Extreme Rotation Estimation in the Wild

Hana Bezalel[1] Dotan Ankri[1] Ruojin Cai[2] Hadar Averbach-Elor[1,2]
[1]Tel Aviv University [2]Cornell University

https://tau-vailab.github.io/ExtremeRotationsInTheWild/

Abstract

We present a technique and benchmark dataset for estimating the relative 3D orientation between a pair of Internet images captured in an extreme setting, where the images have limited or non-overlapping field of views. Prior work targeting extreme rotation estimation assume constrained 3D environments and emulate perspective images by cropping regions from panoramic views. However, real images captured in the wild are highly diverse, exhibiting variation in both appearance and camera intrinsics. In this work, we propose a Transformer-based method for estimating relative rotations in extreme real-world settings, and contribute the **ExtremeLandmarkPairs** *dataset, assembled from scene-level Internet photo collections. Our evaluation demonstrates that our approach succeeds in estimating the relative rotations in a wide variety of extreme-view Internet image pairs, outperforming various baselines, including dedicated rotation estimation techniques and contemporary 3D reconstruction methods.*

1. Introduction

The problem of estimating the relative 3D orientation between a pair of images is embodied in fundamental computer vision tasks, such as camera localization [6, 34, 35] and 3D reconstruction [29, 36, 42]. Establishing pixel correspondences (either explicitly or implicitly) is typically a prerequisite for computing the relative rotation between the images. Correspondences, however, cannot be extracted in extreme settings where the images have little to no overlap. As dense imagery may not necessarily be available for many practical applications, a natural question arises: How can we estimate the relative rotation between non-overlapping RGB images, without the use of additional data (such as depth or temporal information)?

We have recently seen pioneering efforts addressing the task of relative rotation estimation in such extreme non-overlapping settings [8, 12]. Prior work has proposed end-to-end neural architectures, demonstrating that *hidden* cues, such as vanishing points or the directions of cast shadows, can implicitly guide the model for inferring the relative orientation between the images. To facilitate learning and evaluation, datasets constructed from panoramic views were adopted. These datasets emulate perspective views by cropping sub-areas from these panoramas, enabling generation of image pairs with various degrees of overlap. However, while such emulated views perhaps capture some of the challenges associated with extreme-view imagery, are they sufficient for representing real images—particularly, images captured *in the wild*?

Figure 1. Given a pair of (possibly) non-overlapping images captured *in the wild*—*e.g.*, under arbitrary illumination and intrinsic camera parameters—such as the images of the Dam Square in Amsterdam depicted in red and blue boxes above*, our technique estimates the relative 3D rotation between the images. *The background panorama is illustrated for visualization purposes only.

In this paper, we present a new approach that tackles the problem of extreme rotation estimation in the wild. Consider the boxed images in Figure 1. Internet (*i.e.*, in the wild) images may vary due to a wide range of factors, including transient objects, weather conditions, time of day, and the cameras' intrinsic parameters. To explore this problem, we introduce a new dataset, *ExtremeLandmarkPairs* (*ELP*), assembled from publicly-available scene-level Internet image collections. We observe that the set of real extreme-view image pairs is limited, as Internet datasets are typically scene-centric, with nearby cameras commonly

capturing overlapping views. Therefore, to facilitate training, we propose a progressive learning scheme that leverages and augments images cropped from panoramic views, allowing for gradually generalizing the model onto real Internet data. In particular, we construct datasets with varying field of views, that better resemble the distribution of real data, and perform image-level appearance augmentations by leveraging recent advancements in text-to-image Diffusion models [7, 22, 31].

To estimate extreme rotations in the wild, we propose a Transformer-based model that is provided with auxiliary channels, including the spatial distribution of local keypoints and matches and semantic segmentation maps, allowing for better reasoning over real image pairs with little or no overlap. Our results demonstrate that our model can accurately predict the relative rotations for a wide variety of extreme-view image pairs that vary in illumination, dynamic regions, and intrinsic parameters. We conduct extensive experiments, quantifying performance both over real Internet data and also over emulated perspective images cropped from panoramic views. Our evaluation shows that our model significantly improves over strong baselines when considering real images, while achieving comparable performance over emulated perspective image pairs.

2. Related Work

Relative pose estimation is a fundamental task in computer vision, typically studied for overlapping camera views. Traditionally, this task has been divided into two stages: correspondence estimation from local feature matching, followed by geometry-based pose estimation. In recent years, feature matching methods have advanced from using heuristic feature descriptors [5, 27, 32] and RANSAC-based matches [21] to learning-based feature extraction [4, 14–16, 18, 40, 45] and matching techniques [3, 26, 33] , with several methods performing both feature extraction and matching using unified learning-based frameworks [4, 17, 38].

These methods are generally invariant to changes in illumination and appearance, demonstrating robust performance across various scene scales and also over in-the-wild datasets. However, their reliability diminishes in extreme view scenarios due to their dependence on visual overlap.

Extreme-view scenarios lacking local pixelwise correspondences necessitate the use of end-to-end learning-based pose estimation methods which directly infer the 3D relationships and geometry from sparse and extreme-view images. Indeed, large-scale 3D object datasets have paved the way for learning-based methods which estimate camera pose directly from sparse views [20, 25, 37, 43, 44, 50–52]. However, these methods primarily concentrate on object-centric scenes, under experimental settings that typically assume similar lighting and camera intrinsics for the input views. Furthermore, these methods often utilize bounding box inputs defining the object of interest, which is less suitable in the case of images depicting large-scale scenes.

