Deformable ConvNets v2: More Deformable, Better Results

Zhu, Xizhou, Han Hu, Stephen Lin, and Jifeng Dai. "Deformable ConvNets v2: More Deformable, Better Results." *arXiv preprint arXiv:1811.11168* (2018).

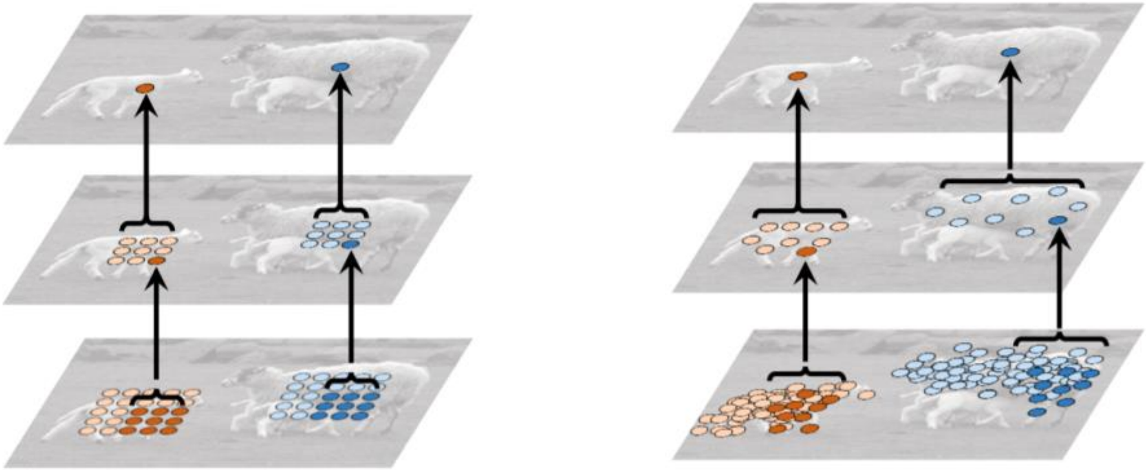
Shared by Tao Kong

Outline

- Quick survey of Deformable ConvNets – 5 pages
- Analysis of Deformable ConvNet v1 Behavior – 6 pages
- Deformable ConvNets V2 – 17 pages

Quick survey of Deformable ConvNets

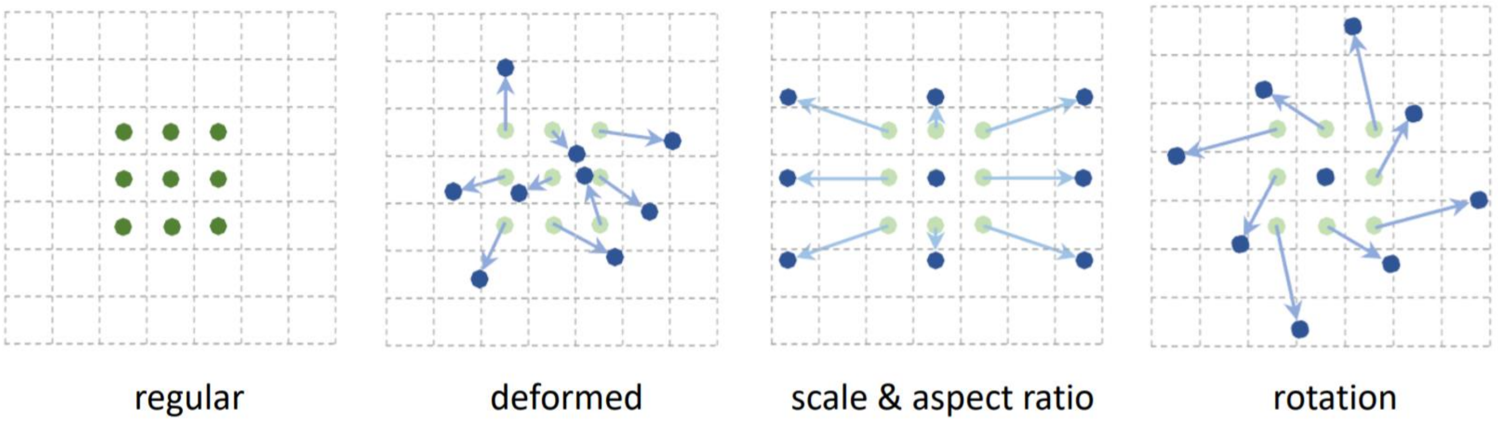
- Geometric variations due to scale, pose, viewpoint and part deformation present a major challenge in object recognition and detection.



