CVPR 2018 Review

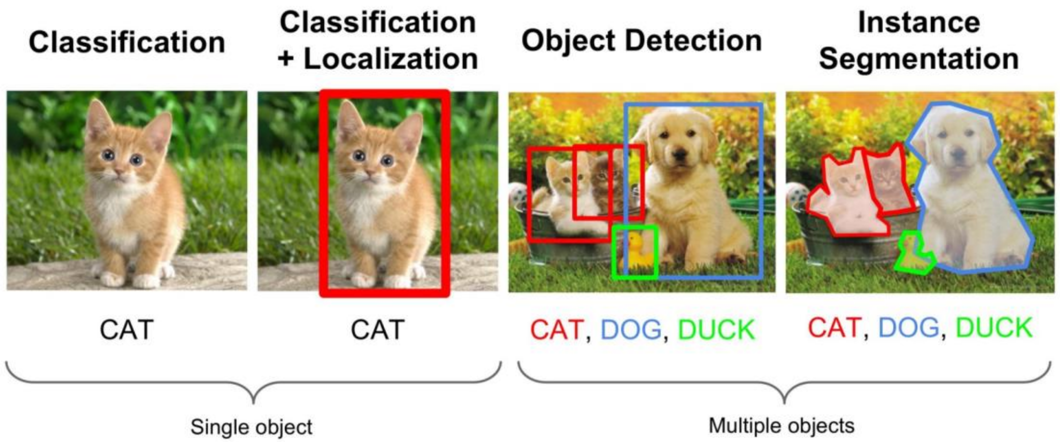
Object Classification/Detection and Issues on Unlabeled or noisy data

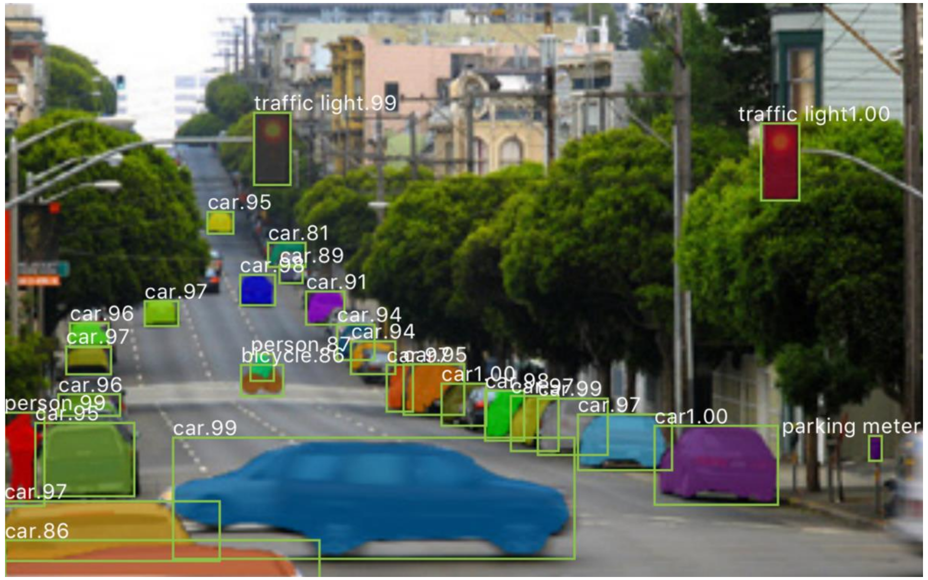
Naver Clova

ML / Dongyoon Han

ML / Sangdoo Yun

Object classification/detection (+segmentation)





Research trend

- Recent classification network architectures (as backbones):
  - ResNet – CVPR 2016
  - ResNeXT, DenseNet, PyramidNet, PolyNet, Xception – CVPR 2017

- Recent classical detection papers (for performance improvement):
  - Faster-RCNN 2015, SSD –ECCV 2016, R-FCN – NIPS 2016,
  - Mask RCNN – ICCV 2017, Deformable Convolution Networks, Feature Pyramid Networks – CVPR 2017,
  - Focal Loss for Dense Object Detection – ICCV 2017.

Research trend

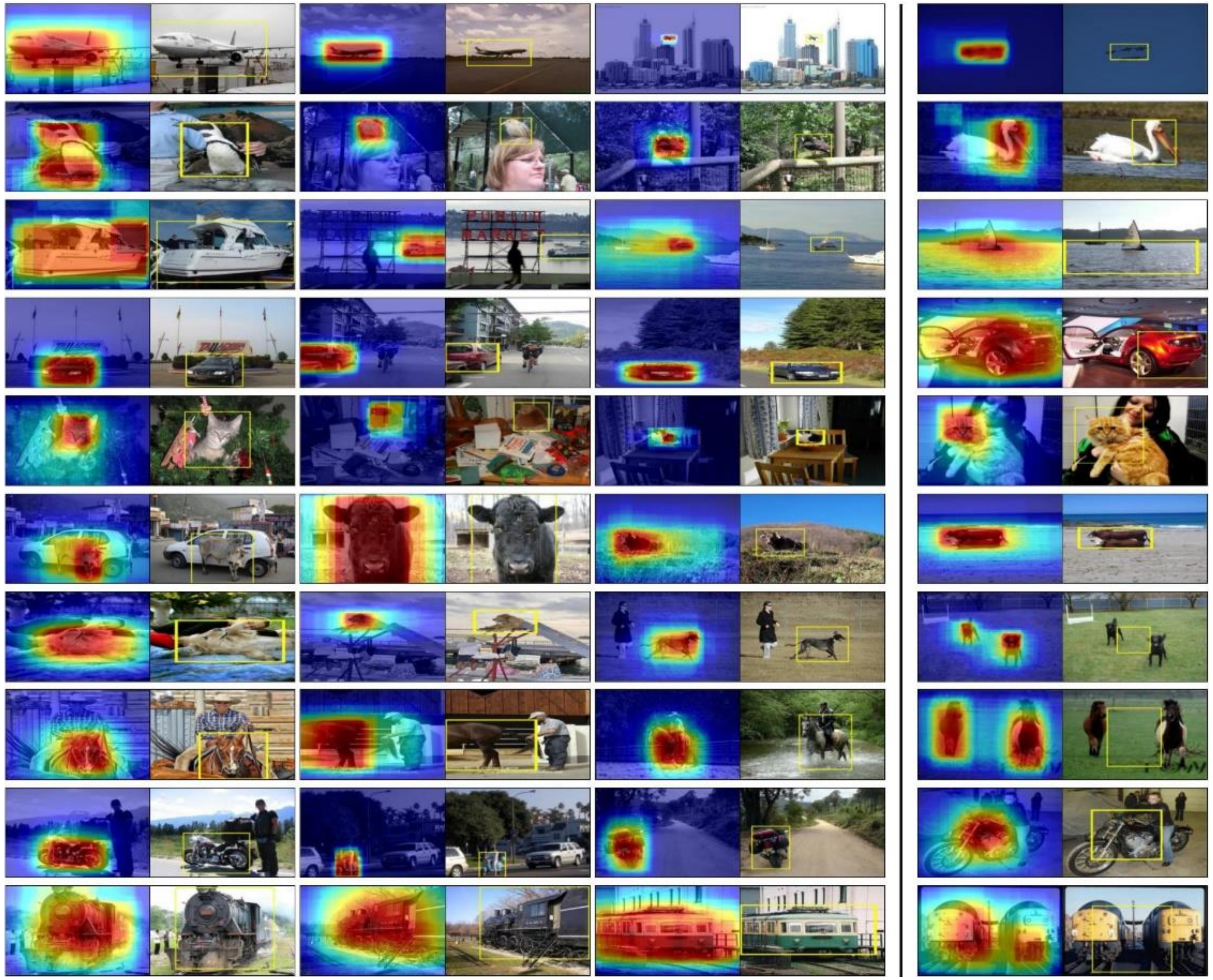
- Few novel classification network architectures (but very incremental)
  - e.g.) Squeeze-and-Excitation Networks – CVPR 2018,
  - E.g.) Deep Layer Aggregation – CVPR 2018.