Several prior works address a sparse view setting at scene-scale. In particular, Chen *et al.* [11] propose to learn a discrete distributions of pose space, Agarwala *et al.* [1] simplify scene reconstruction using a plane representation, and Rockwell *et al.* [30] introduce an inductive bias of the 8-point algorithm into a vision transformer architecture. Recently, models that directly predict pixel-aligned point maps from input pairs (or sparse image collections), such as DUSt3R [46] and Mast3R [23], have demonstrated promising results on pose estimation and scene-scale 3D reconstruction of Internet data with a wide baseline.

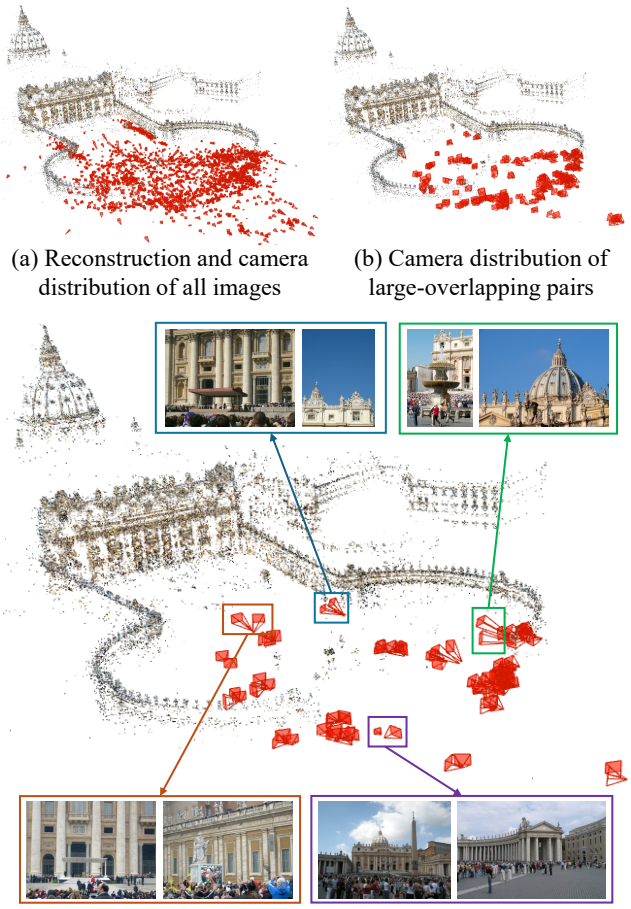
Camera pose estimation for non-overlapping views presents a greater challenge. Earlier efforts [10] explored searching for consistent temporal behavior. Several work utilize pairwise RGB and depth scan data to estimate relative pose among such extreme pairs [48, 49]. Cai et al. [8] tackle pose estimation for non-overlapping views without the use of additional data, by introducing a learning-based network leveraging cross-correlation volume to exploit implicit cues. This correlation volume is later enhanced through the integration of transformer attention modules [12]. Nonetheless, these approaches assume constrained 3D environments, including the assumption of consistent lighting and camera intrinsics, and are designed for camera distribution of emulated perspective views cropped from panoramas. In this work, we aim to address pose estimation for *realistic* in-the-wild non-overlapping image pairs, enhancing the applicability of extreme pose estimation to Internet photos and real-world data.

3. The *ExtremeLandmarkPairs* Dataset

Prior works on extreme pose estimation use panoramic views, cropping from it sub-areas to emulate perspective views [8, 12]. To evaluate and train models on real perspective image pairs, we propose a new benchmark and dataset, *ExtremeLandmarkPairs* (*ELP*), constructed from Internet image pairs from the MegaDepth [24], Cambridge Landmarks [2], and MegaScenes [39] datasets. In this section, we first describe the dataset construction procedure (Section 3.1), and then present details regarding dataset size and train and test splits (Table 1).

3.1. Dataset Construction

To construct a dataset of *real* perspective image pairs with varying degrees of overlap, we leverage available scene-level training data. Existing Internet image collections typically contain camera poses (predicted up to scale), which are determined using Structure-from-Motion (SfM) algorithms, such as COLMAP [36]. In what follows, we describe how we extract real image pairs from this data, which can then be used for training and evaluating models.



(a) Reconstruction and camera distribution of all images

(b) Camera distribution of large-overlapping pairs

(c) Camera distribution of non-overlapping pairs

Figure 2. Camera distribution of the Vatican, Rome scene from the *ExtremeLandmarkPairs* Dataset. We construct a dataset of real perspective image pairs with predominant rotational motion shown in (b) and (c) from the dense imagery reconstruction in (a).

Identifying Pairs with Predominant Rotational Motion. Prior works targeting relative rotation estimation, in particular for non-overlapping views, mostly utilize panoramic views, focusing on image pairs with purely rotational motion. Pairs belonging to *real* image collections almost always contain a non-negligible translation component. Furthermore, unlike in the StreetLearn [28] dataset used by prior work [8, 12] that provides exact translation values between consecutive panoramas which allows for filtering pairs with predominant rotational motions, reconstructed relative poses are only provided up to scale. The scale varies among different reconstructed scenes, and therefore there's no global threshold on the relative translation values which can be used for identifying pairs with predominant rotational motion.

To automatically identify such pairs, we observe that available Internet collections require the existence of dense imagery, to compensate for the vast number of unknowns in the SfM optimization. We therefore construct *mutual* nearest neighbors edge-weighted graphs, with one graph per landmark. In each graph G, nodes $v \in V$ correspond to images, and two images are connected by an edge $e \in E$ if they are both among each other's K nearest neighbors, considering L2 distances between their translations (K is empirically set to 5). Note that images captured from sparser (outlier) regions in space are unlikely to be within the mutual K nearest neighbors of images captured within denser regions, and hence won't be included in G. Finally, we select a subset of image pairs containing relatively small distances from each scene graph G, yielding a set of image pairs with predominant rotational motion; see the supplementary material for additional details.

