Learning with Side Information through Modality Hallucination

Judy Hoffman Saurabh Gupta Trevor Darrell
CVPR 2016

Nan Yang
Supervisor: Benjamin Busam

Computer Aided Medical Procedures
Technische Universität München

15.12.2016

Outline

Introduction

Use available paired RGB-d training data to learn to hallucinate mid-level convolutional features from a RGB image for **object detection** task.

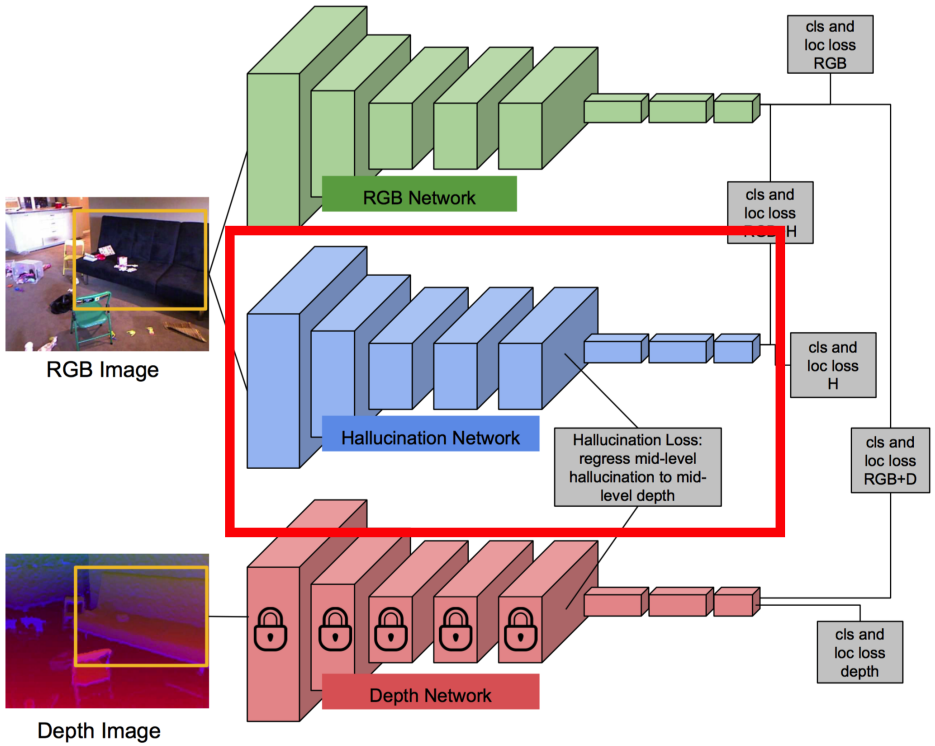


Figure: Network Structure[1]

Introduction

- Task: Object Detection
- Framework: Deep Learning (Convolutional Neuron Networks)
- What's new: Additional ConvNets to extract depth features from RGB image (**modality hallucination**)

Results

RGB object detection:

34.0 mAP (this work) vs 29.9 mAP (Fast R-CNN)

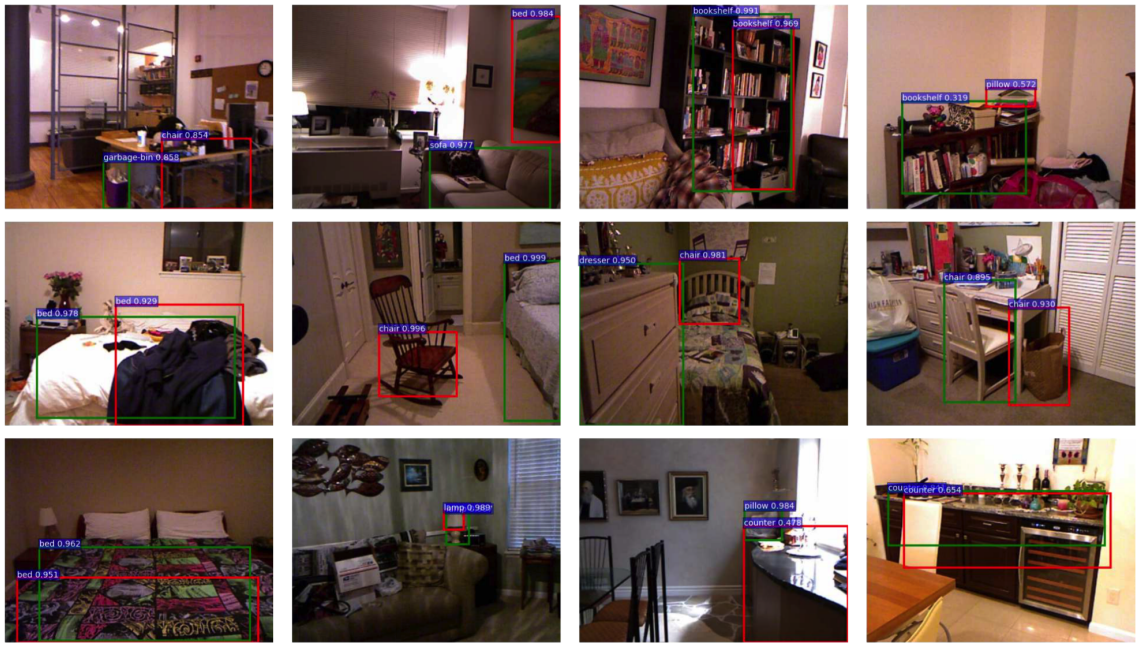


Figure 4: Example Detections on the NYUD2 *test set* where our RGB hallucination network's (green box) top scoring detection for the image is correct while the baseline RGB detector's (red box) top scoring detection is incorrect.

Related Work

- Object detection using deep learning
  - RGB[6]
  - RGB-d[7]
- Transfer Learning[8][9]
- Learning using side information[10]
- Network distillation[11][12]

Related Work

- **Object detection using deep learning**
  - RGB[6]
  - RGB-d[7]
- Transfer Learning[8][9]
- Learning using side information[10]
- Network distillation[11][12]

Object Detection

- Extract region of interest (region proposal)
- Classification of this region
- Result: Region (bounding box coordinates) + Category

Object Detection

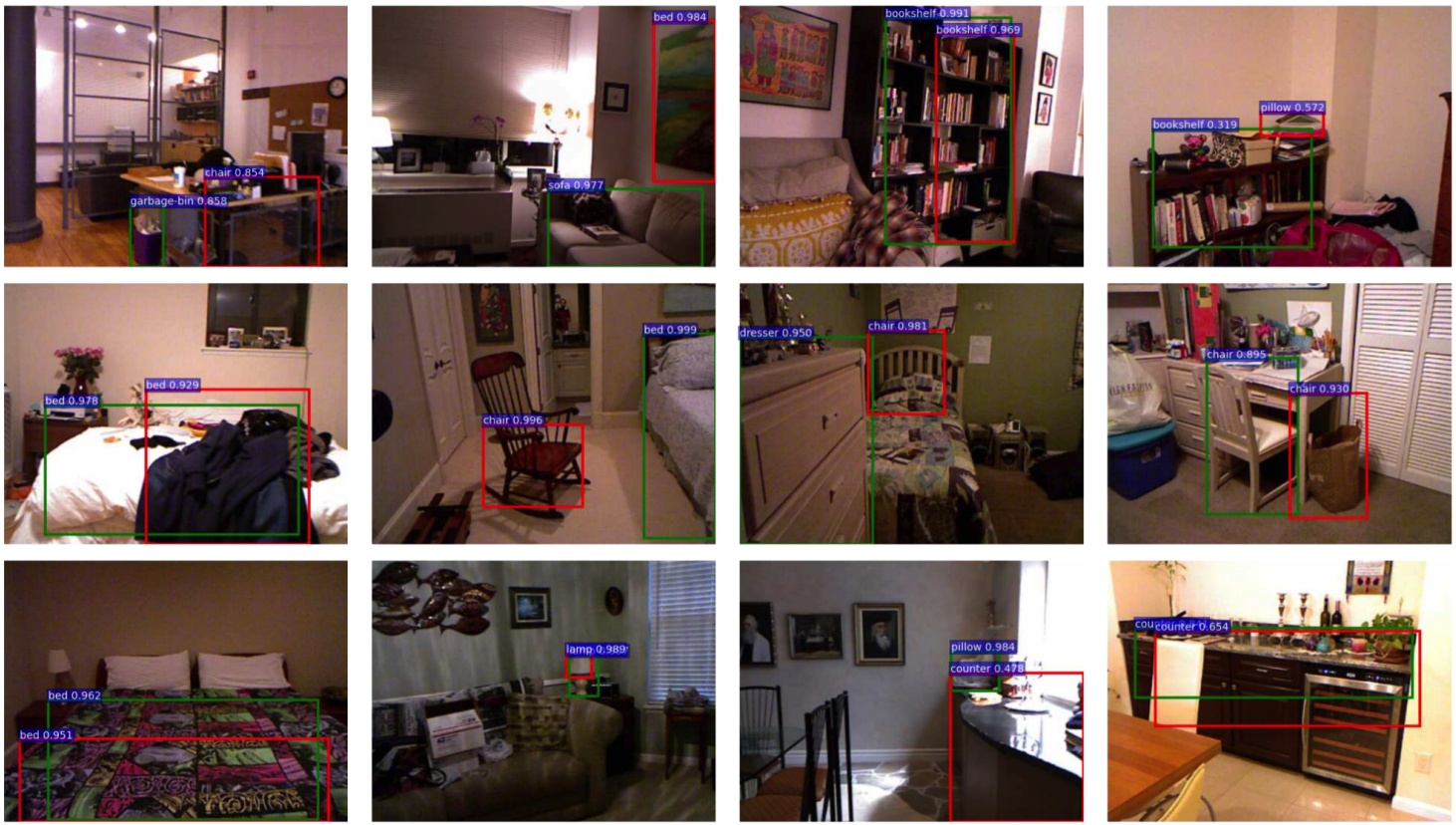
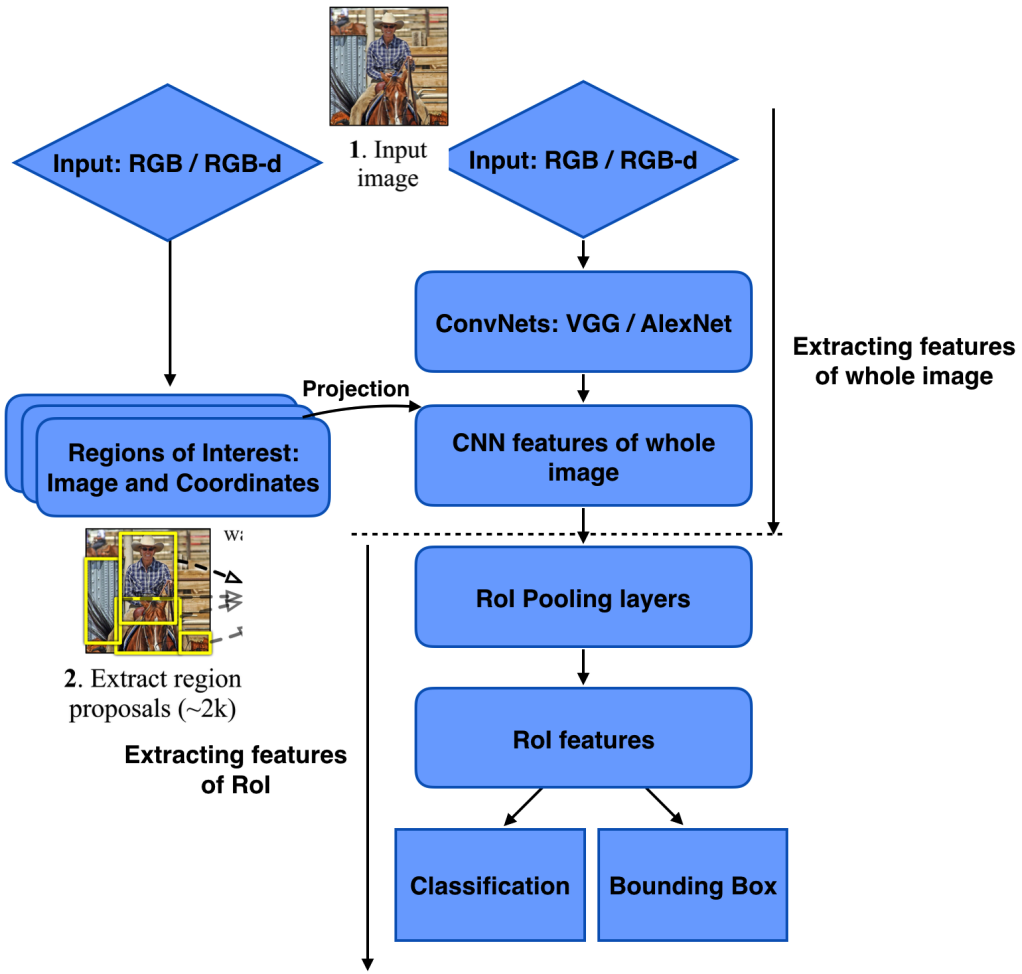


