

End-to-End Learnable Geometric Vision by Backpropagating PnP Optimization

Bo Chen¹ Álvaro Parra¹ Jiewei Cao¹ Nan Li² Tat-Jun Chin¹

¹The University of Adelaide ²Shenzhen University

{bo.chen, alvaro.parrabustos, jiewei.cao, tat-jun.chin}@adelaide.edu.au nan.li@szu.edu.cn

Abstract

Deep networks excel in learning patterns from large amounts of data. On the other hand, many geometric vision tasks are specified as optimization problems. To seamlessly combine deep learning and geometric vision, it is vital to perform learning and geometric optimization end-to-end. Towards this aim, we present BPnP, a novel network module that backpropagates gradients through a Perspective-n-Points (PnP) solver to guide parameter updates of a neural network. Based on implicit differentiation, we show that the gradients of a “self-contained” PnP solver can be derived accurately and efficiently, as if the optimizer block were a differentiable function. We validate BPnP by incorporating it in a deep model that can learn camera intrinsics, camera extrinsics (poses) and 3D structure from training datasets. Further, we develop an end-to-end trainable pipeline for object pose estimation, which achieves greater accuracy by combining feature-based heatmap losses with 2D-3D reprojection errors. Since our approach can be extended to other optimization problems, our work helps to pave the way to perform learnable geometric vision in a principled manner. Our PyTorch implementation of BPnP is available on <http://github.com/BoChenYS/BPnP>.

1. Introduction

The success of deep learning is due in large part to its ability to learn patterns from vast amounts of training data. Applications that have benefited from this ability include object detection and image segmentation [26, 19]. Fundamentally, such problems can often be formulated as classification/regression problems, which facilitates suitable objective functions for backpropagation learning [29].

On the other hand, there are many important computer vision tasks that are traditionally formulated as geometric optimization problems, e.g., camera localization/pose estimation, 3D reconstruction, point set registration. A common property in these optimization problems is the minimization of a residual function (e.g., sum of squared reprojection errors) defined over geometric quantities (e.g.,

6DOF camera poses), which are not immediately amenable to backpropagation learning. This limits the potential of geometric vision tasks to leverage large datasets.

A straightforward solution towards “learnable” geometric vision is to replace the “front end” modules (e.g., image feature detection and matching) using a deep learning alternative [48, 55, 45]. However, this does not allow the “back end” steps (e.g., searching for optimal geometric quantities) to influence the training of the neural network parameters.

On the other extreme, end-to-end methods have been devised [23, 21, 22, 8, 32, 50, 52, 9] that bypass geometric optimization, by using fully connected layers to compute the geometric quantity (e.g., 6DOF camera pose) from a feature map derived from previous layers. However, it has been observed that these methods are equivalent to performing image retrieval [39], which raises questions on their ability to generalize. Also, such end-to-end methods do not explicitly exploit established methods from geometric vision [18], such as solvers for various well-defined tasks.

To benefit from the strengths of deep learning and geometry, it is vital to combine them in a mutually reinforcing manner. One approach is to incorporate a geometric optimization solver in a deep learning architecture, and allow the geometric solver to participate in guiding the updates of the neural network parameters, thereby realising end-to-end learnable geometric vision. The key question is how to compute gradients from a “self-contained” optimizer.

A recent work towards the above goal is differentiable RANSAC [3, 5, 6], which was targeting at the camera localization task. A perspective-n-point (PnP) module was incorporated in a deep architecture, and the derivatives of the PnP solver are calculated using central differences [38] to enable parameter updates in the rest of the pipeline. However, such an approach to compute gradients is inexact and time consuming because, in order to obtain each partial derivative, it requires solving PnP at values that lie to the left and right of the input.

Other approaches to derive gradients from an independent optimization block for backpropagation learning [16, 1] conduct implicit differentiation [2, Chap. 8]. Briefly, in the context of end-to-end learning, the gradient of the opti-

mization routine with respect to the input variables can be computed via partial derivatives of the stationary constraints of the optimization problem (more details in Sec. 3). The gradient can then be backpropagated to the previous layers for parameter updates. A number of motivating examples and applications were explored in [16, 1]. However, larger-scale experiments in the context of specific geometric vision problems, and benchmarking against other end-to-end learning alternatives, were unavailable in [16, 1]. It is worth noting that implicit differentiation of optimization subroutines has been explored previously in several computer vision applications [46, 13, 40] (also earlier in [14, Chap. 5]).

Contributions Our main contribution is a novel network module called *BPnP* that incorporates a PnP solver. BPnP backpropagates the gradients through the PnP “layer” to guide the updates of the neural network weights, thereby achieving end-to-end learning using an established objective function (sum of squared 2D-3D reprojection errors) and solver from a geometric vision problem. Despite incorporating only a PnP solver, we show how BPnP can be used to learn effective deep feature representations for multiple geometric vision tasks (pose estimation, structure-from-motion, camera calibration). We also compare our method against state-of-the-art methods for geometric vision tasks. Fundamentally, our method is based on implicit differentiation; thus our work can be seen as an application of [16, 1] to geometric vision learning.

2. Related works

Backpropagating optimization problems As alluded to above, there are several works that incorporate optimizer blocks in deep neural network architectures, and perform differentiation of the optimization routines for backpropagation learning. A subset of these works address the challenge of incorporating RANSAC in an end-to-end trainable pipeline, such as DSAC [3], ESAC [5], and NG-DSAC [6]. In fact, since these works aim to solve camera localization, they also incorporate a PnP solver in their pipeline. To backpropagate through the PnP solver, they use central differences to compute the partial derivatives. In effect, if the input dimension is n , it requires solving PnP $2n$ times in order to obtain the full Jacobian. Another group of methods applies implicit differentiation [16, 1], which provides an exact and efficient solution for backpropagating through an optimization process. We will describe implicit differentiation in detail later.

Pose estimation from images A target application of our BPnP is pose estimation. Existing works on end-to-end pose estimation [23, 21, 22, 8, 32, 50, 52] usually employ fully connected layers to compute the target output

(pose) using feature maps from previous layers. The output loss function is typically defined using pose metrics (e.g., chordal distance), which are backpropagated using standard differentiation. A recent analysis [39] suggests that what is being performed by these end-to-end networks is akin to learning a set of base poses from the training images, computing a set of weights for the testing image, then predicting the pose as a weighted combination of the base poses. It was further shown that such methods were more related to image retrieval than intrinsically learning to predict pose, hence they may not outperform an image retrieval baseline [39].

