Visual Localization by Learning Objects-of-Interest Dense Match Regression

Philippe Weinzaepfel Gabriela Csurka Yohann Cabon Martin Humenberger
NAVER LABS Europe
firstname.lastname@naverlabs.com

Abstract

We introduce a novel CNN-based approach for visual localization from a single RGB image that relies on densely matching a set of Objects-of-Interest (OOIs). In this paper, we focus on planar objects which are highly descriptive in an environment, such as paintings in museums or logos and storefronts in malls or airports. For each OOI, we define a reference image for which 3D world coordinates are available. Given a query image, our CNN model detects the OOIs, segments them and finds a dense set of 2D-2D matches between each detected OOI and its corresponding reference image. Given these 2D-2D matches, together with the 3D world coordinates of each reference image, we obtain a set of 2D-3D matches from which solving a Perspective-n-Point problem gives a pose estimate. We show that 2D-3D matches for reference images, as well as OOI annotations can be obtained for all training images from a single instance annotation per OOI by leveraging Structure-from-Motion reconstruction. We introduce a novel synthetic dataset, VirtualGallery, which targets challenges such as varying lighting conditions and different occlusion levels. Our results show that our method achieves high precision and is robust to these challenges. We also experiment using the Baidu localization dataset captured in a shopping mall. Our approach is the first deep regression-based method to scale to such a larger environment.

1. Introduction

Visual localization consists in estimating the 6-DoF camera pose from a single RGB image within a given area, also referred to as map. This is particularly valuable if no other localization technique is available, *e.g.* in GPS-denied environments such as indoor locations. Interesting applications include robot navigation [8], self-driving cars and augmented reality (AR) [6, 22]. The main challenges include large viewpoint changes between query and train images, incomplete maps, regions without valuable information (*e.g.* textureless surfaces), symmetric and repetitive elements, varying lighting conditions, structural changes, dynamic objects (*e.g.* people), and scalability to large areas.

Traditional structure-based approaches [17, 18, 24, 26, 27, 34] use feature matching between query image and map, coping with many of the mentioned challenges. However, covering large areas from various viewpoints and under different conditions is not feasible in terms of processing time and memory consumption. To overcome this, image retrieval can be used to accelerate the matching for large-scale localization problems [10, 29, 35]. Recently, methods based on deep learning [2, 16] have shown promising results. PoseNet [16] and its improvements [4, 7, 15, 39] proceed by directly regressing the camera pose from input images. Even if a rough estimate can be obtained, learning a precise localization seems too difficult or would require a large amount of training data to cover diversities both in terms of locations and intrinsic camera parameters. More interestingly, Brachmann *et al.* [2, 3] learn dense 3D scene coordinate regression and solve a Perspective-n-Point (PnP) problem for accurate pose estimation. The CNN is trained end-to-end thanks to a differentiable approximate formulation of RANSAC, called DSAC. Scene coordinate regression-based methods obtain outstanding performance in static environments, *i.e.*, without changes in terms of objects, occlusions or lighting conditions. However, they are restricted in terms of scene scale.

The above mentioned challenges of visual localization become even more important in very large and dynamic environments. Essential assumptions, such as static and unchanged scenes are violated and the maps become outdated quickly while continuous retraining is challenging. This motivated us to design an algorithm inspired by advances on instance recognition [11] that relies on stable, predefined areas, and that can bridge the gap between precise localization and long-term stability in very vivid scenarios. We propose a novel deep learning-based visual localization method that proceeds by finding a set of dense matches between some Objects-of-Interest (OOIs) in query images and the corresponding reference images, *i.e.*, a canonical view of the object for which a set of 2D-3D correspondences is available. We define an Object-of-Interest as a discriminative area within the 3D map which can be reliably detected

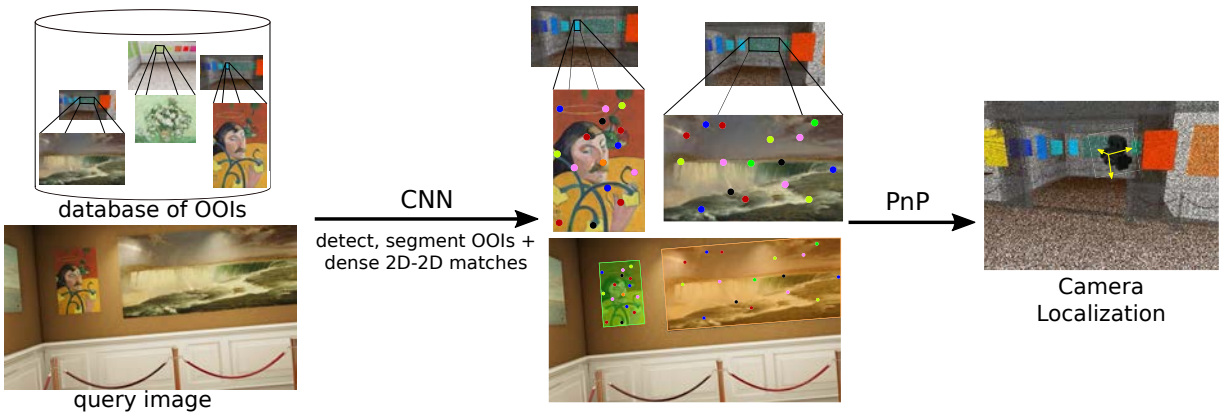

Figure 1. Overview of our pipeline. Given a query image, we use a CNN to first, detect and segment a predefined list of Objects-of-Interest, and second, to regress dense matches to their reference images (dots in color). These reference images contain a dense 2D pixels to 3D coordinates mapping. We thus obtain 2D-3D matches by transitivity and solve a PnP problem to obtain the camera localization.

from multiple viewpoints, partly occluded, and under various lighting conditions. Figure 1 shows an overview of our pipeline. Our model relies on DensePose [9], a recent extension of Mask R-CNN [11] that not only detects and segments humans, but also regresses dense matches between pixels in the image and a mesh surface. In our case, we use it (i) to detect and segment OOIs, and (ii) to obtain a dense set of 2D-2D matches between the detected OOIs and their reference images. Given these matches, together with the 2D-3D correspondences of the reference images, we obtain a set of 2D-3D matches from which camera localization is obtained by solving a PnP problem using RANSAC.

Our method is carefully designed to tackle open challenges of visual localization, it has several advantages, and few limitations with respect to the state of the art. First, reasoning in 2D allows to train the model from few training data: we can artificially generate a rich set of viewpoints for each object with homography data augmentation and we can achieve robustness to lighting changes with color jittering. Second, our method can handle dynamic scenes as long as the OOIs remain static. For instance, one can accurately estimate the pose in a museum with visitors present, even if training data does not contain any humans. Third, if some OOIs are moved, we do not need to retrain the whole network as pose and scene regression-based approaches would require, but we only need to update the 2D-3D mapping of the reference images. Fourth, our method focuses on discriminative objects and thus avoids ambiguous textureless areas. Fifth, our method can scale up to large areas and high numbers of OOIs as object detectors can segment thousands of categories [13]. One clear limitation of our method is that query images without any OOI cannot be localized. However, in many applications such as AR navigation, OOIs are present most of the time and local pose tracking (*e.g.* visual-inertial odometry [1]) can be used in-between. OOI detection is interesting by itself in such an application, *e.g.* to display metadata on paintings in a museum or on shops in malls and airports. Furthermore, in a complex real-world application, OOIs can be used to more easily guide the user to localize successfully. Commands such as 'Take a picture of the closest painting' might be easier to understand than 'Take a picture with sufficient visual information'.

In this paper, we restrict OOIs to planar objects which are common in real-world applications: paintings, posters, store fronts or logos are frequent in environments where localization is challenging such as shopping malls, airports or museums. While the method can be generalized to non-planar objects, considering planar OOIs has several advantages, in addition to homography data augmentation. First, the transformation between any instance of the OOI in a training image and its reference image is a homography, thus allowing to easily propagate dense sets of matches using a few correspondences only. Second, the mapping between 2D pixels in the reference image and 3D world coordinates can be built from a small set of images since planes can be robustly reconstructed in 3D. Finally, planar OOIs allow us to reason in 2D, removing one degree of freedom compared to 3D coordinate regression. Additionally, we show that our method can be used with a minimal amount of manual annotations, being one instance segmentation for each (planar) OOI in any training image.

We demonstrate the strength of our approach on two challenging datasets. The first one is a newly introduced synthetic dataset, VirtualGallery, which represents an art gallery. The second one is the Baidu localization dataset [33] captured in a shopping mall with only a few viewpoints at training and some changes in the scenes at testing, showing the applicability of our approach in complex real-world environments. While our method is accurate on VirtualGallery (even with different lighting conditions and occlusions) achieving a median error of less than 3cm and 0.5°, it also scales up to larger environment such as the Baidu localization dataset. In contrast, deep state-of-the-art regression-based approaches fail in such scenarios.

2. Related Work

Most methods for visual localization can be categorized into four types of approaches: structure-based methods, image retrieval-based methods, pose regression-based methods and coordinate regression-based methods.

Structure-based methods [17, 18, 24, 26, 27, 34] use descriptor matching (*e.g*. SIFT [21]) between 3D points of the map associated with local descriptors and keypoint descriptors in the query image. However, these point features are not able to create a representation which is sufficiently robust to challenging real-world scenarios such as different weather, lighting or environmental conditions. Additionally, they lack the ability to capture global context and require robust aggregation of hundreds of points in order to form a consensus to predict a pose [41].

Image retrieval-based methods [10, 29, 35, 36, 37] match the query image with the images of the map using global descriptors or visual words to obtain the image location from the top retrieved images. The retrieved locations can further be used to either limit the search range within large maps of structure-based approaches [5, 28], or to directly compute the pose between retrieved and query images [42]. These methods allow to speed-up search in large environments, but share similar drawbacks when using structure-based methods for accurate pose computation. InLoc [35] shows recent advances in image retrieval-based methods leveraging dense information. It uses deep features to first retrieve the most similar images, and then to estimate the camera pose within the map. A drawback is the heavy processing load and the need of accurate dense 3D models.

Pose regression-based methods [4, 16, 39] were the first deep learning approaches trained end-to-end for visual localization. They proceed by directly regressing the 6-DoF camera pose from the query image using a CNN, following the seminal PoseNet approach [16]. The method has been extended in several ways, by leveraging video information [7], recurrent neural networks [39], hourglass architecture [23] or a Bayesian CNN to determine the uncertainty of the localization [14]. More recently, Kendall *et al*. [15] replace the naive L2 loss function by a novel loss that relies on scene geometry and reprojection error. Brahmbhatt *et al*. [4] additionally leverage the relative pose between pairs of images at training. Overall, pose regression-based methods have shown robustness to many challenges but remain limited both in accuracy and scale.

