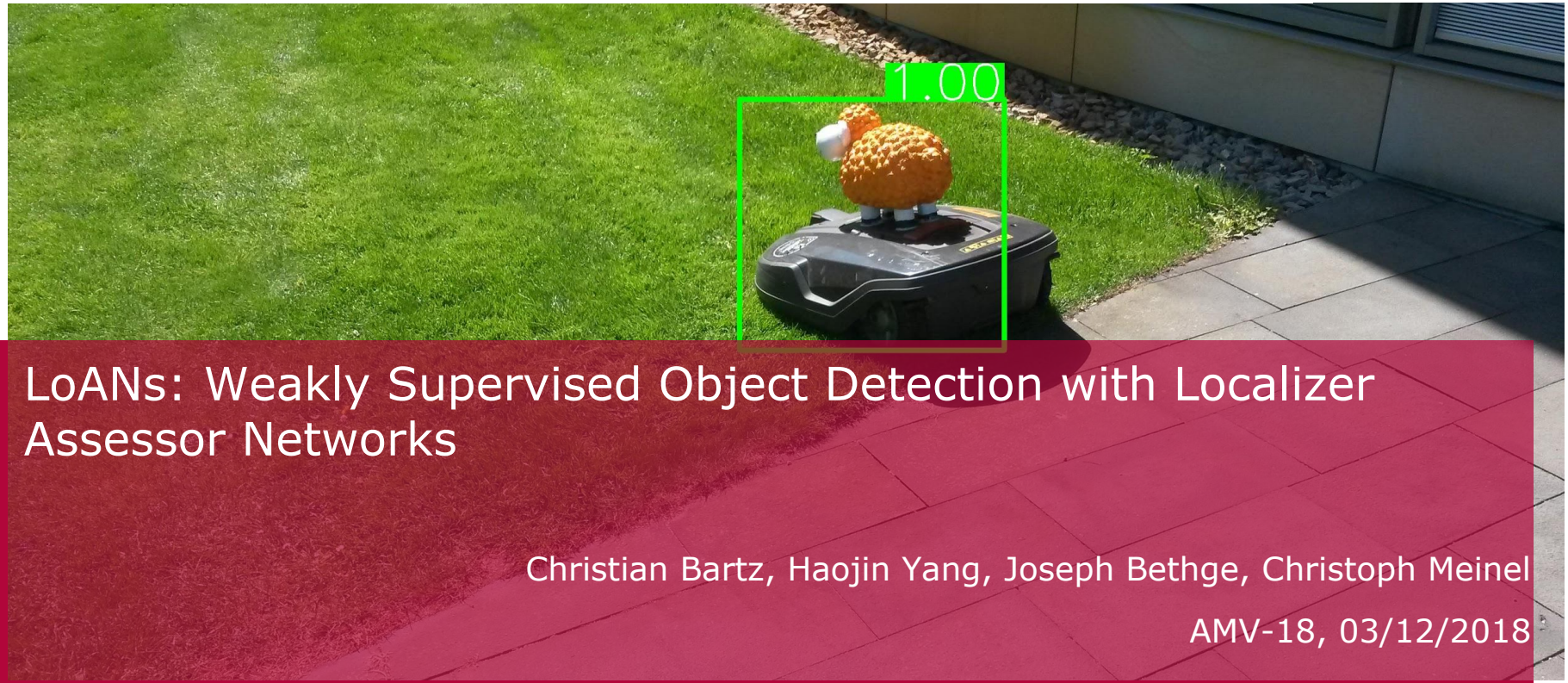
HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam




# LoANs: Weakly Supervised Object Detection with Localizer Assessor Networks

Christian Bartz, Haojin Yang, Joseph Bethge, Christoph Meinel

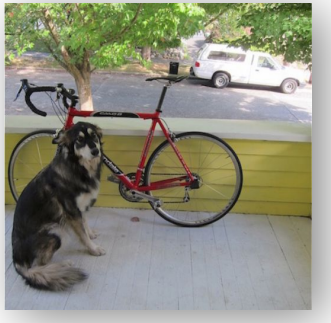
AMV-18, 03/12/2018

# Motivation
# Success Factors of Deep Learning

- availability of hardware for massive parallel computations
- large-scale labeled datasets
  - ImageNet dataset contains more than 14 mio labeled images
  - Youtube-8M dataset contains more than 7 mio labeled videos

# Motivation
# Object Detection

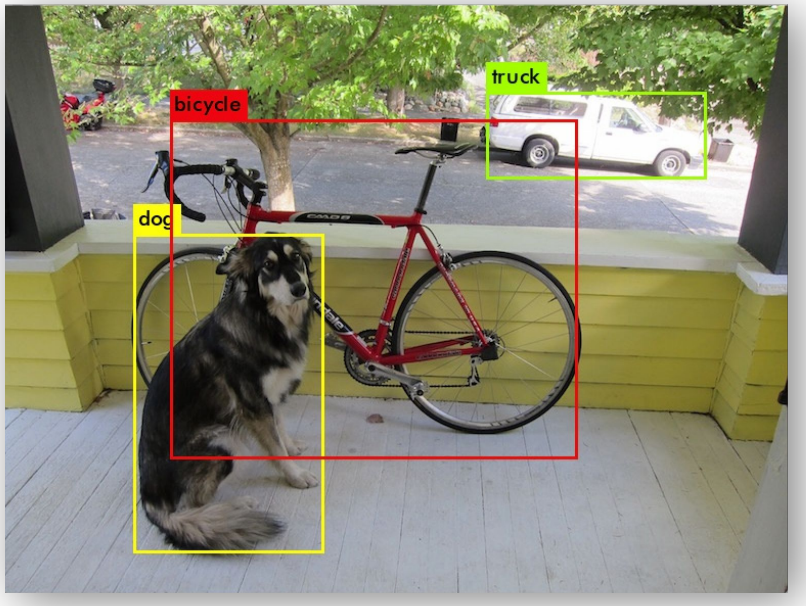
- recent breakthroughs [Redmon16, Liu16, Ren15] use fully annotated datasets
  - labels for locations of objects and classes of objects



+

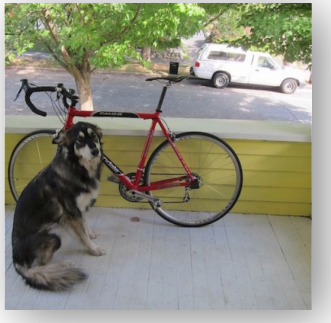
```
[
   {
      "class": "dog",
      "box": [200, 50, 400, 100]
   },
   {
      "class": "bicycle",
      "box": [100, 60, 350, 225]
   },
   {
      "class": "truck",
      "box": [75, 200, 150, 300]
   }
]
```

