Understanding Matters: Semantic-Structural Determined Visual Relocalization for Large Scenes

Jingyi Nie[1] , **Liangliang Cai**[1] , **Qichuan Geng**[2, *] and **Zhong Zhou**[1,3]
[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China
[2]The Information Engineering College, Capital Normal University, Beijing, China
[3]Zhongguancun Laboratory, Beijing, China
{njy_zy2306440, by1906001}@buaa.edu.cn, gengqichuan1989@cnu.edu.cn, zz@buaa.edu.cn

Abstract

Scene Coordinate Regression (SCR) estimates 3D scene coordinates from 2D images, and has become an important approach in visual relocalization. Existing methods exhibit high localization accuracy in small scenes, but still face substantial challenges in large-scale scenes, which usually have significant variations in depth, scale, and occlusion. Although structure-guided scene partitioning is commonly adopted, the over-partitioned elements and large feature variances within subscenes impede the estimation of the 3D coordinates, introducing misleading information for subsequent processing. To address the above-mentioned issues, we propose the Semantic-Structural Determined Visual Relocalization method for SCR, which leverages semantic-structural partition learning and partition-determined pose refinement to better understand the semantic and structural information on large scenes. Firstly, we partition the scene into small subscenes with label assignments, ensuring semantic consistency and structural continuity within each subscene. A classifier is then trained with sampling-based learning to predict these labels. Secondly, the partition predictions are encoded into embeddings and integrated with local features for intra-class compactness and inter-class separation, producing partition-aware features. To further decrease feature variances, we employ a discriminability metric and suppress ambiguous points, improving subsequent computations. Experimental results on the Cambridge Landmarks dataset demonstrate that the proposed method achieves significant improvements with fewer training costs on large-scale scenes, reducing the median error by 38% compared to the state-of-the-art SCR method DSAC*. Code is available: https://gitee.com/VR_NAVE/ss-dvr.

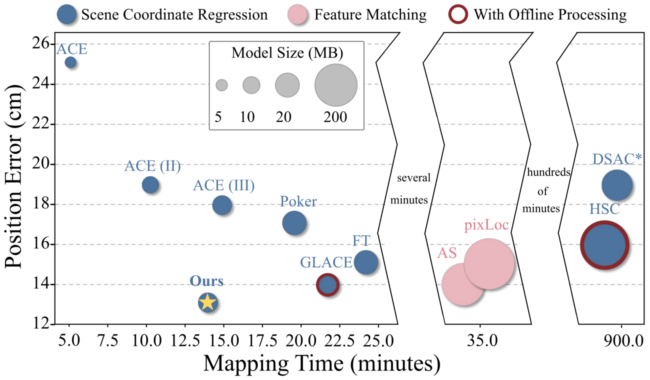

Figure 1: Quantitative comparison of position error and mapping time. We evaluate the state-of-the-art SCR and feature matching methods on the large-scale outdoor dataset Cambridge Landmarks. The position error and training time of these methods are compared. The area of the circles represents the model size. SCR-based methods are marked in blue, and FM-based methods are marked in pink. Methods based on offline processing are marked with red borders.

1 Introduction

Visual relocalization aims to determine the position and orientation of the camera within a known scene by visual cues. This task holds significant promise in various fields, such as robot navigation, augmented reality, and autonomous driving. In particular, image-based Scene Coordinate Regression (SCR) encodes the scene in the weight of neural networks and supports end-to-end inference, showing great potential for a wide range of applications.

With the same motivation, image retrieval [Camposeco *et al.*, 2019a; Revaud *et al.*, 2019] and pose regression [Brahmbhatt *et al.*, 2018; Kendall and Cipolla, 2017] take the entire image as input and extract image-level features for matching [Ruan *et al.*, 2023]. However, global features often lack local structural details, leading to insufficient relocalization accuracy. In contrast, methods based on feature matching [Sarlin *et al.*, 2019; Sarlin *et al.*, 2020a; Sattler *et al.*, 2016a], which build explicit 3D geometry using structure-from-motion (SfM) [Schonberger and Frahm, 2016], achieve excellent performance. However, they require significant storage and extended processing time, ranging from several minutes to hours. Furthermore, they introduce privacy risks. With recent advances in deep learning, SCR

[*] Corresponding author.

researches [Brachmann *et al.*, 2023; Shotton *et al.*, 2013; Wang *et al.*, 2024] employ models to regress the 3D scene coordinates corresponding to each pixel in an image, subsequently estimating the camera pose using the Perspective-n-Point (PnP) algorithm. Consequently, only the raw query image is required during inference, eliminating the need to access previously trained data.

Due to constrained receptive fields, SCR methods struggle to differentiate similar local features in large-scale scenes, leading to decreased performance. Recent SCR methods focus on improving performance in such scenarios. ACE [Brachmann *et al.*, 2023] trains four models (Poker) for partitioned large-scale scenes. HSC [Li *et al.*, 2020] partitions the scene into smaller parts with hierarchical labels based on structural information, guiding the coarse-to-fine relocalization process. GLACE [Wang *et al.*, 2024] encodes a single image into a global feature to distinguish similar features across different images. However, the methods above, which primarily focus on structural information, partition large scenes into small subscenes solely based on spatial distance, without considering scene understanding. This results in over-partitioned elements and high feature variances within subscenes. Semantic information plays a crucial role in scene understanding, guiding more effective scene partitioning.

