Semantic Object Segmentation in Cultural Sites using Real and Synthetic Data

Francesco Ragusa[1,2], Daniele Di Mauro[1], Alfio Palermo[1], Antonino Furnari[1], Giovanni Maria Farinella[1]
[1]FPV@IPLAB - Department of Mathematics and Computer Science, University of Catania, Italy
[2]XDG - XENIA Progetti s.r.l., Acicastello, Catania, Italy
francesco.ragusa@unict.it,{dimauro, furnari, gfarinella}@dmi.unict.it

***Abstract*—We consider the problem of object segmentation in cultural sites. Since collecting and labeling large datasets of real images is challenging, we investigate whether the use of synthetic images can be useful to achieve good segmentation performance on real data. To perform the study, we collected a new dataset comprising both real and synthetic images of 24 artworks in a cultural site. The synthetic images have been automatically generated from the 3D model of the considered cultural site using a tool developed for that purpose. Real and synthetic images have been labeled for the task of semantic segmentation of artworks. We compare three different approaches to perform object segmentation exploiting real and synthetic data. The experimental results point out that the use of synthetic data helps to improve the performances of segmentation algorithms when tested on real images. Satisfactory performance is achieved exploiting semantic segmentation together with image-to-image translation and including a small amount of real data during training. To encourage research on the topic, we publicly release the proposed dataset at the following url: https://iplab.dmi.unict.it/EGO-CH-OBJ-SEG/.**

I. Introduction

Wearable devices equipped with a camera, such as smart glasses, can be used in cultural sites to develop user-centered applications able 1) to provide information on what is being observed by the visitor [1], [2] and 2) to understand where the visitors spend their time in the site as well as which artworks they pay more attention to [3]. In order to develop such applications, it is important to process images collected from the visitor's point of view and recognize which artworks are in the scene. We argue that, since artworks such as statues or small objects can have irregular shapes, performing object recognition at the bounding box level is not sufficient (see Figure 1 for an example). Hence we consider the problem of segmenting artworks with pixel-level accuracy. Different semantic segmentation algorithms can be exploited to deal with this problem [4], [5], [6]. Since in cultural sites objects have to be recognized at the instance level (e.g., a specific painting or statue) rather than at the category level (e.g., a generic painting or statue), these approaches need to be trained on images depicting the specific artworks they should be able to recognize. This requires the collection and labeling of large amounts data, which generally demands a conspicuous effort in terms of time. Moreover, egocentric data can be difficult to acquire due to privacy issues. Previous works have proposed to leverage synthetic data obtained via simulations to train computer vision algorithms when real data is not readily available [7], [8], [9]. The main advantage of these approaches is that, once the simulation of the target environment is available, automated tools can be used to collect arbitrarily-sized datasets of labeled synthetic images with a small effort.

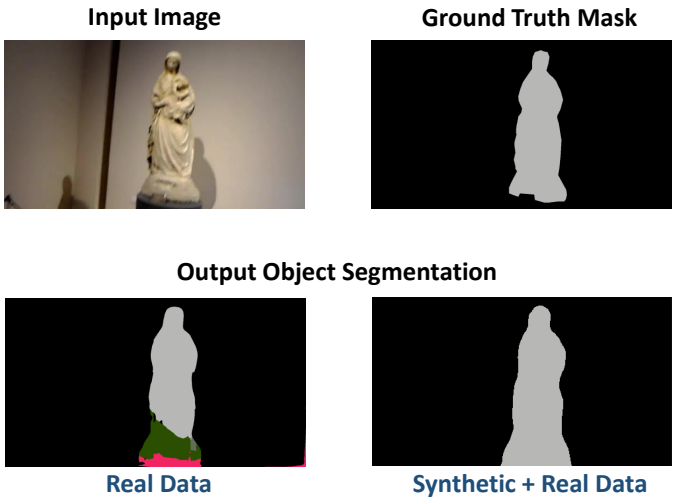

Fig. 1. Example of semantic segmentation obtained using a limited amount of labeled real data. As can be seen, adding synthetic data to the pipeline allows to greatly improve segmentation accuracy.

In this work, we investigate whether the use of both real and synthetic data can help to improve the performance of semantic segmentation algorithms in the context of cultural sites (see Figure 1). To address the analysis, we collected a new dataset comprising both real and synthetic images of 24 artworks located in 11 different environments of a cultural site located in Siracusa, Italy. The real images have been collected by different visitors using wearable devices and labeled with the semantic segmentation mask of each artwork. The synthetic images have been generated from a 3D model of the cultural site acquired using a Matterport scanner[1]. To obtain labels for the synthetic images, we have developed a tool for the automatic collection of paired images and semantic masks of artworks in the 3D model. The proposed dataset comprises 5588 real images and 24000 synthetic images labelled with segmentation masks. To the best of our knowledge, the proposed dataset is the only one comprising both labeled synthetic and real images of artworks in a cultural site.

