2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

SAMLoc: Structure-Aware Constraints With Multi-Task Distillation for Long-Term Visual Localization

Jian Ning[1], Yunzhou Zhang[1*], Xinge Zhao[1], Sonya Coleman[2], Kunmo Li[1], Dermot Kerr[2]

***Abstract*— Real-time and robust long-term visual localization is a crucial technology for autonomous driving. Season and illumination variance make this problem more challenging. At present, most of excellent visual localization algorithms cannot run in real-time on devices with limited computing resources. In this paper, we propose SAMLoc, a structure-aware and self-supervised visual localization system, for fast and robust 6-DoF localization. To obtain structural features in the scene, we propose local and global structure-aware constraints using edge information. Then, we integrate the structure-aware constraints into the hierarchical localization network of multi-task distillation, which significantly reduces the feature extraction time while ensuring localization accuracy. As a result, real-time and robust large-scale localization can be achieved on mobile devices. Experimental results on public datasets show that our system can achieve high localization accuracy and have satisfactory real-time performance. Compared with several state-of-the-art visual localization systems, our framework achieves a competitive localization performance.**

I. Introduction

Real-time large-scale visual localization is a classic problem in computer vision and one of the key steps for autonomous driving. In recent years, the requirements for the accuracy and speed of visual localization have increased with the continuous maturity of autonomous driving technology, especially in some challenging scenes, such as illumination, weather, or seasonal variance. At present, the common visual localization methods include 2D-2D, End-to-End, 2D-3D and other methods. Compared with traditional visual localization methods [1][2], deep learning-based methods [3][4][13] have obvious advantages.

2D-2D visual localization usually adopts image retrieval methods [4][7][11]. Such approaches usually build global feature descriptors of deep learning and perform place recognition. However, as a result it is difficult to determine the exact pose of the image when the scenes are similar or the perspective changes. Recent work [4] fused semantic and geometric information into 2D-2D localization, and Hu et al. [8][24] introduced domain adaptation, these methods significantly improve the effect of localization, but fine localization accuracy still needs to be improved.

*The corresponding author of this paper

[1]Jian Ning, Yunzhou Zhang, Xinge Zhao and Kunmo Li are with College of Information Science and Engineering, Northeastern University, Shenyang 110819, China. zhangyunzhou@mail.neu.edu.cn

[2]Sonya Coleman and Dermot Kerr are with School of Computing and Intelligent Systems, Ulster University, N. Ireland, UK.

This work was supported by National Natural Science Foundation of China (No. 61973066), Major Science and Technology Projects of Liaoning Province(No. 2021JH1/10400049), Fundamental Research Funds for the Central Universities(N2004022).

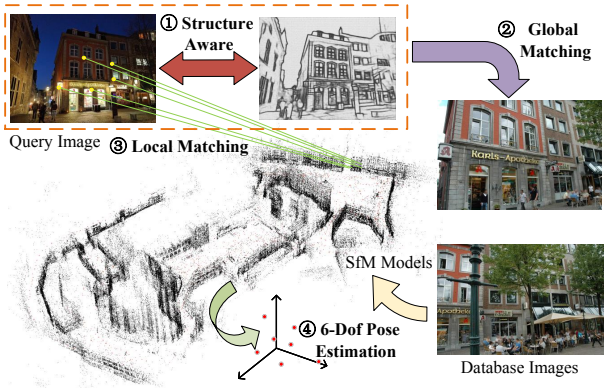


Fig. 1. **Structure-Aware With Multi-task Distillation for Hierarchical Localization.** The global and local descriptors were obtained under structure-aware constraints, then global search was used to estimate the approximate pose, and local feature matching was used to estimate the fine 6-DoF pose.

End-to-End methods [9][10] have noticeable improvements compared with the methods based on image retrieval, but such methods only obtain information from the input image. Since the pose estimation process has no geometric constraints, the generalization performance is usually not strong. It generally needs to be retrained in a new scene to adapt to it. Additionally, in real driving environments, ground-truth is often hard to obtain, especially in challenging outdoor scenarios.

2D-3D-based methods [5][12][13] are still the popular approaches to visual localization, although many excellent localization systems [14][15][21] have emerged so far, which have good localization accuracy. Nevertheless, when the background environment becomes very large, the process of matching can take a long time. In edge devices with limited computing resources, most models are limited by the huge amount of parameters and the huge volume of calculations in the 2D-3D matching process. Sarlin [17] proposed a hierarchical localization paradigm and multi-task distillation model that not only takes the model accuracy into account but also maintains high efficiency on mobile devices.