- Classical detection papers are rare:
  - e.g.) Cascade R-CNN – CVPR 2018.

- Many sub-tasks under the classical object detection task:
  - Weakly-supervised object detection,
  - Object detection for multi-task learning,
  - Others.

Interesting papers

Weakly-supervised object detection



Weakly-supervised object detection

- Weakly ~~ (total 39 works published at CVPR2018)
- This tasks does not use bounding box annotations for training.
- Interesting paper:
  - Zigzag Learning for Weakly Supervised Object Detection –Zhang et al.
    - A proposed measure + curriculum learning-based method:

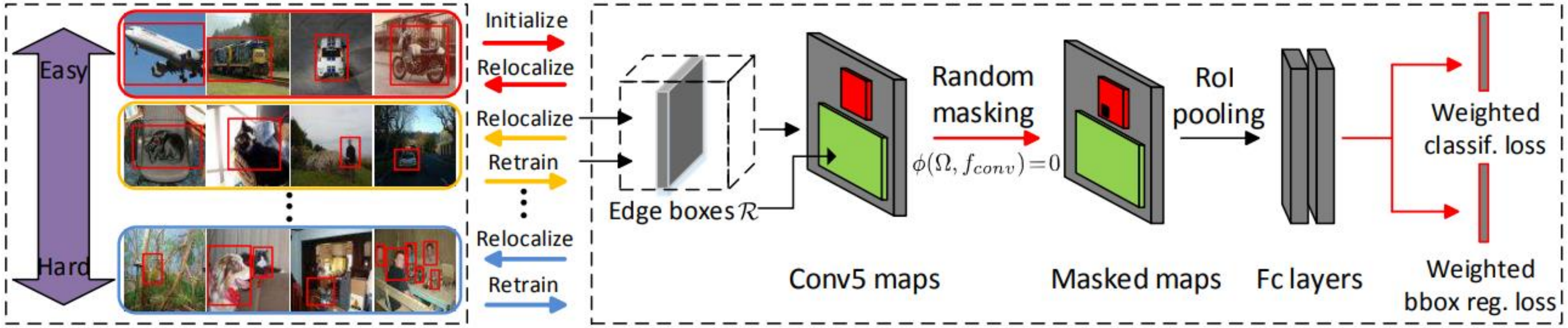
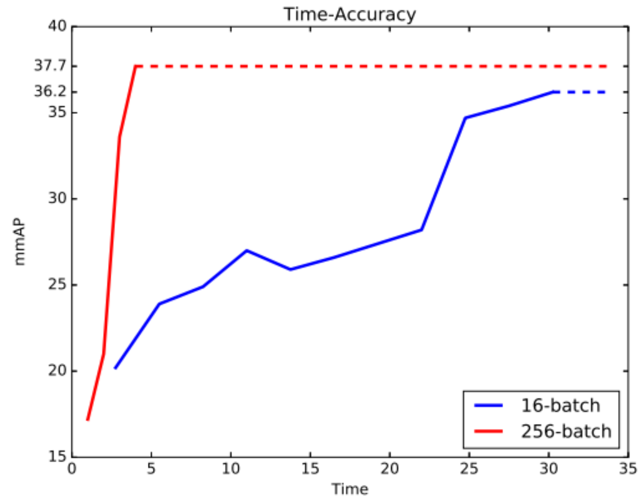
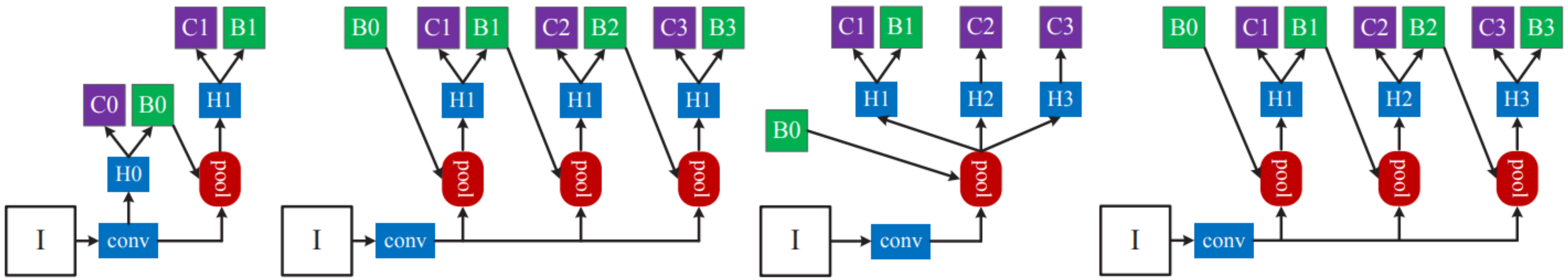


Figure 2. Architecture of our proposed zigzag detection network. We first estimate the image difficulty with mean Accumulated Energy Scores (mEAS), organizing training images in an easy-to-difficult order. Then we introduce a masking strategy over the last convolutional feature maps of fast RCNN framework, which enhances the generalization ability of the model.

A Powerful Object Detector

- Datasets: Pascal VOC or COCO datasets.
- Interesting papers:
  - MegDet: A Large Mini-Batch Object detector – Peng et al
    - Large batch training with Faster-RCNN + Cross-GPU BatchNorm
  - Cascade R-CNN: Delving Into High Quality Object Detection
    - Successively performing bounding box regression and classification





(a) Faster R-CNN (b) Iterative BBox at inference (c) Integral Loss (d) Cascade R-CNN

