Compositional Convolutional Neural Networks: A Robust and Interpretable Model for Object Recognition under Occlusion

Slides by Adam Kortylewski

Qing Liu

Yihong Sun

Angtian Wang

Ju He

Alan Yuille

Overview

- Generalization under Partial Occlusion

- A Deep Architecture with Innate Robustness to Partial Occlusion
  - A Generative Compositional Model of Neural Features
  - Robustness to Occlusion and Occluder Localization

- Robust Object Detection under Occlusion with CompositionalNets
  - Disentanglement of Context and Object Representation

- Conclusion

Motivation – Generalization under occlusion is important

- In natural images objects are surrounded and partially occluded by other objects
- Occluders are highly variable in terms of shape and texture -> **exponential complexity**
- Vision systems must generalize in exponentially complex domains

Motivation – A Fundamental Limitation of Deep Nets

- DCNNs do not generalize when trained with non-occluded data




Occ. Area	**0%**	**30%**	**50%**	**70%**	Avg
VGG -16	99.1	88.7	78.8	63.0	82.4

- What if we train with lots of augmented data?

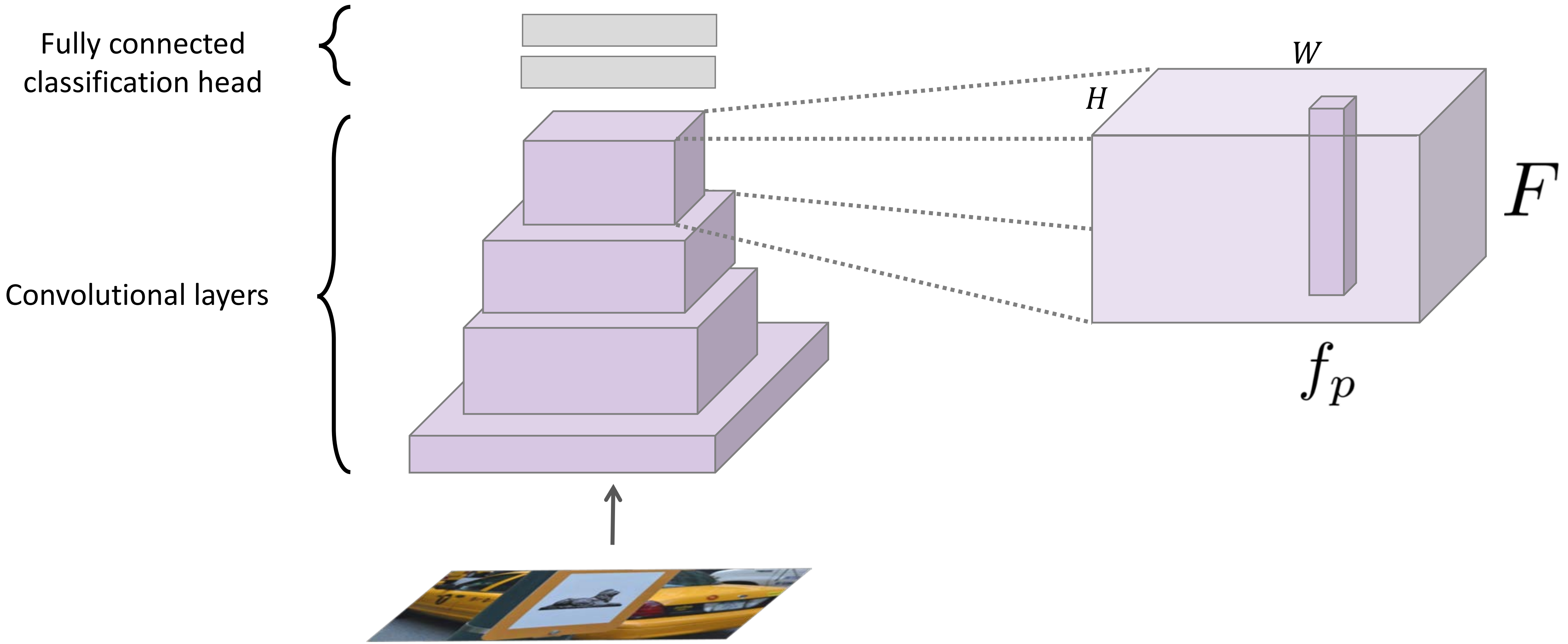


Occ. Area	**0%**	**30%**	**50%**	**70%**	Avg
VGG-16-Augmented	99.3	92.3	89.9	80.8	90.6

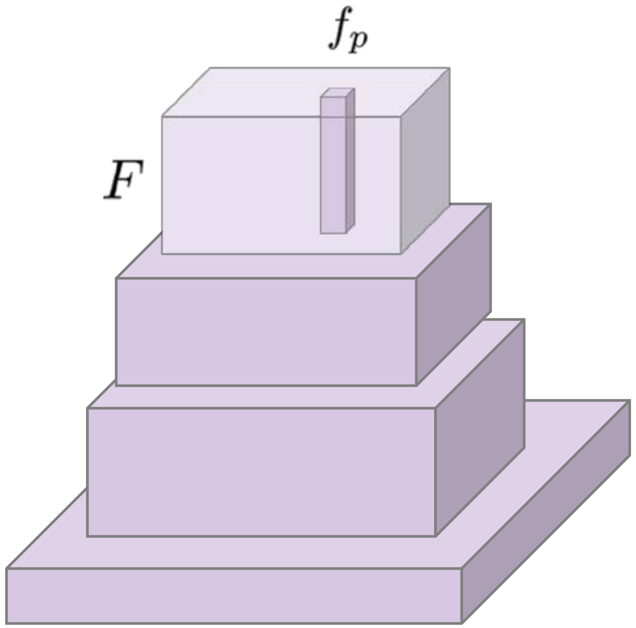
Overview

- Generalization under Partial Occlusion
- **A Deep Architecture with Innate Robustness to Partial Occlusion**
  - Generative Compositional Model of Neural Features
  - Robustness to occlusion and occluder localization
- Robust Object Detection under Occlusion with CompositionalNets
  - Disentanglement of Context and Object Representation
- Conclusion

