Leveraging Equivariant Features for Absolute Pose Regression

Mohamed Adel Musallam
mohamed.ali@uni.lu

Vincent Gaudillière
vincent.gaudilliere@uni.lu

Miguel Ortiz del Castillo
miguel.ortizdelcastillo@uni.lu

Kassem Al Ismaeil
kassem.alismaeil@gmail.com

Djamila Aouada
djamila.aouada@uni.lu

Interdisciplinary Center for Security, Reliability and Trust (SnT)
University of Luxembourg, Luxembourg

Abstract

While end-to-end approaches have achieved state-of-the-art performance in many perception tasks, they are not yet able to compete with 3D geometry-based methods in pose estimation. Moreover, absolute pose regression has been shown to be more related to image retrieval. As a result, we hypothesize that the statistical features learned by classical Convolutional Neural Networks do not carry enough geometric information to reliably solve this inherently geometric task. In this paper, we demonstrate how a translation and rotation equivariant Convolutional Neural Network directly induces representations of camera motions into the feature space. We then show that this geometric property allows for implicitly augmenting the training data under a whole group of image plane-preserving transformations. Therefore, we argue that directly learning equivariant features is preferable than learning data-intensive intermediate representations. Comprehensive experimental validation demonstrates that our lightweight model outperforms existing ones on standard datasets.[1]

1. Introduction

In computer vision, camera pose estimation, and its reference frame inverse, *i.e.*, object pose estimation, have been extensively studied over the last decades [38, 42, 54].

Traditionally, pose estimation has been addressed using 3D geometry. In practice, a set of 2D-3D feature correspondences is generated, then statistically leveraged to recover the camera pose [18, 39, 49, 64]. More recently, direct Absolute Pose Regression (APR) approaches have been introduced, drawing upon early successes of deep learning [1].

[1]This work was funded by the Luxembourg National Research Fund (FNR), under the project reference BRIDGES2020/IS/14755859/MEET-A/Aouada, and by LMO (https://www.lmo.space).

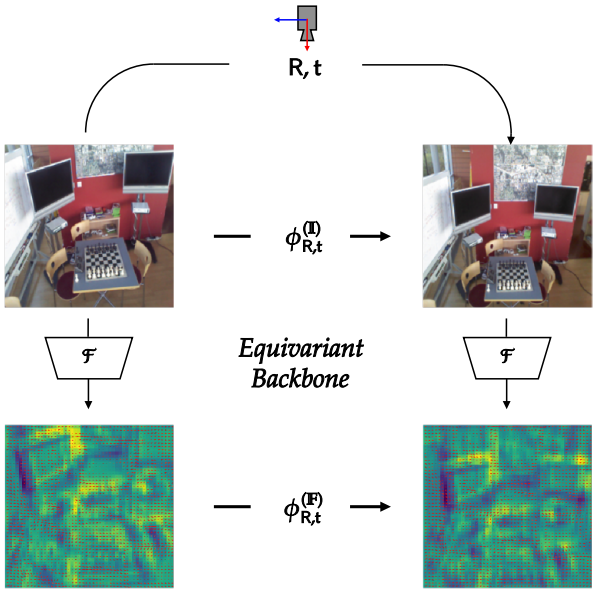

Figure 1. **Illustration of our approach** - Our method adopts a translation and rotation-equivariant convolutional neural network to extract geometry-aware features that directly encode camera planar motions $\mathsf{R}, \mathbf{t}$. While camera moves, equivariance of the proposed feature extractor $\mathcal{F}$ leads to explicit image ($\phi^{(\mathbb{I})}_{\mathsf{R},\mathbf{t}}$) and feature ($\phi^{(\mathbb{F})}_{\mathsf{R},\mathbf{t}}$) changes. This property is leveraged to propose a more efficient solution to the absolute pose regression problem.

These methods consist in directly mapping an image to its pose using a suitably trained Convolutional Neural Network (CNN). Therefore, end-to-end trainable methods have the advantage of providing fully differentiable results, enabling the optimization of all parameters in a comprehensive manner. Moreover, predictions are achieved at a steady speed and power consumption, whereas RANdom SAmple Consensus (RANSAC)-based methods [18] are less predictable

and likely to suffer from an efficiency drop when the inlier rate is low. However, state-of-the-art APR methods have been proven theoretically and shown experimentally to have a lower accuracy compared to 3D structure-based approaches [50]. Indeed, the former are more closely related to image retrieval than to 3D structure [50].

The questions we ask in this work are: *Why do current APR methods fall short in accuracy ? How can they reach their full potential?* Our hypothesis is that there is a lack of exploitation of the geometric properties of data. This happens typically at the level of the feature extraction layers commonly used in classical deep learning approaches. Specifically, we posit that in the case of APR and pose estimation, having a representation which is *equivariant* to the group of rigid motions, *i.e.,* rotations and translations in 3D, may be an effective way to boost the network performance. This should play the role of an implicit data augmentation by means of group equivariance, and in turn alleviate the need for an explicit data augmentation for training.
Indeed, recently, there has been a growing interest in designing more geometric models that are equivariant to groups of such transformations. These approaches leverage theoretical contributions from group theory, representation theory, harmonic analysis and fundamental deep learning [8–10, 12, 21, 22, 61]. More specifically, group-equivariant neural networks, or Group-equivariant CNNs (G-CNNs), are part of the broader and promising field of geometric deep learning [4], that aims to exploit any underlying geometric relationship that can exist within the data. In particular, the special Euclidean groups in 2 and 3 dimensions, denoted as $\mathrm{SE}(2)$ and $\mathrm{SE}(3)$ and encompassing respective rigid motions, are of particular interest in 3D computer vision [13, 61].

Despite the conceptual advances they represent, to the best of our knowledge, the use of deep equivariant features in the APR context is still considerably unexplored. This paper proposes, for the first time, to investigate and justify the use of deep equivariant features for solving APR (see Figure 1).

Contributions. Our contributions are summarized below:
(1) A formulation of how an equivariant CNN induces representations of planar camera motions, lying in $\mathrm{SE}(2)$, directly into the feature space. (Section 4.1)
(2) An intuitive explanation is provided as to how $\mathrm{SE}(2)$-equivariant features can be leveraged to recover any camera motion lying in $\mathrm{SE}(3)$. (Section 4.2)
(3) A lightweight equivariant pose regression model, referred to as *E-PoseNet*, is introduced. (Section 5)
(4) Extensive experimental evaluation of E-PoseNet showing its competitive performance as compared to existing APR methods on standard datasets. (Section 6)

Paper organization. An overview of state-of-the-art APR methods and current exploitations of deep equivariant features is given in Section 2. Section 3 presents the formal definition of equivariance along with the formulation of APR. The theoretical justification as to how $\mathrm{SE}(2)$-equivariant features can explicitly encode planar camera motions is presented in Section 4, whereas the full pose regression pipeline is introduced in Section 5. An extensive experimental validation is given in Section 6 along with a discussion of limitations. Section 7 concludes the paper.

