RelPose: Predicting Probabilistic Relative Rotation for Single Objects in the Wild

Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani

Carnegie Mellon University, Pittsburgh PA 15213, USA
jasonyzhang@cmu.edu

Abstract. We describe a data-driven method for inferring the camera viewpoints given multiple images of an arbitrary object. This task is a core component of classic geometric pipelines such as SfM and SLAM, and also serves as a vital pre-processing requirement for contemporary neural approaches (e.g. NeRF) to object reconstruction and view synthesis. In contrast to existing correspondence-driven methods that do not perform well given sparse views, we propose a top-down prediction based approach for estimating camera viewpoints. Our key technical insight is the use of an energy-based formulation for representing distributions over relative camera rotations, thus allowing us to explicitly represent multiple camera modes arising from object symmetries or views. Leveraging these relative predictions, we jointly estimate a consistent set of camera rotations from multiple images. We show that our approach outperforms state-of-the-art SfM and SLAM methods given sparse images on both seen and unseen categories. Further, our probabilistic approach significantly outperforms directly regressing relative poses, suggesting that modeling multimodality is important for coherent joint reconstruction. We demonstrate that our system can be a stepping stone toward in-the-wild reconstruction from multi-view datasets. The project page with code and videos can be found at jasonyzhang.com/relpose.

1 Introduction

Recovering 3D from 2D images of an object has been a central task in vision across decades. Given multiple views, structure-from-motion (SfM) based meth-

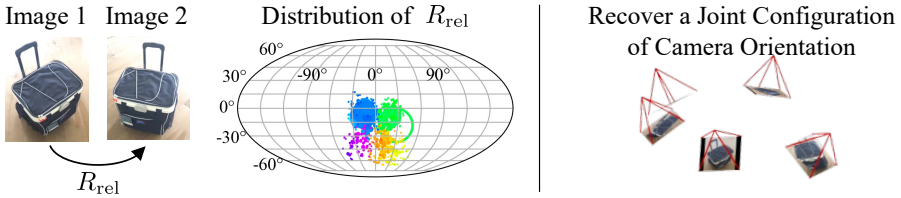

Fig. 1: **Probabilistic Camera Rotation Estimation for Generic Objects**. *Left*: Given two images of the same object, we predict a conditional distribution of relative camera viewpoint (rotation) that effectively handles symmetries and pose ambiguities. *Right*: Given a set of images, our approach outputs a configuration of camera rotations.

ods can infer a 3D representation of the underlying instance while also associating each image with a camera viewpoint. However, these correspondence-driven methods cannot robustly handle sparsely sampled images that minimally overlap, and typically require many (>20) images for a 360-degree 3D inference. Unfortunately, this requirement of densely sampled views can be prohibitive—online marketplaces often have only a few images per instance, and a user casually reconstructing a novel object would also find capturing such views tedious. Although the recently emerging neural 3D reconstruction techniques also typically leverage similarly dense views, some works have shown promise that a far smaller number of images can suffice for high-quality 3D reconstruction. These successes have however still relied on precisely [60] or approximately [20, 27, 82] known camera viewpoints for inference. To apply these methods at scale, we must therefore answer a fundamental question—*given sparsely sampled images of a generic object, how can we obtain the associated camera viewpoints*?

Existing methods do not provide a conclusive answer to this question. On the one hand, bottom-up correspondence-based techniques are not robustly applicable for sparse-view inference. On the other, recent neural multi-view methods can optimize already known approximate camera poses but provide no mechanism to obtain these to begin with. In this work, our goal is to fill this void and develop a method that, given a small number of unposed images of a generic object, can associate them with (approximate) camera viewpoints. Towards this goal, we focus on inferring the camera rotation matrices corresponding to each input image and propose a top-down approach to predict these. However, we note that the 'absolute' rotation is not well-defined given an image of a generic object—it assumes a 'canonical' pose which is not always known a-priori (e.g. what is an identity rotation for a pen? or a plant?). In contrast, the *relative* rotation between two views is well-defined even if a canonical pose for the instance is not. Thus, instead of adopting the common paradigm of single-image based pose prediction, we learn to estimate the relative pose given a pair of input images. We propose a system that leverages such pairwise predictions to then infer a consistent set of global rotations given multiple images of a generic object.

A key technical question that we consider is regarding the formulation of such pairwise pose estimation. Given two informative views of a rotationally asymmetric object, a regression-based approach may be able to accurately predict their relative transformation. The general case however, can be more challenging — given two views of a cup but with the handle only visible in one, the relative pose is ambiguous given just the two images. To allow capturing this uncertainty, we formulate an energy-based relative pose prediction network that, given two images *and* a candidate relative rotation, outputs an energy corresponding to the (unnormalized) log-probability of the hypothesis. This probabilistic estimation of relative pose not only makes the learning more stable, but more importantly, provides a mechanism to estimate a *joint distribution* over viewpoints given multiple images. We show that optimizing rotations to improve this joint likelihood yields coherent poses given multiple images and leads to significant improvements over naive approaches that do not consider the joint likelihoods.

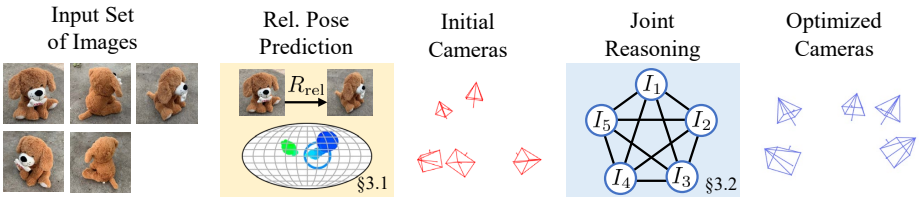

Fig. 2: **Overview.** From a set of images, we aim to recover corresponding camera poses (rotations). To do this, we train a pairwise pose predictor that takes in two images and a candidate relative rotation and predicts energy. By repeatedly querying this network, we recover a probability distribution over conditional relative rotations (see Sec. 3.1). We use these pairwise distributions to induce a joint likelihood over the camera transformations across multiple images, and iteratively improve an initial estimate by maximizing this likelihood (see Sec. 3.2).

We train our system using instances from over 40 commonplace object categories, and find that not only can it infer accurate (relative) poses for novel instances of these classes, it even generalizes to instances from unseen categories. Our approach can thus be viewed as a stepping stone toward sparse-view 3D reconstruction of generic objects; just as classical techniques provide precise camera poses that (neural) multi-view reconstruction methods can leverage, our work provides a similar, albeit coarser, output that can be used to initialize inference in current (and future) sparse-view reconstruction methods. While our system only outputs camera rotations, we note that a reasonable corresponding translation can be easily initialized assuming object-facing viewpoints, and we show that this suffices in practice for bootstrapping sparse-view reconstruction.