Other pose estimation approaches that combine deep learning with geometric optimization (PnP solver) [35, 37, 47, 36, 10] adopt a two-stage strategy: first learn to predict the 2D landmarks or fiducial points from the input image, then perform pose estimation by solving PnP on the 2D-3D correspondences. While the first stage can benefit from the regularities existing in a training dataset, the second stage (PnP solving) which encodes the fundamental geometric properties of the problem do not influence the learning in the first stage. Contrast this to our BPnP which seamlessly connects both stages, and allows the PnP optimizer to guide the weight updates in the first stage (in addition to standard keypoint or landmark regression losses).

Depth estimation and 3D reconstruction There exist many works that employ deep networks to learn to predict depth or 3D structure from input images in an end-to-end fashion. Some of these works [17, 11, 24] can only impose constraints on pairs of images, while others [57, 49] learn the structure and the motion in different network branches and do not impose explicit geometric constraints. Also, many of such works [12, 27, 31, 30, 28, 54, 51] require training datasets with ground truth depth labels, which can be expensive to obtain. The proposed BPnP may help to alleviate this shortcoming; as we will show in Sec. 4.2, a simple structure-from-motion (SfM) framework that utilizes BPnP can jointly optimize using multiple views (not just two), explicitly impose geometric constraints, and learn structure and motion in an unsupervised fashion without depth labels or ground truth 3D structures.

3. Backpropagating a PnP solver (BPnP)

Let g denote a PnP solver in the form of a “function”

$$\mathbf{y} = g(\mathbf{x}, \mathbf{z}, \mathbf{K}), \quad (1)$$

which returns the 6DOF pose \mathbf{y} of a camera with intrinsic matrix $\mathbf{K} \in \mathbb{R}^{3 \times 3}$ from n 2D-3D correspondences

$$\mathbf{x} = [\mathbf{x}_1^T \quad \mathbf{x}_2^T \quad \dots \quad \mathbf{x}_n^T]^T \in \mathbb{R}^{2n \times 1}, \quad (2)$$

$$\mathbf{z} = [\mathbf{z}_1^T \quad \mathbf{z}_2^T \quad \dots \quad \mathbf{z}_n^T]^T \in \mathbb{R}^{3n \times 1}, \quad (3)$$

where $(\mathbf{x}_i, \mathbf{z}_i)$ is the i -th correspondence. Let $\pi(\cdot | \mathbf{y}, \mathbf{K})$ be a projective transformation of 3D points onto the image plane with pose \mathbf{y} and camera intrinsics \mathbf{K} . Intrinsically, the “evaluation” of g requires solving the optimization problem

$$\mathbf{y} = \arg \min_{\mathbf{y} \in SE(3)} \sum_{i=1}^n \|\mathbf{r}_i\|_2^2, \quad (4)$$

where

$$\mathbf{r}_i = \mathbf{x}_i - \boldsymbol{\pi}_i \quad (5)$$

is the reprojection error of the i -th correspondence and

$$\boldsymbol{\pi}_i = \pi(\mathbf{z}_i | \mathbf{y}, \mathbf{K}) \quad (6)$$

is the projection of 3D point \mathbf{z}_i on the image plane. We introduce the shorthand

$$\boldsymbol{\pi} := [\boldsymbol{\pi}_1^T, \dots, \boldsymbol{\pi}_n^T]^T, \quad (7)$$

thus (4) can be rewritten as

$$\mathbf{y} = \arg \min_{\mathbf{y} \in SE(3)} \|\mathbf{x} - \boldsymbol{\pi}\|_2^2. \quad (8)$$

The choice of formulation (8) will be justified in Sec. 3.3.

Our ultimate goal is to incorporate g in a learnable model, where \mathbf{x} , \mathbf{z} and \mathbf{K} can be the (intermediate) outputs of a deep network. Moreover, the solver for (8) should be used to participate in the learning of the network parameters. To this end, we need to treat g as if it were a differentiable function, such that its “gradients” can be backpropagated to the rest of the network. In this section, we show how this can be achieved via implicit differentiation.

3.1. The Implicit Function Theorem (IFT)

Theorem 1 ([25]) *Let $f : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^m$ be a continuously differentiable function with input $(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^n \times \mathbb{R}^m$. If a point $(\mathbf{a}^*, \mathbf{b}^*)$ satisfies*

$$f(\mathbf{a}^*, \mathbf{b}^*) = \mathbf{0} \quad (9)$$

and the Jacobian matrix $\frac{\partial f}{\partial \mathbf{b}}(\mathbf{a}^, \mathbf{b}^*)$ is invertible, then there exists an open set $U \subset \mathbb{R}^n$ such that $\mathbf{a}^* \in U$ and an unique continuously differentiable function $g(\mathbf{a}) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that $\mathbf{b}^* = g(\mathbf{a}^*)$ and*

$$f(\mathbf{a}', g(\mathbf{a}')) = \mathbf{0}, \forall \mathbf{a}' \in U. \quad (10)$$

Moreover, for all $\mathbf{a}' \in U$, the Jacobian matrix $\frac{\partial g}{\partial \mathbf{a}}(\mathbf{a}')$ is given by

$$\frac{\partial g}{\partial \mathbf{a}}(\mathbf{a}') = - \left[\frac{\partial f}{\partial \mathbf{b}}(\mathbf{a}', g(\mathbf{a}')) \right]^{-1} \left[\frac{\partial f}{\partial \mathbf{a}}(\mathbf{a}', g(\mathbf{a}')) \right]. \quad (11)$$

The IFT allows computing the derivatives of a function g with respect to its input \mathbf{a} without an explicit form of the function, but with a function f constraining \mathbf{a} and $g(\mathbf{a})$.

3.2. Constructing the constraint function f

To invoke the IFT for implicit differentiation, we first need to define the constraint function $f(\mathbf{a}, \mathbf{b})$ such that Eq. (9) is upheld. For our problem, we use all four variables \mathbf{x} , \mathbf{y} , \mathbf{z} and \mathbf{K} to construct f . But we treat f as a two variables function $f(\mathbf{a}, \mathbf{b})$, in which \mathbf{a} takes values in $\{\mathbf{x}, \mathbf{z}, \mathbf{K}\}$ - depending on which partial derivative to obtain - and $\mathbf{b} = \mathbf{y}$ (*i.e.*, the output pose of g).

