Deformation Modeling in ConvNets

Jifeng Dai

Visual Computing Group

Microsoft Research Asia

Content

- Background
- Spatial Transformer Networks
- Deformable ConvNets v1
- Deformable ConvNets v2
- Related Work
- Conclusion

Modeling Spatial Transformations

- A long standing problem in computer vision

Part deformation:

Scale:

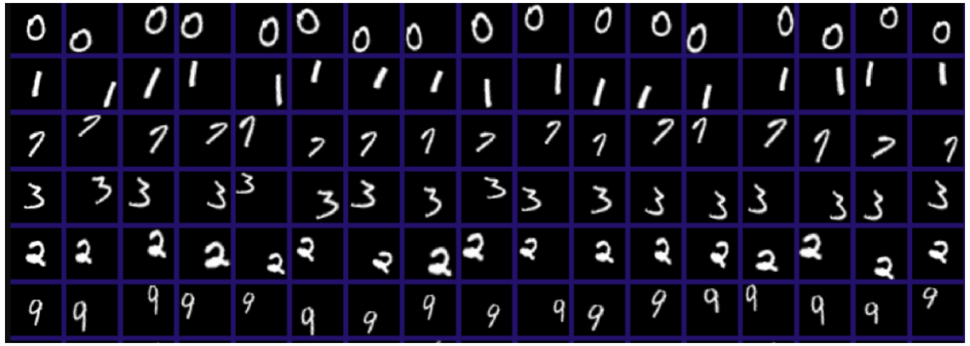
Viewpoint variation:

Intra-class variation:

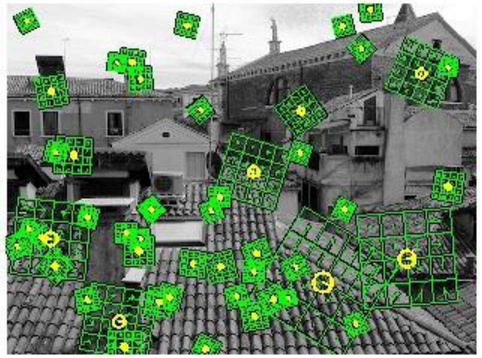
(Some examples are taken from Li Fei-fei's course CS223B, 2009-2010.)

Traditional Approaches

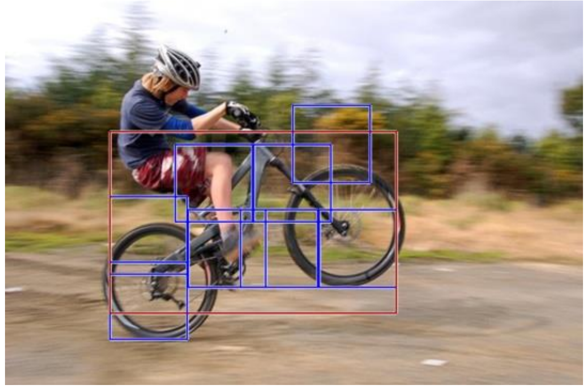
- 1) To build training datasets with sufficient desired variations



- 2) To use transformation-invariant features and algorithms



Scale Invariant Feature Transform (SIFT)

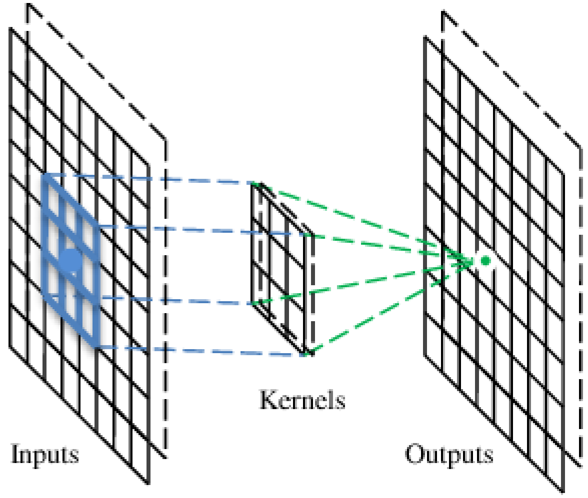


Deformable Part-based Model (DPM)

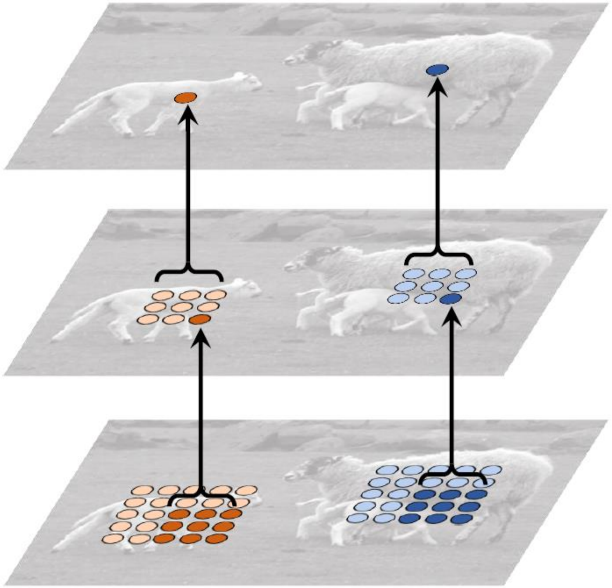
- Drawbacks: geometric transformations are assumed fixed and known, hand-crafted design of invariant features and algorithms