2 Related Work

Structure-from-Motion (SfM). At a high level, structure-from-motion aims to recover 3D geometry and camera parameters from image sets. This is done classically by computing local image features [2,21,30,66], finding matches across images [31], and then estimating and verifying epipolar geometry using bundle adjustment [67]. Later works have scaled up the SfM pipeline using sequential algorithms, demonstrating results on hundreds or even thousands of images [18, 52, 54, 55, 58].

The advent of deep learning has augmented various stages of the classical SfM pipeline. Better feature descriptors [14, 15, 46, 49, 57, 72, 79] and improved featured matching [9, 16, 29, 53, 68] have significantly outperformed their hand-crafted counterparts. BA-Net [63] and DeepSFM [75] have even replaced the bundle-adjustment process by optimizing over a cost volume. Most recently, Pixel-Perfect SfM [28] uses a featuremetric error to post-process camera poses to achieve sub-pixel accuracy.

While these methods can achieve excellent localization, all these approaches are bottom-up: beginning with local features that are matched across images. However, matching features requires sufficient overlap between images, which

may not be possible given wide baseline views. While our work also aims to localize camera poses given image sets, our approach fundamentally differs because it is top-down and does not rely on low-level correspondences.

Simultaneous Localization and Mapping (SLAM). Related is the task of Monocular SLAM, which aims to localize and map the surroundings from a video stream. Indirect SLAM methods, similar to SfM, match local features across different images to localize the camera [5, 37, 38, 51]. Direct SLAM methods, on the other hand, define a geometric objective function to directly optimize over a photometric error [11, 17, 56, 87].

There have also been various attempts to introduce deep learning into SLAM pipelines. As with SfM, learned feature descriptors and matching have helped improve accuracy on SLAM subproblems and increased robustness. End-to-end deep SLAM methods [40, 73, 74, 84] have improved the robustness of SLAM compared to classical methods, but have generally not closed the gap on performance. One notable exception is the recent DROID-SLAM [64], which combines the robustness of learning-based SLAM with the accuracy of classical SLAM.

These approaches all assume *sequential* streams and generally rely on matching or otherwise incorporating temporal locality between neighboring frames. We do not make any assumptions about the order of the image inputs nor the amount of overlap between nearby frames.

Single-view Pose Prediction. The task of predicting a (6-DoF) pose from a single image has a long and storied history, the surface of which can barely be scratched in this section. Unlike relative pose between multiple images, the (absolute) pose given a single image is only well-defined if there exists a canonical coordinate system. Most single-view pose prediction approaches therefore deal with a fixed set of categories, each of which has a canonical coordinate system defined *a priori* [4, 7, 23, 24, 26, 39, 42, 43, 59, 65, 71, 77]. Other methods that are category-agnostic take in a 3D mesh or point cloud as input, which provides a local coordinate system [44, 76, 78, 81].

Perhaps most relevant to us are approaches that not only predict pose but also model inherent uncertainty in the pose prediction [3, 10, 12, 13, 19, 25, 33, 36, 39, 45, 47, 61]. Like our approach, VpDR-Net [41] uses relative poses as supervision but still predicts absolute pose (with a unimodal Gaussian uncertainty model). Implicit-PDF [39] is the most similar approach to ours and served as an inspiration. Similar to our approach, Implicit-PDF uses a neural network to implicitly represent probability using an energy-based formulation which elegantly handles symmetries and multimodal distributions. Unlike our approach, Implicit-PDF (and all other single-view pose prediction methods) predict *absolute* pose, which does not exist in general for generic or novel categories. Instead, we model probability distributions over relative pose given pairs of images.

Learning-based Relative Pose Prediction. When considering generic scenes, prior works have investigated the task of relative pose prediction given two images. However, these supervised [69] or self-supervised [32, 70, 80, 85] methods typically consider prediction of motion between consecutive frames and are not easily adapted to wide-baseline prediction. While some approaches have investi-

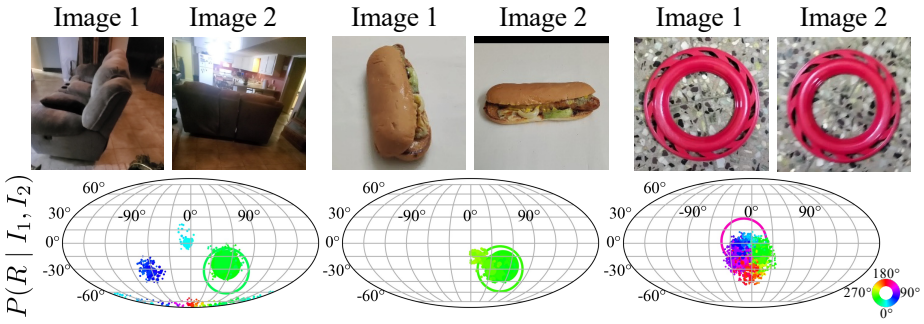

Fig. 3: **Predicted conditional distribution of image pairs from unseen categories.** Here, we visualize the predicted conditional distribution of image pairs. Inspired by [39], we visualize the rotation distribution (Alg. 1) by plotting yaw as latitude, pitch as longitude, and roll as the color. The size of each circle is proportional to the probability of that rotation. We omit rotations with negligible probability. The center of the open circle represents the ground truth. We can see that network predicts 4 modes for the couch images, corresponding roughly to 90 degree increments, with the greatest probability assigned to the correct 90 degree rotation. The relative pose of the hot dog is unambiguous and thus only has one mode. While the relative pose for the frisbee has close to no pitch or yaw, the roll remains ambiguous, hence the variety in colors. See the supplement for a visualization of how to interpret the relative rotations.

gated wide baseline prediction [1, 34], regression-based inference does not effectively capture uncertainty unlike our energy-based model. Perhaps most similar to ours is DirectionNet [8] which also predicts a camera distribution for wide baseline views. While DirectionNet only uses the expected value of the distribution and thus ignores symmetry, we take advantage of multimodal distributions to improve our joint pose estimation.

3 Method

Given a set of N images $\{I_1, \dots I_N\}$ depicting a *generic* object in-the-wild, we aim to recover a set of N rotation matrices $\{R_1, \dots, R_N\}$ such that rotation matrix R_i corresponds to the viewpoint of the camera used to take image i. Note that while we do not model translation, it can be easily initialized using object-facing viewpoints for 3D object reconstruction [27, 82] or a pose graph for SLAM [6]. We are primarily interested in settings with only sparse views and wide baselines. While bottom-up correspondence based techniques can reliably recover camera pose given dense views, they do not adapt well to sparse views with minimal overlap. We instead propose a prediction-based top-down approach that can learn and exploit the global structure directly.