To uphold Eq. (9), we exploit the stationary constraint of the optimization process. Denote the objective function of the PnP solver g as

$$o(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{K}) = \sum_{i=1}^n \|\mathbf{r}_i\|_2^2. \quad (12)$$

Since the output pose \mathbf{y} of a PnP solver is a local optimum for the objective function, a stationary constraint can be established by taking the first order derivative of the objective with respect to \mathbf{y} , *i.e.*,

$$\frac{\partial o(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{K})}{\partial \mathbf{y}} \Big|_{\mathbf{y}=g(\mathbf{x}, \mathbf{z}, \mathbf{K})} = \mathbf{0}. \quad (13)$$

Given an output pose from a PnP solver $\mathbf{y} = [y_1, \dots, y_m]^T$, we construct f based on Eq. (13), which can be written as

$$f(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{K}) = [f_1, \dots, f_m]^T, \quad (14)$$

where for all $j \in \{1, \dots, m\}$,

$$f_j = \frac{\partial o(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{K})}{\partial y_j} \quad (15)$$

$$= 2 \sum_{i=1}^n \langle \mathbf{r}_i, \frac{\partial \mathbf{r}_i}{\partial y_j} \rangle \quad (16)$$

$$= \sum_{i=1}^n \langle \mathbf{r}_i, \mathbf{c}_{ij} \rangle \quad (17)$$

with

$$\mathbf{c}_{ij} = -2 \frac{\partial \boldsymbol{\pi}_i}{\partial y_j}. \quad (18)$$

3.3. Forward and backward pass

Our PnP formulation (8) for g essentially performs least squares (LS) estimation, which is not robust towards outliers (egregious errors in \mathbf{x} , \mathbf{z} and \mathbf{K}). Alternatively, we could apply a more robust objective, such as incorporating an M-estimator [56] or maximizing the number of inliers [15]. However, our results suggest that LS is actually more appropriate, since its sensitivity to errors in the input measurements encourages the learning to quickly converge to parameters that do not yield outliers in \mathbf{x} , \mathbf{z} and \mathbf{K} . In

contrast, a robust objective would block the error signals of the outliers, causing the learning process to be unstable.

Given (8), the choice of the solver remains. To conduct implicit differentiation, we need not solve (8) exactly, since (13) is simply the stationary condition of (8), which is satisfied by any local minimum. To this end, we apply the Levenberg-Marquardt (LM) algorithm (as implemented in the SOLVEPNP_ITERATIVE method in OpenCV [7]), which guarantees local convergence. As an iterative algorithm, LM requires initialization $\mathbf{y}^{(0)}$ in solving (8). We make explicit this dependence by rewriting (1) as

$$\mathbf{y} = g(\mathbf{x}, \mathbf{z}, \mathbf{K}, \mathbf{y}^{(0)}). \quad (19)$$

We obtain the initial pose $\mathbf{y}^{(0)}$ with RANSAC if it is not provided.

In the backward pass, we first construct f as described in Sec. 3.2 to then obtain the Jacobians of g with respect to each of its inputs as

$$\frac{\partial g}{\partial \mathbf{x}} = - \left[\frac{\partial f}{\partial \mathbf{y}} \right]^{-1} \left[\frac{\partial f}{\partial \mathbf{x}} \right], \quad (20)$$

$$\frac{\partial g}{\partial \mathbf{z}} = - \left[\frac{\partial f}{\partial \mathbf{y}} \right]^{-1} \left[\frac{\partial f}{\partial \mathbf{z}} \right], \quad (21)$$

$$\frac{\partial g}{\partial \mathbf{K}} = - \left[\frac{\partial f}{\partial \mathbf{y}} \right]^{-1} \left[\frac{\partial f}{\partial \mathbf{K}} \right]. \quad (22)$$

Given the output gradient $\nabla \mathbf{y}$, BPnP returns the input gradients

$$\nabla \mathbf{x} = \left[\frac{\partial g}{\partial \mathbf{x}} \right]^T \nabla \mathbf{y}, \quad (23)$$

$$\nabla \mathbf{z} = \left[\frac{\partial g}{\partial \mathbf{z}} \right]^T \nabla \mathbf{y}, \quad (24)$$

$$\nabla \mathbf{K} = \left[\frac{\partial g}{\partial \mathbf{K}} \right]^T \nabla \mathbf{y}. \quad (25)$$

3.4. Implementation notes

The number of dimensions of \mathbf{y} , i.e., m , is dependant on the parameterization of $SO(3)$ within the pose. For example, $m = 6$ for the axis-angle representation, $m = 7$ for the quaternion representation, and $m = 12$ for the rotation matrix representation. Experimentally we found the axis-angle representation leads to the best result, possibly since then $m = 6$ is equal to the degrees of freedom.

We compute the partial derivatives in Eqs. (18), (20), (21), and (22) using the Pytorch autograd package [34].

4. End-to-end learning with BPnP

BPnP enables important geometric vision tasks to be solved using deep networks and PnP optimization in an end-to-end manner. Here, we explore BPnP for pose estimation,

Algorithm 1 Pose estimation.

```

1:  $\mathbf{y} \leftarrow$  Identity pose.
2: Randomly initialize  $\theta$ 
3: while loss  $\ell$  has not converged do
4:    $\mathbf{x} \leftarrow h(I; \theta)$ .
5:    $\mathbf{y} \leftarrow g(\mathbf{x}, \mathbf{z}, \mathbf{K}, \mathbf{y})$ .
6:    $\ell \leftarrow l(\mathbf{x}, \mathbf{y})$ .
7:    $\theta \leftarrow \theta - \alpha \frac{\partial \ell}{\partial \theta}$ . (Backpropagate through PnP)
8: end while

```

SfM and camera calibration, and report encouraging initial results. These results empirically validate the correctness of the Jacobians $\frac{\partial g}{\partial \mathbf{x}}$, $\frac{\partial g}{\partial \mathbf{z}}$ and $\frac{\partial g}{\partial \mathbf{K}}$ of the PnP solver g obtained using implicit differentiation in Sec. 3.2.

This section is intended mainly to be illustrative; in Sec. 5, we will develop a state-of-the-art object pose estimation method based on BPnP and also report more comprehensive experiments and benchmarks.