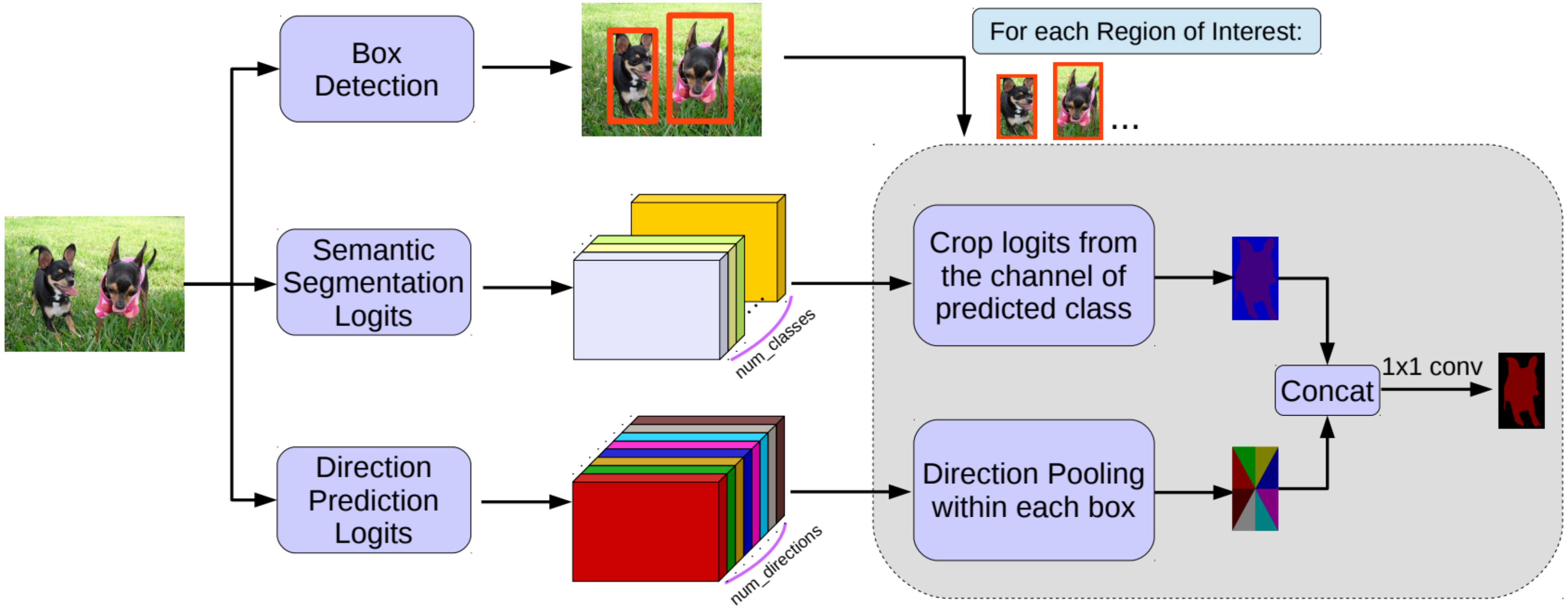
Figure 3. The architectures of different frameworks. "I" is input image, "conv" backbone convolutions, "pool" region-wise feature extraction, "H" network head, "B" bounding box, and "C" classification. "B0" is proposals in all architectures.

Object detection for multi-task learning

- Interesting papers:
  - MaskLab: Instance Segmentation by Refining Object Detection With Semantic and Direction Features – Chen et al (Google Inc)
    - **Box detection** + semantic sementation + direction prediction.
  - Detecting and Recognizing Human-Object Interactions - Gkioxari et al (FAIR).
    - Joint learning to **detect people** and objects concerning the task (human, verb, object) triplets, and fusing it (InteractNet).
  - Detect-and-Track: Efficient Pose Estimation in Videos – Girdhar et al (CMU).
    - Human detection by **3d-Mask-RCNN** and link the detections for video understanding.

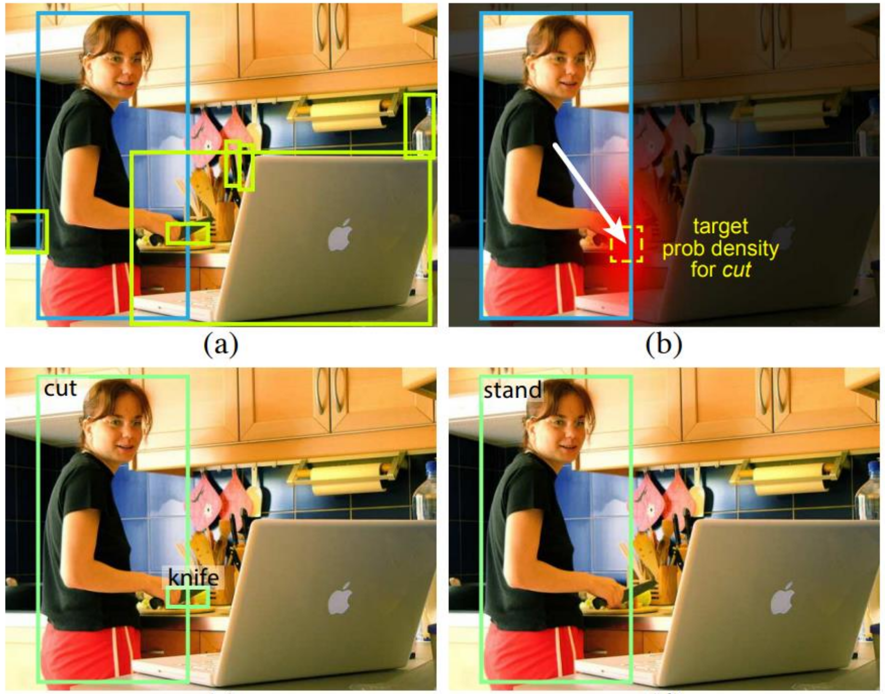
Object detection for multi-task learning

- Interesting papers:
  - MaskLab: Instance Segmentation by Refining Object Detection With Semantic and Direction Features – Chen et al (Google Inc)
    - **Box detection** + semantic sementation + direction prediction.



Object detection for multi-task learning

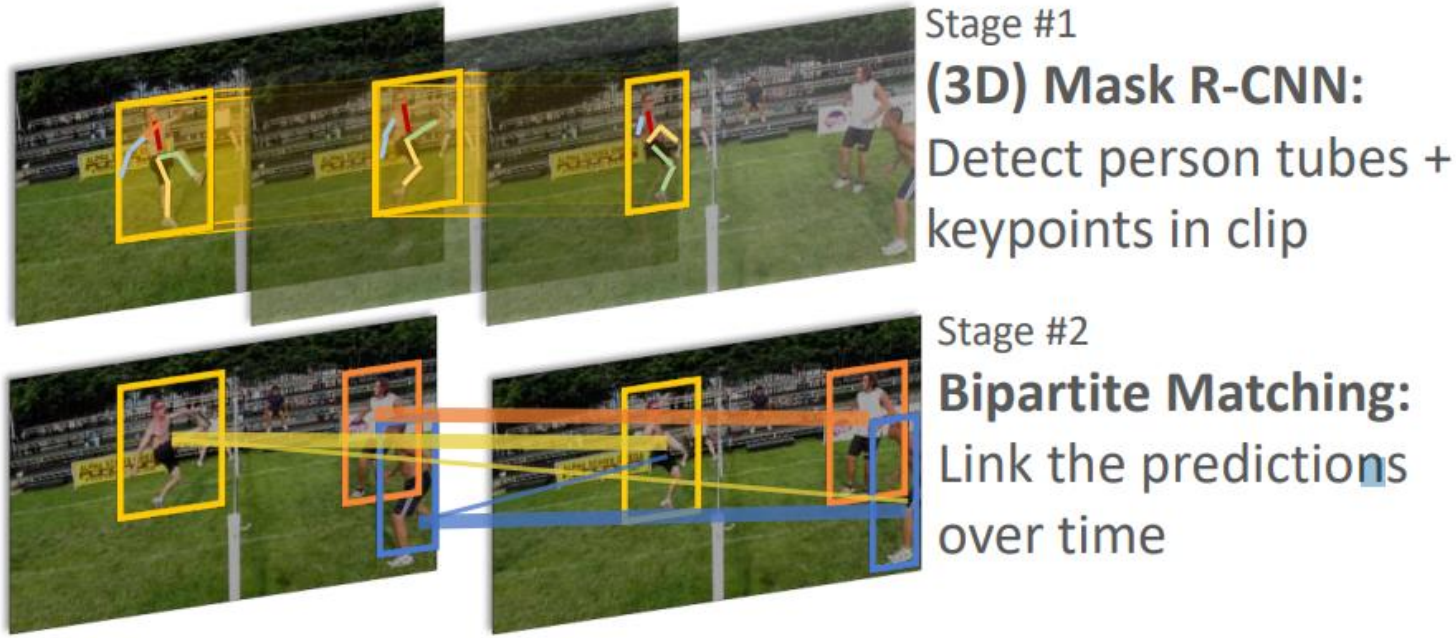
- Interesting papers:
  - Detecting and Recognizing Human-Object Interactions - Gkioxari et al (FAIR).
    - Joint learning to **detect people** and objects concerning the task (human, verb, object) triplets, and fusing it (InteractNet).



