IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

SACNet: A Scattered Attention-based Network with Feature Compensator for Visual Localization

Ke Wang[1,2,*], Zhiqiang Jiang[1,2,*], Kun Dai[1], Tao Xie[1], Ducheng Jin[1], Ruifeng Li[1], Lijun Zhao[1], Xiao Chen[3]

***Abstract*— Visual localization, an integral component of a vast array of computer applications, has been effectively resolved by scene coordinate regression (SCoRe) methods. However, due to the limited receptive field of convolutional neural networks (CNNs), current SCoRe methods have difficulty in distinguishing comparable image patches in sparse texture scenes, thus impairing localization performance. Recently, Transformer exhibits remarkable capability in modelling long-range dependencies, which provides a remedy to the aforementioned problem. Whereas the Transformer alleviates the deficiencies of CNNs, the quadratically computational cost of Transformer leaves it incapable of handling intensive regression tasks, such as scene coordinates prediction. Towards this end, we introduce SACNet, a sparse attention-based network for efficient and accurate visual localization. We overhaul the core designs of vanilla Transformer and further propose a multiple scattered Transformer (MST) with linear complexity. MST consists of a multiple scattered attention (MSA) layer and a filtered feed-forward network (F-FFN). The MSA layer calculates the attention matrix along the channel dimension and adaptively retains the most profitable attention values for feature consolidation such that the consolidated features can better foster scene coordinate regression. F-FFN utilizes a gate mechanism that suppresses less pertinent features, where multi-scale depth-wise convolutions are further used to promote the information flow. After MST, SACNet develops a feature compensator (FC) that combines local geometry features with global context information to predict element-wise soft attention mask, thus enabling the network to adaptively reconcile the importance of local and global-aware local features. Extensive experimental results demonstrate that SACNet noticeably surpasses the cutting-edge methods on several datasets.**

I. Introduction

Visual localization, which endeavours to estimate the 6-DOF camera pose of a query RGB image, is an essential component in various computer vision systems.

Current learning-based methods can be separated into absolute pose regression (APR) [1], [2], scene coordinate regression (SCoRe) [3]–[8], and relative pose regression (RPR) [9]. Recently, SCoRe methods, which use convolutional neural networks (CNNs) to regress scene coordinates and a PnP algorithm to recover 6-DOF pose, have exhibited extraordinary localization performance in small static scenes. However, CNN-based SCoRe methods [3], [4], [10] possess an inherent limitation in differentiating similar image patches. More concretely, owing to the limited receptive field of CNNs, the existing SCoRe methods [11], [12] extract identical feature representations for similar pixels when confronted with extensive sparse texture regions in real scenes. After that, such similar features are employed to predict diverse scene coordinates, thus leading to inferior localization performance. Recently, vision Transformer has garnered great attention in computer vision tasks owing to its ability in integrating global information. With the global receptive fields, Transformer can effectively eliminate the ambiguity induced by visual similarity between image patches. For example, as shown in Fig. 1, the similar image patches can be distinguished according to their relative positional information to the corners. However, two barriers exclude Transformer from being effectively integrated into SCoRe methods.

Fig. 1: Points A and B with similar appearances can be distinguished according to their position to the corners (marked in red), indicating the significance of global receptive fields.

(i) Vanilla Transformer calculates the dot product of each query vector with all key vectors to determine their similarities, which are used to retrieve message from value vectors, hence realizing message exchange. However, the computational cost of vanilla Transformer increases quadratically with respect to the number of image patches, rendering it incapable of handling intensive scene coordinate regression tasks. On the other hand, recently, several attention mechanism have proved that uninformative features should be restrained to emphasize crucial features. Nevertheless, vanilla Transformer computes all attention correlations according to all query-key pairings. Smaller weights signify lower correlations between query-key pairings, which make the network capture uninformative message from the value vectors, resulting in redundant message exchange and disturbing the feature interactions process.

(ii) The inductive bias of the CNNs has been demonstrated to be extremely critical for the network to model the scene information [13]. Thus, how to blend the local features

Corresponding author: Tao Xie, Lijun Zhao, and Ruifeng Li. *: These authors contributed equally to this work.

This letter was recommended for publication by Editor Cesar Cadena Lerma upon evaluation of the reviewers' comments. This work was in part by National Natural Science Foundation of China under Grant 62176072, and was supported in part by the open research fund of anhui province key laboratory of machine vision inspection (KLMVI-2023-HIT-12).

[1]State Key Laboratory of Robotics and System, Harbin Institute of Technology, Harbin 150006, China

[2]Zhengzhou Research Institute, Harbin Institute of Technology, Zhengzhou 450000, China.

[3]Wuhu HIT Robot Industry Technology Research Institute CO.,LTD, Wuhu 241000, China.

generated by CNNs and the global features manufactured by Transformer remains further investigation.

To tackle these issues, we introduce SACNet, a novel Transformer-based SCoRe method that effectively mitigates the visual similarity with manageable computational cost, and models the correlation between local and global-aware local features in an adaptive manner. For the first issue about high computational cost and redundant message exchange in vanilla Transformer, we develop a multiple scattered Transformer (MST) that is comprised of multiple scattered attention (MSA) layer and a filtered feed-forward network (F-FFN). More specifically, inspired by [14], MSA calculates the attention matrix across the feature dimension, i.e., computing cross-covariance over channels to establish an attention map that encodes the global information implicitly. In this way, MSA possesses linear complexity in terms of image patches numbers, thus facilitating an effective analysis of images with high resolutions. Besides, different sparse attention matrices are constructed to integrate diverse global context information, which are added together to ensure that the network can flexibly alleviate the influence of the redundant information. After MSA, F-FFN utilizes a gate mechanism to suppress less pertinent features, where multi-scale depth-wise convolutions are further used to promote the information flow. For the second issue in terms of the adaptively blend of local features and global-aware local features, we develop a feature compensator (FC) that first integrates the local geometry features into global context information to generate discriminative feature representation for scene coordinate regression. After that, FC predicts element-wise soft attention masks, which adaptively identify the relative importance of local features and global-aware local features. The key insight lies in that local geometrical information of pixels in a texture-less region are somewhat ambiguous owing to the identical visual similarity, whereas global-aware local features that model the long-range dependencies are favorable. In contrast, for pixels located at texture-rich areas, the local geometrical context is sufficiently discriminative to regress scene coordinates. We undertake extensive experiments on 7-Scenes [10], 12-Scenes [15], LIVI [6], and Cambridge [1] datasets, which demonstrates the superiority of our method to realize accurate visual localization.

