Unsupervised Domain Adaptation for Object Detection in Cultural Sites

Giovanni Pasqualino, Antonino Furnari, Giovanni Maria Farinella
FPV@IPLAB - Department of Mathematics and Computer Science, University of Catania, Italy
giovanni.pasqualino@studium.unict.it,{furnari, gfarinella}@dmi.unict.it

***Abstract*—The ability to detect objects in cultural sites from the egocentric point of view of the user can enable interesting applications for both the visitors and the manager of the site. Unfortunately, current object detection algorithms have to be trained on large amounts of labeled data, the collection of which is costly and time-consuming. While synthetic data generated from the 3D model of the cultural site can be used to train object detection algorithms, a significant drop in performance is generally observed when such algorithms are deployed to work with real images. In this paper, we consider the problem of unsupervised domain adaptation for object detection in cultural sites. Specifically, we assume the availability of synthetic labeled images and real unlabeled images for training. To study the problem, we propose a dataset containing 75244 synthetic and 2190 real images with annotations for 16 different artworks. We hence investigate different domain adaptation techniques based on image-to-image translation and feature alignment. Our analysis points out that such techniques can be useful to address the domain adaptation issue, while there is still plenty of space for improvement on the proposed dataset. We release the dataset at our web page to encourage research on this challenging topic: https://iplab.dmi.unict.it/EGO-CH-OBJ-ADAPT/.**

I. Introduction

The development of increasingly powerful smart wearable devices allows the exploitation of machine learning and computer vision algorithms to support user applications both in indoor and outdoor environments [1], [2], [3]. For example, in a museum, smart glasses can be used to guide the user in the site and to show additional information related to the observed artworks [4]. At the same time, the analysis of images and videos acquired through these devices can provide insights into the visitors' behavior and preferences, which can be later exploited by the manager of the site to improve the offered services [5]. Such applications require algorithms able to recognize the observed objects not only at the category level (e.g., if an object is a painting and a statue), but also at the instance level (e.g., which particular artworks are in the scene). Object detection algorithms have to be trained on large dataset of labeled images. However, collecting and labeling such data can be very expensive, especially in the case of instance-level object detection in cultural sites. Being able to automate this process can significantly reduce costs and human effort.

A recent work [6] proposed to reduce the cost of creating large datasets of labeled images for visitors' behavior understanding in cultural sites by generating synthetic images with a dedicated tool. The approach proposed in [6] requires the acquisition of a 3D model of the cultural site through a 3D scanner such as Matterport[1]. The 3D model is hence labeled to specify the position of the artworks in the cultural site. A tool is then used to simulate a virtual agent navigating the site and generate synthetic images which are automatically associated to the related ground truth annotations for object detection. This procedure allows to capture images of the artworks in their natural environments. Figure 1 compares an example of a synthetic image acquired using this procedure with a real image acquired in the same cultural site. The synthetic data generated using this procedure can be used to train the object recognition algorithms at the instance level. However, such algorithms tend to have limited performance on real data when trained only on synthetic data, due to the domain shift induced by the two data sources. Recent studies have highlighted that, in the presence of labeled images from the training domain and unlabeled images from the test domain, unsupervised domain adaptation techniques can be used to reduce the domain shift [7], [8], [9], [10], [11], [12], [13].

Fig. 1: The results of a Faster-RCNN model trained on synthetic data on (left) a synthetic image and (right) a real image. As can be noted, the algorithm is not able to detect the artwork in the real image.

In this paper, we consider the problem of unsupervised domain adaptation for object detection in cultural sites. Specifically, we assume the availability of labeled synthetic images and unlabeled real images. Since real images do not need to be manually annotated, this assumption greatly reduces labeling costs. To study the problem, we collected a dataset of real and synthetic images labeled for the object detection problem. The real images have been collected in a cultural site through wearable devices and labeled by humans [14], whereas the synthetic images and their labels have been generated from a 3D model of the same cultural site using the procedure proposed in [6]. Hence, we have analyzed the problem of unsupervised domain adaptation training the models only with

[1]https://matterport.com/

synthetic data and testing them on real data. We considered CycleGAN [15] as a baseline technique for unsupervised domain adaptation and performed experiments with respect to different parameters, such as the employed object detection algorithm (Faster-RCNN [16] or RetinaNet [17]), the direction of image translation (real to synthetic or synthetic to real), and the number of epochs required for training in order to avoid overfitting to the source domain. Additionally, we compared the considered pipeline based on image-to-image translation with DA-Faster RCNN [9], a state of the art approach based on adversarial feature alignment. Our analysis points out that: 1) object detection models trained on synthetic images tend to rapidly overfit to the source domain – hence early stopping can be used as a criterion to reduce the domain gap, 2) RetinaNet is less sensitive to domain shift than Faster RCNN, 3) domain adaptation through image-to-image translation is most effective when models are trained on synthetic images translated to the real domain, 4) domain adaptation approaches based on feature alignment can be combined with image-to-image translation to improve performance. The experiments show that a simple RetinaNet baseline trained on synthetic images transformed to the real domain achieves an mAP of 55.54%, which consists in an improvement of $+41.1\%$ with respect to a naive baseline with no adaptation, $+29.51\%$ with respect to a Faster RCNN model adapted using the same technique, and $+22.34\%$ with respect to the DA-Faster RCNN approach [9] combined with image-to-image translation.

