# VS-Net: Voting with Segmentation for Visual Localization

Zhaoyang Huang[1,2*] Han Zhou[1*] Yijin Li[1] Bangbang Yang[1] Yan Xu[2]
Xiaowei Zhou[1] Hujun Bao[1] Guofeng Zhang[1†] Hongsheng Li[2,3]

[1]State Key Lab of CAD&CG, Zhejiang University[‡]
[2]CUHK-SenseTime Joint Laboratory, The Chinese University of Hong Kong
[3]School of CST, Xidian University

## Abstract

*Visual localization is of great importance in robotics and computer vision. Recently, scene coordinate regression based methods have shown good performance in visual localization in small static scenes. However, it still estimates camera poses from many inferior scene coordinates. To address this problem, we propose a novel visual localization framework that establishes 2D-to-3D correspondences between the query image and the 3D map with a series of learnable scene-specific landmarks. In the landmark generation stage, the 3D surfaces of the target scene are over-segmented into mosaic patches whose centers are regarded as the scene-specific landmarks. To robustly and accurately recover the scene-specific landmarks, we propose the Voting with Segmentation Network (VS-Net) to segment the pixels into different landmark patches with a segmentation branch and estimate the landmark locations within each patch with a landmark location voting branch. Since the number of landmarks in a scene may reach up to 5000, training a segmentation network with such a large number of classes is both computation and memory costly for the commonly used cross-entropy loss. We propose a novel prototype-based triplet loss with hard negative mining, which is able to train semantic segmentation networks with a large number of labels efficiently. Our proposed VS-Net is extensively tested on multiple public benchmarks and can outperform state-of-the-art visual localization methods. Code and models are available at https://github.com/zju3dv/VS-Net.*

## 1. Introduction

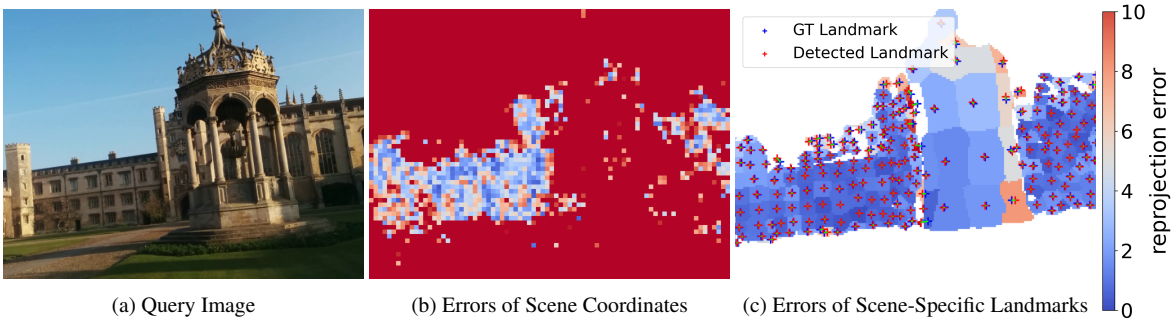
Localization [58, 44, 40] is a pivotal technique in many real-world applications, such as Augmented Reality (AR), Virtual Reality (VR), robotics, etc. With the popularity and low cost of visual cameras, visual localization has attracted widespread attention from the research community.

Recently, scene coordinate regression based methods [8, 7, 25], which learn neural networks to predict dense scene coordinates of a query image and recover the camera pose through RANSAC-PnP [18], dominate visual localization and achieve state-of-the-art localization accuracy in small static scenes. Compared with classical feature-based visual localization frameworks [16, 26, 61, 44] relying on identified map points from Structure-from-Motion (SfM) techniques, it only requires to estimate 2D-to-3D correspondences and can be benefited from high-precision sensors. Although scene coordinates construct dense 2D-3D correspondences, most of them are unable to recover reliable camera poses. In dynamic environments, there could exist moving objects and varying lighting conditions which raise the outlier ratio and increase the probability of choosing an erroneous pose with RANSAC algorithms. In addition, even after outlier rejection with RANSAC, there might exist inferior scene coordinates that lead to inaccurate localization.

In the hope of estimating camera poses more robustly and accurately, we propose *Voting with Segmentation Network (VS-Net)* to identify and localize a series of scene-specific landmarks through a *Voting-by-Segmentation* framework. In contrast with scene coordinate regression methods that predict pixel-wise dense 3D scene coordinates, the proposed framework only estimates a small quantity of scene-specific landmarks (or 2D-3D correspondences) that are of much higher accuracy.

Unlike feature-based visual localization methods, where landmarks are directly extracted from the images according to certain rules, we manually specify a series of scene-

*Zhaoyang Huang and Han Zhou assert equal contributions.
†Corresponding author: Guofeng Zhang.
‡The authors from Zhejiang University are also affiliated with ZJU-SenseTime Joint Lab of 3D Vision. This work was partially supported by NSF of China(Nos. 61822310 and 61932003), Centre for Perceptual and Interactive Intelligence Limited, the General Research Fund through the Research Grants Council of Hong Kong under Grants (Nos. 14208417 and 14207319), and CUHK Strategic Fund.




(a) Query Image (b) Errors of Scene Coordinates (c) Errors of Scene-Specific Landmarks

Figure 1: Reprojection errors of 2D-to-3D correspondences of scene coordinates and scene-specific landmarks. (a) The query image. (b) The reprojection errors of dense scene coordinates predicted by the regression-only network [25]. (c) The reprojection errors of scene-specific landmarks and their surrounding patches by the proposed method. Pixels belonging to the same landmark are painted with the same color representing the landmark' reprojection error. The white pixels in (c) are filetered by our voting-by-segmentation algorithm

specific landmarks from each scene's reconstructed 3D surfaces. The 3D surface of a scene is first uniformly divided into a series of 3D patches, and we define the centers of the 3D patches as the 3D scene-specific landmarks. Given a new image obtained from a new viewpoint, we aim to identify the 3D scene-specific landmarks' projections on the 2D image. The Voting-by-Segmentation framework with the VS-Net casts the landmark localization problem as a combination of patch-based landmark segmentation coupled with pixel-wise direction voting problem. Each pixel in the image is first segmented into one of the pre-defined patches (landmarks) and the pixels classified into the $k$th landmark are responsible for estimating the corresponding landmark's 2D location. To achieve the goal, the proposed VS-Net also estimates a 2D directional vector at each pixel location, which is trained to point towards the pixel's corresponding landmark. For a given patch, such predicted directional vectors can be treated as directional votes. With a RANSAC algorithm, for each predicted patch, the accurate 2D landmark location can be accurately estimated. In contrast to existing scene coordinate regression methods, in our proposed framework, pixels or regions that are poorly segmented with erroneous patch labels and directional votes can be robustly filtered out as those pixels have low voting consistency. Therefore, this strategy ensures that the survived landmarks are of high accuracy and the inferior pixels would not jeopardize the accuracy of camera pose estimation. It results in fewer landmarks with lower outlier ratios and reprojection errors than scene coordinate regression methods (Fig. 1).