To address the aforementioned challenges, we propose a Semantic-Structural Determined Visual Relocalization method. Specifically, semantic-structural scene partition learning is applied to divide the large scene into smaller subscenes, maintaining semantic consistency and structural continuity within each subscene through label assignments. Then, accelerated partition localization is performed with sampling-based learning. To better leverage the semantic and structural information in the labels and to avoid errors in relocalization caused by ambiguous areas, e.g. repetitive textures and flat regions, we put forward partition-determined pose refinement. Partition labels are encoded into embeddings and integrated with local features next, creating partition-aware features, and guiding the further discriminable point selection. Our contribution can be summarized as follows:

i) We propose a Semantic-Structural Determined Visual Relocalization method for scene coordinate regression. It partitions the scene into small parts with labels, ensuring semantic consistency and structural continuity, guiding further accelerated partition localization with sampling-based learning.

ii) We design a partition-determined pose refinement method to reduce the inaccuracy of pose estimation, which selects determinable points based on discriminability scores derived from partition-aware features.

iii) Our method achieves the state-of-the-art performance on Cambridge Landmark in a short training time when depth maps are provided, and can still achieve the competitive performance without depth maps.

2 Related Work

Typical methods of visual relocalization can be divided into four categories, e.g. image retrieval, pose regression, feature matching, and scene coordinate regression.

Image Retrieval. A set of mapping images with known poses and global descriptors serves as the known environment in retrieval-based methods. Given a query image, these methods search for the most similar image in the database using global descriptor matching [Arandjelovic *et al.*, 2016; Revaud *et al.*, 2019; Torii *et al.*, 2015], and approximate the pose of the query image based on the top retrieved images [Camposeco *et al.*, 2019a]. Since image-level descriptors are used, retrieval-based methods can scale to large scenes. Recently, some researches have combined retrieval methods with structure-based techniques and relative pose estimation.

Pose Regression. Instead of descriptor matching, pose regression [Brahmbhatt *et al.*, 2018; Kendall and Cipolla, 2017; Kendall *et al.*, 2015; Shavit *et al.*, 2021; Revaud *et al.*, 2019; Winkelbauer *et al.*, 2021] uses neural networks to predict the absolute pose of the query image directly, namely absolute pose regression (APR). Some methods predict the relative pose between the query image and a mapping image. PoseNet [Kendall *et al.*, 2015] regresses the pose from a query image with a neural network, but it takes several hours to train the model. Several variants of it have introduced improvements, ranging from the loss function [Kendall and Cipolla, 2017] to the network architecture [Walch *et al.*, 2017; Wang *et al.*, 2020; Shavit *et al.*, 2021]. However, recent researches indicate that pose regression methods are more similar to image retrieval than to feature matching, resulting in their performance being surpassed by feature matching.

Feature Matching (FM). Methods based on feature matching [Sarlin *et al.*, 2019; Sarlin *et al.*, 2020a; Sattler *et al.*, 2016a], also known as structure-based, perform well for visual relocalization. They establish 2D-3D correspondences between pixels in query images and 3D coordinates in the scene, usually employing descriptor matching. Thus, they need to reconstruct the known environment as a 3D point cloud through Structure-from-Motion (SfM), so that points can have several feature descriptors from different viewpoints. To scale to large scenes, commonly they need a large amount of storage and a long time for feature extraction. Some recent researches work on handling the storage problems. For instance, techniques such as storing fewer descriptors [Sattler *et al.*, 2016b], compressing descriptors [Yang *et al.*, 2022], and combining image retrieval with descriptor matching [Arandjelovic *et al.*, 2016; Torii *et al.*, 2015] are commonly used. Several approaches try to directly match the descriptors with the point cloud or mesh [Zhou *et al.*, 2022a]. Given all of the above, large storage requirements and long descriptor generation time are unavoidable.

Scene Coordinate Regression. This family of methods regresses 3D coordinates for the 2D pixels in the query image instead of learning the entire pipeline. These methods do not rely on traditional feature detection and descriptor matching. Instead, they implicitly encode the scene information within a neural network [Brachmann *et al.*, 2023; Brachmann *et al.*, 2017; Brachmann and Rother, 2018; Brachmann and Rother, 2019; Brachmann and Rother, 2021; Cavallari *et al.*, 2019a; Dong *et al.*, 2022; Li *et al.*, 2020] or random forests [Brachmann *et al.*, 2016; Cavallari *et al.*, 2017; Cavallari *et al.*, 2019b; Shotton *et al.*, 2013]. By doing so, they leverage learned representations of the environment to predict the 3D coordinates directly from the in-