Currently, approaches to visual localization achieve excellent accuracy, but the speed needs to be improved. The multi-task distillation method can greatly improve the speed of feature extraction, but there is also a loss in accuracy to some extent. Edge features can accurately extract the structural information in the scene [18][19][23]. In [16], edge information is used to track camera motion to improve the accuracy of pose estimation. Also, [20] fused edge features with optical flow estimation to solve the boundary flow

estimation problem for consecutive frames. In [22], edge information was also used to improve the accuracy of place recognition by employing knowledge distillation.

In this paper, we argue that the model should pay more attention to structural features in the scene while learning feature extraction. We combine the structure-aware process with multi-task distillation, and use it for hierarchical localization (Fig.1). Our model can maintain high speed, accuracy, and generalization ability even in challenging scenarios.

Our main contributions are as follows:

- We propose SAMLoc, a self-supervised hierarchical localization framework, based on multi-task distillation, by using structure-aware constraints for long-term visual localization.
- We propose a structure-aware module and integrate it into local and global structure-aware constraints that capture robust features in the scene. Furthermore, the reason why it can be robust and efficient is also explained.
- We demonstrate the effectiveness of our method and the excellent generalization of our model through extensive experiments, and realize real-time operation on mobile devices.

II. Related Works

A. Edge Feature and Structure-Aware

The edge information contains stable structural features in the scene, it is not affected by conditions such as illumination, so it has a good ability to resist environmental changes. In [22], edge information was used to assist the encoder to capture stable structural information during the feature decoupling process utilizing knowledge distillation. Inspired by [18] and [26], [19] proposed a deep structural model for extracting image edge information and generating thin edges. In this method, the encoders of each main block of the network are output separately to obtain edge information containing different orientations, and the outputs of each sub-block are connected to ensure that each depth block can retain the edge features. In our method, the multi-level edge outputs are divided into shallow and deep layers, and integrated as local structure-aware and global structure-aware modules, respectively, making it suitable for multi-task distillation processes.

B. Multi-task distillation with hierarchical localization

Traditional 2D-3D visual localization methods rely on the direct matching of query images to SfM models [14][15], which are computationally expensive and difficult to run in real-time on mobile devices. Yang et al. [13] proposed to achieve efficient 2D-3D matching by compressing the SfM model. Sarlin [17] proposed the paradigm of hierarchical localization. This method significantly improves the operating efficiency. Multi-task distillation [25] is also applied to visual localization, making the model lightweight enough to be easily deployed on mobile devices. This self-supervised framework can greatly reduce the cost of data training because it does not require ground-truth, which has great advantages and potential. Our method incorporates structure-aware constraints into hierarchical localization of multi-task distillation, which greatly reduces the feature extraction time for hierarchical localization while maintaining accuracy.

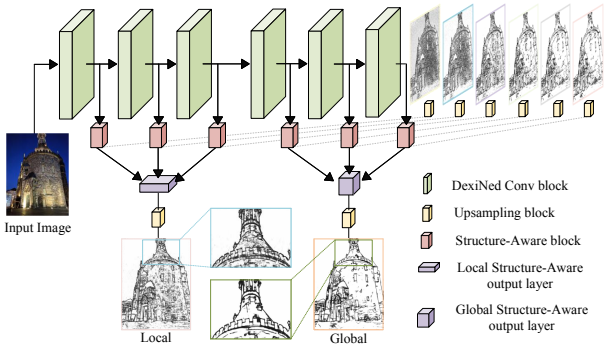


Fig. 2. **Local and global structure-aware constraints** focus on the detailed texture and robust features of the input image, respectively, and obtain visualization through upsampling layers. They are output from the output layer.

III. Method

We design a structure-aware module for scene feature extraction and combine it with a hierarchical localization network based on multi-task distillation to achieve real-time robust visual localization.

A. Structure-Aware Constraint

We are inspired by the DexiNed [19] edge detection network with multi-channel outputs. Image pixels can change with light and seasonal factors, but the contour information of objects such as houses and street lamps is stable and unchanged, thus the edge information is very reliable. We improve based on the DexiNed network as shown in Fig.2.

To adapt to the multi-task distillation network, we propose the concepts of local structure-aware constraints and global structure-aware constraints. We divide the network into two parts, the first part is used to acquire detailed structural texture, and the second part is used to capture the structure information of the scene.

Compared with the original network, our improvements are as follows:

- We insert a side structure-aware module before the upsampling block of each side output to improve the structure-aware capability of the network. It consists of a 1×1 Conv layer and an Upsampling Block[19].
- We use the output ports of the first three blocks and the last three blocks of the network as the output of the structure-aware constraint through 1×1 convolution respectively, and then obtain the local structure-aware and global structure-aware visualization images through the upsampling block.
- We remove the last fusion layer of the original network because it reduces the training effect of the structure-aware module and we modify the loss function.