The basic building block of our prediction system (visualized in Fig. 3) is a pairwise pose predictor that infers *relative* camera orientations given pairs of images. However, symmetries in objects and possibly uninformative viewpoints make this an inherently uncertain prediction task. To allow capturing this un-

certainty, we propose an energy-based approach that models the *multi-modal distribution* over relative poses given two images.

Given the predicted distributions over pairwise relative rotations, we show that these can be leveraged to induce a *joint* distribution over the rotations. Starting with a greedy initialization, we present a coordinate-ascent approach that jointly reasons over and improves the set of inferred rotations. We describe our approach for modeling probability distributions over relative poses between two images in Sec. 3.1, and build on this in Sec. 3.2 to recover a joint set of poses across multiple images. Finally, we discuss implementation details in Sec. 3.3.

3.1 Estimating Pair-wise Relative Rotations

Given a pair of images depicting an arbitrary object, we aim to predict a distribution over the relative rotation corresponding to the camera transformation between the two views. As there may be ambiguities when inferring the relative pose given two images, we introduce a formulation that can model uncertainty.

procedure PAIRWISEDISTRIBUTION(I_1, I_2)
  `queries` ←SAMPLEROTATIONSUNIF(50000)
  `energies` ← $f(I_1, I_2,$ `queries`$)$
  `probs` ← SOFTMAX(`energies`)
  return `queries`, `probs`
end procedure

Algorithm 1: **Pseudo-code for recovering a pairwise distribution**. We describe how to recover the distribution of the relative pose given images.

Energy-based Formulation. We wish to model the conditional distribution over a relative rotation matrix R given input images I_1 and I_2: $P(R \mid I_1, I_2)$. Inspired by recent work on *implicitly* representing the distribution over rotations using a neural network [39], we propose using an energy-based relative pose estimator. More specifically, we train a network $f(R, I_1, I_2)$ that learns to predict the energy, or the unnormalized joint log-probability, $P(R, I_1, I_2) = \alpha \exp f(R, I_1, I_2)$ where α is the constant of integration. From the product rule, we can recover the conditional probability as a function of f:

$$P(R \mid I_1, I_2) = \frac{P(R, I_1, I_2)}{P(I_1, I_2)} \approx \frac{\alpha \exp f(R, I_1, I_2)}{\sum_{R'} \alpha \exp f(R', I_1, I_2)} = \frac{\exp f(R, I_1, I_2)}{\sum_{R'} \exp f(R', I_1, I_2)} \quad (1)$$

We marginalize over rotations to avoid having to compute α (see Alg. 1), but note that the number of sampled rotations should be large for the approximation to be accurate. It is therefore important to use a lightweight network f since it is queried once per sampled rotation in the denominator.

Training. We train our network by maximizing the log-likelihood of the conditional distribution, or equivalently minimizing the negative log-likelihood:

$$\mathcal{L} = -\log P(R_1^\top R_2 \mid I_1, I_2) \quad (2)$$

where R_1 and R_2 are the ground truth poses of I_1 and I_2 respectively. Note that while the 'absolute' poses (R_1, R_2) are in an arbitrary coordinate system (depending on e.g. SLAM system outputs), the relative pose $R_1^\top R_2$ between two

views is agnostic to this incidental canonical frame. Following (1), we sample multiple candidate rotation matrices to compute the conditional probability.

Inference. Recovering the optimal transformation from the pose of I_1 to I_2 amounts to optimizing f over the space of rotations:

$$R^* = \underset{R \in \mathbf{SO}(3)}{\arg\max}\, P(R \mid I_1, I_2) = \underset{R \in \mathbf{SO}(3)}{\arg\max}\, f(R, I_1, I_2) \tag{3}$$

In practice, the loss landscape of f is often un-smooth, so we find that sampling and scoring rotations based on f to be more effective than gradient ascent.

We can also compute the conditional distribution of the relative rotation from I_1 to I_2 by sampling rotations over $\mathbf{SO}(3)$. The probability associated with each rotation can be computed using a softmax function, as described Alg. 1 and derived in (1). Inspired by [39], we can visualize the distribution of rotations by projecting the rotation matrices on a 2-sphere using pitch and yaw and coloring the rotation based on roll. See Fig. 3 and the supplement for sample results.

3.2 Recovering Joint Poses

In the previous section, we describe an energy-based relative pose predictor conditioned on pairs of images. Using this network, we recover a coherent set of rotations when given a set of images.

Greedy Initialization. Given predictions for relative rotations between every pair of images, we aim to associate each image with an absolute rotation. However, as the relative poses are invariant up to a global rotation, we can treat the pose of the first image as the identity matrix: $R_1 = I$. We note that the rotations for the other images can be uniquely induced given any $N - 1$ relative rotations that span a tree.

procedure CoordAsc(Images $\{I_i\}_N$)
  $\{R_i\}_N \leftarrow$ InitializeRotations($\{I_i\}_N$)
  for $t \in 1, \ldots,$ Num Iterations **do**
    $k \leftarrow$ RandomInteger(N)
    ▷ R'_k (Q×3×3): Q replacements for R_k
    $R'_k \leftarrow$SampleRotationsUnif(Q=250000)
    `energs` $\leftarrow$ Zeros(Q)
    for $i \in 1, \ldots, N$ **and** $i \neq k$ **do**
      $R \leftarrow$ Repeat(R_i, Q) ▷ 3×3→Q×3×3
      `energs` $\leftarrow$ `energs` $+ f(I_i, I_k, R^\top R'_k)$
      `energs` $\leftarrow$ `energs` $+ f(I_k, I_i, R'^\top_k R)$
    end for
    $R_k \leftarrow R'_k$[ArgMax(`energs`)]
  end for
end procedure

Algorithm 2: **Pseudo-code for joint inference using relative pose predictor**. We describe how to recover the joint poses given n images via coordinate ascent.

Sequential Chain. Perhaps the simplest way to construct such a tree is to treat the images as part of an ordered sequence. Given $R_1 = I$, all subsequent poses can be computed by using the best scoring relative pose from the previous image: $R_i = R_{i-1} R^*_{(i-1)\to i}$, denoting $R_{i\to j}$ as the relative rotation matrix $R_i^\top R_j$. However, this assumes that the images are captured sequentially (e.g. in a video) and may not be applicable for settings such as online marketplaces.

Maximum Spanning Tree. We improve over the naive linear chain by recognizing that some pairs of images may produce more confident predictions. Given N images, we construct a directed graph with $N \cdot (N-1)$ edges, where the weight of edge $(i, j) = P(R^*_{i\to j}|I_i, I_j)$. We then construct a Maximum Spanning Tree (MST) that covers all images with the most confident set of relative rotations.

Reasoning over all images jointly. Both of the previous methods, which select a subset of edges, do not perform any joint reasoning and discard all but the highest scoring mode for each pair of images. Instead, we can take advantage of our energy-based formulation to enforce global consistency.