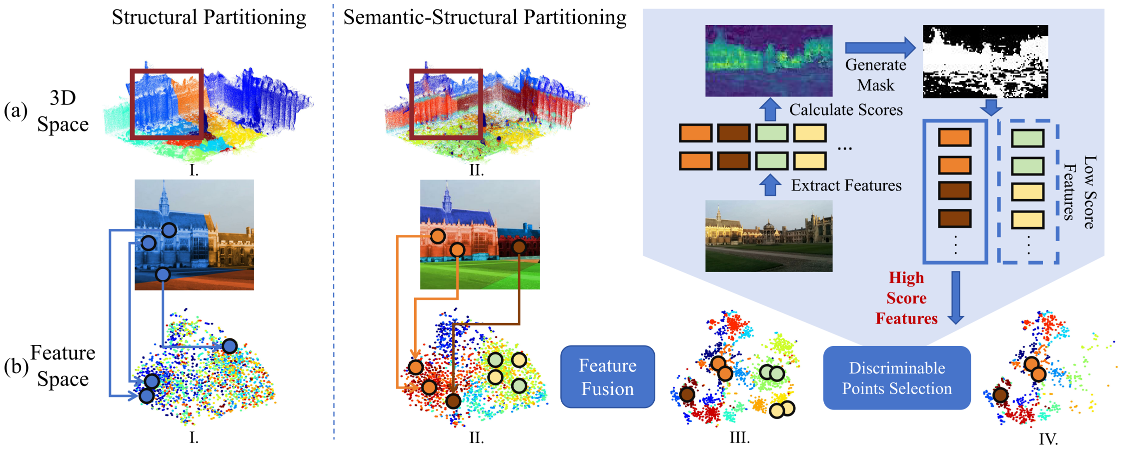

Figure 2: Overview of our method. Left: The structural partitioning strategy. Features with unified labels exhibit a large variance in the feature space. Right: Our semantic-structural partitioning approach partitions the scene based on semantic information while ensuring structural continuity within each subscene. We use t-SNE to map the features from the high-dimensional space to a 2D space to observe the relationships between the features. Features labeled by structural partitioning are scattered and disordered, while features labeled by semantic-structural partitioning form more compact and consistent clusters. We design a feature fusion module to keep intra-class compactness and inter-class separation, and a discriminability judge module to choose features with high scores for the camera pose estimation.

put image, without the need for explicit feature matching or descriptor-based techniques. These approaches can be considered privacy-preserving, as the implicit map cannot be regenerated without access to the training images of the scene. However, although SCR-based methods achieve high accuracy in small-scale scenes, their performance tends to degrade in large-scale environments. Existing methods, such as DSAC* [Brachmann and Rother, 2021], achieve state-of-the-art accuracy but require several hours of model training. Recently, ACE [Brachmann *et al.*, 2023] proposed a model that achieved high accuracy in a short time of training. However, its performance deteriorates in large-scale environments, requiring the assembly of more separate models for effective learning of such scenes.

3 Method

Our Semantic-Structural Determined Visual Relocalization method comprises two components. Semantic-Structural Partition Learning partitions the large scene into smaller subscenes, with labels assigned to each. This is followed by accelerated partition localization with sampling-based learning. The partition-determined pose refinement enhances features with partition labels and selects discriminable points.

3.1 Semantic-Structural Partition Learning

In this section, the scene is partitioned into small parts preserving semantic consistency and structural continuity, which are assigned labels. Subsequently, a label classifier is trained on the sampled feature set.

Semantic-Structural Label Assignment. Recent works have shown that in large scenes with many similar local features, it is difficult for SCR to form an effective mapping from features to 3D coordinates. Previous studies, such as HSC [Li *et al.*, 2020], in Fig. 2 (a-I), employ scene partitioning strategies that disrupt the semantic information of the scene, leading to over-partitioned elements and large feature variances in each subscene. Apparently, a scene partitioning strategy that groups features with similar semantic information into the same category can maintain semantic consistency within the feature set. We propose our scene partitioning and point labeling strategy next. Semantic features are extracted from the images with the backbone.

$$f_i = \mathcal{F}_B(\mathcal{P}_i; w_B), \tag{1}$$

where $\mathcal{P}_i \in \mathbb{R}^{C_I \times H_\mathcal{P} \times W_\mathcal{P}}$ is a patch of image, and $\mathcal{F}_B$ is the backbone that extracts semantic features $f_i \in \mathbb{R}^{C_f}$ from image patch $\mathcal{P}_i$ with network parameters w_B.

To ensure semantic consistency, we perform semantic-level clustering of the features. Large-scale scenes often contain many similar local patches. The limited ability of the network to differentiate features from these patches constraints its performance. Since these patches may be far apart in the scene, incorrect matching of similar patches can lead to significant errors, emphasizing the necessity of structural continuity. We partition the large scene into smaller subscenes, ensuring both semantic consistency and structural continuity within each subscene. We take the semantic feature f_i and the corresponding 3D coordinates y_i of the point as input and output a label vector that determines the cluster assignment.

Recent work [Li *et al.*, 2020] proposes a hierarchical label assignment approach. During the inference stage, although the subscene label is correctly predicted, the predicted region label has a decisive impact on the final result. Our strategy assigns two labels to each point simultaneously, ensuring the independence of the two partitions. This decoupling approach reduces the dependence on a single label and enhances the robustness of the results. The process is shown as follows:

$$(c_1, c_2)_i = \arg\min_{(k_1,k_2)} \|(f_i, y_i) - (\mu_1(k_1), \mu_2(k_2))\|^2, \tag{2}$$

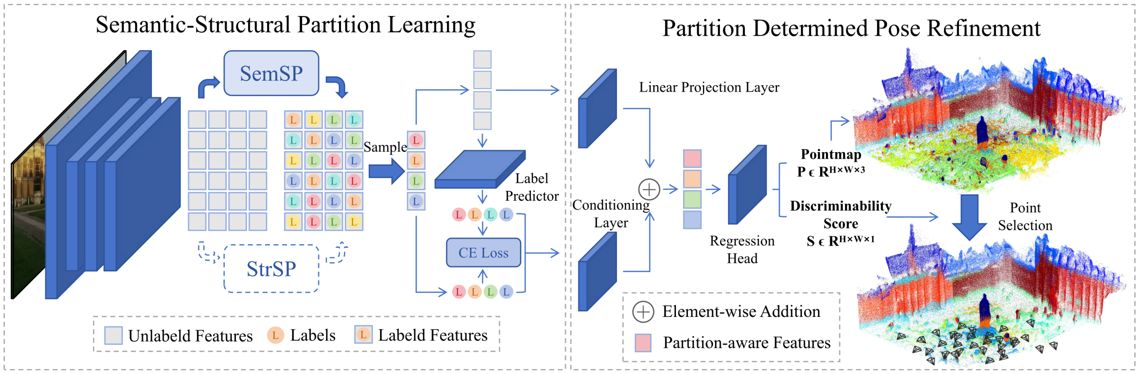

Figure 3: The architecture of our network. $SemSP$ represents semantic scene partitioning. $StrSP$ represents structural scene partitioning. When depth maps are unavailable, scene partitioning can be performed using only semantic scene partitioning.

