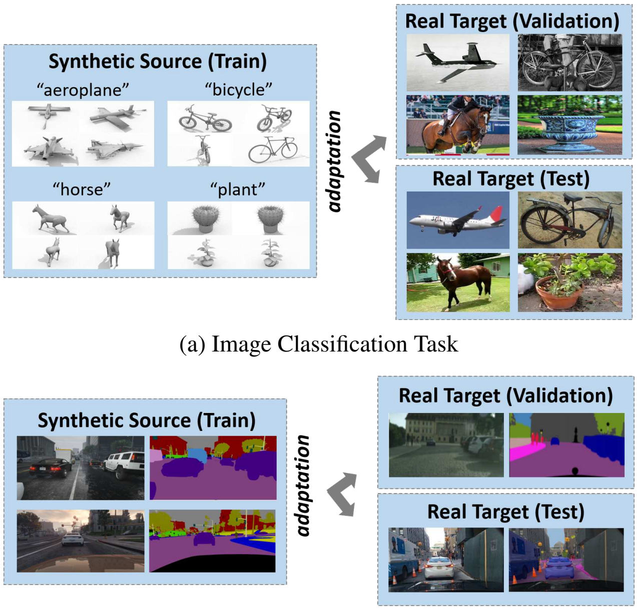
VisDA: A Synthetic-to-Real Benchmark for Visual Domain Adaptation

Xingchao Peng[1], Ben Usman[1], Neela Kaushik[1], Dequan Wang[2], Judy Hoffman[2], and Kate Saenko[1]
xpeng,usmn,nkaushik,saenko@bu.edu, jhoffman,dqwang@eecs.berkeley.edu
[1]Department of Computer Science, Boston University
[2]EECS, University of California Berkeley

1. Introduction

The success of machine learning methods on visual recognition tasks is highly dependent on access to large labeled datasets. However, real training images are expensive to collect and annotate for both computer vision and robotic applications. The synthetic images are easy to generate but model performance often drops significantly on data from a new deployment domain, a problem known as *dataset shift*, or *dataset bias*. Changes in the visual domain can include lighting, camera pose and background variation, as well as general changes in how the image data is collected. While this problem has been studied extensively in the *domain adaptation* literature, progress has been limited by the lack of large-scale challenge benchmarks.

Several benchmark datasets have been collected and used to evaluate visual domain adaptation, most notable are summarized in Table 1. Majority of popular benchmarks lack task diversity: the most common cross-domain datasets focus on the image classification task, i.e. digits of different styles, objects [16] or faces [18] under varying conditions. One issue about these benchmark is the small scale. Modern computer vision methods require a lot of training data, while cross-domain datasets such as Office Dataset [16] only contain several hundred of images. Besides the small size of these benchmarks, another problem is the relatively small domain shifts, such as the shift between two different sensors (DSLR vs Webcam in the Office dataset [16]). Other tasks such as detection [12], structure prediction [15, 5, 14] and sequence labeling [6] have been relatively overlooked.



(a) Image Classification Task

(b) Semantic Image Segmentation Task

Figure 1: (Best viewed in color) The VisDA dataset aims to test models ability to perform *unsupervised domain adaptation*, i.e. to transfer knowledge from a large labeled source domain to an unlabeled target domain. It contains a challenging simulation-to-real domain shift and consists of two tasks: (a) classification and (b) semantic segmentation. For each task we provide data from *three distinct* domains: train (source), validation (target) and test (target), therefore challenging domain adaptation methods' ability to perform well out-of-the-box on unseen domains without manual hyper-parameters tuning.