4.1. Pose estimation

Given a known sparse 3D object structure \mathbf{z} and known camera intrinsics \mathbf{K} , a function h (a deep network, e.g., CNN, with trainable parameters θ) maps input image I to a set of 2D image coordinates \mathbf{x} corresponding to \mathbf{z} , before $g(\mathbf{x}, \mathbf{z}, \mathbf{K})$ is invoked to calculate the object pose \mathbf{y} . Our goal is to train h to accomplish this task. Since the main purpose of this section is to validate BPnP, it is sufficient to consider a “fixed input” scenario where there is only one training image I with ground truth pose \mathbf{y}^* .

Algorithm 1 describes the algorithm for this task. The loss function $l(\cdot)$ has the form

$$l(\mathbf{x}, \mathbf{y}) = \|\pi(\mathbf{z}|\mathbf{y}, \mathbf{K}) - \pi(\mathbf{z}|\mathbf{y}^*, \mathbf{K})\|_2^2 + \lambda R(\mathbf{x}, \mathbf{y}), \quad (26)$$

which is the sum of squared errors between the projection of \mathbf{z} using the ground truth pose \mathbf{y}^* and current pose \mathbf{y} from PnP (which in turn depends on \mathbf{x}), plus a regularization term

$$R(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \pi(\mathbf{z}|\mathbf{y}, \mathbf{K})\|_2^2. \quad (27)$$

The regularization ensures convergence of the estimated image coordinates \mathbf{x} to the desired positions (note that the first error component does not impose constraints on \mathbf{x}).

A main distinguishing feature of Algorithm 1 is that one of the gradient flow of ℓ is calculated w.r.t. $\mathbf{y} = g(\mathbf{x}, \mathbf{z}, \mathbf{K})$ before the gradient of \mathbf{y} is computed w.r.t. to \mathbf{x} which is then backpropagated to update θ :

$$\frac{\partial \ell}{\partial \theta} = \frac{\partial \ell}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \theta} + \frac{\partial \ell}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \theta}. \quad (28)$$

The implicit differentiation of $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ follows Eq. (20).

Figs. 1 and 2 illustrate Algorithm 1 on a synthetic example with $n = 8$ landmarks, respectively for the cases

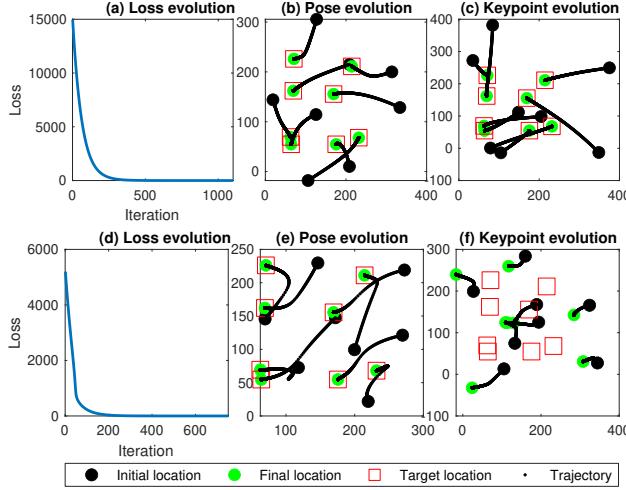


Figure 1. Sample run of Algorithm 1 on synthetic data with $n = 8$ landmarks and $h(I; \theta) = \theta$, where $\theta \in \mathbb{R}^{8 \times 2}$. The first and second row has $\lambda = 1$ and $\lambda = 0$ respectively. Left column: loss curve. Middle column: evolution of y presented as $\pi(z|y, K)$. Right column: the evolution of predicted keypoints x . Red square markers represent the target locations $\pi(z|y^*, K)$.

- $h(I; \theta) = \theta$, i.e., the parameters θ are directly output as the predicted 2D keypoint coordinates x ; and
- $h(I; \theta)$ is a modified VGG-11 [43] network that outputs the 2D keypoints x for I .

The experiments show that the loss ℓ is successfully minimized and the output pose y converges to the target pose y^* —this is a clear indication of the validity of (20).

The experiments also demonstrate the usefulness of the regularization term. While the output pose y will converge to y^* with or without $R(x, y)$, the output of h (the predicted keypoints x) can converge away from the desired positions $\pi(z|y^*, K)$ without regularization.

4.2. SfM with calibrated cameras

Let $\{\mathbf{x}^{(j)}\}_{j=1}^N$ indicate a set of 2D image features corresponding to n 3D points z associated/tracked across N frames $\{I_j\}_{j=1}^N$. Following (2), each $\mathbf{x}^{(j)}$ is a vector of 2D coordinates; however, z may not be fully observed in I_j , thus $\mathbf{x}^{(j)}$ could contain fewer than n 2D coordinates. Let

$$z^{(j)} = \mathcal{S}(z|x^{(j)}) \quad (29)$$

indicate the selection of the 3D points z that are seen in I_j . Given $\{\mathbf{x}^{(j)}\}_{j=1}^N$ and the camera intrinsics for each frame (assumed to be constant K without loss of generality), we aim to estimate the 3D structure $z \in \mathbb{R}^{3n \times 1}$ and camera poses $\{y^{(j)}\}_{j=1}^N$ corresponding to the N frames.

Our end-to-end method estimates the 3D structure

$$z = h(\mathbf{1}^\otimes; \theta) \quad (30)$$

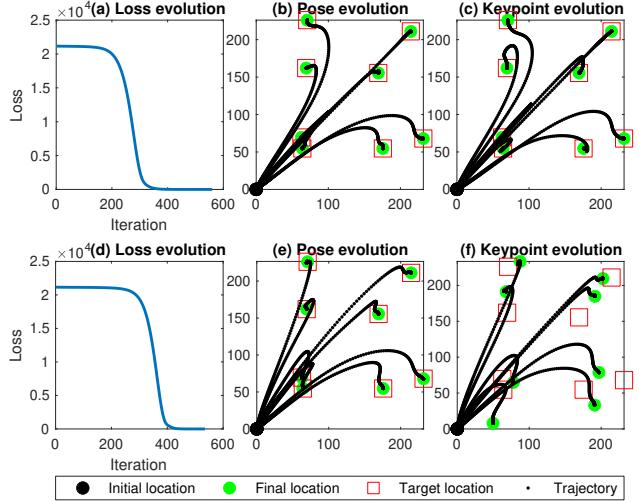


Figure 2. Same experiment as in Fig. 1 except that h is a modified VGG-11 [43] network which outputs the 2D keypoints x .

