Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization

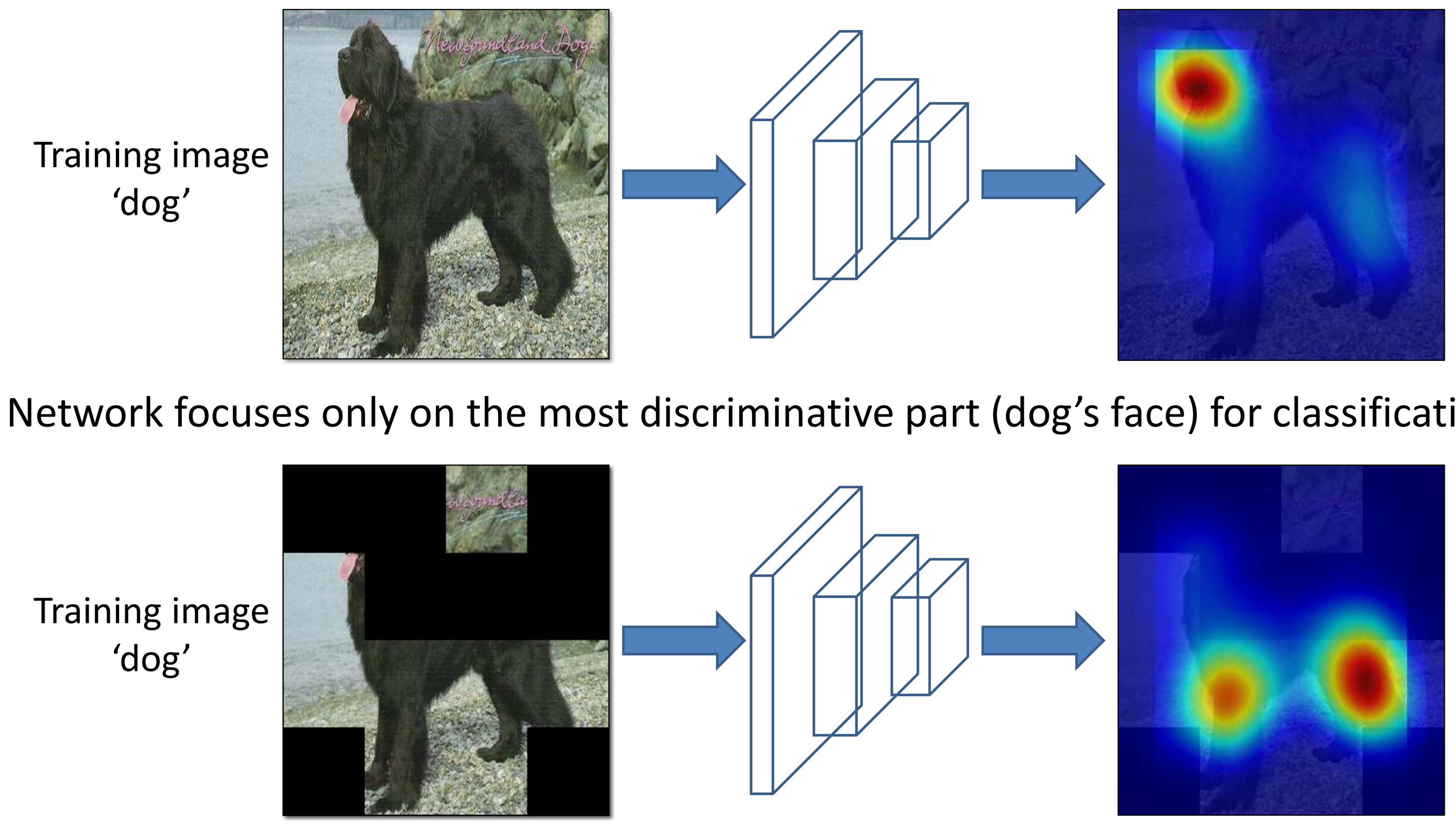
UC DAVIS UNIVERSITY OF CALIFORNIA

Krishna Kumar Singh   Yong Jae Lee

ICCV17

Motivation and Idea:

Goal: Improve object localization capability of image classification networks.