We use X and Y to represent the input image and the groundtruth of the edge, respectively. Our edge network extractor has six side output layers and each of them generates

predictions through the Upsampling Block. We use $w^{(i)}$ to represent the weight of the output of the i-th layer, then the total weight can be expressed as $W = (w^{(1)}, w^{(2)}, \ldots w^{(6)})$. The side output layer follows the method in [19], and the loss function of each output layer is set as:

$$L^i_{\text{output}}\left(W, w^i\right) = -\varepsilon \sum_{j \in Y+} \log f\left(y_j = 1 \mid X; W, w^i\right) - (1-\varepsilon) \sum_{j \in Y-} \log f\left(y_j = 0 \mid X; W, w^i\right) \quad (1)$$

For our structure-aware module, we use cross-entropy as the local and global structure-aware loss function, thus the final loss function is as follows:

$$L = \left(\sum_{i=1}^{6} \alpha_i L^i_{\text{output}}\left(W, w^i\right)\right) + \beta CrossEntropy\left(Y, \hat{Y}^l_{\text{pred}}\right) + \gamma CrossEntropy\left(Y, \hat{Y}^g_{\text{pred}}\right) \quad (2)$$

where Y^+ and Y^- represent edge pixels and non-edge pixels respectively, $\delta = |Y^-|/|Y^+|$, $f(y_j = 1|X; W, w^i) = sigmoid(a^i_j)$ is the activation value of the sigmoid function at pixel j, $\hat{Y}^l_{pred}$ and $\hat{Y}^g_{pred}$ represent the prediction outputs of local structure awareness and global structure awareness, respectively, α, β, and γ are the hyperparameters that we use to adjust the weights. Our structure-aware constraint module will be added to the hierarchical localization network of multi-task distillation for knowledge distillation which will be mentioned in Section III-B.

The visualization results of local structure awareness and global structure awareness are shown in the enlarged view of Fig.2. The left side is the local structure perception output, which extracts features from the shallow layers of the network, thus it pays more attention to the detailed texture of the scene. The right side is the global structure-aware output, which has stable and robust structural information from the deep layers of the network, so it has the characteristics of good resistance to environmental changes. Our local structure-aware constraint and global structure-aware constraint modules are output through the output layer in Fig.2 and added to the multi-task distillation network in Section III-B. We demonstrate the role of structure-aware through experiments in Section IV-C.

B. Multi-task Distillation and Hierarchical Localization

We demonstrate that the majority of the time for hierarchical localization is spent on feature extraction through experiments in Section IV-E. In order to greatly improve the localization speed while maintaining the localization accuracy, we adopt a scheme that combines structure-aware multi-task distillation based on DexiNed [19] and HFNet [17]. Multi-task distillation improves efficiency by identifying multiple objectives through sharing information, while structure-aware constraints improve prediction accuracy.

Our localization method is divided into two parts: offline map construction and online visual localization. First, the local and global features of the database images are extracted through SAMLoc, and then the SfM map is constructed

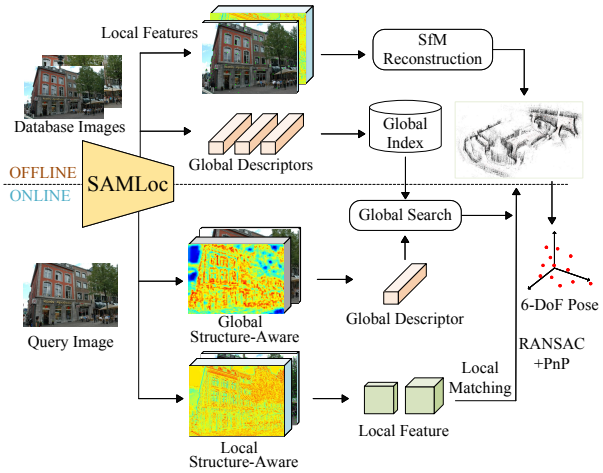


Fig. 3. **Hierarchical localization framework for structure-aware multi-task distillation.** We use structure awareness and multi-task distillation to improve the speed and accuracy of matching based on the hierarchical localization network.

offline by using the local features, and the global descriptors of the database images are built into a global index. Then when we input the query image, the global features and local features are obtained through the structure-aware module. We use the global features to perform rough matching through the use of KNN (K-Nearest Neighbor) and database index matching, and then use the obtained local features and SfM maps to perform 2D-3D matching through RANSAC [27] and PnP methods. Finally, we obtain the 6-DoF pose. Our network framework is shown in Fig.3.

Our encoder uses MobileNetV3-Large [28] as the backbone network, and as the local feature and global feature teacher models we use Superpoint [29] and NetVLAD [11] respectively. Fig.4 depicts our network framework.