Figure 4: Example Detections on the NYUD2 *test set* where our RGB hallucination network's (green box) top scoring detection for the image is correct while the baseline RGB detector's (red box) top scoring detection is incorrect.

Object Detection Pipeline (Fast R-CNN[6])



Object Detection Pipeline (Fast R-CNN[6])

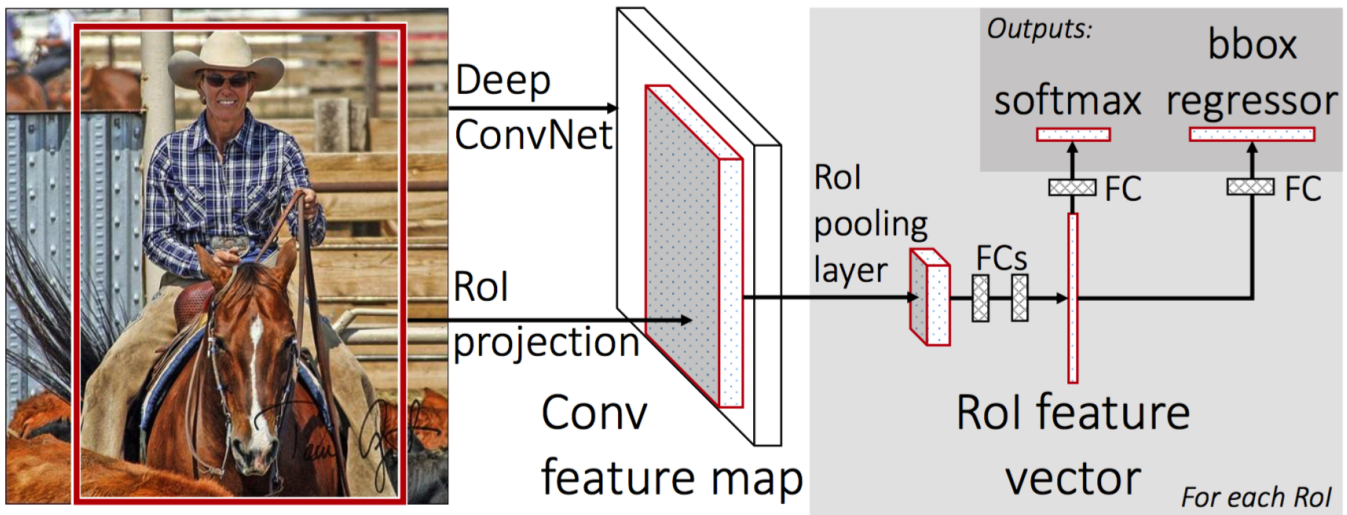


Figure: F-RCNN Structure[6]

State of the art

- Fast R-CNN (Mean Average Precision: 29.9%)[6]
- RGB-d Fast R-CNN (Mean Average Precision: **44.4%**)[7]

Observation

Object detection can be improved significanlty with
RBD-d images (additional modality, complementary information).

Problem

Not all RBG images have corresponding depth information.

Idea

How about extracting depth features from RGB images?

Related Work

- Object detection using deep learning
  - RGB[6]
  - RGB-d[7]
- **Transfer Learning**[8][9]
- **Learning using side information**[10]
- **Network distillation**[11][12]

Transfer Learning[8][9]

- Transfer learning and domain adaptation which learns to share information from one task (**Obj detection using depth features**) to another (**Obj detection with RGB features**).
- Previous work used depth information at training time to inform RGB test time detection by learning transformations into a **single representation** for the joint modality space.
- This work focuses on learning an **additional RGB representation**, which is the **hallucination network**.

Learning using side information[10][13]

- Learning algorithm has additional knowledge at training time, whether meta data or an **additional modality**.
- Use this extra information to inform training of a **stronger model** than could be produced otherwise, e.g. surface normals at training time could produce detection improvement.

Network distillation[11][12]

- Network Distillation: The output from one network (**depth network**) is used as the target probability distribution for a new network (**hallucination network**). Input can be different.
- This approach can also be seen as using distillation to learn representations on RGB images by **transferring supervision** from paired depth images, by employing joint training.

Hallucination Network (Training time)

Sharing depth modality infomation with RGB through hallucination network.

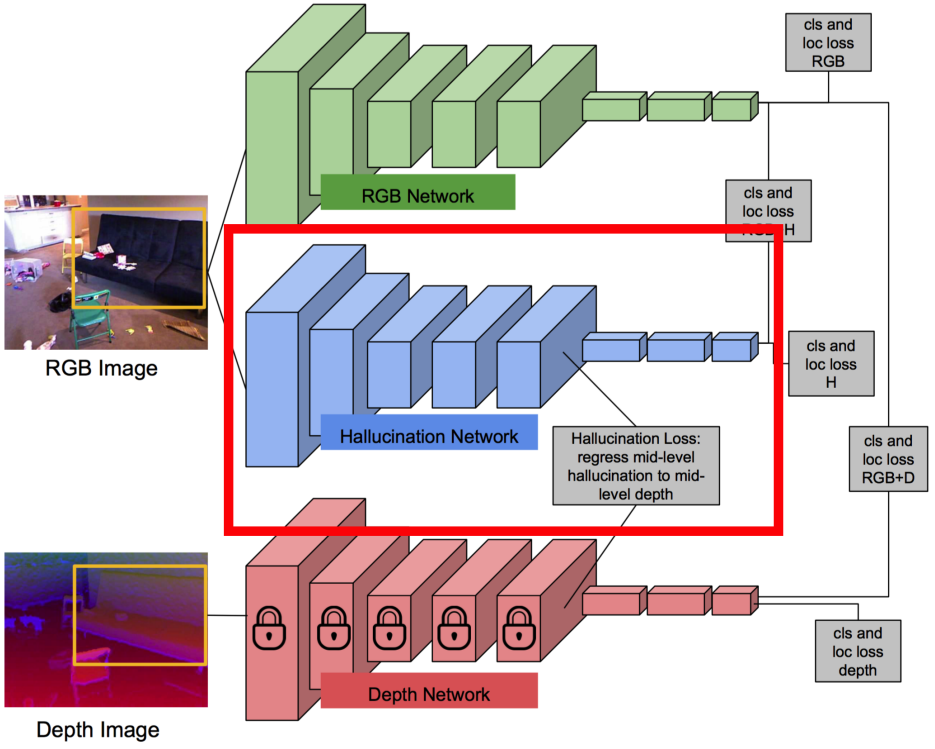


Figure: Network Structure[1]

Hallucination Network (Test time)

Given **only a RGB image** and regions of interest, pass through both the RGB network and the hallucination network to produce **two scores** per category, per region, which we **average** and take the softmax to produce our final predictions.

