Long-Term Visual Localization with Semantic Enhanced Global Retrieval

Hongrui Chen[1], Yuan Xiong[1], Jingru Wang[1], and Zhong Zhou * [1]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China

***Abstract*—Visual localization under varying conditions such as changes in illumination, season and weather is a fundamental task for applications such as autonomous navigation. In this paper, we present a novel method of using semantic information for global image retrieval. By exploiting the distribution of different classes in a semantic scene, the discriminative features of the scene's structure layout is embedded into a normalized vector that can be used for retrieval, i.e. semantic retrieval. Color image retrieval is based on low-level visual features extracted by algorithms or Convolutional Neural Networks (CNNs), while semantic retrieval is based on high-level semantic features which are robust in scene appearance variations. By combining semantic retrieval with color image retrieval in the global retrieval step, we show that these two methods can complement with each other and significantly improve the localization performance. Experiments on the challenging CMU Seasons dataset show that our method is robust across large variations of appearance and achieves state-of-the-art localization performance.**

I. Introduction

Visual localization is a fundamental task for autonomous driving and is especially favorable in scenarios where Global Positioning System (GPS) is unavailable. Estimating the accurate 6-Degree-of-Freedom (DoF) pose of the camera within an existing 3D map is a basic requirement for applications such as autonomous navigation [1], [2], Augmented Reality (AR) [1], Structure-from-Motion (SfM) [3], and Simultaneous Localization and Mapping (SLAM) [1], [4]. Current leading approaches tend to exploit correspondences between 2D features found in a query image and 3D points or structures in a scene model [5], [6]. Image retrieval techniques are often included in a hierarchical pipeline [5], [7], 2D-3D correspondences are then established between retrieved images and the query image. With these 2D-3D matches, the camera pose of the query image can be estimated using an n-point-pose solver inside a Random Sample Consensus (RANSAC) loop. The result is heavily depending on the correctness of the retrieved images and the stability of visual information extracted from the environment. Since the environment is frequently changed , maintaining the robustness of a localization system over changing conditions is still a challenge.

In this paper, we propose to leverage recent advances in semantic segmentation of images, and design a semantic localization framework based on the hierarchical localization paradigm. Existing global image retrieval only includes color image retrieval methods that suffer from variations in light, weather and season. The core idea is to embed the high-level semantic information for retrieval and allow us to handle changes in scene appearance under different conditions. One main challenge is the dynamic labels in semantic segmentations which have a significant impact on the understanding of the scene. Conventional approaches mainly remove dynamic labels or classify them as invalid information, but these methods break the scene's structure layout and seriously hamper the feature extraction. Instead, we propose to modify semantic segmentations to convert the dynamic content into static, and thus recover a static semantic layout that can be better utilized for embedding.

In summary, the main contributions of this paper can be presented as follows:

- Semantic Inpainting Network (SI-GAN) is proposed to convert semantic images that have dynamic objects into those with complete static objects. This alleviates the impact of occlusions and provides extra information of the scene.
- Semantic enhanced global retrieval method is proposed which consists of Score-Map Embedding (SME) and Interval Selection (IS). SME embeds the static semantic segmentation of an image recovered by SI-GAN, and generates a normalized vector for semantic retrieval. IS is furthur conducted to refine retrieval results using the sequential information from the dataset.
- Experiments conducted on the fashionable CMU Seasons dataset show that our approach achieves state-of-the-art performance with an outstanding robustness under challenging conditions with large seasonal variations.

II. Related Work

In recent years, image-based localization has been widely studied. In this section we review other works that related to visual localization with and without semantic segmentation of images.

Non-semantic visual localization only uses the low-level visual features to establish correspondences between the query image and the scene. The relations between different categories in the current viewport are not concerned, so that it is very sensitive to environmental and conditional change. Some researches propose learning-based methods which directly estimate the absolute pose of a query image using a Convolutional Neural Network (CNN) [8], [9], or indirectly estimate the relative pose of a query image in accord with