Superpoint designed a self-supervised network model to extract pixel-accurate feature points and descriptors simultaneously by using encoding and decoding, as well as a strategy for enhancing the repeated extraction of feature points called Homographic Adaptation, whose generalization is proved excellent. NetVLAD is widely used in the field of image retrieval. It takes a convolutional neural network as the basic feature extraction structure and outputs global descriptors through the NetVLAD layer, which has great advantages in image retrieval. However, it has a large number of parameters and high computational costs; it performs well but is difficult to run in real-time on mobile devices. Therefore, we use the knowledge distillation method to distill the part to MobileNetV3. Although there is a loss in accuracy, the speed can be greatly improved, which can be proved by our experiments in Section IV-A and Section IV-E.

Our local feature extraction branch is at layer 7 of MobileNet, while global feature extraction is at layer 18 of MobileNet, and we also add structure-aware constraints in front of the local and global feature heads respectively. We add local and global structure-aware constraints to the 4th and 11th layers of the network and keep the structure of the backbone network unchanged, which maximizes the speed

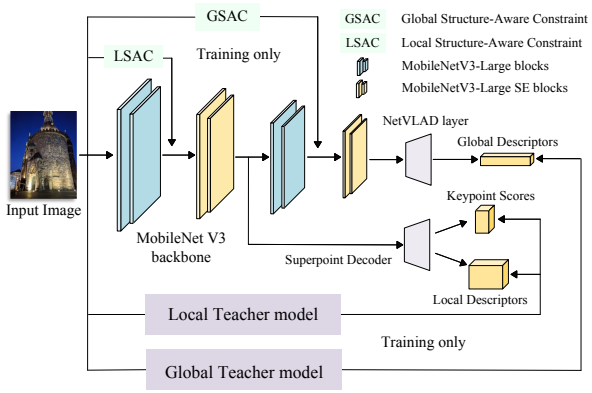


Fig. 4. **Our backbone network** used MobileNetV3. We introduced local and global structure-aware constraints into a hierarchical localization network with multi-task distillation. Our model can better learn features from teacher models under structure-aware constraints.

of feature extraction.

Our loss function for structure-aware and multi-task distillation is divided into local loss and global loss. Eq.(3) is the local loss function, Eq.(4) is the global loss, and Eq.(5) is the total loss function. The parameter e^{-w_i} in front of the formula represents the weight of each loss, d represents the descriptor, s represents the structure-aware constraint, the subscript s represents the student model, t represents the teacher model, the superscript l and g represent local and global respectively, and k represents key points.

$$\begin{aligned} L_{\text{local}} =& e^{w_1}\left\|d_s^l - d_{t_1}^l\right\|_2^2 + \frac{e^{w_2}}{\sqrt{e^{w_1}+e^{w_3}}}\left\|s_s^l - s_{t_1}^l\right\|_2^2 \\ &+ e^{w_3} CrossEntropy\left(k_s, k_{t_3}\right) \end{aligned} \tag{3}$$

The local loss function (Eq.3) consists of three terms, which are local descriptor distillation loss, local structure-aware constraint loss, and keypoint score.

$$L_{\text{global}} = e^{w_4}\left\|d_s^g - d_{t_1}^g\right\|_2^2 + \frac{e^{w_5}}{\sqrt{e^{w_1}+e^{w_3}+e^{w_4}}}\left\|s_s^g - s_{t_1}^g\right\|_2^2 \tag{4}$$

The global loss function (Eq.4) is divided into two categories, the global descriptor distillation loss, and the global structure-aware constraint loss. The proposed total loss function (Eq.5) is as follows, where w_i is the regularization term for each loss.

$$L = L_{\text{local}} + L_{\text{global}} + \sum^{n} w_i \tag{5}$$

This hierarchical localization method with structure awareness and multi-task distillation is very convenient to train since all the knowledge comes from the teacher model. Hence, this method greatly simplifies the training task while ensuring the structural features extracted into the network, thus our model can balance speed and accuracy.

C. Training Process

Our model can be applied to most visual localization datasets. In addition, our training method is simple and can be directly applied to other datasets for testing without any additional training, even in challenging datasets.

The following is an introduction to our training dataset and training methods:

BIPED (Barcelona Images for Perceptual Edge Detection) dataset is a detailed annotated edge dataset proposed by [19]. It contains 250 outdoor images that have been carefully verified and cross-checked by experts for accurate labeling.

MS COCO [30] dataset is a large-scale dataset with rich image resources, including a variety of common object categories and labeled instances. The images contain rich object categories and rich scene features.

