Learning Shape-based Representation for Visual Localization in Extremely Changing Conditions

Hae-Gon Jeon[1], Sunghoon Im[2], Jean Oh[3] and Martial Hebert[3]

***Abstract*— Visual localization is an important task for applications such as navigation and augmented reality, but is a challenging problem when there are changes in scene appearances through day, seasons, or environments. In this paper, we present a convolutional neural network (CNN)-based approach for visual localization across normal to drastic appearance variations such as pre- and post-disaster cases. Our approach aims to address two key challenges: (1) to reduce the biases based on scene textures as in traditional CNNs, our model learns a shape-based representation by training on stylized images; (2) to make the model robust against layout changes, our approach uses the estimated dominant planes of query images as approximate scene coordinates. Our method is evaluated on various scenes including a simulated disaster dataset to demonstrate the effectiveness of our method in significant changes of scene layout. Experimental results show that our method provides reliable camera pose predictions in various changing conditions.**

I. Introduction

Localizing an agent's location based on the images that the agent views is an important problem in various applications including navigation, augmented reality, and robotics. This problem is known as *visual localization*, where the goal is to infer the 6 degree-of-freedom (DoF) camera poses consisting of translational and rotational information from a query image with respect to a scene model. Previous works [1], [2], [3], [4] on visual localization have shown promising performance by establishing scene coordinates [1] that map image patches to corresponding dense 3D points in a scene model. The studies show that camera pose estimation based on local features works well under a constraint of similar viewing conditions.

Such a constraint, however, can be too restrictive in practice. Scene appearances or geometry may change over time due to a variety of causes such as object movements, variations in illumination, and seasons. Such changes result in a significant performance drop of the visual localization because the scene coordinates may appear inconsistent. As a solution to the changing scene appearance problem, high-level scene abstraction from semantic information is utilized to find the matches between a query image and corresponding points [5], [6], [7]. These approaches still suffer from the scene geometric changes that can increase the variations of semantic labels.

Convolutional neural networks (CNNs) have demonstrated some capacity to address visual localization by leveraging context information inferred from the scene [8], [9], [10], [11], especially when large-scale datasets [8] are available. Unfortunately, according to a recent study in [12], features from CNNs can be highly biased toward recognizing textures rather than shapes . Such a bias can be a more serious issue in those problem domains where large variation in appearance is prevalent, for instance, in the post-disaster environments. Motivated by the idea of reducing the texture bias in [12], we develop an architecture that is robust to appearance variances.

In this paper, we present an end-to-end CNN-based visual localization approach that is robust to scene appearance and geometry changes. Our ideas in building the CNN are twofold: 1) increasing shape bias, and 2) estimating dominant planes of scenes such as road, floor, building, wall, and ceiling. Our CNN is trained on both reference images and stylized images to increase shape bias as in [12]. We further analyze and discuss which seed images for style transfer are useful for this task. In addition, our CNN computes dominant planes of scenes with 3D depth information, which acts as a simple version of scene coordinates and allows to handle scene geometry changes. This is made possible through the use of a plane structure-induced loss represented by 3D information [13] to estimate the planes.

With this end-to-end network for visual localization, we obtain state-of-the-art results over various scenes. In particular, our network shows a promising performance on a challenging synthetic dataset representing disaster scenarios such as buildings on fire and destruction from earthquake in [14]. Ablation studies also indicate that each of these technical contributions leads to appreciable improvements in camera poses prediction.

II. Related Works

We predict a 6-DoF camera pose from an image taken across large appearance variations. We refer the readers to [15] for a comprehensive review of visual localization.

Visual Localization in Changing Conditions A key issue of visual localization in changing conditions is to match new (query) images to previously recorded (database) ones.

[1]Hae-Gon Jeon is affiliated with both AI Graduate School and the School of Electrical Engineering and Computer Science, GIST, Gwangju 61005, South Korea haegonj@gist.ac.kr

[2]Sunghoon Im is with the Information and Communication Engineering, DGIST, Daegu 42988, South Korea (*Corresponding author*) sunghoonim@dgist.ac.kr

[3]Jean Oh and Martial Hebert are with the Robotics Institute of Carnegie Mellon University, Pittsburgh, PA 15213, USA

Acknowledgement This work is in part supported by the Air Force Office of Scientific Research under award number FA2386-17-1-4660, Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST)), GIST Research Institute (GRI) grant funded by the GIST in 2020, and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2020R1C1C1012635).

