SegLoc: Learning Segmentation-based Representations for Privacy-Preserving Visual Localization

Maxime Pietrantoni[1,2] Martin Humenberger[3] Torsten Sattler[2]
Gabriela Csurka[3]
[1] Faculty of Electrical Engineering, Czech Technical University in Prague
[2] Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague
[3] NAVER LABS Europe
{firstname.lastname}@cvut.cz,{firstname.lastname}@naverlabs.com

Abstract

Inspired by properties of semantic segmentation, in this paper we investigate how to leverage robust image segmentation in the context of privacy-preserving visual localization. We propose a new localization framework, SegLoc, that leverages image segmentation to create robust, compact, and privacy-preserving scene representations, i.e., 3D maps. We build upon the correspondence-supervised, fine-grained segmentation approach from [42], making it more robust by learning a set of cluster labels with discriminative clustering, additional consistency regularization terms and we jointly learn a global image representation along with a dense local representation. In our localization pipeline, the former will be used for retrieving the most similar images, the latter to refine the retrieved poses by minimizing the label inconsistency between the 3D points of the map and their projection onto the query image. In various experiments, we show that our proposed representation allows to achieve (close-to) state-of-the-art pose estimation results while only using a compact 3D map that does not contain enough information about the original images for an attacker to reconstruct personal information.

1. Introduction

Visual localization is the problem of estimating the precise camera pose – position and orientation – from which the image was taken in a known scene. It is a core component of systems such as self-driving cars [31], autonomous robots [49], and mixed-reality applications [4, 53].

Traditionally, visual localization algorithms rely on a 3D scene representation of the target area, which can be a 3D point cloud map [29, 34, 35, 45, 46, 66, 68, 69, 73, 79], *e.g.* , from Structure-from-Motion (SfM), or a learned 3D representation [9, 10, 14, 37, 38, 71, 76]. The representation is typically derived from reference images with known camera poses. Depending on the application scenario, these maps

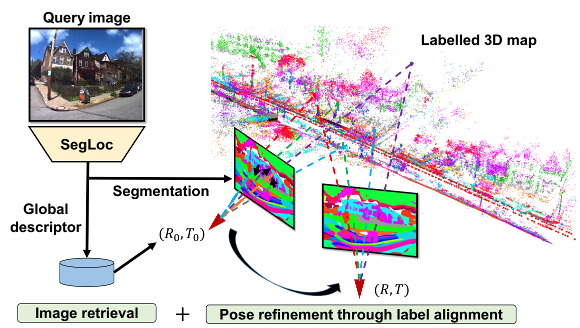

Figure 1. The SegLoc localization pipeline: Our model jointly creates a robust global descriptor used to retrieve an initial pose $(\mathbf{R}_0, \mathbf{T}_0)$ and dense local representations used to obtain the refined pose $(\mathbf{R}, \mathbf{T})$ by maximizing the label consistency between the reprojected 3D points and the query image.

need to be stored in the cloud, which raises important questions about **memory consumption** and **privacy preservation**. It is possible to reconstruct images from maps that contain local image features [62], amongst the most widely used for scene representation.

To tackle the above challenges that feature-based approaches may face, inspired by semantic-based [48, 82] and segmentation-based [42] approaches, we propose a visual localization pipeline where robust segmentations are used as the sole cue for localization, yielding reduced storage requirements (compared to using local features) while increasing privacy-preservation. Our proposed localization pipeline, called SegLoc, follows standard structure based-localization pipelines [34, 66] that represent the scene via a 3D model: first, image retrieval based on a compact image representation is used to coarsely localize a query image. Given such an initial pose estimate, the camera pose is refined by aligning the query image to the 3D map. Contrary to prior work that is based on extracting features directly from images, we derive a more abstract representation in the form of a robust dense segmentation based on a set of clusters learned in a self-supervised manner. As illustrated in Figure 1, we use this segmentation to both extract a global descriptor for image retrieval and for pose refinement. The pose is refined by maximizing the label consistency between the predictions in the query image and a set of labeled 3D points in the scene.

Such an approach has multiple advantages. First, our model is able to learn representations which are **robust to seasonal or appearance changes**. Similar to semantic segmentations, which are invariant to viewing conditions as the semantic meaning of regions do not change, our representation is trained such that the same 3D point is mapped to the same label regardless of viewing conditions. Second, it results in **low storage requirements**, as instead of storing high-dimensional feature descriptors, for each 3D point we only keep its label. Finally, it allows privacy-preserving visual localization [15,22,28,78], as it creates a non-injective mapping from multiple images showing similar objects with different appearances to similar labels. While, ensuring user privacy comes at the cost of reduced pose accuracy [19,98], our method comes close to state-of-the-art results with a better accuracy *vs*. memory *vs*. privacy trade-off.

To summarize, our **first contribution** is a new localization framework, called **SegLoc**, that extends the idea [41, 42] of learning robust fine-grained image segmentations in a self-supervised manner. To that end, we leverage discriminative clustering while putting more emphasis on representation learning. Furthermore, we derive a full localization pipeline, where our model jointly learns global image representation to retrieve images for pose initialization, and dense local representations for building a compact 3D map – an order of magnitude smaller compared to feature-based approaches – and to perform privacy-preserving pose refinement. As a **second contribution**, we draw a connection between segmentation-based representations and privacy-preserving localization, opening up viable alternatives to keypoint-based methods within the accuracy-privacy-memory trade-off. We evaluate our approach in multiple indoor and outdoor environments while quantitatively measuring privacy through detailed experiments.

2. Related Work

Semantic-based Visual Localization. Semantic segmentation is used in structure-based localization methods as a way to facilitate feature selection or matching [2, 39, 40, 54, 56, 72], to filter 2D-3D matches by maximizing the semantic consistency between 2D images and 3D models [42, 74, 83] or to improve keypoint tracking [48]. In these works, the pre-trained segmentation model is used mainly to filter matches or to improve SfM/VO, hence they still rely on keypoints descriptors. Similar to FGSN [42], our model learns robust image segmentation in a self-supervised manner exploiting label consistency between matched keypoints. Contrary to FGSN, our model provides both global and local representations, resulting in a full localization pipeline.

Semantics-based retrieval and pose approximation. To cope with extreme environmental, seasonal, and illumination changes in place recognition and image retrieval, several methods leverage image-to-image translation to handle the domain shift between the database and query images [1, 32, 33, 63, 80, 96, 97]. Other methods directly aim at obtaining image representation by leveraging weather-invariant semantic [7, 26, 82, 94] or geometric information [59, 60]. In particular, [82] describes images through histograms of semantic classes from pre-trained semantic segmentation, while LoST [26] performs a semantic-based pooling of convolutional features. Closest to our work, [32] and [58] train global representations within a deep metric learning framework and utilize semantic segmentation as an auxiliary task to infuse semantic information. Instead, we learn a finite set of (not necessarily semantic) classes to perform image segmentation from which we build local and global representations.

Pose refinement. Pose refinement approaches obtain an accurate camera pose estimate from an initial approximate pose via image alignment. In contrast to early methods based on handcrafted features [55] or pixel intensities [23], more recent methods learn deep features through direct feature alignment suitable for such pose refinement [67, 87] or cast the camera pose localization as a metric learning problem. [82] proposed a semantic-based pose refinement relying on a pre-trained model, handcrafted global descriptors, and a geometric prior. Our pose refinement is inspired by PixLoc [67], except that instead of using multi-scale deep features, we align 1D features (labels) by minimizing a reprojection error as a function of label inconsistency.

