Depth Privileged Object Detection in Indoor Scenes via Deformation Hallucination

Zhijie Zhang, Yan Liu, Junjie Chen, Li Niu, Liqing Zhang

MoE Key Lab of Artificial Intelligence, Department of Computer Science and Engineering, Shanghai Jiao Tong University

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Contents

Contents

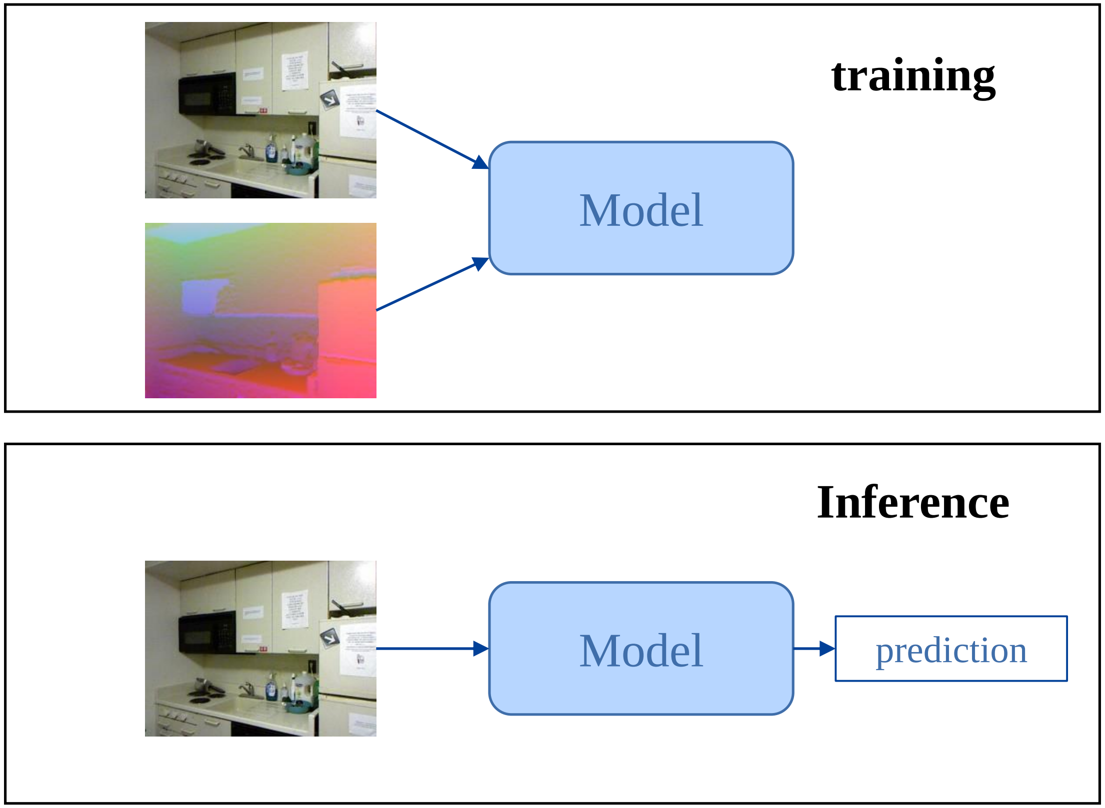
上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Motivations

- Cluttered objects remain difficult to be detected due to the large variations (e.g., occlusion and illumination)
- Depth information can provide additional geometric cues to complement RGB image
- While RGB image capturing devices are pervasive, depth capturing devices are much less prevalent.