Spatial transformations in CNNs

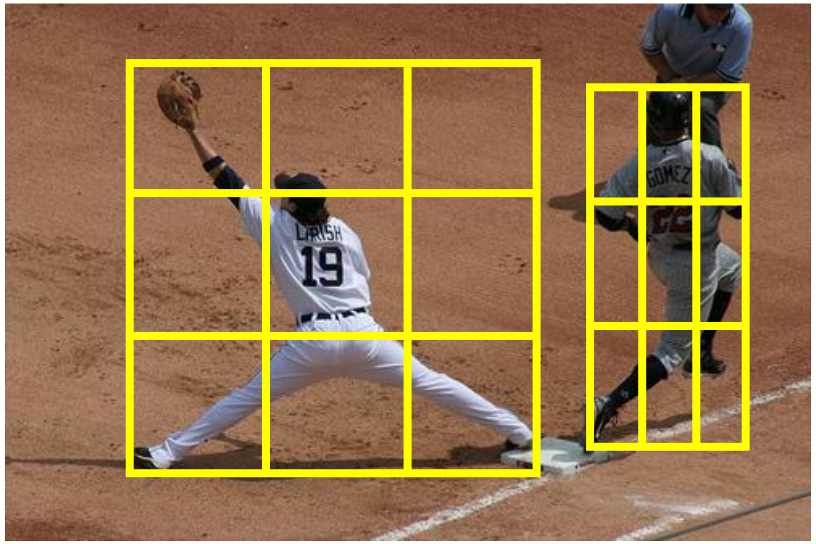
- Regular CNNs are inherently limited to model large unknown transformations
    - The limitation originates from the fixed geometric structures of CNN modules



regular convolution



2 layers of regular convolution



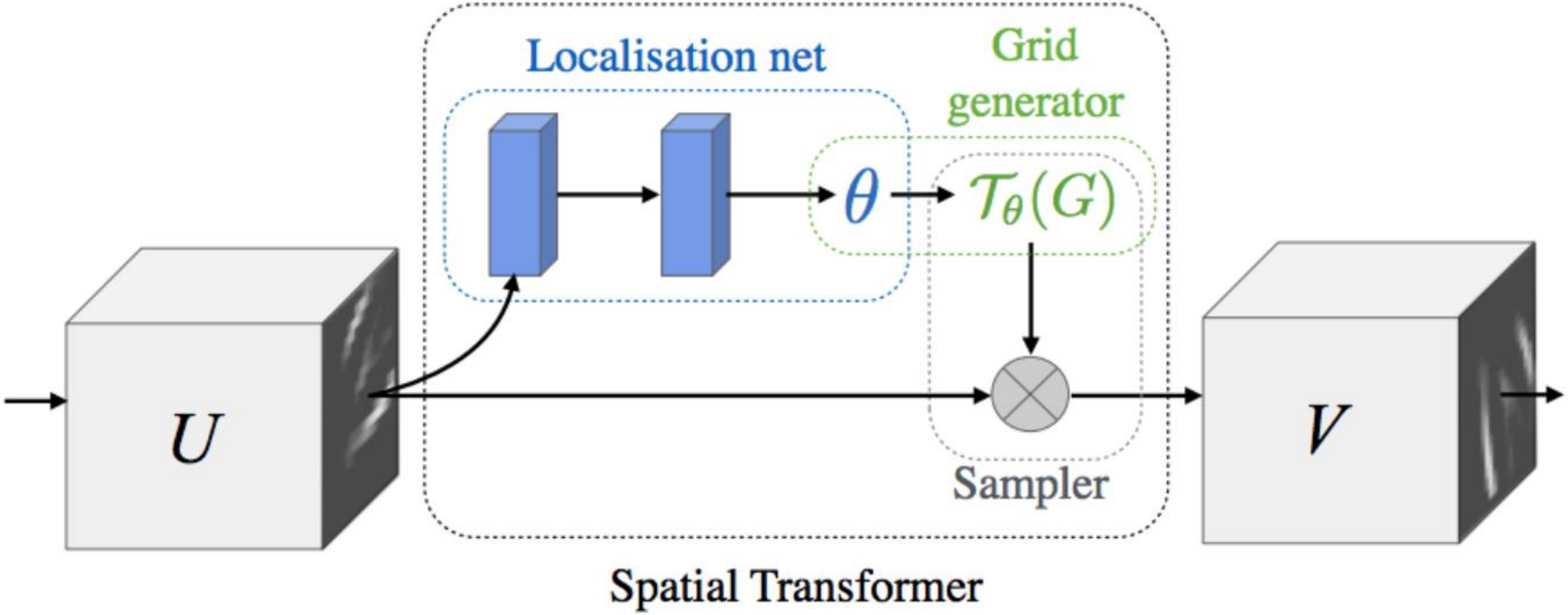
regular RoI Pooling

Content

- Background
- Spatial Transformer Networks
- Deformable ConvNets v1
- Deformable ConvNets v2
- Related Work
- Conclusion

Max Jaderberg, Karen Simonyan, Andrew Zisserman, Koray Kavukcuoglu. Spatial Transformer Networks. NIPS 2015.