Privacy-preserving localization. Storing the 3D maps on the cloud or sending images or descriptors from mobile devices to a server raise the question of privacy. As shown in [62], detailed and recognizable images of the scene can be obtained from sparse 3D point clouds with local descriptors. Geometry-based matching methods [12, 98] do not rely on visual descriptors to localize, hence they are less subject to privacy issues. Still, [62] shows that depth and color are sometimes sufficient to recover details in a scene. Learning-based pose regression or scene point regression models [8, 38, 88, 91] do not explicitly store the 3D map and, thus, partially avoid the privacy issues. Yet, according to [51], since these models memorize the scenes quite well, model inversion is often possible. Given a set of pre-selected scene landmarks, [19] learns to detect them and to regress the associated bearing vectors used by geometric camera pose estimation. However, this method still faces the same scaling issues as regression methods.

To address privacy, [75, 77] propose to transform 3D point clouds into 3D line clouds, thus obfuscating the scene geometry. However, according to [15], a significant amount of information about the scene geometry is preserved in these line clouds, allowing to (approximately) recover image content. [28] propose a cloud-based mapping solution to preserve the privacy of users by hiding critical content of the input images. As the recovered pose may also be considered as sensitive, [27] perform a partial estimation of a 3DoF pose on a single dimension against a partial map. These partials maps are distributed on distinct servers so that the 6DoF pose can only be recovered on the user side. However, they do not tackle the privacy of the partial maps directly. In addition, these approaches do not consider private information contained in the query images, which could, *e.g.* , allow an attacker to track individuals and to study their behavior. Concerning privacy preservation of the query, [20, 21] show that it is possible to reconstruct the original image from local image features. To address this, [22] propose to obfuscate the appearance of the original image by lifting the descriptors to affine subspaces. [78] propose to replace 2D points in the query image with randomly oriented 2D lines passing through the given point. This allows them to address privacy of both the query and the map. By relying on class labels, where only labels are kept in the map and hence making it impossible to recover fine details that could reveal private information, our SegLoc representations jointly tackles query and map privacy.

3. The SegLoc Model

Our goal is to jointly learn local and global representations for visual localization. Inspired by the invariance of semantic class labels to viewing conditions, we propose a robust image segmentation method based on a set of clusters uncovered in a self-supervised manner. To make the segmentation robust to viewpoint and appearance changes, we train our model on an ensemble of image pairs taken from different viewpoints and at different points in time with a set of automatically extracted keypoint correspondences between them (see Supplementary). We assume a pre-trained encoder providing initial dense representations, which are grouped into K prototypes, where K is the number of clusters / labels / classes representing the desired segmentation granularity. They are used to initialize the segmentation heads and the discriminative clustering step.

Hence, the main ingredients of our model are: dense segmentation as representations learned with discriminative clustering (Sec. 3.1), three additional consistency regularization terms (Sec. 3.2), and global image representations trained with a multi-similarity pairwise loss (Sec. 3.3).

3.1. Multi-scale dense representation

The segmentation network has a hierarchical structure and uses a hybrid-DPT [64] like encoder-decoder module as backbone such that the output of each level is the input of the next one. The resolutions of the output decoded feature maps $\mathbf{F}_l \in \mathbb{R}^{D\times H_l \times W_l}$ progressively increases. Each feature map $\mathbf{F}_l$ is further processed by a classification head h_{μ_l} parametrized by μ_l which predicts a set of yielding segmentation heatmaps $\mathbf{P}_l^k \in \mathbb{R}^{H_l \times W_l}$ – with per pixel class likelihoods – corresponding the k^{th} cluster. The tensor concatenating the K maps, denoted by $\mathbf{P_l} \in \mathbf{R^{K\times H_l \times W_l}}$, is an abstract representation of the image. As the decoder outputs higher resolution feature maps the encoded information becomes finer. We thus operate on four complementary distinct metric spaces and classification spaces ($l \in 1..4$), learning four distinct cluster sets, one per level. During pose refinement, we hierarchically use these maps from coarser to finer to leverage information from different levels of granularity. For pose approximation only the finer segmentation is used to compute the global representation while we use the four segmentations for pose refinement. In the following, we drop the level notation l as the described steps are applied on each level without distinction.

Discriminative clustering. For clustering, we rely on a Deep Discriminative Clustering (DDC) framework [18, 36, 92] as it focuses on learning the boundaries between clusters rather than explicitly modelling the data distribution, hence casting the clustering task as a classification problem. Following [18], we use an auxiliary target to supervise the training by minimizing the Kullback-Leibler (KL) divergence between the predicted distributions $\mathbf{P}$ and target distributions $\mathbf{Q}$. To avoid degenerated solutions, [18] use a regularization term that minimizes $KL(\mathbf{d}^q \| \mathbf{d}^u)$ between the empirical label distribution $\mathbf{d^q}$ defined as the soft frequency of cluster assignments in the target distribution and the uniform distribution $\mathbf{d^u}$ to enforce a balanced cluster assignments. We instead rely on the data itself to directly estimate an empirical label distribution $\mathbf{d}^p$ [1].

We add an entropy term $H(\mathbf{Q})$ that encourages peaked target distributions and minimize the clustering objective:

$$\mathcal{L}_{DC} = KL(\mathbf{Q}\|\mathbf{P}) + KL(\mathbf{d}^q\|\mathbf{d}^p) + H(\mathbf{Q}) \quad , \qquad (1)$$

where $\mathbf{d}^q_k = \sum_i^{HWB} q_{ik}$ and B is the batch size. As this objective depends both on the target distributions $\mathbf{Q}$ and the network parameters, it is minimized by alternating the following two sub-steps in every batch:

1. Update target distribution: With network parameters fixed, the following closed-form solution minimizes the cost function Eq. (1) in a batch of size B:

$$q_{i_b k} = \frac{d^p_k p^2_{i_b k} / (\sum_{b'=1}^{B} \sum_{i_{b'}=1}^{HW} p^2_{i'_b k})^{\frac{1}{2}}}{\sum_{k=1}^{K} d^p_k p^2_{i_b k} / (\sum_{b'=1}^{B} \sum_{i_{b'}=1}^{HW} p^2_{i'_b k})^{\frac{1}{2}}} \quad . \qquad (2)$$

[1] Enforcing uniform prior is not desirable as information is unbalanced in the dataset and even more in a batch.

2. Update model parameters: With target distributions fixed, minimizing the cost function accounts to minimizing the following per-pixel cross entropy loss:

$$\mathcal{L}_{CE} = -\frac{1}{HWB}\sum_{b=1}^{B}\sum_{i_b=1}^{HW}\sum_{k=1}^{K} q_{i_b k}\log(\sigma(p_{i_b k})) \quad , \tag{3}$$

with σ being the softmax function.

The model is self-supervised by the auxiliary target distributions $\mathbf{Q}$, where q_{ik} are computed from the initial class predictions p_{ik}. However, these predictions are not reliable at the beginning of the training process. Therefore, during the first epoch, instead of using Eq. (2) to update $\mathbf{Q}$, we rely on some initial prototypes (cluster centers) $\mathbf{c}_k$. Specifically, we compute soft class assignments w.r.t. the associated cluster for each pixel x_i using a Student's t-distribution [86]:

$$q_{ik} = \frac{(1+\|\mathbf{F}_i - \mathbf{c}_k\|_2^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{j=1}^{K}(1+\|\mathbf{F}_i - \mathbf{c}_j\|_2^2/\alpha)^{-\frac{\alpha+1}{2}}} \quad , \tag{4}$$

using the corresponding feature vectors $\mathbf{F_i}$ and $\alpha = 1$ as in [92]. Using Eq. (4) instead of Eq. (2) in the first epoch acts not only as initialization, but allows also to distill underlying prior knowledge (see details in the Supplementary), helping the learning process to be more efficient.

3.2. Consistency regularization

Aiming to learn dense segmentations robust to photometric changes while being equivariant to viewpoint changes and to avoid overfitting, we propose the following three consistency regularization losses.

