DASGIL: Domain Adaptation for Semantic and Geometric-aware Image-based Localization

Hanjiang Hu, Ming Cheng, Zhe Liu, Hesheng Wang*, Senior Member, IEEE

Abstract—Long-Term visual localization under changing environments is a challenging problem in autonomous driving and mobile robotics due to season, illumination variance, etc. Image retrieval for localization is an efficient and effective solution to the problem. In this paper, we propose a novel multi-task architecture to fuse the geometric and semantic information into the multi-scale latent embedding representation for visual place recognition. To use the high-quality ground truths without any human effort, depth and segmentation generator model is trained on virtual synthetic dataset and domain adaptation is adopted from synthetic to real-world dataset. The multi-scale model presents the strong generalization ability on real-world KITTI dataset though trained on the virtual KITTI 2 dataset. The proposed approach is validated on the Extended CMU-Seasons dataset through a series of crucial comparison experiments, where our performance outperforms state-of-the-art baselines for retrieval-based localization under the challenging environment.

Index Terms—Visual localization, image retrieval, representation learning, domain adaptation.

I. INTRODUCTION

VISUAL localization plays an essential role in visual perception in mobile robots and outdoor self-driving vehicles [1]–[4], especially for long-term SLAM systems, in which environmental factors including illumination, weather and seasonal change have significant influence on the precision of visual localization [5]. Image retrieval, as a convenient technique for visual localization, has been shown to be an effective method [6] by proving the query image to be the linear combination of retrieved images from database under specific distance metric. Consequently, the key to the retrieval-based localization is to find the most similar images to the query image from the database, which prepares for local high-accuracy 6-DoF camera pose regression [7], [8] by giving global course retrieval images [9].

Traditional local descriptors like SIFT [10], BRIEF [11], ORB [12], BRISK [13], etc. work satisfactorily in image-based localization without much variance of environmental conditions. However, due to the dependence on image pixels, these man-made local features are not robust under drastically varying conditions. Global feature shows impressive advantages on visual localization, i.e. VLAD-like feature [14], [15] and DenseVLAD [16] gives impressive performance on long-term image-based localization. In the meanwhile, with the great development of deep neural networks, especially CNN in computer vision, the deep and dense features extracted from CNN have been used in image-based localization in changing environments [17], showing that shallow feature vectors are promising in illumination change while deeper ones go for robustness to change of view point. As feature goes deeper, the semantic information has been more extracted and therefore could be used to generate depth map [18], [19] or semantic segmentation map [20], [21].

Since the higher-level context is intuitively robust to the changing pixel caused by environmental variance, recent work [22] leverages the geometric information with auxiliary depth map for image-based localization. Besides, [23], [24] introduce the semantic segmentation for the improvement of visual localization. However, it is effort-cost and time-consuming to obtain the depth and semantic segmentation maps for real-world images. And consequently, synthetic data and domain adaptation have drawn significant attention in recent years, aiming to use the high-quality groundtruth at the least cost by alleviating the gap between virtual and real images. Furthermore, [25] shows that the complementary depth map can boost semantic segmentation through domain adaptation.

To efficiently leverage the geometric and semantic information for visual localization with zero labor costs, we propose DASGIL, a novel domain adaptation with semantic and geometric information for image-based localization. The proposed method adopts the one-encoder-two-decoder multi-task architecture, fusing geometric and semantic information to multi-scale latent features through the shared encoder called Fusion Feature Extractor. The fused features of both virtual and real images share the same distribution in multi-layer-feature adversarial training, adapting from synthetic images to real-world images. Based on the fused multi-scale features, representation learning for place recognition is accomplished through multi-scale triplet loss and the retrieval process is based on the multi-scale features, which is more effective and efficient. For the experiments, we train the model on Virtual KITTI 2 dataset but test it on Extended CMU-Seasons dataset for retrieval-based localization, and our results are better than state-of-the-art methods under various regional environments, vegetation conditions and weather conditions. In summary, our work makes the following contributions:

- We propose a novel and state-of-the-art approach, DASGIL, fusing semantic and geometric information into
latent features through a multi-task architecture of depth prediction and semantic segmentation.

- The multi-scale latent embedding is introduced through adversarial training for domain adaptation from synthetic to real-world images with zero labor cost but high-quality groundtruth.
- Fused representation learning for place recognition is adopted through multi-scale triplet loss and features from multiple scales are applied in retrieval process as well.
- A series of comparison experiments have been conducted to validate the effectiveness of every proposed module in DASGIL. And our approach outperforms state-of-the-art image-based localization baselines on the Extended CMU-Seasons dataset though supervisely trained on Virtual KITTI 2 dataset.

We structure the rest of this paper as follows. Section II analyzes the related work. The architecture of DASGIL and the pipeline for image-based localization are introduced in Section III and Section IV respectively. Section V introduces the training details, the experimental results and ablation study. Finally, in Section VII we draw our conclusions and give some suggestions for future work.