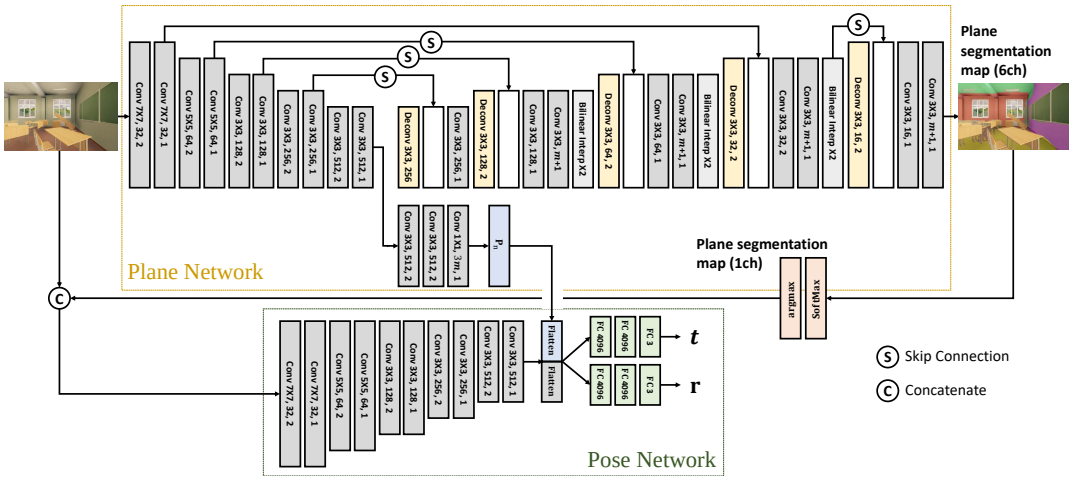


Fig. 1: An overview of the proposed visual localization network. All convolution and deconvolution layers have ReLU except for the prediction layers. The numbers in each block represent a filter size, a channel size, and stride.

In [16], a directed acyclic data association graph with a hand-crafted descriptor is built to formulate this problem as a minimum cost network flow problem. This network flow has been shown to provide a higher gain when integrated with more robust image descriptions from CNNs. In [17], an illumination invariant binary description of image sequences over seasonal changes is utilized and its similarity is computed in terms of the Hamming distance between the binary descriptors of database and query images.

The use of semantic information is another approach to predict the location information of a model [5], [6], [7]. In [5], a CNN model achieves robust global localization by learning visually discriminant regions of interest. The CNN then aggregates the output features from both a reference image and a query image prior to matching them. Using a depth map and a semantic map of a query image, variational CNN captures high-level geometry and semantic information for scene completion [6]. A descriptor from the high-level information is embedded in the Euclidean space and can be matched efficiently. In [7], a semantic consistency score is proposed to match semantic labels between a scene model and a query image. As a last step, a weighted RANSAC refines an initial estimated poses from the score.

CNN-based Visual Localization Starting from PoseNet [8] that regresses 6-DoF camera poses directly from GoogLeNet [18], visual localization, also known as camera pose relocalization, has been actively studied in robotics field. This work is extended by adopting a stochastic Dropout [19] in [9], and a geometric loss minimizing reprojection errors in [20]. Long Short-Term Memory units are incorporated into CNNs to learn structured feature correlations in [21], and to exploit temporal dependencies in input video streams [22]. MapNet [23] takes a data-driven approach to bring geometric constraints used in SLAM or structure from motion into a CNN framework.

For more reliable camera poses prediction, recent works take advantage of the concept of scene coordinates [24], [11], [10]. In [24], a dual-stream CNN encodes input color and depth images separately to minimize a weighted Euclidean distance between ground-truth (GT) and predicted camera poses. A differentiable RANSAC (DSAC) proposed in [11] enables an end-to-end CNN with the scene coordinates from single input images assumption to optimize its initial prediction. In [10], a CNN regresses scene coordinates using RGB-D images. Initial predictions of the CNN are refined by a differentiable module for counting the number of inliers of 2D-3D matches. CNN-based approaches, however, can exhibit strong texture bias of images [12].

Lastly, an approach using global descriptors such as NetVLAD [25] may be able to encode distinctive information when local features of CNNs fail to find a match between a query and a reference; however, its performance on the visual localization problem has not been fully evaluated. Because NetVLAD was developed for an image retrieval task, it does not predict the 6-DoF camera pose. In the benchmark [26], NetVLAD instead takes the pose of a retrieved image in the database as the pose of a query image.