=



# Motivation
# Weakly Supervised Object Detection

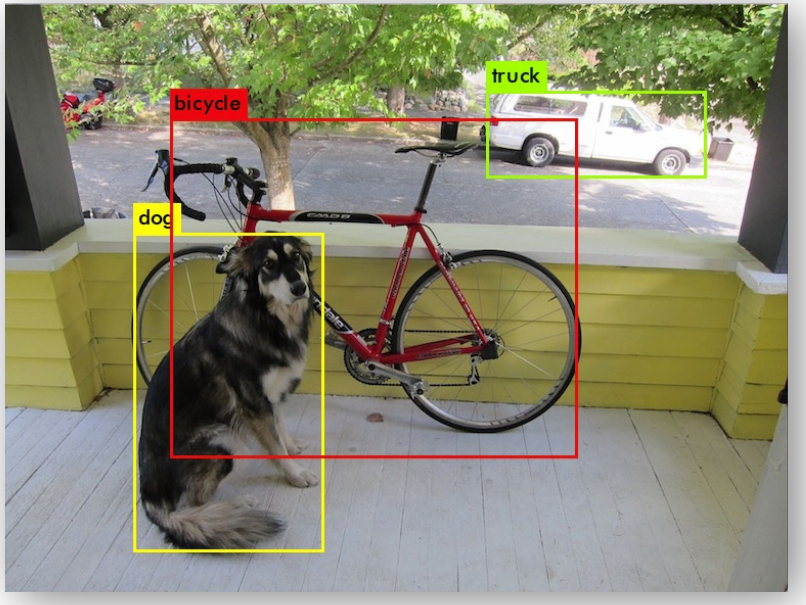
- recent methods [Wei18, Tang18] only use class labels
  - leverage implicit localization capability of feature extractor



+

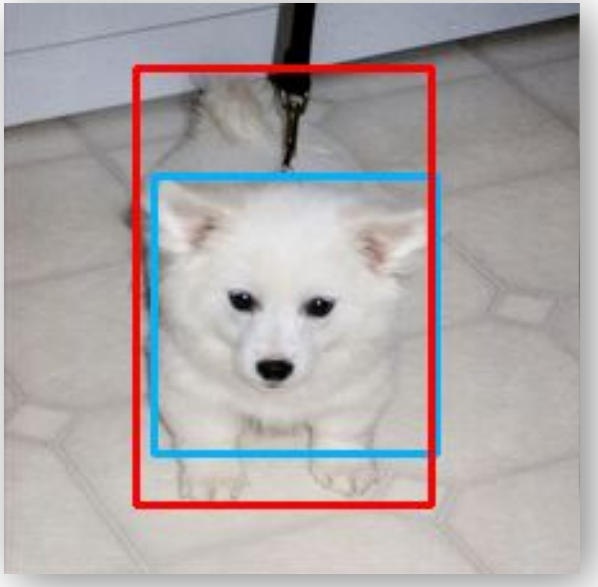
```
[
    {"class": "dog"},
    {"class": "bicycle"},
    {"class": "truck"}
]
```

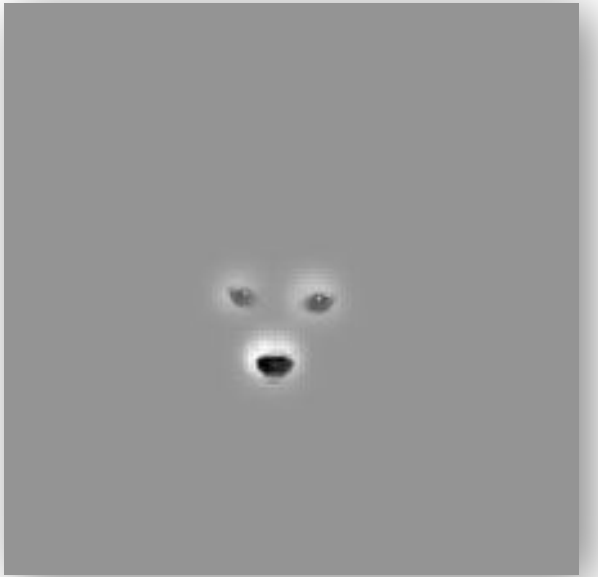
=



# Motivation
# Weakly Supervised Object Detection

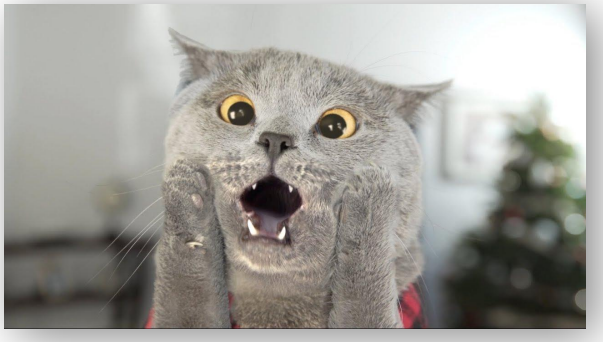
- current weakly supervised methods have inherent problems