II. RELATED WORK

A. Domain Adaptation for Segmentation and Depth Prediction

In the applications of visual perception in autonomous driving and mobile robotics, monocular depth prediction and semantic segmentation are significantly important and have been constantly studied in recent years. The early studies leverage man-made descriptors or probabilistic graph models [26]–[28] for depth prediction or semantic segmentation. Since deep convolutional neural networks (DCNN) boost the performance of visual perception, enormous studies show promising results for monocular depth estimation [18], [19], [29], [30] and semantic segmentation [21], [31], [32], where fully convolutional neural networks (FCNNs) are introduced to enable end-to-end training.

However, the groundtruth of depth map and segmentation map in outdoor scenarios are often time-consuming, labor-cost and expensive to obtain, which constrains the development and performance of supervised learning methods. Some unsupervised or weakly-supervised methods are proposed for depth prediction [33]–[35] and semantic segmentation [23], [24], [36], with the assistance of left-right consistency [33] or egomotion pose constraint [37] for depth prediction and image level tags or 2D-2D points [24] for semantic segmentation.

Besides, the virtual synthetic datasets [38]–[40] are developed to deal with such issue as well, where the image sequences are under changing environments with perfectly high-quality depth map and segmentation map as groundtruth. However, as the models are trained on these virtual datasets, there occurs a concern about the generalization ability to the real-world images with the domain gap.

Domain adaptation refers to the generalization ability to a new different dataset for a model trained on one dataset. Some previous works focus on domain-invariant deep latent feature learning [41] in which [41] mitigates the discrepancy between source and target domain to make the feature distribution identical by minimizing Maximum Mean Discrepancy (MMD). While adversarial loss [42] is also commonly used to decrease the discrepancy, through training the discriminator to discriminate the representation feature from source or target domain.

As for the domain adaptation in the task of depth prediction [43]–[45] and semantic segmentation [46], [47], lots of work focus on image translation [43], [45], which largely depends on the performance of translation [48] and is hard to cover all the changing environments, so they not suitable for place recognition in various conditions. [49] proposes to train the mid-feature to be consistent between two domains while it only copes with single-layer feature, lacking the adaptation of multi-scale features.

B. Long-term Place Recognition and Localization

Outdoor visual place recognition has been studied for many years and could be directly used for visual localization in autonomous driving or loop closure detection of SLAM, in which the most similar images are retrieved from database for query images. Traditional local feature descriptors are aggregated for image retrieval [50], [51], and have successfully addressed most cases of loop closure detection in real-time V-SLAM [52], [53] without huge environmental changes. VLAD [14] is the most successful man-made feature for place recognition and it has been extended to different versions. NetVLAD [15] extract deep features through VLAD-like network architecture. DenseVLAD [16] presents retrieval results through extracting multi-scale SIFT descriptor with VLAD [14] under drastic perceptual changes.

Since convolutional neural networks (CNNs) has successfully addressed many tasks in computer vision, long-term visual place recognition and localization has developed significantly aided with CNN. Image translation-based methods seem to be the most direct way to solve the problem [54], [55], where images are transferred across different domains based on generative adversarial networks (GANs) [56]. ToDayGAN [57] similarly translates night-images to day-images and uses DenseVLAD for following retrieval. Jenicek et al. [58] proposes to use U-Net to obtain photometric normalization image and then find deep embedding for retrieval. However, generalization ability is limited for translation-based methods because the accuracy of image-level retrieval largely depends on the quality of translated image compared to latent-feature retrieval of ours.

To cope with the challenging perceptual change, many recent works follow the pipeline of learning the robust deep representation through neural networks together with semantic [59]–[62], geometric [22], [63], context-aware information [64], [65], etc. However, these works need auxiliary information which is effort-cost to obtain in most cases. Instead of explicit image translation, feature learning is promising for image retrieval [68]–[70]. Consequently, domain adaptation at the level of feature map seem prospective for place recognition assisted with geometric and semantic information.
III. DASGIL ARCHITECTURE

A. Architecture Overview

As shown in Figure. [1] our proposed DASGIL adopts one-encoder-two-decoders architecture, including one shared fusion feature extractor $E$ and two map generators $G_S, G_D$ for semantic segmentation and depth prediction, respectively. The extracted multi-scale features from $E$ are used to generate target depth map and segmentation map given the virtual image input $I_V \sim p_V(I)$. To diminish the domain gap between synthetic images and real-world images, the adversarial training is incorporated through the multi-scale feature discriminator $D$, resulting in the same distribution of multi-scale features from the inputs of both the real images $I_R \sim p_R(I)$ and virtual images $I_V \sim p_V(I)$.

B. Fusion Feature Extractor

To extract geometric and semantic information from the input RGB image ($I$), we use a shared encoder ($E$) to accomplish this task. Since the depth map and segmentation map are both based on the extracted features through $E$, the extractor fuses geometric information and semantic information into the multi-scale features. Besides, deep features of multiple scales are extracted from $E$ through all the convolution layers for the skip connection to decoders as U-Net-like models [71], [72] do. This structure instructs the model to obtain and use different levels of features containing geometric and semantic information, which assists the the generation of depth map and segmentation map.

C. Depth Map And Segmentation Map Generator

Two map generation networks ($G_D, G_S$) with the same structure are used to generate depth map and segmentation map, respectively. In order to instruct the model to generate depth map and segmentation map based on the multi-scale features extracted from $E$, we use skip-connections from $E$ to $G_D$ and $G_S$, as Figure. [2] shows. Notice that the segmentation map and depth map are decoded from the same fusion features extracted from the shared encoder $E$. The depth map and segmentation map generated are shown in Figure. [2].

