Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 1243041, 12 pages
https://doi.org/10.1155/2022/1243041

Research Article

A Sparse Feature Matching Model Using a Transformer towards Large-View Indoor Visual Localization

Ning Li,[1,2] Weiping Tu,[1,2] and Haojun Ai[3]

[1]*National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China*
[2]*Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China*
[3]*School of Cyber Science and Engineering, Wuhan University, Hubei 430072, China*

Correspondence should be addressed to Weiping Tu; tuweiping@whu.edu.cn and Haojun Ai; aihj@whu.edu.cn

Received 12 May 2022; Accepted 21 June 2022; Published 4 July 2022

Academic Editor: Yuanlong Cao

Accurate indoor visual localization has been a challenging task under large-view scenes with wide baselines and weak texture images, where it is difficult to accomplish accurate image matching. To address the problem of sparse image features mismatching, we develop a coarse-to-fine feature matching model using a transformer, termed MSFA-T, which assigns the corresponding semantic labels to image features for an incipient coarse matching. To avoid the anomalous scoring of sparse feature interrelationship in the attention assigning phase, we propose a multiscale forward attention mechanism that decomposes the similarity-based features to learn the specificity of sparse features, the influence of position-independence on sparse features is reduced and the performance of the fine image matching in visual localization is effectively improved. We conduct extensive experiments on the challenging datasets; the results show that our model achieves image matching with an average 79.8% probability of the area under the cumulative curve of the corner point error, which outperforms the related state-of-the-art algorithms by an improvement of 13% probability at 1 m accuracy for the image-based visual localization in large view scenes.

1. Introduction

Obtaining an accurate indoor location is a key for location-based services such as seamless indoor-outdoor integrated navigation and multimedia information push in smart cities and augmented/virtual reality applications [1]. The demand for the high-precision location-based services in large indoor spaces is also becoming increasingly urgent.

Visual indoor localization is currently the mainstream solution under the premise of high precision location [2]. The localization accuracy estimated with visual information exceeds that with wireless radio frequency (RF) signals, IMU, and geomagnetic signals. The RF signal is affected by multipath effects and signal fading, while IMU suffers from error accumulation, and they cannot compete with robust visual localization. Visual localization, i.e., estimating the camera pose by query image matching to the scene model, is a core problem under a large-view condition in computer vision. In the absolute pose estimation of a camera, it is necessary to estimate the pose in an indoor coordinate system using the information provided by the image database and 3D point clouds. The main challenge for the image-based [3–5] or structure-based [6, 7] indoor visual localization methods is to obtain the exact image feature matching (i.e., find the feature points corresponding to the query image from the candidate images) and complete the homography constraint in optimal camera pose estimation [8, 9]. However, in complex indoor scenes, especially images in large-view scenes with long viewing distances and wide baselines, which contain the visual information with sparse features, feature distortion or partially occluded makes it difficult to accomplish accurate feature matching of the visual localization. Similarly, some viewpoint changes in a wide range of viewing conditions lead to acute perspective distortion, which results in a little scene structural overlap between the query and the candidate images. Because image matching

focuses on the small part of an image [10], the variability of the scale and rotation of local features makes feature matching in large view scenes highly ambiguous and unable to accomplish accurate visual localization.

The precise correspondence of image features between the query and the database is a key to visual localization under a wide range of viewing conditions. The accuracy of image matching in such scenes can be improved by visual semantic information and spatial context [11]. The works [12, 13] extracted scene semantic information for consistency matching and used the geometric and semantic understanding of the scene to learn the new generative descriptors for positioning under failed scenes. These methods are able to eliminate the influences of illumination and occlusion for visual long-time localization. However, the accuracy of geometric descriptors [14] and semantic segmentation models [15] needs to be further improved for getting accurate geometric features and semantic annotation of the large-view indoor scene 3D model. For the image sparse feature matching of visual localization in large view scenes, the attention-based matching algorithm provides a promising approach [16, 17], the translational and rotational invariance of features is learned to enhance the expression of sparse features, and the different matching strategies are accomplished through different attention weights assignment, which can solve the ambiguous problem of feature matching in a large view scene with crossviewpoint. In the existing methods of self-attention weights and crossattention weights [17–19], the anomalous attention weights of the sparse feature points under weak texture scenes are prone to occur for location-independent feature points (i.e., feature points prone to distortion) because they are not subject to any constraints, leading to the pervasive weak texture image matching errors in large view and increasing visual localization errors.

To tackle the above challenges, we investigate the problems of ambiguous matching of sparse features and anomaly weights for visual feature correlation under large view scenes. We develop a coarse-to-fine feature matching model to remove the dependence on appearance-based reliable feature matching and reduce the effects of the large view and viewpoint changes. As shown in Figure 1, a key insight of our method is to learn the self-correlation among the image sparse features and crosscorrelation among the features on different image patches through semantic correlation and forward multiscale attention mechanism, which reduces the influence of image distortion and improves the matching accuracy of sparse feature points under a wide range of viewing conditions. The key contributions are summarized as follows:

(1) We develop a novel coarse to fine feature matching network with a transformer, termed MFSA-T, which solves the problem of sparse feature matching in large view scenes. Meanwhile, semantic match consistency and position correlation are exploited to improve the robustness of the refined matching model

(2) We propose a multiscale forward attention mechanism to solve the anomaly score of sparse feature point interrelationship and the attention weight on different image patches. This mechanism enables our network to decompose the similarity features to learn the specificity, which improves the matching accuracy of the sparse features in weak texture regions and refines the visual localization in large view scenes

(3) We achieve an average correct matching rate of 79.8% in large view scenes and reduce the localization error by 9.5% in wide baseline scenes of the public datasets, which outperforms the state-of-the-art image matching algorithms. The performance of image-based visual localization algorithms using the MFSA-T model in large-view scenarios is successfully improved

The rest of the paper is organized as follows. Section 2 discusses the existing studies related to this research and Section 3 illustrates the method regarding the developed sparse feature matching network in large views scenes. Finally, the experiment results along with their analysis and the summarization of the developments are discussed in Section 4 and Section 5, respectively.

2. Related Work

Robust visual localization in large view scenes is an essential problem in computer vision. The solution of this problem in difficult situations is not only a challenging task but also highly relevant in practice, such as augmented reality, multimedia information push, and autonomous robots. Large view scenes with extreme viewpoint changes, a wide baseline of view, and weak textures lead to acute perspective distortion and frequently bring on the few common matching parts between the query and the database. These challenges in visual localization attract a large number of researchers to investigate different visual problems [20]. In this section, we review and summarize the research on issues related to visual localization in a large view scene.