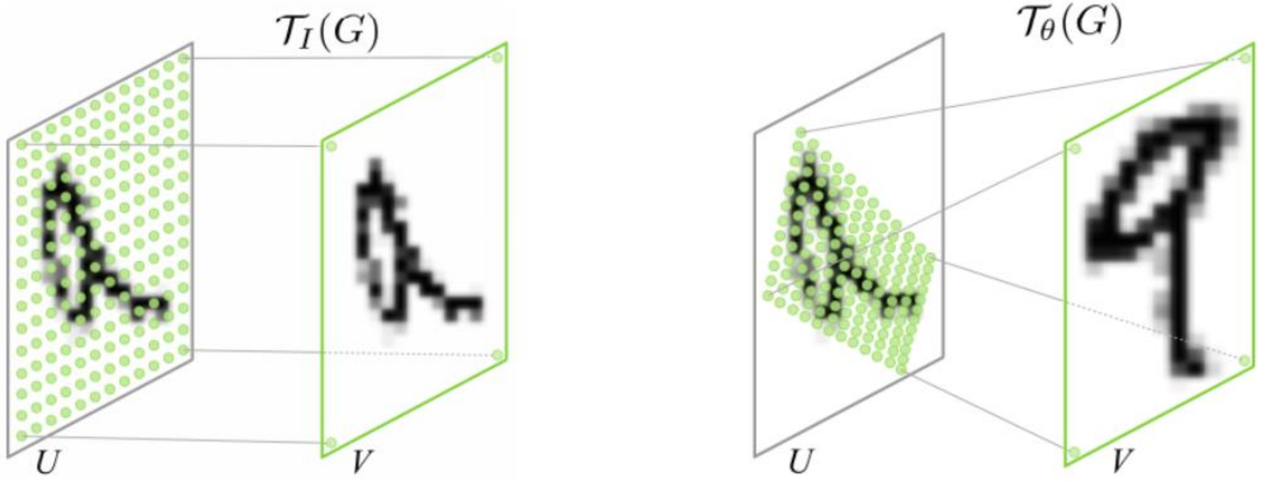
Spatial Transformer Networks



Spatial Transformer Networks

- Parameterized Sampling Grid

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$$

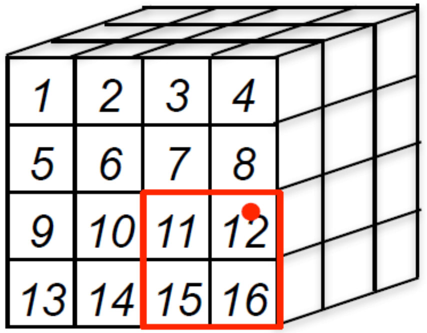


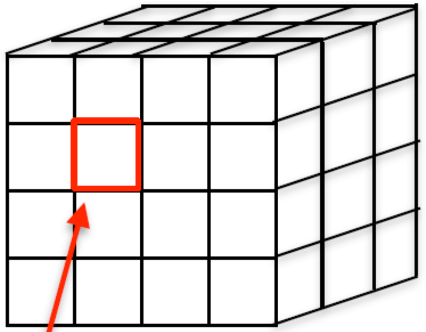
(a) (b)

Spatial Transformer Networks

- Differentiable Image Sampling

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|)$$

$U =$ 

$V =$ 

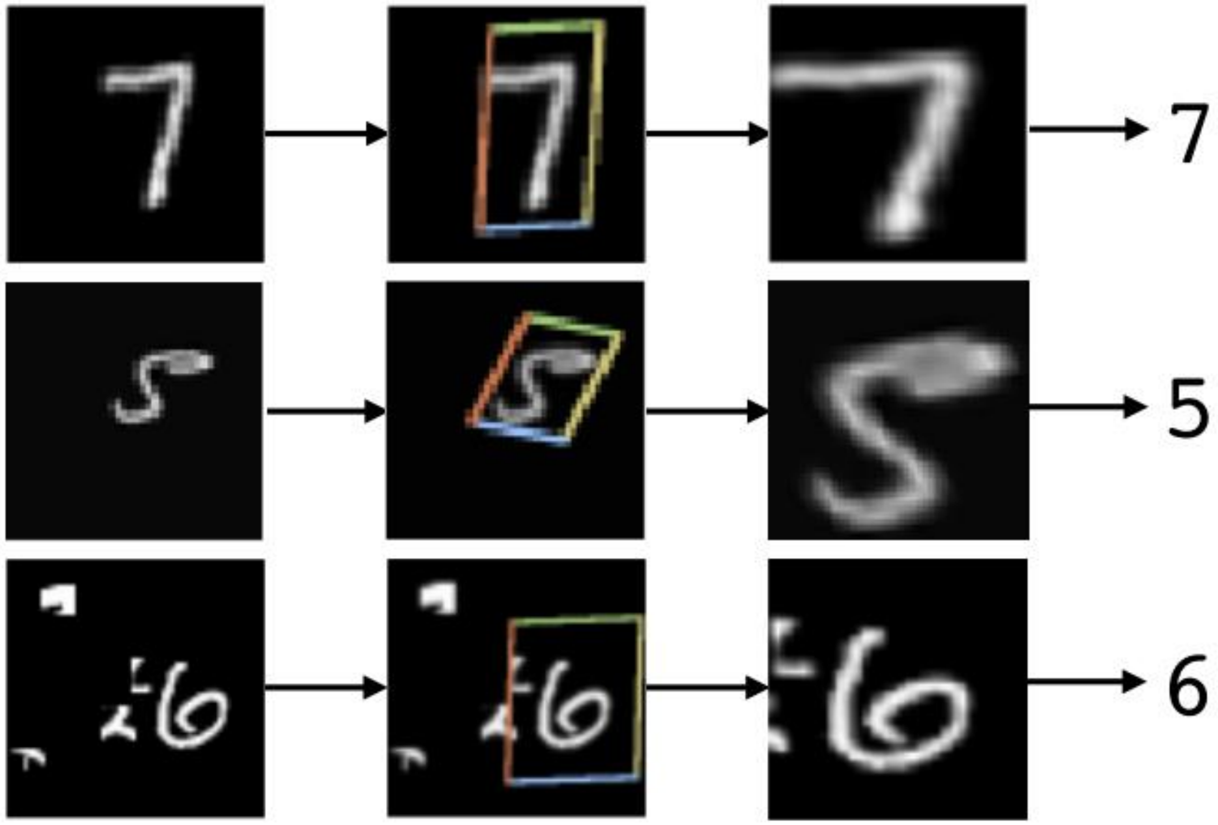
$x_i^s = 3.7 \quad y_i^s = 3.1$

$i = 6$

$$\boxed{0.27 \times 11 + 0.63 \times 12 + 0.03 \times 15 + 0.07 \times 16 = 12.1}$$

Spatial Transformer Networks

- Learning a global, parametric transformation on feature maps
  - Prefixed transformation family, infeasible for complex vision tasks



Content

- Background
- Spatial Transformer Networks
- Deformable ConvNets v1
- Deformable ConvNets v2
- Related Work
- Conclusion

Deformable Convolutional Networks. Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, Yichen Wei. ICCV 2017.

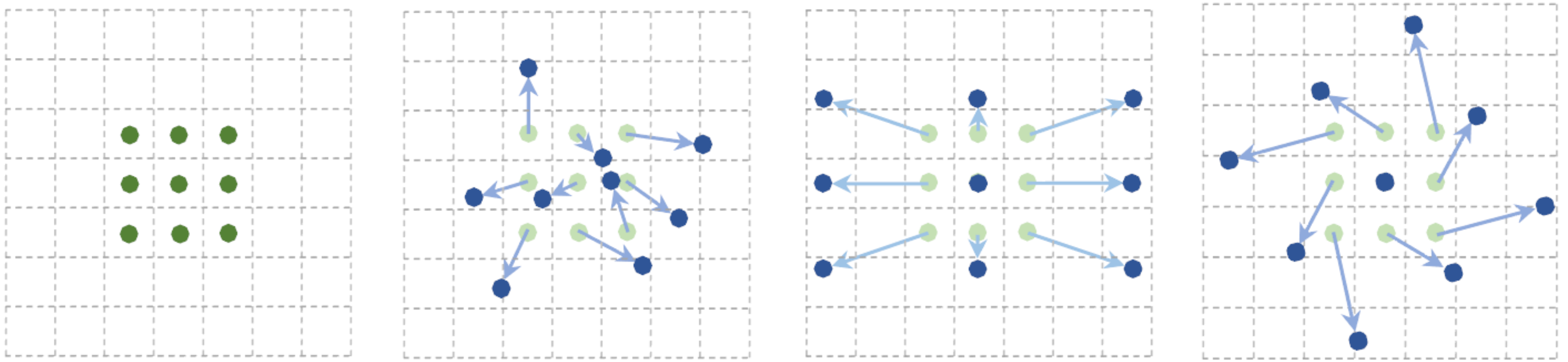
Highlights

- **Enabling effective modeling of spatial transformation** in ConvNets
- **No additional supervision** for learning spatial transformation
- **Significant accuracy improvements** on sophisticated vision tasks

