Adaptiope: A Modern Benchmark for Unsupervised Domain Adaptation

Tobias Ringwald
Karlsruhe Institute of Technology
tobias.ringwald@kit.edu

Rainer Stiefelhagen
Karlsruhe Institute of Technology
rainer.stiefelhagen@kit.edu

Abstract

Unsupervised domain adaptation (UDA) deals with the adaptation process of a given source domain with labeled training data to a target domain for which only unannotated data is available. This is a challenging task as the domain shift leads to degraded performance on the target domain data if not addressed. In this paper, we analyze commonly used UDA classification datasets and discover systematic problems with regard to dataset setup, ground truth ambiguity and annotation quality. We manually clean the most popular UDA dataset in the research area (Office-31) and quantify the negative effects of inaccurate annotations through thorough experiments. Based on these insights, we collect the Adaptiope dataset – a large scale, diverse UDA dataset with synthetic, product and real world data – and show that its transfer tasks provide a challenge even when considering recent UDA algorithms. Our datasets are available at `https://gitlab.com/tringwald/adaptiope`.

1. Introduction

Training a modern convolutional neural network (CNN) requires a vast amount of labeled data in order to learn millions of weight parameters. While annotated data is often readily available for a given source domain, this might not be the case for the actual domain of interest (target domain). In the medical area, for example, synthetic images (*e.g.* 3D renderings) could be used for training while inference on real patient data (*e.g.* CT scans) is the actual objective. This would enable the accumulation of larger amounts of training data while also being cheaper and less critical from a privacy perspective. Unfortunately, current CNN classification methods are unable to handle large domain gaps such as synthetic (source) to real (target) transfer tasks, which leads to a degraded performance when trying to apply a source-trained model to the target domain.

Unsupervised domain adaptation (UDA) seeks to address this domain shift problem and strives for accurate predictions on the target domain even in the absence of target labels. Multiple different methods have been presented in recent years that tackle the UDA problem from different directions. At the image level, UDA methods usually try to infer an image-to-image mapping between the domains by the means of image-to-image translation or style transfer, thereby enabling transfer into a common image distribution where straightforward classification is possible [10, 1]. At the feature level, common approaches rely on minimizing divergence or distance measures between the different domains such as the Kullback-Leibler divergence [21] or the maximum mean discrepancy (MMD) [12]. Other works also considered domain discriminative training for enforcing either domain-specific or domain-invariant feature representations [7, 25, 32].

Independent of the chosen approach, key component of every algorithm is its fair and unbiased evaluation on a given target domain after training with both source and (unlabeled) target domain data. For this, the ground truth target domain labels have to be accurate and unambiguous as

(a) bicycle helmet (b) backpack (c) hot glue gun (d) diving fins (e) tyrannosaurus (f) acoustic guitar (g) boxing gloves (h) over-ear headphones

Figure 1: Example images from eight randomly picked classes of our proposed Adaptiope dataset. From left to right: product, real life and synthetic domain.

Dataset	#C	#D	#Images			Avg. Size [px]	Balanced?	Synthetic Domain?
			Min.	Max.	Overall			
Office-31 [26]	31	3	7	100	4,110	418×418	×	×
Syn2Real-C (VisDA 2017) [24]	12	2	2,075	17,360	207,785	355×219	×	✓
Office-Home [28]	65	4	15	99	15,588	727×650	×	×
Digits [15, 8, 6]	10	3	542	6,000	127,291	27×27	×	×
ImageCLEF-DA [3]	12	4	50	50	2,400	443×357	✓	×
Refurbished Office-31 (ours)	31	3	7	100	4,110	556×548	×	×
Modern Office-31 (ours)	31	4	7	100	7,210	781×777	×	✓
Adaptiope (ours)	123	3	100	100	36,900	1,186×1,144	✓	✓

Table 1: Most commonly used unsupervised domain adaptation datasets (see Figure 2) in comparison with our proposed datasets. *Min.* and *Max.* denote the minimum and maximum number of images available concerning all classes in all domains.

otherwise the performance of a given classifier will be misjudged. Unfortunately, the opposite is oftentimes the case with popular UDA datasets. In this paper, we analyze frequently used UDA datasets with regard to annotation quality and uncover systematic problems in the annotation and data selection process. In order to validate our findings, we manually clean the most popular dataset in the research area – Office-31 [26] – and quantify the negative effect of inaccurate annotations through experiments.

Given our insights w.r.t. the annotation processes, we collect a new dataset called Adaptiope, which does not suffer from the aforementioned problems. Adaptiope is one of the biggest, most diverse datasets for unsupervised domain adaptation and offers 123 classes in 3 different domains. Due to the practical relevance for real world application, we also focus on the difficult synthetic to real UDA task: Adaptiope contains a large scale synthetic domain with multiple hundred different 3D models in addition to a product image and real life domain. Example images are shown in Figure 1. To summarize, our contributions are as follows:

- We analyze commonly used UDA datasets and uncover systematic problems w.r.t. annotation quality.
- We clean the most popular Office-31 [26] dataset and validate our findings through thorough experiments.
- We propose Adaptiope, a large scale UDA dataset with 123 diverse classes that offers a synthetic, product and real life domain.
- Based on Adaptiope's synthetic domain and our cleaned Office-31 version, we construct Modern Office-31 – an updated Office-31 version.