The main contributions of this work can be summarized as:

- We propose SACNet, a novel SCoRe method that leverages Transformer to effectively distinguish similar image patches, thus elevating the localization performance.
- We develop a multiple scattered Transformer including a multiple scattered attention (MSA) layer and a filtered feed-forward (F-FFN) network. MSA layer calculates the attention matrix along channel dimensions to reduce computational cost and builds multiple sparse attention matrices to realize effective message propagation. F-FFN utilizes a gate mechanism to filter less pertinent features, where multi-scale depth-wise convolutions are further used to promote the information flow.
- We introduce a feature compensator (FC) that adaptively fuses local features and global-aware local features to enhance the feature representation and model capability of the network.
- Comprehensive experiments reveal that SACNet achieves impressive performances on several benchmarks with comparable computational cost.

II. Related Work

A. Structure-based Visual Localization

Structure-based approaches [16], [17] aim to determine the correspondences between 2D pixel locations and 3D scene coordinates when given an image query, typically accompanied by a PnP algorithm to estimate the 6-DOF pose. Recently, advanced methods in this field have introduced image retrieval to reduce the searching space. After that, state-of-the-art feature matching methods, such as SuperGlue [18], are applied to establish reliable 2D-2D matches, thereby generating accurate 2D-3D correspondences. The incorporation of efficient image retrieval and feature matching techniques noticeably elevates the applicability and robustness of structure-based methods in large-scale scenes. Experiments demonstrate that the structure-based approaches achieve excellent localization performance in large-scale scenes and exhibit impressive generalization to unknown scenes. However, they consume massive memory space to store pre-built 3D scene models and corresponding descriptors, making them expensive to deploy to resource-constrained devices.

B. Learning-based Visual Localization

The learning-based visual localization methods can be typically divided into relative pose regression (RPR) [9], absolute pose regression (APR) [1], [2], and scene coordinate regression (SCoRe) methods [3]–[8]. The RPR methods learn the similarity between images in query and in database to determine the 3D poses of the camera, which takes up a significant amount of time. The APR methods provide a straightforward but effective end-to-end pipeline to predict camera pose directly, which, however, realizes inferior localization precison. The SCoRe methods directly regress the 2D-3D correspondence of pixels in image, followed by a PnP algorithm to calculate the 3D camera pose. Mainstream SCoRe works [3], [4], [6], [8], [19], [20] use CNNs to identify more accurate coordinate regression. HSCNet [4] designs a hierarchical network architecture to regress scene coordinates. EAAINet [5] concurrently extracts the absolute features and relative features of images to alleviate the ambiguous caused by similar image patches. Though model parameters are tied to a specific scene, the SCoRe pipeline achieves the best performance in small and medium scenes without storing databases like structure-based methods [20]. However, it also inevitably receives the inherent shortcomings of CNN, that is local receptive field, hindering long range information convergence. Hence, exploring the efficient global information aggregation in scene coordinate regression framework is meaningful.

III. Methodology

A. Overall

As shown in Fig. 2, SACNet initially utilizes a CNN-based backbone to generate local feature representations

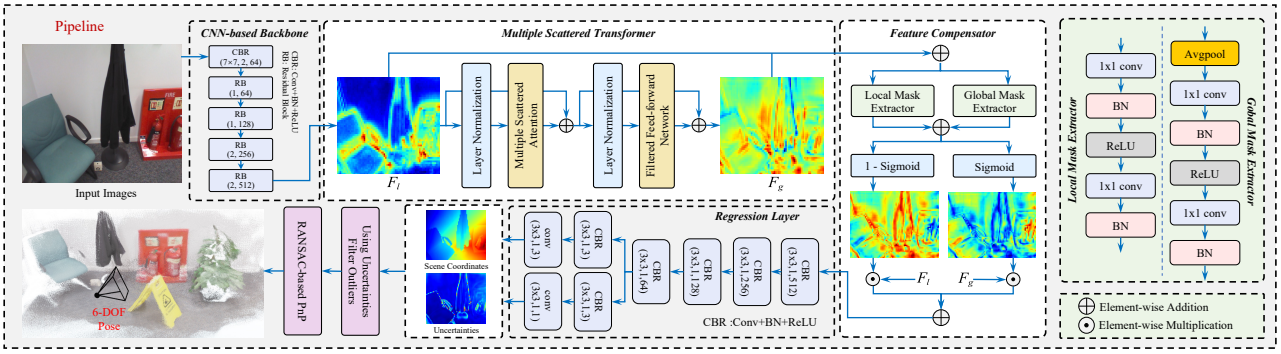

Fig. 2: Overall pipeline of SACNet. A CNN-based backbone and a multiple scattered Transformer is utilized to extract local features and global-aware local features. A a regression layer is leveraged to regress the 3D scene coordinates, followed by a RANSAC-based PnP algorithm to recover the 6-DOF camera pose.

during the feature extraction stage. Then, SACNet proposes a multiple scattered Transformer (MST) that calculates the multiple sparse attention matrices so as to implicitly encodes the global information and mitigate redundant information exchange. Subsequently, MST develops a filtered feed-forward network (F-FFN), which utilizes two parallel multi-scale convolution paths along with GELU activation function as gate mechanism to further filter relevant features. After MST, a feature compensator (FC) is introduced to predict element-wise soft attention mask to adaptively merge local and global-aware local features and establish their correlation, thereby boosting the feature representation and model capability of the network. Finally, SACNet leverages a regression layer to predict dense 3D scene coordinates and 1D uncertainties, followed by a PnP algorithm to regress poses.