Extracting Level of Overlap. Following prior work [8, 12], we are interested in training and evaluating models according to three different categories: *Large*, *Small* and *None*, indicating image pairs with a varying amount of overlap. However, unlike prior work that use cropped images with a fixed 90° FoV, Internet images contain varying FoV values. Thus, the relative rotation angle is not sufficient for extracting the pair's overlap level.

Denote the FoV values of image $i \in [1, 2]$ as $[\text{fov}_x^i, \text{fov}_y^i]$. We can parameterize a 3D rotation matrix $\mathbf{R}$ using three Euler angles $[\alpha, \beta, \gamma]$, denoting the relative roll, pitch and yaw angles, respectively:

$$\mathbf{R}(\alpha, \beta, \gamma) = \mathbf{R}_z(\alpha)\mathbf{R}_y(\beta)\mathbf{R}_x(\gamma), \tag{1}$$

following Dense Correlation Volumes' coordinate system. We use the following conditions to determine the overlap level o:

$$o = \begin{cases} \textit{Large} & |\gamma| < \frac{\text{fov}_x^1 + \text{fov}_x^2}{4} \wedge |\beta| < \frac{\text{fov}_y^1 + \text{fov}_y^2}{4} \\ \textit{None} & |\gamma| > \frac{\text{fov}_x^1 + \text{fov}_x^2}{2} \wedge |\beta| > \frac{\text{fov}_y^1 + \text{fov}_y^2}{2} \\ \textit{Small} & \textit{else} \end{cases} \tag{2}$$

In other words, pairs with relative yaw and pitch angles that are smaller than a quarter of the average corresponding FoV values are considered as pairs with a large overlap ratio. Likewise, pairs with relative yaw and pitch angles that are larger than half the average corresponding FoV values are considered non-overlapping pairs. Pairs in-between these conditions are considered pairs with a small overlap ratio.

Additional Filtering. In the scene-scale datasets we explore, large FoV disparities could result in one image focusing on specific architectural details like a statue, while the other captures a much broader scene perspective, further complicating the problem of estimating the relative rotations. To extract pairs consistent in scale, we limit the difference between the FoV values to be at most 5°. Furthermore, as most images contain small roll values, we rotate the scenes to match the gravity axis and horizontal axis

Subset	Source	#Pairs				
		#Scenes	Large	Small	None	Total
Train	[39]	5883	33430	13684	29481	76595
Validation	[39]	515	3398	710	4130	8238
Validation Balanced	[39]	177	92	55	707	854
*s*ELP	[2]	6	2512	827	1961	5300
*w*ELP	[24]	17	2700	829	643	4172

Table 1. ***ExtremeLandmarkPairs* Dataset Statistics**. Above, we report the number of image pairs extracted for each overlap level, split into train and test (with *s*ELP denoting a *single* camera setting and *w*ELP denoting the "in the *wild*" setting).

and filtered images with rolls exceeding 10°. Finally, we excluded images captured from an aerial perspective, by filtering images with translation along y-axis exceeding a global threshold, empirically set to 1.

3.2. Dataset Size and Splits

We apply the procedure described above over landmarks from the MegaScenes [39] dataset to create a training set and a validation set. As we focus on outdoor scenes in our work, we filter reconstructions that capture indoor scenes. We obtain a total of nearly 34K non-overlapping pairs originating from over 2K unique landmarks. The validation set was also balanced the set over the overall angle.

For evaluation, we create two test sets, to separately examine image pairs captured in a *single* camera setting with constant illumination (*s*ELP) and image pairs captured in the *wild* (*w*ELP). Image pairs in the *s*ELP test set contain images from the Cambridge Landmarks [2] dataset, which contains videos capturing six different landmarks. Image pairs in the *w*ELP test set contain images from the MegaDepth dataset [24], which contains Internet photos from Flickr for a set of large-scale landmarks.

As there is some overlap between the landmarks in MegaScenes [39] and MegaDepth [24], we performed additional filtering, ensuring no overlap exists between the train and test sets. Furthermore, we filter test pairs if one of the images is *highly* transient—*i.e.*, if transient objects dominate the image. We quantify this using a pretrained segmentation model [19], filtering images containing transient objects in over 40% of the pixels. We also manually validate non-overlapping image pairs, allowing to further filter unlabeled objects, such as a stage or a market stand. Finally, for non-overlapping image pairs, we observed that the overall relative rotations are highly imbalanced, and therefore we balance this set by the overall relative angle. Table 1 summarizes the number of image pairs and landmarks used for both training and test.

4. Method

Given a pair of Internet images with (possibly) extreme relative motion, we estimate the relative rotation $\mathbf{R}$ between the images. Following prior works on extreme rotation estimation [8, 12], we assume a camera-centric setting, where the two cameras have limited translation. However, our approach departs from prior works by operating on outdoor images captured by a crowd of photographers, with varying intrinsic parameters as well as appearances—*e.g.*, due to illumination changes and dynamic objects.

Our model (detailed in Section 4.1; see Figure 3 for an overview) outputs three discrete Euler angles, denoting the relative roll, pitch and yaw angles. As illustrated in prior work [8], this parameterization enables using a simple cross-entropy loss for training. In Section 4.2, we describe our progressive learning scheme, allowing for gradually adapting the model to extreme Internet imagery.