2. Related Work

Methods. In recent years, many different domain adaptation methods have been presented. At the feature level, discriminative training has been used for a long time: One of the first uses was proposed by Ganin *et al*. [7, 8] in their domain-adversarial neural network (DANN). DANN is based on a domain classifier that – in conjunction with a novel gradient-reversal layer (RevGrad) – enforces domain-invariant representations at the feature level. Xie *et al*. [30] later picked up this idea in their moving semantic transfer network (MSTN), which complements the domain-discriminative training with a centroid-based alignment. Both DANN and MSTN were chosen as base architecture for the robust spherical domain adaptation (RSDA) method [9] which extends these approaches with a robust pseudo-label loss and a spherical classifier. The recently proposed SymNets [32] by Zhang *et al*. extend the idea of discriminative training to a two-level adversarial training setup that aims at aligning both joint distributions of features and categories. Furthermore, they use a symmetric design of source and target domain classifiers with an additional shared classifier stage.

Instead of discriminative training, distribution distance and divergence measures have been successfully applied for unsupervised domain adaptation and proved to be effective in aligning the different distributions of source and target domain features. For example, Meng *et al*. [21] used the Kullback-Leibler divergence while Long *et al*. [19] employ the Maximum Mean Discrepancy (MMD) to achieve domain adaptation. Recently, this MMD-based approach was extended by Kang *et al*. [12] in their Contrastive Adaptation Network (CAN). Here, inter-class and intra-class MMD is applied in their proposed CDD loss during an iterative clustering and training regime, which can successfully align the source and target domain features.

Datasets. Image classification is one of the most extensively studied problems in computer vision research and thus offers a large amount of different benchmark datasets. For unsupervised domain adaptation, however, it is difficult to create new multi-domain benchmarks by re-using

existing datasets because the class set has to be identical between the domains. Earlier UDA research thus focused on simplistic, common tasks such as handwritten character recognition across domains, for which existing datasets with a shared class set already existed. With regard to digits (ten classes, 0-9), these datasets are usually MNIST [15], MNIST-M [8] and USPS [6] for handwritten digits or SVHN [22] for real life images of house numbers. Due to the small amount of classes and simplicity of the adaptation task, new benchmarks such as the popular Office-31 dataset [26] were created and superseded the almost solved digits transfer tasks. Office-31 contains images of 31 object categories from an everyday office setting in 3 different domains: Amazon (product images) as well as DSLR and Webcam pictures. While providing a more difficult challenge than abovementioned digit tasks, the dataset is still restricted to an Office setting with basic categories such as *pen* and *scissors*. Another issue is the small scale of the dataset: Modern CNN architectures require a lot of training data, while Office only offers approximately 4,000 images. Especially the DSLR domain is problematic in this regard as it contains less than 500 images with the *ruler* class having only 7 images available for training and evaluation. Similar arguments apply to other UDA datasets such as ImageCLEF-DA [3] with only 2,400 images overall and the Office-Home [28] dataset with only 15 images for some classes. With regard to dataset scale, the Syn2Real-C (VisDA 2017) [24] dataset tries to address this problem by offering over 200,000 images in a more difficult synthetic to real UDA setup. Despite of its size, however, VisDA 2017 only contains 12 object categories that are also highly distinguishable (*e.g*. horse and airplane). Recent research [12] showed that results on VisDA 2017 already approach the upper bound of a target domain oracle [24, 12]. Therefore, this dataset also ceases to provide a challenge for recently proposed UDA methods.

For these reasons, one of our main contributions in this work is a novel, large scale UDA dataset with 123 classes and a difficult synthetic to real transfer task that we will show to provide a new challenge for modern unsupervised domain adaptation algorithms.

3. Problems in Common UDA Datasets

In this section, we provide an in-depth analysis on commonly used unsupervised domain adaptation datasets and their problems. For this, we first establish the most popular datasets by inspecting 14 recent CVPR 2020 publications in the UDA research area [31, 13, 27, 4, 23, 16, 11, 9, 20, 29, 17, 14, 18, 5]. The resulting top 5 datasets and their usage percentage are visualized in Figure 2. Further statistics (*e.g*. number of classes and domains) for these datasets are available in Table 1. We observe that most research focuses on Office-31 [26], VisDA 2017 [24], Digits [15, 8, 6], Office-Home [28] and ImageCLEF-DA [3] with the overwhelming majority (78% of analyzed papers) reporting results on the Office-31 dataset. Due to its high popularity, Office-31 will therefore be the main subject of further analysis. Similarities to other datasets such as VisDA 2017 [24] will also be pointed out when applicable.

