Object discover and self-supervised approach for robotics

Jungseok Hong, Ph.D candidate in CS

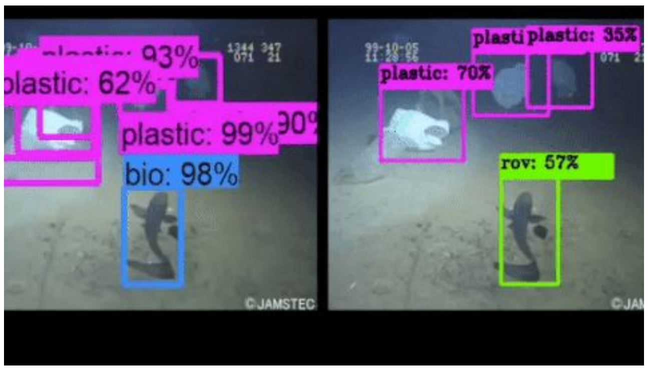
University of Minnesota
Driven to Discover℠

Topics to cover

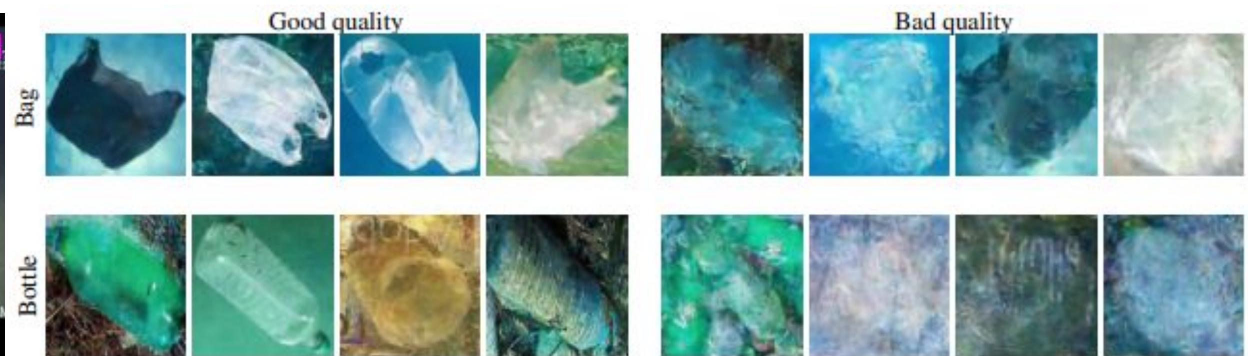
- Motivation
- Objectness
- Non-maximum suppression (NMS)
- NMS alternatives
- Self-supervised learning for robotics (RSS workshop-based)
- Related research
- Resources

Motivation

Trash Detection (ICRA 2019)



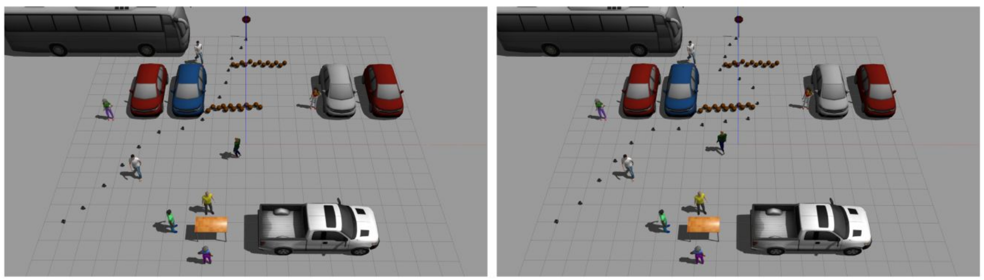
Generated Trash (ICRA 2020)