4.1. Model

We extract image features using a pretrained LoFTR model [38]. In contrast to the features extracted using common convolutional neural networks pretrained on ImageNet [13], LoFTR is a Transformer-based model trained on Internet pairs, with the goal of extracting local feature matches – a setting and task which is highly related to the one we address in our work, thus enabling extraction of better (*i.e.*, more relevant) features.

As we are interested in designing a network that can *also* reason over image pairs with little or no overlap, we combine the extracted features with additional *auxiliary* channels; see Figure 3 (bottom left). These include keypoint and pairwise matches masks, utilized previously for disambiguating images for similar structures [9]. Intuitively, knowledge over pairwise matches can assist the model in cases of small overlap and for generalizing across different camera intrinsic properties. We also incorporate a segmentation map as an additional auxiliary channel, which segments images into several categories (such as sky, building, road and sidewalk). This channel allows for identifying additional cues, such as the skyline or transient objects, which can aid in determining the rotation for non-overlapping pairs. In Section 5, we demonstrate the benefit of incorporating these auxiliary channels in our model.

We then reshape extracted features (and auxiliary channels) to tokens, concatenating these image tokens with learnable Euler angle position embeddings. These tokens are processed by our *Rotation Estimation Transformer* module, which uses a transformer decoder architecture [41]. The output Euler angle tokens obtain information from image features and auxiliary channels within transformer attention modules. These tokens are then processed by three different prediction heads (one per angle, denoted as *MLP* in Figure 3).Each output prediction head uses as input averaged image tokens and one of the output Euler angle tokens, which provides the model with additional angle-specific information, allowing for achieving improved performance,

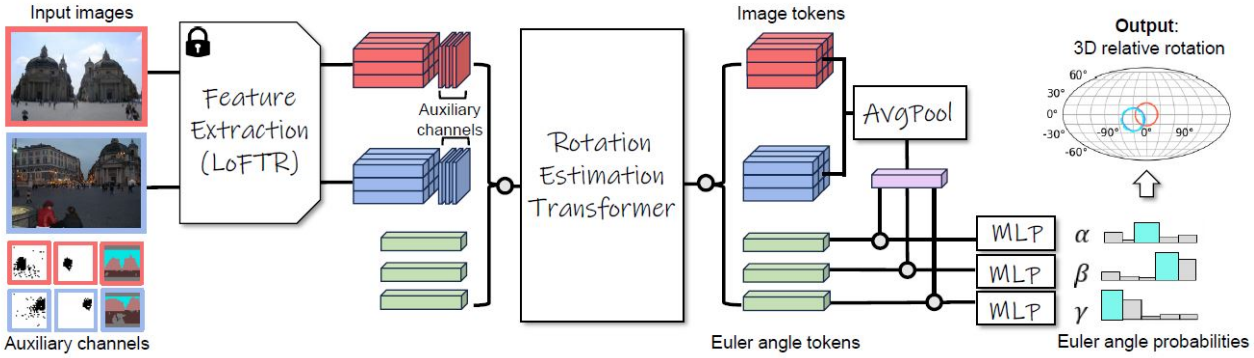


Figure 3. **Method architecture.** Given a pair of input Internet images, we extract image features using pretrained LoFTR. These features are combined with auxiliary channels, including keypoint and pairwise matches masks, and segmentation maps (visualized on the bottom left). These image features are reshaped into tokens and concatenated with Euler angle position embeddings, which are then processed by our Rotation Estimation Transformer module. The output Euler angle tokens and averaged image tokens are concatenated and processed by MLPs to predict the probability distribution of Euler angles, representing the relative 3D rotation between the input images.

as we show in the supplementary material. The prediction heads output a probability distribution over $N = 360$ bins, capturing an angle in the range $[-180°, 180°]$.

4.2. Learning

As detailed in Section 3, we assemble real image pairs from Internet imagery, which can be used for training and evaluating models. However, even with our proposed *ExtremeLandmarkPairs* dataset, the set of real image pairs is limited—*e.g.*, only ∼36K non-overlapping image pairs are extracted, as available Internet datasets are typically scene-centric, with nearby cameras usually capturing overlapping views. Therefore, in what follows, we propose a *progressive* learning scheme, which leverages panoramic images, and allows for gradually generalizing the model to images captured in-the-wild. All learning stages are optimized using a cross-entropy loss for each Euler angle prediction. Additional details are provided in the supplementary material.

Initialization. We begin by training our model on the perspective views cropped from panoramic images, using the data created by Cai et al [8]. Specifically, we use the image pairs cropped from panoramic images included in the StreetLearn [28] dataset, depicting various streets in Manhattan. As further detailed in Cai et al [8], this training set includes roughly 1M image pairs sampled from the same panorama, split according to the overlap level.

Training with Data Augmentations. We observe that the dataset constructed by prior work could be augmented to better capture the distribution of image pairs captured in-the-wild. In particular, we focus on two types of data augmentations, which we elaborate on next: (i) field of view (FoV) augmentations (denoted henceforth as ΔFoV) and (ii) image-level appearance augmentations (denoted henceforth as ΔIm). In Section 5, we demonstrate how both types

Figure 4. **Augmenting perspective images cropped from panoramic views with image-level appearance modifications.** Given an input image (left) and a target text prompt *"Make it* $\langle w \rangle$*"* ($\langle w \rangle$ is specified above), we use a conditional Diffusion model [7] to create semantic appearance augmentations which modify both the global image characteristics as well as local image regions.

of augmentations improve the model's ability in generalizing to real Internet scenes.