Algorithm 2 SfM with calibrated cameras.

```

1:  $\mathbf{y}^{(j)} \leftarrow$  Identity pose for  $j = 1, \dots, N$ .
2: Randomly initialize  $\theta$ 
3: while loss  $\ell$  has not converged do
4:    $\mathbf{z} \leftarrow h(\mathbf{1}^\otimes; \theta)$ .
5:    $\mathbf{z}^{(j)} \leftarrow \mathcal{S}(\mathbf{z}|\mathbf{x}^{(j)})$ , for  $j = 1, \dots, N$ 
6:    $\mathbf{y}^{(j)} \leftarrow g(\mathbf{x}^{(j)}, \mathbf{z}^{(j)}, K, \mathbf{y}^{(j)})$ , for  $j = 1, \dots, N$ .
7:    $\ell \leftarrow l(\{\mathbf{y}^{(j)}\}_{j=1}^N, \mathbf{z})$ .
8:    $\theta \leftarrow \theta - \alpha \frac{\partial \ell}{\partial \theta}$ . (Backpropagate through PnP)
9: end while

```

using a deep network h (a modified VGG-11 [43]) with the input fixed to a 1-tensor (more on this below); see Algorithm 2. Note that the algorithm makes use of the PnP subroutine to estimate each camera pose given the current z estimate. The loss function $l(\cdot)$ has the form

$$l(\{\mathbf{y}^{(j)}\}_{j=1}^N, \mathbf{z}) = \sum_{j=1}^N \|\mathbf{x}^{(j)} - \pi(z^{(j)}|\mathbf{y}^{(j)}, K)\|_2^2, \quad (31)$$

which is simply the sum of squared reprojection errors across all frames. Again, a unique feature of our pipeline is the backpropagation of the loss through the PnP solver to update network parameters θ .

$$\frac{\partial \ell}{\partial \theta} = \sum_{j=1}^N \left(\frac{\partial l}{\partial z^{(j)}} \frac{\partial z^{(j)}}{\partial \theta} + \frac{\partial l}{\partial y^{(j)}} \frac{\partial y^{(j)}}{\partial z^{(j)}} \frac{\partial z^{(j)}}{\partial \theta} \right) \quad (32)$$

The implicit differentiation of $\frac{\partial y^{(j)}}{\partial z^{(j)}}$ follows Eq. (21).

Fig. 3 illustrates the results of Algorithm 2 on a synthetic dataset with $n = 1000$ points on a 3D object seen in $N = 12$ images (about half of the 3D points are seen in each image). Starting from a random initialization of θ

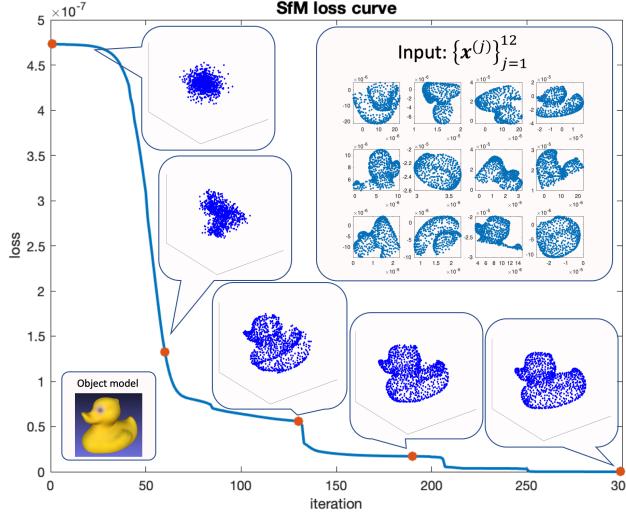


Figure 3. SfM result with Algorithm 2. The mesh of the object has $n = 1000$ points \mathbf{z}^* , which were projected to $N = 12$ different views to obtain $\{\mathbf{x}^{(j)}\}_{j=1}^N$ (about half of the 3D points are seen in each view). The function h is a modified VGG-11 network [43] which outputs the 3D structure \mathbf{z} from a fixed input of 1-tensor. We depict the output structure \mathbf{z} at various steps. A movie of this reconstruction is provided in the supplementary material.

(which leads to a poor initial \mathbf{z}), the method is able to successfully reduce the loss and recover the 3D structure and camera poses. Fig 4 shows the result from another dataset.

Effectively, our tests show that a generic deep model (VGG-11 with fixed input (30)) is able to “encode” the 3D structure \mathbf{z} of the object in the network weights, even though the network is not designed using principles from multiple view geometry. Again, our aim in this section is mainly illustrative, and Algorithm 2 is not intended to replace established SfM algorithms, e.g., [42, 41]. However, the results again indicate the correctness of the steps in Sec. 3.

4.3. Camera calibration

In the previous examples, the intrinsic matrix \mathbf{K} is assumed known and only \mathbf{x} and/or \mathbf{z} are estimated. Here in our final example, given \mathbf{x} and \mathbf{z} (2D-3D correspondences), our aim is to estimate \mathbf{K} of the form

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \quad (33)$$

where f_x and f_y define the focal length, and c_x and c_y locate the principal point of the image.

We assume $[f_x, f_y, c_x, c_y]^T \in [0, 1000]^4$. Under our BPnP approach, we train a simple neural network

$$[f_x, f_y, c_x, c_y]^T = h(\boldsymbol{\theta}) = 1000 \text{ sigmoid}(\boldsymbol{\theta}) \quad (34)$$

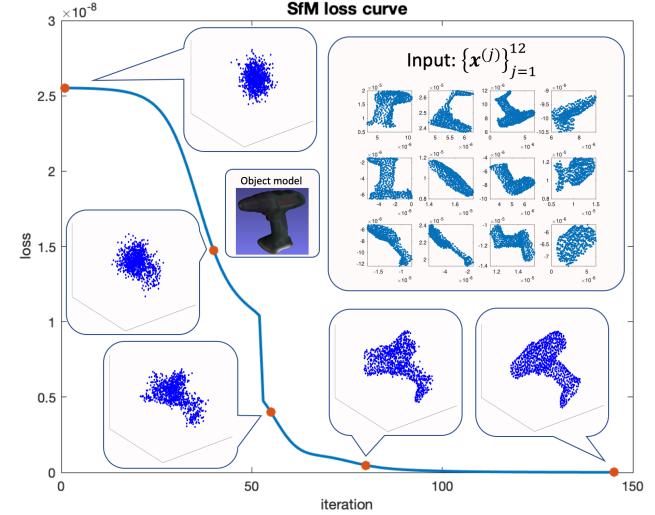


Figure 4. SfM result with a different object which has $n = 1000$ points \mathbf{z}^* . All settings are the same as in Fig. 3. A movie of this reconstruction is provided in the supplementary material.