[1]https://matterport.com/

Exploiting the proposed dataset, we performed a preliminary experimental analysis to investigate whether the use of synthetic data can help to improve the performance of semantic segmentation algorithms when applied to real data. This has been tackled by assessing 1) the effect of pre-training the segmentation algorithms on synthetic data, and then fine-tuning them on different amounts of real data; 2) the effect of combining fine-tuning with image-to-image translation to reduce the domain shift between synthetic and real images. Our analysis points out that the use of synthetic data helps to improve performance when the segmentation algorithm is used on real images. Moreover, using only a small amount of real data during training, the investigated approach coupled with image-to-image translation outperforms the results obtained by the semantic segmentation algorithm trained only on real data.

The contributions of this work are as follows:

- We propose a novel dataset comprising both synthetic and real images of 24 artworks in a cultural site. The images have been labeled with semantic masks of the depicted artworks. To the best of our knowledge, this dataset is the first of its kind. We release it publicly to encourage research on this topic. A preliminary experimental analysis to assess the usefulness of synthetic data to improve the performance of semantic segmentation on real data. The proposed analysis also provides useful baseline results to highlight the potential of the proposed dataset.

The remainder of the paper is organized as follows. Section II gives an overview of related works. The proposed dataset is presented in Section III. Section IV describes the semantic segmentation methods compared in this study. Details about experimental settings and results are given in Section V. We conclude the paper discussing insights for future works in Section VI.

II. Related Work

Our work is related to different lines of research. These include the use of egocentric vision in cultural sites, approaches to semantic segmentation, datasets for semantic segmentation and object detection, and approaches to domain adaptation. The following sections discuss the relevant works belonging to these research lines.

A. Understanding Visitor's Behavior.

Previous works have investigated the use of wearable devices to augment the fruition of cultural sites by the visitors [1]. Other works have focused on analyzing the attention of the visitors in the site [2]. The authors of [3] developed a system to localize the visitors and to understand what they pay attention to in a cultural site. This information is inferred from the analysis of egocentric videos and helps the manager of the cultural site to analyze its performance in order to improve the services offered to visitors. Object classification and artwork recognition have been exploited in [10], [11] to contextual information and perform user profiling during visit to the museum. Other works in this context have proposed to improve the user experience and the fruition of multimedia materials through semi-automatic interaction by a smart audio guide [11].

In this paper, we focus on the problem of understanding which artworks are present in egocentric images using both real and synthetic images. Since detecting the object of interest at the bounding box level can be limited in the context of cultural heritage (because the shape and size of artworks has high variability) we focus on the task of semantic object segmentation at the pixel-level.

B. Semantic Segmentation

The aim of a semantic segmentation algorithm is to assign a class label to each pixel of a given input image. Several approaches to semantic segmentation have been proposed through the years. In particular, recent works train CNNs for pixel-wise classification in a fully supervised fashion. Among the most notable approaches, the authors of [6] proposed fully convolutional networks, which generalize CNNs for image classification to perform semantic segmentation. The authors of [12] introduced SegNet, an encoder-decoder architecture based on VGG [13]. The authors of [14] investigated the use of up-sampled convolutional filters to enlarge receptive fields, spatial pyramid pooling to segment objects at multiple scales, and probabilistic graphic models to improve the localization of object boundaries. The authors of [15] introduced RefineNet to exploit multi-level features in a recursive manner to generate high-resolution semantic feature maps. The authors of [5] designed a pyramid scene parsing network (PSPNet) to spatially enhance pixel-level features using global pyramid pooling. The authors of [16] introduced a Semantic Prediction Guidance (SPG) which learns how to re-weight local features across prediction stages. The authors of [17] presented SceneAdapt, a scene-based domain adaptation approach for semantic segmentation.

In this work, we investigate whether the use of synthetic data is beneficial to improve the performance of semantic segmentation algorithms on real data. Specifically, we consider PSPNet [5] as a state-of-the-art semantic segmentation baseline for our experiments.

