Robust Long Tail Object Detection

PhD Comprehensive Exam

Presenter
Sulabh Shrestha
sshres2@masonlive.gmu.edu

Committee Members
Prof. Jyh Ming Lien (Chair)
Prof. Jana Kosecka (Advisor)
Prof. Zoran Duric
Prof. Yutam Gingold

Object Detection

→ Object detection is the task of finding objects in an image

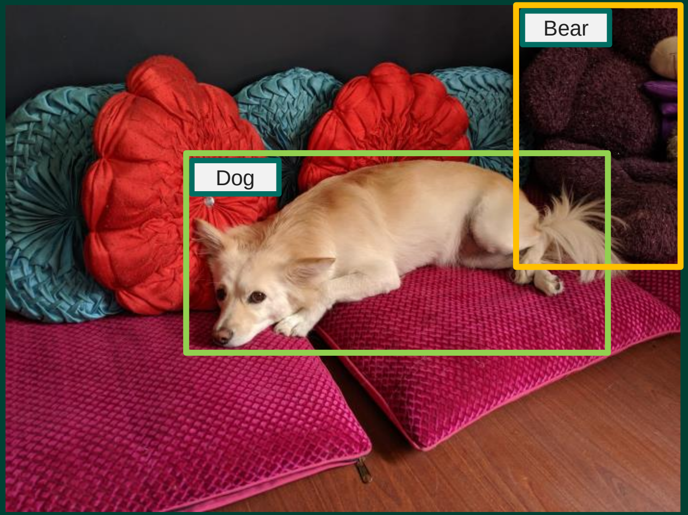
→ For each of the N_{obj} objects in the image

- ○ Bounding box $\{x_i, y_i, w_i, h_i\}$ $i \in N_{obj}$
- ○ Class label c_i $c_i \in [1, C]$

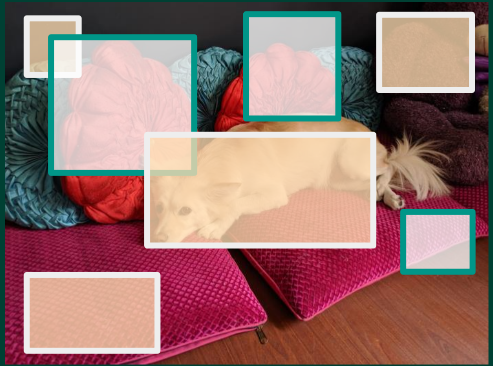
→ Need to search over a large search space

- ○ Possible location $(x_i, y_i) \in Image$
- ○ Possible scale and shape or aspect-ratio (w_i, h_i)
- ○ $N_{location} \times N_{area} \times N_{aspect-ratio}$

→ Classification over all possible options is costly



Object Detection Approaches



→ Efficiently generate a smaller set of hypotheses boxes likely to contain objects (object proposals)

→ Calculate features which can be used for classification

→ Prior to deep learning

- ○ Objects Proposals generation
  - ■ Bottom-up segmentation. E.g. Selective Search[6]
  - ■ Sliding window on edges. E.g. EdgeBoxes[1]
- ○ Classification
  - ■ Compute hand engineered features. E.g. SIFT[2]
  - ■ Use a classifier. E.g. SVM[3]

→ With the advent of deep learning

- ○ Features are learned using convolutional neural networks (CNN)
- ○ Both proposal generation and classification use the learned feature
  - ■ Extract features for each proposal
    - ● CNN variants: VGG[4], ResNet[5]
  - ■ Use a classifier on each feature
    - ● Fully Connected Layers

1. L. Zitnick and P. Dollar. Edge boxes: Locating object proposals from edges, ECCV 2014
2. D. G. Lowe. Object recognition from local scale-invariant features, ICCV 1999
3. C. Cortes, V. Vapnik. Support-vector networks, Mach Learn 1995
4. K. Simonyan, A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR 2015
5. K. He, X. Zhang, S. Ren and J. Sun. Deep Residual Learning for Image Recognition, CVPR 2016
6. J. Uijlings, T. Gevers, K. Sande, S. Arnold. Selective Search for Object Recognition. IJCV 104

