CHALMERS
UNIVERSITY OF TECHNOLOGY

Pose Proposal Critic: Robust Pose Refinement by Learning Reprojection Errors

Pose Proposal Critic: Robust Pose Refinement by Learning Reprojection Errors

Lucas Brynte
brynte@chalmers.se
Fredrik Kahl
fredrik.kahl@chalmers.se
Chalmers University of Technology
Gothenburg, SWEDEN

Abstract

In recent years, considerable progress has been made for the task of rigid object pose estimation from a single RGB-image, but achieving robustness to partial occlusions remains a challenging problem. Pose refinement via rendering has shown promise in order to achieve improved results, in particular, when data is scarce.

In this paper we focus our attention on pose refinement, and show how to push the state-of-the-art further in the case of partial occlusions. The proposed pose refinement method leverages on a simplified learning task, where a CNN is trained to estimate the reprojection error between an *observed* and a *rendered* image. We experiment by training on purely synthetic data as well as a mixture of synthetic and real data. Current state-of-the-art results are outperformed for two out of three metrics on the Occlusion LINEMOD benchmark, while performing on-par for the final metric.

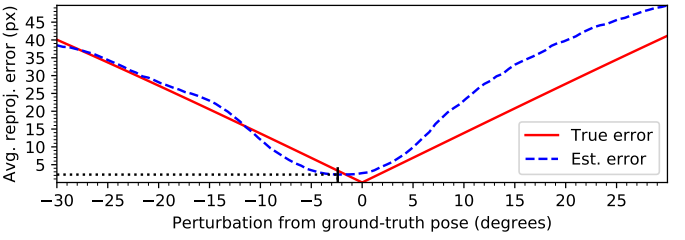
1 Introduction

Accurately estimating the 3D location and orientation of an object from a single image, a.k.a. rigid object pose estimation, has many important real world applications, such as robotic manipulation, augmented reality and autonomous driving. Although the problem has commonly been addressed by exploiting RGB-D cameras, e.g., [1], this introduces an increased cost of hardware, sensitivity to sunlight, and unreliable / missing depth measurements for reflective / transparent objects. In recent years, more attention has been put to RGB-only pose estimation, and although considerable progress has been made, a major challenge remains in achieving robustness to partial occlusions. To this end, rendering-based pose refinement methods have shown promise in order to achieve improved results, but their full potential remains unexplored.

In this paper we revisit pose refinement via rendering, and focus specifically on how to further improve on the robustness of such methods, in particular with respect to partial occlusions. Our method can hence be used to refine the estimates of any pose algorithm and we will give several experimental demonstrations that this is indeed achieved for different algorithms. Naturally, we will also compare to other refinement methods.

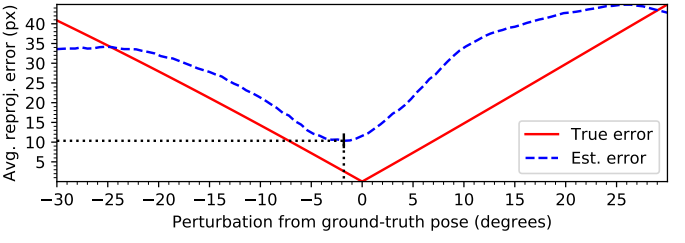
Shared among contemporary rendering-based pose refinement methods [14, 16, 25] is the approach of feeding an observed as well as a synthetically rendered image as input to a CNN model, which is trained to predict the relative pose of an object between the two images. Our

(a)



(b)

(c)



(d)

Figure 1: Estimated average reprojection error of our network for the *cat* object in two test frames of Occlusion LINEMOD [1]. A rotational perturbation is applied for a fixed axis in the camera coordinate frame, in the range of $[-30, 30]$ degrees. The estimated minimum is marked in the figure. (a-b) An unoccluded example. (c-d) An example with occlusion.

key insight is that rendering-based pose refinement is possible without explicitly regressing to the parameter vector of the relative pose. Instead, estimating an error function of the relative pose is enough, since minimization of said error function w.r.t. pose can be done during inference. Figure 1 shows error function estimates for two test frames. Although a larger bias is observed in the occluded case, the estimated minimum is still close to ground-truth. Our main contributions can be summarized as follows:

- A novel pose refinement method, which works well without real training data.
- Robustness to partial occlusions, by the implicit nature of our method, making it insensitive to over- or under-estimations of the error function.
- State-of-the-art results for two out of three metrics on Occlusion LINEMOD [1].