In sum, the contributions of this paper are as follows: 1) we introduce a dataset to study the problem of unsupervised domain adaptation in the context of cultural sites; 2) we benchmark different solutions to address the problem of unsupervised domain adaptation for object detection and provide baseline results on the dataset; 3) we compare the performance of approaches to unsupervised domain adaptation for object detection based on image-to-image translation and feature alignment; 4) we discuss the limits of the considered techniques and give insights into future research directions.

The reminder of the paper is organized as follows. Related works are discussed in Section II. The dataset is presented in Section III. The compared domain adaptation approaches are detailed in Section IV. Results are discussed in Section V. Section VI concludes the paper.

II. Related Work

Our work is related to different lines of research, including the use of egocentric vision in cultural sites, image-to-image translation techniques for domain adaptation, and unsupervised domain adaptation for object detection. The following sections discuss the relevant works belonging to these research lines.

A. Egocentric vision in cultural sites

Previous works focused on the use of wearable devices to improve the fruition of artworks and the user experience in cultural sites [4], [18], [19]. To enable such applications, it is necessary to design algorithms capable of detecting and recognizing the desired objects. The authors of [14] proposed a dataset of egocentric videos acquired using a Microsoft HoloLens device in two cultural sites located in Italy. The authors of [6] proposed a tool to generate automatically labeled synthetic egocentric visual data of a real cultural site. The authors of [20] released an offline system for the automatic detection of visual attention and the identification of salient items. The system uses head mounted cameras and combines multiple state-of-the-art techniques addressing different computer vision tasks such as tracking, image matching and retrieval. In [21], the problem of localizing visitors in a cultural site from egocentric images has been investigated. This can be useful to provide assistance to the visitors and help the manager of the site to understand their behaviour, inferring for instance where they spend more time. The authors of [22] presented and evaluated a method to analyze social images of tourist routes. Our work focuses on the problem of detecting objects in cultural sites. Since collecting large datasets of labeled images is challenging, we investigate techniques to leverage synthetic images for training the algorithms.

B. Domain adaptation

Domain adaptation techniques aim at reducing the domain gap between training and testing images, which are usually referred to as source and target domain. Such domain gap is in generally large when training and test data are acquired with different settings, as it is the case of real and synthetic images. Recent works have investigated the problem of domain adaptation using deep learning. The authors of [23] introduces a framework to tackle domain adaptation using deep neural networks. The authors of [24] proposes to embed images onto two spaces: a domain-invariant content space and a domain-specific attribute space. The authors of [7] introduced an approach which combines discriminative modeling, untied weight sharing and a GAN loss. The authors of [8] proposes a new model which adapts features at both pixel and feature-level without requiring aligned pairs. In [25], it has been presented an approach to unsupervised domain adaptation with a deep architecture that requires only labeled data for the source domain and unlabeld data for the target domain.

C. Image-to-image translation

When the images from the source and target domain are visually different, image-to-image translation techniques can be effectively used to reduce the domain gap. The goal of image-to-image translation algorithms is to transform images from the source domain to make them visually similar to those belonging to the target domain without changing their content. One of the first studies on image-to-image translation has been introduced in [26]. The authors of this work proposed a framework for processing images by examples, called "image analogies". The authors of [27] presented a new approach to image-to-image translation using conditional adversarial networks to learn a mapping between a pair of input-output images. A recent work proposed CycleGAN [15], a method to learn a mapping between two domains in the absence of paired example. Specifically, given a source domain X and a

target domain Y, CycleGAN learns a mapping $G : X \rightarrow Y$ from images belonging to X and Y, even when no relationship between image pairs is known. Since no pairing is required, this approach is particularly useful to perform domain adaptation. In our work, we explore the feasibility of image-to-image translation as an unsupervised domain adaptation method for instance-based object detection in cultural sites.