A Generative Model of Neural Feature Activations



A Generative Model of Neural Feature Activations

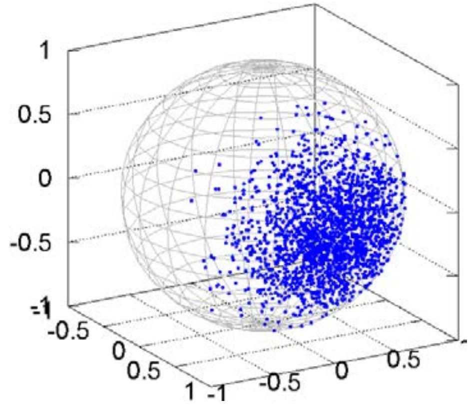


$$p(F|\Theta_y) = \sum_m \nu^m p(F|\theta_y^m)$$

$$p(F|\theta_y^m) = \prod_p p(f_p|\mathcal{A}_{p,y}^m, \Lambda)$$

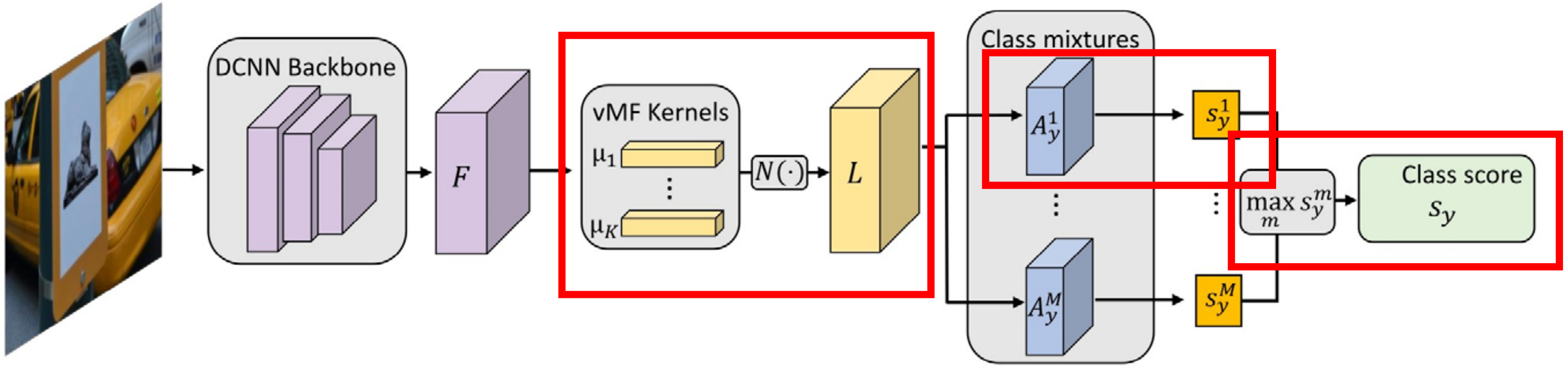
$$p(f_p|\mathcal{A}_{p,y}^m, \Lambda) = \sum_k \alpha_{p,k,y}^m p(f_p|\lambda_k), \quad \lambda_k = \{\mu_k, \sigma_k\}$$

$$p(f_p|\lambda_k) = \frac{e^{\sigma_k \mu_k^T f_p}}{Z(\sigma_k)}, \|f_p\| = 1, \|\mu_k\| = 1$$



Y labels object class
P labels position in the image
fp are the feature vectors at p
m label the mixture (viewpoint)
alpha's,lambda's,mu's are parameters which are learnt.

Inference as Feed-Forward Neural Network

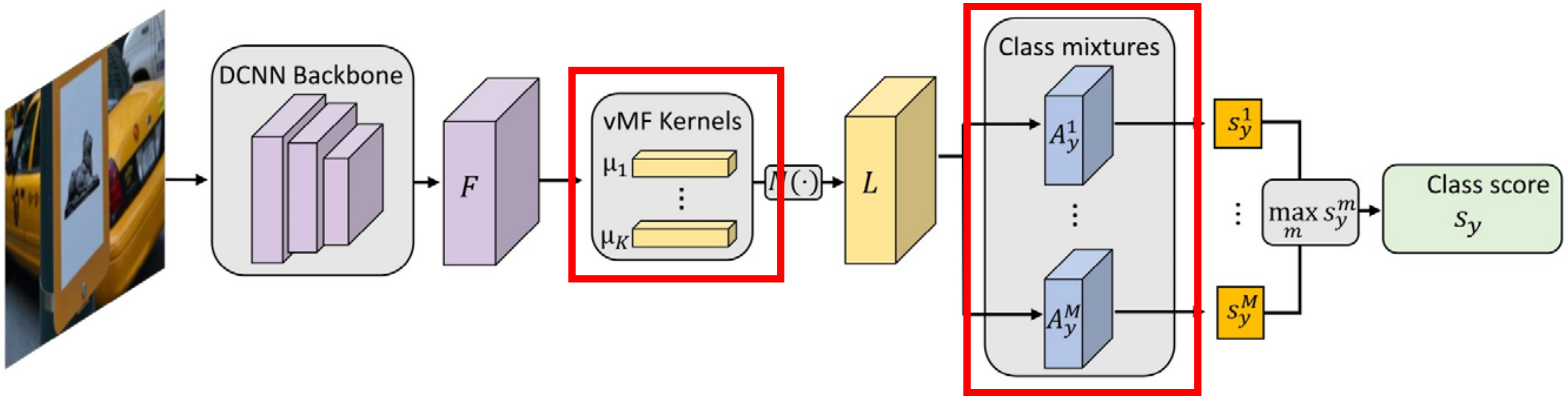


1. vMF likelihood: $$p(f_p|\lambda_k) = \frac{e^{\sigma_k \mu_k^T f_p}}{Z(\sigma_k)}, \|f_p\| = 1, \|\mu_k\| = 1$$

2. Mixture likelihoods: $$p(F|\theta_y^m) = \prod_p \sum_k \alpha_{p,k,y}^m p(f_p|\lambda_k)$$

3. Class score: $$p(F|\Theta_y) = \sum_m \nu^m p(F|\theta_y^m), \qquad \nu^m \in \{0,1\}, \sum_m \nu^m = 1$$

Learning the Model Parameters with Backpropagation

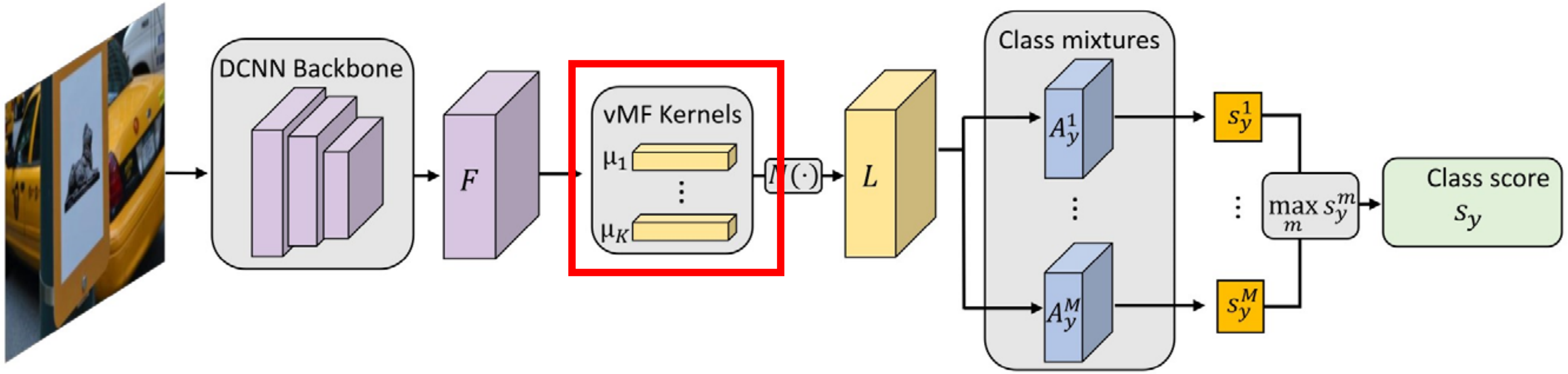


$$\mathcal{L} = \mathcal{L}_{class}(y, y') + \gamma_1 \mathcal{L}_{weight}(W) + \gamma_2 \mathcal{L}_{vmf}(F, \Lambda) + \gamma_3 \mathcal{L}_{mix}(F, \mathcal{A}_y)$$

$$\mathcal{L}_{vmf}(F, \Lambda) = -\sum_p \max_k \log p(f_p|\mu_k) = C \sum_p \min_k \mu_k^T f_p$$