III. Approach

Our network aims to estimate 6-DoF camera poses from a single image in changing conditions. To do this, we increase the shape bias of a CNN and use a geometric consistency between database and query images by predicting the dominant planes of scenes with its parameters. The overall structure is depicted in Fig. 1. We first describe the technical details of our approach to training with stylized images to reduce shape bias in Section III-A. We then discuss the estimation of geometric properties in the form of dominant planes (Section III-B), followed by the estimation of 6D pose from the plane distribution (Section III-C).

A. Increasing the Shape Bias of CNNs

As seen in a benchmark for visual localization [26], the task becomes more challenging as the appearance gap between the images in the database and queries increases, resulting in a lack of sufficient feature matches.

Fig. 2: Performance changes according to the seed image selections. (a) Examples of DISC dataset for simulating before and after building on fire. (b) Seed images with its corresponding contextual loss. The images are categorized based on averaged contextual loss between each seed image and all query images. (c) The x-axis represents combinations of seed images in (b) and the y-axis are units for translation (m) and rotation error (degree), respectively. The horizontal bars represent the errors when the seed images are not used

Conventional end-to-end CNN-based visual localization methods [8], [9], [23], [21] work well in the environments with relatively few appearance changes, but their performance degrades significantly when there are high variations in shape changes.

Such degradation is consistent with the findings from recent research in [12] that CNNs for object recognition tend to learn a representation that is biased toward textures over shapes. CNNs with strong shape representation can be achieved by training a set of images with various textures. By adding variations in texture to the training data, the idea is to train a model that reduces the texture bias and in turn focuses more on shape features. The textures come from different image classes and are synthesized by style transfer.

Following this idea, we develop a visual localization approach that is robust in changing conditions in the context of relocalization. We generate a training dataset using a conventional style transfer method [27] to train our network toward high shape bias of scenes.

Intuitively, our network could perform more effectively if we account for prior knowledge of the variations between the database and the query image. For example, using nighttime images as a seed of the style transfer should be more beneficial for day-night changes when compared to random images. Although selecting seed images based on prior knowledge on a target domain is useful, we also note that it is not required for reducing texture bias.

To validate this idea, we analyze the sensitivity of the prediction accuracy to the types of seed images. Here, we use a synthetic disaster dataset in [14], which will be described in Sec. IV-A. We categorize the seed images that we can use for stylization based on the contextual loss [28] between a seed and a reference images, which measures the similarity between non-aligned images in Fig. 2(b). Seed images in the first row in Fig. 2(b) are searched based on prior knowledge of the query scene[1], and the other images are randomly selected. With these stylization seed images, we generate 12 different combinations and train our network using each of the combinations as seeds for the stylization. As shown in Fig. 2(c), we observe that our network achieves significant performance improvement regardless of any combination of seed images, when compared to the baseline that does not utilize stylized images during training (blue and orange color bars). In addition, since the images with low contextual loss help capture the variation loss of query images as well as increasing shape representations, they contribute to the performance improvement, but the gain is not significant.

[1]We use a keyword "building on fire" for Google image search.

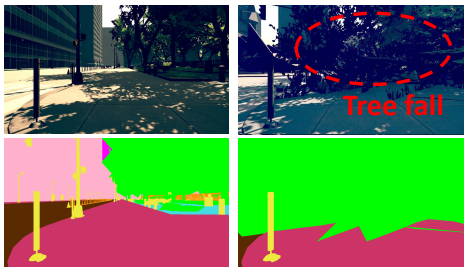
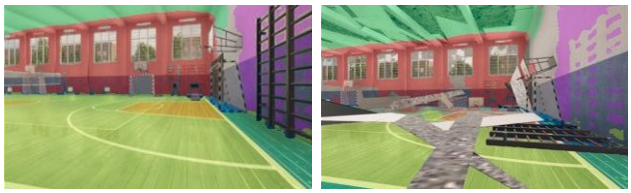


Fig. 3: An example of semantic label changes caused by geometry changes. (Green: vegetation, Plum: sidewalk, Brown: road, Pink: building, Yellow: streetlamp)

B. Dominant Plane Estimation

Previous visual localization methods for changing conditions assume that scene geometry between the database and the query images is consistent. Since the positions of buildings and vegetation do not change, counting feature matches between its semantic labels produces good results [6], [7] in Fig. 3(a). However, geometry changes might cause low semantic consistency matches as shown in Fig. 3(b).