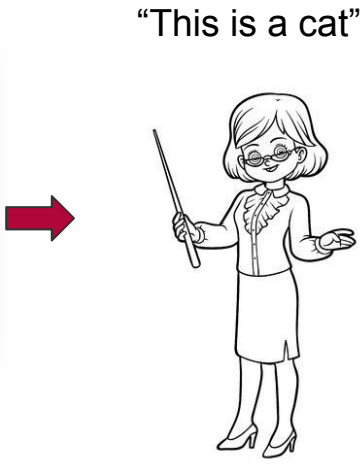




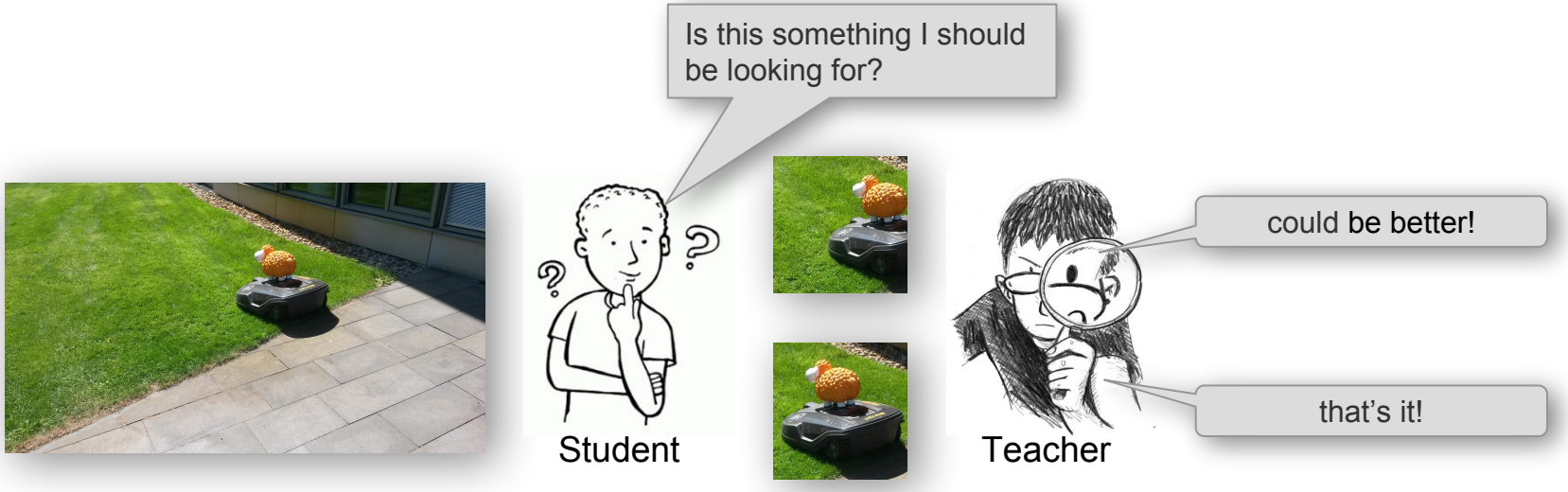
# Motivation
# Teacher Student Networks

- knowledge transfer [Hinton15, Chen16] between networks does not need labels for training the student