D. Unsupervised domain adaptation for object detection

Many works on domain adaptation for object detection are based on adversarial paradigm [28] to learn domain-invariant features. An example of this approach is presented in [9]. In that work, adversarial learning is used to align features at both image-level and object instance-level. The authors of [10] proposed a feature alignment algorithm based on strong local alignment and weak global alignment. The authors of [12] mined discriminative regions that are directly pertinent to object detection and focused on aligning them across different domains. The authors of [11] presented a multi-level domain adaptive model to simultaneously align the distributions of local and global features. In our work, we compare feature alignment approaches for domain adaptation to those based on image-to-image translation, in the context of instance-based object detection in cultural sites. We also study the impact of combining the two approaches on the proposed dataset.

III. Dataset

We built the proposed dataset using publicly available tools and data. We used the tool proposed in [6] to generate 75244 synthetic images of 16 artworks from the 3D model of the real indoor cultural site "Galleria Regionale di Palazzo Bellomo", located in Siracusa, Italy[2]. The selected artworks cover a variety of object types which can be found in a museum, including paintings, statues and books (see Figure 2). The 3D model of the site has been acquired using a Matterport 3D scanner[3]. The tool provided in [6] allows to annotate the 3D model of the building to specify the location of each of the artworks of interest. Hence, it is possible to simulate an agent walking inside the cultural site which observe the artworks. Each image collected during the simulated visit is automatically associated to a semantic mask, which allows to obtain bounding box annotations for each image. Figure 2 reports sample images of the 16 artworks present in the dataset of synthetic images.

Real images of the same artworks are taken from the EGO-CH dataset proposed in [14]. This dataset includes videos of 70 subjects visiting two cultural sites. The videos have been acquired using a Microsoft HoloLens device and contain 176999 images with bounding box annotations and over 200 points of interest. The bounding box labels of EGO-CH have been produced manually by human annotators. We consider the subset of EGO-CH collected in the "Galleria Regionale di Palazzo Bellomo" cultural site and select images containing the 16 artworks annotated in the synthetic images. This amounts to 2190 real images in total. Figure 3 reports sample images of the 16 artworks present in the dataset of real images.

To perform the experiments, we split both the real and synthetic images into training and test sets. In particular, we use 1502 real images and 51284 synthetic images for training as well as 688 real images and 23960 synthetic images for test. We make sure that real images from the same video and synthetic images from the same simulated navigation fall entirely either in the training or in the test set. The proposed dataset can be used to study synthetic-to-real domain adaptation for object detection algorithms in the context of cultural sites. To encourage research in this field, we publicly release the dataset at the following url: https://iplab.dmi.unict.it/EGO-CH-OBJ-ADAPT/.

IV. Methods

In this section, we present the compared methods: 1) baseline approaches without adaptation, 2) approaches performing domain adaptation through image-to-image translation, 3) approaches performing domain adaptation through feature alignment and 4) approaches combining feature alignment and image-to-image translation.

A. Baseline approaches without adaptation

These baselines do not include any domain adaptation module. Specifically, we consider three different baselines: 1) the model is trained on synthetic images and tested on synthetic images (no domain shift), 2) the model is trained on synthetic images and tested on real images (no adaptation), 3) the model is trained on real images and tested on real images (an "oracle" model which has access to labels from the target domain). We investigate these baseline to quantify the loss in performance due to the domain shift between synthetic and real images and to produce naive baseline results for unsupervised domain adaptation. We trained two different object detection algorithms, the single-stage object detector RetinaNet [17] and the two-stage object detector Faster RCNN [16]. It is worth noting that most previous works on domain adaptation for object detection concentrated on Faster RCNN [9], [10], [11], [12]. Interestingly, our experiments show that RetinaNet is more robust to the domain shift on the considered dataset.

B. Domain adaptation through image-to-image translation

A popular approach to reduce the domain gap between synthetic and real images is to use image-to-image translation techniques [7], [8], [13], [23]. We consider another set of baselines which use CycleGAN [15] to perform domain adaptation. Specifically, we train a CycleGAN model to perform image-to-image translation between real and synthetic images using training images from both domains. We hence consider two approaches: 1) translating real images to synthetic and 2) translating synthetic images to real. In the first case, we train the object detection models using labeled synthetic images. Inference is performed on real images transformed to the synthetic domain using CycleGAN. In the second case, we train the object detection models on the labeled synthetic