where y_i is the 3D coordinates corresponding to the feature, $k_1 \in \{1, \ldots, n_{k_1}\}$ represents the index of semantic clustering, $k_2 \in \{1, \ldots, n_{k_2}\}$ represents the index of structural clustering, $\mu_1(k_1)$ is the k_1-th semantic cluster center and $\mu_2(k_2)$ is the k_2-th structural cluster center, and $(c_1, c_2)_i$ represents a 2-dimensional label vector.

Standard k-means algorithms require significant time to cluster high-dimensional features. We use mini-batch k-means for semantic consistency partitioning on features and standard k-means for structural continuity partitioning on coordinates. Here is the process of mini-batch k-means:

$$\mu_1^{\text{new}}(k_1) = \mu_1^{\text{old}}(k_1) + \frac{1}{\beta} \sum_{x_i \in B_k} \left(f_i - \mu_1^{\text{old}}(k1) \right), \quad (3)$$

$$c_1(i) = \arg\min_{k_1} \|f_i - \mu_1(k_1)\|^2 \quad \text{for } f_i \in f_B, \quad (4)$$

where f_B is a batch of semantic features. We randomly select n_{k_1} initial cluster centers $\mu_1(1), \mu_1(2), \ldots, \mu_1(n_{k_1})$. Then we iteratively perform a random selection of small subsets to assign each feature to the nearest feature centroid and update the centroids based on the average of the features in the mini-batch. This process repeats for several iterations or until the centroids stabilize. The mini-batch k-means uses only a mini-batch of data to incrementally update the centroids, while the centroids of standard k-means are recalculated by taking the mean of all data in the cluster.

$$\mu_2^{new}(k_2) = \frac{1}{|P_{str}(k_2)|} \sum_{y_i \in P_{str}(k_2)} y_i, \quad (5)$$

where $P_{str}(k_2) = \{y_i | c_2(i) = k_2\}$ denotes the 3D coordinate set of points with the structural label k_2. For the category with semantic label k_1 and structural label k_2, we compute the average of the 3D coordinates.

Accelerated Partition Localization with Sampling-based Learning. The features surrounding a point are more likely to belong to the same category as the point itself, both at the structural and semantic levels. Therefore, it is possible to predict the labels of other local points on the basis of the label of a given point. Furthermore, partitioning all points in the scene incurs significant time overhead. In HSC [Li *et al.*, 2020], labels are assigned to every pixel in all images offline, resulting in poor transferability and high computational cost. Inspired by [Brachmann *et al.*, 2023], we perform a sampling operation on the features of an image.

$$F = \bigcup_{i=1}^{N} S(\mathcal{B}(I)), \quad (6)$$

where N is the number of training images, S represents the sampling process, and $\mathcal{B}$ is the feature extractor.

In addition, we train a small classifier on the sampled feature set F to predict the label of a feature:

$$\hat{o}_i = \mathcal{F}_L(f_i; w_L), \text{ and } f_i \in F, \quad (7)$$

where $\mathcal{F}_L$ is a small MLP head. And it is trained with the cross-entropy loss:

$$\mathcal{L}_c = -\sum_i (o_i)^\top log\hat{o}_i, \quad (8)$$

where o_i denotes the one-hot label of pixel i. $\hat{o}_i$ denotes the corresponding label probabilities.

3.2 Partition-determined Pose Refinement

In this section, we present the process of pose refinement, which includes partition-aware feature enhancement and discriminable point selection for pose refinement.

Partition-aware Feature Enhancement. The quality of semantic features significantly affects the accuracy of SCR models [Nguyen *et al.*, 2024]. Although we assign semantic and structural labels to each feature in Equation 2, the clusters formed of semantic features with the same label are still not compact enough, as shown in Fig. 2 (b-II). As a result, semantic features from different categories remain relatively close in the feature space. The overlap reduces their separability, thereby affecting discriminability. To mitigate this, we introduce a feature fusion module that encodes labels into structure-semantic partition-aware embeddings and fuse them with the semantic features f_i extracted from the image, to promote intra-class compactness and inter-class separation:

$$\hat{f}_i = \phi(f_i, y_i), \quad (9)$$

		Mapping Time	Mapping Size	7 Scenes (D-SLAM poses)	12 Scenes (D-SLAM poses)
FM	AS (SIFT) [2016a]	~1.5h	~200MB	68.7%	99.6%
	D.VLAD+R2D2 [2020]		~1GB	77.6%	99.7%
	hLoc (SP+SG) [2019]		~2GB	76.8%	99.8%
	pixLoc [2021]		~1GB	75.7%	N/A
SCR (w/ Depth)	DSAC* (Full) [2021]	15h	28MB	**84.0%**	99.2%
	DSAC* (Tiny) [2021]	11h	4MB	70.0%	83.1%
	SANet [2019]	~2.3 min	~550MB	68.2%	N/A
	SRC [2022]	2 min†	40MB	55.2%	N/A
SCR	DSAC* (Full) [2021]	15h	28MB	81.1%	98.8%
	DSAC* (Tiny) [2021]	11h	4MB	69.1%	81.6%
	GLACE [2024]	23 min	9MB	81.4%	99.6%
	ACE [2023]	4 min	4MB	80.8%	99.6%
	Ours w/o depth	4 min	9MB	81.3%	**99.8%**
	Ours w/ depth	13 min	14MB	82.5%	**99.8%**

