REVERSE DOMAIN ADAPTATION FOR INDOOR CAMERA POSE REGRESSION

Debaditya Acharya[1,*], Kourosh Khoshelham [2,3]

[1] Geospatial Science, RMIT University, City Campus, Melbourne, Victoria, 3000, Australia
[2] Department of Infrastructure Engineering, The University of Melbourne, Parkville, Victoria 3010, Australia
[3] Building 4.0 CRC, Caulfield East, Victoria 3145, Australia
debaditya.acharya@rmit.edu.au, k.khoshelham@unimelb.edu.au

KEY WORDS: Domain adaptation, GAN, deep learning, Indoor localization, 3D building models, camera pose regression

ABSTRACT:

Synthetic images have been used to mitigate the scarcity of annotated data for training deep learning approaches, followed by domain adaptation that reduces the gap between synthetic and real images. One such approach is using Generative Adversarial Networks (GANs) such as CycleGAN to bridge the domain gap where the synthetic images are translated into real-looking synthetic images that are used to train the deep learning models. In this article, we explore the less intuitive alternate strategy for domain adaption in the reverse direction; i.e., real-to-synthetic adaptation. We train the deep learning models with synthetic data directly, and then during inference we apply domain adaptation to convert the real images to synthetic-looking real images using CycleGAN. This strategy reduces the amount of data conversion required during the training, can potentially generate artefact-free images compared to the harder synthetic-to-real case, and can improve the performance of deep learning models. We demonstrate the success of this strategy in indoor localisation by experimenting with camera pose regression. The experimental results indicate an improvement in localisation accuracy is observed with the proposed domain adaptation as compared to the synthetic-to-real adaptation.

1. INTRODUCTION

Deep learning has been successfully applied in several computer vision tasks, where the success of these algorithms is dependent on the availability of large volumes of manually annotated training images. However, this requirement becomes the major constraint where those annotated images are scarce. The lack of annotated data has motivated researchers to use existing data from the source domain of synthetic images to perform similar tasks in the target domain of real images. These synthetic images are usually simulated from the virtual worlds that are created using 3D object models (textured or texture-less). However, due to the difference (gap) between the real and synthetic images, the deep learning models trained with synthetic images directly do not perform well on real images.

Domain adaptation in this context is the learning of the domain gap, to reduce the dataset bias created due to the image differences. Generative models such as CycleGAN (Zhu et al., 2017) have been used to reduce the gap between the real and the synthetic images, without requiring correspondence between the synthetic and real images. CycleGAN learns the data distribution of the two domains in an unsupervised manner to generate novel samples and has been widely used for unpaired image-image translation tasks, such as translating synthetic images to real-looking synthetic images. This synthetic-to-real domain adaptation approach has been applied in various tasks such as object detection (Ramamonjison et al., 2021), depth prediction (Zhao et al., 2020), scene understanding and camera pose estimation (Chen et al., 2022), image enhancement (Shao et al., 2020), and medical science (Mahmood et al., 2018). CycleGAN uses cycle consistency using which the domain adaptation is performed in both directions, i.e. from source to target and vice versa, and as such, two possible ways of domain adaptation are possible: synthetic-to-real and real-to-synthetic (Nikolenko, 2021). Several studies have explored the reverse domain adaptation, i.e. real-to-synthetic strategy, and reported either comparable results or better results, highlighting advantages over the synthetic-to-real strategy (Zhao et al., 2020; Petsiuk et al., 2022; Rojtberg et al., 2020; Shoman et al., 2020). The improvement obtained by using real-to-synthetic adaptation is due to the problem simplification, where the complex real image is simplified into a synthetic image as compared to synthetic-to-real adaptation, where CycleGAN has to hallucinate textures derived from real images to fill synthetic images. However, the real-to-synthetic adaptation strategy is an unexplored territory for approaches that perform camera pose regression by learning completely from synthetic images that are obtained from texture-less 3D models.

Camera pose regression involves the estimation of camera pose from a single image and approaches like PoseNet (Kendall et al., 2015) and its subsequent variants (Kendall and Cipolla, 2017; Walch et al., 2017; Clark et al., 2017) have been successfully used to perform the task using deep learning. Training PoseNet requires images with known camera poses. Recent works have used synthetic images obtained from existing texture-less 3D building models to train PoseNet-like models and their variants, and applied simple domain adaptation strategies utilizing image edges (Acharya et al., 2019a,b), segmentation (Acharya et al., 2022), shallow image features (Ha et al., 2018), adversarial learning (Li et al., 2022), transformer models (Kim and Kim, 2022), and recently GANs (Chen et al., 2022). However, these methods have not been shown to perform accurate localisation, and the reported accuracies (approximately 1 meter) are far from comparable with approaches that use real images.

In this paper, we explore the synthetic-real domain adaptation problem for indoor camera pose regression. Specifically, We investigate the reverse real-to-synthetic adaptation strategy using CycleGAN and compare its performance for pose regression with the more common synthetic-to-real adaptation. The

* Corresponding author.

contributions of our work are the following:

1. We train CycleGAN to perform domain adaptation for camera pose regression using three visualisations of synthetic images that are obtained from a texture-less 3D model.
2. We evaluate the performance of reverse domain adaptation strategy, i.e., real-to-synthetic, for camera pose regression and compare it with the synthetic-to-real adaptation.
3. We improve the accuracy of camera pose regression by two-fold as compared to the state-of-the-art results.

In the following, we describe the background of real-synthetic domain adaptation and works related to camera pose estimation in Section 2. In Section 3, we describe the methodology where we use CycleGAN for adaptations in both directions. Section 4 presents the experiments and the results, which is followed by a discussion of limitations and directions for future research in Section 5. Conclusions are presented in Section 6.

