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”

\[
y = g(x, z, K),
\]

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

\[
x = [x_1^T \ x_2^T \ \ldots \ x_n^T]^T \in \mathbb{R}^{2n \times 1},
\]

\[
z = [z_1^T \ z_2^T \ \ldots \ z_n^T]^T \in \mathbb{R}^{3n \times 1},
\]

8101
where \((x_i, z_i)\) is the \(i\)-th correspondence. Let \(\pi(\cdot; y, K)\) be a projective transformation of 3D points onto the image plane with pose \(y\) and camera intrinsics \(K\). Intrinsically, the “evaluation” of \(g\) requires solving the optimization problem

\[
y = \arg\min_{y \in SE(3)} \sum_{i=1}^{n} \|r_i\|_2^2,
\]

where

\[
r_i = x_i - \pi_i
\]
is the reprojection error of the \(i\)-th correspondence and

\[
\pi_i = \pi(z_i; y, K)
\]
is the projection of 3D point \(z_i\) on the image plane. We introduce the shorthand

\[
\pi := [\pi_1^T, \ldots, \pi_n^T]^T,
\]
thus (4) can be rewritten as

\[
y = \arg\min_{y \in SE(3)} \|x - \pi\|_2^2.
\]
The choice of formulation (8) will be justified in Sec. 3.3.

Our ultimate goal is to incorporate \(g\) in a learnable model, where \(x\), \(z\) and \(K\) 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 \(f : \mathbb{R}^{n+m} \to \mathbb{R}^m\) be a continuously differentiable function with input \((a, b) \in \mathbb{R}^n \times \mathbb{R}^m\). If a point \((a^*, b^*)\) satisfies

\[
f(a^*, b^*) = 0
\]
and the Jacobian matrix \(\frac{\partial f}{\partial \xi}(a^*, b^*)\) is invertible, then there exists an open set \(U \subset \mathbb{R}^n\) such that \(a^* \in U\) and an unique continuously differentiable function \(g(a) : \mathbb{R}^n \to \mathbb{R}^m\) such that \(b^* = g(a^*)\) and

\[
f(a', g(a')) = 0, \forall a' \in U.
\]

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

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

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

### 3.2. Constructing the constraint function \(f\)

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

\[
o(x, y, z, K) = \sum_{i=1}^{n} \|r_i\|_2^2.
\]

Since the output pose \(y\) 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 \(y\), i.e.,

\[
\frac{\partial o(x, y, z, K)}{\partial y} \bigg|_{y=g(x, z, K)} = 0.
\]

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

\[
f(x, y, z, K) = [f_1, \ldots, f_m]^T,
\]
where for all \(j \in \{1, \ldots, m\},

\[
f_j = \frac{\partial o(x, y, z, K)}{\partial y_j}
\]
\[
= 2 \sum_{i=1}^{n} \langle r_i, \frac{\partial r_i}{\partial y_j} \rangle
\]
\[
= \sum_{i=1}^{n} \langle r_i, c_{ij} \rangle
\]
with

\[
c_{ij} = -2 \frac{\partial \pi_i}{\partial y_j}.
\]

### 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 \(x\), \(z\) and \(K\)). 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 \(x\), \(z\) and \(K\).
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 \texttt{SOLVEPNP\_ITERATIVE} method in OpenCV [7]), which guarantees local convergence. As an iterative algorithm, LM requires initialization \( y^{(0)} \) in solving (8). We make explicit this dependence by rewriting (1) as

\[
y = g(x, z, K, y^{(0)}). \tag{19}
\]

We obtain the initial pose \( y^{(0)} \) 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

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

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

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

Given the output gradient \( \nabla y \), BPnP returns the input gradients

\[
\nabla x = \left[ \frac{\partial g}{\partial x} \right]^T \nabla y, \tag{23}
\]

\[
\nabla z = \left[ \frac{\partial g}{\partial z} \right]^T \nabla y, \tag{24}
\]

\[
\nabla K = \left[ \frac{\partial g}{\partial K} \right]^T \nabla y. \tag{25}
\]

