Unsupervised Domain Adaptation for 6DOF Indoor Localization

Daniele Di Mauro[1][a], Antonino Furnari[1][b], Giovanni Signorello[2][c]
and Giovanni Maria Farinella[1][d]
[1]*Department of Mathematics and Computer Science, University of Catania, Piazza Università 2, Catania, Italy*
[2]*CUTGANA, University of Catania, Piazza Università 2, Catania, Italy*
{*dimauro, furnari, gfarinella*}*@dmi.unict.it, g.signorello@unict.it*

Keywords: Domain Adaptation, Localization, 6DOF, Camera Pose Estimation.

Abstract: Visual Localization is gathering more and more attention in computer vision due to the spread of wearable cameras (e.g. smart glasses) and to the increase of general interest in autonomous vehicles and robots. Unfortunately, current localization algorithms rely on large amounts of labeled training data collected in the specific target environment in which the system needs to work. Data collection and labeling in this context is difficult and time-consuming. Moreover, the process has to be repeated when the system is adapted to a new environment. In this work, we consider a scenario in which the target environment has been scanned to obtain a 3D model of the scene suitable to generate large quantities of synthetic data automatically paired with localization labels. We hence investigate the use of Unsupervised Domain Adaptation techniques exploiting labeled synthetic data and unlabeled real data to train localization algorithms. To carry out the study, we introduce a new dataset composed of synthetic and real images labeled with their 6-DOF poses collected in four different indoor rooms which is available at https://iplab.dmi.unict.it/EGO-CH-LOC-UDA. A new method based on self-supervision and attention modules is hence proposed and tested on the proposed dataset. Results show that our method improves over baselines and state-of-the-art algorithms tackling similar domain adaptation tasks.

1 INTRODUCTION

The topic of visual localization is central in Computer Vision due to the increasing use of smartphones and smart glasses, as well as due to its applicability in contexts such as autonomous vehicles and robotics. Being able to locate the position of a device in an environment is an important and often necessary ability to solve other complex tasks such as understanding which future actions are possible, determining how to reach specific places, or providing assistance to the user (Häne et al., 2017; Gupta et al., 2017; Ragusa et al., 2020b). While visual localization can be performed both in indoor and outdoor environments, it is particularly relevant in indoor scenarios where GPS systems can not be used and infrastructures such as WI-FI or bluetooth receivers are not always feasible to install (e.g., museums, archaeological sites). Visual localization can be performed at different levels of granularity, depending on the application. For example, a navigation system may require the precise location of the user within a building, whereas a wearable contextual assistant may need to recognize only in which room a certain action of the user is taking place (Ortis et al., 2017).

In this work, we focus on accurate indoor visual localization through the estimation of the 6 Degrees of Freedom (6-DOF) pose of the camera carried by the user. Popular approaches to tackle this task require the collection and labeling of large datasets of images in the target environment (Kendall et al., 2015; Melekhov et al., 2017b). While images can be easily collected with a moving camera, labeling is performed by attaching a 6-DOF pose to each frame using structure from motion techniques (Schönberger and Frahm, 2016; Schönberger et al., 2016b), which often requires the manual intervention of experts. As a result, creating datasets for training visual localization algorithms is time-consuming and expensive.

In this work, we investigate the use of Unsupervised Domain Adaptation approaches (Tzeng et al., 2017; Hoffman et al., 2018) to exploit labeled syn-

[a] https://orcid.org/0000-0002-4286-2050
[b] https://orcid.org/0000-0001-6911-0302
[c] https://orcid.org/0000-0002-5140-4975
[d] https://orcid.org/0000-0002-6034-0432