[2]http://www.regione.sicilia.it/beniculturali/palazzobellomo/

[3]https://matterport.com/

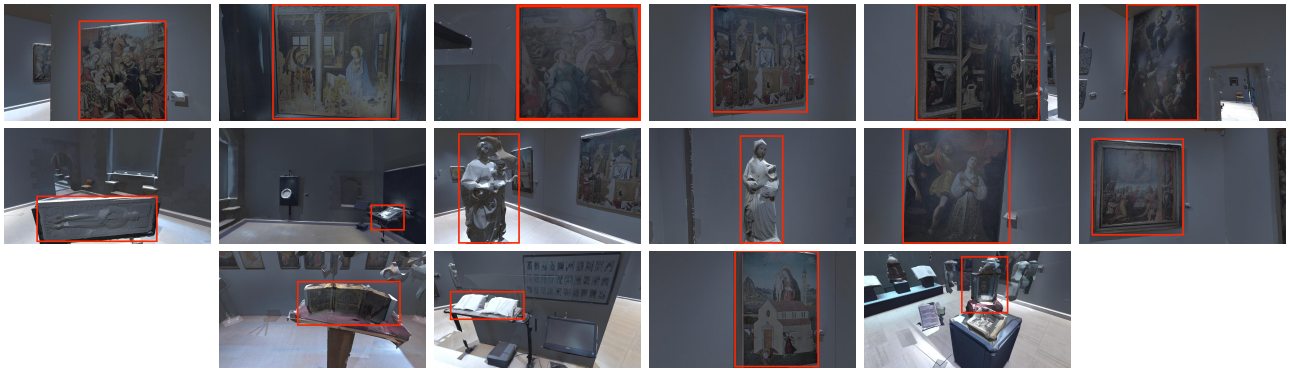

Fig. 2: Sample synthetic images of the 16 artworks of our dataset. See Fig. 3 to compare synthetic vs real images.

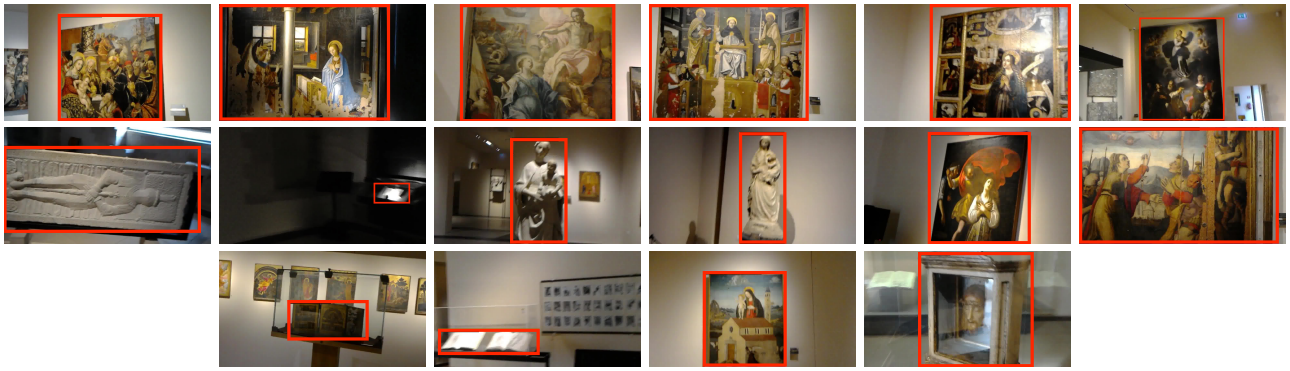

Fig. 3: Sample real images of the 16 artworks of our dataset. See Fig. 2 to compare real vs synthetic images.

TABLE I: Performance of Faster RCNN and RetinaNet when trained and tested on images coming from the same domain.

	mAP	
Model	Synthetic	Real
Faster RCNN	**93.08%**	92.04%
RetinaNet	91.67%	**92.15%**

images transformed to the real domain using CycleGAN. Inference is performed on real test images using the trained model directly.

C. Domain adaptation through feature alignment

Many approaches to unsupervised domain adaptation through object detection use adversarial feature alignment to reduce the domain gap between synthetic and real images [9], [10], [11]. To compare the aforementioned approaches based on image-to-image translation with respect to methods relying on feature alignment, we consider the method proposed in [9], which is based on adversarial feature alignment both at the image and object level. To investigate the benefits of combining feature alignment and image-to-image translation, we also pair this approach with CycleGAN using the same methods discussed in Section IV-B.

V. Experimental Results

In this section, we report and analyze the results obtained by the compared methods on the proposed dataset and discuss the computational times required to train the methods.

A. Baseline approaches without adaptation

Table I reports the results of Faster RCNN and RetinaNet when trained and tested on the source domain. As can be noted, both algorithms achieve very good performance when there is no domain shift between training and test data. The methods have been trained using the Detectron2 [29] implementation[4] for 62K iterations with a batch size of 4 and a learning rate of 0.0002 for the 30K iterations, which is multiplied by 0.1 for the remaining iterations. All the other parameters have been set to default values. These results suggest that, in the absence of domain shift, the problem of instance object detection in cultural sites can be solved if a sufficient amount of data is provided.