The patch-based landmark segmentation in our VS-Net requires assigning pre-defined patch labels, *i.e.*, landmark IDs, to pixels. However, the number of patches or landmarks in a scene can reach tens of thousands. Directly adopting the conventional cross-entropy loss for multi-class segmentation requires huge memory and computational costs as the number of parameters in the classification layer increases proportionally to the number of patches. We propose prototype-based triplet loss to address this problem, which avoids computing complete label scores by developing pixel-wise triplet loss with prototypes. Moreover, prototype-based triplet loss improves the training efficiency by online mining informative negative prototypes.

In summary, our proposed approach has the following major contributions: (1) We propose the novel VS-Net framework that casts the problem of visual localization from scene-specific landmarks as a voting-by-segmentation problem. Camera poses estimated from the proposed scene-specific landmarks are shown to be more robust and accurate. (2) We propose the prototype-based triplet loss for patch-based landmark segmentation with a large number of classes, which shows competitive segmentation accuracy while saving much computation and memory. To our best knowledge, we are the first to address the problem of a large number of classes in image segmentation. (3) The VS-Net significantly outperforms previous scene coordinate regression methods and representative SfM-based visual localization methods on both the popular 7Scenes dataset and the Cambridge Landmarks dataset.

## 2. Related Works

**Visual Localization.** Visual localization aims at estimating 6-DoF camera pose in the map built beforehand for a query image. Traditional visual localization frameworks [4, 16, 26, 61, 44, 12] build a map by SfM techniques [62, 1, 67, 46, 55] with general feature detectors and descriptors [30, 6, 43, 35, 15, 17, 28, 41]. Given a query image, they extract the same 2D features and match them to the 3D features in the map via descriptors. The capability of the feature detector and the feature descriptor is of great importance in this framework because it affects both the

map quality and the establishment of the 2D-3D correspondences in a query image, which determines the localization accuracy. Many feature detectors and descriptors have been proposed, such as handcrafted features [30, 6, 31, 43, 22] and learned features [35, 15, 17, 41, 19, 53]. In SfM-based visual localization systems, the 3D feature points are reconstructed with triangulation according to multiple associated observations. They are always messy in that an ideal 3D point may be represented by different feature points that are not matched and merged because of large viewpoint or scale change, which may impact the following localization.

With the development of deep learning, training a scene-specific neural network to encode the map and localize an image from it becomes an alternative visual localization approach. Neural pose regression [24, 23, 10, 21, 52, 59] learns to directly predict parameters of a camera pose from an image, which are not competitive with other visual localization frameworks in accuracy. Another method is to predict scene coordinates [7, 8, 49, 51, 11, 9, 25, 66] as an intermediate representation and estimate the camera pose through a RANSAC-PnP [18, 49] algorithm, which achieves state-of-the-art localization performance in small and medium scenes. Recently, many works extend this pipeline for better localization accuracy. Brachmann *et al*. [7, 8] learn scene coordinate regression with differential RANSAC. Li *et al*. [25] hierarchically predicts scene coordinates. Zhou *et al*. [66] improves the regression by using temporal information. Weinzaepfel *et al*. [54] propose to localize from Objects-of-Interest, which is so coarse-grained and requires annotations.

**Keypoint-based Object Pose Estimation.** Keypoint is widely utilized as an intermediate representation in object pose estimation [38, 20, 34, 37, 50]. Many of them showed that keypoint-based pose estimation outperforms direct pose regression and object-customized keypoints are better than general features. Inspired by these works, we propose to learn to find scene-specific landmarks for visual localization. Recently, PVNet [38] significantly improves robustness and accuracy of object pose estimation by detecting keypoints with pixel-wise votes, inspired by which, we propose to detect scene-specific landmarks with pixel-wise votes.

**Semantic Segmentation and Large-scale Classification.** Semantic segmentation [29, 42, 60, 14, 64], which predicts pixel-wise labels according to a set of semantic categories, is a long-standing topic in computer vision and has been widely discussed in the past decades. Unfortunately, pixel-wise cross-entropy loss adopted by previous methods devours a lot of memory and computation when the number of categories is large. Furthermore, the classifier matrix can not be learned effectively due to the large variance of gradients [57]. Large-scale classification is also encountered in many other tasks, such as person re-identification [57, 63, 65, 27], face recognition [48, 56], *etc*. Online instance matching (OIM) [57] loss and proxy-based metric learning [5, 32, 39] share a similar idea that maintains a memory bank containing a feature prototype for each label. However, the number of pixels in an image is quite large, and they still compute scores between each data point and each label, which runs into the same situation of cross-entropy loss. To tackle this issue, we propose the prototype-based triplet loss that simultaneously maintains a prototype for each class and a network to predict class labels by imposing pixel-wise triplet loss on prototypes.

## 3. Method

State-of-the-art visual localization methods for small-scale scenes are dominated by scene coordinate regression based methods [7, 8] that establish dense 2D-to-3D correspondences (scene coordinates) between each pixel in an input query image and the 3D surface points of a scene. However, a large portion of predicted scene coordinates shows high re-projection errors, which increase the chance of localization failure and deteriorate the localization accuracy of the follow-up RANSAC-PnP algorithms. To tackle the issues, we propose the Voting with Segmentation Network (VS-Net) to identify a series of scene-specific landmarks (Fig. 2) and establish their correspondences to the 3D map for achieving accurate localization. The scene-specific landmarks are sparsely and directly defined from a scene's 3D surfaces. Given different viewpoints of the training images, we can project the scene-specific landmarks and their surrounding surface patches to the image planes to identify their corresponding pixels in the images. In this way, we obtain the pixels of the surrounding patches of each landmark in the multiple training images. The problem of localizing the scene-specific landmarks from the images can be cast as 2D patch-based landmark segmentation and pixel-wise landmark location voting.

