Virtual to Real Unsupervised Domain Adaptation for Image-Based Localization in Cultural Sites (Supplementary Material)

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

Fig. 1: A real (left) and a virtual (right) image of the same environment.

I. Introduction

This document is intended for the convenience of the reader and reports additional information on both the proposed dataset and qualitative results of proposed method for unsupervised domain adaptation for Image-Based Localization (IBL).

The reader is referred to the manuscript and to our web page https://iplab.dmi.unict.it/DomainAdaptationLocalization/ for further information.

II. Dataset

The virtual dataset contains 4 simulated navigations acquired at 5 frames per seconds. The data has been divided in a training set, including the 2^{nd} and 3^{rd} navigation ($51,284$ frames), a validation set, including the 4^{th} navigation ($23,960$ frames), and a test set, including the 1^{st} navigation ($24,525$ frames). Virtual images are very different from images collected in the real environment due to the lack of photo-realistic detail which characterizes the 3D reconstruction (see Fig. 1).

To align the real dataset with the virtual one, we extracted frames from each video at 5 fps and discarded all frames not comprised in the 1^{st} floor of the building in order to comply with the aforementioned virtual images. Real images have been divided into training, validation and test sets making sure that frames from a given video fall entirely in one of the three sets. In total, we considered $24,357$ real images for training (videos $1-6$), $10,288$ images for validation (videos $9-10$) and $12,008$ images for test (videos $7-8$). Fig. 2 reports the map of the site and a pair of real and virtual samples for each of the 11 rooms. We publicly release the proposed dataset[1] to encourage research on unsupervised domain adaptation for image-based localization in cultural sites. Together with the dataset, we release the pre-extracted mid-level representations.

III. Results

Table I reports the overall accuracy and per-class F_1 obtained by the best performing methods. These are also compared with an approach using ToDayGAN rather than CycleGAN to perform image-to-image translation (RGB + ToDayGAN). The table shows that, differently from CycleGAN, ToDayGAN does not allow to achieve any improvement over the RGB baseline (accuracy of 19.80% versus 22.58%) in our experiments.

Fig. 3 reports some qualitative results of the compared approaches. Each example is composed of two rows. The first row reports the query image, transformed considering the specific pipeline adopted by each method. The second row reports the images from the virtual training set which have been associated as closest to the queries. Predictions with a

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

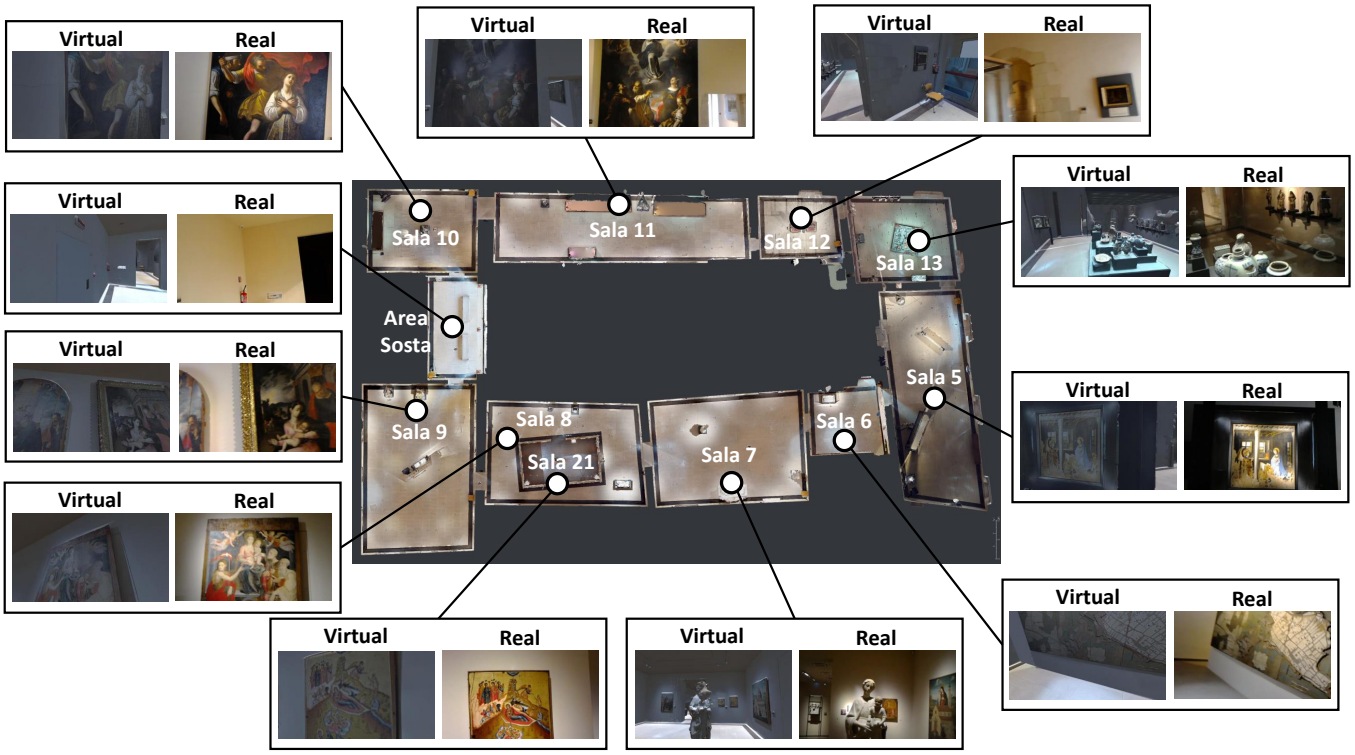

Fig. 2: Map and example pairs of images from the proposed dataset.