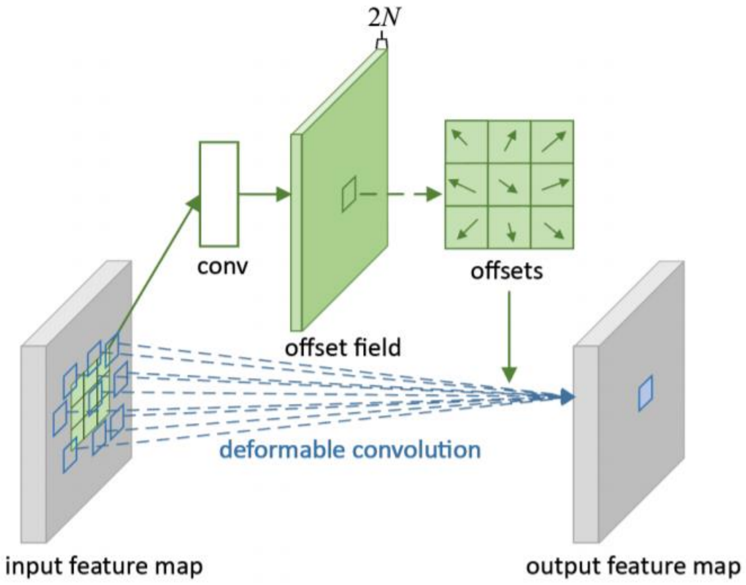
(a) standard convolution (b) deformable convolution

Quick survey of Deformable ConvNets

Learning to deform the sampling locations in the convolution/RoI Pooling modules



Quick survey of Deformable ConvNets



Regular convolution

$$\mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n)$$

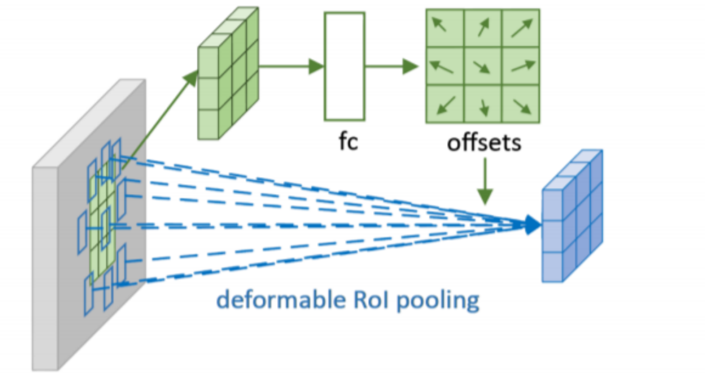
Deformable convolution

$$\mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n + \Delta\mathbf{p}_n)$$

where $\Delta\mathbf{p}_n$ is generated by a sibling branch of regular convolution

The grid sampling locations of standard convolution are each offset by displacements learned with respect to the preceding feature maps.

Quick survey of Deformable ConvNets



deformable RoI Pooling

Regular RoI pooling

$$\mathbf{y}(i,j) = \sum_{\mathbf{p} \in bin(i,j)} \mathbf{x}(\mathbf{p}_0 + \mathbf{p})/n_{ij}$$

Deformable RoI pooling

$$\mathbf{y}(i,j) = \sum_{\mathbf{p} \in bin(i,j)} \mathbf{x}(\mathbf{p}_0 + \mathbf{p} + \Delta\mathbf{p}_{ij})/n_{ij}$$

where $\Delta\mathbf{p}_{ij}$ is generated by a sibling fc branch

Offsets are learned for the bin positions in RoIpooling

Quick survey of Deformable ConvNets

- Same input & output as the plain versions
  - Regular convolution -> deformable convolution
  - Regular RoI pooling -> deformable RoI pooling
- End-to-end trainable
- Gives the network more ability to adapt its feature representation to the configuration of an object, specifically by deforming its sampling and pooling patterns to fit the object’s structure

ConvNet v1 **Behavior on Spatial Deformation**

- Effective receptive fields
  - whose values are calculated as the gradient of the node response with respect to intensity perturbations of each image pixel

- Effective sampling / bin locations
  - the gradient of the network node with respect to the sampling / bin locations

- Error-bounded saliency regions
  - the smallest image region giving the same response as the full image, within a small error bound.

The spatial support of network nodes

Effective sampling locations

Error-bounded saliency regions

(a) regular conv

The spatial support of network nodes

Effective sampling locations

Error-bounded saliency regions

(b) deformable conv@conv5 stage (DCNv1)