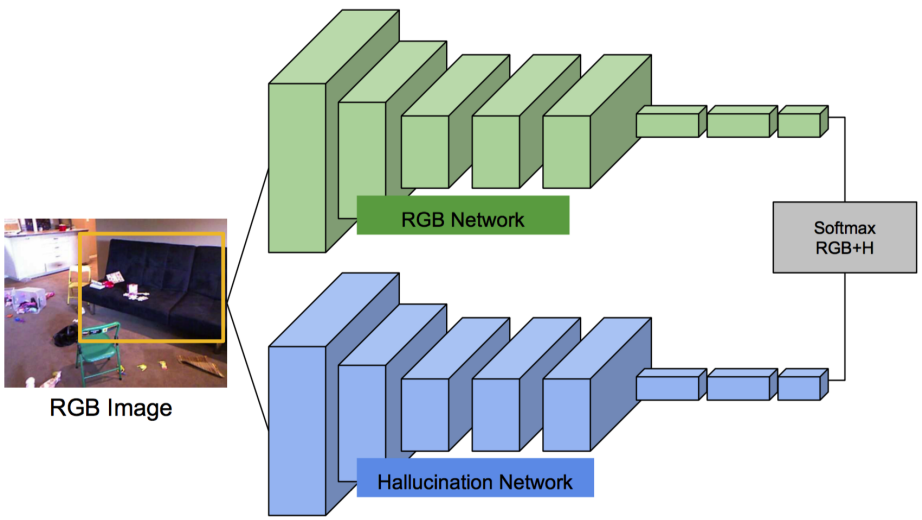


Figure: Test Network Structure[2]

Explanation

1. A convolutional neuron network (**Hallucination network**) whose input is a **RGB image**.

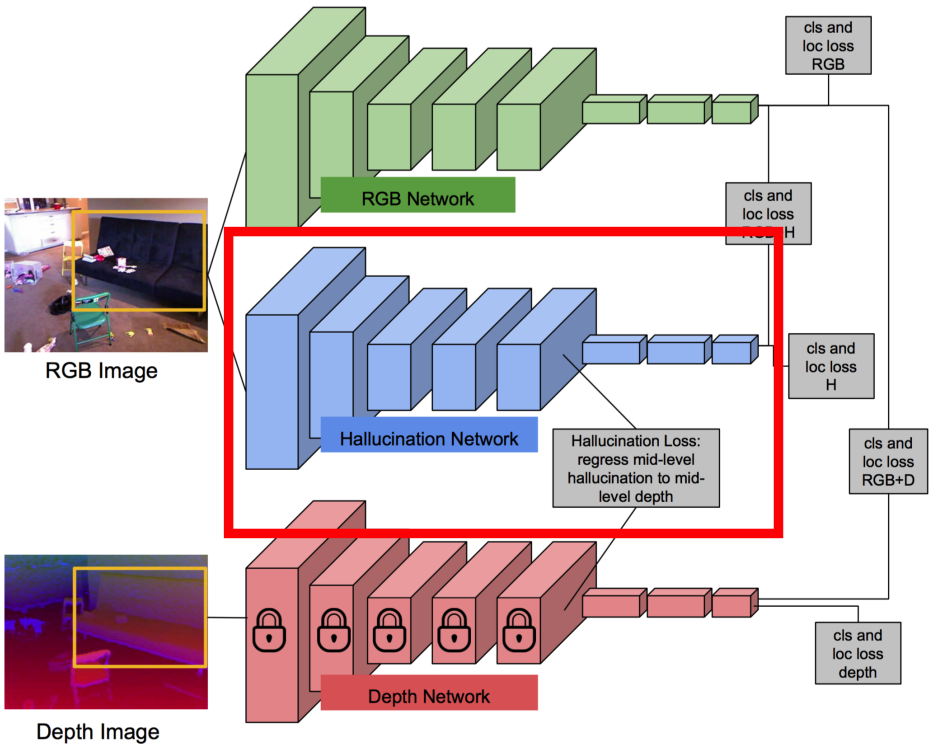


Figure: Network Structure[1]

Explanation

2. The **features maps** generated by this network is as identical as possible to the ones generated by depth ConvNets.

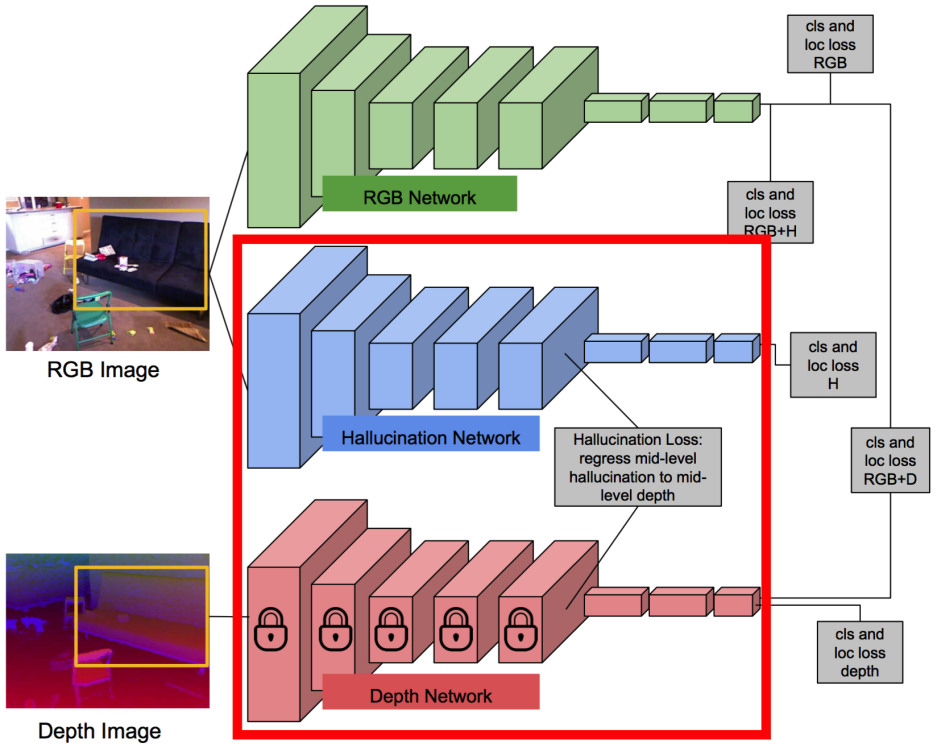


Figure: Network Structure[1]

Explanation

3. A network whose input is RBG image generates depth features.

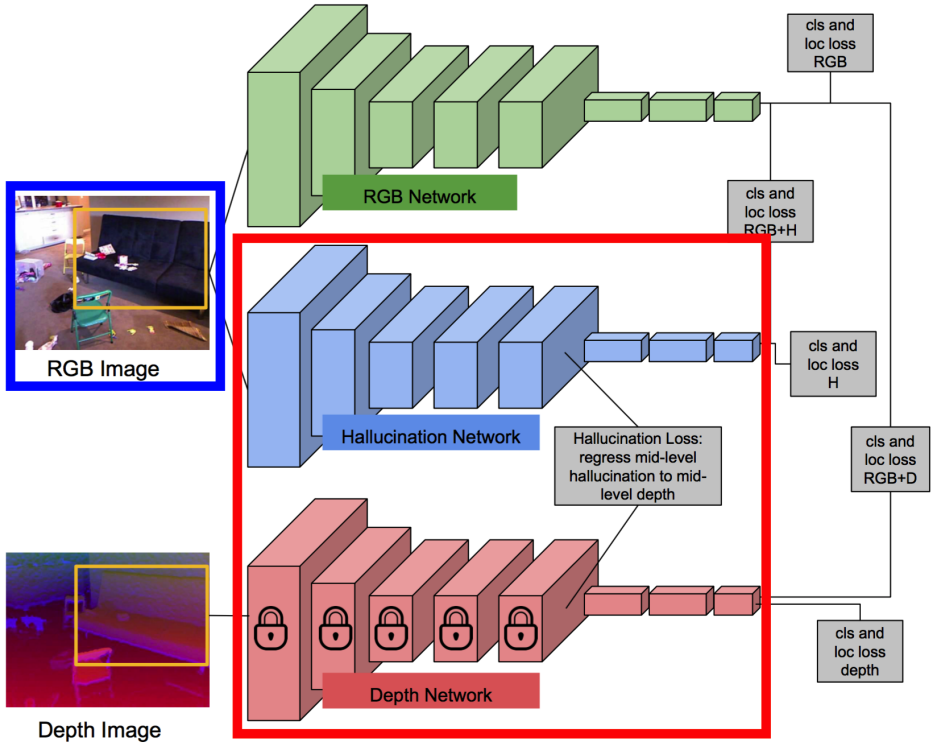


Figure: Network Structure[1]

Explanation

4. Similar feature maps but different convolutional kernels (weights) because of different input modalities.

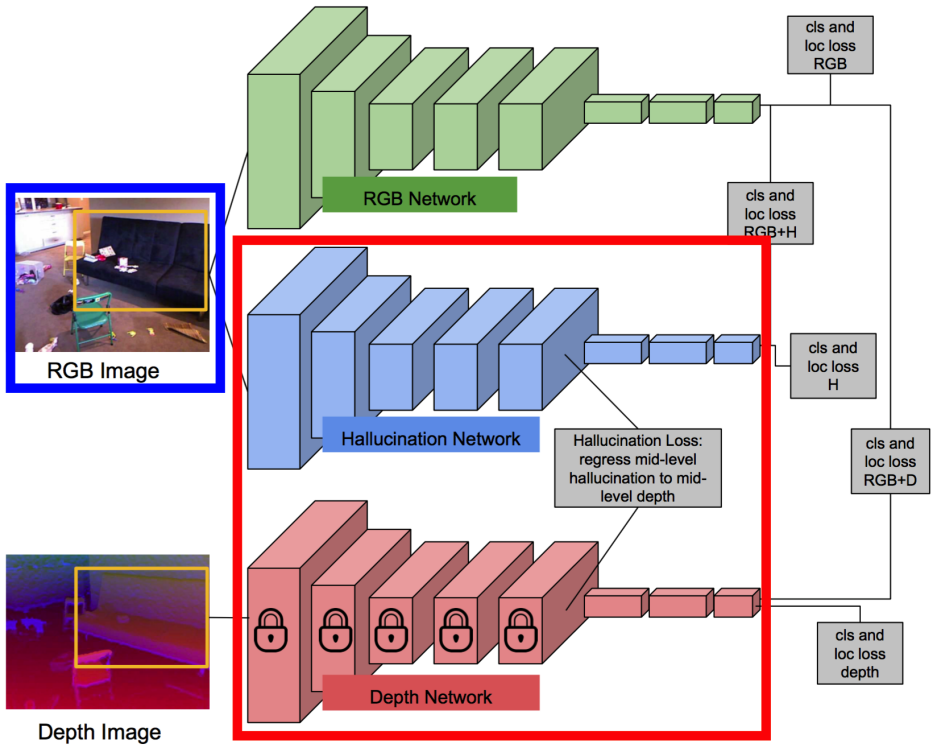


Figure: Network Structure[1]

Architecture Definition

Multi channels ConvNets:

- RGB Network
- Hallucination Network
- Depth Network

The base network structures can be AlexNet / VGG (to extract features).