e-mail: zz@buaa.edu.cn

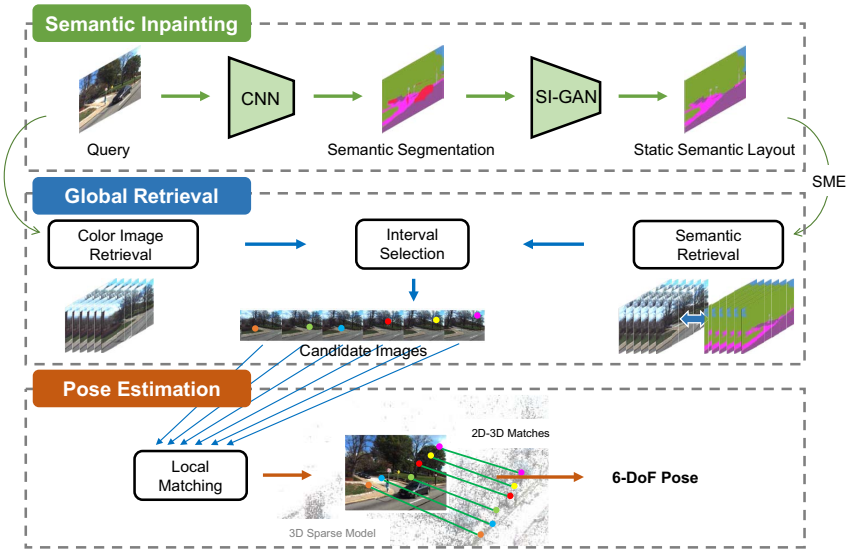

Fig. 1. Illustration of the semantic localization framework proposed in this paper. We divide our framework into three parts: semantic inpainting, global retrieval and pose estimation.

images in the database with ground truth camera poses [10]. These methods rely on extensive training data and their performance are not satisfied. Derived from Structure from Motion (SfM), structure-based methods tend to exploit correspondences between 2D pixels of a query image and 3D points in a 3D model, and use bundle adjustment to solve and refine the camera pose. The 2D-3D matches are established by machine learning techniques [11], [12], or by conventional feature mapping [13]. Image-retrieval based methods [14] tend to represent the pose of a query image using the pose of the best matched candidate retrieved from an image database with known pose. These methods are scale-invariant and robust to condition change. Hierarchical methods follow a coarse-to-fine localization paradigm [5], [7], [15], [16]. Image retrieval methods are used as the coarse step to find a group of candidate images in the image database, and structure-based approaches are applied to the retrieval result for refinement. Hierarchical methods can greatly improve localization accuracy, but the result is heavily relied on the performance of the retrieval method.

Semantic visual localization focuses on the high-level structure features found on image semantic segmentations for localization [6], [17]–[19]. These features neglect details from the scene and describe the image in a macro perspective. Carl Toft, Carl Olsson and Fredrik Kahl [6] created a 3D semantic map using the contour of different semantic labels and projected the map to the current scene to recursivly refine the estimated pose. Instead of projecting the 3D model to a 2D plane, Johannes L. Schonberger, Marc Pollefeys, Andreas Geiger and Torsten Sattler [19] refined the pose estimation by adding 3D-3D voxel match in their semantic map. Image semantic segmentation can also be used for outlier rejection. Some works [20], [21] use the semantic context to directly

TABLE I
CLASS CATEGORIES USED IN SEMANTIC INPAINTING.

label	state	
0	static	unlabeled, others
1	static	sidewalk
2	static	building
3	static	wall, fence
4	static	pole, light, sign
5	static	vegetation, terrain
6	static	sky
7	static	road
8	dynamic	person, rider, car, truck, bus, caravan, trailer, train, motorcycle, bicycle

filter the match of local features. Others [5], [17], [22] use the Semantic Match Consistency (SMC) [17] as a soft outlier rejection method, which projects the 3D semantic map to the hypothesized plane according to the estimated camera pose. The consistency ratio between the semantic labelling of projected structures is then used for biasing sampling in the RANSAC procedure.

Our semantic localization framework combines the hierarchical localization pipeline with image semantic segmentation. In contrast to the previously discussed approaches, which use semantics to improve local feature matching or pose estimation, our approach focuses on the image retrieval stage. Experiments are furthur conducted to study the importance of retrieved images for visual localization. With the combination of semantic features and local visual features, our approach is robust to environmental changes and achieves state-of-the-art performance on the CMU Seasons dataset.

III. Semantic Localization Framework

The main purpose of our approach is to estimate the global pose precisely and improve the robustness and accuracy of localization by using semantic information. Our semantic localization framework follows a hierarchical localization paradigm, shown in Fig. 1. In the first step, our SI-GAN recovers the static semantic layout of the query image. Then, a semantic enhanced global search retrieves candidate images and IS is applied to refine the retrieval result. Finally, a matching-estimation process computes the pose iteratively.

A. Semantic Inpainting

Given a set of label classes $C = S \cup M$, where C is the classes in segmentation, S is the classes for static objects, and M is the classes for dynamic objects. Given a segmentation with labels in C, the semantic inpainting procedure converts the dynamic labels with values in M into appropriate static labels with values in S. Similar to [23], [24], we use a GAN to solve this problem, which is our SI-GAN.