Table II reports the performance of the two object detection algorithms when trained on synthetic data and tested on real data. We report the performance obtained on the test set after different amounts of training iterations to study whether the algorithms tend to overfit to the source domain. Note that there is a significant drop in performances of both methods due the domain shift (compare with respect to Table I). The results highlight that models trained for few iterations generalize better than models trained for more iterations. This suggests that, after an initial stage in which the model learns to extract representations useful for both domains, there is

[4]https://github.com/facebookresearch/detectron2

TABLE II: Performance of Faster RCNN and RetinaNet on real images after different amounts of training iterations on synthetic images.

	Training Iterations						
Model	6K	12K	22K	32K	42K	52K	62K
F. RCNN	2.27%	**9.67%**	5.79%	3.58%	3.33%	3.81%	3.62%
RetinaNet	9.83%	**14.44%**	13.22%	12.31%	12.09%	12.44%	11.97%

Fig. 4: Examples of successful (top) and failed (bottom two) transformations from real to synthetic. Each example includes the input real image, the transformed image, and a reference synthetic image containing the same object.

a tendency to overfit to the source domain. Hence, fully trained models tends to extract features which are too specific to the synthetic images and less useful in the domain of real images. Indeed the mAP of Faster RCNN trained for only 12K iterations is 9.67%, which is about 6% better than 3.62% related to the mAP of the same model trained for 62K iterations. A similar trend is observed for RetinaNet, which obtains an mAP of 14.44% after 12K iterations, compared to only 11.97% after 62K iterations. It is worth noting that, when no domain adaptation technique is considered, RetinaNet tends to generalize better than Faster RCNN, achieving an mAP of 14.44%, which is about 5% better than the mAP of obtained by Faster RCNN (9.67%).

B. Domain adaptation through image-to-image translation

Table III reports the results obtained transforming test images from real to synthetic using CycleGAN. We consider the Faster RCNN and RetinaNet trained both for 62K iterations and 12K iterations to assess the influence of overfitting to the source domain in the presence of domain adaptation techniques. We compare results obtained considering CycleGAN models trained for different amounts of epochs with the results obtained without adaptation (N.A. - results reported from Table II). All models significantly improve their performance when CycleGAN is used to transform test images (compare N.A. with any of the results obtained using CycleGAN). RetinaNet obtains better results than Faster RCNN when the models have been trained for 62K iterations (mAP of 34.15% vs 28.25%). Also, RetinaNet trained for 62K iterations

TABLE III: Results obtained transforming real images to synthetic at test time. The models have been trained on synthetic images. N.A. stands for No Adaptation.

		Training epochs for CycleGAN					
Model (iter)	N.A.	10	20	30	40	50	60
F. RCNN (62K)	3.62%	25.16%	25.49%	25.51%	26.68%	27.65%	**28.25%**
RetinaNet (62K)	11.97%	27.30%	32.14%	**34.15%**	32.66%	32.79%	32.82%
F. RCNN (12K)	9.67%	29.93%	32.84%	33.95%	31.45%	**34.19%**	31.58%
RetinaNet (12K)	14.44%	34.51%	35.45%	34.84%	35.34%	**35.76%**	35.74%

TABLE IV: Results obtained training the models on synthetic images transformed to real and tested on real images. N.A. stands for No Adaptation.

		Training epochs for CycleGAN					
Model	N.A.	10	20	30	40	50	60
F. RCNN	9.67%	18.76%	20.92%	21.22%	23.17%	24.45%	**26.03%**
RetinaNet	14.44%	40.13%	44.29%	46.05%	47.89%	49.96%	**55.54%**

achieves the best performance when CycleGAN has been trained for just 30 epochs, while Faster RCNN trained for 62K iterations achieves its best performance after 60 epochs. When considering models trained for 12K iterations (and hence less influenced by overfitting), Faster RCNN and RetinaNet obtain similar performance (34.19% vs 35.76%), with a peak in the case in which CycleGAN has been trained for 50 epochs. These results suggest that Faster RCNN is more sensitive to overfitting to the source domain with respect to RetinaNet. Indeed, using an instance of CycleGAN trained for more epochs or choosing a model less influenced by overfitting greatly affects the performance of Faster RCNN, while the performance of RetinaNet tends to be more stable. The main limitations of this approach are given by the limited ability of the CycleGAN model to perform translation from the real to the synthetic domain. Figure 4 reports three examples of real to synthetic translation. Each example includes the input real image, the transformed image and a reference synthetic image which contains the same object. The top row presents a case in which the translation was successful, whereas the bottom two rows report failure examples. In particular, in the second row the image is not correctly transformed due to light reflection, while in the third row texture is corrupted in the transformed image.