2.1. Feature-Based Localization. The mainstream visual localization algorithms for large-scale complex indoor scenes use local feature matching of the query image with the 3D model from the structure from motion (SFM) [21] of the scene, such as SIFT [22] and FREAK [23]; the homography matrix formed by the corresponding features after RANSAC filtering is solved by perspective-n-point (PnP) [24] to estimate the pose of the query image [9]. To eliminate the influence of viewpoint changes and weak textures in large-view scenes, the geometric features of the scene are utilized in [25] to complete the regional correspondence of the scene and the multiple scales local correspondence of the same ratio. This type of traditional descriptor matching usually uses region priority matching or efficient sparse feature association, which is typically a direct matching scheme. But the robustness of this type of method decreased dramatically due to visual distortion occurring in large-view scenes; the localization performance is substantially reduced.

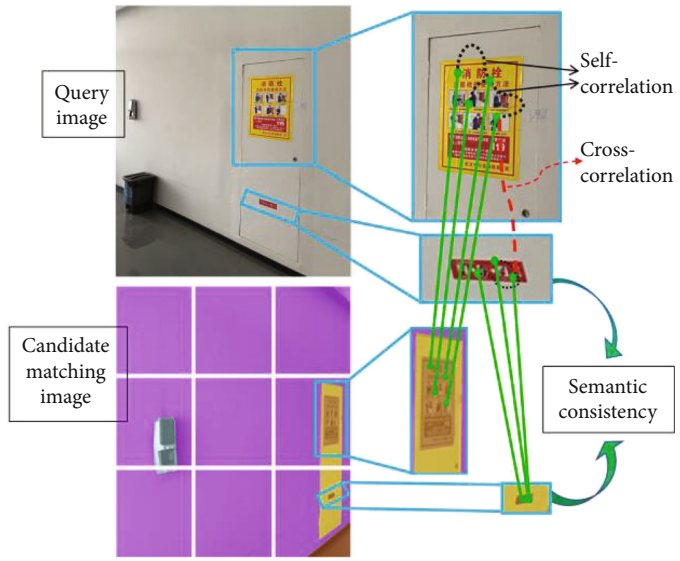

Figure 1: Schematic of sparse feature matching in weak texture scene with viewpoint change.

Camposeco et al. [26] proposed geometric outlier filtering to remove the mismatching relationship of features caused by a viewpoint change in large view scenes. The optimal camera pose estimation result in a large scene is the one with the most votes [27, 28], searching for the covisibility information between a query image and database images [7, 29, 30], retrieving the structural overlap region to eliminate the influence of the wide baseline scenes, and seeking the key frames and local matching features of query images [31], which can effectively remove the influence of viewpoint changes in the visual localization. To ensure the credibility of visual localization results, Taira et al. [10, 32] proposed the pose verification and incorporated scene geometric and semantic information for a trained pose verification model that generates a pose-score similar to the query image by a fractional regression convolutional neural network (CNN).

2.2. Visual Semantic Localization. Visual semantic features have richer scene information and object class information than traditional features, which is more robust to visual information distortion in large views [33]. In recent years, visual semantic information has been used in indoor positioning with promising results [34]. An extended structure-based method was proposed in [12] by combining image features and semantic understanding of the scene in the camera pose estimation stage of the query image. The method uses the geometric outlier filtering [27] and scene semantic labels to deal with the wide range of viewing scenarios where it is hard to seek the correct correspondences of image features. Toft et al. [35] proposed a sparse 3D point cloud model composed of scene curves and pixel-wise semantic labellings of the query image to enhance the image features discrimination for visual localization. Another semantic localization strategy is to include the image semantic information in the feature matching process of the visual localization algorithm [13, 36, 37], i.e., detecting and matching semantic features of the scene images. The latter type of semantic localization method only provides an additional weak semantic feature information does not solve the problem of seeking enough correct matches in wide baseline scenes, which motivates our work.

In contrast to the approaches previously discussed, our method focuses on the image feature matching stage of visual localization. Our model combines the sparse features in large-view scenes and the corresponding semantic information into a single confidence feature and learns discriminative and crosscorrelation of features, which completes accurate image matching to improve visual localization accuracy in wide baseline and long-range view scenes.

2.3. Learning-Based Feature Matching Network. Recent works show that the learning-based image matching network significantly improves matching performance [17]. Learning-based feature matching models can be divided into two categories. A common strategy of the first category is to learn the translation invariance and rotation invariance of feature descriptors [16, 38, 39] to enhance the representation of image features. A trainable single-image matching CNN was proposed in [40], which is a dense feature descriptor as well as a feature detector. The obtained keypoints by trainable CNN are more robust and stable than their traditional counterparts. The second category of approaches mainly focuses on different matching strategies for image features; for instance, a universal dense correspondence network was proposed [41] for geometric and visual semantic matching of images. Sarlin et al. [18] proposed a sparse feature matching model with GNN (graph neural network), which completes feature matching by self-attention and crossattention. A pixel dense matching network with a transformer was proposed in COTR [19], which selects query interest points and retrieves sparse counterparts between images to obtain local and global prior information by iteratively estimating scaling around the points. The same

self-attention and crossattention layers are used in LoFTR [17], a coarse-to-fine image matching model, where the steps of sequentially performing image feature detection, description, and matching are replaced by a pipeline using coarse-to-fine image feature matching. This algorithm conducts pixel dense matching at the coarse granularity and then refines the matching at fine granularity, which improves the image matching accuracy for weak texture scenes. However, if feature points are position-independent, they have similar background features (e.g., walls with weak texture or untextured corridors); some models [17–19] miscalculate the image attention weights and cannot complete accurate matching.

In contrast to the above, we focus on the precise correspondences of image features of the matching stage in visual localization. We propose multiscale forward attention to improve the self-correlation and crosscorrelation of sparse features for the anomalous scores of attention weighting of sparse feature points in large view scenes. We establish a coarse-to-fine feature matching model using a transformer network to better the image feature matching accuracy in extreme viewpoints.