Our idea to solve this problem is to relax the assumption such that, although the positions of semantic features at the fine-grained level can change, a high-level, abstract view may still be retained between the database and the query images. This intuition leads us to use the dominant plane information of the scenes as input to the pose estimation network in Fig. 1. According to a recent work in SLAM in [29], camera poses are computed by using only structural features that have useful geometric information of parallelism, orthogonality, and/or coplanarity in the scene.



(a) (b)

Fig. 4: (a) Before geometry changes. (b) After geometry changes with plane estimation results. Each estimated plane is colorized for visualization.

To extract the structural features–i.e., in our case, a family of salient planes in the scene–we use an encoder-decoder network with skip connections and multiple levels of scales for plane estimation [13]. This network predicts plane segmentation maps and plane parameters. As in [30], an input image is passed through the encoder to produce a set of high-level feature maps and then the decoder upsamples the feature maps via a series of deconvolution layers to infer the plane segmentation maps with $m+1$ channels including the non-planar class where m is a hyper-parameter specifying the number of planes. The multiple scales allow the network to abstract feature maps and the skip connections help preserve high-level information. Another branch of the network infers plane parameters $\mathbf{p}_n = \{a_n, b_n, c_n\}$ where n is the plane index. A 3D point $\mathbf{X}$ belongs to plane n if $\mathbf{p}_n^T X = 1$. This branch shares the same encoder for the plane segmentation maps and has two fully connected layers whose output is $1 \times 3m$ of the m planes.

For this network, we use a depth-induced loss L_P in [13]:

$$L_P = \sum_{n=1}^{m} \sum_{\mathbf{x}} \Psi_n(\mathbf{x}) \left\| \mathbf{p}_n \cdot \mathbf{X} - 1 \right\|_2 + R(\Psi_n(\mathbf{x})), \quad (1)$$

$$\text{where} \quad \mathbf{X} = D(\mathbf{x})\mathbf{K}^{-1}\mathbf{x} \quad \text{and}$$

$$R(\Psi_n) = \sum_{\mathbf{x}} -H(\mathbf{x}) \cdot \log\left(\sum_{n=1}^{m} \Psi_n(\mathbf{x})\right) - (1 - H(\mathbf{x})) \cdot \log\left(\sum_{n=1}^{m} \Psi_n(\mathbf{x})\right),$$

where $\mathbf{x} = [x, y, 1]^T$ is a pixel of input images in homogeneous coordinates, D is a depth map corresponding to an input image, $\mathbf{X}$ is a 3D point, and Ψ indicates the probability of pixel $\mathbf{x}$ belonging to the n-th plane. $\|\cdot\|_2$ is an L_2 norm. $\mathbf{K}$ is an intrinsic matrix of a camera used[2]. R is a cross entropy-based regularization term which is minimized with constant label 1 at each pixel. We define dominant planes including building, road, sidewalk, wall and ceiling as planar class, and $H(\mathbf{x}) = 1$ if a pixel $\mathbf{x}$ belongs to the planar class.

The predicted plane segmentation maps and parameters represent the 3D information of scenes. Since the descriptors allow the CNNs to consider large plane parts of scenes that are less likely to suffer from minor geometric changes, e.g., the floor plane in Fig. 4, we hypothesize that the descriptors based on dominant plane information can make the pose-prediction CNN more robust to geometry changes.

[2]We assume that $\mathbf{K}$ is known in this paper.

C. 6-DoF Camera Pose Prediction

With the estimated plane segmentation maps and its parameters, we predict 6-DoF poses from the pose network. The pose network has a similar structure as the encoder of the plane network, but takes the image and the plane segmentation map as inputs. We reduce $m+1$ channels of the plane segmentation map to one without loss of information by using the softmax and argmax operations. We then pass a concatenation of the image and that one-channel plane segmentation map through the convolution layers.

As illustrated in Fig. 4, the 2D representation of the plan segmentation map cannot fully capture the scene geometry errors of the 3D space, which can cause structural ambiguities of scenes. In order to augment the 2D segmentation map with additional 3D information, we embed the plane parameters $\mathbf{p}_n = \{a_n, b_n, c_n\}$ for each plane n into the encoded feature vector in a similar manner presented in [31].