Algorithm 3 Camera calibration.

```

1:  $\mathbf{y} \leftarrow$  Identity pose.
2: Randomly initialize  $\boldsymbol{\theta}$ 
3: while loss  $\ell$  has not converged do
4:    $[f_x, f_y, c_x, c_y]^T \leftarrow h(\boldsymbol{\theta})$ .
5:    $\mathbf{K} \leftarrow [[f_x, 0, 0]^T [0, f_y, 0]^T [c_x, c_y, 1]^T]$ .
6:    $\mathbf{y} \leftarrow g(\mathbf{x}, \mathbf{z}, \mathbf{K}, \mathbf{y})$ .
7:    $\ell \leftarrow l(\mathbf{K}, \mathbf{y})$ .
8:    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \frac{\partial \ell}{\partial \boldsymbol{\theta}}$ . (Backpropagate through PnP)
9: end while

```

to learn the parameters from correspondences \mathbf{x} and \mathbf{z} , where $\boldsymbol{\theta} \in \mathbb{R}^4$. Algorithm 3 summarizes a BPnP approach to learn the parameters $\boldsymbol{\theta}$ of h . The loss function is simply the sum of squared reprojection errors

$$l(\mathbf{K}, \mathbf{y}) = \|\mathbf{x} - \pi(\mathbf{z} | \mathbf{y}, \mathbf{K})\|_2^2, \quad (35)$$

which is backpropagated through the PnP solver via

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}} = \frac{\partial l}{\partial \mathbf{K}} \frac{\partial \mathbf{K}}{\partial \boldsymbol{\theta}} + \frac{\partial l}{\partial \mathbf{y}} \frac{\partial g}{\partial \mathbf{K}} \frac{\partial \mathbf{K}}{\partial \boldsymbol{\theta}}. \quad (36)$$

The implicit differentiation of $\frac{\partial g}{\partial \mathbf{K}}$ follows Eq. (22).

Fig. 5 illustrates the result of Algorithm 3 using the ground truth correspondences \mathbf{x}, \mathbf{z} in Fig. 1 as input. The correct intrinsic parameters are $f_x^* = 800, f_y^* = 700, c_x^* = 400, c_y^* = 300$, which the algorithm can clearly achieve as the loss converges to 0.

5. Object pose estimation with BPnP

We apply BPnP to build a model that regresses the object pose directly from an input image, not through fully con-

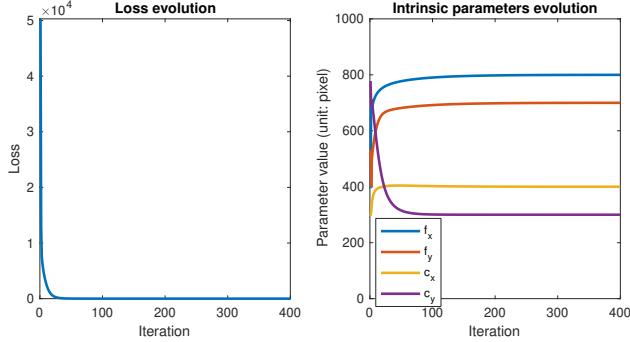


Figure 5. Camera calibration using Algorithm 3, based on the ground truth correspondences \mathbf{x}, \mathbf{z} in Fig. 1. Left: loss curve. Right: the intrinsic parameters which converge to the ground truth.

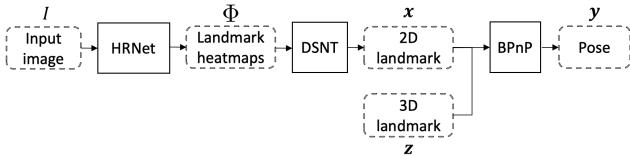


Figure 6. The pipeline of our object pose estimation network.

nected layers, but through projective geometric optimization while remaining end-to-end trainable. Our model is unique in that it simultaneously learns from feature-based loss and geometric constraints in a seamless pipeline.

The pipeline of the proposed model is depicted in Fig. 6. We use HRNet [44] as the backbone to predict the landmark heatmaps Φ . We then use the Differentiable Spatial to Numerical Transform (DSNT) [33] to convert heatmaps Φ to 2D landmark coordinates \mathbf{x} . Finally, BPnP obtains the pose \mathbf{y} from the 2D landmarks \mathbf{x} and the 3D structural landmarks of the object \mathbf{z} .

Let Φ^* denote the ground truth heatmaps constructed with the ground truth 2D landmarks \mathbf{x}^* . We define a heatmap loss

$$\ell_h = \text{MSE}(\Phi, \Phi^*), \quad (37)$$

where $\text{MSE}(\cdot, \cdot)$ is the mean squared error function; a pose loss

$$\ell_p = \|\pi(\mathbf{z} | \mathbf{y}, \mathbf{K}) - \mathbf{x}^*\|_F^2 + R(\mathbf{x}, \mathbf{y}); \quad (38)$$

and a mixture loss

$$\ell_m = \ell_h + \beta \|\pi(\mathbf{z} | \mathbf{y}, \mathbf{K}) - \mathbf{x}^*\|_F^2, \quad (39)$$

for training the model respectively. The regularization term $R(\mathbf{x}, \mathbf{y})$ is defined in Eq. (27). Note that in the mixture loss $R(\mathbf{x}, \mathbf{y})$ is unnecessary because the heatmap loss ℓ_h acts as a regularization term. We set the balancing weight β to 0.0002 in the experiments.

We apply our pipeline on the LINEMOD [20] dataset. For each object we

- obtain a 3D model representation consisting of 15 landmarks by using the Farthest Point Sampling (FPS) [36] over the original object mesh,
- randomly reserve 400 images as the test set and set the remaining (about 800, depending on the object) as the training set, and
- train a model to predict the 6DOF object pose from the input image.

