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: \( \text{score}(t_f) = \sum \text{IOU}(v_i, t_f) \times \text{SIM}(r_i, v_f) \)

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:

<table>
<thead>
<tr>
<th>Method</th>
<th>VOC 2007</th>
<th>VOC 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aero</td>
<td>Bird</td>
</tr>
<tr>
<td>Wang et al., 2014</td>
<td>48.9</td>
<td>26.1</td>
</tr>
<tr>
<td>Cinbis et al., 2015</td>
<td>39.3</td>
<td>28.8</td>
</tr>
<tr>
<td>Ours</td>
<td>53.9</td>
<td>37.7</td>
</tr>
</tbody>
</table>

Ablation Study (VOC 2007):

<table>
<thead>
<tr>
<th>Method</th>
<th>Aero</th>
<th>Bird</th>
<th>Boat</th>
<th>Car</th>
<th>Cat</th>
<th>Cow</th>
<th>Dog</th>
<th>Horse</th>
<th>Mbike</th>
<th>Train</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial pseudo GT</td>
<td>43.4</td>
<td>30.5</td>
<td>11.9</td>
<td>50.2</td>
<td>39.6</td>
<td>16.7</td>
<td>31.6</td>
<td>36.7</td>
<td>42.2</td>
<td>40.7</td>
<td>34.4</td>
</tr>
<tr>
<td>Updated pseudo GT (UGT)</td>
<td>48.0</td>
<td>34.2</td>
<td>12.7</td>
<td>51.3</td>
<td>43.0</td>
<td>21.9</td>
<td>33.4</td>
<td>39.1</td>
<td>43.8</td>
<td>42.2</td>
<td>36.9</td>
</tr>
<tr>
<td>UGT + box-reg</td>
<td>50.7</td>
<td>36.6</td>
<td>13.4</td>
<td>53.1</td>
<td>50.8</td>
<td>21.6</td>
<td>37.6</td>
<td>44.0</td>
<td>46.2</td>
<td>43.4</td>
<td>38.7</td>
</tr>
<tr>
<td>UGT + box-reg + finetune</td>
<td>53.9</td>
<td>37.7</td>
<td>13.7</td>
<td>56.6</td>
<td>51.3</td>
<td>24.0</td>
<td>38.5</td>
<td>47.9</td>
<td>47.0</td>
<td>48.4</td>
<td>41.9</td>
</tr>
</tbody>
</table>

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.