The encoded feature is passed to the two separate, fully-connected pose-prediction layers to generate the position of the camera $\mathbf{t}$ and its orientation $\mathbf{r}$ encoded as the Euler angles [21]. We minimize a pose loss function L_{RT} as below:

$$L_{RT} = \|\mathbf{t} - \hat{\mathbf{t}}\|_2 + \gamma \|\mathbf{r} - \hat{\mathbf{r}}\|_2, \quad (2)$$

where $\hat{\mathbf{t}}$ and $\hat{\mathbf{r}}$ are ground-truth position and orientation, respectively. γ determines the relative weight of the orientation error with respect to the positional error.

Finally, the loss function L of the whole network is defined as a weighted sum of the plane loss and the pose loss:

$$L = L_P + \lambda L_{RT}, \quad (3)$$

where λ is a scale factor between the plane estimation error and the pose estimation error.

D. Implementation details

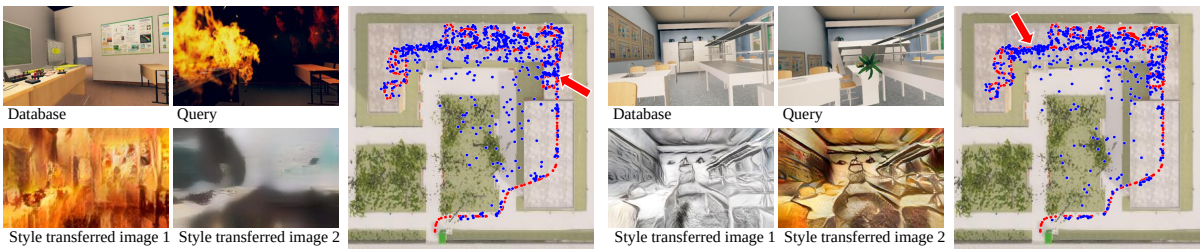
In the training procedure, we use color images with 192×320 resolution and its corresponding depth maps, rotation and translation information. We train our model from scratch for 1M iterations in total and minimize the loss function in Eq.3 with the ADAM optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The batch size is 4 and the learning rate is 10^{-4} in all the iterations. We use Tensorflow [32] to implement our network on one NVIDIA 1080Ti GPU, which takes one day. For data augmentation, we perform spatial transformations including rotation, translation, cropping, zooming and chromatic transformations such as color, contrast, brightness for all training images. A forward pass of our network takes about 0.008 seconds for pose prediction on an image. We set the hyper-parameters m and γ to 5 and 250 in all the experiments.

IV. Experiments

We conduct experiments on various datasets including seasonal, appearance, and geometry changes to demonstrate the effectiveness of the proposed method. For our comparisons, we choose state-of-the-art visual localizations

Scene (Spatial Extent)	Situation	PoseNet [8]	SCoRF [1]	DSAC [11]	Ours
House (10m×10m)	Building on fire	2.78m, 78.122° 0.0 / 0.0 / 12.20	0.23m, 3.58° 0.0 / 0.0 / 0.0	8.19m, 75.43° 0.0 / 0.0 / 24.39	0.36m, 1.43° 34.72 / 69.44 / 95.83
	Destruction	0.67m, 18.55° 0.0 / 9.76 / 26.83	0.12m, 1.98° 0.0 / 0.00 / 0.0	4.83m, 62.21° 0.0 / 0.0 / 33.65	0.22m, 1.29° 33.33 / 65.71 / 94.00
Office1 (5m×9m)	Building on fire	2.28m, 52.15° 0.0 / 0.0 / 18.06	0.31m, 3.43° 0.0 / 0.0 / 0.0	17.17m, 67.69° 0.0 / 0.0 / 0.0	0.29m, 2.96° 18.82 / 66.08 / 91.94
	Destruction	2.79m, 38.89° 0.0 / 0.0 / 29.17	0.32m, 4.14° 0.0 / 0.0 / 0.0	11.16m, 60.06° 0.0 / 0.0 / 19.44	0.49m, 1.87° 15.44 / 60.37 / 89.65
Office2 (18m×20m)	Building on fire	4.23m, 81.55° 0.0 / 0.0 / 11.11	0.38m, 5.41° 0.0 / 0.0 / 0.0	14.60m, 61.93° 0.0 / 0.0 / 0.55	0.33m, 2.69° 19.00 / 67.80 / 91.36
	Destruction	4.42m, 64.08° 0.0 / 0.0 / 5.00	0.29m, 5.67° 0.0 / 0.0 / 0.0	17.29m, 61.75° 0.0 / 0.0 / 9.44	0.29m, 1.18° 42.58 / 70.74 / 98.02
School (90m×80m)	Building on fire	23.58m, 79.14° 0.0 / 0.0 / 0.27	3.08m, 23.07° 0.0 / 0.0 / 0.0	61.50m, 68.96° 0.0 / 0.0 / 0.0	1.38m, 8.59° 3.62 / 15.82 / 74.13
	Destruction	8.66m, 33.72° 0.27 / 2.95 / 28.42	2.26m, 12.11° 0.0 / 0.0 / 0.0	55.97m, 60.94° 0.0 / 0.0 / 0.26	1.30m, 7.16° 5.36 / 18.36 / 72.65