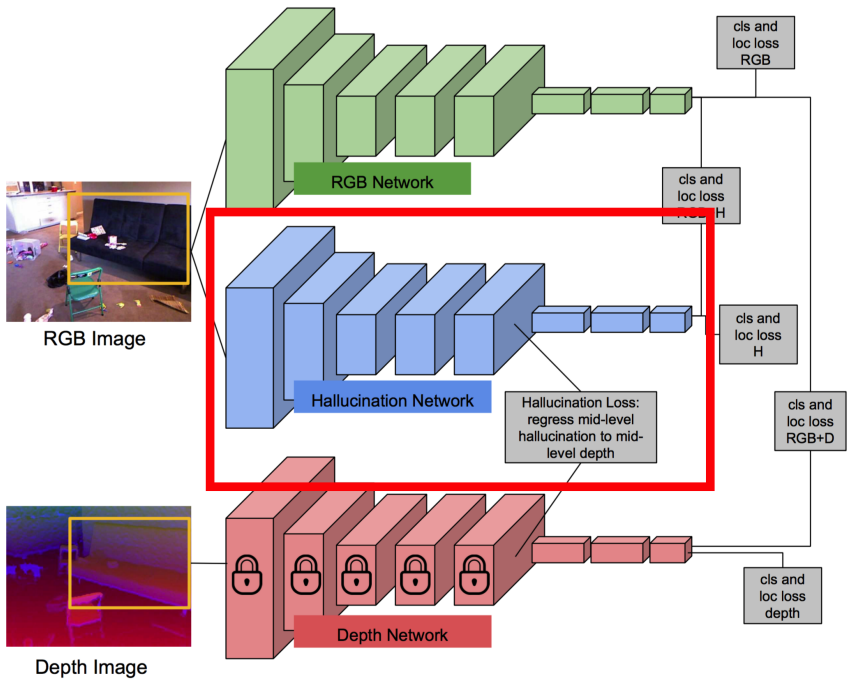


Figure: Network Structure[1]

Modality Hallucination Model

Training time

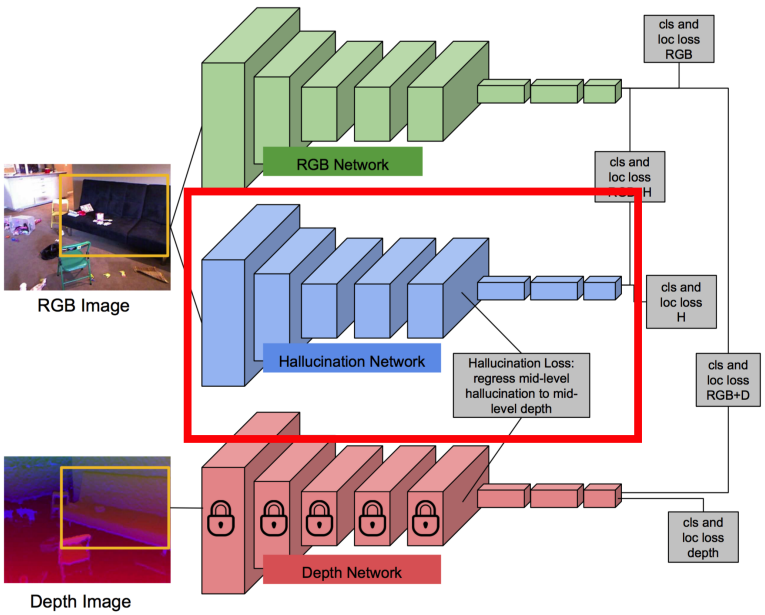


Figure: Network Structure[1]

Testing time

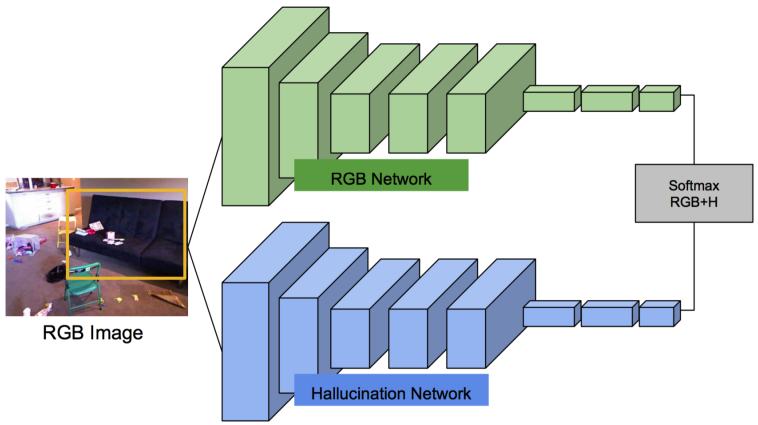
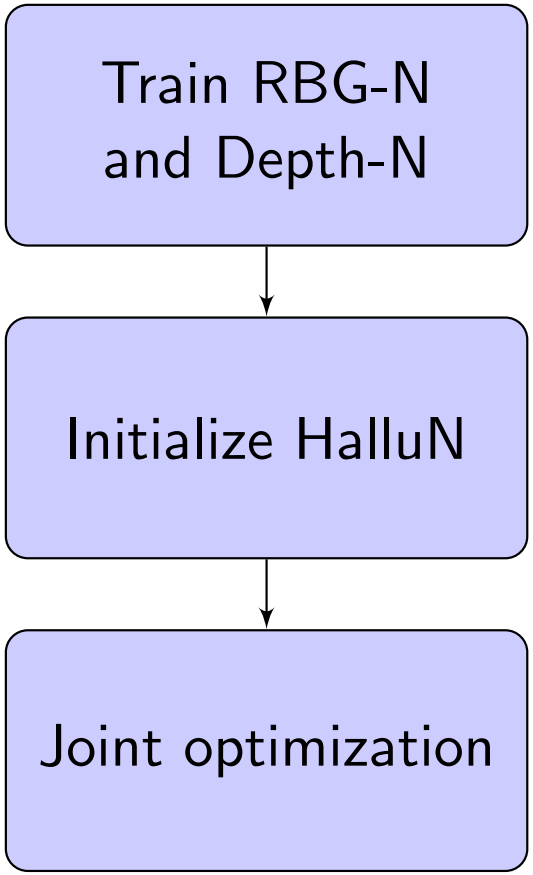


Figure: Test Network Structure[2]

Training Pipeline



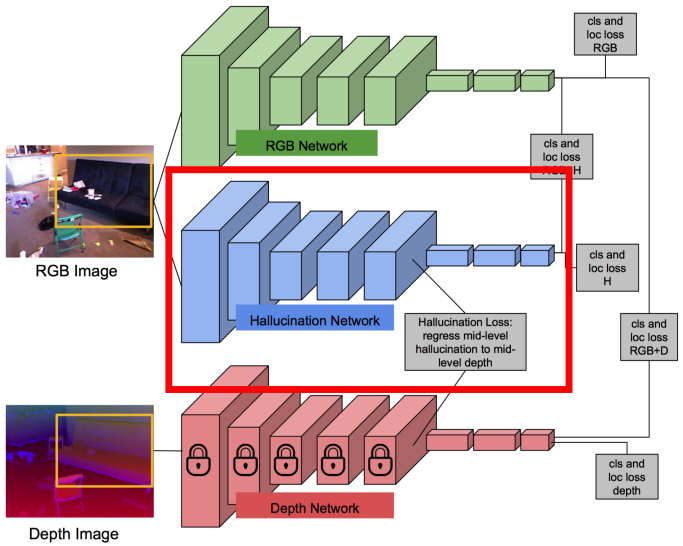
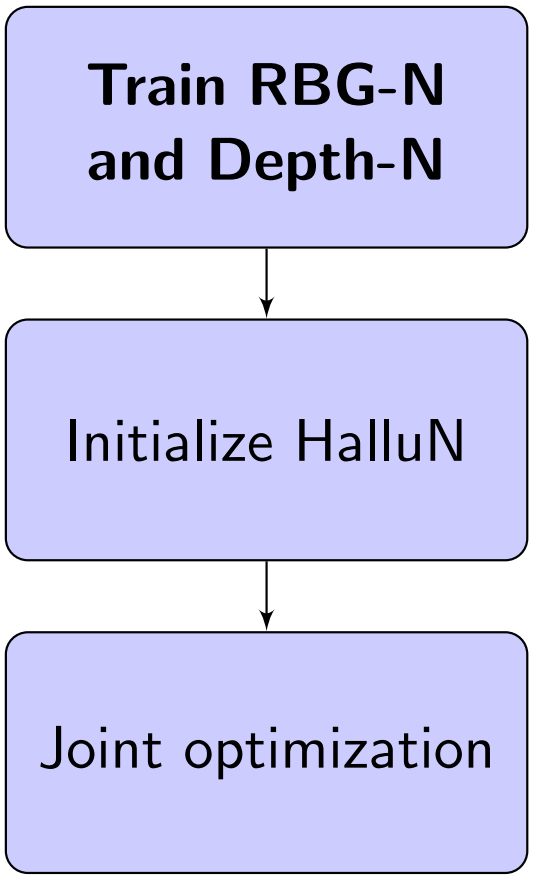