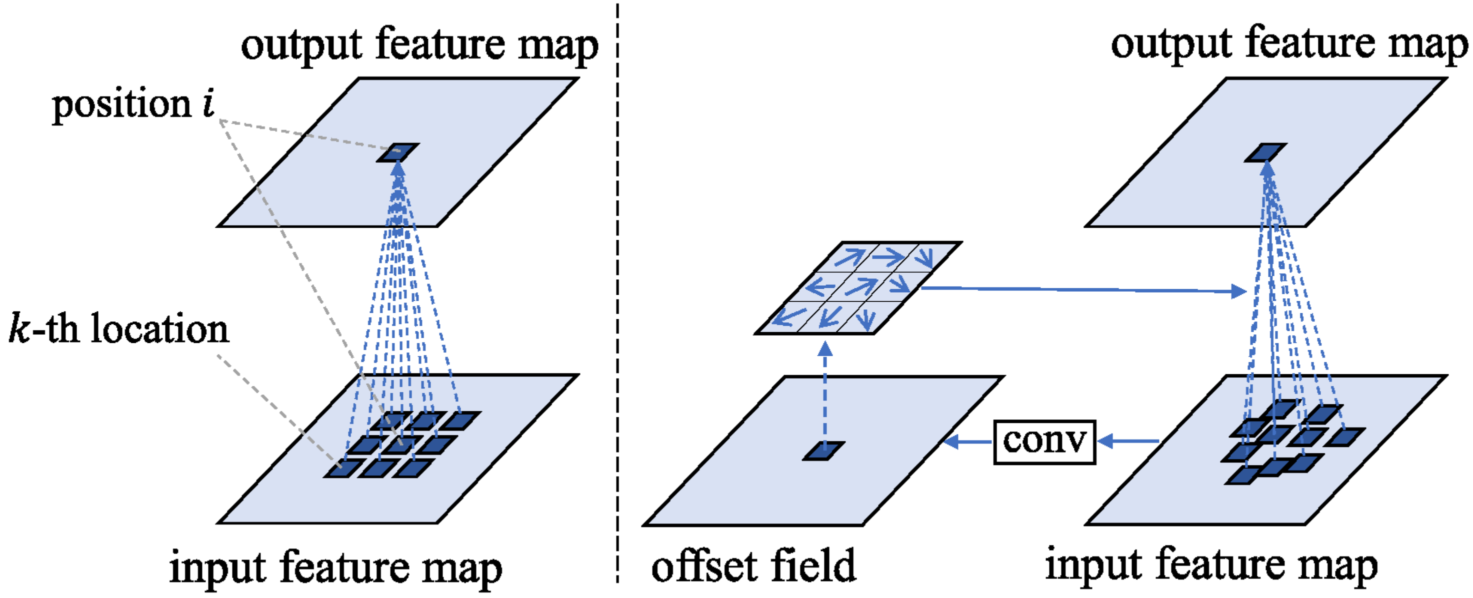
Learning Using Privileged Information

- Privileged information: can be accessed in training stage while be not available in inference stage.