TABLE I: Comparison results on [14] for simulating severe conditions. The first row of each block is a position (m) and a rotation (°) error, and the second row means the accuracy measures with the high-/medium-/low- precisions. Database for each situation of House, Office1, Office2, and School contains 118, 216, 515, and 2123 images, respectively. The number of query images is 41, 72, 180, and 746, respectively.



(a) Building on fire (b) Destruction

Fig. 5: Results from our network on School scene in [14]. In the maps, red and blue dots represent ground-truth positions and the estimated positions, respectively. The red arrows mean positions of the database, query and style transferred images.

Fig. 6: Failure cases of our network.

for PoseNet [8] which has an end-to-end CNN structure, SCoRF [1] which uses scene coordinates, and DSAC [11] which is a CNN-based approach with a differentiable RANSAC process.

In our evaluation, we use common quantitative measures of pose estimation: the median of position and rotation error [8], [9] and pose accuracy with three thresholds: high-precision (0.25m, 2°), medium-precision (0.5m, 5°), coarse-precision (5m, 10°) as in [26]. The translation error is measured in terms of the Euclidean distance $\|\mathbf{t}_{gt} - \mathbf{t}_{est}\|_2$ between a ground-truth position $\mathbf{t}_{gt}$ and a estimated position $\mathbf{t}_{est}$. The rotation error is measured via the absolution orientation error $|\alpha|$ in degree. The $|\alpha|$ is computed by $2\cos|\alpha| = Tr(\mathbf{R}_{gt}^{-1}\mathbf{R}_{est}) - 1$ where Tr is a trace operation of a matrix and $\mathbf{R}$ is a rotation matrix of an orientation $\mathbf{r}$.

A. Severe Conditions

With the simulated disaster scenarios dataset in [14], we compare our method with the comparison methods. Table I summarizes quantitative results. Through this experiment, we observe that a CNN-based approach such as PoseNet [8] has difficulty in predicting accurate camera poses. CNN features from PoseNet are useful for scene discrimination in textureless regions, but are biased toward scene textures. The texture-biased features result in severe rotational error.

We are not able to obtain competitive results as well. In severe changing conditions, SCoRF [1] fails to produce reliable results even though ground-truth depth maps are used as input. We observe an initial set of camera pose hypotheses was not generated due to the difficulty in matching 2D-3D correspondences.

The poor pose estimation results of DSAC [11] is noticeable. One main reason for the lowest performance is that DSAC fails to generate a good initial scene coordinate of query images from a CNN. We observe that the scene coordinates between database and query images is unmatched. As a result, computing reliable geometry of scenes is necessary to account for the changes in scene structures.

Our network tackles this problem by taking advantage of CNN's feature and scene geometry with dominant planes. Particularly, since our network has high shape representation of scenes, it is robust to severe changing conditions. In Fig. 5, we show pose prediction results from our network with input database and style transferred images. Although database images share only small scene structures with the query images, the style transferred images enable to learn shapes of whole images. However, there are several inaccurate prediction points in Fig. 5. The errors come from large