Formally, let I^a, I^b be an image pair with the corresponding l2-normalized feature maps $\mathbf{F}^a = f_\theta(I^a)$ and $\mathbf{F}^b = f_\theta(I^b)$ respectively. We denote the set of automatically obtained 2D-2D keypoint correspondences by $\{x^a_{u_l}, x^b_{v_l}\}_{l=1}^{L}$, where $x^a_{u_l}$ and $x^b_{v_l}$ are the keypoint locations in the feature maps $\mathbf{F}^a$ respectively $\mathbf{F}^b$. We define the following losses:

Correspondence consistency loss. Similar to [42], to enforce consistency between pairs of segmentations we use a *correspondence consistency loss*:

$$\mathcal{L}_{CC} = -\frac{1}{2L}\sum_{l=1}^{L} \mathbb{1}^\top_{s^b_{v_l}}\log(\sigma(\mathbf{p}^a_{u_l})) + \mathbb{1}^\top_{s^a_{u_l}}\log(\sigma(\mathbf{p}^b_{v_l})) \quad ,$$

where $\mathbf{p}^a_{u_l} = h_\mu(\mathbf{F}^a_{u_l})$, $\mathbf{p}^b_{v_l} = h_\mu(\mathbf{F}^b_{v_l})$, $\mathbb{1}_k$ is the one-hot vector with all zero values except at position k, $s^a_{u_l}$ and $s^b_{v_l}$ are the hard-assigned cluster labels to the keypoints $x^a_{u_l}$ and $x^b_{v_l}$ based on their distance to the prototypes $\{\mathbf{c}_k\}_{k=1}^{K}$ based on: $s^a_{u_l} = \text{argmax}_k\, \mathbf{c}_k^\top \mathbf{F}^a_{u_l}$ and $s^b_{v_l} = \text{argmax}_k\, \mathbf{c}_k^\top \mathbf{F}^b_{v_l}$.

Prototypical cross contrastive loss. To constrain the feature space to ensure separability between the implicitly defined classes and to improve intra-class compactness, we define the *prototypical cross contrastive loss*, inspired by the ProtoNCE [44], as follows:

$$\mathcal{L}_{PC} = \frac{1}{2L}\sum_{l=1}^{L}\log\left(\frac{1}{Z}\exp\left(\frac{\mathbf{c}^\top_{s^b_{v_l}}\mathbf{F}^a_{u_l}}{\phi_{s^b_{v_l}}} + \frac{\mathbf{c}^\top_{s^a_{u_l}}\mathbf{F}^b_{v_l}}{\phi_{s^a_{u_l}}}\right)\right) \quad ,$$

with $Z = \left(\sum_k \exp\left(\mathbf{c}_k^\top \mathbf{F}^a_{u_l}/\phi_k\right)\right)\left(\sum_k \exp\left(\mathbf{c}_k^\top \mathbf{F}^b_{v_l}/\phi_k\right)\right)$, ϕ_k being the concentration of the prototype $\mathbf{c_k}$ defined as the average feature distance to the prototype within the cluster k and it acts as a scaling factor preventing cluster collapse. This loss incorporates in the feature space the structure conveyed by the prototypes.

Contrastive feature consistency loss. To exploit the relationships between keypoints, we adapt the supervised contrastive loss [85] to enforce *feature consistency*:

$$\mathcal{L}_{FC} = -\frac{1}{L}\sum_{l=1}^{L}\log\frac{\exp(\mathbf{F}^{a\top}_{u_l}\mathbf{F}^b_{v_l}/\tau)}{\sum_{j=1}^{L}\exp(\mathbf{F}^{a\top}_{u_l}\mathbf{F}^b_j/\tau)} \quad . \tag{5}$$

The anchor/positive pairs are provided by the pixel-to-pixel correspondences, while the negatives are obtained by sampling amongst the other keypoints in the set $\{x^b_{v_j}, j \neq l\}$. This loss forces the features of corresponding keypoints to be similar, hence facilitating the subsequent clustering.

3.3. Segmentation-based global descriptor

To fully leverage our segmentations, we propose to compute a global image representation by applying a pooling operator on the segmentation heatmap instead of the feature maps. We use the Generalized Pooling Operator (GPO) [16], which generalizes over different pooling strategies to learn the most appropriate pooling strategy to describe the content. Given a heatmap's channel $\mathbf{P}^k \in \mathbb{R}^{H\times W}$, it is defined as a weighted sum over sorted features:

$$v^k = \sum_{o=1}^{HW}\theta_o\psi_o^d \quad \text{where} \quad \sum_{o=1}^{HW}\theta_o = 1 \quad , \tag{6}$$

where v^k is the k^{th} element of the output feature vector, ψ^k_o is the o^{th} element from the ordered descending lists of the values in the in the heatmap's channel; $\mathbf{P}^k$ and the weights θ_o are shared between the channels. We use the higher resolution heatmap from the last level of the decoder as input. The segmentation labels provide a much weaker signal compared to features, we thus opted for spatial pooling to increase discriminativeness instead of using a permutation-invariant pooling strategy such as [3]. We divide the image into M overlapping sliding sub-windows and apply pooling within each sub-window. The corresponding features are then concatenated yielding a global representation of dimension MK. While this implies lower robustness to viewpoint change, in practice we find it sufficient since subsequent pose refinement requires an initial pose close enough to the true pose to enable convergence (*c.f.* Fig. A.5 in the Supplementary). PCA+whitening postprocessing is applied to reduce the dimension to 4096.

Multi-similarity loss. We consider as global training objective the multi-similarity loss [90]. Given an anchor image I_j^a, we denote the corresponding positive respectively negative image sets by $\mathcal{N}_n^+ = \{I_j^+\}$ and $\mathcal{N}_n^- = \{I_j^-\}$ and the corresponding similarities, computed between the pooled global representations by S_{jn}^+ and S_{jn}^-. The *multi-similarity loss* is then defined as:

$$\mathcal{L}_{MS} = \frac{1}{N}\sum_{n=1}^{N}\sum_{\rho\in\{+,-\}}\frac{1}{\alpha^\rho}\log\left(1+\sum_{I_j^\rho\in\mathcal{N}_n^\rho} e^{\rho\alpha^\rho(\lambda-S_{jn}^\rho)}\right)$$

where α^+, α^- and λ are hyper-parameters. We use image pairs contained in our dataset as anchor/positive pair. According to standard practices, the rest of positive/negative samples are mined from $\{I_{n'}^a, I_{n'}^b\}_{n'\neq n}$ through a semi-hard mining scheme based [24, 30] on features distances and image positions (see details in the Supplementary).

4. The SegLoc Localization Pipeline

This section describes our two step localization pipeline using SegLoc (see Figure 1). We use a 3D representation of the environment that consists of a set of reference images with their corresponding camera pose and SegLoc global descriptor as well a sparse 3D model. Each 3D point is associated to a cluster label which is the index of its assigned cluster instead of a visual descriptor. First, given a query image, the most relevant database images are retrieved based on global descriptor similarity. Second, we refine the approximated pose that was derived from the retrieved images. The proposed pose refinement process works as follows. Let $(\mathbf{R}_0, \mathbf{T}_0)$ be the approximate pose obtained using the poses of the top-k retrieved similar images and let $\mathcal{X} = \{(\mathbf{X}_m, y_m)\}$ be the set of labeled 3D points visible in the top-k images, where $\mathbf{X}_m$ represents its 3D coordinates and y_m the associated label. Inspired by image/features alignment [6, 23, 67] methods, in order to refine the initial approximate pose, we propose to use a geometric optimization approach. However, in contrast to existing methods, to find the query camera pose $(\mathbf{R}, \mathbf{T})$, we neither rely on the reference image nor on complex features. Instead, we minimize the inconsistency between the reprojected 3D labels and the predicted segmentation map in the query image defined as:

$$E(\mathbf{R}, \mathbf{T}) = \sum_{\mathcal{X}} w_m \rho(|\mathbf{p}_m - \mathbb{1}_{y_m}|) \ , \qquad (7)$$

where $\mathbb{1}_k$ is the one-hot vector with all zero values except at position y_m, $\mathbf{p_m}$ is the class probability vector for $\mathbf{x_m} = \mathbf{K}(\mathbf{R}\mathbf{X_m} + \mathbf{T})$, $\mathbf{K}$ being the query camera matrix, $(\mathbf{R}, \mathbf{T})$ the estimated pose, and w_m are learned weights for outdoor environments or weights derived from edge detectors for indoor environments (details in Supplementary).

