Virtual to Real Unsupervised Domain Adaptation for Image-Based Localization in Cultural Sites

Santi Andrea Orlando
DMI, University of Catania
DWORD - XENIA progetti s.r.l.
Catania, Italy
santi.orlando@unict.it

Antonino Furnari
DMI, University of Catania
Catania, Italy
furnari@dmi.unict.it

Giovanni Maria Farinella
DMI, University of Catania
Catania, Italy
gfarinella@dmi.unict.it

***Abstract*—The ability to localize the visitors of a cultural site from egocentric images can allow applications to understand where people go and what they pay attention to in the site. Current pipelines to tackle the problem require the collection and labeling of large amounts of images, which is challenging, especially in large-scale indoor environments. On the contrary, virtual images of a cultural site can be generated and automatically labeled using dedicated tools with minimum effort. In this paper, we investigate whether unsupervised domain adaptation techniques can be used to train localization models on labeled virtual data and unlabeled real data, and deploy them to work with real images. To perform this study, we propose a new dataset of both real and virtual images acquired in a cultural site which are labeled for room-based localization as well as for 3 DOF camera pose estimation. We hence compare two approaches to unsupervised domain adaptation: mid-level representations and image-to-image translation. Our analysis shows that both approaches can be used to reduce the domain gap arising from the different data sources and that the proposed dataset is a challenging benchmark for unsupervised domain adaptation for image-based localization.**

***Index Terms*—Image-Based Localization, Domain Adaptation, Cultural Heritage**

I. Introduction

Localizing the visitors in a cultural site can enable useful services for both the visitors and the manager of the site [1]. Indeed, on the one hand, the location of the visitor can be used to assist navigation within the site and provide additional information on the visited environments. On the other hand, analyzing location information collected from multiple visitors can allow the manager to analyze where people go and what they observe to measure the performance of the site and improve its services. While Global Positioning System (GPS) can be an option for coarse localization in outdoor places [2], accurate localization can be obtained both outdoors [3] and indoors [4] using image-based localization.

Different approaches to image-based localization exist [5], [6]. However, all of them require the collection and labeling of large datasets of images of the environment in which localization is to be performed. Structure from Motion techniques [7] can be employed both outdoors and indoors to reconstruct a 3D model of the scene and hence attach a 6DOF pose to each image, as discussed in [8], [9]. However, the aforementioned framework require the supervision of experts, which can be expensive in terms of time and resources. Recent works have proposed to leverage virtual data obtained from a 3D reconstruction of the environment [10], [11] to train image-based localization algorithms. The advantage of such approach is that, once the real environment has been modeled, it is straightforward and inexpensive to generate large amounts of labeled data useful for training. However, the domain of virtual images is very different from the one of images collected in the real environment, due to the lack of photo-realistic detail which characterizes the 3D reconstruction (see first column of Fig. 1). To fill this domain gap, previous works have used both virtual and real images with domain adaptation techniques such as image-to-image translation [11]. Since labeling even a small amount of real data for localization is challenging, it would be ideal to train the models using *labeled virtual data and unlabeled real data*, which is referred as unsupervised domain adaptation problem [12], [13].

In this paper, we investigate the problem of unsupervised virtual to real domain adaptation for image-based localization in indoor cultural sites. To perform the study, we collected a dataset of real and virtual images labeled for localization. The data has been collected in the museum "Galleria Regionale di Palazzo Bellomo" located in Siracusa, Italy[1]. Each image of the dataset has been labeled according to the room in which the image has been acquired. This allows to investigate room-based localization algorithms. A subset of the data has also been labeled with the 3 Degrees of Freedom (3DOF) pose of the camera. Using the collected dataset, we investigated the use of different approaches to unsupervised domain adaptation for image-based localization considering two levels of granularity: room-based localization and 3DOF pose estimation. We considered an image-based localization pipeline which performs image retrieval using a Triplet network trained on virtual data and tested on real data. We hence explored two main approaches to reduce the domain gap. The former consists in training the Triplet network on mid-level representations [14], such as 2D/3D edges, 2D/3D keypoints and estimated depth. The latter exploits image-to-image translation techniques (e.g., CycleGAN [15] and ToDayGAN [16]) to translate the images from the real to the virtual domain, in order to be able to re-