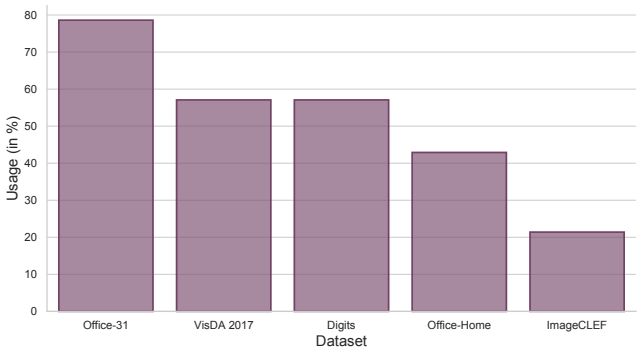


Figure 2: Top 5 commonly used unsupervised domain adaptation datasets for image classification concerning recent CVPR 2020 publications: Office-31 [26], VisDA 2017 [24], Digits [15, 6, 8], Office-Home [28] and ImageCLEF [3].

Office-31 [26] is a classification dataset for UDA tasks and contains 4,110 images of 31 commonplace office objects in three different domains: *Amazon* (product images) with 2,817 images, 498 *DSLR* images of objects from different viewpoints and similarly 795 images from a *Webcam*. Given these 3 domains, the dataset offers 6 possible transfer tasks (A→D, A→W, ...) for which a model is trained with labeled source domain data (A in this case) and adapted to the unlabeled target domain data. Finally, a method's UDA performance is measured as classification accuracy on the target domain for a single transfer task and as an overall average over all transfer tasks (see Section 5).

While images for the DSLR and Webcam domains were manually collected, Office-31's Amazon domain was crawled from the Amazon online store. However, product images often do not correspond to search terms leading to an inherent label noise problem with such an automated data collection process. During our analysis of the Office-31 dataset, we noticed five major problem areas w.r.t. the Amazon domain and also the overall dataset:

Annotation and Data Quality. This was mostly reflected in three different ways: (i) Some images are only vaguely related to the object category, because they were offered as additional services on the product page (extended warranty option icon in Figure 3b). (ii) Some images show an actual object, however, with an incorrect annotation (see *e.g*. laptop labeled as backpack in Figure 3a). (iii) Many images were found to be exact duplicates of other instances in the Amazon domain. Overall, we found 71 duplicate images which constitute about 2.5% of the available data.

Ambiguous Ground Truth. Many images in the dataset show multiple objects of the class set in the same frame, *e.g.* a keyboard, mouse, monitor or speakers in front of a computer (see Figure 3e and 3f). However, only a single label is given. This leads to problems when training with these annotations (using the domain as source) but even more so when evaluating a model on these annotations (using the domain as target), because only one label is considered to be correct. We found this to be especially problematic when the non-annotated objects are dominant with regard to the occupied image area, *e.g.* the *monitor* in Figure 3f while the image is annotated as *computer*.

Class Definitions. Some of the classes follow a different definition between the domains. For the *desk chair* class, the DSLR and Webcam domains show only office chairs (see Figure 3i) while the Amazon domain contains a broader definition of seating options (see Figure 3j). We argue that this does not constitute the classical domain shift problem but instead represents an entirely new class. Similarly, the *bike helmet* class includes motorcycle helmets in the Amazon domain while the DSLR and Webcam domains only depict bicycle helmets.

Dataset Imbalance. Office-31 is highly imbalanced w.r.t. available training and evaluation data. For certain classes and domains, only 7 images are available while there are 100 for other classes (see Table 1). This leads to skewed evaluation results as a model can ignore whole classes without being penalized significantly.

Domain Leakage. Many images in the Amazon domain (product images) leak information w.r.t. other domains, due to showing the object in a real life setting (see Figure 3m) instead of the usual product setting with white background (see Figure 3n). This significantly weakens the *unsupervised* domain adaptation task because source domain labels are available during training.

Unfortunately, these problems are not specific to Office-31 [26] and also occur in other datasets. In VisDA 2017 [24], for example, many images are labeled as *car* although they show other vehicles such as *trucks* (see Figure 3c) or unrecognizable blurry content (see Figure 3d). Many images in VisDA 2017 also show multiple objects, *e.g.* a *motorcycle* or *horse* in front of a *car* (see Figure 3g and 3h) although all 3 objects are part of the label set and the image only has a single annotation. The knife class of the real life domain consists mostly of kitchen knives (see Figure 3k) while the synthetic domain also includes cleavers and axes (see Figure 3l). Furthermore, some synthetic images leak information about the real life domain by including color although those images are supposed to be grayscale (see Figure 3o and 3p). Similar problems can also be observed in Office-Home [28] and ImageCLEF-DA [3].

Recent neural network architectures and training setups have been shown to be moderately robust w.r.t. label errors and low quality images as long as the majority of the *training* data is unaffected. This, however, does not hold when the *evaluation* dataset contains errors. Here, a model could predict an objectively correct label for a given image while being penalized if the prediction does not correspond to an errorenous annotation. For example, a model would be penalized for classifying Figure 3h as *horse* instead of the ground truth *car* annotation. Given N domains $\mathcal{D}$ in an UDA dataset, it is common to construct $N \times (N-1)$ transfer tasks ($\{\mathcal{D}_1 \rightarrow \mathcal{D}_2, \mathcal{D}_2 \rightarrow \mathcal{D}_1, ...\}$), thereby using every domain as source and target domain at least once. Considering this construction process, we hypothesize that a domain $\tilde{\mathcal{D}}$ with aforementioned problems (such as label noise and ambiguous ground truth) will have a high impact on the results when used for evaluation ($* \rightarrow \tilde{\mathcal{D}}$) while being less influential when used for training ($\tilde{\mathcal{D}} \rightarrow *$).