3. Method

To address the matching ambiguity in the image matching phase of visual localization under large-view scenes, we propose a novel coarse to fine sparse feature matching network using a transformer, named MSFA-T, which is also suitable for other applications based on image matching such as object tracking and object retrieval. The structure of our model is shown in Figure 2, ε is the D2-Net [40] model used to extract the CNN features and the positions, $I - I'$ are the input images, and M is the mapping of feature map F and semantic segmentation map S.

Our goal is to train a coarse-to-fine sparse feature matching model that can output optimal geometric constraints for the visual localization algorithm in large-view scenes. First, we obtain the semantic features of the query image and candidate images by SETR [15], which can perceive the large view scene with failed localization in scene recognition. The scene semantic features are embedded into the sparse feature points of the image for coarse feature matching. Our model obtains the spatial locations of the feature points with similar semantics to learn the interrelationship of different feature points and decompose the similarity features, which completes a coarse matching of image features and the division of image patches with semantic classes and solves the problem of the misclassification of feature points under distorted view. We propose a multiscale forward attention mechanism (MSFA) embedded into a transformer to compute the attention weights of sparse features at different positions and to motivate the model to learn the self-correlation of features with the same semantic information and the crosscorrelation of features with different semantic information. MSFA module deals with the problem that the image distortion at a long-viewing distance produces anomalous scores on the attention weights of image features. The main specific constraints derived from the computation of feature vectors by the neural network (NN) are also executed. Finally, the transformer output vector is decoded by a multilayer perceptron (MLP) to obtain a confidence feature matrix for accurate image matching. Our model provides the optimal geometric constraints for visual localization.

After expanding the feature patches obtained by coarse matching to one-dimensional vector, we add positional encoding. We use the general linear positional encoding in transformers following DETR [42]; the positional encoding gives each feature patch unique position information to ensure that the transformed vectors of the sparse features become position dependent. This process enables our model to resist the influence of weak texture regions. The fused position-encoded feature vectors are fed into the transformer, and their weights are obtained according to our proposed multiscale forward attention module for computing confidence features.

3.1. Semantic Mapping for Coarse Matching. The mapping of semantic maps to image patches assigns different semantic labels to the sparse features of images, which facilitates the calculation of the self-correlation of feature points with the same semantic information and the crosscorrelation between different semantic feature points (as shown in Figure 1); meanwhile, it provides a priori information for coarse image matching. The incorrect matching of image features in weak texture scenes is significantly reduced. The specific computational details are as follows.

Semantic class labels are constructed by performing pixel-level semantic segmentation on all images [33], semantic label S_c is assigned to its same semantic class of image patches. The feature map after semantic mapping is defined as:

$$M = \left\{ \left(F_{i,j}, S_{c,(i,j)} \right) \right\}_{i=1,j=1}^{N}, \tag{1}$$

where each feature point and its image coordinates are defined as $F_{i,j}$, its semantic label class and corresponding semantic label position are defined as $S_{c,(i,j)}$, and N is the total number of feature patches.

As each feature patch is given a semantic label, we compare the observed feature patches and the corresponding semantic labels between the query and the database for scoring semantic consistency to obtain a coarse region where the feature patches are located. The semantic matching score is defined as

$$Y_{F_{i,j} \leftrightarrow F'_{i,j}} = \left\{ S_c \in \mathbb{R}^2, \left\| S_c \ominus S'_C \right\|^{-1} \right\}. \tag{2}$$

For the retrieved coarse match region, we crop it out with a partial window of size $m \times n$, as shown in the blue-boxed area in Figure 1. Coarse matching outputs of the same semantic region can be used with the dual-softmax operator, which is also the optimal transport layer in SuperGlue [18], as the output results can all be matched differentiable. The local window of a coarse matching region is refined to a

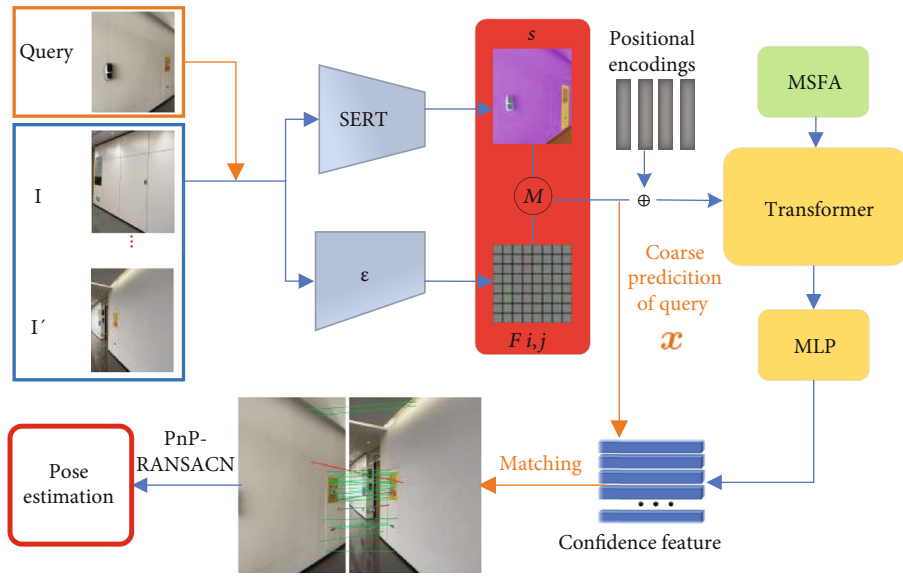

Figure 2: Overview of the proposed method, an image feature extraction, and matching structure.

subpixel level, the center of the query window is fixed as the query feature, and then, the distance between the feature in the window of the candidate image and the center of the feature in the query window is calculated; the image patches with highest scores are used as the accurate predicted match prediction of final image feature.

3.2. Multiscale Forward Attention Module. Feature matching models with a transformer calculate the attention weights of different feature points to enhance the correlation and uniqueness of feature points [17, 18], which reduces the feature matching errors in weak texture scenes. However, the type of methods are prone to abnormal scores in the process of local attention weights, which makes the larger deviation of relevant scores between neighboring feature points and leads to position-independent feature point matching errors in weak textures. We propose the multiscale forward attention module, MSFA, as shown in Figure 3. This module uses the self-attention weights at previous moment to smooth the anomaly scores at current moment and constrains the previous moment attention weights to optimize the forward attention model for the purpose of adaptive smoothing. A multiscale model, then, is introduced to different feature units for obtaining the target feature vectors with different characteristics, which solves the problem of attention weight anomaly scores.