We train each model with three different losses (ℓ_h , ℓ_p , and ℓ_m), for 120 epochs each. To assist convergence, when training the model with ℓ_p and ℓ_m , we first train with ℓ_h for the first 10 epochs leaving the remaining 110 epochs to train the target loss.

We evaluate our method with the following two metrics.

Average 3D distance of model points (ADD) [20] This is the percentage of accurately predicted poses in the test set. We consider a predicted pose as accurate if the average distance between the 3D model points expressed in the predicted coordinate system and that expressed in the ground truth coordinate system is less than 10% of the model diameter. For symmetric objects we use the ADD-S [53] metric instead which is based on the closest point distance.

2D projection [4]. Mean distance between 2D keypoints projected with the estimated pose and those projected with ground truth pose. An estimated pose is considered correct if this distance is less than a threshold ψ .

Table 1 summarizes the results of our experiments. In terms of the ADD(-S) metric, the model trained with ℓ_h performs considerably better than the one with ℓ_p . As expected, heatmaps can exploit richer spatial features than coordinates. However, the mixture loss achieves the highest accuracy, which suggests that heatmap loss benefits from additional correction signals from the pose loss.

In terms of the 2D projection metric, all methods perform similarly, with an average accuracy of at least 99%. To better distinguish the performances amongst different loss functions, we tighten the positive threshold ψ from the standard 5 pixels to 2 pixels. Consistent with the ADD(-S) result, the mixture loss outperformed pure heatmap loss on training an object pose estimation model. Visualization of a random subset from the test results is shown in Fig. 7.

We provide the result of the current state-of-the-art PVNet [36] as a reference. Overall, models trained with ℓ_h and ℓ_m have higher average test accuracy than PVNet, in terms of both the ADD(-S) metric and the 2D projection metric. We remind the reader to be aware of several factors while comparing the performances: we use a different backbone from PVNet; our train-test numbers of images are about 800-400 while about 20200-1000 in PVNet. Because

Model	ADD(-S)			2D projection with $\psi = 5$			2D projection with $\psi = 2$				
	Ours		PVNet	Ours		PVNet	Ours				
	ℓ_h	ℓ_p		ℓ_h	ℓ_p		ℓ_h	ℓ_p	ℓ_m		
ape	74.00	56.75	74.75	43.62	99.50	99.50	99.50	99.23	90.00	86.75	93.75
benchwise	98.50	98.00	99.00	99.90	99.25	98.75	99.25	99.81	86.00	82.75	86.00
cam	96.25	83.75	96.25	86.86	98.75	98.50	99.25	99.21	91.50	81.50	90.50
can	97.00	94.75	98.00	95.47	99.50	99.50	99.75	99.90	93.25	89.00	92.75
cat	93.00	85.25	94.25	79.34	99.50	99.50	99.50	99.30	96.75	95.25	96.75
driller	98.50	98.00	99.25	96.43	99.00	98.50	98.50	96.92	83.75	81.50	84.50
duck	76.25	49.25	78.50	52.58	99.00	99.25	99.00	98.02	88.50	84.00	91.50
eggbox	95.75	93.25	96.50	99.15	99.50	99.25	99.50	99.34	92.75	93.25	92.50
glue	87.50	76.25	90.00	95.66	99.50	99.50	99.50	98.45	93.50	90.00	94.00
holepuncher	89.50	80.25	91.50	81.92	99.75	99.50	99.75	100.00	92.25	90.25	91.75
iron	97.75	96.50	97.75	98.88	98.50	98.25	98.50	99.18	86.75	79.75	87.25
lamp	99.75	98.50	99.75	99.33	98.75	98.00	98.50	98.27	85.00	83.25	86.75
phone	95.75	96.00	97.00	92.41	99.25	99.00	99.25	99.42	91.50	88.00	92.25
average	92.27	85.12	93.27	86.27	99.21	99.00	99.21	99.00	90.12	86.56	90.79

Table 1. Test accuracy on the LINEMOD dataset in terms of the ADD(-S) metric (columns 2-5) and the 2D projection metric with $\psi = 5$ pixels (columns 6-9) and $\psi = 2$ pixels (columns 10-12). Objects eggbox and glue are considered as symmetric objects and the ADD-S metric is used.

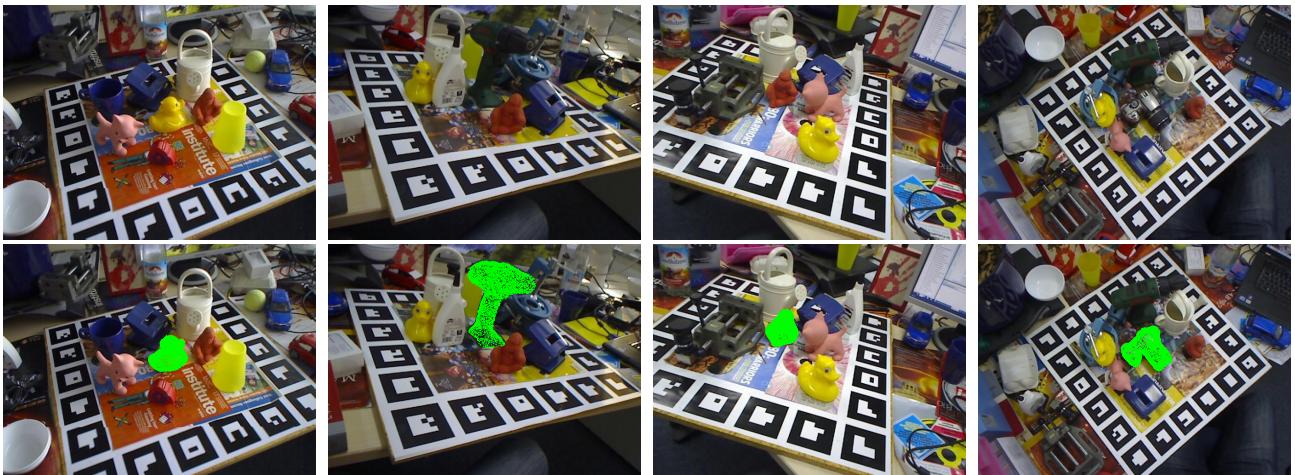


Figure 7. Random sample of test results of the proposed model trained with the mixture loss ℓ_m . The first row are the test images for predicting the pose of the central object. The second row shows the regressed alignment for each query object.

we require the ground truth pose label for training, we did not use any data augmentation such as cropping, rotation or affine transformation.