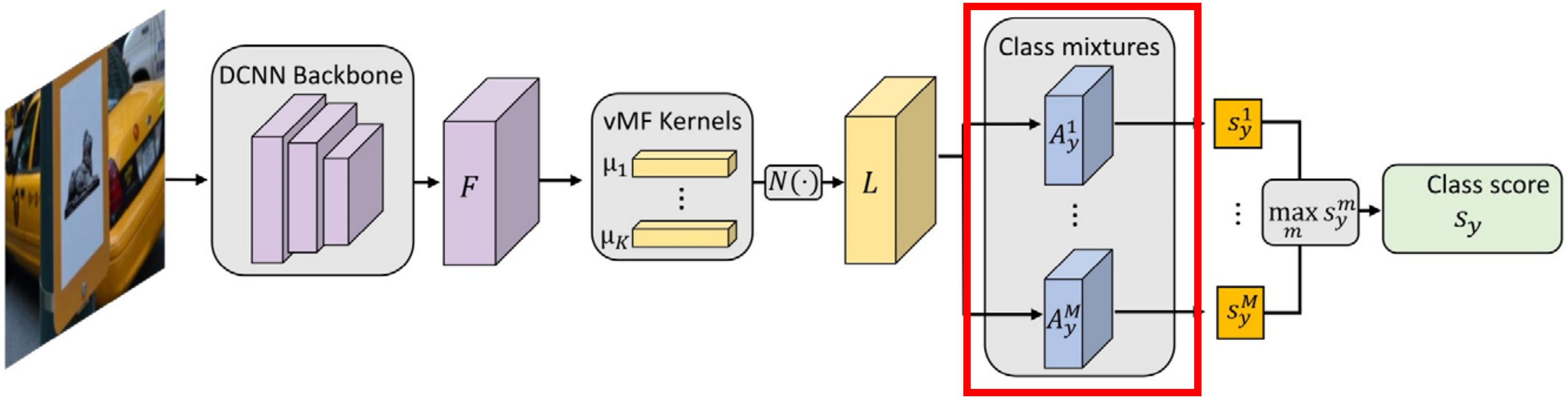
$$\mathcal{L}_{mix}(F, \mathcal{A}_y) = -\sum_p \log\Big[\sum_k \alpha^m_{p,k,y} p(f_p|\mu_k)\Big]$$

Explainability - vMF Kernels resemble „part detectors“



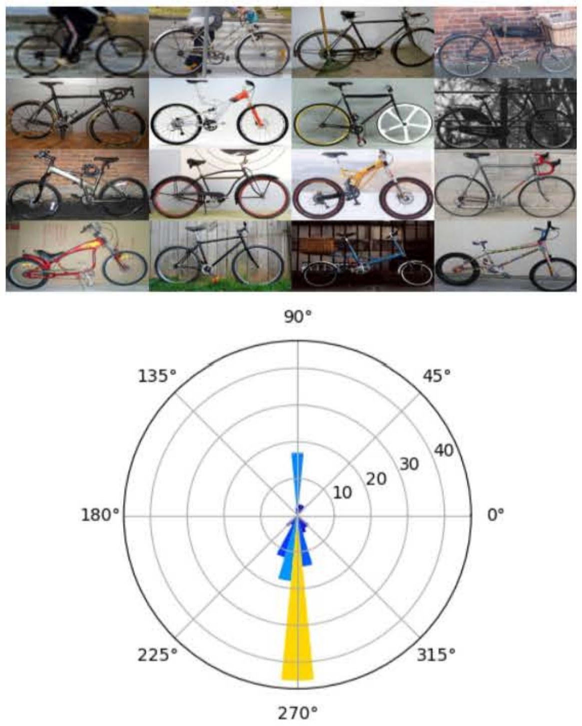
- Image patterns with highest likelihood:

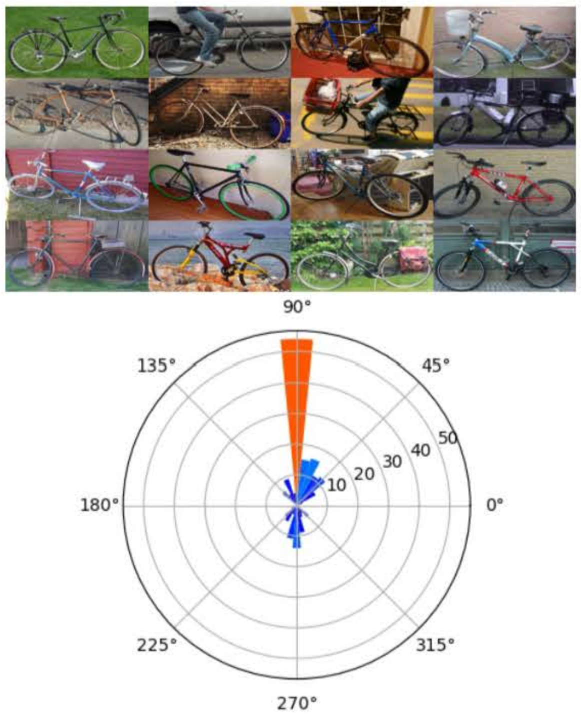
Explainability – Mixture components model object pose

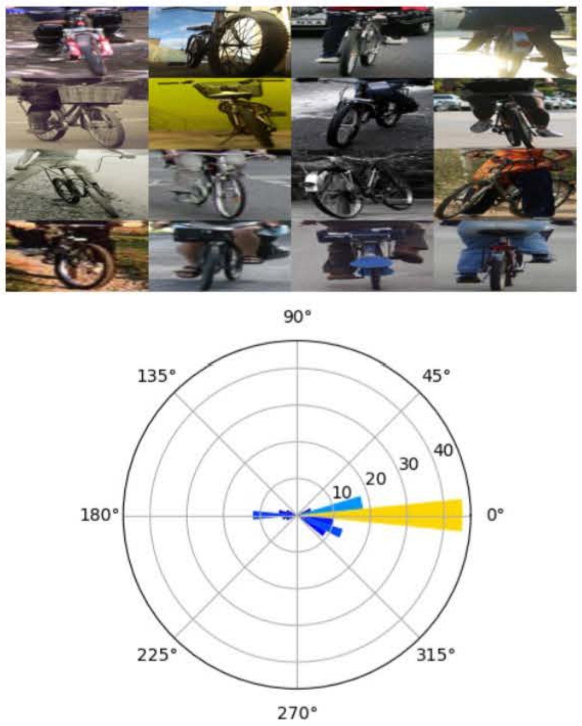


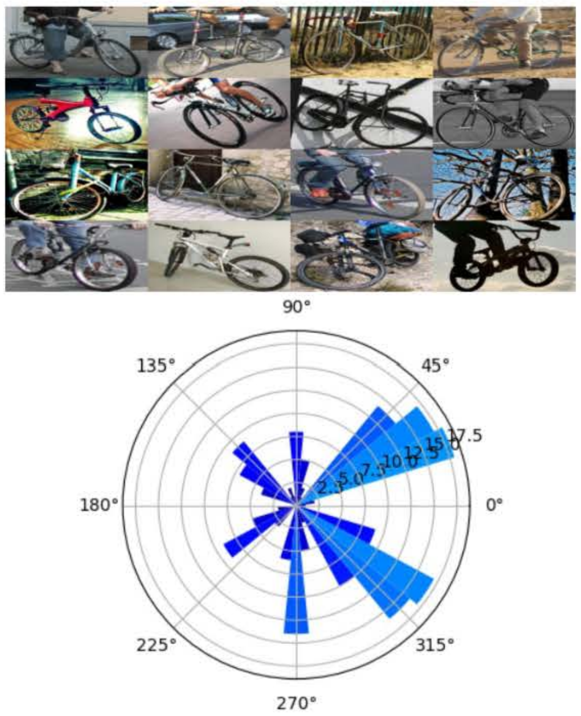
- Images with highest likelihood for mixture components:

Explainability – Mixture components model object pose

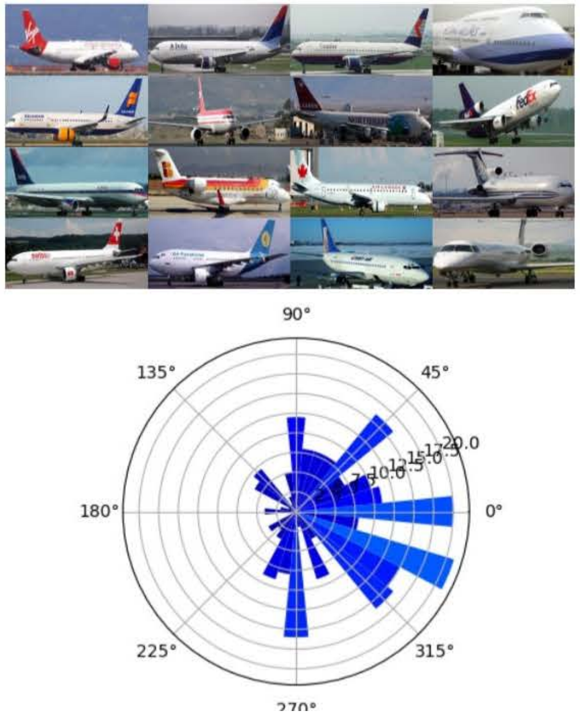


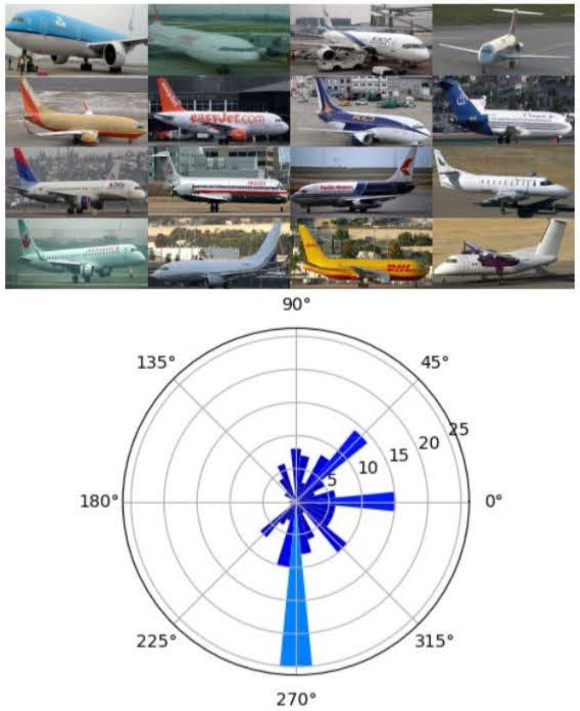


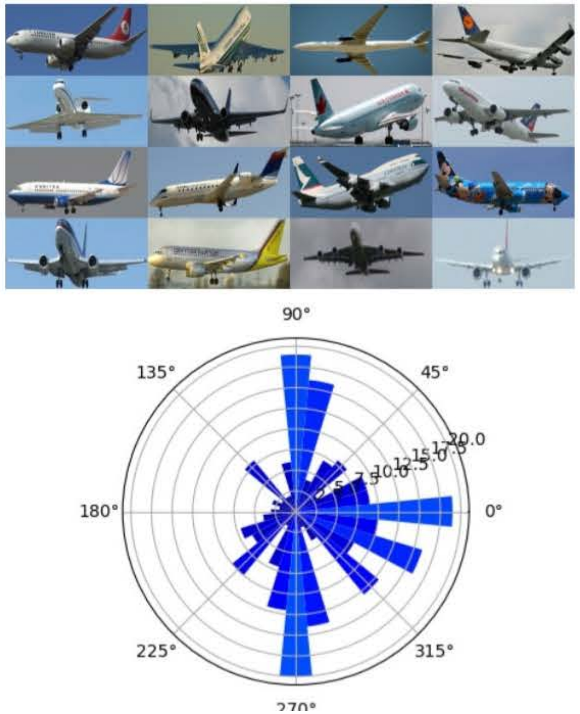


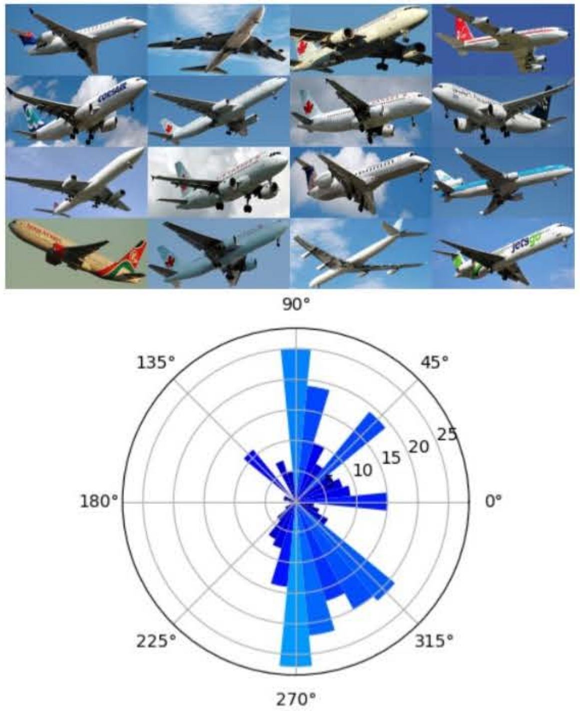


Explainability – Mixture components model object pose





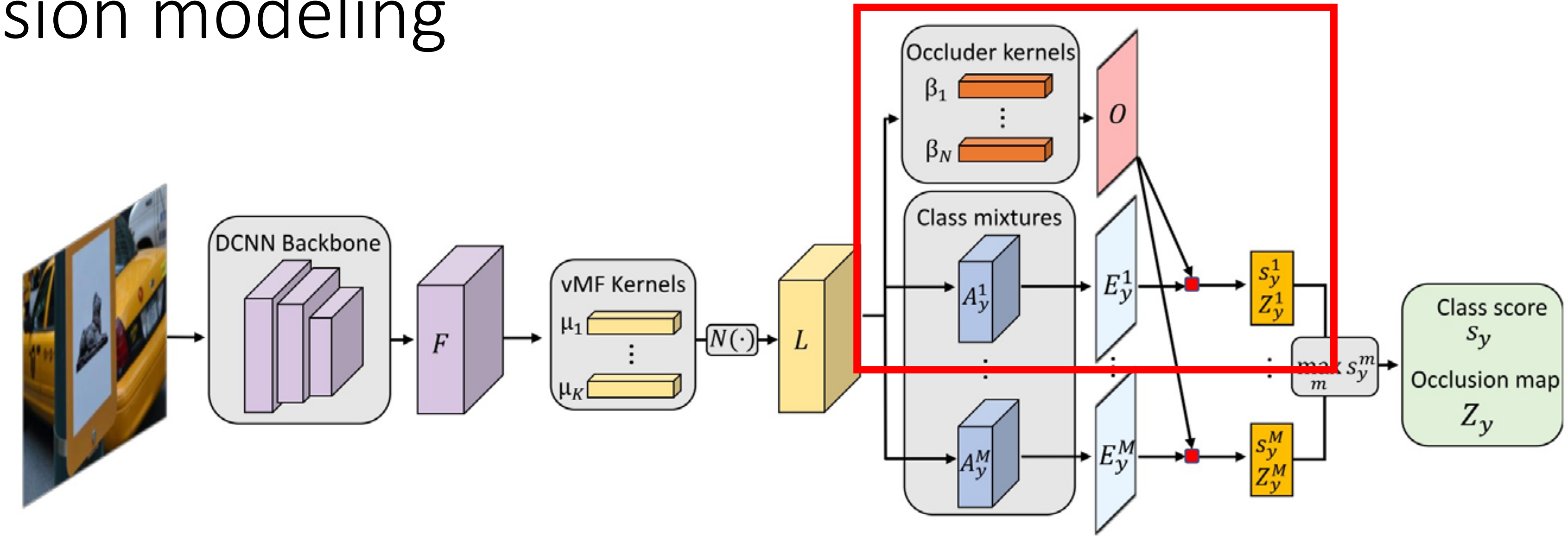




Overview

- Generalization under Partial Occlusion
- A Deep Architecture with Innate Robustness to Partial Occlusion
  - Generative Compositional Model of Neural Features
  - **Robustness to occlusion and occluder localization**
- Robust Object Detection under Occlusion with CompositionalNets
  - Disentanglement of Context and Object Representation
- Conclusion