Code is available at https://github.com/msracver/Deformable-ConvNets

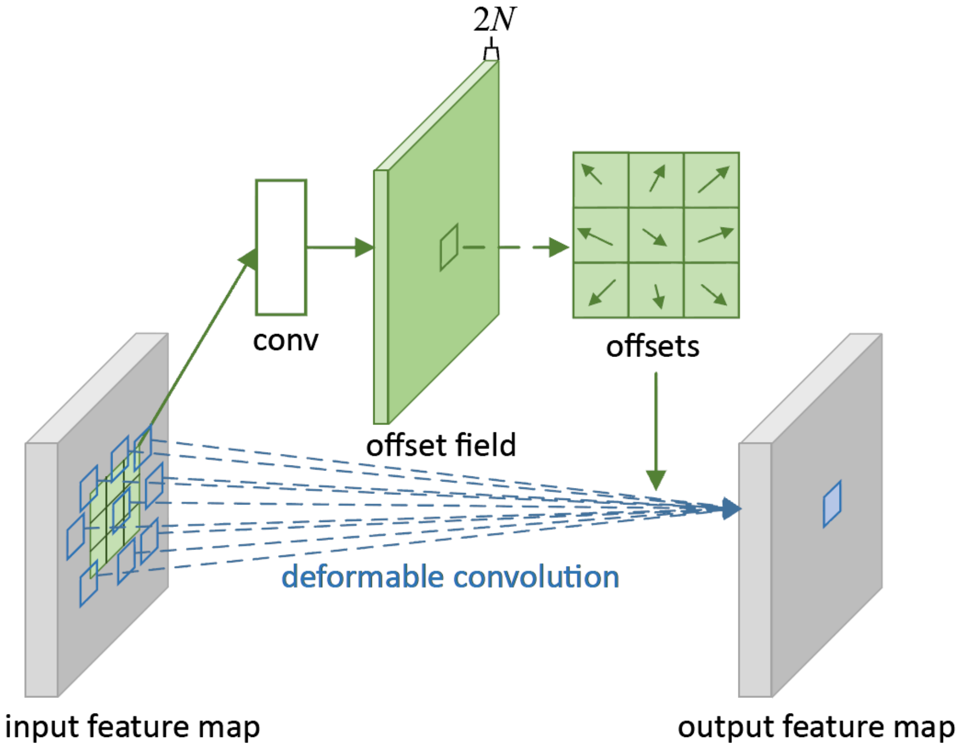
Deformable Convolution

- Local, dense, non-parametric transformation
  - Learning to deform the sampling locations in the convolution/RoI Pooling modules



regular   deformed   scale & aspect ratio   rotation

Deformable Convolution



Regular convolution

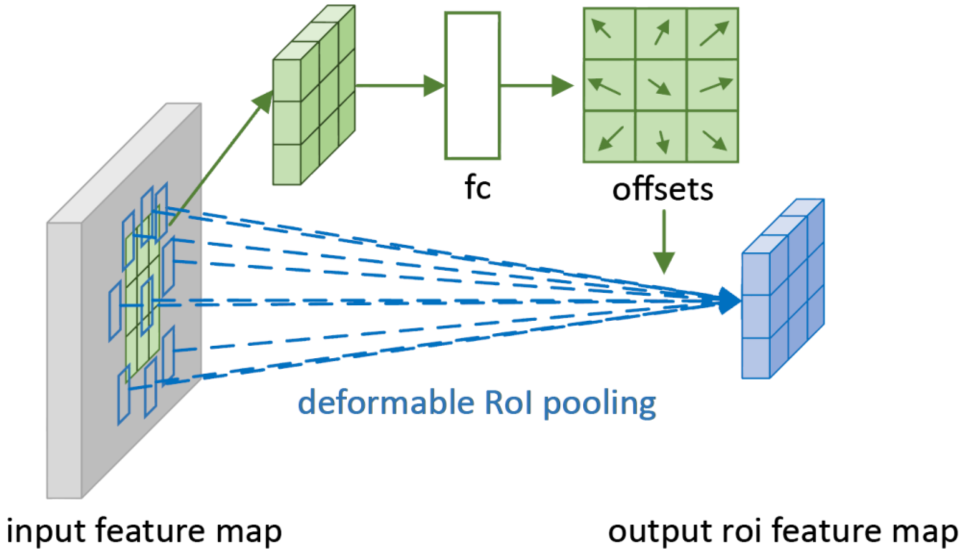
$$\mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n)$$