Our pose refinement pipeline takes as input (a) an object CAD model and (b) an initial pose estimate, referred to as a "pose proposal". The pose proposal is assumed to be obtained from another method, and is fed to the refinement pipeline, consisting of three parts:

1. Synthetic rendering of the detected object under the pose proposal.
2. Estimation of the average reprojection error of all model points, when projected into the image using the ground truth pose as well as the pose proposal.
3. Iterative refinement of the 6D pose estimate by minimizing the reprojection error.

We will refer to our method as Pose Proposal Critic (PPC), since the heart of the method involves judging the quality of a pose proposal.

2 Related Work

Similarly to the work of Kendall *et al*. for camera localization [12], the rigid object pose estimation methods of Xiang *et al*. [24] and Do *et al*. [2] estimate the pose by directly regressing the pose parameters. The most successful methods for rigid object pose estimation do however make use of a two-stage pipeline, where 2D-3D correspondences are first established, and the object pose is then retrieved by solving the corresponding camera resectioning problem [4]. A common approach has been to regress 2D locations of a discrete set of object keypoints, projected into the image, yielding a sparse set of correspondences [11, 21, 22]. Other methods instead output heatmaps in order to encode said keypoint locations [17, 19].

Among the sparse correspondence methods, Oberweger *et al*. [17] stand out in that they address the problem of partial occlusions very carefully. They show that occluders typically have a corrupting effect on CNN activations, far beyond the occluded region itself, and that training with occluded samples might not help to overcome this problem. Instead they resort to limiting the receptive field of their keypoint detector, as a crude but effective way to suppress the impact of occluders.

Dense correspondence methods on the other hand, are inherently more robust to noise in correspondence estimates. Pixel-wise regression on the corresponding (object frame) 3D coordinates has been proposed by [7, 15, 18], while Zakharov *et al*. [25] take a different approach and discretize the object surface into smaller segments, and then perform classification on which segment is visible in which pixel. Hu *et al*. [8] use a sparse set of keypoints, but yet leverage on dense correspondences due to redundantly regressing a number of 2D locations for each object keypoint. Peng *et al*. [20] take a similar approach, but simplify the output space by regressing to the direction from each pixel to each of the projected keypoints, but not the corresponding distance. Pair-wise sampling of pixel-wise predictions then yields votes for keypoint locations.

The method of Peng *et al*. [20] does indeed prove robust to partial occlusions, and yields accurate estimates of rotation as well as lateral translation on the Occlusion LINEMOD benchmark. Consequently, the results are state-of-the-art for the depth-insensitive metric based on reprojection errors. Nevertheless, their depth estimates are still not accurate, and suffer in the presence of partial occlusion. The rendering-based pose refinement method of DeepIM (Li *et al*. [14]) does however perform well on Occlusion LINEMOD for all common metrics, and in particular gives a huge boost in depth estimation accuracy, yielding state-of-the-art results for the metric based on matching point clouds in 3D, and suggesting that rendering-based pose refinement is a powerful tool for accurate pose estimation in the presence of partial occlusion. We will experimentally compare to DeepIM and show how one can achieve significantly improved results for partial occlusions.

Moreover, we point out that while a multitude of approaches for increasing robustness, especially to partial occlusions, has been observed among correspondence-based pose estimation methods, we have not yet seen any directed efforts to address these issues in the literature of rendering-based pose refinement.

When it comes to rendering-based pose refinement, early work was done by Tjaden and Schömer [23], proposing a segmentation pipeline based on hand-crafted features, and iterative alignment of silhouettes. Rad and Lepetit [21] also apply rendering-based refinement as part of the BB8 pose estimation pipeline, improving on initial estimates. BB8 is based on sparse correspondences (8 bounding box corners), and a refinement CNN is trained to regress the reprojection errors for each of the bounding box corners. Refinement is then carried out on the correspondences themselves, yielding an updated camera resectioning problem to be