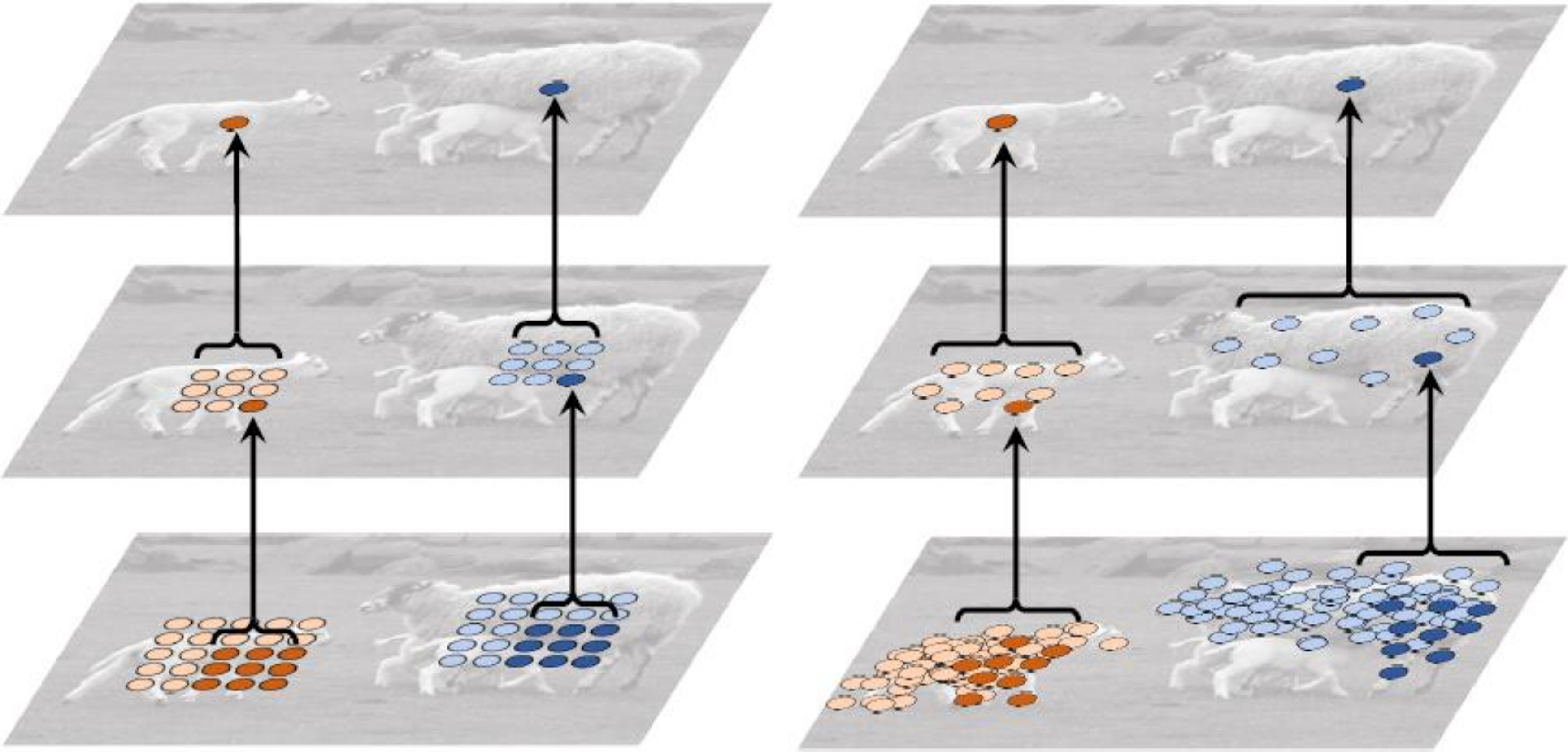


Deformable Convolutional Networks

Standard Convolution **Deformable Convolution**

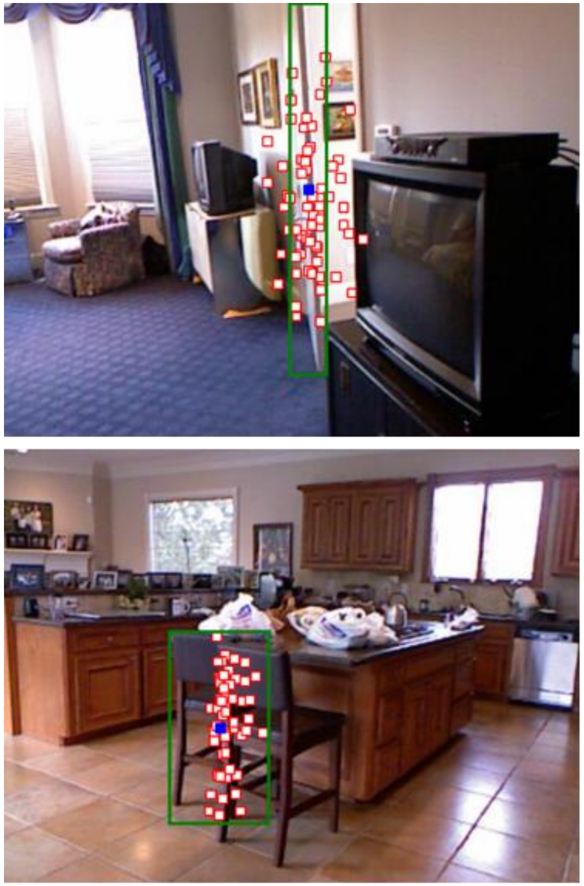


Deformable Convolutional Networks