2. BACKGROUND

CycleGAN has been used in the past to perform real-synthetic translations for many computer vision tasks, including camera pose estimation, where these synthetic images are usually obtained by the graphical rendering of textured or texture-less 3D models of objects, or a 3D scene. However, in the field of camera pose regression, very limited works exist that utilise texture-less 3D building models for localisation. In the following, we review the works that utilised CycleGAN and related approaches that perform real-synthetic domain adaptation.

2.1 Real-synthetic domain adaptation using GANs

Several works have utilised real-to-synthetic conversion, and some have reported improvements compared with the synthetic-to-real adaptation. Zhao et al. (2020) remove clutter and novel objects from the real images using CycleGAN and use the synthetic images to predict depth, and conclude that cleaned real-looking synthetic images are accurate for depth predictions. Atapour-Abarghouei and Breckon (2018) transform real images into synthetic depth maps using unpaired image stylisation similar to CycleGAN utilising cycle consistency and report good results compared to the current approaches. In medical science, synthetic cystoscopic environments are used to generate synthetic depth maps which are used to adapt the real images using GANs (Mahmood et al., 2018). In the field of image enhancement, de-raining (Chen and Wang, 2022) and dehazing (Shao et al., 2020), real-to-synthetic translations have been used to improve the estimations. Moreover, such reverse adaptation also finds applications in object detection, such as detecting objects in paintings (Ramamonjison et al., 2021) or identifying vehicles at nighttime using daytime images (Lin et al., 2020), and in point cloud classification (Cardace et al., 2023).

2.2 Synthetic image generation from 3D models

3D texture-less object models have also been used to generate synthetic images, which have been stylised using CycleGAN or related approaches using a real-to-synthetic strategy. Petsiuk et al. (2022) use of synthetic images generated from 3D object models for performing semantic segmentation using CycleGAN for applications in additive manufacturing. The authors reported that utilising real-to-synthetic translation resulted in better results as the translation reduced the effects of saturation and the reflections of the real images and incidental filament strings. Planche et al. (2019) propose SynDA, an encoder-decoder network that maps real to the synthetic images that are generated from texture-less synthetic CAD models, and can perform tasks such as image segmentation. Pasqualino et al. (2021) used a CycleGAN and 3D model of the museum to generate synthetic images, and subsequently used translated images for object detection in cultural sites, experimenting both adaptation strategies. Rojtberg et al. (2020) use a CycleGAN-based pipeline for object pose detection, and report better accuracy for the reverse domain adaptation (real-to-synthetic case), and the reported accuracies comparable with a baseline experiment using augmented real image backgrounds. Likewise, Rad et al. (2018) use a feature mapping network using real-synthetic image pairs for 3D object pose detection from real images.

2.3 Camera pose estimation using textured 3D models

CycleGAN and related approaches have been used for camera pose estimation using synthetic images from 'textured' 3D models, which are derived from Structure-from-motion (SfM). Yang et al. (2021a) propose frustum intersection over union to calculate real-synthetic pairs similarity, and then use relative camera pose regression for estimating the absolute camera pose. CycleGAN was utilised for increasing the number of synthetic images using a 3D textured model. Langerman et al. (2021) propose the use of Contrastive Unpaired Translation (CUT) for performing unpaired image-image translation of synthetic images that are generated from coloured point clouds of the scene, and report their approach's performance is comparable with existing fully supervised techniques. Some works have reported that the real-synthetic strategy resulted in better performance for camera pose estimation. Yang et al. (2021b) train RCPNet using real and real-looking synthetic images derived from an outdoor SfM model and CycleGAN for relative camera pose estimation, and demonstrate a single trained model can be used across scenes, without requiring retraining. Shoman et al. (2020) propose CNN-based feature alignment called real-to-synthetic feature transformation (REST) where they convert the real features to synthetic features and then match them against the accumulated database of synthetic images. The authors experimented with different lighting conditions and concluded that real-to-synthetic is a better adaptation strategy as the synthetic scene is a simplification of the real scene, and hence easy to synthesize, and can be performed using a simple network.

2.4 Camera pose regression using texture-less 3D models

PoseNet is a popular camera pose regression approach which utilises images with known camera poses estimated using SfM approaches. Although several improvements to PoseNet have been proposed (Clark et al., 2017; Kendall and Cipolla, 2017), the primary challenge of the approach is the requirement of performing SfM reconstruction of the scene. The widespread availability of 3D building models that can be derived from Building Information Modelling (BIM) has motivated some works to generate synthetic images from those texture-less models and utilise them to perform camera pose regression by domain adaptation. Ha et al. (2018) utilise shallow CNN features for synthetic-real domain adaptation, and perform retrieval of the nearest synthetic image to estimate a coarse camera pose. Baek et al. (2019) extend the work of Ha et al. (2018) for and develop a facility management interface using augmented reality.