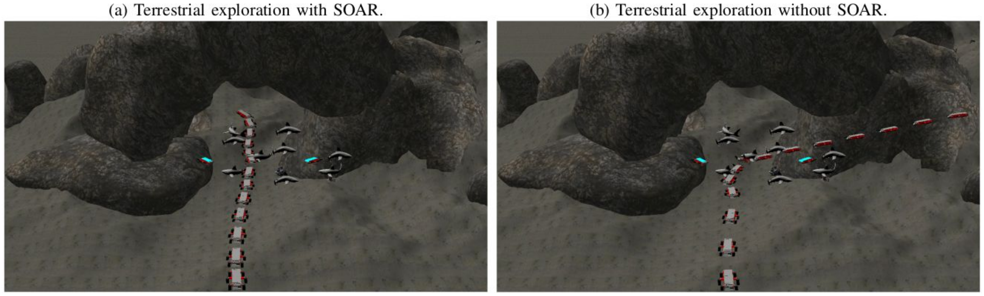
TrashCan 1.0 An Instance-Segmentation Labeled Dataset of Trash Observations (7,212 images)

Motivation

Semantically-aware obstacle avoidance (ICRA 2021)

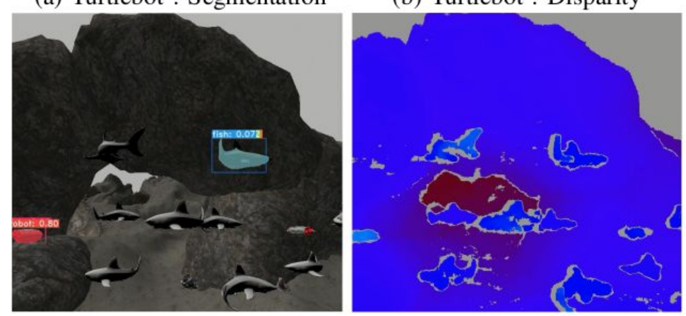


(a) Terrestrial exploration with SOAR. (b) Terrestrial exploration without SOAR.



(c) Underwater exploration with SOAR. (d) Underwater exploration without SOAR.

(a) Turtlebot : Segmentation (b) Turtlebot : Disparity



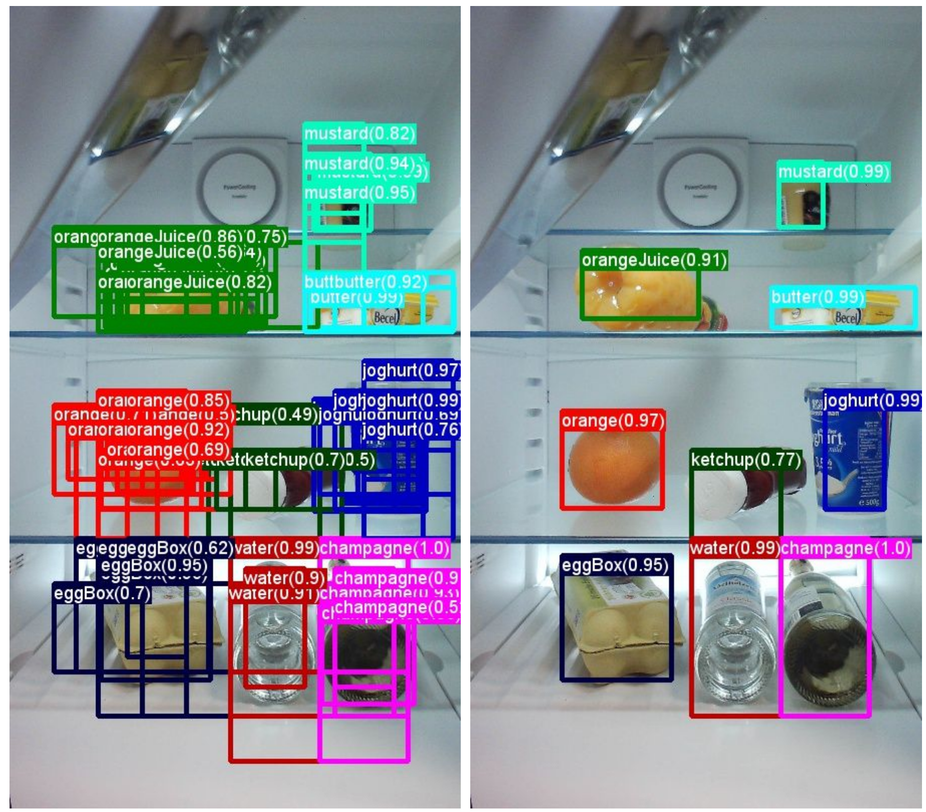
(c) Aqua : Segmentation (d) Aqua : Disparity

What is Objectness?