For the semantic information, the DeepLabv3+ [25] with model trained on Cityscapes dataset [26] is used for image segmentation. Cityscapes has more than 30 classes in total and we divide them into 9 subcategories(see Table I) which contains 8 static categories and 1 dynamic categorie. Clustering together similar objects would not only balance the proportion of each categories but also remove redundant ones.

SI-GAN. Similar to recent advances in image inpainting, a coarse-to-fine generative pipeline consists of two steps: a coarse encoder-decoder network recovering local information of an image, and a refined encoder-decoder network using the local information as a reference to generate a final result. Our proposed model extends the EdgeConnect Network [27]: an edge-model is included as the coarse model to recover the full contour of the masked semantic image, then an inpaint-model is employed as the refined model to generate more reasonable static semantic layout guided by the preceding full contour. The original edge-model remains unchanged, and modifications are only applied to the inpaint-model. For the inpaint-model, the encoding layers are replaced with Mobile Blocks [28] and the CBAM [29] attention module is added to the decoding layer. The modified inpaint-model is constituted by 19 convolutional layers: 1 start h-swish layer with 6 Mobile Block layers as the encoder, 4 atrous convolutional layers followed by 2 standard ones as the middle, a final upsampling block of 6 layers with CBAM attention module as the decoder. CBAM is only used before upsampling layers. Through these modifications, the parameters of the inpaint-model are reduced by 53.48%. Different from color image inpainting that outputs continuous value to predict the color, semantic labels are discrete in distribution, so One-Hot Encoding is used to convert the input semantic data to a 9-dimensional tensor, the inpaint-model will output an 8-dimensional tensor. Additionally, the original pixel-wise reconstruction loss, style loss and perceptual loss are replaced by softmax cross-entropy loss, which casts the problem from a regression task to a classification one.

The incomplete One-Hot encoded semantic data $\tilde{S}_{gt}$ conditioned by a composite edge map C_{comp} is used as the input for our inpaint-model. C_{comp} is the composite edge map get from edge-model [27], and $\tilde{S}_{gt}$ can be computed by

$$\tilde{S}_{gt} = epd(S_{gt}) \odot (1 - M_k) \tag{1}$$

where S_{gt} is the ground truth static semantic segmentation and epd denotes the expand dim operation to add the dimension for dynamic class, M_k is the mask as a pre-condition (1 for dynamic labels, 0 for static labels), and $\odot$ denotes the Hadamard product. The network G returns a semantic segmentation S_{pred}, with dynamic labels being replaced:

$$S_{pred} = G(\tilde{S}_{gt}, C_{comp}, M_k) \tag{2}$$

The inpaint-model is trained over a joint loss that consists of a cross entropy loss L_{ce}, adversarial loss L_{adv}, and feature-matching loss L_{fm} [30].

$$\min_G \max_D L_{GAN} = \lambda_{ce} L_{ce} + \lambda_{adv} L_{adv} + \lambda_{fm} L_{fm} \tag{3}$$

where $\lambda_{ce}, \lambda_{adv}, \lambda_{fm}$ are regularization parameters. The adversarial loss is define as:

$$L_{adv} = E[\log D(S_{gt}, C_{comp}) + \log(1 - D(S_{pred}, C_{comp}))] \tag{4}$$

where D is the discriminator. The feature matching loss L_{fm} is defined as:

$$L_{fm} = E[\sum_{l=1}^{n} \frac{1}{H_l W_l} \sum_{h,w} \left\| D^{(l)}(S_{gt}) - D^{(l)}(S_{pred}) \right\|_1] \tag{5}$$

where l is the feature layer of the discriminator, H_l and W_l represent width and height respectively, and $H_l W_l$ denotes the number of elements in layer l. $D^{(l)}$ is the activation in the l'th layer of the discriminator. For our experiments, we choose $\lambda_{ce} = 1$, $\lambda_{adv} = 0.1$ and $\lambda_{fm} = 1$.

The edge-model and inpaint-model are trained on our own Semantic Inpainting dataset along with irregular mask dataset provided by Liu et al. [31]. The Semantic Inpainting dataset is generated using CARLA [32], and the label is adjusted to make it suitable for our work. During the training procedure, we randomly swap the input labels to let the network focuses more on the structure of semantic layout.