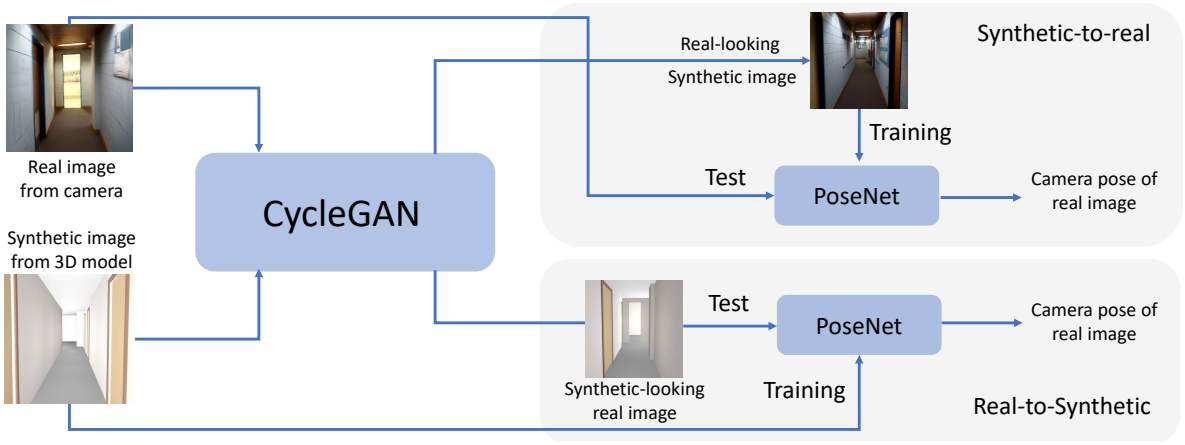

Figure 1. Workflow of the proposed method.

Acharya et al. (2019a) use edge maps for performing synthetic-real domain adaptation. Their subsequent works involved modelling the uncertainty (Acharya et al., 2019b), utilising spatiotemporal information utilising LSTM (Acharya et al., 2020), and intermediate domain representation using segmented images (Acharya et al., 2022). Li et al. (2022) utilised adversarial learning to align the synthetic and the real features derived from CNN. Kim and Kim (2022) perform the domain adaptation utilising the attention property of the recent transformer-based models to identify the most relevant features for performing the task. Chen et al. (2022) perform synthetic-real domain adaptation using CycleGAN, use point features such as HOG and SIFT for matching the real and the synthetic images, and subsequently use the 3D depth information from the BIM to perform a Perspective-n-Point (PnP) camera pose estimation. This is the only work that used CycleGAN for translating synthetic images (obtained from a texture-less 3D model) to real images for the task. The best-reported accuracy of the state-of-the-art approach is 1.12 meters (Acharya et al., 2022).

2.5 Limitations of existing methods

The review of the literature reveals two research gaps. Firstly, the reverse domain adaptation, i.e. the conversion of real images to synthetic has been explored for camera pose estimation, but using 'textured' 3D models derived from SfM (Yang et al., 2021a; Langerman et al., 2021; Yang et al., 2021b; Shoman et al., 2020). Therefore, the synthetic images used in those experiments, contain texture from the real world. Moreover, the reverse domain adaptation strategy is not explored with approaches that perform camera pose regression using synthetic images obtained from texture-less 3D building models (Ha et al., 2018; Acharya et al., 2019a; Li et al., 2022; Kim and Kim, 2022; Chen et al., 2022). Secondly, the reviewed works on camera pose regression using synthetic images suffer from low accuracy. The best-reported accuracy is 1.12 meters (Acharya et al., 2022), which is nowhere near the accuracy of PoseNet trained on real images, which is approximately 0.4 meters for the 7Scenes indoor dataset (Kendall et al., 2015). This indicates that in the existing works, the domain adaptation is not achieved properly, or the synthetic images are not effective for the task.

3. METHODOLOGY

Figure 1 shows the workflow of the proposed approach where we use CycleGAN for synthetic-to-real adaptation. This process generates real-looking synthetic images. We also perform real-to-synthetic adaptation, which generates synthetic-looking real images. We train two PoseNets, one for the synthetic-to-real adaptation and one for the reverse real-to-synthetic adaptation. We train Synthetic-to-real PoseNet with real-looking synthetic images and test it with real images, whereas for the Real-to-synthetic PoseNet, we train the network directly with synthetic images, and test it with real-looking synthetic images.

3.1 Image-image translation using CycleGAN

CycleGAN consists of a generator and a discriminator, where the generators generate the fake real images (real-looking synthetic images) and the discriminator distinguishes the fake from real images. However, to encode the outputs in a meaningful way, CycleGAN utilises cycle consistency that uses real-looking synthetic images to reconstruct back the synthetic images using another generator and a discriminator, and adversarial loss is used to train the model. At the end of the training, the generators learn to generate real-looking synthetic images and synthetic-looking real images.

CycleGAN is a generative model that has been widely used for translating synthetic images to real-looking synthetic images (and vice versa). A relevant question in this aspect will be whether the images generated by CycleGAN reduce the domain gap between the synthetic images that are generated from a texture-less 3D building model, and real images containing building structures, novel objects and repeating texture. Therefore, we used deep CNN features and visualised the domain gap using t-distributed stochastic neighbour embedding or t-SNE (Van der Maaten and Hinton, 2008).

Another relevant question is whether the direction of domain adaptation (synthetic-real vs real-synthetic) has an effect on reducing the domain gap, as several studies have reported the reverse direction is more efficient. Therefore, to quantify the measure of the similarity of the translated images in both directions, we use Fréchet inception distance or FID Heusel et al. (2017) which is a widely used measure of the similarity of two image domains using deep CNN image features.

3.2 Training PoseNet with translated images

PoseNet follows the idea of fine-tuning a pre-trained GoogleNet model with real images and known camera poses, and during inference, it maps a test image to its corresponding camera pose.

Previous studies have established that synthetic images or their intermediate representations such as edge maps and segmented images, or adversarial learning can be used to train PoseNet, and during inference, real test images can be used for pose regression. Chen et al. (2022) used CycleGAN to reduce the domain gap between the synthetic and the real images. However, we believe that the synthetic to real image translation results in more image artefacts that are not suitable for matching using point features, such as HOG or SIFT. Therefore, we propose fine-tuning PoseNet with real-looking synthetic images directly (Figure 2 third row) and testing with real images (Figure 2 (a)), as we believe the image features derived from the 'deep' layers of the CNN will be robust to the artefacts generated by real-looking synthetic images obtained from CycleGAN.