We use different scale convolution filter (S-CF) on the basis of the multiheaded self-attention model to obtain feature units at different positions. We then model the feature units with different weights and calculate the interrelationships between different feature units. In addition, the target feature vectors with different weights are spliced and fused by a NN. Finally, the transformer output vector is decoded by a multilayer perceptron to obtain the confidence feature matrix. The specific calculation process is formulated as follows.

By smoothing the self-attention weights from the previous moment to the current moment, the new attention score of the current moment is $\bar{A}_{i,j}$.

$$\bar{A}_{i,j} = A_{i,j} \bullet \left(\sum_{t=0}^{l-1} \bar{A}_{i-t,j-1} \right), \tag{3}$$

where $A_{i,j}$ is the self-attention score at position i and moment j, $0 \le A_{i,j} \le 1$. The computation of attention weights is to select relevant information by measuring the similarity between query (Q) and each key (K); its output vector is a weighted sum of values with similarity scores. We use the dot product to weight the input features, which can be expressed as $A = \text{softmax}(Q \bullet K^T) \odot V$. l is the one-dimensional vector expanded by the input vector and positional encoding, and $\bar{A}_{i-t,j-1}$ is the forward attention score at the position $i - t$ of the previous moment.

After normalizing the forward attention weights at different positions using the softmax function, the anomalies at the current moment are smoothed by the self-attention weights to eliminate the anomalous scores of the attention weights, ensuring the continuity between the attention weights of different feature units at the previous and next moments. We note that the influence degree of single forward moment attention weight on n forward vectors is not consistent. It is not consistent for the attention weights of vectors at different moments; therefore, new constraint information needs to be added to the forward n vectors to improve the effect of smoothing anomalous attention score. We use a NN to generate a constraint factor φ_j to dynamically control the influence of the attention scores corresponding to different vectors in the previous moment on

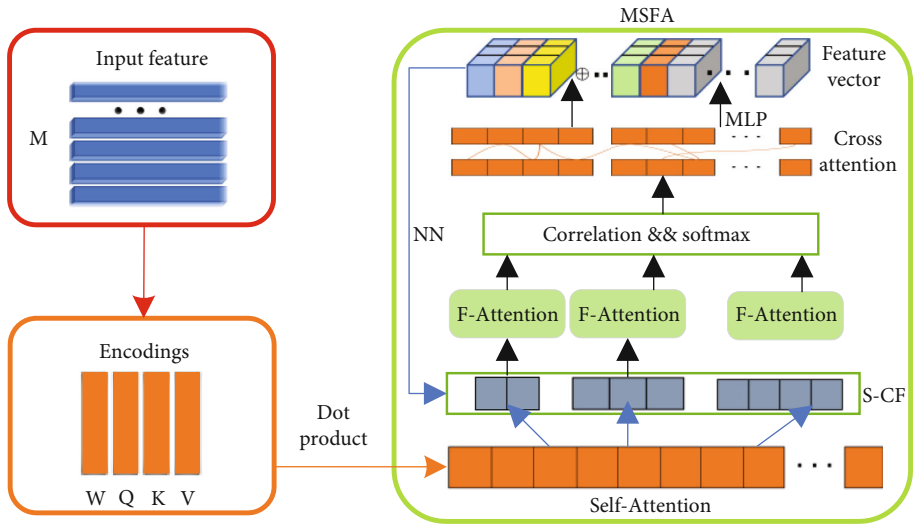

Figure 3: Multiscale forward attention module.

the different vectors in the current moment. The constraint factor φ_j is:

$$\varphi_j = NN\left(q_{j-1}, v_{j-1}, \bar{M}_{j-1}\right), \tag{4}$$

where q_{j-1} is the MLP decoder state at the previous moment, v_{j-1} is the target vector at the previous moment, $\bar{M}_{j-1}$ is the vectors sequence output from the decoder, and NN is a neural network model containing an implicit layer and a Sigmoid activation function. The constraint factor φ_j can add effective constraint information to the attention weight score at the previous moment; thus, the importance related to the vectors with higher attention anomaly scores will be reduced. By dynamically adjusting the importance of the attention weight score at the previous moment, one can optimally smooth that the abnormal attention score at the current moment is achieved. The smoothing function is shown by

$$\bar{A}_{i,j} = A_{i,j} \bullet \left(\sum_{t=0}^{l-1} \varphi_j \bullet \bar{A}_{i-t,j-1}\right). \tag{5}$$

The softmax function is used to normalize $\bar{A}_{i,j}$ so that the attention weights of vector units important at the previous moment are better learned at the current moment. Figure 4 shows the attention weights of the learned image features. Adaptive smoothing of the abnormal attention scores at the current moment is achieved by constraint factors to better align the vector positions of the model. MSFA-T assigns significant attention weights to the union distribution of sparse features in weak texture scenarios, which focuses on significant markings, structure information, object types, or feature location to learn the correlation of sparse feature points within the local regions of semantic consistency. It learns to ignore dynamic objects like pedestrians and repeated patterns like the corridor or wall.

The multiscale forward attention mechanism is used to solve the problem of anomalous attention weights of some feature vectors caused by a low degree of the model representation in weak texture scenes. Different from the multiheaded attention mechanism, we use different sizes of convolutional filters to calculate the respective scores of attention weights for each layer of the multiheaded attention model. The change patterns of feature vectors at different moments are obtained to model the vector units at different scales. Compared to using a single-scale filter in modeling the fixed vector units, the multiscale attention mechanism can extract deeper and richer feature information. In the multiscale model, convolution is computed for the forward attention score $\bar{A}_{j-1}$ using different sizes of convolution filters S-CF as follows.

$$f_j = \left\{C_k * \bar{A}_{j-1}, k = 1, \cdots, 4\right\}, \tag{6}$$

where C is the convolution operation and k is the convolution kernel size. As the image features are expanded as one-dimensional vectors and the positional encoding is also one-dimensional data, the one-dimensional convolutional filtering of different sizes corresponds to sliding windows of different sizes. Sliding on the vectors ensures that the vector units included each time can constitute a feature unit, thus preventing the same feature unit from being assigned different attention weights. The forward attention score of the convolution result f_j is calculated to obtain the target vector of K different feature units, which are finally stitched and integrated by one full connect to obtain the confidence feature matrix with more discriminative and correlative feature model representation.