B. Global Retrieval

The Global Retrieval step is designed to retrieve images that are similar to the query image and provide candidate images for pose estimation. There are two retrieval pipelines in our framework: color image retrieval pipeline and semantic retrieval pipeline. Color image retrieval is performed by matching the query image q with database images using global descriptors generated by NetVLAD [14], and generates a group of candidate images R_V which represent potential places in the map. Semantic retrieval is performed by matching the query static semantic image S_{pred} with static semantic database images using semantic descriptors generated by SME, and also generates a set of candidate images R_S. In both pipelines, the normalized L_2 distance of descriptors is used to differentiate potential candidates. The retrieved images from

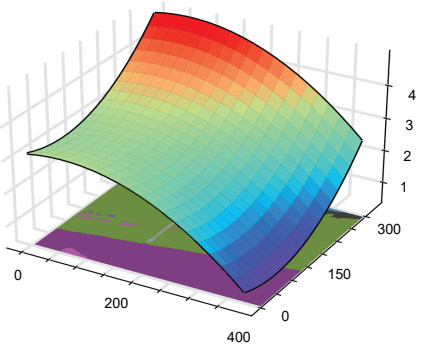

Fig. 2. A score map is a discrete two-dimensional distribution for grading each pixel of the static semantic image. The score increases from blue to red in the map.

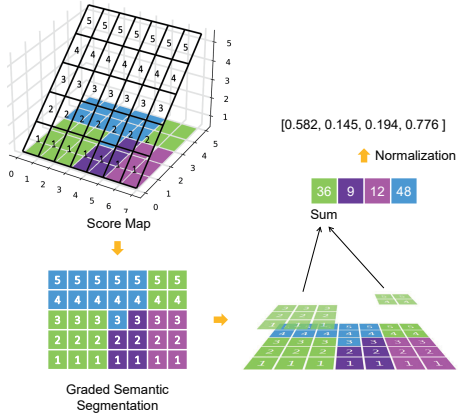

Fig. 3. A brief example of SME. The size of the semantic segmentation is $7 * 5$ and different colors represent different semantic classes. The number in the grid cell represents the grade defined by the score map, the score map in this example is simple and can be described as a 2D plane. SME caculates the total score for each class in the semantic segmentation with a score map and generates a normalized vector.

R_V and R_S are then combined by IS to provide a more robust group of candidate images for pose estimation.

Score-Map Embedding (SME). Since the static semantic segmentation of image S_{pred} contains less visual information for feature extraction, we make our decision based on the common knowledge of spatial distribution of objects (eg. The sky is on the top right corner, the building is on the left, no people in the scene, etc.). This distribution represents the discriminative layout of the scene which is useful for localization. Motivated by this, we propose the Score-Map Embedding method to extract this layout information. The score map M_s is a discrete two-dimensional distribution that is used as a map to guide the sampling process on the semantic image. Fig. 2 is an illustration of a score map and Fig. 3 is a brief example of how SME works. Each position p in the score map is assigned a $score_p$ which represents the score for that position. For each class label c in S_{pred}, we collect its total score TS_c by summing up scores for all related cells in M_s. These two operations inherently contain the position and quantity information of each label class:

$$TS_c = \sum_{p \in M_s} (\mathbb{I}_p score_p) \tag{6}$$

where $\mathbb{I}_p$ is the indicator function, which is 1 when the class in position p is the same as c. These scores are then combined into a vector with normalization, which is the embedding result v_{ms} for the given score map. Different score maps represent different distributions that focus on different layout features of the static semantic image. Thus, multiple score maps can be added together to generate a more descriptive descriptor. In our work, we use 4 score maps to generate a 32-dimentional descriptor as a representation of a semantic image.

Interval Selection (IS). Given two sets of images R_V and R_S from the retrieval stage which represent candidates from the color image retrieval pipeline and semantic retrieval pipeline respectively. Since the database consists of image sequences with consecutive camera poses, each query image matches a sub sequence of images from the database. We first place these candidates in an order and build a long query sequence, then we use IS to pick a sub sequence slice that contains the query image with highest probability.

The illustration of our method is shown in Fig. 4. Each image in the database is assigned an ID representing its relative position in the database sequence. A fixed-sized interval window is used to slide through the query sequence to find the interval with the minimun span calculated by subtracting left window ID from right window ID. Candidate images within this minimun interval will be selected for pose estimation. This method works because similar images are most likely to be close to each other in the sequence, and an effective retrieval system will provide accurate result which would fall in the same region. The value of the window size is determined by the dataset attributes and the practical demands. Firstly, in the sequential datasets a place can appear in several continuous frames, and the maximum number of related frames determine the upper bound of the window size. Secondly, large window size contains more candidate images for pose estimation, which means more time for computation. In our work, we set the window size to 10.