2. Related Work

The goal of this paper is to exploit the power of equivariant features in the context of APR. Therefore, we split this section into: (1) a review of the relevant literature on APR, and (2) an overview of recent deep equivariant feature extraction methods applied to computer vision problems.

Absolute Pose Regression. Since the rise of deep learning and CNNs in the early 2010's, many works have explored the application of CNNs for APR. This began with the introduction of PoseNet by Kendal *et al.* [35], who used the GoogLeNet model [55] as a feature extraction backbone coupled with a regression head to estimate the translation and rotation vectors. Most of the subsequent improvements lie in changes in the feature extraction architecture [3, 35, 62], modified objective functions [32, 43, 57], and additional intermediate representations [25, 27].

In [50], Sattler *et al.* provide an in-depth analysis of existing works on APR [6, 34, 44, 62, 63]. In particular, they show that structure-based and image retrieval methods are more accurate than APR. Moreover, they demonstrate that APR algorithms do not explicitly leverage knowledge about projective geometry. Instead, they learn a mapping between image content and camera poses directly from the data, and in the form of a set of base poses such that all training samples can be expressed as a linear combination of these reference entities. Wang *et al.* [60] proposed an approach to integrate dense correspondence-based intermediate geometric representations within an end-to-end trainable pipeline. However, this method still relies on classical (non-equivariant) features, and thereby requires a significant amount of data for generalization. Furthermore, methods such as [25, 27, 60] propose to learn intermediate representations that are indirectly equivariant, such as segmentation masks, object detections, and depth or normal maps. However, this comes at the cost of parameter redundancy. This core observation suggests that directly learning equivariant features may be a valuable direction to improve the accuracy of pose estimation while reducing the number of model parameters.

Deep Equivariant Features. There is a rich history in computer vision on the design of hand-crafted equivariant features (*e.g.,* Scale-Invariant Feature Transform (SIFT) [41], Oriented filters [65], Steerable filters [19],

Rotation-equivariant Fields of Experts (R-FoE) [51], Lie groups-based filters [17,46]). In the deep learning literature, convolutional layers [37] have been proven to be equivariant to image shifting, while max-pooling layers are only invariant to small shifts of the input image [20].
Although convolutional layers are inherently equivariant to translation, there is a significant amount of spatial information regarding the inputs that is not encoded by CNNs in a precise fashion [28, 31]. More specifically, local and global poolings, if added to CNNs, render translation information unrecoverable, discarding the foregoing equivariance [40].
A recent investigation shows that many neurons in CNNs learn slightly transformed (*e.g.,* rotated) versions of the same basic feature [47]. These are especially common in early vision, *e.g.,* in curve detectors, high-low frequency detectors, and line detectors.

There have been attempts to extend the G-CNNs to wider groups of transformations. In [5, 48], Mallat *et al.* extended CNNs to be equivariant to $\mathrm{SE}(2)$ using scattering transform with predefined wavelets. In [2, 29], Bekkers *et al.* also extended CNNs to be equivariant to the $\mathrm{SE}(2)$ group via B-splines. In [9], Cohen *et al.* proposed group convolutions network equivariant to the p4m discrete group via $90°$ rotations and flips, where they demonstrated the effectiveness of group convolutions for classification task.
More recently, the use of equivariant features has been investigated for solving various computer vision tasks such as 3D point cloud analysis [8], aerial object detection [22] and 2D tracking [21]. In [14], Esteves *et al.* proposed to use projection and embedding from 2D images into a spherical CNN latent space to estimate the relative orientations of the object. Similarly, Zhang *et al.* proposed to use spherical CNN for learning camera pose estimation in omnidirectional localization [66]. However, to the best of our knowledge, equivariant features have not yet been explicitly leveraged in the context of APR for single 2D input image, which is the very focus of this paper.

3. Preliminaries

This section provides the necessary mathematical background. First, we introduce the notions of invariant and equivariant features. Then, we present the general framework for APR, and finally show the added value of relying on equivariant features in this context.

Notation. The following notation will be adopted: vectors and column images are denoted by boldface lowercase letters $\mathbf{x}$, matrices by uppercase letters X, scalars by italic letters x or X, functions as $\mathcal{X}$ and spaces as $\mathbb{X}$. The special orthogonal group, the Euclidean group, and the special Euclidean group, of dimension n, are denoted as $\mathrm{SO}(n)$, $\mathrm{E}(n)$ and $\mathrm{SE}(n)$, respectively.

Invariant and Equivariant Features.

Given an image $\mathbf{x}$, captured by a camera, an APR method $\mathcal{P}$ predicts the 6-Degrees-of-Freedom (6-DoF) pose, *i.e.,* position and orientation, of the camera with respect to its environment.
Let us denote by $\mathbb{I} \subset \mathbb{R}^m$ the linear space of vectorized m-dimensional images (or image regions), and by $\mathbb{F} \subset \mathbb{R}^n$ the latent space of features – with dimension n. Considering a CNN-based feature extraction function $\mathcal{F}$, we write:

$$\begin{aligned} \mathcal{F}: \ & \mathbb{I} \to \mathbb{F} \\ & \mathbf{x} \mapsto \mathcal{F}(\mathbf{x}). \end{aligned}$$

Given $\mathfrak{G}$, a generic group of transformations and $\mathfrak{g}$, an element of $\mathfrak{G}$, we denote by $\phi_{\mathfrak{g}}^{(\mathbb{I})}$ and $\phi_{\mathfrak{g}}^{(\mathbb{F})}$ the actions of $\mathfrak{g}$ into the image and feature spaces, respectively.

Definition 1 *$\mathcal{F}$ is invariant to $\mathfrak{G}$ if and only if*

$$\forall \mathfrak{g} \in \mathfrak{G}, \forall \mathbf{x} \in \mathbb{I}, \quad \mathcal{F}(\phi_{\mathfrak{g}}^{(\mathbb{I})}\mathbf{x}) = \mathcal{F}(\mathbf{x}). \tag{1}$$

Definition 2 *$\mathcal{F}$ is equivariant to $\mathfrak{G}$ if and only if*

$$\forall \mathfrak{g} \in \mathfrak{G}, \forall \mathbf{x} \in \mathbb{I}, \quad \mathcal{F}(\phi_{\mathfrak{g}}^{(\mathbb{I})}\mathbf{x}) = \phi_{\mathfrak{g}}^{(\mathbb{F})}\mathcal{F}(\mathbf{x}). \tag{2}$$

Note that invariance can be seen as a special case of equivariance where $\phi_{\mathfrak{g}}^{(\mathbb{F})} = \mathcal{I}$, the identity mapping, $\forall \mathfrak{g} \in \mathfrak{G}$.