D. Discriminator of Multi-layer GAN

Since the feature extractor $E$ is trained on the virtual synthetic images $I_V$ while tested on real-world images $I_R$ for image localization, the extracted mid-features fused with geometric and semantic information must be distribution-consistent for both virtual images and real-world images. For the domain adaptation from synthetic domain to real-world domain, the adversarial training strategy is adopted in the multi-scale latent embedding space. The multi-scale features are concatenated and then go through a batchnorm layer before fed into the feature discriminator $D$. The proposed feature discriminator consists of three fully connected layers as a two-class classifier to determine whether the latent feature is from...
real-world domain $R$ or virtual domain $V$. Different from existing domain adaptation work \cite{43, 45} which build the discriminator at image or single-scale feature level, we build the discriminator at the level of features with multiple scales, as shown in Figure. \[4\] This multi-layer discriminator structure allows the model to recognize $R$ and $V$ from multiple levels, enabling the model to better utilize the fusion information extracted. The effectiveness of this multi-layer discriminator is validated in $V\sim D$ and also shown in Figure. \[4\]

IV. DASGIL PIPELINE FOR IMAGE-BASED LOCALIZATION

For the image retrieval of localization, the multi-task architecture model is trained to learn the fusion feature representation, incorporating the geometric and semantic information into the latent representation. The learning pipeline and training losses are introduced first, and then the image-based localization process is presented.

A. Domain Adaptation for Multi-task Training

The overall architecture is designed for multi-task, i.e. generating both semantic segmentation map and depth map at the same time. The reconstructed depth loss and segmentation loss are introduced to instruct the extractor $E$ and generators $G_S, G_D$ to learn latent embedding features and generate the two target maps for virtual input images in Figure. \[4\] Besides, the adversarial loss is to assure that both of the real-world images and virtual images could extract the same-distributed features through the same extractor $E$ in Figure. \[4\]

1) Multi-scale Depth Reconstruction Loss: Inspired by \cite{30}, we construct the multi-scale reconstruction loss for the generation of depth map given the virtual images $I_V \sim p_V(I)$. After the forward propagation of $E, G_D$, we utilize the results obtained from multiple layers of $G_D$ and then resize the ground truth of depth map $Depth_{GT}$ into corresponding sizes. Then we compute L1 Loss between the features of $G_D$ and ground truth $Depth_{GT}$ at each level. Finally we add the losses together to get the reconstruction loss for depth map.

$$L_D = \mathbb{E}_{I_V \sim p_V(I)} \sum_i ||D(G(E(I_V))), Depth_{GT}||_1$$ (1)

where $i$ refers to the $i^{th}$ layer of extractor $E$ and generator $G_D$, and $n$ is the total layers involved, while $Depth_{GT}$ with subscript represents the ground truths which have been resized to the size of $i^{th}$ layer. Note that the involved layers are with large scale and resolution.

2) Cross Entropy Segmentation Loss: As for segmentation map, we use cross entropy loss to train the segmentation generator model. The generation of segmentation map is the classification at pixel level for the discrete classification instead of continuous regression. Similar to the image forward propagation for depth reconstruction, we apply score regression and the soft-max layer for the outputs at multiple layers of generator $G_S$. The cross entropy loss is adopted for every pixel at multiple layers.

$$L_S = \mathbb{E}_{I_V \sim p_V(I)} \sum_i (-\sum_{c=1}^{M} Seg_{GT}, log(p_c))$$ (2)

where $i$ refers to the $i^{th}$ layer of extractor $E$ and generator $G_S$, and $n$ is the total layers involved. $p_c$ denotes the probability for class $c$ pixel-wise while $Seg_{GT}$ is the ground truth one-shot label for class $c$, and there are $M$ classes in all.

3) Multi-scale Feature Adversarial Loss: Instead of constructing GAN loss at image level which has been done by most of the existing work in domain adaptation, we construct GAN loss at multiple scales of features and force them to have consistent distribution so as to make full use of the features...
extracted through each layer in $E$ as shown in Figure. [3] Because the features of all layers in $E$ are skip-connected to $G_D$ and $G_S$, it is of great necessity to consider all features when constructing our discriminator and thus the distribution of generated depth map and segmentation map is consistent between virtual and real-world domains.

Suppose the image from virtual domain as $I_V \sim p_V(I)$, image from real-world domain as $I_R \sim p_R(I)$, we use the LSGAN [73] form instead of the original form [56].

$$L_{GAN} = \mathbb{E}_{I_V \sim p_V(I), I_R \sim p_R(I)}[0.5 \times (D(E(I_R) - 1)^2 + (3) \quad D(E(I_V)^2)] + \mathbb{E}_{I_V \sim p_V(I)}[0.5 \times D(E(I_V) - 1)^2]$$

where $D$ is the multi-layer feature discriminator described in III-C and $E$ is the fusion feature extractor described in III-B

Note that the first term only optimizes the discriminator and the second one only optimizes the generator in an adversarial training manner. We validate the effectiveness of our multi-layer adversarial loss in Section. V-D and also shown in Figure. [4]

### B. Representation Learning for Place Recognition

While the multi-task model is trained to fuse the geometric and semantic information into the latent embedding feature, the key point of retrieval-based localization is to learn the robust representations for database and query images. Also for the deep metric learning task, the fusion triplet loss is calculated using concatenated features extractor from $E$ in multiple scales.