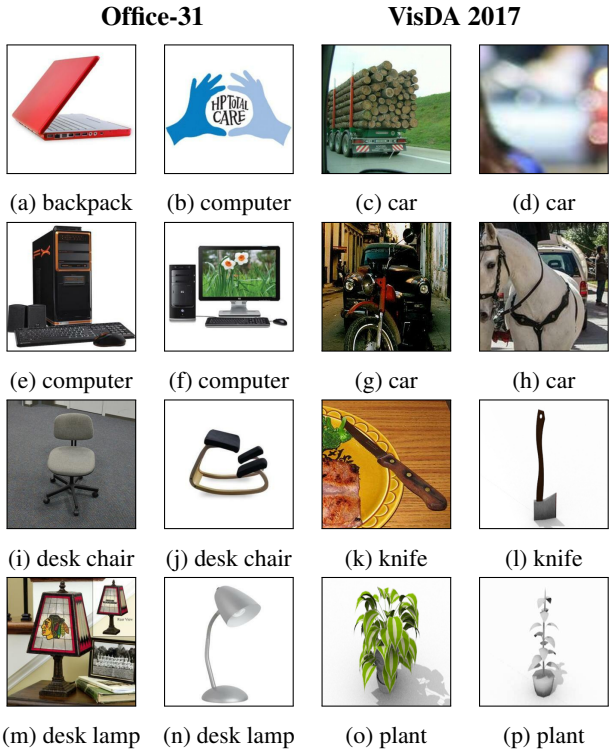


Figure 3: Different problems types in UDA datasets. From top to bottom row: Annotation and data quality, ambiguous ground truth, different class definitions and domain leakage. The image caption denotes the ground truth annotation.

4. Proposed Datasets

4.1. Adaptiope

In this section, we now introduce Adaptiope – our new, diverse UDA classification dataset with synthetic data. Adaptiope contains 36,900 images from 123 classes in the

3 domains product image, real life and synthetic. Example images for eight randomly picked classes are shown in Figure 1. Additionally, major dataset statistics are shown in Table 1. The main goal of Adaptiope was to fix the problems mentioned in Section 3 and to provide a better benchmarking opportunity for UDA algorithms: Unlike other UDA datasets, Adaptiope is balanced – every class contains exactly 100 unique images per domain. This inhibits many problems during evaluation and forces a model to treat every class equally, as incorrect predictions are weighted evenly for all classes. Nevertheless, Adaptiope can also be used to research domain adaptation in imbalanced settings by selectively removing images. On the converse, a dataset such as Office-31 could only be balanced by collecting new data because some classes only contain 7 images.

Images for the product and real life domain were crawled from `amazon.com` and then manually cleaned by human annotators. Here, customer review images were used for the real life domain while the actual article image was used for the product domain. To guarantee a cleanly annotated dataset, we took utmost care to prevent multiple classes from being visible in the same frame by either cropping or replacing the whole image during the collection process. For the synthetic domain, freely available 3D models were collected from sites such as `sketchfab.com` or created by the authors themselves. The models were then ray-traced using the Blender [2] rendering engine. Similar to [24], 3D models were rendered without textures. A total of 615 models (5 per class) were positioned on a shadow catcher surface under a light source and automatically rendered from 20 different camera angles (see supplementary material). For some objects, the camera views were manually filtered, as many classes (*e.g*. a bookcase) would not be distinguishable when seen directly from the top. When compared to the only other UDA dataset with synthetic 3D renderings – VisDA 2017 [24] – we have a multiple times larger synthetic image size (see Table 1) with a resolution of 1080×1080px whereas VisDA provides only 384×216px renderings. VisDA's synthetic renderings also suffer from visual noise caused by undersampling during ray tracing and are based on rather simplistic 3D models (see Figure 4b). We therefore also focused on higher quality 3D models and better rendering settings.

Classes for our dataset were chosen based on the availability and distinguishability of 3D models. Whenever possible, any overlap with classes from ImageNet was avoided in order to prevent leakage from pretrained models. In total, Adaptiope offers 123 diverse classes, which is more than 10 times as many as the only other UDA dataset with synthetic data, VisDA 2017 [24] (12 classes). Moreover, our 123 classes are also a superset of the classes from Office-31, which will be further discussed in Section 4.3. Overall, Adaptiope's classes were designed to be more fine-grained than any other UDA dataset: While many classes in other datasets are trivially distinguishable (*e.g*. airplane and horse in VisDA 2017 [24]), Adaptiope offers much more detailed class definitions such as in-ear vs. over-ear headphones, acoustic vs. electrical guitar and bicycle helmet vs. motorcycle helmet. A full list of all 123 classes is provided in the supplementary material.