The spatial support of network nodes

aligned
RoIpooling

Effective sampling
locations

Error-bounded
saliency regions

(a) aligned RoIpooling, with deformable conv@conv5 stage

The spatial support of network nodes

deformable RoIpooling

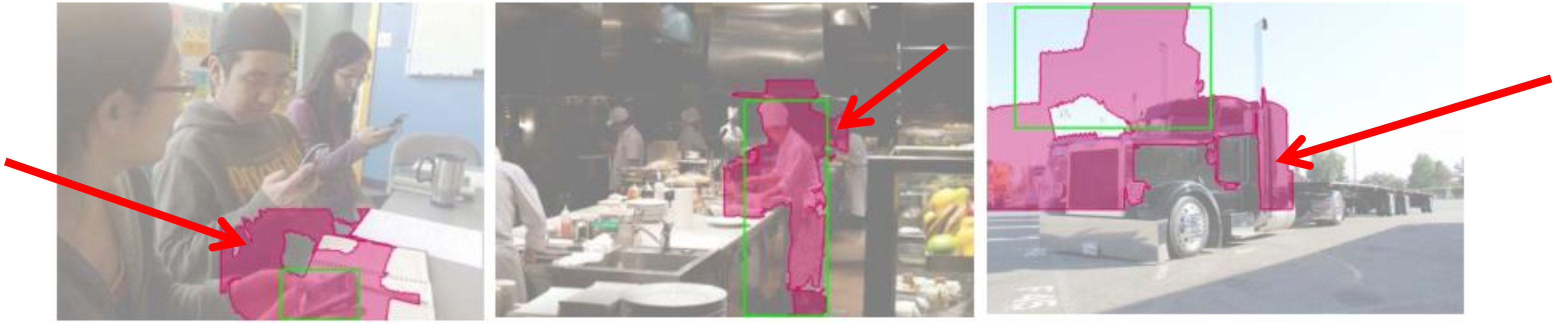
Effective sampling locations

Error-bounded saliency regions

(b) deformable RoIpooling, with deformable conv@conv5 stage (DCNv1)

Observations

- The error-bounded saliency regions in both aligned RoIpooling and Deformable RoIpooling are not fully focused on the object foreground, which suggests that image content outside of the RoI affects the prediction result.
- Spatial support of DCN-v1 may extend beyond the region of interest.



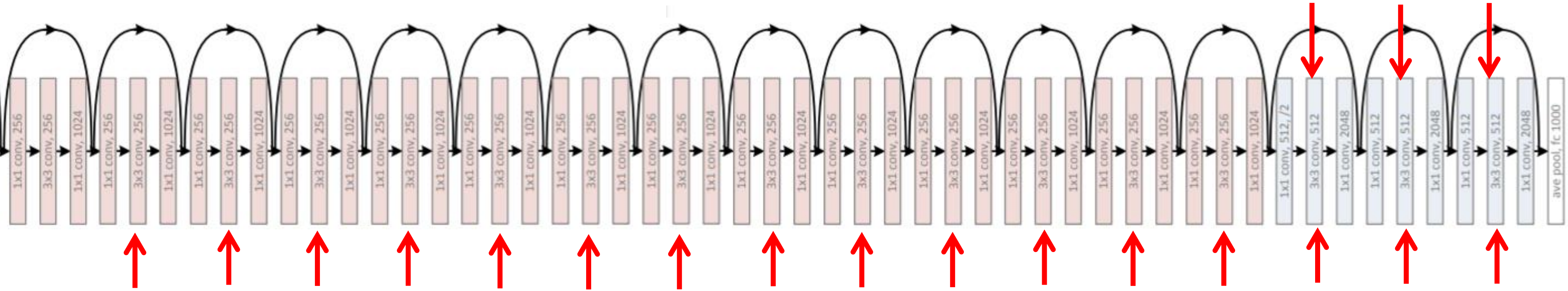
(b) deformable RoIpooling, with deformable conv@conv5 stage (DCNv1)

Deformable ConvNets V2

- Stacking More Deformable Conv Layers
  - the expanded use of deformable convolution layers within the network.
- Modulated Deformable Modules
  - each sample not only undergoes a learned offset, but is also modulated by a learned feature amplitude
- Better training: R-CNN Feature Mimicking
  - learns features unaffected by irrelevant information outside the region of interest.

Stacking More Deformable Conv Layers

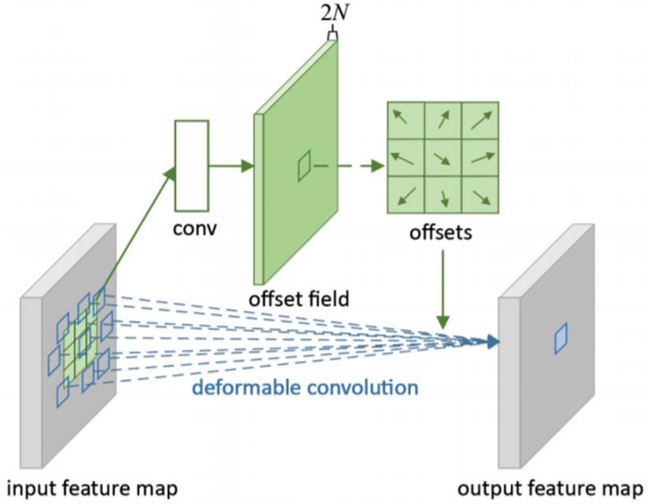
DCN-v1: 3 deform layers at stage 5



DCN-v2: each 3*3 Conv in stage 3, 4 and 5 is replaced with deform
13 layers for ResNet-50/ 30 layers for ResNet-101

by stacking more deformable conv layers, the geometric transformation modeling capability of the entire network can be further strengthened.

Deformable Modules: DCN-v1



$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k)$$

$p_k \in \{(-1,-1), (-1,0), \ldots, (1,1)\}$ defines a 3×3 convolutional kernel

$x(p)$: The features at location p from the input feature maps x

$y(p)$: The features at location p for the output feature maps y

Δp_k : Offset for x and y directions, real number with unconstrained range.