Table IV reports the results obtained training Faster RCNN and RetinaNet on synthetic images transformed to real. We compare the results obtained training CycleGAN for different amounts of epochs with the best results obtained with no adaptation (best results in Table II). Both models outperform the no adaptation baselines (compare N.A. to the other results on the same row in Table IV). Interestingly, Faster RCNN obtains a better performance when the images are transformed from real to synthetic (mAP of 34.19% in Table III versus 26.03% in Table IV). On the contrary, RetinaNet obtains much better performance when trained on images transformed from synthetic to real (55.54% in Table IV vs 35.76% in Table III).

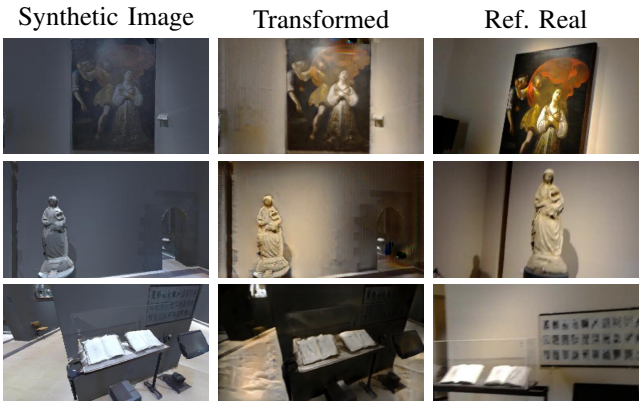

Fig. 5: Examples of successful (top two) and failed (bottom) transformations from synthetic to real. Each example includes the input synthetic image, the transformed image, and a reference real image containing the same object.

This result seems to confirm the observation that RetinaNet is more robust to the domain shift. Although RetinaNet and Faster RCNN achieve the best performance using images translated by CycleGAN trained for 60 epochs, their behavior is expected to remain unchanged even for a greater number of CycleGAN training epochs. Furthermore, as reported in Section V-E on computational analysis, training CycleGAN for a greater number of epochs becomes uninteresting as the training times become too long. The first two rows of Figure 5 show examples of good translation. The last row shows an example of unsuccessful translation, in which textures are corrupted.

C. Domain adaptation through feature alignment

We assess the performance of DA-Faster RCNN [23], an approach to unsupervised domain adaptation for object detection based on adversarial feature alignment. This algorithm can be trained using labeled synthetic images and unlabeled real images. We further combine this approach with two forms of domain adaptation through image-to-image translation: 1) real to synthetic (Real2Syn), and 2) synthetic to real (Syn2Real). Please note that this approach is similar in spirit to previous methods combining image-to-image-translation and feature alignment such as CyCADA [8]. The fist approach transform the unlabeled training set to synthetic during the training and then test the algorithm on real images transformed to synthetic. The second approach consists in training the algorithm with synthetic labeled images transformed to real and unlabeled real images. Results in Table V show that DA-Faster RCNN improves over the baseline Faster RCNN with no adaptation (12.94% versus 9.67% in Table II). DA-Faster RCNN benefits from image-to-image translation, both when images are transformed form real to synthetic at test time (mAP of 19.88% versus 12.94% with no translation) and when images are transformed from synthetic to real at test time (mAP of 32.20% versus 12.94% with no translation).

TABLE V: Results of DA-Faster RCNN combined with two image-to-image translation approaches.

	image-to-image translation		
Model	None	Real2Syn	Syn2Real
DA-Faster RCNN	12.94%	19.88%	**33.20%**

TABLE VI: Summary of the performance of the compared methods.

Object Detector	Adaptation	mAP
Faster RCNN	None	9.67%
RetinaNet	None	14.44%
Faster RCNN	Real2Syn (Test set)	34.19%
RetinaNet	Real2Syn (Test set)	35.76%
Faster RCNN	Syn2Real (labeled Training set)	26.03%
RetinaNet	Syn2Real (labeled Training set)	**55.54%**
DA-Faster RCNN	Feat.Align.	12.94%
DA-Faster RCNN	Feat.Align.+Real2Syn (Test set and Unlabeled Training set)	19.88%
DA-Faster RCNN	Feat.Align.+Syn2Real (labeled Training set)	33.20%

D. Summary table and qualitative results

Table VI summarizes the performance of all the compared methods. The table reports best performance in those cases in which multiple results are available for different training iterations or epochs. The reported numbers can serve as baseline results for the proposed dataset.

Figure 6 reports qualitative results of six of the compared models. Faster RCNN with no adaptation cannot detect objects in many examples, whereas, in some cases, it also shows false positive detections. DA-Faster RCNN with only feature alignment adaptation and RetinaNet with no adaptation detect many more objects but they are still subject to false positives and miss-classifications. Faster RCNN coupled with Syn2Real adaptation correctly detects objects in the "easy" samples (first two rows), with some troubles when there are more objects and occlusions (the other samples). DA-Faster RCNN with feature alignment and Syn2Real adaptation does not achieve accurate detections in the fourth example and struggles with the last two challenging examples. RetinaNet with Syn2Real adaptation detects correctly all the objects in the the first four examples. In the last two examples shown in Figure 6, all methods struggle to detect objects, which highlights that the proposed dataset presents challenging examples, and hence there is still space for improvement.