During the training phase, for all pixels corresponding to the same surrounding surface patch of a landmark, their outputs are required to predict the same segmentation label (landmark ID) via patch-based landmark segmentation of the proposed VS-Net. Another landmark location branch is introduced to make each pixel responsible for estimating the 2D location of its corresponding landmark by outputting the directional vectors pointing towards the landmark's 2D projection.

For inference, given a new input image, we obtain the landmark segmentation map and the landmark location voting map from the VS-Net. The 2D-to-3D landmark correspondences can then be established based on the landmark segmentation and location voting maps. Unlike outlier 2D-to-3D correspondences from scene coordinate regression methods that can only be rejected by RANSAC PnP algorithms, landmarks from our proposed approach

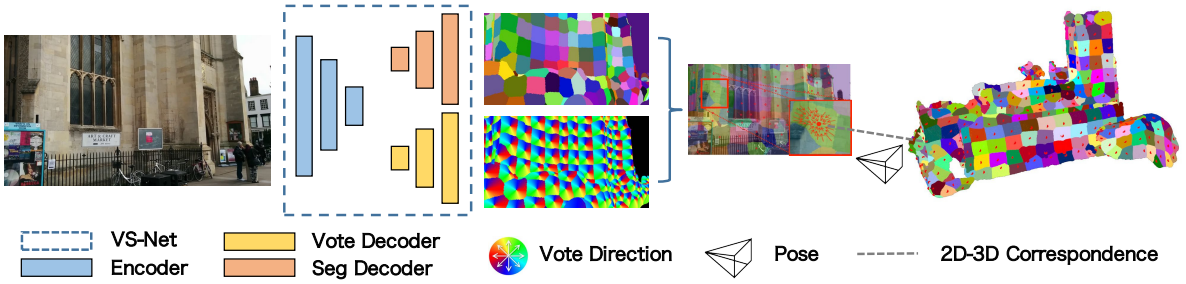



Figure 2: Visual localization by VS-Net. There are two decoder branches respectively predicting a landmark segmentation map and a pixel-wise voting map, from which we can detect the location and labels of landmarks. After establishing 2D-3D correspondences according to the landmark labels, we can estimate the 6-DoF camera pose with a standard RANSAC-PnP.

that do not have high enough voting confidence would be directly dropped, which prevent estimating camera poses from poorly localized landmarks (Fig. 1). Furthermore, the correspondences built upon scene coordinate methods can be easily influenced by unstable predictions, while minor disturbed votes do not deteriorate the accuracy of the voted landmark locations in our method because they would be filtered by the within-patch RANSAC intersection algorithm.

### 3.1. Creation of Scene-specific 3D Landmarks

Given each scene for visual localization, we can obtain the scene's 3D surfaces from existing 3D reconstruction algorithms, such as multiview stereo [47], Kinect fusion [33], *etc*. The proposed scene-specific 3D landmarks are created based on the reconstructed 3D surfaces. We partition the 3D surfaces into a series of 3D patches with the 3D over-segmentation algorithm, Supervoxel [36]. The center points of the $n$ over-segmented 3D patches $\{\mathbf{q}_1, \ldots, \mathbf{q}_n\} \in \mathbb{R}^3$ are chosen as the scene-specific landmarks for localization. As Supervoxel produces patches of similar sizes, the generated landmarks are mostly uniformly scattered on the 3D surfaces, which can provide enough landmarks from different viewpoints and therefore benefit localization robustness.

Given the training images along with camera poses of a scene, the 3D scene-specific landmarks $\mathbf{q}_1, \ldots, \mathbf{q}_n$, and their associated 3D patches can be projected to the 2D images. For each image, we can generate a landmark segmentation map $\mathbf{S} \in \mathbb{Z}^{H \times W}$ and a landmark location voting map $\mathbf{d} \in \mathbb{R}^{H \times W \times 2}$. For patch-based landmark segmentation, each pixel $i$ with coordinate $\mathbf{p}_i = (u_i, v_i)$ is assigned the landmark label (ID) $j$ determined by the projection of the 3D patches. If a pixel corresponds to some region that is not covered by the projected surfaces, such as the sky or distant objects, a background label 0 is assigned to it to represent that this pixel is noneffective for visual localization.

For landmark location voting, we first compute a landmark $\mathbf{q}_j$'s projected 2D location $\mathbf{l}_j = \mathcal{P}(\mathbf{q}_j, \mathbf{K}, \mathbf{C}) \in \mathbb{R}^2$ by projecting the 3D landmark according to the camera intrinsic matrix $\mathbf{K}$ and the camera pose parameters $\mathbf{C}$. Each pixel $i$ belonging to landmark $j$'s patch is responsible for predicting the 2D directional vector $\mathbf{d}_i \in \mathbb{R}^2$ pointing towards the $j$'s 2D projection, *i.e.*

$$\mathbf{d}_i = (\mathbf{l}_j - \mathbf{p}_i)/||\mathbf{l}_j - \mathbf{p}_i||_2, \quad (1)$$

where $\mathbf{d}_i$ is a normalized 2D vector denoting the landmark $j$'s direction.

After defining the ground-truth landmark segmentation and location voting maps, $\mathbf{S}$ and $\mathbf{d}$, we can supervise the proposed VS-Net to predict the two maps. After training, VS-Net can predict the two maps for a query image, from which we can establish accurate 2D-to-3D correspondences for achieving robust visual localization.

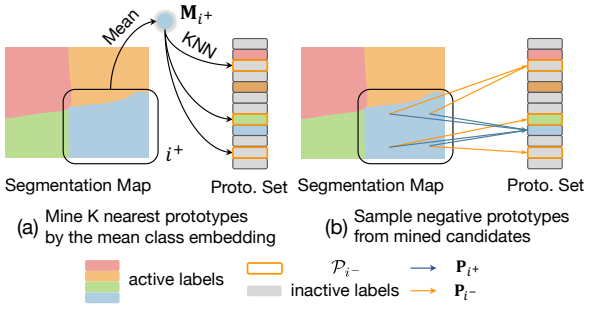



Figure 3: Prototype-based triplet loss. (a) Given a pixel $i$ with its mean class embeddings $\mathbf{M}_{i^+}$ calculated with pixels belonging to the same label $i^+$, we find its $k$ negative nearest neighbors in the prototype set. (b) We formulate pixel-wise prototype-based triplets where positive prototype is $\mathbf{P}'_{i^+}$ and the negative prototype is sampled from the mined $K$ neighbors $\mathcal{P}_{i^-}$.