B. CNN-based Backbone

We adopt ResNet18 [21] with small tweaks as the backbone to obtain local feature representation. Specifically, we discard the max pooling layer, average pooling layer and fully connected layer of ResNet18. Besides, in accordance with previous works [5], [19], we set the strides of the four Resblocks to $\{1,1,2,2\}$ so that enabling the final output feature map F_l to possess dimension with $\mathbb{R}^{C\times H/8\times W/8}$.

C. Multiple Scattered Transformer (MST)

Recently, substantial studies [4], [5], [19] reveal that the CNN-based SCoRe methods have difficulty in distinguishing similar image patches in sparse texture areas, thus resulting in inferior localization performance. Transformer has exhibited impressive capability in integrating global context, which supplies a remedy to this issue. With the global receptive fields, Transformer can competently eradicate ambiguity caused by visual similarity between image patches. For instance, comparable image patches can be distinguished based on their translational information in relation to corners or edges. However, the quadratic increase in computational cost with the number of image patches and the excessive interaction of redundant information impede the practical applicability of Transformers to dense regression tasks. Towards this end, we overhaul the core designs of vanilla Transformer and propose multiple scattered Transformer (MST) including a multiple scattered attention (MSA) layer and a filtered feed-forward network (F-FFN), which exploits to noticeably reduce the computational cost and alleviate the issue of redundant information passing.

Multiple Scattered Attention (MSA) Layer. The proposed MSA commences by performing a point-wise convolution to expand the dimension of the input feature map F_l, followed by three separate depth-wise convolutions with different kernel sizes to extract multi-scale local features, which are then integrated together to generate intermediate features. After that, MSA leverages point-wise convolution to squeeze the channel dimension, with the resulted features divided into three parts to acquire the query, key, and value vectors $Q,K,V \in \mathbb{R}^{N\times C}$, where $N = H/8\times W/8$, which can be formulated as:

$$\begin{aligned} F_m &= C_{p1}(F_l), \\ [\tilde{Q},\tilde{K},\tilde{V}] &= C_{p2}([D_3(F_m)||D_5(F_m)||D_7(F_m)]), \\ Q,K,V &= I2S(\tilde{Q}), I2S(\tilde{K}), I2S(\tilde{V}), \end{aligned} \tag{1}$$

where $C_{p1}(\cdot)$ and $C_{p2}(\cdot)$ mean point-wise convolution, $D_3(\cdot)$, $D_5(\cdot)$, and $D_7(\cdot)$ represent depth-wise convolution with kernel sizes of 3, 5, 7, and $I2S(\cdot)$ means converting images to sequences.

Sparse Attention Matrix. As illustrated in Fig. 3, MSA normalizes the query vector Q and key vector K to stabilize the training process, which can be formulated as:

$$\bar{Q} = Q/||Q||_2^N, \quad \bar{K} = K/||K||_2^N, \tag{2}$$

where $||\cdot||_2^N$ means calculating the l_2 norm along the token dimension.

Inspired by [14], we calculate the attention matrix $M \in \mathbb{R}^{C\times C}$ along feature dimensions, altering the computation from $\mathcal{O}(N^2C)$ to $\mathcal{O}(C^2N)$, where the global context can be implicitly encoded by the computed attention matrix [14]. In this way, the computational cost would be significantly decreased given that $C \ll N$ and the attention matrix can be formulated as:

$$M = \bar{K}^T\bar{Q}, \quad M \in \mathbb{R}^{C\times C}. \tag{3}$$

Although using the attention matrix to retrieve information from value vectors can provide valuable context and facilitate information integration, the propagation of redundant information is introduced and can be detrimental to the localization performance. To alleviate redundant message passing, MSA constructs three sparse attention matrices to

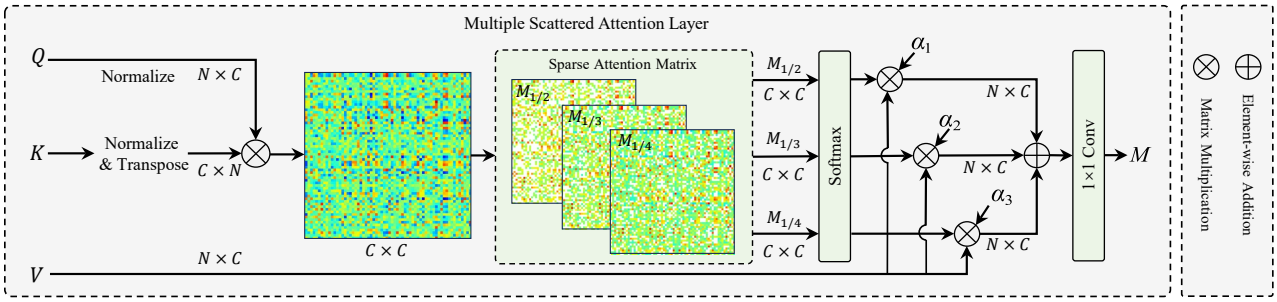

Fig. 3: Structure of multiple scattered attention layer. The attention matrix is calculated over the channel dimension to decrease the computational cost. Multiple scattered attention matrices are also calculated to alleviate redundant message passing.

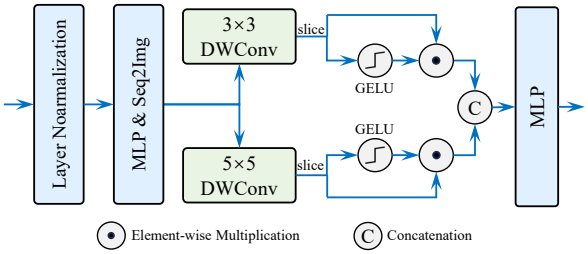

Fig. 4: Structure of Filtered Feed-forward Network. GELU activation function is utilized as gate mechanism to further filter less pertinent features.

selectively attend to different regions of the input, while minimizing the interference of the redundant information. Specifically, we discard $1/2$, $1/3$, and $1/4$ values that have the lowest values across each row of the M matrix, indicating that we replace their probabilities with zero. Then, we derive the sparse attention matrices $M_{1/2}$, $M_{1/3}$, and $M_{1/4}$ and utilize them to retrieve message from the value vectors V, which is formulated by:

$$\begin{aligned} ME_{1/2} &= V \cdot Softmax(M_{1/2}/\sqrt{C}), \\ ME_{1/3} &= V \cdot Softmax(M_{1/3}/\sqrt{C}), \\ ME_{1/4} &= V \cdot Softmax(M_{1/4}/\sqrt{C}). \end{aligned} \tag{4}$$