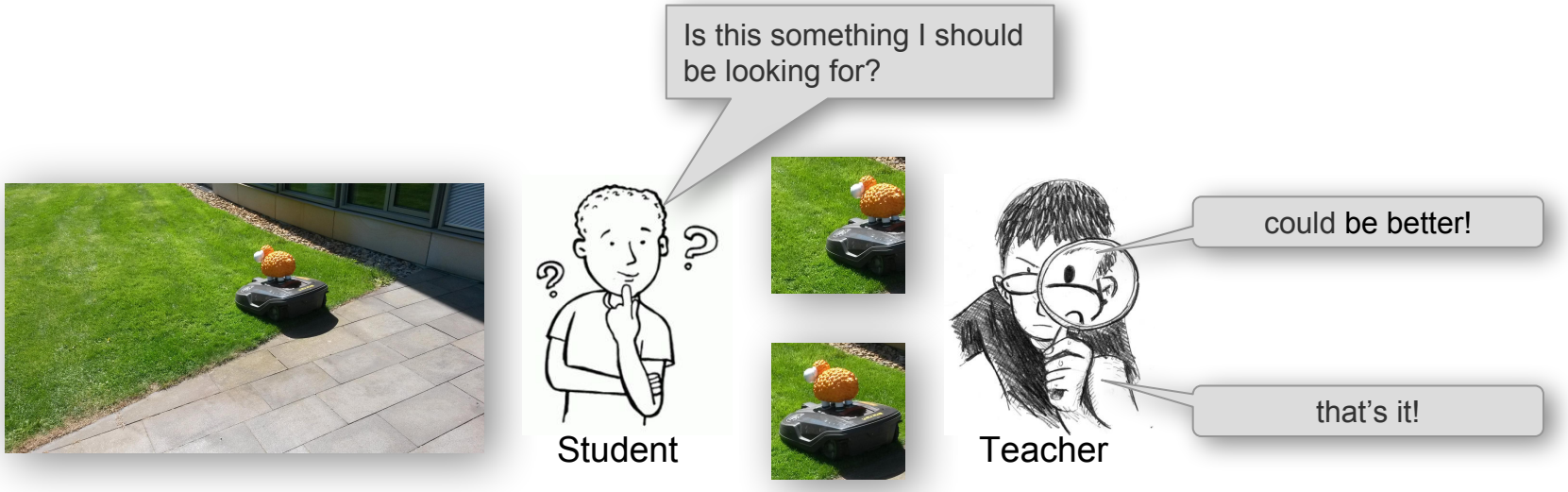



# Motivation
# Idea

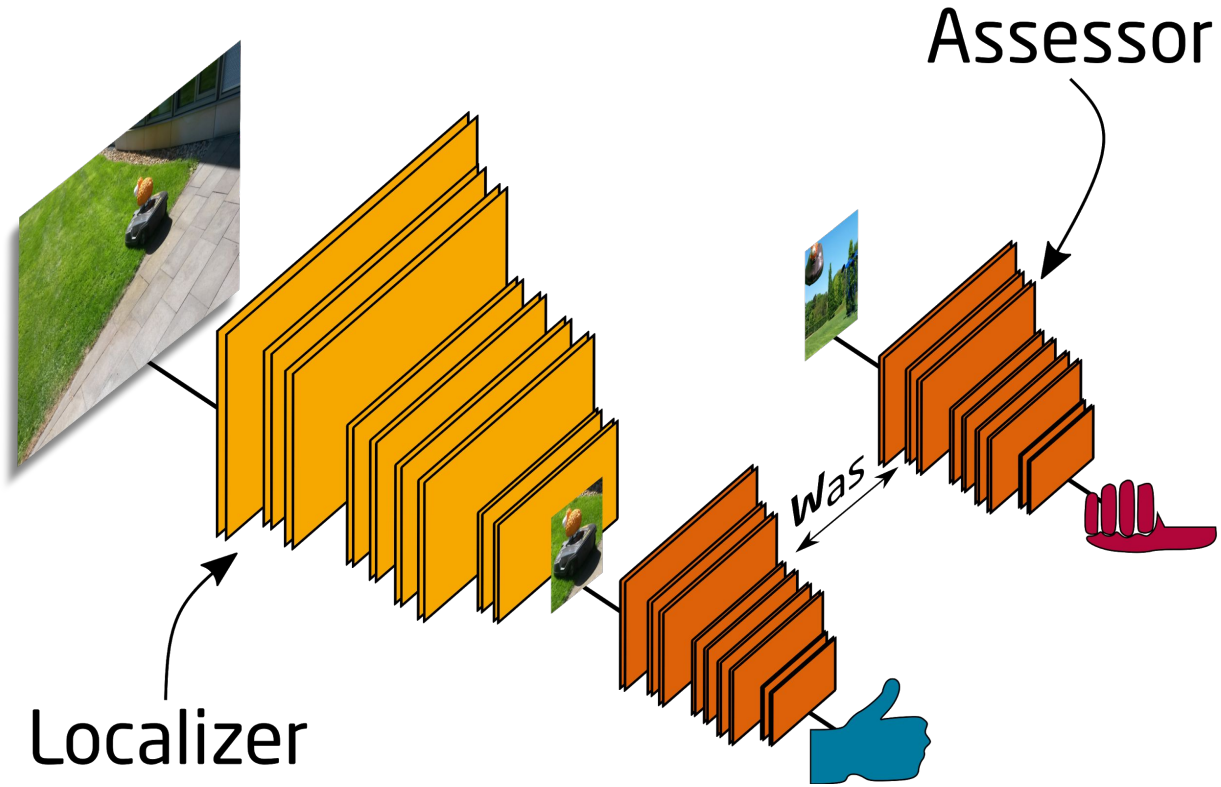


# Motivation
# Idea

- no labels for bounding boxes necessary
- no model pre-trained on ImageNet necessary
- can use artificial data for training teacher model

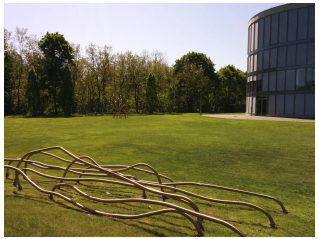


# Proposed System
## Localizer Assessor Networks



# Proposed System
## Assessor

- predict intersection over union (IOU) of image crop and shown object
- trained on synthetically generated data
- needs background and template images