(a) Adaptiope's *fighter jet* class (b) VisDA's *airplane* class

Figure 4: Comparison of synthetic image quality in our proposed Adaptiope dataset (left) and VisDA 2017 (right). VisDA offers only simplistic 3D models that exhibit visual noise due to suboptimal rendering settings.

4.2. Refurbished Office-31

In Section 3, we stated major problems with regard to the Office-31 dataset. To quantify the extend of these problems through experiments, we propose a refurbished version of the Office-31 dataset. We apply our insights and points of criticism to manually clean all affected images in Office-31's Amazon domain and replace them with newly gathered product images from the Amazon.com online store. We replaced a total of 834 of the 2,817 images in the Amazon domain. Detailed replacement statistics are available in the supplementary material. During the replacement process, we tried to closely match the distribution of the original images to provide a fair comparison between the dataset versions. This also implies that the total number of images remains unchanged, *i.e*. 2,817 images in the Amazon domain. As the DSLR and Webcam domains were mostly unaffected by abovementioned problems, we include their original versions in our refurbished dataset for a total of 3 domains and 6 transfer tasks. Further information about the refurbished version can be found in Table 1.

4.3. Modern Office-31

Building upon our refurbished Office-31 version and our newly proposed Adaptiope dataset, we also introduce Modern Office-31. Modern Office-31 is supposed to be a drop-in replacement for the popular Office-31 [26] dataset using the same 31 classes. However, a synthetic domain is supplemented to provide a new challenge for UDA algorithms. As Adaptiope's 123 classes are a superset of the 31 classes in Office-31, we easily construct this domain by utilizing Adaptiope's synthetic 3D renderings. This results in a new synthetic domain containing 3,100 images. Furthermore, we adopt the original Office-31 webcam images in order to provide a real life domain. Finally, we employ our cleaned

Method	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
Source only	81.8±0.9	75.1±0.8	62.6±0.9	96.3±0.2	63.9±0.5	99.1±0.3	79.8±0.6
RSDA-DANN	91.1±0.6	92.5±1.2	73.1±1.5	98.8±0.2	75.4±0.5	99.9±0.1	88.5±0.1
RSDA-MSTN	94.9±0.5	94.9±0.7	76.5±0.4	99.1±0.2	78.3±0.4	100.0±0.0	90.6±0.1
SymNet	93.5±0.5	90.8±0.6	74.9±0.2	98.0±0.1	72.0±0.3	99.8±0.0	88.2±0.2
CAN	94.9±0.1	95.2±0.4	76.9±0.6	98.5±0.2	75.1±0.9	99.7±0.2	90.1±0.4

Table 2: Classification accuracy (in %) on the **original Office-31** [26] dataset for four recent UDA methods using ResNet-50.

Method	A_{ref}→D	A_{ref}→W	D→A_{ref}	D→W	W→A_{ref}	W→D	Avg.	ΔAvg.
Source only	79.2±0.6	76.8±1.0	73.5±1.2	96.3±0.2	74.1±0.5	99.1±0.3	83.2±0.6	+3.4
RSDA-DANN	90.9±1.3	91.8±0.5	87.3±0.6	98.8±0.2	90.5±0.9	99.9±0.1	93.2±0.5	+4.7
RSDA-MSTN	93.2±1.0	92.2±0.3	91.7±0.9	99.1±0.2	93.0±0.6	100.0±0.0	94.9±0.4	+4.3
SymNet	92.4±0.4	91.0±0.2	90.6±0.4	98.0±0.1	89.2±0.4	99.8±0.0	93.5±0.1	+5.3
CAN	94.4±0.3	92.8±0.5	92.3±0.8	98.5±0.2	90.9±1.0	99.7±0.2	94.8±0.2	+4.7

Table 3: Classification accuracy (in %) for our proposed **Refurbished Office-31** dataset with a refurbished Amazon domain (denoted as A_{ref}) using ResNet-50. Compared to the original dataset, results improve by a huge margin when using the cleaned Amazon domain for evaluation ($*$→A_{ref}) instead of the noisy original version. ΔAvg. is calculated w.r.t. Table 2.

Amazon domain (see Section 4.2) for an additional product image domain. Overall, our Modern Office-31 dataset provides 7,210 images in the 3 domains Amazon, Synthetic and Webcam resulting in 6 possible transfer tasks. Further statistics can be found in Table 1.

5. Experiments

5.1. Setup

We validate our findings on four different datasets: (1) The original Office-31 [26] dataset with domains Amazon, DSLR and Webcam. (2) Our Refurbished Office-31 dataset with a cleaned Amazon domain as well as the original DSLR and Webcam domains. (3) Our proposed Modern Office-31 dataset consisting of our cleaned Amazon domain, a new synthetic domain and the original webcam domain. (4) Our proposed Adaptiope dataset with domains Product, Real Life and Synthetic. For further information and statistics concerning these datasets, please refer to Section 3 and 4 as well as Table 1.