Equivariant Features for APR. Sattler *et al.* proposed the following formulation for the pose function $\mathcal{P}$ [50]:

$$\mathcal{P}(\mathbf{x}) = \mathbf{b} + \mathsf{P}.\mathcal{E}(\mathcal{F}(\mathbf{x})), \tag{3}$$

where the feature extractor $\mathcal{F}$ is first applied to the image $\mathbf{x}$ followed by a non-linear embedding of the features $\mathcal{E}$ lifting them to a higher-dimensional space. Then, a linear projection into the space of camera poses, represented by a matrix P, is applied. Finally, a bias term $\mathbf{b}$ is added.
As presented in Section 2, the work in [50] demonstrated that classical APR is more closely related to pose approximation via image retrieval than to accurate pose estimation leveraging the 3D structure, thus the accuracy gap.

Our hypothesis is that this is likely due the lack of geometric information carried by classical CNN features. Indeed, the perceptive power of classical CNNs can most often be considered as a statistical phenomenon, whereas pose estimation is a geometrical problem.

In this work, we thus propose to replace the classical convolutional layers of the feature extractor $\mathcal{F}$ by their group-equivariant counterparts [9], then to assess how this affects both the accuracy and data efficiency of the model. Therefore, assuming that $\mathcal{F}$ is equivariant to $\mathfrak{G}$, *i.e.*, verifies **Definition 2**, and applying the transformation $\phi_{\mathfrak{g}}^{(\mathbb{I})}$ to the image $\mathbf{x}$, the pose regression function $\mathcal{P}$ in (3) becomes:

$$\mathcal{P}(\phi_{\mathfrak{g}}^{(\mathbb{I})}\mathbf{x}) = \mathbf{b} + \mathsf{P} \cdot \mathcal{E}\left(\phi_{\mathfrak{g}}^{(\mathbb{F})}\mathbf{v}\right), \tag{4}$$

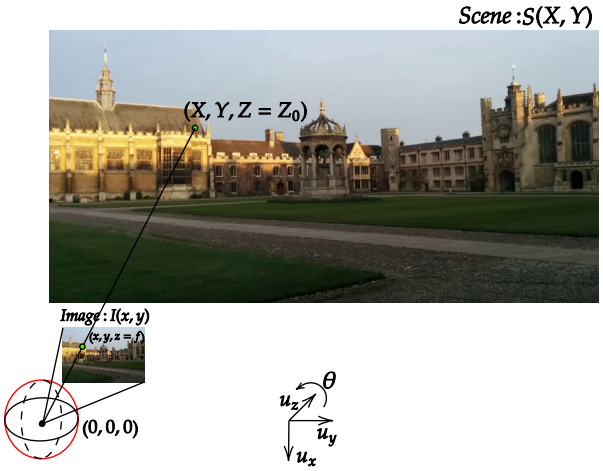

Figure 2. **Illustration** – Scene plane ($Z = Z_0$) - Parallel image plane ($Z = f$) - Camera center $(0, 0, 0)$. The scene is defined as the set of light intensity values $\mathsf{S}(X, Y)$ within the plane $Z = Z_0$. Rays of lights are then projected to the camera center. Intersections of projected rays with the image plane (considered as infinite) define image intensity values $\mathsf{I}(x, y)$. Camera motions are restricted to $\mathrm{SE}(2)$, *i.e.* translation along $\mathbf{u}_x$, $\mathbf{u}_y$ and rotation around $\mathbf{u}_z$ (characterized by roll angle θ).

where $\mathbf{v} = \mathcal{F}(\mathbf{x})$. This suggests that any action of $\mathfrak{G}$ on the image has a direct effect in the latent space and that, in particular, camera motion transformations of images, *i.e.* changes in camera pose, explicitly induce actions on the feature vector $\mathbf{v}$, and implicitly on the regressed pose.
Considering $\mathfrak{G}$ as $\mathrm{SE}(2)$ or $\mathrm{SE}(3)$, we posit that such equivariant features will help improve the performance of APR.

4. Pose from SE(2)-Equivariant Features

We consider a piecewise planar scene, where the scene planes are parallel to the image plane. We then consider camera motions that locally preserve the latter.

4.1. SE(2)-Equivariant Features

We herein restrict camera motions to those of the $\mathrm{SE}(2)$ group, *i.e.,* planar translations and rotations within the image plane (Figure 2). With that, we analyse the effects of planar camera motions on the image and feature spaces, assuming that the feature extractor $\mathcal{F}$ is equivariant to $\mathrm{SE}(2)$.

Effects of Camera Planar Motions on Images. Following the notations introduced in Figure 2, rotating the camera with a roll angle θ is equivalent to rotating the scene around $\mathbf{u}_z$ (camera viewing direction) with angle $-\theta$ [30].
Similarly, translating the camera center along $\mathbf{u}_x$ and $\mathbf{u}_y$ is equivalent to translating the scene in the opposite direction. Let us denote any rigid motion of the camera along its image plane (*i.e.,* in $\mathrm{SE}(2)$) by $\mathsf{R}, \mathbf{t}$, where $\mathbf{t}$ is a planar translation

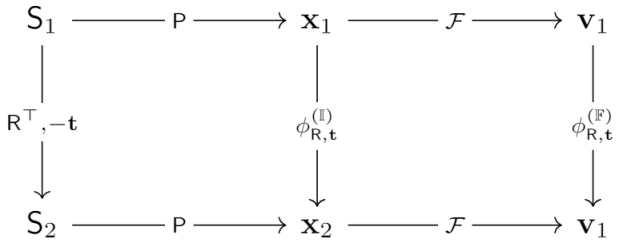

Figure 3. **Equivariance map –** Camera planar motion transformations of planar scenes ($\mathsf{S}_1 \mapsto \mathsf{S}_2$, first column), images ($\mathbf{x}_1 \mapsto \mathbf{x}_2$, second column) and features ($\mathbf{v}_1 \mapsto \mathbf{v}_2$, third column) induce representations of $\mathrm{SE}(2)$. In other words, these transformations commute with scene projector P and feature extractor $\mathcal{F}$.

and R a planar rotation. The effect of this motion on any point $\mathbf{p}$ of the scene is obtained by applying $-\mathbf{t}$ then $\mathsf{R}^\top$ such that $\mathbf{p}' = \mathsf{R}^\top(\mathbf{p} - \mathbf{t})$.
In 3D, considering $\mathbf{t} = (T_X, T_Y, 0)^\top$, we have

$$\mathbf{p}' = \begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) & 0 \\ -\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X - T_X \\ Y - T_Y \\ Z \end{pmatrix}. \tag{5}$$