1) Multi-scale Fusion Triplet Loss: We propose a Multi-scale Fusion Triplet Loss to instruct the model to learn specific representation, as shown in Figure. [5] Different from [22] which generates depth maps and extracts geometric information explicitly through another depth encoder before calculating the triplet loss, we improve this strategy by fusing both geometric information and semantic information explicitly for representation learning. A triplet loss involves an anchor virtual image ($q_i \sim q_V(I)$), a positive sample ($q_{i+}$) representing the same scene as the anchor and a negative sample ($q_{i-}$) which is unrelated to the anchor image. The triplet loss ($L_T(q_i, q_{i+}, q_{i-}, m)$) is shown in the formula below:

$$L_T = \mathbb{E}_{q_i, q_{i+}, q_{i-} \sim p_V(I)}[\max(0, 1 - \frac{||q_i - q_{i-}||_2}{m + ||q_i - q_{i+}||_2})] \quad (4)$$

where $m$ represents the margin how the distance of negative pairs is larger than that of positive pairs.

Considering the impact of features at different levels on the final task, retrieval-based localization, we improve the formula (4) and propose a multi-scale triplet loss in formula (5)

$$L_{mul-T} = \mathbb{E}_{q_i, q_{i+}, q_{i-} \sim p_V(I)}[\sum_{l=L_m}^{L_k} \mathcal{C}_T]

= \mathbb{E}_{q \sim p_V(I)}[\sum_{l=L_m}^{L_k} \max(0, 1 - \frac{||q_i - q_{i-}||_2}{m + ||q_i - q_{i+}||_2})] \quad (5)$$

where $L_m$ and $L_k$ refer to the $m^{th}$ and $k^{th}$ layers of $E$, respectively.

Unlike using the concatenated feature of RGB images and depth maps for triplet loss in [22], we only use features from the shared encoder $E$ to construct the triplet loss, where geometric and semantic information having been fused in $E$ implicitly. This shows that the multi-scale features from $E$ have already contained the information of input image, depth map and segmentation map, indicating that the model has enough ability of recognizing different places through these features. Therefore, there is no need to construct the triplet loss by extracting features from generated depth map and segmentation map explicitly. The effectiveness of multi-scale fusion triplet loss is validated in Section. V-D.
2) **Total Training Loss**: In order to simultaneously train the domain adaptation for semantic segmentation, depth prediction, and representation learning, all the losses are combined to be a total losses shown in \( \mathcal{L}_{\text{Total}} \) weighted by \( \lambda_T, \lambda_{GAN}, \lambda_D, \lambda_S \) respectively.

\[
\mathcal{L}_{\text{Total}} = \lambda_T \mathcal{L}_{\text{mul-T}} + \lambda_{GAN} \mathcal{L}_{GAN} + \lambda_D \mathcal{L}_D + \lambda_S \mathcal{L}_S
\]

Note that the GAN loss actually contains two optimizing processes, following the adversarial training pipeline. Other losses, Equation 1, Equation 2, Equation 5 are added in the generation optimizing process, while only discrimination loss in the GAN loss Equation 3 is involved in the discrimination optimizing process.

C. **Image Retrieval for Localization**

For image-based localization procedure, the feature representations of database should be built first. For each query image, we will find the one in the database with the most similarity as the image retrieval result, as Figure 6 shows.

As for the representation of database, every image in the database goes through Fusion Feature Extractor \( \mathcal{E} \) to extract multi-scale fusion feature \( f_{db} \), resulting in the fusion features \( F_{db} \) of all images in the database is built. For the query image, the same procedure is done to obtain the multi-scale fusion feature \( f_q \).

To measure the similarity between \( f_q \) and \( f_{db} \) as well as taking different scales of features into consideration, we apply L1 measure metric to find the least distance for retrieval. Note that the measurement metric used in the triplet loss is L2, which is stricter to train the model to learning the best representation.

Due to the previous work [17] where the mid-layer feature is the most suitable for image representation for place recognition, instead of using features extracted from all the layers of the extractor \( \mathcal{E} \), we only apply the ones from middle layers of the extractor. The middle-layer features combine the advantages of both the deep and shallow features together, maximally instruct the model to measure the similarity between two images. According to [17], although higher-level feature is more robust to the environmental variance and change of viewpoint than low-level feature, it cannot retrieve the high-precision results since the deep semantic information is not suitable to distinguish the slight place difference. Therefore, the middle-layer features are chosen to calculate the final similarity.

As for image-based localization task, for every query image, the database image which has the least L1 distance summed by multi-scale features to the query is considered to be the retrieval result. After traversing all the images in the query set, we finally obtain the image-based localization result.

V. **Experiments**

This section introduces the experiments, including the dataset introduction, implementation details, experiment results and ablation studies. Our code and pretrained models are available at [https://github.com/HanjiangHu/DASGIL](https://github.com/HanjiangHu/DASGIL).