For evaluation, we select four recent state-of-the-art unsupervised domain adaptation algorithms. Results are obtained using RSDA [9] in conjunction with DANN [7, 8] and MSTN [30], SymNets [32] and CAN [12]. These methods utilize different approaches such as domain discriminative training or distribution-based loss functions and are described further in Section 2.

For every transfer task, we report *source only* results by training a plain model on the source domain and evaluating it on the target domain without applying any domain adaptation techniques. This constitutes a lower bound as no target domain images are used during training. We obtain these results by adding a single linear layer to the respective backbone architecture and train for 4,000 mini-batch iterations using SGD with momentum of 0.9, learning rate 5×10^{-4} and a batch size of 64. In general, every result is reported as mean and standard deviation over 3 runs.

5.2. Results

Refurbished Office-31. We start by reporting baseline results on the original Office-31 [26] dataset for our selected UDA algorithms in Table 2. Our reproduced results closely match the results reported in the respective publications [9, 32, 12] which verifies our basic experimental setup.

We continue by examining our main hypothesis w.r.t. dataset and annotation quality of the original Office-31 [26] dataset. For this, we utilize our Refurbished Office-31 dataset. Please recall from above that this version is similar to the original dataset with the exception of the Amazon domain, which was cleaned of problematic images w.r.t. our insights in Section 3. We evaluate the same four UDA algorithms on the refurbished version and show the results in Table 3. We note three key observations:

Webcam and DSLR. Results for D→W and W→D remain constant as these domains were not modified in the refurbished dataset version.

Amazon as source. All evaluated methods achieve similar results as the original version when using the refurbished Amazon domain as source domain (*i.e.* A_{ref}→ $*$ tasks). This confirms that we closely matched the original image distribution when replacing images. Additionally, cleaning the Amazon domain did not make the domain adaptation

Method	A_{ref}→S	A_{ref}→W	W→A_{ref}	W→S	S→A_{ref}	S→W	Avg.
Source only	51.3±0.4	76.8±1.0	74.1±0.5	51.3±1.0	14.9±3.5	8.0±1.5	46.1±1.3
RSDA-DANN	76.1±0.5	91.8±0.5	90.5±0.9	70.4±1.0	80.8±0.3	83.1±2.8	82.1±0.7
RSDA-MSTN	82.0±0.6	92.2±0.3	93.0±0.6	76.3±0.7	90.0±1.5	86.2±2.9	86.6±0.3
SymNet	65.9±1.0	91.0±0.2	89.2±0.4	56.5±1.4	86.8±2.1	82.2±1.2	78.6±0.6
CAN	79.1±3.0	92.8±0.5	90.9±1.0	77.9±1.6	91.2±0.5	89.7±0.5	86.9±0.7

Table 4: Classification accuracy (in %) for our proposed **Modern Office-31** dataset using ResNet-50.

Backbone	Method	P→R	P→S	R→P	R→S	S→P	S→R	Avg.
ResNet-50	Source only	63.6±0.4	26.7±1.7	85.3±0.3	27.6±0.3	7.6±0.8	2.0±0.2	35.5±0.6
ResNet-50	RSDA-DANN	78.6±0.1	48.5±0.7	90.0±0.0	43.9±0.9	63.2±1.1	37.0±1.2	60.2±0.2
ResNet-50	RSDA-MSTN	73.8±0.2	59.2±0.2	87.5±0.2	50.3±0.5	69.5±0.6	44.6±1.1	64.2±0.4
ResNet-50	SymNet	81.4±0.3	53.1±0.6	92.3±0.2	49.2±1.0	69.6±0.5	44.9±1.5	65.1±0.2
ResNet-50	CAN	81.5±0.3	72.3±0.3	92.2±0.2	69.3±0.9	74.9±1.1	60.7±0.6	75.2±0.2
ResNet-101	Source only	67.7±0.6	35.3±0.9	86.0±0.4	28.2±2.2	8.9±1.0	2.2±0.6	38.1±0.9
ResNet-101	CAN	83.5±0.2	75.7±0.2	93.4±0.2	73.6±0.2	78.3±0.2	63.1±0.8	77.9±0.1
ResNeXt-50_{32x4d}	Source only	68.4±0.5	34.2±1.4	87.7±0.2	29.2±0.3	8.5±4.3	4.7±0.2	38.8±0.4
ResNeXt-50_{32x4d}	CAN	83.8±0.2	73.1±1.1	93.1±0.1	73.0±0.6	79.1±1.4	64.5±1.4	77.8±0.2

Table 5: Classification accuracy (in %) for our proposed **Adaptiope** dataset using different backbone architectures.

task easier. In fact, for some methods, the accuracy slightly decreases regarding those transfer tasks.

Amazon as target. All methods improve by a large margin w.r.t. to the original version when using the refurbished Amazon domain for evaluation (*i.e.* $*$→A_{ref} tasks). This is in accordance with our hypothesis about annotation errors and image quality in the original Office-31 dataset. While UDA algorithms can usually compensate some label noise in the source domain annotations, they cannot account for label errors during evaluation. This is reflected in large gains of up to 17.2% accuracy (for SymNet W→A_{ref} compared to W→A), which are consistent for all four methods.