We present the Synthetic-to-Real Visual Domain Adaptation (VisDA) Benchmark , a large-scale testbed for unsupervised domain adaptation across visual domains. As showed in Figure 1, the VisDA dataset is focused on the simulation-to-reality shift and has two associated tasks: image classification and image segmentation. The goal in both tracks is to first train a model on simulated, synthetic data in the source domain and then adapt it to perform well on real image data in the unlabeled test domain. Our dataset is the largest one to date for cross-domain object classification, with over 280K images across 12 categories in the combined training, validation and testing domains. The image segmentation dataset is also large-scale with over 30K images across 18 categories in the three domains. We compare VisDA to existing cross-domain adaptation datasets and provide a baseline performance analysis using various

OBJECT CLASSIFICATION

Dataset	Examples	Classes	Domains
COIL20 [11]	1,440	20	1 (tools)
Office [16]	1,410	31	3 (office)
Caltech [2]	1,123	10	1 (office)
CAD-office [12]	775	20	1 (office)
Cross-Dataset [22]	70,000+	40	12 (mixed)
VisDA-C	280,157	12	3 (mixed)

SEMANTIC SEGMENTATION

Dataset	Examples	Classes	Domains
SYNTHIA-subset [15]	9,400	12	1 (city)
CityScapes [5]	5,000	34	1 (city)
GTA5 [14]	24,966	18	1 (city)
VisDA-S	31,466	18	3 (city)

Table 1: Comparison of VisDA to existing cross-domain datasets used for domain adaptation experiments, with corresponding numbers of classes and domains. Datasets that share object categories can be combined to form cross-domain benchmarks.

domain adaptation models that are currently popular in the field.

The VisDA dataset focuses on the domain shift from simulated to real imagery–a challenging shift that has many practical applications in robotics and computer vision. This type of "synthetic-to-real" domain shift is important in many real-world situations when labeled imagery is difficult or expensive to collect (autonomous decision making in robotics, medical imaging, etc.), whereas synthetic rendering pipeline can produce virtually infinite amounts of labeled data. For this reason we generated the largest cross-domain synthetic-to-real object classification dataset to date with over 280K images in the combined training, validation and testing sets. For the semantic segmentation track we augmented existing datasets for a total of approximately 30k images across three domains.

2. VisDA-C: Classification Dataset

The *VisDA Classification (VisDA-C)* dataset provides a large-scale testbed for studying unsupervised domain adaptation in image classification. The dataset contains three splits (domains), each with the same 12 object categories:

- **training domain (source):** synthetic renderings of 3D models from different angles and with different lighting conditions,
- **validation domain (target):** a real-image domain consisting of images cropped from the Microsoft COCO dataset [9],
- **testing domain (target):** a real-image domain consisting of images cropped from the Youtube Bounding Box dataset [13]

We use different target domains for the validation and test splits to prevent hyper-parameter tuning on the test data.

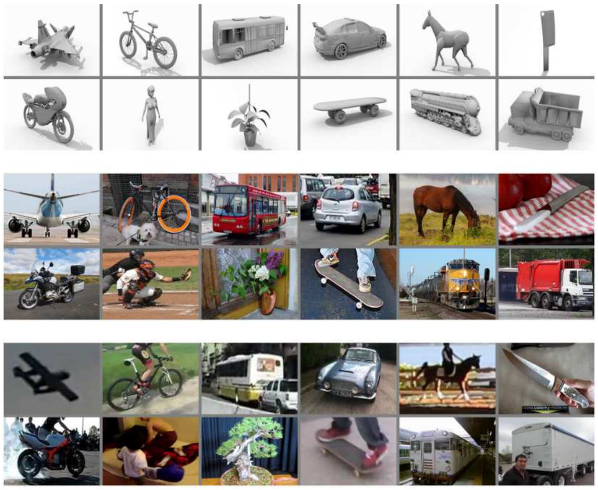


Figure 2: Sample images from the VISDA-C dataset. The top group shows synthetically rendered images (source domain), the middle group shows objects cropped from COCO dataset [9] using their bounding boxes (validation target domain), and the bottom group shows similarly cropped images from YouTube-BB dataset [13] (test target domain).

Unsupervised domain adaptation is usually done in a *transductive* manner, meaning that unlabeled test data is actively used to train the model. However, it is not possible to tune hyper-parameter on the test data, since it has no labels. Despite this fact, the lack of established validation sets often leads to poor experimental protocols where the labeled test set is used for this purpose. In our benchmark, we provide a validation set to mimic the more realistic deployment scenario where the target domain is unknown at training time and test labels are not available for hyper-parameter tuning. This setup also discourages algorithms that are designed to handle a specific target domain. It is important to mention that the validation and test sets are *different* domains, so over-tuning to one can potentially degrade performance on another.