With the MSFA module, not only can we get refined attention scores by modeling feature units at different scales but also smooth outliers by using normal attention scores from the previous moment to effective elimination of abnormal attention scores to complete the exact feature matching.

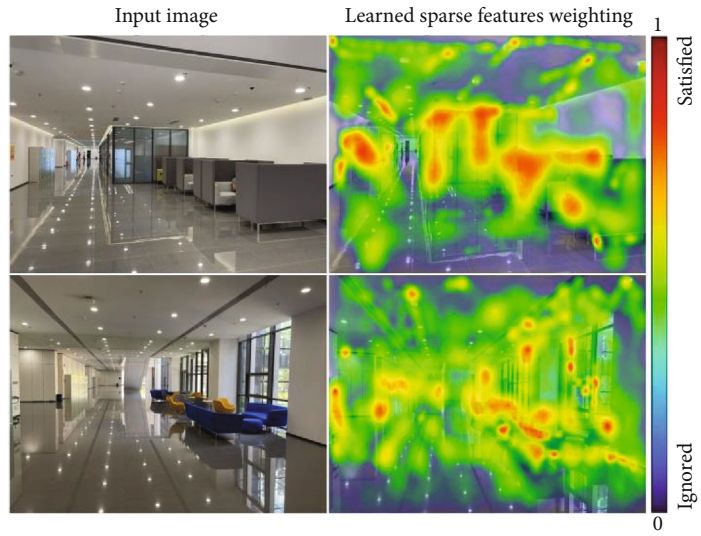

Figure 4: Heat maps of sparse feature attention weights in large view scenes.

3.3. Refined Matching and Loss Function. We obtain the exact matching prediction of a query image by selecting the matching terms based on the confidence threshold θ_m and mutual nearest neighbor (MNN) criteria. The matching process is as follows. We first calculate the score matrix Y between the output features by

$$Y_{i,j} = \tau^{-1}\left\{F_{i,j}, F'_{i,j}\right\}, \tag{7}$$

where the τ^{-1} is the scale factor of the matching, then softmax and MNN on the two dimensions of the score matrix Y are applied to predict the matching probability M_c. We denote the refined matching M_c as:

$$M_c = \{\forall(i,j) \mid MNN\left[\text{softmax}(Y_i, \cdot)_j \cdot \text{softmax}\left(Y_{\cdot,j}\right)_i\right] > \theta_m\}. \tag{8}$$

The final loss function includes both coarse-level loss and fine-level loss, i.e., $L = L_c + L_f$. In coarse matching, each feature point $F_{i,j}$ is directly compared for the score of semantic label consistency and distance difference; its generated variance $\sigma^2(i)$ is calculated by the position error to measure its uncertainty. The weighted loss function of the coarse-level matching is:

$$L_c = \sum_{i,j} \sigma^2(i)^{-1} \left\| S_c \ominus S'_C \right\|_2. \tag{9}$$

The fine-level loss function is generated from the negative log-likelihood loss on matrix M_c obtained by the dual-softmax operator. The feature matching is performed using MNN, so that the loss function is:

$$L_f = -\sum_{i,j} \log M_c(i,j). \tag{10}$$

In the localization phase, the output in feature matching with MSFA-T model is used to form a homography matrix using an efficient association algorithm for feature maps and 3D point clouds [30]. The pose of a query image is finally solved by PnP-RANSAC [9].

4. Implementation Details and Experiment Results

In this section, we present the training implementation details of our model, evaluate the image matching accuracy of MSFA-T compared with the state-of-the-art methods, and assess the role of the MSFA-T model in the visual localization systems.

4.1. Datasets. Training data: we train our image matching model MSFA-T on the ScanNet [43] dataset and the MegaDepth [44] dataset. ScanNet is an RGB-D indoor scene dataset that contains a series of views in 1513 indoor scenes annotated with 3D camera poses and semantic segmentations. MegaDepth dataset provides a large number of large-view images and corresponding dense depth maps generated by SFM [21], which includes large variations in appearance of scenes and viewpoint changes of a camera. The above datasets are required to learn translational invariance and rotational invariance models to improve the robustness of the model for large view scenes. Existing accuracy of the depth maps is sufficient to learn accurate local features [19] in the large view scenes, reducing the influence of weak texture scenes.

Testing data for image matching and visual localization: we used the image matching challenge HPatches dataset [45] to test the matching accuracy of our model for large-view scene images, as well as its robustness to viewpoint changes, long-view distance, and weak texture scenes. HPatches is a challenging dataset for image matching, which contains wide-baseline stereo images, long-range views images, and weak texture images. In addition, we used the InLoc [10]

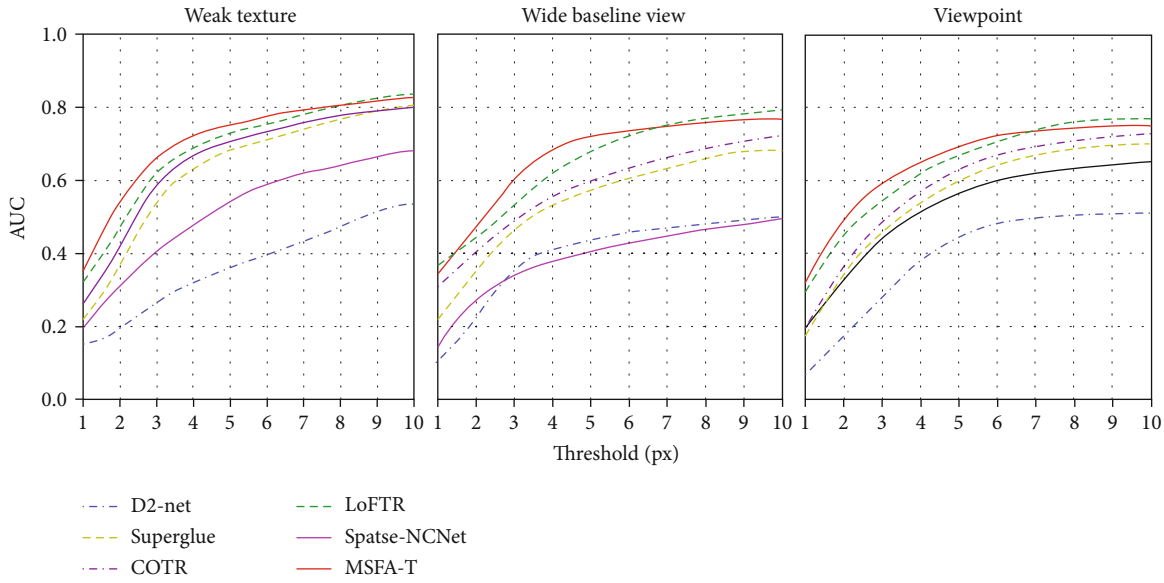

Figure 5: Evaluation on HPatches for image matching.