### 3.2. Voting-with-Segmentation Network with Prototype-based Triplet Loss

The proposed Voting-with-Segmentation Network (VS-Net) consists of an image encoder to encode the image

into visual feature maps, a segmentation decoder to predict patch-based landmark segmentation map, and a voting decoder for generating the landmark voting map. The two maps are responsible for estimating landmarks' 2D locations as detailed below. In contrast with scene coordinate regression that relies on a neural network with a small receptive field to avoid overfitting, as discussed by Brachmann *et al*. [25], the prediction of the pixel-wise landmark labels and landmark directional votes can benefit from contextual information. We use DeepLabv3 [13], which enlarges the receptive field with atrous spatial pyramid pooling, as the backbone of our VS-Net.

**Patch-based landmark segmentation with prototype-based triplet loss.** Conventional semantic segmentation tasks generally adopt the cross-entropy loss to supervise complete categorical confidence vectors of all predicted pixels. However, our landmark segmentation requires ~~to~~ output segmentation maps with a large number of classes (landmarks) to effectively model each scene. A common scene in the 7Scenes dataset can consist of up to $5000$ landmarks. Simply supervising a $5000$-class segmentation map of size $640 \times 480$ with the cross-entropy loss requires $34.3$ GFLOPS and $5.7$ GB memory, which can easily drain computational resources of even modern GPUs.

To address this issue, we propose a novel prototype-based triplet segmentation loss with online hard negative sampling to supervise semantic segmentation with a large number of classes. It maintains and updates a set $\mathbf{P}$ of learnable class prototype embeddings, each of which is responsible for a semantic class, and $\mathbf{P}_j$ denotes the $j$th class's embedding. Intuitively, pixel embeddings of the $j$th class should be close to $\mathbf{P}_j$ and be far away from other classes' prototypes. Our proposed loss is designed based on the triplet loss with an online negative sampling scheme.

Specifically, given a pixel-wise embedding map $\mathbf{E}$ output by the segmentation branch of VS-Net and the class prototype set $\mathbf{P}$, the individual embeddings are first $L2$ normalized and are then optimized to minimize the following prototype-based triplet loss for each pixel $i$'s embedding $\mathbf{E}_i$,

$$\mathcal{L}_{\text{seg}} = \sum_{\text{all } i} \max(0, m + \text{sim}(\mathbf{E}_i, \mathbf{P}_{i^-}) - \text{sim}(\mathbf{E}_i, \mathbf{P}_{i^+})), \tag{2}$$

where $\text{sim}(a, b) = \frac{\boldsymbol{a}^T \boldsymbol{b}}{\|\boldsymbol{b}\| \cdot \|\boldsymbol{b}\|}$ measures the cosine similarity between a pixel embedding and a class prototype embedding, $m$ represents the margin of the triplet loss, $\mathbf{P}_{i^+}$ denotes the ground-truth (positive) class's prototype embedding corresponding to pixel $i$, and $\mathbf{P}_{i^-}$ denotes a sampled non-corresponding (negative) class prototype embedding of $i$ (to be discussed below).

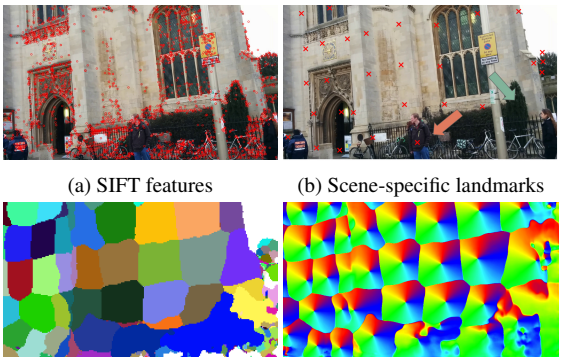
For each pixel $i$, how to determine its negative-class prototype embedding $\mathbf{P}_{i^-}$ in the above prototype-based triplet loss has crucial impacts on the final performance and randomly sampling negative classes would make the training over-simplified. Given an input image, we observe that the number of active landmarks (*i.e*. at least one pixel in the image belonging to the landmarks) is limited. In addition, pixels belonging to the same patch of a landmark are spatially close to each other and would share similar hard negative prototypes because they have similar embeddings. We, therefore, propose to mine representative negative classes for each active landmark, and each pixel randomly samples a negative class from the mined class set to form representative triplets.

Specifically, given a pixel $i$ with an active landmark (class) index $i^+$, we first retrieve all pixel embeddings corresponding to the landmark $i^+$ in the input image and take their average to obtain the landmark's mean class embedding $\mathbf{M}_{i^+}$ from the current image. The mean class embedding is then used to retrieve the $k$-nearest-neighbor negative prototypes $\mathcal{P}_{i^-}$ from the prototype embedding set. Such $k$NN negative prototypes can be considered as hard negative classes. The pixel $i$'s single negative prototype embedding $\mathbf{P}_{i^-}$ to be used in the triplet loss (Eq. (2)) is uniformly sampled from the $k$NN negative prototype set (Fig. 3).

The proposed prototype-based triplet loss is much more efficient than the conventional cross-entropy loss when used for supervising semantic segmentation as it only computes complete class scores for active landmarks rather than for all pixels. With an input image of size $640 \times 480$, the conventional cross-entropy loss costs 36.9 GFLOPS and 5.7GB memory. In contrast, if there are 100 active labels in an image, our proposed prototype-based triplet loss costs only 26.7 MFLOPS and $3.08$ MB memory, where the $k$NN hard negative search costs $12.0$ MFLOPS and $1.91$ MB memory, and the triplet loss itself only costs $14.7$ MFLOPS and $1.17$ MB memory (Tab. 1). OIM loss [57] is a popular loss for supervising large-scale classification problems. However, for each sample, it still needs to compute scores of belonging to all classes. As each image has a large number of pixels, it is still impractical to adopt the OIM loss in semantic segmentation.