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

use models trained only on virtual images. We also explored a combination of the two approaches which uses image-to-image translation to further adapt mid-level representations. The contribution of this work is two-fold: 1) We propose and publicly release a dataset of virtual and real images collected in a real cultural site. Please visit our web page for further information[2]. The dataset has been labeled for image-based localization, and hence it can be used to study the problem of unsupervised domain adaptation for image-based localization in cultural sites. To the best of our knowledge, this is the first dataset available to the community to study the considered problem. 2) We benchmark two main approaches to unsupervised domain adaptation for image-based localization: the use of mid-level representations [14] and the use of image-to-image translation techniques [16], [11].

The remainder of the paper is organized as follows: Section II discusses the related work. Section III describes the proposed dataset for unsupervised domain adaptation for image-based localization in cultural sites. Section IV introduces the considered methods. Section V reports the experiments and analyzes the results. Section VI concludes the paper and discusses future works.

II. Related Work

A. Image-Based Localization

Image-based localization consists in recognizing from which location a given query image has been acquired [5]. This problem can be tackled at different levels of granularity, such as class-based localization [4] (e.g., recognizing the room in which a given image has been acquired) and camera pose estimation [8] (e.g., inferring the 6DOF camera pose from which a given image has been collected).

Image-based localization can be addressed using different approaches. Some methods use a CNN to solve localization through camera pose regression [8], [17]. Other approaches use 2D-3D matches between images and a 3D map of the environment to estimate the camera pose [18]. Another family of algorithms uses image retrieval to recognize the place in which an image has been collected with compact and efficient descriptors [6]. In this case, the camera pose is approximated with the one of the most similar image contained in a labeled database of training images [19], [9]. Recent work [5] demonstrates that existing methods for image-based localization performing camera pose regression do not yet achieve performance comparable to methods based on image retrieval and structure. All the aforementioned methods perform localization directly in the real domain. Our approach instead tries to adapt the real domain to the virtual one using unpaired domain adaptation.

B. Domain Adaptation

The goal of domain adaptation is to improve the performance of an algorithm trained on a source domain, when it is tested on data coming from a target domain. Some approaches to domain adaptation use image-to-image translation to make images from the source and target domain look more similar. Examples of such forms of adaptation include the transferring of the artistic style of an image to another [20], as well other forms of translation, such as day to night and aerial image to map. To perform style transfer a dataset of paired examples from the two domains is given. A recent work [15] introduced CycleGAN, an approach which can transform images from a domain to another relying on unpaired samples from the two domains for training. Other approaches have focused on adapating features both at the pixel-level and at the feature-level, extending adaptation to other tasks such as digit classification and semantic segmentation [11]. Limited work has also been done in the context of domain adaptation for image-based localization. The authors of [16] investigated visual localization from driving images translating daytime images to nighttime ones in order to allow the model to work on both lightning conditions. The authors of [11] proposed to use real images to create a 3D model of the scene in order to render virtual images from different points of view. Virtual images are used as a way to augment data to train the localization algorithm. The gap between real and virtual images is bridged using image-to-image translation techniques.

In this paper, we investigate the suitability of image-to-image translation and mid-level representations as a means to perform domain adaptation for image-based localization. Differently from [11], we perform image-to-image translation from the real to the virtual domain in order to re-use models trained only on virtual data.

C. Mid-Level representations

While high-level representations carrying semantic information are mainstream in computer vision, mid-level representations such as edges or depth have been less studied. Recent works have investigated the relationships between different visual tasks [14], proposing a taxonomy of abstract visual representations. Such abstract representations, also referred to as "mid-level representations", have been found useful to improve generalization in learning visuomotor tasks [21]. In this work, we investigate the suitability of mid-level representations to handle the problem of unsupervised domain adaptation for image-based localization.