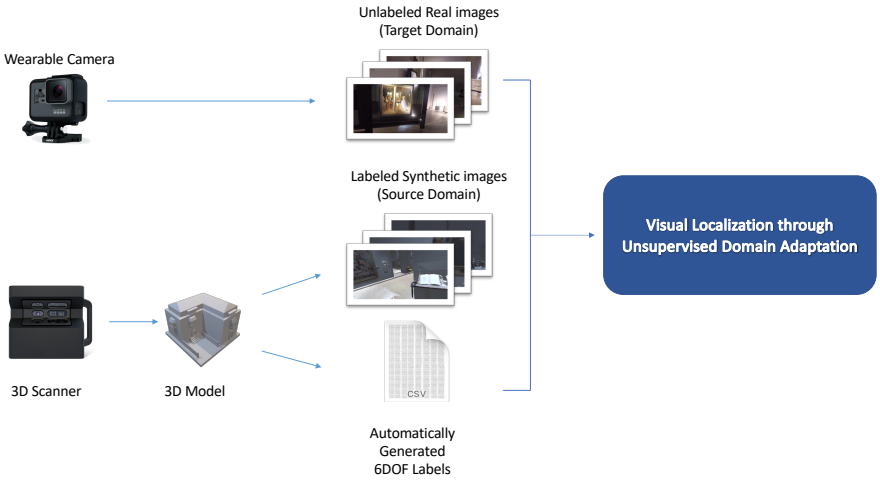

Figure 1: A scheme of the considered domain adaptation pipeline. We use weareable cameras to collect unlabeled real images ad a scanner to obtain a 3D model of the environment. The 3D model is used to generate synthetic images which are automatically labeled with their 6-DOF camera poses. This data is used to train a visual localization approach through unsupervised domain adaptation techniques. At test time, the localization algorithm is required to work on real images. Note that this pipeline does not require any manual annotation.

thetic data and unlabeled real data for the training of localization algorithms. Specifically, we consider a scenario in which a 3D model of the environment has been acquired using a scanner such as Matterport 3D[1]. The 3D model can be used to simulate an agent navigating the environment and automatically obtain synthetic labeled images as proposed in (Orlando et al., 2020). We also assume that real images of the same environment have been acquired but not labeled. Since labeling is the most expensive step, acquiring this data is significantly less expensive. We hence investigate the use of unsupervised domain adaptation techniques that take the labeled synthetic data and the unlabeled real data as input to learn how to perform localization in real test data. Figure 1 shows a scheme of the considered pipeline. The main contributions of this work are as follows:

1. we investigate the novel task of unsupervised domain adaptation for 6DOF camera pose estimation for visual localization in indoor scenarios;
2. we propose a first dataset to study the considered problem. The dataset has been acquired in 4 different rooms of a cultural heritage site and contains synthetic and real data which has been labeled with 6-DOF camera poses for algorithm evaluation and comparison. We publicly release the dataset to encourage research in this domain;
3. we propose a new approach based on self-supervision and attention modules that outperforms baselines and state-of-the-art approaches.

[1]https://matterport.com/

2 RELATED WORK

Visual Localization. Visual localization approaches based on monocular RGB images can be grouped in two major classes: methods based on classification and methods based on camera pose estimation. In turn, camera pose estimation can be obtained through image retrieval, direct regression or exploiting 2D-3D matchings. Classification based localization is obtained through the discretization of the space in cells and training of a classifier to assign the correct cell to a given image. Some classification-based methods use Bag of Words representations (Ishihara et al., 2017; Cao and Snavely, 2013), whereas others are based on CNNs (Weyand et al., 2016). These approaches are not designed to estimate the accurate position and orientation of the camera. The authors of (Furnari et al., 2016) recognize user-specified personal locations from first person videos by considering visual localization as an "open-set" classification problem where locations of interest for the user have to be recognized, while the ones not defined by the user have to be rejected. Other works (Starner et al., 1998) threat localization as a "closed-set" classification problem in which only known rooms are considered.

Approaches based on image retrieval (Sattler et al., 2016; Torii et al., 2015; Weyand et al., 2016) approximate the location of a test image assigning it the pose of the most similar one in the training set.