Rather than cropping perspective images with a fixed 90° FoV as was done in prior work, we analyze the FoV values of the images belonging to the *ELP* training set. We compute the mean and standard-deviation values, denoted as μ and σ, respectively. We then construct new perspective images by sampling from a Gaussian distribution $G(\mu, a \cdot \sigma)$ that resembles the distribution of real data, setting $a = 1.5$ for obtaining a more diverse set, which also bears higher similarity to the perspective images used during initialization. We also allow for FoV differences (of up to 5°) between the two images paired together. Additionally, rather than providing the model with the full content within these regions, we construct crops with various aspect ratios to fur-

ther emulate real image pairs.

In-the-wild images differ not only in their intrinsic parameters, but also in their appearance – due to presence of varying illumination and dynamic objects. Therefore, in addition to augmenting the dataset with images of various field of views, we also perform image-level appearance augmentations, leveraging recent advancements in text-to-image Diffusion models [7, 22, 31]. In particular, we apply the conditional InstructPix2Pix [7] model on a subset of our data, using a set of text prompts of the format *"Make it* $\langle w \rangle$*"*, where $\langle w \rangle$ captures diverse appearance modifications. See Figure 4 for a few illustrative examples; note how some of these modify the global illumination (*e.g.*, altering the time of day or the season) while others yield local edits (such as adding more people to make it *"a busy street"*). We performed additional filtering on the augmented set, to avoid significant edits that also altered the structure of the scene; see the supplementary material for additional details.

Training on Real Data. Finally, we finetune the model over real image pairs from our proposed *ELP* dataset. To prioritize learning over extreme image pairs, we first pass a batch of non-overlapping image pairs, followed by a batch of overlapping image pairs.

5. Results and Evaluation

To validate our method, we compare our model with prior relative rotation estimation methods on the *s*ELP and *w*ELP test sets. In particular, our experiments seek to answer the following questions:

- How well do previous methods perform on our proposed task of extreme rotation estimation in the wild?
- How important is our progressive training scheme, and to what extent can the improved performance be attributed to training on the *ExtremeLandmarkPairs* dataset?
- What is the impact of our design choices?

We also present qualitative results over different overlap levels in Figure 5. These results illustrate our model's performance, and its robustness to varying illumination conditions and transient objects (such as the screen and the beer cup in the rightmost examples). Please refer to the supplementary material for many more qualitative results, including visualizations for baseline methods.

5.1. Baselines

We compare our method to correspondence-based techniques, dedicated (end-to-end) relative rotation estimation techniques as well as recent relative pose estimation techniques. Specifically, we consider two correspondence-based methods: SIFT [27] and LoFTR [38]. These methods extract keypoint matches to compute an essential matrix using the RANSAC algorithm and calculate the rotation that stems from the essential matrix. We also compare against two prior works specifically targeting extreme rotation estimation using end-to-end deep learning frameworks: DenseCorrVol [8] and CascadedAtt [12]. These methods are trained and evaluated on images sampled from panoramic views from StreetLearn [28]. Results for CascadedAtt are reported over a reproduced model, as a pre-trained model is not provided; additional details are provided in the supplementary material.

We also consider 8PointViT [30], which estimates the relative pose between two images with a Vision Transformer that is modified to be close to the eight point algorithm. Following prior work [12], we only report results over overlapping settings for this baseline. Furthermore, as it assumes a single camera setting, we evaluate it only on the *s*ELP test set. Finally, we consider Dust3R [46], a recent technique for dense and unconstrained stereo 3D reconstruction. While this work does not specifically target the setting of extreme rotation estimation, we examine to what extent they can be adapted for this task in the wild. Note that Dust3R was trained on MegaDepth. Since our *w*ELP test set is constructed from image pairs from this dataset, it is not entirely fair comparison; therefore, Dust3R's results on the *w*ELP test set are highlighted in gray.

5.2. Evaluation Metrics

For each image pair, we compute the geodesic error, defined as follows:

$$\text{err} = \arccos\left(\frac{tr(\mathbf{R}^T\mathbf{R}^*) - 1}{2}\right) \tag{3}$$

where $\mathbf{R}$ is the predicted rotation matrix, and $\mathbf{R}^*$ is the ground truth relative rotation matrix. We report the median geodesic error (MGE) and relative rotation accuracy (RRA) for each test set and overlap ratio. RRA is reported over a predefined threshold τ, indicating the percentage of image pairs with relative rotation error below τ. We report this metric for $\tau = 15^\circ$ (RRA_{15}) and $\tau = 30^\circ$ (RRA_{30}).

Additionally, as we discretize the space of rotations and estimate three rotation angles, each predicted as a probability distribution over N bins, we also report the performance of the *Top 5* predictions in our ablation study. The *Top 5* prediction reports the lowest error, considering the top-5 peaks in the relative yaw prediction only, as we observe that errors are mostly a function of the relative yaw angles.

5.3. Quantitative Comparison

The main quantitative results comparing our method to other methods on rotation estimation in the wild are reported in Table 2. In the supplementary material, we conduct an additional evaluation in the more constrained setting examined by prior work targeting extreme rotation estimation (*i.e.*, images cropped from panoramas).