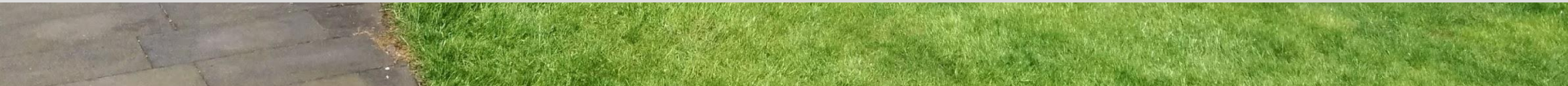




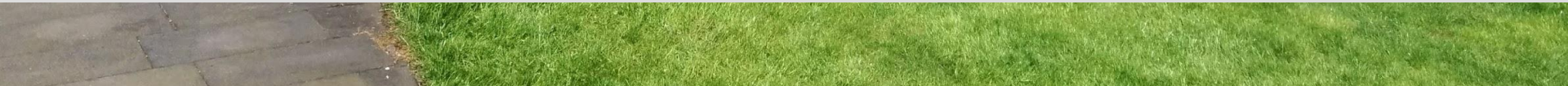


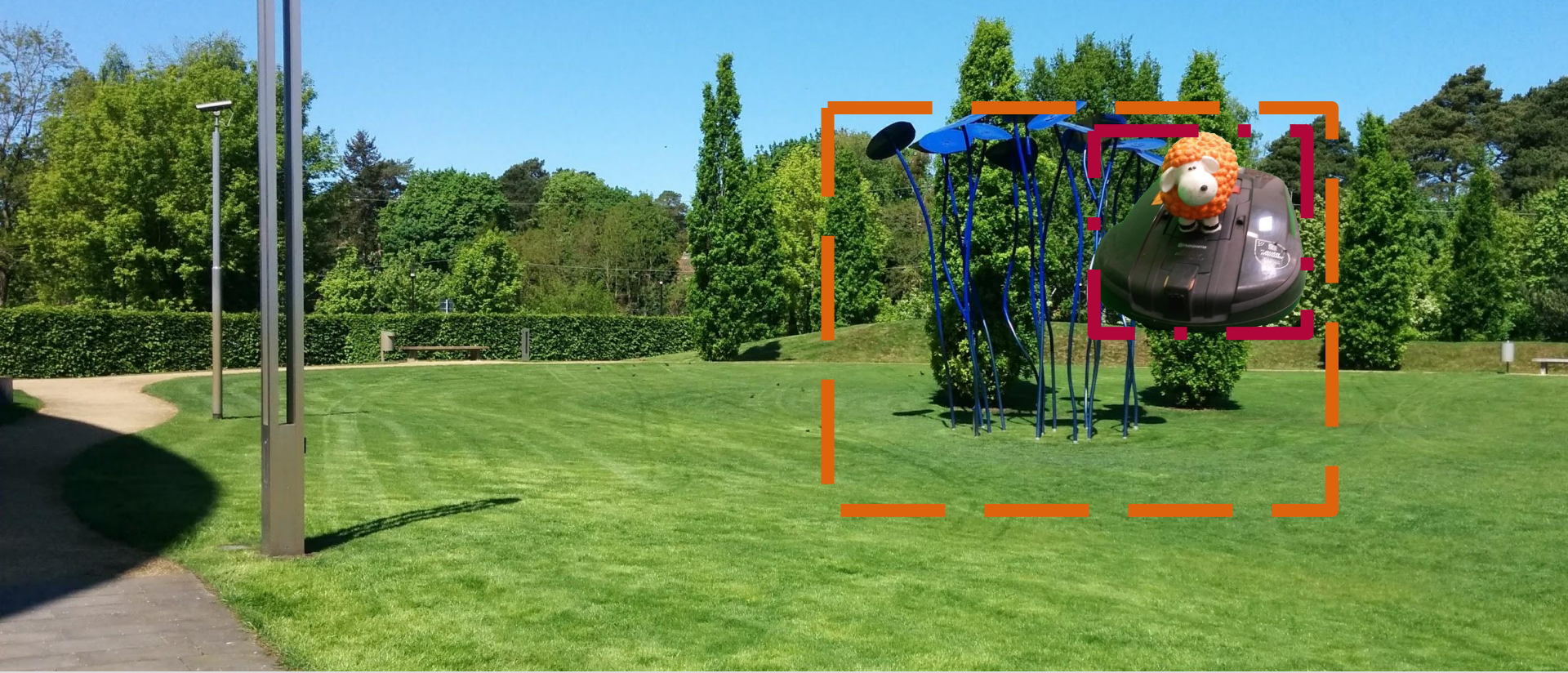



Step 0: select a background image

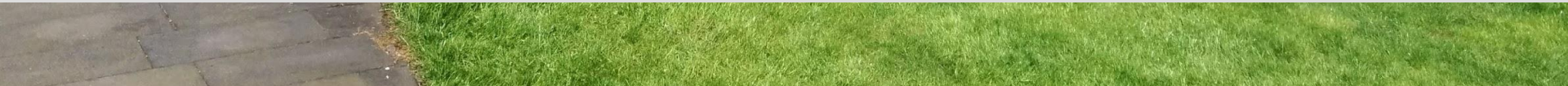


Step 1: place template image at random location in background image





Step 2: find a box with desired intersection over union



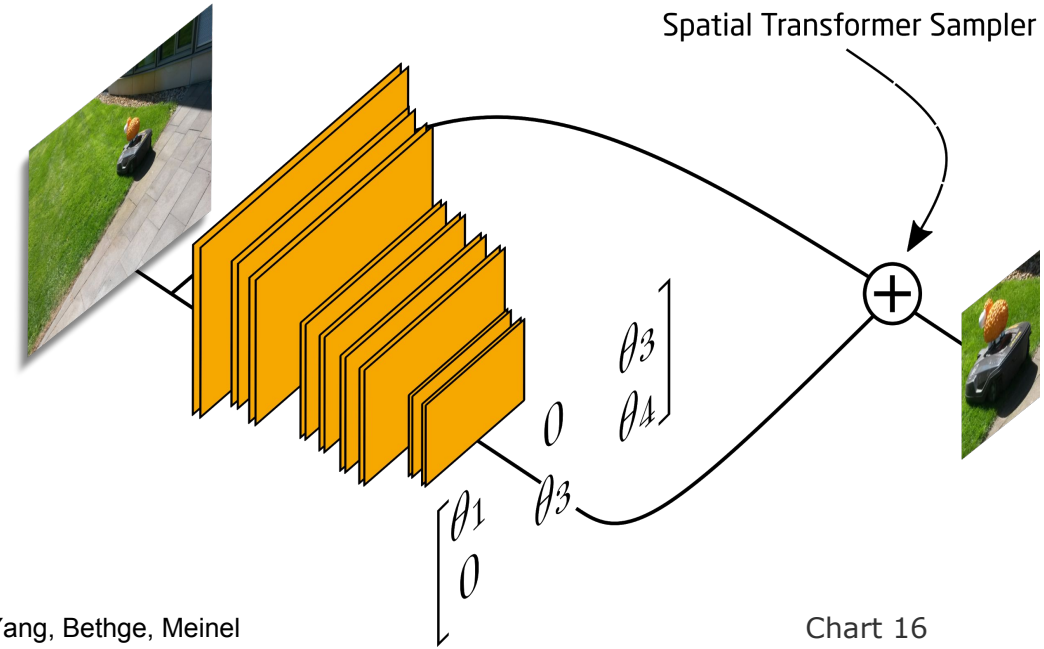
Step 3: crop box and resize image

# Proposed System
## Assessor

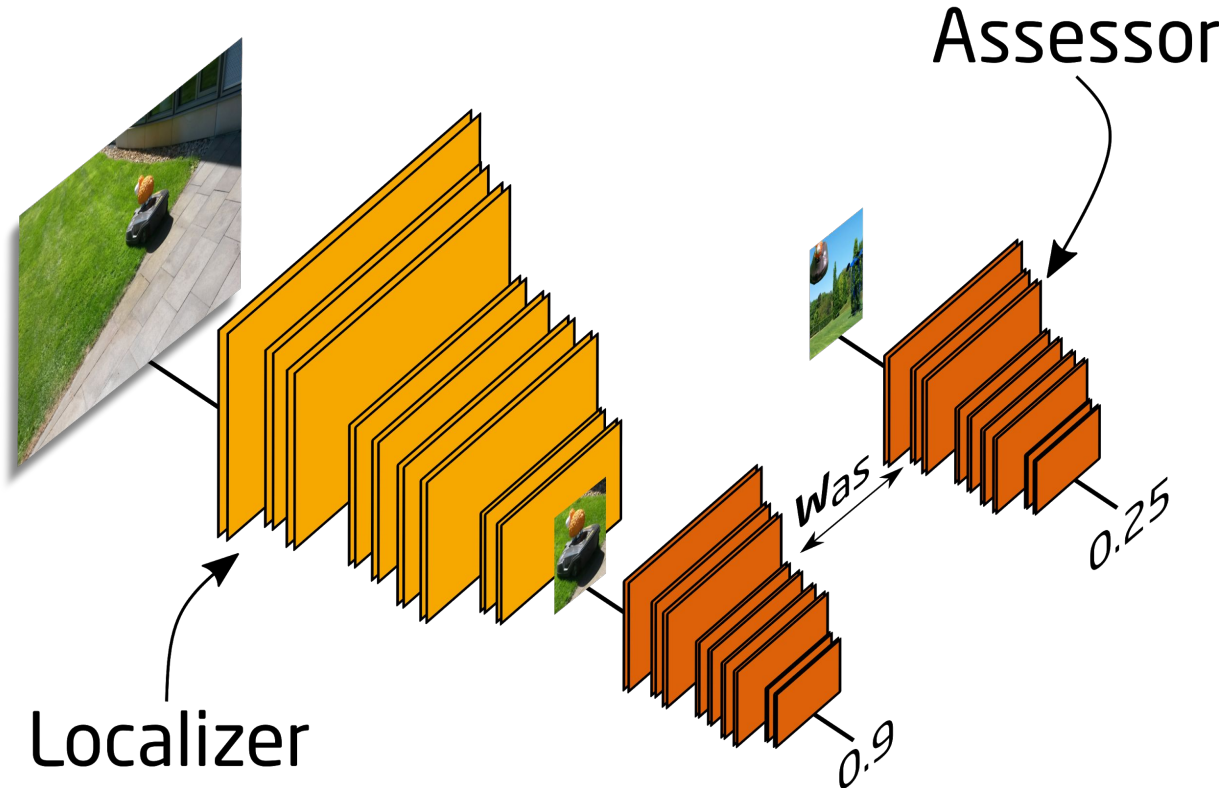
- save image and intersection over union of crop with bounding box of object