Approaches to localization through regression from monocular images are based on absolute pose

regression or relative pose prediction. In the first class of approaches, a CNN is trained to predict camera poses directly from input images. Popular approaches based on direct camera pose regression first extract features using a backbone CNN, then embed features in a high-dimensional space. The learned embedding space is hence used to regress the camera pose (Kendall et al., 2015; Melekhov et al., 2017a; Radwan et al., 2018). The second line of approaches are based on relative camera pose regression. These methods have the advantage that a relative pose regression network can be trained to generalize on multiple scenes. Such approaches try to predict the pose of a test image relative to one or more training images (Balntas et al., 2018; Saha et al., 2018).

Approaches based on 2D-3D matchings are currently the state of the art for localization. These methods rely on establishing matchings between 2D pixels positions in the image and 3D scene coordinates. Matchings are established by using a descriptor matching algorithm or by regressing 3D coordinates from image patches (Brachmann and Rother, 2018; Taira et al., 2018). Despite their accuracy these methods currently do not scale well to city-scale environments, especially when they have to be executed in real time.

In this work, we focus on approaches based on direct camera pose estimation and show how they can be extended with unsupervised domain adaptation approaches.

Unsupervised Domain Adaptation. Despite huge amounts of unlabeled data are generated and made available in many domains, the cost of data labeling is still high. To avoid this pitfall, several alternative solutions have been proposed in order to exploit huge amounts of unlabeled data for training. We focus on Unsupervised Domain Adaptation (Ganin and Lempitsky, 2015; Gong et al., 2012) which leverages labeled data available in a source domain to improve performance in an unlabeled target domain. In general, we assume that the label set defined on the target domain is identical to the one defined on the source domain, whereas source and target domains are assumed to be related but not identical. When the distributions of the source and the target domains do not match, performance can be poor at testing time. This difference in distribution is called *domain-shift* (Saenko et al., 2010). A main cause of domain shift is the change in data acquisition conditions, e.g. background, position, use of images of different kind, e.g. photographs vs clip-art. There are several algorithms for Domain Adaptation, both based on handcrafted features and based on deep learning. In the last years, architectures designed to tackle domain adaptation are increasing for different tasks such as classification(Tzeng et al., 2017; Hoffman et al., 2018), semantic segmentation (Di Mauro et al., 2020; Hoffman et al., 2018; Ragusa et al., 2020a), and object detection (Pasqualino et al., 2020). A first class of methods is based on the minimization of discrepancy measures, such as the Maximum Mean Discrepancy (MMD) defined between corresponding activations from two streams of a Siamese architecture (Long et al., 2017; Rozantsev et al., 2018). Some approaches use adversarial losses to learn domain-invariant representations. We distinguish between adversarial discriminative models which encourage domain confusion through an adversarial objective with respect to a domain discriminator (Ganin et al., 2016; Tzeng et al., 2017), and adversarial generative models which combine the discriminative model with a generative component based on GANs (Goodfellow et al., 2014; Hoffman et al., 2018) . Other methods are based on data reconstruction through an encoder-decoder architecture. These approaches jointly learn source label predictions and unsupervised target data reconstruction alternating between unsupervised and supervised training (Zeiler et al., 2010; Di Mauro et al., 2020; Ghifary et al., 2016).

In this work, we reduce the discrepancy between the source and target domains, aligning their features via self-supervised tasks.

Generation of Synthetic Data. Recent advances in computer graphics and game engines allow to generate photo-realistic virtual worlds with realistic and physically consistent events and actions. This has increased the use of virtual worlds for synthetic data generation in conjunction with domain adaptation models. Most popular virtual worlds have been especially designed for autonomous driving applications such as SYNTHIA (Ros et al., 2016) or for robot agents training such as HABITAT (Savva et al., 2019). In most cases, the synthetic data is used alongside the real data during the training of the models. Domain Adaptation techniques may further assist in adapting the trained model with virtual (source) data to real (target) data, especially when labeled real data (Ros et al., 2016) are not available, or scarce. In (Orlando et al., 2020) a tool has been developed to collect synthetic visual data for localization purpose and to automatically tag data. The tool simulates a virtual agent navigating the 3D model and automatically captures images along with the associated camera poses and semantic masks showing the location of the artworks. In this work we use the aforementioned tool to produce the synthetic part of our dataset.