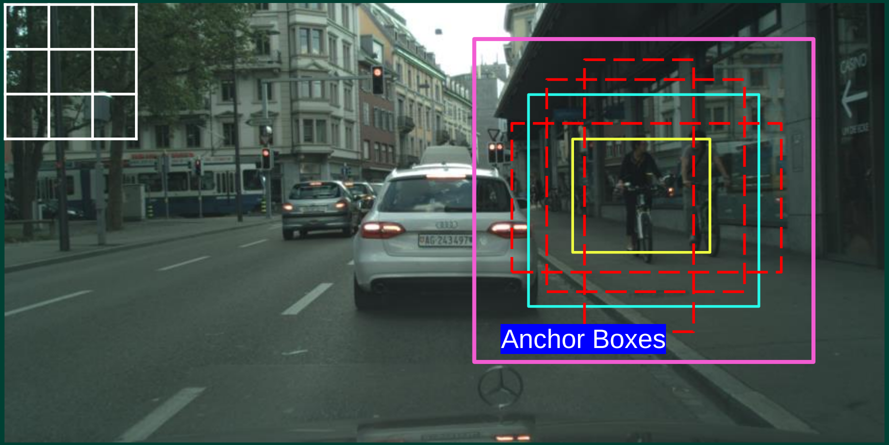
Feature Extraction

→ Backbone Networks

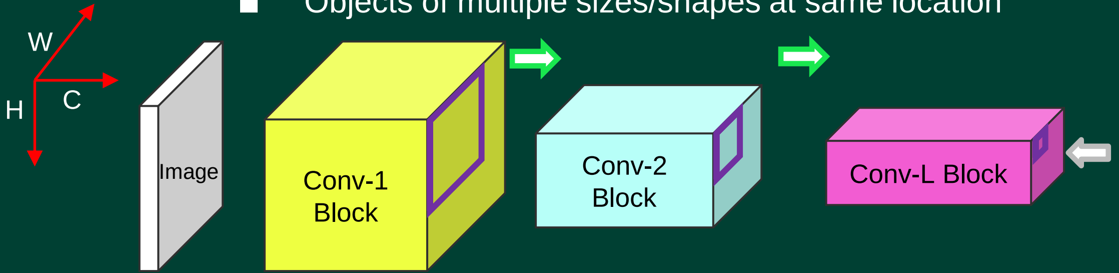
- Variants of CNN. E.g. VGG[2], ResNet[3]

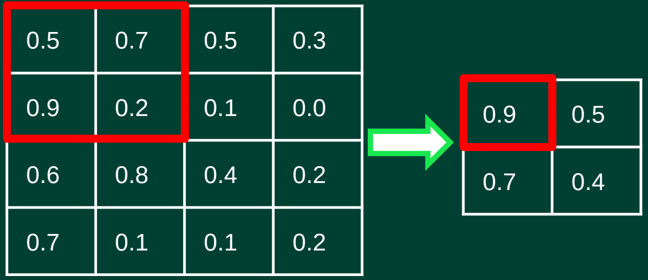
→ Networks models

- Blocks of CNN
- Pooling in between blocks of CNN. E.g. Max
- Deeper features have lower spatial resolution
- Deeper features have higher larger view (receptive field) of input image
  - Objects of multiple sizes/shapes at same location



CityScapes Dataset [1]





0.5	0.7	0.5	0.3
0.9	0.2	0.1	0.0
0.6	0.8	0.4	0.2
0.7	0.1	0.1	0.2

0.9	0.5
0.7	0.4

Max Pooling

1. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding, CVPR 2016
2. K. Simonyan, A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR 2015
3. K. He, X. Zhang, S. Ren and J. Sun. Deep Residual Learning for Image Recognition, CVPR 2016