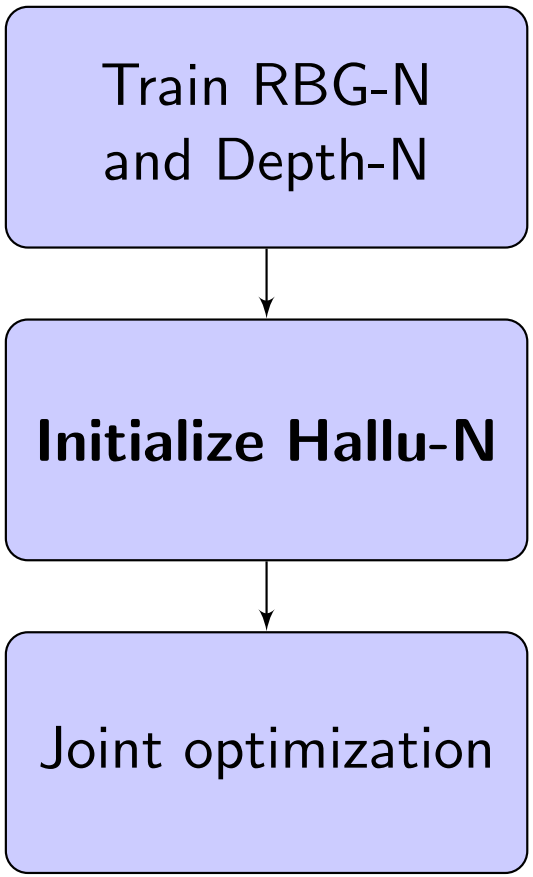
Figure: Network Structure[1]

Training Pipeline



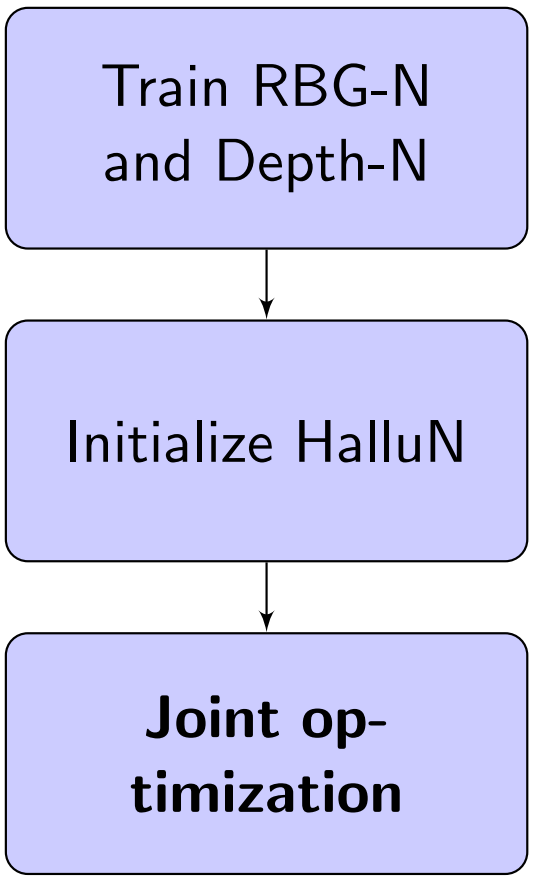
RGB and depth network are independently trained using the Fast R-CNN algorithm with the corresponding image input.

Training Pipeline



The hallucination network parameters are initialized with the learned **depth network weights**.

Training Pipeline



Multi-task Optimization with **hallucination loss**.

Hallucination Loss

Activations after some layer, l, should be similar between the hallucination and depth networks.

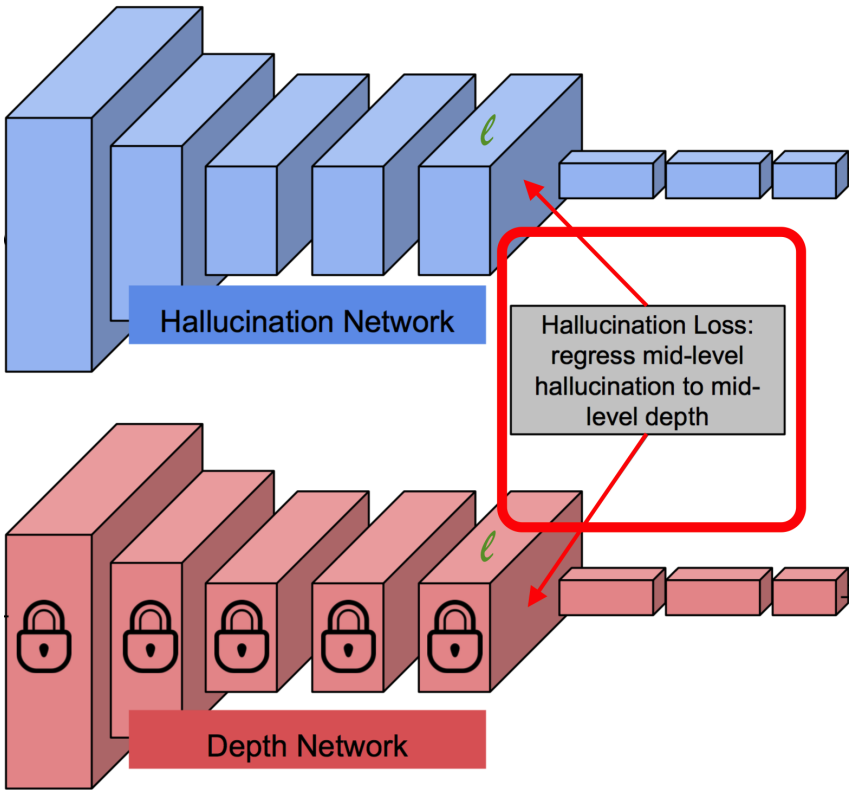


Figure: Network Structure[1]

Hallucination Loss

Euclidean loss: $L_{hallucinate}(l) = \left\| \sigma(A_l^{dNet}) - \sigma(A_l^{hNet}) \right\|_2^2$
where $\sigma(x) = 1/(1 + \exp(-x))$ (sigmoid function) and $l = layer$

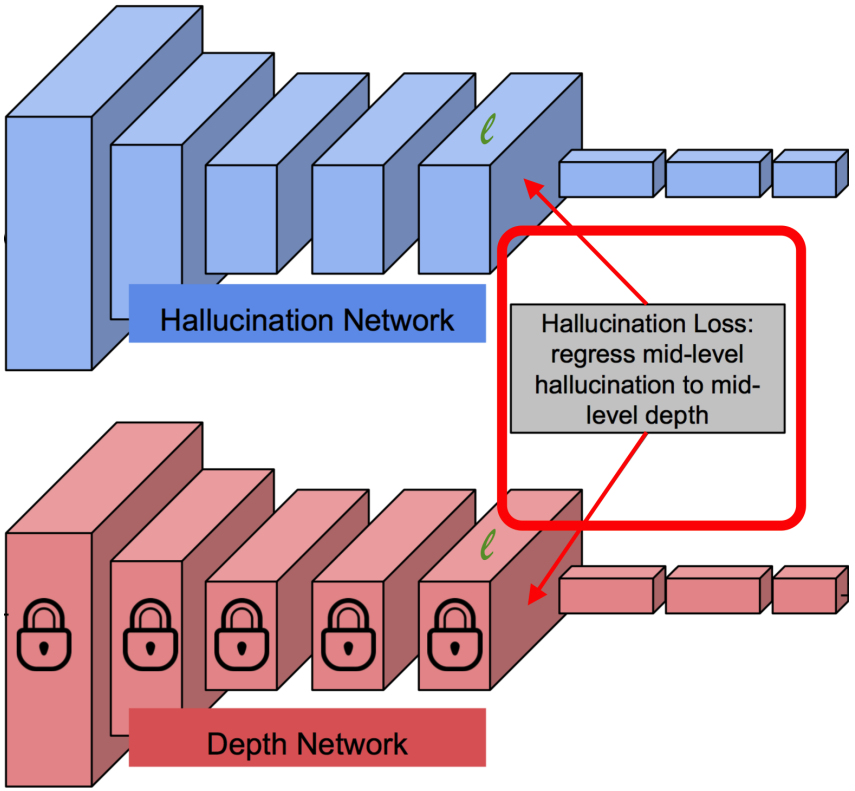


Figure: Network Structure[1]