Table 1: **7Scenes and 12Scenes Results.** We report the percentage of frames below a 5cm, 5° pose error. Best results in **bold** for the "SCR" group, second best results underlined. We list the time and size needed for mapping.

where ϕ represents the fusion function and $\hat{f}_i$ represents the feature f_i fused with the embedding. Specifically, ϕ is a small network module composed of MLP layers.

Discriminable Points Selection for Pose Refinement. Some fused features extracted from repetitive textures or flat regions still struggle to be mapped. We present a pixel-wise discriminability score $\hat{s}_i$, which is generated by our prediction head, along with the predicted 3D coordinate $\hat{y}_i$:

$$\hat{y}_i, \hat{s}_i = \mathcal{L}_H(\hat{f}_i; w_H), \tag{10}$$

where $\mathcal{L}_H$ is a regression head. We define $s_i = 1 + e^{\hat{s}}$ to ensure that it is always greater than 1. The fused feature with a higher discriminability score is considered to have effective scene information, leading to an accurate prediction result. Due to the complexity of large-scale scenes, we find that using s_i to represent the discriminability of the mapping from features to 3D coordinates results in a large overall error. Therefore, we combine s_i with the reprojection error $\hat{\ell}$ to ensure that s_i represents the discriminability of a point along the line connecting the imaging point and the actual pixel coordinates of that point:

$$\hat{\ell} = ||x_i - Kh^{-1}\hat{y}_i||_1, \tag{11}$$

where h is the pose of the camera and K is the camera calibration matrix. $\hat{y}_i$ denotes the predicted 3D scene coordinates and x_i denotes the ground-truth pixel coordinates.

$$\ell_i = s_i\hat{\ell}_i + \alpha ln\frac{1}{s_i}, \tag{12}$$

where $\hat{\ell}_i$ is the reprojection loss of pixel i, and α is a hyperparameter that controls the regularization term.

The partial derivatives of Equation 12 can be computed with respect to its two parameters:

$$\frac{\partial \ell_i}{\partial s_i} = \hat{\ell}_i - \frac{\alpha}{s_i}, \tag{13}$$

$$\frac{\partial \ell_i}{\partial \hat{\ell}_i} = s_i, \tag{14}$$

where s_i is always greater than one to ensure that model training continually reduces $\hat{\ell}_i$ and to prevent the gradient from vanishing. And the Equation 13 is positive only when $s_i > \frac{\alpha}{\hat{\ell}_i}$. It reaches its minimum when $s_i = \frac{\alpha}{\hat{\ell}_i}$. In other words, after the training process is complete, for the input features, the output should have a large s_i and a small $\hat{\ell}_i$ or a small s_i and a large $\hat{\ell}_i$. Therefore, during the prediction process, s_i can be used to assess the reliability of the feature f_i. However, since s_i varies depending on the magnitude of the loss, it cannot directly reflect the reliability of the prediction. For features with very small s_i, the probability of high reliability is very low. Therefore, an appropriate threshold needs to be defined to filter out points with low discriminability. This helps improve accuracy. In the regression head, the prediction layers of $\hat{y}_i$ and $\hat{s}_i$ share almost all parameters.

3.3 Optimization

Combined with Equation 12, the reprojection loss is calculated as follows:

$$\mathcal{L}_{Rep} = \sum_{i \in P} \ell_i, \tag{15}$$

where P is a batch of points from the sampled point set. However, reprojection loss, which is based solely on observations of the same landmark from multiple viewpoints, is insufficient for the model to encode the map information within its parameters. Euclidean loss, which represents the Euclidean distance between predicted and ground-truth 3D coordinates, imposes strong constraints for coordinate prediction:

$$\mathcal{L}_{Euc} = \sum_{i \in P} ||y_i - \hat{y}_i||_2. \tag{16}$$

In reprojection loss, the 2D distances of 3D points are calculated after they are projected onto the camera plane, while the discrepancy between the predicted and true 3D coordinates is minimized by Euclidean loss. The combination of these two types of losses eliminates the need for implicit triangulation. The total loss function is given by:

$$\mathcal{L} = \mathcal{L}_{Rep} + \beta\mathcal{L}_{Euc} + \gamma\mathcal{L}_c. \tag{17}$$