Scene	PoseNet [8]	SCoRF [1]	DSAC [11]	Ours
Synthia1	56.31m, 115.80° 0.0 / 0.0 / 0.0	117.98m, 87.54° 0.0 / 0.0 / 0.0	63.69m, 66.46° 0.0 / 0.0 / 8.55	2.24m, 6.80° 7.76 / 27.87 / 66.85
Synthia2	28.48m, 117.91° 0.0 / 0.0 / 0.0	137.02m, 109.87° 0.0 / 0.0 / 0.0	57.84m, 57.79° 0.0 / 0.0 / 4.61	2.37m, 7.44° 6.23 / 30.24 / 67.52
Synthia4	36.06 m, 114.14° 0.0 / 0.0 / 0.0	130.42m, 108.00° 0.0 / 0.0 / 0.0	85.58m, 70.21° 0.0 / 0.0 / 1.26	3.25m, 8.15° 3.60 / 22.19 / 64.39
Synthia5	26.79m, 119.04° 0.0 / 0.0 / 0.0	126.49m, 129.73° 0.0 / 0.0 / 0.0	60.00m, 61.28° 0.0 / 0.0 / 3.99	1.99m, 6.32° 8.15 / 37.05 / 75.26

TABLE II: Comparison results on four scenes of SYNTHIA dataset. Each scene has 5000 images consisting of spring, summer and fall seasons for training procedure and 100 images of winter season for test.

	Buildings on fire	Destruction
# of ST: None	9.80m, 26.97° 0.0 / 9.64 / 20.71	10.73m, 29.68° 0.0 / 4.19 / 28.15
# of ST: 2	1.57m, 9.14° 7.02 / 48.76 / 78.55	1.62m, 10.36° 6.79 / 45.27 / 76.04
# of ST: 3	1.02m, 6.84° 10.52 / 55.43 / 83.61	0.98m, 6.71° 12.82 / 54.39 / 86.12
# of ST: 4	1.06m, 6.77° 10.11 / 56.03 / 83.99	0.98m, 6.87° 13.09 / 55.63 / 87.01
# of ST: 5	0.98m, 6.50° 11.08 / 56.01 / 84.90	0.97m, 6.73° 12.93 / 54.27 / 86.88
w/o plane parameter	1.51m, 10.03° 7.69 / 47.99 / 78.34	1.52m, 10.88° 6.20 / 46.83 / 74.75
w/o plane map	2.03m, 11.58° 8.14 / 40.20 / 79.81	2.24m / 14.39° 3.95 / 37.26 / 65.33

TABLE III: Ablation Experiment on our buildings on fire and destruction dataset. ST means a style transferred image.

occlusion and texture disappearances as shown in Fig. 6.

B. Seasonal Changes

SYNTHIA dataset [33] provides various conditions in a 3D virtual space such as weather, time, and seasonal changes. Among them, winter season in the dataset presents the most challenging conditions due to scene textures covered in snow. In addition, since the dataset has many dynamic objects such as cars and people, we could test how robust our network is in the presence of them.

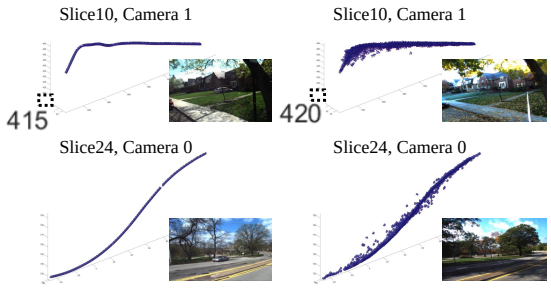
We separate the dataset into database consisting of spring, summer, and fall seasons, and query with the winter season. Each pair of database and query images has 5000 images and 100 images, respectively. Since the dataset simulates realistic traffic conditions, a camera on a virtual car frequently stops during image capture. As a result, the dataset has many images captured in similar view-points. We filter these redundant images using a simple heuristic based on a distance at least $1m$ between two sequential frames.

As shown in Table II, our network outperformed the state-of-the-art methods in seasonal changes. The comparison methods did not work well, especially in very snowy areas. This result confirms the tendency in Sec. IV-A. Since SYNTHIA has larger spatial extent than the School set in Sec. IV-A, the performance drop is inevitable. Many repeated scene elements such as buildings and roads in SYNTHIA dataset also lead to significant performance drop of SCoRF.

C. Ablation Study

An extensive ablation study was carried out to demonstrate the effect of different component on our network. We summarize the results in Table III.

We first examine the performance of our network with respect to the number of style transferred images for one



(a) Database (b) Predicted

Fig. 7: Real-world results on CMU-Seasons dataset. (a) Camera poses of database. (b) Predicted camera poses.

database image. We set the same number of epochs. As shown in Table III, the greater the number of images is, the better the results are. This is because the network learns various situations through more style transferred images. Even though the performance improvement plateaus when three or more style transferred images are used, more images will be beneficial if more various situations in query images exist. Note that we used three seed images to generate style transferred images for all experiments in this work.