E. Computational Resources Analysis

Table VII reports the time required to train the models, using a single NVIDIA® Tesla® K80. The best performing method is based on CycleGAN. Unfortunately, training CycleGAN for 60 epochs requires about 61 days on our dataset. Using a method based only on feature alignment requires much less time but involves a significant decrease in performance (12.94%) achieved by DA-Faster RCNN in

Fig. 6: Qualitative results of six of the compared models.

TABLE VII: Training time required for each Neural Networks.

Model	Hours (Days)
RetinaNet (12K iterations)	~ 10 (~ 0.5)
RetinaNet (62K iterations)	~ 65 (~ 3)
Faster RCNN (62K iterations)	~ 131 (~ 5.5)
DA-Faster RCNN	~ 142 (~ 6)
CycleGAN	~ 1470 (~ 61)
CycleGAN + RetinaNet	~ 1535 (~ 64)
CycleGAN + Faster RCNN	~ 1601 (~ 66)
CycleGAN + DA-Faster RCNN	~ 1612 (~ 67)

Table VI vs RetinaNet + CycleGAN, which achieves 55.54% in Table VI. In comparison, training RetinaNet for 12K iteration requires only 10 hours and hleads to better performance (mAP of 14.44% in Table VI). Training the combination of DA-Faster RCNN and CycleGAN require about 67 days and allows to achieve a mAP 33.20% whereas the combination of RetinaNet and CycleGAN requires 64 days and leads to a mAP of 55.54%

F. Discussion

The presented results in this section lead to the following findings. 1) Object detection algorithms trained on synthetic images tend to overfit to the source domain (Table II). This can be mitigated by performing early stopping, e.g., selecting the models trained for 12K iterations rather than those trained for 62K iterations. 2) RetinaNet exhibits less sensitiveness to the domain shift between synthetic and real data, as compared to Faster RCNN. This is confirmed by the results in Table III, in which RetinaNet achieves better performance for a reasonable training epochs of CycleGAN in the presence of an overfitted model (i.e., a model trained for 62K iterations) and a non-fully trained CycleGAN model (i.e., trained for just 30 epochs). The better performance obtained by RetinaNet when the training images are translated from synthetic to real (Table IV) confirm this observation. 3) image-to-image translation is particularly effective as a domain adaptation technique when the models are trained on labeled synthetic images transformed to the real domain (Table IV), as compared to transforming real images to the synthetic domain at test time (Table III). This seems to be due to the limited ability of CycleGAN to transform real images to synthetic ones (Figure 4), whereas the inverse transformation seems is in general more successful (Figure 5). 4) Approaches to unsupervised domain adaptation for object detection can be combined with image-to-image translation for improved performance, as outlined in Table V. In this case, once the number of CycleGAN training epochs is fixed, choosing the right object detector allows to gain an improvement of $\sim 20\%$ as shown in Table VI. Again, the training times required by the object detectors are negligible compared to the time required by CycleGAN. For this reason, the proposed dataset is a challenging benchmark for unsupervised domain adaptation in the context of object detection in terms of

accuracy and computational resources, as it is highlighted by the results shown in Tables VI and VII.

VI. Conclusion

We have considered the challenging problem of unsupervised domain adaptation for object detection in cultural sites. To investigate the performance of methods in this domain, we proposed a dataset comprising 75244 synthetic and 2190 real images of 16 artworks in a cultural site. We hence investigated different approaches to unsupervised domain adaptation for object detection on the proposed dataset. The performed analysis provided different baseline results for the benchmark dataset. We hope that this will encourage future research on the topic. Our results suggest that single-stage detector such as RetinaNet, combined with image-to-image translation allow to significantly reduce the domain gap between synthetic and real images and generally, performs better than DA-Faster RCNN and Fast RCNN combined with CycleGAN. However, this comes at the cost of longer training times due to the optimization of CycleGAN. Also, approaches based on feature alignment and image-to-image translation techniques can be combined for improved performance but, choosing the right object detector for the dataset, can significantly improve the performance. Future work may extend the analysis to other application scenarios beyond cultural heritage and will explore the possibility of combining feature alignment and image-to-image translation techniques in a single-stage object detection algorithm to improve performance on the considered benchmark.