**Pixel-wise voting for landmark location.** Given the segmentation map $\mathbf{S}$ generated from the above-introduced segmentation decoder, each pixel $i$ in the input image is either assigned a landmark label $\mathbf{S}_i$ or a noneffective label $0$ denoting too distant objects or regions (*e.g*., sky). We introduce another voting decoder for determining landmarks' projected 2D locations in the given image. The decoder outputs a directional voting map $\mathbf{d}$, where each pixel $i$ outputs a 2D directional vector $\mathbf{d}_i$ (Eq. (1)) pointing towards its corresponding landmark's 2D location (according to $\mathbf{S}_i$). The voting decoder is supervised with the following loss,

$$\mathcal{L}_{\text{vote}}(i) = \sum_{\text{all } i} \mathbf{1}(\mathbf{S}_i \neq 0) ||\hat{\mathbf{d}}_i - \mathbf{d}_i||_1, \tag{3}$$



(a) SIFT features (b) Scene-specific landmarks

(c) Landmark segmentation map (d) Landmark voting map

Figure 4: Comparison of (a) SIFT features and (b) the proposed scene-specific landmarks. (c-d) The scene-specific landmarks in (b) are obtained based on (c) landmark segmentation map and (d) landmark location voting map.

where $\mathbf{1}$ denotes the indicator function, and $\mathbf{d}_i$ and $\hat{\mathbf{d}}_i$ are ground-truth and predicted voting directions of pixel $i$.

**Overall loss function.** The overall loss $\mathcal{L}_{\text{overall}}$ is the combination of the patch-based landmark segmentation loss and landmark direction voting loss,

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{seg}}(i) + \lambda \mathcal{L}_{\text{vote}}(i), \tag{4}$$

where $\lambda$ weights the contributions of the loss terms.

**Localization with landmark segmentation and voting maps.** In the localization stage, pixels that are predicted to have the same landmark label in the landmark segmentation map are grouped together and we estimate its corresponding landmark location by computing the intersection of the landmark directional votes from the predicted voting map, which is dubbed the voting-by-segmentation algorithm.

Particularly, given the segmentation map, we first filter out landmark patches whose sizes are smaller than a threshold $T_s$ because too small landmark segments are generally unstable. The initial estimation of the 2D location $\hat{\mathbf{l}}_j$ of the landmark $j$ is computed from RANSAC with a vote intersection model [38], which generates multiple landmark location hypotheses by computing intersections of two randomly sampled directional votes and choosing the hypothesis having the most inlier votes. Then, the locations are further refined by an iterative EM-like algorithm. In the E-step, we collect inlier directional votes for the landmark $j$ from the surrounding circular region of the current $\hat{\mathbf{l}}_j^{(t)}$. In the M-step, we adopt the least-square method introduced by Antonio *et al*. [2] to compute the updated landmark location $\hat{\mathbf{l}}_j^{(t+1)}$ from the votes in the circular region. During the iterations, a voted landmark not supported by enough directional votes, indicating low voting consistency, would be dropped.

There are inevitable some disturbed pixels and some disordered regions caused by environmental noise or unfaithful surfaces. As shown in Fig. 4, the landmarks generated by our voting-by-segmentation algorithm achieve high accuracy and robustness against these distracting factors because we can accurately detect landmark locations by filtering disturbed pixel votes (pointed by the red arrow) and further reject unstable regions (pointed by the green arrow) in advance by checking the voting consistency. In contrast, cluttered SIFT features can easily result in erroneous matches, and the detected locations are easily disturbed on locally unstable regions, *e.g*. trees. Finally, all the estimated 2D landmarks in the query image naturally associate with the 3D landmarks in the scene, and the camera pose can be reliably estimated with standard RANSAC-PnP algorithm.

| | cross entropy | proto. triplet | |
|---|---|---|---|
| | Total | $k$NN | Triplet |
| Computation | 36.9 GF | 12.0 MF | 14.7 MF |
| Memory | 5.7 GB | 1.91 MB | 1.17 MB |

Table 1: Computation and memory cost comparison. GF and MF denotes GFLOPS and MFLOPS.

## 4. Experiments

In this section, we conduct a comparison with state-of-the-art methods and perform an ablation study to investigate individual components. Implementation details and extra results can be found in the supplementary materials.

### 4.1. Dataset

We evaluate our VS-Net on two standard visual localization benchmark datasets. (1) *Microsoft 7-Scenes Dataset* consists of seven static indoor scenes, which provides data recorded by a Kinect RGB-D sensor. The 3D surfaces, along with the camera poses, are computed from KinectFusion [33]. (2) *Cambridge Landmarks Dataset* contains six urban scenes. The images are collected by a smartphone and the camera poses are recovered from SfM. The *GreatCourt* and the *King'sCollege* are two challenging scenes that are affected by varying lighting conditions and dynamic objects. We reconstruct a dense 3D surface through multi-view stereo for each scene with given camera poses.

### 4.2. Comparison with State-of-the-arts

Previous visual localization systems that achieve good performance are SfM-based frameworks [16, 26, 61, 44] and scene coordinate regression frameworks [25, 8]. We compare VS-Net with these two frameworks on the 7-Scenes dataset and Cambridge Landmarks dataset. We do not present the results of neural pose regression [24, 23, 10, 21] because their pose accuracies are not competitive enough. There are representative SfM-based visual localization methods [45, 44]. Active Search [45] utilizes SIFT features, which is the state-of-the-art method using hand-crafted features with a priority-based matching algorithm.