Hallucination Loss

Asymmetric transfer of information: **0 learning rate** lower than l

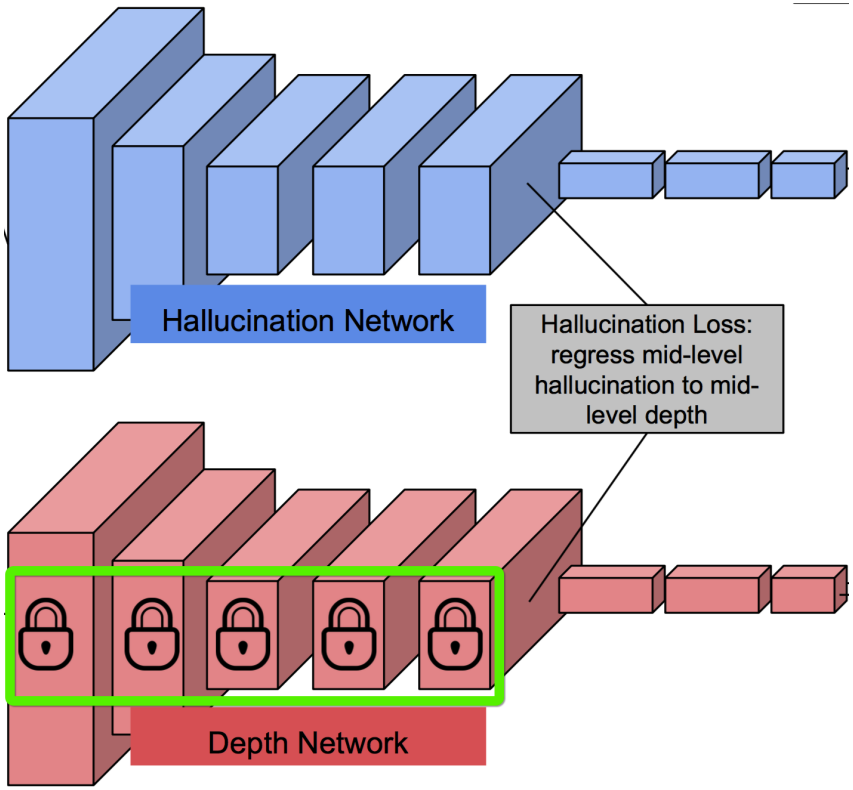
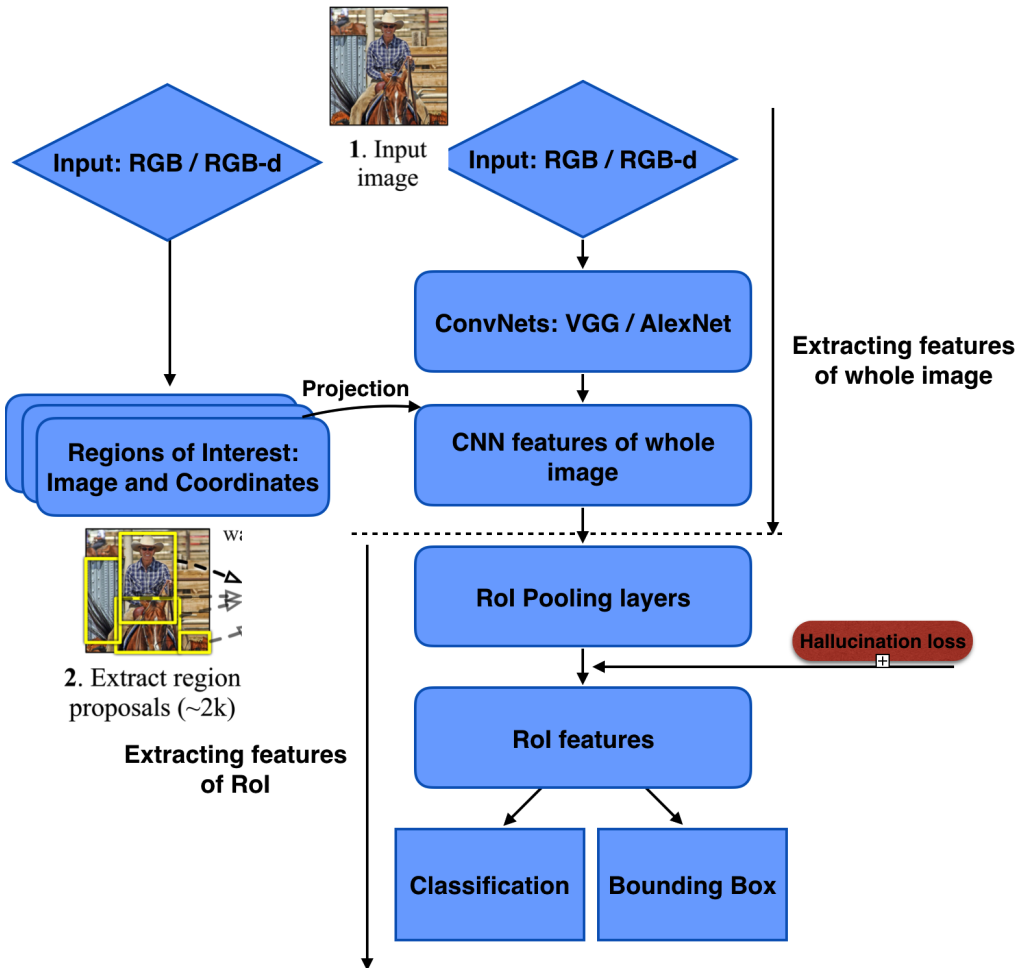


Figure: Network Structure[1]

Hallucination Loss



Multi-task Optmization

11 total losses:

- 5 softmax cross-entropy losses for **classification**
- 5 smooth L1 losses for **bounding box** coordinates regression
- 1 hallucination loss

Multi-task Optmization

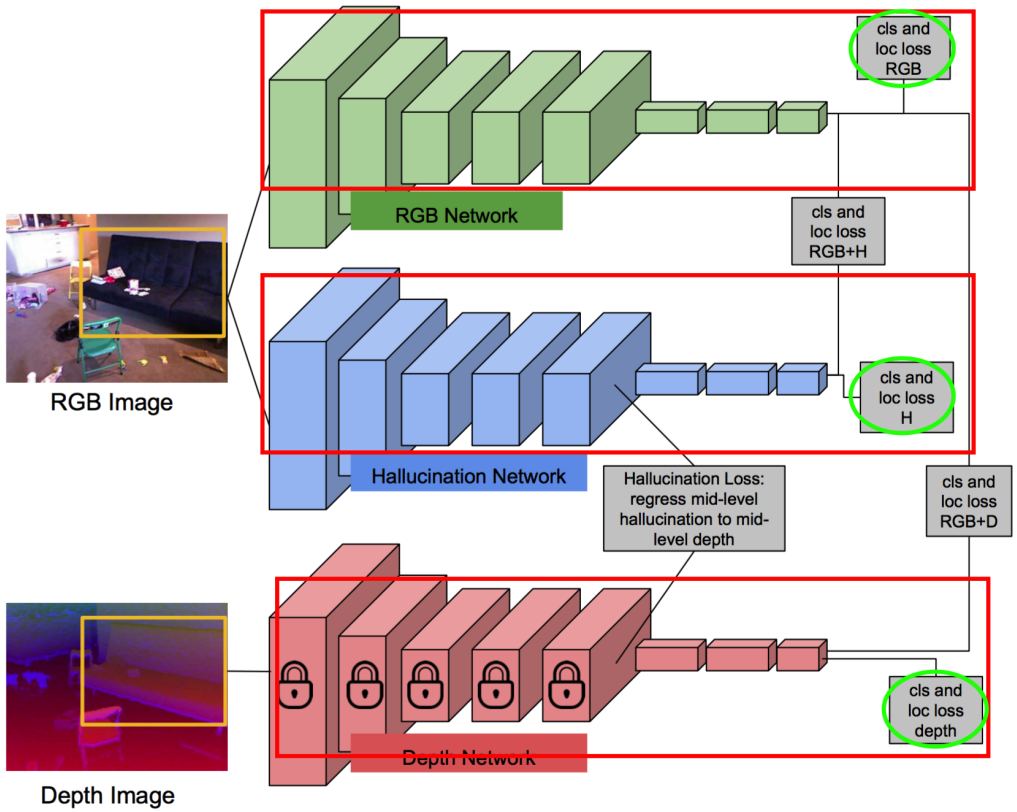


Figure: Network Structure[1]

Multi-task Optmization

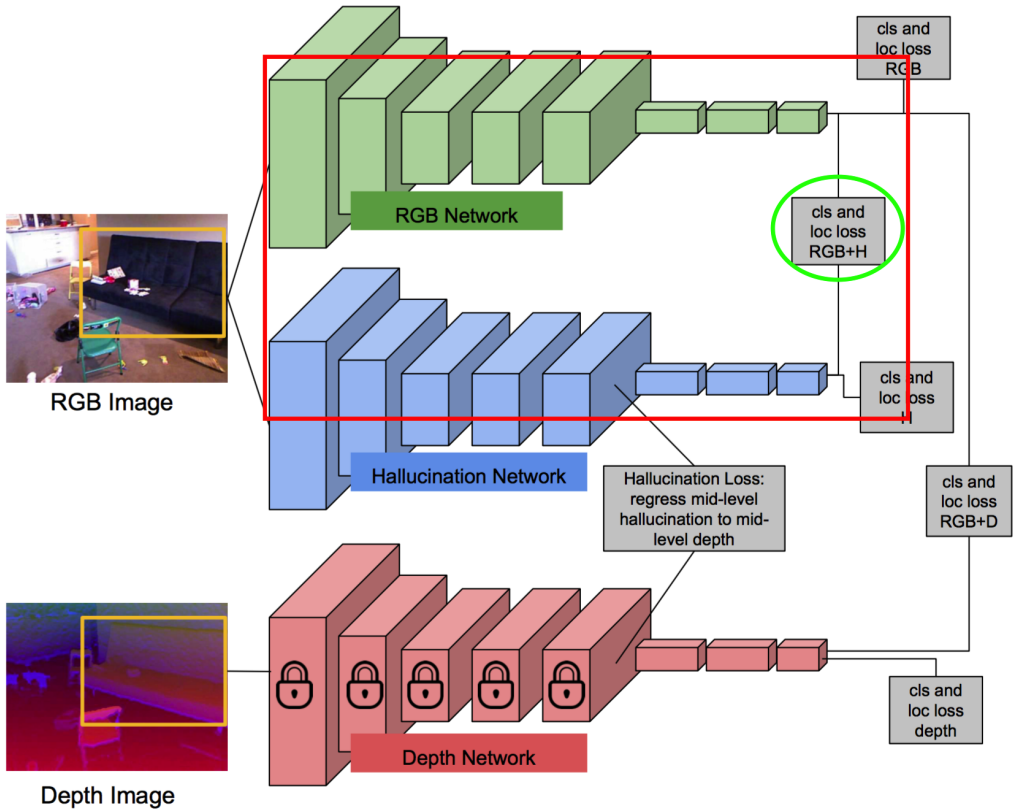


Figure: Network Structure[1]

