Track and Transfer: Watching Videos to Simulate Strong Human Supervision for Weakly-Supervised Object Detection

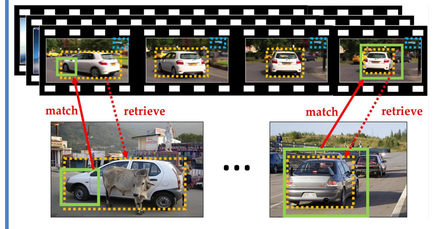
UC DAVIS UNIVERSITY OF CALIFORNIA

Krishna Kumar Singh **Fanyi Xiao** **Yong Jae Lee**

CVPR 2016

Motivation and Idea:

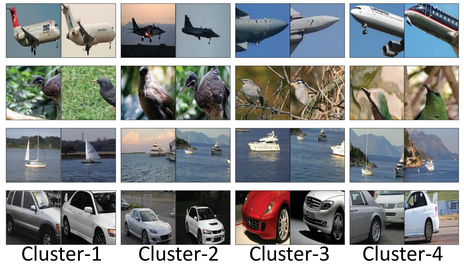
Train detectors with inexpensive image and video level tags, without using any localization information



- Existing weakly-supervised methods mine discriminative visual patterns, but they correspond to object-parts or include co-occurring background
- **Key Idea**: Transfer tracked object boxes from videos as a substitute for strong human supervision to obtain more precise boxes

Approach:

Mining discriminative positive visual regions:

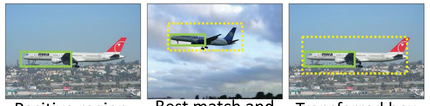


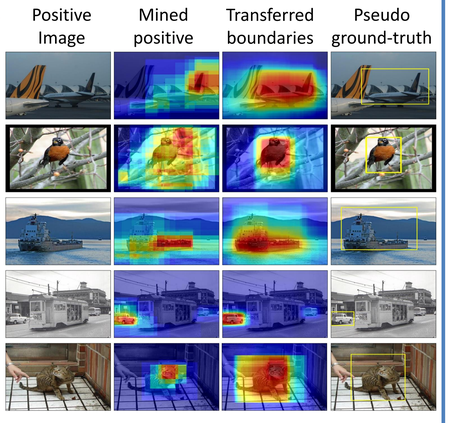
- Find discriminative object proposals by clustering in pre-trained Alexnet *pool-5* feature space and ranking clusters based on class consistency
- Most clusters do not tightly fit object due to scene clutter, occlusion, and intra class appearance variation

Unsupervised video object tracking:

- Apply appearance and motion based unsupervised tracking method [1]
- For each video choose 9 highest ranked tracks
- Select highest scoring box: $score\left(t_j^f\right) = \Sigma_i IOU\left(v_i^f, t_j^f\right)\ X\ sim\left(r_i, v_i^f\right)$

Transferring tracked object boxes:





- For each positive region, find *n* best matching video regions and return corresponding tracked object boxes to the image
- Create 4-dimensional hough space where each box casts a vote for its coordinates

Training an object detector:

- Train R-CNN using discovered pseudo ground-truth boxes
- Perform latent SVM update to improve the pseudo GT boxes
- Fine-tune the R-CNN model to update the features using pseudo GT boxes

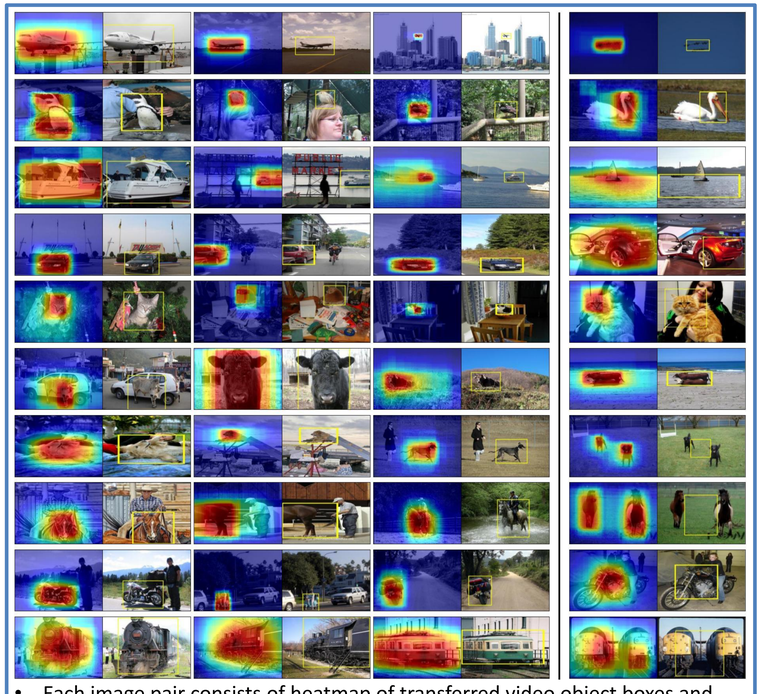
Quantitative Results:

VOC 2007	*Aero*	*Bird*	*Boat*	*Car*	*Cat*	*Cow*	*Dog*	*Horse*	*Mbike*	*Train*	*mAP*
Wang et al., 2014	48.9	26.1	11.3	40.9	34.7	**34.7**	34.4	35.4	**52.7**	34.8	35.4
Cinbis et al., 2015	39.3	28.8	**20.4**	47.9	22.1	33.5	29.2	38.5	47.9	41.0	34.9
Ours	**53.9**	**37.7**	13.7	**56.6**	**51.3**	24.0	**38.5**	**47.9**	47.0	**48.4**	**41.9**

VOC 2010	*Aero*	*Bird*	*Boat*	*Car*	*Cat*	*Cow*	*Dog*	*Horse*	*Mbike*	*Train*	*mAP*
Cinbis et al., 2015	44.6	25.5	**14.1**	36.3	23.2	26.1	29.2	36.0	**54.3**	31.2	32.1
Ours	**53.5**	**37.5**	8.0	**44.2**	**49.4**	**33.7**	**43.8**	**42.5**	47.6	**40.6**	**40.1**

Ablation Study (VOC 2007)	*Aero*	*Bird*	*Boat*	*Car*	*Cat*	*Cow*	*Dog*	*Horse*	*Mbike*	*Train*	*mAP*
Initial pseudo GT	43.4	30.5	11.9	50.2	39.6	16.7	31.6	36.7	42.2	40.7	34.4
Updated pseudo GT (UGT)	48.0	34.2	12.2	51.3	43.0	21.9	33.4	39.1	43.8	42.2	36.9
UGT + bbox-reg	50.7	36.6	13.4	53.1	50.8	21.6	37.6	44.0	46.1	43.4	39.7
UGT + bbox-reg + finetune	**53.9**	**37.7**	**13.7**	**56.6**	**51.3**	**24.0**	**38.5**	**47.9**	**47.0**	**48.4**	**41.9**

Qualitative Results:



- Each image pair consists of heatmap of transferred video object boxes and final selected pseudo GT box; last column shows some failure cases
- Our method accurately localizes the object in many images

Conclusion:

- A novel weakly-supervised object detection framework that tracks and transfers object boxes from weakly-labeled videos to images
- State-of-the-art-results on PASCAL 2007 and 2010 datasets for the 10 categories of the YouTube-Objects dataset

Acknowledgement: This work was supported in part by an Amazon Web Services Education Research Grant and GPUs donated by NVIDIA

Reference: [1] F. Xiao and Y. J. Lee. Track and segment: An iterative unsupervised approach for video object proposals. In CVPR, 2016.