III. Dataset

The proposed dataset contains real and virtual images of the indoor cultural site Galleria Regionale di Palazzo Bellomo, located in Siracusa, Italy. Each image is labeled according to the room in which it has been acquired. A subset of the data has also been labeled with the related 3DOF camera pose. This consists in images contained in 4 of the 11 rooms. Please see the supplementary material for details about the dataset.[3] Together with the dataset, we release the pre-extracted mid-level representations. The following sections detail how virtual

[2]https://iplab.dmi.unict.it/DomainAdaptationLocalization/

[3]https://iplab.dmi.unict.it/DomainAdaptationLocalization/Supplementary_Material_IPAS2020.pdf

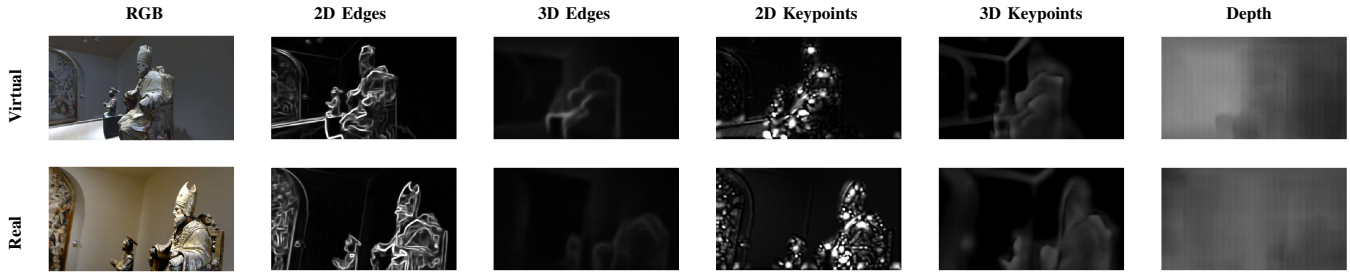

Fig. 1. Examples of mid-level representations with the corresponding RGB images. We report pairs of virtual and real images.

and real images have been collected and how 3DOF camera pose labels have been attached to the images.

A. Virtual images

We consider the dataset of virtual images proposed in [22]. The data has been generated from the 3D model of the cultural site Palazzo Bellomo, which has been reconstructed using a Matterport 3D scanner[4]. Specifically, the authors of [22] designed a tool which simulates a virtual agent visiting the site. The tool has hence been used to generate images of the 1^{st} floor of the building with the related ground truth labels. In particular, each image is associated to its 3DOF camera pose and a label indicating in which of the 11 rooms of the site the image has been collected.

B. Real images

To obtain real images of the same cultural site, we resort to the EGO-CH dataset [23]. This dataset contains videos of subjects visiting two different cultural sites acquired with a Microsoft HoloLens devices. In this paper, we consider the 10 video sequences collected by visitors in the Galleria Regionale di Palazzo Bellomo cultural site. Each frame of the videos has been labeled to specify in which room the visitor was located at the moment of the acquisition. All frames are associated with a room-level label which specifies in which of the 11 rooms the image has been acquired.

3DOF labels for real images: While the virtual images have been automatically labeled with both the 3DOF camera pose and the room-based label, the real data contained in [23] only contains room-based labels. To be able to perform experiments with unsupervised domain adaptation also at the level of camera pose estimation, we considered camera pose labels for a subset of 4 of the 11 rooms of the considered cultural site: Sala 5, Sala 7, Sala 9 and Sala 13. Camera poses have been obtained performing a 3D reconstruction of each room through structure from motion using COLMAP [7]. We focused on a subset of the rooms since, as observed in past works [9], applying structure from motion to images of large-scale building is challenging. We fed the structure from motion algorithm with $1,766$ frames for Sala 5, $1,597$ for Sala 7, $1,570$ for Sala 9, and 837 for Sala 13. We used the "geo-registration" function of COLMAP to register the 3D reconstructions with the 3D model of the building acquired from [22]. This has been done by feeding the model with real images of known camera pose obtained form the high resolution scans produced by Matterport. The process returned 932 labeled frames as not all images could be properly attached to the model by the structure from motion algorithm. Note that we use these 932 labeled images for test only.