| | SSL | SfM | | Scene Coordinate | | |
|---|---|---|---|---|---|---|
| | VS-Net (Ours) | AS [45] | HF-Net [44] | HSC-Net [25] | Reg [25] | DSAC++ [8] |
| Chess | **1.5**cm, **0.5°** | 4cm, 1.96° | 2.6cm, 0.9° | 2.1cm, 0.7° | 2.1cm, 1.0° | **1.5**cm, **0.5°** |
| Fire | **1.9**cm, **0.8°** | 3cm, 1.53° | 2.7cm, 1.0° | 2.2cm, 0.9° | 2.4cm, 0.9° | 2.0cm, 0.9° |
| Heads | **1.2**cm, **0.7°** | 2cm, 1.45° | 1.4cm, 0.9° | **1.2**cm, 0.9° | **1.2**cm, 0.8° | 1.3cm, 0.8° |
| Office | **2.1**cm, **0.6°** | 9cm, 3.61° | 4.3cm, 1.2° | 2.7cm, 0.8° | 3.1cm, 0.9° | 2.6cm, 0.7° |
| Pumpkin | **3.7**cm, **1.0°** | 8cm, 3.10° | 5.8cm, 1.6° | 4.0cm, **1.0°** | 4.3cm, 1.1° | 4.3cm, 1.1° |
| Kitchen | **3.6**cm, **1.1°** | 7cm, 3.37° | 5.3cm, 1.6° | 4.0cm, 1.8° | 4.5cm, 1.4° | 3.8cm, **1.1°** |
| Stairs | **2.8**cm, **0.8°** | 3cm, 2.22° | 7.2cm, 1.9° | 3.1cm, **0.8°** | 3.8cm, 0.9° | 9.1cm, 2.5° |
| Avg | **2.4**cm, **0.8°** | 5.1cm, 2.5° | 4.2cm, 1.3° | 2.7cm, 1.0° | 3.1cm, 1.0° | 3.5cm, 1.1° |
| GreatCourt | **0.22**m, **0.1°** | - | 0.76m, 0.3° | 0.28m, 0.2° | 1.25m, 0.6° | 0.40m, 0.2° |
| KingsCollege | **0.16**m, **0.2°** | 0.42m, 0.55° | 0.34m, 0.4° | 0.18m, 0.3° | 0.21m, 0.3° | 0.18m, 0.3° |
| OldHospital | **0.16**m, **0.3°** | 0.44m, 1.01° | 0.43m, 0.6° | 0.19m, **0.3°** | 0.21m, **0.3°** | 0.20m, **0.3°** |
| ShopFacade | **0.06**m, **0.3°** | 0.12m, 0.40° | 0.09m, 0.4° | **0.06**m, **0.3°** | **0.06**m, **0.3°** | **0.06**m, **0.3°** |
| St.MarysChurch | **0.08**m, **0.3°** | 0.19m, 0.54° | 0.16m, 0.5° | 0.09m, **0.3°** | 0.16m, 0.5° | 0.13m, 0.4° |
| Avg | **0.136**m, **0.24°** | - | 0.356m, 0.31° | 0.160m, 0.28° | 0.378m, 0.40° | 0.194m, 0.3° |

Table 2: Visual localization accuracy of state-of-the-art methods. We evaluate the localization performance by median positional error and angular error. The bar (-) means Active Search fails in the GreatCourt.

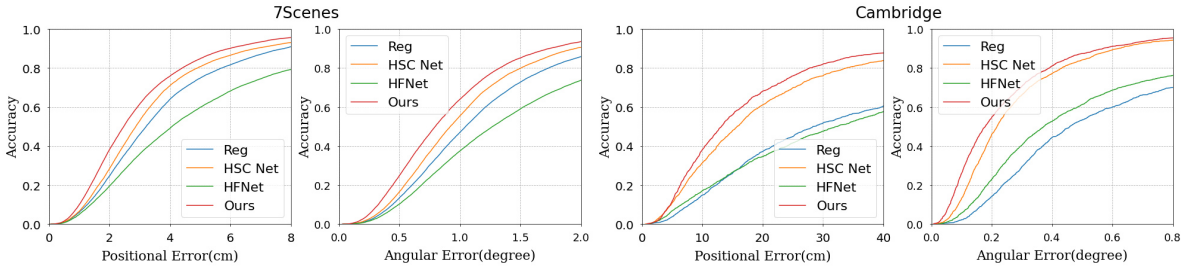



Figure 5: Cumulative pose error distribution of representative methods. For each dataset, we combine the poses of all scenes together, and count the ratio of poses under an increasing error threshold.

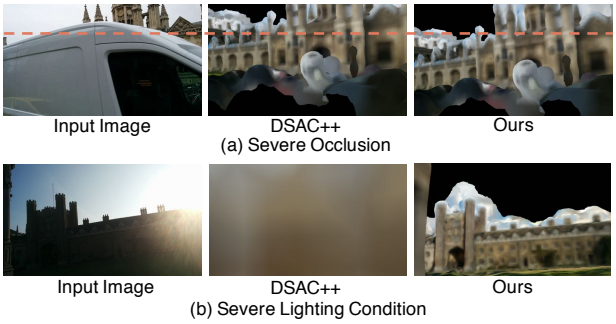



Figure 6: Localization in tough environments. We quantitatively compare VS-Net and DSAC++ by projecting the model into the original frames with estimated camera poses.

HF-Net [44] computes camera poses with NetVLAD [3] and SuperPoint [15], which are learned image features and local features, respectively. For scene coordinate regression methods, we select Reg [25], DSAC++ [8], and HSC-Net [25] for comparison. Reg [25] is a regression-only method, which directly regresses the scene coordinates from a query image. It is regarded as the baseline in scene coordinate regression methods. DSAC++ [8] designs a pose hypothesis selection algorithm based on Reg. HSC-Net [25] is a state-of-the-art scene coordinate regression method, which predicts hierarchical scene coordinates to improve the localization performance.

We compare the localization accuracy of VS-Net with the above mentioned methods. Positional error and angular error are the main metrics for evaluating pose accuracies. Table 2 presents the median of pose errors in each individual scene. SfM-based methods, including Active Search and HF-Net, produce mediocre results because the general-purpose features are not accurate enough. VS-Net achieves better performances in all scenes. Even compared with improved scene coordinate regression methods (HSC-Net and DSAC++), VS-Net still outperforms them in most scenes. Fig. 5 shows the cumulative distributions of overall pose error across scenes, which illustrate that the holistic performance of VS-Net is better than the others. Moreover, VS-Net is able to obtain high-quality poses even though running into challenging cases (Fig. 6) while DSAC++, an improved scene coordinate regression method, fails.

| | Cross Entropy (FLOPS/Bytes) | Proto. Triplet (Avg FLOPS/Bytes) | | |
|---|---|---|---|---|
| | | A.L. | w/o $k$NN | w/ $k$NN |
| Chess | n/a (30G/5G) | 470 | 0.45 (0.01G/1M) | 0.80 (0.06G/9M) |
| Fire | n/a (35G/6G) | 662 | 0.34 (0.01G/1M) | 0.69 (0.10G/14M) |
| Heads | n/a (69G/12G) | 826 | 0.25 (0.01G/1M) | 0.58 (0.21G/32M) |
| Office | n/a (34G/6G) | 409 | 0.4 (0.01G/1M) | 0.75 (0.06G/9M) |
| Pumpkin | n/a (32G/5G) | 519 | 0.43 (0.01G/1M) | 0.61 (0.07G/10M) |
| Kitchen | n/a (44G/7G) | 496 | 0.26 (0.01G/1M) | 0.58 (0.09G/13M) |
| Stairs | n/a (95G/16G) | 224 | 0.45 (0.01G/1M) | 0.68 (0.09G/13M) |