C. Datasets for Object Segmentation and Detection

Different datasets such as Pascal VOC [18] and COCO [19] have been proposed to explore the problem of semantic object segmentation. Despite these datasets are useful benchmarks to design algorithms, when objects have to be recognized at the instance level, as it is the case of cultural sites, it is necessary to fine-tune such algorithms with domain-specific data. Hence, is is required to collect and manually label images of the specific objects of interest, as done in [20]. This procedure is generally laborious and expensive. The availability of synthetic data would in principle enable to train semantic segmentation models at a lower cost. Synthetic datasets have been proposed in the past considering virtual 3D environments to generate semantic labels in a simple way [21], [22]. Some works have considered photo-realistic images obtained through augmented reality [23], whereas others have generated synthetic clones

Fig. 2. Examples of real images depicting the 24 artworks along with the annotated segmentation masks.

from a small amount of real data [24]. However, such datasets do not generally include both real and synthetic images of the same object annotated with semantic masks at the instance level. To fill this gap and allow to study the effectiveness of synthetic data for semantic object segmentation in cultural sites, we propose a dataset containing 29588 synthetic and real images labelled with semantic masks related to 24 artworks collected in a real cultural site.

D. Approaches to Domain Adaptation

The aim of domain adaptation techniques is to reduce the performance drop caused by the misalignment between the distribution of training test data. Such misalignment is very common in those case in which training and test data have been collected with different settings, as it is the case of synthetic and real images. Domain adaptation has been studied using both classic approaches [25], [26], [27] and deep learning [28], [29], [30], [31], [32]. A common way to reduce the domain shift when the input images are visually different, is to use image-to-image translation techniques [33] inspired by Generative Adversarial Networks (GANs) [34]. In particular, CycleGAN [33] allows to perform image translation from a source domain to a target domain without the need of paired examples, which is particularly useful for domain adaptation. In our work, we investigate whether the use of image-to-image translation through CycleGAN [33] can be useful to reduce the domain shift between synthetic and real images, when combined with PSPNet [5] for semantic segmentation.

III. Dataset

The proposed dataset contains real and synthetic images depicting 24 artworks located in 11 different environments of the Galleria Regionale Palazzo Bellomo[2]. All images are paired with semantic segmentation masks indicating the presence of artworks at pixel-level. The following sections detail the collection and labeling process for images belonging to the real and synthetic domains. The dataset is publicly available at the following link: https://iplab.dmi.unict.it/EGO-CH-OBJ-SEG/.

A. Real Images

We consider real images depicting artworks belonging to the EGO-CH dataset [20]. EGO-CH has been acquired using a Microsoft HoloLens device across two different cultural sites. For each of the cultural sites, EGO-CH contains training videos collected by instructed operators and test videos collected by subjects visiting the cultural site. We concentrate on the subset of the data acquired in the Galleria Regionale di Palazzo Bellomo cultural site and select 24 artworks for our study. Since images of EGO-CH are annotated only with bounding boxes, we have manually labelled 5588 images. Specifically we annotated 4740 images from the training set of [20] and 848 images from its test set. In this paper, 170 images of the 848 images are used for validation, whereas the remaining 678 are used for test. We used VGG annotation tool [35] to obtain all annotations. Table I reports some details about the dataset and summarizes the number of training and test images belonging to the dataset, whereas Figure 2 reports

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

examples of real images for each of the 24 artworks together with the associated segmentation masks. Class labels and the related number of manually annotated segmentation masks are reported in Table II.

B. Synthetic Images

To automatically obtain a large number of synthetic images with the related semantic mask annotations, we developed a tool based on Blender [36]. Given a 3D model of a cultural site, the tool allows to manually label the artworks in the 3D coordinate system. It then automatically generates RGB images of the artworks acquired from multiple points of view, together with the related segmentation masks. To generate the synthetic images, we used the 3D model of Palazzo Bellomo acquired in [37] (see Figure 3), which is the same scenario where real images have been acquired. Using the developed tool, we generated 12000 training images, 1200 validation images and 10800 test images (see Table I for a summary). In particular, we generated 1000 images for each considered artwork. Figure 4 shows examples of synthetic images of the 24 artworks with the related segmentation masks.

IV. Methods

We compare three approaches to perform object segmentation using real and synthetic data. All approaches consider PSPNet [5] as a baseline semantic segmentation algorithm.

A. PSPNet_R

This approach consists in training and testing PSPNet on real data. To assess the amount of labeled real data needed to obtain reasonable performance, we train the model with different amounts of training data, namely $5\%, 10\%, 25\%, 50\%$ and 100%.

B. PSPNet_S+R

This approach uses both synthetic and real data during training. The training phase is composed of two stages. In the first stage, PSPNet is trained using only synthetic data. In the second stage, we fine-tune the model obtained at the first stage using real data. The obtained model is then tested on the real images of test set. Similarly to the PSPNet_R approach, we consider different amounts of real data to train the model in the second stage: $0\%, 5\%, 10\%, 25\%, 50\%$ and 100%. The 0% indicates that the model is trained only on synthetic data (first stage training only).