(a) (b)
(c) (d)

Object detection for multi-task learning

- Interesting papers:
  - Detect-and-Track: Efficient Pose Estimation in Videos – Girdhar et al (CMU).
    - Human detection by **3d-Mask-RCNN** and link the detections for video understanding.



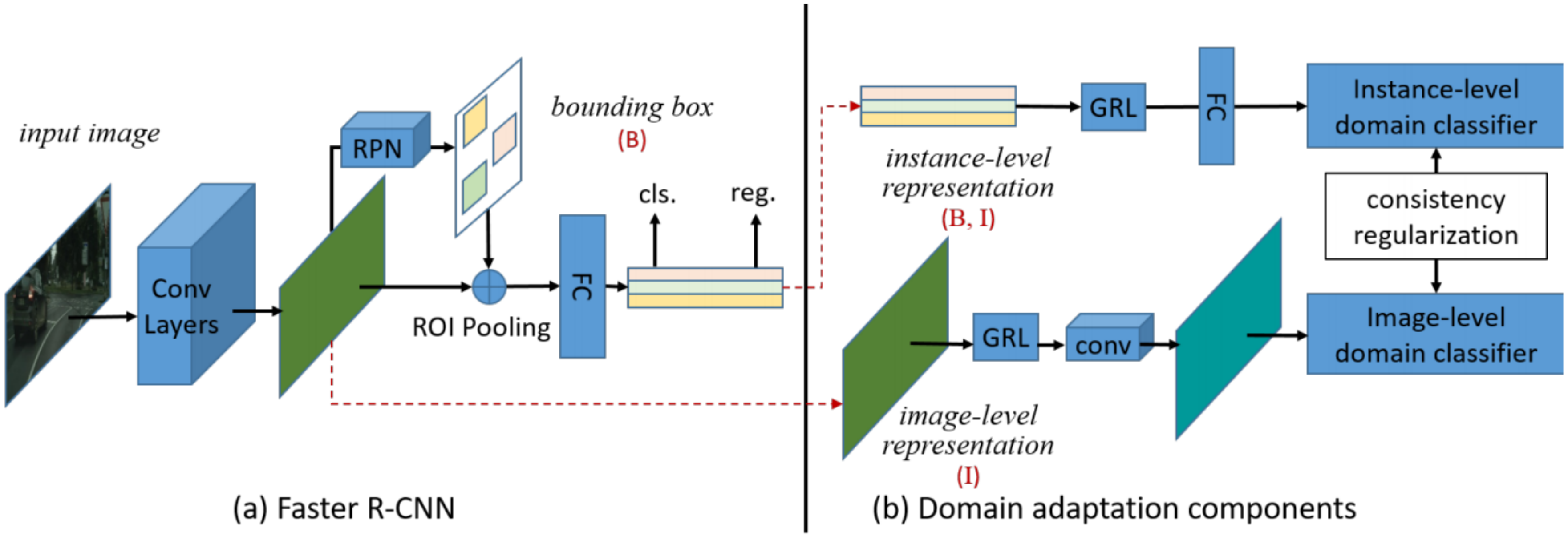
Others

- Interesting papers:
  - Domain Adaptive Faster R-CNN for Object Detection in the Wild
    - **Faster R-CNN** + two domain adaptation components, on image level and instance level, to reduce the domain discrepancy (for autonomous driving).

  - R-FCN-3000 at 30fps: Decoupling Detection and Classification
    - Decoupled **R-FCN** (+ another classification network) to learn 3000-classes + novel classes.

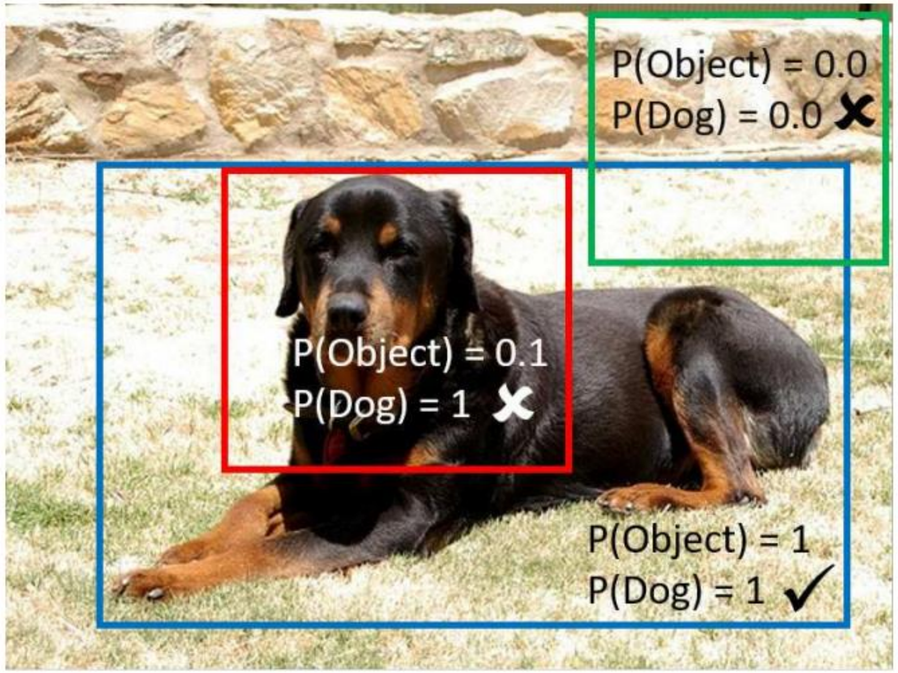
Others

- Interesting papers:
  - Domain Adaptive Faster R-CNN for Object Detection in the Wild
    - **Faster R-CNN** + two domain adaptation components, on image level and instance level, to reduce the domain discrepancy (for autonomous driving).



Others

- Interesting papers:
  - R-FCN-3000 at 30fps: Decoupling Detection and Classification
    - Decoupled **R-FCN** (+ another classification network) to learn 3000-classes + novel classes.