Given our pairwise conditional probabilities, we can define a joint distribution over the set of rotations:

$$P\left(\{R_i\}_{i=1}^N \mid \{I_i\}_{i=1}^N\right) = \alpha \exp\Big(\sum_{(i,j)\in\mathcal{P}} f(R_{i\to j} \mid I_i, I_j)\Big) \tag{4}$$

where $\mathcal{P} = \{(i, j) \mid (i, j) \in [N] \times [N], i \neq j\}$ is the $N(N-1)$ set of pairwise permutations and α is the normalizing constant. Intuitively, this corresponds to the distribution modeled by a factor graph with a potential function corresponding to each pairwise edge.

We then aim to find the most likely set of rotations $\{R_1, \ldots, R_N\}$ under this conditional joint distribution (assuming $R_1 = I$). While it is not feasible to analytically obtain the global maxima, we adopt an optimization-based approach and iteratively improve the current estimate. More specifically, we initialize the set of poses with the greedy MST solution and at each iteration, we randomly select a rotation R_k to update. Assuming fixed values for $\{R_i\}_{i\neq k}$, we then search for the rotation R_k under the conditional distribution that maximizes the overall likelihood. We show in supplementary that this in fact corresponds to computing the most likely hypothesis under the distribution $P(R'_k \mid \{R_i\}_{i\neq k}, \{I_i\}_i)$:

$$\log P(R'_k \mid \{R_i\}_{i\neq k}, \{I_i\}_i) = \sum_{i\neq k} \left(f(R_{i\to k'}, I_i, I_k) + f(R_{k'\to i}, I_k, I_i)\right) + C \tag{5}$$

Analogous to our approach for finding the optimal solution for a single relative rotation, we sample multiple hypotheses for the rotation R_k, and select the hypothesis that maximizes (5). We find that this search-based block coordinate ascent helps us consistently improve over the initial solution while avoiding the local optima that a continuous optimization is susceptible to. We provide pseudocode in Alg. 2 and visualize one iteration of coordinate ascent in Fig. 4.

3.3 Implementation Details

Network Architecture. We use a ResNet-50 [22] with anti-aliasing [83] to extract image features. We use a lightweight 3-layer MLP that takes in a concatenation of 2 sets of image features and a rotation matrix to predict energy. We use positional encoding [35, 62] directly on flattened 3×3 rotation matrix, similar to [39]. See the supplement for architecture diagrams.

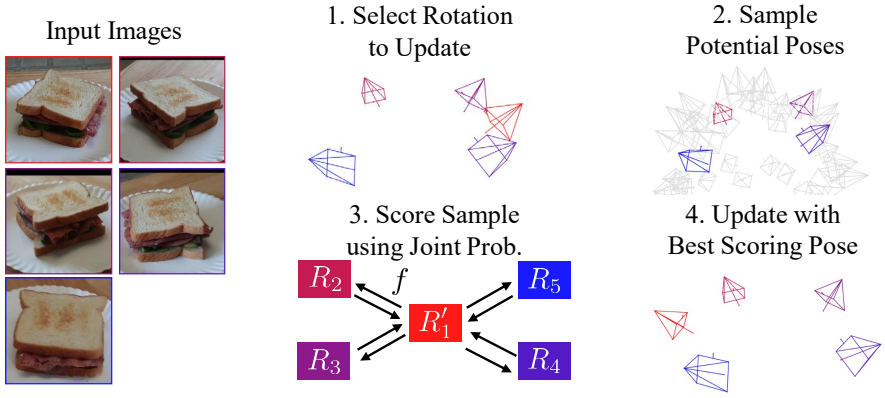

Fig. 4: **Recovering Joint Poses with Coordinate Ascent.** Given a set of images $\{I_1, \ldots, I_N\}$, we initialize a set of corresponding poses $\{R_1, \ldots, R_N\}$. During each iteration of coordinate ascent, we: 1) randomly select one pose R_k to update (the red camera in this case); 2) sample a large number (250k) of candidate poses; 3) score each pose according to the joint distribution conditioned on the other poses and images (5); and 4) update with the highest scoring pose. See Sec. 3.2 for more detail.

Number of Rotation Samples. We use the equivolumetric sampling in [39] to compute query rotations (37k total rotations) during training. For each iteration of coordinate ascent, we randomly sample 250k rotation matrices. For visualizing distributions, we randomly sample 50k rotations.

Runtime. We train the pairwise estimator with a batch size of 64 images for approximately 2 days on 4 NVIDIA 2080TI GPUs. Inference for 20 images takes around 1-2 seconds to construct an MST and around 2 minutes for 200 iterations of coordinate ascent on a single 2080TI. Note that the runtime of the coordinate ascent scales linearly with the number of images.

4 Evaluation

4.1 Experimental Setup

Dataset. We train and test on the Common Objects in 3D dataset (CO3D) [48], a large-scale dataset consisting of turntable-style videos of 51 common object categories. We train on the subset of the dataset that has camera poses, which were acquired by running COLMAP [54] over all frames of the video.

To train our network, we sample random frames and their associated camera poses from each video sequence. We train on 12,299 video sequences (from the `train-known` split) from 41 categories, holding out 10 categories to test generalization. We evaluate on 1,711 video sequences (from the `test-known` split) over all 41 trained categories (seen) as well as the 10 held out categories (unseen). The 10 held out categories are: `ball`, `book`, `couch`, `frisbee`, `hotdog`, `kite`, `remote`, `sandwich`, `skateboard`, and `suitcase`. We selected these categories randomly after excluding some of the categories with the most training images.

