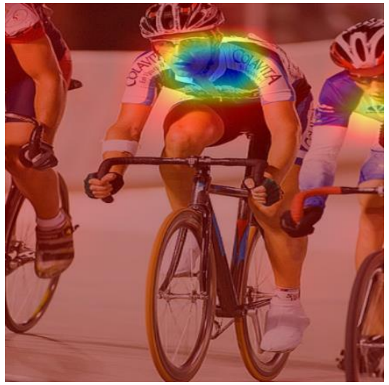
bicycle



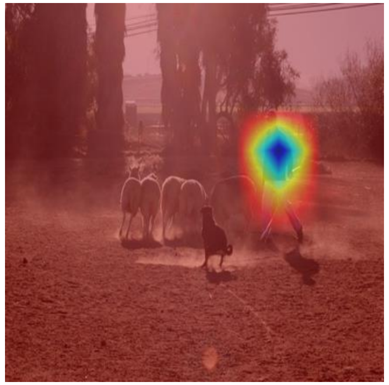
dog



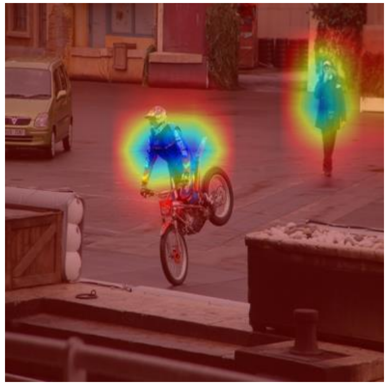
motorbike



person



person



person

Implementation

- If you want to train the model, I briefly explain about how to train and what to prepare in github page.
- All given results are also in github page as a README.md
- If you have any question about the implementations or need a pretrained model for PASCAL-VOC 2012 dataset, contact me and I’ll give any help.

- Thank you