C. PSPNet_S+R+CycleGAN

This approach uses image to image translation to reduce the domain shift between real and synthetic images. We have used CycleGAN [33] for this purpose. First, CycleGAN is trained to perform image to image translation between real and synthetic images. Then PSPNet is trained in two stages. In the first stage, it is trained using only synthetic data. In the second stage, real images are transformed to synthetic with CycleGAN and then the network obtained in the first stage is fine-tuned with the traslated images. The model is hence tested on real images transformed to synthetic using CycleGAN. As in the previous cases, we consider different amounts of real training data: $0\%, 5\%, 10\%, 25\%, 50\%$ and 100%.

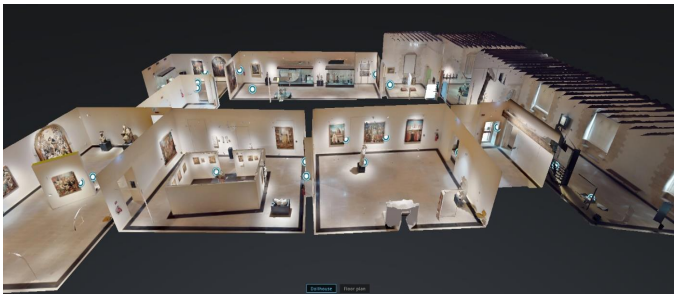

Fig. 3. 3D model of the Galleria Regionale di Palazzo Bellomo acquired in [37].

V. Experiments and Results

We train all segmentation models with a learning rate of *0.005*, weight decay of *0.0001* and momentum of *0.9*. Each model has been trained for *30* epochs. We selected the best epoch considering the Mean Intersection over Union (*Mean IoU*) on the Validation set. CycleGAN has been trained using the standard parameters suggested in [33].

Table III and Figure 5 report the performances of the three compared approaches on real test data. We evaluated our methods using standard evaluation measures adopted in semantic segmentation benchmarks [18]. The global accuracy (**Accuracy**) counts the fraction of pixels which have have been correctly classified. The per-class accuracy (**Class Accuracy**) is the mean of the accuracy values obtained independently for each class. The mean Intersection over Union (**Mean IoU**) is the average of the IoU values between predicted and ground truth segmentation masks, computed independently for each class. The frequency weighted average Jaccard Index (**FWAVACC**) is similar to **Mean IoU**, but per-class IoU values are aggregated using a weighted average based on the number of pixels in each class. While the former two evaluation measures assess the ability to roughly localize objects, the latter two measure how accurate are the predicted semantic masks at the object boundaries.

The results shown in Table III and Figure 5 highlight that using only real data allows to achieve limited performance. Indeed PSPNet_R achieves a maximum Accuracy of 83.51%, a maximum Class Accuracy of 63.15%, a maximum Mean IoU of 47.15%, and a maximum FWAVACC of 72.76%. Using only synthetic data is not sufficient to achieve satisfactory performance on real data (see PSPNet_S+R with 0% real training data in Table III). For instance PSPNet_S+R achieves a Class Accuracy of only 8.45% and a Mean IoU of only 5.50% when trained only on synthetic data. Instead, due to the fact that PSPNet_S+R+CycleGAN was trained using also images belonging to the real domain to learn the synthetic-real translation, it achieves better performance with 0% of real data

TABLE I

DETAILS ABOUT THE PROPOSED DATASET, INCLUDING THE NUMBER OF REAL AND SYNTHETIC TRAINING, VALIDATION AND TEST IMAGES.

	Resolution	#Artworks	#Environments	Segmentation Masks	Training Images	Val. Images	Test Images	All Images
Real	1280x720	24	11	5624[1]	4740 (85%)	170 (3%)	678 (12%)	5580
Synthetic	1280x720	24	11	24000	12000 (50%)	1200 (5%)	10800 (45%)	24000

[1] The number of segmentation masks is greater than the number of images due to the fact that some images have multiple annotations.

TABLE II

CLASSES RELATED TO THE DATASET WITH NUMBER OF ANNOTATION MASKS. EACH POINT OF INTEREST IS DENOTED BY AN ID IN THE FORM X.Y (E.G.,2.1) WHERE "X" DENOTES THE ENVIRONMENT IN WHICH THE ARTWORK IS LOCATED AND "Y" IDENTIFIES THE ARTWORK.

ID	Class	Annotations	ID	Class	Annotations	ID	Class	Annotations
2.1	Acquasantiera	244	5.1	Annunciazione	303	9.1	AdorazionedeiMagi	230
2.3	LastraconLeoni	248	5.2	LibroD'OreMin.	253	9.2	S.ElenaCost.eMadon.	247
3.1	MadonnainTrono	237	5.3	LastraG.Cabastida	307	9.3	TaccuinidiDisegni	212
3.2	FrammentoS.Leo	186	5.4	MadonnadelCard.	223	10.1	MartirioS.Lucia	196
4.1	MadonnainTrono	245	7.1	DisputaS.Tomm.	200	10.2	VoltodiCristo	210
4.2	MonumentoE.d'Aragona	222	7.2	TraslazioneS.Casa	279	11.1	MiracolodiS.Orsola	250
4.3	Trasf.Cristo	233	7.3	MadonnacolBam.	231	11.2	Immacolata	219
4.4	Piatti	208	8.1	ImmacolataConc.	245	21.1	StoriedellaGenesi	196

Fig. 4. Samples of synthetic images of the 24 artworks with the related segmentation masks automatically generated using the developed tool.