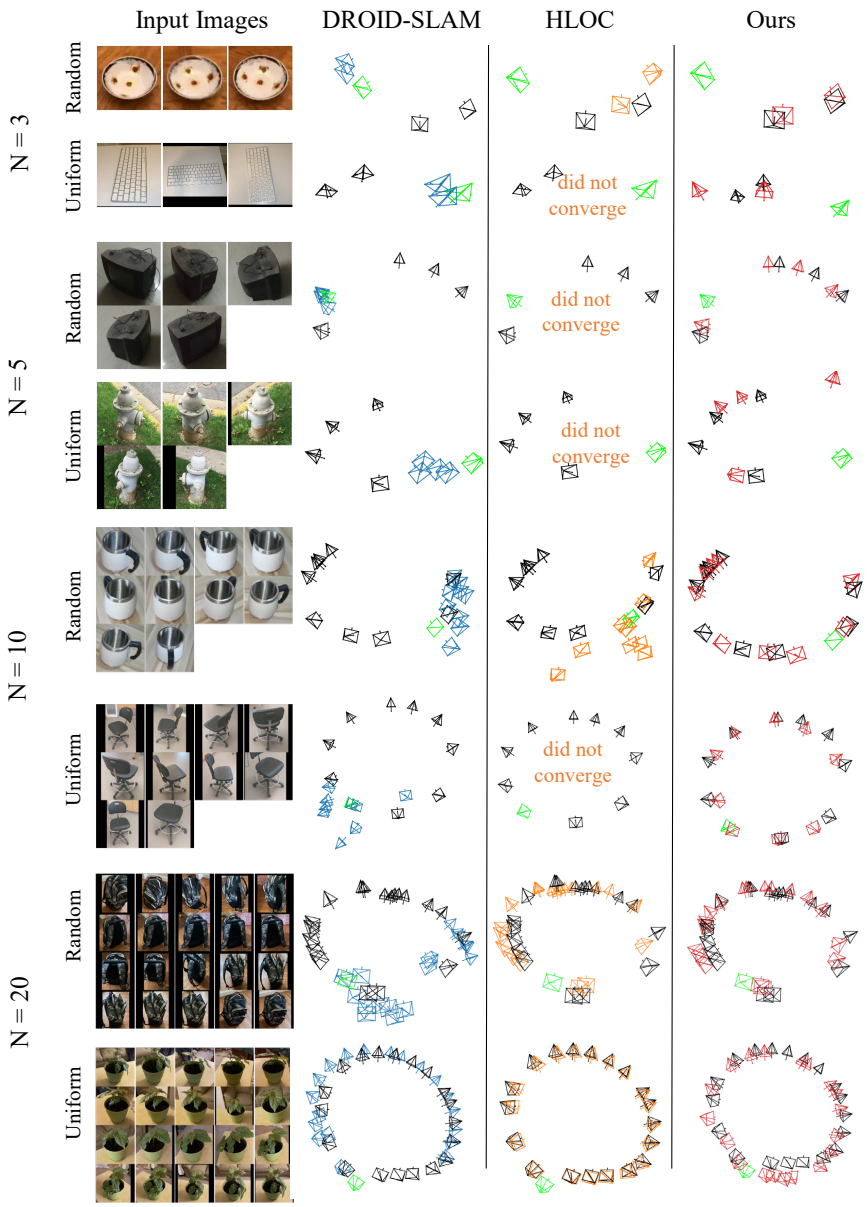

Fig. 5: **Qualitative Comparison of Recovered Camera Poses with Baselines.** We visualize the camera poses (rotations) predicted by DROID-SLAM, COLMAP with SuperPoint/SuperGlue, and our method given sparse image frames. The black cameras correspond to the ground truth. We only visualize the rotations predicted by each method, and set the translation such that the object center is a fixed distance away along the camera axis. As the poses are agnostic to a global rotation, we align the predicted cameras across all methods to the ground truth coordinate system by setting the recovered camera pose for the first image to the corresponding ground truth (visualized in green). Odd rows correspond to randomly sampled image frames, while even rows correspond to uniformly-spaced image frames.

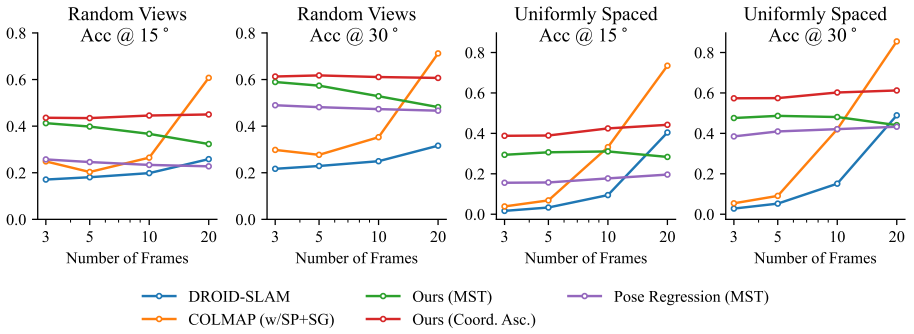

Fig. 6: **Mean Accuracy on *Seen* Categories.** We evaluate our approach against competitive SLAM (DROID-SLAM) and SfM (COLMAP with SuperPoint + SuperGlue) baselines in sparse-view settings. We also train a direct relative rotation predictor (Pose Regression) that is not probabilistic and uses the MST generated by our method to recover joint pose. We consider both randomly sampling and uniformly spacing frames from a video sequence. We report the proportion of pairwise relative poses that are within 15 and 30 degrees of the ground truth, averaged over all seen categories. We find that our approach shines with fewer views because it does not rely on correspondences and thus can handle wide baseline views. Correspondence-based approach need about 20 images to begin to work.

Task and Metrics. We consider the task of sparse-view camera pose estimation with $N = 3$, 5, 10, and 20 images, subsampled from a video sequence. This is highly challenging, especially when $N \leq 10$, because the ground truth camera poses have wide baselines.

We consider two possible ways to select N frames from a video sequence. First, we can randomly sample a set of N indices per video sequence (Random). Alternatively, we can use N uniformly-spaced frame indices (Uniform). We note that because CO3D video sequences are commonly taken in a turntable fashion, the uniformly spaced sampling strategy may be more representative of real world distributions of sparse view image sets. We report metrics on both task setups.

Because the global transformation of the camera poses is ambiguous, we evaluate each pair of relative rotations. For each of the $N(N-1)$ pairs, we compare the angular difference between the relative predicted rotation and the relative ground truth rotation using Rodrigues' formula [50]. We report the proportion of relative rotations that are within 15 and 30 degrees of the ground truth. We note that rotation errors within this range are relatively easy to handle by downstream 3D reconstruction tasks (See Fig. 10 for an example).

Baselines. We compare against DROID-SLAM [64], a current state-of-the-art SLAM approach that incorporates learning in an optimization framework. Note that DROID-SLAM requires trajectories and camera intrinsics. Thus, we provide the DROID-SLAM baseline with sorted frame indices and intrinsics, but do not provide these to any other method.

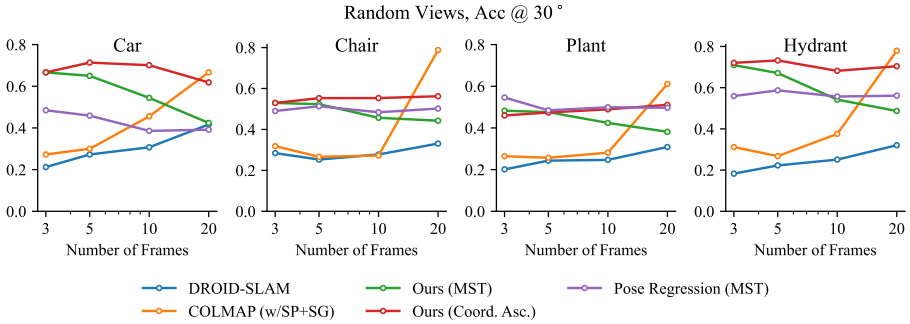

Fig. 7: **Accuracy on Subset of *Seen* Categories.** Here we compare all approaches on a representative subset of seen categories. We find that direct regression of relative poses (purple) struggles more on categories with symmetry (Car, Hydrant) than categories without symmetry (Chair, Plant), suggesting that multimodal prediction is important for resolving ambiguity.