Modulated Deformable Modules: DCN-v2

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \boxed{\Delta m_k}$$

$p_k \in \{(-1,-1), (-1,0), \ldots, (1,1)\}$ defines a 3×3 convolutional kernel

$x(p)$: The features at location *p* from the input feature maps *x*

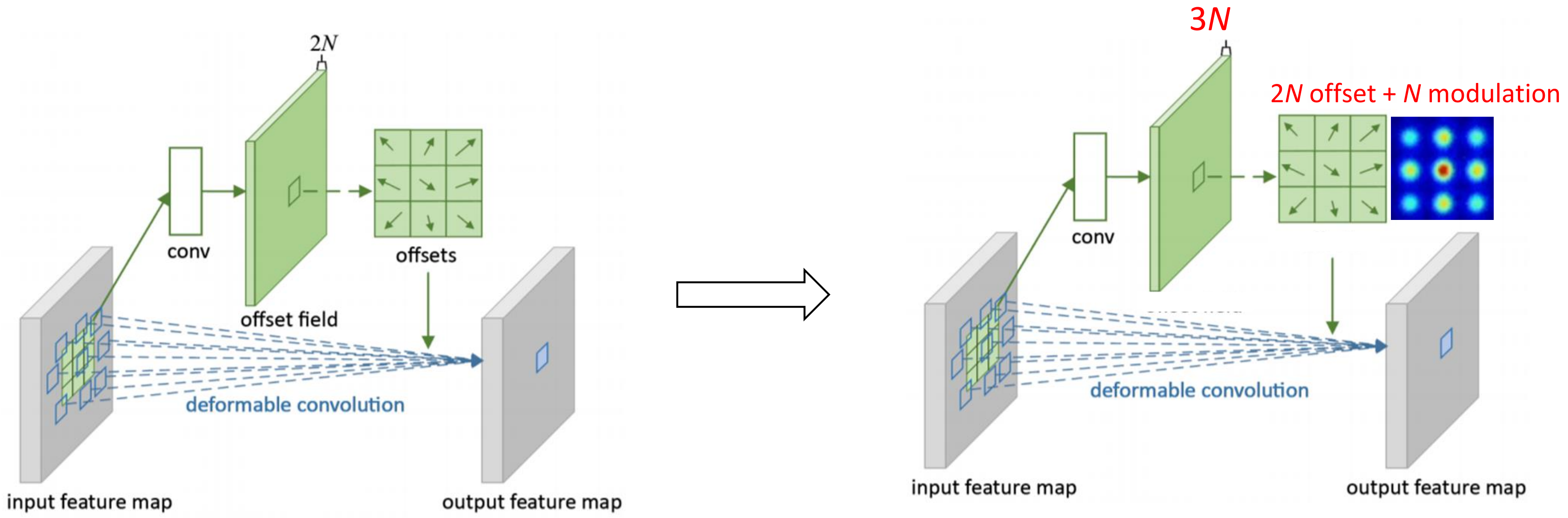
$y(p)$: The features at location *p* for the output feature maps *y*

Δp_k : Offset for x and y directions, real number with unconstrained range.

Δm_k : Modulation scalar lies in the range [0, 1], using sigmoid activation

Modulated Deformable Modules: DCN-v2

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k$$



With modulation, the Deformable ConvNets modules can not only adjust offsets in perceiving input features, but also modulate the input feature amplitudes/weights from different spatial locations.

Ablation studies

method	setting (shorter side 800)	Faster R-CNN						Mask R-CNN			
		AP^{bbox}	AP^{bbox}_S	AP^{bbox}_M	AP^{bbox}_L	param	FLOP	AP^{bbox}	AP^{mask}	param	FLOP
baseline	regular (RoIpooling)	32.8	13.6	37.2	48.7	51.3M	196.8G	-	-	-	-
	regular (aligned RoIpooling)	35.6	18.2	40.3	48.7	51.3M	196.8G	37.8	33.4	39.5M	303.5G
	dconv@c5 + dpool (DCNv1)	38.2	19.1	42.2	54.0	52.7M	198.9G	40.3	35.0	40.9M	304.9G
enriched deformation	dconv@c5	37.6	19.3	41.4	52.6	51.5M	197.7G	39.9	34.9	39.8M	303.7G
	dconv@c4∼c5	39.2	19.9	43.4	55.5	51.7M	198.7G	41.2	36.1	40.0M	304.7G
	dconv@c3∼c5	39.5	21.0	43.5	55.6	51.8M	200.0G	41.5	36.4	40.1M	306.0G
	dconv@c3∼c5 + dpool	40.0	21.1	44.6	56.3	53.0M	201.2G	41.8	36.4	41.3M	307.2G
	mdconv@c3∼c5 + mdpool	40.8	21.3	45.0	58.5	65.5M	214.7G	42.7	37.0	53.8M	320.3G

Table 2. Ablation study on enriched deformation modeling. The input images are of shorter side 800 pixels. Results are reported on the COCO 2017 validation set.