	Acc.	A. Sosta	Sala 10	Sala 11	Sala 12	Sala 13	Sala 21	Sala 5	Sala 6	Sala 7	Sala 8	Sala 9
RGB (no adaptation)	22.58%	8.72%	13.06%	0.56%	15.14%	2.41%	31.96%	55.43%	2.30%	18.21%	0.88%	26.57%
RGB + ToDayGAN	19.80%	2.96%	18.41%	28.02%	11.43%	14.69%	25.45%	19.01%	7.47%	21.14%	9.72%	22.12%
RGB + 2D Edges + 3D Kp. + 3D Edges	28.57%	11.92%	19.72%	13.87%	20.76%	16.30%	46.89%	56.52%	13.46%	27.95%	10.93%	29.88%
RGB + CycleGAN	50.74%	16.54%	39.20%	62.47%	32.81%	**26.38%**	58.70%	**64.56%**	**19.35%**	53.86%	39.19%	48.76%
RGB + 2D Edges + CycleGAN	**52.98%**	**21.68%**	**41.99%**	**64.25%**	**39.09%**	23.29%	**66.42%**	64.39%	18.95%	**59.70%**	**42.10%**	**48.95%**

TABLE I: Accuracy and F_1 scores obtained by the compared methods on each class.

position error below $1\ m$ and orientation error below to $45°$ are highlighted in green. Methods tend to perform well when an object is clearly depicted in the query image, as it is the case of the first example. Image-to-image translation is necessary in the second example in which the domain shift is more evident. Indeed, this example is not correctly localized by methods not incorporating image-to-image translation. Mid-level representations such as edges, in addition to image-to-image translation, are particularly useful when the query image contains geometrical structure, as shown in the last example.

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.

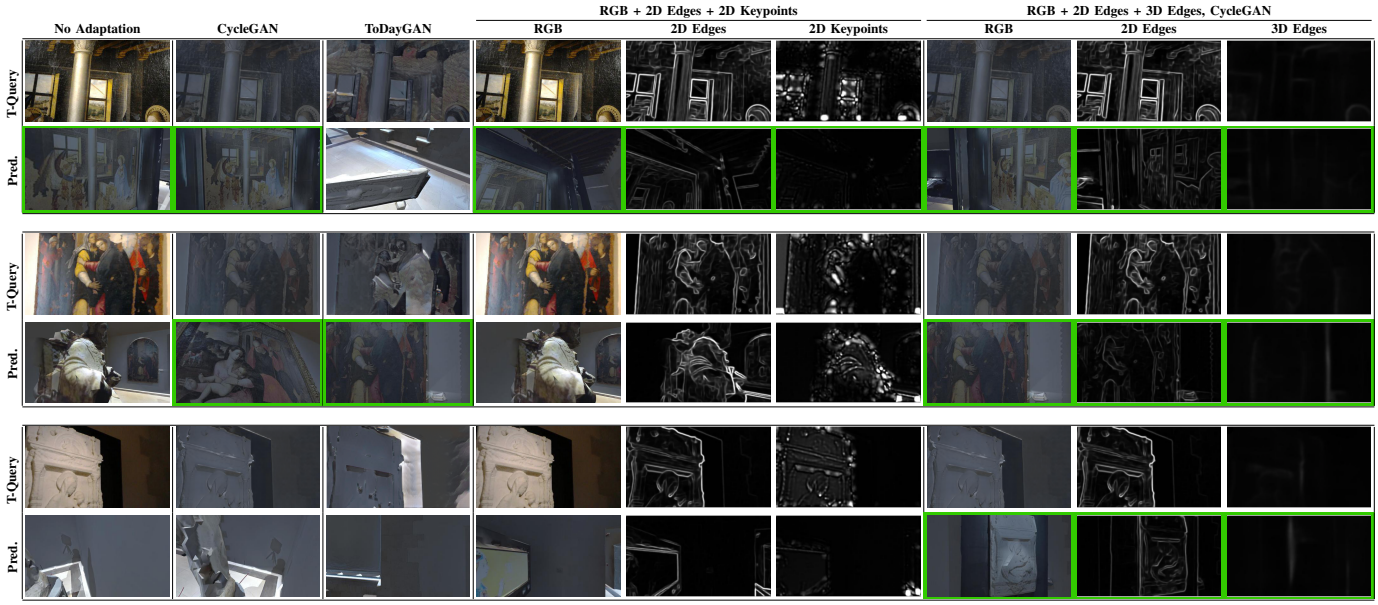

Fig. 3: Qualitative results obtained by the compared methods. T-Query rows are related to the transformed images used as query. Correct predictions are highlighted in green.