		Mapping Time	Map Size	Cambridge Landmarks					Average
				Court	King's	Hospital	Shop	St. Mary's	(cm / °)
FM	AS (SIFT) [2016a]	∼35min	∼200MB	24/0.1	13/0.2	20/0.4	4/0.2	8/0.3	14/0.2
	hLoc (SP+SG) [2019; 2020b]		∼800MB	16/0.1	12/0.2	15/0.3	4/0.2	7/0.2	11/0.2
	pixLoc [2021]		∼600MB	30/0.1	14/0.2	16/0.3	5/0.2	10/0.3	15/0.2
	GoMatch [2022b]		∼12MB	N/A	25/0.6	283/8.1	48/4.8	335/9.9	N/A
	HybridSC [2019b]		∼1MB	N/A	81/0.6	75/1.0	19/0.5	50/0.5	N/A
APR	PoseNet17 [2017]	4 – 24h	50MB	683/3.5	88/1.0	320/3.3	88/3.8	157/3.3	267/3.0
	MS-Transformer [2021]	∼7h	∼18MB	N/A	83/1.5	181/2.4	86/3.1	162/4.0	N/A
SCR w/ Depth	DSAC* (Full) [2021]	15h	28MB	49/0.3	**15/0.3**	21/0.4	5/0.3	13/0.4	21/0.3
	SANet [2019]	∼1min	∼260MB	328/2.0	32/0.5	32/0.5	10/0.5	16/0.6	84/0.8
	SRC [2022]	2 min†	40MB	81/0.5	39/0.7	38/0.5	19/1.0	31/1.0	42/0.7
SCR	DSAC* (Full) [2021]	15h	28MB	34/0.2	18/0.3	21/0.4	5/0.3	15/0.6	19/0.4
	DSAC* (Tiny) [2021]	11h	4MB	98/0.5	27/0.4	33/0.6	11/0.5	56/1.8	45/0.8
	GLACE [2024]	23 min	13MB	**19/0.1**	19/0.3	17/0.5	**4/0.3**	**9/0.3**	14/0.3
	Poker (4 ACE Ensemble) [2023]	16 min	16MB	28/0.1	18/0.3	25/0.5	5/0.3	**9/0.3**	17/0.3
	ACE [2023]	4 min	4MB	43/0.2	28/0.4	31/0.6	5/0.3	18/0.6	25/0.4
	Ours w/o Depth	4 min	9MB	27/0.1	26/0.3	27/0.5	5/0.3	20/0.6	21/0.4
	Ours w/ Depth	13 min	14MB	20/0.1	**15/0.3**	**15/0.5**	5/0.3	10/0.3	**13/0.3**

Table 2: **Cambridge Landmarks Results.** We report median rotation and position errors. Best results in **bold** for the "SCR" group, second best results underlined.

The correctness of the label prediction is crucial to localization performance, and thus a large value γ should be set. We set $\beta = 0.1$ to ensure that the model does not overly focus on fitting the ground-truth 3D coordinates.

4 Experiment

4.1 Datasets

We conduct our experiments on 7Scenes [Shotton *et al.*, 2013], 12Scenes [Valentin *et al.*, 2016], Cambridge Landmarks [Kendall *et al.*, 2015] and Wayspots [Brachmann *et al.*, 2023]. Experiments are conducted on the first three datasets with RGB and RGB-D. On Wayspots, experiments are conducted with RGB. 7Scenes and 12Scenes are indoor relocalization datasets. They contain 7 and 12 indoor scenes respectively. For each scene, the dataset contains a set of RGB images along with their camera poses and depth maps.

Cambridge is an outdoor relocalization dataset. It contains 5 outdoor scenes. For each scene, the scale is large, reaching up to hundreds of meters. It also includes factors such as lighting, pedestrians, cars, and other influences, consistent with real-world conditions. The dataset is captured using a handheld device, resulting in complex motion trajectories.

Wayspots is a dataset with 10 small outdoor scenes, curated from a publicly available corpus of phone scans. The ground truth poses are reconstructed using SfM, and the original phone trajectories are registered to the SfM poses. Our method outperforms ACE and DSAC* without depth maps.

4.2 Implementation

Architecture. We implement our method in PyTorch, using the backbone of ACE [Brachmann *et al.*, 2023] as the feature extractor. Based on the ACE prediction head, we add a label prediction head and a feature fusion module. Semantic labels are generated using mini-batch k-means, while structural labels are generated using k-means. We use a 6-layer MLP head as the label classifier. It takes the features generated by the feature extractor as input and outputs the labels for each feature. The feature fusion module consists of several MLP layers for feature alignment, as well as a conditioning layer [Li *et al.*, 2020] that performs linear modulation at each position. The image features and label embeddings are simply added together to form the partition-aware features.

Scene	DSAC* (Full)	DSAC* (Tiny)	ACE	Ours
Cubes	83.9%	68.7%	**97.0%**	96.7%
Bears	82.6%	73.1%	80.7%	**89.1%**
Winter Sign	0.2%	0.3%	1.0%	**1.6%**
Inscription	**54.1%**	41.3%	49.0%	50.1%
The Rock	**100%**	99.8%	**100%**	**100%**
Tendrils	25.1%	19.6%	34.9%	**40.6%**
Map	**56.7%**	53.3%	56.5%	55.9%
Square Bench	69.5%	60.3%	66.7%	**71.2%**
Statue	0.0%	0.0%	0.0%	0.0%
Lawn	34.7%	20.0%	35.8%	**45.4%**
Average	50.7%	43.6%	52.2%	**55.1%**

Table 3: **Wayspots Datasets Results.** We show accuracy as the percentage of frames with pose error below 10cm, 5°. Best results in **bold**.