# Proposed System
## Localizer

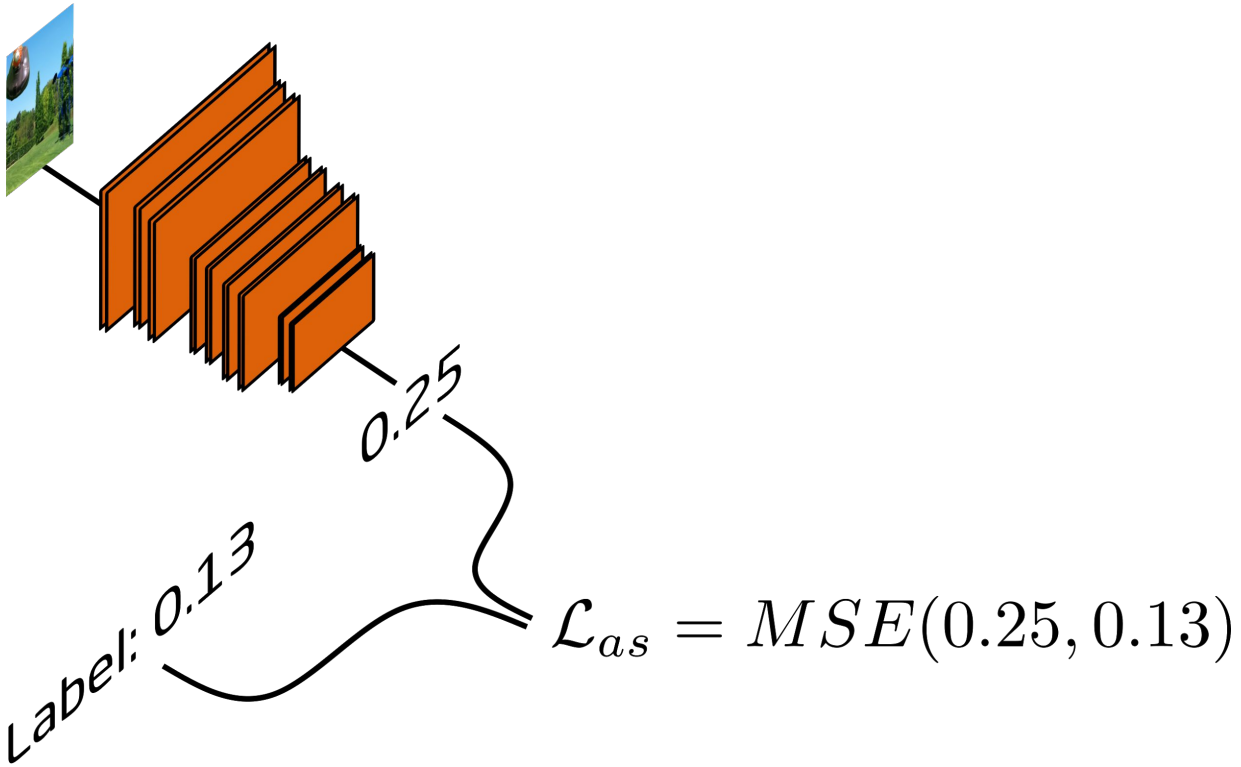
- predict a image region that is likely to contain target object
- crop image region with a spatial transformer [Jaderberg15]
- trained on unlabeled data
- entirely supervised by assessor



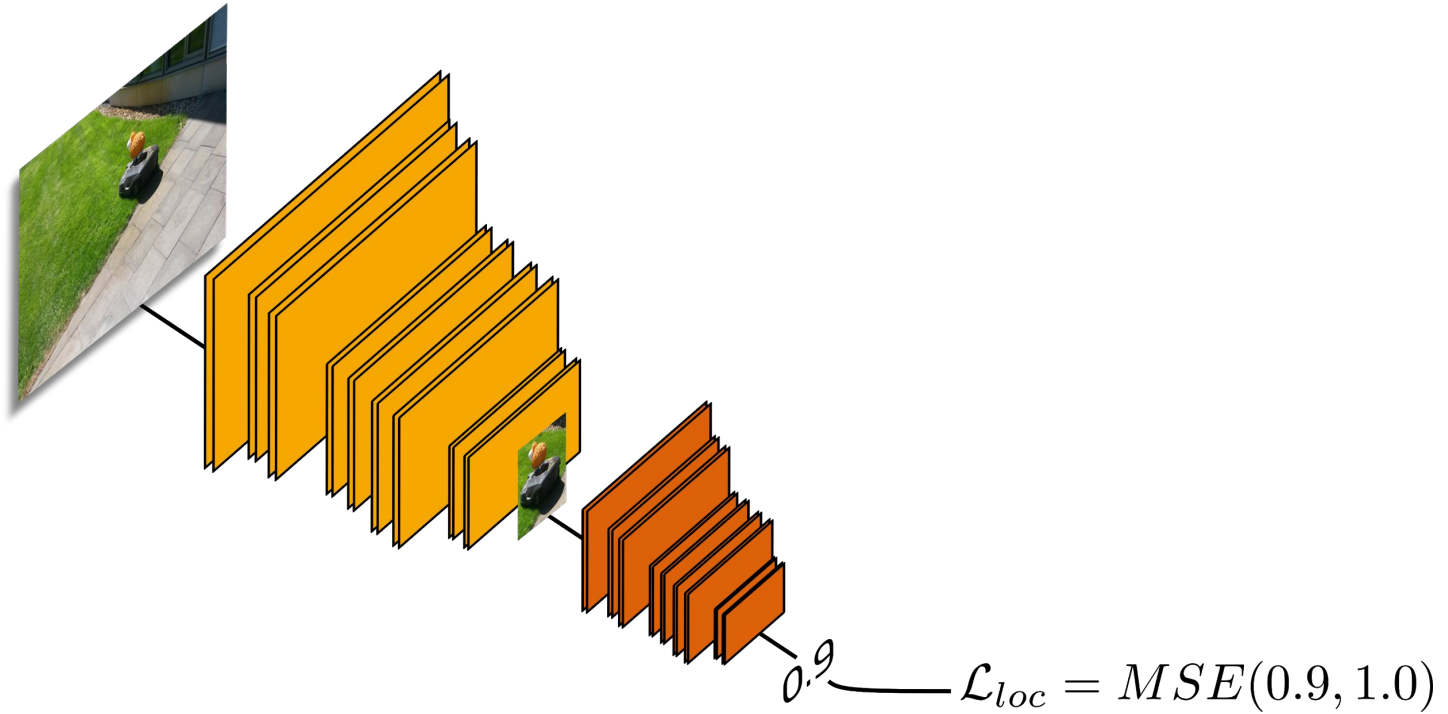
# Proposed System
## Training of Both Networks