Scene coordinate regression-based methods [2, 32, 38] proceed by regressing dense 3D coordinates and estimating the pose using a PnP solver with RANSAC. While random forests were used in the past [32, 38], Brachmann *et al*. [2] recently obtain an extremely accurate pose estimation by training a CNN to densely regress 3D coordinates. They additionally introduce a differentiable approximation of RANSAC, called DSAC, allowing end-to-end training for visual localization at the cost of multi-step training, with depth data required for the first step. DSAC++ [3] is an improvement of the method where real depth data is not mandatory and can be replaced by a depth prior. The network is still trained in multiple steps: first based on depth data or the prior; second, based on minimization of the reprojection error; and finally based on camera localization error using the DSAC module. DSAC++ obtains outstanding performance on datasets of relatively small scale with constant lighting conditions and little dynamics. However, the method does not converge for larger scenes. In contrast, our method is built upon an object detection pipeline, which scales to large environments. Compared to DSAC++, since we consider planar objects, we can regress 2D matches instead of 3D coordinates, which removes one degree of freedom, and allows homography data augmentation.

3. Visual Localization Pipeline

In this section, we first present an overview of our approach in Section 3.1. Next, we introduce details about our CNN model for segmenting OOIs and dense matching (Section 3.2). Finally, Section 3.3 explains how SfM maps can be leveraged to train our approach from weak supervision.

3.1. Visual localization from detected OOIs

Let $\mathcal{O}$ be the set of OOIs, and $|\mathcal{O}|$ the number of OOI classes. Our method relies on reference images: each OOI $o \in \mathcal{O}$ is associated with a canonical view, *i.e*., an image $\mathcal{I}_o$ where o is fully visible. We assume for now that each OOI is unique in the environment and that the mapping M_o between 2D pixels $\mathbf{p}'$ in the reference image and the corresponding 3D points in the world $M_o(\mathbf{p}')$ is known.

Given a query image, our CNN outputs a list of detections. Each detection consists of (a) a bounding box with a class label o, *i.e*., the id of the detected OOI, and a confidence score, (b) a segmentation mask, and (c) a set of 2D-2D matches $\{\mathbf{q} \rightarrow \mathbf{q}'\}$ between pixels $\mathbf{q}$ in the query image and pixels $\mathbf{q}'$ in the reference image $\mathcal{I}_o$ of the object-of-interest o, see Figure 1. By transitivity, we apply the mapping M_o to the matched pixels in the reference image and obtain for each detection a list of matches between 2D pixels in the query image and 3D points in the world coordinates: $\{\mathbf{q} \rightarrow \mathcal{M}_o(\mathbf{q}')\}$.

Given the list of 2D-3D matches for all detections in the query image and the intrinsic camera parameters, we estimate the 6-DoF camera pose by solving a Perspective-n-Point problem using RANSAC.

Note that if there is no detection in a query image, we do not have matches and, thus, we are not able to perform localization. However, in venues such as museums or airports, OOIs can be found in most of the images. Moreover, in real-world applications, localization is used in conjunc-

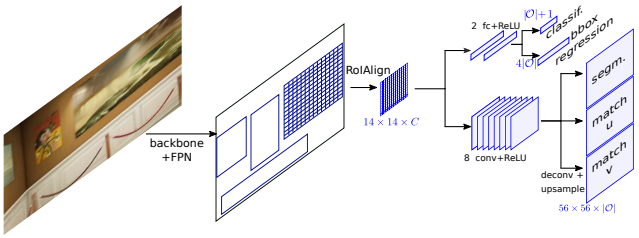

Figure 2. Overview of the network architecture to detect and segment OOIs as well as to obtain dense matches with respect to the reference images.

tion with local pose tracking and, thus, precision of the system is more important than coverage.

In summary, the only components learned in our approach are detection and dense matching between a query image and the reference images of the OOIs.

Handling non-unique OOIs. So far, we assumed that each OOI is unique, *i.e.*, it appears only once in the environment. While most OOIs are highly discriminative, some of them can have multiple instances in an environment, *e.g.* logos in a shopping mall. In this case, as a detector cannot differentiate them, we aggregate the same OOIs in a single class and train the model with a common reference image. The mapping M_o is not unique anymore since there exist multiple 3D coordinate candidates for the same reference image, *i.e.*, one per instance present in the environment. Given that in practice we have a limited number of duplicates per OOI and a limited number of detections per image, we solve the PnP problem for the best combination of possible coordinates using geometric plausibility checks. In detail, we minimize the sum of reprojection errors of the OOI 3D center points in the query image. Ideally, the 3D center point lies in the middle of the segmented detection. We ignore noise due to incomplete detections of OOIs.

3.2. Detecting OOIs and matching to references

We follow DensePose [9], an extension of Mask R-CNN [11], designed for finding dense correspondences between any point on a human body and the corresponding point on the surface mesh. For each box generated by the region proposal networks (RPN) [25], C-dimensional convolutional features are estimated with a fixed resolution of 14×14 using the RoIAlign layer. The Feature Pyramid Networks (FPN) improvement [19] is used to better handle small objects. Box features are fed to two branches. One of them is designed to predict the class score (human *vs*. non-human in their case) and to perform class-specific box coordinate regression, following Faster R-CNN design [25]. The other branch, which is fully-convolutional, predicts a per-pixel human body part label and per-pixel correspondences (*i.e.*, two coordinates) with the corresponding mesh surface. In practice, the CNN predicts segmentation and correspondences on a dense grid of size 56×56 in each box which is then interpolated to obtain a per-pixel prediction.

In our case, given the box features after RoIAlign, we use a similar CNN model as DensePose, see Figure 2. One branch predicts OOI scores and regresses bounding boxes, the only difference being that we have $|\mathcal{O}| + 1$ classes (including the background class) instead of 2. The second branch predicts different tasks: (a) binary segmentation for each OOI, (b) OOI-specific u and v reference image coordinate regression. At training time several losses are combined. In addition to the FPN loss for box proposals, we use the cross-entropy loss for box classification. For the ground-truth class we use a smooth-L1 loss on its box regressor, the cross-entropy loss for the 56×56 mask predictor, and smooth-L1 losses for the $u-$ and $v-$regressors. Training requires a ground-truth mask and matches for every pixel. In Section 3.3, we explain how such annotations can be automatically obtained from minimal annotation. At test time, we keep the box detections with classification score above 0.5 and keep the matches for the points within the segmentation mask.

Implementation details. We use FPN [19] with both ResNet50 [12] and ResNeXt101-32x8d [40] backbones. The branch that predicts segmentation and match regression follows the Mask R-CNN architecture: it consists of 8 convolutions and ReLU layers before a final convolutional layer of each task. We train the network for 500k iterations, starting with a learning rate of 0.00125 on 1 GPU and dividing it by 10 after 300k and 420k iterations. We use SGD as optimizer with a momentum of 0.9 and a weight decay of 0.0001. To make all regressors proceeding at the same scale, we normalize the reference coordinates in $[0, 1]$.

Data augmentation. As our CNN regresses matches only in 2D, we apply homography data augmentation on all input images. To generate plausible viewpoints, we compute a random displacement limited by 33% of the image size for each of the 4 corners and fit the corresponding homography. We do not use flip data augmentation as OOIs (logos, paintings, posters) are left-right ordered. We study the impact of using color jittering (brightness, contrast, saturation) for robustness against changing lighting conditions in our experiments (Section 5).

3.3. Weakly-supervised OOI annotation

Key of our approach is the minimal amount of manual annotations required, thanks to a propagation algorithm that leverages a SfM reconstruction obtained with COLMAP [30]. The only annotation to provide is one segmentation mask for each planar OOI, see the blue mask in the middle of Figure 3. The reference image for this OOI is defined by the annotated mask.

Using the set of 2D to 3D matches from SfM, we label 3D points that match with 2D pixels in the annotated OOI segmentation, see blue lines and dots in Figure 3. We

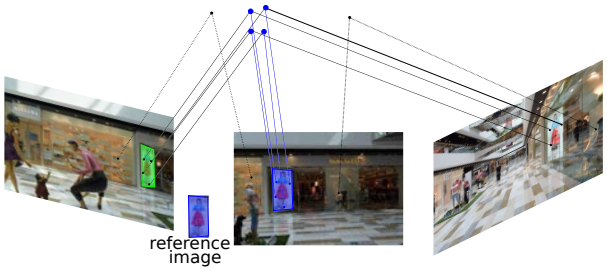

Figure 3. The bounding box of the manual mask annotation in blue (middle frame) defines the reference image for the OOI. We propagate the OOI to the other training images (green mask in the left image) using the 3D positions of keypoints within the annotation (blue mask). Propagation may fail if not enough matches are present (right image).

propagate the label to all training images which contain observations of these 3D points. If there are at least 4 matches in an image, we can fit a homography given that the OOI o is planar. For more robustness to noise, we only consider regions with a minimum of 7 matches and use RANSAC-based homography estimation. This homography is used to propagate the mask annotation as well as the dense 2D-2D matches, see the green mask on the left image of Figure 3.

Either due to a low number of matches leading to missing propagation, see right image in Figure 3, or because of noisy matches in the SfM model, the propagation is not always successful. Fortunately, the CNN model is, to some extent, robust against noisy or missing labels.

Handling non-unique OOIs. For non-unique OOIs, such as a logo appearing multiple times in a shopping mall, we annotate one segmentation mask for each instance of the OOI and we apply our propagation method on each instance independently. As mentioned above, it would be impossible for any detector to differentiate between the different instances. Thus, we merge them into a single OOI class and compute a homography between the reference images of the different instances and the dedicated main reference image using SIFT descriptor [21] matches. As main reference for the class (OOI), we select the reference image with the highest number of 3D matches. Since the regressed 2D-2D matches correspond to the main class, we additionally apply a perspective transform using the computed intra-class homographies, see Section 3.1.

4. Datasets

The VirtualGallery dataset. We introduce a new synthetic dataset to study the applicability of our approach and to furthermore measure the impact of varying lighting conditions and occlusions on different localization methods. It consists of a scene containing 3-4 rooms, see Figure 4 (left), in which 42 free-of-use famous paintings[1] are placed on the

[1] https://images.nga.gov/

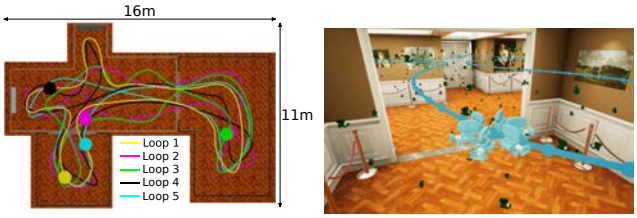

Figure 4. *Left:* floorplan of the art gallery with the different training loops. *Right:* a training loop with the 6 cameras at a fixed height (cyan) and test cameras at different plausible places.