DCN-v1:
Adding deformable convolution to stage 5 improves ~2% AP, compared with regular counterpart

Ablation studies

method	setting (shorter side 800)	Faster R-CNN						Mask R-CNN			
		AP^{bbox}	AP^{bbox}_S	AP^{bbox}_M	AP^{bbox}_L	param	FLOP	AP^{bbox}	AP^{mask}	param	FLOP
baseline	regular (RoIpooling)	32.8	13.6	37.2	48.7	51.3M	196.8G	-	-	-	-
	regular (aligned RoIpooling)	35.6	18.2	40.3	48.7	51.3M	196.8G	37.8	33.4	39.5M	303.5G
	dconv@c5 + dpool (DCNv1)	38.2	19.1	42.2	54.0	52.7M	198.9G	40.3	35.0	40.9M	304.9G
enriched deformation	dconv@c5	37.6	19.3	41.4	52.6	51.5M	197.7G	39.9	34.9	39.8M	303.7G
	dconv@c4∼c5	39.2	19.9	43.4	55.5	51.7M	198.7G	41.2	36.1	40.0M	304.7G
	dconv@c3∼c5	39.5	21.0	43.5	55.6	51.8M	200.0G	41.5	36.4	40.1M	306.0G
	dconv@c3∼c5 + dpool	40.0	21.1	44.6	56.3	53.0M	201.2G	41.8	36.4	41.3M	307.2G
	mdconv@c3∼c5 + mdpool	40.8	21.3	45.0	58.5	65.5M	214.7G	42.7	37.0	53.8M	320.3G

Table 2. Ablation study on enriched deformation modeling. The input images are of shorter side 800 pixels. Results are reported on the COCO 2017 validation set.

More stages improves another ~2% AP

Ablation studies

method	setting (shorter side 800)	Faster R-CNN						Mask R-CNN			
		AP^{bbox}	AP^{bbox}_S	AP^{bbox}_M	AP^{bbox}_L	param	FLOP	AP^{bbox}	AP^{mask}	param	FLOP
baseline	regular (RoIpooling)	32.8	13.6	37.2	48.7	51.3M	196.8G	-	-	-	-
	regular (aligned RoIpooling)	35.6	18.2	40.3	48.7	51.3M	196.8G	37.8	33.4	39.5M	303.5G
	dconv@c5 + dpool (DCNv1)	38.2	19.1	42.2	54.0	52.7M	198.9G	40.3	35.0	40.9M	304.9G
enriched deformation	dconv@c5	37.6	19.3	41.4	52.6	51.5M	197.7G	39.9	34.9	39.8M	303.7G
	dconv@c4∼c5	39.2	19.9	43.4	55.5	51.7M	198.7G	41.2	36.1	40.0M	304.7G
	dconv@c3∼c5	39.5	21.0	43.5	55.6	51.8M	200.0G	41.5	36.4	40.1M	306.0G
	dconv@c3∼c5 + dpool	40.0	21.1	44.6	56.3	53.0M	201.2G	41.8	36.4	41.3M	307.2G
	mdconv@c3∼c5 + mdpool	40.8	21.3	45.0	58.5	65.5M	214.7G	42.7	37.0	53.8M	320.3G

Table 2. Ablation study on enriched deformation modeling. The input images are of shorter side 800 pixels. Results are reported on the COCO 2017 validation set.