C. Pose Estimation

The Global Retrieval step provides a group of candidate images with the minimun interval span. We match the features in the query image with the features in candidate images using local feature matching methods, SuperPoint [33] and SuperGlue [34]. The 3D model of the scene is built offline using the database images through a SfM pipeline [3], so each database image contains a set of 3D points in the model. Notice that the features of candidate images only contain those used in the SfM procedure and corresponds to a 3D point in the 3D model, thus the correspondences between 2D keypoints in the query image and the 3D points contained in the model are established. We finally feed all the matches to a RANSAC PnP solver to estimate the camera pose.

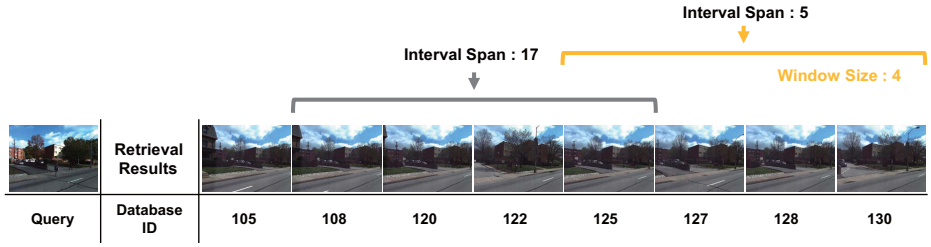

Fig. 4. The IS uses a fixed-sized sliding window to filter candidate images with the minimun interval span. Each database image is assigned an ID so that interval span can be calculated by subtracting window left ID from window right ID.

TABLE II
PERPIXEL ACCURACY ON THE MICC DATASET.

Method	Perpixel accuracy(%)
NS [35]	63.79
LB's Method [24]	70.58
SI-GAN(ours)	**70.78**

IV. EVALUATION

In this section we present experimental evaluations of the proposed methods. We first evaluate the SI-GAN to test the capability of our inpainting network. Then we evaluate the localization performance of our semantic localization framework to examine the applicability of SME and IS to large-scale localization problems. Finally we conduct ablation study on five different localization runs to furthur analysis the contribution of SME and IS for localization.

A. SI-GAN Evaluation

Dataset. We test our model on the MICC-SRI Semantic Road Inpatining dataset [24], a virtual dataset generated with CARLA driving simulator. The dataset contains 11,913 pairs of perfectly aligned frames with and without dynamic objects.

Results. The input segmentation is obtained from Deeplabv3+ and the pixel-wise accuracy is calculated as the criterion for performance. We compare our model with traditional image inpainting method Navier-stokes (NS) [35] and Lorenzo Berlincioni's method [24]. Results in Table II show that our method achieves the best performance. Fig. 5 shows the inpainting result for different methods. Fig. 6 shows a sample of images generated by our model, the input image is form real world and the colors of edge map is reversed for visualization. Since SI-GAN considers the edge information of the scene, it can better deal with the boundary between different semantic classes.

B. Localization Evaluation

Dataset. Since our semantic localization framework is constrained to sequential datasets, we evaluate our method on the CMU Seasons dataset [36], which is a sequential dataset especially for long-term visual localization. The images were captured by two front-facing cameras mounted on a car. The dataset depicts urban, suburban, and park scenes in the area of Pittsburgh, and is recorded over a period of one year, which contains challenging conditions under varying seasonal change. The whole dataset is split into 17 slices, and a 3D reconstructed model is provided. We evaluate our semantic localization framework with SME and IS, named SemSeq. The color image retrieval pipeline and semantic retrieval pipeline

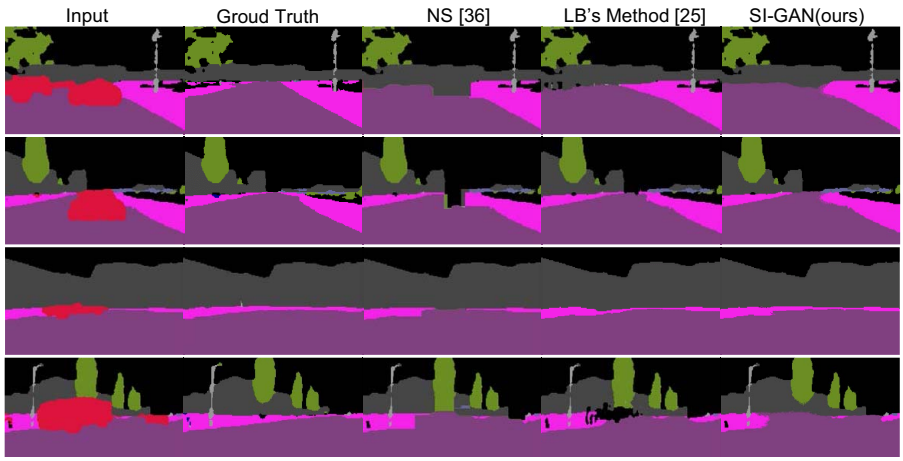

Fig. 5. Inpainting results on the MICC-SRI dataset.