3.4. Implementation notes

The number of dimensions of \( y \), 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, SFM and camera calibration, and report encouraging initial results. These results empirically validate the correctness of the Jacobians \( \frac{\partial g}{\partial x}, \frac{\partial g}{\partial z} \) and \( \frac{\partial g}{\partial K} \) 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 \( z \) and known camera intrinsics \( K \), a function \( h \) (a deep network, e.g., CNN, with trainable parameters \( \theta \)) maps input image \( I \) to a set of 2D image coordinates \( x \) corresponding to \( z \), before \( g(x, z, K) \) is invoked to calculate the object pose \( y \). 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 \( y^* \).

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

\[
l(x, y) = \| \pi(z|y, K) - \pi(z|y^*, K) \|^2_2 + \lambda R(x, y), \tag{26}
\]

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

\[
R(x, y) = \| x - \pi(z|y, K) \|^2_2. \tag{27}
\]

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

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

\[
\frac{\partial \ell}{\partial \theta} = \frac{\partial l}{\partial y} \frac{\partial g}{\partial x} \frac{\partial \theta}{\partial h} + \frac{\partial l}{\partial h} \frac{\partial \theta}{\partial \theta}. \tag{28}
\]

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

Figs. 1 and 2 illustrate Algorithm 1 on a synthetic example with \( n = 8 \) landmarks, respectively for the cases
4.2. SfM with calibrated cameras

Let \( \{x^{(j)}\}_{j=1}^{N} \) 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 \( x^{(j)} \) is a vector of 2D coordinates; however, \( z \) may not be fully observed in \( I_j \), thus \( x^{(j)} \) could contain fewer than \( n \) 2D coordinates. Let

\[
z^{(j)} = S(z|x^{(j)})
\]

(29)

indicate the selection of the 3D points \( z \) that are seen in \( I_j \). Given \( \{x^{(j)}\}_{j=1}^{N} \) 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

\[
z = h(1^\otimes; \theta)
\]

(30)

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

\[
l(\{y^{(j)}\}_{j=1}^{N}, z) = \sum_{j=1}^{N} \| x^{(j)} - \pi(z^{(j)}|y^{(j)}, K) \|_2^2,
\]

(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 \( \theta \).

\[
\frac{\partial l}{\partial \theta} = \sum_{j=1}^{N} \left( \frac{\partial l}{\partial z^{(j)}} \frac{\partial z^{(j)}}{\partial \theta} + \frac{\partial l}{\partial y^{(j)}} \frac{\partial y^{(j)}}{\partial z^{(j)}} \frac{\partial z^{(j)}}{\partial \theta} \right)
\]

(32)

The implicit differentiation of \( \frac{\partial l}{\partial \theta} \) 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 \( \theta \)
which leads to a poor initial $z$), 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 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 $K$ is assumed known and only $x$ and/or $z$ are estimated. Here in our final example, given $x$ and $z$ (2D-3D correspondences), our aim is to estimate $K$ of the form

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

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

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

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

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

which is backpropagated through the PnP solver via

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

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

Fig. 5 illustrates the result of Algorithm 3 using the ground truth correspondences $x$, $z$ 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-
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 (\(\ell_h\), \(\ell_p\), and \(\ell_m\)), for 120 epochs each. To assist convergence, when training the model with \(\ell_p\) and \(\ell_m\), we first train with \(\ell_h\) 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 \(\psi\).

Table 1 summarizes the results of our experiments. In terms of the ADD(-S) metric, the model trained with \(\ell_h\) performs considerably better than the one with \(\ell_p\). 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 \(\psi\) 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 \(\ell_h\) and \(\ell_m\) 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...
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.

6. Conclusions

We present BPnP, a novel approach for performing backpropagation 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


[38] Clarence Hudson Richardson. An introduction to the calculus of finite differences. Van Nostrand, 1954. 4