Table 3: Segmentation accuracy on *7-Scenes* by our VS-Net with different segmentation losses. n/a denotes that cross entropy loss alone already occupies too much memory even for a single image and cannot be used in practice. A.L. denotes the average active labels of images in the scene.

| Size | 1.50m | **1.75m** | 2.00m | 2.25m | 2.50m | 2.75m | 3.00m |
|---|---|---|---|---|---|---|---|
| Num. | 7418 | **5333** | 4089 | 3278 | 2603 | 2099 | 1804 |
| Pos. | 16cm | **15cm** | 16cm | 15cm | 17cm | 17cm | 18cm |
| Ang. | 0.3° | **0.2°** | 0.2° | 0.3° | 0.3° | 0.3° | 0.3° |
| Size | 10.0cm | 12.5cm | **15.0cm** | 17.5cm | 20.0cm | 25.0cm | 30.0cm |
| Num. | 10918 | 6501 | **4330** | 3013 | 2280 | 1409 | 925 |
| Pos. | 1.54cm | 1.52cm | **1.52cm** | 1.52cm | 1.57cm | 1.58cm | 1.65cm |
| Ang. | 0.54° | 0.53° | **0.50°** | 0.50° | 0.54° | 0.54° | 0.55° |

Table 4: Localization accuracy in *King's College* (above) and *chess* (below) with different patch sizes.

### 4.3. Ablation study

**Scene-specific landmarks vs. scene coordinates.** We propose a new 2D-to-3D correspondence representation, the scene-specific landmarks, to replace the pixel-wise scene coordinates in deep learning based methods [25, 8]. To compare these two representations, we remove the vote decoder in our VS-Net and directly regress pixel-wise scene coordinates with our segmentation decoder, which is similar to the Reg [25], and keep other settings the same. Its average median errors of camera poses in the Microsoft 7-Scenes dataset and Cambridge Landmark dataset are 36.5cm/16° and 99cm/1.7°, while our scene-specific landmarks achieves 2.4cm/0.8° and 14cm/0.24°. It is also far worse than Reg because the large receptive field of our VS-Net impacts scene coordinate regression.

**Localization accuracy with different patch sizes.** Our landmarks are generated through 3D surface over-segmentation, where patch size is a hyper-parameter that determines the number and distribution of landmarks in VS-Net. To explore the influence of different patch sizes, we train and evaluate VS-Net with the landmark segmentation map and pixel-wise voting map predicted with different patch sizes. We show the corresponding landmark numbers and median pose errors of different patch sizes in Table 4. The best choice of patch size for *King's College* and *chess* are 1.75m and 15cm, respectively. Both a smaller size and a larger size would deteriorate the localization accuracy but do not severely impact it, which indicates that VS-Net is not very sensitive to the patch size.

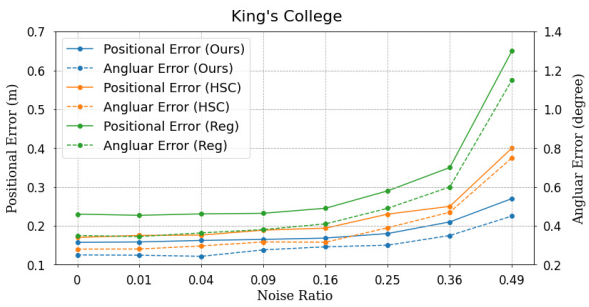



Figure 7: Localization pose errors of different methods when a noisy occlusion block of different sizes (noise ratios) is introduced into query images.

**Localization with challenging occlusions.** To evaluate the robustness of visual localization systems against environmental noise, we use a noisy occlusion block that contains 3-channel random noise ranging from 0 to 255 to cover a portion of the query image randomly. The height and width of the noise blocks are set as 10%-70% of the image, which correspond to 1%-49% noise ratio and can indicate different levels of noise interference. As shown in Fig. 7, we compare the median pose error of our VS-Net with those of HSC-Net [25] and Reg [25] under different noise ratios. VS-Net consistently results in lower pose errors.

**Landmark segmentation with prototype-based triplet loss.** To address the problem of the too large number of landmarks in our landmark segmentation sub-task, we propose the prototype-based triplet loss for our VS-Net. We conduct an experiment in Table 3 to compare segmentation loss functions on the 7-Scenes dataset. The conventional cross entropy loss does not work for VS-Net because it consumes much computation and memory. We present its theoretical computational and memory costs for a single image of size $640 \times 480$ in the braces. We also test using the prototype-based triplet loss for landmark segmentation but without the $k$NN negative mining, which is able to train VS-Net but achieves inferior performances. Our complete prototype-based triplet has low latency and computation complexity while maintaining superior performance.

## 5. Conclusion

In this paper, we have proposed a novel visual localization framework that represents the map by patches and landmarks, and design a neural network VS-Net to detect the scene-specific landmarks on images. The experiments on the public datasets demonstrate the effectiveness of the proposed framework. Utilizing hierarchical spatial structure and temporal information has been proved beneficial in both SfM-based methods and scene coordinate regression. Exploring how to improve scene-specific landmarks with these strategies will be the direction of our future work.

## References


[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.

[2] Franklin Antonio. Faster line segment intersection. In *Graphics Gems III (IBM Version)*, pages 199–202. Elsevier, 1992.

[3] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[4] Clemens Arth, Daniel Wagner, Manfred Klopschitz, Arnold Irschara, and Dieter Schmalstieg. Wide area localization on mobile phones. In *2009 8th ieee international symposium on mixed and augmented reality*, pages 73–82. IEEE, 2009.

[5] Nicolas Aziere and Sinisa Todorovic. Ensemble deep manifold similarity learning using hard proxies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7299–7307, 2019.

[6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *Proceedings of the European conference on computer vision*, pages 404–417. Springer, 2006.

[7] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6684–6692, 2017.

[8] Eric Brachmann and Carsten Rother. Learning less is more-6d camera localization via 3d surface regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4654–4662, 2018.

[9] Eric Brachmann and Carsten Rother. Expert sample consensus applied to camera re-localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7525–7534, 2019.