[4]https://matterport.com/

IV. Methods

A. Image-based localization pipeline

To investigate the effect of unsupervised domain adaptation techniques, we consider a simple image-based localization pipeline based on image retrieval. The considered pipeline assumes that training and test images are represented using two representation functions Ψ_{train} and Ψ_{test}. While Ψ_{train} and Ψ_{test} can be the same representation function, in general we assume that $\Psi_{train} \neq \Psi_{test}$ since training and test images belong to different distributions (domain shift due different acquisition modality). The goal of Ψ_{train} and Ψ_{test} is to map training and test images to a representation space in which images of nearby locations have a small distance, whereas images of distant locations have a large distance. Given a query image I_q, localization is performed using a nearest neighbor search, i.e., we assign to I_q the label of the closest training sample in the representation space induced by Ψ_{train} and Ψ_{test}. It should be noted that, since training samples are virtual images, they contain both room-based and 3DOF camera pose labels. Hence, this approach is used to perform localization at both levels of granularity.

B. Representation function Ψ

In practice, to benchmark the performance of mid-level representations and image-to-image translation techniques for domain adaptation, we define a general structure for the representation functions Ψ_{train} and Ψ_{test}, as illustrated in Figure 2. The generic representation function Ψ can contain different modules, as discussed in the following sections. To benchmark the different modules, we perform experiments with number of ablations of this structure including and excluding branches and modules.

1) Mid-level representation branches: We use a series of branches to extract mid-level representations from the input RGB image through representation functions M (e.g., $M_{2D\ Edges}$ in Figure 2). One of such branches implements an identity function which passes the unprocessed RGB image to

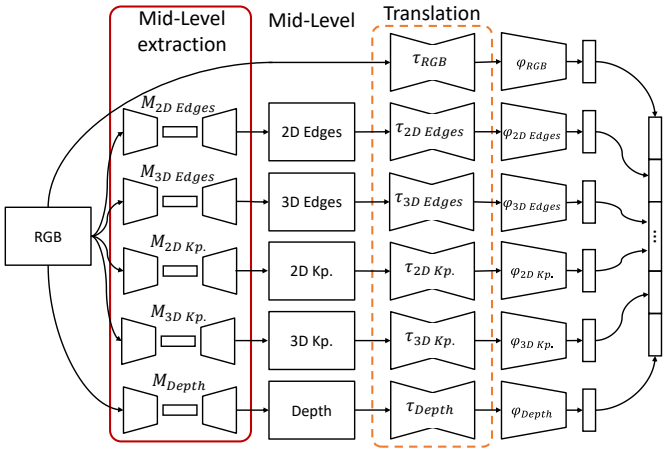

Fig. 2. General structure for the representation functions Ψ.

the next layer. The other branches extract five different mid-level representations: 2D Edges, 3D Edges, 2D Keypoints, 3D Keypoints and Depth [14]. Examples of mid-level representations extracted for both virtual and real images are shown in Fig. 1.

As can be noted, some of these representations (such as 2D Edges and 2D Keypoints possess) some degree of invariance with respect to the image source (real or virtual). Note that each considered mid-level representation is a gray-scale image.

2) Image-to-image translation modules: An optional series of image-to-image translation modules (one per branch) is used to transform the RGB images or mid-level representations from the real to the virtual domain. This module can be included for test images, which belong to the real domain, but it is always excluded for training images, which belong to the virtual one. This is referred in Fig. 2 as translation.

3) Embedding functions and concatenation layer: A series of embedding functions are finally used to map the input image/representation (optionally translated) to the embedding space. These embedding functions are specific to the input modality. For instance, φ_{RGB} is used to process RGB images, whereas $\varphi_{2DEdges}$ is used to process 2D Edges. The embedding vectors obtained at the end of each branch are hence concatenated to obtain the final representation.

