Addressing Challenging Place Recognition Tasks using Generative Adversarial Networks

Yasir Latif, Ravi Garg, Michael Milford and Ian Reid

Abstract—Place recognition is an essential component of any Simultaneous Localization and Mapping (SLAM) system. Correct place recognition is a difficult perception task in cases where there is significant appearance change as the same place might look very different in the morning and at night or over different seasons. This work addresses place recognition using a two-step (generative and discriminative) approach. Using a pair of coupled Generative Adversarial Networks (GANs), we show that it is possible to generate the appearance of one domain (such as summer) from another (such as winter) without needing image to image correspondences. We identify these relationships considering sets of images in the two domains without knowing the instance-to-instance correspondence. In the process, we learn meaningful feature spaces, the distances in which can be used for the task of place recognition. Experiments show that learned feature correspond to visual space and can be effectively used for place recognition across seasons.

I. INTRODUCTION

The problem of Simultaneous Localization And Mapping (SLAM) has matured enough for it to move out of a laboratory setting and into the real world. This has allowed SLAM algorithms to operate over longer time periods and potentially leading to the ideal scenario of operating over the life time of an autonomous robot.

One of the essential components for effective operation of any SLAM system is place recognition or loop closure detection as it allows a mobile robot to limit its overall uncertainty and reduce the overall error in the map by associating the current pose of the robot to one of poses in the past.

Based on the timescale of operation, any SLAM algorithm can be divided into two components: a sequential inference subsystem which estimates the current pose of the robot in relation to poses in the recent past and a loop closing component that operates “out of sequences” and is responsible for detecting revisits. Under normal circumstances, appearance change is small and gradual in the frame-to-frame timescale and poses no challenge for the sequential part, however, for a loop closing algorithm appearance change is accumulative and can be significant over a long operational time frame. Place recognition in presence of changes in appearance of the same place, such as those that occur over seasons or between day and night, is a difficult perception problem and is the subject of this work.

This work addresses the problem of place recognition as an image translation task; if we have observed a place under various illumination conditions, can we generate what the current place would look like under different (previously observed) conditions? By observing the same place under different conditions, we aim to estimate a non-linear mapping that can convert from one illumination conditions to another. Hence, the problem is reduced to that of domain translation, that is, given an image in one domain (for example summer), generate images in a corresponding domain (for example winter), such that the structure remains the same, but the
transient conditions vary. For this task, we use a variant of Generative Adversarial Networks (GANs) \cite{goodfellow2014generative} and are able to generate realistic images in the target domain from images in the source domain. For such a task, the normal approach is to train a network using images with known correspondences in the two domains, which is time consuming and might not always be possible. In this work, the relationship between the two domains is learned from sets of images in each domain without requiring one-to-one image correspondences across the domains. These generated images can be used directly for place recognition, however, we show that the features learned for each domain are more stable than pixel-wise differences in the image space, and therefore more useful for place recognition tasks. Some of the images for winter-summer translation are depicted in Fig. 1. The first and third rows contain real images, while the other two rows contain images generated from the corresponding real images.

The rest of the paper is organized as follows: in Section II we present an overview for different approaches that have been devised for place recognition under extreme weather changes and present the relevant literature for GANs as well. In Section III we present the specific network used in this work and briefly justify why it is correct choice for the task. Experiments are then presented that highlight the properties of the learned features and their performance for the place recognition task. Finally, conclusions and future work is presented.

II. RELATED WORK

Place recognition with cameras is a well studied problem and a good overview can be found in \cite{wu2017benchmarking}. Traditionally, the problem of visual loop closure detection has been addressed using feature-based approaches, where, the extracted features serve as a description of the structure being observed by the robot \cite{lowe1999object, bay2008surf}. Such methods operate very well within the short time scale with limited appearance or structural changes. They would, however, fail in situations where the change in the environment is greater than the invariances (such a rotation, scale, illumination etc.) provided by the underlying feature descriptor.

A. Visual Place Recognition

Visual place recognition is the task of localizing a robot in an environment using information from a single camera. It is a well studied problem and can be split into two dominant approaches; feature based, that represent a particular image as a set of key features extracted from interesting location in the image, and image-based, that instead of extracting features, tries to reason using all the image information \cite{wu2017benchmarking}. Each approach has it own limitations: feature-based methods may fail under illumination changes while image-based methods are sensitive to viewpoint changes.