Figure 2: Scans of the considered rooms of the cultural site.

Table 1: Dataset Splits.

	Real			Simulated		
	Train	**Test**	**Val**	**Train**	**Test**	**Val**
Room 1	561	373	252	8221	4078	4154
Room 2	562	305	233	6299	3280	3081
Room 3	405	253	321	10493	3204	3493
Room 4	128	88	65	2049	1096	989
Total	1656	1019	871	27062	11658	11718

3 DATASET

The dataset was acquired in the Bellomo Palace Regional Gallery, which is a museum sited in Syracuse, Italy. We recorded 10 videos of subjects visiting the museum with a GoPro Hero 4 wearable camera and Matterport 3D to obtain a 3D scan of the environment. We collected data in 4 rooms of the building (Figure 2). The considered rooms offer a good representation of what can be found in a museum because they contain statues, paintings and items behind display cases. We used the tool proposed in (Orlando et al., 2020) to simulate a virtual agent navigating the 4 rooms from the 3D model of the environment. Specifically, we produced 4 different videos of simulated navigations. Due to the simulated nature of the navigations, the generated images were automatically associated to their 6-DOF camera pose.

To label real images collected through the GoPro Hero 4 camera, we reconstructed each room using COLMAP (Schönberger and Frahm, 2016). The reconstructed models have then been aligned to the related Matterport 3D models using a reference set of localized images obtained through Matterport 3D. The alignment has been performed using the Manhattan world alignment functionality of COLMAP. Through this procedure each real and synthetic image has been annotated with 6-DOF camera pose including 3 spatial coordinates and 4 values to represent the rotation as a quaternion. Figure 3 shows some examples of the acquired data. Real and simulated images have been split into train, test and validation sets as shown in Table 1. These splits have been defined as follows: for synthetic data, all frames extracted from first navigation video have been used for test, whereas frames from the second and third videos have been used for train and frames from the fourth video have been used for validation; for the real data, all frames extracted from the first to sixth video are in the train split, frames from the seventh and eighth videos have been used for test split and frames from the nineth and tenth videos have been used for validation split.

4 METHOD

The proposed method learns to perform image-based localization following an unsupervised domain adaptation approach, i.e, only labeled synthetic images and unlabeled real images are used at training time, whereas real labeled images are used for evaluation only. Figure 4 shows a schema of the proposed model which includes a ResNet backbone for feature extraction, a regression branch with an attention module and a self-supervised branch. The regression branch is designed to predict the 6-DOF camera pose directly from the input image as in PoseNet (Kendall et al., 2015), whereas the self-supervised branch encourages the learned features to be consistent across domains. The regression branch is composed of an attention module and a regression head. The dual attention module follows the design proposed in (Fu et al., 2019) and is composed of: a Position Atten-

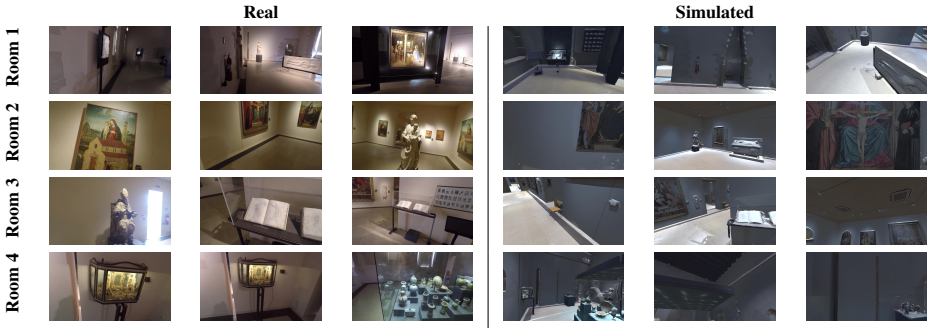

Figure 3: Three examples of real (left) and synthetic (right) images from each room.