Figure 5. Test samples from VirtualGallery. *First row:* test images with various viewpoints. *Second and third rows:* different lighting conditions. *Fourth row:* different human densities (occlusions).

walls. The scene was created with the Unity software, allowing to extract ground-truth information such as depth, semantic and instance segmentations, 2D-2D and 2D-3D correspondences, together with the rendered images.

We consider a realistic scenario that simulates the scene captured by a robot for training and photos taken by visitors for testing. The camera setup consists of 6 cameras in a 360° configuration at a fixed height of 165cm, see Figure 4 (right). The robot follows 5 different loops inside the gallery with pictures taken roughly every 20cm, resulting in about 250 images for each camera, see Figure 4 (left).

At test time, we sample random positions, orientations and focal lengths, ensuring that viewpoints (a) are plausible and realistic (in terms of orientation, height and distance to the wall), and (b) span the entire scene. This covers the additional challenges of viewpoint changes and varying intrinsic camera parameters between training and test images.

To study robustness to lighting conditions, we generate the scene using 6 different lighting configurations with significant variations between them, both at training and test time, see second and third rows of Figure 5. To evaluate robustness to occluders such as visitors, we generate test images which contain randomly placed human body models, see last row of Figure 5. The test set con-

Figure 6. Examples of training images from the Baidu localization dataset with propagated mask annotations.

sists of 496 images that are rendered for each of the 6 lighting conditions and for 4 different densities of humans present in the scene (including the empty case). The dataset can be found at http://www.europe.naverlabs.com/Research/3D-Vision/Virtual-Gallery-Dataset.

Objects-of-Interest. We use each painting as an object-of-interest, each one being unique in the scene. We use the original image downloaded from the website as reference image. We obtain ground-truth segmentation masks and 2D-2D correspondences for each image using Unity. We also get the position and size where the painting is placed in the scene, thus providing the 2D-3D mapping functions. Note that, among the test images, 5 out of 496 do not contain any OOI, and 23 (respectively 48) additional ones have no OOI visible by more than 50% (respectively 80%).

The Baidu localization dataset [33]. It consists of images captured in a Chinese shopping mall covering many challenges for visual localization such as reflective and transparent surfaces, moving people and repetitive structures. It contains 689 images captured with DSLR cameras as training set and over 2000 cell phone photos taken by different users a few months later as test set. The test images contain significant viewpoint changes compared to the training images that are all taken parallel or perpendicular with respect to the main corridor. All images were semi-automatically registered to the coordinate system defined by a LIDAR scanner. We use the provided camera poses, both for training (3D reconstruction of the OOIs and annotation propagation) and testing (ground-truth evaluation of the results). We did not use the LIDAR data, even if it would potentially improve the 3D reconstruction quality of our OOIs.

Objects-of-Interest. We manually annotated one segmentation mask for 220 instances from 164 classes representing different types of planar objects such as logos or posters on storefronts. We then propagated these annotations to all training images, see Section 3.3. This real-world dataset is more challenging than VirtualGallery, thus, some OOI propagations are noisy. Figure 6 shows examples of training images with masks around the OOIs after propagation.

5. Experimental results

We evaluate our approach on VirtualGallery (Section 5.1) and on the Baidu localization dataset (Section 5.2).

5.1. Experiments on VirtualGallery

We use different train/test scenarios to evaluate some variants of our method and to study the robustness of state-of-the-art approaches to lighting conditions and occlusions. In the experiments below, we use the ground-truth correspondences obtained from Unity. We experimented with manual annotations and obtained similar performance.

Impact of data augmentation. We first study the impact of homography data augmentation at training. We train models with ResNet50 (resp. ResNext101) backbone on the first loop of the standard lighting condition, and report the percentages of successfully localized images with the standard lighting condition and no humans in Figure 7 (plain blue and dotted blue, resp. black, curves). Homography data augmentation significantly improves the performance, specifically for highly accurate localization: the ratio of localized images within 5cm and 5° increases from 25% to 69% with ResNet50 backbone. The impact is less significant at higher error thresholds with an improvement from 72% to 88% at 25cm and 5°. Homography data augmentation at training allows to generate more viewpoints of the OOIs and, thus, to better detect and match OOIs captured from unknown viewpoints at test time.

Impact of regressing dense 2D-2D matches. We now compare our approach that relies on 2D-2D dense correspondences to a few variants. First, we directly regress the 8-DoF homography parameters (OOIs-homog) between the detected OOI and its reference image and generate the dense set of 2D-2D matches afterwards. Second, we directly regress 3D world coordinates for each OOI instance (OOIs-2D-3D) without using the reference image. Performances with the same train/test protocol as above are reported in Figure 7. Regressing the 8 parameters of the homography leads to a drop of performance, only 70% of the images could be successfully localized within 25cm and 5°. CNNs have difficulties regressing parameters of transformations where small differences can lead to significant changes of the result. The 3D variant performs reasonably well, with roughly 70% of images localized within 10cm and 5°, and 81% within 25cm and 5°. However, our method with 2D reference images outperforms the 3D variant, specifically for low position error thresholds. Our 2D reference images ensure that all 3D points are on the planar object, while the 3D variant adds one extra and unnecessary degree of freedom. We finally replace the ResNet50 backbone by ResNeXt101-32x8d and obtain a more precise localization at low thresholds (8% additional images are localized within 5cm and 5°); the difference becomes marginal at higher thresholds.

Comparison to the state of the art. We now compare our approach to a SfM method (COLMAP [30, 31]), PoseNet with geometric loss [15], and DSAC++ trained with a 3D model [3]. All methods are trained on all loops from all

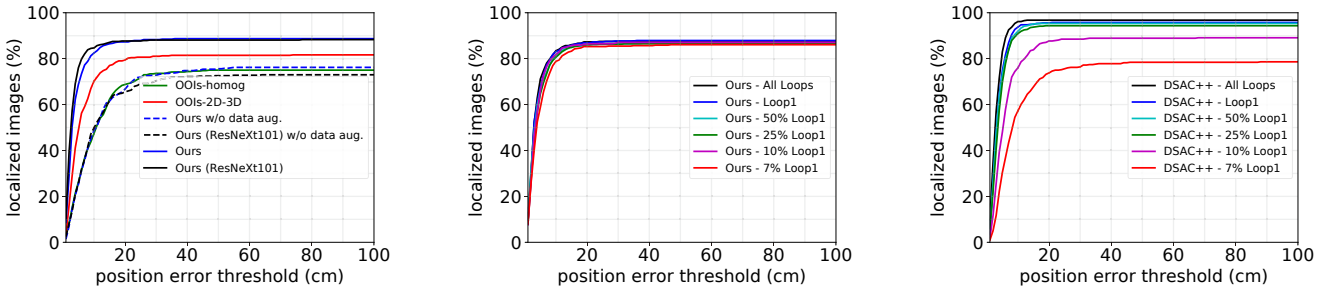

Figure 7. Percentages of localized images on the VirtualGallery test set with the standard lighting condition without humans for varying position error thresholds and a fixed orientation error threshold of 5°. *Left:* comparison between variants of our approach trained on the first loop of the standard lighting condition. *Middle and Right:* robustness to a lower amount of training data (from the standard lighting condition) for our approach (*Middle*) and DSAC++ (*Right*).

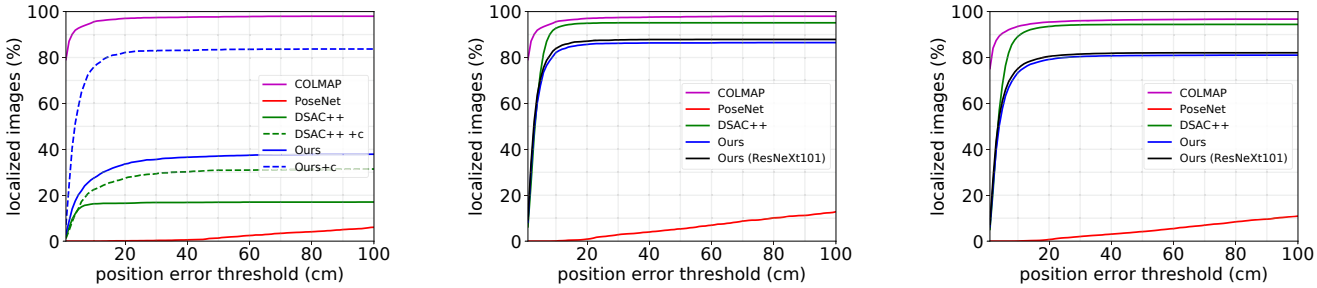

Figure 8. Robustness to lighting conditions and occlusions with the percentages of localized images on VirtualGallery test set for varying position error threshold and a fixed 5° orientation error threshold. *Left:* training on the standard lighting condition, testing on all lighting conditions without human (+c indicates training with color jittering). *Middle:* training on all lighting conditions (except COLMAP), testing on all lighting conditions without humans. *Right:* training on all lighting conditions (except COLMAP), testing on all lighting conditions with all human occlusion levels.

lighting conditions and tested on all lighting conditions without human occlusion, except COLMAP which is only trained using the standard lighting. The percentages of successfully localized images at a given error threshold are reported in Figure 8 (middle). The performance of PoseNet with geometric loss [15] is quite low because the training data does not contain enough variation: all images are captured at the same height, with 0° roll and pitch. Consequently, it learns this training data bias which is not valid anymore on the test data. COLMAP performs best with about 95% of the images localized within 10cm and 5°. DSAC++ localizes 75% of images within 5cm and 5°. Our approach performs similarly at low thresholds achieving about the same percentage of images successfully localized within 5cm and 5°. At higher thresholds, our method saturates earlier (around 88% with ResNeXt101 backbone) than DSAC++. We find that our approach fails to detect OOIs in cases where they are poorly visible (see top right example of Figure 5), thus we cannot localize such images. In contrast, DSAC++ can better handle this case as it relies not only on the OOIs but on the entire image. Overall, our method still computes highly accurate localization in standard cases where at least one OOI is present. We achieve a median error of less than 3cm, which is slightly lower than DSAC++.

Impact of the quantity of training data. Figure 7 (middle) shows the performance of our method when reducing the amount of training data. The percentages of localized images is roughly constant, even when training on only 1 image out of 15 (7%) of the first loop. In contrast, DSAC++, see Figure 7 (right), shows a larger drop of performance when training on a few images, highlighting the robustness of our approach to a small amount of training data. We did not observe significant difference with COLMAP, as two views of each point are sufficient to build the map of this comparably easy dataset for structure-based methods.

Robustness to lighting conditions. To study the robustness to different lighting conditions, we compare the average performance over the test sets with all lighting conditions without human, when training on all loops of the standard lighting condition only, see Figure 8 (left) *vs*. training on all loops and all lighting conditions, see Figure 8 (middle).