Several approaches have been introduced to address the problem of visual place recognition under extreme appearance changes. These approaches can be divided into two broad categories: methods that try to mitigate the effect of visual change by learning illumination invariant descriptors, and methods that learn how to transform images from one visual domain to another.

SeqSLAM \cite{wu2017benchmarking} shows that even though a single image might not contain enough information, sequences of images, can lead to successful place recognition under extreme weather changes. SeqSLAM represents images as mean and variance normalized patches before reasoning about their similarity. Along similar lines, \cite{wu2017benchmarking} formulate the problem of matching sequences as a network flow problem. They use HOG features over a grid in an image as the illumination invariant representation of the image. Similarly, \cite{wu2017semantic} shows that U-SURF \cite{wu2017semantic} is more robust to illumination changes than SIFT and SURF which, together with epipolar constraints, can be used to close loops across different illumination conditions. The effect of illumination change over the span on a day is studied in various works \cite{wu2017benchmarking,wu2017benchmarking}. It is shown that U-SIFT \cite{wu2017benchmarking} has the greatest illumination invariance among traditional feature descriptors.

Instead of whole images or features, several structure based methods have also been proposed that use edges in the images are the illumination invariant description \cite{wu2017benchmarking,wu2017benchmarking}. Techniques such as shadow removal \cite{wu2017benchmarking} has shown remarkable improvement by removing unwanted artifacts from images. Additionally, features from Convolutional Neural Networks (CNNs) are also successfully used for the task of place recognition due to their invariance to illumination and viewpoint changes \cite{wu2017benchmarking}.

This work takes a different approach towards the problem of place recognition. We take a generative approach, that is, we aim to generate how a place would look after a given appearance transformation. If we know the conditions in which a previous image was captured, we want the ability to transform the current image in such a way that it resembles the image captured under conditions of interest.

B. Generative Adversarial Networks

GANs \cite{goodfellow2014generative} have been successful applied to the task of domain specific image generation. In its simplest form, a GAN consists of two components: a generator $G$, which randomly samples from a latent space and aims to generate a realistic image that resembles images from a domain being learned (such as face, cars, rooms, etc), and a discriminator $D$ whose task is to correctly discriminate between real and generated (fake) images. These two components aim to beat each other, that is the generator aims to produce realistic images so it can fool the discriminator and the discriminator aims correctly identify which image is real and which is generated. In a single pass through such a network, a random sample from a known distribution is converted to an image by the generator, which, along with a real image from the domain, is shown to the discriminator. The classification error between the real and generate image serves as an update for the discriminator and the generator, one trying to minimize it and the other trying to maximize it. As a result of such an adversarial training, the generator learns a mapping from the latent space to the image domain under consideration. The latent space serves as a representation
for the images in the domain which is shown to be smooth (each point in the space correspondence to some image) and support vector operations [15].

The common analogy to explain how GANs work is the contest between a forger who produces fake banknotes and a bank which is tries to detect them. They are both locked in an adversarial zero sum game: where success for one means a loss for the other. The better the bank gets at detecting the fake banknotes, the better the forger has to defeat it. Similarly, the more adapted the forger gets, the better the bank has to be in order to detect the fakes banknotes. Over time, they reach an equilibrium, in which the forger is the best it can be at producing fake banknotes, and the bank becomes the best it can be at detecting the forgeries.

There is significant of literature regarding GANs and their various applications, however, we are particularly interested in the types of networks that perform transfer. One such work is DiscoGAN [9], which discovers mappings between two image domains using unpaired examples from both domains. They show that a pair of coupled GANs can discover relationships between styles of handbags and shoes and discover common attributes (such as orientation) between images of cars and human faces, etc. Along similar lines, CycleGAN [21] enforces cyclic-consistency to learn one-to-one mapping between domains and show translation from satellite images to maps. If the two domain have paired information, pix2pix [8], can be used to learn the translation from one domain to another.

