Our approach bypasses labor-intensive labeling and generates images annotation effort to collect and label a large number of training sources. Conventional object detection requires a high-cost manual labeling effort to collect and label a large number of training images. Two sets of virtual images were generated from random real ImageNet images; Our approach bypasses labor-intensive labeling and generates models directly from 3D models downloaded from the web.

**Virtual Images**

Two sets of virtual images were generated in 3ds Max: Virtual: background and texture from random real ImageNet images; Virtual-Gray: uniform gray texture with white background.

**RESULTS**

<table>
<thead>
<tr>
<th></th>
<th>Virtual</th>
<th>Virtual-Gray</th>
<th>Amazon</th>
<th>DSLR</th>
<th>PASCAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtual</td>
<td>30.8 (0.4)</td>
<td>16.5 (0.6)</td>
<td>24.1 (0.6)</td>
<td>28.7 (0.2)</td>
<td>10.7 (0.9)</td>
</tr>
<tr>
<td>Virtual-Gray</td>
<td>32.3 (0.6)</td>
<td>32.3 (0.6)</td>
<td>27.3 (0.6)</td>
<td>32.7 (0.6)</td>
<td>17.9 (0.7)</td>
</tr>
<tr>
<td>Amazon</td>
<td>39.9 (0.6)</td>
<td>30.0 (2.0)</td>
<td>39.2 (0.4)</td>
<td>37.9 (0.4)</td>
<td>18.8 (0.4)</td>
</tr>
<tr>
<td>DSLR</td>
<td>88.2 (10.2)</td>
<td>62.1 (1.0)</td>
<td>68.1 (0.6)</td>
<td>68.5 (0.1)</td>
<td>37.7 (0.5)</td>
</tr>
</tbody>
</table>

**3D MODELS**

- Collected models for 20 objects in “Office” dataset
- Selected 2 models per category manually from first page of results
- 15 poses per model were generated by randomly rotating the original model from 0 to 20 degrees in each of the three axes

**OVERVIEW**

- Conventional object detection requires a high-cost manual annotation effort to collect and label a large number of training images
- Our approach bypasses labor-intensive labeling and generates models directly from 3D models downloaded from the web

- Discriminative scoring function $f_c$ for image $I$ and some region $b$
  
  $$f_c(I,b) = w^T \phi(I,b)$$

- Keep regions with high score
- In our case, $f_c$ is learned via Linear Discriminant Analysis (LDA)
  
  $$w = \Sigma^{-1} (\mu - \mu_t)$$

- Features are HOG (could use others)

**EFFECT OF MISMATCHED STATISTICS**

Mean bicycle de-correlated with mismatched-domain covariance(left) vs. with same domain covariance(right).

**MAIN IDEA**

(a) Applying a linear classifier $w$ learned by LDA to source data $x$ is equivalent to (b) applying classifier $w + \Sigma_s^{-1}w$ to de-correlated point $\Sigma_s^{-1}x$.

(c) However, target point $u$ may still be correlated after $\Sigma_t^{-1}u$, hurting performance.

(d) Our method uses target-specific covariance $T$ to obtain properly de-correlated $\hat{w}$

If $T$ is target covariance, $S$ is source covariance, then

$$f_c(u) = w^T u = \frac{1}{2} (\mu - \mu_t)$$

Also, if source and target are the same, this reduces to $\Sigma_s^{-1}$. Note, this corresponds to different whitening operation $\Sigma_s^{-1}$.

**SAMPLE MODELS**

- Synthetic vs. Real

**CONCLUSION**

This paper demonstrates that virtual data rendered from freely available 3D models could be a promising new way to train object detectors on a large scale. In our experiments, detectors trained on virtual data and adapted to real-image statistics perform comparably to detectors trained on real image datasets, including ImageNet. Interestingly, our results showed that non-photorealistic data works just as well as attempts to render more realistic images. The objects in our evaluation were mostly rigid man-made objects; in future work we plan to include more non-rigid objects and more categories.

**ACKNOWLEDGMENTS**

This research was supported by NSF award #1212928 and by DARPA.