Additionally, we believe the effects of the artefacts and other novel objects in the scene will be reduced during the reverse domain adaptation. Consequently, we fine-tune PoseNet directly with synthetic images (Figure 2 second row), and during inference, we used synthetic-looking real images (Figure 2 fourth row). We selected the Syn-Car images for our experiments with synthetic-real adaptations, whereas Syn-pho-real-tex images for real-synthetic adaptations based on the best FID scores of the image translations by CycleGAN.

4. EXPERIMENTS AND RESULTS

We used the publicly available PyTorch implementation of CycleGAN[1] and used all the default setting, except for turning off the image flipping, which improved the quality of the translated images. We trained the CycleGAN and PoseNet models on a Tesla P100 GPU having 12GB memory.

4.1 Dataset details

We use the only available indoor cross-domain dataset called Unimelb Corridor that has been used by the previous studies (Acharya et al., 2019a; Li et al., 2022; Kim and Kim, 2022). As shown in Figure 1 (a) - (d), this dataset contains real images of a university corridor and synthetic images of the same corridor obtained from a texture-less 3D building model of the corridor. Three different visualisations of synthetic images are generated by moving a virtual camera: Syn-car visualisation stands for Synthetic Cartoonish, Syn-pho-real stands for Synthetic Photorealistic, and Syn-pho-real-tex stands for Synthetic Photorealistic Textured, where Syn-pho-real-tex contains synthetic repeating textures. The other cross-domain datasets such as VKITTI are not suitable for the task, as they are paired synthetic-real pairs in addition to being outdoors. We used a subset of the dataset and removed some of the redundant frames both from the real and synthetic images towards the end, and added some real images throughout the trajectory.

4.2 Exploring domain adaptation using CycleGAN

We trained three CycleGAN models with Syn-car, Syn-pho-real and Syn-pho-real-tex datasets and real images. We did not use any real-synthetic image pairs for training, nor did we use the ground truth of the real images during training, thereby performing unsupervised image-image translations. As in the CycleGAN framework, there is no objective measure to identify the best model during training, we utilised t-SNE plots of the test images that were not used during the training to qualitatively understand the progress of domain adaptation.

[1] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

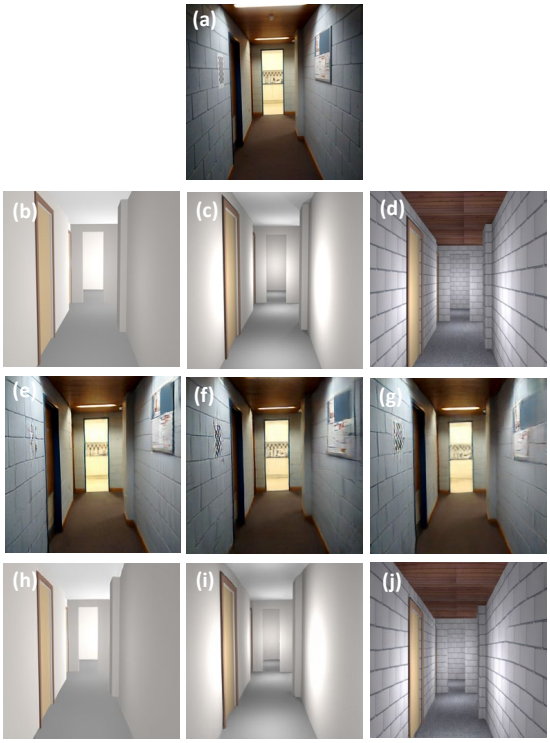

Figure 2. Dataset description. (a) Real image (b) - (d) Synthetic images obtained from the 3D building model called Syn-Car, Syn-pho-real and Syn-pho-real- tex respectively. (e) - (g) Real-looking synthetic images generated by CycleGAN using Syn- Car, Syn- pho-real and Syn-pho-real-tex respectively. (h) - (j) Synthetic-looking real images generated from the real image.

Figure 3 shows the t-SNE plots of the three synthetic image visualisations before and after domain adaptation by CycleGAN, where the blue and red filled points represent the real and synthetic images respectively. We observe for all three visualisations, the real and the synthetic images are well separated, therefore, demonstrating the initial domain gap before adaptation. We also plot the real-looking synthetic images during different epochs of training for the synthetic-to-real adaptation represented by triangles for different epochs of training. We observe that the real-looking synthetic images move closer to the real images in the embedding space, demonstrating the domain adaptation. Additionally, we notice the images generated from the later epochs (e.g. green and black triangles) are closer to the blue points compared to the early epochs. For the reverse domain adaptation, we plot the synthetic-looking real images for different epochs of training represented by diamonds. Likewise the synthetic-to-real case, we observe that the images generated from the CycleGAN model at later epochs are closer to the red synthetic points. Interestingly, the distribution of the points for the reverse domain adaptation case is closer compared to the synthetic images (red points), compared to the distribution of the points for the synthetic-to-real case, the spread of which is wider. These distribution patterns in the t-SNE figures show a correlation between what is observed in Table 1, and show that reverse domain adaptation is better compared to the synthetic-to-real case, although does not provide any quantitative insights regarding which synthetic image visualisation is better.