III. PLACE RECOGNITION USING GANs

The holy grail for place recognition methods is to find a suitable representation space for images in which these properties hold: (a) distance between images depends on the observed structure and not the appearance, (b) the representation is view-point invariant, or at least view-point aware, so that images that observe the same structure from different viewpoints are placed close to each other, and (c) the distances in this space are meaningful, that is, the larger the distance between two images in this space, the less likely it is that they have common structure (in short, a vector space in which the triangular inequality holds based on the structure being observed). If such a representation is found, then place recognition simplifies to the nearest neighbour search because all other distractors (such as illumination, viewpoint, etc) have been taken into consideration and normalized. We show that by using GANs, something akin to such a representation can be learned.

We first describe the useful properties of GANs which make them ideal for the task of place recognition and describe the architecture of the system that is used in this work.

A vanilla GAN takes as input a random sample from a latent space and generates images of a particular domain (as shown in Fig. [3], however, this is not useful for our task, as we are interested in encoding an image into the latent space, and a normal GAN does not have the provision for it. Therefore, we need a generator that can take as input an image, instead of random noise, and generate an image. In order to do that, a simple encoder-decoder network is used as a generator paired with a discriminator to learn the representation for a single domain (an autoencoder), or to translate from one domain to another by having different input to the generator and discriminator.

For the translation task, there are infinitely many ways that the source domain can be mapped to a target domain, since the GAN maps one distribution to another and is not aware of instance-level relationships. The problem is constrained by using two such GANs coupled with each other and simultaneously learning to translate both ways, from domain A to B and vice versa. This allows us to have cyclic constraints [21,9], that is, an image from domain A, after translation to domain B and retranslated back to domain A, should match the original image.

We provide a general overview of such a system below. Assuming that there exists a generator function $G_B$ that can map an image in domain A, $x_A$ to an image in domain B such that $x_{AB} = G_B(x_A)$, where $x_{AB}$ means an image from domain A translate to domain B. Additionally, another generator $G_A$ maps in the reverse direction, that is, $x_{BA} = G_A(x_B)$. In order to be able to translate between the two domains, we need to learn these two generators. However,
Fig. 5: Coupled GANs for learning domain translation: One GAN consists of \( G_A \) and \( D_A \) and the other of \( G_B \) and \( D_B \). The dashed generators represent reevaluation using the existing generators. Each domain contains its own discriminator which learns a domain specific feature for images. Lighter arrows represent the cyclic constraint.

without any further restriction, there are infinitely many of functions that can satisfy these requirements. Instead of mapping from one distribution to another, we want a one-to-one mapping between the two domain. As proposed in [9], we use the cyclic consistency to restrain the system to do one-to-one mapping between the two domains by minimizing the cyclic reconstruction loss:

\[
\arg \min_{G_A,G_B} \| x_A - x_{ABA} \|_2^2 \tag{1}
\]

and

\[
\arg \min_{G_A,G_B} \| x_B - x_{BAB} \|_2^2 \tag{2}
\]

which forces the two generators to be inverses of each other, that is \( G_A(G_B(x_A)) \rightarrow x_A \) and similarly \( G_B(G_A(x_B)) \rightarrow x_B \).

Discriminators \( D_A \) and \( D_B \) work in each domain and try to discriminate between \( x_A \), \( x_{BA} \) and \( x_B \) and \( x_{AB} \) respectively. We design the discriminator so that it can learn a feature space near the end of the network. Fig. [5] provides an overview our set up. The generator is each domain translate for the other domain and in each domain there is a discriminator to generate between real and fake images. Each generated image is then fed into the other generator to get an image in the original domain. The gray lines represent cyclic pixel-wise Mean Squared Error constraint.

A. Network architecture

Fig. [3] shows the architecture of the generator and discriminator part of the networks. The generator consists of a set of convolution and deconvolution layers as shown in Fig. [3]. In order to preserve edge information, we use skip connections in the network so that useful information about edges can flow easily through the network. The bottleneck layer is the encoding space of the generator which is then used to generate the image in the other domain.

The discriminator consists of the encoder part of the generator with a slight difference. We use a fixed dimensional fully connected layer at the end of the discriminator. This serves as the feature layer in the discriminator. In the experimental section, we show that this layer learns informative features that can be used for place recognition.

B. Place Recognition

We consider the case where we have seen the same place in two different weather conditions, that data from which can be used to train for the image translation task. Each domain (summer, winter) has a set of training images from which the relation between the domains can be learned. At test time, we assume that the current conditions are known and we want to do place recognition in the other domain. We can generate images in the other domain using the appropriate generator, which ideally would give us the exact image. However, this assumes the same viewpoint for both images.