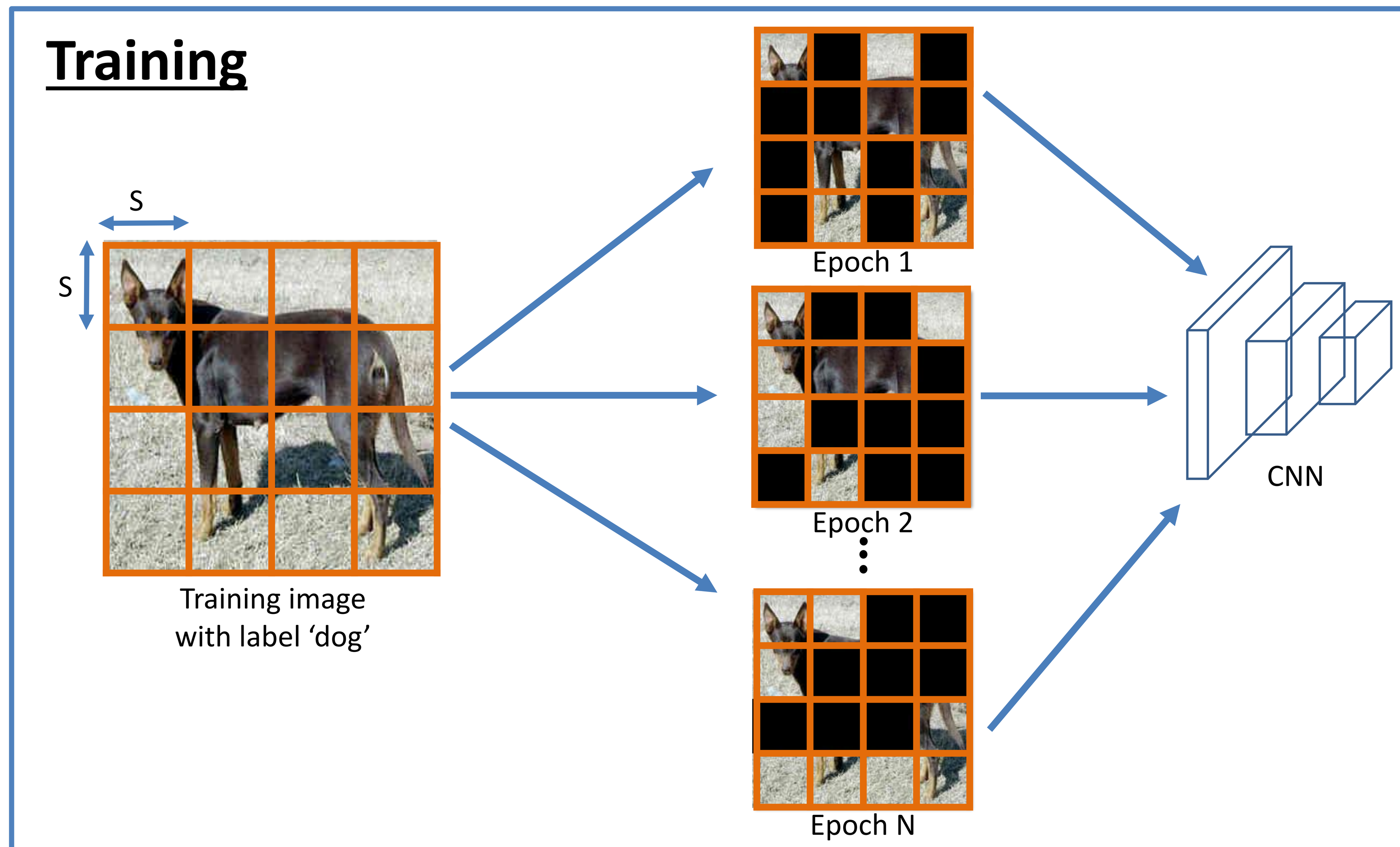
Network focuses only on the most discriminative part (dog's face) for classification

Hide patches to force the network to *seek* other relevant parts

- Existing object localization methods (e.g., Oquab 2015, Zhou 2016) tend to localize only the most discriminative part.
- Masking image pixels has been used for visualizing CNNs (Zeiler 2014), semantic segmentation (Wei 2017), and generating occlusion training examples (Wang 2017).

