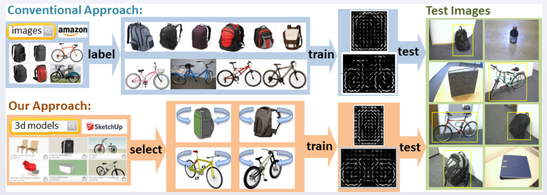
From Virtual to Reality: Fast Adaptation of Virtual Object Detectors to Real Domains

Baochen Sun, Kate Saenko

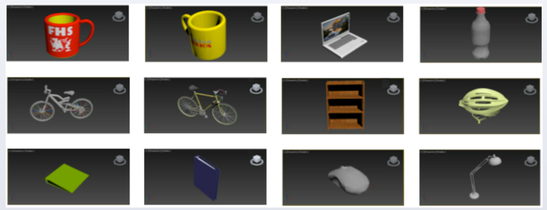
University of Massachusetts Lowell

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

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

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.

DETECTION APPROACH

- Discriminative scoring function f_w for image I and some region b

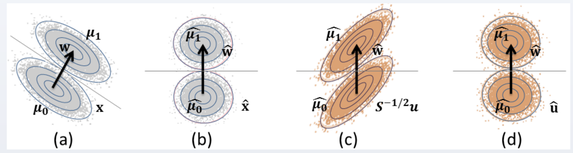
$$f_w(I,b) = w^T \phi(I,b)$$

- Keep regions with high score
- In our case, f_w is learned via Linear Discriminant Analysis (LDA)

$$w = \Sigma^{-1}(\mu_1 - \mu_0)$$

- Features are HOG (could use others)

MAIN IDEA



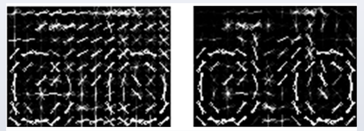
(a) Applying a linear classifier w learned by LDA to source data x is equivalent to (b) applying classifier $\hat{w} = \Sigma^{-1/2} w$ to de-correlated points $\Sigma^{-1/2} x$. (c) However, target points u may still be correlated after $\Sigma^{-1/2} \mu$, hurting performance. (d) Our method uses target-specific covariance T to obtain properly de-correlated $\hat{u}$.

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

$$\begin{aligned} f_{\hat{\mathbf{w}}}(\mathbf{u}) &= \hat{\mathbf{w}}^T \hat{\mathbf{u}} \\ &= (\mathbf{S}^{-1/2}(\mu_1 - \mu_0))^T (\mathbf{T}^{-1/2}\mathbf{u}) \\ &= ((\mathbf{T}^{-1/2})^T \mathbf{S}^{-1/2}(\mu_1 - \mu_0))^T \mathbf{u} \end{aligned}$$

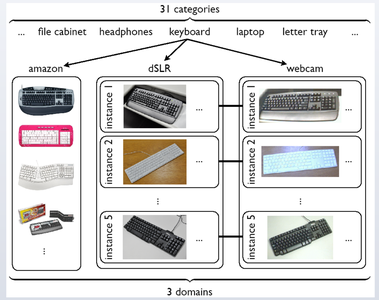
Note, this corresponds to different whitening operation $(T^{-1/2})^T (S^{-1/2})$
Also, if source and target are the same, this reduces to S^{-1}

EFFECT OF MISMATCHED STATISTICS



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

DADASET



Training

Domain	No. Train
Amazon	20
DSLR	8
Virtual(-Gray)	30
Imagenet	150-2000

Test

Webcam

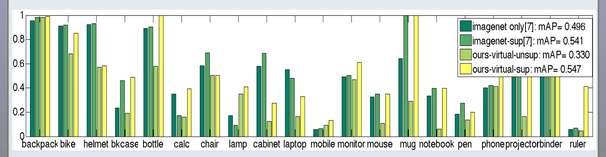
RESULTS

	Virtual	Virtual-Gray	Amazon	DSLR	PASCAL
Virtual	30.8 (*0.1*)	16.5 (*1.0*)	24.1 (*0.6*)	28.3 (*0.2*)	10.7 (*0.5*)
Virtual-Gray	32.3 (*0.6*)	32.3 (*0.5*)	27.3 (*0.8*)	32.7 (*0.6*)	17.9 (*0.7*)
Amazon	39.9 (*0.4*)	30.0 (*1.0*)	39.2 (*0.4*)	37.9 (*0.4*)	18.6 (*0.6*)
DSLR	68.2 (*0.2*)	62.1 (*1.0*)	68.1 (*0.6*)	66.5 (*0.1*)	37.7 (*0.5*)

MAP of detectors trained on positive examples from each row's source domain and background statistics from each column's domain. The average distance between each set of background statistics(each column) to the true source(each row) and target(webcam) statistics is shown in parentheses.

Source	Source-only [10]	UnsupAdapt-Ours	SupAdapt [7]	SupAdapt-Ours
Virtual	10.7	27.9	30.7	45.2
Virtual-Gray	17.9	33.0	35.0	54.7
Amazon	18.6	38.9	35.8	53.0
DSLR	37.7	67.1	42.9	71.4

Comparison of the source-only [10] and supervised-adapted model of [7] with our unsupervised-adapted and supervised adapted models. Mean AP across categories is reported on the webcam test data, using different source domains for training.

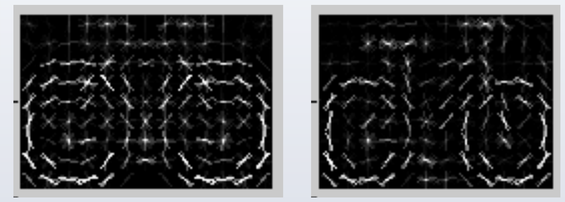


Comparison of unsupervised and supervised adaptation of virtual detectors using our method with the results of training on ImageNet and supervised adaptation from ImageNet reported in [7]. Our supervised-adapted detectors achieve comparable performance despite not using any real source training data, and using only 3 positive images for adaptation, and even outperform ImageNet significantly for several categories (c.f. *ruler*).

[7] Daniel Goehring, Judy Hoffman, Erik Rodner, Kate Saenko, and Trevor Darrell. Interactive adaptation of real-time object detectors. In International Conference on Robotics and Automation (ICRA), 2014.

[10] Bharath Hariharan, Jitendra Malik, and Deva Ramanan. Discriminative decorrelation for clustering and classification. In Computer Vision–ECCV 2012.

SAMPLE MODELS



Synthetic vs. Real

Synthetic "bike" performs better on Webcam target images.

SAMPLE DETECTIONS

Sample detections of our unsupervised adaptation approach of DSLR domain.

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.