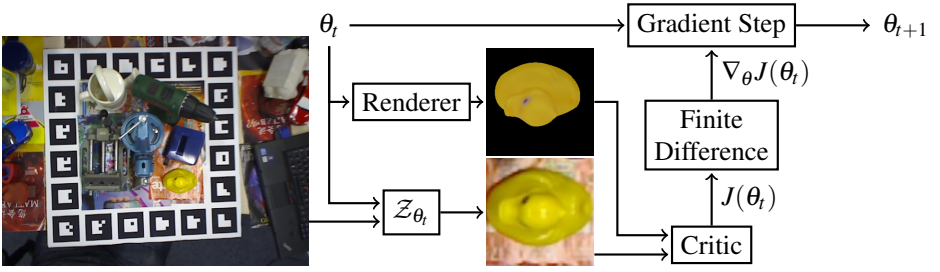


Figure 2: Illustration of the pose refinement pipeline. Given a pose proposal θ_t, a synthetic image is rendered, while the zoom-in operator $\mathcal{Z}_{\theta_t}$ zooms in on a corresponding image patch in the observed image. A neural network, pretrained for optical flow, acts as a critic by comparing both image patches and estimating the average reprojection error. Finally, finite differences approximate the gradient, and a gradient step yields a new estimate θ_{t+1}.

solved. In contrast, our method instead estimates the average reprojection error over all model points and refinement is done directly on the pose.

Manhardt *et al*. [16], Li *et al*. [14] and Zakharov *et al*. [25], all propose a CNN-based refinement pipeline, where the model is trained to learn the relative pose between an observed image and a synthetically rendered image under a pose proposal. The main difference between their approaches and ours is that we instead choose to learn an error function of the relative pose. Among these methods, [14] is the only one that handles partial occlusions well. The results of [25] seem competitive at a first glance, but evaluation is only carried out on the frames for which the 2D object detector successfully detected the object of interest, and furthermore parts of the Occlusion LINEMOD dataset were used for training, which does not allow for a fair comparison.

3 Method

In this section, the three main parts of our pose refinement method will be described in detail.

The core idea of our approach is that even though neural networks have an amazing capacity to learn difficult estimation tasks, the learning problem should be kept as simple as possible. Given a pose proposal, the task of our network is to determine how good the proposal is with respect to the ground truth. So, instead of learning the pose parameters directly, it is only required for the network to act as a critic of different proposals. To further simplify the task, we render a synthetic image using the pose proposal, and then the network only needs to determine if the *rendered* image is similar to the *observed* image or not. As a measure of similarity, we use the average reprojection error of object CAD model points. Then, at inference, the objective is to find the pose parameters with lowest predicted reprojection error, resulting in a minimization problem which can be solved with standard optimization techniques. For an overview of the pipeline, see Figure 2.

It is assumed that intrinsic camera parameters are known for the *observed* images, and that a three-dimensional CAD model of the object of interest is available.

3.1 Part I: Rendering the Object Under a Pose Proposal

Similar to previous work [14, 16, 21, 25], we render a synthetic image of a detected object based on the suggested pose proposal. Rendering is done on the GPU using OpenGL with Lambertian shading and the light source at the camera center. The background is kept black.

Rather than using all of the *observed* image directly, we zoom in on the detected object. The zoom-in operation is dynamic and controlled by the current pose proposal, yielding square image patches centered at the projection of the object center. The size of the corresponding image patches (*observed* and *rendered*) is chosen as 1.2 times the projection of the object diameter, and furthermore, the *observed* patch is bilinearly upsampled to 512×512 pixels.

For future reference, let $\mathcal{Z}_\theta$ denote the zoom-in operator for pose proposal θ, acting on an *observed* image, I_{obs}, resulting in an image patch $P_{obs} = \mathcal{Z}_\theta I_{obs}$. We will denote the *rendered* image patch by P_{rend}. For performance reasons, the patch is rendered at 256×256 resolution, and then bilinearly upsampled to 512×512.

3.2 Part II: Learning Average Reprojection Error

We train a CNN model to take the *observed* and *rendered* image patches as input, and output an estimate of the average reprojection error, i.e., the average image distance between the projected CAD model points using the ground truth and the pose proposal, respectively.