In light of the fact that different images contain various levels of redundant information, it is reasonable to reconcile different retrieved message. Thus, we implement three learnable weighting factors to adaptively integrate the retrieved messages, which can be expressed as:

$$M = C_{p3}(\alpha_1 ME_{1/2} + \alpha_2 ME_{1/3} + \alpha_3 ME_{1/4}), \tag{5}$$

where α_1, α_2, and α_3 represent the learnable weighting factors, and $C_{p3}(\cdot)$ denotes the point-wise convolution. Thus, our proposed MSA layer offers more flexible and adaptive filtering for the retrieved messages.

Filtered Feed-forward Network (F-FFN). As illustrated in Fig. 4, given the retrieved message M, our proposed F-FFN implements a gate mechanism to filter out less relevant features, wherein the gate mechanism is structured as element-wise product of two concurrent pathways of transformation layers, a particular one of which is triggered with the GELU activation function as the weighting factor to emphasize the crucial information. We also construct two depth-wise convolutions with kernel sizes of 3 and 5 to further promote the local feature representation. In this way, this process can be formulated as:

$$\begin{aligned} M_m &= S2I(MLP(LN(M))), \\ [M^1_{m1}, M^1_{m2}] &= D_3(M_m), \quad M^1 = \Phi(M^1_{m1}) \odot M^1_{m2}, \\ [M^2_{m1}, M^2_{m2}] &= D_5(M_m), \quad M^2 = \Phi(M^2_{m1}) \odot M^2_{m2}, \end{aligned} \tag{6}$$

where $\Phi(\cdot)$ means GELU nonlinear activation function and $\odot$ represents element-wise product.

Ultimately, we concatenate M^1 and M^2 along the channel dimension, utilize the linear transformation layer to alter feature dimension, and convert the sequences to discriminative global-aware local features $F_g \in \mathbb{R}^{C\times H/8\times W/8}$ for the visual localization task.

D. Feature Compensator (FC)

It has been revealed that the inductive bias of CNNs is critical for the network to model scene information [13]. In contrast, the Transformer possesses difficulty in acquiring local geometry features. How to fuse the local features and global-aware local features remains further investigation. To tackle this issue, we propose feature compensator (FC) that adaptively amalgamates local and global-aware local features to augment the overall modeling capabilities of the network. The key insight stems from the fact that local geometrical of pixels in texture-less regions are somewhat ambiguous owing to their visual similarity, whereas global-aware local features that model the long-range dependencies are preferable. In contrast, for pixels located in texture-rich areas, the local geometrical context is distinct enough to enable scene coordinate regression.

As shown in Fig. 2, FC first performs element-wise addition to fuse local features F_l and global-aware local features F_g, thereby effectively capitalising on the previously provided context. Following that, FC designs two separate branches to generate element-wise soft attention masks from local and global perspective, respectively. Specifically, for the local mask extractor, FC first leverages point-wise convolution to squeeze channel dimension, followed by RELU activation function to improve the non-linearity of the network. Subsequently, FC leverages another point-wise convolution to expand the channel dimension, generating $F_{ln} \in \mathbb{R}^{C\times H/8\times W/8}$. For the global mask extractor, FC utilizes a global average pooling layer, two point-wise convolutions, and a RELU activation function to extract $F_{gn} \in \mathbb{R}^{C\times 1\times 1}$. After that, we

TABLE I: Localization experiment on **7-Scenes dataset**. The median translational error (m), rotational error (°), and 5cm-5° accuracy (%) are reported.

7-Scenes	Chess		Fire		Heads		Office		Pumpkin		Redkitchen		Stairs		Average	
—	**Med. Err.**	**Acc.**	**Med. Err.**	**Acc.**	**Med. Err.**	**Acc.**	**Med. Err.**	**Acc.**	**Med. Err.**	**Acc.**	**Med. Err.**	**Acc.**	**Med. Err.**	**Acc.**	**Med. Err.**	**Acc.**
HLoc [16]	0.020, 0.85	—	0.020, 0.94	—	0.010, 0.75	—	0.030, 0.92	—	0.050, 1.30	—	0.040, 1.40	—	0.050, 1.47	—	0.031, 1.09	73.10
MS-Transformer [2]	0.110, 4.66	—	0.240, 9.60	—	0.140, 12.19	—	0.170, 5.66	—	0.180, 4.44	—	0.170, 5.94	—	0.260, 8.45	—	0.181, 7.28	—
SCoordNet [3]	0.019, 0.63	—	0.023, 0.91	—	0.018, 1.26	—	0.026, 0.73	—	0.039, 1.09	—	0.039, 1.18	—	0.037, 1.06	—	0.029, 0.98	—
DSAC* [8]	0.020, 1.10	—	0.020, 1.24	—	0.010, 1.82	—	0.030, 1.15	—	0.040, 1.34	—	0.040, 1.68	—	0.030, 1.16	—	0.027, 1.36	85.20
VSNet [20]	**0.015, 0.50**	—	0.019, 0.80	—	0.012, 0.70	—	0.021, **0.60**	—	0.037, 1.00	—	0.036, 1.10	—	0.028, 0.80	—	0.024, 0.79	—
FDANet [19]	0.018, 0.64	95.70	0.018, 0.73	96.10	0.013, 1.07	99.20	0.026, 0.75	88.08	0.036, 0.91	65.65	0.034, 1.03	78.32	0.041, 1.14	62.80	0.026, 0.89	83.69
HSCNet [4]	0.020, 0.70	97.50	0.020, 0.90	96.70	0.010, 0.90	**100.00**	0.030, 0.80	86.50	0.040, 1.00	59.90	0.040, 1.20	65.50	0.030, 0.80	87.50	0.027, 0.90	84.80
HSCNet++ [22]	0.020, 0.63	—	0.020, 0.79	—	0.010, 0.80	—	**0.020**, 0.65	—	0.030, 0.85	—	0.030, 1.09	—	0.030, 0.83	—	0.023, 0.81	88.70
EAAINet [5]	0.017, 0.58	**97.60**	0.019, 0.70	97.95	0.013, 0.93	**100.00**	0.023, 0.65	**90.40**	0.030, 0.92	68.35	0.034, 1.00	80.28	0.039, 1.16	75.40	0.025, 0.84	87.14
OFVL-MS50 [6]	**0.015**, **0.50**	97.10	**0.015**, **0.59**	**99.40**	**0.008, 0.56**	**100.00**	0.023, 0.63	89.53	0.030, 0.86	68.80	0.031, 0.99	**81.48**	0.026, 0.76	84.70	0.021, 0.69	88.72
ACE [7]	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	80.80
SACNet (ours)	**0.015**, 0.53	97.10	0.019, 0.71	92.90	0.009, 0.59	99.80	0.026, 0.65	89.25	**0.023, 0.65**	**81.95**	**0.029, 0.93**	80.46	**0.017, 0.40**	**87.00**	**0.020, 0.64**	**89.78**