- One of the requirements of self-supervised learning for robotics
- A robot should be able to find “objects” first.
- Objectness: Finding image regions that contain object-like characteristics
- How to find objects?
  - State-of-the-art object detection methods use an object proposal algorithm (OPA) to generate general object proposals (GOPs)
  - Each GOP consists of two elements: a bounding box (b) and an objectness confidence score (o)
  - The GOPs are typically applied to a classifier, which then assigns them with an object class.

NMS

- NMS has been used as one of the key components of object detectors.
- NMS selects bounding boxes with the highest score and suppress ones that have a high overlap with each bounding box.
- The overlap measure is “Intersection-over-Union (IoU)” threshold to a predefined value.
    - Greedy NMS
    - soft-NMS
    - matrix-NMS



https://docs.microsoft.com/en-us/cognitive-toolkit/object-detection-using-fast-r-cnn

Issues of using NMS

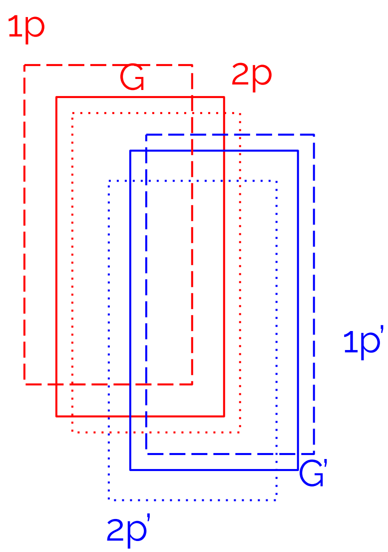
- Due to the nature of NMS, NMS only yields one bounding box if proposals are highly overlapped.
    - This is fine when objects are not occlude each other but it will be problematic for crowded scenes.
- Most parts in object detectors are end-to-end trainable, but NMS still remains as hand-crafted.
- a higher threshold (more FP) and a lower threshold (more missed detections)

Figure 1. Illustration of greedy-NMS results of different thresholds. The blue box shows the missing object, while the red ones highlight false positives. The bounding boxes in (b) are generated using Faster R-CNN. In a crowd scene, a lower NMS threshold may remove true positives (c) while a higher NMS threshold may increase false positives (d). The threshold for visualization is above 0.3.

How to improve NMS?

- Propose losses to produce tighter predictions.
  - Additional penalties are introduced to generate more compact bounding boxes.
- RepLoss (CVPR 2018)
  - Propose a bounding box regression loss designed for crowd scenes.
  - push each proposal to reach its designed target.
  - Keep each proposal away from other nearby objects.
- AggLoss (ECCV 2018)
  - Propose a loss term to enforce proposals locate compactly to the designated ground truth object.

How to improve NMS?

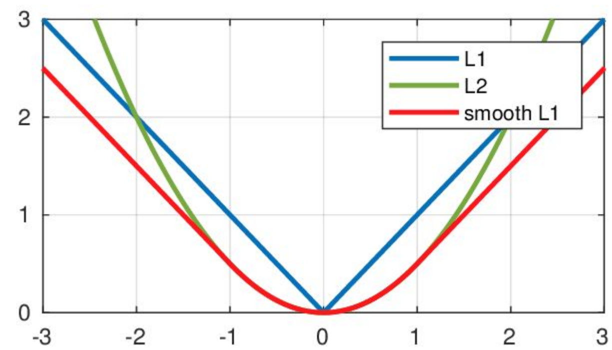


Repulsion Loss = Attraction Term - Repulsion Term

Dist(1p, G)+Dist(2p,G)+Dist(1p',G')+Dist(2p',G')
-Dist(1p, G')-Dist(2p, G')-Dist(1p',G)-Dist(2p',G)
-Dist(1p,1p')-Dist(1p,2p')-Dist(2p,1p')-Dist(2p, 2p')

Agg Loss = Dist(Avg(1p,2p), G) + Dist(Avg(1p',2p'), G')

Dist = IoU, Intersection over Ground-truth (IoG), Smooth L1 loss



How to improve NMS?