Training. To accelerate the training process, we implement mini-batch k-means with CUDA. The batch size is set to 400K, and the maximum number of iterations is 200. We cluster 64 categories at the structural level and 2 classes at the semantic level. To better adapt to the scale of the scene, we define α in Equation 9 as 20, which yields relatively good results for both large and small scenes. Additionally, we set β in Equation 15 to 0.1. γ is defined as 100 to encour-

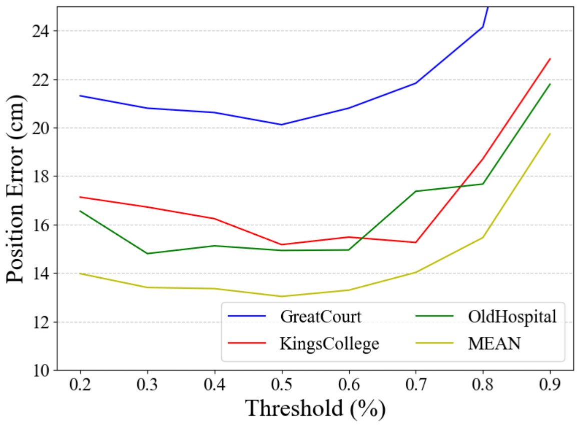

Figure 4: Ablation of discriminability scores on Cambridge dataset. The accuracy varies with changes in the threshold of discriminability scores. The horizontal axis represents the score threshold, while the vertical axis represents the accuracy. The threshold is expressed as a percentage (%). We present our results on the $GreatCourt$, $KingsCollege$, and $OldHospital$, along with the $MEAN$ performance across the Cambridge dataset.

age the network to focus on label prediction accuracy. As described in Section 3.2, the 0.51 percentile of all discriminability scores is used as the threshold for feature filtering. We compare mapping times of ACE, GLACE, DSAC* and ours on NVIDIA GeForce RTX 2080 Ti. The results of other methods come from ACE [Brachmann *et al.*, 2023].

4.3 Analysis

7 Scenes and 12 Scenes. As shown in Tab. 1, our approach achieves higher accuracy compared to ACE [Brachmann *et al.*, 2023] and GLACE [Wang *et al.*, 2024]. Mappings with a time of less than a quarter of an hour are labeled green, while ones smaller than 15 MB are labeled green. Our method achieves the best performance of these methods on 12Scenes, regardless of the existence of depth maps. Additionally, on 7Scenes, our method demonstrates a significant time advantage over DSAC*, while also achieving competitive results.

Cambridge Landmarks. The advantages of our method are fully demonstrated in large-scale scenes. As shown in Tab. 2, our method outperforms state-of-the-art SCR methods [Brachmann *et al.*, 2023; Wang *et al.*, 2024] and is competitive with FM methods. It is worth noting that effectively utilizing depth maps in large-scale scenes presents significant challenges, with DSAC* achieving poor performance when depth maps are provided. The results show that we effectively combine Euclidean distance loss and reprojection loss. Using Euclidean distances as a complement and enhancement to the reprojection error has yielded favorable results.

Wayspots. We also evaluate our method on a small-scale outdoor dataset without depth maps. The main challenge of this dataset is that the scenes contain a large amount of repetitive textures or textureless regions. As shown in Tab. 3, on six of these scenes, we achieved the best performance, with an accuracy improvement of almost 10% points in the $Lawn$ scene. On average, our method achieves an improvement of 2.9% over ACE, while maintaining a training time of 4 minutes.

4.4 Ablation Study

We also conduct ablation studies on the main design choices, modules, and training strategies of our approach on the Cambridge Landmarks dataset.

Structural Partition	Semantic Partition	Feature Fusion	Dis. Score	Average (cm / °)
✓	✓	✓	✓	13.0/0.3
×	✓	✓	✓	18.3/0.4
✓	×	✓	✓	15.1/0.3
✓	✓	×	✓	15.3/0.3
✓	✓	✓	×	16.1/0.3

Table 4: Ablation studies on structural partition, semantic partition, feature fusion and discriminability score. We report median position and rotation errors on Cambridge Landmarks.

Scene Partitioning Strategies. As shown in Tab. 4, we compare our method w/ and w/o structural and semantic partitioning. Without the structural partitioning strategy, performance suffers due to structural discontinuity. Without the semantic partitioning strategy, performance shows a slight decline. Our semantic-structural partitioning strategy preserves both semantic consistency and structural continuity within each subscene, achieving the best results.

Feature Fusion Module. As shown in Tab. 4, with the feature fusion module, the prediction error decreases from 16.1 cm to 13.0 cm.

Thresholds for Discriminability Scores. As shown in Fig. 4, the threshold is expressed as a percentage (%) to prevent inconsistencies in discriminability scores in different scenes. As the threshold increases, the position error initially decreases and then increases. Here, we provide a reasonable explanation. When the threshold is small, only ambiguous points are defined as invalid. Excluding these points enhances the pose estimation. However, as the threshold increases, many highly distinguishable points are excluded, leading to instability in the results. As shown in Fig. 2, where $c = 0.51$, invalid points are concentrated in repetitive textures and flat regions such as grass and sky, while valid points are concentrated in highly distinguishable areas, such as buildings. In the other two scenes of the Cambridge Landmark dataset, $ShopFacade$ and $StMarysChurch$, the error does not exhibit a clear trend with score variation. However, the error does not decrease. With the discriminability score, our method achieves better results.

5 Conclusion

We propose a novel scene coordinate regression method named Understanding Matters: Semantic-Structural Determined Visual Relocalization for Large Scenes. We partition the scene into small parts, ensuring semantic consistency and structural continuity, further leading the accelerated partition localization with sampling-based learning. Partition-determined pose refinement is then put forward to reduce the inaccuracy of pose estimation. Our method supports RGB-D and RGB camera relocalization in large-scale scenes. We can achieve the state-of-the-art performance within 13 minutes.