We also compare with a state-of-the-art structure-from-motion pipeline that uses COLMAP [54] with SuperPoint feature extraction [14] and SuperGlue matching [53]. We used the implementation provided by [52]. For instances for which COLMAP does not converge or is unable to localize some cameras, we treat the missing poses as identity rotation for evaluation. We note that DROID-SLAM also outputs approximate identity rotations when the optimization fails.

Ablations. In the spirit of learning-based solutions that directly regress pose, we train a network that predicts relative rotation directly given two images. Similar to our energy-based predictor, we pass the concatenated images features from a ResNet-50 into an MLP. We double the number of layers from 3 to 6 and add a skip connection to give this network increased capacity. Rotations are predicted using the 6D rotation representation [86]. See the supplement for additional architecture details. The relative pose regressor cannot directly predict poses for more than two images. To recover sets of poses from sets of images, we use the MST graph recovered by our method to link the pairs of relative rotations (we find that this performs better than linking the relative rotations sequentially).

To demonstrate the benefits of joint reasoning, we additionally report the performance of our method using the greedy Maximum Spanning Tree (MST) solution. The performance of the sequential solution is in the supplement.

4.2 Quantitative Evaluation

We evaluate all approaches on sparse-view camera pose estimation by averaging over all seen categories in Fig. 6. We find that our approach outperforms all baselines for $N \leq 10$ images. Correspondence-based approaches (DROID-SLAM and COLMAP) do not work until roughly 20 images, at which point image frames have sufficient overlap for local correspondences. However, real world multi-view data (e.g. marketplace images) typically have much fewer images. We find that

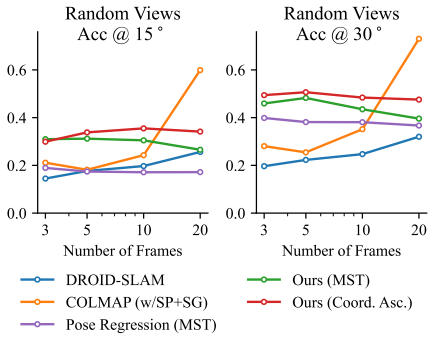

Fig. 8: **Mean Accuracy on *Unseen* Categories.** We evaluate our approach on held out categories from CO3D.

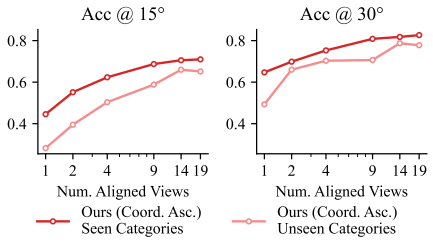

Fig. 9: **Novel View Registration.** Here, we evaluate the task of registering a new view given previously aligned cameras. We find that adding more views improves performance, suggesting that additional views reduce ambiguity.

coordinate ascent helps our approach scale with more image frames whereas the greedy maximum spanning tree accumulates errors with more frames.

Directly predicting relative poses does not perform well, possibly because pose regression cannot model multiple modes, which is important for symmetrical objects. We visualize the performance for four categories in Fig. 7. We find that the performance gap between our approach and direct regression is larger for objects with some symmetry (car, hydrant) than for objects without symmetry (chair, plant). Moreover, unlike our energy-based approach that models a joint distribution, a regression-based method does not allow similar joint reasoning.

We also test the generalization of our approach for *unseen* categories in Fig. 8. We still find that our method significantly outperforms all other approaches with sparse view ($N \leq 10$) even for never-before-seen object categories, indicating its ability to handle generic objects beyond training. The per-category evaluation for both seen and unseen categories are in the supplement.

Novel View Registration. In our standard SfM-inspired task setup, we aim to recover N camera poses given N images. Intuitively, adding images reduces ambiguity, but recovering additional cameras is also more challenging. To disambiguate between the two, we evaluate the task of registering new views given previously aligned images in Fig. 9. Given $N+1$ images, of which N have aligned cameras, we use our energy-based regressor to recover the remaining camera (equivalent to one iteration of coordinate ascent). We find that adding images improves accuracy, suggesting that additional views can reduce ambiguity.

4.3 Qualitative Results

We show qualitative results on the outputs of our pairwise predictor in Fig. 3. The visualized distributions suggest that our model is learning useful information about symmetry and can model multiple modes even for unseen categories.

We visualize predicted camera poses for DROID-SLAM, COLMAP, and our method with coordinate ascent in Fig. 5. Unable to bridge the domain gap from

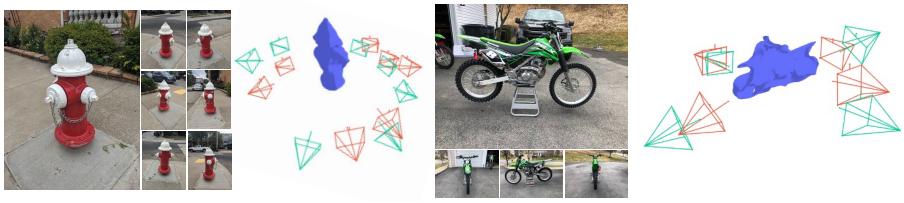

Fig. 10: **Initializing 3D NeRS Reconstruction using Predicted Cameras.** NeRS [82] is a representative 3D reconstruction approach that takes noisy cameras as initialization and jointly optimizes object shape, appearance, and camera poses. We run our method with coordinate ascent on 7 input images of a fire hydrant and 4 input images of a motorbike to obtain the camera initialization (green), which we provide to NeRS. NeRS then finetunes the cameras (red) and outputs a 3D reconstruction.

narrow baseline video frames, DROID-SLAM often gets stuck in the trajectory. Although COLMAP sometimes fails to converge, it performs well for N=20. Our approach consistently outputs plausible interpretations but is unable to achieve *precise* localization. See supplementary for visualizations on randomly selected sequences and more category-specific discussion.

We also validate that our camera pose estimations can be used for downstream 3D reconstruction. We use our camera poses to initialize NeRS [82], a representative sparse-view surface-based approach that requires a (noisy) camera initialization. Using our cameras, we successfully reconstruct a 3D model of a fire hydrant from 7 images and a motorbike from 4 images in Fig. 10. Note that the camera pose initialization in the original NeRS paper was manually selected.