(a) standard convolution (b) deformable convolution

Deformation in Different Modalities



(a) RGB Deformation

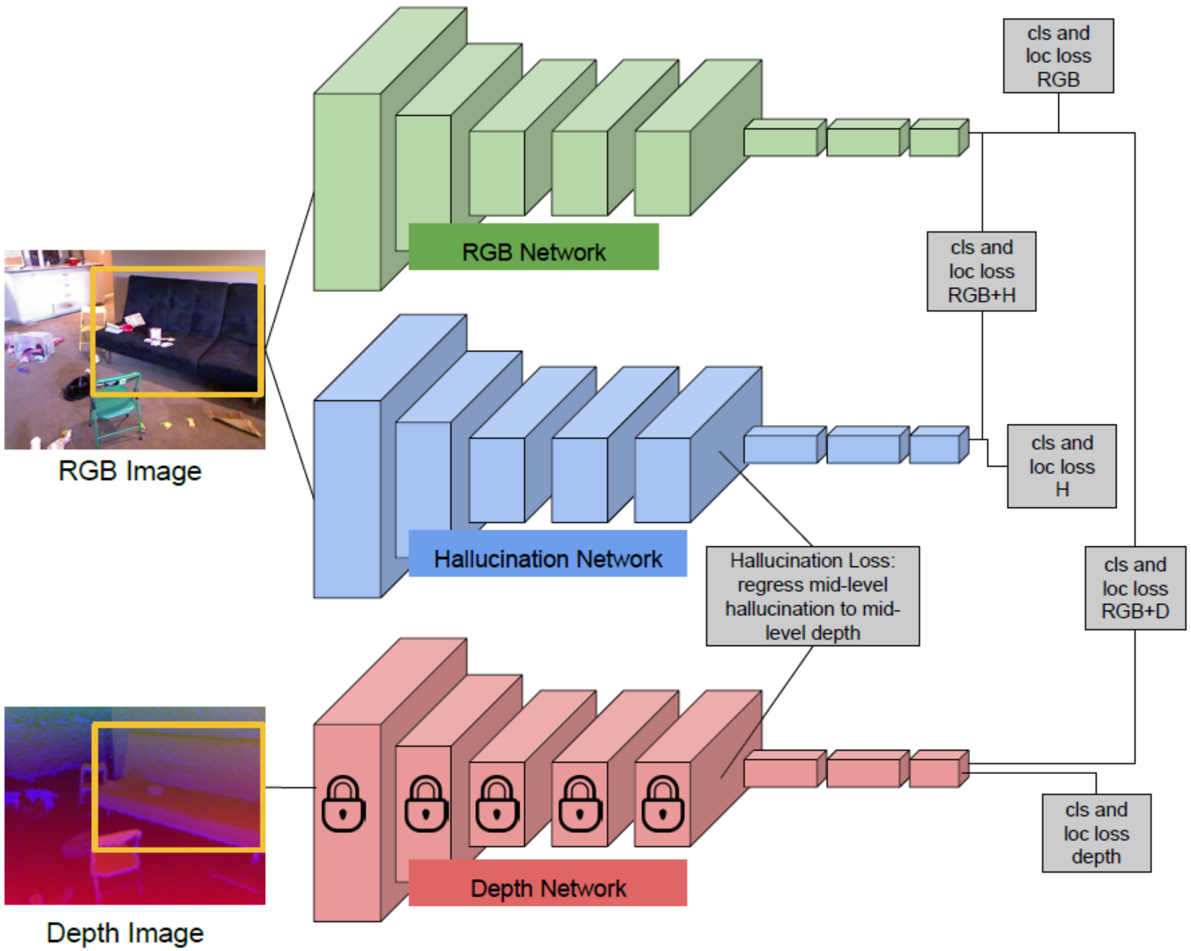
(b) Depth Deformation

Contents

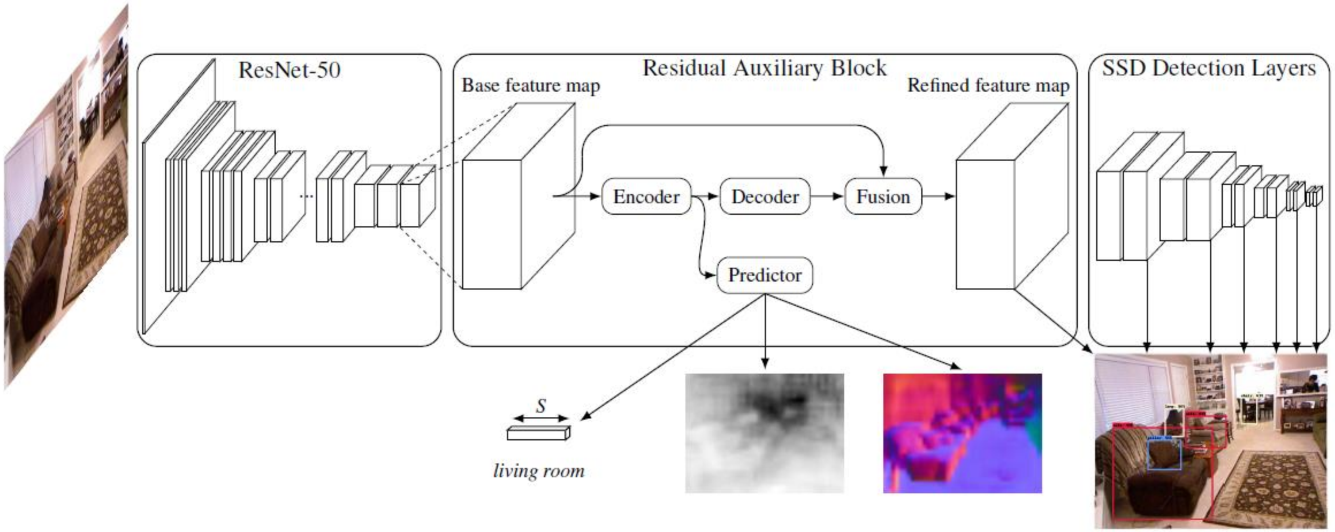