Another interesting paper

- Squeeze excitation networks

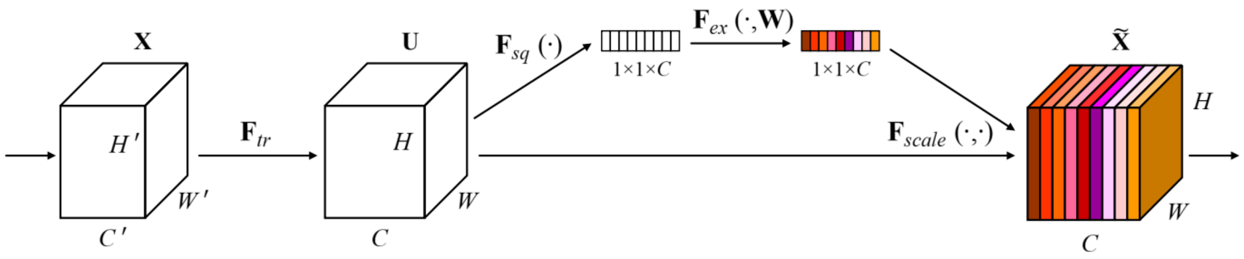
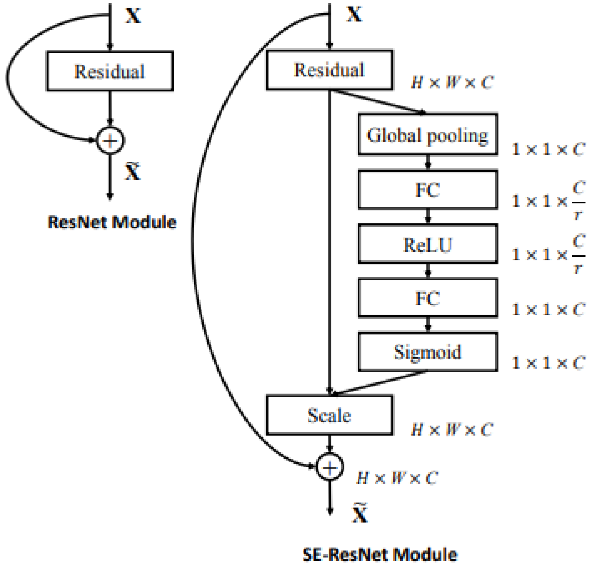


Figure 1: A Squeeze-and-Excitation block.

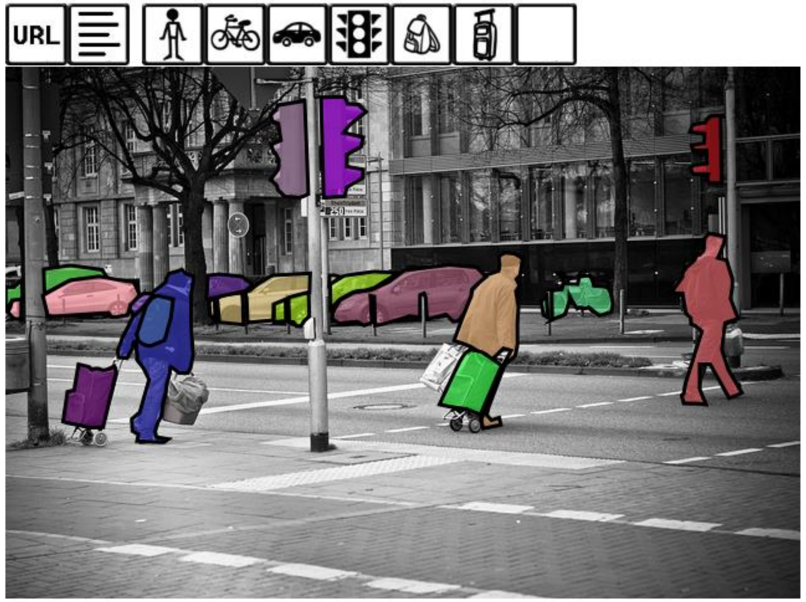


	224×224		320×320 / 299×299	
	top-1 err.	top-5 err.	top-1 err.	top-5 err.
ResNet-152 [10]	23.0	6.7	21.3	5.5
ResNet-200 [11]	21.7	5.8	20.1	4.8
Inception-v3 [44]	-	-	21.2	5.6
Inception-v4 [42]	-	-	20.0	5.0
Inception-ResNet-v2 [42]	-	-	19.9	4.9
ResNeXt-101 (64×4d) [47]	20.4	5.3	19.1	4.4
DenseNet-264 [14]	22.15	6.12	-	-
Attention-92 [46]	-	-	19.5	4.8
Very Deep PolyNet [51] †	-	-	18.71	4.25
PyramidNet-200 [8]	20.1	5.4	19.2	4.7
DPN-131 [5]	19.93	5.12	18.55	4.16
SENet-154	**18.68**	**4.47**	**17.28**	**3.79**
NASNet-A (6@4032) [55] †	-	-	17.3‡	3.8‡
SENet-154 (post-challenge)	-	-	**16.88**‡	**3.58**‡

Issues on unlabeled or noisy data

Why is it important?

- Even though there are plenty of labeled data (MS-COCO: 200K)
- Still, there are a huge number of raw data
- Annotating these data is too expensive