C. Training of the representation function

The representation functions Ψ are trained in a piece-wise fashion. In particular, we use the approach proposed in [14] to implement the M functions to extract mid-level representations. The Depth modality considered in this paper is equivalent to the mid-level representation referred to as Euclidean Distance in [14]. The image-to-image translation modules are implemented using either CycleGAN [15] or ToDayGAN [16]. The last is a day-to-night translation approach specifically designed for localization. We considered ToDayGAN in our experiments as it has been proposed for image-to-image translation for localization, even if in different context of night-to-day adaptation. These algorithms are trained separately to translate either RGB images or mid-level representations between the real and virtual domain.

The embedding functions φ are trained following the approach discussed in [22]. Specifically, φ is defined as a CNN. The embedding function is hence trained using a Triplet architecture [24] to learn an embedding space suitable for location-based image retrieval. Training samples consist of triplets of images (x_i, x_i^+, x_i^-) containing an "anchor" image x_i, a "similar" image x_i^+ and a "dissimilar" image x_i^-. Given an anchor image x_i, we randomly sample two images x_i^+ and x_i^- such that:

$$\begin{aligned}(d_o(x_i, x_i^+) <= th_o) \wedge (d_p(x_i, x_i^+) <= th_p)\\ (d_o(x_i, x_i^-) > th_o) \vee (d_p(x_i, x_i^-) > th_p)\end{aligned} \quad (1)$$

where $d_o(x_1, x_2)$ and $d_p(x_1, x_2)$ are respectively the distances between the orientations and positions of the two images x_1 and x_2. th_o and th_p are two thresholds used to establish when two images are to be considered similar or dissimilar. Following [22] we set $th_o = 45°$ and $th_p = 0.5\ m$. As the architecture for φ, we use an InceptionV3 CNN [25] pre-trained on Imagenet with the last classification layer removed. To train the model with mid-level representations, we adapt the first convolutional layer of the CNN to process gray-scale images. The triplet architecture is hence trained using the Margin Ranking Loss:

$$L_\varphi(x_i, x_i^+, x_i^-) = \max(0, d(\varphi(x_i), \varphi(x_i^+)) - d(\varphi(x_i), \varphi(x_i^-)) + m) \quad (2)$$

where d is the Euclidean distance and m is the margin hyper-parameter, which we set to $m = 0.2$.

V. Experimental Settings and Results

A. Experimental Settings

We extracted mid-level representations from both real and synthetic images using the models provided in [14]. We processed all images at the resolution of 464×256 pixels. We trained all triplet architectures using stochastic gradient descent with momentum equal to 0.5 and a learning rate of 0.001. All models have been trained for 100 epochs. As a post-hoc early stopping strategy, we selected the epoch that gives the best accuracy on the validation set.

CycleGAN and ToDayGAN have been trained for 35 epochs, randomly cropping images to 224×224 pixels for data augmentation. At test time, images are processed at 464×256 pixels. To balance the number of samples from each domain, we sub-sampled the virtual samples by a factor of 2.

Each component of the vectors obtained using the representation functions Ψ_{train} has been normalized to have zero mean and unit standard deviation. The same normalization has been applied to the vectors obtained using Ψ_{test}.

B. Room-Based Localization Results

Table I reports the results obtained considering different combinations of RGB images and mid-level representations. We report results including and excluding image translation through CycleGAN from the pipeline. Best results per-section are reported in bold numbers. As can be observed, using

TABLE I

ROOM-BASED LOCALIZATION RESULTS CONSIDERING DIFFERENT COMBINATIONS OF RGB IMAGES AND MID-LEVEL REPRESENTATIONS, WITH AND WITHOUT IMAGE-TO-IMAGE TRANSLATION.

		CycleGAN	
	Representation	**No**	**Yes**
Baseline	RGB	22.58%	50.74%
Single mid-level representation	Depth	15.63%	15.37%
	2D Keypoints	18.43%	13.33%
	2D Edges	18.45%	**38.07%**
	3D Edges	18.52%	18.32%
	3D Keypoints	**19.83%**	19.44%
Combination of two repr.	RGB + Depth	21.47%	47.02%
	RGB + 2D Keypoints	22.62%	29.68%
	RGB + 2D Edges	24.26%	**52.98%**
	RGB + 3D Edges	24.76%	46.98%
	RGB + 3D Keypoints	**25.03%**	47.21%
Combination of trhee repr.	RGB + 2D Edges + Depth	24.17%	50.60%
	RGB + 2D Edges + 2D Kp.	22.20%	42.98%
	RGB + 2D Edges + 3D Edges	25.71%	**50.88%**
	RGB + 2D Edges + 3D Kp.	**26.17%**	50.17%
Four repr.	RGB + 2D Edges + 3D Kp. + 3D Edges	**28.57%**	48.46%
ToDayGAN	RGB	19.80%	