add F_{ln} and F_{gn} together and resort to a sigmoid activation function to acquire element-wise soft attention mask $W_g \in \mathbb{R}^{C\times H/8\times W/8}$ for F_g. To endow the network with the capability to reconcile global-aware local features F_g and local features F_l, we define the soft attention mask for F_l as $W_l = 1 - W_g$. Ultimately, we obtain the discriminative features $F_f \in \mathbb{R}^{C\times H/8\times W/8}$ as:

$$F_f = W_l \odot F_l + W_g \odot F_g. \tag{7}$$

The attention masks enables the network to selectively highlight pertinent features and suppress unuseful message in both the local and global contexts, as shown in **??**.

E. Pose Estimation

As shown in Fig. 2, a regression layer is employed to concurrently predict 3D scene coordinates $C = \{(x_i, y_i, z_i)|i = 1,2,...,N\}$ along with 1D uncertainties $U = \{u_i|i = 1,2,...,N\}$. The regression layer generates N groups of 2D pixels to 3D scene coordinates correspondences. Since uncertainties measure the predicted errors, we discard the predicted scenes coordinates with uncertainties larger than β. Ultimately, we utilize a PnP algorithm to regress 6-DOF camera pose T.

F. Loss

In accordance with the approach outlined in [3], [5], [19], we maximize the distribution of the predicted scene coordinates and formulate the loss function as:

$$L = \frac{1}{N}\sum_{i=1}^{N}(3log u_i + \frac{||c_i - \hat{c}_i||_2^2}{2u_i^2}), \tag{8}$$

where $c_i = (x_i, y_i, z_i)$ indicates the ground-truth scene coordinates of the i-th pixel, $\hat{c}_i = (\hat{x}_i, \hat{y}_i, \hat{z}_i)$ denotes the predicted scene coordinates of the i-th pixel, and u_i means the predicted uncertainty for the i-th pixel.

IV. Experiments

A. Datasets and Experimental Settings

Dataset. We conduct the visual localization experiments under 7-Scenes [10], 12-Scenes [15], LIVI [6], and Cambridge [1] datasets. The 7-Scenes dataset contains seven indoor environments, each represented by 500-1000 image video sequences. All scenes are acquired using a Kinect RGB-D camera with a resolution of 640×480. The 12-Scenes dataset comprises RGB-D images for 12 rooms in 4 large scenes. The RGB-D images are obtained using the depth sensor and iPad color camera. The LIVI dataset is capture by a robot armed with a RealSense D455 camera and a VLP-16 laser radar in four indoor scenes. Ground truth poses are generated by a LiDAR-based SLAM system A-LOAM. Substantial sparse texture, severe motion blur and lighting interference in these scenes make visual localization in LIVI dataset difficult. The Cambridge dataset is a large-scale outdoor localization dataset. It contains RGB images and SfM reconstructions from six outdoor scenes.

Implemantation Details. The output dimensions of four residual blocks in backbone are set to $64, 128, 256, 512$, respectively. The threshold β is set to 0.2 for filtering outliers.

Training Strategy. For 7-Scenes, 12-Scenes, LIVL, and Cambridge datasets, it takes about 3, 3, 6, and 9 hours for 40K, 40K, 80K, and 80K iterations to train SACNet on 8 Tesla V100 GPUs. We use an Adamw solver and with an initial learning rate of 0.0016. The batch size is set to 4 for each GPU. A cosine annealing schedule is selected to adjust learning rate. For 7-Scenes, LIVL, and Cambridge datasets, we adopt the same data augmentation as HSCNet [4] to avoid overfitting. We discard data augmentation for 12-Scenes since the training trajectory and testing trajectory are very close.

Evaluation Metrics. The evaluation metrics include: (i) the median translational and rotational errors of predicted poses; (ii) the 5cm-5° accuracy: the percentage of images with translational and rotational errors less than 5cm and 5°.

B. Comparison with the State-of-the-art Methods

Experiment on 7-Scenes dataset. As illustrated in Table I, we compare SACNet with typical structure-based approach (HLoc), absolute pose regression approach (MS-Transformer), and scene coordinate regression approaches (SCoordNet, DSAC*, VSNet, FDANet, HSCNet, HSC-Net++, EAAINet, OFVL-MS50, and ACE). Overall, SACNet surpasses all state-of-the-art methods by a significant margin under all metrics. Notably, different from scene-specific absolute pose regression and scene coordinate regression methods, the scene-agnostic structure-based approaches (such as HLoc) possess excellent generalization, which means that they can be directly deployed in unknown scenes without retraining. Compared to HLoc, SACNet delivers exceptional localization performance, revealing that the superiority of scene coordinate regression method for visual localization in

TABLE II: Localization experiment on **12-Scenes dataset**. The median translational error (m), rotational error (°), and 5cm-5° accuracy (%) are reported.