(Class Accuracy of 53.93% and Mean IoU of 39.43%) respect to PSPNet_S+R.

Using 10% of the real data to fine-tune the model (PSPNet_S+R), allows to obtain a Class Accuracy of 57.87% and a Mean IoU of 40.87%, which are comparable to the performances of PSPNet trained on 50% of real data (PSPNet_R), i.e. Class Accuracy 59.40% and Mean IoU 44.37%. This suggests that pre-training with synthetic data helps the model to achieve good performance on real images with less real data for the training procedure. Importantly, when pre-training on synthetic images, we need to label much less data (10% vs 50%) to obtain similar performance. In general, the curves in Figure 5 show that PSPNet_S+R needs less real training data to achieve reasonable performance, as compared to PSPNet_R.

Adding image-to-image translation to the pipeline (PSPNet_S+R+CycleGAN) allows to outperform all the other approaches even using only 5% of real data for fine-tuning. Indeed, PSPNet_S+R+CycleGAN obtains an Accuracy of 87.82% and a FWAVACC of 79.85% using 5% of real data, which outperforms the best result obtained using 100% of the real data in the PSPNet_R baseline (i.e., 83.51% and 72.76%). Moreover, adding the total amount of real data (100%) PSPNet_S+R+CycleGAN obtains a Class Accuracy of 81.22% and a Mean IoU of 68.20%, which significantly outperform

TABLE III
RESULTS OF THE COMPARED METHODS ON REAL TEST DATA.

	Real Training Data	Accuracy%	Class Accuracy%	Mean IoU%	FWAVACC%
PSPNet_R	5%	71.10	43.73	29.47	56.96
	10%	76.28	46.66	31.88	62.49
	25%	80.95	58.54	43.47	68.86
	50%	82.47	59.40	44.37	70.86
	100%	83.51	63.15	47.15	72.76
PSPNet_S+R	0%	58.32	8.45	5.50	35.60
	5%	70.18	42.38	27.06	56.54
	10%	80.23	57.87	40.87	67.71
	25%	82.14	58.55	45.03	69.90
	50%	83.07	65.09	47.80	72.51
	100%	83.70	59.02	47.06	72.00
PSPNet_S+R+CycleGAN	0%	80.52	53.93	39.43	67.77
	5%	87.82	77.90	59.49	79.85
	10%	88.58	81.67	66.19	80.45
	25%	88.62	79.91	60.93	80.57
	50%	90.23	78.72	68.25	82.44
	100%	90.23	81.22	68.20	82.77

TABLE IV
RESULTS OF PSPNET_S+R ON THE SYNTHETIC DATA

	chunk	Accuracy%	Class Accuracy%	Mean IoU%	FWAVACC%
PSPNet_S+R	0%	95.31	88.10	81.48	91.20
	5%	77.88	58.39	39.14	69.28
	10%	80.32	60.94	43.44	71.44
	25%	86.55	63.76	50.15	77.48
	50%	83.15	62.93	47.08	74.33
	100%	86.63	59.64	49.35	76.90

the results achieved by PSPNet_R (63.15% and 47.15%). The curves in Figure 5 show that PSPNet_S+R+CycleGAN achieves much better results using the same amounts of real data.

Figure 6 reports some qualitative results of the compared approaches. For each example we show the input RGB image, the ground truth segmentation mask and the results obtained by the compared methods when using different amounts of real training data. As can be observed, PSPNet_S+R produces better segmentation masks with less real training data compared to the PSPNet_R, which is trained only using real data (1^{st} example). Moreover, using only 5% of real data (second column) PSPNet_S+R+CycleGAN achieves very accurate segmentation masks as compared to the other two approaches. The second example shows two objects in the scene, but only one of them belongs to the 24 chosen artworks. Both PSPNet_R and PSPNet_S+R wrongly predict a mask for the object in the background, whereas PSPNet_S+R+CycleGAN can segment the correct artwork using only 5% of real data.

Table IV finally reports the results obtained by PSPNet_S+R on the synthetic test data when different amounts of real training data are used to fine-tune the model. As can be noted, fine-tuning greatly impacts performance according to all measures on the synthetic domain. This suggests that the model tends to overfit to the domain of synthetic data during pre-training and similarly, to the domain of real images during fine-tuning. Ideally, semantic segmentation methods should retain good performance on both domains. We leave the exploration of this issue to future works.