tion Module (PAM) and a Channel Attention Module (CAM). PAM selectively aggregates features at each position through a weighted sum of features at all positions. In this way, similar features will be related to each other regardless of their spatial distances. As explained in (Fu et al., 2019), given a local feature map $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$, the PAM module uses convolution layers to generate three new feature maps $\mathbf{B},\mathbf{C},\mathbf{D} \in \mathbb{R}^{C\times H\times W}$. The values of a spatial attention map $\mathbf{S} \in \mathbb{R}^{N\times N}$, where $N = H \times W$, are computed as:

$$s_{ji} = \frac{exp(B_i \cdot C_j)}{\sum_{i=1}^{N} exp(B_i \cdot C_j)} \tag{1}$$

where i and j index the spatial locations of $\mathbf{B}$ and $\mathbf{C}$. The final values of the output $\mathbf{E} \in \mathbb{R}^{C\times H\times W}$ are computed as:

$$E_j = \alpha \sum_{i=1}^{N} (s_{ji} D_i) + A_j \tag{2}$$

where α is a scaling parameter initialized as 0 and optimized at training time. The CAM module selectively emphasizes interdependent channel maps by integrating associated features among all channel maps. The values of the attention map $\mathbf{X} \in \mathbb{R}^{C\times C}$ are directly computed from the original feature maps $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$ as follows:

$$x_{ji} = \frac{exp(A_i \cdot A_j)}{\sum_{i=1}^{C} exp(A_i \cdot A_j)} \tag{3}$$

where x_{ji} measures the i^{th} channel's impact on the j^{th} channel. The values of the final output $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ are computed as follows:

$$F_j = \beta \sum_{i=1}^{C} (x_{ji} A_i) + A_j \tag{4}$$

where β is initialized to 0 and optimized at training time. The outputs of the two modules are finally summed.

To learn features which are consistent across domains, the self-supervised branch introduces an auxiliary task which can be performed on both domains simultaneously without the need for supervised labels. We considered 2 different tasks: the classification of affine transformation applied to the image and the detection of the presence of an overlap between a pair of images. For the affine transformation task, we pre-computed 36 affine transformations. Each affine transformation is defined by a combination of angle, translation and shear. The classes are a subset of classes generated varying the rotation angle between $0°$ and $90°$, the translation parameter between -10 pixels and 10 pixels at a step of 5 pixels, and the shear between -10 pixels and 10 pixels at a step of 5 pixels. In our tests, a higher number of classes does not seem to affect results. At training time, we apply a random transformation to the input image. The self supervised module is hence trained to recognize which affine transformation was applied. This self-supervised task encourages the backbone to extract features which allow to recognize geometrical variations.

The overlap detection task is performed with a siamese network with a shared backbone, which classifies a pair of images as overlapping or not overlapping. The task encourages to learn features which allow to understand if there is a common part between two different images. Overlapping image pairs are identified at batch level. In this case, a batch has to be composed by an even number of images. A pair is labeled as overlapping if it is detected at least one match using the OpenCV Flann based matcher.

We train the model with the MSE loss to regress position and orientation, while the cross-entropy loss is used to train the self-supervised classification task and detection overlap (yes/no binary classification). At each training iteration, the backbone weights are updated first on the self-supervised task, then on the localization task.