Depth Privileged Object Detection

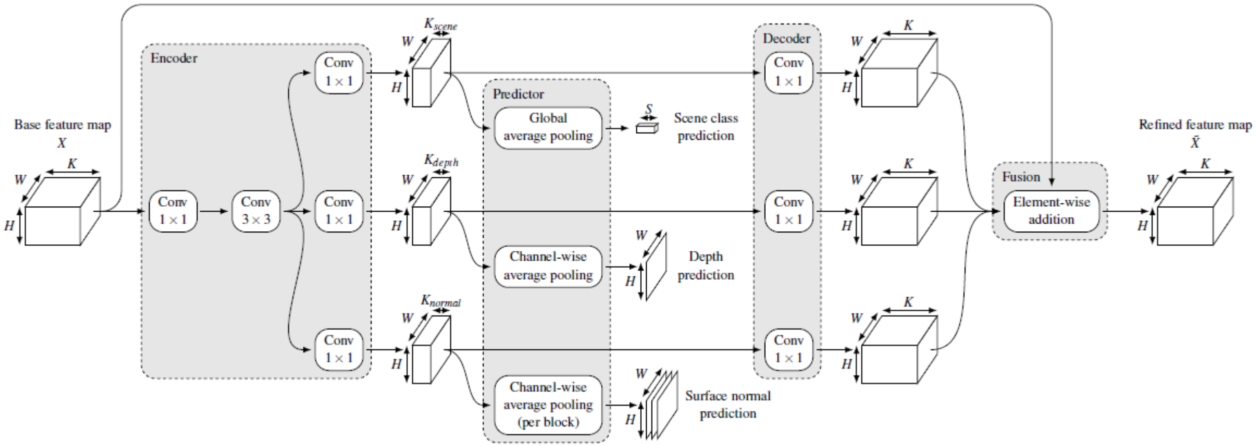
- Research lines:
  - Modality hallucination: HallucitantionNet , Cao et al. (2016)
  - Multi-task learning: ROCK