Occlusion modeling



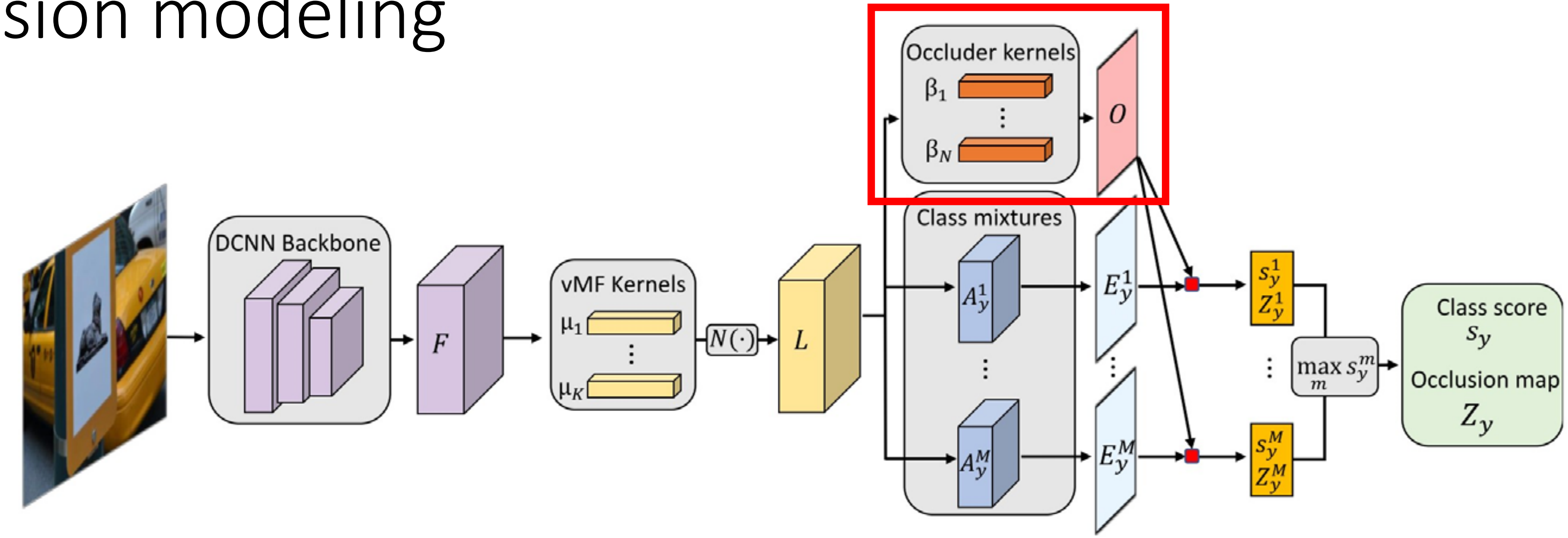
- We introduce an outlier model:

$$p(F|\theta_y^m, \beta) = \prod_p \underline{p(f_p, z_p^m = 0)}^{1-z_p^m} \underline{p(f_p, z_p^m = 1)}^{z_p^m}, \quad \mathcal{Z}^m = \{z_p^m \in \{0,1\} | p \in \mathcal{P}\}$$

$$\underline{p(f_p, z_p^m = 1)} = p(f_p|\beta, \Lambda)\ p(z_p^m = 1),$$

$$\underline{p(f_p, z_p^m = 0)} = p(f_p|\mathcal{A}_{p,y}^m, \Lambda)\ (1 - p(z_p^m = 1)).$$

Occlusion modeling

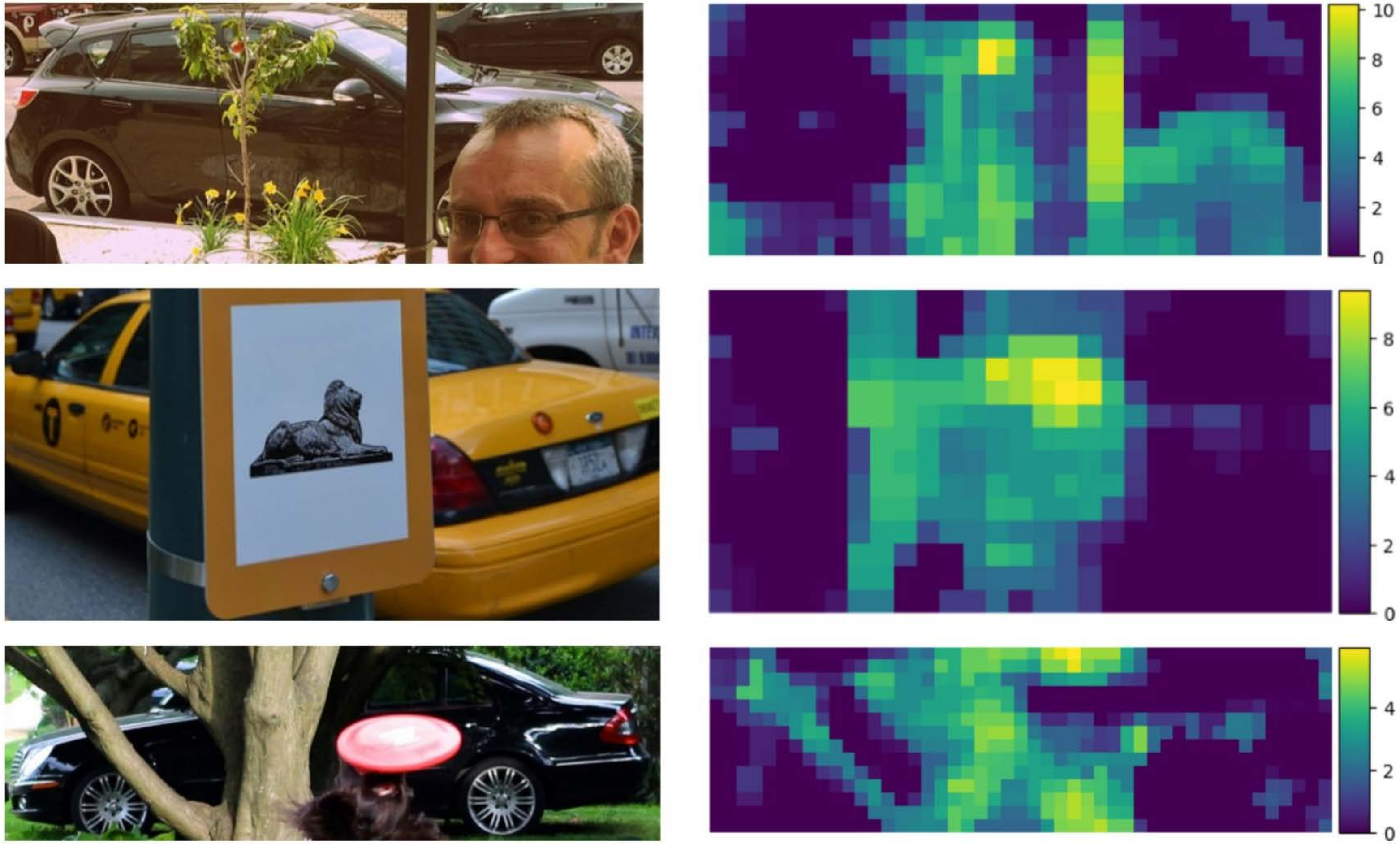


- We introduce an outlier model:

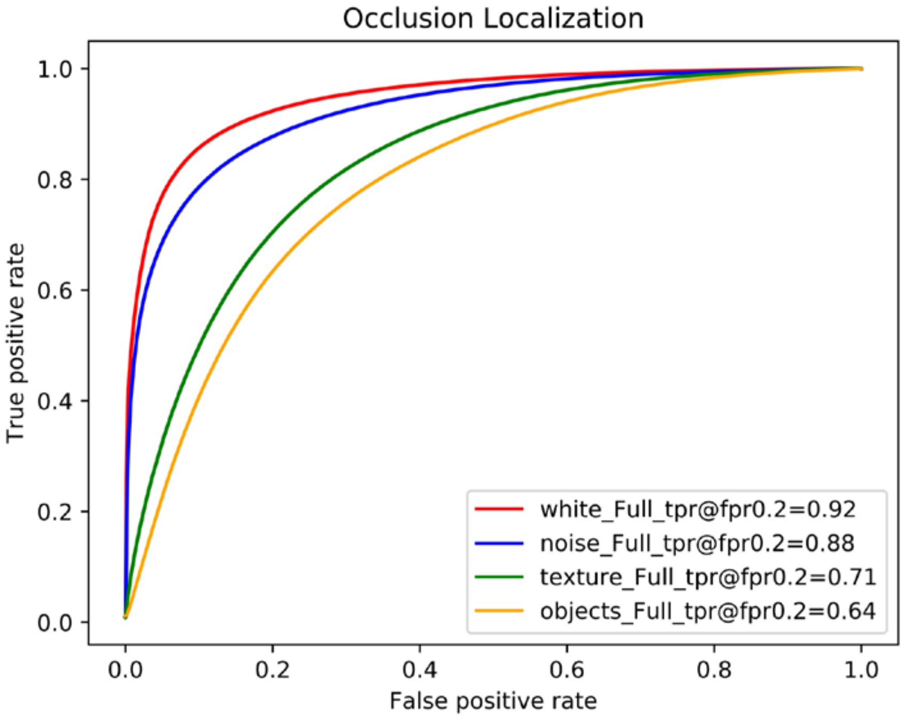
$$p(F|\theta_y^m, \beta) = \prod_p p(f_p, z_p^m = 0)^{1-z_p^m} \underline{p(f_p, z_p^m = 1)}^{z_p^m}, \quad \mathcal{Z}^m = \{z_p^m \in \{0,1\} | p \in \mathcal{P}\}$$

- A simple model of how the object does not look like:

Competition between object and outlier model



Quantitative Evaluation of Occluder Localization

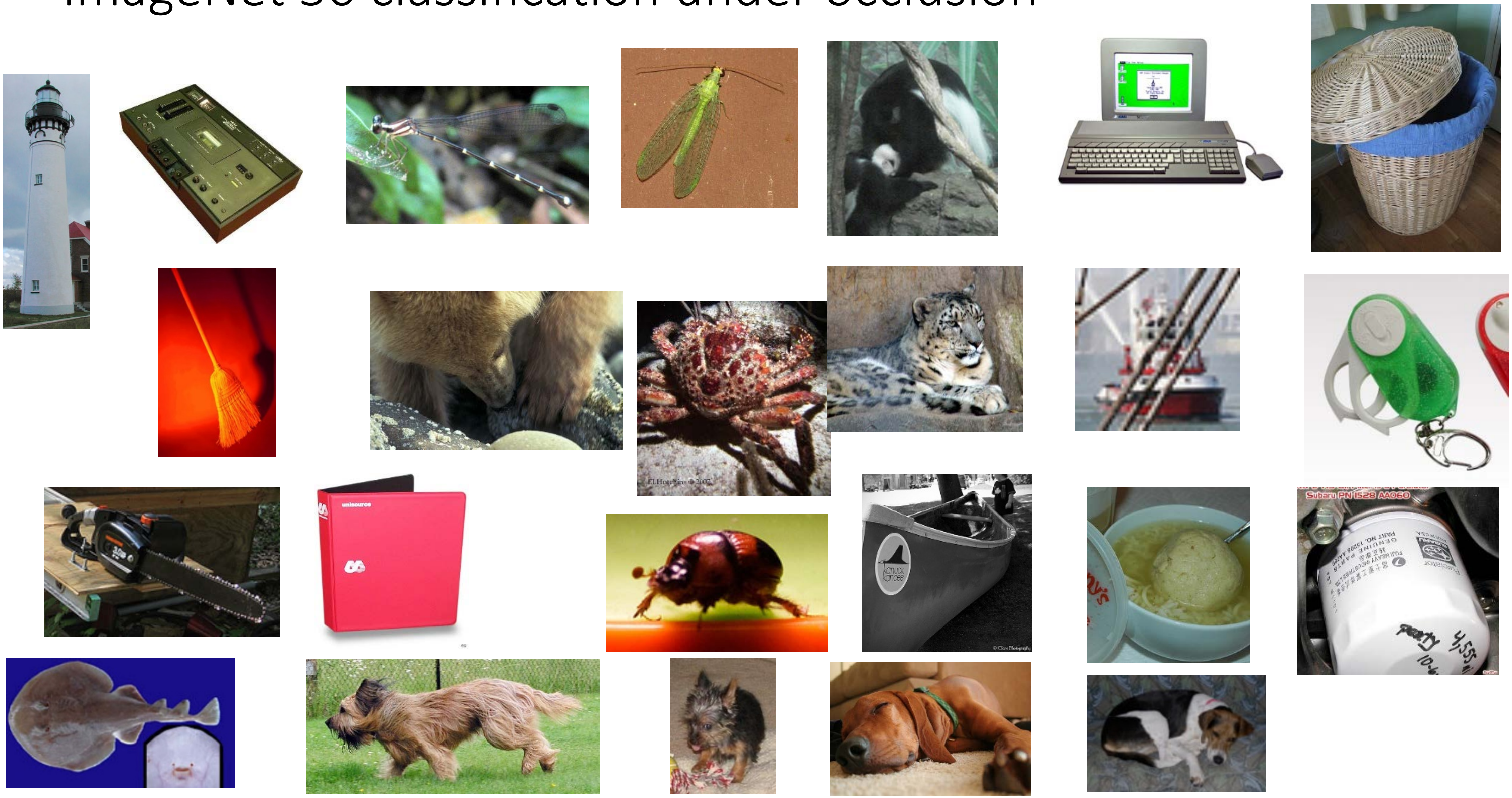


CompNets can classify partially occluded vehicles robustly

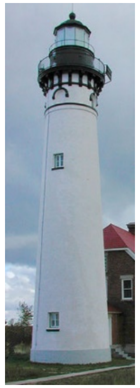



Occ. Area	**L0**	**L1**	**L2**	**L3**	Avg
VGG	97.8	86.8	79.1	60.3	81.0
ResNet50	98.5	89.6	84.9	71.2	86.1
ResNext	98.7	90.7	85.9	75.3	87.7