The performance of PoseNet drops by about 10% at 1m and 5° when training on only one lighting condition, despite color jittering at training. The performance of COLMAP stays constant even if we train using the standard lighting condition only (SfM does not work well with multiple images from identical poses, as in our training data with different lighting conditions). The high quality images and the unique patterns are easy to handle for illumination invariant SIFT descriptors. The performance of DSAC++ without any color data augmentation drops significantly when training on one lighting condition; in detail, by a factor of around

Figure 9. *Left:* input images with overlaid detections of masks. *Right:* for each detected class, we warp the reference image according to the homography fitted from the regressed matches.

6, which is the number of different lighting conditions. This means that only the images with the same lighting as training are localized, but almost none of the others. However, when adding color jittering to DSAC++ at training, the performance slightly increases.

The performance of our method (plain blue curve, on the left plot) is significantly higher than DSAC++ and PoseNet when training on one lighting condition, with about 30% of the images localized with an error below 25cm and 5°, showing that we overfit less on the training lighting condition. We also try to incorporate color jittering at training (dotted blue curve on the left plot) and obtain a significant increase of performance, achieving almost 85% of localized images with an error below 25cm and 5°. This performance is pretty close to the one obtained when training on all lighting conditions (middle plot), which can be considered as an upper bound of achievable results. In practice, this means that for real-world applications our method can be used even if only one lighting condition is present in the training data.

Robustness to occlusions. To study robustness to occlusions, we compare the performance of all methods when training on all loops and all lighting conditions, and testing on (a) all lighting conditions without visitors, see Figure 8 (middle), and (b) all lighting conditions and various visitor densities, see Figure 8 (right). All methods show a slight drop of performance. For our approach, the decrease of performance mostly comes from images where (a) only one painting is present and (b) the OOI is mostly occluded. This causes the OOI detection to fail. In most cases however, our approach is robust despite not having seen any human at training. Figure 9 shows examples of instance segmentation (left) in presence of humans. To visualize the quality of the matches, we fitted homographies between the test image and the reference images and warped the reference images onto the query image plane. We observe that the masks and the matches remain accurate.

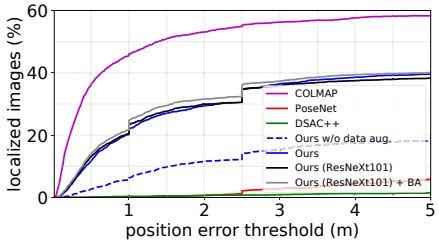

Figure 10. Percentages of localized images on the Baidu localization dataset for varying position error thresholds and an orientation error threshold of 5° for pos. error below 1m, 10° for pos. error between 1m and 2.5m, and 20° for pos. error above 2.5m.

5.2. Experiments on the Baidu localization dataset

Figure 10 presents results on the Baidu localization dataset [33]. This benchmark represents a realistic scenario which makes it extremely challenging: (a) the training data is limited to 689 images, captured in a mall of around 240m, (b) training and test images have different cameras and viewpoints, and (c) the environment has some changes in terms of lighting conditions and dynamic objects. Deep state-of-the-art approaches perform poorly, with less than 2% of images localized within 2m and 10°. COLMAP is able to localize more images, with about 45% at 1m and 5° and 58% at 5m and 20°. Our method is the first deep regression-based method able to compete with structured-based methods on this dataset. We successfully localize about 25% of the images within 1m and 10° and almost 40% within 5m and 20°. To further increase accuracy, we ran a non-linear least square optimization (sparse bundle adjustment [20]) as post processing (grey curve) obtaining a performance increase of about 2%. Figure 10 again highlights the benefit of homography data augmentation at training (plain *vs*. dotted blue curves). We are not able to localize about 10% of the query images where no OOI is detected.

6. Conclusion

We proposed a novel approach for visual localization that relies on densely matching a set of Objects-of-Interest. It is the first deep regression-based localization method that scales to large environments like the one of the Baidu localization dataset. Furthermore, our approach achieves high accuracy on smaller datasets like the newly introduced VirtualGallery. Since our method relies on OOIs, localization is only possible in the presence of OOIs. This putative drawback is a core characteristic of our approach because it enables the use of visual localization in fast changing and dynamic environments. For this to be true we assume that in such a situation at least the selected OOIs remain stable or can be tracked when being moved or changed. The learning component of our approach allows to increase robustness against changing lighting conditions and viewpoint variations. Future work include automatic mining of OOIs as well as generalizing to non-planar OOIs.

References

[1] Michael Bloesch, Michael Burri, Sammy Omari, Marco Hutter, and Roland Siegwart. Iterated extended kalman filter based visual-inertial odometry using direct photometric feedback. *IJRR*, 2017. 2

[2] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. DSAC-differentiable RANSAC for camera localization. In *CVPR*, 2017. 1, 3

[3] Eric Brachmann and Carsten Rother. Learning less is more - 6D camera localization via 3D surface regression. In *CVPR*, 2018. 1, 3, 6

[4] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *CVPR*, 2018. 1, 3

[5] Federico Camposeco, Andrea Cohen, Marc Pollefeys, and Torsten Sattler. Hybrid camera pose estimation. In *CVPR*, 2018. 3

[6] Robert Castle, Georg Klein, and David W Murray. Video-rate localization in multiple maps for wearable augmented reality. In *International Symposium on Wearable Computers*, 2008. 1

[7] Ronald Clark, Sen Wang, Andrew Markham, Niki Trigoni, and Hongkai Wen. Vidloc: A deep spatio-temporal model for 6-dof video-clip relocalization. In *CVPR*, 2017. 1, 3

[8] Mark Cummins and Paul Newman. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *IJRR*, 2008. 1

[9] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2, 4

[10] Qiang Hao, Rui Cai, Zhiwei Li, Lei Zhang, Yanwei Pang, and Feng Wu. 3D visual phrases for landmark recognition. In *CVPR*, 2012. 1, 3

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 1, 2, 4

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[13] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *CVPR*, 2018. 2

[14] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *ICRA*, 2016. 3

[15] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *CVPR*, 2017. 1, 3, 6, 7

[16] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*, 2015. 1, 3

[17] Yunpeng Li, Noah Snavely, and Dan Huttenlocher. Location recognition using prioritized feature matching. In *ECCV*, 2010. 1, 3

[18] Yunpeng Li, Noah Snavely, Dan Huttenlocher, and Pascal Fua. Worldwide pose estimation using 3D point clouds. In *ECCV*, 2012. 1, 3

[19] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 4

[20] Manolis Lourakis and Antonis Argyros. The design and implementation of a generic sparse bundle adjustment software package based on the levenberg-marquardt algorithm. Technical report, Institute of Computer Science-FORTH, Heraklion, Crete , 2004. 8

[21] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 3, 5

[22] Simon Lynen, Torsten Sattler, Michael Bosse, Joel A Hesch, Marc Pollefeys, and Roland Siegwart. Get out of my lab: Large-scale, real-time visual-inertial localization. In *Robotics: Science and Systems*, 2015. 1

[23] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Image-based localization using hourglass networks. *ICCV workshops*, 2017. 3

[24] Pierre Moulon, Pascal Monasse, and Renaud Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *ICCV*, 2013. 1, 3

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 4

[26] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Improving image-based localization by active correspondence search. In *ECCV*, 2012. 1, 3

[27] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE Trans. PAMI*, 2017. 1, 3

[28] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. Are large-scale 3D models really necessary for accurate visual localization? In *CVPR*, 2017. 3

[29] Grant Schindler, Matthew Brown, and Richard Szeliski. City-scale location recognition. In *CVPR*, 2007. 1, 3

[30] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 4, 6

[31] Johannes Lutz Schönberger, True Price, Torsten Sattler, Jan-Michael Frahm, and Marc Pollefeys. A vote-and-verify strategy for fast spatial verification in image retrieval. In *ACCV*, 2016. 6

[32] 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. 3

[33] Xun Sun, Yuanfan Xie, Pei Luo, and Liang Wang. A dataset for benchmarking image-based localization. In *CVPR*, 2017. 2, 6, 8

[34] Linus Svärm, Olof Enqvist, Magnus Oskarsson, and Fredrik Kahl. Accurate localization and pose estimation for large 3D models. In *CVPR*, 2014. 1, 3

[35] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *CVPR*, 2018. 1, 3

[36] Akihiko Torii, Josef Sivic, Masatoshi Okutomi, and Thomas Pajdla. Visual place recognition with repetitive structures. *IEEE Trans. PAMI*, 2015. 3

[37] Akihiko Torii, Josef Sivic, and Thomas Pajdla. Visual localization by linear combination of image descriptors. In *ICCV Workshop*, 2011. 3

[38] Julien Valentin, Matthias Nießner, Jamie Shotton, Andrew Fitzgibbon, Shahram Izadi, and Philip HS Torr. Exploiting uncertainty in regression forests for accurate camera relocalization. In *CVPR*, 2015. 3

[39] Florian Walch, Caner Hazirbas, Laura Leal-Taixe, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using LSTMs for structured feature correlation. In *ICCV*, 2017. 1, 3

[40] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 4

[41] Bernhard Zeisl, Torsten Sattler, and Marc Pollefeys. Camera pose voting for largescale image-based localization. In *ICCV*, 2015. 3

[42] Wei Zhang and Jana Kosecka. Image based localization in urban environments. In *International Symposium on on 3D Data Processing, Visualization, and Transmission (3DPVT)*, 2006. 3

CVPR2019 Paper Translation

姓名：彭伟聪

学号：2017300769

班号：10011701

通过回归感兴趣物体的稠密匹配进行视觉定位

Philippe Weinzaepfel Gabriela Csurka Yohann Cabon
Martin Humenberger
NAVER LABS Europe
firstname.lastname@naverlabs.com

摘要

我们将介绍一个新方法，用于使用仅仅一张 *RGB* 图像进行视觉定位，这种方法依赖于稠密匹配一组感兴趣物体（*Objects-of-Interest*，下称 *OOI*）。在本文中，我们关注的是在环境中具有高度描述性的平面物体，例如说博物馆中的绘画及商场或者机场中的标志和店面。对于每一个 *OOI*，我们定义了一张可得到三维世界坐标的参考图像。对于检索图像，我们的 *CNN* 模型先检测，分割 *OOI* 并找到一组在检测出的 *OOI* 和它对应的参考图像的 *2D-2D* 的密集匹配。从这些 *2D-2D* 的匹配，以及每个参考图像的三维世界坐标，我们得到了一组可通过求解多点透视问题（*Perspective-n-Point*）获得位姿估计的 *2D-3D* 的匹配。利用运动恢复结构（*Structure-from-Motion*）从每个 *OOI* 的单个实例标注可以获得这些关于参考图像的 *2D-3D* 匹配，和所有 *OOI* 的标注。我们引进了一个新的综合数据集——“*VirtualGallery*”，它面向不同光照条件及不同的遮挡水平下的挑战。我们的结果表明我们的方法取得了很高的精确度，而且对于上述的挑战也足够鲁棒。同时，我们也在购物中心采集的百度定位数据集进行了实验。我们的方法是第一个可以扩展到如此大环境的，基于深度回归的方法。