indoor dataset to test the improvement of MSFA-T in the accuracy of the localization algorithm and also used our previous work EfiLoc [9] and related localization algorithms, Image Bimodal Localization [46] and HAIL [47] for comparison. We compared only the image localization modules of the above localization algorithms. InLoc provides the large-scale indoor data based on two Washington University buildings, including 356 pieces of 4032×3024 query images and 9972 pieces of 1600×1200 database images that contain the scenes with wide baselines and weak textures. Thus, the localization based on such a dataset is a challenge considering the complexity of the indoor wide range of view scenes.

4.2. Implementation Details. We used a share backbone ResNet [48] architecture and a semantic segmentation SETR model [15] to initialize the CNN feature extraction network and semantic segmentation network, respectively. We used the feature map after the fourth downsampling layer of size $16 \times 16 \times 1024$ in the residual network with a convolutional kernel of size 1×1, the initial learning rate of $1 * 10^{-3}$ and a batch size of 64. For the transformer, we used the same number for layers of encoder and decoder; each encoder layer contains a self-attention layer and a multiscale forward-attention layer to ensure that accurate learning weights are assigned to each feature patch to enhance the self-correlation of image features. Each decoder layer contains the corresponding encoder-decoder attention layers without self-attention layers, which prevent the mutual communication between query points in order to enhance the relevant communication between query points and candidate points. Finally, we used 3-layer MLP to decode the vector output from the transformer and obtain the confidence matrix for query matching. We evaluate the performance of image-based visual localization systems [9, 46, 47] that use our image matching model and compared their localization accuracy under different scenarios.

Table 1: Evaluation on HPatches image pairs.

Method	AUC			#matches
	3 px	5 px	10 px	
D2-net	23.2	35.9	53.6	0.2 K
SuperGlue	53.9	68.3	81.7	0.6 K
COTR	62.8	67.9	80.6	1.0 K
Sparse-NCNet	48.9	54.2	67.1	1.0 K
LoFTR	65.9	75.6	**84.6**	1.0 K
MS-T	28.5	48.6	52.7	1.0 K
MFA-T	47.7	62.8	73.9	1.0 K
MSFA-T	**68.5**	**76.9**	83.5	1.0 K

4.3. Experiment Results. Image matching: to evaluate the performance of our model, we compared it with the state-of-the-art models, D2-Net [40], COTR [19], SuperGlue [18], Sparse-NCNet [49], and LoFTR [17]. D2-Net is a detector-based local feature matching network that uses a describe-and-detect methodology. The detection of D2-Net is postponed until a more reliable image feature is available and done jointly with the image description. SuperGlue is a detector-based local feature matcher, which uses self-attention and crossattention to improve the matching accuracy of image feature points (SuperPoint [16]). COTR, Sparse-NCNet, and LoFTR are detector-free matchers models, which have no local feature keypoints and directly output the dense matching result of the image. In addition, in order to confirm the important roles in assigning semantic features and multiscale forward attention mechanism to image CNN features in our model, we trained MS-T model, i.e., MSFA-T without multiscale forward attention mechanism, and MFA-T model, i.e., MSFA-T without semantic feature fusion module, respectively. We design ablation experiments to test their image matching performances in

Figure 6: Comparison results of image matching in large view scenes.

comparison with the related state-of-the-art matching algorithms.

For the matching challenge on the HPatches dataset, we restricted the number of image keypoints rather than the correct matching rate. For the image local feature matching algorithm, we restricted the extraction to a maximum of 2 K features with mutual nearest neighbors as the matches phase. For detector-free methods, which directly output the matches, we restricted the matches results with a maximum of 1 K outputting matches. Meanwhile, we used the initial default hyperparameters in the original matching algorithm implementation for all the baselines. Figure 5 shows the comparison of image matching results for wide-baseline view, weak texture, and viewpoint changes. For each method, we show the mean number of mutual nearest neighbor matches per image at different matching thresholds. From the comparison results, our method outperforms the other methods for the matching threshold below 7 pixels, especially in indoor scenes with weak textures and wide baseline views. Our approach makes the coarse-to-fine matching process that from semantic consistency matching to sparse features with the same semantic labels play an important role. Our multiscale forward attention overcomes the problem of anomalous scoring of sparse feature weights in weak texture scenes, which enhances the self-correlation and crosscorrelation of these features, improving the overall performance of the model.

The overall evaluation results on the HPatches dataset are shown in Table 1. We report the area under the cumulative curve (AUC) of corner error in image matching with the threshold of corner error being 3, 5, and 10 pixels, respectively. The AUC of the corner error as a function of the matching threshold in percentage is shown. Bold values in the table indicate the best results for that particular experiment. Our method has higher matching accuracy, especially for the error thresholds of 3 and 5 pixels in weak texture scenes.

Our MSFA-T matching model achieves the optimal performance with the error threshold values of 3 and 5 pixels, respectively. LoFTR achieves the optimal matching result with the error threshold value of 10 pixels because it uses the good matches at a fine level. In contrast, we fused the scene semantic features with the image CNN features so that the model filters out some semantic conflicting sparse features to ensure the refined matches of the images in complex large views.

We also perform ablation experiments on models MS-T and MFA-T. The MS-T model without the multiscale forward attention mechanism shows some sparse feature matching errors in weak texture scenes with wide baselines, which is due to the attention weight learning anomaly on position-independent features in this scene, causing the correlation between the features to be misallocated. The MFA-T model without the semantic feature fusion module shows the matching errors of some different types of objects due to the lack of the sparse features with semantic label information in wide baseline scenes and viewpoint change scenes. The MSFA-T model, which uses both the semantic information fusion mechanism and the multiscale forward attention mechanism, shows optimal matching results in large viewpoint scenes. The performances of the above models with the error threshold values of 3, 5, and 10 pixels, respectively, are shown in Table 1. These experimental results demonstrate the effectiveness of the coarse-to-fine network (semantic correspondence coarse matching to fine matching of features with the same semantic information) and multiscale forward attention mechanism proposed in this paper for refined image matching and also show the robustness of our method for large view scenes. The partial image matching schematic of our method with different module on indoor image pairs is shown in Figure 6. The green color indicates the correct match with a probability close to 1, in contrast, the lower the probability, the closer the color to red. MSFA-T achieves the best matches and fewer mismatches, which successfully copes with the image matching in weak texture areas and wide baseline views.