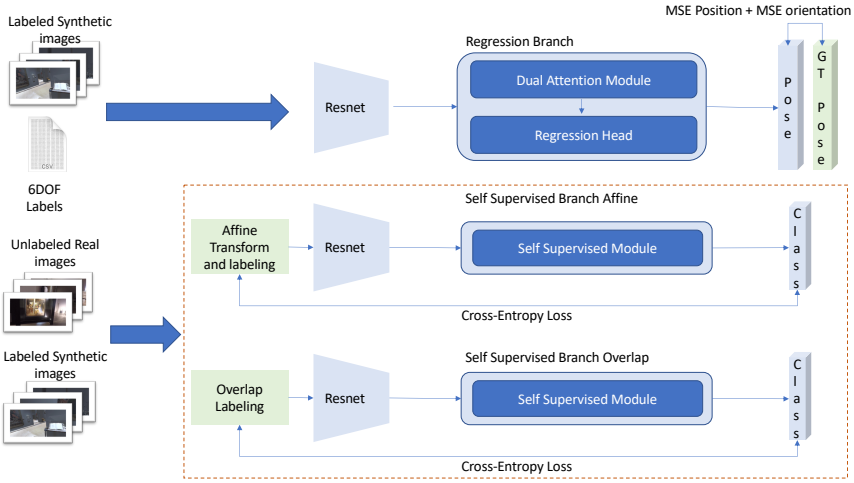

Figure 4: The proposed approach composed of a ResNet backbone, a regression branch with a dual attention module and a self supervised head to perform unsupervised domain adaptation. Backbone weights are shared. For the self-supervised task which detects overlap, batch size has to be even.

5 EXPERIMENTAL SETTINGS AND RESULTS

We compare our method with the following baselines:

1. a SIFT-based image retrieval approach, "Vote And Verify" implemented in the COLMAP software (Schönberger et al., 2016a);
2. the PoseNet approach to localization (Kendall et al., 2015);
3. two naive domain adaptation methods based on CylcleGAN (Zhu et al., 2017): transforming real images to look like synthetic ones and using PoseNet trained over synthetic images. We also considered transforming synthetic images to look like real, train a new model, and test over real images;
4. ADDA, a domain adaptation method which focuses on features adaptation (Tzeng et al., 2017);
5. CyCada, a domain adaptation method which focuses on adaptation on both features and input data (Hoffman et al., 2018).

All the networks have been trained for 500 epochs on images of each room. The best performing epoch on the test set has been chosen as a form of early stopping. The method is implemented using PyTorch and tested on a system with two NVIDIA GeForce Titan X Pascal GPUs with 12GB GDDR5X RAM. We performed experiments training and testing on the same domain, as well as training on simulated images and testing on real images. We trained each model independently on each room and averaged the results obtained across rooms.

Table 2 reports the results obtained training and testing methods on real data. In the table, we show average position error, average quaternion error and average Euler angle errors. We can see that best results are obtained using the classic image retrieval technique methods. Our method shows better results than PoseNet. We think that the use of the attention head and self-supervised task improves the embedding space. The classic image retrieval "Vote And Verify" is still better performing than our method, probably because the big vocabulary used (1 million visual words) results in a better way to index images. Table 3 reports the results obtained training and testing approaches on the simulated domain. Finally, Table 4 reports results of the domain adaptation methods. Specifically we compare our method to style transfer at test time, where real images are transformed to "synthetic", style transfer at train time, where PoseNet is trained with synthetic data styled as real, ADDA (Tzeng et al., 2017) and CyCada (Hoffman et al., 2018). These latter two approaches achieve even worse results than model when has not been adapted. Both these approaches fail to solve the problem in this formulation. Our method reduces the position error by 44.08% in the best test in comparison to PoseNet, using the overlap detection task and has similar results on orientation, and reduces the error by 5.97% in comparison to style transfer at training time, it is worth to note that our method is faster at training time, it takes one third of time to train our method instead of training CycleGAN, translate the images and finally training PoseNet.

Table 2: Real vs. Real on first room, average results over four models trained and tested on each room.