Following classical projection rules [23], image coordinates (x, y) are then given by $x = f\frac{X}{Z_0}$ and $y = f\frac{Y}{Z_0}$, where f is the distance from the camera center to the image plane, and Z_0 is the distance to the scene plane.
Multiplying (5) by $\frac{f}{Z_0}$, then restricting the coordinates to the first two ones gives:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x - t_x \\ y - t_y \end{pmatrix}, \tag{6}$$

where $t_x = f\frac{T_X}{Z_0}$, and $t_y = f\frac{T_Y}{Z_0}$. By denoting $\mathsf{R}^{(2)}$ the 2D rotation matrix of angle θ and $\mathbf{t}^{(2)} = (t_x, t_y)^\top$, the effect of any planar camera motion $\mathsf{R}, \mathbf{t}$ on any point $\mathbf{p}^{(2)}$ of the image is thus given by $\mathbf{p}^{(2)\prime} = \mathsf{R}^{(2)\top}(\mathbf{p}^{(2)} - \mathbf{t}^{(2)})$, where $\mathbf{p}^{(2)\prime}$ is the image of $\mathbf{p}^{(2)}$ under the transformation.
We finally denote $\phi^{(\mathbb{I})}_{\mathsf{R},\mathbf{t}}$ the effect of the camera motion on an image $\mathbf{x}_1$, resulting in another image $\mathbf{x}_2$, *i.e.,* $\mathbf{x}_2 = \phi^{(\mathbb{I})}_{\mathsf{R},\mathbf{t}}\mathbf{x}_1$.

In what follows, we prove that the image transformation due to planar motions of the camera commute with the projection operator P (Figure 3). Indeed, applying a planar motion $\mathsf{R}_1, \mathbf{t}_1$ followed by a second one $\mathsf{R}_2, \mathbf{t}_2$ to the camera, has the following effect[2]: $\mathbf{p}' = \mathsf{R}_3^\top(\mathbf{p} - \mathbf{t}_3)$, where $\mathsf{R}_3 = \mathsf{R}_2\mathsf{R}_1$ and $\mathbf{t}_3 = \mathbf{t}_1 + \mathsf{R}_1\mathbf{t}_2$.
Similarly, one can easily observe that combining two camera motions has a similar effect on any point $\mathbf{p}^{(2)}$ of the image such that $\mathbf{p}^{(2)\prime} = \mathsf{R}_3^{(2)\top}(\mathbf{p}^{(2)} - \mathbf{t}_3^{(2)})$ where $\mathsf{R}_3^{(2)} =$

[2]Please refer to the supplementary material for further details

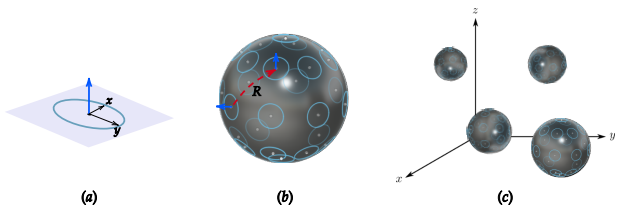

Figure 4. **From planar to 3D motions –** (Figure reproduced from [11]): (a) $\mathbb{S}(1)$, (b) SO(3)/SO(2) $\simeq \mathbb{S}(2)$, (c) SE(3) $= \mathbb{R}^3 \rtimes$ SO(3).

$\mathsf{R}_2^{(2)}\mathsf{R}_1^{(2)}$, and $\mathbf{t}_3^{(2)} = \mathbf{t}_1^{(2)} + \mathsf{R}_1^{(2)}\mathbf{t}_2^{(2)}$. Therefore,

$$\phi^{(\mathbb{I})}_{(\mathsf{R}_2,\mathbf{t}_2)\circ(\mathsf{R}_1,\mathbf{t}_1)} = \phi^{(\mathbb{I})}_{\mathsf{R}_3,\mathbf{t}_3} = \phi^{(\mathbb{I})}_{\mathsf{R}_2,\mathbf{t}_2} \circ \phi^{(\mathbb{I})}_{\mathsf{R}_1,\mathbf{t}_1}. \quad (7)$$

This proves that the correspondence from $\mathsf{R}, \mathbf{t}$ to $\phi^{(\mathbb{I})}_{\mathsf{R},\mathbf{t}}$ is a group homomorphism from SE(2). In other words, the set of $\phi^{(\mathbb{I})}_{\mathsf{R},\mathbf{t}}$, where $\mathsf{R}, \mathbf{t} \in$ SE(2), is the image of a representation of SE(2) into the image space.

Effects of Camera Planar Motions on Features. We herein consider an SE(2)-equivariant CNN-based feature extractor $\mathcal{F}$. For the sake of clarity and simplicity, we discard the discreteness of numerical images and consider their supports as continuous.

Classical convolutional layers are only equivariant to the translation group $(\mathbb{R}^2, +)$. Indeed, at each layer l, a conventional CNN takes as input a stack of intermediate feature maps $\mathbf{v}^{(l)} : \mathbb{R}^2 \to \mathbb{R}^{K^{(l)}}$ and convolves it with a set of $K^{(l+1)}$ filters $\psi^{(l)} : \mathbb{R}^2 \to \mathbb{R}^{K^{(l)}}$. Therefore we have

$$\forall \mathbf{t}^{(2)} \in \mathbb{R}^2, \quad \left(\left(\phi_{\mathbf{t}^{(2)}}\mathbf{v}\right) * \psi^{(l)}\right)(.) = \left(\phi_{\mathbf{t}^{(2)}}\left(\mathbf{v} * \psi^{(l)}\right)\right)(.), \quad (8)$$

where $\phi_{\mathbf{t}^{(2)}}$ are images of $\mathbf{t}^{(2)}$ under representations of $(\mathbb{R}^2, +)$. In other words, if the input image is translated, the output feature map translates in the same way. However, in general, the same is not true for rotations, *i.e.* if the input image is rotated, the output feature map will not be rotated accordingly. The work in [61] has extended CNN equivariance to the SE(2) group, *i.e.* the group of continuous rotations and translations in $\mathbb{R}^2$, the image domain.

By replacing the translation group equivariance of classical CNNs by equivariance to SE(2), such particular CNNs can then be characterized by the following equation:

$$\forall \mathsf{R}^{(2)}, \mathbf{t}^{(2)} \in \mathrm{SE}(2),$$
$$\left(\left(\phi_{\mathsf{R}^{(2)},\mathbf{t}^{(2)}}\mathbf{v}\right) * \psi^{(l)}\right)(.) = \left(\phi_{\mathsf{R}^{(2)},\mathbf{t}^{(2)}}\left(\mathbf{v} * \psi^{(l)}\right)\right)(.), \quad (9)$$

where $\phi_{\mathsf{R}^{(2)},\mathbf{t}^{(2)}}$ are images of $\mathsf{R}^{(2)}, \mathbf{t}^{(2)}$ under representations of SE(2). In particular, considering the last convolutional layer output, we obtain that feature extraction $\mathcal{F}$ and Euclidean transformations of images commute.