Indoor visual localization: accurate localization of indoor vision relies on robust image matching algorithms; therefore, we used the MSFA-T model in the image matching phase of indoor localization in indoor large view scenes to evaluate the localization performance of EfiLoc and related state-of-the-art visual localization algorithms [46, 47]. Similarly, we compare these localization algorithms using the MSFA-T model with original localization algorithms. EfiLoc-MSFA-T denotes the EfiLoc localization algorithm that uses the MSFA-T model, the same for others, e.g., IBL-MSFA-T

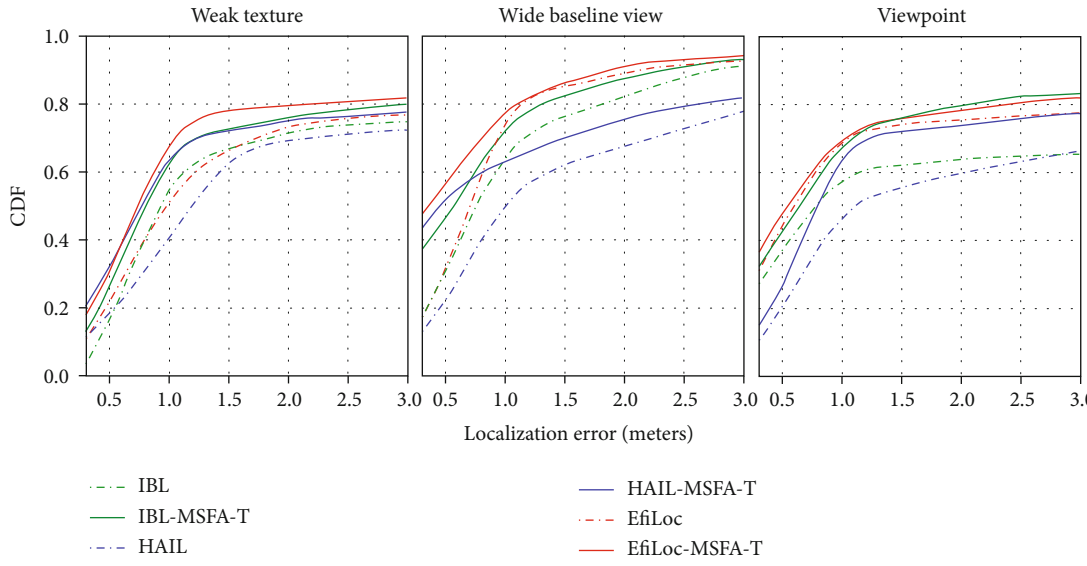

Figure 7: Localization error comparison.

denotes the Image Bimodal Localization with MSFA-T and HAIL with-MSFA-T (HAIL-MSFA-T). The comparison of the cumulative error function (CDF) of these positioning methods is shown in Figure 7.

The improvement of the visual localization performance is achieved by using the MSFA-T model instead of the image matching module in the original localization algorithm. The correct localized queries rate of the original localization algorithms (dashed lines in Figure 7) with the MSFA-T image matching model (solid lines) have different degrees of improvement with different influencing factors. At a localization error of 1 m, the performance improvement rate of the correct localized queries is 12% and 9% for IBL-MSFA-T and EfiLoc-MSFA-T, respectively. The general localization performance is most improved with HAIL. This is because HAIL uses the filtered SIFT feature keypoints that cannot accomplish robust image feature matching in the challenging scenarios described above, especially in indoor scenes with weak textures and viewpoint changes. This also demonstrates that our image matching model can successfully improve the performance of visual localization in large viewpoint scenes.

5. Conclusion

In this paper, we propose a model MSFA-T, a robust sparse feature matching network with a transformer, which accomplishes accurate image matching in visual localization in large view indoor scenes. MSFA-T successfully solves the problems of viewpoint distortion and weak textures using the image semantic information and the optimal confidence features. In addition, to deal with the problems of interrelationship and attention weight anomaly score of sparse feature points on different image patches, we use the transformer with our MSFA module for learning the specificity and correlation of the sparse features, which improves the matching accuracy of the sparse features in weak textures regions to enhance refined visual localization in large view scenes. MSFA-T accomplishes an average 79.8% probability of the AUC of the corner error in large view scenes, which outperforms the related state-of-the-art image matching algorithms. Moreover, our model improves on average the localization accuracy of image-based visual localization by 11.2% on the InLoc dataset. We believe the MFSA-T model takes a promising step toward refined image matching to improve a practical smartphone indoor localization services.

Data Availability

The data underlying the results presented in the study are available within the manuscript or directly access these publicly available datasets according to the references [10, 43–45].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant no. 61971316).

References

[1] Y. Cao, R. Ji, L. Ji, G. Lei, H. Wang, and X. Shao, "l^2-MPTCP: a learning-driven latency-aware multipath transport scheme for industrial internet applications," *IEEE Transactions on Industrial Informatics*, 2022.

[2] J. Xu, E. Dong, Q. Ma, C. Wu, and Z. Yang, "Smartphone-based indoor visual navigation with leader-follower mode," *ACM Transactions on Sensor Networks (TOSN)*, vol. 17, no. 2, pp. 1–22, 2021.

[3] Z. Chen, A. Jacobson, N. Sünderhauf et al., "Deep learning features at scale for visual place recognition," in *2017 IEEE*

International Conference on Robotics and Automation (ICRA), pp. 3223–3230, May 2017.

[4] T. Sattler, A. Torii, J. Sivic et al., "Are large-scale 3d models really necessary for accurate visual localization?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1637–1646, Honolulu,USA, July 2017.

[5] X. Bai, M. Huang, N. R. Prasad, and A. D. Mihovska, "A survey of image-based indoor localization using deep learning," in *2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC)*, pp. 1–6, Lisbon,Portugal, November 2019.