TABLE II

3DOF LOCALIZATION RESULTS CONSIDERING DIFFERENT COMBINATIONS OF RGB IMAGES AND MID-LEVEL REPRESENTATIONS, WITH AND WITHOUT IMAGE-TO-IMAGE TRANSLATION.

		Position Error (m) (avg/median)		Orientation Error (°) (avg/median)	
	Representation	**CycleGAN**			
		No	**Yes**	**No**	**Yes**
Baseline	RGB	3.82/3.75	3.23/2.68	70.13/55.01	52.71/38.38
Single mid-level representation	Depth	4.90/4.67	4.94/4.81	87.41/89.29	59.62/50.47
	2D Kp.	4.29/4.20	4.73/4.49	84.87/85.34	59.62/50.47
	2D Edges	**3.84/3.48**	**3.75/3.14**	**74.51/62.01**	64.87/50.64
	3D Edges	4.25/4.03	4.27/3.86	82.29/76.80	**52.55/45.93**
	3D Kp.	4.18/3.61	4.06/3.49	77.49/70.06	59.62/50.47
Combination of two repr.	RGB + Depth	3.73/3.38	3.52/2.94	72.81/60.15	59.84/45.90
	RGB + 2D Kp.	3.72/3.52	4.11/3.73	81.35/76.81	67.18/50.29
	RGB + 2D Edges	**3.57**/3.53	3.27/**2.51**	72.64/63.65	54.67/**38.77**
	RGB + 3D Edges	3.67/3.21	**3.23**/2.61	**72.62/58.39**	**58.01**/42.62
	RGB + 3D Kp.	3.75/**3.15**	3.46/2.87	73.14/58.94	59.60/45.56
Combination of trhee repr.	RGB + 2D Edges + Depth	3.57/3.17	3.27/2.60	69.48/60.90	56.52/46.79
	RGB + 2D Edges + 2D Kp.	**3.49/3.13**	3.28/2.73	75.71/69.11	60.38/44.23
	RGB + 2D Edges + 3D Edges	3.59/3.32	**3.20/2.46**	68.35/56.75	**55.55**/41.87
	RGB + 2D Edges + 3D Kp.	3.57/3.26	3.24/2.51	**67.76/52.84**	55.84/**40.90**

only RGB images with no image-to-image translation allows to obtain limited performance (22.58%). This is due to the domain gap between virtual and real images. Using a single mid-level representation leads to worse performance, with best results achieved when considering 3D Keypoints (19.83%). This is reasonable as mid-level representations tend to discard low-level and textural information which could be useful for retrieval (see Fig. 1). Interestingly, combining RGB images with one of the considered mid-level representations leads to improved performance in almost all cases, with RGB + 3D Keypoints obtaining an accuracy of 25.03%, which is an improvement of +5.2% over 3D Keypoints and +2.45% over RGB. The limited performance obtained when combining RGB and Depth may be due to the limited performance of the network proposed in [14] to accurately estimate depth on the provided dataset (see Fig. 1). Overall, these results suggest that mid-level representations tend to extract domain-invariant information which is complementary to the one exploited by the triplet network when processing RGB images. We also investigate the effect of combining RGB with 2D Edges and other mid-level representations. Combinations of three representations lead to further gains, with RGB + 2D Edges + 3D Keypoints achieving 26.17% (+3.59% with respect to RGB). The table further shows that combining RGB with 2D Edges, 3D Keypoints and 3D Edges brings an accuracy of 28.57%, which is +5.99% with respect to RGB. The rightmost column of Table I shows that including image-to-image translation with CycleGAN in the pipeline allows to greatly reduce the domain gap. Indeed, the RGB baseline with image-to-image translation obtains an accuracy of 50.74%, a +28.16% boost with respect to the same baseline without image-to-image translation (22.58%). Table I also shows that, among the mid-level representations, only 2D Edges benefit from image-to-image translation through CycleGAN. The method based on 2D Edges with image-to-image translation obtains an accuracy of 38.07%, which is a +19.62% boost with respect to 2D Edges without image-to-image translation (18.45%). As a result, the combination of RGB + 2D Edges is the one benefiting most from image-to-image translation. Indeed, this approach achieves an accuracy of 52.98%, a +30.4% improvement with respect to the RGB baseline and a +2.24% improvement with respect to RGB + CycleGAN. ToDayGAN does not allow to achieve any improvement over the RGB baseline (accuracy of 19.80% versus 22.58%). This may be due to the fact that ToDayGAN has been specifically designed to perform translation between day and night images, whereas CycleGAN has a more general architecture.