1. 引言

视觉定位在于从一个给定的空间或称为地图，获得一张 RGB 图像来估计一个六自由度的相机位姿。如果没有其他可行的定位技术的情形，例如室内定位等强拒止（GPS-denied）环境，这是一个特别有价值的定位技术。视觉定位有趣的应用包括机器人导航 [8]，自动驾驶以及增强现实（augmented reality，AR）[6, 22]。主要的挑战包括在检索图片和训练图片中大的视角的变化、不完整的地图、缺少有用信息的区域（例如无纹理的表面）、对称和重复的元素、变化的光照条件、结构变化、动态对象（例如人）及大面积的尺度变化。

传统的基于结构的方法 [17, 18, 24, 26, 27, 34] 在检索图片和地图间进行特征匹配，来应对很多提及的挑战。然而，就处理时间和内存开销而言，大面积进行不同视角及不同场景下的特征匹配是不可行的。为了克服这些问题，可以使用图像检索来加速大规模定位的匹配问题。[10, 29, 35] 最近，基于深度学习的方法 [2, 16] 显示出了很有希望的结果。PoseNet [16] 及其改进 [4, 7, 15, 39] 通过直接从输入图像回归相机位姿。即使可以获得粗略估计，学习精确定位似乎过于困难，或者需要大量的训练数据来覆盖包括相机外参和相机内参的多样性。更有趣的是，Brachmann 等人 [2, 3]，提出通过学习稠密三维坐标回归来解决多点透视问题（PnP）问题以进行精确的位姿估计。由于 RANSAC 存在一个可微分的近似公式 DSAC，这样的 CNN 模型得以被端到端的训练。这种基于场景坐标回归的方法在静态环境中，例如对象、遮挡和光照条件没有变化的环境，展现了出色的性能。然而，这些方法在场景规模上受限。

上述关于视觉定位的挑战在规模非常大且动态的场景下变得更加关键。在这种场景下，诸如静态且不变的场景等基本假设会被违背，且连续重新训练模型具有挑战性的同时，地图也会更快失效。这促使我们设计了一种灵感来自实例分割 [11] 的算法，该算法依赖于稳定的预定义区域，能在相当动态的场景中平衡精确定位与长期稳定性之间的冲突。我们提出了一种新

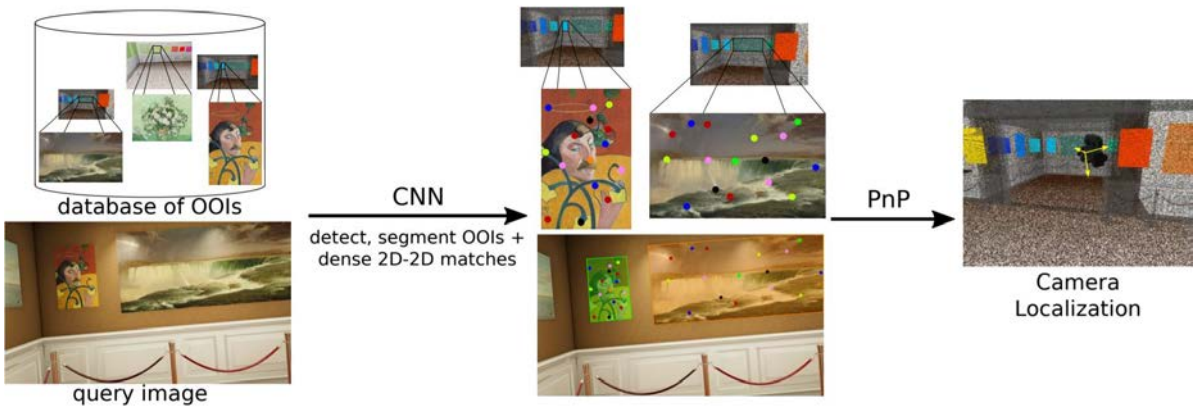

图 1. 以下是我们的流程概述。对给定检索图像，我们首先使用 CNN 模型进行检测和分割预定义的感兴趣物体。紧接着，通过回归得到与他们参考图像（彩色点）的稠密匹配。这些参考图像包含稠密的 2D 像素到 3D 坐标的映射。因此，我们可以利用传递性获得 2D-3D 匹配并解决多点透视问题来获得相机位姿。

颖的基于深度学习的视觉定位方法，这种方法通过在检索图片中的一些 OOI 和 OOI 相应的参考图像之间找到一组稠密匹配来进行，即一个物体的规范视角可以获得一组 2D-3D 对应关系。我们将 OOI 定义为 3D 地图内的，可以从多个视点，部分遮挡以及在各种照明条件下可靠地检测的有区分度的区域。图1展示了我们流程的概述。我们的模型依赖于 DensePose [9]，这是 Mask R-CNN [11] 的最新扩展，不仅可以检测和分割人类，还可以回归出图像中像素和网格曲面之间的稠密匹配。在我们的例子中，我们使用它（i）来检测和分割 OOIs，以及（ii）在检测到的 OOI 和它们的参考图像之间获得一组稠密的 2D-2D 匹配。从这些匹配以及与参考图像的 2D-3D 对应关系中，我们可以获得一组 2D-3D 匹配，并中通过使用 RANSAC 算法解决多点透视问题获得相机定位。

我们的方法经过精心设计，以应对视觉定位的公开挑战，它具有如下几个优点，而且几乎可以称得上是最先进的方法。第一，在 2D 进行推理可以使用少量数据来训练模型：我们可以通过基于单应矩阵的数据增强手工地为每个对象生成众多的视角，同时，我们可以通过颜色抖动增强对光照变化的鲁棒性。第二，只要 OOI 保持静止，我们的方法就可以处理动态场景。例如，即使训练数据不包含任何人，也可以准确地估计出充满游客的博物馆中的相机位姿。第三，如果移动一些 OOI，相比基于位姿和场景回归的方法，我们不需要重新训练整个网络，我们只需要更新参考图像的 2D-3D 映射。第四，我们的方法着重于判别对象，（相比于基于结构的方法）避免了无纹理区域带来的影响。第五，因为目标探测器可以分割出数千个类别 [13]，所以我们的方法可以应用到更大的定位区域和更多的 OOI。但是，我们方法的一个明显缺陷在于在检索图像不含任何 OOI 时，定位将会失效。然而，在诸如 AR 导航的许多应用中，在大多数时间里视野里都存在 OOI，而且在其中可使用局部姿态跟踪（例如，视觉里程计 [1]）OOI 检测本身在这样的应用中是很有意思的，例如，其可用于在博物馆或购物中心和机场里的的商店中诠释绘画。此外，在复杂的实际应用程序中，OOI 可用于更轻松地引导实际用户进行定位。诸如“拍摄最近画的照片”之类的命令可能比“拍摄具有足够视觉信息的照片”更容易让用户理解。

在本文中，我们将 OOI 限制在实际应用中常见平面物体：例如绘画，海报，店面或徽标，这些在定位在例如购物中心，机场或博物馆等具有挑战性的环境中经常出现。虽然该方法可以推广到非平面物体，但是平面 OOI 除了在基于单应矩阵的数据增强有优势之外，还具有以下几个优点。首先，训练图像中的 OOI 实例与其参考图像之间的变换是单应的，因此只需要很少的对应关系就能得到稠密的匹配。其次，参考图像的 2D 像素与 3D 世界坐标之间的映射可以从少量的图像获得，因为可以在 3D 中鲁棒地重建平面。最后，平面 OOI 能让我们在 2D 中推理，与直接在 3D 坐标上进行回归相比，减少了一个自由度。此外，我们的方法可以仅仅使

用像是在任意训练图片中的一个对于（平面）OOI 的实例分割那样的少量人工标注。

我们在两个富有挑战性的数据集上了验证了我们方法的优势。第一个是新引入的合成数据集 Virtual-Gallery，它代表了在艺术画廊下的场景。第二个是在购物中心采集的百度定位数据集 [33]，我们在训练模型时只有很少的视角，但是在测试时对场景增加一定改动。这展现了我们的方法对复杂的现实环境的适应性。虽然我们的方法在 VirtualGallery 上是准确的（即使在不同的光照条件和遮挡），实现了小于 3cm 和 0.5° 的中误差，它也可以扩展到更大的环境，例如百度定位数据集。相比之下，最好的基于深度学习回归的方法在这种情况下失效了。

2. 相关工作

大多数视觉定位方法可以分为四种类型的方法：基于结构的方法，基于图像检索的方法，基于相机位姿回归的方法和基于坐标回归的方法

基于结构的方法 [17, 18, 24, 26, 27, 34] 在与局部特征点相关联的 3 维点地图和检索图像中的描述符使用描述符进行匹配（例如 SIFT [21]）。然而，这些点特征不能足够鲁棒地应对诸如不同天气，照明或环境条件的场景带来的挑战。此外，这种方法缺乏获得全局上下文的能力，而且需要鲁棒地获得数百个点的以用于预测位姿 [41]。

基于图像检索的方法 [10, 29, 35–37] 使用全局描述符或视觉词将检索图像与地图的图像进行匹配，用于自顶向下检索图像以进行定位。检索到的位置可以进一步用于限制基于结构的方法的搜索范围 [5, 28]，或者直接用于计算检索图像和检索图像之间的位姿 [42]。这些方法可以在大型环境中加速搜索，但在使用基于结构的方法进行精确的位姿计算时具有与基于结构的方法类似的缺点。InLoc [35] 代表了利用稠密信息的基于图像检索的方法的最新进展。它首先使用深度特征来检索最相似的图像，然后估计在地图中相机的位姿。但是它的缺点是巨大的计算负担以及需要精确的稠密 3D 模型。

基于位姿回归的方法 [4, 16, 39] 是第一批针对视觉定位进行端到端训练的深度学习方法。他们追随开创性的 PoseNet [16] 方法，使用 CNN 直接从检索图像中回归 6 自由度的相机位姿。这种方法在很多方面进行了衍生，比如使用视频信息 [7]，递归神经网络 [39]，沙漏结构 [23] 及贝叶斯 CNN，以减少定位的不确定度 [14]。最近，Kendall 等 [15] 又将原始的 L2 损失函数替换成一个依赖场景几何和重投影误差的新型损失函数。Brahmbhatt [4] 等利用训练时图像之间的相对位姿来提升。总体而言，基于位姿回归的方法已展现出在许多挑战的鲁棒性，但其在准确性和规模方面仍然受到限制。