A. **Experimental Setup**

1) **Datasets for Training and Testing**: KITTI [74] dataset captures rural areas and roadway with multiple outdoor objects in the scenes by driving around the mid-size city of Karlsruhe. However, the ground truths of depth map and semantic segmentation map is not accurate and hard to obtain from original Lidar Scanner and camera devices. Virtual KITTI 2 [38], [40] dataset is a synthetic dataset including five different sequences cloned from real-world KITTI dataset with different camera angle-views (15 degree left, 15 degree right, 30 degree left, 30 degree right) and weather conditions (clone, fog, morning, overcast, rain, sunset). Besides, it contains high-quality ground truths of depth map and semantic segmentation map without any human effort. Due to the multi-conditional image sequences and large-quantity ground truths with high
quality of Virtual KITTI 2, the representation learning multi-task model could be well trained on it together with real-world images in KITTI dataset.

The CMU Seasons dataset [5] is derived from the CMU Visual localization dataset [75]. It is recorded over a year along a 9 kilometers long route including urban, suburban, and park areas in Pittsburgh, USA. The left and right images are captured from cameras on both sides of a car. The dataset has 11 query environments and 1 reference environment and is challenging for the huge change of foliage. The Extended CMU Seasons dataset is extended from the above dataset and owns 40% more images. Therefore, the Extended CMU dataset is more challenging than the original one [5].

2) Evaluation Metrics: We follow the evaluation method on the benchmark website [5] and use its measurement metrics to evaluate the performance of image-retrieval localization. The benchmark on the evaluation website is for visual localization with three levels of precision: high, medium and coarse precision, i.e. (0.25m, 2°), (0.5m, 5°) and (5m, 10°), respectively. The percentage of pose error within each precision is counted to evaluate the performance. Notice that the database images are sunny + No Foliage while other query images are under the combination of Cloudy, Overcast, Low Sun, Sunny, Snow and Foliage, Mixed Foliage, No Foliage under all urban, suburban and park areas.

3) Baselines of Image-based Localization: In the experiment, we choose the several state-of-the-art image retrieval-based localization baselines as follows.

- **NetVLAD** [76] extracts deep features in VLAD-like networks and uses these to retrieve target images.
- **DenseVLAD** [77] uses multi-scale dense VLAD of SIFT descriptors for image retrieval in a traditional manner.
- **DIFL** [69] learns the domain-invariant feature as representation for retrieval through a self-supervised image-to-image translation architecture.
- **Xin et al.** [66] proposes a Landmark Localization Network to localize the discriminative visual areas that benefit the similarity measurement, which gives the best results currently.
- **WASABI** [62] retrieves images from the semantic edge wavelet transform through a global image description with the semantic and topological information.

**B. Implementation Details**

The virtual images for depth prediction, semantic segmentation and triplet representation learning are from the virtual KITTI 2 dataset, while both of virtual and real-world KITTI images are involved in the GAN loss for domain adaptation. For the retrieval-based localization, the test dataset is the Extended CMU Seasons dataset. Therefore, different datasets are chosen for training and testing respectively and the generalization ability is validated as well across multiple datasets.

The original RGB images as well as the ground truths of depth map and segmentation maps are cropped to 256 x 1024. The batch size is set as 4 and learning rate is 0.005 for ADAM stochastic gradient descent algorithm.

| Methods      | Park   | Suburban | Urban   |
|--------------|--------|----------|---------|
| NetVLAD [76] | 2.6/ 59 | 3.7/ 74  | 12.2/ 89|
| DenseVLAD [77]| 5.2/ 62 | 5.3/ 73  | 14.7/ 83|
| DIFL [69]    | 6.6/ 70 | 4.3/ 60  | 17.3/ 42|
| Xin et al. [66]| 6.6/ 70 | 3.8/ 60  | 17.3/ 42|
| WASABI [62]  | 7.5/ 60 | 3.8/ 60  | 17.3/ 42|

The RGB images are input into an eight-layer encoder and two decoders for the generation of depth map and segmentation map. Skip connection is applied on all eight layers while the Multi-scale Depth Reconstruction Loss and Cross Entropy Segmentation Loss only involve the last four layers.

For the domain adaptation at the feature level, the feature discriminator is a three-layer fully linear neural network with dimension as 1004800, 64, 64, 1. Due to the eight-layer skip connection structure, the multi-layer features from the encoder are flattened and concatenated, going through a batchNorm layer right before fed to discriminator.

Since there are different variations on camera angles and environments for any image sequence in the virtual KITTI 2 dataset, the positively paired images are within 5-image interval along a sequence but from different environments while the negatively paired ones are randomly chosen and flipped. The features from the middle-four layers are used to construct the multi-scale triplet loss, while features only from the fifth and sixth layers are used as representation for image retrieval. The margins are set to be 1 for the features at the third, fourth, fifth and sixth layer in the multi-scale triplet loss.