Multi-task Optmization

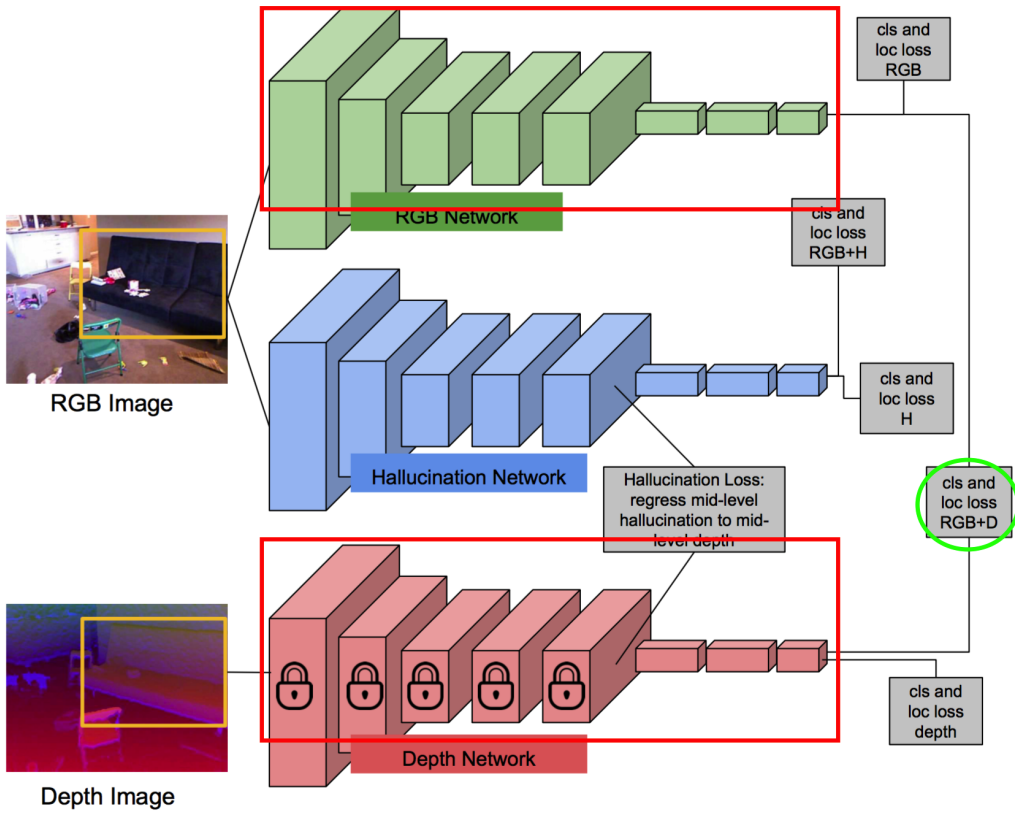


Figure: Network Structure[1]

Multi-task Optmization

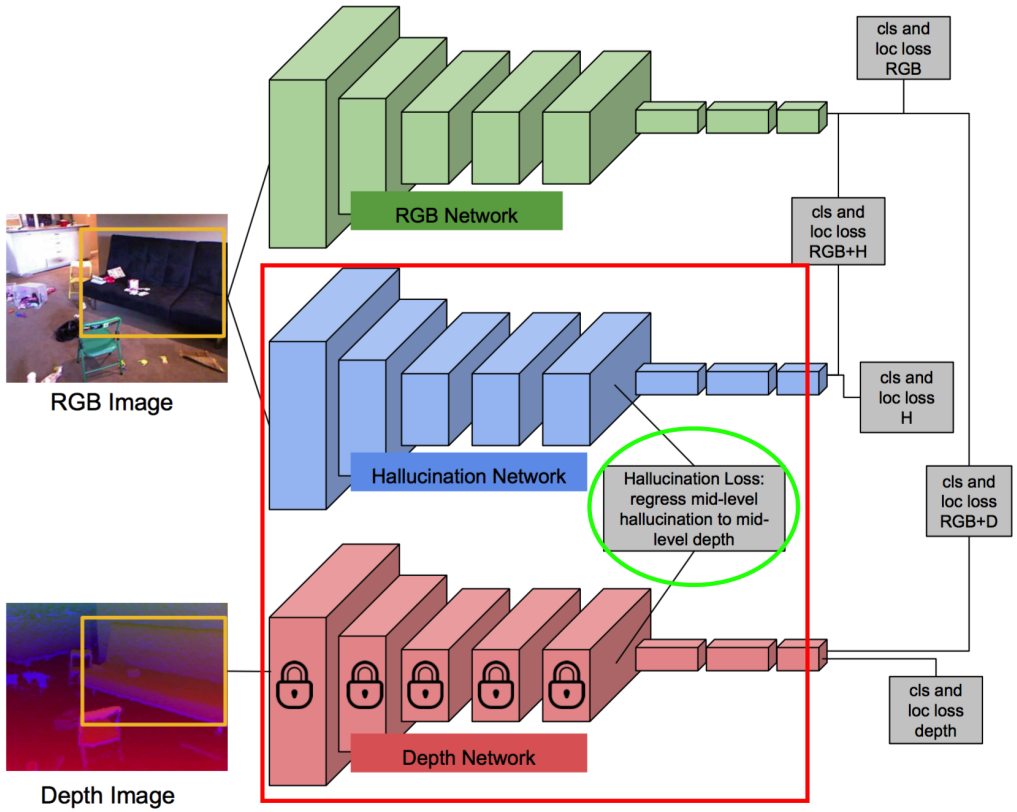


Figure: Network Structure[1]

Balancing objectives

$$
\begin{aligned}
\mathcal{L} \quad = \quad & \gamma \mathcal{L}_{\text{hallucinate}} \\
& + \alpha \left[\mathcal{L}_{\text{loc}}^{\text{dNet}} + \mathcal{L}_{\text{loc}}^{\text{rNet}} + \mathcal{L}_{\text{loc}}^{\text{hNet}} + \mathcal{L}_{\text{loc}}^{\text{rdNet}} + \mathcal{L}_{\text{loc}}^{\text{rhNet}} \right] \\
& + \beta \left[\mathcal{L}_{\text{cls}}^{\text{dNet}} + \mathcal{L}_{\text{cls}}^{\text{rNet}} + \mathcal{L}_{\text{cls}}^{\text{hNet}} + \mathcal{L}_{\text{cls}}^{\text{rdNet}} + \mathcal{L}_{\text{cls}}^{\text{rhNet}} \right]
\end{aligned}
$$

$$\alpha = 0.5, \beta = 1, \gamma = ?$$

γ is heuristically set to **10x** the contribution of any of the other losses

Experiments

- Base Network: AlexNet, VGG
- Region Proposals: Multiscale Combinatorial Grouping (MCG)
- SGD Hyper-parameters: LR = 0.001, Momentum = 0.9, Weight decay = 0.0005, T = 10 (Threshold of gradient glipping)

NYUD2 Detection Evalutation

A - AlexNet, V - VGG

method	btub	bed	bshelf	box	chair	counter	desk	door	dresser	gbin	lamp	monitor	nstand	pillow	sink	sofa	table	tv	toilet	mAP
RGB only [10] (A)	7.5	50.6	**36.8**	1.4	30.2	34.9	10.8	21.5	27.8	16.9	26.0	32.6	20.6	**25.1**	31.6	**36.7**	14.8	25.1	54.6	26.6
RGB ensemble (A-A)	10.5	53.7	33.6	1.6	32.0	34.8	12.2	20.8	34.5	19.6	28.6	45.7	28.5	24.4	31.4	34.7	14.5	**34.0**	**56.1**	29.0
Our Net (A-RGB, A-H)	**13.9**	**56.1**	34.4	**1.9**	**32.9**	**40.5**	**12.9**	**22.6**	**37.4**	**22.0**	**28.9**	**46.2**	**31.9**	22.9	**34.2**	34.2	**19.4**	33.2	53.6	**30.5**
RGB only [10] (V)	15.6	59.4	38.2	1.9	33.8	36.3	12.1	24.5	31.6	18.6	25.5	46.5	30.1	20.6	30.3	40.5	19.5	**37.8**	45.7	29.9
RGB ensemble (A-V)	14.8	60.4	**43.1**	**2.1**	36.4	40.7	13.3	**27.1**	35.5	20.8	29.9	**52.9**	**33.5**	26.2	33.0	44.4	19.9	36.7	50.2	32.7
Our Net (A-RGB, V-H)	**16.8**	**62.3**	41.8	**2.1**	**37.3**	**43.4**	**15.4**	24.4	**39.1**	**22.4**	**30.3**	46.6	30.9	**27.0**	**42.9**	**46.2**	**22.2**	34.1	**60.4**	**34.0**

Figure: NYUD2 evaluation[3]

NYUD2 Detection Evalutation

A - AlexNet, V - VGG

method	mAP
RGB only [10] (A)	26.6
RGB ensemble (A-A)	29.0
Our Net (A-RGB, A-H)	**30.5**
RGB only [10] (V)	29.9
RGB ensemble (A-V)	32.7
Our Net (A-RGB, V-H)	**34.0**

Figure: NYUD2 evaluation[3]

How to initialize the hallucination net?

Initial Weights	bathtub	bed	bshelf	box	chair	counter	desk	door	dresser	gbin	lamp	monitor	nstand	pillow	sink	sofa	table	tv	toilet	mAP
RGB	7.5	50.4	9.9	**0.9**	**26.2**	**24.9**	5.8	**15.8**	13.0	**29.8**	12.0	43.1	20.9	14.7	17.9	25.3	**15.1**	32.5	59.1	22.4
depth	9.9	**52.4**	**14.9**	**0.9**	24.9	24.4	4.3	15.3	18.1	24.1	**14.8**	**45.8**	**27.2**	**18.5**	**21.3**	**29.0**	13.7	**33.6**	**66.4**	**24.2**
random	**10.5**	47.6	12.3	0.6	23.5	20.2	**6.0**	13.0	**19.3**	12.0	13.3	42.8	12.8	12.1	13.6	23.0	13.9	28.6	61.5	20.3

Figure: Initialization evaluation[4]

How to initialize the hallucination net?

Initial Weights	mAP
RGB	22.4
depth	**24.2**
random	20.3

Figure: Initialization evaluation[4]

What did the hallucination net learn?

roi-pool5 activations from corresponding regions in the test image which have highest final detection scores

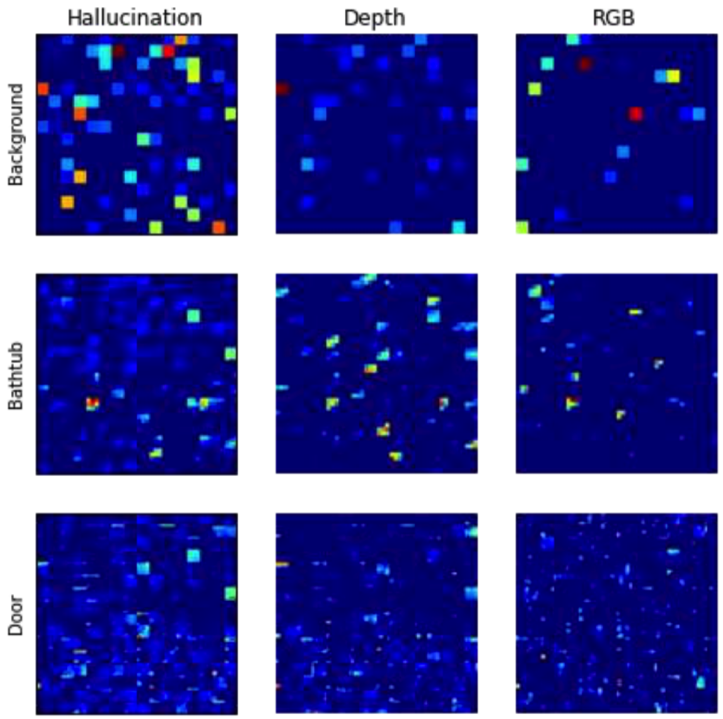


Figure: Activation of RoI Pool5[5]

What did the hallucination net learn?