HallucinationNet



ROCK


ResNet-50
Residual Auxiliary Block
Base feature map
Refined feature map
SSD Detection Layers
Encoder
Decoder
Fusion
Predictor
S
living room

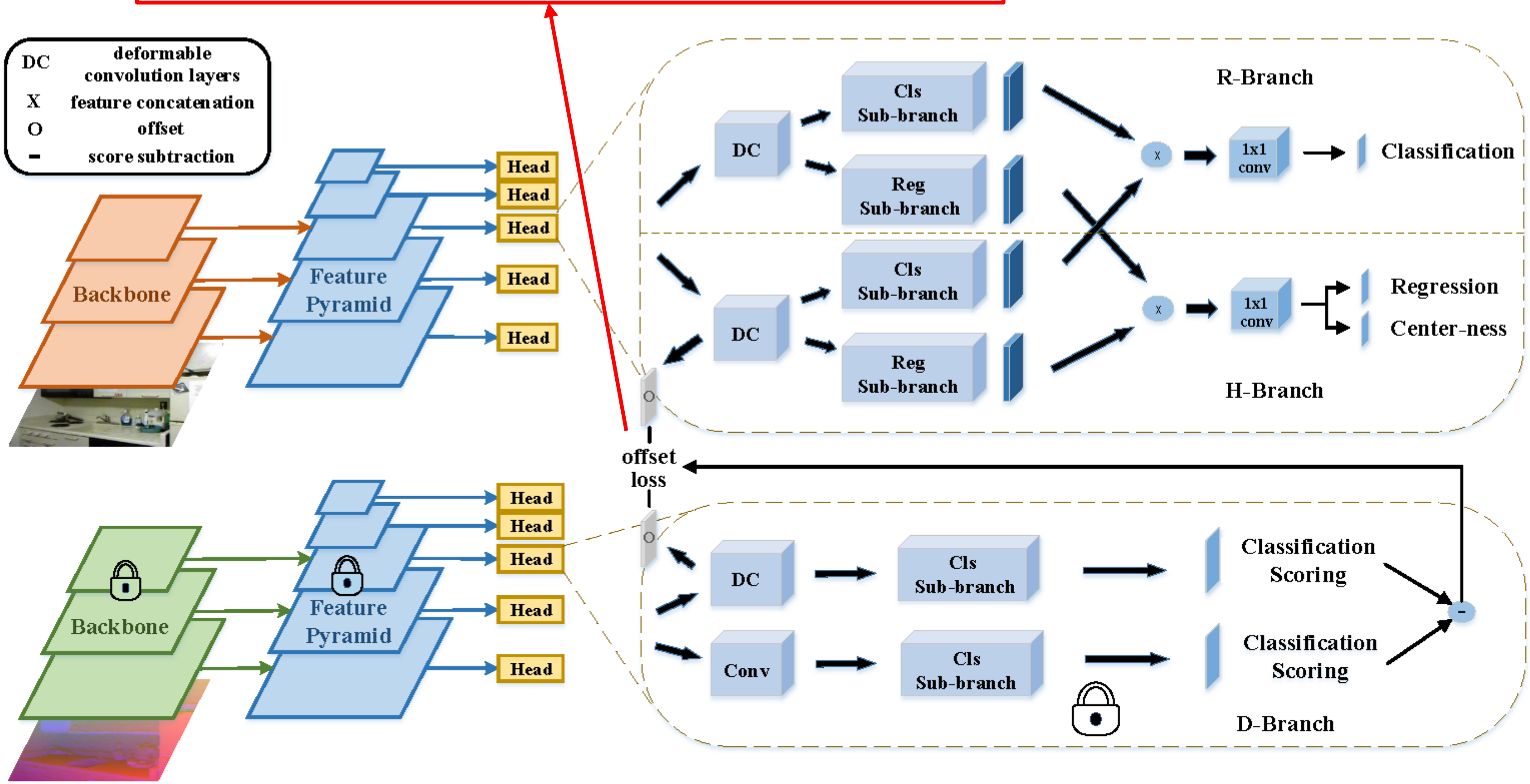

Base feature map
X
Encoder
Conv 1 × 1
Conv 3 × 3
Conv 1 × 1
Conv 1 × 1
Conv 1 × 1
Predictor
Global average pooling
Scene class prediction
Channel-wise average pooling
Depth prediction
Channel-wise average pooling (per block)
Surface normal prediction
Decoder
Conv 1 × 1
Conv 1 × 1
Conv 1 × 1
Fusion
Element-wise addition
Refined feature map

Contents

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Framework

$$L_o = \sum_{s=1}^{S} \sum_{m=1}^{M} \sum_{i=1}^{h_s * w_s} \sum_{k=1}^{K^2} \left\| \Delta l_{s,m,i,k}^{D} - \Delta l_{s,m,i,k}^{H} \right\|_2^2$$