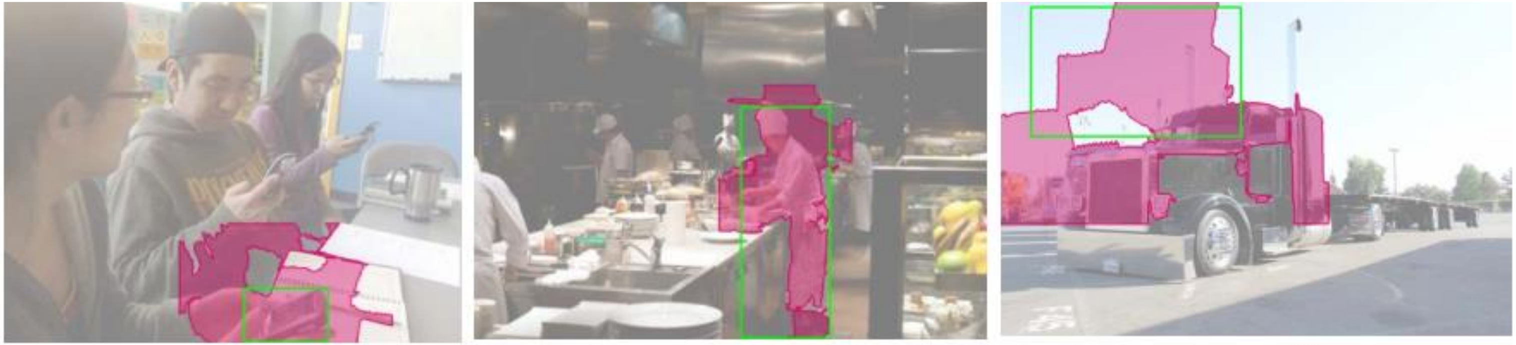
Deform RoI-Pooling: +0.5%

Modulated deform convolution + pooling: +0.8%

Most gains come from stacking more deformable layers: ~+2%

R-CNN Feature Mimicking

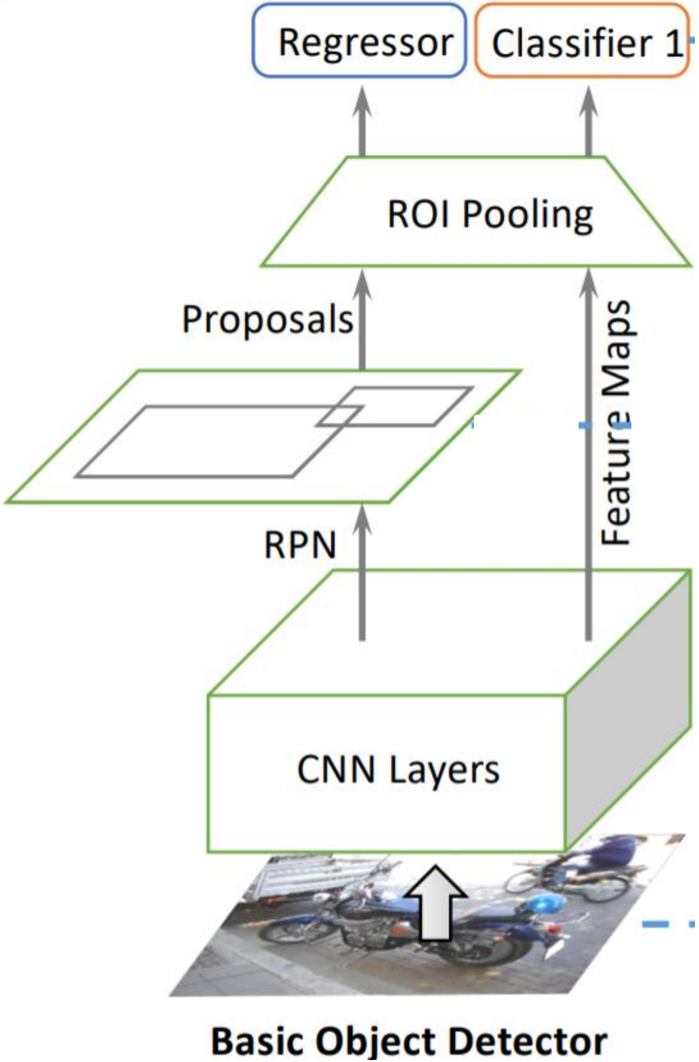
- Image content outside of the RoI may affect the extracted features and consequently degrade the final results of object detection.
- Such representations cannot be learned well through the standard Faster R-CNN training procedure. Additional guidance is needed to steer the training



(b) deformable RoIpooling, with deformable conv@conv5 stage (DCNv1)

R-CNN Feature Mimicking

- Why?
- Deep features at each region may have information that outside the region.



R-CNN Feature Mimicking

$$L_{\text{mimic}} = \sum_{b \in \Omega} [1 - \cos(f_{\text{RCNN}}(b), f_{\text{FRCNN}}(b))],$$

- At training, the network parameters between the corresponding modules in the R-CNN and the Faster R-CNN branches are shared

- In inference, only the Faster R-CNN network is applied on the test images.

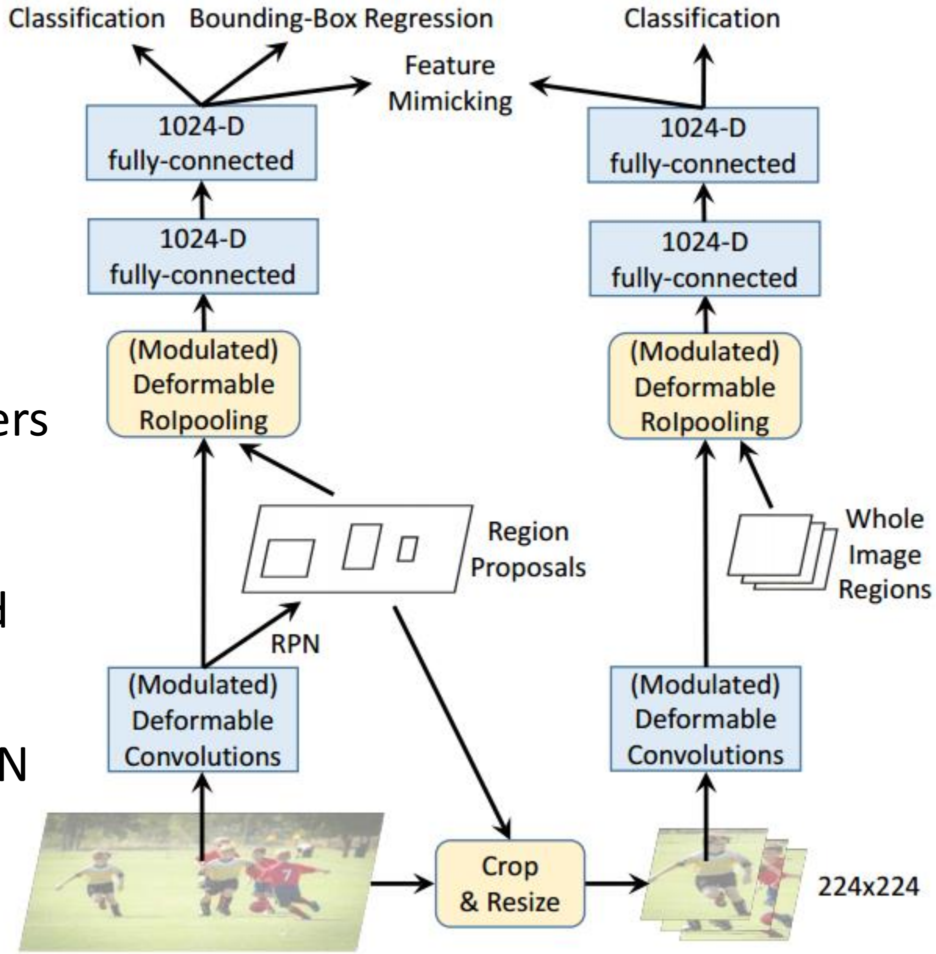


Figure 3. Network training with R-CNN feature mimicking.