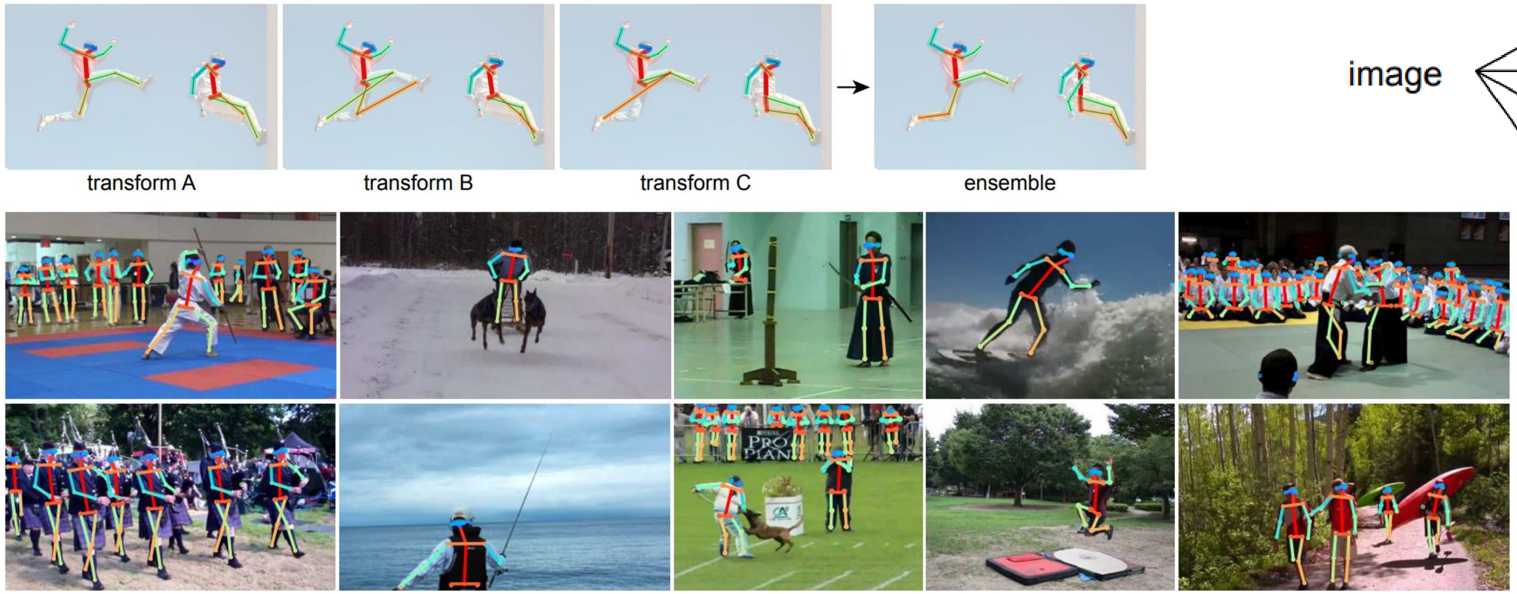


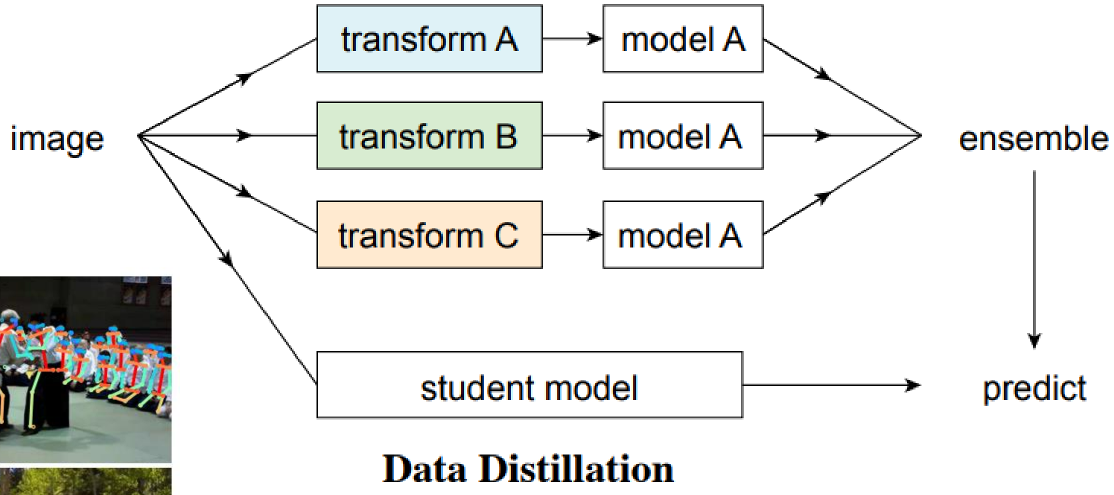
Making pseudo annotation of unlabeled data

- **Pseudo Mask** Augmented Object Detection (Zhao et al. MS Rearch)
- Cross-Domain **Weakly-Supervised** Object Detection through Progressive Domain Adaptation (Inoue et al., Univ of Tokyo)
- Improving Landmark Localization with **Semi-Supervised** Learning
- **Data Distillation**: Towards Omni-Supervised Learning (Honary et al., MILA)
- Towards Human-Machine Cooperation: **Self-supervised Sample Mining** for Object Detection (Wang et al., SYU & Sensetime)

Making pseudo annotation of unlabeled data

- Base model is trained with labeled data
- Unlabeled data -> various **transform** (augmentation)
  -> Prediction using base model
- Ensemble the outputs -> **Pseudo ground truth**