As illustrated in Table 2, SIFT and LoFTR, which are correspondence-based methods, exhibit some robustness

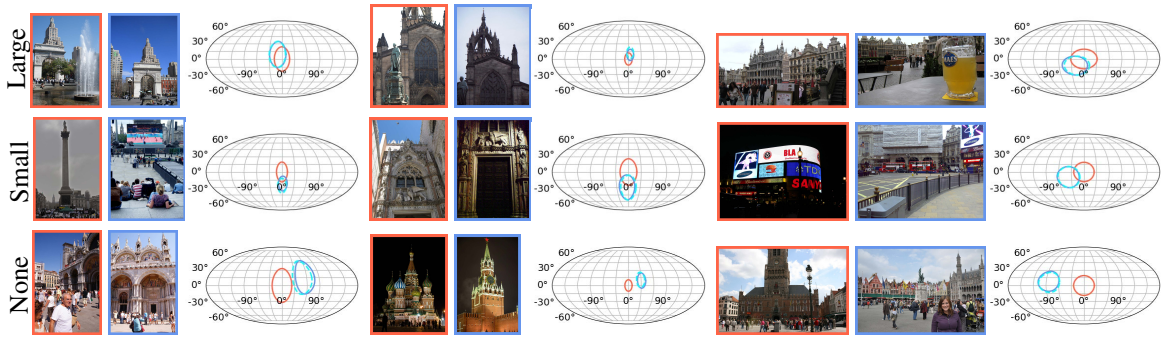


Figure 5. **Qualitative results on the *w*ELP test set**. We visualize the results of our model over different overlap levels, where the images on the left serve as the reference points, and their coordinate system determines the relative rotation, which defines the images on the right. The ellipsoids representing the ground truth are color-coded to match their respective images, with the estimated relative rotation illustrated by a cyan dashed line. As illustrated by the examples above, our method can accurately predict relative rotations for diverse image pairs containing varying appearances and intrinsic parameters. Please refer to the supplementary material for additional qualitative results.

	Method	*s*ELP MGE↓	*s*ELP RRA_{15}↑	*s*ELP RRA_{30}↑	*w*ELP MGE↓	*w*ELP RRA_{15}↑	*w*ELP RRA_{30}↑
Large	SIFT* [27]	1.95	92.3	95.3	2.94	74.6	80.8
	LoFTR* [38]	1.76	97.2	99.1	**2.13**	85.2	93.8
	DenseCorrVol [8]	98.51	25.9	33.3	120.53	7.0	13.0
	CascadedAtt [12]	29.75	42.7	50.0	170.62	7.3	9.2
	8PointViT [30]	22.33	31.9	64.8	–	–	–
	Dust3R [46]	**0.77**	**99.7**	**99.9**	1.01	98.4	99.2
	Ours	2.45	96.7	96.8	2.41	**97.5**	**97.9**
Small	SIFT* [27]	5.07	64.7	71.3	7.27	61.4	68.4
	LoFTR* [38]	2.70	81.5	93.6	6.80	66.6	81.2
	DenseCorrVol [8]	143.47	2.8	9.4	125.73	3.1	9.4
	CascadedAtt [12]	148.44	0.0	3.0	139.14	2.7	4.4
	8PointViT [30]	51.30	1.7	12.8	–	–	–
	Dust3R [46]	**1.96**	**95.9**	**94.6**	2.80	89.8	94.4
	Ours	4.35	88.3	89.0	**4.47**	**87.2**	**91.6**
None	SIFT* [27]	121.94	2.7	5.4	122.84	0.0	2.0
	LoFTR* [38]	-	-	-	56.54	0.0	33.0
	DenseCorrVol [8]	77.10	9.0	27.0	82.04	2.9	13.7
	CascadedAtt [12]	69.69	8.4	23.1	78.60	7.5	20.8
	Dust3R [46]	114.33	19.8	23.9	81.21	15.4	26.9
	Ours	**13.62**	**52.7**	**59.7**	**26.97**	**36.1**	**50.7**

Table 2. **Rotation Estimation in the Wild.** We evaluate performance over the *s*ELP and *w*ELP test sets, separately considering Large (top), Small (middle) and Non-overlapping (bottom) pairs. * indicates median errors are computed only over successful image pairs, for which these algorithms output a pose estimate (failure over more than 50% of the test pairs is shown in gray). Note that Dust3R was trained on images from *w*ELP.

when handling highly overlapping Internet images, compared to methods that trained on images cropped from panoramas, achieving a median error of less than 3° for Large overlap pairs in both test sets. However, these methods rely on image overlap and may not always provide an output of estimated camera pose, as the geometric verification requires a sufficient number of detected inliers. Therefore, they struggle to produce reliable matches in cases of limited overlapping regions, as also observed in prior work focusing on extreme scenarios [8, 12].

DenseCorrVol, CascadedAtt and 8PointViT exhibit relatively poor performance on the *ELP* test sets, illustrating that models trained on images sampled from panoramas cannot easily generalize to Internet photos. In the supplementary material, we show that our model significantly outperforms these prior work even when trained on the same dataset (*i.e.*, images with a constant FoV and illumination cropped from panoramas). As further discussed there, we believe this gap can be partially attributed to the usage of a pretrained LoFTR feature extractor, which is capable of encoding knowledge between Internet image pairs (which vary in their intrinsics and appearance).

Dust3R operates on Internet datasets without the need for calibration and performs exceptionally well on overlapping pairs, achieving the lowest median error and highest success rates for Large and Small overlap categories for *s*ELP test set. However, this method is designed for overlapping pairs as it initializes the model using pretrained CroCo [47], which is trained to address cross-view completion problem from two overlapping views. Furthermore, its output consists of a unified dense point cloud for each pair of images. Due to its design for overlapping pairs, its performance on non-overlapping views is extremely low, also on the *w*ELP test set, which contains images from scenes included in its training set. Moreover, in terms of size, Dust3R (with a DPT head) contains 577 million parameters. In contrast, our model is significantly more compact, with only 80 million parameters. As shown in the table, despite its smaller size, our model yields significantly improved performance over non-overlapping pairs on both test sets.