Consequently, to quantify the domain adaptation, we used FID to measure the similarity of the translated images throughout the training. We do this for both the synthetic-real adaptation

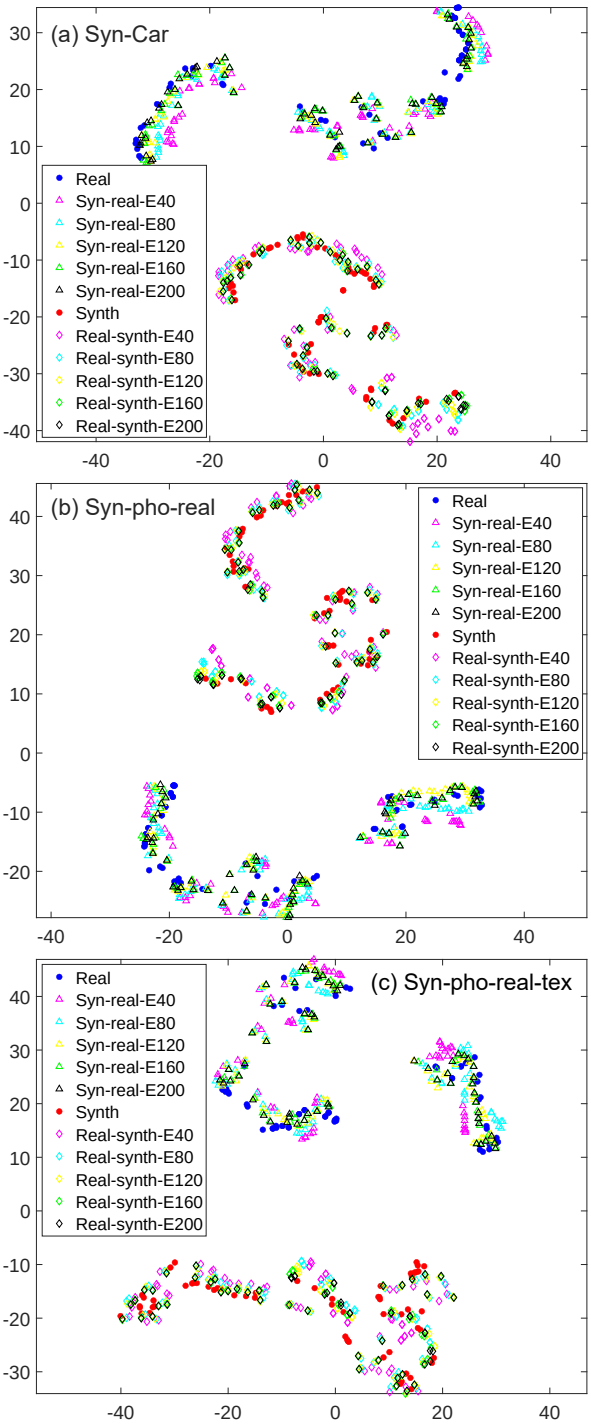

Figure 3. Visualisation of the t-SNE using ResNet50 features during the different epochs of the adaptation of (a) Syn-car (b) Syn-pho-real and (c) Syn-pho-real-tex datasets.

and reverse adaptation as well as to identify the better domain adaptation direction. Table 1 shows the results of the experiments, where we perform the evaluation at an interval of 40 epochs for the 200 epochs training. Epoch 0 refers to the case before domain adaptation was performed. From Table 1, we observe that the lowest synthetic-to-real image similarity measured by FID is for Syn-Car images, achieved at 200 epochs, and we use these real-looking synthetic images from this CycleGAN model for the synthetic-to-real experiments. Whereas for the real-to-synthetic translations, the lowest FID is observed for Syn-pho-real-tex images at 200 epochs, and we use these synthetic-looking real images translated from this CycleGAN model for the reverse domain adaptation experiments.

Figure 2 (e) - (g) shows the results of the synthetic-to-real translations resulting in real-looking synthetic images generated from the corresponding images in the top row (Figure 2 (b) - (d)) using the CycleGAN models trained for 200 epochs. From Figure 2 (e) - (g) we observe that all the translated synthetic images look realistic, containing the textures from the real images, but, containing some distortions and artefacts. For instance, the posters on the walls are not reconstructed properly, and the brick texture of the walls is noisy. Also, most of the images are blurry, the translation for the Syn-pho-real-tex images being the most. Additionally, we observe the edges of the building structures, especially the door frames are not straight. Figure 2 (h) - (j) shows the real-to-synthetic translations resulting in synthetic-looking real images generated from real images (Figure 2 (a)) using the CycleGAN model trained for 200 epochs. From Figure 2 (h) - (j) we observe that the real-to-synthetic translations for all three visualisations are very similar to the original synthetic images, and are sharp as compared to the real images, with some distortions, specifically for Syn-pho-real-tex images, where the brick textures are distorted. Similar to the synthetic-to-real translations, we also observe some distortions near the door frames (see Figure 2 (i)).

4.3 Evaluation of Synthetic-to-real pose regression

The qualitative and quantitative analysis shows that good domain adaptation in terms of image similarity has been achieved, and we believe this should improve the pose regression performance of PoseNet when being trained with adapted images. To identify the pose regression performance of the synthetic-to-real domain adaptation, we trained PoseNet with real-looking synthetic images that were generated from the CycleGAN model trained for 200 epochs on Syn-car images (Figure 2 (e)), and call it Synthetic-to-real PoseNet. The camera pose of the synthetic images was used during training, and no camera pose of real images was used. During inference, real test images taken from the camera were used, which regressed the camera pose in the coordinate system of the 3D building model.

Figure 4 (a) shows the estimated points by Synthetic-to-real PoseNet for the real images, where the colour of the points represents the error magnitude. The red line shows the ground truth

Table 1. Fréchet inception distances (FID) of the different synthetic images during the training course of CycleGAN for both domain adaptation directions (lower is better).