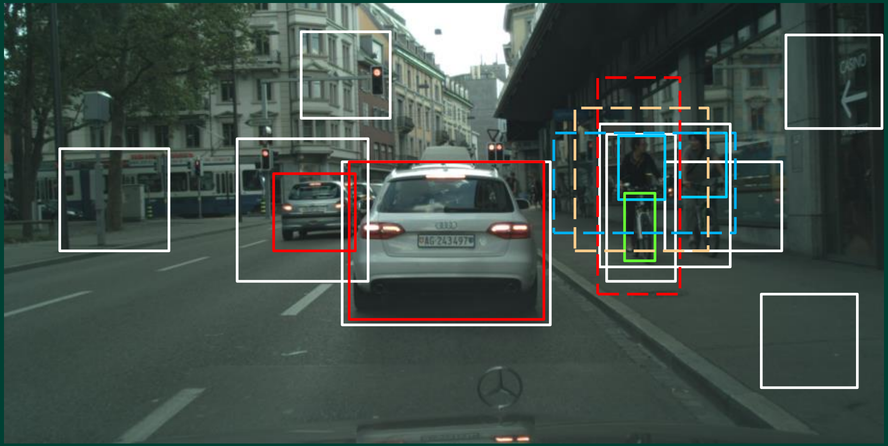
Two Stage Detectors

→ Region Proposal Network

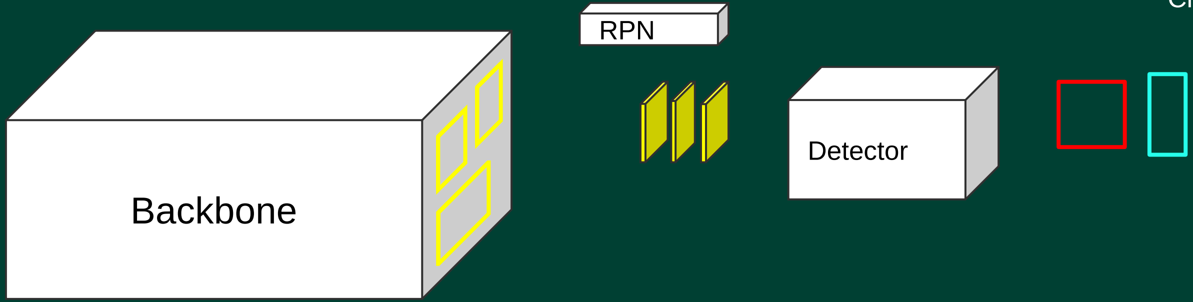
- Use feature extracted by Backbone
- Make multiple predictions at uniform grid
  - With reference to multiple **Anchor Boxes**
- Get object proposals

→ Detector Head

- Extract features for each proposal
- Classify and refine



CityScapes Dataset [1]



Eg: Faster-RCNN[2]

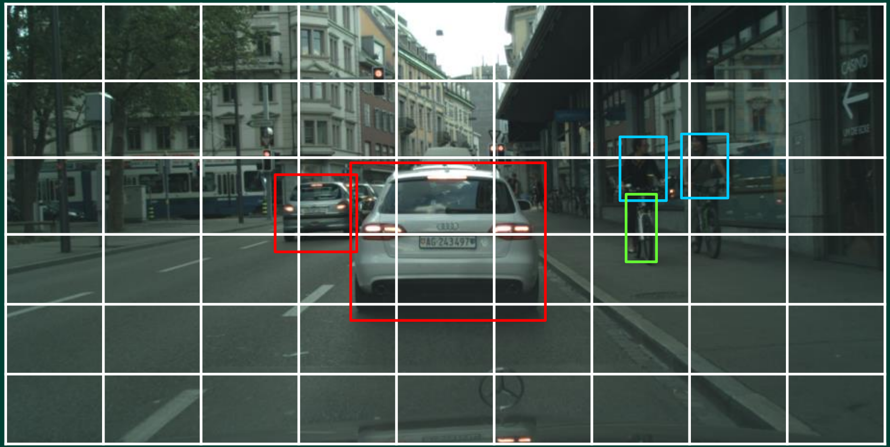
1. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding, CVPR 2016
2. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NIPS 2015

Single Stage Detectors

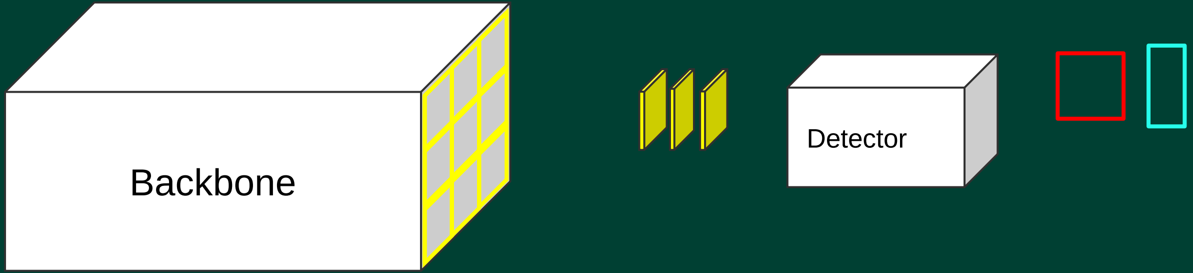
→ Backbone Network

→ Detector Head

- ○ Divide image into uniform grids
- ○ Use features corresponding to each grid
- ○ Make multiple predictions at each grid
  - ■ With reference to **Anchor Boxes**



CityScapes Dataset [1]