Training Domain: Synthetic Dataset The synthetic dataset was generated by rendering 3D models of the same object categories as in the real data from different angles and under different lighting conditions. We obtained 1,907 models in total and generated 152,397 synthetic images. We used four main sources of models that are indicated with a *sec* prefix of the corresponding image filename. These four sources include manually chosen subsets of ShapenetCore [3], NTU 3D [4], SHREC 2010 [23] with some labels retrieved from TSB [21] and our own collection of 3D CAD models from 3D Warehouse SketchUp.

We used twenty different camera yaw and pitch combinations with four different light directions per model. The lighting setup consists of ambient and sun light sources in 1:3 proportion. Objects were rotated, scaled and translated

Training Domain (CAD-synthetic) → Validation Domain (MS COCO)

Method	Train	Test	aero	bike	bus	car	horse	knife	mbike	person	plant	skbrd	train	truck	Mean	Source	Gain
DAN [10]	syn	real	71.0	47.4	67.3	31.9	61.4	49.9	72.1	36.1	64.7	28	70.6	19	51.62	28.12	83.6%
D-CORAL [20]	syn	real	76.5	31.8	60.2	35.3	45.7	48.4	55	28.9	56.4	28.2	60.9	19.1	45.53	28.12	61.91%
Source (AlexNet)	syn	real	53.5	3.7	50.1	52.2	27.9	14.9	27.6	2.9	25.8	10.5	64.4	3.9	28.12	-	-
Oracle (AlexNet)	syn	syn	100	100	99.8	99.9	100	99.9	99.8	100	100	100	99.9	99.7	99.92	-	-
Oracle (AlexNet)	real	real	94.9	83.2	83.1	86.5	93.9	91.8	90.9	86.6	94.9	88.9	87	65.4	87.26	-	-

Training Domain (CAD-synthetic) → Testing Domain (YT-BB)

Method	Train	Test	aero	bike	bus	car	horse	knife	mbike	person	plant	skbrd	train	truck	Mean	Source	Gain
DAN	syn	real	55.4	18.4	59.9	68.6	55.2	41.4	63.4	30.3	78.8	23.0	62.8	40.1	49.78	30.81	61.57%
D-CORAL	syn	real	62.5	21.7	66.4	64.7	31.1	36.6	54.3	24.9	73.8	30.0	43.4	34.1	45.29	30.81	47.0%
Source (AlexNet)	syn	real	46.5	0.8	59.2	82.7	21.0	14.4	23.2	1.0	46.1	17.2	47.8	9.8	30.81	-	-
Oracle (AlexNet)	real	real	94.5	84.4	90.1	95.5	93.2	95.1	90.4	90.1	95.7	89.5	94.6	91.8	92.08	-	-
Oracle (ResNext)	real	real	96.2	89.3	92.8	98.3	94.8	95.7	90.7	92.0	95.9	86.0	94.9	93.5	93.40	-	-

Table 2: **Baseline results for the classification track**. We show per-category accuracy for models trained using various adaptation methods. The top table reports performance of models adapted to *validation* domain of VisDA-C. The bottom one reports adaptation performance for to the *test* domain. First column indicates either the method used for adaptation, or a special setup (a source only case with no adaptation; a classifier trained using oracle label values). The tables show domain adaptation algorithms (DAN [10] and D-CORAL [20]) can improve the results by roughly 20 percent. Gain column indicates relative improvement over source model.

to match the floor plane, duplicate faces and vertices were removed, and the camera was automatically positioned to capture the entire object with a margin around it. For textured models, we also rendered their un-textured versions with a plain grey albedo. In total, we generated 152,397 synthetic images to form the synthetic source domain.

Validation Domain: MS COCO. The validation dataset for the classification track is built from the Microsoft COCO [9] *Training* and *Validation* splits. In total, the MS COCO dataset contains 174,011 images. We used annotations provided by the COCO dataset to find and crop relevant object in each image. All images were padded by retaining an additional ~50% of its cropped height and width (i.e. by dividing the height and width by $\sqrt{2}$). Padded image patches whose height or width was under 70 pixels were excluded to avoid extreme image transformations on later stages. In total, we collected 55,388 object images that fall into the chosen twelve categories. We took all images from each of twelve categories with the exception of the "person" category, which was reduced to 4,000 images in order to balance the overall number of images per category (the original "person" category has more than 120k images).