Methods	HSCNet [4]	DSAC* [8]	EAAINet [5]	ACE [7]	SACNet (ours)
Med. Err.	**0.011**, 0.50	—	0.014, 0.39	—	0.015, **0.34**
Acc.	99.3	99.1	99.5	**99.6**	99.1

small-scale static scenes. In addition, SACNet exceeds the cutting-edge scene coordinate regression methods DSAC*, FDANet, HSCNet, HSCNet++, EAAINet, OFVL-MS50, and ACE by 4.58%, 6.09%, 4.98%, 1.08%, 2.64%, 1.06%, and 8.98%, exhibiting the superior localization performance of SACNet.

Experiment on 12-Scenes dataset. As shown in Table II, since the training trajectory and testing trajectory are close, all methods achieve impressive localization accuracy. SACNet realizes minimal rotational errors, further demonstrating the superiority of SACNet.

Experiment on LIVL dataset. To fully verify the localization performance of SACNet, we further perform experiments on LIVL dataset. As reported in Table III, SACNet attains 5cm-5 accuracy of 43.41%, exceeding the cutting-edge methods significantly. Specifically, SACNet outperforms the CNN-based techniques SCoordNet, FDANet, EAAINet, and OFVL-MS50 with the improvement of 27.84%, 25.23%, 16.84%, and 9.62%. Notably, Floor and Parking-Lot1 scenes contain large areas of grounds and walls with substantial sparse textures, which poses an inevitable challenge for achieving accurate visual localization. In these specific scenes, SACNet noticeably surpasses all state-of-the-art methods in terms of all evaluation metrics, which demonstrates the effectiveness of leveraging global receptive fields to distinguish similar image patches, hence elevating localization accuracy.

Experiment on Cambridge dataset. In addition to indoor scenes, we conduct an outdoor localization experiment on a Cambridge dataset. As shown in Table IV, HLoc achieves exceptional performance in almost all outdoor scene, while as a structure-based method, HLoc consumes significant memory space, as detailed in Section II-A. Since the SfM reconstruction of Great Court scene contains numerous outliers, the localization errors of SACNet in this specific scene is high. Overall, SACNet achieves higher localization precision than DSAC*, OFVL-MS34 and ACE while exhibiting comparable performance to other SCoRe techniques.

C. Experiment Analysis

We conduct all ablation experiments on 7-Scenes dataset.

Experiment on images with sparse or abundant textures. To showcase the efficacy of global receptive fields used in SACNet in distinguishing similar image patches, we first choose the pumpkin scene from 7-Scenes dataset and the parking-lot1 scene from LIVL dataset since these two scenes contain numerous images with sparse textures. Then, we divide all images in each scene into two distinct categories, i.e., texture-rich images and texture-less images. After that, we adopt the SACNet and a CNN-based technique EAAINet to execute visual localization on these two sequences. Technically, we calculate the gray-level co-occurrence matrix (GLCM) for each image to extract the contrast and entropy, whose sum is defined as a texture score. According to texture scores, we categorize all images into texture-rich or texture-less images. As depicted in Table V, SACNet exhibits superior performance than EAAINet in texture-less scenes, as opposed to texture-rich scenes. Specifically, SACNet surpasses EAAINet by 18.5% when processing texture-less images from the pumpkin scene, surpassing the 8.7% enhancement for texture-rich images. For images from the parking-lot1 scene, SACNet exhibits a performance improvement of 23.29% for texture-less images, outstripping the 15.53% increment for texture-rich images. Experimental results demonstrate that SACNet with global receptive fields displays more powerful capability to discern similar image patches.

Ablation study for proposed modules. We go ahead with our baseline that directly utilizes the modified ResNet18cd as feature extractor, leverages the regression layer to regress 3D coordinates, and finally employs the PnP to recover camera pose, with results reported in Table VI. The baseline obtains 0.029/0.99 median pose error and 80.13% accuracy. Then, we integrate the proposed multiple scattered attention layer into the baseline and utilize the output feature of the layer to regress scene coordinates. We can see that the median pose errors are decreased to 0.025/0.81 and the accuracy is enhanced to 84.76%, revealing the significance of global context information in distinguishing similar image patches. After that, we further integrate the proposed filtered feed-forward network into it, further decreasing the median pose errors to 0.023/0.74 and boost the accuracy to 87.35%. Finally, when incorporating the proposed feature compensator into the baseline, the median pose error is deceased to 0.020/0.64 and the accuracy is improved to 89.78%.

Ablation study for feature compensator (FC). To assess the efficacy of the proposed feature compensator (FC), as shown in Table VII, we perform experiments based on different model variants. EXP1: Removing FC leads to 2.43% performance degradation. EXP2, EXP3: Applying the local / global branch causes 1.51% / 1.06% increase, respectively. Besides, integrating both local and global branches enables more adaptive fusion of local features and global-aware local features, resulting in 89.78% localization accuracy. EXP4: We substitute FC with two 1×1 convolutional layers with the same parameters as FC to prove that the elaborated structure of FC, rather than the increase of model parameters, is the key reason for the improved localization accuracy. Replacing FC with convolutional layers noticeably reduces the localization precision by 1.6%, proving the rationality of the elaborated architecture of FC.

Ablation study for multiple scattered attention layer (MSA). As depicted in Table VIII, we conduct various ablation studies to verify the superiority of the proposed MSA. EXP1: We compare the proposed MSA with the vanilla attention. Experimental results show that the MSA achieves better localization accuracy while requiring significantly less inference speed and training memory. EXP2: We contrast

TABLE III: Localization experiment on **LIVL**. The median translational error (m), rotational error (°), and 5cm-5° accuracy (%) are reported.