Another point of comparison is the features learned by the discriminator. We can go from \( x_A \) to \( x_{AB} \) and compare the features from \( D_B \) for the pair \( X_B \) and \( X_{AB} \), where the first the image from domain A, and second is an image from domain B translated to A.

In the experiment section, we present results for place recognition in the feature space and explore its properties.

IV. EXPERIMENTS

To show the performance of our system, we use the Norland dataset. The dataset has been recorded onboard a train in Norway, over four different seasons: summer, winter, autumn and spring. The footage consists of about 10 hours of video that has been synchronized using GPS information.

We generate frames from the video at 2Hz leading to about just over 70 thousand images for each sequence. We split the dataset in half and use the first half as the test set and train on a subset of the images from the second half.

For training, we divide the training set into half again and use the first half of summer and the later half of the winter images as training examples. This is done to avoid any actual

\footnote{http://nrkbeta.no/2013/01/15/nordlandsbanen- minute-by-minute-season-by-season/}
Fig. 6: Distances in the feature space of summer images. Distance have been truncated at a maximum to show the pattern around the main diagonal. Block diagonal appear where the train either remains stationary or moves through a tunnel.

pairs being seen by the network. The training set finally contains about 17000 summer and as many winter images. All images are rescaled to a size of $64 \times 64$. We train on color images instead of grayscale as it contains more information about the weather conditions. All experiment were carried out on a commodity GPU (GeForce GTX 980). At test time, translation and feature generation takes less than 10ms per image on the GPU.

A. Learned Features

We first investigate if the features learned in the discriminator correlate to visual appearance. If the space correlated to visual appearance, images close to each other in appearance, should have small distance between them. We extract features in each domain for the test set, and compute the cosine distance between them. The images are sampled regularly from the set of summer image for the purpose of visualization.

It can be seen in Fig. 6 that as we move away from the main diagonal, the distances become larger. Very large distances have been clipped to show the structure around the main diagonal. The block diagonal elements appear when the train remain stationary or goes thorough a tunnel. In both cases, there is no visual change and the features lie very close to each other.

In order to verify this visually, we randomly select some of the images in summer set and show their nearest-neighbours in the feature space (Fig. 7). It can be seen that indeed similar images are close to each other in this space.

B. Place recognition across summer and winter

We use the first 1000 images (at 1Hz) from the Norland sequence as the test case for this experiment. We extract features for summer images and translate winter images to summer followed by feature extraction. Both features are extracted in the same domain. A single feature, even in this learned space, is not able to localize with sufficient accuracy. Therefore, we to use sequences of features for place recognition. Our implementation of these features is very simple compared to SeqSLAM [12].

In order to use sequences, we compose the corresponding features from the sequence of images, that is, for a sequence length $n$ starting at time $t$, the feature $f_{t-n+1:t} = [f_{t-n+1}^T f_{t-n}^T \ldots f_t^T]^T$, in other words, simply stacking features for single images, gives us a feature for the sequence.

For matching, we use cosine distances, after normalizing each feature to unit length. The results for various sequence length are given in Fig. 8.

This experiment shows that these learned features, along with the generator that translates from winter to summer, can be effectively used for place recognition under severe illumination changes. For comparison with SeqSLAM on the same dataset, we refer the reader to [18](Fig. 4), where the performance of SeqSLAM is shown for various sequence lengths. Our method is able to achieve full precision with a much smaller sequence length.

C. Conclusions and Future Work

In this work, we explored the possibility of using GANs for the task of image translation and subsequently using the discriminator features for place recognitions. We show that features encode visually similar images close in the space and can be used for place recognition using generated images in
Fellowship FL130100102 to IR.

under project DP130104413, the ARC Centre of Excellence
learn an embedding for edges in the image, which can then
results are shown in Fig. 9. Such a setup would allow us to
and learn how to associate an image to its edges, some initial
structure. We can represent structure using lines in the image
into domains are trivially available for each domain, such as
have to be learned. This thought has motivated us to look
many different conditions all pairs of possible translations
have to be learned. This thought has motivated us to look
into domains are trivially available for each domain, such as
structure. We can represent structure using lines in the image
and learn how to associate an image to its edges, some initial
results are shown in Fig. 9. Such a setup would allow us to
learn an embedding for edges in the image, which can then
can be used to match edges between different weather conditions.