R-CNN Feature Mimicking

setting	regions to mimic	Faster R-CNN	Mask R-CNN	
		AP^{bbox}	AP^{bbox}	AP^{mask}
mdconv3∼5 + mdpool	None	41.7	43.1	37.3
	FG & BG	42.1	43.4	37.6
	BG Only	41.7	43.3	37.5
	FG Only	43.1	44.3	38.3
regular	None	34.7	36.6	32.2
	FG Only	35.0	36.8	32.3

Table 3. Ablation study on R-CNN feature mimicking. Results are reported on the COCO 2017 validation set.

R-CNN Feature Mimicking

setting	regions to mimic	Faster R-CNN	Mask R-CNN	
		AP^{bbox}	AP^{bbox}	AP^{mask}
mdconv3∼5 + mdpool	None	41.7	43.1	37.3
	FG & BG	42.1	43.4	37.6
	BG Only	41.7	43.3	37.5
	FG Only	43.1	44.3	38.3
regular	None	34.7	36.6	32.2
	FG Only	35.0	36.8	32.3

1.4% improvements

Table 3. Ablation study on R-CNN feature mimicking. Results are reported on the COCO 2017 validation set.

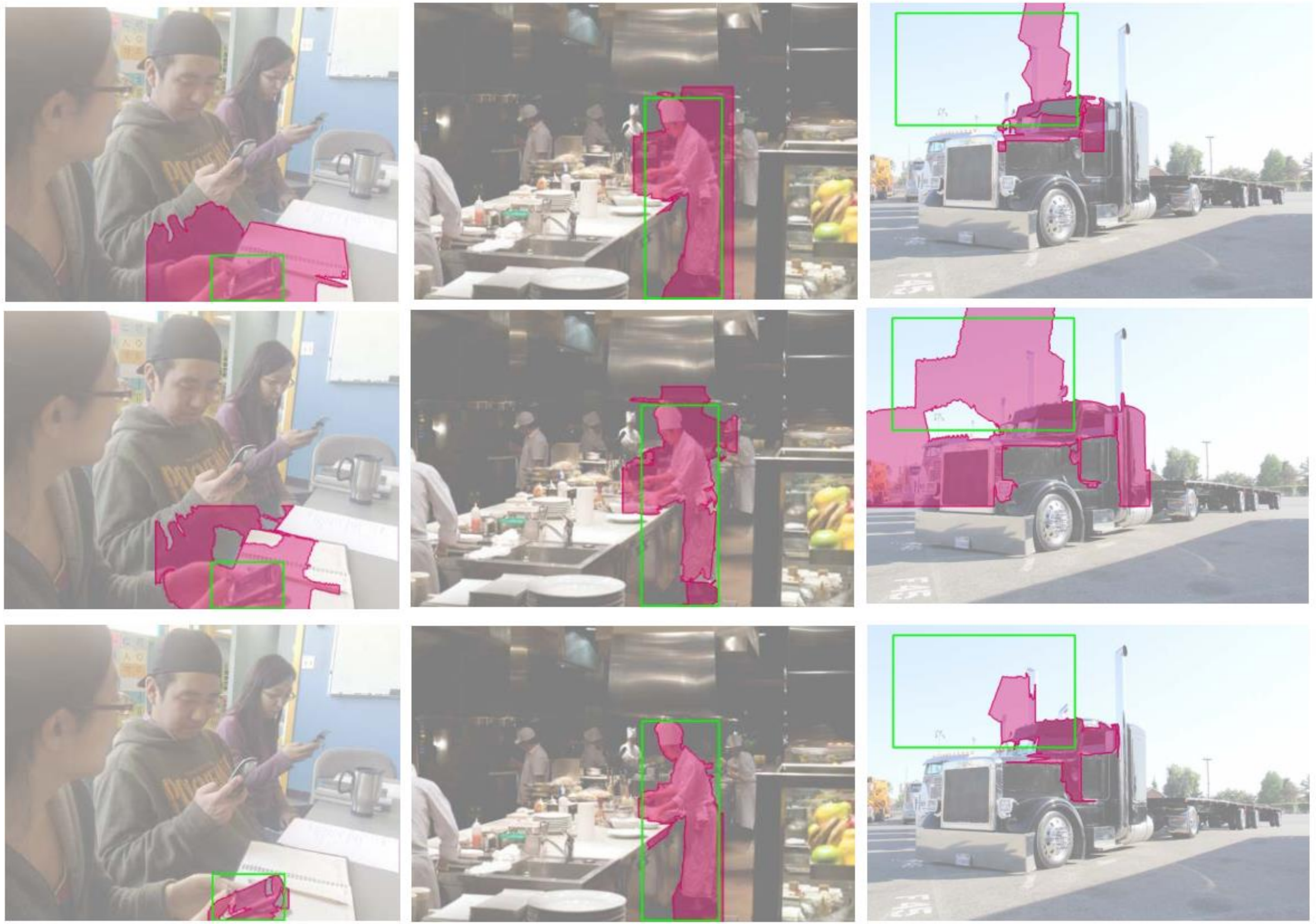
R-CNN Feature Mimicking

setting	regions to mimic	Faster R-CNN	Mask R-CNN	
		AP^{bbox}	AP^{bbox}	AP^{mask}
mdconv3∼5 + mdpool	None	41.7	43.1	37.3
	FG & BG	42.1	43.4	37.6
	BG Only	41.7	43.3	37.5
	FG Only	43.1	44.3	38.3
regular	None	34.7	36.6	32.2
	FG Only	35.0	36.8	32.3

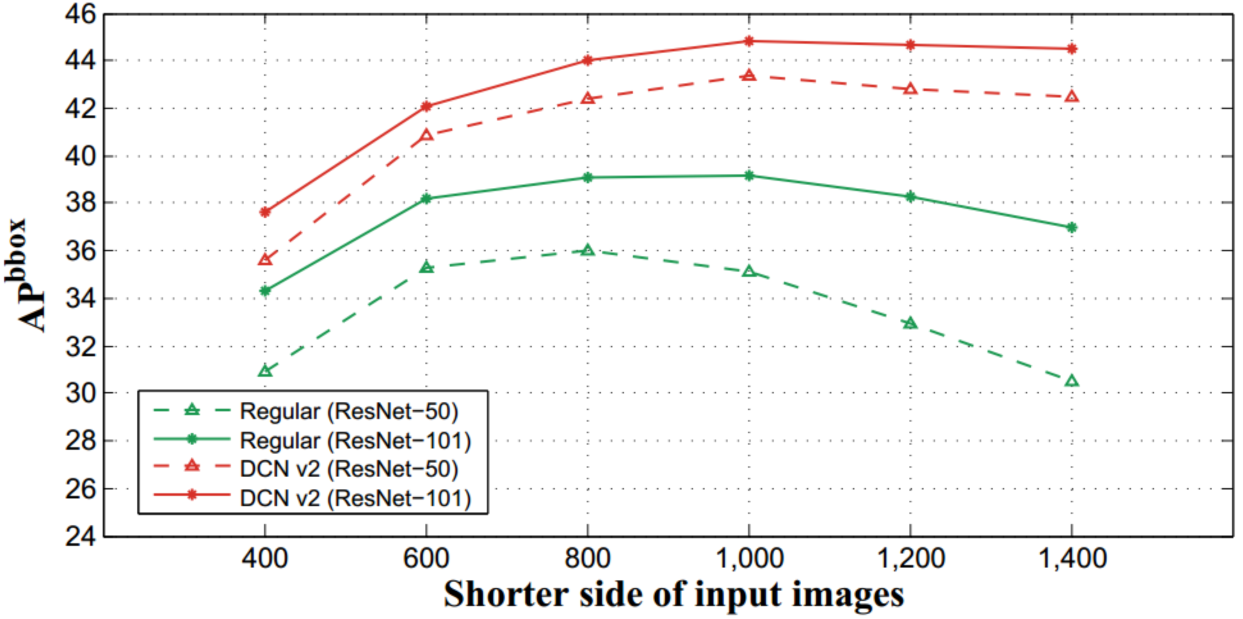