# Proposed System
## Training of Assessor



# Proposed System
## Training of Localizer



# Experiments
# Sheep Dataset

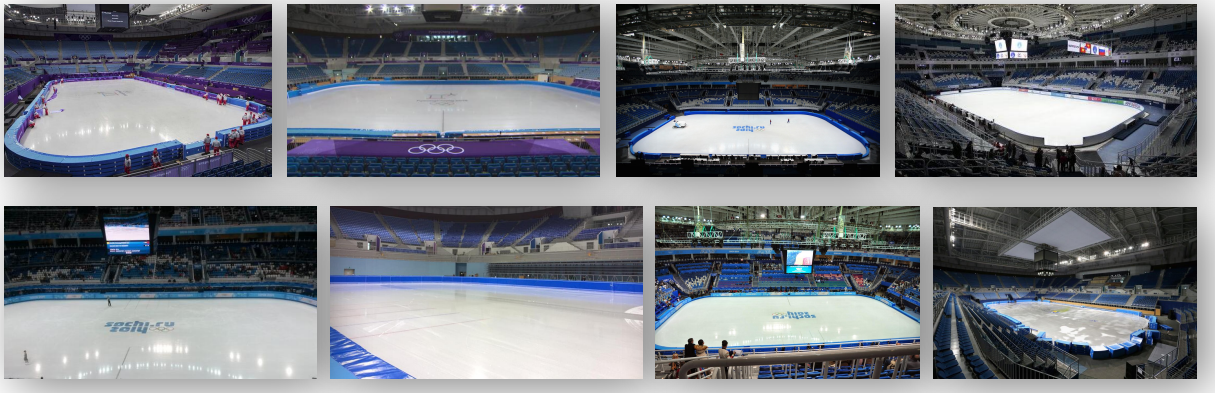
- 8,320 fully annotated images for training localizer
  - full annotation enables comparison to fully supervised approach
- 10,000 images for training assessor

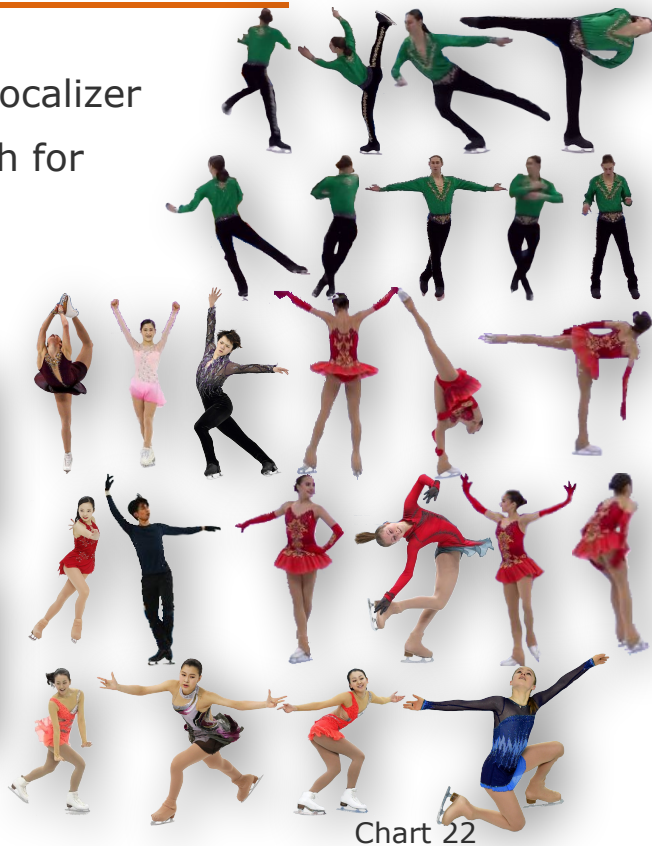
| Method | 224 x 224 | 300 x 300 | 512 x 512 |
|---|---|---|---|
| SSD [Liu16] | - | 0.887 | 0.969 |
| ResNet-18 | 0.887 | 0.937 | 0.967 |
| ResNet-50 | 0.959 | 0.958 | 0.976 |


SSD Threshold 0.3
LoANs Threshold 0.3
SSD Threshold 0.001
LoANs Theshold 0.0

# Experiments
# Figure Skating Dataset

- 4 YouTube videos are enough to create train dataset for localizer
- 8 background images and 25 template images are enough for generation of assessor dataset

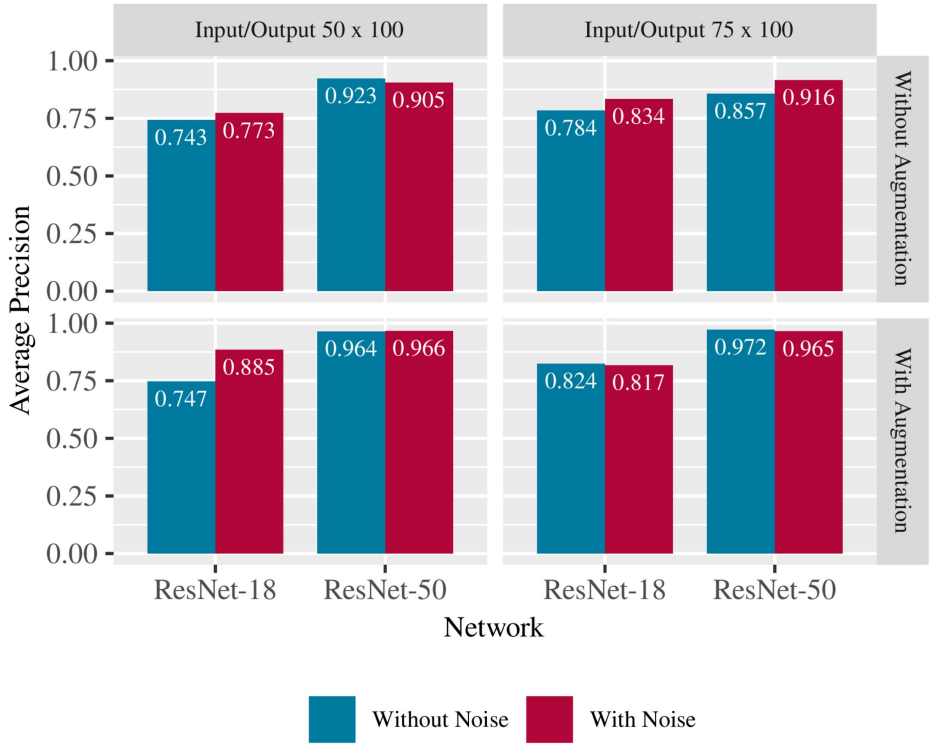




# Experiments
# Figure Skating Dataset

- YouTube videos contain a lot of noisy images
- experimented with noisy data and without noise
  - dataset without noise only contains 48% of the number of original images