We initialize $(\mathbf{R}, \mathbf{T})$ by $(\mathbf{R_0}, \mathbf{T_0})$ and iteratively refine the pose by minimizing he objective Eq. (7) with the Levenberg-Marquart [43, 52] algorithm, ρ being a Cauchy robust cost function $\rho(x) = \frac{\psi^2}{2} log(1 + \frac{x}{\psi}^2)$ with $\psi = 0.1$. We hierarchically apply this refinement procedure with regard to the 3D model using coarser to finer segmentations.

5. Experimental Evaluation

Training and test data. We train and evaluate our model in both indoor and outdoor scenes. Outdoors, we created an extended version of the Cross-Seasons Correspondences dataset [41] (built upon the training slices of ECMU [5, 81]), including more diverse intra-seasons image pairs and larger viewpoint changes between image pairs (see Supplementary). Indoors, we use the challenging Indoor6 dataset [19] from which we sample pairs of co-visible images captured under different conditions and compute their correspondences based on geometry. For in-domain evaluation, we use the test sets of ECMU and Indoor6, for evaluating the generalization ability, we use RobotCar Seasons (RC) [50, 70] and Cambridge Landmarks [38]. Our models use 100 classes throughout the whole experimental section.

Evaluation protocol. To measure pose accuracy, we follow [61, 70] and compute the position and rotation errors between the estimated query pose and the ground truth pose. We report the percentage of query images localized within fine (.25m, 2°), medium (.5m, 5°) and coarse (5m, 10°) thresholds for outdoor environments and median translation and rotation errors in (cm/°) as well as the localization recall at (5cm, 5°) for indoor environments.

5.1. Results and ablative study

Pose approximation results. Tab. 1 compares our global descriptor against four popular global representations used for localization [61], DELG [13], APGeM [65], DenseVLAD [84] and NetVLAD [3]. Additionally, we include global descriptors that implicitly leverage semantic information: DASGIL [32], DIFL-FCL [33] and LVLPR [93]. On ECMU, our global descriptor significantly outperforms all existing representations, demonstrating the discriminativeness of the learned segmentations. Furthermore, SegLoc performs very well in day conditions of the RC dataset, despite not being trained on it. The drop in performance between day and night images on RC can be explained by the fact that our training set does not include nighttime photos. Still, SegLoc is only outperformed by LVLPR on the coarser thresholds, which was trained on RC.

Pose refinement results. Tab. 2 compares our pose refinement approach against PixLoc [67] (also trained on

Model	Trained on	ECMU Seasons: Urban	ECMU Seasons: Suburban	ECMU Seasons: Park	RC Seasons: Day	RC Seasons: Night
DELG [13]	GL18 [57]	7.8 / 19.9 / 73.7	2.5 / 9.6 / 66.2	1.5 / 6.0 / 43.3	4.9 / 20.5 / 85.4	0.1 / 1.6 / 22.7
AP GeM [65]	GL18 [57]	8.0 / 20.6 / 74.7	2.7 / 10.0 / 63.8	1.3 / 5.6 / 41.4	6.1 / 22.3 / 90.1	0.5 / 3.0 / 28.4
DenseVLAD [84]	24/7 Tokyo [84]	12.2 / 29.2 / 74.7	4.7 / 17.3 / 73.0	4.5 / 18.0 / 62.6	7.4 / 29.7 / 90.1	1.1 / 5.5 / 24.9
NetVLAD [3]	Pitts30k [3]	10.3 / 25.7 / 78.1	3.3 / 12.2 / 68.9	2.3 / 9.3 / 53.6	6.2 / 26.2 / **91.5**	0.3 / 2.1 / 16.1
DIFL-FCL [33]	ECMU [5, 81]	14.8 / 35.1 / 79.6	5.6 / 18.2 / 69.8	6.1 / 20.7 / 69.1	7.6 / 26.2 / 75.9	2.5 / 6.5 / 15.8
LVLPR [93]	RC [50, 70]	17.3 / 42.5 / 89.0	5.8 / 19.4 / 76.1	6.6 / 23.1 / 73.0	7.9 / 30.0 / 85.9	**4.1 / 15.7 / 59.1**
DASGIL [32]	Virtual KITTI [25]	17.4 / 42.0 / 91.1	6.7 / 22.1 / 88.5	7.9 / 26.9 / 83.5	8.7 / 30.7 / 81.3	1.7 / 4.6 / 20.7
SegLoc	ECMU [5, 81]	**21.5 / 51.7 / 96.5**	**8.7 / 28.5 / 92.6**	**10.0 / 35.9 / 92.7**	**9.5 / 37.0** / 90.3	1.8 / 10.5 / 33.3

Table 1. Pose approximation (PA) results obtained with the pose of the top-1 retrieved images using different global representations.

		Memory (GB)	Reconstruction quality: PSNR ($\uparrow$)	LPIPS ($\downarrow$)	SSIM ($\uparrow$)	MAE ($\downarrow$)	Localization Accuracy: Urban (%)	Suburban (%)	Park (%)
ECMU	SegLoc NV 1L	**0.102**	**15.66**	**0.46**	**0.63**	**0.11**	76.3 / 83.0 / 92.8	63.0 / 71.1 / 82.6	45.4 / 53.1 / 67.5
	SegLoc NV						81.9 / 87.4 / 93.9	66.7 / 72.9 / 81.4	50.9 / 57.8 / 68.3
	SegLoc 1L						81.1 / 87.3 / 93.6	65.9 / 73.8 / 81.2	66.9 / 76.0 / 85.6
	SegLoc						88.0 / **93.2 / 97.2**	**83.7 / 89.2 / 93.4**	**80.5 / 87.5 / 93.1**
	PixLoc NV (descriptors)	9.313	21.85	0.28	0.83	0.06	**88.3** / 90.4 / 93.7	79.6 / 81.1 / 85.2	61.0 / 62.5 / 69.4
	PixLoc Oracle	9.313	21.85	0.28	0.83	0.06	92.8 / 95.1 / 98.5	91.9 / 93.4 / 95.8	84.0 / 85.8 / 90.9
							Day	Night	
RC	SegLoc NV 1L	**0.211**	**14.10**	**0.43**	**0.69**	**0.15**	39.8 / 67.5 / 92.4	3.2 / 8.1 / 33.4	
	SegLoc NV						43.2 / 70.1 / 92.6	6.3 / 14.8 / 42.7	
	Segloc 1L						42.6 / 68.9 / 89.8	4.2 / 10.6 / 29.4	
	SegLoc						44.2 / 70.2 / 90.1	7.7 / 16.4 / 32.2	
	PixLoc NV (descriptors)	19.603	18.50	0.33	0.82	0.09	**52.7 / 77.5 / 93.9**	**12.0 / 20.7 / 45.4**	
	PixLoc Oracle	19.603	18.50	0.33	0.82	0.09	55.8 / 80.8 / 96.4	23.6 / 40.3 / 77.8	

Table 2. Comparing cluster-based SegLoc with feature-based PixLoc (both trained on ECMU) on the pose refinement (PR) task in terms of pose accuracy, memory requirements, and privacy of the underlying 3D map representation. Privacy is evaluated by recovering images from the point clouds used by both methods (worse image quality implies a higher level of privacy). For a better comparison we also evaluate SegLoc using NetVLAD for retrieval (NV). We further provide results when instead of querying with the full label distribution, we only use a single label per pixel (1L). Recall that in SegLoc, the 3D point are always represented with a single label.