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

Deformable RoI Pooling



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

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

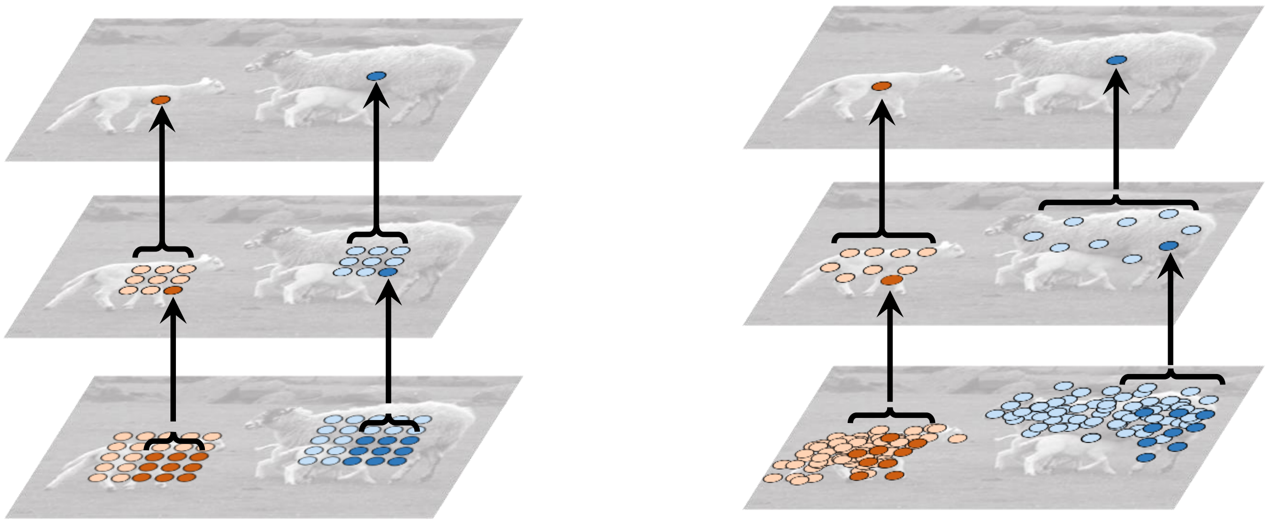
where $\Delta\mathbf{p}_{ij}$ is generated by a sibling fc branch

Deformable ConvNets

- Same input & output as the plain versions
  - Regular convolution -> deformable convolution
  - Regular RoI pooling -> deformable RoI pooling

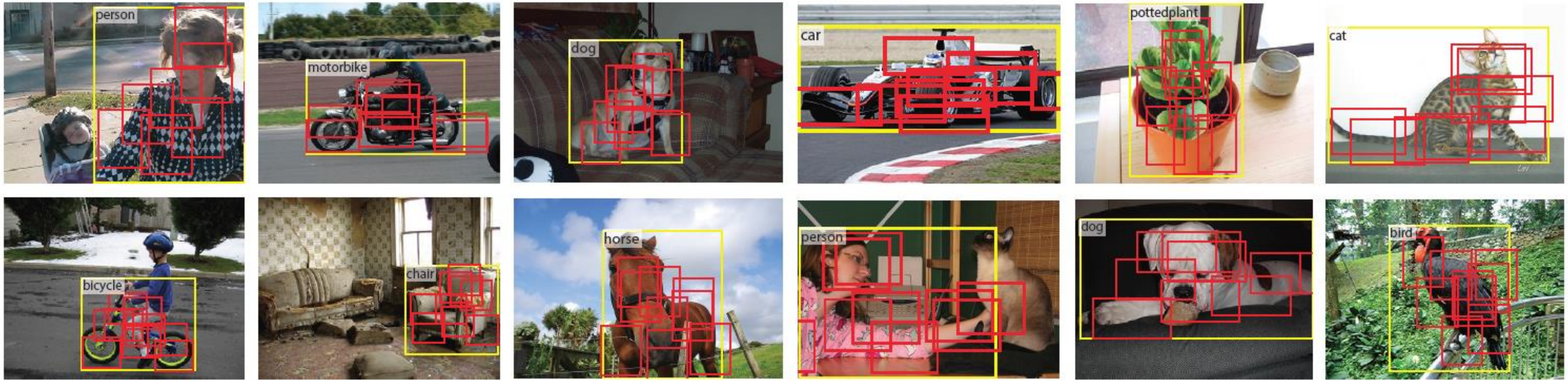
- End-to-end trainable without additional supervision

Sampling Locations of Deformable Convolution



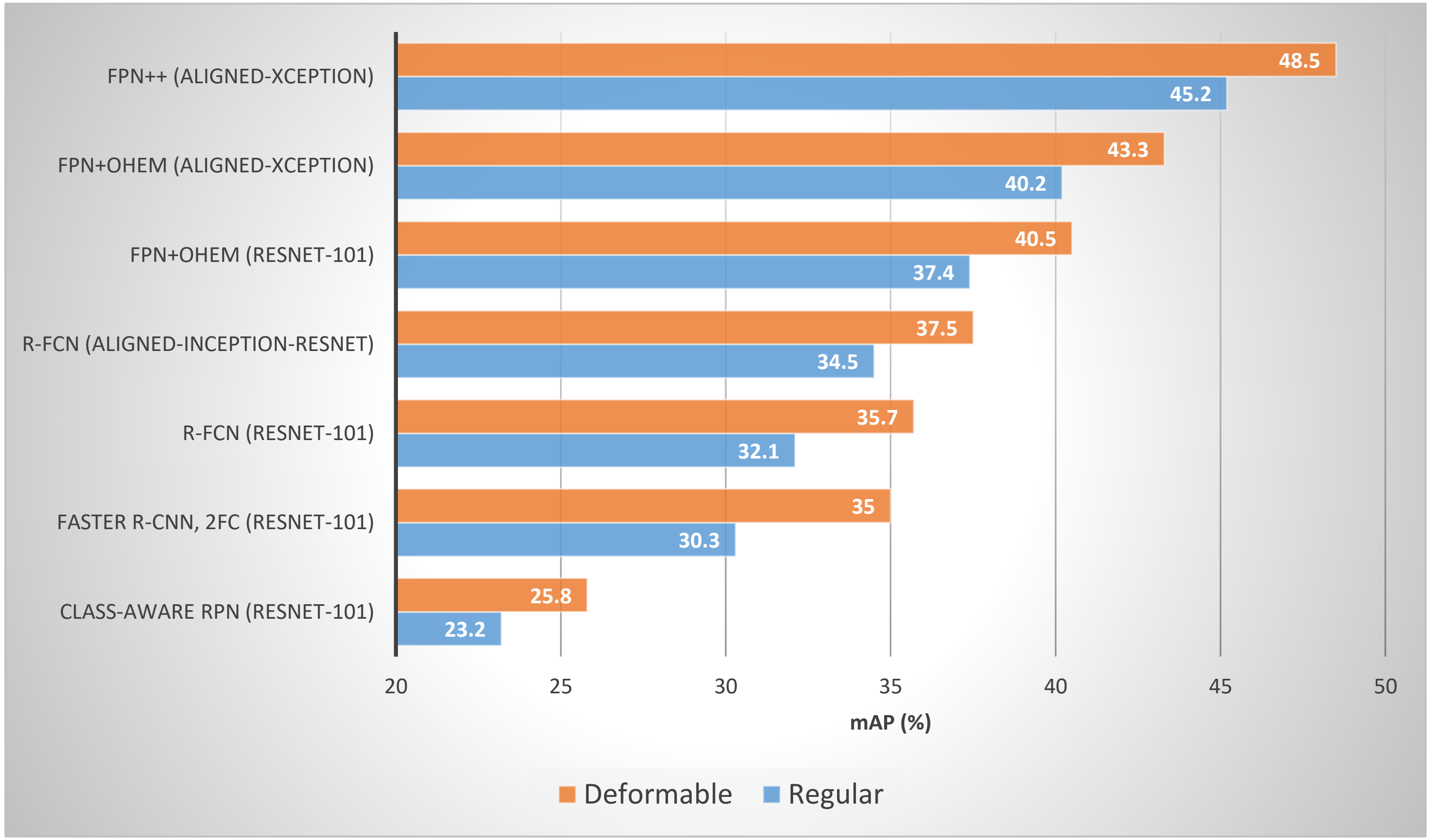
(a) standard convolution (b) deformable convolution