- NMS designed to handle occlusions.
- Adaptive NMS (CVPR 2019)
    - Propose dynamic suppression idea. (addressed the issue mentioned earlier)
    - The threshold
        - increases as instances gather and occlude each other
        - decreases when instances appear separately.
    - Predict the object density score (or crowdedness) online with a separate subnet and uses it as an adaptive threshold for NMS.
    - This adaptively adjusts up the threshold in crowded regions with a high crowdedness score.
    - Crowdedness estimation could be a problem.

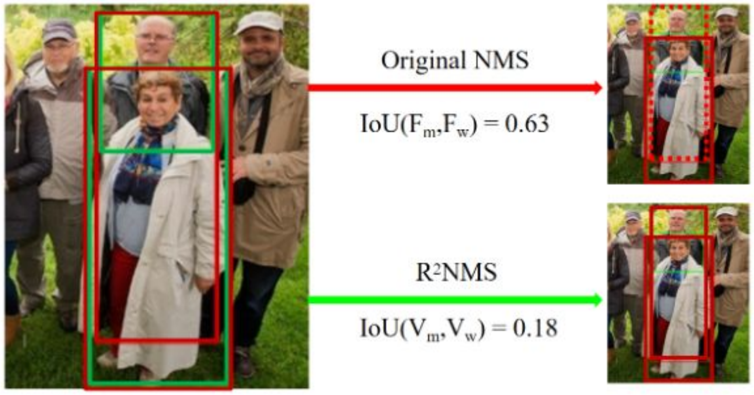
How to improve NMS?

- Double Anchor (? 2019)
  - Use prior knowledge: body and head are connected
- Useful for human detection.
- Usually head has a smaller scale, less overlap, and a better view in real-world images (compared to the body)
  - more robust to pose variations and crowd occlusions.
- The network predicts a head box and a body box with a confidence score.
- Then a joint NMS method uses a weighted score from both head bbox score and body bbox score, and boxes with a lower score will be suppressed if either the body overlap or the head overlap exceeds a certain threshold

How to improve NMS?

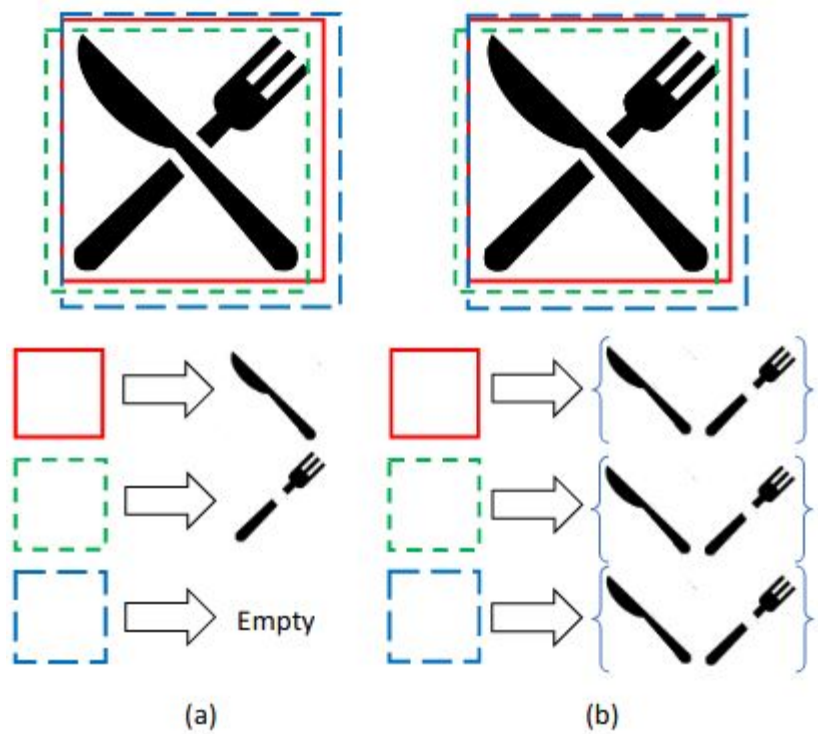
Paired RPN: generate a paris of proposals from the same anchor

- R2NMS
- Red: full body predictions
- Green: visible body predictions.
- Red solid represents the preserved bboxes while red dotted bbox indicates the reduced true positive bbox.