ECMU), a state-of-the-art pose refinement method. We report pose accuracy results for ECMU and RC together with the storage requirements of the two methods. We further analyze both approaches in terms of map privacy. To this end, we train the map inversion approach from [22, 62] to recover images from the point clouds used by both PixLoc and our approach (details in Sec. 5.2). Comparing the recovered images with the original ones, we report the PSNR, LPIPS [95], SSIM, and MAE metrics. Low scores are desired for PSNR and SSIM (resp. high scores for LPIPS and MAE) as this means the images cannot be well reconstructed. For visualizations and a memory consumption report, see Supplementary. The results verify our claim that our approach is more privacy-preserving than feature-based methods as the image reconstruction results are significantly worse (see also Sec. 5.2). Furthermore, our approach requires significantly less storage space.

Concerning accuracy, using solely SegLoc as a full pipeline (using a single model to compute representations for both retrieval and pose refinement) outperforms PixLoc on ECMU. Particularly in the "suburban" and "park" scenes, where it is hard to find stable and reliable local features in scenes dominated by vegetation – especially under seasonal changes, our representations are more robust. In the case of "park", SegLoc even outperforms PixLoc on coarser accuracy levels, even if PixLoc uses an "oracle" ranking. These gains are partially due to our robust global representation. When using NetVLAD to initialize our poses (SegLoc NV), our results are below the PixLoc performance. This comes at no surprise given that PixLoc uses high-dimensional features that store significantly more information (see Tab. 2), confirming the observation made in [19, 98] that privacy-preservation comes at the cost of decreased pose accuracy.

With a limited semantic gap as in RC (or Cambridge Landmarks) with regard to the ECMU training set, our approach is still able to significantly refine initial poses as shown in Tab. 2 and Tab. 4. However, our approach is data-driven and uncovers a set of clusters without human supervision or dense annotations. While being somewhat interpretable, these clusters remain tied to the semantic space of the training dataset. Without explicit domain adaptation, we thus cannot expect strong generalization capabilities. This explains the gap between SegLoc and PixLoc results on RC Seasons. Improving generalization, *e.g.*, by training on more data, is an interesting direction for future work.

Comparison to privacy preserving methods. Next, we compare against recent privacy-preserving visual localization methods, DSAC* [11], GoMatch [98], and NBE+SLD [19] on the Indoor6 [19] (Tab. 3, all methods are also trained on Indoor6 dataset) and Cambridge Landmarks [38] (Tab. 4). By design these methods do not scale to large outdoor environments, so we did not include them in our outdoor comparisons (ECMU and RC Seasons). On Indoor6, SegLoc significantly outperforms DSAC*, but in some

	Privacy Preserving	Memory (MB)	scene1	scene2a	scene3	scene4a	scene5	scene6
			Median pose error (cm.) (↓), Median angle error (°) (↓), Recall at 5cm/5° (%) (↑)					
DSAC* [11]	✓	27	12.3/2.06/18.7	7.9/0.9/28.0	13.1/2.34/19.7	**3.7**/0.95/60.8	40.7/6.72/10.6	6.0/1.40/44.3
NBE+SLD [19]	✓	132	6.5/0.9/38.4	7.2/0.68/32.7	4.4/0.91/**53.0**	3.8/**0.94**/**66.5**	6.0/0.91/40.0	5.0/0.99/**50.5**
SegLoc	✓	161	**3.9**/**0.72**/**51.0**	**3.2**/**0.37**/**56.4**	**4.2**/**0.86**/41.8	6.6/1.27/33.84	**5.1**/**0.81**/**43.1**	**3.5**/**0.78**/34.5

Table 3. Localization results on Indoor6 compared in terms of memory footprint (required to store the map), and localization accuracy.

Model	MB	King's	Old	Shop	St. Mary's
GoMatch ✓ [98]	48	0.25/0.64	2.83/8.14	0.48/4.77	3.35/9.94
SegLoc ✓	**23**	**0.24**/**0.26**	**0.36**/**0.52**	**0.11**/**0.34**	**0.17**/**0.46**
PixLoc [67]	3545	0.14/0.24	0.16/0.32	0.05/0.23	0.10/0.34

Table 4. Comparison of median position and orientation errors (m./°) on Cambridge Landmarks [38]. We outperform the privacy-preserving GoMatch [98] approach in all metrics. For reference, we include the non-privacy-preserving PixLoc method, which requires more than two order of magnitude more memory.

	urban	suburban	park
DPT Hybrid [64]	4.7/14.9/88.2	3.9/14.0/57.3	3.4/8.7/50.0
FGSN [42]	26.8/51.5/92.3	15.1/32.1/68.2	10.3/22.8/**66.0**
SegLoc	**43.4**/**63.4**/**92.6**	**27.4**/**42.0**/**69.1**	**17.0**/**31.0**/65.7

Table 5. Pose refinement results on ECMU when varying the segmentation models.

cases it performs much worse than NBE+SLD. NBE+SLD trains a model for each scene learning to detect a set of scene dependent landmarks. In contrast, we train and use a single model for all scenes. On Cambridge Landmarks, SegLoc (trained on ECMU) consistently outperforms GoMatch (trained on MegaDepth [47]), improving up to an order of magnitude in position accuracy.

Ablation study. Tab. 5 compares the discriminativeness and consistency of our segmentations to the pre-trained semantic segmentations from DPT Hybrid [64] and the segementations from FGSN [42] (also trained on ECMU). For a fair comparison, we used a single scale SegLoc model with lower input resolution (480x480) and one segmentation layer instead of four, which explains the gap between the numbers in Tab. 5 and Tab. 2. We use all three segmentations in our refinement procedure with the same initial poses. Our approach, which uses a discriminative clustering framework combined with joint global/local representation learning, significantly outperforms both baselines.

Furthermore in Tab. 6, we study the impact of the individual components of our approach on its performance. As can be seen, all losses contribute to the overall performance and improve upon the core model using only the discriminative clustering (first row). A more detailed analysis and further ablation studies can be found in the Supplementary.

5.2. Accuracy vs privacy trade-off

As seen in our experiments above, and as already mentioned in [19, 98], ensuring user privacy comes at the cost of reduced pose accuracy. In this section, we explore this trade-off in more details. Inspired by [17], we quantify how privacy-preserving our approach is in comparison to PixLoc

	MS/CC/PC/FC	urban	suburban	park
PA	×/×/×/×	9.0 / 24.1 / 82.3	3.5 / 13.0 / 65.9	2.0 / 8.4 / 49.0
	✓/×/×/×	12.1 / 31.3 / 89.9	5.0 / 17.9 / 78.7	4.2 / 16.2 / 68.36
	✓/✓/✓/×	11.3 / 29.8 / 90.6	4.5 / 17.3 / 80.7	3.8 / 14.9 / 71.7
	✓/×/✓/✓	13.2 / 34.1 / 91.4	5.6 / 20.0 / 83.2	5.3 / 20.5 / 82.6
	✓/✓/ × / ✓	15.4 / 38.7 / 93.1	6.0 / 21.1 / 86.1	6.9 / 25.8 / 85.5
	× /✓/✓/✓	14.6 / 37.1 / 91.7	6.0 / 21.6 / 85.3	6.8 / 25.3 / 86.7
	✓/✓/✓/✓	**15.7** / **39.4** / **93.2**	**6.4** / **22.5** / **86.6**	**7.4** / **27.1** / **87.6**
PR	×/×/×/×	33.6 / 51.1 / 82.9	26.7 / 40.4 / 66.4	12.1 / 20.7 / 48.7
	✓/× /×/×	40.0 / 60.7 / 89.4	31.7 / 50.1 / 79.0	16.4 / 30.6 / 66.6
	×✓/✓/✓	42.8 / 64.2 / 92.4	**34.9** / **55.8** / 86.0	**31.8** / **54.7** / 87.0
	✓/✓/✓/✓	**44.8** / **64.5** / **93.7**	33.9 / 55.7 / **86.9**	29.8 / 52.1 / **87.5**

Table 6. Ablation study for pose approximation (PA) and pose refinement (PR). We disable different losses and used an input size of 480x480 pixels and one feature level during training.