Adaptation direction	Image visualization	Fréchet inception distance (FID) at					
		0 epoch	40 epochs	80 epochs	120 epochs	160 epochs	200 epochs
Synthetic-to-real	Syn-car	210.76	108.15	77.30	66.70	64.04	**59.20**
	Syn-pho-real	187.83	104.34	80.42	65.28	63.29	59.29
	Syn-pho-real-tex	112.89	110.58	87.77	72.06	69.29	65.84
Real-to-synthetic	Syn-car	210.76	88.13	58.30	51.02	44.58	43.68
	Syn-pho-real	187.83	63.72	48.42	39.67	37.43	38.50
	Syn-pho-real-tex	112.89	54.86	42.94	36.49	33.24	**30.47**

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-W1-2023-453-2023 | © Author(s) 2023. CC BY 4.0 License.

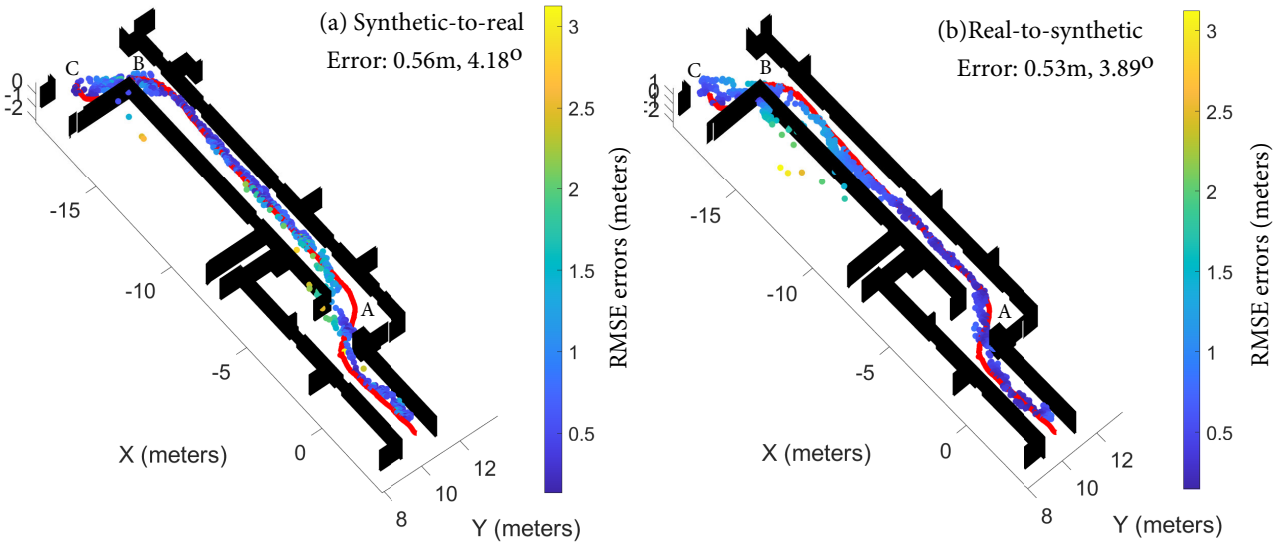

Figure 4. The estimated camera locations by (a) Synthetic-to-real PoseNet and (b) Real-to-synthetic PoseNet. The red line shows the ground truth trajectory of the camera. The accuracies are shown in the top right, and the colour of the points shows the errors.

trajectory which was generated using SfM reconstruction of the scene, and none of these camera poses was used during training Synthetic-to-real PoseNet. The median error of localisation is 0.56 meters and 4.18°. From Figure 4 (a), we notice that most of the points are predicted very close to the ground truth trajectory having errors lower than 1 meter (dark blue points), whereas some points were predicted wrongly (yellow points), containing high errors. We also notice slightly inconsistent results near Points A, B and C of the trajectory, where the predicted locations are not precise. The larger errors near these points could be a result of low lighting causing image blur and the differences between the real scene and the 3D building model.

4.4 Evaluation of Real-to-synthetic pose regression

To identify the pose regression performance of the reverse domain adaptation case, we trained PoseNet with Syn-pho-real-tex synthetic images directly using the camera poses of these synthetic images, and we call this model as Real-to-synthetic PoseNet. During inference, we used the CycleGAN model trained with Syn-pho-real-tex images trained for 200 epochs to convert the real images to synthetic-looking real images. Figure 4 (b) shows the estimated points by Real-to-synthetic PoseNet, and we observe there is a slight improvement in the accuracy of camera pose regression observed, with median errors of localisation as 0.53 meters and 3.89°. Compared to the synthetic-to-real case, we observe that the predictions are better near Point A of the trajectory, indicating the better quality of the synthetic-looking real images derived from the real images during inference. However, the distribution of the estimated points is worse near Points B and C of the trajectory.

Motion blur due to low lighting and the structural differences between the scene and the 3D building model are two main factors influencing the performance of pose regression. For the synthetic-to-real adaptation, the synthetic images are adapted to look like real images, and this can to some extent cover the gap between the structural difference between the scene and the real world. However, for the reverse domain adaptation, as PoseNet is trained directly with the synthetic images, it predicts ambiguous results during inference when being tested with real-looking synthetic images that have structural variations. In addition, in the synthetic-to-real adaptation case, the frames that are close to the walls contain texture, which can help the network in the pose regression task. But in the reverse domain adaptation case, the absence of texture results in ambiguous pose estimations. Nevertheless, the results of pose estimation with reverse domain adaptation suggest that the cleaner synthetic-looking real images achieve a higher localisation accuracy.