Finally, the camera motion transformations of both images and features induce representations of SE(2). As a result, the image and feature spaces explicitly encode the planar motions of the camera.

4.2. From SE(2) to SE(3)

After demonstrating how an SE(2)-equivariant CNN can induce representations of planar camera motions directly into the feature space, we herein discuss how these features, which are equivariant to planar camera motions, *i.e.,* in SE(2), are leveraged for general pose regression in SE(3).

Indeed, SE(2)-equivariant features extracted with $\mathcal{F}$, are now to be mapped to camera poses in SE(3) via $\gamma := \mathsf{P} \cdot \mathcal{E}$. This is the last step towards finding $\mathcal{P}(\mathbf{x})$, as defined in equation (3) [3].

The SE(3) group may be written as a semidirect product SE(3) $= \mathbb{R}^3 \rtimes$ SO(3), and similary for SE(2) $= \mathbb{R}^2 \rtimes$ SO(2). We may therefore restrict the discussion to the mapping $\gamma^* :$ SO(2) $\to$ SO(3) [4, 11]. We rely on the observation that the quotient space SO(3)/SO(2) $\simeq \mathbb{S}(2)$ is the sphere in 3D, where $\simeq$ denotes a homeomorphism.

For every point on $\mathbb{S}(2)$, it is possible to move to another point via a rotation. $\mathbb{S}(2)$ is consequently a homogeneous space for SO(3), and SO(3) can be seen as a bundle of elements of $\mathbb{S}(1)$, *i.e.,* planar circles on $\mathbb{S}(2)$, for which a continuous mapping γ^* exists [11]. These mappings, γ^*, directly relate to γ by composing with translations. Figure 4 illustrates how the planar rotations around a fixed axis (Figure 4(a)) can be viewed as local patches on the sphere in 3D (Figure 4(b)); thus, relating to rotations in 3D, and finally how this may be generalized to full rigid motions by translation as shown in Figure 4(c). The mapping γ is learned as part of the end-to-end APR.

Intuitively, one can interpret this as an approximation of the space of camera poses by a finite set of learned ones, with feature equivariance used to generalize and extend the coverage within the space. Indeed, an SE(2)-equivariant model is capable of generalizing from each learned pose to every poses that preserve the image plane (*i.e.* z-rotated and x,y-translated versions of the original camera). In other words, instead of learning some cropped image planes like classical CNNs do, relying on an SE(2)-equivariant CNN rather consists in learning several infinite image planes, therefore providing a denser coverage of the scene space.

5. Proposed *E-PoseNet*

This section gives an overview of our proposed equivariant pose regression model, *E-PoseNet*.

To be able to assess how explicitly encoding pose information into the feature space can result in a more accurate

[3] For the sake of clarity, and without loss of generality, we drop the bias $\mathbf{b}$ in this discussion.

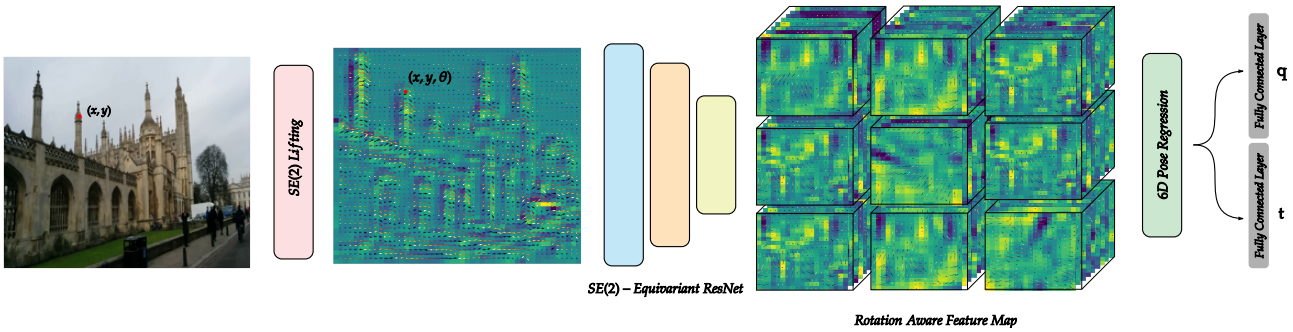

Figure 5. **E-PoseNet -** Our pose regression pipeline leverages a roto-translation equivariant ResNet18 [24] backbone, two fully connected Multilayer Perceptrons (MLP) for lifting the features to a higher dimensional space, followed by two branches for separately regressing the position and orientation of the camera.

and data-efficient pose regressor, we proceed from the architecture of PoseNet [35]. We follow the same pipeline, except that we substitute the GoogLeNet backbone by an $\mathrm{SE}(2)$-equivariant [61] version of ResNet [24], to extract both translation and rotation-equivariant features. The resulting model is presented in Figure 5.

Network Architecture. *E-PoseNet* is composed of a roto-translation equivariant ResNet18 backbone, two fully connected Multilayer Perceptrons (MLP) for lifting the features to a higher dimensional space, followed by two branches for separately regressing the position and orientation of the camera. Each branch consists of an independent fully-connected MLP head.

SE(2)-Equivariant Backbone. Our backbone, *i.e.,* feature extractor $\mathcal{F}$, is an $\mathrm{SE}(2)$ roto-translation equivariant version of ResNet. Specifically, we use the *e2cnn* [61] implementation for $\mathrm{E}(2)$-equivariant convolution, pooling, normalization, and non linearities, to build an equivariant ResNet18. To decrease the computational cost, we discretize the $\mathrm{SE}(2)$ group making the model only equivariant to the $\left(\mathbb{R}^2, +\right) \rtimes C_N$ group, meaning all translations in $\mathbb{R}^2$ and rotations by angles multiple of $\frac{2\pi}{N}$. Extracted features are now rotation-equivariant feature maps V with the size $(K \times N \times H \times W)$, where K is the number of channels, N the number of feature orientations (for our model we used $N = 8$), and H, W respectively, the height and width.
In addition to classical translation information, obtained features thus encode rotation information that can enhance the pose regression. Furthermore, equivariance to broader transformations constrains the network in a way that can aid generalization, especially due to the weights shared under image rotations [9]. Finally, this rotation-equivariant ResNet shows a significant reduction in model size, about $1/N$ parameters compared to the regular ResNet architecture, to obtain the same feature size. Indeed, the size of classical feature maps is in the form $(K \times H \times W)$.