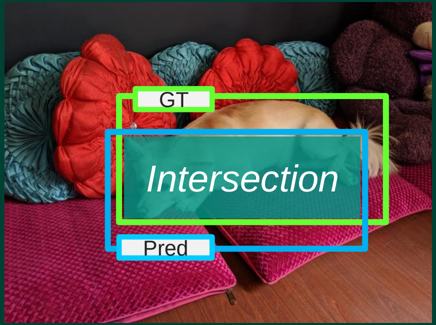


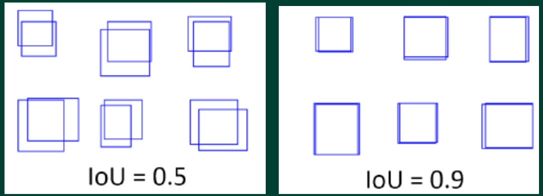
Eg: SSD[2], RetinaNet[3]

1. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding, CVPR 2016
2. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector, ECCV 2016
3. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal Loss for dense object detection, ICCV 2017

Evaluation of Object Detectors

- → Predictions are matched using IOU threshold
  - ○ Intersection over Union
    - ■ Between 2 boxes
    - ■ $IOU = \frac{area(Box_{GT} \cap Box_{Pred})}{area(Box_{GT} \cup Box_{Pred})}$
- → Average Precision
  - ○ Change threshold of confidence
  - ○ Area under the precision/recall curve
- → Mean Average Precision
  - ○ Average AP for all classes at IOU = K (mAP_K)
  - ○ Average AP for all classes at IOU = 0.5:0.05:0.95 (mAP_{coco})





EdgeBoxes [1]

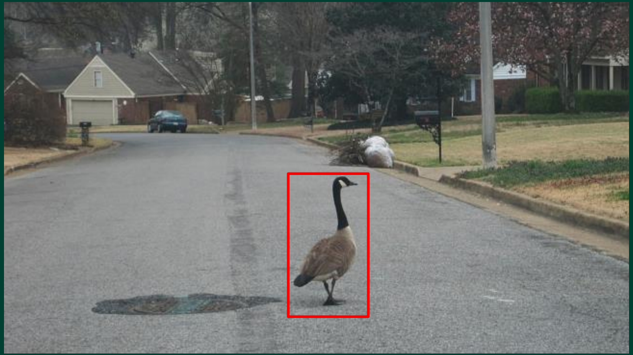
[1] L. Zitnick and P. Dollar. Edge boxes: Locating object proposals from edges, ECCV 2014

Analysis of Object Detectors

- → Strengths
  - Works very well on large datasets of images
    - COCO dataset has 80 classes
    - Collection of >200K Internet images
    - State of the Art methods get AP ~60
- → Improvement Space
  - Works in a closed-set assumption
  - Novel objects handling
    - More suitable in Two-Stage detectors as this step closely relates with getting candidate proposals
  - Does not work well when appearances change
    - Appearance features do not generalize to novel views

Model	**AP_{50}**	**AP**
Faster RCNN[1]	59.1	36.2
RetinaNet[2]	59.1	39.1

Object detector evaluation on COCO[3] dataset



Road Anomaly Dataset [3]

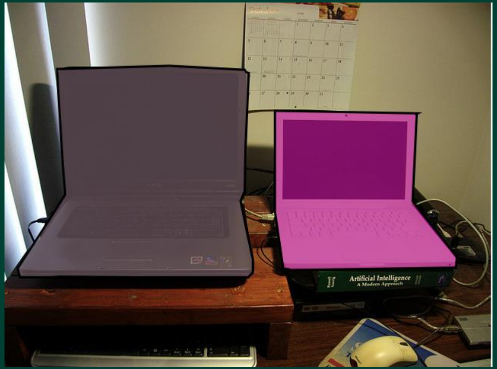
1. S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NIPS 2015
2. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal Loss for dense object detection, ICCV 2017
3. K. Lis, K. Nakka, P. Fua, and M. Salzmann. Detecting the unexpected via image resynthesis, ICCV 2019

But what signifies an OBJECT?