Positive Position Transfer

- The depth deformation is not always reliable, transferring noisy deformation is likely to degrade the performance.

- Positive samples, which are determined by ATSS algorithm, has a major influence on the results (classification, regression)

$$L_o^p = \sum_{s=1}^{S} \sum_{m=1}^{M} \sum_{i=1}^{h_s * w_s} \sum_{k=1}^{K^2} P_{s,m,i} \left\| \Delta l_{s,m,i,k}^{\mathrm{D}} - \Delta l_{s,m,i,k}^{H} \right\|_2^2$$

Avoid Negative Transfer

- Classification scores (with/without dcn) can reflect the quality of deformation

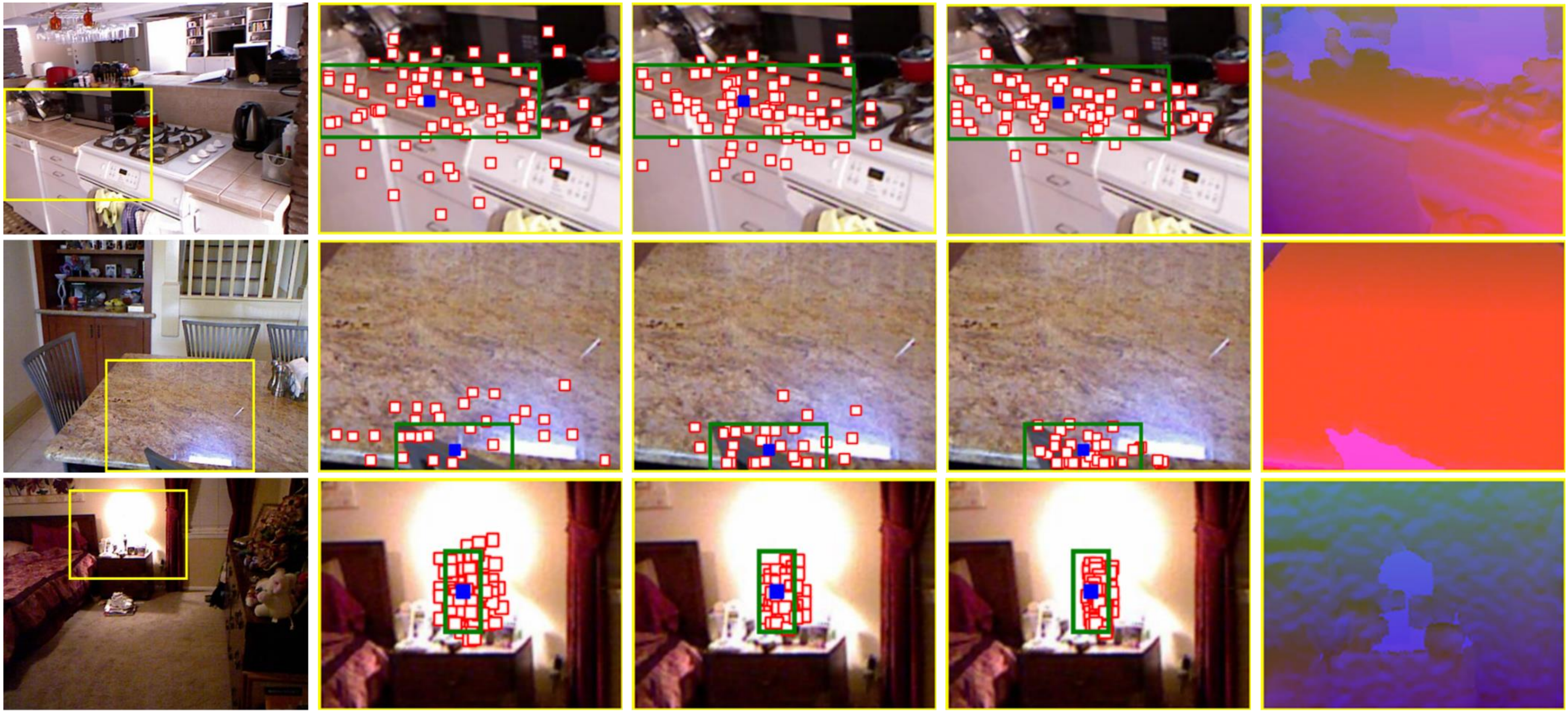
- we calculate the loss weights at different positions in previous DeformConvs by tracking their contributions to the classification score improvement.