	Object	Segloc	Pixloc
		Precision/Recall/AP@.5/AP (%) (↑)	
Outdoor	Person	57 / 2 / 2 / 1	81 / 23 / 22 / 12
	Car	58 / 22 / 15 / 6	84 / 51 / 49 / 32
	Truck	92 / 8 / 8 / 5	75/ 32 / 28 / 19
Indoor	Bed	27 / 5 / 3 / 2	29 / 13 / 8 / 7
	Chair	64 / 6 / 5 / 2	61 / 10 / 9 / 5
	TV	21 / 7 / 3 / 1	47 / 19 / 15 / 11
	Refrigerator	46 / 13 / 12 / 8	66 / 16 / 15 / 14

Table 7. Detection of privacy sensitive classes on reconstructed images (ECMU / Indoor6 test sets)

		Fine accuracy (.25m,2°) (%)	PSNR (↑)	LPIPS (↓)	SSIM (↑)	MAE (↓)
Outdoor	SegLoc	60.3	14.44	0.42	0.53	0.14
	SegLoc top10	66,5	17.37	0.32	0.70	0.09
	SegLoc full	66,5	19.63	0.31	0.77	0.07
	Pixloc	76.3	22.00	0.23	0.85	0.05
		Recall (5cm,5°) (%)	PSNR (↑)	LPIPS (↓)	SSIM (↑)	MAE (↓)
Indoor	SegLoc	33.4	14.13	0.49	0.69	0.16
	SegLoc top10	37.01	16.47	0.35	0.77	0.12
	SegLoc full	37.01	17.90	0.26	0.80	0.10
	Pixloc	32.14	25.30	0.10	0.90	0.05

Table 8. Dense reconstruction from query input representations.

(the most similar in terms of pose refinement).

For any visual localization service, the user sends a query to a server and the latter performs visual localization against a stored database model and returns the 6DoF pose to the user. In this context, we define privacy as the inability for an attacker to recover critical details of the scene either from the query or the database. Qualitatively, we measure this degree of privacy through the output of feature/SfM model inversion approaches [22, 62] that recover images from 3D models or image representations. Quantitatively, we measure privacy using image reconstruction metrics and by quantifying the ability of detecting privacy-sensitive objects from the recovered images using an object detector.

Recovering map images. Given the reference 3D models

Figure 2. Left to right: original image, reconstructions from SegLoc and PixLoc 3D models. We also show the yolov7 detections.

of ECMU, RC, and Indoor6, we train an inversion model [62] per dataset to recover images from these sparse SfM models. We learn inversion models for both SfM models where a 3D point is associated to a PixLoc descriptor or to a single SegLoc cluster label. We evaluate the models on reference 3D models of testing slices (unseen during training). Reconstruction metrics reported in Tab. 2 and visualizations provided in Fig. 2 show that SegLoc is both qualitatively and quantitatively more privacy-preserving than PixLoc (see *e.g.* in the top right example the reconstructed buildings and white car from PixLoc features).

Detecting sensitive areas. Reconstruction quality metrics are image level and do not evaluate what happens for particular objects. Therefore, we evaluate also discernability of sensitive classes (pedestrians, cars, indoor furniture) in the reconstructed images. To that end, we first evaluate the yolov7 [89] object detector on the original database images of ECMU and Indoor6. We use these detections as ground truth and try to detect the same classes from the reconstructed images. IoU metrics are reported in Tab. 7 and the corresponding bounding boxes shown in Fig. 2. While reconstructed images from SegLoc maps preserve the overall structure (which is encoded in the boundaries between classes), they do not contain recognizable details. On the contrary, PixLoc's maps allows the detector to recover fine details on previously unseen images. Note furthermore that even when an object is "reconstructed" in an image, the details such as color or brand is not discovered (see *e.g.* the white car reconstructed as a dark one in Fig. 2 top row).

The privacy of the query. Finally, some scenarios might require that the query sent to the server be privacy preserving. As query, SegLoc can either use the dense segmentation, part of it, or a single label representation while PixLoc uses the dense feature map. Given ECMU and Indoor6 database images, we train a dense inversion model adapted from [62] to invert the aforementioned input representations. In Tab. 8, we report reconstruction results and associated localization performances for both ECMU and Indoor6 testing sets. Using a one-hot query guarantees a high level of privacy while increasing the amount of encoded information facilitates localization at the cost of lowering privacy.

6. Conclusion

This paper explored to what extent robust segmentations based on a set of clusters may be used as an alternative intermediate representation for visual localization. Given the increasing concerns about privacy and storage requirements, such representations promise a competitive discriminativeness-privacy-memory trade-off. To address this, we proposed a novel method SegLoc that jointly learns global image descriptor and dense local representations by uncovering underlying clusters in a weakly-supervised manner. Such classes enable 3D representations of the environment that are an order of magnitude lighter than feature descriptor-based 3D maps. Despite the loss of information induced by using segmentations with a finite number of classes, we show that our method comes close to the performance of state-of-the-art feature based-methods on outdoor and indoor environments. Furthermore, we explicitly establish a connection between robust segmentation-based localization and privacy-preserving localization, showing that our representations offer an excellent trade-off between pose accuracy, privacy preservation, and memory requirements, opening new perspectives for visual localization.

Acknowledgments. Maxime and Torsten received funding from NAVER LABS Europe.

References

[1] Asha Anoosheh, Torsten Sattler, Radu Timofte, Marc Pollefeys, and Luc Van Gool. Night-to-day Image Translation for Retrieval-based Localization. In *ICRA*, 2019. 2

[2] Relja Arandjelović, , and Andrew Zisserman. Visual Vocabulary with a Semantic Twist. In *ACCV*, 2014. 2

[3] Relja Arandjelović, Petr Gronát, Akihiko Torii, Tomáš Pajdla, and Josef Sivic. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In *CVPR*, 2016. 4, 5, 6

[4] Clemens Arth, Daniel Wagner, Manfred Klopschitz, Arnold Irschara, and Dieter Schmalstieg. Wide Area Localization on Mobile Phones. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* , 2009. 1

[5] Hernan Badino, Daniel Huber, and Takeo T. Kanade. Visual Topometric Localization. In *IEEE Intelligent Vehicles Symposium (IVS)*, 2011. 5, 6

[6] Simon Baker and Iain Matthews. Lucas-Kanade 20 Years On: A Unifying Framework. *International Journal of Computer Vision (IJCV)*, 56(3):221–255, 2004. 5

[7] Assia Benbihi, Stéphanie Arravechia, Matthieu Geist, and Cédric Pradalier. Image-Based Place Recognition on Bucolic Environment Across Seasons From Semantic Edge Description . In *ICRA*, 2020. 2

[8] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. DSAC - Differentiable RANSAC for Camera Localization. In *CVPR*, 2017. 2

[9] Eric Brachmann and Carsten Rother. Expert Sample Consensus Applied to Camera Re-Localization. In *ICCV*, 2019. 1

[10] Eric Brachmann and Carsten Rother. Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC, 2020. 1

[11] 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. 6, 7

[12] Dylan Campbell, Liu Liu, and Stephen Gould. Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization. In *European Conference on Computer Vision*, pages 244–261. Springer, 2020. 2