Part Offsets in Deformable RoI Pooling



Object Detection on COCO (Test-dev)

- Deformable ConvNets v.s. regular ConvNets
  - Noticeable improvements for varies baselines
  - Marginal parameter & computation overhead



Content

- Background
- Spatial Transformer Networks
- Deformable ConvNets v1
- Deformable ConvNets v2
- Related Work
- Conclusion

Xizhou Zhu, Han Hu, Stephen Lin, Jifeng Dai , Deformable ConvNets v2: More Deformable, Better Results. CVPR, 2019.

Highlights

- Better understanding of deformation modeling in CNNs
- Reformulation of Deformable ConvNets to strengthen its deformation modeling capability
- To harness the enhanced modeling capability, guide network training via R-CNN feature mimicking

Core operators are available at https://github.com/msracver/Deformable-ConvNets

Analysis of Deformable ConvNet Behavior

- DCN v1 visualization: theoretical spatial support (sampling / bin location only)

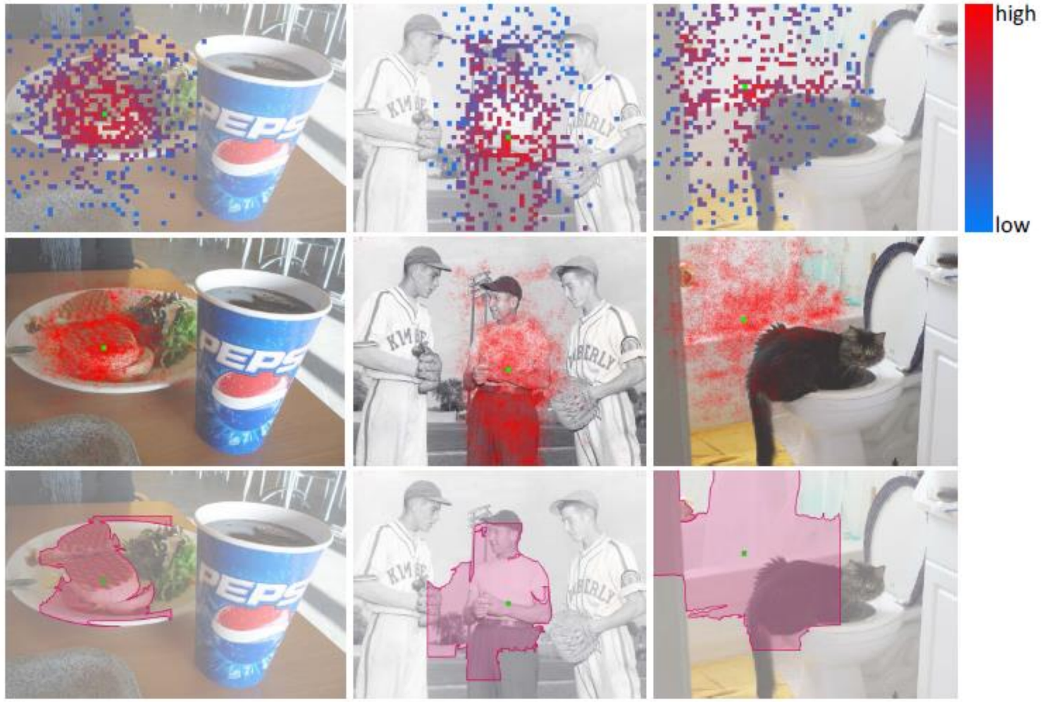
- DCN v2 visualization: effective spatial support (sampling / bin location & learnable network weights)
  - Effective sampling / bin locations
  - Effective receptive fields [Luo et al., NIPS 2016]
  - Error-bounded saliency regions

$$\begin{aligned} &\min \|M\|_1 \\ \text{s.t.} \quad & L_{\text{rec}}(\mathcal{N}(\mathbf{I}), \mathcal{N}(\mathbf{I} \odot M)) < \epsilon, \end{aligned}$$

Analysis of Deformable ConvNet Behavior

- Spatial support of nodes in the last layer of the conv5 stage of ResNet-50
  - Regular ConvNets can model geometric variations to some extent.
  - By introducing deformable convolution, the network's ability to model geometric transformation is considerably enhanced, **but still lacks**.