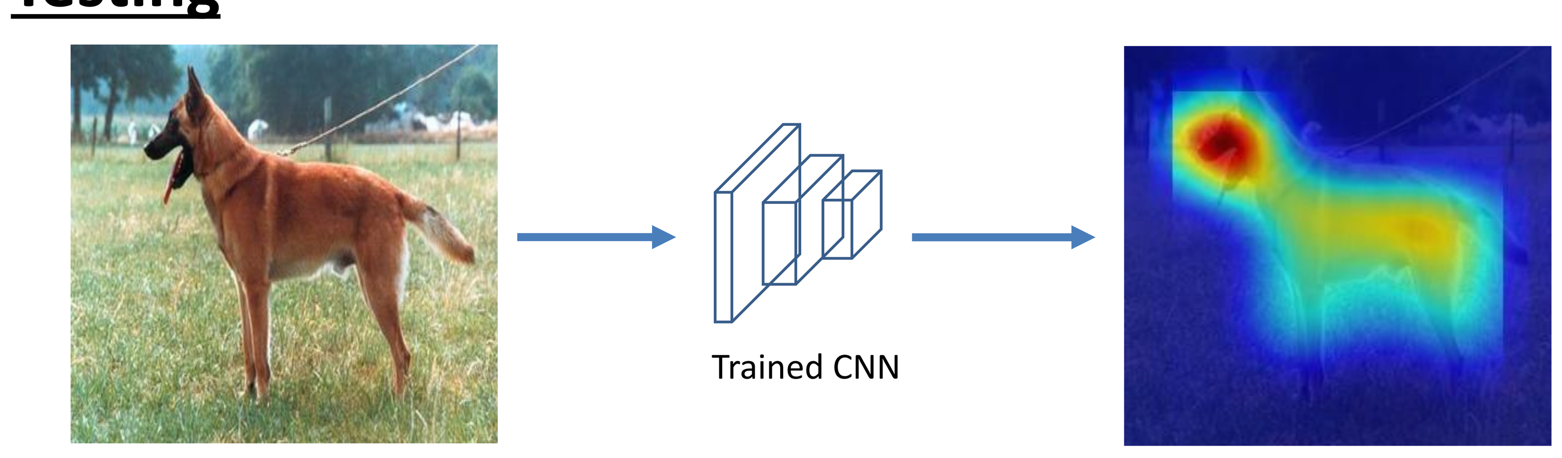
Approach:

Training



- For the same image, we randomly hide a different set of patches in each training epoch.
- This allows the network to learn multiple relevant object parts for each image.

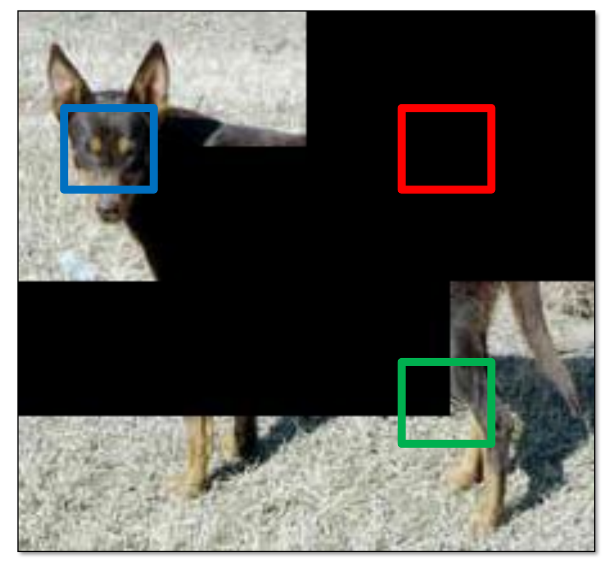
Testing



Class Activation Map (CAM)
Predicted label: 'dog'

- During testing, the full image without any hidden patches is given as input.

Assigning value to hidden pixels:

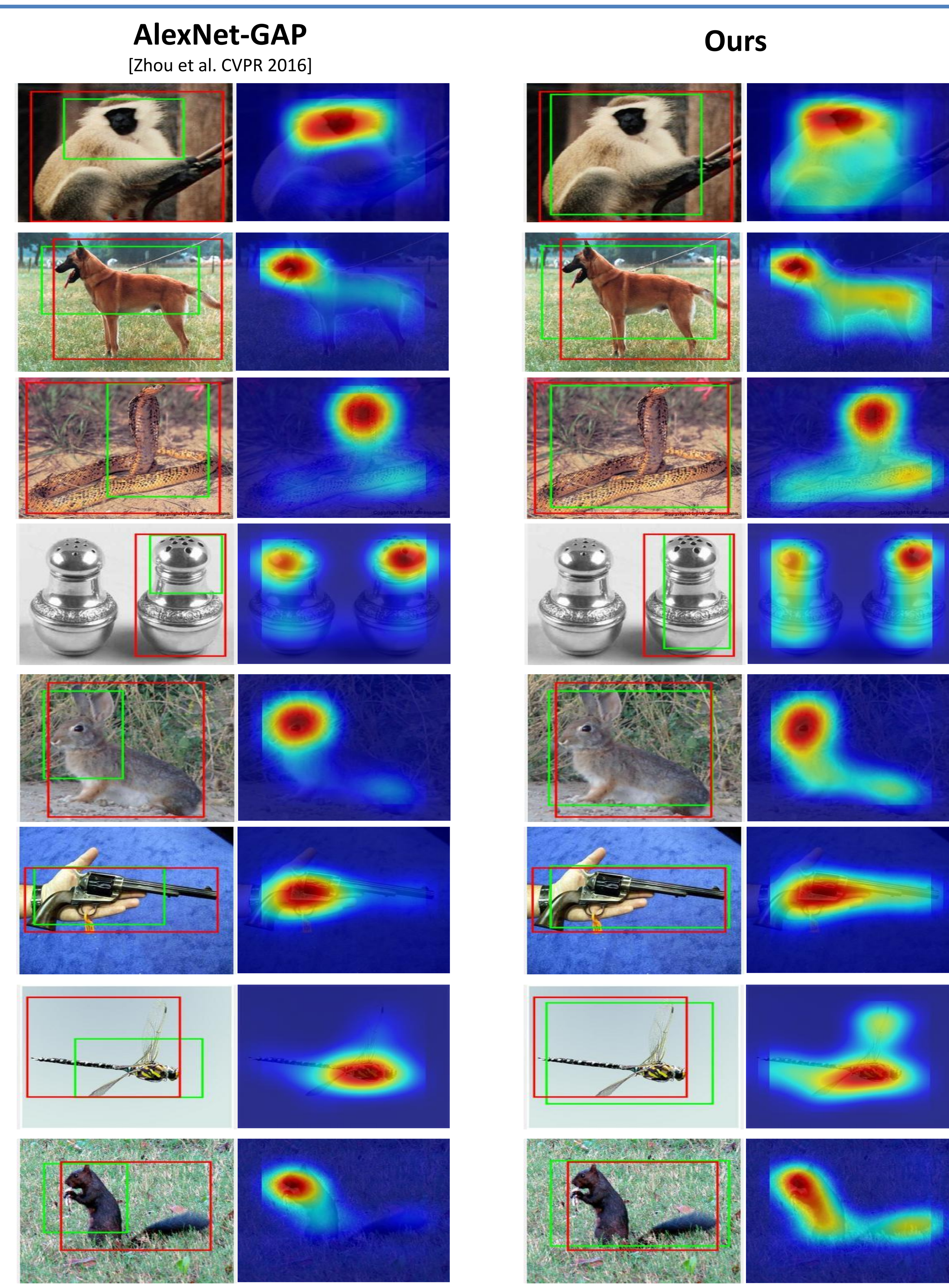


- Patches are hidden only during training; during testing full image is given as input.
- Activations of 1st conv layer will have different distribution during training and testing.
- Assigning μ (mean RGB value of all pixels in dataset) to each hidden pixel ensures same activation (in expectation) during training and testing:

$$\mathbb{E}[\sum_{i=1}^{k\times k} \mathbf{w}_i^\top \mathbf{x}_i] = \sum_{i=1}^{k\times k} \mathbf{w}_i^\top \mu$$