5 Discussion

We presented a prediction-based approach for estimating camera rotations given (a sparse set of) images of a generic object. Our energy-based formulation allows capturing the underlying uncertainty in relative poses, while also enabling joint reasoning over multiple images. We believe our system's robustness under sparse views can allow it to serve as a stepping stone for initializing (neural) reconstruction methods in the wild, but also note that there are several open challenges. First, our work reasoned about the joint distribution using only pairwise potentials and developing efficient higher-order energy models may further improve performance. Moreover, while we outperform existing techniques given sparse-views, the correspondence-driven methods are more accurate given a large number of views, and we hope future efforts can unify the two approaches. Finally, our approach may not be directly applicable to reasoning about camera transformations for arbitrary scenes as modeling camera translation would be more important compared to object-centric images.

Acknowledgements. We would like to thank Gengshan Yang, Jonathon Luiten, Brian Okorn, and Elliot Wu for helpful feedback and discussion. This work was supported in part by the NSF GFRP (Grant No. DGE1745016), Singapore DSTA, and CMU Argo AI Center for Autonomous Vehicle Research.

References

1. Balntas, V., Li, S., Prisacariu, V.: RelocNet: Continuous Metric Learning Relocalisation using Neural Nets. In: ECCV (2018) 5
2. Bay, H., Tuytelaars, T., Gool, L.V.: SURF: Speeded Up Robust Features. In: ECCV (2006) 3
3. Brachmann, E., Michel, F., Krull, A., Yang, M.Y., Gumhold, S., et al.: Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image. In: CVPR (2016) 4
4. Bukschat, Y., Vetter, M.: EfficientPose: An Efficient, Accurate and Scalable End-to-end 6D Multi Object Pose Estimation Approach. arXiv:2011.04307 (2020) 4
5. Campos, C., Elvira, R., Gómez, J.J., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM. T-RO **37**(6), 1874–1890 (2021) 4
6. Carlone, L., Tron, R., Daniilidis, K., Dellaert, F.: Initialization techniques for 3D SLAM: A survey on rotation estimation and its use in pose graph optimization. ICRA (2015) 5
7. Chen, B., Chin, T.J., Klimavicius, M.: Occlusion-Robust Object Pose Estimation with Holistic Representation. In: WACV (2022) 4
8. Chen, K., Snavely, N., Makadia, A.: Wide-Baseline Relative Camera Pose Estimation with Directional Learning. In: CVPR (2021) 5
9. Choy, C.B., Gwak, J., Savarese, S., Chandraker, M.: Universal Correspondence Network. NeurIPS (2016) 3
10. Corona, E., Kundu, K., Fidler, S.: Pose Estimation for Objects with Rotational Symmetry. In: IROS (2018) 4
11. Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O.: MonoSLAM: Real-time Single Camera SLAM. TPAMI **29**(6), 1052–1067 (2007) 4
12. Deng, X., Mousavian, A., Xiang, Y., Xia, F., Bretl, T., Fox, D.: PoseRBPF: A Rao-Blackwellized Particle Filter for 6D Object Pose Tracking. In: RSS (2019) 4
13. Deng, X., Xiang, Y., Mousavian, A., Eppner, C., Bretl, T., Fox, D.: Self-supervised 6D Object Pose Estimation for Robot Manipulation. In: ICRA (2020) 4
14. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-supervised Interest Point Detection and Description. In: CVPR-W (2018) 3, 12
15. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In: CVPR (2019) 3
16. Dusmanu, M., Schönberger, J.L., Pollefeys, M.: Multi-view Optimization of Local Feature Geometry. In: ECCV (2020) 3
17. Engel, J., Koltun, V., Cremers, D.: Direct Sparse Odometry. TPAMI (2018) 4
18. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Towards Internet-scale Multi-view Stereo. In: CVPR (2010) 3
19. Gilitschenski, I., Sahoo, R., Schwarting, W., Amini, A., Karaman, S., Rus, D.: Deep Orientation Uncertainty Learning based on a Bingham Loss. In: ICLR (2019) 4
20. Goel, S., Gkioxari, G., Malik, J.: Differentiable Stereopsis: Meshes from multiple views using differentiable rendering. In: CVPR (2022) 2
21. Harris, C., Stephens, M.: A Combined Corner and Edge Detector. In: Alvey Vision Conference (1988) 3
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR (2016) 8
23. Iwase, S., Liu, X., Khirodkar, R., Yokota, R., Kitani, K.M.: RePOSE: Fast 6D Object Pose Refinement via Deep Texture Rendering. In: ICCV (2021) 4

24. Kehl, W., Manhardt, F., Tombari, F., Ilic, S., Navab, N.: SSD-6D: Making RGB-based 3D Detection and 6D Pose Estimation Great Again. In: ICCV (2017) 4
25. Kendall, A., Cipolla, R.: Modelling Uncertainty in Deep Learning for Camera Relocalization. In: ICRA (2016) 4
26. Kendall, A., Grimes, M., Cipolla, R.: PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In: ICCV (2015) 4
27. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: BARF: Bundle-Adjusting Neural Radiance Fields. In: ICCV (2021) 2, 5
28. Lindenberger, P., Sarlin, P.E., Larsson, V., Pollefeys, M.: Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. In: ICCV (2021) 3
29. Liu, C., Yuen, J., Torralba, A.: SIFT Flow: Dense Correspondence Across Scenes and Its Applications. TPAMI **33**(5), 978–994 (2010) 3
30. Lowe, D.G.: Distinctive Image Features from Scale-invariant Keypoints. IJCV **60**(2), 91–110 (2004) 3
31. Lucas, B.D., Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision. In: IJCAI (1981) 3
32. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised Learning of Depth and Ego-Motion From Monocular Video Using 3D Geometric Constraints. In: CVPR (2018) 4
33. Manhardt, F., Arroyo, D.M., Rupprecht, C., Busam, B., Birdal, T., Navab, N., Tombari, F.: Explaining the Ambiguity of Object Detection and 6D Pose from Visual Data. In: ICCV (2019) 4
34. Melekhov, I., Ylioinas, J., Kannala, J., Rahtu, E.: Relative Camera Pose Estimation Using Convolutional Neural Networks. In: ACIVS (2017) 5
35. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In: ECCV (2020) 8
36. Mohlin, D., Sullivan, J., Bianchi, G.: Probabilistic Orientation Estimation with Matrix Fisher Distributions. NeurIPS (2020) 4
37. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: A Versatile and Accurate Monocular SLAM System. T-RO **31**(5), 1147–1163 (2015) 4
38. Mur-Artal, R., Tardós, J.D.: ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. T-RO **33**(5), 1255–1262 (2017) 4
39. Murphy, K.A., Esteves, C., Jampani, V., Ramalingam, S., Makadia, A.: Implicit-PDF: Non-Parametric Representation of Probability Distributions on the Rotation Manifold. In: ICML (2021) 4, 5, 6, 7, 8, 9
40. Newcombe, R.A., Lovegrove, S.J., Davison, A.J.: DTAM: Dense Tracking and Mapping in Real-time. In: ICCV (2011) 4
41. Novotny, D., Larlus, D., Vedaldi, A.: Learning 3D Object Categories by Looking Around Them. In: ICCV (2017) 4
42. Novotny, D., Ravi, N., Graham, B., Neverova, N., Vedaldi, A.: C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion. In: ICCV (2019) 4
43. Oberweger, M., Rad, M., Lepetit, V.: Making Deep Heatmaps Robust to Partial Occlusions for 3D Object Pose Estimation. In: ECCV (2018) 4
44. Okorn, B., Gu, Q., Hebert, M., Held, D.: ZePHyR: Zero-shot Pose Hypothesis Scoring. In: ICRA (2021) 4
45. Okorn, B., Xu, M., Hebert, M., Held, D.: Learning Orientation Distributions for Object Pose Estimation. In: IROS (2020) 4
46. Pautrat, R., Larsson, V., Oswald, M.R., Pollefeys, M.: Online Invariance Selection for Local Feature Descriptors. In: ECCV (2020) 3