Google Landmarks [31] dataset is the largest landmark recognition dataset in the world, containing over one million images of tens of thousands of landmarks. It contains rich daytime urban building information, which helps us to extract robust features from buildings.

Berkeley Deep Drive [32] dataset consists of more than 100K videos with various labels, which contain rich road information and a large amount of night scene information. Such blurred images at night are very effective for improving the localization of night-time scenes.

The network training is divided into two stages: structure-aware module training and multi-task distillation training. The advantage of doing this is that the training process can be performed simultaneously, the network parameters can be optimized quickly, and the debugging time can be greatly shortened. Our experiments are performed using Tensorflow [33].

For the structure-aware module, we perform data augmentation on 200 images in the BIPED dataset and input images are resized to 640 by 480 pixels. The training process uses the Adam optimizer with Batchsize=8, learning rate of 10^{-4}, and convergence of about 130K iterations. Our model is trained from scratch and there is no pre-trained model; training on a GTX 1080Ti took approximately 2 days.

For the multi-task distillation part, we selected the MS COCO dataset of approximately 120K unlabeled images, 110K daytime city building images from the Google Landmarks dataset, and approximately 30K blurred night scene images from the BDD dataset were used for training. All the images were randomly selected from the dataset and resized to 640 by 480 pixels for the network. The local features are trained with grayscale images, and the global features are trained with RGB images. The training process uses random Gaussian noise and random brightness and contrast changes. The training time for our model on a GTX 1080Ti is approximately 1 day. Our initial weight adopts the pre-trained model of Imagenet [34] and we use the RMSProp optimizer [35] with batchsize =16, and the initial learning rate is 10^{-3}, which is successively divided by 10 at the iterations of 70K, 100K, and 120K.

IV. Experience

We evaluate the component modules and the entire network of SAMLoc on the visual localization website [6]. All of our visual localization tests are performed by generalization without any retraining.

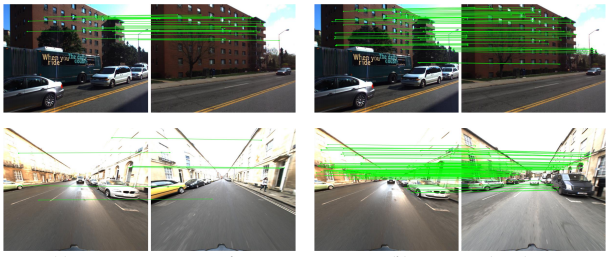
A. Visual localization of season and weather changes

We evaluate our model using the CMU Season dataset [38] and the RobotCar Season dataset [39]. Here, X+Y

TABLE I

THE RECALL [%] AT DIFFERENT DISTANCE AND ORIENTATION THRESHOLDS, HIGHLIGHTING FOR EACH OF THEM THE BEST (RED) AND SECOND-BEST (BLUE) METHODS ON CMU SEASON DATASET AND ROBOTCAR SEASON DATASET.

distance[m] orient.[deg]	CMU Season Dataset				RobotCar Season Dataset			
	urban	suburban	low sun	no foliage	overcast winter	rain	dusk	dawn
	0.25/0.50/5.0 2/5/10	0.25/0.50/5.0 2/5/10	0.25/0.50/5.0 2/5/10	0.25/0.50/5.0 2/5/10	0.25/0.50/5.0 2/5/10	0.25/0.50/5.0 2/5/10	0.25/0.50/5.0 2/5/10	0.50/0.50/5.0 2/5/10
AS[14]	68.9/75.7/83.4	36.2/44.4/56.0	46.8/55.0/66.3	65.6/74.9/84.8	33.1/71.5/93.8	51.3/79.8/96.9	44.7/74.6/**95.9**	36.2/68.9/89.4
CSL[15]	36.7/42.0/53.1	8.6/11.7/21.1	22.6/27.4/38.8	36.5/43.2/57.5	39.5/75.9/92.3	59.9/83.1/**97.6**	56.6/**82.7**/**95.9**	47.2/73.3/**90.1**
NetVLAD[11]	17.4/40.3/93.2	7.6/21.0/80.5	10.1/25.7/77.7	11.8/29.2/82.0	2.8 /25.9/92.6	9.0/ 35.9/96.0	7.4/29.7/92.9	6.2/22.8/82.6
DISAM[8]	22.7/46.4/85.4	11.3/27.2/71.9	16.2/37.7/79.3	15.5/36.3/77.6	5.1/25.4/65.9	8.1/ 29.0/75.1	8.6/28.7/69.3	9.7/29.0/60.7
HFnet[17]	90.3/93.0/96.1	71.6/77.9/86.8	72.3/77.9/85.3	80.6/85.3/90.2	49.7/73.1/90.0	60.6/**85.3**/**97.1**	53.6/81.5/94.2	48.9/**74.5**/89.9
PixLoc(E2E)[10]	78.3/81.8/94.6	–	–	–	52.8/**78.7**/95.1	**63.7**/84.3/96.7	**57.4**/80.5/93.9	**49.9**/72.7/89.9
NV+SP(teacher)	**92.0**/**94.7**/**97.9**	**73.8**/**80.5**/**91.0**	**75.3**/**81.6**/**90.1**	**83.7**/**88.5**/**93.6**	**53.6**/**78.5**/**96.2**	**59.6**/84.8/**97.1**	**54.8**/**83.0**/**96.2**	48.4/73.5/**90.1**
SAMLoc(ours)	**91.2**/**94.0**/**97.3**	**73.4**/**80.7**/**91.1**	**74.0**/**80.8**/**89.0**	**82.6**/**88.2**/**93.7**	**55.4**/78.2/**96.4**	58.7/**85.0**/**97.6**	**54.8**/82.0/95.2	**50.1**/**77.0**/**94.6**