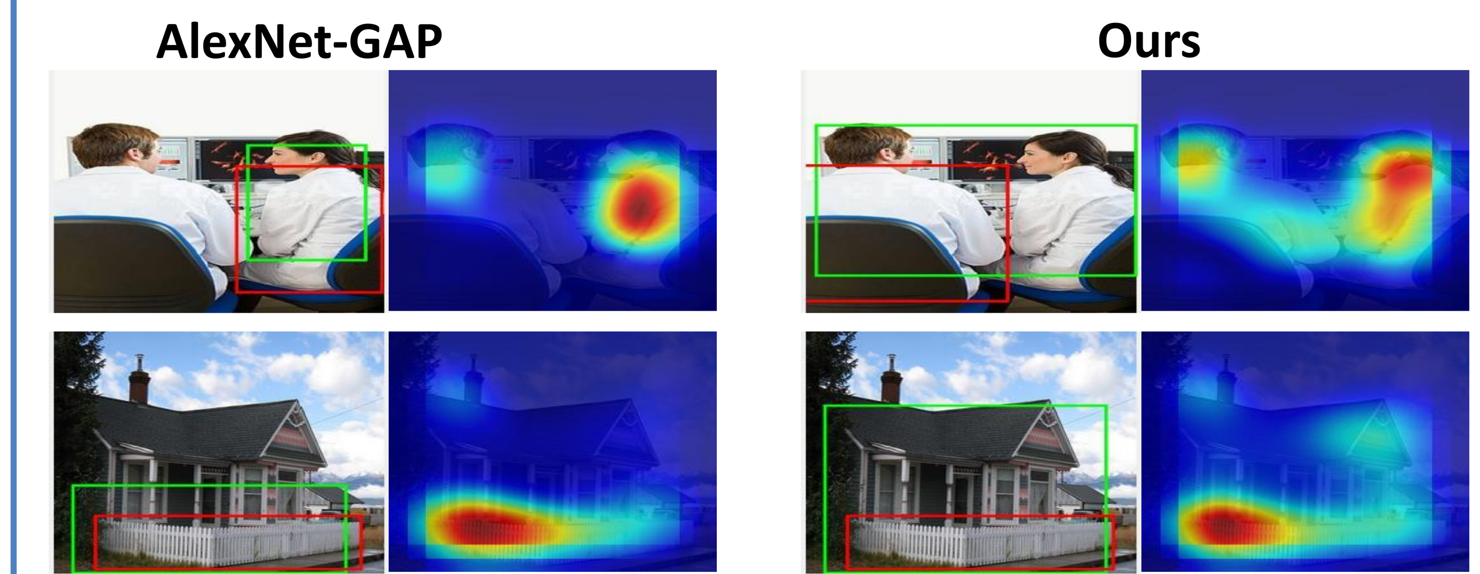
Inside visible patch | Inside hidden patch | Partially in hidden patch

Qualitative Results:



- For each image, we show the bounding box and CAM (Class Activation Map) obtained by AlexNet-GAP (left) and our method (right).
- Ground-truth and predicted boxes are in red and green, respectively.
- Our approach localizes more relevant parts.

Failure Cases:



- Merges spatially close instances together (first row).
- Localizes co-occurring context of a class (second row).

Quantitative Results:

Object localization results on ImageNet validation data:

Methods	GT-known Loc (AlexNet)	Top-1 Loc (AlexNet)	GT-known Loc (GoogLeNet)	Top-1 Loc (GoogLeNet)
Backprop [Simonyan 2014]	-	34.83	-	38.69
GAP [Zhou 2016]	54.90	36.25	58.41	43.60
Ours	**58.68**	**37.65**	**60.29**	**45.21**

- Hiding patches during training leads to better object localization results.
- Our approach generalizes across different networks.

Comparison with dropout:

Methods	GT-known Loc	Top-1 Loc
Ours	**58.68**	**37.65**
AlexNet-dropout-trainonly	42.17	7.65
AlexNet-dropout-traintest	53.48	31.68

Results with different patch sizes:

Methods	GT-known Loc	Top-1 Loc
AlexNet-GAP	54.90	36.25
AlexNet-HaS-16	57.86	36.77
AlexNet-HaS-32	**58.75**	37.33
AlexNet-HaS-44	58.55	37.54
AlexNet-HaS-56	58.43	37.34
AlexNet-HaS-mix	58.68	**37.65**

- In Dropout, units in a layer are dropped randomly, while in our work, contiguous image regions are dropped.
- Our method of hiding patches performs better than dropout on input image.
- Hiding patches of mixed sizes gives best *Top-1 Loc* accuracy.

Action localization results on THUMOS 2014:

Methods	IOU thresh = 0.1	0.2	0.3	0.4	0.5
Video-GAP	34.23	25.68	17.72	11.00	6.11
Ours	**36.44**	**27.84**	**19.49**	**12.66**	**6.84**

- Hiding the frames during training leads to better action localization results.

Pre-training with Hide-and-Seek for image segmentation :

Methods	Pixel acc.	Mean acc.	Mean IU	f.w. IU
AlexNet	85.58	63.01	48.00	76.26
AlexNet (with Hide and Seek)	**86.24**	**63.58**	**49.31**	**77.11**

- Pre-training the AlexNet with Hide-and-Seek gives better segmentation results.

Conclusions:

- Simple idea of Hide-and-Seek to improve weakly-supervised object and action localization.
- Only need to change the input without modifying the network.
- Generalizes to multiple network architectures, input data, and tasks.

Acknowledgement: This work was supported in part by Intel Corp, Amazon Web Services Cloud Credits for Research, and GPUs donated by NVIDIA.