[10] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2018.

[11] Ignas Budvytis, Marvin Teichmann, Tomas Vojir, and Roberto Cipolla. Large scale joint semantic re-localisation and scene understanding via globally unique instance coordinate regression. *arXiv preprint arXiv:1909.10239*, 2019.

[12] Federico Camposeco, Andrea Cohen, Marc Pollefeys, and Torsten Sattler. Hybrid scene compression for visual localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7653–7662, 2019.

[13] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[14] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[15] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018.

[16] Michael Donoser and Dieter Schmalstieg. Discriminative feature-to-point matching in image-based localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 516–523, 2014.

[17] Mihai Dusmanu, Ignacio Rocco, Tomás Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable CNN for joint description and detection of local features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8092–8101, 2019.

[18] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[19] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. *arXiv preprint arXiv:2006.03926*, 2020.

[20] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11632–11641, 2020.

[21] Zhaoyang Huang, Yan Xu, Jianping Shi, Xiaowei Zhou, Hujun Bao, and Guofeng Zhang. Prior guided dropout for robust visual localization in dynamic environments. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2791–2800, 2019.

[22] Marco Imperoli and Alberto Pretto. Active detection and localization of textureless objects in cluttered environments. *arXiv preprint arXiv:1603.07022*, 2016.

[23] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5974–5983, 2017.

[24] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015.

[25] Xiaotian Li, Shuzhe Wang, Yi Zhao, Jakob Verbeek, and Juho Kannala. Hierarchical scene coordinate classification and regression for visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11983–11992, 2020.

[26] Yunpeng Li, Noah Snavely, and Daniel P Huttenlocher. Location recognition using prioritized feature matching. In *European conference on computer vision*, pages 791–804. Springer, 2010.

[27] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. Unsupervised person re-identification via softened

similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3390–3399, 2020.

[28] Yuan Liu, Zehong Shen, Zhixuan Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou. Gift: Learning transformation-invariant dense visual descriptors via group cnns. In *Advances in Neural Information Processing Systems*, pages 6990–7001, 2019.

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[30] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[31] Jean-Michel Morel and Guoshen Yu. Asift: A new framework for fully affine invariant image comparison. *SIAM journal on imaging sciences*, 2(2):438–469, 2009.

[32] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 360–368, 2017.

[33] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, volume 11, pages 127–136, 2011.

[34] Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Making deep heatmaps robust to partial occlusions for 3d object pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–134, 2018.

[35] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: learning local features from images. In *Advances in neural information processing systems*, pages 6234–6244, 2018.

[36] Jeremie Papon, Alexey Abramov, Markus Schoeler, and Florentin Wörgötter. Voxel cloud connectivity segmentation - supervoxels for point clouds. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, Portland, Oregon, June 22-27 2013.

[37] Georgios Pavlakos, Xiaowei Zhou, Aaron Chan, Konstantinos G Derpanis, and Kostas Daniilidis. 6-dof object pose from semantic keypoints. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 2011–2018. IEEE, 2017.

[38] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019.

[39] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. Softtriple loss: Deep metric learning without triplet sampling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6450–6458, 2019.

[40] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018.

[41] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. In *Advances in Neural Information Processing Systems*, pages 12405–12415, 2019.

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[43] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *Proceedings of the IEEE international conference on Computer Vision (ICCV)*, pages 2564–2571. IEEE, 2011.

[44] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019.

[45] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Improving image-based localization by active correspondence search. In *European conference on computer vision*, pages 752–765. Springer, 2012.

[46] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

[47] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[48] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[49] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2937, 2013.

[50] Chen Song, Jiaru Song, and Qixing Huang. Hybridpose: 6d object pose estimation under hybrid representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 431–440, 2020.

[51] Julien Valentin, Matthias Nießner, Jamie Shotton, Andrew Fitzgibbon, Shahram Izadi, and Philip HS Torr. Exploiting uncertainty in regression forests for accurate camera relocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4400–4408, 2015.

[52] Bing Wang, Changhao Chen, Chris Xiaoxuan Lu, Peijun Zhao, Niki Trigoni, and Andrew Markham. Atloc: Attention guided camera localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10393–10401, 2020.

[53] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. *arXiv preprint arXiv:2004.13324*, 2020.

[54] Philippe Weinzaepfel, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. Visual localization by learning objects-of-interest dense match regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5634–5643, 2019.

[55] Changchang Wu et al. Visualsfm: A visual structure from motion system. 2011.

[56] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2848, 2017.

[57] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017.

[58] Yan Xu, Zhaoyang Huang, Kwan-Yee Lin, Xinge Zhu, Jianping Shi, Hujun Bao, Guofeng Zhang, and Hongsheng Li. Selfvoxelo: Self-supervised lidar odometry with voxel-based deep neural networks. *Conference on Robot Learning*, 2020.

[59] Fei Xue, Xin Wu, Shaojun Cai, and Junqiu Wang. Learning multi-view camera relocalization with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11372–11381. IEEE, 2020.

[60] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[61] Bernhard Zeisl, Torsten Sattler, and Marc Pollefeys. Camera pose voting for large-scale image-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2704–2712, 2015.

[62] Guofeng Zhang, Zilong Dong, Jiaya Jia, Tien-Tsin Wong, and Hujun Bao. Efficient non-consecutive feature tracking for structure-from-motion. In *European Conference on Computer Vision*, pages 422–435. Springer, 2010.

[63] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose-invariant embedding for deep person re-identification. *IEEE Transactions on Image Processing*, 28(9):4500–4509, 2019.

[64] Zilong Zhong, Zhong Qiu Lin, Rene Bidart, Xiaodan Hu, Ibrahim Ben Daya, Zhifeng Li, Wei-Shi Zheng, Jonathan Li, and Alexander Wong. Squeeze-and-attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13065–13074, 2020.

[65] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[66] Lei Zhou, Zixin Luo, Tianwei Shen, Jiahui Zhang, Mingmin Zhen, Yao Yao, Tian Fang, and Long Quan. Kfnet: Learning temporal camera relocalization using kalman filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4919–4928, 2020.

[67] Siyu Zhu, Tianwei Shen, Lei Zhou, Runze Zhang, Jinglu Wang, Tian Fang, and Long Quan. Parallel structure from motion from local increment to global averaging. *arXiv preprint arXiv:1702.08601*, 2017.