6. Conclusions

We present BPnP, a novel approach for performing back-propagation through a PnP solver. BPnP leverages on implicit differentiation to address computing the non-explicit gradient of this geometric optimization process. We validate our approach in three fundamental geometric optimization problems (pose estimation, structure from motion, and camera calibration). Furthermore, we developed an end-to-end trainable object pose estimation pipeline with BPnP,

which outperforms the current state-of-the-art. Our experiments show that exploiting 2D-3D geometry constraints improves the performance of a feature-based training scheme.

The proposed BPnP opens a door to vast possibilities for designing new models. We believe the ability to incorporate geometric optimization in end-to-end pipelines will further boost the learning power and promote innovations in various computer vision tasks.

Acknowledgement

This work was supported by ARC LP160100495 and the Australian Institute for Machine Learning. Nan Li is sponsored by NSF China (11601378) and Tencent-SZU Fund.

References

- [1] Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *ICML*, 2017. 1, 2
- [2] K. G. Binmore and J. Davies. *Calculus*. Cambridge University Press, 1983. 1
- [3] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *CVPR*, 2017. 1, 2
- [4] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In *CVPR*, 2016. 7
- [5] Eric Brachmann and Carsten Rother. Expert sample consensus applied to camera re-localization. In *ICCV*, 2019. 1, 2
- [6] Eric Brachmann and Carsten Rother. Neural-guided ransac: Learning where to sample model hypotheses. In *ICCV*, 2019. 1, 2
- [7] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 4
- [8] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *CVPR*, 2018. 1, 2
- [9] Ming Cai, Chunhua Shen, and Ian D Reid. A hybrid probabilistic model for camera relocalization. In *BMVC*, 2018. 1
- [10] Bo Chen, Jiewei Cao, Alvaro Parra, and Tat-Jun Chin. Satellite pose estimation with deep landmark regression and non-linear pose refinement. In *ICCVW*, 2019. 2
- [11] Ronald Clark, Michael Bloesch, Jan Czarnowski, Stefan Leutenegger, and Andrew J. Davison. Learning to solve non-linear least squares for monocular stereo. In *ECCV*, 2018. 2
- [12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014. 2
- [13] A. Eriksson and A. van den Hengel. Efficient computation of robust low-rank matrix approximations in the presence of missing data using the l_1 norm. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 2
- [14] O. Faugeras. *Three-dimensional computer vision*. MIT Press, 1993. 2
- [15] 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*, 24(6):381–395, 1981. 3
- [16] Stephen Gould, Basura Fernando, Anoop Cherian, Peter Anderson, Rodrigo Santa Cruz, and Edison Guo. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *arXiv preprint arXiv:1607.05447*, 2016. 1, 2
- [17] Ankur Handa, Michael Blösch, Viorica Patrascu, Simon Stent, John McCormac, and Andrew J. Davison. gvnn: Neural network library for geometric computer vision. In *ECVW*, 2016. 2
- [18] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1
- [19] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1
- [20] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *ACCV*, 2012. 7
- [21] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *ICRA*, 2016. 1, 2
- [22] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *CVPR*, 2017. 1, 2
- [23] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*, 2015. 1, 2
- [24] Alex Kendall, Hayk Martirosyan, Saumitra Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, 2017. 2
- [25] Steven G. Krantz and Harold R. Parks. *The Implicit Function Theorem, History, Theory, and Applications*. Birkhäuser, New York, NY, 2012. 3
- [26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012. 1
- [27] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *CVPR*, 2014. 2
- [28] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016. 2
- [29] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 1
- [30] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, 2015. 2
- [31] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian D. Reid. Learning depth from single monocular images using deep convolutional neural fields. *TPAMI*, 38(10):2024–2039, 2016. 2
- [32] Tayyab Naseer and Wolfram Burgard. Deep regression for monocular camera-based 6-dof global localization in outdoor environments. In *IROS*, 2017. 1, 2
- [33] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*, 2018. 7
- [34] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS*, 2017. 4
- [35] Georgios Pavlakos, Xiaowei Zhou, Aaron Chan, Konstantinos G Derpanis, and Kostas Daniilidis. 6-dof object pose from semantic keypoints. In *ICRA*, 2017. 2

- [36] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Jun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, 2019. 2, 7
- [37] Mahdi Rad and Vincent Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *ICCV*, 2017. 2
- [38] Clarence Hudson Richardson. *An introduction to the calculus of finite differences*. Van Nostrand, 1954. 1
- [39] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. Understanding the limitations of cnn-based absolute camera pose regression. In *CVPR*, 2019. 1, 2
- [40] U. Schmidt and S. Roth. Shrinkage fields for effective image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2
- [41] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *CVPR*, 2016. 6
- [42] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *ECCV*, 2016. 6
- [43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5, 6
- [44] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 7
- [45] Supasorn Suwajanakorn, Noah Snavely, Jonathan J Tompson, and Mohammad Norouzi. Discovery of latent 3d key-points via end-to-end geometric reasoning. In *NIPS*, 2018. 1
- [46] M. F. Tappen, C. Liu, E. H. Adelson, and W. T. Freeman. Learning Gaussian conditional random fields for low-level vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 2
- [47] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *CVPR*, 2018. 2
- [48] James Thewlis, Shuai Zheng, Philip HS Torr, and Andrea Vedaldi. Fully-trainable deep matching. In *BMVC*, 2016. 1
- [49] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Niklaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *CVPR*, 2017. 2
- [50] Florian Walch, Caner Hazirbas, Laura Leal-Taixe, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using lstms for structured feature correlation. In *ICCV*, 2017. 1, 2
- [51] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L. Yuille. Towards unified depth and semantic prediction from a single image. In *CVPR*, 2015. 2
- [52] Jian Wu, Liwei Ma, and Xiaolin Hu. Delving deeper into convolutional neural networks for camera relocalization. In *ICRA*, 2017. 1, 2
- [53] 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 (RSS)*, 2018. 7
- [54] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *CVPR*, 2017. 2
- [55] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *ECCV*, 2016. 1
- [56] Zhengyou Zhang. Parameter estimation techniques: a tutorial with application to conic fitting. *Image and Vision Computing*, 15(1):59 – 76, 1997. 3
- [57] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 2