(a) NetVLAD+Superpoint (b) SAMLoc(ours)

Fig. 5. Example of hierarchical localization matching of image pairs from the CMU season dataset (top row) and RobotCar season dataset (bottom row). The left column is NV+SP, the right column is SAMLoc (ours)

denotes hierarchical localization with X(Y) as global(local) descriptors. Active Search (AS) [14] and City Scale Localization (CSL) [15] are both 2D-3D direct matching methods. NetVLAD [11] and DISAM [8] are image retrieval based methods. PixLoc (E2E) [10] is an end-to-end visual localization method, and we do not have a SfM model for the CMU dataset. HFNet [17] and SceneSqueezer [13] are efficient 2D-3D matching methods, the latter only has results using the Aachen Day-Night dataset [36] so it is only compared in Section IV-B. NV+SP is our teacher model, which uses NetVLAD for global search and SuperPoint for local matching. SAMLoc (Ours) achieves a level similar to the teacher model in several aspects, but our feature extraction time is greatly reduced, which will be discussed in Experiment IV-E. Our localization results are shown in Tab. I. We also show the visual localization results of our method and that of the teacher model in Fig.5.

We demonstrate the strong generalization performance of our model. Under the structure-aware constraints, our multi-task distillation result is only approximately 1% loss compared with the teacher model, and our method achieves almost comparable strength to the teacher model with a lightweight model.

B. Visual localization of day-night changes

We evaluated using the Aachen Day-Night dataset [36], which contains 4328 day-time database images of the old

TABLE II

EVALUATE DAY-NIGHT CHANGES ON THE AACHEN DATASET

distance[m] orient.[deg]	Aachen	
	day	night
	0.25/0.50/5.0 2/5/10	0.50/1.0/5.0 2/5/10
AS[14]	57.3 / 83.7 / **96.6**	28.6 / 37.8 / 51.0
CSL[15]	52.3 / 80.0 / 94.3	29.6 / 40.8 / 56.1
NetVLAD[11]	0.0 / 0.2 / 18.9	0.0 / 0.0 / 14.3
HFnet[17]	76.2 / 85.4 / 91.9	**62.2** / 73.5 / 81.6
PixLoc(E2E)[10]	61.7 / 67.6 / 74.8	46.9 / 53.1 / 64.3
SceneSqueezer[13]	75.5 / **89.7** / **96.2**	50.0 / 67.3 / 78.6
NV+SP(teacher)	**79.6** / **87.1** / 93.8	**64.3** / **75.5** / **88.8**
SAMLoc(ours)	**79.0** / 86.8 / 93.0	61.2 / **77.6** / **84.7**

city, and also 824 day-time and 98 night-time query images. Our SfM model is reconstructed using COLMAP [37].

Our experimental results are shown in Tab.II, and DISAM does not provide a pre-trained model for this scenario. As the Aachen dataset contains rich features, most of the algorithms have good results using the day-time images. Although we are limited by the size of the model, we still show competitive results. Using the night-time scenes, the results of most algorithms are not as good as in the day-time due to the drastic change in the scene, however the results for our method are still stable at night. Fig.6 shows our structure-aware heatmap and visualized edge feature map.

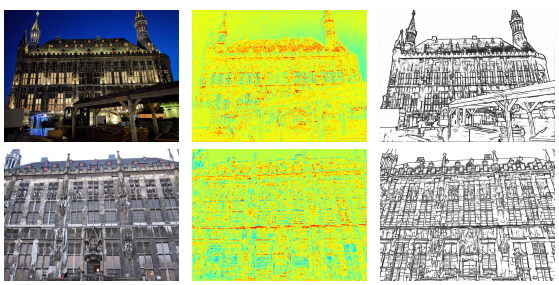


Fig. 6. Example of visualized structure aware from the Aachen Day-Night Dataset. The left column is the query image and the database image, the middle is the structure-aware heatmap, and the right is the visual edge feature map. Our edge features are hardly affected by illumination.