47. Prokudin, S., Gehler, P., Nowozin, S.: Deep Directional Statistics: Pose Estimation with Uncertainty Quantification. In: ECCV (2018) 4
48. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction. In: ICCV (2021) 9
49. Revaud, J., De Souza, C., Humenberger, M., Weinzaepfel, P.: R2D2: Reliable and Repeatable Detector and Descriptor. NeurIPS (2019) 3
50. Rodrigues, O.: Des lois géométriques qui régissent les déplacements d'un système solide dans l'espace, et de la variation des coordonnées provenant de ces déplacements considérés indépendamment des causes qui peuvent les produire. Journal de Mathématiques Pures et Appliquées **5** (1840) 11
51. Rosinol, A., Abate, M., Chang, Y., Carlone, L.: Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping. In: ICRA (2020) 4
52. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In: CVPR (2019) 3, 12
53. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning Feature Matching with Graph Neural Networks. In: CVPR (2020) 3, 12
54. Schönberger, J.L., Frahm, J.M.: Structure-from-Motion Revisited. In: CVPR (2016) 3, 9, 12
55. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise View Selection for Unstructured Multi-View Stereo. In: ECCV (2016) 3
56. Schops, T., Sattler, T., Pollefeys, M.: BAD SLAM: Bundle Adjusted Direct RGB-D SLAM. In: CVPR (2019) 4
57. Simonyan, K., Vedaldi, A., Zisserman, A.: Learning Local Feature Descriptors Using Convex Optimisation. TPAMI **36**(8), 1573–1585 (2014) 3
58. Snavely, N., Seitz, S.M., Szeliski, R.: Photo Tourism: Exploring Photo Collections in 3D. In: SIGGRAPH. ACM (2006) 3
59. Song, C., Song, J., Huang, Q.: Hybridpose: 6D Object Pose Estimation under Hybrid Representations. In: CVPR (2020) 4
60. Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J.B., Freeman, W.T.: Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling. In: CVPR (2018) 2
61. Sundermeyer, M., Marton, Z.C., Durner, M., Brucker, M., Triebel, R.: Implicit 3D Orientation Learning for 6D Object Detection from RGB Images. In: ECCV (2018) 4
62. Tancik, M., Srinivasan, P.P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J.T., Ng, R.: Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. NeurIPS (2020) 8
63. Tang, C., Tan, P.: BA-Net: Dense Bundle Adjustment Network. In: ICLR (2019) 3
64. Teed, Z., Deng, J.: DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. NeurIPS (2021) 4, 11
65. Tekin, B., Sinha, S.N., Fua, P.: Real-Time Seamless Single Shot 6D Object Pose Prediction. In: CVPR (2018) 4
66. Tola, E., Lepetit, V., Fua, P.: Daisy: An Efficient Dense Descriptor Applied to Wide-baseline Stereo. TPAMI **32**(5), 815–830 (2009) 3
67. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle Adjustment—A Modern Synthesis. In: International workshop on vision algorithms (1999) 3
68. Truong, P., Danelljan, M., Timofte, R.: GLU-Net: Global-Local Universal Network for Dense Flow and Correspondences. In: CVPR (2020) 3

69. Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: DeMoN: Depth and Motion Network for Learning Monocular Stereo. In: CVPR (2017) 4
70. Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R., Fragkiadaki, K.: SfM-Net: Learning of Structure and Motion from Video. arXiv:1704.07804 (2017) 4
71. Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., Savarese, S.: DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. In: CVPR (2019) 4
72. Wang, Q., Zhou, X., Hariharan, B., Snavely, N.: Learning Feature Descriptors Using Camera Pose Supervision. In: ECCV (2020) 3
73. Wang, S., Clark, R., Wen, H., Trigoni, N.: DeepVO: Towards End-to-End Visual Odometry with Deep Recurrent Convolutional Neural Networks. In: ICRA (2017) 4
74. Wang, W., Hu, Y., Scherer, S.: TartanVO: A Generalizable Learning-based VO. In: CoRL (2020) 4
75. Wei, X., Zhang, Y., Li, Z., Fu, Y., Xue, X.: DeepSFM: Structure From Motion Via Deep Bundle Adjustment. In: ECCV (2020) 3
76. Wong, J.M., Kee, V., Le, T., Wagner, S., Mariottini, G.L., Schneider, A., Hamilton, L., Chipalkatty, R., Hebert, M., Johnson, D.M., et al.: SegICP: Integrated Deep Semantic Segmentation and Pose Estimation. In: IROS (2017) 4
77. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In: RSS (2018) 4
78. Xiao, Y., Qiu, X., Langlois, P., Aubry, M., Marlet, R.: Pose from Shape: Deep Pose Estimation for Arbitrary 3D Objects. In: BMVC (2019) 4
79. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: LIFT: Learned Invariant Feature Transform. In: ECCV (2016) 3
80. Yin, Z., Shi, J.: GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. In: CVPR (2018) 4
81. Zhang, J.Y., Pepose, S., Joo, H., Ramanan, D., Malik, J., Kanazawa, A.: Perceiving 3D Human-Object Spatial Arrangements from a Single Image in the Wild. In: ECCV (2020) 4
82. Zhang, J.Y., Yang, G., Tulsiani, S., Ramanan, D.: NeRS: Neural Reflectance Surfaces for Sparse-view 3D Reconstruction in the Wild. In: NeurIPS (2021) 2, 5, 14
83. Zhang, R.: Making Convolutional Networks Shift-Invariant Again. In: ICML (2019) 8
84. Zhou, H., Ummenhofer, B., Brox, T.: DeepTAM: Deep Tracking and Mapping. In: ECCV (2018) 4
85. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised Learning of Depth and Ego-Motion From Video. In: CVPR (2017) 4
86. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR (2019) 12
87. Zubizarreta, J., Aguinaga, I., Montiel, J.M.M.: Direct sparse mapping. T-RO (2020) 4