**C. Experimental Results**

We conduct a series of experiments under different regional environments, vegetation conditions and weather conditions. The results of other baselines are from the long-term visual-ization benchmark website https://www.visuallocalization.net/benchmark

1) Results In Different Regional Environments: We have validated the effectiveness of our model under various regional environments, as shown in Table. I. Due to the variety of regional environments in the real world, i.e. urban area, suburban area, and other areas, it is of great necessity to validate the effectiveness of the model in various areas. Besides, the scenario of different environments varies significantly. For example, in the urban area, there exist lots of cars, roads and modern buildings while there are not many trees. In park area, however, trees occupy most of the image but there are almost
no buildings and cars. Therefore, the model should be robust for place recognition in these different areas.

From Table. I we can conclude that our model behaves almost the best among all of the baselines under three regional environments: urban, suburban and park, except on only one evaluation result (0.5m, 5°) in urban area, where ours is in the second place.

Particularly, in suburban area, our model performs 11.1% higher than state-of-the-art one under the coarse precision. In the task of retrieval-based localization, the improvement in coarse precision is much more important than that in high precision because there is numerous methods for finer pose regression for localization [9], [78], [79]. Because there are many trees and other static objects in park and suburban area, the usage of geometric and semantic information is much more helpful for place recognition. As for urban area, our model performs the best except under medium precision. Because there are too many cars and other dynamic objects in the urban area which could move across environments, geometric and semantic information would be affected and not consistent even for the same place under changing environments. These factors make our model perform not as satisfactory as in suburban area or park area. However, for coarse-precision localization, our model is robust to dynamic objects, making ours the best compared with other baselines.

2) Results Under Different Vegetation Conditions: We’ve also tested our model under different foliage conditions under all the areas of urban, suburban and park, and the results are shown in Table. III. Since the reference vegetation is No Foliage, the results under different foliage conditions indicate the effectiveness of our model under various foliage conditions, including foliage and mixed foliage. Our model performs the best among all baselines under these vegetation conditions, resulting in the robustness to the huge change of vegetation.

Among the foliage conditions, our model performs the best when there is mixed foliage in the environment, which is also the most difficult one for place recognition for the reason that the species, location and amount of the foliage are the most complex. Due to the geometric and semantic information we’ve used for this task, our model has robustness to the change of foliage, with multi-scale deep features.

3) Result Under Different Weather Conditions: Besides the experiments on the change of regional environment and vegetation conditions, we also validate our model under different weather conditions for the images in all the areas, including low sun, cloudy, overcast and snow. Note that the reference images in the database are under sunny condition. The result are shown in Table. III with the same evaluation method as in Section. V-C1.

Our model outperforms the other state-of-the-art baselines under most the weather conditions, and especially on the coarse precision, ours leads the second baseline by 5% at least, which has great significance for image-based localization. Note that the training set Virtual KITTI 2 does not contain snow condition, but our model still performs satisfactorily under the snow condition on medium and coarse precision, showing the strong generalization ability of our model. Due to the geometric and semantic information we’ve used, our model has the robustness to the weather change. Although weather and illumination conditions are different across different environments, the depth and segmentation maps remain the same for the same place.

D. Ablation Study

The effectiveness of the modules in the architecture of DASGIL as well as the methods of training and testing are validated through a series of comparison experiments in ablation study section.

Table. IV shows the impact of different modules in DASGIL, including the generation module of both depth map and segmentation map as well as the types of discriminator of GAN module. It could be concluded that all of the Depth generation module, Segmentation generation module and GAN module are effective and indispensable in the proposed DASGIL framework. Furthermore, the Multiple GAN performs significantly better than Single GAN (best result from the 5th layer) for the reason that features from all layers used in the skip connection are adapted from synthetic domain to real-world domain. Also, Multiple Triplet Loss (3rd to 6th layers) improves the results compared to the best Single Triplet Loss (5th layer) for triplet loss, confirming the claim in [V-B1].

VI. Conclusion

In this paper, we propose a novel multi-task architecture, DASGIL, to fully extract geometric and semantic information for retrieval-based localization. Our method implements domain adaptation from synthetic to real-world images and fuses the features from original image, depth maps and semantic segmentation maps. We construct depth reconstruction loss and cross entropy loss for depth prediction and semantic segmentation respectively. As for the place recognition, the multi-scale triplet loss is proposed as well. Besides, the experiments are conducted on the Extended CMU Seasons dataset to present the performance on image-based localization, resulting in outperforming state-of-the-art baselines for retrieval-based localization under changing environments. However, our model is not that robust to moving objects, on which we will focus in the future.

REFERENCES

[1] L. Svarn, O. Enqvist, M. Oskarsson, and F. Kahl, “Accurate localization and pose estimation for large 3d models,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 532–539.
[2] T. Sattler, B. Leibe, and L. Kobbelt, “Efficient & effective prioritized matching for large-scale image-based localization,” IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 9, pp. 1744–1756, 2016.
[3] X. Gao, R. Wang, N. Demmel, and D. Cremers, “Ldsoc: Direct sparse odometry with loop closure,” in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018, pp. 2198–2204.
[4] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, “Dynaslam: Tracking, mapping, and inquiring in dynamic scenes,” IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 4076–4083, 2018.
TABLE III
RESULTS OF DIFFERENT WEATHER CONDITIONS
DATABASE REFERENCE IS UNDER SUNNY