[13] Bingyi Cao, André Araujo, and Jack Sim. Unifying Deep Local and Global Features for Image Search. In *ECCV*, 2020. 5, 6

[14] Tommaso Cavallari, Stuart Golodetz, Nicholas A. Lord, Julien Valentin, Luigi Di Stefano, and Philip H.S. Torr. On-The-Fly Adaptation of Regression Forests for Online Camera Relocalisation. In *CVPR*, 2017. 1

[15] Kunal Chelani, Fredrik Kahl, and Torsten Sattler. How Privacy-Preserving Are Line Clouds? Recovering Scene Details From 3D Lines. In *CVPR*, 2021. 2, 3

[16] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the Best Pooling Strategy for Visual Semantic Embedding. In *CVPR*, 2021. 4

[17] Deeksha Dangwal, Vincent T. Lee, Hyo Jin Kim, Meghan Shen, Tianwei andc Cowan, Rajvi Shah, Caroline Trippel, Brandon Reagen, Timothy Sherwood, Vasileios Balntas, Armin Alaghi, and Eddy Ilg. Analysis and Mitigations of Reverse Engineering Attacks on Local Feature Descriptors. In *BMVC*, 2021. 7

[18] Kamran Ghasedi Dizaji, Amirhossein Herandi, Cheng Deng, Weidong Cai, and Heng Huang. Deep Clustering via Joint Convolutional Autoencoder Embedding and Relative Entropy Minimization. In *ICCV*, 2017. 3

[19] Tien Do, Ondrej Miksik, Joseph DeGol, Hyun Soo Park, and Sudipta N Sinha. Learning To Detect Scene Landmarks for Camera Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11132–11142, 2022. 2, 5, 6, 7

[20] Alexey Dosovitskiy and Thomas Brox. Generating Images with Perceptual Similarity Metrics based on Deep Networks. In *NeurIPS*, 2016. 3

[21] Alexey Dosovitskiy and Thomas Brox. Inverting Visual Representations with Convolutional Networks. In *CVPR*, 2016. 3

[22] Mihai Dusmanu, Johannes L. Schönberger, Sudipta N. Sinha, and Marc Pollefeys. Privacy-Preserving Image Features via Adversarial Affine Subspace Embeddings. In *CVPR*, 2021. 2, 3, 6, 7

[23] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct Sparse Odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 40(3):611–625, 2017. 2, 5

[24] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 5

[25] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. In *CVPR*, 2016. 6

[26] Sourav Garg, Niko Suenderhauf, and Michael Milford. LoST? Appearance-Invariant Place Recognition for Opposite Viewpoints using Visual Semantics. 2018. 2

[27] Marcel Geppert, Viktor Larsson, Johannes L Schönberger, and Marc Pollefeys. Privacy Preserving Partial Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17337–17347, 2022. 3

[28] Marcel Geppert, Viktor Larsson, Pablo Speciale, Johannes L. Schönberger, and Marc Pollefeys. Privacy Preserving Structure-from-Motion. In *ECCV*, 2020. 2, 3

[29] Hugo Germain, Guillaume Bourmaud, and Vincent Lepetit. Sparse-to-Dense Hypercolumn Matching for Long-Term Visual Localization. In *3DV*, 2019. 1

[30] Ben Harwood, Vijay Kumar BG, Gustavo Carneiro, Ian Reid, and Tom Drummond. Smart mining for deep metric learning. In *CVPR*, pages 2821–2829, 2017. 5

[31] Lionel Heng, Benjamin Choi, Zhaopeng Cui, Marcel Geppert, Sixing Hu, Benson Kuan, Peidong Liu, Rang Nguyen, Ye Chuan Yeo, Andreas Geiger, Gim Hee Lee, Marc Pollefeys, and Torsten Sattler. Project AutoVision: Localization and 3D Scene Perception for an Autonomous Vehicle with a Multi-Camera System. In *ICRA*, 2019. 1

[32] Hanjiang Hu, Zhijian Qiao, Ming Cheng, Zhe Liu, and Hesheng Wang. DASGIL: Domain Adaptation for Semantic and Geometric-aware Image-based Localization. *IEEE Transactions on Image Processing (TIP)*, 30:1342–1353, 2021. 2, 5, 6

[33] Hanjiang Hu, Hesheng Wang, Zhe Liu, Chenguang Yang, Weidong Chen, and Le Xie. Retrieval-based Localization Based on Domain-invariant Feature Learning under Changing Environments. In *IROS*, 2019. 2, 5, 6

[34] Martin Humenberger, Yohann Cabon, Nicolas Guerin, Julien Morat, Jérôme Revaud, Philippe Rerole, Noé Pion, Cesar de Souza, Vincent Leroy, and Gabriela Csurka. Robust Image Retrieval-based Visual Localization using Kapture. arXiv:2007.13867, 2020. 1

[35] Arnold Irschara, Christopher Zach, Jan-Michael Frahm, and Horst Bischof. From Structure-from-Motion Point Clouds to Fast Location Recognition. In *CVPR*, 2009. 1

[36] Mohammed Jabi, Marco Pedersoli, Amar Mitiche, and Ismail Ben Ayed. Deep clustering: On the link between discriminative models and k-means. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1887–1896, 2019. 3

[37] Alex Kendall and Roberto Cipolla. Geometric Loss Functions for Camera Pose Regression with Deep Learning. In *CVPR*, 2017. 1

[38] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: a Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *ICCV*, 2015. 1, 2, 5, 6, 7

[39] Jan Knopp, Josef Sivic, and Tomáš Pajdla. Avoiding Confusing Features in Place Recognition. In *ECCV*, 2010. 2

[40] Nikolay Kobyshev, Hayko Riemenschneider, and Luc Van Gool. Matching Features Correctly through Semantic Understanding. In *3DV*, 2014. 2

[41] Mans Larsson, Erik Stenborg, Lars Hammarstrand, Marc Pollefeys, Torsten Sattler, and Fredrik Kahl. A Cross-Season Correspondence Dataset for Robust Semantic Segmentation. In *CVPR*, 2019. 2, 5

[42] Mans Larsson, Erik Stenborg, Carl Toft, Lars Hammarstrand, Torsten Sattler, and Fredrik Kahl. Fine-Grained Segmentation Networks: Self-Supervised Segmentation for Improved Long-Term Visual Localization. In *ICCV*, 2019. 1, 2, 4, 7

[43] Kenneth Levenberg. A Method for the Solution of Certain Non-linear Problems in Least Squares. *Quarterly of Applied Mathematics*, 2(2):164–189, 1944. 5

[44] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical Contrastive Learning of Unsupervised Representations. In *ICLR*, 2021. 4

[45] Yunpeng Li, Noah Snavely, and Dan Huttenlocher. Location Recognition Using Prioritized Feature Matching. In *ECCV*, 2010. 1

[46] Yunpeng Li, Noah Snavely, Dan Huttenlocher, and Pascal Fua. Worldwide Pose Estimation Using 3D Point Clouds. In *ECCV*, 2012. 1

[47] Zhengqi Li and Noah Snavely. MegaDepth: Learning Single-View Depth Prediction from Internet Photo. In *CVPR*, 2018. 7

[48] Konstantinos-Nektarios Lianos, Johannes L. Schönberger, Marc Pollefeys, and Torsten Sattler. VSO: Visual Semantic Odometry. In *ECCV*, 2018. 1, 2

[49] Hyon Lim, Sudipta N. Sinha, Michael F. Cohen, Matt Uyttendaele, and H. Jin Kim. Real-time Monocular Image-based 6-DoF Localization. *International Journal of Robotics Research*, 34(4–5):476–492, 2015. 1

[50] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *International Journal of Robotics Research*, 36(1):3–15, 2017. 5, 6

[51] Aravindh Mahendran and Andrea Vedaldi. Understanding Deep Image Representations by Inverting Them. In *CVPR*, 2015. 2