C. 3DOF Camera Pose Estimation Results

In this section, we compare the performance of the methods on the subset of data with 3DOF camera pose labels. Table II reports the results obtained considering different combinations of RGB images and mid-level representations. We also report results with and without CycleGAN. As can be noted from the table, differently from room-based localization, performance on 3DOF localization is still very limited, with large position and orientation errors. This highlights that the considered problem is challenging and that the dataset is an interesting benchmark for unsupervised domain adaptation for 3DOF localization.

From Table II we can observe that, when CycleGAN is not included in the pipeline, using only mid-level representations generally performs worse than using only RGB inputs, with the exception of 2D Edges, which perform comparably to RGB (mean/median position errors of $3.84m/3.48m$, versus $3.82m/3.75m$) and worse for orientation errors ($74.51°/62.01°$ vs $70.13°/55.01°$). Combining mid-level representations with RGB allows to improve results over the

TABLE III
SUMMARY OF 3DOF LOCALIZATION RESULTS.

Adaptation mean	Position Error(m)		Orientation Error(∘)		Improvements(m)		Improvements(∘)	
	Avg.	Median	Avg.	Median	Avg.	Median	Avg.	Median
RGB (no adaptation)	3.82	3.75	70.13	55.01	/ /	/ /	/ /	/ /
RGB + CycleGAN	3.23	2.68	**52.71**	**38.38**	−0.59	−1.07	**− 31.75**	**− 16.63**
RGB + ToDayGAN	4.52	3.81	78.95	73.13	+ 0.70	+ 0.06	+ 8.82	+ 18.12
RGB + 2D Edges + 2D Kp.	3.49	3.13	68.35	56.75	−0.33	−0.62	− 1.78	+ 1.74
RGB + 2D Edges + 3D Edges + CycleGAN	**3.20**	**2.46**	55.55	41.87	**−0.62**	**−1.29**	− 14.58	− 13.14

RGB baseline. Indeed, RGB + 2D Edges obtains a mean position error of $3.57m$ and RGB + 3D Keypoints obtains a median position error of $3.15m$ (versus $3.82m/3.75m$ of RGB). Combining RGB with two mid-level representations leads to further improvements, with RGB + 2D Edges + 2D Keypoints obtaining position errors of $3.49m/3.13m$, while orientation error improves using RGB + 2D Edges + 3D Keypoints with $67.76°/52.84°$.

Including CycleGAN in the pipeline allows to improve 3DOF localization when RGB images are considered. Indeed, with CycleGAN, position errors get to $3.23m/2.68m$, versus $3.82m/3.75m$ without CycleGAN and orientation errors improve to $52.71°/38.38°$, versus $70.13°/55.01°$. Similarly to what observed with room-based localization, CycleGAN helps only with some mid-level representations such as 2D Edges (position errors of $3.75m/3.14m$ vs $3.84m/3.48m$ without CycleGAN). Using CycleGAN in combination with RGB and more than one mid-level representations is beneficial. Indeed, best results are obtained by RGB + 2D Edges + 3D Edges, which obtains position errors of $3.20m/2.46m$ and orientation error of $55.55°/41.87°$. Table III summarizes the results of the best performing approaches from Table II. We also compare these approaches with RGB + ToDayGAN. Results show that using CycleGAN allows to reduce position estimation errors by $0.59m/1.07m$. Similarly, mid-level representations allow to reduce position errors by $0.33m/0.62m$. Combining the two approaches finally allows to improve over the RGB baseline by $0.62m/1.29m$. Interestingly, the best improvements for orientation are of $31.75°/16.63°$ by using RGB+CycleGAN, while mid-level representations improve the performances only by $1.78°$ with worse median error. Also using CycleGAN with mid level representation helps to reduce orientation error obtaining $14.58°/13.14°$.