Acknowledgement

This research is supported by the project VALUE (CUP G69J18001060007) granted by PO FESR 2014/2020 - Azione 1.1.5, and by Piano della Ricerca 2016-2018 linea di Intervento 2 of DMI, University of Catania. The authors would like to thank Regione Siciliana Assessorato dei Beni Culturali dell'Identità Siciliana - Dipartimento dei Beni Culturali e dell'Identità Siciliana and Polo regionale di Siracusa per i siti culturali - Galleria Regionale di Palazzo Bellomo.

References

[1] S. Alletto, D. Abati, G. Serra, and R. Cucchiara, "Exploring architectural details through a wearable egocentric vision device," *Sensors*, vol. 16, p. 237, 02 2016.

[2] S. Alletto, D. Abati, G. Serra, and R. Cucchiara, "Wearable vision for retrieving architectural details in augmented tourist experiences," in *2015 7th International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN)*, June 2015, pp. 134–139.

[3] P. Buonincontri and A. Marasco, "Enhancing cultural heritage experiences with smart technologies: An integrated experiential framework," *European Journal of Tourism Research*, vol. 17, pp. 83–101, 02 2017.

[4] L. Seidenari, C. Baecchi, T. Uricchio, A. Ferracani, M. Bertini, and A. Del Bimbo, "Deep artwork detection and retrieval for automatic context-aware audio guides," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 13, pp. 1–21, 06 2017.

[5] G. Farinella, G. Signorello, S. Battiato, A. Furnari, F. Ragusa, R. Leonardi, E. Ragusa, E. Scuderi, A. Lopes, L. Santo, and M. Samarotto, *VEDI: Vision Exploitation for Data Interpretation*, 09 2019, pp. 753–763.

[6] S. Orlando, A. Furnari, and G. M. Farinella, "Egocentric visitor localization and artwork detection incultural sites using synthetic data," *Pattern Recognition Letters*, 2020.

[7] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," *CoRR*, vol. abs/1702.05464, 2017.

[8] J. Hoffman, E. Tzeng, T. Park, J. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," *CoRR*, vol. abs/1711.03213, 2017.

[9] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster r-cnn for object detection in the wild," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[10] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," *CoRR*, vol. abs/1812.04798, 2018.

[11] R. Xie, F. Yu, J. Wang, Y. Wang, and L. Zhang, "Multi-level domain adaptive learning for cross-domain detection," *CoRR*, vol. abs/1907.11484, 2019.

[12] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[13] M. Mancini, L. Porzi, S. R. Bulo, B. Caputo, and E. Ricci, "Inferring latent domains for unsupervised deep domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[14] F. Ragusa, A. Furnari, S. Battiato, G. Signorello, and G. M. Farinella, "Ego-CH: Dataset and fundamental tasks for visitors behavioral understanding using egocentric vision," *Pattern Recognition Letters*, 2020.

[15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[16] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.

[17] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *CoRR*, vol. abs/1708.02002, 2017. [Online]. Available: http://arxiv.org/abs/1708.02002

[18] R. Cucchiara and A. Del Bimbo, "Visions for augmented cultural heritage experience," *IEEE Multimedia*, vol. 21, 01 2014.

[19] M. Portaz, M. Kohl, G. Quénot, and J. Chevallet, "Fully convolutional network and region proposal for instance identification with egocentric vision," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017, pp. 2383–2391.

[20] A. S. Razavian, O. Aghazadeh, J. Sullivan, and S. Carlsson, "Estimating attention in exhibitions using wearable cameras," in *Proceedings of the 2014 22nd International Conference on Pattern Recognition*, ser. ICPR '14. USA: IEEE Computer Society, 2014, p. 2691–2696. [Online]. Available: https://doi.org/10.1109/ICPR.2014.465

[21] F. Ragusa, A. Furnari, S. Battiato, G. Signorello, and G. Farinella, "Egocentric visitors localization in cultural sites," *Journal on Computing and Cultural Heritage*, vol. 12, pp. 1–19, 04 2019.

[22] G. Signorello, G. M. Farinella, L. Di Silvestro, A. Torrisi, and G. Gallo, "Exploring geo-tagged photos to assess spatial patterns of visitors in protected areas : the case of park of etna (italy)," in *VGI Geovisual Analytics Workshop*, D. Burghardt, S. Chen, G. Andrienko, N. Andrienko, R. Purves, and A. Diehl, Eds., 2018. [Online]. Available: http://bdva.net/2018/index.php/vgi-geovisual-analytics-workshop/

[23] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4500–4509.

[24] H. Lee, H. Tseng, Q. Mao, J. Huang, Y. Lu, M. Singh, and M. Yang, "DRIT++: diverse image-to-image translation via disentangled representations," *CoRR*, vol. abs/1905.01270, 2019. [Online]. Available: http://arxiv.org/abs/1905.01270

[25] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," 2014.

[26] A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, "Image analogies," in *SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press, 2001, pp. 327–340.

[27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[28] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[29] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.