LIVL	SCoordNet [3]	FDANet [8]	EAAINet [5]	OFVL-MS34 [6]	OFVL-MS50 [6]	SACNet (ours)
Metrics	Med. Err. Acc	Med. Err. Acc	Med. Err. Acc	Med. Err. Acc	Med. Err. Acc	Med. Err. Acc
K-544	0.171, 2.12 9.86	0.143, 1.89 12.64	0.109, 2.45 25.63	0.071, 1.01 42.53	**0.050**, **0.81 49.91**	0.101, 1.04 41.19
Floor	0.208, 1.94 20.31	0.167, 1.59 23.87	0.121, 1.18 26.51	0.147, 1.48 25.54	0.148, 1.37 30.72	**0.056**, **0.90 46.14**
Parking-Lot1	0.353, 2.97 11.28	0.291, 2.75 13.31	0.228, 1.45 29.31	0.278, 2.08 25.72	0.265, 2.48 26.17	**0.059**, **0.59 48.71**
Parking-Lot2	0.184, 2.13 20.82	0.138, 1.61 22.89	0.107, 1.23 24.83	0.095, 1.17 29.03	0.107, 1.02 28.34	**0.088**, **0.75 37.60**
Average	0.229, 2.29 15.57	0.185, 1.96 18.18	0.141, 1.58 26.57	0.147, 1.43 30.71	0.142, 1.42 33.79	**0.076**, **0.82 43.41**

TABLE IV: Localization experiment on **Cambridge**. The median translational error (m) and rotational error (°) are reported.

Cambridge	HLoc [16]	DSAC++ [11]	DSAC* [8]	VS-Net [20]	HSCNet [4]	HSCNet++ [22]	OFVL-MS34 [6]	ACE [7]	SACNet (ours)
Metrics	Med. Err	Med. Err	Med. Err	Med. Err	Med. Err	Med. Err	Med. Err	Med. Err	Med. Err
Great Court	**0.16**, 0.11	0.40, 0.20	0.49, 0.30	0.22, **0.10**	0.28, 0.20	0.39, 0.23	0.46, 0.31	0.43, 0.20	0.47, 0.32
K.College	**0.12, 0.20**	0.18, 0.30	0.15, 0.30	0.16, **0.20**	0.18, 0.30	0.19, 0.34	0.28, 0.53	0.28, 0.40	0.17, 0.30
Old Hospital	**0.15**, **0.20**	0.30, 0.29	0.21, 0.40	0.16, 0.30	0.19, 0.30	0.20, 0.31	0.25, 0.49	0.31, 0.60	0.18, 0.30
Shop Facade	**0.04, 0.20**	0.06, 0.30	0.05, 0.30	0.06, 0.30	0.06, 0.24	0.06, 0.24	0.16, 0.56	0.05, 0.30	0.06, 0.30
St M.Church	**0.07, 0.21**	0.13, 0.40	0.13, 0.40	0.08, 0.30	0.09, 0.30	0.09, 0.27	0.24, 0.61	0.18, 0.60	0.12, 0.34
Average	**0.11, 0.20**	0.19, 0.30	0.21, 0.34	0.14, 0.24	0.16, 0.28	0.19, 0.28	0.28, 0.50	0.25, 0.40	0.20, 0.31

TABLE V: Performance comparisons on texture-less images and texture-rich images.

Scenes	Image Type	EAAINet [5]	SACNet
7-Scenes Pumpkin	texture-rich images	0.026, 0.78, 80.60	**0.019, 0.49, 89.30**
	texture-less images	0.042, 1.21, 56.10	**0.028, 0.86, 74.60**
	all images	0.030, 0.92, 68.35	**0.023, 0.65, 81.95**
LIVL Parking-lot1	texture-rich images	0.093, 0.86, 45.38	**0.021, 0.27, 60.91**
	texture-less images	0.527, 4.43, 13.26	**0.285, 2.49, 36.55**
	all images	0.228, 1.45, 29.31	**0.059, 0.59, 48.71**

TABLE VI: The effect of proposed modules. MSA represents the proposed multiple scattered attention block. F-FFN denotes the proposed filtered feed-forward network. FC is the proposed feature compensator.

Methods	Baseline	MSA	F-FFN	FC	Med. Err.	Acc.
EXP1	✓	✗	✗	✗	0.029, 0.99	80.13
EXP2	✓	✓	✗	✗	0.025, 0.81	84.76
EXP3	✓	✓	✓	✗	0.023, 0.74	87.35
EXP4	✓	✓	✓	✓	**0.020**, **0.64**	**89.78**

TABLE VII: Ablation study for different feature compensator (FC) variants. CL means convolutional layers.

Methods	Variants	Med. Err.	Acc.
EXP1	w/o FC	0.023, 0.74	87.35
EXP2	w/o Local branch	0.022, 0.69	88.41
EXP3	w/o Global branch	0.021, 0.67	88.86
EXP4	replace FC with CL	0.021, 0.70	88.18
EXP5	w FC (ours)	**0.020**, **0.64**	**89.78**

TABLE VIII: Ablation study for Multiple Scattered Attention (MSA) on **7-Scenes dataset**. SA and LA denote naive self-attention and linear attention. TM denotes training memory.

Methods	Variants	Med. Err.	Acc.	Runtime	TM
EXP1	replace MSA with SA	0.021, 0.66	89.49	138.7*ms*	16.2G
EXP2	replace MSA with LA	0.023, 0.74	88.07	**34.9***ms*	**2.9G**
EXP3	w/o Sparse Attention Matrix	0.023, 0.70	88.56	35.8*ms*	3.1G
EXP4	w MSA (ours)	**0.020**, **0.64**	**89.78**	39.1*ms*	3.1G

MSA with the classic linear attention. It can be observed that MSA elevates the localization accuracy by 1.71% at a slight expense of 4.2ms runtime and 0.2G training memory increasing, indicating that the structure of MSA is more suitable for scene coordinates prediction. EXP3: Removing multiple sparse attention matrices and just utilizing a dense attention matrix M to retrieve message results in a clear 1.22% accuracy drop, highlighting the significance of constructing sparse attention matrices in alleviating redundant message passing.