Testing Domain: YouTube Bounding Boxes. Due to the overlap in object category labels with the other two domains, we chose the YouTube Bounding Boxes (YT-BB) dataset [13] to construct the test domain. Compared to the validation domain (MS COCO), the image resolution in YT-BB is much lower, because they are frames extracted from YouTube videos. The original YT-BB dataset contains segments extracted from 240,000 videos and approximately 5.6 million bounding box annotations for 23 categories of tracked objects. We extracted 72,372 frame crops that fall into one of our twelve categories and satisfy the size constraints.

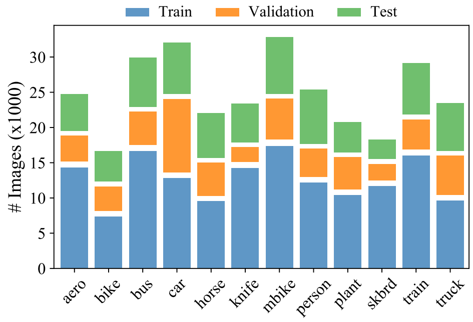


Table 3: Number of images per category in VisDA Classification training, validation and testing domains.

Baseline experiments We evaluate two existing domain adaptation algorithms as baselines. ***DAN*** (Deep Adaptation Network) [10] learns transferable features by training deep models with Maximum Mean Discrepancy [17] loss to align the feature distribution of source domain to target domain. In our implementation, the network architecture of ***DAN*** is extended from AlexNet [8], which consists of 5 convolutional layers (*conv1 - conv5*) and 3 fully connected layers (*fc6 - fc8*) and ***Deep CORAL*** (Deep Correlation Alignment) [19] performs deep model adaptation by matching the second-order statistics of feature distributions. The domain discrepancy is then defined as the squared Frobenius norm $d(S,T) = \|\mathrm{Cov}_S - \mathrm{Cov}_T\|_F^2$, where $\mathrm{Cov}_S, \mathrm{Cov}_T$ are the covariance matrices of feature vectors from the source and target domain, respectively.

Baseline Results. Baseline results on the validation domain for classification are shown in Table 2. "Oracle" or in-domain AlexNet performance for training and testing on the synthetic domain reaches 99.92% accuracy, and training and testing on the real validation domain leads to 87.62%. This supervised learning performance provides a loose upper bound for our adaptation algorithms. As far as unadapted source-only results on the validation dataset,

GTA → CityScapes Validation Domain

Method	road	sidewalk	building	wall	fence	pole	t light	t sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
Source (Dilation F.E.)	30.6	21.2	44.8	10.1	4.4	15.4	12.4	1.7	75.1	13.5	58.1	38.0	0.2	67.5	9.4	5.0	0.0	0.0	0.0	21.4
Oracle (Dilation F.E.)	96.2	76.0	88.4	32.5	46.4	53.5	52.0	68.7	88.6	46.6	91.0	74.8	46.0	90.5	46.9	58.0	44.7	45.2	70.3	64.0

GTA → Nexar Test Domain

Method	road	sidewalk	building	wall	fence	pole	t light	t sign	veg	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
Source (Dilation F.E.)	40.7	19.2	42.3	4.2	20.0	21.8	26.0	13.4	68.0	19.6	84.7	32.4	5.8	59.0	10.3	9.8	1.6	13.5	0.0	25.9
FCN-in-the-wild [7]	57.8	20.1	51.0	6.5	14.1	20.4	26.7	13.7	66.1	22.2	88.9	34.1	13.2	63.2	10.2	7.1	2.0	18.7	0.0	28.2

Table 4: **Baseline results for the segmentation track**. The top shows IoU results from [7] for the source Dilation Front End model and its oracle performance on the validation CityScapes domain. The domain adaptation method presented in [7] achieves 27.1 mIoU. The bottom evaluates the same source and adapted models on the test Nexar domain, and shows results obtained by the top three challenge teams.