ACKNOWLEDGMENT

This work was supported by Australian Research Council
under project DP130104413, the ARC Centre of Excellence
for Robotic Vision C E140100016, and through a Laureate
Fellowship FL130100102 to IR.

REFERENCES

[1] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded
Up Robust Features. Computer vision–ECCV 2006,
pages 404–417, 2006.
[2] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuza,
J. Neira, I. Reid, and J. J. Leonard. Past, present,
and future of simultaneous localization and mapping:
Toward the robust-perception age. IEEE Transactions
on Robotics, 32(6):1309–1332, 2016.
[3] P. Corke, R. Paul, W. Churchill, and P. Newman. Dealing
with shadows: Capturing intrinsic scene appearance
for image-based outdoor localisation. In Intelligent
Robots and Systems (IROS), 2013 IEEE/RSJ Interna-
tional Conference on, pages 2085–2092. IEEE, 2013.
[4] M. Cummins and P. Newman. Fab-map: Probabilistic
localization and mapping in the space of appearance.
The International Journal of Robotics Research,
27(6):647–665, 2008.
[5] E. Eade and T. Drummond. Edge landmarks in monoc-
ular slam. In BMVC, pages 7–16, 2006.
[6] D. Gálvez-López and J. D. Tardos. Bags of bi-
ary words for fast place recognition in image se-
quencies. IEEE Transactions on Robotics, 28(5):1188–
1197, 2012.
[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu,
D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in neural
information processing systems, pages 2672–2680, 2014.
[8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-
image translation with conditional adversarial networks.
arxiv, 2016.
[9] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim. Learning
to discover cross-domain relations with generative
adversarial networks. arXiv:1703.05192, 2017.
[10] D. G. Lowe. Object recognition from local scale-
invariant features. In Computer vision, 1999. The pro-
cedings of the seventh IEEE international conference
on, volume 2, pages 1150–1157. Ieee, 1999.
[11] S. Lowry, N. Sündorhauf, P. Newman, J. J. Leonard,
D. Cox, P. Corke, and M. J. Milford. Visual place
recognition: A survey. IEEE Transactions on Robotics,
32(1):1–19, 2016.
[12] M. J. Milford and G. F. Wyeth. Seqslam: Visual route-
based navigation for sunny summer days and stormy
winter nights. In Robotics and Automation (ICRA),
2012 IEEE International Conference on, pages 1643–
1649. IEEE, 2012.
[13] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss.
Robust visual robot localization across seasons using
network flows. In Twenty-Eighth AAAI Conference on
Artificial Intelligence, 2014.
[14] S. Nuske, J. Roberts, and G. Wyeth. Robust outdoor
visual localization using a three-dimensional-edge map.
Journal of Field Robotics, 26(9):728–756, 2009.
[15] A. Radford, L. Metz, and S. Chintala. Unsupervised
representation learning with deep convolutional gener-
ative adversarial networks. arXiv:1511.06434, 2015.
[16] P. Ross, A. English, D. Ball, B. Upcroft, G. Wyeth,
and P. Corke. A novel method for analysing lighting
variance. In Australian Conference on Robotics and
Automation, 2013.
[17] P. Ross, A. English, D. Ball, and P. Corke. A method
to quantify a descriptor’s illumination variance. 2014.
[18] N. Sündorhauf, P. Neubert, and P. Protzel. Are we
there yet? challenging seqslam on a 3000 km journey
across all four seasons. In Proceedings of Workshop on
Long-Term Autonomy, IEEE International Conference
on Robotics and Automation (ICRA), 2013.
[19] N. Sündorhauf, S. Shirazi, F. Dayoub, B. Upcroft,
and M. Milford. On the performance of convnet features
for place recognition. In Intelligent Robots and Systems
(IROS), 2015 IEEE/RSJ International Conference on,
pages 4297–4304. IEEE, 2015.
[20] C. Valgren and A. J. Lilienthal. Sift, surf & seasons:
Appearance-based long-term localization in outdoor
environments. Robotics and Autonomous Systems, 58
(2):149–156, 2010.
[21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired
image-to-image translation using cycle-consistent ad-
versarial networks. arXiv:1703.10593, 2017.