D. Inference speed and model size comparison

To further validate the effectiveness of SACNet, we disclose model parameters (#Params) and inference speed of different approaches running on a single NVIDIA TITAN RTX GPU. When measuring inference speed, we execute the testing code 1000 times and publish the mean time to eradicate randomness. As reported in Table IX, SACNet requires significantly less runtime and parameters than HSCNet and VSNet. Despite having a comparatively larger memory footprint than FDANet and EAAINet, SACNet possesses faster inference time and noticeably surpasses them in terms of localization accuracy by 6.09% and 2.64%, respectively. In comparison to the efficient and compact techniques DSAC* and ACE, SACNet is less effective in inference speed and memory usage. However, concerning localization accuracy on the 7-Scenes dataset, SACNet prominently surpasses DSAC* and ACE by 4.58% and 8.98%, respectively. Therefore, SACNet is more suitable for situations that prioritize accuracy over runtime and storage efficiency. Moreover, we investigate the reason for small parameters but long runtime of EAAINet. We conduct a detailed analysis of the time consumption for each module of EAAINet, identifying that the retrieving operation takes up most of the runtime. Specifically, the interval sampling module proposed by EAAINet retrieves 150 minimum values from each row of a distance matrix with dimension of $\mathscr{R}^{300\times300}$, consuming approximately 9.3ms, which accounts for almost 49.47%

TABLE IX: **Inference speed, parameters, and accuracy on 7-Scenes dataset** comparisons.

	Runtime (SCoRe)	Runtime (PnP)	Runtime (total)	Accuracy	#Params
HSCNet [4]	51.1*ms*	30.6*ms*	81.7*ms*	84.80	165M
VSNet [20]	783.2*ms*	35.5*ms*	818.7*ms*	—	236M
FDANet [19]	16.2*ms*	30.1*ms*	46.3*ms*	83.69	97M
EAAINet [5]	18.8*ms*	27.4*ms*	46.2*ms*	87.14	76M
DSAC* [8]	**4.0***ms*	30.4*ms*	**34.4***ms*	85.20	28M
ACE [7]	6.1*ms*	32.4*ms*	38.5*ms*	80.80	**4M**
SACNet	12.9*ms*	**26.2***ms*	39.1*ms*	**89.78**	99M

of the total SCoRe runtime (18.8ms). Despite utilizing retrieving operation to construct sparse attention matrices, the multiple scattered attention layer (MSA) of SACNet employs a multi-head attention mechanism to reduce the dimension of attention matrix from $\mathscr{R}^{512\times512}$ to $\mathscr{R}^{64\times64}$, hence significantly declining the runtime. Experimental results exhibit that MSA only takes 0.8ms to construct sparse attention matrices. Since retrieving operation does not occupy model parameters, EAAINet takes a longer runtime than SACNet.

E. Limitaion

In contrast to current excellent visual localization methodologies, such as ACE, our approach presents limitations in terms of both model size and training duration. Exemplified by ACE, designing a compact and efficient technique is crucial for real-time application in real-world. In the future, it will be a meaningful route to investigate a more lightweight Transformer-based visual localization framework to significantly reduce model complexity and training time.

V. CONCLUSION

We propose SACNet, a novel Transformer-based SCoRe method that achieves precise visual localization. Specifically, SACNet proposes a multiple scattered Transformer (MST) that calculates multiple scattered attention matrices along the channel dimension to reduce computational complexity and relieve redundant information passing concurrently. In addition, SACNet utilizes gate mechanism in the filtered feed-forward network to further alleviate redundant information. After that, SACNet proposes a feature compensator (FC) that predicts element-wise soft attention masks for local features and global-aware local features, hence enabling network adaptively reconcile the importance of different features. Comprehensive experiments demonstrate that SACNet achieves impressive performance in several benchmarks.

REFERENCES

[1] 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.

[2] Y. Shavit, R. Ferens, and Y. Keller, "Learning multi-scene absolute pose regression with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2733–2742.

[3] L. Zhou, Z. Luo, T. Shen, J. Zhang, M. Zhen, Y. Yao, T. Fang, and L. Quan, "Kfnet: Learning temporal camera relocalization using kalman filtering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4919–4928.

[4] X. Li, S. Wang, Y. Zhao, J. Verbeek, and J. Kannala, "Hierarchical scene coordinate classification and regression for visual localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 983–11 992.

[5] K. Dai, T. Xie, K. Wang, Z. Jiang, D. Liu, R. Li, and J. Wang, "Eaainet: An element-wise attention network with global affinity information for accurate indoor visual localization," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3166–3173, 2023.

[6] T. Xie, K. Dai, S. Lu, K. Wang, Z. Jiang, J. Gao, D. Liu, J. Xu, L. Zhao, and R. Li, "Ofvl-ms: Once for visual localization across multiple indoor scenes," *arXiv preprint arXiv:2308.11928*, 2023.

[7] E. Brachmann, T. Cavallari, and V. A. Prisacariu, "Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5044–5053.

[8] 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.

[9] Y. Abouelnaga, M. Bui, and S. Ilic, "Distillpose: lightweight camera localization using auxiliary learning," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 7919–7924.

[10] 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.

[11] E. Brachmann and C. Rother, "Learning less is more-6d camera localization via 3d surface regression," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4654–4662.

[12] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "Dsac-differentiable ransac for camera localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6684–6692.

[13] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[14] A. Ali, H. Touvron, M. Caron, P. Bojanowski, M. Douze, A. Joulin, I. Laptev, N. Neverova, G. Synnaeve, J. Verbeek *et al.*, "Xcit: Cross-covariance image transformers," *Advances in neural information processing systems*, vol. 34, pp. 20 014–20 027, 2021.

[15] J. Valentin, A. Dai, M. Nießner, P. Kohli, P. Torr, S. Izadi, and C. Keskin, "Learning to navigate the energy landscape," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 323–332.

[16] 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.

[17] 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.

[18] 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.

[19] T. Xie, K. Dai, K. Wang, R. Li, J. Wang, X. Tang, and L. Zhao, "A deep feature aggregation network for accurate indoor camera localization," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3687–3694, 2022.

[20] Z. Huang, H. Zhou, Y. Li, B. Yang, Y. Xu, X. Zhou, H. Bao, G. Zhang, and H. Li, "Vs-net: Voting with segmentation for visual localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6101–6111.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] S. Wang, Z. Laskar, I. Melekhov, X. Li, Y. Zhao, G. Tolias, and J. Kannala, "Hscnet++: Hierarchical scene coordinate classification and regression for visual localization with transformer," *arXiv preprint arXiv:2305.03595*, 2023.