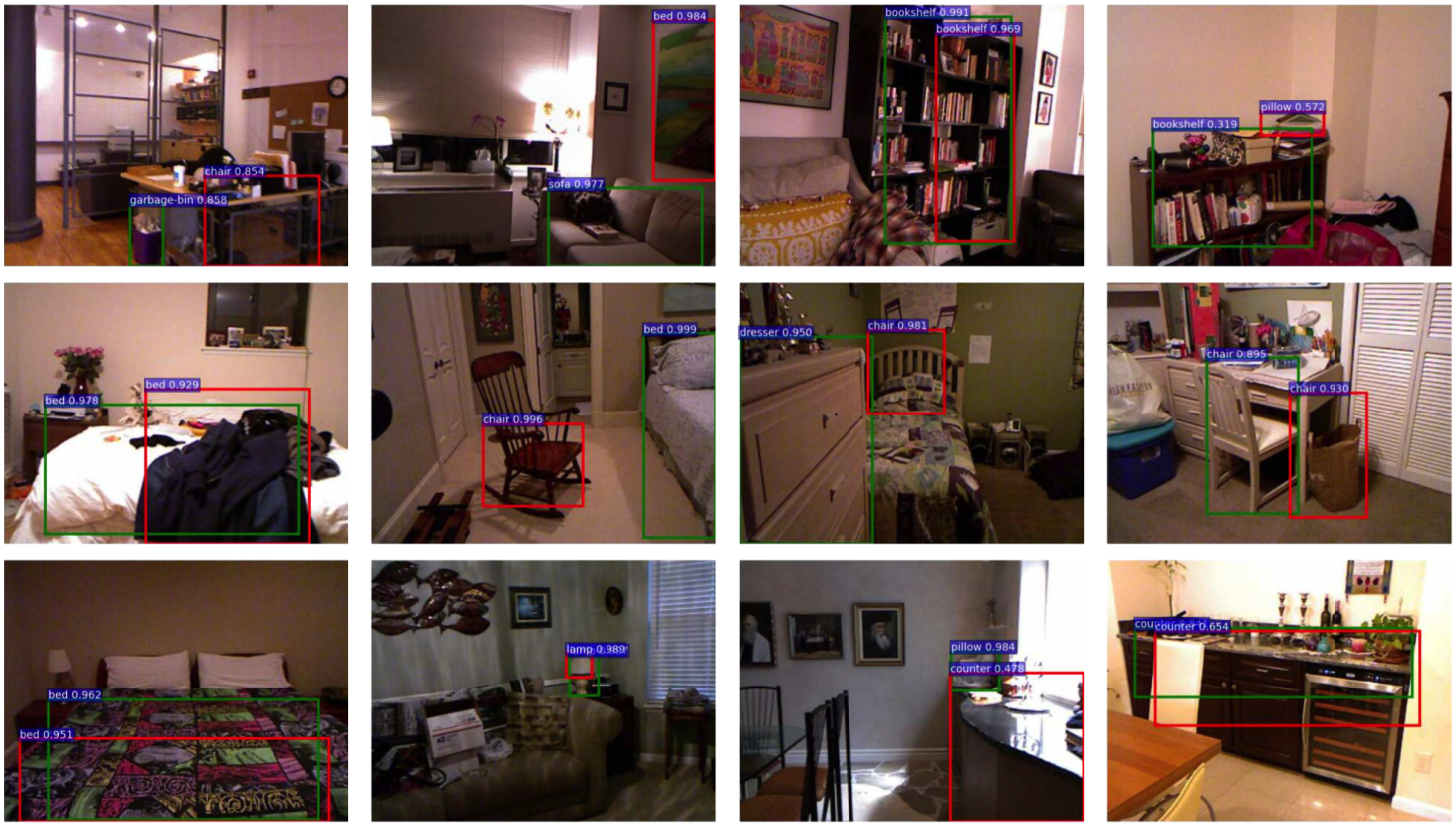
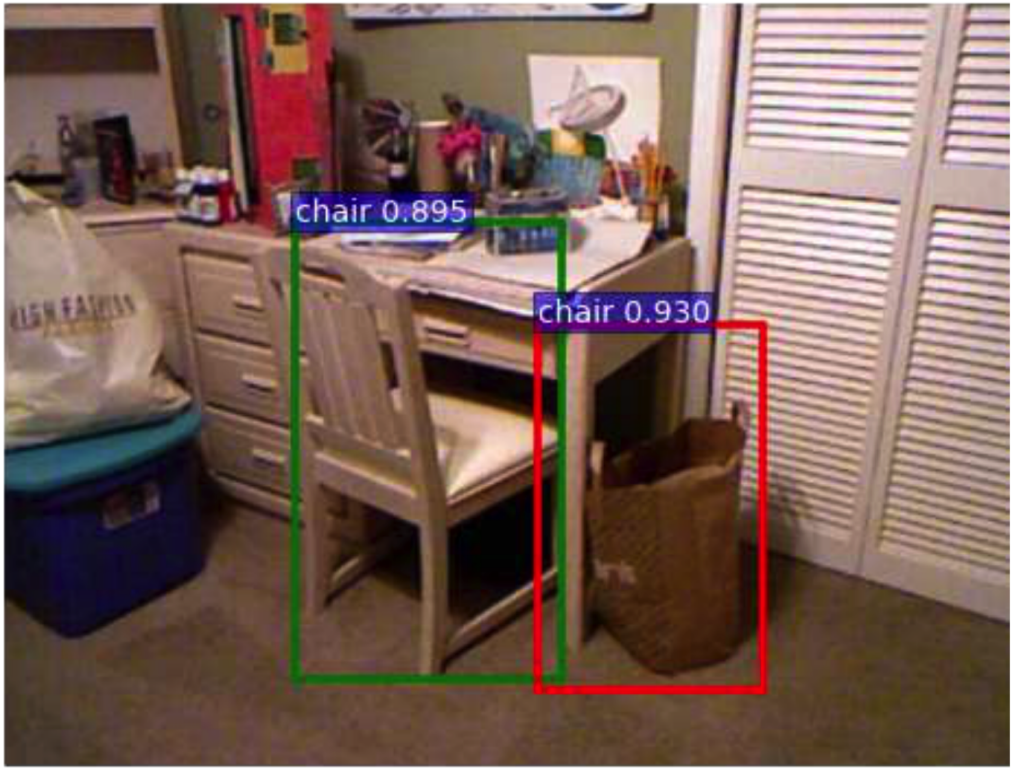


Figure 4: Example Detections on the NYUD2 *test set* where our RGB hallucination network's (green box) top scoring detection for the image is correct while the baseline RGB detector's (red box) top scoring detection is incorrect.

What did the hallucination net learn?

What did the hallucination net learn?



But it's not perfect.

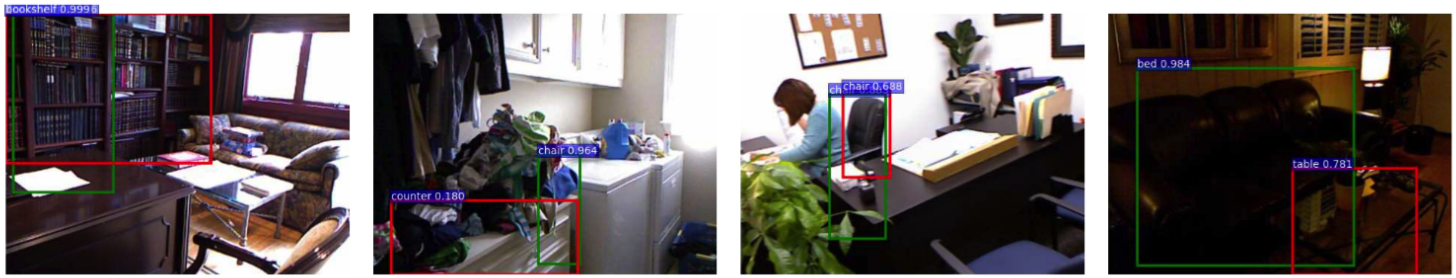


Figure 5: Example Detections on the NYUD2 *test set* where our RGB hallucination network's (green box) top scoring detection for the image is a false positive while the baseline RGB detector's (red box) top scoring detection is a true positive.

But it's not perfect.

Conclusion

- A novel technique for extracting **unseen modality** features.
- Outperforms the corresponding Fast R-CNN RGB detection models on the NYUD2 dataset.

Sources

- [1][2][3][4][5]: J. Hoffman, S. Gupta, T. Darrell, Learning with Side Information through Modality Hallucination
- [6]: R. Girshick, Fast R-CNN
- [7]: S. Gupta, R. Girshick, P. Arbelaez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation.
- [8]: L. Chen, W. Li, and D. Xu. Recognizing rgb images by learning from rgb-d data.
- [9]: L.Spinello and K.O.Arras, Leveraging rgb-d data: Adaptive fusion and domain adaptation for object detection.
- [10]: V. Vapnik and A. Vashist, A new learning paradigm: Learning using privileged information.
- [11]: G.Hinton,O.Vinyals, andJ.Dean, Distillingtheknowledge in a neural network.
- [12]: S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer.
- [13]: A. Shrivastava and A. Gupta. Building part-based object detectors via 3d geometry. In ICCV, 2013.

Thanks.