VI. CONCLUSION

We have considered the problem of unsupervised domain adaptation for image-based localization in cultural sites. To perform the study, we have proposed a dataset of virtual and real images collected in a cultural site. The images are labeled for room-based localization and 3DOF camera pose estimation. Using the proposed dataset, we performed an analysis to assess the suitability of mid-level representations and image-to-image translation approaches to perform unsupervised domain adaptation for image-based localization. The results highlight that both techniques allow to obtain promising results, while the dataset is a challenging benchmark to study unsupervised domain adaptation for image-based localization.

ACKNOWLEDGMENT

This research is supported by XENIA Progetti - DWORD, by the project VALUE - Visual Analysis for Localization and Understanding of Environments (N. 08CT6209090207, CUP G69J18001060007) granted by PO FESR 2014/2020 - Azione 1.1.5 - "Sostegno all'avanzamento tecnologico delle imprese attraverso il finanziamento di linee pilota e azioni di validazione precoce dei prodotti e di dimostrazioni su larga scala", 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] G. M. 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," in *ICIAP*, 2019.

[2] A. R. Zamir and M. Shah, "Accurate image localization based on google maps street view," in *European Conference on Computer Vision*. Springer, 2010, pp. 255–268.

[3] F. L. Milotta, A. Furnari, S. Battiato, G. Signorello, and G. M. Farinella, "Egocentric visitors localization in natural sites," *Journal of Visual Communication and Image Representation*, vol. 65, p. 102664, 2019.

[4] F. Ragusa, A. Furnari, S. Battiato, G. Signorello, and G. M. Farinella, "Egocentric visitors localization in cultural sites," *JOCCH*, 2019.

[5] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixe, "Understanding the limitations of cnn-based absolute camera pose regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3302–3312.

[6] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition." *PAMI*, vol. 40, no. 6, pp. 1437–1451, 2018.

[7] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4104–4113.

[8] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *ICCV*, 2015, pp. 2938–2946.

[9] E. Spera, A. Furnari, S. Battiato, and G. M. Farinella, "Egocentric shopping cart localization," in *ICPR*, 2018, pp. 2277–2282.

[10] S. A. Orlando, A. Furnari, S. Battiato, and G. M. Farinella, "Image based localization with simulated egocentric navigations," in *VISAPP*, 2019.

[11] M. S. Mueller, T. Sattler, M. Pollefeys, and B. Jutzi, "Image-to-image translation for enhanced feature matching, image retrieval and visual localization." *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 4, 2019.

[12] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *ArXiv*, vol. abs/1703.00848, 2017.

[13] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," *arXiv preprint arXiv:1711.03213*, 2017.

[14] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3712–3722.

[15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[16] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool, "Night-to-day image translation for retrieval-based localization," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5958–5964.

[17] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using lstms for structured feature correlation," in *ICCV*, 2017, pp. 627–637.

[18] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "Inloc: Indoor visual localization with dense matching and view synthesis," in *CVPR*, 2018, pp. 7199–7209.

[19] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys, "Large-scale location recognition and the geometric burstiness problem," in *CVPR*, 2016, pp. 1582–1590.

[20] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[21] A. Sax, J. O. Zhang, B. Emi, A. Zamir, S. Savarese, L. Guibas, and J. Malik, "Learning to navigate using mid-level visual priors," *arXiv preprint arXiv:1912.11121*, 2019.

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

[23] 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.

[24] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*, 2015.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.