[52] Donald W. Marquardt. An Algorithm for Least-squares Estimation of Nonlinear Parameters. *Journal of Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963. 5

[53] Sven Middelberg, Torsten Sattler, Ole Untzelmann, and Leif Kobbelt. Scalable 6-DoF Localization on Mobile Devices. In *ECCV*, 2014. 1

[54] Arsalan Mousavian, Jana Košecká, and Jyh-Ming Lien. Semantically Guided Location Recognition for Outdoors Scenes. In *ICRA*, 2015. 2

[55] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D. Tardos. ORB-SLAM: a Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 2

[56] Tayyab Naseer, Gabriel L. Oliveira, Thomas Brox, and Wolfram Burgard. Semantics-aware Visual Localization under Challenging Perceptual Conditions. In *ICRA*, 2017. 2

[57] Hyeonwoo Noh, André Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-Scale Image Retrieval with Attentive Deep Local Features. In *ICCV*, 2017. 6

[58] Valerio Paolicelli, Antonio Tavera, Gabriele Berton, Carlo Masone, and Barbara Caputo. Learning Semantics for Visual Place Recognition through Multi-Scale Attention. arXiv:2201.09701, 2022. 2

[59] Nathan Piasco, Désiré Sidibé, Valérie Gouet-Brunet, and Cédric Demonceaux. Learning Scene Geometry for Visual Localization in Challenging Conditions. In *ICRA*, 2019. 2

[60] Nathan Piasco, Désiré Sidibé, Valérie Gouet-Brunet, and Cédric Demonceaux. Improving Image Description with Auxiliary Modality for Visual Localization in Challenging Conditions. *International Journal of Computer Vision (IJCV)*, 129:185–202, 2021. 2

[61] Noé Pion, Martin Humenberger, Gabriela Csurka, Yohann Cabon, and Torsten Sattler. Benchmarking Image Retrieval for Visual Localization. In *3DV*, 2020. 5

[62] Francesco Pittaluga, Sanjeev J Koppal, Sing Bing Kang, and Sudipta N Sinha. Revealing scenes by inverting structure from motion reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 145–154, 2019. 1, 2, 6, 7, 8

[63] Horia Porav, Tom Bruls, and Paul Newman. Don't Worry About the Weather: Unsupervised Condition-Dependent Domain Adaptation. In *IEEE Intelligent Transportation Systems Conference*, 2019. 2

[64] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. In *ICCV*, 2021. 3, 7

[65] Jerome Revaud, Jon Almazan, Rafael Sampaio de Rezende, and Cesar Roberto de Souza. Learning with Average Precision: Training Image Retrieval with a Listwise Loss. In *ICCV*, 2019. 5, 6

[66] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *CVPR*, 2019. 1

[67] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. Back to the Feature: Learning Robust Camera Localization From Pixels To Pose. In *CVPR*, 2021. 2, 5, 7

[68] Torsten Sattler, Michal Havlena, Filip Radenović, Konrad Schindler, and Marc Pollefeys. Hyperpoints and Fine Vocabularies for Large-Scale Location Recognition. In *ICCV*, 2015. 1

[69] 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 (PAMI)*, 39(9):1744–1756, 2017. 1

[70] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DoF Outdoor Visual Localization in Changing Conditions. In *CVPR*, 2018. 5, 6

[71] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixé. Understanding the Limitations of CNN-based Absolute Camera Pose Regression. In *CVPR*, 2019. 1

[72] Matthias Schorghuber, Daniel Steininger, Yohann Cabon, Martin Humenberger, and Margrit Gelautz. SLAMANTIC - Leveraging Semantics to Improve VSLAM in Dynamic Environments. In *ICCV Workshops* , Oct 2019. 2

[73] Stephen Se, David G. Lowe, and J. J. Little. Global Localization using Distinctive Visual Features. In *IROS*, 2002. 1

[74] Tianxin Shi, Shuhan Shen, Xiang Gao, and Lingjie Zhu. Visual Localization using Sparse Semantic 3D Map. 2019. 2

[75] Mikiya Shibuya, Shinya Sumikura, and Ken Sakurada. Privacy Preserving Visual SLAM. In *ECCV*, 2020. 2

[76] 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 *CVPR*, 2013. 1

[77] Pablo Speciale, Johannes L. Schönberger, Sing Bing Kang, Sudipta N. Sinha, and Marc Pollefeys. Privacy Preserving Image-Based Localization. In *CVPR*, 2019. 2

[78] Pablo Speciale, Johannes L. Schönberger, Sudipta N. Sinha, and Marc Pollefeys. Privacy Preserving Image Queries for Camera Localization. In *ICCV*, 2019. 2, 3

[79] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomáš Pajdla, and Torii Akihiko. InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 43(4):1293–1307, 2021. 1

[80] Li Tang, Yue Wang, Qianhui Luo, Xiaqing Ding, and Rong Xiong. Adversarial Feature Disentanglement for Place Recognition Across Changing Appearance. In *ICRA*, 2020. 2

[81] Carl Toft, Will Maddern, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, et al. Long-term visual localization revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 5, 6

[82] Carl Toft, Carl Olsson, and Fredrik Kahl. Long-term 3D Localization and Pose from Semantic Labellings. In *ICCV Workshops* , 2017. 1, 2

[83] Carl Toft, Erik Stenborg, Lars Hammarstrand, Lucas Brynte, Marc Pollefeys, Torsten Sattler, and Fredrik Kahl. Semantic Match Consistency for Long-term Visual Localization. In *ECCV*, 2018. 2

[84] Akihiko Torii, Relja Arandjelović, Josef Sivic, Masatoshi Okutomi, and Tomáš Pajdla. 24/7 Place Recognition by View Synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 40(2):257–271, 2018. 5, 6

[85] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748, 2018. 4

[86] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, 9, 2008. 4

[87] Lukas Von Stumberg, Patrick Wenzel, Qadeer Khan, and Daniel Cremers. Gn-net: The gauss-newton loss for multi-weather relocalization. *IEEE Robotics and Automation Letters*, 5(2):890–897, 2020. 2

[88] Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based Localization Using LSTMs for Structured Feature Correlation. In *ICCV*, 2017. 2

[89] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. 8

[90] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. Multi-Similarity Loss with General Pair Weighting For Deep Metric Learning. In *CVPR*, 2019. 5

[91] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo Geolocation with Convolutional Neural Networks. In *ECCV*, 2016. 2

[92] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised Deep Embedding for Clustering Analysis. In *ICML*, 2016. 3, 4

[93] Zhe Xin, Yinghao Cai, Tao Lu, Xiaoxia Xing, Shaojun Cai, Jixiang Zhang, Yiping Yang, and Yanqing Wang. Localizing Discriminative Visual Landmarks for Place Recognition. In *ICRA*, 2019. 5, 6

[94] Xin Yu, Sagar Chaturvedi, Chen Feng, Yuichi Taguchi, Teng-Yok Lee, Clinton Fernandes, and Srikumar Ramalingam. VLASE: Vehicle Localization by Aggregating Semantic Edges. In *IROS*, 2018. 2

[95] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[96] Ziqiang Zheng, Yang Wu, Xinran Han, and Jianbo Shi. ForkGAN: Seeing into the Rainy Night. In *ECCV*, 2020. 2

[97] Huabing Zhou, Jiayi Ma, Chiu C. Tan, Yanduo Zhang, and Haibin Ling. Cross-Weather Image Alignment via Latent Generative Model With Intensity Consistency. *IEEE Transactions on Image Processing (TIP)*, 29(3):5216–5228, 2020. 2

[98] Qunjie Zhou, Sergio Agostinho, Aljosa Osep, and Laura Leal-Taixe. Is Geometry Enough for Matching in Visual Localization? *arXiv preprint arXiv:2203.12979*, 2022. 2, 6, 7