5.4. Ablation study

We conduct various ablations analyzing the effect of our progressive training scheme and other design choices,also reporting performance over the *Top 5* predictions (consid-

Overlap	Train data	Top 1			Top 5		
		MGE ↓	RRA_{15} ↑	RRA_{30}↑	MGE ↓	RRA_{15}↑	RRA_{30}↑
Large	[8]	13.65	35.4	73.5	12.22	61.4	84.7
	+ΔFoV	4.61	79.7	81.1	4.41	90.3	98.9
	+ΔIm	4.46	90.4	92.4	4.43	94.3	99.1
	+*ELP*	**2.41**	**97.5**	**97.9**	**2.41**	**98.4**	**99.4**
Small	[8]	55.28	3.7	29.1	29.83	15.0	50.3
	+ΔFoV	12.91	56.2	68.2	10.97	66.0	85.4
	+ΔIm	11.46	62.5	80.6	10.73	68.0	91.0
	+*ELP*	**4.47**	**87.2**	**91.6**	**4.24**	**91.1**	**97.2**
None	[8]	74.94	12.8	25.3	25.11	26.1	58.8
	+ΔFoV	61.62	25.0	38.4	16.82	44.2	75.0
	+ΔIm	68.31	25.0	36.1	16.21	45.7	78.2
	+*ELP*	**26.97**	**36.1**	**50.7**	**12.85**	**57.1**	**85.8**

Table 3. Ablation study, evaluating the effect of our progressive training scheme over the *w*ELP test set. All experiments start with the cropped panoramas used in Cai *et al.* [8].

Overlap	KP	SM	Top 1			Top 5		
			MGE ↓	RRA_{15}↑	RRA_{30}↑	MGE ↓	RRA_{15}↑	RRA_{30}↑
Large	×	×	**2.18**	97.4	98.1	**2.18**	97.4	98.1
	×	✓	2.30	97.0	97.4	2.30	**98.5**	99.4
	✓	×	2.44	**97.6**	**98.3**	2.31	98.4	**99.7**
	✓	✓	2.41	97.5	97.9	2.41	98.4	99.4
Small	×	×	4.50	87.9	91.6	4.50	87.9	91.7
	×	✓	4.49	**88.1**	92.0	4.46	91.2	96.7
	✓	×	**4.41**	87.5	**92.2**	4.32	**91.9**	**97.6**
	✓	✓	4.47	87.2	91.6	**4.24**	91.1	97.2
None	×	×	48.81	34.0	44.1	**12.56**	**57.5**	84.6
	×	✓	43.07	31.2	44.2	13.99	53.5	83.2
	✓	×	41.39	35.3	46.8	13.04	56.9	**86.2**
	✓	✓	**26.97**	**36.1**	**50.7**	12.85	57.1	85.8

Table 4. Ablation study, evaluating the effect of the auxiliary channels added as input to our network. We train models without the keypoints and matches (KP) and without the segmentation maps (SM), comparing to our model over the *w*ELP test set , after using a validation split that is balanced over the overall angle.

ering the top-5 peaks in the relative yaw prediction, as detailed in Section 5.2). Additional ablations, including an analysis of architectural components, are provided in the supplementary material.

The effect of our progressive training scheme. We conduct multiple experiments ablating the effect of our progressive training scheme. Our training process consists of four stages: initialization (following prior work [8]), incorporating multiple FoVs (+ΔFoV), training with image-level appearance augmentations (+ΔIm), and training with *ExtremeLandmarkPairs* pairs (+*ELP*).

Table 3 illustrates the impact of each training stage on the results. As can be observed from the table, each training stage further refines the model's performance, often in a significant manner. For instance, the median error in the *w*ELP small overlap test set decreased from 55.3° to 12.9° when FoV augmentations were added. Additionally, the *ExtremeLandmarkPairs* training set plays a crucial rule in finalizing our training process, yielding a significant reduction in the median error (more than half) for small and non-overlapping scenarios in the *w*ELP test set. While this median error remains relatively high for non-overlapping pairs, we observe that the top-5 scores show significant improvements, *e.g.*, reducing the median error from 27.0° to 12.8°. This demonstrates that the model has learned this knowledge, although it cannot necessarily be recovered by the largest peak alone. In the supplementary material, we demonstrate that the intermediate training stages are indeed important and that the improved performance cannot be obtained with the *ExtremeLandmarkPairs* training set alone.

Moreover, evidenced in Table 3, our model demonstrates strong generalization capabilities even when trained exclusively on panorama-cropped images. The model generalizes to the *w*ELP test set significantly better (e.g., for Large overlap cases, our model achieves an MGE of 13.65° compared to values higher than 120° achieved by baselines), demonstrating that the improvement is not just due to our progressive training scheme. We further examined architectural differences by conducting an additional ablation study, applying our progressive training scheme to the baseline models (see Table 9 in the supplementary materials). This experiment revealed that prior models are not directly suitable for real-world applications, as these baseline models showed significantly poorer performance across all metrics.

The effect of adding auxiliary channels. We ablate the effect of adding auxiliary in Table 4, training models without keypoints and matches (KP) or the segmentation map (SM) provided as additional inputs. As illustrated in the table, for non-overlapping cases, these auxiliary channels boost performance almost across all metrics. In particular, both channels play a role in reducing the errors (reducing the median error from over 40° to 27.0°).

6. Conclusion

We present a method and benchmark dataset for estimating relative 3D rotations between pairs of (possibly) non-overlapping RGB images. Our approach extends prior work addressing extreme rotations to real-world data that exhibit variation in both appearance in intrinsic camera parameters. While our model shows promising results on real-world Internet image pairs, it also highlights the inherent difficulty of the underlying task, suggesting that considerable progress can be achieved with future techniques that leverage our dataset. Our paired data could also potentially serve for exploring the challenging task of estimating extreme translations in real-world settings. Future work could also consider incorporating more views for enhancing performance in such extreme non-overlapping scenarios.