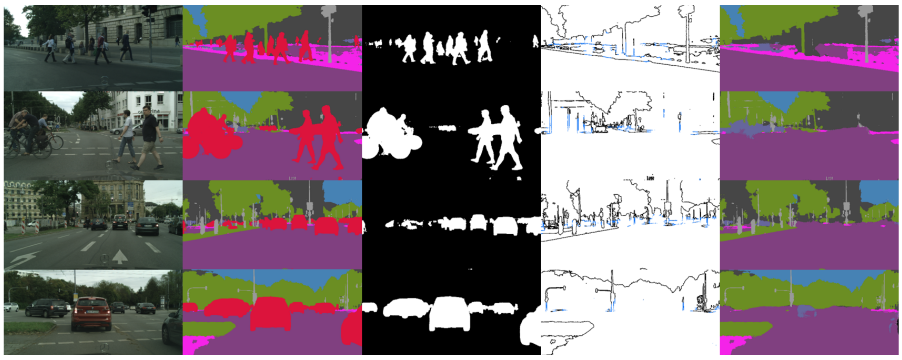

Fig. 6. The experimental result of SI-GAN on real world scenes. (Left to Right) Original image, original semantic segmentation, input mask, generated edge, semantic inpainting result.

TABLE III
EVALUATION OF THE LOCALIZATION ON THE CMU SEASONS DATASETS. WE LIST THE RECALL [%] AT DIFFERENT DISTANCE AND ORIENTATION THRESHOLDS, THE **BEST** RESULT IS HIGHLIGHTED.

	CMU Seasons		
	urban	suburban	park
distance [m]	.25/.50/5.0	.25/.50/5.0	.25/.50/5.0
orient. [deg]	2/5/10	2/5/10	2/5/10
AS [37]	68.9 / 75.7 / 83.4	36.2 / 44.4 / 56.0	24.8 / 31.1 / 41.5
CSL [13]	36.7 / 42.0 / 53.1	8.6 / 11.7 / 21.1	7.0 / 9.6 / 17.0
DenseVLAD [38]	22.2 / 48.6 / 92.8	9.8 / 26.6 / 85.2	10.3 / 27.1 / 77.0
NetVLAD [14]	17.4 / 40.3 / 93.2	7.6 / 21.0 / 80.5	5.6 / 15.7 / 65.8
SMC [17]	75.0 / 82.1 / 87.8	44.0 / 53.6 / 63.7	30.0 / 37.9 / 48.2
NV+SP [7]	91.7 / 94.6 / 97.7	74.5 / 81.5 / 91.3	54.3 / 62.5 / 79.0
SemSeq(ours)	**96.4 / 97.5 / 98.4**	**90.1 / 92.9 / 95.7**	**77.4 / 82.3 / 87.5**

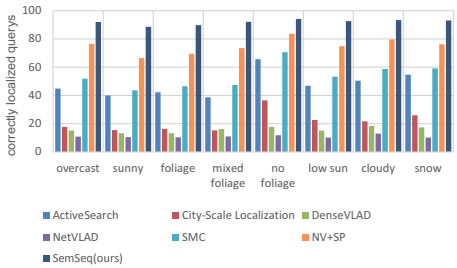

Fig. 7. Localization recall[%] on the CMU Seasons dataset under different conditions at threshold 0.25m, 2°.

both contribute 10 candidate images in the global retrieval step, and these 20 candidates are filtered by IS with a window size of 10. The results are reported as the percentage of query images which were localized within three given translation and rotation thresholds, as defined by the benchmark [36]. We compare our method against several localization approaches. More concretely, we compare against ActiveSearch (AS) [37] and the City-Scale Localization (CSL) [13] which are 2D-3D direct matching methods. In addition, we compare against DenseVLAD [38] and NetVLAD [14] which are image retrieval methods. We also consider some recently introduced methods. The Semantic Match Consistency (SMC) [17] relies on a 3D semantic map, and use semantic match score to weight the point selection in the RANSAC procedure. Hierarchical Localization [7] is a robust localization architecture that representing the current state-of-the-art in terms of accuracy. Other recent works are not concerned because they either need a trainable dataset [15] or rely on the stability of the scene [16] which is contradicted with the basis of CMU seasons dataset. Table III shows the localization result for different methods in different scenes. Fig. 7 shows the Localization result under different conditions at threshold 0.25m, 2°.