TABLE III

RESULTS OF ABLATION EXPERIMENTS ON THE CMU SEASON DATASET

distance[m] orient.[deg]			CMU Season Dataset		
			urban	suburban	park
-p	-g	-l	0.25/0.50/5.0 2/5/10	0.25/0.50/5.0 2/5/10	0.25/0.50/5.0 2/5/10
×	×	×	88.2/91.0/94.0	68.0/73.9/82.3	46.7/53.9/67.1
×	×	✓	88.8/91.4/94.5	68.6/74.9/83.6	47.8/55.5/69.3
×	✓	×	89.0/91.8/95.0	69.0/75.1/83.5	48.4/56.0/69.4
×	✓	✓	90.4/93.4/96.8	71.7/78.9/90.0	50.3/59.2/75.7
✓	×	×	89.2/92.0/95.2	69.3/76.2/85.8	47.6/55.7/69.9
✓	✓	×	90.7/93.6/96.8	72.3/79.5/89.9	50.8/59.6/76.3
✓	×	✓	90.9/93.8/97.2	72.7/80.1/90.7	51.4/60.2/77.4
✓	✓	✓	**91.2/94.0/97.3**	**73.4/80.7/91.1**	**51.9/61.0/77.8**

C. Ablation Study on Structure-Aware Constraints

Through Tab.III, we evaluate the impact of the structure-aware modules and the pre-trained models on visual localization performance, where $-p$ means use pre-trained model, $-g$ means use of global structure-aware constraints, and $-l$ means use of local structure-aware constraints.

We can see that adding global and local awareness modules and using pre-trained models are optimal. Using the CMU dataset, it can be seen that the effect is the best in the urban scene, and the overall performance is relatively low in the park. This is because the increase in structural features, such as stable houses and roads, increases the effect of structural perception and the localization accuracy. In addition, the pre-training model is also helpful for the localization accuracy, and the network effect using the pre-training model is better.

TABLE IV

QUANTITATIVE EXPERIMENTS ON THE EFFECT OF MULTI-TASK DISTILLATION USING THE AACHEN DATASET

distance[m] orient.[deg]	Aachen	
	day	night
	0.25/0.50/5.0 2/5/10	0.25/0.50/5.0 2/5/10
NV+SP	79.6/87.1/93.8	64.3/75.5/88.8
NV+SAMLoc	78.6/86.2/93.2	64.3/77.6/88.8
SAMLoc+SP	77.9/86.5/92.2	69.4/78.6/85.7
SAMLoc+SAMLoc	79.0/86.8/92.2	61.2/77.6/84.7

D. Quantitative assessment of multi-task distillation

Tab.IV examines the effectiveness of our structure-aware approach for multi-task distillation. We evaluated the impact of different predictors within the hierarchical localization framework with the Aachen dataset. We separately compare the effect of using our teacher model NV+SP and our network SAMLoc, and we demonstrate the effect of our structure-aware with multi-task distillation network replacements between hierarchical localization components.

By comparing the night-time scenes of NV+SP and SAMLoc+SP, we can see that the latter is better than the former, indicating that our global structure-aware distillation method has better resistance to illumination variance. By comparing NV+SAMLoc and NV+SP, it can be seen that our local feature extraction achieves a similar effect to SuperPoint. By comparing the daytime scenes of our SAMLoc is not as effective as NV+SP, and we analyze that the reason may be that the scene itself has rich features, but the size of the network model limits us, so it is slightly lower than the teacher model.

TABLE V

RESULTS OF RUNNING TIME(MS) ON AACHEN DATASET

Aachen Day-time							
distance[m] orient.[deg]	feature extraction		localization				total
	global	local	global	Covis.	local	pnp	
AS	-234-		-	-	-104-		338
PixLoc(E2E)	- 1955 -						1955
HFNet	15		7	4	9	9	**44**
NV+SP	110	50	7	4	9	9	189
SAMLoc	20		7	4	9	9	**49**
Aachen Night-time							
AS	-236-		-	-	-118-		354
PixLoc(E2E)	- 2062 -						2062
HFNet	15		7	4	9	9	**44**
NV+SP	110	50	7	4	9	9	189
SAMLoc(ours)	20		7	4	9	9	**49**

E. Runtime Evaluation

Both our training and evaluation devices use an Intel Core i7-11700 CPU, 32GB RAM, and a GTX1080Ti. We compared the runtime of the 2D-3D and end-to-end approaches. As can be seen from Tab.V, the hierarchical localization method is very efficient, but it is still difficult to run in real time on mobile devices. The pre-trained model of SceneSqueezer is not available, so we do not provide a comparison with it here. CSL takes three orders of magnitude more time than our approach, hence it is not included.