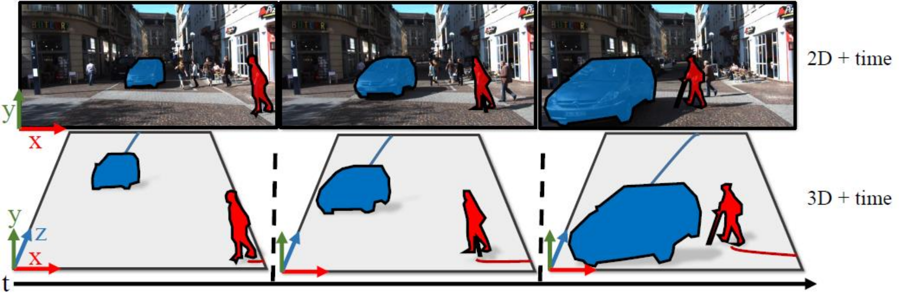
→ Close contours

→ Depth Discontinuity

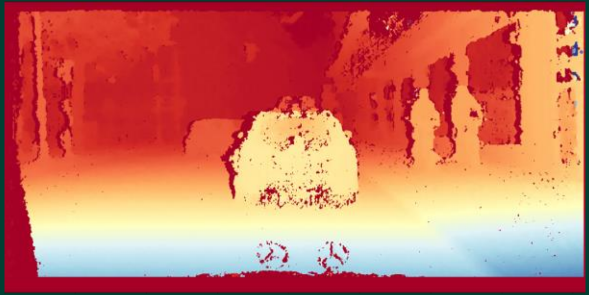
→ Coherence over time and space



Contours [2]



Spatio-Temporal Coherence [3]



Depth [1]

1. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding, CVPR 2016
2. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and L. Zitnick. Microsoft COCO: Common Objects in Context, ECCV 2014
3. A. Osep, P. Voigtlaender, M. Weber, J. Luiten, and B. Leibe. 4d generic video object proposals, ICRA 2020

Object Cues: Contours

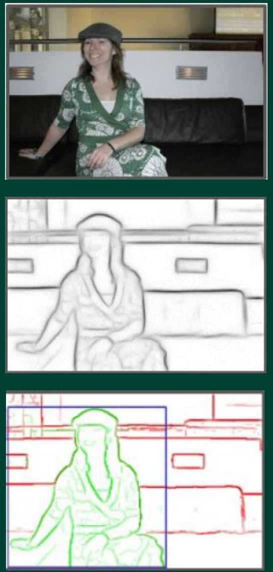
→ Objects have closed boundaries/contours

→ Edge Boxes[1]

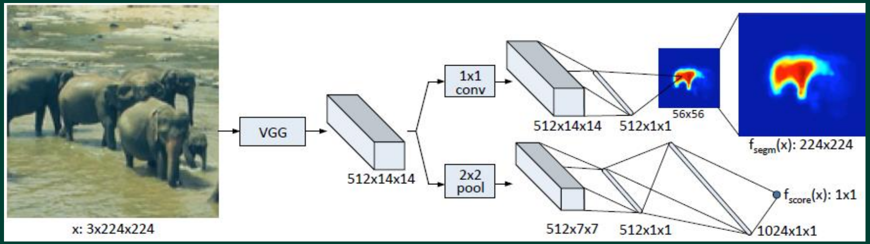
- Idea: Score box using number of wholly enclosed contours
- Extract edges and group them based on 8-connected and orientation
- Sliding window across image
- Find edge groups that stay within the window boundary
- Score based on sum of magnitudes of such edges

→ DeepMask[2]

- Extract features using CNN
- Slide window across image
- Classify as object and predict binary masks
- The contour cue enters as mask supervision



EdgeBoxes [1]



DeepMask [2]

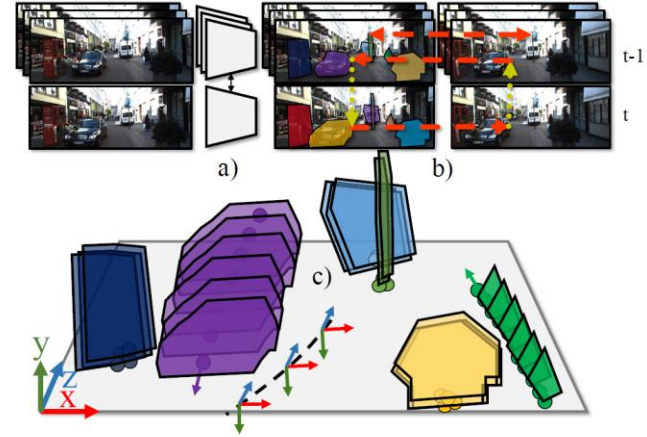
1. L. Zitnick and P. Dollar. Edge boxes: Locating object proposals from edges, ECCV 2014
2. P. O. Pinheiro, R. Collobert, and P. Dollar. Learning to segment object candidates, NIPS 2015