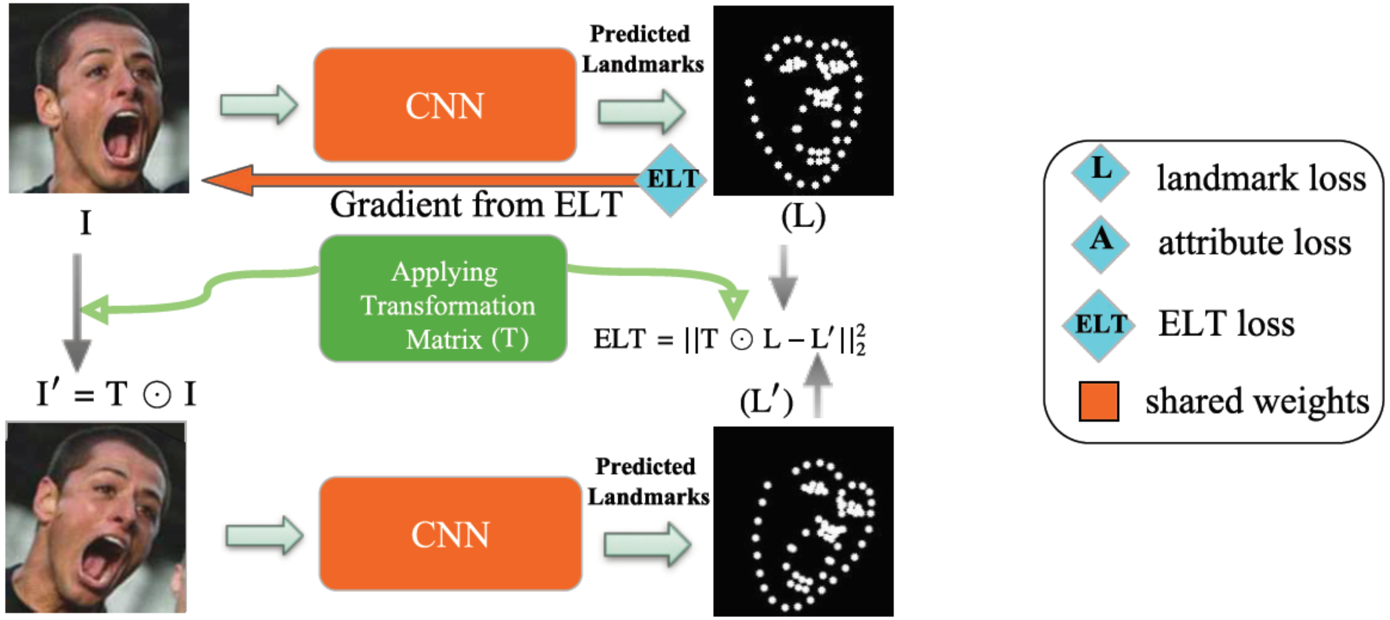




Data Distillation: Towards Omni-Supervised Learning, CVPR 201

Making pseudo annotation of unlabeled data

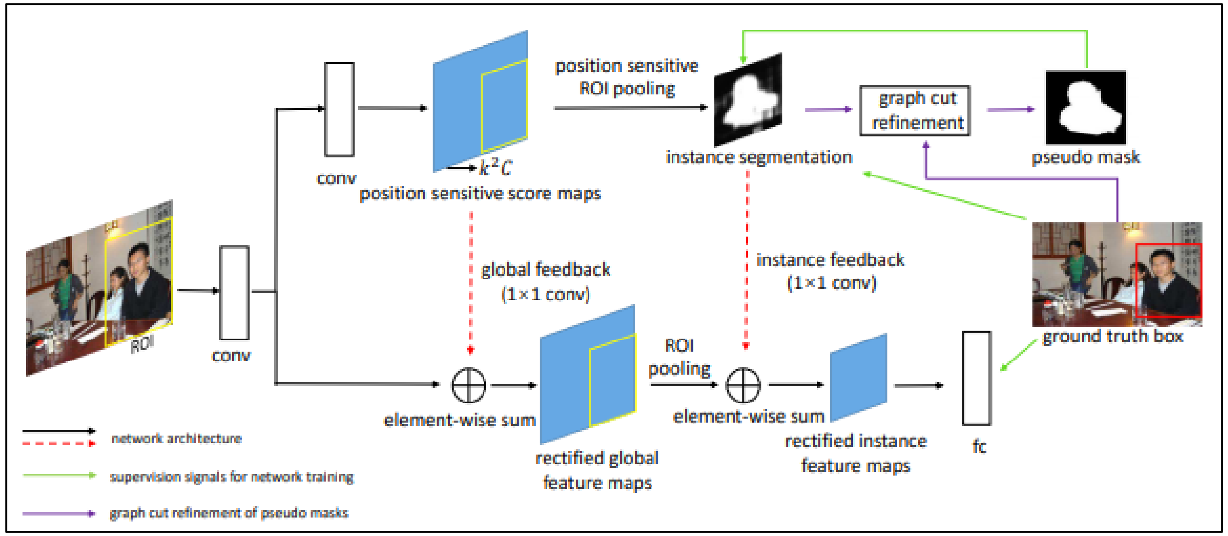
- Base model is trained with labeled data
- Landmark results w/ and w/o transformation augmentation are consistent, then the predicted landmark locations are used as pseudo ground truth

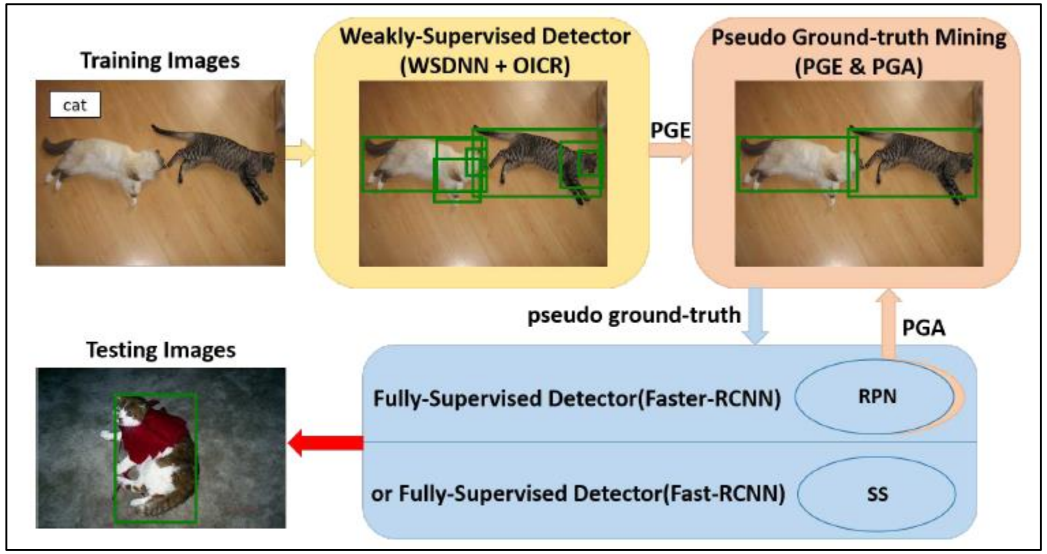


Improving Landmark Localization with Semi-Supervised Learning, CVPR

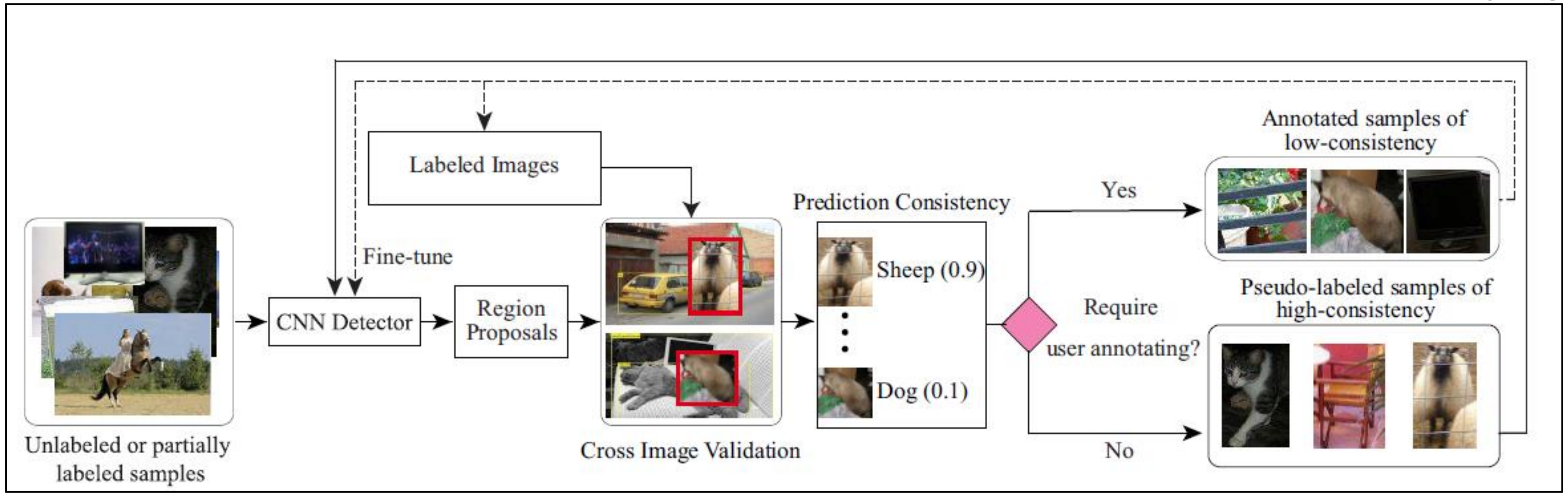
Making pseudo annotation of unlabeled data

Pseudo Mask Augmented Object Detection, CVPR 2018





W2F: A Weakly-Supervised to Fully-Supervised Framework for Object Detection, CVPR 2018