How to improve NMS?

- A single anchor + multiple prediction
- CrowdDet (CVPR 2020)
- Predict multiple detections per anchor for crowd detection.
- The predicted boxes from the same anchor are expected to infer the same set of instances (not distinguishing individual instances as in the single prediction paradigm in most object detectors).



How to improve NMS?

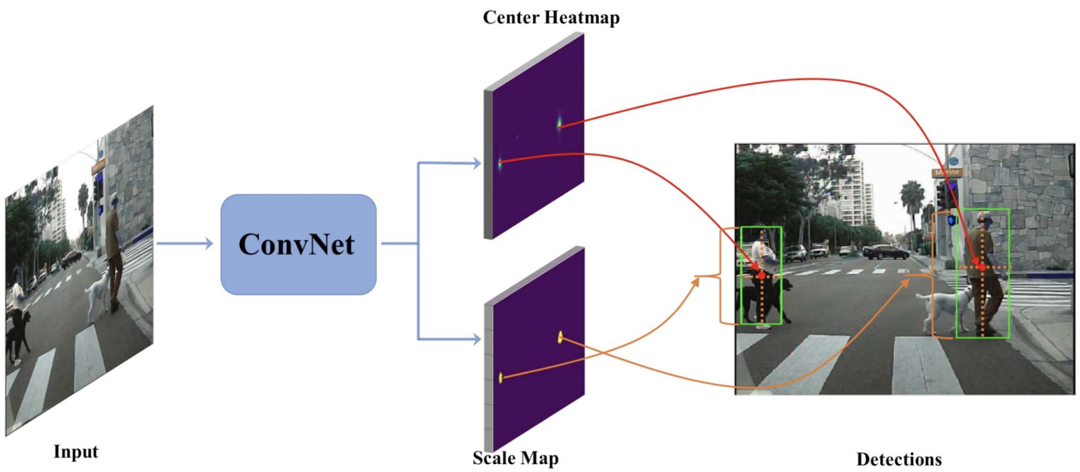
- A modified set NMS largely follows the normal NMS procedure but skips suppression for prediction coming from the same anchor.
- EMD (earth mover's distance) loss is used to select the best matching one with the smallest loss
- Add dummy boxes whose class label is regarded as background

NMS variations

Relation Net, PedHunter, affinity propagation clustering, Hashing-based NMS, Fast-NMS, Learning NMS, Seq-NMS, DeepParts, Fitness NMS, cluster NMS, GossipNet NMS,... etc

NMS alternatives

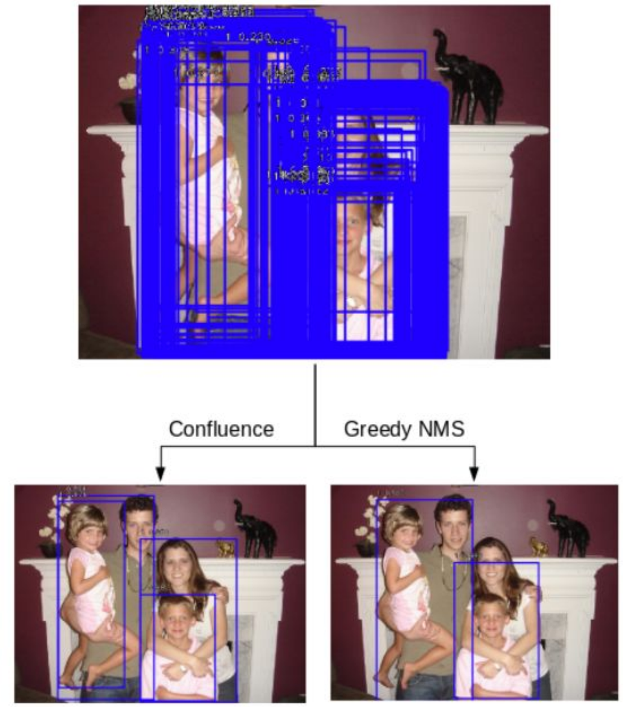
- NMS-free, anchor-free
- CenterNet (? 2019, PR-241: objects as points)
    - Object as a single point
- CSP (CVPR 2019)