We maximize the improved localization accuracy based on real-time localization through a structure-aware multi-task distillation scheme. Compared with the teacher model, we speed up feature extraction by a factor of 7 with extremely low loss of feature extraction accuracy. Experiments show that both our method and HFNet achieve real-time localization in resource-constrained devices, but our localization accuracy is superior by comparison.

V. CONCLUSION

In this paper, we propose a self-supervised structure-aware constraints with multi-task distillation visual localization method. Our method introduces structural information into the multi-task distillation process to constrain the network to learn robust information in the scene. We explain why and how to introduce structure-aware constraints, and demonstrate through experiments that our method generalizes well even in challenging scenarios. Through this multi-task distillation method, we can even achieve better localization accuracy than the teacher model in some parts while increasing the feature extraction speed by 7 times. In future work, we will further consider reducing feature extraction time as well as compressing map models to achieve more efficient matching.

REFERENCES

[1] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[2] H. Lim, S. N. Sinha, M. F. Cohen, M. Uyttendaele, and H. J. Kim, "Real-time monocular image-based 6-dof localization," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 476–492, 2015.

[3] 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.

[4] H. Hu, Z. Qiao, M. Cheng, Z. Liu, and H. Wang, "Dasgil: Domain adaptation for semantic and geometric-aware image-based localization," *IEEE Transactions on Image Processing*, vol. 30, pp. 1342–1353, 2020.

[5] E. Brachmann and C. Rother, "Visual camera re-localization from rgb and rgb-d images using dsac," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5847–5865, 2021.

[6] 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.

[7] P. Yin, L. Xu, X. Li, C. Yin, Y. Li, R. A. Srivatsan, L. Li, J. Ji, and Y. He, "A multi-domain feature learning method for visual place recognition," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 319–324.

[8] H. Hu, H. Wang, Z. Liu, and W. Chen, "Domain-invariant similarity activation map contrastive learning for retrieval-based long-term visual localization," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 2, pp. 313–328, 2021.

[9] 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.

[10] P.-E. Sarlin, A. Unagar, M. Larsson, H. Germain, C. Toft, V. Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl *et al.*, "Back to the feature: Learning robust camera localization from pixels to pose," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3247–3257.

[11] 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.

[12] H. Germain, G. Bourmaud, and V. Lepetit, "S2dnet: Learning accurate correspondences for sparse-to-dense feature matching," *arXiv preprint arXiv:2004.01673*, 2020.

[13] L. Yang, R. Shrestha, W. Li, S. Liu, G. Zhang, Z. Cui, and P. Tan, "Scenesqueezer: Learning to compress scene for camera relocalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8259–8268.

[14] 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.

[15] 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.

[16] M. Ramamonjisoa, Y. Du, and V. Lepetit, "Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 648–14 657.

[17] 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.

[18] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.

[19] X. S. Poma, E. Riba, and A. Sappa, "Dense extreme inception network: Towards a robust cnn model for edge detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1923–1932.

[20] P. Lei, F. Li, and S. Todorovic, "Boundary flow: A siamese network that predicts boundary motion without training on motion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3282–3290.

[21] 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.

[22] C. Qin, Y. Zhang, Y. Liu, D. Zhu, S. A. Coleman, and D. Kerr, "Structure-aware feature disentanglement with knowledge transfer for appearance-changing place recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[23] A. P. Kelm, V. S. Rao, and U. Zölzer, "Object contour and edge detection with refinecontournet," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2019, pp. 246–258.

[24] H. Hu, H. Wang, Z. Liu, C. Yang, W. Chen, and L. Xie, "Retrieval-based localization based on domain-invariant feature learning under changing environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 3684–3689.

[25] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.

[26] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[27] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[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] 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.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[31] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3456–3465.

[32] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[33] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "{TensorFlow}: a system for {Large-Scale} machine learning," in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, pp. 265–283.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[35] T. Tieleman, G. Hinton *et al.*, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[36] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, "Image retrieval for image-based localization revisited." in *BMVC*, vol. 1, no. 2, 2012, p. 4.

[37] 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.

[38] A. Bansal, H. Badino, and D. Huber, "Understanding how camera configuration and environmental conditions affect appearance-based localization," in *2014 IEEE Intelligent Vehicles Symposium Proceedings*. IEEE, 2014, pp. 800–807.

[39] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.