Overall, the average accuracy increased by over 4% for all methods (see ΔAvg. in Table 3). This is mostly owed to transfer tasks that use the refurbished Amazon domain as target domain, which provides a fairer, noiseless evaluation.

In Figure 5, we provide further analysis in the form of confusion matrices using the RSDA-MSTN algorithm. For the W→A task, the confusion matrix clearly demonstrates the presence of annotation errors and image quality problems in the original Amazon domain as discussed in Section 3. For example, the *computer* class is often confused with the *monitor* class as they are often shown together in the same image while only one label is provided (see *e.g.* Figure 3f). Similarly, the *punchers* and *stapler* classes contain a lot of label noise and unrecognizable images due to which less than 22% of those images were classified correctly – even when considering a SOTA algorithm like RSDA-MSTN. Additionally, we found that the most confused classes correlate with the amount of images that were replaced in our refurbished Amazon domain (see supplementary). For the W→A_{ref} task, a much more distinct diagonal can be observed due to the rectification of annotation errors and replacement of low quality images. Problematic classes – such as *e.g. computer* – profit hugely from our cleaning effort and gain more than 56% in accuracy. The remaining off-diagonal non-zero values constitute the actual domain adaptation challenge and are not caused by incorrect annotations anymore. This effect can also be observed for other UDA algorithms and is shown for CAN [12] in the supplementary material.

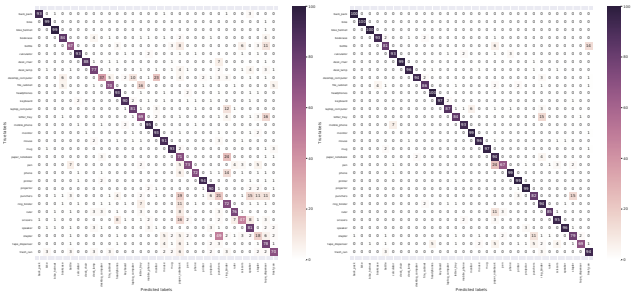
We conclude that our refurbished Amazon domain provides a much cleaner, fairer evaluation ($*$ →A_{ref}) while not weakening the UDA challenge in A_{ref}→ $*$ cases. This confirms our main hypothesis about the negative influence of label noise on the evaluation of Office-31 transfer tasks.

Modern Office-31. Furthermore, we evaluate our *Modern Office-31* dataset as drop-in replacement for Office-31 [26]. This dataset was introduced in Section 4.3 and builds upon our refurbished Amazon domain and synthetic images from Adaptiope. Results for the four selected UDA algorithms are shown in Table 4. Compared to Office-31 [26] and Refurbished Office-31 (see Table 2 and 3), Modern Office-31 provides a much greater challenge to the evaluated algorithms where even the best method (CAN [12]) drops on average 3.2% w.r.t. Office-31 and 7.9% w.r.t. Refurbished Office-31. Note that this challenge is also reflected in the *source only* results, which

are already surpassing 79.8% for Office-31 without even applying any domain adaptation techniques. In this regard, Modern Office-31 is also much more challenging with only 46.1% accuracy after the *source only* training.

Adaptiope. We continue by evaluating our proposed Adaptiope dataset with product, real and synthetic domains. Results are shown in Table 5 for a ResNet-50 backbone (23M parameters). Again, RSDA [9], SymNet [32] and CAN [12] are selected for evaluation. Evidently, Adaptiope's 123 classes provide a much greater challenge than all previously evaluated datasets. A large part of this challenge is posed by the transfer tasks involving the synthetic domain. Here, RSDA and SymNet only achieve less than 60% in the P→S task, less than 51% in the R→S task and less than 70% in the S→P task. Especially the synthetic to real task exhibits a strong domain gap: All evaluated algorithms achieve less than 61% on Adaptiope's challenging S→R transfer task.
Overall, CAN [12] was the best performing method on Adaptiope with an average accuracy of 75.2%. In order to control for the employed backbone architecture, we also evaluate CAN with a ResNet-101 (42M parameters) and ResNeXt-50_{32x4d} (23M parameters) feature extractor stage and show results in Table 5. Our results indicate that Adaptiope's domain adaptation challenge cannot simply be solved by adding more parameters (ResNet-101) or using more complicated architectures (ResNeXt): While the average accuracy marginally increases, Adaptiope's hardest transfer task S→R remains at less than 65% even when almost doubling the number of parameters.