ImageNet 50 classification under occlusion



ImageNet 50 classification under occlusion









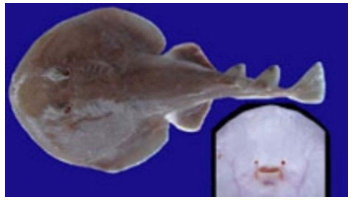






ImageNet under Occlusion

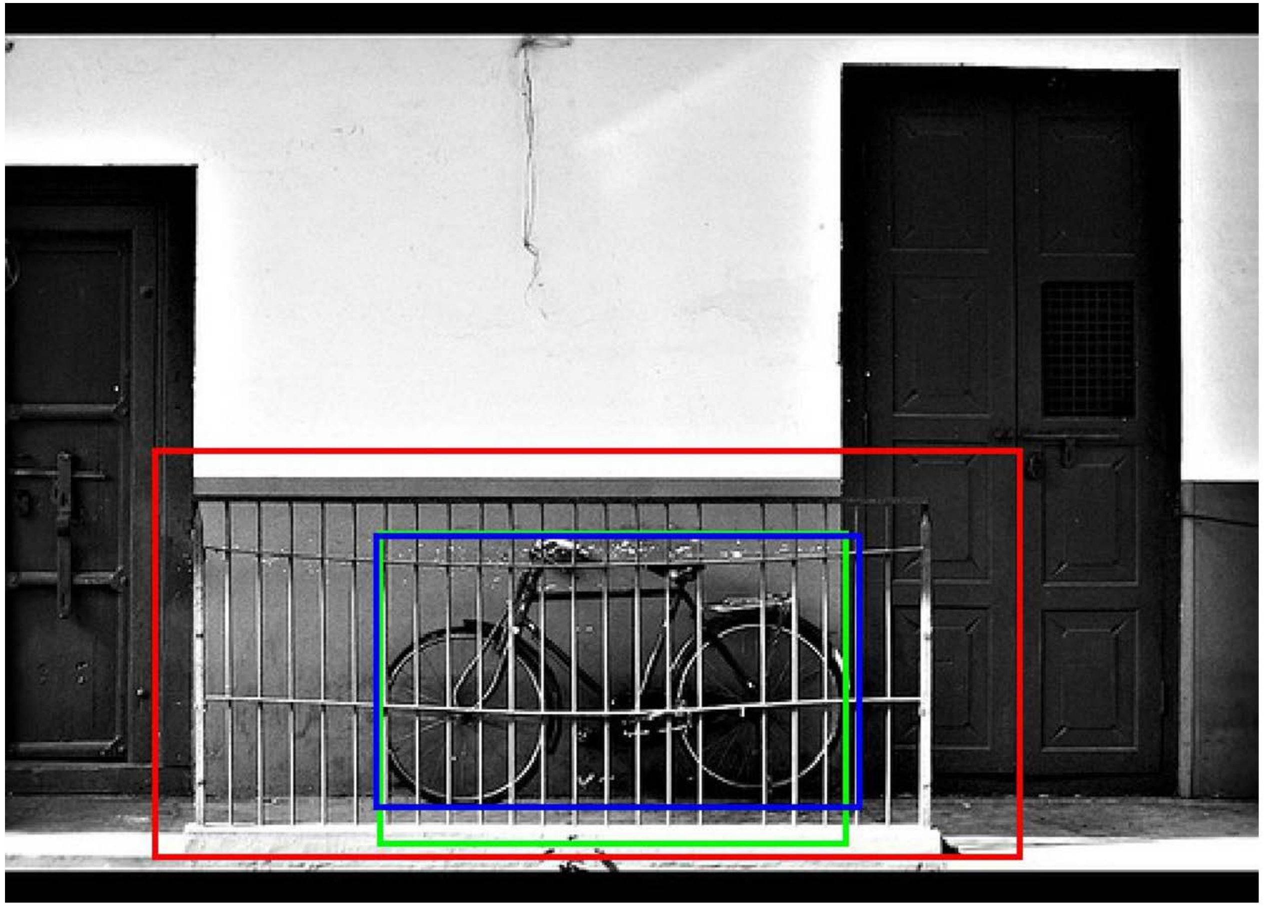
Occ. Area	0%	30%	50%	70%	Avg
ResNext	**98.4**	69.3	48.7	31	61.9
CompNet-ResNext	96.3	**76.6**	**60.1**	**45.5**	**69.6**



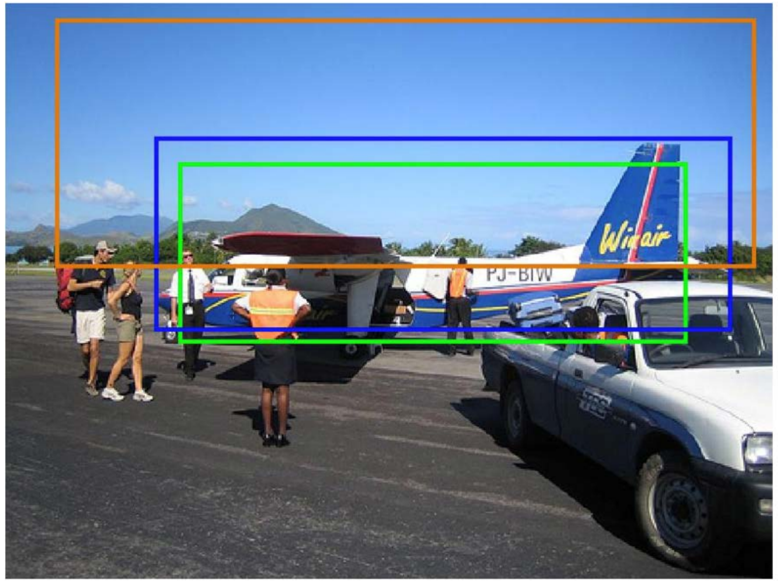
Overview

- Generalization under Partial Occlusion
- A Deep Architecture with Innate Robustness to Partial Occlusion
  - Generative Compositional Model of Neural Features
  - Robustness to occlusion and occluder localization
- **Robust Object Detection under Occlusion with CompositionalNets**
  - Disentanglement of Context and Object Representation
- Conclusion

DCNNs for object detection also do not generalize well



Context has too much influence when object is occluded

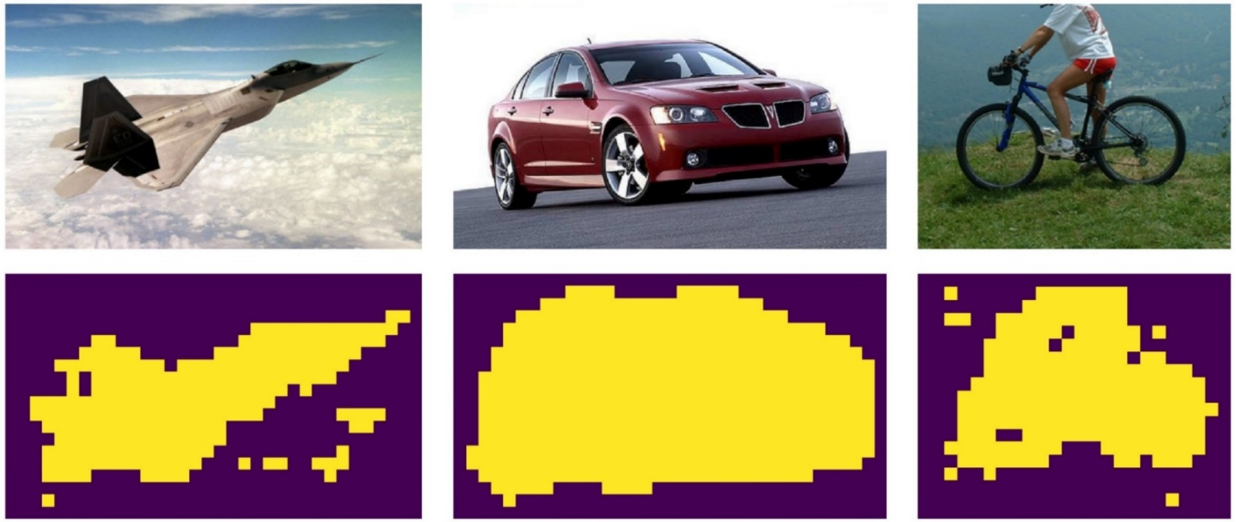


Seperate the representation of context and object

- We introduce a context-aware object model:

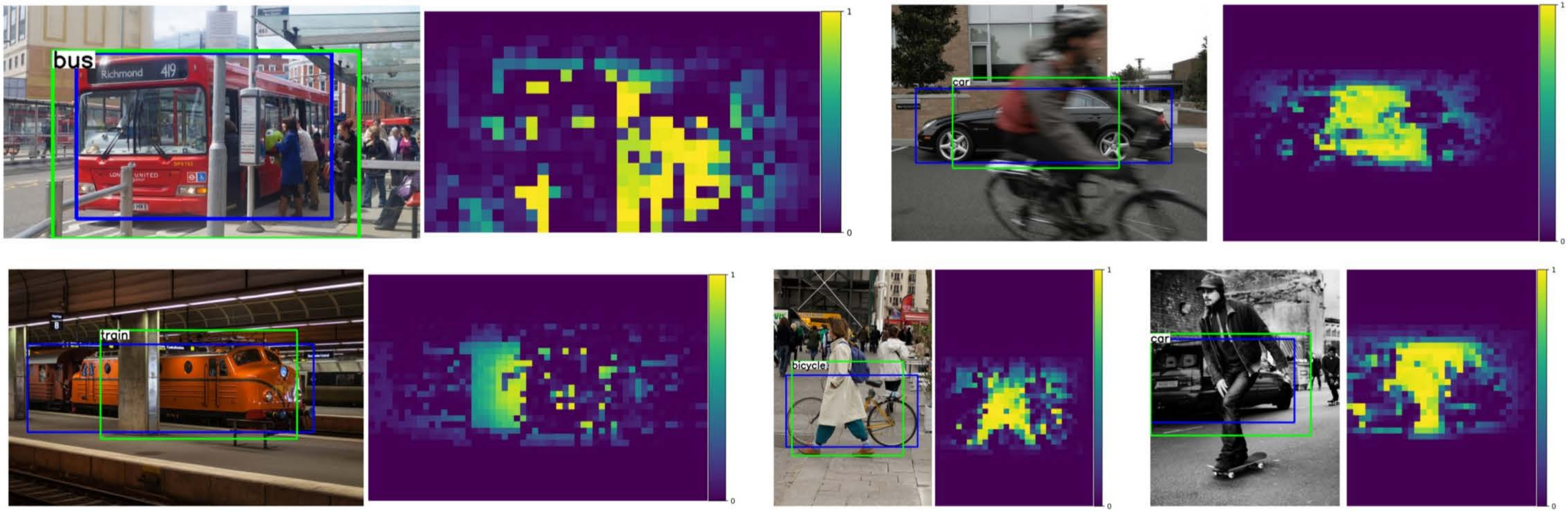
$$p(f_p|\mathcal{A}^m_{p,y}, \chi^m_{p,y}, \Lambda) = \omega\, p(f_p|\chi^m_{p,y}, \Lambda) + (1-\omega) p(f_p|\mathcal{A}^m_{p,y}, \Lambda)$$

- Segment the image during training:



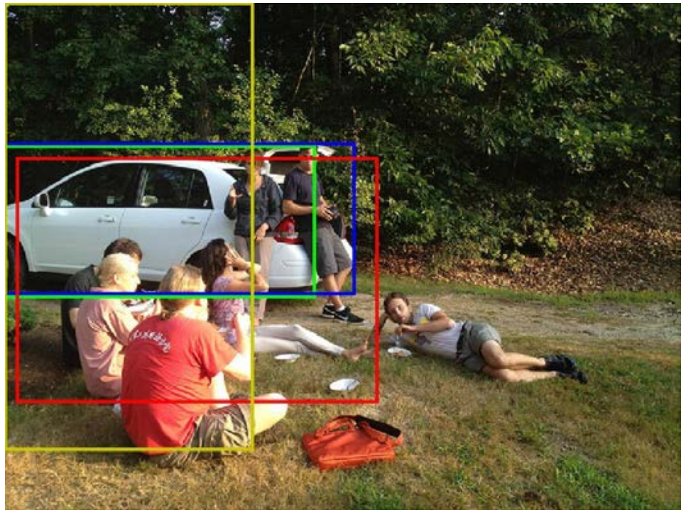
Context-awareness Improves Localization

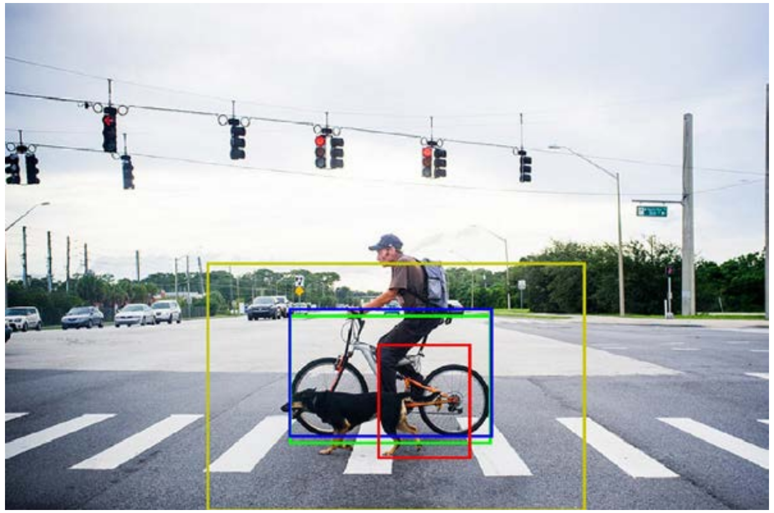
Explainability- Occluder localization in Object detection

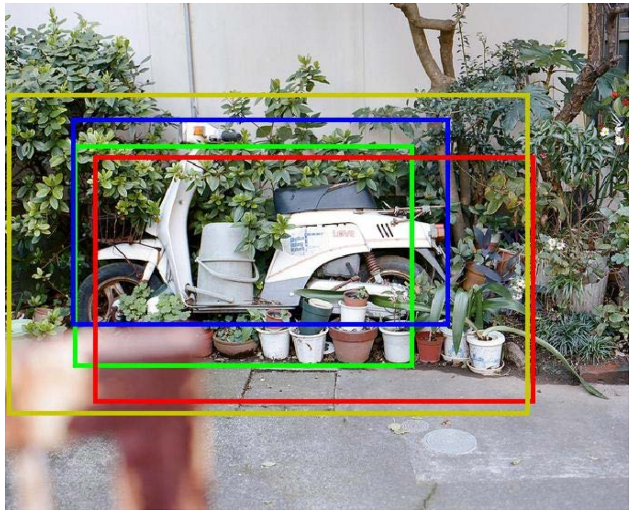


Detection Results

method	light occ.	heavy occ.
Faster R-CNN	73.8	55.2
Faster R-CNN with reg.	74.4	56.3
Faster R-CNN with occ.	77.6	62.4
CA-CompNet via BBV $\omega = 0.5$	78.6	76.2
CA-CompNet via BBV$\omega = 0.2$	**87.9**	**78.2**
CA-CompNet via BBV $\omega = 0$	85.6	75.9







Conclusion

- Partial occlusion introduces exponential complexity in the data
- The complexity gap can be overcome by introducing prior knowledge about compositionality, partial occlusion and context into the neural architecture
- Generalization beyond the training data in terms of partial occlusion & context
- Retain high discriminative performance due to end-to-end training
- Future work: Articulated objects, 3D geometry, top-down reasoning, scale, …