VI. CONCLUSION

We have considered the problem of object segmentation in cultural sites. Starting from the assumption that manually labeling images with semantic masks is expensive and time-consuming, we have studied whether the availability of large amounts of synthetic images can allow to improve performance on real images. To perform this study, we have collected and released a new dataset of real and synthetic images related to 24 artworks in a cultural site. We have hence compared three approaches to semantic segmentation which use both real and synthetic images. Results highlight that synthetic images can be beneficial to improve performance on real data, especially when coupled with image-to-image translation techniques, to reduce the domain shift arising from the two different data sources. The proposed dataset can also be used to study the problem of unsupervised domain adaptation for semantic object segmentation, which assumes the unavailability of real training data. Interestingly, all the compared approaches perform poorly under this assumption. Future work will be devoted to study unsupervised domain adaptation approaches on the proposed dataset.

ACKNOWLEDGMENT

This research is supported by MIUR - Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 - Dottorati Innovativi a Caratterizzazione Industriale XXXIII CICLO, by the project VALUE - Visual Analysis for Localization and Understanding of Environments (N. 08CT6209090207, CUP G69J18001060007) granted by PO FESR 2014/2020 -

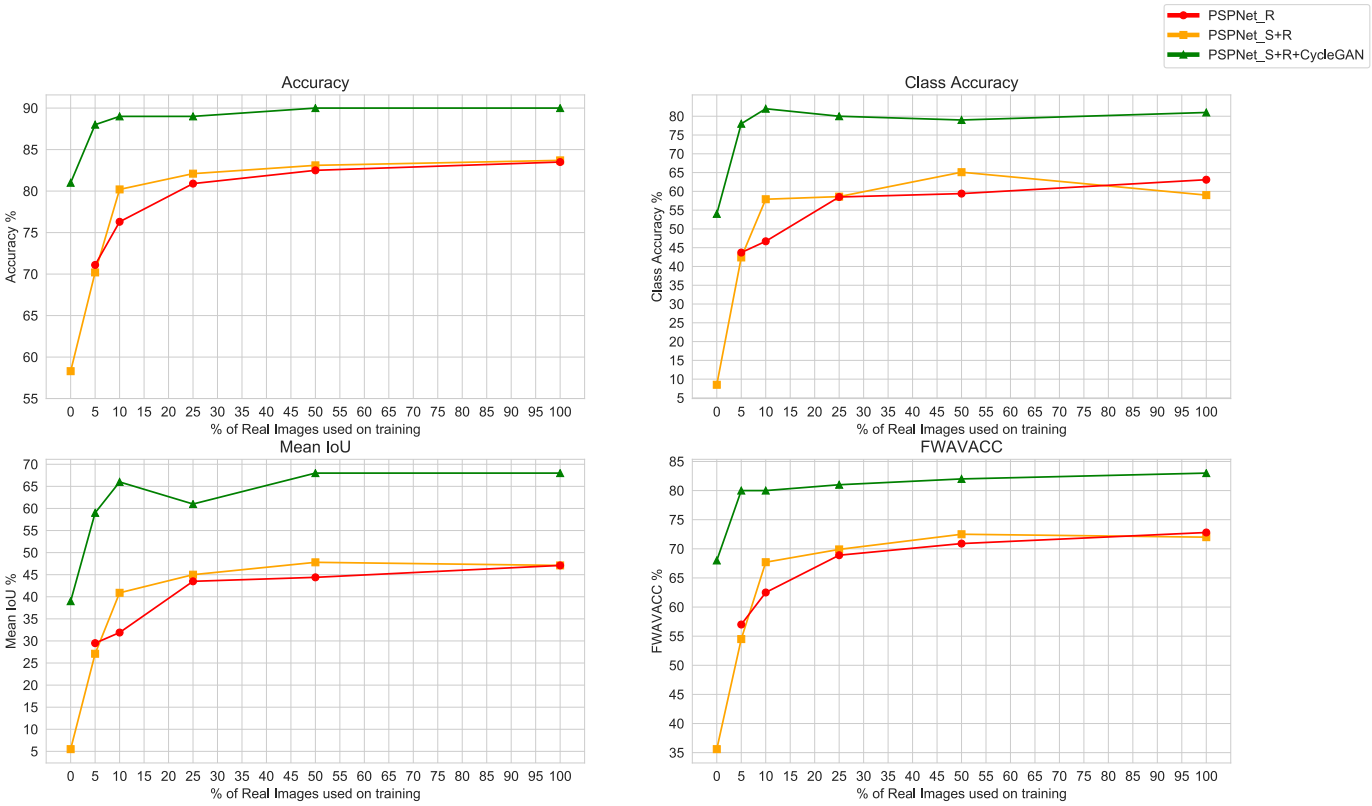

Fig. 5. Comparison using Real data, Synthetic+Real data, CycleGAN. The range of values in the y axes is different for visualization purpose.