Acknowledgments This work was partially supported by ISF (grant number 2510/23).

References

[1] Samir Agarwala, Linyi Jin, Chris Rockwell, and David F Fouhey. Planeformers: From sparse view planes to 3d reconstruction. In *European Conference on Computer Vision*, pages 192–209. Springer, 2022. 2

[2] Roberto Cipolla Alex Kendall, Matthew Grimes. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV -International Conference on Computer Vision*, pages 2938–2946, 2015. 2, 4

[3] Lionel Baboud, Martin Čadík, Elmar Eisemann, and Hans-Peter Seidel. Automatic photo-to-terrain alignment for the annotation of mountain pictures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 41–48. IEEE, 2011. 2

[4] Vassileios Balntas, Shuda Li, and Victor Prisacariu. Relocnet: Continuous metric learning relocalisation using neural nets. In *European Conference on Computer Vision*, pages 751–767. Springer, 2018. 2

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *ECCV*, pages 404–417. Springer, 2006. 2

[6] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6684–6692, 2017. 1

[7] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR IEEE / CVF Computer Vision and Pattern*, pages 18392–18402, 2023. 2, 5, 6

[8] Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme rotation estimation using dense correlation volumes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14566–14575, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[9] Ruojin Cai, Joseph Tung, Qianqian Wang, Hadar Averbuch-Elor, Bharath Hariharan, and Noah Snavely. Doppelgangers: Learning to disambiguate images of similar structures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 34–44, 2023. 4

[10] Yaron Caspi and Michal Irani. Aligning non-overlapping sequences. *IJCV*, 48(1):39–51, 2002. 2

[11] Kefan Chen, Noah Snavely, and Ameesh Makadia. Wide-baseline relative camera pose estimation with directional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3258–3268, 2021. 2

[12] Shay Dekel, Yosi Keller, and Martin Cadik. Estimating extreme 3d image rotations using cascaded attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2588–2598, 2024. 1, 2, 3, 4, 6, 7

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[14] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2

[15] Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, and Ping Luo. Camnet: Coarse-to-fine retrieval for camera relocalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2871–2880, 2019.

[16] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In *CVPR*, 2019. 2

[17] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023. 2

[18] Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Dedode: Detect, don't describe–describe, don't detect for local feature matching. *arXiv preprint arXiv:2308.08479*, 2023. 2

[19] Xie Enze, Wang Wenhai, Yu Zhiding, Anandkumar Anima, M. Alvarez Jose, and Luo Ping. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Conference on Neural Information Processing Systems (NEURIPS)*, 2021. 4

[20] Zhiwen Fan, Panwang Pan, Peihao Wang, Yifan Jiang, Dejia Xu, Hanwen Jiang, and Zhangyang Wang. Pope: 6-dof promptable pose estimation of any object, in any scene, with one reference. *arXiv preprint arXiv:2305.15727*, 2023. 2

[21] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM (CACM)*, 24(6):381–395, 1981. 2

[22] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 6

[23] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r, 2024. 2

[24] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 2, 4

[25] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *arXiv preprint arXiv:2305.04926*, 2023. 2

[26] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching atokens are frozen. *arXiv preprint arXiv:2306.13643*, 2023. 2

[27] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2, 6, 7

[28] Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray

Kavukcuoglu, Andrew Zisserman, et al. The streetlearn environment and dataset. *arXiv preprint arXiv:1903.01292*, 2019. 3, 5, 6

[29] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion*. *Acta Numerica*, 26:305–364, 2017. 1

[30] Chris Rockwell, Justin Johnson, and David F Fouhey. The 8-point algorithm as an inductive bias for relative pose prediction by vits. In *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2022. 2, 6, 7

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 6

[32] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2

[33] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2

[34] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, et al. Back to the feature: Learning robust camera localization from pixels to pose. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3247–3257, 2021. 1

[35] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8601–8610, 2018. 1

[36] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1, 2

[37] Samarth Sinha, Jason Y Zhang, Andrea Tagliasacchi, Igor Gilitschenski, and David B Lindell. Sparsepose: Sparse-view camera pose regression and refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21349–21359, 2023. 2

[38] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 2, 4, 6, 7

[39] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. *arXiv preprint arXiv:2406.11819*, 2024. 2, 4

[40] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 2

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[42] Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z Jane Wang, and Rabab Ward. Multi-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5722–5731, 2021. 1

[43] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9773–9783, 2023. 2

[44] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023. 2

[45] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 757–774. Springer, 2020. 2

[46] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. *arXiv preprint arXiv:2312.14132*, 2023. 2, 6, 7

[47] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022. 7

[48] Zhenpei Yang, Jeffrey Z Pan, Linjie Luo, Xiaowei Zhou, Kristen Grauman, and Qixing Huang. Extreme relative pose estimation for rgb-d scans via scene completion. In *CVPR*, 2019. 2

[49] Zhenpei Yang, Siming Yan, and Qixing Huang. Extreme relative pose network under hybrid representations. In *CVPR*, 2020. 2

[50] Zhenpei Yang, Zhile Ren, Miguel Angel Bautista, Zaiwei Zhang, Qi Shan, and Qixing Huang. Fvor: Robust joint shape and pose optimization for few-view object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2497–2507, 2022. 2

[51] Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *European Conference on Computer Vision*, pages 592–611. Springer, 2022.

[52] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. *arXiv preprint arXiv:2402.14817*, 2024. 2