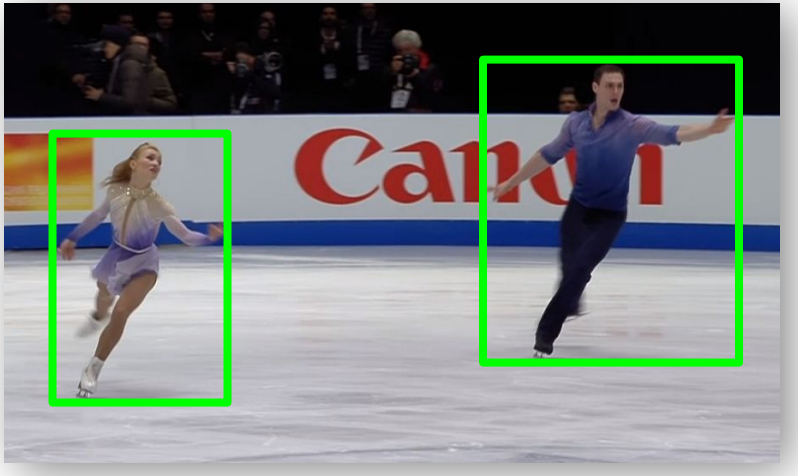
# Experiments
## Results on Figure Skating Dataset

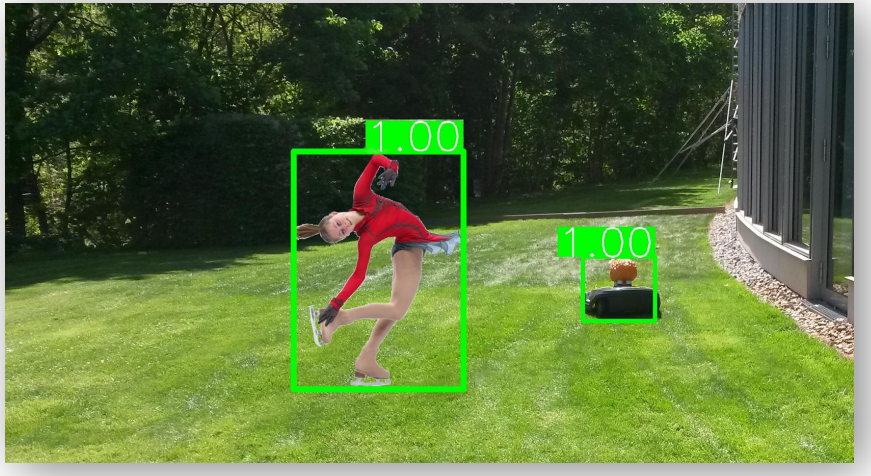


# Limitations and Future Work

- approach only works for one class at a time
- only works with images containing a single object
- we have no means to determine whether system detected object or not

**What we would like to have in the future:**

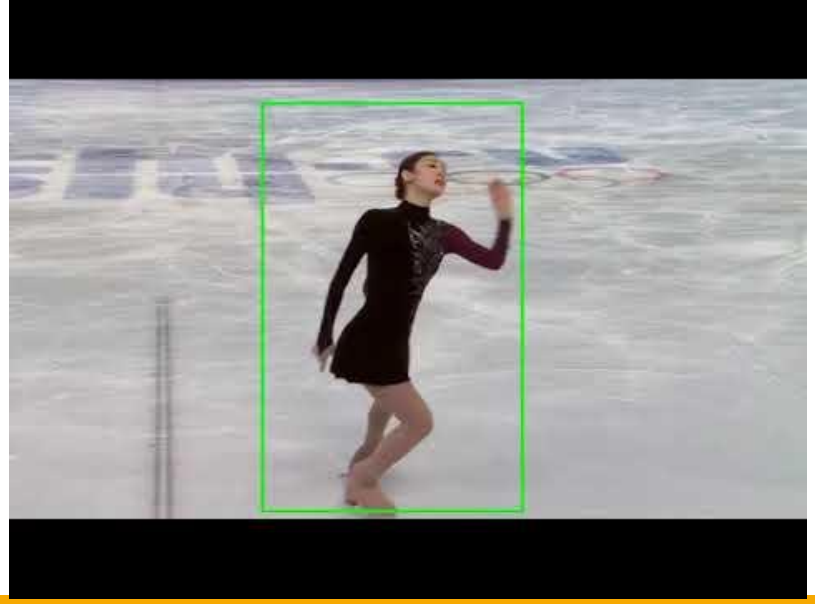




# Conclusion

- presented a novel approach for weakly supervised detection
- should be simple and cost efficient to create training data for specialized systems
- code, models and datasets are available online:

https://github.com/Bartzi/loans



HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

# Thank you for your attention

Thank you for your attention

# References


- [Redmon16] - Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- [Liu16] - Liu, Wei, et al. "Ssd: Single shot multibox detector." *European conference on computer vision*. Springer, Cham, 2016.
- [Ren15] - Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks
- [Wei18] - Wei, Yunchao, et al. "TS2C: tight box mining with surrounding segmentation context for weakly supervised object detection." European Conference on Computer Vision. Springer, Cham, 2018.

# References


- [Tang18] - Tang, Peng, et al. "Weakly supervised region proposal network and object detection." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
- [Hinton15] - Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the Knowledge in a Neural Network." NIPS Deep Learning and Representation Learning Workshop. 2015.
- [Chen16] - Chen, Tianqi, Ian Goodfellow, and Jonathon Shlens. "Net2net: Accelerating learning via knowledge transfer." International Conference on Learning Representations. 2016.

# References


- [Jaderberg15] - Jaderberg, Max, Karen Simonyan, and Andrew Zisserman. "Spatial transformer networks." Advances in neural information processing systems. 2015.