Object Cues: Spatio-Temporal Coherence

→ Objects persist over a sequence of frames

→ 4D Generic Video Object Proposals[1]

- Find objects over consecutive frames in a stereo video
- Re-purpose Mask RCNN[2]
  - Train Mask RCNN for object vs non-object classification
  - All classes combined into single object class
  - Better generalization to new objects
  - Mask supervision
- Predict object masks for each frame in a video sequence
- Track each object mask over space time
  - Compute 3D velocity and position
  - Find IOU between predicted mask and translated mask



Spatio-Temporal Coherence [1]

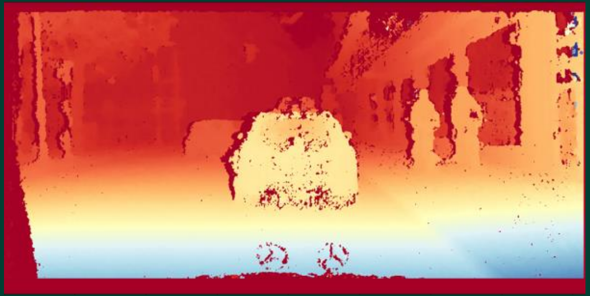
1. A. Osep, P. Voigtlaender, M. Weber, J. Luiten, and B. Leibe. 4d generic video object proposals, ICRA 2020
2. K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. ICCV 2017

Object Cues: Depth Discontinuity

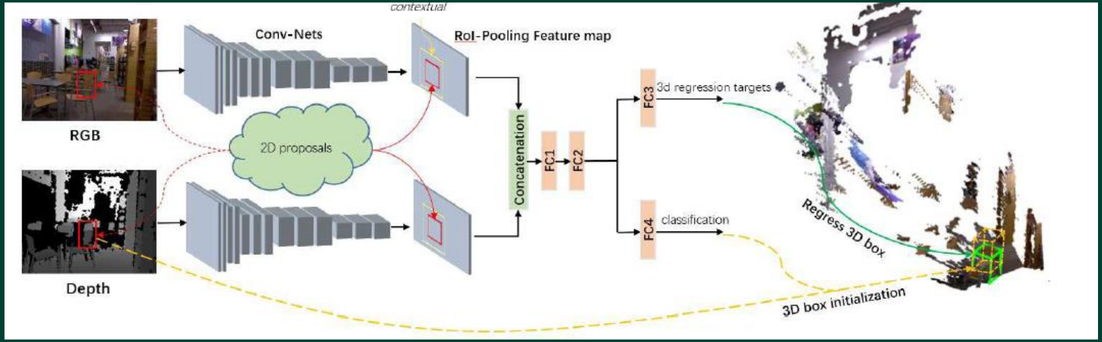
→ Objects have depth discontinuity

→ Amodal Detection of 3D Objects[2]

- Extract features from RGB and Depth Image in parallel
- Get 2D proposals
- Concatenate features for proposal from both modalities
- Classify and refine



Depth from CityScapes [1]

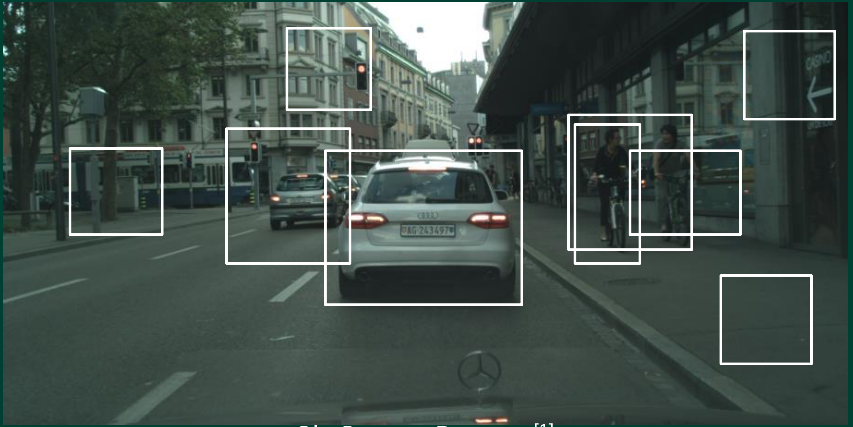


RGB and Depth Feature Extraction [2]

1. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding, CVPR 2016
2. Z. Deng and L. Jan Latecki. Amodal detection of 3d objects: Inferring 3d bounding boxes from 2d ones in rgb-depth images, CVPR 2017

Object Cues for Proposals

- → Multiple cues can be utilized to generate proposals for objects
- → Use uncertainty in classifier for getting novel objects
  - ○ Sigmoid confidence values not well suited for this
  - ○ Other uncertainty estimation techniques
    - ■ E.g. Dropout sampling[2]
- → Uncertainties arises from both novel objects as well as background regions
  - ○ How can these be handled properly?