Loss Function. To regress camera poses, we use the loss function introduced in [34], and defined as:

$$\mathcal{L}_{\mathcal{P}} = \mathcal{L}_{\mathbf{t}} \exp\left(-s_{\mathbf{t}}\right) + s_{\mathbf{t}} + \mathcal{L}_{\mathsf{R}} \exp\left(-s_{\mathsf{R}}\right) + s_{\mathsf{R}}, \quad (10)$$

where the position loss $\mathcal{L}_{\mathbf{t}} = \left\|\mathbf{t}_0 - \mathbf{t}\right\|_2$, and the orientation loss $\mathcal{L}_{\mathsf{R}} = \left\|\mathbf{q}_0 - \frac{\mathbf{q}}{\|\mathbf{q}\|_2}\right\|_2$, are computed from predicted ($\mathbf{q}$,$\mathbf{t}$) and groundtruth ($\mathbf{q}_0$,$\mathbf{t}_0$) camera poses, considering the quaternion representation for orientations. $s_{\mathbf{t}}$, s_{R} are learned parameters.

6. Experiments and Analysis

The proposed method aims to improve APR accuracy by utilizing an equivariant feature extraction backbone able to learn geometry-aware feature maps. We first show the effect of $\mathrm{SE}(2)$-equivariant models on rotated feature maps using samples from the T-Less dataset [26]. Then, we benchmark our proposed *E-PoseNet* on two datasets for both indoor and outdoor camera localization.

Equivariance analysis on T-Less. In this study, we use a sequence of 'object 5' from the T-less training dataset [26]. With only one textureless symmetric object present in the scene and undergoing continuous rotations, this sequence represents an ideal case for testing the impact of the different rotation parametrization and channeling. To assess the effect of equivariance, our backbone is made of 10 convolution layers with kernel size equal to 2, ELU non-linearity and Max Pooling downsampling every two layers with kernel size equal to 2. Different degrees of equivariance were tested, namely, Classical CNN "translation equivariant", Equivariant 90° (N=4), Equivariant 45° (N=8), Equivariant 18° (N=20), Equivariant 10° (N=36), and finally the Equivariant SO(2). Equivariant models are generated based on *e2cnn* implementation [61]. We trained the model on one sequence only, without any data augmentation and for 100 epochs. Number of parameters, optimizer, learning rate and random seeds were fixed.

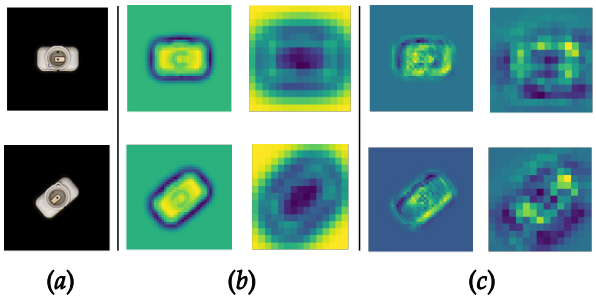

Figure 6. **Extracted feature maps -** difference between: (b) equivariant CNN and (c) classical CNN. The used samples (a) are from the T-LESS Dataset [26].

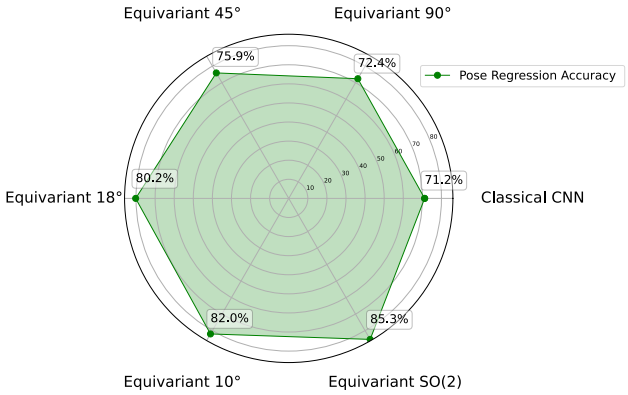

Figure 7. **Equivariant models comparison -** On sequence of 'object 5' from the T-less train dataset [26], we can see the increase in the model accuracy when increasing the equivariant group.

Figure 7 reports the proportion of samples for which the predicted pose error is below 10cm, 10°. We observed that increasing the level of equivariance, *i.e.,* decreasing the discrete sampling angle, leads to increasing the performance of the pose estimation model. Furthermore, the best reported performance has been achieved by the $\mathrm{SO}(2)$ continuous rotation equivariance. The metric used here does not follow the standard T-LESS metric since it is only used for model variants comparison.

The difference between rotation-equivariant and classical CNN features is highlighted in Figure 6. By using images with different orientations (Figure $6(a)$) as input, the same transformation links extracted feature maps from different stages of the model (b). In contrast, this is not the case with the classical CNN where the obtained feature maps are not rotated versions of each other (c).

Datasets.

Cambridge Landmarks – We use this dataset [35] to evaluate the performance of *E-PoseNet* in outdoor camera relocalization. It is a large scale dataset taken around Cambridge University, containing original videos labelled with 6-DoF camera poses and a visual reconstruction of the scene (spatial extent of $\sim 900 - 5500m^2$). We train and evaluate *E-PoseNet* on four scenes (see Table 1). Furthermore, a few samples are used to visualize the obtained *E-PoseNet* feature fields. Figure 8 shows that they directly support representations of $\mathrm{SE}(2)$ and are therefore enriched with some notion of orientation, visualized in a vector field form. On the contrary classical CNNs do not encode geometric information directly into their feature space.

Figure 8. **Feature visualization for Cambridge Landmarks -** Visualization of samples images from the Cambridge Landmarks dataset [35] (left), along with their respective $\mathrm{SE}(2)$ group representations learned by *E-PoseNet* (right).

7-Scenes – For indoor camera localization, we use the 7-Scenes dataset [53] which is a collection of tracked RGB-D camera frames for indoor scenes with a spatial extent of $\sim 1{-}10m^2$. Only RGB images are used in our experiments. Finally, the two datasets present various challenges, *i.e.,* occlusion, reflections, motion blur, lighting conditions, repetitive textures, and variations in viewpoint and trajectory.