AlexNet trained on the synthetic source domain and tested on the real domain obtains 28.12% accuracy, a significant drop from in-domain performance. This provides a measure of how much the domain shift affects the model. Among the tested domain adaptation algorithms, Deep CORAL improves the cross-domain performance from 28.12% to 45.53% and DAN further boosts the result to 51.62%. While their overall performance is not at the level of in-domain training, they achieve large relative improvements over the base model through unsupervised domain adaptation, improving it by 83.6% and 61.9% respectively.

In-domain oracle and source-only performance of AlexNet was similar on the test dataset to the validation dataset. Oracle performance of AlexNet is 92.08% and ResNext-152 improves the result to 93.40%. Source AlexNet achieves 30.81% mean accuracy, and DAN and Deep CORAL improve the result to 49.78% and 45.29%, respectively. As a base model, AlexNet has relatively low performance due to its simpler architecture, compared to more recent CNNs. However, the relative improvement of domain adaptation algorithms (*i.e.* DAN and Deep CORAL) is still large.

3. VisDA-S: Semantic Segmentation

The goal of our *VisDA2017 Segmentation (VisDA-S)* benchmark is to test adaptation between synthetic and real dashcam footage for semantic image segmentation. The training data includes pixel-level semantic annotations for 19 classes. We also provide validation and testing data, following the same protocol as for classification:

- **training domain (source):** synthetic dashcam renderings from the GTA5 dataset along with semantic segmentation labels,
- **validation domain (target):** a real-world collection of dashcam images from the CityScapes dataset along with semantic segmentation labels to be used for validating the unsupervised adaptation performance,
- **test domain (target):** a different set of unlabeled, real-world images from the new Nexar dashcam dataset.

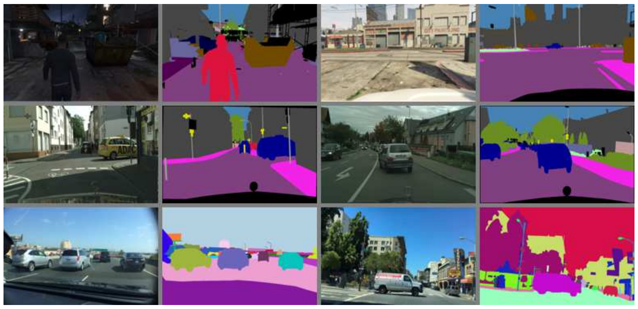


Figure 3: Images in the VisDA-S dataset. The first row shows the synthetic GTA5 images (training domain), the second row shows the images from CityScapes dataset (validation domain), the last row shows the images from Nexar dataset (test domain).

The training and validation domain datasets used here are the same as those used in Hoffman et al (2016) [7] for their work in synthetic to real adaptation in semantic segmentation tasks.

Training Domain: Synthetic GTA5. The images in the segmentation training come from the GTA5 dataset. GTA5 consists of 24,966 high quality labeled frames from the photorealistic, open-world computer game, Grand Theft Auto V (GTA5). The frames are synthesized from a fictional city modeled off of Los Angeles, CA and are in high-resolution, 1914×1052. All semantic segmentation labels used in the GTA5 dataset have a counterpart in the CityScapes category list for adaptation. See Figure 3 for sample training domain data.

Validation Domain: Real CityScapes. Data in the segmentation validation domain comes from the CityScapes dataset. CityScapes contains 5,000 dashcam photos separated by the individual European cities from which they were taken, with a breakdown of 2,975 training, 500 validation and 1,525 test images. Images are in high resolution, 2048 × 1024. In total, the CityScapes dataset has 34 semantic segmentation categories, of which we are interested in the 19 that overlap with the synthetic GTA5 dataset. See Figure 3 for sample validation domain data.

Test Domain: Real DashCam Images. Dashcam photos in the test domain were taken from a dataset recently released by Berkeley Deep Drive and Nexar [1]. They were collected using the Nexar dashcam interface and manually annotated with segmentation labels. We use 1500 images of size 1280 × 720 available with annotations corresponding to the 19 categories matching GTA5 and CityScapes. Note that this data along with the annotations is part of a larger data collection effort by Berkeley Deep Drive (BDD). See Figure 3 for sample test domain data.