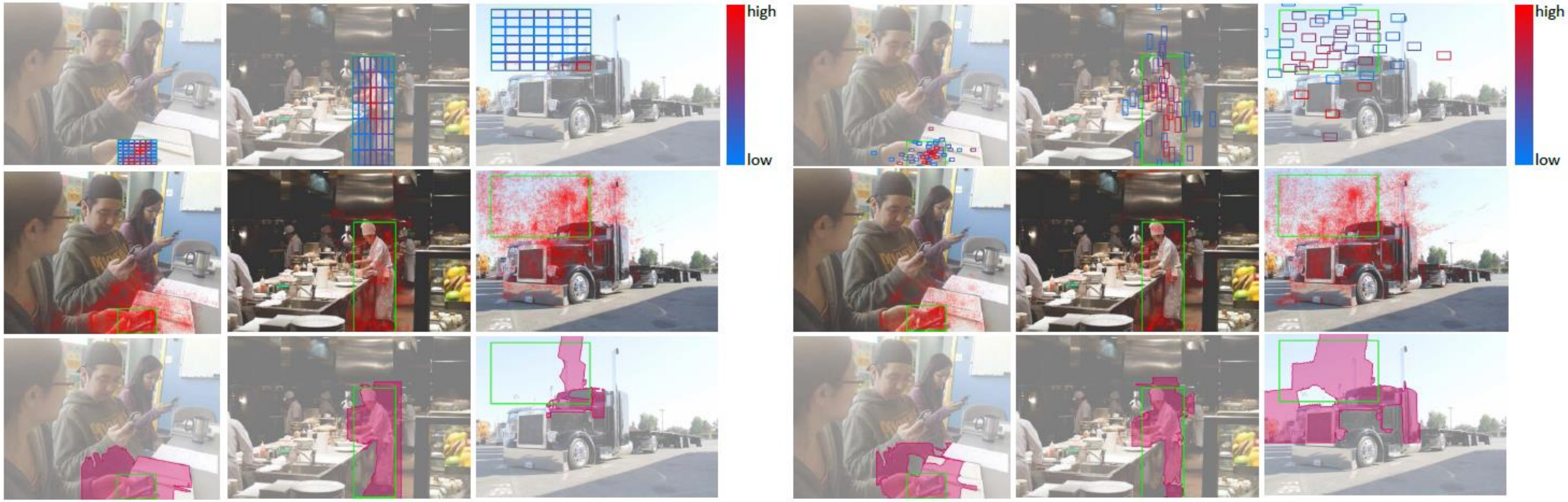
(a) regular conv



(b) deformable conv@conv5 stage (DCNv1)

Analysis of Deformable ConvNet Behavior

- Spatial support of the 2fc node in the per-RoI detection head
  - By introducing deformable RoI pooling, the network's ability to model geometric transformation is enhanced, **but still lacks**.



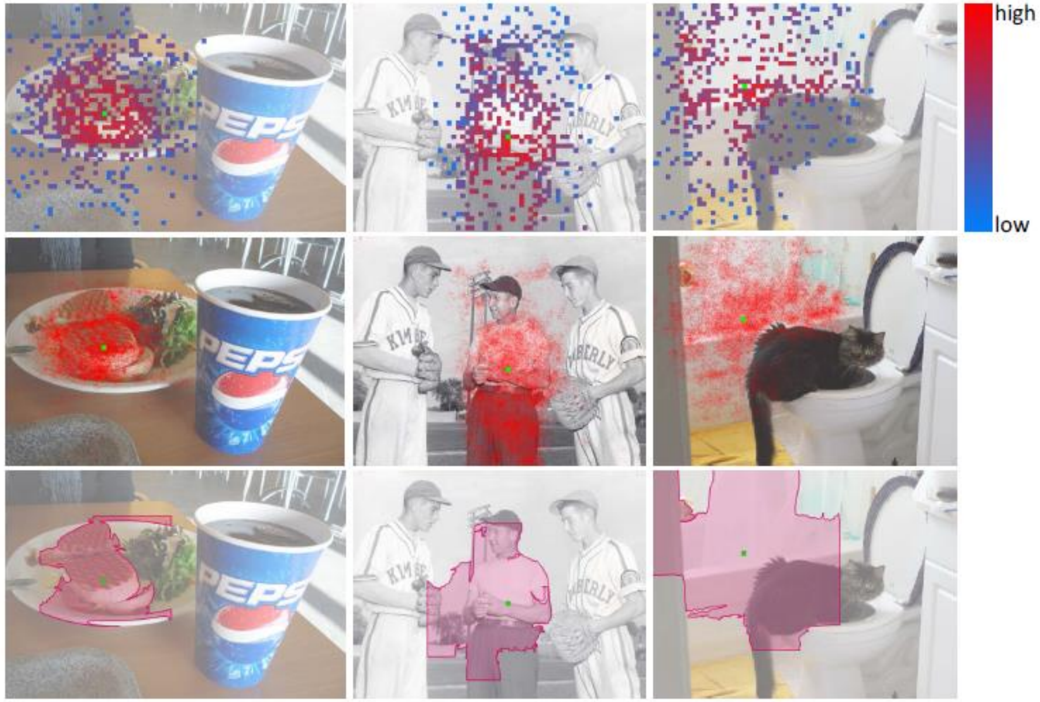
(a) aligned RoIpooling, with deformable conv@conv5 stage (b) deformable RoIpooling, with deformable conv@conv5 stage (DCNv1)

Analysis of Deformable ConvNet Behavior

- Observations
  - Regular ConvNets can model geometric variations to some extent.
  - By introducing deformable convolution & deformable RoI pooling, the network's ability to model geometric transformation is considerably enhanced, **but still lacks**.
  - The three presented types of spatial support visualizations are more informative than the sampling locations used in Deformable ConvNets v1 paper.
- What's next?
  - To upgrade Deformable ConvNets so that they can better focus on pertinent image content and deliver greater accuracy

Stacking More Deformable Conv Layers

- To strengthen the geometric transformation modeling capability of the entire network



(b) deformable conv@conv5 stage (DCNv1)

(c) modulated deformable conv@conv3∼5 stages (DCNv2)

Modulated Deformable Modules

- Not only adjust offsets in perceiving input features, but also modulate the input feature amplitudes from different spatial locations / bins