Comparative Analysis of Camera Pose Regressors. We compare the performance of *E-PoseNet* with state-of-the-art APR methods for camera localization in both outdoor and indoor scenes. First, we tested the performance on the Cambridge Landmarks dataset, for which we provide the median position and orientation errors in Table 1. We also compare the performance of *E-PoseNet* with respect to the state-of-the-art monocular pose regressors reporting on the 7-Scenes dataset. Table 2 contains the results. From the results on both datasets, and in comparison with APR methods, we

	King's College	*Old Hospital*	*Shop Facade*	*St. Mary*
DenseVLAD + Inter. (Baseline) [56]	1.48/4.45	2.68/4.63	0.90/4.32	1.62/6.06
PoseNet (PN) [35]	1.92/5.40	2.31/5.38	1.46/8.08	2.65/8.48
PN learned weights [34]	0.99/**1.06**	2.17/2.94	1.05/3.97	1.49/3.43
BayesianPN [33]	1.74/4.06	2.57/5.14	1.25/7.54	2.11/8.38
LSTM-PN [58]	0.99/3.65	1.51/4.29	1.18/7.44	1.52/6.68
SVS-Pose [45]	1.06/2.81	1.50/4.03	0.63/5.73	2.11/8.11
GPoseNet [7]	1.61/2.29	2.62/3.89	1.14/5.73	2.93/6.46
MapNet [3]	1.07/1.89	1.94/3.91	1.49/4.22	2.00/4.53
IRPNet [52]	1.18/2.19	1.87/3.38	0.72/3.47	1.87/4.94
MS-Transformer [15]	0.83/1.47	1.81/**2.39**	0.86/3.07	1.62/3.99
TransPoseNet [16]	**0.60**/2.43	1.45/3.08	**0.55**/3.49	1.09/4.99
E-PoseNet (Ours)	0.95/1.63	**1.43**/2.64	0.60/**2.78**	**1.00/3.16**

Table 1. **Comparative analysis of pose regressors on Cambridge Landmarks dataset (outdoor localization) [35] -** We report the median position/orientation error in meters/degrees for each method. Best results are highlighted in bold.

	Chess	*Fire*	*Heads*	*Office*	*Pumpkin*	*Kitchen*	*Stairs*
DenseVLAD + Inter. [56]	0.18/10.0	0.33/12.4	0.15/14.3	0.25/10.1	0.26/9.42	0.27/11.1	**0.24**/14.7
PoseNet (PN) [35]	0.32/8.12	0.47/14.4	0.29/12.0	0.48/7.68	0.47/8.42	0.59/8.64	0.47/13.8
PN learned weights [34]	0.14/4.50	0.27/11.8	0.18/12.1	0.20/5.77	0.25/4.82	0.24/5.52	0.37/10.6
BayesianPN [33]	0.37/7.24	0.43/13.7	0.31/12.0	0.48/8.04	0.61/7.08	0.58/7.54	0.48/13.1
LSTM-PN [58]	0.24/5.77	0.34/11.9	0.21/13.7	0.30/8.08	0.33/7.0	0.37/8.83	0.40/13.7
GPoseNet [7]	0.20/7.11	0.38/12.3	0.21/13.8	0.28/8.83	0.37/6.94	0.35/8.15	0.37/12.5
GeoPoseNet [34]	0.13/4.48	0.27/11.3	0.17/13.0	0.19/5.55	0.26/4.75	0.23/5.35	0.35/12.4
MapNet [3]	**0.08**/3.25	0.27/11.7	0.18/13.3	0.17/**5.15**	0.22/4.02	0.23/**4.93**	0.30/12.1
IRPNet [52]	0.13/5.64	0.25/9.67	0.15/13.1	0.24/6.33	0.22/5.78	0.30/7.29	0.34/11.6
AttLoc [59]	0.10/4.07	0.25/11.4	0.16/11.8	0.17/5.34	0.21/4.37	0.23/5.42	0.26/10.5
MS-Transformer [15]	0.11/4.66	0.24/**9.60**	0.14/12.2	0.17/5.66	0.18/4.44	**0.17**/5.94	0.26/8.45
TransPoseNet [16]	**0.08**/5.68	0.24/10.6	**0.13**/12.7	0.17/6.34	0.17/5.60	0.19/6.75	0.30/**7.02**
E-PoseNet (Ours)	**0.08/2.57**	**0.21**/11.0	0.16/**10.3**	**0.15**/6.80	**0.16/3.82**	0.20/6.81	**0.24**/9.92

Table 2. **Comparative analysis of pose regressors on the 7-Scenes dataset (indoor localization) [53] -** We report the median position/orientation error in meters/degrees for each method. Best results are highlighted in bold.

conclude that the proposed *E-PoseNet* achieves the lowest location error across all the outdoor and indoor scenes, and the lowest orientation error across the majority of them. It also competes with most recent transformer-based architectures [15, 16] on these datasets.

Implementation Details. We tested different architectures for the equivariant backbone, with ResNet18 being the most suitable model in our experiments, from both model size and performance perspectives. We trained our model for $5 - 10$k epochs using Adam optimizer [36], with $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-5}$ and a batch size of 256. During the training phase, we rescaled the image so that its smaller length is 256 pixels followed by a random 224×224 crop. No further data augmentation was used.

Limitations. While we focus on introducing equivariant operations for the feature extraction part of the APR pipeline, the following stages (*i.e.* embedding, regression) do not have the same property, resulting in breaking the equivariance of the overall pipeline. Another limitation of the proposed APR model is the longer time required for equivariant CNN models as compared to classical CNN ones. Note that this is only during training, while the inference time is similar for both types of models.

7. Conclusions

This paper presents a new direction for the problem of camera pose regression leveraging equivariant features to encode more geometric information about the input image. By using an $\mathrm{SE}(2)$-equivariant feature extractor, our model is able to outperform existing methods on both outdoor and indoor benchmarks. Furthermore, we conclude that the equivariant properties of deep learning models that are used for geometric reasoning offer a promising direction for reaching the potential of absolute pose regression.

References

[1] Mahbubul Alam, Manar D. Samad, L. Vidyaratne, Alexander Glandon, and Khan M. Iftekharuddin. Survey on deep neural networks in speech and vision systems. *Neurocomputing*, 417:302–321, 2020. 1

[2] Erik J. Bekkers, Remco Duits, Tos Berendschot, and Bart M. ter Haar Romeny. A multi-orientation analysis approach to retinal vessel tracking. *J. Math. Imaging Vis.*, 49(3):583–610, 2014. 3

[3] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2018. 2, 8

[4] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021. 2, 5

[5] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1872–1886, 2013. 3

[6] Ming Cai, Chunhua Shen, and Ian Reid. A hybrid probabilistic model for camera relocalization. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 238, 2018. 2

[7] Ming Cai, Chunhua Shen, and Ian Reid. A hybrid probabilistic model for camera relocalization. 2019. 8

[8] Haiwei Chen, Shichen Liu, Weikai Chen, Hao Li, and Randall Hill. Equivariant point network for 3d point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14514–14523, June 2021. 2, 3

[9] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016. 2, 3, 6

[10] Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. 2

[11] Taco S. Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9142–9153, 2019. 5

[12] Taco S. Cohen and Max Welling. Steerable cnns. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. 2