Results. As can be seen, our proposed method leads to significant improvement in localization performance for all scenes and all conditions on the CMU Seasons dataset. In the most challenging park scene, our localization accuracy increased by 23.1% than NV+SP under the 0.25m and 2° threshold. Overall, SemSeq sets a new state-of-the-art on the CMU Seasons dataset and shows that our approach is both more accurate and more robust.

Ablation Study. In Table IV, we present an ablation study of our approach on the CMU Seasons dataset. We test the performance of localization on five different runs with 10 candidate images from different methods:

- Candidate images from retrieval result of NetVLAD, named NV.
- Candidate images from semantic retrieval, named SEM.
- Candidate images from 20 coarse candidates from NetVLAD filtered by IS with a window size of 10, named NV-IS.
- Candidate images from 20 coarse candidates from semantic retrieval filtered by IS with a window size of 10, named SEM-IS.
- Candidate images from 10 NetVLAD retrieval candidates mixed with 10 semantic retrieval candidates filtered by IS, named NV-SEM-IS, which is the same condition as SemSeq.

From SEM, we note that only with semantic retrieval pipeline, our localization framework can still work, which proves the effectiveness of SME. Comparing NV with NV-IS and SEM with SEM-IS, we can see the power of IS which leads to significant improvement in localization performance for all conditions. IS uses the sequential information from the dataset to refine the retrieval results, thus we can collect a

Fig. 8. Candidate images from three localization runs. NV represents localization with NetVLAD retrieval, SEM represents localization with semantic retrieval, and NV-SEM-IS combines the results of NV and SEM by IS. The number in the image is the ID representing its relative position in the database sequence. Improper candidates are filtered out after IS, for example 215, 134.

TABLE IV
ABLATION STUDY ON THE CMU SEASONS DATASET. WE COMPARE THE RECALL[%] OF LOCALIZATION FOR DIFFERENT METHODS.

	Distance[m] / Orient.[deg]	NV	SEM	NV-IS	SEM-IS	**NV-SEM-IS (SemSeq)**
Urban	0.25 / 2	93.5	73.2	95.7	75.5	**96.4**
	0.5 / 5	94.8	75.4	96.8	77.7	**97.5**
	5 / 10	96.0	78.3	97.7	80.3	**98.4**
Suburban	0.25 / 2	83.1	52.5	88.5	56.2	**90.1**
	0.5 / 5	86.1	55.9	91.6	59.6	**92.9**
	5 / 10	89.0	60.9	94.5	64.5	**95.7**
Park	0.25 / 2	66.9	45.6	75.9	48.4	**77.4**
	0.5 / 5	72.0	50.2	80.9	53.0	**82.3**
	5 / 10	78.1	56.7	86.1	59.3	**87.5**

set of images that contains more accurate and robust 2D-3D matches. NV-IS performs better than SEM-IS, beacuse in most cases visual information is more accurate and more abundant than semantic layout information. However, in some cases it does not perform well as expected. Fig. 8 shows some counter-examples of candidate images from different localization runs that semantic retrieval performs better than NetVLAD retrieval. The first example is mainly because lacking strong visual features, the scene is full of trees under changes in season and lighting which makes it harder for robust feature extraction. The second example is mainly because dynamic occlusions, dynamic objects can affect the features extracted from the scene thus influence the retrieval result. However, in both conditions, the static semantic layout is stable which leads to more reasonable retrieval results for the semantic retrieval pipeline. So IS fuses the retrieval results from these two retrieval pipelines and improves the performance of NetVLAD retrieval in hard cases using semantic retrieval. This proves that visual information and layout information can complement with each other, and that is why NV-SEM-IS performs better than NV-IS and SEM-IS.

We also provide extra information from NV-SEM-IS that can show the effectiveness of SME. The 10 candidates from IS are a mix of two retrieval pipelines, so we calculate the ratio of candidates from semantic retrieval pipeline. The ratio is used as a brief indication for the contribution of semantic retrieval, and we get an average result of 43.48%.

V. CONCLUSION

In this paper, we have presented a robust and accurate method for long-term visual localization. Our method outperforms state-of-the-art localization approaches on the challenging CMU Seasons dataset that contains substantial appearance variations across weather conditions and seasons. Our semantic localization framework can leverage correspondences between candidate images to provide robust 2D-3D matches. We demonstrate the enhancement of retrieval pipeline that utilizes structured spatial information from semantic image segmentations, combined with the color-image-based pipeline. Experimental results show that the proposed method achieves remarkable improvement over existing approaches in terms of camera pose estimation.