| Methods       | Overcast     | Low sun     | Cloudy      | Snow        |
|---------------|--------------|-------------|-------------|-------------|
| NetVLAD [69]  | 6.7 / 19.1 / 76.3 | 5.5 / 17.5 / 71.3 | 6.8 / 20.1 / 79.5 | 5.9 / 16.4 / 88.0 |
| DenseVLAD [77] | 8.4 / 23.3 / 72.1 | 8.3 / 26.1 / 76.0 | 9.3 / 27.3 / 80.5 | 8.3 / 29.0 / 78.9 |
| DIFL [69]     | 9.7 / 25.3 / 70.9 | 8.7 / 25.3 / 74.4 | 8.8 / 24.7 / 76.9 | 7.4 / 26.7 / 73.5 |
| Xin et al. [66] | 11.5 / 30.8 / 80.8 | 9.3 / 27.6 / 79.2 | 9.4 / 28.0 / 83.7 | 7.6 / 27.6 / 75.9 |
| WASABI [62]   | 5.4 / 15.8 / 70.8 | 4.2 / 14.0 / 62.1 | 5.1 / 15.3 / 71.0 | 3.4 / 13.2 / 58.0 |

DASGIL (ours) | 11.9 / 31.6 / 87.9 | 10.1 / 30.3 / 87.1 | 10.2 / 29.3 / 90.2 | 8.1 / 30.3 / 83.2 |

TABLE IV
ABLATION STUDIES: MODULES IN DASGIL

| Depth1 | Segmentation2 | Single GAN3 | Multiple GAN4 | Single Triplet Loss5 | Multiple Triplet Loss6 |
|--------|--------------|------------|--------------|---------------------|----------------------|
| ✓      | ×            | ✓          | ✓            | ✓                   | ✓                    |
| ×      | ✓            | ✓          | ×            | ×                   | ×                    |
| ✓      | ✓            | ✓          | ✓            | ✓                   | ✓                    |
| ✓      | ✓            | ✓          | ×            | ×                   | ×                    |
| ✓      | ✓            | ✓          | ✓            | ×                   | ×                    |
| ✓      | ✓            | ✓          | ✓            | ✓                   | ✓                    |
| ✓      | ✓            | ✓          | ✓            | ✓                   | ✓                    |

1. Depth refers to whether depth map generator module is used in DASGIL.
2. Segmentation refers to whether segmentation map generator module is used in DASGIL.
3. Single GAN represents that only single-layer feature is used when training discriminator, where the 5th layer feature is the best among all the 8 layers.
4. Multiple GAN represents that multi-layer features are used when training discriminator of the GAN.
5. Single Triplet Loss represents that only the feature from the 5th layer of the extractor is used to construct the triplet loss, which is the best single-layer result among all the 8 layers.
6. Multiple Triplet Loss represents that features from multiple layers of the extractor are used to construct the triplet loss.

[5] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic et al., “Benchmarking 6dof outdoor visual localization in changing conditions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8601–8610.
[6] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixe, “Understanding the limitations of cnn-based absolute camera pose regression,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3302–3312.
[7] N.-T. Tran, D.-K. Le Tan, A.-D. Doan, T.-T. Do, T.-A. Bui, M. Tan, and N.-M. Cheung, “On-device scalable image-based localization via prioritized cascade search and fast one-many ransac,” IEEE Transactions on Image Processing, vol. 28, no. 4, pp. 1675–1690, 2019.
[8] Y. Feng, L. Fan, and Y. Wu, “Fast localization in large-scale environments using supervised indexing of binary features,” IEEE Transactions on Image Processing, vol. 25, no. 1, pp. 343–358, 2015.
[9] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, “From coarse to fine: Robust hierarchical localization at large scale,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12716–12725.
[10] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.
[11] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, “Brief: Binary robust independent elementary features,” in European conference on computer vision. Springer, 2010, pp. 778–792.
[12] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, “Orb: An efficient alternative to sift or surf,” in ICCV, vol. 11, no. 1. Citeseer, 2011, p. 2.
[13] S. Leutenegger, M. Chli, and R. Siegwart, “Brisk: Binary robust invariant scalable keypoints,” in 2011 IEEE international conference on computer vision (ICCV). Ieee, 2011, pp. 2548–2555.
[14] H. Jégou, M. Douze, C. Schmid, and P. Pérez, “Aggregating local descriptors into a compact image representation,” in CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2010, pp. 3304–3311.
[15] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “Netvlad: Cnn architecture for weakly supervised place recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1437–1451, 2017.
[16] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, “2d/7 place recognition by view synthesis.” IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 2, pp. 257–271, 2018.
[17] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Ucspcroft, and M. Milford, “On the performance of convnet features for place recognition,” in 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2015, pp. 4297–4304.
[18] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in Advances in neural information processing systems, 2014, pp. 2366–2374.
[19] X. Wang, D. Fouhey, and A. Gupta, “Designing deep networks for surface normal estimation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 539–547.
[20] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.
[21] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[22] N. Piasco, D. Sidibé, V. Gouet-Brunet, and C. Demonceaux, “Learning scene geometry for visual localization in challenging conditions,” in 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 9094–9100.
[23] M. Larsson, E. Stenborg, C. Toft, L. Hammarstrand, T. Sattler, and F. Kahl, “Fine-grained segmentation networks: Self-supervised segmentation for improved long-term visual localization,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 31–41.
[24] M. Larsson, E. Stenborg, L. Hammarstrand, M. Pollefeys, T. Sattler, and F. Kahl, “A cross-season correspondence dataset for robust semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9532–9542.
[25] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, “Dada: Depth-aware domain adaptation in semantic segmentation,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7364–7373.
[26] A. Saxena, S. H. Chung, and A. Y. Ng, “Learning depth from single monocular images,” in Advances in neural information processing systems, 2006, pp. 1161–1168.
[27] A. Saxena, M. Sun, and A. Y. Ng, “Make3d: Learning 3d scene structure from a single still image,” IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 5, pp. 824–840, 2008.
[28] J. Xiao and L. Quan, “Multiple view semantic segmentation for street view images,” in 2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 686–693.
[29] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep ordinal regression network for monocular depth estimation,” in Proceedings...
of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2002–2011.