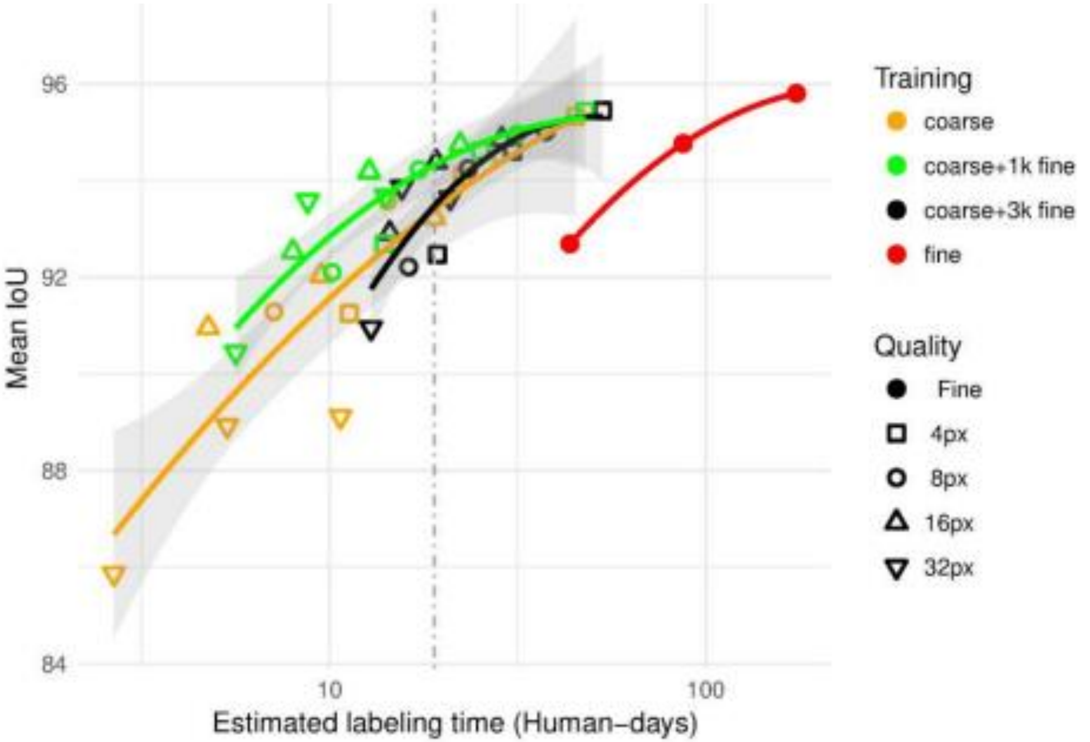
Towards Human-Machine Cooperation: Self-supervised Sample Mining for Object Detection, CVPR 2018

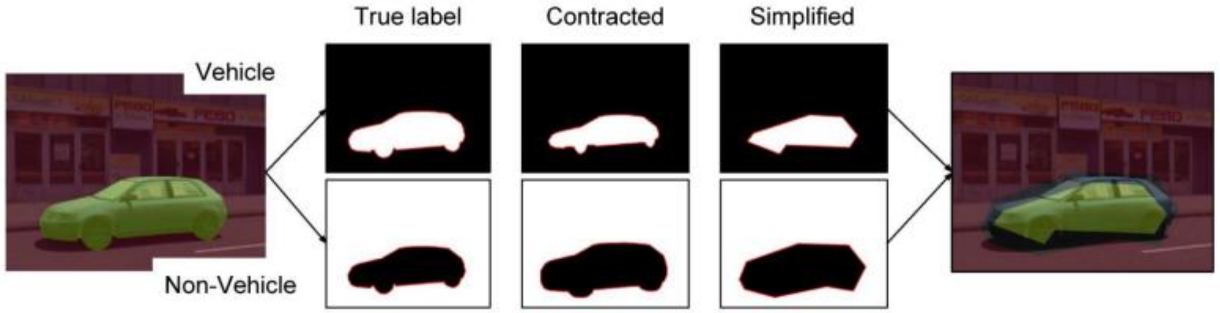
Coarse or noisy data annotation

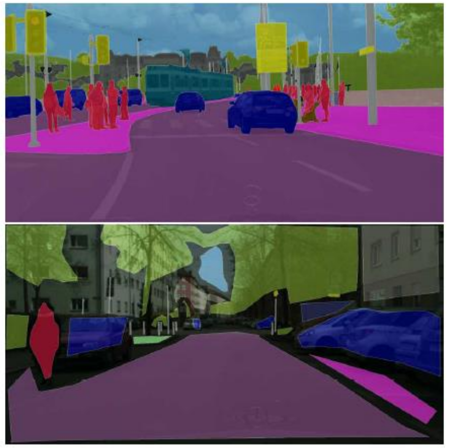
- On the importance of **label quality** for semantic segmentation (Zlateski et al., MIT)
- Learning from **noisy web data** with category-level supervision (Niu et al., Rice Univ)
- Deep Unsupervised Saliency Detection: A Multiple **Noisy Labeling** Perspective (Zhang et al., NPU)
- A Generative Adversarial Approach for Zero-Shot Learning From **Noisy Texts** (Zhu et al., Rutgers Univ)
- Separating Self-Experession and Visual Content in Hashtag Supervision (Veit et al., Cornell Univ & FAIR)

Coarse data annotation

- Large number of coarse annotation is better than few fine annotations





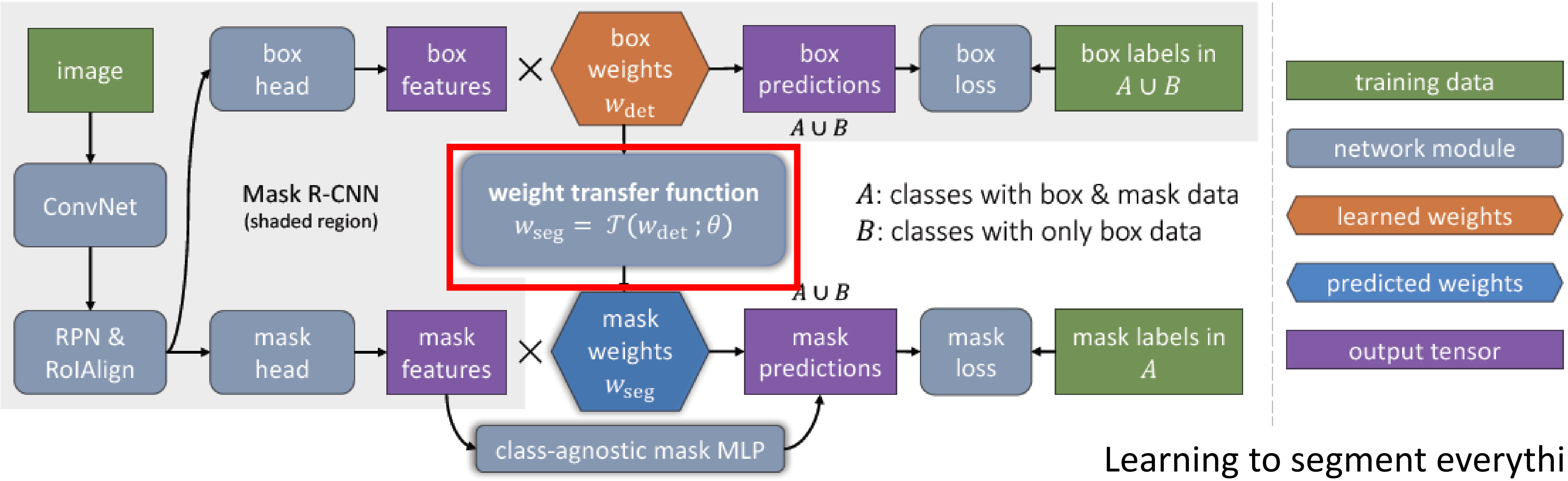


ure 1: A finely annotated (top) and a corsely annotated (bottom) image from the CityScape's dataset.

On the importance of label quality for semantic segmentation, CVPR 2018

Transfer weights between tasks

- Segmentation and detection annotation: 80 classes
- Only detection annotation: ~3000 classes (no segmentation)
- Train relationship between detection and segmentation tasks. (80 classes)
- Weights for detection is transferred to segmentation task



Learning to segment everything, CVPR 2018