Let f be a neural network, and $f(\mathcal{Z}_\theta I_{obs}, P_{rend}(\theta))$ be its the reprojection error estimate between the *observed* image patch $\mathcal{Z}_\theta I_{obs}$ and the *rendered* image patch $P_{rend}(\theta)$, for pose proposal θ. If $\mathcal{P}_\theta$ denotes the projection of a 3D point onto the image patch using pose θ[1], then the reprojection error to be estimated is given by

$$\frac{1}{M}\sum_{i=1}^{M}\left|\left|\mathcal{P}_{\hat{\theta}}(R_{\hat{\theta}}p_i + t_{\hat{\theta}}) - \mathcal{P}_{\hat{\theta}}(R_{\theta^*}p_i + t_{\theta^*})\right|\right|_2, \tag{1}$$

where $\hat{\theta}$ and θ^* are the estimated and true poses, respectively, and p_i are the M object model points. Figure 1 shows the estimated error function for two test frames of Occlusion LINEMOD, one in which the object is partially occluded.

The reprojection error is measured in image patch pixels, i.e. after zoom-in rather than before. Estimating the reprojection error before zoom-in would require the network to estimate (and rescale with) the absolute depth, which would introduce an unnecessary complication. The reason we choose the reprojection error is that we expect it to be relatively easy to infer from image pairs without a lot of high level reasoning, and thus providing a relatively easy learning task. Furthermore, the reprojection error is quite related to optical flow, and should fit particularly well with a pretrained optical flow backbone.

3.2.1 Network Architecture and Training Details

Like [14] we use a pretrained FlowNetSimple optical flow network as backbone (the FlowNet 2.0 version from Ilg *et al.* [9], rather than the original model from Dosovitskiy *et al.* [3]). The encoder output feature maps are flattened and fed through three fully-connected layers, constituting our main branch. In contrast to [14], standard ReLU (rather than leaky-ReLU)

[1] Note that the projection operator itself depends on the pose, due to the dependence of the zoomed-in image patch, and thus the effective intrinsic camera parameters, on the estimated object position.

activation functions are used. Each hidden layer has 1024 neurons, where dropout is applied with 30 % probability. A single output neuron then represents the reprojection error estimate.

We also follow [14] in adding an auxiliary branch for foreground / background segmentation, through an extra 1-channel 3×3 convolutional layer next to the optical flow prediction at level 4 (a.k.a. `flow4`). The optical flow prediction itself is however disregarded, since [14] showed only a minor boost from including this auxiliary task. Finally, unlike [14], no segmentation is provided as network input, since it proved unnecessary. Re-training their network without segmentation input did not cause any performance drop.

We train on the loss function $\mathcal{L} = 0.01 \cdot \mathcal{L}_{reproj} + 0.3 \cdot \mathcal{L}_{seg}$, where $\mathcal{L}_{reproj}$ is the L_1 error between the true and estimated average reprojection error, and $\mathcal{L}_{seg}$ is the binary cross-entropy loss of the foreground / background segmentation, averaged over all pixels. Furthermore, the target for average reprojection error is saturated at 50 px, making sure that very large perturbation samples will not introduce a disturbance into the training, letting the network focus on achieving high precision within the range of reasonable perturbations.

One model is trained for each object, for 75 epochs with 42 batches sampled in each epoch. The learning rate is initialized to $5 \cdot 10^{-5}$, and multiplied by 0.3 every 10 epochs. The batch size is 12 and L_2 regularization is applied with weight decay $5 \cdot 10^{-4}$.

For further details on the implementation, we refer to the supplementary material. How to sample pose proposals during training is covered in Section A.1, while Section A.2 describes data augmentation strategies, and in particular how to generate synthetic training examples.

3.3 Part III: Minimizing Reprojection Error

Once the network f is trained for estimating reprojection errors, we may apply it for refining an initial pose proposal θ_0 of our object of interest in the *observed* image I_{obs}.

Let the compound function $J(\theta) = f(\mathcal{Z}_\theta I_{obs}, P_{rend}(\theta))$ encapsulate the operations of rendering, zoom-in and the CNN itself, which leads to the optimization problem: $\min_\theta J(\theta)$. We minimize J locally, initializing at θ_0. Gradient-based optimization is carried out and although analytical differentiation is a tempting approach in the light of differentiable renderers such as [10], we observed noisy behavior in J, and we instead apply numerical differentiation for robustly estimating $\nabla_\theta J$.