Acknowledgments

This paper is supported by the Science and Technology Project of Hainan Provincial Department of Transportation (Grant No.HNJTT-KXC-2024-3-22-02), the National Natural Science Foundation of China (Grant No. 62272018, 62206184), and Zhongguancun Laboratory.

References

[Arandjelovic *et al.*, 2016] 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.

[Brachmann and Rother, 2018] 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.

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

[Brachmann and Rother, 2021] Eric Brachmann and Carsten Rother. Visual camera re-localization from rgb and rgb-d images using dsac. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5847–5865, 2021.

[Brachmann *et al.*, 2016] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, et al. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3364–3372, 2016.

[Brachmann *et al.*, 2017] 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.

[Brachmann *et al.*, 2023] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5044–5053, 2023.

[Brahmbhatt *et al.*, 2018] 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.

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

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

[Cavallari *et al.*, 2017] Tommaso Cavallari, Stuart Golodetz, Nicholas A Lord, Julien Valentin, Luigi Di Stefano, and Philip HS Torr. On-the-fly adaptation of regression forests for online camera relocalisation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4457–4466, 2017.

[Cavallari *et al.*, 2019a] Tommaso Cavallari, Luca Bertinetto, Jishnu Mukhoti, Philip Torr, and Stuart Golodetz. Let's take this online: Adapting scene coordinate regression network predictions for online rgb-d camera relocalisation. In *2019 International Conference on 3D Vision (3DV)*, pages 564–573. IEEE, 2019.

[Cavallari *et al.*, 2019b] Tommaso Cavallari, Stuart Golodetz, Nicholas A Lord, Julien Valentin, Victor A Prisacariu, Luigi Di Stefano, and Philip HS Torr. Real-time rgb-d camera pose estimation in novel scenes using a relocalisation cascade. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2465–2477, 2019.

[Dong *et al.*, 2022] Siyan Dong, Shuzhe Wang, Yixin Zhuang, Juho Kannala, Marc Pollefeys, and Baoquan Chen. Visual localization via few-shot scene region classification. In *2022 International Conference on 3D Vision (3DV)*, pages 393–402. IEEE, 2022.

[Humenberger *et al.*, 2020] Martin Humenberger, Yohann Cabon, Nicolas Guerin, Julien Morat, Vincent Leroy, Jérôme Revaud, Philippe Rerole, Noé Pion, Cesar de Souza, and Gabriela Csurka. Robust image retrieval-based visual localization using kapture. *arXiv preprint arXiv:2007.13867*, 2020.

[Kendall and Cipolla, 2017] 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.

[Kendall *et al.*, 2015] 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.

[Li *et al.*, 2020] 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.

[Nguyen *et al.*, 2024] Son Tung Nguyen, Alejandro Fontan, Michael Milford, and Tobias Fischer. Focustune: Tuning visual localization through focus-guided sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3606–3615, 2024.

[Revaud *et al.*, 2019] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5107–5116, 2019.

[Ruan *et al.*, 2023] Jiahao Ruan, Li He, Yisheng Guan, and Hong Zhang. Combining scene coordinate regression and absolute pose regression for visual relocalization. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11749–11755. IEEE, 2023.

[Sarlin *et al.*, 2019] 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/CVF conference on computer vision and pattern recognition*, pages 12716–12725, 2019.

[Sarlin *et al.*, 2020a] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[Sarlin *et al.*, 2020b] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[Sarlin *et al.*, 2021] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, et al. Back to the feature: Learning robust camera localization from pixels to pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3247–3257, 2021.

[Sattler *et al.*, 2016a] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1744–1756, 2016.

[Sattler *et al.*, 2016b] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1744–1756, 2016.

[Schonberger and Frahm, 2016] 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.

[Shavit *et al.*, 2021] Yoli Shavit, Ron Ferens, and Yosi Keller. Learning multi-scene absolute pose regression with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2733–2742, 2021.

[Shotton *et al.*, 2013] 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.

[Torii *et al.*, 2015] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1808–1817, 2015.

[Valentin *et al.*, 2016] Julien Valentin, Angela Dai, Matthias Nießner, Pushmeet Kohli, Philip Torr, Shahram Izadi, and Cem Keskin. Learning to navigate the energy landscape. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 323–332. IEEE, 2016.

[Walch *et al.*, 2017] Florian Walch, Caner Hazirbas, Laura Leal-Taixe, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using lstms for structured feature correlation. In *Proceedings of the IEEE international conference on computer vision*, pages 627–637, 2017.

[Wang *et al.*, 2020] 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.

[Wang *et al.*, 2024] Fangjinhua Wang, Xudong Jiang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Glace: Global local accelerated coordinate encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21562–21571, 2024.

[Winkelbauer *et al.*, 2021] Dominik Winkelbauer, Maximilian Denninger, and Rudolph Triebel. Learning to localize in new environments from synthetic training data. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5840–5846. IEEE, 2021.

[Yang *et al.*, 2019] Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, and Ping Tan. Sanet: Scene agnostic network for camera localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 42–51, 2019.

[Yang *et al.*, 2022] Luwei Yang, Rakesh Shrestha, Wenbo Li, Shuaicheng Liu, Guofeng Zhang, Zhaopeng Cui, and Ping Tan. Scenesqueezer: Learning to compress scene for camera relocalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8259–8268, 2022.

[Zhou *et al.*, 2022a] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022.

[Zhou *et al.*, 2022b] Qunjie Zhou, Sérgio Agostinho, Aljoša Ošep, and Laura Leal-Taixé. Is geometry enough for matching in visual localization? In *European Conference on Computer Vision*, pages 407–425. Springer, 2022.