Domain Adaptation Algorithms. For details on the domain adaptation algorithms applied to this domain shift, we refer the reader to the original work that performed adaptation from GTA5 (synthetic) to CityScapes (real) in [7]. The authors use the front-end dilated fully convolutional network as the baseline model. The method for domain adaptive semantic segmentation consists of both global and category specific adaptation techniques. Please see section 3 (Fully Convolutional Adaptation Models) in [7] for detailed information about these techniques and their implementation. In all experiments, the Intersection over Union (IoU) evaluation metric is used to determine per-category segmentation performance.

Baseline Results. Please refer to Table 4 and Section 4.2.1 in Hoffman *et al.* [7] for full experimental results and discussion of semantic segmentation performance in GTA5→CityScapes adaptation. Some relevant results are replicated here. In summary, the front-end dilation source achieves a mean IoU (mIoU) of 21.6 over all semantic categories on the val domain, compared to oracle mIoU of 64.0. The adaptation method in [7] improves mIoU to 25.5. A similar performance improvement is seen when adapting the GTA5 model to our challenge test domain.

4. Conclusion

In this paper, we introduce a large scale synthetic-to-real dataset for unsupervised domain adaptation. We highly encourage researchers to work on adaptation methods that *do not rely on the supervised pre-training*, because there are plenty of important domains, such as robotic simulation, that seriously lack labeled data, and therefore might greatly benefit from synthetic-to-real domain adaptation. Large scale synthetic-to-real datasets as the one described in this paper present an experimental setup designed for figuring out how to train these models without supervised pre-training, and therefore working on methods that perform well in practical domains that do not have large labeled datasets using simulated data, which is becoming more and more important these days. As of now the no-pretrain setup poses a substantial challenge for existing domain adaptation methods and solving it would greatly benefit the research community.

References

[1] http://data-bdd.berkeley.edu. 5

[2] F. M. Carlucci, L. Porzi, B. Caputo, E. Ricci, and S. R. Bulò. Autodial: Automatic domain alignment layers. *arXiv preprint arXiv:1704.08082*, 2017. 2

[3] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 2

[4] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, and M. Ouhyoung. On visual similarity based 3d model retrieval. In *Computer graphics forum*, volume 22, pages 223–232. Wiley Online Library, 2003. 2

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[6] A. Graves et al. *Supervised sequence labelling with recurrent neural networks*, volume 385. Springer. 1

[7] J. Hoffman, D. Wang, F. Yu, and T. Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 4, 5

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3

[9] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 2, 3

[10] M. Long and J. Wang. Learning transferable features with deep adaptation networks. *CoRR, abs/1502.02791*, 1:2, 2015. 3

[11] S. A. Nene, S. K. Nayar, H. Murase, et al. Columbia object image library (coil-20). 1996. 2

[12] X. Peng, B. Sun, K. Ali, and K. Saenko. Learning deep object detectors from 3d models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1278–1286, 2015. 1, 2

[13] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. *CoRR*, abs/1702.00824, 2017. 2, 3

[14] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pages 102–118. Springer, 2016. 1, 2

[15] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016. 1, 2

[16] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010*, pages 213–226. Springer, 2010. 1, 2

[17] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013. 3

[18] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 53–58. IEEE, 2002. 1

[19] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. *arXiv preprint arXiv:1511.05547*, 2015. 3

[20] B. Sun and K. Saenko. Deep CORAL: correlation alignment for deep domain adaptation. *CoRR*, abs/1607.01719, 2016. 3

[21] A. Tatsuma, H. Koyanagi, and M. Aono. A large-scale shape benchmark for 3d object retrieval: Toyohashi shape benchmark. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–10. IEEE, 2012. 2

[22] T. Tommasi, T. Tuytelaars, and B. Caputo. A testbed for cross-dataset analysis. *Computer Vision - ECCV 2014 Workshops. ECCV 2014*, 2014. 2

[23] T. P. Vanamali, A. Godil, H. Dutagaci, T. Furuya, Z. Lian, and R. Ohbuchi. Shrec'10 track: Generic 3d warehouse. In *Proceedings of the 3rd Eurographics Conference on 3D Object Retrieval*, 3DOR '10, pages 93–100, Aire-la-Ville, Switzerland, Switzerland, 2010. Eurographics Association. 2