[13] Carlos Esteves, Avneesh Sud, Zhengyi Luo, Kostas Daniilidis, and Ameesh Makadia. Cross-domain 3d equivariant image embeddings. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1812–1822. PMLR, 2019. 2

[14] Carlos Esteves, Avneesh Sud, Zhengyi Luo, Kostas Daniilidis, and Ameesh Makadia. Cross-domain 3d equivariant image embeddings. In *International Conference on Machine Learning*, pages 1812–1822. PMLR, 2019. 3

[15] Shavit et al. Learning multi-scene absolute pose regression with transformers. In *ICCV 2021*. 8

[16] Shavit et al. Paying attention to activation maps in camera pose regression. *CoRR*, abs/2103.11477, 2021. 8

[17] Mario Ferraro and Terry M Caelli. Lie transformation groups, integral transforms, and invariant pattern recognition. *Spatial Vision*, 1994. 3

[18] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 1

[19] William T. Freeman and Edward H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(9):891–906, 1991. 2

[20] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 3

[21] Deepak K. Gupta, Devanshu Arya, and Efstratios Gavves. Rotation equivariant siamese networks for tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12362–12371, June 2021. 2, 3

[22] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2786–2795, June 2021. 2, 3

[23] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 4

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016. 6

[25] Tomas Hodan, Daniel Barath, and Jiri Matas. Epos: Estimating 6d pose of objects with symmetries. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[26] Tomáš Hodaň, Pavel Haluza, Štěpán Obdržálek, Jiří Matas, Manolis Lourakis, and Xenophon Zabulis. T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017. 6, 7

[27] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Segmentation-driven 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[28] Md Amirul Islam, Matthew Kowal, Sen Jia, Konstantinos G Derpanis, and Neil DB Bruce. Position, padding and predictions: A deeper look at position information in cnns. *arXiv preprint arXiv:2101.12322*, 2021. 3

[29] Michiel Janssen, A. J. E. M. Janssen, Erik J. Bekkers, J. Olivan Bescos, and Remco Duits. Design and processing of invertible orientation scores of 3d images. *J. Math. Imaging Vis.*, 60(9):1427–1458, 2018. 3

[30] Kenichi Kanatani. *Group Theoretical Methods in Image Understanding*. Springer-Verlag, Berlin, Heidelberg, 1990. 4

[31] Osman Semih Kayhan and Jan C. van Gemert. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[32] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, May 16-21, 2016*, pages 4762–4769, 2016. 2

[33] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *2016 IEEE international conference on Robotics and Automation (ICRA)*, pages 4762–4769. IEEE, 2016. 8

[34] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6555–6564. IEEE Computer Society, 2017. 2, 6, 8

[35] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2938–2946. IEEE Computer Society, 2015. 2, 6, 7, 8

[36] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 8

[37] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems 2, [NIPS Conference, Denver, Colorado, USA, November 27-30, 1989]*, pages 396–404, 1989. 3

[38] Vincent Lepetit and Pascal Fua. Monocular model-based 3d tracking of rigid objects: A survey. *Found. Trends Comput. Graph. Vis.*, 1(1), 2005. 1

[39] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate $O(n)$ solution to the pnp problem. *Int. J. Comput. Vis.*, 81(2):155–166, 2009. 1

[40] Qianli Liao and Tomaso Poggio. Exact equivariance, disentanglement and invariance of transformations. Technical report, 2017. 3

[41] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004. 2

[42] Éric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: A hands-on survey. *IEEE Trans. Vis. Comput. Graph.*, 22(12):2633–2651, 2016. 1

[43] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Image-based localization using hourglass networks. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pages 870–877. IEEE Computer Society, 2017. 2

[44] Tayyab Naseer and Wolfram Burgard. Deep regression for monocular camera-based 6-dof global localization in outdoor environments. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017*, pages 1525–1530. IEEE, 2017. 2

[45] Tayyab Naseer and Wolfram Burgard. Deep regression for monocular camera-based 6-dof global localization in outdoor environments. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1525–1530. IEEE, 2017. 8

[46] Klas Nordberg and Gösta H. Granlund. Equivariance and invariance-an approach based on lie groups. In *Proceedings 1996 International Conference on Image Processing, Lausanne, Switzerland, September 16-19, 1996*, pages 181–184, 1996. 3

[47] Chris Olah, Nick Cammarata, Chelsea Voss, Ludwig Schubert, and Gabriel Goh. Naturally occurring equivariance in neural networks. *Distill*, 2020. https://distill.pub/2020/circuits/equivariance. 3

[48] Edouard Oyallon and Stephane Mallat. Deep roto-translation scattering for object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 3

[49] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7667–7676. IEEE, 2019. 1

[50] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. Understanding the limitations of cnn-based absolute camera pose regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3

[51] Uwe Schmidt and Stefan Roth. Learning rotation-aware features: From invariant priors to equivariant descriptors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012. 3

[52] Yoli Shavit and Ron Ferens. Do we really need scene-specific pose encoders? In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3186–3192. IEEE, 2021. 8

[53] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 7, 8

[54] Liangchen Song, Gang Yu, Junsong Yuan, and Zicheng Liu. Human pose estimation and its application to action recognition: A survey. *Journal of Visual Communication and Image Representation*, 76:103055, 2021. 1

[55] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent

Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[56] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2015. 8

[57] Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using lstms for structured feature correlation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 627–637, 2017. 2

[58] Florian Walch, Caner Hazirbas, Laura Leal-Taixe, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using lstms for structured feature correlation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 627–637, 2017. 8

[59] Bing Wang, Changhao Chen, Chris Xiaoxuan Lu, Peijun Zhao, Niki Trigoni, and Andrew Markham. Atloc: Attention guided camera localization. *arXiv preprint arXiv:1909.03557*, 2019. 8

[60] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16611–16621, June 2021. 2

[61] Maurice Weiler and Gabriele Cesa. General e(2)-equivariant steerable cnns. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14334–14345, 2019. 2, 5, 6

[62] Jian Wu, Liwei Ma, and Xiaolin Hu. Delving deeper into convolutional neural networks for camera relocalization. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5644–5651, 2017. 2

[63] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018. 2

[64] Chi Xu, Lilian Zhang, Li Cheng, and Reinhard Koch. Pose estimation from line correspondences: A complete analysis and a series of solutions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1209–1222, 2017. 1

[65] Jerry Jun Yokono and Tomaso A. Poggio. Oriented filters for object recognition: an empirical study. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2004), May 17-19, 2004, Seoul, Korea*, pages 755–760, 2004. 2

[66] Chao Zhang, Ignas Budvytis, Stephan Liwicki, and Roberto Cipolla. Rotation equivariant orientation estimation for omnidirectional localization. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 3