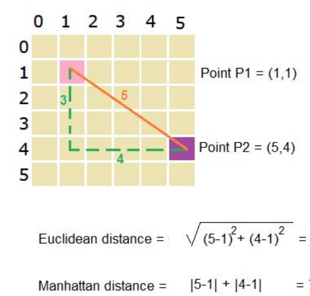


NMS alternatives

Confluence (? 2020)

- Works well for crowded images
- Similar to Greedy NMS but use different score/metrics
- Sort candidate boxes by confluence score (based on Manhattan distance)
- Remove duplicated boxes by using normalized Manhattan distance (Greedy NMS use IoU)





https://prismoskills.appspot.com/lessons/2D_and_3D_Puzzles/Chapter_05_-_Distance_between_points.jsp

Self-supervised learning for Robotics (RSS workshop)

What is the problem with current approaches?

- ImageNet : 1M labels for 5 years, Facebook generates >600M images per day
- Simulation is 1 task, tons of interactions, but in reality babies do 1000s of tasks in parallel with less structure

Self-supervised learning for Robotics (RSS workshop)

- Self-supervised learning: Supervised learning without labelling the data - Learn embeddings, automatic labelling.
- (+) large data collection is feasible, in real world it leads to better experimental design and engineering.
- (-) structure of the problem needs to be known and consistent, labelling mechanism needed.
- The front end prettiness of robotics vs hidden behind the scenes challenges in robotics, and self-supervision may mitigate this.

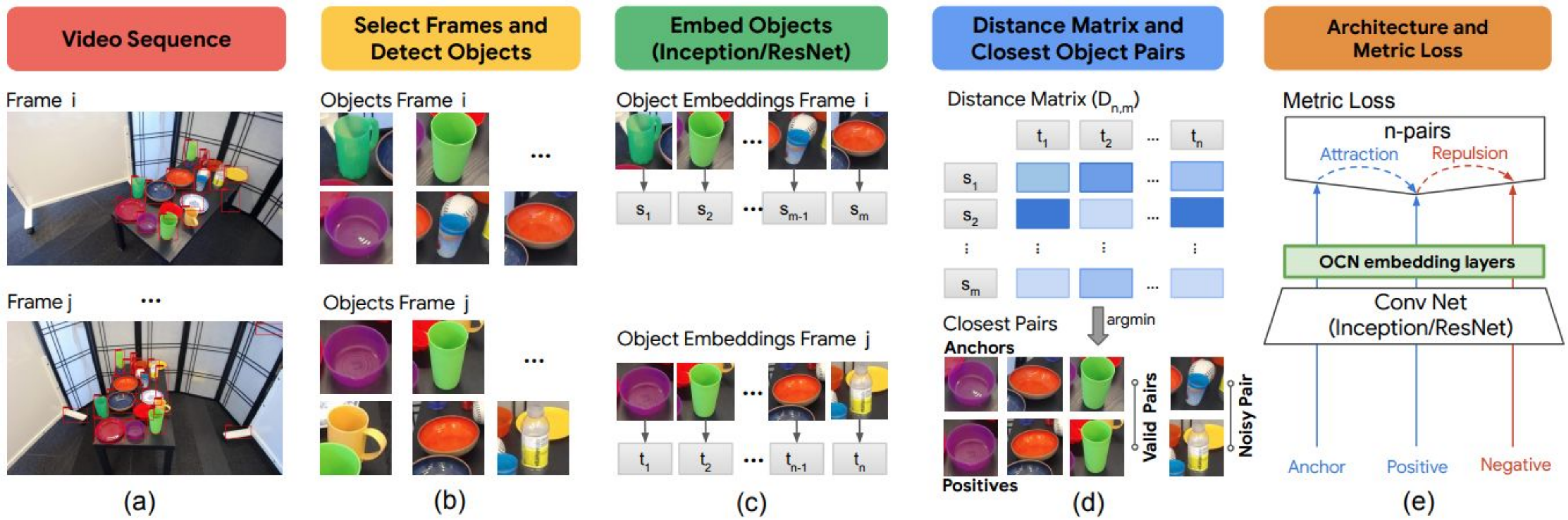
Related research

Online Learning of Object Representations by Appearance Space Feature Alignment (ICRA 2020)