It is beyond the representation capability of regular ConvNets to focus features on the object foreground, and thus this cannot be learned regular

DCN-v1

DCN-v2



Final results: COCO object detection benchmark



(a) AP^{bbox} for all objects

DCN-ResNet-50 better AP than that of ResNet-50

Final results: ImageNet classification

backbone	method	top-1 acc (%)	top-5 acc (%)	param	FLOP
ResNet-50	regular	76.5	93.1	26.6M	4.1G
	DCNv1	76.6	93.2	26.8M	4.1G
	DCNv2	78.2	94.0	27.4M	4.3G
ResNet-101	regular	78.4	94.2	45.5M	7.8G
	DCNv1	78.4	94.2	45.8M	7.8G
	DCNv2	79.2	94.6	47.4M	8.2G
ResNeXt-101	regular	78.8	94.4	45.1M	8.0G
	DCNv1	78.9	94.4	45.6M	8.0G
	DCNv2	79.8	94.8	49.0M	8.7G

1% improvements

Summary

- The authors observe that the learned offset in DCN-v1 may extend well beyond the region of interest, causing features to be influenced by irrelevant image content.
- Several improvements on DCN-v1:
  - More deform layers (+2%), modulated term(+0.8%), and feature mimicking(+1.4)
- Leading results on several tasks:
  - Image classification (ImageNet)
  - object detection(ImageNet/COCO/VOC)
  - instance/semantic segmentation(COCO/VOC)
- Op: https://github.com/msracver/Deformable-ConvNets