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] R. Cucchiara, A. Del Bimbo, Visions for augmented cultural heritage experience, IEEE MultiMedia 21 (1) (2014) 74–82.

[2] A. S. Razavian, O. Aghazadeh, J. Sullivan, S. Carlsson, Estimating attention in exhibitions using wearable cameras, ICPR (2014).

[3] G. M. Farinella, G. Signorello, S. Battiato, A. Furnari, F. Ragusa, R. Leonardi, E. Ragusa, E. Scuderi, A. Lopes, L. Santo, M. Samarotto, Vedi: Vision exploitation for data interpretation, in: International Conference on Image Analysis and Processing (ICIAP), 2019.

[4] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, arXiv preprint arXiv:1703.06870 (2017).

[5] Z. Hengshuang, S. Jianping, Q. Xiaojuan, W. Xiaogang, J. Jiaya, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[6] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 640–651. doi:10.1109/TPAMI.2016.2572683.

[7] F. E. Nowruzi, P. Kapoor, D. Kolhatkar, F. A. Hassanat, R. Laganière, J. Rebut, How much real data do we actually need: Analyzing object detection performance using synthetic and real data, CoRR abs/1907.07061 (2019). arXiv:1907.07061.
URL http://arxiv.org/abs/1907.07061

[8] J. Zhang, Z. Chen, J. Huang, L. Lin, D. Zhang, Few-shot structured domain adaptation for virtual-to-real scene parsing, in: The IEEE International Conference on Computer Vision (ICCV) Workshops, 2019.

[9] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, S. Birchfield, Training deep networks with synthetic data: Bridging the reality gap by domain randomization, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018.

[10] G. Taverriti, S. Lombini, L. Seidenari, M. Bertini, A. Del Bimbo, Real-time wearable computer vision system for improved museum experience, in: ACM Multimedia, 2016, pp. 703–704.

[11] L. Seidenari, C. Baecchi, T. Uricchio, A. Ferracani, M. Bertini, A. D. Bimbo, Deep artwork detection and retrieval for automatic context-aware audio guides, TOMM 13 (3s) (2017) 35.

[12] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 2481–2495.

[13] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR abs/1409.1556 (2014).

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs., CoRR abs/1606.00915 (2016).

[15] G. Lin, A. Milan, C. Shen, I. D. Reid, Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, IEEE CVPR (2016).