- Robots can automatically collect data once deployed
- Robots can achieve multiple views of the same objects
- Supervised models can't detect new objects
- Faster R-CNN for finding objects (objectness)
- ResNet-50 to extract features from images
- Use N-pair loss
- Inner product of (anchor, positive)-pair to be larger than all (anchor, negative)-pairs.

$$\mathcal{L}_{N-pair}(\{(x_i, x_i^+)\}_{i=1}^N; f) = \frac{1}{N}\sum_{i=1}^{N} log\Big(1 + \sum_{j \neq i} exp(f_i^\intercal f_j^+ - f_i^\intercal f_i^+)\Big)$$

Related research



Related research

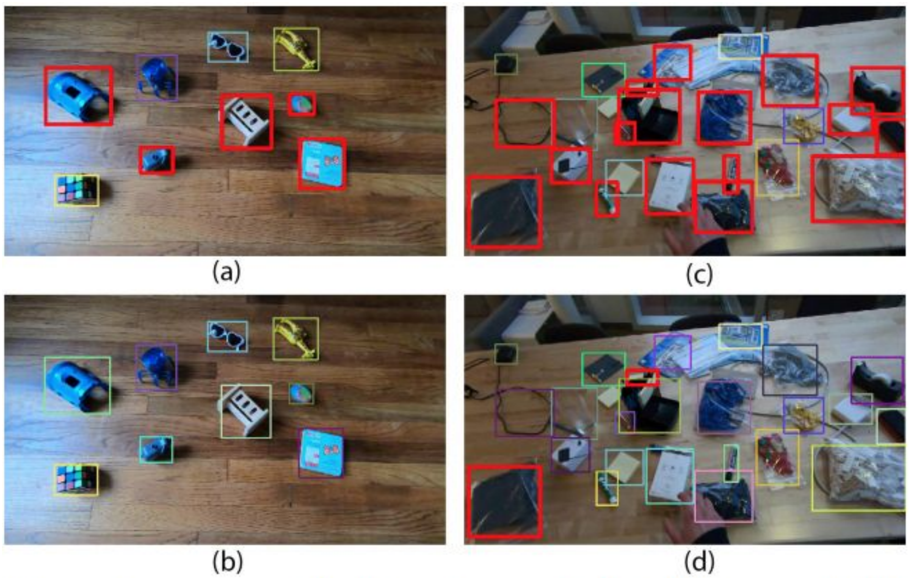
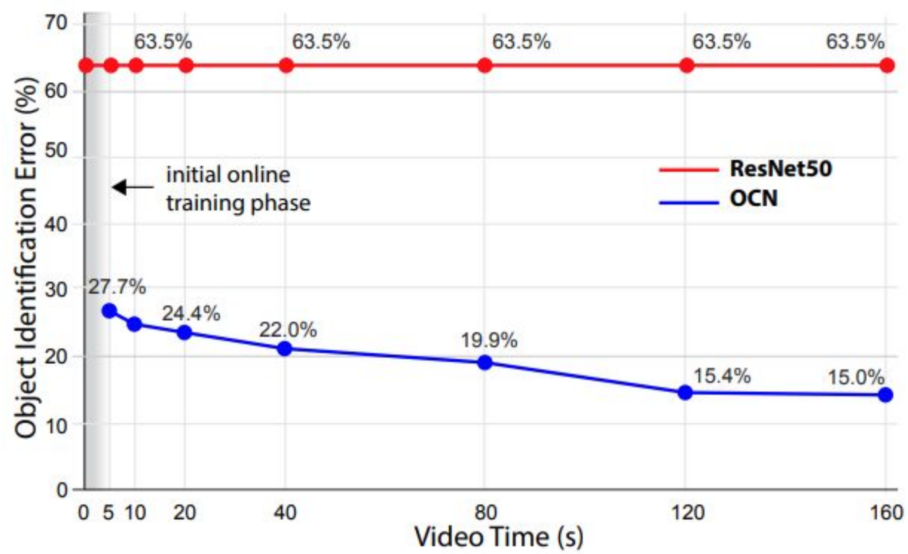