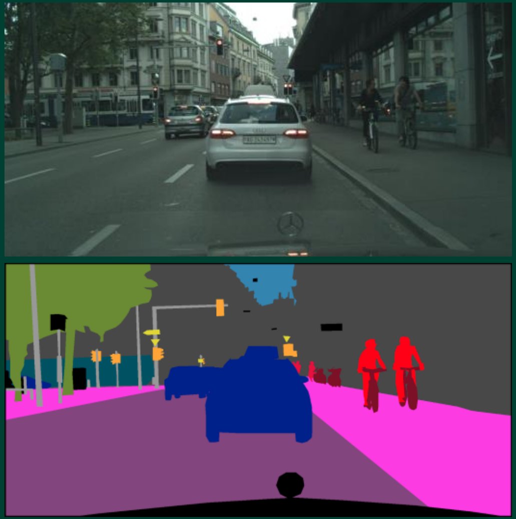


CityScapes Dataset [1]

1. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding, CVPR 2016
2. D. Miller, L. Nicholson, F. Dayoub, and N. Sunderhauf. Dropout sampling for robust object detection in open-set conditions, ICRA 2018

Semantic Segmentation

- Image consists of
  - Thing classes
    - Limited in extent in space and can be enclosed
    - Have distinct instances
    - E.g. Car, Bicycle, Person
  - Stuff classes
    - All other classes which cannot be enclosed
    - E.g. Road, Sky, Vegetation
- Object detection helps understand Thing classes
- Understanding Stuff classes can aid object detectors
  - Proposals that contain only stuff pixels can be filtered



CityScapes Dataset [1]

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding, CVPR 2016

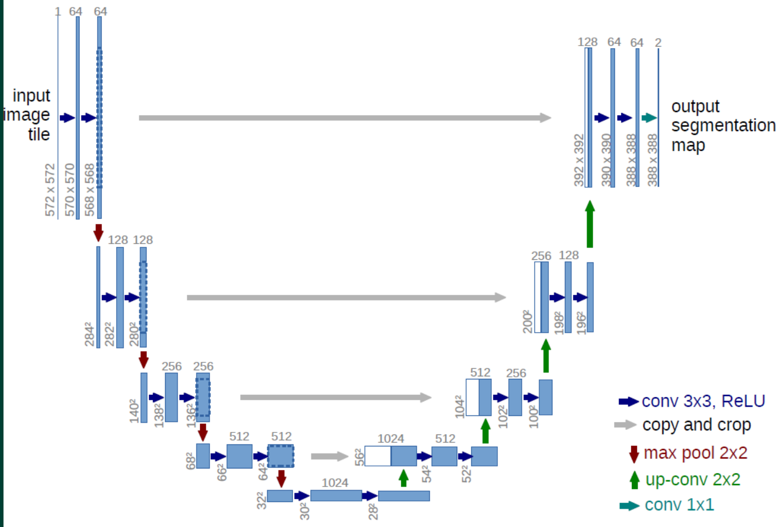
Semantic Segmentation

→ Pixel level classification

- Both stuff and thing classes

→ U-Net Model[1]

- Architecture
  - Repeated blocks of 2-layer 3x3 CNN
  - Max pooling between blocks for down-sampling in left half
  - Up-convolutions for up-sampling in right half
  - Final layer has 1x1 convolution to convert into channels = number of classes
- Lower-level features from left half are concatenated with higher level counterpart from right half
- Loss function
  - $\mathrm{E} = \sum_{x \in \Omega} t_l \log(p_l(x))$



U-Net Architecture [1]

[1] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, MICCAI 2015

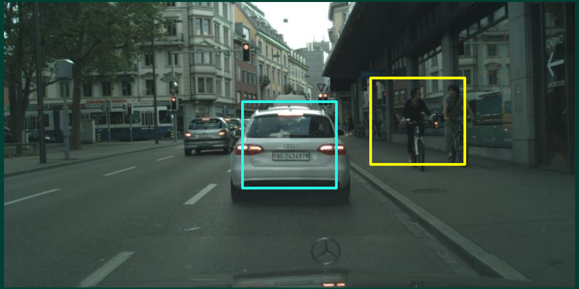
Semantic Segmentation

→ Input region size of features used for each pixel classification is fixed by the architecture