For parameterizing the rotation, we take advantage of the Lie Algebra of SO(3). The initial rotation R_0 is used as a reference point and the parameterization is $R(\theta_r) = e^{A(\theta_r)} R_0$, where the parameters θ_r constitute the three elements of the 3×3 skew-symmetric matrix $A(\theta_r)$. The translation is split into two parts. The lateral translation θ_l represents the deviation from the projection of the initial position in pixels, i.e. $\mathcal{P}_{\theta_0}(t) - \mathcal{P}_{\theta_0}(t_0)$, where θ_0 denotes the initial pose proposal, and t_0 denotes the translation part specifically. The depth is parameterized as $d(\theta_d) = e^{\theta_d} d_0$, where d_0 is the initial depth estimate.

3.3.1 Optimization Scheme

It needs to be stressed that there may be spurious local minima in J, and for this reason care should be taken during the optimization. For better control over the procedure, we let decoupled optimizers run in parallel for the rotation / depth / lateral translation parameters, with different hyperparameters and step size decay schedules. We apply in total 100 iterations.

In order to handle non-convex and noisy behavior of J, we use stochastic optimization. In particular, the Adam optimizer [13] proved effective for handling the fact that J may be quite steep in the vicinity of the optimum, yet quite flat farther away from the optimum. This

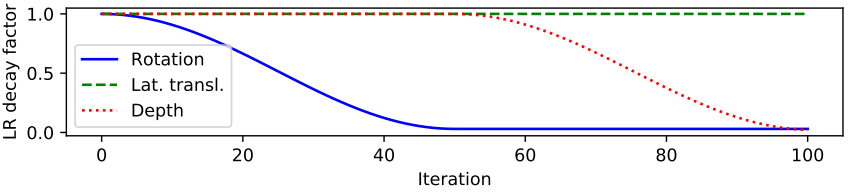


Figure 3: Step size decay schedules for the different parameters θ_r, θ_l and θ_d during inference. The decay value is relative to the respective initial step sizes.

property does otherwise risk the optimizer taking too far steps when encountering "steep" points or easily getting stuck in local minima, if the step size is too high or low, respectively.

Consider for now the optimization w.r.t. rotation and depth. The optimization is roughly carried out in two phases, first w.r.t. rotation and then depth, with a smooth transition between the two. This sequential strategy is due to two reasons: (1) Although optimization w.r.t. rotation works well despite a sub-optimal depth estimate, keeping θ_d fixed reduces noise for the moment estimates of the optimizer. (2) For precise depth estimation, a good estimate of the other parameters is crucial. The reprojection error to be estimated, as well as the *rendered* image itself, is much less sensitive to depth perturbations than to the other pose parameters, and focusing specifically on θ_d in the final stage helps to improve depth estimation.

Optimization w.r.t. lateral translation proves relatively easy, and less coupled with the other parameters, i.e. a reasonable minimum may be found despite e.g. a poor rotation estimate. Particularly fast convergence of the lateral translation is desirable for the converse reason, that optimization w.r.t. the other parameters is coupled with the lateral translation estimate, and may not work well unless this is adequate. Luckily, convergence of θ_l is achieved in just a couple of iterations when using a plain SGD optimizer with momentum 0.5 rather than Adam. The step size w.r.t. θ_l is set constantly to 1.

The step size decay schedule for all parameters is illustrated in Figure 3 and the exponential decay rates for the moment estimation of Adam were set to $(\beta_1, \beta_2) = (0.6, 0.9)$ for θ_r, and $(\beta_1, \beta_2) = (0.4, 0.9)$ for θ_d. The step sizes used for finite differences were 0.01, 1.0 and 0.005 for θ_r, θ_l, and θ_d, respectively.

4 Experiments

4.1 Datasets and Training Data

Experiments are carried out on LINEMOD as well as Occlusion LINEMOD.

LINEMOD is a standard benchmark for rigid object pose estimation and was introduced by Hinterstoisser *et al.* [5]. The dataset consists of 15 object CAD models along with 15 RGB-D image sequences of an indoor scene where objects are laid out on a table with cluttered background. For each sequence there is a corresponding object of interest put at the center. Although depth images are provided, it is also a common benchmark for RGB-only pose estimation. As two of the objects suffer from low quality CAD models, they are commonly excluded from evaluation and we follow the same practice.