Finally, we conduct upper bound experiments similar to the "oracle" setup in [24]. Here, a single domain is chosen as both source and target. Note that these are not domain adaptation experiments but instead used to determine a theoretical upper bound of any algorithm on that dataset. This can also be seen as a proxy measure for dataset cleanness. The more noise a dataset contains, the more complicated the objective becomes. A model would need an increased number of parameters in order to approximate this more complex function and also fit the noisy examples. Results are reported in Table 6 for different network architectures.
Adaptiope was carefully crafted in order to contain as little noise as possible. This is clearly reflected in our results, which approach 100% accuracy for all transfer tasks and backbones. We also apply the same experimental setup to VisDA 2017 [24], as its synthetic to real task is similar to Adaptiope's S→R task. While VisDA's synthetic domain only has minor issues, we observe subpar results for the real domain. With Alexnet, only 85.7% accuracy can be achieved on the Real→Real task while all transfer tasks of Adaptiope converge towards 100%. Note that this is neither related to dataset nor model size: VisDA's synthetic domain (152,397 images) is three times bigger than its real domain (55,388 images) but can be fitted by all three models – unlike the real domain. Furthermore, not even a large architecture such as ResNet-101 with 42M parameters can successfully fit the real domain of VisDA 2017. Instead, these results are caused by the inherent annotation and image quality problems discussed in Section 3, once again confirming the need for clean UDA datasets such as Adaptiope.



(a) W→A (b) W→A_{ref}

Figure 5: Confusion matrices for the W→A and W→A_{ref} tasks using RSDA-MSTN. The refurbished Amazon domain exhibits a much more distinct diagonal due to the rectification of annotation errors.

Dataset	Task	Alexnet	ResNet-50	ResNet-101
VisDA 2017 [24]	Synth→Synth	98.3±0.2	99.8±0.0	99.9±0.0
VisDA 2017 [24]	Real→Real	85.7±0.4	94.7±0.1	95.8±0.2
Adaptiope (ours)	P→P	99.8±0.1	99.7±0.1	99.9±0.0
Adaptiope (ours)	R→R	99.5±0.1	99.0±0.0	99.8±0.0
Adaptiope (ours)	S→S	99.6±0.1	99.9±0.1	99.9±0.1

Table 6: Upper bound accuracy (in %) for different backbones and datasets. Note that these are **not** domain adaptation tasks but instead proxy measures for dataset cleanness.

6. Conclusion

In this paper, we analyze commonly used image classification datasets for unsupervised domain adaptation. We find systematic problems w.r.t. data and annotation quality in the most popular Office-31 dataset and undertake a cleaning effort following key insights gained from our analysis. We quantify the negative influence of the problematic Amazon domain in Office-31 with the help of our proposed Refurbished Office-31 dataset. Our results indicate that prior research was limited by poor annotation quality and evaluation results strongly improve when correcting label errors and noisy images. Following these insights, we propose our novel Modern Office-31 and Adaptiope datasets in order to provide a better benchmarking opportunity and new challenge for UDA algorithms. Both proposed datasets are publicly available to enable further research on the challenging synthetic to real domain adaptation task.

References

[1] Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2810, 2018.

[2] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, 2019.

[3] Barbara Caputo, Henning Müller, Jesus Martinez-Gomez, Mauricio Villegas, Burak Acar, Novi Patricia, Neda Marvasti, Suzan Üsküdarlı, Roberto Paredes, Miguel Cazorla, et al. Imageclef 2014: Overview and analysis of the results. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 192–211. Springer, 2014.

[4] Zhihong Chen, Chao Chen, Zhaowei Cheng, Boyuan Jiang, Ke Fang, and Xinyu Jin. Selective transfer with reinforced transfer network for partial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12706–12714, 2020.

[5] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12455–12464, 2020.

[6] John S Denker, WR Gardner, Hans Peter Graf, Donnie Henderson, Richard E Howard, W Hubbard, Lawrence D Jackel, Henry S Baird, and Isabelle Guyon. Neural network recognizer for hand-written zip code digits. In *Advances in neural information processing systems*, pages 323–331, 1989.

[7] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1180–1189. JMLR.org, 2015.

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[9] Xiang Gu, Jian Sun, and Zongben Xu. Spherical space domain adaptation with robust pseudo-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9101–9110, 2020.

[10] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018.

[11] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Unsupervised domain adaptation with hierarchical gradient synchronization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4043–4052, 2020.

[12] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019.

[13] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4544–4553, 2020.

[14] Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, R Venkatesh Babu, et al. Towards inheritable models for open-set domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12376–12385, 2020.

[15] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[16] Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13936–13944, 2020.

[17] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2020.

[18] Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X Yu, and Boqing Gong. Open compound domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2020.

[19] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 2200–2207, 2013.

[20] Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9111–9120, 2020.

[21] Zhong Meng, Jinyu Li, Yifan Gong, and Biing-Hwang Juang. Adversarial teacher-student learning for unsupervised domain adaptation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5949–5953. IEEE, 2018.

[22] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[23] Yingwei Pan, Ting Yao, Yehao Li, Chong-Wah Ngo, and Tao Mei. Exploring category-agnostic clusters for open-set domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13867–13875, 2020.

[24] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge, 2017.

[25] Pedro O Pinheiro. Unsupervised domain adaptation with similarity learning. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 8004–8013, 2018.

[26] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[27] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8725–8735, 2020.

[28] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

[29] Sinan Wang, Xinyang Chen, Yunbo Wang, Mingsheng Long, and Jianmin Wang. Progressive adversarial networks for fine-grained domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9213–9222, 2020.

[30] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 5419–5428, 2018.

[31] Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, and Jindong Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4394–4403, 2020.

[32] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5031–5040, 2019.