$$w_{s,m,i} = \exp(\delta \cdot (f_{s,i}^{(w)} - f_{s,i}^{(w/o)}))$$

$$L_o^{pw} = \sum_{s=1}^{S} \sum_{m=1}^{M} \sum_{i=1}^{h_s * w_s} \sum_{k=1}^{K^2} w_{s,m,i} P_{s,m,i} \left\| \Delta l_{s,m,i,k}^{\mathrm{D}} - \Delta l_{s,m,i,k}^{H} \right\|_2^2$$

Contents

Visualization Results



(a) RGB images (b) RGB Deformation (c) Hallucinated Deformation (d) Depth Deformation (e) HHA-encoded images

Compared with SOTA Methods

L_d	L_o	L_o^p	L_o^{pw}	mAP(%)	
				NYUDv2	SUN RGB-D
√				44.01	53.93
√	√			46.26	56.15
√		√		46.50	56.47
√			√	46.88	56.84

Method	mAP(%)	
	NYUDv2	SUN RGB-D
FCOS+ATSS	42.73	52.94
HallucinationNet	45.22	55.35
ROCK	44.89	55.14
Cao et al.(2016)	44.96	55.27
Ours	46.88	56.84

Ablation Study

Configuration of DeformConvs

Configuration	mAP(%)	
	L_d	$L_d + \mu L_o$
B_1	43.66	45.34
B_2	44.17	45.93
B_3	44.43	46.19
H_1	43.35	45.36
H_2	44.04	46.26
H_3	44.21	46.30
B_3+H_2	44.25	46.32

Fusion Strategy Analyses

mAP(%)	After DeformConvs	After sub-branch
Concatenation	46.32	46.88
Addition	45.79	46.15

Evaluation on RGB-only dataset

Model	mAP(%)
RGB-only	38.96
Ours	40.68
RGB-only(ft)	81.01
Ours(ft)	82.57

Depth Privileged Object Detection in Indoor Scenes via Deformation Hallucination

Thanks For Watching!

Zhijie Zhang, Yan Liu, Junjie Chen, Li Niu, Liqing Zhang

MoE Key Lab of Artificial Intelligence, Department of Computer Science and Engineering, Shanghai Jiao Tong University