Even though the final results are accurate, there are still some constraints need to be concerned. First, our IS can only be applied to sequential datasets. Second, the dimension of semantic descriptor is determined by the scene specifically, which means that the amount of score maps is relative to the scale of dataset. In the future work, we expect to extend our IS method with spatial information so that our semantic localization framework can apply to non-sequential datasets.

ACKNOWLEDGEMENT

This work is supported by the National Key R&D Program of China under Grant No.2018YFB2100603, the Natural Science Foundation of China under Grant No. 61872024 and the Strategic Consulting Research Project of Henan Research

Institute of China Engineering Science and Technology Development Strategy under Grant No.2021HENZDA03.

REFERENCES

[1] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[2] M. Bürki, L. Schaupp, M. Dymczyk, R. Dubé, C. Cadena, R. Siegwart, and J. Nieto, "Vizard: Reliable visual localization for autonomous vehicles in urban outdoor environments," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1124–1130.

[3] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[4] H. Zhou, D. Zou, L. Pei, R. Ying, P. Liu, and W. Yu, "Structslam: Visual slam with building structure lines," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 4, pp. 1364–1375, 2015.

[5] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "Inloc: Indoor visual localization with dense matching and view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7199–7209.

[6] C. Toft, C. Olsson, and F. Kahl, "Long-term 3d localization and pose from semantic labellings," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 650–659.

[7] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 716–12 725.

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

[9] F. Xue, X. Wu, S. Cai, and J. Wang, "Learning multi-view camera relocalization with graph neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 11 372–11 381.

[10] Z. Laskar, I. Melekhov, S. Kalia, and J. Kannala, "Camera relocalization by computing pairwise relative poses using convolutional neural network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 929–938.

[11] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2930–2937.

[12] E. Brachmann and C. Rother, "Expert sample consensus applied to camera re-localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7525–7534.

[13] L. Svärm, O. Enqvist, F. Kahl, and M. Oskarsson, "City-scale localization for cameras with known vertical direction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1455–1461, 2016.

[14] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[15] M. Humenberger, Y. Cabon, N. Guerin, J. Morat, J. Revaud, P. Rerole, N. Pion, C. de Souza, V. Leroy, and G. Csurka, "Robust image retrieval-based visual localization using kapture," *arXiv preprint arXiv:2007.13867*, 2020.

[16] Q. Zhou, T. Sattler, and L. Leal-Taixe, "Patch2pix: Epipolar-guided pixel-level correspondences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4669–4678.

[17] C. Toft, E. Stenborg, L. Hammarstrand, L. Brynte, M. Pollefeys, T. Sattler, and F. Kahl, "Semantic match consistency for long-term visual localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 383–399.

[18] T. Shi, S. Shen, X. Gao, and L. Zhu, "Visual localization using sparse semantic 3d map," in *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2019, pp. 315–319.

[19] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic visual localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6896–6906.

[20] N. Kobyshev, H. Riemenschneider, and L. Van Gool, "Matching features correctly through semantic understanding," in *2014 2nd International Conference on 3D Vision*, vol. 1. IEEE, 2014, pp. 472–479.

[21] 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/CVF International Conference on Computer Vision*, 2019, pp. 31–41.

[22] H. Taira, I. Rocco, J. Sedlar, M. Okutomi, J. Sivic, T. Pajdla, T. Sattler, and A. Torii, "Is this the right place? geometric-semantic pose verification for indoor visual localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4373–4383.

[23] Y. Song, C. Yang, Y. Shen, P. Wang, Q. Huang, and C.-C. J. Kuo, "Spg-net: Segmentation prediction and guidance network for image inpainting," *arXiv preprint arXiv:1805.03356*, 2018.

[24] L. Berlincioni, F. Becattini, L. Galteri, L. Seidenari, and A. Del Bimbo, "Road layout understanding by generative adversarial inpainting," in *Inpainting and Denoising Challenges*. Springer, 2019, pp. 111–128.

[25] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[26] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[27] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "Edgeconnect: Generative image inpainting with adversarial edge learning," *arXiv preprint arXiv:1901.00212*, 2019.

[28] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314–1324.

[29] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[30] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[31] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 85–100.

[32] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[33] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[34] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[35] M. Bertalmio, A. L. Bertozzi, and G. Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1. IEEE, 2001, pp. I–I.

[36] 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.

[37] 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.

[38] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1808–1817.