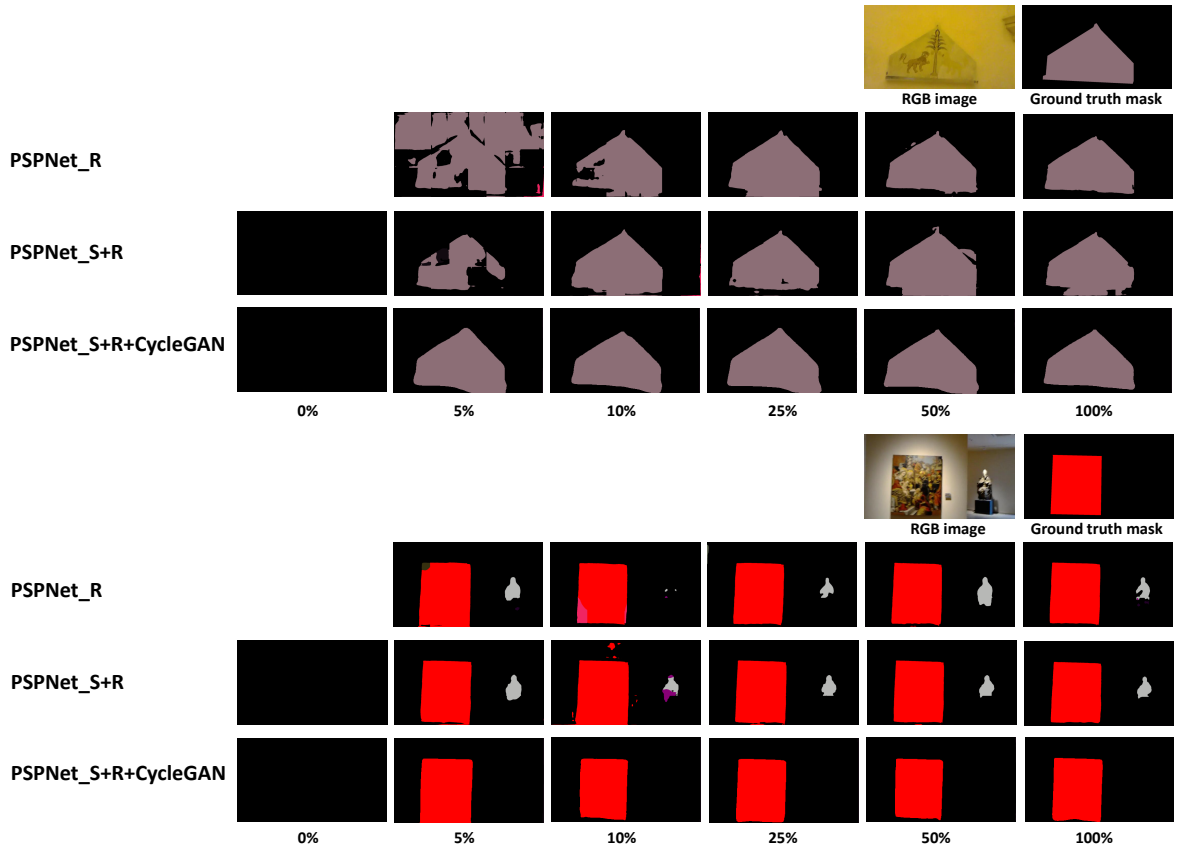

Fig. 6. Qualitative results of the compared approaches using different amount of real data in the training phase. 0% denotes that the model has been trained using only synthetic images.

[16] B. Cheng, L.-C. Chen, Y. Wei, Y. Zhu, Z. Huang, J. Xiong, T. Huang, W.-M. Hwu, H. Shi, Spgnet: Semantic prediction guidance for scene parsing, 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 5217–5227.

[17] D. Di Mauro, A. Furnari, G. Patanè, S. Battiato, G. M. Farinella, Sceneadapt: Scene-based domain adaptation for semantic segmentation using adversarial learning, Pattern Recognition Letters 136 (2020).

[18] M. Everingham, L. Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vision 88 (2) (2010) 303–338. doi:10.1007/s11263-009-0275-4.

[19] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: ECCV, 2014.

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

[21] S. R. Richter, V. Vineet, S. Roth, V. Koltun, Playing for data: Ground truth from computer games, ArXiv abs/1608.02192 (2016).

[22] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, A. M. Lopez, The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes, in: IEEE CVPR, 2016.

[23] H. Abu Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, C. Rother, Augmented reality meets computer vision: Efficient data generation for urban driving scenes, Int. J. Comput. Vision 126 (9) (2018) 961–972. doi:10.1007/s11263-018-1070-x.

[24] A. Gaidon, Q. Wang, Y. Cabon, E. Vig, Virtual worlds as proxy for multi-object tracking analysis, IEEE Conference on Computer Vision and Pattern Recognition (2016).

[25] B. Kulis, K. Saenko, T. Darrell, What you saw is not what you get: Domain adaptation using asymmetric kernel transforms, CVPR 2011 (2011) 1785–1792.

[26] R. Gopalan, R. Li, R. Chellappa, Domain adaptation for object recognition: An unsupervised approach, 2011 International Conference on Computer Vision (2011) 999–1006.

[27] W. Li, Z. Xu, D. Xu, D. Dai, L. V. Gool, Domain generalization and adaptation using low rank exemplar svms., IEEE Trans. Pattern Anal. Mach. Intell. 40 (5) (2018) 1114–1127.

[28] M. Long, Y. Cao, J. Wang, M. I. Jordan, Learning transferable features with deep adaptation networks, in: Proceedings of the 32nd ICML, 2015.

[29] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, W. Li, Deep reconstruction-classification networks for unsupervised domain adaptation, ArXiv abs/1607.03516 (2016).

[30] O. Sener, H. O. Song, A. Saxena, S. Savarese, Learning transferrable representations for unsupervised domain adaptation, in: NIPS, 2016.

[31] D. Li, Y. Yang, Y.-Z. Song, T. M. Hospedales, Deeper, broader and artier domain generalization, 2017 IEEE International Conference on Computer Vision (ICCV) (2017) 5543–5551.

[32] F. M. Carlucci, L. Porzi, B. Caputo, E. Ricci, S. R. Bulò, Autodial: Automatic domain alignment layers, 2017 IEEE International Conference on Computer Vision (ICCV) (2017) 5077–5085.

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

[34] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, Generative adversarial nets, in: NIPS, 2014.

[35] A. Dutta, A. Gupta, A. Zissermann, VGG image annotator (VIA), http://www.robots.ox.ac.uk/ vgg/software/via/ (2016).

[36] B. O. Community, Blender - a 3d modelling and rendering package (2018).
URL http://www.blender.org

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