基于场景坐标回归的方法 [2, 32, 38] 通过回归稠密 3D 坐标和使用带 RANSAC 的多点透视问题求解程序来估计位姿。虽然过去曾使用随机森林 [32, 38]，但 Brachmann 等人 [2] 最近通过训练 CNN 来稠密回归 3D 坐标来获得极其准确的位姿估计。他们还引入了可微分的 RANSAC 近似，称为 DSAC，这允许以多步训练为代价进行视觉定位的端到端训练，其中第一步需要深度信息。DSAC++ [3] 是对该方法的改进，其中真实深度数据不是强制性的，可以用先验深度代替。网络仍然经过多个步骤的训练：其中第一步是基于深度数据或先验的深度信息; 第二步基于最小化重投影误差；最后使用 DSAC 模块的基于相机定位误差。DSAC++ 在恒定光照，不那么动态的相对较小规模的数据集上展现了出色的性能。但是，这种方法在较大的场景很难收敛。相比之下，我们的方法建立在目标检测上，可以扩展到大型环境。与 DSAC ++ 相比，由于我们只考虑平面物体，我们可以回归 2D 匹配而不是 3D 坐标，这消除了一个自由度，并能进行基于单应矩阵的数据增强。

3. 视觉定位流程

在本节中，我们首先在 3.1 节中概述我们的方法。接着，我们将介绍有关用于分割 OOI 和稠密匹配的 CNN 模型的详细信息（第 3.2 节）。最后，我们在第 3.3 节解释了如何利用 SfM 在地图中弱监督地训练。

3.1. 通过检测 OOI 进行视觉定位

设 OOI 的集合为 $\mathcal{O}$，OOI 类别的数量为 $|\mathcal{O}|$。我们的方法依赖于参考图像：每一个 OOI$o \in \mathcal{O}$ 都与一个规范视角相关联，即一张 $o \in \mathcal{O}$ 可见的图片 $\mathcal{I}_o$。我们暂时假设每一个 OOI 是在场景中唯一存在的，并且参考图像中的二维像素 $\mathbf{p}'$ 和世界 $M_o(\mathbf{p}')$ 对应的三维点的映射 M_o 是已知的。

对于给定的检索图像，我们的 CNN 模型输出所有

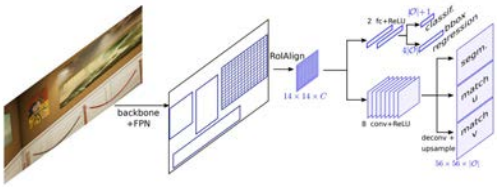

图 2. 网络架构概述，用于检测和分割 OOI 以及获取与之相关的稠密匹配参考图像。

检测结果。每个检测结果包括一个带类别标签（即检测到的 OOI 的 id）O 的包围盒、一个置信度、一个分割掩膜、一组检索图像像素 $\mathbf{q}$ 和 O 的参考图像 $\mathcal{I}_o$ 像素 $\mathbf{q}'$ 之间的 2D-2D 的匹配 $\{\mathbf{q} \rightarrow \mathbf{q}'\}$，见图1 通过利用传递性，我们将映射 M_o 应用于参考图像中的匹配的像素，并为每个检测结果进行检索图像中的 2D 像素与世界坐标中的 3D 点之间的匹配：$\{\mathbf{q} \rightarrow \mathcal{M}_o(\mathbf{q}')\}$。

给定在检索图像中所有检测结果的 2D-3D 匹配和相机内参，我们可以通过使用 RANSAC 算法来求解一个多点透视问题来估计相机内参。

请注意，如果在检索图像中没有任何检测结果，我们无法获得点的匹配，而且因此我们无法进行定位。然而，在诸如博物馆或者机场这样的公共场所中，可以在绝大多数图像中找到 OOI。此外，在现实世界中的实际应用里，全局定位通常会与局部位姿跟踪结合使用，因此整体系统的精度比覆盖范围更重要。

概括起来，我们方法中需要进行学习的只有目标检测和检索图像与参考图像之间的稠密匹配。

处理非唯一的 OOI。到目前为止，我们假设每个 OOI 都是独一无二的，即它在环境中只出现一次。虽然大多数 OOI 都是又相当多区别的，但是其中一些可以在同一环境中具有多个实例，例如在购物中心的 logo。在这种情况下，由于检测器无法区分他们，我们便标记同样的 OOI 为同一个类别并使用共同的参考图像来训练模型。于是映射 M_o 便不再是唯一的，因为对于相同的参考图像存在多个候选三维坐标。鉴于在实践中每个 OOI 的重复次数有限，每个图像检测次数有限，于是我们可以使用几何合理性约束来获得可能坐标的最佳组合，从而求解多点透视问题。具体一点，我们要最小化检索图像中 OOI 的三维中心点的重投影误差之和。理想情况下，OOI 的三维中心点位于分割后检测检测的正中。对于 OOI 的不完整检测而产生的噪音，我们选择置之不理。

3.2. 检测 OOI 及匹配

我们追随 DensePose [9]，一个 Mask R-CNN 的拓展 [11]，它被设计用于寻找人体任意点与表面网格上对应点之间的稠密对应关系。对于区域推荐网络（RPN）[25] 生成的每个盒子，使用 ROIAlign 层，生成 $14 \times 14 \times C$ 维的特征向量。同时，我们使用特征金字塔网络的改进 [19] 来处理小的物体。盒子的特征向量会被送到两个分支。一个分支用于获得每个类别的分数（人类的和非人类的）及像 Faster R-CNN [25] 那样回归带类别的包围盒的坐标。另一个分支是全卷积网络，用于预测出像素级的每部分人体的标签和与之对应的网格曲面坐标（即，两个坐标）。在实践中，这个 CNN 模型预测密集网格上每一个大小为 56×56 的框的的分割和对应关系，然后对其进行插值以获得像素级的预测结果。

在我们的场景中，对于在 RoIAlign 层后得到的框的特征向量，我们使用一个类似 DensePose 结构的 CNN 模型，见图 22。一个分支用于预测 OOI 的分数和回归包围盒，与 DensePose 唯一的不同在于此处我们有 $|\mathcal{O}|+1$ 个类别（包括背景）而不是两类。另一个分支任务有所不同：（a）对每一个 OOI 进行二值分割，（b）回归 OOI 的图像参考坐标 u 和 v。在训练阶段，我们结合了几种损失函数。在区域推荐我们使用的 FPN 损失函数，我们在图像分类时使用交叉熵损失函数。我们在回归框时使用 smooth-L1 损失函数，在预测大小为 56×56 的掩膜时使用交叉熵损失函数，在回归 u 和 v 时使用 smooth-L1 损失函数。进行训练需要标注正确的掩膜和每个像素的匹配。在第3.3节，我们将展示如何自动化的从少量的标注获得大量的标注。在测试时，我们保留分类得分高于 0.5 的包围盒和在分割掩膜中点的匹配。

实现细节。我们将 FPN [19] 与 ResNet50 [12] 和 ResNeXt101-32x8d [40] 主干网一起使用。预测分段和匹配回归的分支遵循 Mask R-CNN 架构：它在每个任务的最后一个卷积层之前由 8 个卷积和 ReLU 层组成。训练网络时候，我们使用一块 GPU，进行 $500k$ 次迭代，初始学习率为 0.00125，在 $300k$ 和 $420k$ 迭代后都将其除以 10。我们使用 SGD 优化器，动量为 0.9，权重衰减为 0.0001。为了使所有回归量以相同的尺度进行，我们将参考坐标标准化至 $[0, 1]$ 范围。

数据增强由于我们的 CNN 仅在 2D 主句中进行回归匹配，因此我们对所有输入图像进行基于单应矩阵的数据增强。为了生成合理的视角下的数据，我们生成 4 个

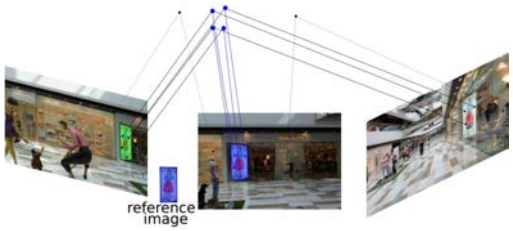

图 3.（中间框）蓝色的手动标注掩膜的包围盒确定了 OOI 的参考图像。我们利用手工标注（蓝色掩膜）内的关键点的 3D 坐标将 OOI 传播到其他训练图像（左图中的绿色掩膜）。如果没有足够的匹配（右图），这样的传播可能会失败。

限制为 33% 图像尺寸的随机位移的顶点，并计算相应的单应矩阵。我们不使用翻转进行数据增强，因为 OOI（徽标，绘画，海报）是遵循从左到右的顺序的。我们研究了在实验中使用颜色抖动（亮度，对比度，饱和度）对不同光照条件下的鲁棒性的影响（见第5节）。

3.3. 弱监督下 OOI 的标注

我们的方法的关键是少量的手工标注，这要归功于 COLMAP [30]，它开创地利用了传播算法进行 SfM。要提供的唯一标注是平面 OOI 的一个分割掩膜，可参见图3中间的蓝色掩膜。这个 OOI 的参考图像由带标注的掩膜确定。

利用来自 SfM 的 2D-3D 的匹配，我们标注与标注好的 OOI 分割里二维点相匹配的三维点，参见图3中的蓝线和点。我们将标签传播到包含这些观测到的三维点的所有训练图像。假设 OOI 是平面的，且图像中至少有 4 个匹配点，我们可以获得一个单应矩阵。为了更好地抑制噪声，我们仅考虑具有最少 7 个匹配的区域并使用基于 RANSAC 的单应矩阵估计。此单应矩阵用于传播掩膜标注以及稠密的 2D-2D 匹配，可参见图3左侧的绿色掩膜。

由于匹配数量较少，可参见图3的右图，及 SfM 模型中的错误匹配，这种传播并不总是成功。幸运的是，CNN 模型在某种程度上可以抵抗嘈杂或丢失的标签。

处理不唯一的 OOI 对于不唯一的 OOI，例如在商场中多次出现的 logo，我们为 OOI 的每个实例标注一个分割掩膜，并且我们在每个实例上独立地使用传播。但如上文所提及的，没有检测器可以区分不同的实例。因此，我们将它们合并到一个 OOI 类中，并计算使用 SIFT 描述符 [21] 在不同实例的参考图像和主参考图像之间

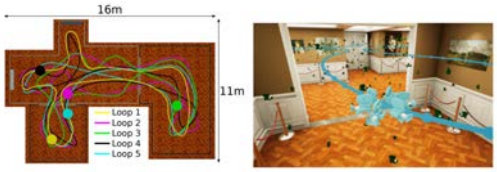

图 4. 左图：艺术画廊的平面图，带有不同的训练环路。右图：一个训练环路，其中 6 个摄像机处于固定高度（青色），测试摄像机位于不同的合理的位置上。

图 5. 来自 VirtualGallery 的样本。第一行：不同视角的测试图像。第二排和第三排：不同的光照条件。第四行：不同的人类密度（遮挡）

的单应矩阵。其中，我们选择具有最多三维匹配的参考图像作为一类（OOI）的主要参考图像。由于回归的 2D-2D 匹配对应于主类，我们还使用已经得到的类内单应矩阵从回归得到的到主类的 2D-2D 匹配进行透视变换，可参见图3.1。