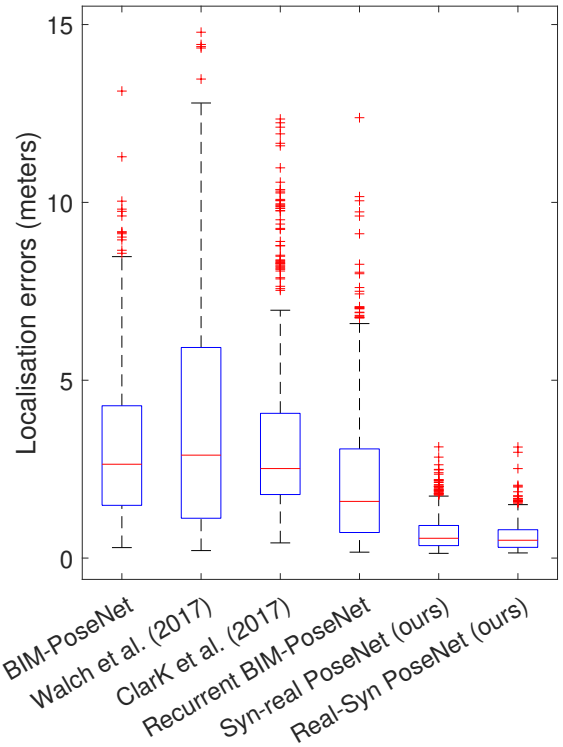

Figure 5. The box plots of the two proposed strategies compared to the existing domain approaches. We observe a significant improvement in the accuracy and the distribution of errors.

Subsequently, to compare the results of the two domain adaptation approaches, we plotted the box plots of the errors (shown in Figure 5), and compared them with the existing approaches. From Figure 5 we observe that both the domain adaptation

strategy results in a lower error of localisation compared to the existing approaches, where the results of the reverse domain adaptation strategy are slightly better. We believe this improvement is a result of the clean synthetic-looking real images containing fewer artefacts that are used during the reverse domain adaptation strategy.

4.5 Comparison with state-of-the-art

Compared to the state-of-the-art domain adaptation results, we see a significant improvement in the accuracy of the proposed approach for both strategies. Compared to Chen et al. (2022), who use CycleGAN (accuracy of 1.34m and 10.29°), our results are improved approximately by 150% in terms of localisation accuracy. Also, compared to the leading results reported by Acharya et al. (2022), we see an improvement of 111% and 55% for location and rotational errors, respectively. We summarise the comparison results in Table 2.

Table 2. Comparison of the proposed Synthetic-to-real PoseNet and Real-to-synthetic PoseNet with existing domain adaptation methods.

Domain adaptation method	Pose regression error (meters, degrees)
Li et al., 2022	1.53, 11.00
Kim et al., 2022	1.37, 6.86
Chen et al., 2022	1.34, 10.29
Acharya et al., 2022	1.12, 6.06
Synthetic-to-real PoseNet (ours)	0.56, 4.18
Real-to-synthetic PoseNet (ours)	**0.53, 3.89**

5. DISCUSSIONS AND FUTURE DIRECTIONS

Artefacts generated in real-looking synthetic images are a major limitation that restricts the applicability of approaches such as Chen et al. (2022) for performing PnP camera pose estimations, as such, using image features of the deep layers of CNN can help to mitigate the challenge. Additionally, the experiments suggested that the synthetic-to-real strategy is sensitive to motion blur.

Using reverse domain adaptation can help us mitigate the challenges of the artefacts and motion blur as we generate a much clean synthetic-looking real image, compared to a real-looking synthetic image. Thus, the experiments indicate that we can achieve a slightly better localisation accuracy with this strategy. However, structural differences between the 3D building model and the actual scene are a major limitation of using a reverse domain adaptation strategy. Additionally, another limitation of this strategy is the loss of the real texture of the scene which will result in ambiguities for areas with similar structural geometry.

One of the possible future directions could be exploring the use of spatiotemporal information (image sequences) instead of performing pose regression with single images, as previous studies have established that the accuracy of the pose regression improves utilising image sequences (Walch et al., 2017; Clark et al., 2017; Acharya et al., 2020). Another possible direction could be exploring whether this domain adaptation can be performed without the need for real images of the scene and whether indoor images from other public datasets can be utilised to train CycleGAN. This framework will only require a 3D building model for performing camera pose estimation. This line of work will also address the challenge of the lack of public datasets for the approaches performing camera pose regression that utilises 3D texture-less building models.

6. CONCLUSIONS

We explore a reverse domain adaptation strategy utilising a texture-less 3D building model by generating synthetic-looking real images with CycleGAN, and compare it with the synthetic-to-real strategy. The results of the experiments with camera pose regression indicate that the generated synthetic-looking real images contain fewer artefacts and are cleaner compared to the real-looking synthetic images. However, some of the limitations of the reverse strategy include the sensitivity to structural differences between the actual scene and the 3D building model, in addition to the ambiguity due to the lack of texture for the frames that are closer to the wall. Despite these limitations, an improvement in the pose regression performance using PoseNet is observed, where the reverse adaptation tackles the problem of motion blur well compared to the synthetic-to-real adaptation. A significant improvement is observed for camera pose regression by performing domain adaptation using CycleGAN, and compared to the existing state-of-the-art domain adaptation approaches, we decrease the localisation errors approximately by a factor of two.

ACKNOWLEDGMENTS

This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne (established with the assistance of ARC LIEF Grant LE170100200). The authors acknowledge the support from Building 4.0 CRC.

References

Acharya, D., Khoshelham, K., Winter, S., 2019a. BIM-PoseNet: Indoor camera localisation using a 3D indoor model and deep learning from synthetic images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150, 245–258.

Acharya, D., Singha Roy, S., Khoshelham, K., Winter, S., 2019b. Modelling uncertainty of single image indoor localisation using a 3D model and deep learning. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, IV-2/W5, 247-254.