	Position Err.	Quaternion Err.	α Err.	β Err.	γ Err.
PoseNet (beta 100)	1.43m	28.54°	32.23°	9.48°	10.12°
Vote-and-Verify	**0.82m**	**22.78°**	27.98°	**7.83°**	**8.23°**
Our w/affine	1.26m	27.73°	**27.53°**	8.90°	9.53°
Our w/overlap	1.28m	27.40°	28.75°	9.12°	9.05°

Table 3: Simulated vs. Simulated average results over four models trained and tested on each room.

	Position Err.	Quaternion Err.	α Err.	β Err.	γ Err.
PoseNet (beta 100)	1.28m	28.72°	62.11°	19.80°	54.24°
Vote-and-Verify	**0.54m**	**14.50°**	**50.66°**	**10.43°**	**43.72°**
Our w/affine	0.89m	43.71°	69.87°	29.59°	66.33°
Our w/overlap	0.82m	43.97°	72.96°	29.28°	69.32°

Table 4: Real vs. Simulated average results over four models trained and tested on each room.

	Position Err.	Quaternion Err.	α Err.	β Err.	γ Err.
PoseNet (beta 100)	3.38m	114.70°	108.38°	45.77°	83.24°
Vote-and-Verify	3.39m	**100.58°**	112.72°	41.62°	61.94°
CycleGan + PoseNet (Test)	2.68m	113.42°	88.82°	41.44°	98.84°
CycleGan + PoseNet (Train)	2.01m	101.22°	99.72°	39.64°	84.61°
ADDA	4.54m	131.07°	80.84°	44.29°	113.73°
CyCada	4.15m	116.56°	109.06°	49.01°	**32.72°**
Our w/affine	1.96m	111.55°	**77.31°**	**32.72°**	109.27°
Our w/overlap	**1.89m**	109.06°	96.30°	36.81°	94.73°

6 CONCLUSION

In this work we have proposed the new problem of unsupervised domain adaptation for 6-DOF localization. We collected a new dataset in a cultural site which is available at https://iplab.dmi.unict.it/EGO-CH-LOC-UDA/. We have introduced a method to exploit synthetic data to learn to regress pose in indoor environments. Results show that the problem is still open. In particular, the results of absolute pose estimation are still underperforming compared to classical image retrieval approaches and domain adaptation approaches are accordingly affected by this. Relative pose estimation could be investigated in the future as a way to reduce localization error as well as models for 2D-3D matching.

ACKNOWLEDGEMENT

This research is supported by XENIA Progetti s.r.l., by project VALUE - Visual Analysis for Localization and Understanding of Environments (N. 08CT6209090207 - CUP G69J18001060007) - PO FESR 2014/2020 - Azione 1.1.5., by Piano della Ricerca 2016-2018 linea di Intervento 1 CHANCE - 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

Balntas, V., Li, S., and Prisacariu, V. (2018). Relocnet: Continuous metric learning relocalisation using neural nets. In *European Conference on Computer Vision (ECCV)*, pages 751–767.

Brachmann, E. and Rother, C. (2018). Learning less is more-6d camera localization via 3d surface regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4654–4662.

Cao, S. and Snavely, N. (2013). Graph-based discriminative learning for location recognition. In *ieee conference on computer vision and pattern recognition*, pages 700–707.

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

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., and Lu, H. (2019). Dual attention network for scene segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154.

Furnari, A., Farinella, G. M., and Battiato, S. (2016). Recognizing personal locations from egocentric videos. *IEEE Transactions on Human-Machine Systems*, 47(1):6–18.

Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Ghifary, M., Kleijn, W. B., Zhang, M., Balduzzi, D., and Li, W. (2016). Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613.

Gong, B., Shi, Y., Sha, F., and Grauman, K. (2012). Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073. IEEE.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Gupta, S., Davidson, J., Levine, S., Sukthankar, R., and Malik, J. (2017). Cognitive mapping and planning for visual navigation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2616–2625.

Häne, C., Heng, L., Lee, G. H., Fraundorfer, F., Furgale, P., Sattler, T., and Pollefeys, M. (2017). 3d visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection. *Image and Vision Computing*, 68:14–27.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., and Darrell, T. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In

International conference on machine learning, pages 1989–1998. PMLR.