4. 数据集

VirtualGallery 数据集我们引入了一个新的合成数据集来验证我们的方法的适用性，并更进一步观测不同光照条件和遮挡对不同定位方法的影响。它由一个 3-4 个房间的场景组成，见图4（左），其中有 42 幅免费使用的名画 [1] 被挂在墙壁上。该场景通过使用 Unity 创建的，能过获得准确的信息，如深度，语义和实例分割，2D-2D 和 2D-3D 匹配，以及渲染后的图像。

我们考虑一个逼真的场景，使用模拟机器人捕获的场景进行训练，并由游客拍摄照片进行测试。机器人在 360° 的范围里设置设置 6 个摄像机，且固定高度为

[1] https://images.nga.gov

165cm，见图4（右）。其在画廊内根据 5 个不同的环路，大约每 20 厘米拍摄一次，每个相机会采集约有 250 张图像，见图4（左）。

在测试时，我们选取随机的位置，方向和焦距，但要确保视点（a）看起来合理且逼真（在方向，高度和到墙的距离方面），以及（b）跨越整个场景。这涵盖了视角变化和训练和测试图像之间不同相机内参的带来的额外挑战。

为了研究在不同光照条件下的鲁棒性，在训练和测试阶段，我们都使用 6 种不同的照明配置生成场景，请参见图5的第二和第三行。为了评估对诸如游客等遮挡物的鲁棒性，我们生成包含随机放置的人体模型的测试图像，参见图5的最后一行。测试集由 496 个图像组成，这些图像在 6 种不同照明条件，以及 4 种不同密度的人（包括没有）情况下进行了渲染。该数据集可在 http://www.europe.naverlabs.com/Research/3D-Vision/Virtual-Gallery-Dataset 中找到。我们使用每幅画作为一个感兴趣的对象，每一幅画在场景中都是独一无二的。我们使用从网站下载的原始图像作为参考图像。我们使用 Unity 获得每个图像的地准确的分割掩膜和 2D-2D 的对应关系。我们还获得绘画放置在场景中的位置和大小，从而能完成 2D-3D 的映射。注意，在测试图像中，496 中的 5 张不包含任何 OOI，并且有 23 张（独立的为 48 张）没有 OOI 可见超过 50%（独立的为 80%）。

百度定位数据集 [33]。它由在中国的购物中心拍摄的图片组成，涵盖了视觉定位的许多挑战，例如反射或透明表面，移动的人和重复的结构。它将 689 张用 DSLR 相机拍摄的图像作为训练集，以及超过 2000 张几个月后由不同用户拍摄的手机照片作为测试集。相比较训练图像均相对于走廊平行或垂直拍摄，测试图像拍摄的视角有显著的变化。所有图像都被半自动地标注到由 LIDAR 获得的坐标系。我们使用这个数据集提供的相机位姿，以用于训练（OOI 的 3D 重建）和测试（结果的真实性评估）。我们没有使用 LIDAR 的数据，即使它可能会提高我们 OOI 的 3D 重建质量。

感兴趣物体。我们为 164 个不同类型的平面对象（例如店面上的徽标或海报）手动标注 220 个实例的分割掩膜。然后我们将这些标注传播到所有训练图像，可参阅第3.3节。这个真实世界的数据集比 VirtualGallery 更具挑战性，因此一些 OOI 传播是有噪声的。图6展示了在传播之后具有围绕 OOI 的掩膜的训练图像的示例。

图 6. 百度定位数据集中的使用传播的掩膜标注的训练数据示例

5. 实验结果

我们在 VirtualGallery（参见第5.1节）和百度定位数据集（参见第5.2节）上评估我们的方法。

5.1. 在 VirtualGallery 上的实验结果

我们使用不同的训练/测试场景来评估我们方法的一些变体，并研究最先进的方法对于照明变化和遮挡的鲁棒性。在下面的实验中，我们使用从 Unity 获得的准确匹配。我们尝试了手动标注并获得了相近的性能。**数据增强对结果的影响**。我们首先研究了基于单应矩阵的数据增强对训练的影响。我们在第一个环路且标准照明条件下，训练基于 ResNet50 主干网的模型，并统计在标准光照且没有人条件下成功定位的图片的百分比，见图7。基于单应矩阵的数据增强显着提高了性能，特别是对于高精度的定位：使用 ResNet50 主干网时，定位精度在 5cm 和 5° 范围内的图像比例从 25%增加到 69%。但是在较高的误差阈值下影响不大，在 25cm 和 5° 时比例只从 72%提高到 88%。训练时的基于单应矩阵的数据增强能生成 OOI 在更多视角下的数据，从而能在测试时更好地检测和匹配从未知视点捕获的 OOI。**稠密 2D-2D 匹配回归对结果的影响**。我们现在将我们的，依赖 2D-2D 稠密匹配的方法与一些变体进行比较。首先，我们直接回归检测到的 OOI 与其参考图像之间的 8 自由度的单应矩阵参数（OOI 到单应矩阵），然后生成一组稠密的 2D-2D 的匹配。然后，我们在不使用参考图像情况下，直接回归每个 OOI 实例（OOIs 到 2D 到 3D）的 3D 世界坐标。与上述相同的训练/测试格式的性能在图7中展示。通过回归获得的单应矩阵的 8 个参数导致性能下降，只有 70%的图像可以在 25cm 和 5° 内成功定位。当较小差异可能导致结果的显着变化的场景下，CNN 模型难以回归变换的参数。直接回归到

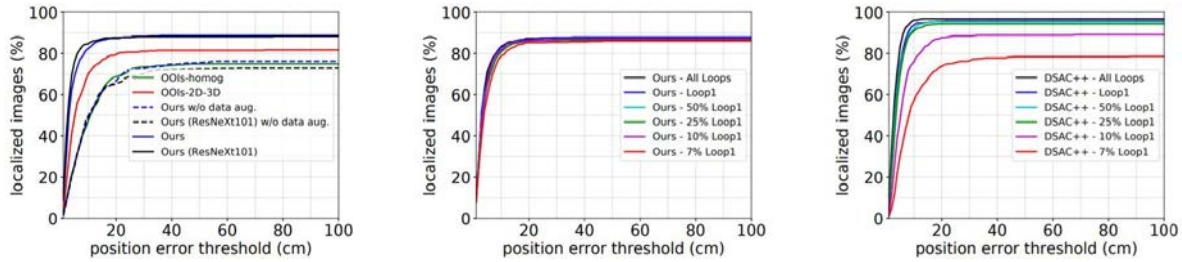

图 7. VirtualGallery 测试集上成功定位的图像的占比，在标准光照，没有人类，方向误差阈值固定为 5°，位置误差阈值为变量。*左图*：在标准的光照，第一个环路上训练的方法之间的比较。*中间和右侧*：我们的方法（中间）和 DSAC++（右侧）对较少的训练数据（标准光照下）的鲁棒性。

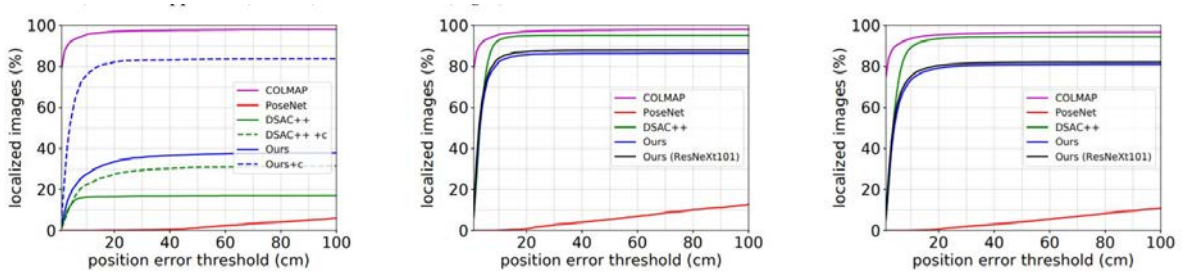

图 8. VirtualBallery 测试集上，在不同的位置误差阈值和固定的 5° 方向误差阈值设置下，定位成功的图像百分比及不同光照和遮挡的鲁棒性。*左*：在标准光照，在没有人的数据中进行训练，测试所有照明条件（+c 表示颜色抖动训练）。*中*：在所有光照（COLMAP 除外）下训练，在没有人的所有不同光照条件下测试。*右*：在所有光照（COLMAP 除外）下训练，在有人的所有不同光照条件下测试。

3D 坐标的变体表现相当不错，大约 70%的图像定位精度在 10cm 和 5° 之内，81%在 25cm 和 5° 之内。然而，我们的使用 2D 参考图像的方法优于这些方法，尤其是在小位置误差阈值的设置下。我们的 2D 参考图像确保所有 3D 点都在平面物体上，而这些方法增加了一个额外且不必要的自由度。我们最终将 ResNeXt101-32x8d 替换 ResNet50 主干网。这使得在低阈值时获得更精确的定位（8%的图像定位不在 5cm 和 5° 内）；在较高的阈值时，与其他方法之间的差异变得微不足道。

与最先进方法的比较。我们现在将我们的方法与 SfM 方法（COLMAP [30,31]），结合几何学损失的 PoseNet [15] 以及通过 3D 模型训练的 DSAC++ [3] 进行比较。所有方法都在所有光照条件的所有循环上进行训练，并在除了仅使用标准照明进行训练的 COLMAP，所有方法都在没有人为遮挡的所有光照条件下进行测试。图8（中间）展示了在给定误差阈值下能成功进行定位的图像所占的百分比。使用几何损失的 PoseNet [15] 性能非常差，因为训练数据没有包含足够多的变化：所有图像都在相同高度捕获，没有侧倾和俯仰。因此，它在训练时候学习到了在测试数据无用的数据偏差。COLMAP 表现最佳，约 95%的图像定位精度在于 10cm 和 5° 内 75%的图像在使用 DSAC++ 是，定位精度在在 5cm 和 5° 之内。我们的方法在低阈值时，表现跟上述方法相近。在阈值为 5cm 和 5° 时，大约有相同比例的图像会被成功定位。在更高的阈值下，我们的方法比 DSAC++ 更快收敛（在使用 ResNeXt101 主干网时约 88%）。我们发现我们的方法在它们看不清楚的情况下无法检测到 OOI（参见图5的右上方示例），因此我们无法对这些图像进行定位。相比之下，DSAC++ 能更好地处理这种情况，因为它不仅依赖于 OOI，而且依赖于整幅图像。总的来说，我们的方法在至少存在一个 OOI 的情况下能进行高精度的定位。我们的中误差小于 3cm，略低于 DSAC++。

训练数据量的影响。图7展示了减少训练数据量时我们方法的性能。成功定位图像的百分比大致恒定，甚至在第一个环路，在只有 $\frac{1}{15}$（约 7%）的数据中训练也是如