Acharya, D., Singha Roy, S., Khoshelham, K., Winter, S., 2020. A Recurrent Deep Network for Estimating the Pose of Real Indoor Images from Synthetic Image Sequences. *Sensors*, 20(19), 5492.

Acharya, D., Tennakoon, R., Muthu, S., Khoshelham, K., Hoseinnezhad, R., Bab-Hadiashar, A., 2022. Single-image localisation using 3D models: Combining hierarchical edge maps and semantic segmentation for domain adaptation. *Automation in Construction*, 136, 104152.

Atapour-Abarghouei, A., Breckon, T. P., 2018. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2800–2810.

Baek, F., Ha, I., Kim, H., 2019. Augmented reality system for facility management using image-based indoor localization. *Automation in construction*, 99, 18–26.

Cardace, A., Spezialetti, R., Ramirez, P. Z., Salti, S., Di Stefano, L., 2023. Self-distillation for unsupervised 3d domain adaptation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4166–4177.

Chen, J., Li, S., Liu, D., Lu, W., 2022. Indoor camera pose estimation via style-transfer 3D models. *Computer-Aided Civil and Infrastructure Engineering*, 37(3), 335–353.

Chen, Z.-B., Wang, Y.-G., 2022. Dtt-net: Dual-domain translation transformer for semi-supervised image deraining. *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, 1621–1625.

Clark, R., Wang, S., Markham, A., Trigoni, N., Wen, H., 2017. Vidloc: A deep spatio-temporal model for 6-dof video-clip relocalization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3.

Ha, I., Kim, H., Park, S., Kim, H., 2018. Image retrieval using BIM and features from pretrained VGG network for indoor localization. *Building and Environment*, 140, 23–31.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Kendall, A., Cipolla, R., 2017. Geometric loss functions for camera pose regression with deep learning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5974–5983.

Kendall, A., Grimes, M., Cipolla, R., 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. *Proceedings of the IEEE international conference on computer vision*, 2938–2946.

Kim, D., Kim, J., 2022. CT-Loc: Cross-domain visual localization with a channel-wise transformer. *Neural Networks*.

Langerman, J., Qiu, Z., Sörös, G., Sebők, D., Wang, Y., Huang, H., 2021. Domain Adaptation of Networks for Camera Pose Estimation: Learning Camera Pose Estimation Without Pose Labels. *arXiv preprint arXiv:2111.14741*.

Li, Q., Cao, R., Zhu, J., Hou, X., Liu, J., Jia, S., Li, Q., Qiu, G., 2022. Improving synthetic 3D model-aided indoor image localization via domain adaptation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183, 66–78.

Lin, C.-T., Huang, S.-W., Wu, Y.-Y., Lai, S.-H., 2020. GAN-based day-to-night image style transfer for nighttime vehicle detection. *IEEE Transactions on Intelligent Transportation Systems*, 22(2), 951–963.

Mahmood, F., Chen, R., Durr, N. J., 2018. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *IEEE transactions on medical imaging*, 37(12), 2572–2581.

Nikolenko, S. I., 2021. Synthetic-to-real domain adaptation and refinement. *Synthetic data for deep learning*, Springer, 235–268.

Pasqualino, G., Furnari, A., Signorello, G., Farinella, G. M., 2021. An unsupervised domain adaptation scheme for single-stage artwork recognition in cultural sites. *Image and Vision Computing*, 107, 104098.

Petsiuk, A., Singh, H., Dadhwal, H., Pearce, J. M., 2022. Synthetic-to-real Composite Semantic Segmentation in Additive Manufacturing. *arXiv preprint arXiv:2210.07466*.

Planche, B., Zakharov, S., Wu, Z., Hutter, A., Kosch, H., Ilic, S., 2019. Seeing beyond appearance-mapping real images into geometrical domains for unsupervised cad-based recognition. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2579–2586.

Rad, M., Oberweger, M., Lepetit, V., 2018. Feature mapping for learning fast and accurate 3d pose inference from synthetic images. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4663–4672.

Ramamonjison, R., Banitalebi-Dehkordi, A., Kang, X., Bai, X., Zhang, Y., 2021. Simrod: A simple adaptation method for robust object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3570–3579.

Rojtberg, P., Pöllabauer, T., Kuijper, A., 2020. Style-transfer gans for bridging the domain gap in synthetic pose estimator training. *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, IEEE, 188–195.

Shao, Y., Li, L., Ren, W., Gao, C., Sang, N., 2020. Domain adaptation for image dehazing. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2808–2817.

Shoman, S., Mashita, T., Plopski, A., Ratsamee, P., Uranishi, Y., 2020. Real-to-Synthetic Feature Transform for Illumination Invariant Camera Localization. *IEEE Computer Graphics and Applications*, 42(1), 47–55.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Walch, F., Hazirbas, C., Leal-Taixe, L., Sattler, T., Hilsenbeck, S., Cremers, D., 2017. Image-based localization using lstms for structured feature correlation. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 627–637.

Yang, C., Liu, Y., Zell, A., 2021a. Learning-based camera relocalization with domain adaptation via image-to-image translation. *2021 International Conference on Unmanned Aircraft Systems (ICUAS)*, IEEE, 1047–1054.

Yang, C., Liu, Y., Zell, A., 2021b. Relative camera pose estimation using synthetic data with domain adaptation via cycle-consistent adversarial networks. *Journal of Intelligent & Robotic Systems*, 102(4), 79.

Zhao, Y., Kong, S., Shin, D., Fowlkes, C., 2020. Domain decluttering: Simplifying images to mitigate synthetic-real domain shift and improve depth estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3330–3340.

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