Ishihara, T., Vongkulbhisal, J., Kitani, K. M., and Asakawa, C. (2017). Beacon-guided structure from motion for smartphone-based navigation. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 769–777. IEEE.

Kendall, A., Grimes, M., and Cipolla, R. (2015). Posenet: A convolutional network for real-time 6-dof camera relocalization. In *IEEE international conference on computer vision*, pages 2938–2946.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. (2017). Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR.

Melekhov, I., Ylioinas, J., Kannala, J., and Rahtu, E. (2017a). Image-based localization using hourglass networks. In *IEEE International Conference on Computer Vision*, pages 879–886.

Melekhov, I., Ylioinas, J., Kannala, J., and Rahtu, E. (2017b). Relative camera pose estimation using convolutional neural networks. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 675–687. Springer.

Orlando, S. A., Furnari, A., and Farinella, G. M. (2020). Egocentric visitor localization and artwork detection in cultural sites using synthetic data. *Pattern Recognition Letters*, 133:17–24.

Ortis, A., Farinella, G. M., D'Amico, V., Addesso, L., Torrisi, G., and Battiato, S. (2017). Organizing egocentric videos of daily living activities. *Pattern Recognition*, 72:207–218.

Pasqualino, G., Furnari, A., Signorello, G., and Farinella, G. M. (2020). Synthetic to real unsupervised domain adaptation for single-stage artwork recognition in cultural sites. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE.

Radwan, N., Valada, A., and Burgard, W. (2018). Vlocnet++: Deep multitask learning for semantic visual localization and odometry. *IEEE Robotics and Automation Letters*, 3(4):4407–4414.

Ragusa, F., Di Mauro, D., Palermo, A., Furnari, A., and Farinella, G. M. (2020a). Semantic object segmentation in cultural sites using real and synthetic data. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE.

Ragusa, F., Furnari, A., Battiato, S., Signorello, G., and Farinella, G. M. (2020b). EGO-CH: Dataset and fundamental tasks for visitors behavioral understanding using egocentric vision. *Pattern Recognition Letters*, 131:150–157.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *IEEE conference on computer vision and pattern recognition*, pages 3234–3243.

Rozantsev, A., Salzmann, M., and Fua, P. (2018). Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):801–814.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. (2010). Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226.

Saha, S., Varma, G., and Jawahar, C. (2018). Improved visual relocalization by discovering anchor points. *arXiv preprint arXiv:1811.04370*.

Sattler, T., Havlena, M., Schindler, K., and Pollefeys, M. (2016). Large-scale location recognition and the geometric burstiness problem. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1582–1590.

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., and Batra, D. (2019). Habitat: A platform for embodied ai research. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Schönberger, J. L. and Frahm, J.-M. (2016). Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schönberger, J. L., Price, T., Sattler, T., Frahm, J.-M., and Pollefeys, M. (2016a). A vote-and-verify strategy for fast spatial verification in image retrieval. In *Asian Conference on Computer Vision (ACCV)*.

Schönberger, J. L., Zheng, E., Pollefeys, M., and Frahm, J.-M. (2016b). Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*.

Starner, T., Schiele, B., and Pentland, A. (1998). Visual contextual awareness in wearable computing. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No. 98EX215)*, pages 50–57. IEEE.

Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., and Torii, A. (2018). Inloc: Indoor visual localization with dense matching and view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209.

Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., and Pajdla, T. (2015). 24/7 place recognition by view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *IEEE conference on computer vision and pattern recognition*, pages 7167–7176.

Weyand, T., Kostrikov, I., and Philbin, J. (2016). Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*, pages 37–55. Springer.

Zeiler, M. D., Krishnan, D., Taylor, G. W., and Fergus, R. (2010). Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition*, pages 2528–2535. IEEE.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE international conference on computer vision*, pages 2223–2232.