[30] C. Godard, O. M. Aodha, M. Firman, and G. J. Brostow, “Digging into self-supervised monocular depth estimation,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3828–3838.

[31] M. Larsson, A. Arnab, S. Zheng, P. Torr, and F. Khalid, “Revisiting deep structure from motion models for pixel-level labeling with gradient-based inference,” SIAM Journal on Imaging Sciences, vol. 11, no. 4, pp. 2610–2628, 2018.

[32] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, “Semantic image segmentation via deep parsing network,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1377–1385.

[33] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 270–279.

[34] J. Watson, M. Firman, G. J. Brostow, and D. Turkulkhambetov, “Self-supervised monocular depth hints,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2162–2171.

[35] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, “Semantic image segmentation via deep parsing network,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 340–349.

[36] J. Dai, K. He, and J. Sun, “Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1635–1643.

[37] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1851–1858.

[38] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, “Virtual worlds as proxy for multi-object tracking analysis,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4340–4349.

[39] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, “The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3234–3243.

[40] Y. Cabon, N. Murray, and M. Hennenberger, “Virtual kitti 2,” arXiv preprint arXiv:2001.10771, 2020.

[41] M. Long, G. Ding, J. Wang, J. Sun, Y. Guo, and P. S. Yu, “Transfer sparse coding for robust image representation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 407–414.

[42] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, “Deep reconstruction-classification networks for unsupervised domain adaptation,” in European Conference on Computer Vision. Springer, 2016, pp. 597–613.

[43] A. Atapour-Abarghouei and T. P. Breckon, “Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2800–2810.

[44] J. Nath Kundu, P. Krishna Uppala, A. Pahuja, and R. Venkatesh Babu, “AdaDepth: Unsupervised content congruent adaptation for depth estimation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2656–2665.

[45] S. Zhao, H. Fu, M. Gong, and D. Tao, “Geometry-aware symmetric domain adaptation for monocular depth estimation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9788–9798.

[46] S. Sankaranarayanan, Y. Balaji, A. Jain, S. Nam Lim, and R. Chellappa, “Learning from synthetic data: Addressing domain shift for semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3752–3761.

[47] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, “Image to image translation for domain adaptation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4500–4509.

[48] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.

[49] C. Godard, T.-J. Cham, and J. Cai, “T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 767–783.
Hanjiang Hu received the B.Eng degree in Mechanical Engineering from Shanghai Jiao Tong University, Shanghai, China, in 2018. He is currently working toward the M.S. degree in Control Science and Engineering at Shanghai Jiao Tong University, Shanghai, China. His current research interests include visual perception and localization, autonomous driving, computer vision and machine learning.

Ming Cheng is currently pursuing the B.S. degree in Department of Automation, Shanghai Jiao Tong University, Shanghai, China. He is now doing research at Autonomous Robot Lab, Shanghai Jiao Tong University. His research interests are deep learning, computer vision and autonomous driving.

Zhe Liu received his B.S. degree in Automation from Tianjin University, Tianjin, China, in 2010, and Ph.D. degree in Control Technology and Control Engineering from Shanghai Jiao Tong University, Shanghai, China, in 2016. From 2017 to 2020, he was a Post-Doctoral Fellow with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong. He is currently a Research Associate with the Department of Computer Science and Technology, University of Cambridge. His research interests include autonomous mobile robot, multirobot cooperation and autonomous driving system.

Hesheng Wang (SM’15) received the B.Eng. degree in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 2002, and the M.Phil. and Ph.D. degrees in automation and computer-aided engineering from The Chinese University of Hong Kong, Hong Kong, in 2004 and 2007, respectively. He was a Post-Doctoral Fellow and Research Assistant with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, from 2007 to 2009. He is currently a Professor with the Department of Automation, Shanghai Jiao Tong University, Shanghai, China. His current research interests include visual servoing, service robot, adaptive robot control, and autonomous driving. Dr. Wang is an Associate Editor of Assembly Automation and the International Journal of Humanoid Robotics, a Technical Editor of the IEEE/ASME TRANSACTIONS ON MECHATRONICS. He served as an Associate Editor of the IEEE TRANSACTIONS ON ROBOTICS from 2015 to 2019. He was the General Chair of the IEEE RCAR 2016, and the Program Chair of the IEEE ROBIO 2014 and IEEE/ASME AIM 2019.