此。相比之下，DSAC++，见图7，其在较少图像上进行训练时性能下降，这体现了我们方法面对少量训练数据时的鲁棒性。我们没有观察到与 COLMAP 的显着差异，因为使用基于结构的方法时，在这种简单的数据集中，每个点的两个视角可以有效地构建映射。

对光照的鲁棒性。为了研究不同光照条件下的鲁棒性，在标准光照下，见图8（左），和所有光照条件下，见图8（中），我们在带有所有光照条件下且无人体的测试集上比较他们的平均性能。

尽管在训练中使用了颜色抖动进行数据增强，但在仅一种光照条件下进行训练时，PoseNet 的性能在阈值为 1m 和 5° 时下降了约 10%。但是当仅使用标准照明条件进行训练，COLMAP 的性能也保持基本不变（SfM 不能很好地处理来自相同位姿的多个图像，如我们在不同光照条件下的训练数据）。对于照明不变的 SIFT 描述符，高质量图像和独特纹理是很容易进行处理的。在仅一种光照条件下进行训练时，没有任何色彩抖动进行数据增强的 DSAC++ 的性能会显着下降；准确地说，下降为原来的六分之一，而不同照明条件的数量也是 6。这意味着只有与训练时相同光照的图像才能被定位，但是在其他光照条件下几乎没有图像能被定位。然而，在训练时向 DSAC++ 添加颜色抖动，性能会略有提高。

在仅使用一个光照条件的数据下进行训练时，我们的方法（左侧图中蓝色曲线）的性能显著高于 DSAC++ 和 PoseNet，大约 30%的图像定位误差低于 25cm 和 5°，这表明在减少训练照明条件时候，我们没有过拟合。我们还尝试在训练中加入颜色抖动（左图中的点蓝色虚线），并获得显着的性能提升，获得了近 85%的图像定位误差低于 25cm 和 5°。这样的性能非常接近于在所有照明条件（中间图）上进行训练时获得的性能，这可以被视为所有可实现结果的上限。这意味着对于实际应用，即使训练数据中仅存在一个光照条件，也可以使用我们的方法。

对遮挡的鲁棒性。为了研究对遮挡的鲁棒性，我们比较了所有方法在所有环路和所有照明条件下进行训练时的性能，并测试了（a）没有游客的所有光照条件，见图8（中），以及（b）所有照明条件和各种游客密度，见图8（右）。所有方法的性能都轻微下降。对于我们的方法，性能的降低主要原因来自图像本身，即（a）仅存在一幅画或（b）OOI 被大部分遮挡。这导致 OOI 检测失效。然而在大多数情况下，即使在训练时没有任何人的遮挡，我们的方法仍然鲁棒。图9展示了在人类存在下的实例分割（左）的示例。为了可视化匹配的质量，我们拟合测试图像和参考图像之间的单应矩阵，并将参考图像扭曲到查询图像平面上。我们观察到掩膜和匹配仍然准确。

图 9. *左*：输入的带有遮盖掩膜的图像。*右*：对于每个检测到的类别，我们根据回归匹配拟合得到的单应矩阵来变换参考图像。

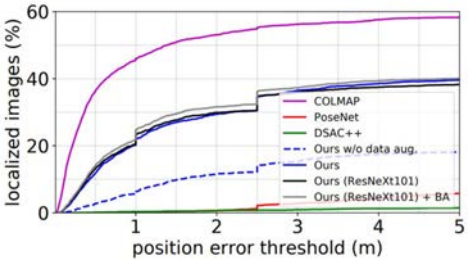

图 10. 改变位置误差阈值，固定角度误差阈值为 5°，百度定位数据集上的定位成功图像的百分比。位置误差低于 1m，角度误差低于 10°。位置误差在 1m 到 2.5m 之间，角度误差在 20° 之间。位置误差超过 2.5m。

5.2. 在百度定位数据集上的实验结果

10展示了在百度定位数据集 [33] 上的结果。这个数据集代表了一个现实的场景下的基准，这使得它极具挑战性：（a）训练数据仅有在 240 米左右的商场拍摄的 689 张图像，（b）训练和测试图像是在不同相机和视角拍摄的，以及（c）环境在光照条件和动态物体方面发生了一些变化。基于深度学弟的最先进方法表现不佳，只有不到 2%的图像定位误差在 2m 和 10° 之内。COLMAP 能够定位更多图像，在 1m 和 5 度时约为 45%，在 5m 和 20° 时为 58%。我们的方法是第一个能够在此数据集上与基于结构化的方法竞争的基于深度回

归的方法。我们成功地将大约 25%的图像在 1m 和 10° 内的误差阈值定位，在 5m 和 20° 内几乎达到了 40%。为了进一步提高准确度，我们进行了非线性最小二乘优化（稀疏光束平差法 [20]）作为后端（灰色曲线），获得了约 2%的性能提升。图10再次强调了基于单应矩阵进行数据增强的好处（普通蓝色曲线与蓝色虚线曲线）。我们无法定位大约 10%的未检测到 OOI 的检索图像。

6. 结论

我们提出了一种新型视觉定位方法，它依赖于稠密匹配一组感兴趣的对象。这是第一个基于深度回归的，可扩展到百度定位数据集这类环境较大时的定位方法。此外，我们的方法在较小的数据集上做到了高精度，例如新推出的 VirtualGallery。由于我们的方法依赖于 OOI，因此只有在 OOI 存在的情况下才能进行定位。因为它使得能够在快速变化和动态环境中使用视觉定位，所以依赖 OOI 这个缺点是我们的方法的核心特征。为此，我们假设在这种情况下，至少所选择的 OOI 可以保持稳定或者在移动或改变时可以被跟踪。我们的方法中需要进行学习的组件可以通过改变照明条件和视角变化来增强鲁棒性。未来的工作包括自动挖掘 OOI 以及推广到非平面 OOI。

参考文献

[1] Michael Bloesch, Michael Burri, Sammy Omari, Marco Hutter, and Roland Siegwart. Iterated extended kalman filter based visual-inertial odometry using direct photometric feedback. *The International Journal of Robotics Research*, 36(10):1053–1072, 2017.

[2] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac —differentiable ransac for camera localization. pages 2492–2500, 2017.

[3] Eric Brachmann and Carsten Rother. Learning less is more - 6d camera localization via 3d surface regression. pages 4654–4662, 2018.

[4] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. pages 2616–2625, 2018.

[5] Federico Camposeco, Andrea Cohen, Marc Pollefeys, and Torsten Sattler. Hybrid camera pose estimation. pages 136–144, 2018.

[6] Robert Oliver Castle, Georg Klein, and David W Murray. Video-rate localization in multiple maps for wearable augmented reality. pages 15–22, 2008.

[7] Ronald Clark, Sen Wang, Andrew Markham, Niki Trigoni, and Hongkai Wen. Vidloc: A deep spatio-temporal model for 6-dof video-clip relocalization. pages 2652–2660, 2017.

[8] Mark Joseph Cummins and Paul Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008.

[9] Riza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. pages 7297–7306, 2018.

[10] Qiang Hao, Rui Cai, Zhiwei Li, Lei Zhang, Yanwei Pang, and Feng Wu. 3d visual phrases for landmark recognition. pages 3594–3601, 2012.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross B Girshick. Mask r-cnn. pages 2980–2988, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 2016.

[13] Ronghang Hu, Piotr Dollar, Kaiming He, Trevor Darrell, and Ross B Girshick. Learning to segment every thing. pages 4233–4241, 2018.

[14] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. pages 4762–4769, 2016.

[15] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. pages 6555–6564, 2017.

[16] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. pages 2938–2946, 2015.

[17] Yunpeng Li, Noah Snavely, and Daniel P Huttenlocher. Location recognition using prioritized feature matching. pages 791–804, 2010.

[18] Yunpeng Li, Noah Snavely, Daniel P Huttenlocher, and Pascal Fua. Worldwide pose estimation using 3d point clouds. pages 15–29, 2012.

[19] Tsungyi Lin, Piotr Dollar, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. pages 936–944, 2017.

[20] Manolis I. A. Lourakis, Manolis I. A. Lourakis, Antonis A. Argyros, and Antonis A. Argyros. The design and implemen-

tation of a generic sparse bundle adjustment software package based on the levenberg-marquardt algorithm. Technical report, 2004.

[21] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[22] Simon Lynen, Torsten Sattler, Michael Bosse, Joel A Hesch, Marc Pollefeys, and Roland Siegwart. Get out of my lab: Large-scale, real-time visual-inertial localization. 11:37, 2015.

[23] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Image-based localization using hourglass networks. pages 870–877, 2017.

[24] Pierre Moulon, Pascal Monasse, and Renaud Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. pages 3248–3255, 2013.

[25] Shaoqing Ren, Kaiming He, Ross B Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv: Computer Vision and Pattern Recognition*, 2015.

[26] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Improving image-based localization by active correspondence search. pages 752–765, 2012.

[27] 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, 2017.

[28] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. Are large-scale 3d models really necessary for accurate visual localization? pages 6175–6184, 2017.

[29] Grant Schindler, Matthew Brown, and Richard Szeliski. City-scale location recognition. pages 1–7, 2007.

[30] Johannes L Schonberger and Jan Michael Frahm. Structure-from-motion revisited. pages 4104–4113, 2016.

[31] Johannes L Schonberger, True Price, Torsten Sattler, Jan Michael Frahm, and Marc Pollefeys. A vote-and-verify strategy for fast spatial verification in image retrieval. pages 321–337, 2016.

[32] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew W Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. pages 2930–2937, 2013.

[33] Xun Sun, Yuanfan Xie, Pei Luo, and Liang Wang. A dataset for benchmarking image-based localization. pages 5641–5649, 2017.

[34] Linus Svarm, Olof Enqvist, Magnus Oskarsson, and Fredrik Kahl. Accurate localization and pose estimation for large 3d models. pages 532–539, 2014.

[35] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. pages 7199–7209, 2018.

[36] Akihiko Torii, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. Visual place recognition with repetitive structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2346–2359, 2015.

[37] Akihiko Torii, Josef Sivic, and Tomas Pajdla. Visual localization by linear combination of image descriptors. pages 102–109, 2011.

[38] Julien Valentin, Matthias Niebner, Jamie Shotton, Andrew W Fitzgibbon, Shahram Izadi, and Philip H S Torr. Exploiting uncertainty in regression forests for accurate camera relocalization. pages 4400–4408, 2015.

[39] Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using lstms for structured feature correlation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 627–637, 2016.

[40] Saining Xie, Ross B Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. pages 5987–5995, 2017.

[41] Bernhard Zeisl, Torsten Sattler, and Marc Pollefeys. Camera pose voting for large-scale image-based localization. pages 2704–2712, 2015.

[42] Wenjun Zhang and Jana Kosecka. Image based localization in urban environments. pages 33–40, 2006.