  - Modulated deformable Convolution

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

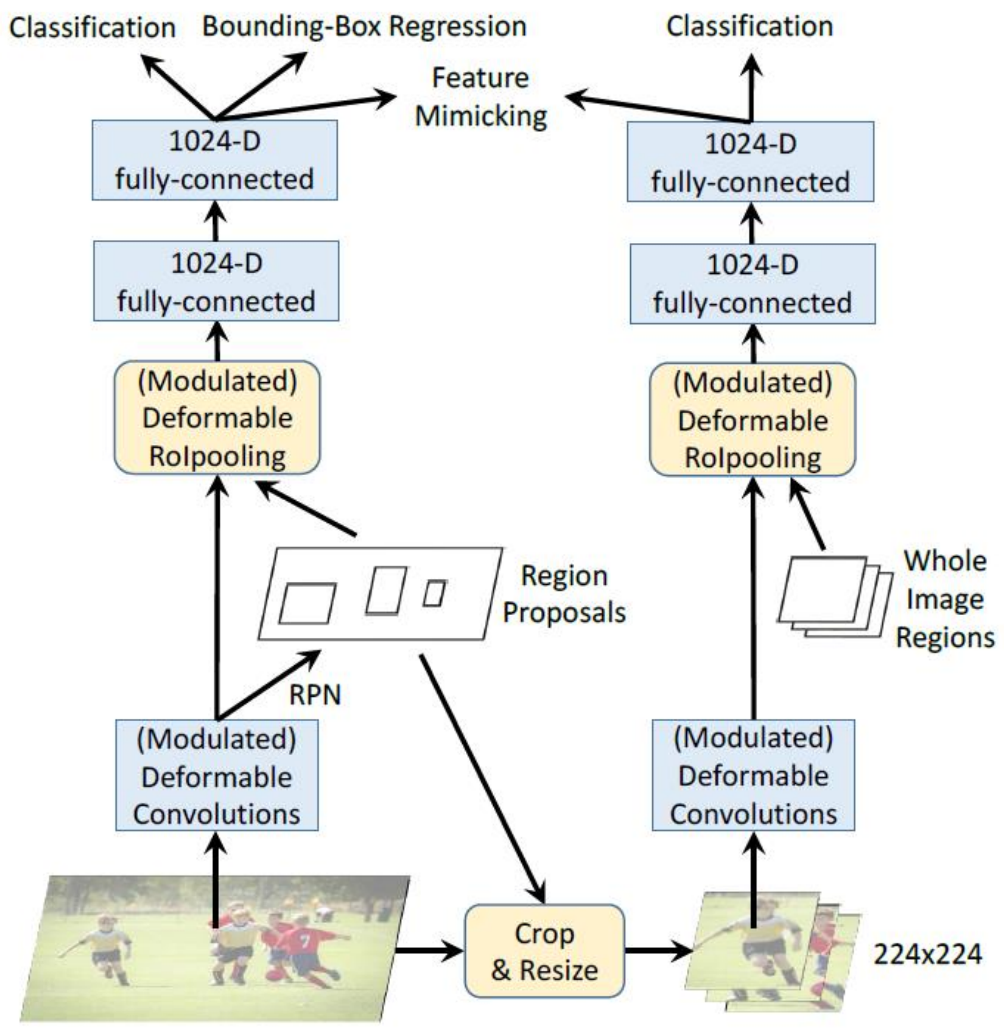
  - Modulated deformable RoIpooling

$$y(k) = \sum_{j=1}^{n_k} x(p_{kj} + \Delta p_k) \cdot \Delta m_k / n_k,$$

R-CNN Feature Mimicking

- Motivation
  - Even with the strong geometry modeling capability, the spatial support of the per-RoI node can still not focus on the RoI
  - Additional guidance is needed to steer the training

R-CNN Feature Mimicking



- Applied at training time only, no additional overhead for inference
- Feature mimicking loss enforced on sampled positive RoIs

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

R-CNN Feature Mimicking

(c) modulated deformable RoIpooling, with modulated deformable conv@conv3∼5 stages

(d) with R-CNN feature mimicking on setting (c) (DCNv2)

Ablation Experiments on Enriched Deformation

- Stacking more deformable conv layers and exploitation of modulation mechanism effectively improve the accuracy

method	setting (shorter side 1000)	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)	32.1	14.9	37.5	44.4	51.3M	326.7G	-	-	-	-
	regular (aligned RoIpooling)	34.7	19.3	39.5	45.3	51.3M	326.7G	36.6	32.2	39.5M	447.5G
	dconv@c5 + dpool (DCNv1)	38.0	20.7	41.8	52.2	52.7M	328.2G	40.4	35.3	40.9M	449.0G
enriched deformation	dconv@c5	37.4	20.0	40.9	51.0	51.5M	327.1G	40.2	35.1	39.8M	447.8G
	dconv@c4~c5	40.0	21.4	43.8	55.3	51.7M	328.6G	41.8	36.8	40.0M	449.4G
	dconv@c3~c5	40.4	21.6	44.2	56.2	51.8M	330.6G	42.2	37.0	40.1M	451.4G
	dconv@c3~c5 + dpool	41.0	22.0	45.1	56.6	53.0M	331.8G	42.4	37.0	41.3M	452.5G
	mdconv@c3~c5 + mdpool	41.7	22.2	45.8	58.7	65.5M	346.2G	43.1	37.3	53.8M	461.1G

Table 1. Ablation study on enriched deformation modeling. The input images are of shorter side 1,000 pixels (default in paper). In the setting column, "(m)dconv" and "(m)dpool" stand for (modulated) deformable convolution and (modulated) deformable RoIpooling, respectively. Also, "dconv@c3~c5" stands for applying deformable conv layers at stages conv3~conv5, for example. Results are reported on the COCO 2017 validation set.

Ablation Experiments of 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.

Content

- Background
- Spatial Transformer Networks
- Deformable ConvNets v1
- Deformable ConvNets v2
- Related Work
- Conclusion

Related Work

- Deformation Modeling
  - SIFT [Lowe, ICCV 1999] , ORB [Rublee et al., ICCV 2011], DPM [Felzenszwalb et al., TPAMI 2010]
  - Spatial Transformer Networks [Jaderberg et al., NIPS 2015], DeepID-Net [Ouyang et al., CVPR 2015], etc.
- Relation Networks and Attention Modules
  - Relation Modules in NLP [Gehring et al., ACL 2017], physical system modeling [Battaglia et al., NIPS 2016]
  - Relation networks for object detection [Hu et al., CVPR 2018], non-local networks [Wang et al., CVPR 2018], Learning region features for object detection [Gu et al., ECCV 2018]

Related Work

- Spatial Support Manipulation
  - Atrous convolution [Chen et al., ICLR 2015], active convolution [Jeon and Kim, CVPR 2017], multi-path network [Zagoruyko et al., BMVC 2016]

- Network Mimicking and Distillation
  - [Ba and Caruana, NIPS 2014], [Hinton et al., STAT 2015], [Li et al., CVPR 2017]

Content

- Background
- Spatial Transformer Networks
- Deformable ConvNets v1
- Deformable ConvNets v2
- Related Work
- Conclusion

Conclusion

- Standard CNNs are not very well equipped to model deformations, and transformations of the objects.
- Spatial Transformer Networks and Deformable ConvNets enabled effective modeling of geometric deformation in CNNs
- Open questions:
  - More effective manner to capture geometric deformation
  - Disentangle different factors in geometric deformation
  - Many more…

Q & A