[6] X. Chen and G. Fan, "Egocentric Indoor Localization From Coplanar Two-Line Room Layouts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1549–1559, New Orleans, Louisiana, June 2022.

[7] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua, "Worldwide pose estimation using 3d point clouds," in *European conference on computer vision*, pp. 15–29, Springer, Berlin, 2012.

[8] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2d-to-3d matching," in *International Conference on Computer Vision*, pp. 667–674, Barcelona Spain, November 2011.

[9] N. Li and H. Ai, "EfiLoc: large-scale visual indoor localization with efficient correlation between sparse features and 3D points," *The Visual Computer*, vol. 38, no. 6, pp. 2091–2106, 2022.

[10] H. Taira, M. Okutomi, T. Sattler et al., "InLoc: indoor visual localization with dense matching and view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7199–7209, Salt Lake City, Utah, 2018.

[11] F. Gu, X. Hu, M. Ramezani et al., "Indoor localization improved by spatial context—a survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 3, pp. 1–35, 2020.

[12] C. Toft, E. Stenborg, L. Hammarstrand et al., "Semantic match consistency for long-term visual localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 383–399, Munich, Germany, October 2018.

[13] 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*, pp. 6896–6906, Salt Lake City, Utah, 2018.

[14] N. Atanasov, S. L. Bowman, K. Daniilidis, and G. J. Pappas, "A unifying view of geometry, semantics, and data association in SLAM," *IJCAI*, pp. 5204–5208, 2018.

[15] S. Zheng, J. Lu, H. Zhao et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6881–6890, 2021.

[16] 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*, pp. 224–236, Salt Lake City, UT, USA, December 2018.

[17] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8922–8931, 2021.

[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*, pp. 4938–4947, 2020.

[19] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: efficiently bridging cnn and transformer for 3d medical image segmentation," *International conference on medical image computing and computer-assisted intervention*, pp. 171–180, Springer, Cham, 2021.

[20] X. Xin, J. Jiang, and Y. Zou, "A review of visual-based localization," in *Proceedings of the 2019 International Conference on Robotics, Intelligent Control and Artificial Intelligence*, pp. 94–105, Shanghai, China, September 2019.

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

[22] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[23] A. Alahi, R. Ortiz, and P. Vandergheynst, "Freak: fast retina keypoint," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 510–517, Providence USA, June 2012.

[24] F. Youyang, W. Qing, Y. Yuan, and Y. Chao, "Robust improvement solution to perspective-n-point problem," *International Journal of Advanced Robotic Systems*, vol. 16, no. 6, article 172988141988570, 2019.

[25] X. Zuo, X. Xie, Y. Liu, and G. Huang, "Robust visual SLAM with point and line features," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1775–1782, Vancouver Canada, September 2017.

[26] F. Camposeco, T. Sattler, A. Cohen, A. Geiger, and M. Pollefeys, "Toroidal constraints for two-point localization under high outlier ratios," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4545–4553, USA, July 2017.

[27] 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, 2017.

[28] B. Zeisl, T. Sattler, and M. Pollefeys, "Camera pose voting for large-scale image-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2704–2712, Santiago, Chile, 2015.

[29] G. Lu and J. Song, "3D image-based indoor localization joint with WiFi positioning," in *Proceedings of the ACM on International Conference on Multimedia Retrieval*, pp. 465–472, New York, June 2018.

[30] 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, 2017.

[31] 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*, pp. 12716–12725, Long Beach, CA,USA, 2019.

[32] H. Taira, I. Rocco, J. Sedlar et al., "Is this the right place? Geometric-semantic pose verification for indoor visual localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4373–4383, Long Beach, USA, 2019.

[33] E. Stenborg, C. Toft, and L. Hammarstrand, "Long-term visual localization using semantically segmented images," in *In*

international conference on robotics and automation (ICRA), pp. 6484–6490, Brisbane Australia, May 2018.

[34] M. Sualeh and G. W. Kim, "Simultaneous localization and mapping in the epoch of semantics: a survey," *International Journal of Control, Automation and Systems*, vol. 17, no. 3, pp. 729–742, 2019.

[35] 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*, pp. 650–659, Honolulu, HI, USA, 2017.

[36] N. Atanasov, M. Zhu, K. Daniilidis, and G. J. Pappas, "Localization from semantic observations via the matrix permanent," *The International Journal of Robotics Research*, vol. 35, no. 1-3, pp. 73–99, 2016.

[37] F. Yu, J. Xiao, and T. Funkhouser, "Semantic alignment of LiDAR data at city scale," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1722–1731, Boston, USA, 2015.

[38] P. H. Chen, Z. X. Luo, Z. K. Huang, C. Yang, and K. W. Chen, "IF-Net: an illumination-invariant feature network," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8630–8636, Paris, May 2020.

[39] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: learned invariant feature transform," in *European conference on computer vision*, pp. 467–483, Springer, Cham, 2016.

[40] M. Dusmanu, I. Rocco, T. Pajdla et al., "D2-net: a trainable cnn for joint description and detection of local features," in *In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 8092–8101, Long Beach, CA, USA, 2019.

[41] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker, "Universal correspondence network," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[42] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, Cham, 2020.

[43] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, Honolulu, HI, USA, 2017.

[44] Z. Li and N. Snavely, "Megadepth: learning single-view depth prediction from internet photos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2041–2050, Salt Lake City, UT, USA, 2018.

[45] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "HPatches: a benchmark and evaluation of handcrafted and learned local descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5173–5182, Honolulu, HI, USA, 2017.

[46] M. D. Redžić, C. Laoudias, and I. Kyriakides, "Image and wlan bimodal integration for indoor user localization," *IEEE Transactions on Mobile Computing*, vol. 19, no. 5, pp. 1109–1122, 2019.

[47] Q. Niu, M. Li, S. He, C. Gao, S. H. Gary Chan, and X. Luo, "Resource-efficient and automated image-based indoor localization," *ACM Transactions on Sensor Networks (TOSN)*, vol. 15, no. 2, pp. 1–31, 2019.

[48] 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*, pp. 770–778, Las Vegas, USA, 2016.

[49] I. Rocco, R. Arandjelović, and J. Sivic, "Efficient neighbourhood consensus networks via submanifold sparse convolutions," in *European conference on computer vision*, pp. 605–621, Springer, Cham, 2020.