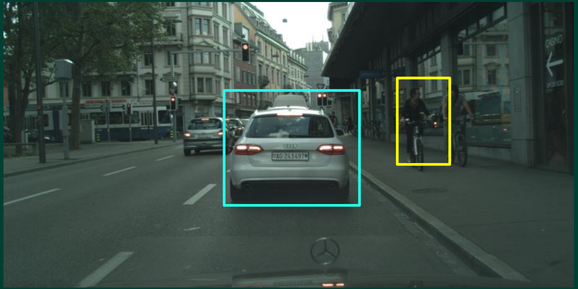
- Models cannot reason about the extent of all objects

→ Object detectors extract features of variable input region that corresponds to the whole object through region proposals

→ Both compliment each other



Semantic Segmentation Features[1]



Object Detection Features[1]

1. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding, CVPR 2016

More problems for Novel Object Detection

→ Lack of fixed evaluation criteria

- ROC Curve
- OpenSet error[1]
- mAP

→ Lack of fixed setup to evaluate on

- Lost and Found
- Road Anomaly
- Hand labelled objects in existing dataset

1. D. Miller, L. Nicholson, F. Dayoub, and N. Sunderhauf. Dropout sampling for robust object detection in open-set conditions, ICRA 2018

Application of Novel Object detection

→ Ability to handle novel objects without their labels

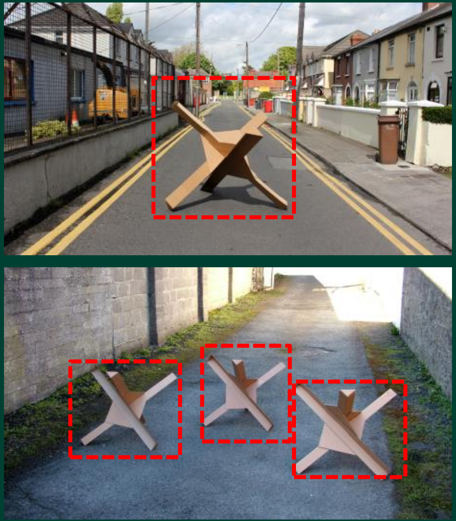
→ Autonomous Driving

- Flag unknown obstacles

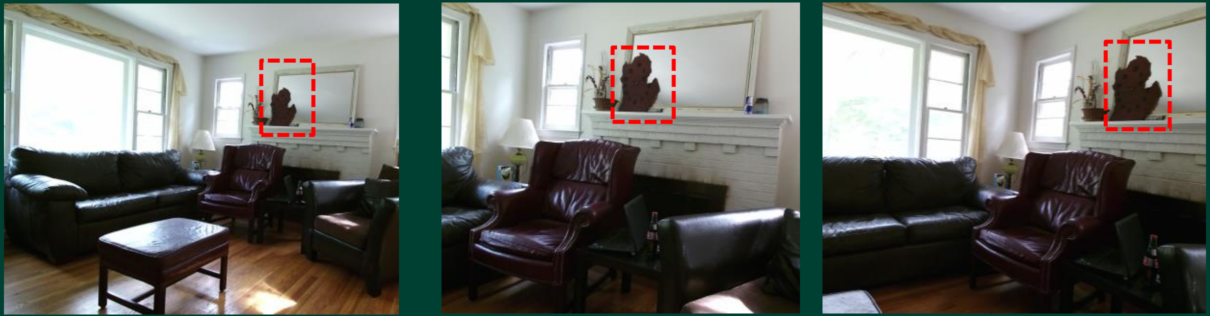
→ Indoor Robot Learning

- Object discovery using detection over time to gather examples for novel objects

→ Self-supervised learning



Road Anomaly Dataset [1]



Active Vision Dataset [2]

1. K. Lis, K. Nakka, P. Fua, and M. Salzmann. Detecting the unexpected via image resynthesis, ICCV 2019
2. P. Ammirato, P. Poirson, E. Park, J. Kosecka, and A. C. Berg. A dataset for developing and benchmarking active vision, ICRA 2017

Summary

→ Various object cues can be used for better proposal generation and thus for novel object detection

→ Semantic segmentation can augment object detectors

→ Need to design a method that better utilizes all cues

→ Need for better setup of novel object detection evaluation

→ Novel object detection applications

- ○ Better proposals and detection in general
- ○ Object discovery
- ○ Self-supervised learning