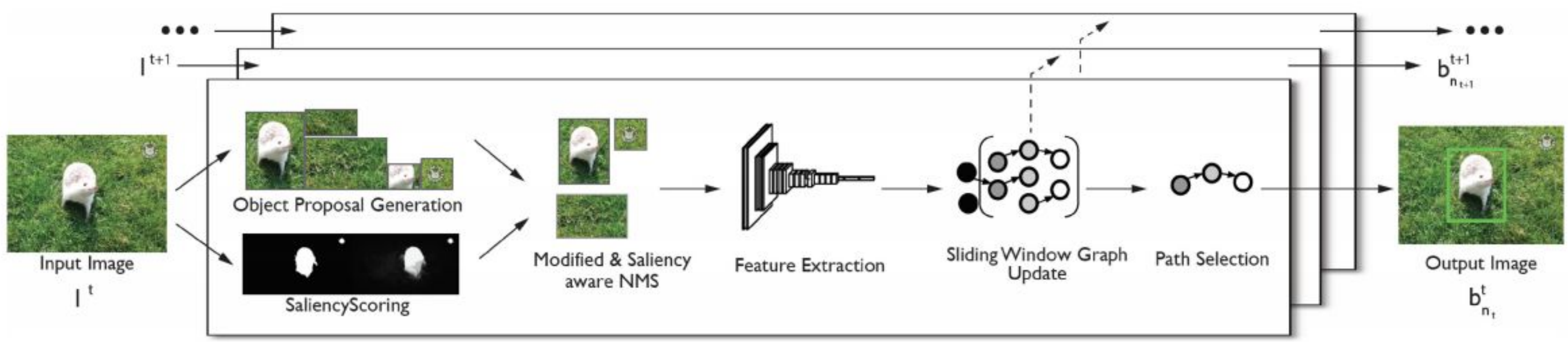
Fig. 6. Comparison of identifying objects with ResNet50 (a, c) and OCN (b, d) embeddings for the environments kids room and challenging. Red bounding boxes indicate a mismatch of ground truth and associated index



Related research

Unseen Salient Object Discovery for Monocular Robot Vision (RAL 2020)

- Unsupervised Foraging of Objects (UFO), a novel, unsupervised, salient object discovery method designed for monocular robot vision.
- Use a spatiotemporal stream of RGB images
- Object proposal (DeepMask with N=100)
- Saliency Scoring (Minimum Barrier Distance (MBD) Transform)



Related research

- Add all overlapping neighbors
- Feature extraction
    - CNN architectures (AlexNet, VGG19, ResNet, and InceptionV3) tested but used VGG-19 for its simplicity
    - Extracted from the last FC layer

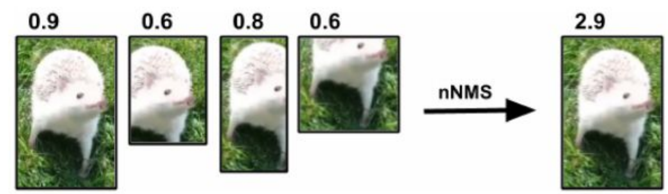


Fig. 2. In modified non-maximum suppression (mNMS), the strongest bounding box is assigned with the cumulative sum of the scores of all overlapping neighbors.

Thoughts

- Self-supervised models still rely on pretrained models (supervised) for various stages.
- Best way to extract features?
- Most self-supervised learning algorithms for robotics are focused on manipulation tasks due to the “interaction” components.

References

https://www.brainlinks-braintools.uni-freiburg.de/rss20-ssrl/

https://towardsdatascience.com/deep-learning-based-object-detection-in-crowded-scenes-1c9fddbd7bc4

Thank you