We next compare our network with and without the plane information. It is shown that including the plane information leads to significant performance improvements. In particular, plane segmentation maps remove unnecessary features of query images by focusing on locally preserved regions from destruction effects. The plane parameters also help to enhance the performance because plane parameters account for the geometric uncertainty of predicted plane segmentation maps, learning 3D scene information from the tensor of embedding features.

V. Discussion

In this paper, we have presented a convolutional neural network for visual localization robust to changing conditions. To do this, we have increased shape representations of the proposed network by training database images with various textures generated by a style transfer. In addition, our network predicts dominant planes with its corresponding parameters to establish image to 3D correspondences in severe geometric changes. We demonstrate the effectiveness of the proposed network and its components through several evaluations on various scenes.

However, there are still rooms for improvement of the proposed network. When testing our network on a real-world dataset [26], we observe a scale issue of predicted positions as shown in Fig. 7. According to a recent analysis in [34], CNN-based visual localizations learn to predict 6-DoF poses based on linear combinations of database poses. The real-world dataset provides fewer database images and has larger positional gaps between sequential frames than the synthetic dataset [14], [33], which causes the scale issue. We think that an adaptation of a differentiable RANSAC based on dominant plane information can be a way to solve this issue.

REFERENCES

[1] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in rgb-d images," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[2] E. Brachmann, F. Michel, A. Krull, M. Ying Yang, S. Gumhold, *et al.*, "Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] J. Valentin, M. Nießner, J. Shotton, A. Fitzgibbon, S. Izadi, and P. H. Torr, "Exploiting uncertainty in regression forests for accurate camera relocalization," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] A. Guzman-Rivera, P. Kohli, B. Glocker, J. Shotton, T. Sharp, A. Fitzgibbon, and S. Izadi, "Multi-output learning for camera relocalization," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[5] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard, "Semantics-aware visual localization under challenging perceptual conditions," in *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[6] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic visual localization," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] C. Toft, E. Stenborg, L. Hammarstrand, L. Brynte, M. Pollefeys, T. Sattler, and F. Kahl, "Semantic match consistency for long-term visual localization," in *Proc. of European Conference on Computer Vision (ECCV)*, 2018.

[8] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, 2015.

[9] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," in *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[10] E. Brachmann and C. Rother, "Learning less is more-6d camera localization via 3d surface regression," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[11] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "Dsac-differentiable ransac for camera localization," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] R. Geirhos, P. Rubisch, C. Michaelis, F. Wichmann, W. Brendel, and M. Bethge, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in *Proc. of International Conference on Learning Representations (ICLR)*, 2019.

[13] F. Yang and Z. Zhou, "Recovering 3d planes from a single image via convolutional neural networks," in *Proc. of European Conference on Computer Vision (ECCV)*, 2018.

[14] H.-G. Jeon, S. Im, B.-U. Lee, D.-G. Choi, M. Hebert, and I. S. Kweon, "Disc: A large-scale virtual dataset for simulating disaster scenarios," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[15] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[16] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *Proc. of Association for the Advancement of Artificial Intelligence (AAAI)*, 2014.

[17] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, and E. Romera, "Towards life-long visual localization using an efficient matching of binary sequences from images," in *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2015.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[19] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proc. of International Conference on Machine Learning (ICML)*, 2016.

[20] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using lstms for structured feature correlation," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, 2017.

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

[23] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz, "Geometry-aware learning of maps for camera localization," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[24] R. Li, Q. Liu, J. Gui, D. Gu, and H. Hu, "Indoor relocalization in challenging environments with dual-stream convolutional neural networks," *IEEE Transactions on Automation Science and Engineering (TASE)*, vol. 15, no. 2, pp. 651–662, 2018.

[25] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[26] 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 *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[27] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, 2017.

[28] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. of European Conference on Computer Vision (ECCV)*, 2018.

[29] H. Li, Y. Xing, J. Zhao, J.-C. Bazin, and Y. Liu, "Leveraging structural regularity of atlanta world for monocular slam," in *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[30] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, 2015.

[31] L. He, G. Wang, and Z. Hu, "Learning depth from single images with deep neural network embedding focal length," *IEEE Transactions on Image Processing (TIP)*, vol. 27, no. 9, pp. 4676–4689, 2018.

[32] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[33] 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 *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[34] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixe, "Understanding the limitations of cnn-based absolute camera pose regression," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.