The Occlusion LINEMOD dataset was produced by Brachmann *et al.* [1] by taking one of the LINEMOD sequences and annotating the pose of the surrounding 8 objects. While the

central object is typically unoccluded, the surrounding objects are often partially occluded, resulting in a challenging dataset. The central object is not part of the benchmark.

For experiments on Occlusion LINEMOD we train on real images from LINEMOD, as has conventionally been done in the literature, while the Occlusion LINEMOD images are only used for testing. With 33 % probability we sample a real training image, and with 67 % probability a synthetic one. In the synthetic case, there is a 50 % probability that 2 occluding objects are rendered, and a 50 % probability that no occluders are rendered. We carry out additional experiments on Occlusion LINEMOD where the model is trained only on synthetic data, still with 50 % of the samples being occluded.

For experiments on LINEMOD, we split training and test data exactly as [14], with the same ~ 200 samples for training and 1000 samples for test. With 50 % probability we sample a real training image and with 50 % probability a synthetic one, but without any occluders.

The objects known as `eggbox` and `glue`, present in both datasets, are conventionally considered symmetric w.r.t. a 180 degree rotation, but it can be argued whether to consider them as actual symmetries. Nevertheless, for these objects we duplicate the initial pose proposals with their 180 degree rotated equivalents and refine the pose using both initializations. In the end, the iterate with the lower estimated error is chosen.

4.2 Evaluation Metrics for Pose Refinement

For evaluation of our pose refinement method, we use three conventional metrics, explained in the following. All of them are defined as the percentage of annotated object instances for which the pose is correctly estimated, i.e., the recall according to the specific ways of quantifying the error.

The average distance metric ADD-0.1D [5] is the percentage of object instances for which the object point cloud, when transformed with the estimated pose as well as the ground-truth pose, has an average distance less than 10 % of the diameter of the object. The ADD-S-0.1D metric [5] is closely related, and only differs in that the closest point distance is used, rather than the distance between corresponding points. In general ADD-0.1D is used, but ADD-S-0.1D is used for objects that are considered symmetrical, and we let ADD(-S)-0.1D refer to the two of them together. The REPROJ-5PX metric is similar to ADD-0.1D, but differs in that the transformed point clouds are projected into the image before the mean distance is computed. The acceptance threshold is set to 5 pixels. Finally, the 5CM/5° metric accepts a pose estimate if the rotational and translational components differ from their ground-truth equivalents by at most 5 degrees and 5 cm, respectively.

4.2.1 "Symmetric" Objects and Faulty Annotations

When it comes to the REPROJ-5PX and 5CM/5° metrics, they do typically not take symmetries into account. This is not a huge problem for the LINEMOD dataset, partly because the high correlation between training and test data may help resolve any potential symmetries, and partly because, as pointed out earlier, none of the objects are truly symmetrical.

For Occlusion LINEMOD however, the `eggbox` object is unfortunately annotated according to the supposedly equivalent 180 degrees rotated pose, in all but the first 396 frames. For this reason, above mentioned metrics make little sense. Li *et al*. [14] do however modify these metrics in order to evaluate against the most beneficial of all proposed symmetries, which makes much more sense given the circumstances. We follow their proposal and

perform the evaluation on Occlusion LINEMOD w.r.t. these symmetrically aware metrics, which we will refer to as REPROJ-S-5PX and 5CM/5°-S.

4.3 Pose Refinement Results

Here we present our main pose refinement results. For a comparison of different backbone networks and detailed per-object results, we refer the reader to Section B.1 and Section B.3 in the supplementary material. Illustrations of refinement iterates are also available, in Section B.2 as well as in the supplied video.

State-of-the-art comparisons on the Occlusion LINEMOD dataset are given in Table 1. DeepIM [14] used initializations from PoseCNN [24], but as these predictions are not publicly available, we instead rely on initial pose proposals from PVNet [20]. The evaluation of PVNet was carried out by us and is based on the `clean-pvnet` implementation and pre-trained models[2]. Note that although the symmetry-aware REPROJ-S-5PX metric should be used on Occlusion LINEMOD (see Section 4.2.1), Oberweger *et al.* [17] report their results based on the REPROJ-5PX metric. We also want to mention that CDPN [15] perform well on Occlusion LINEMOD, but no quantitative numbers are reported.

	Oberweger *et al.* [17]	PVNet [20]	PoseCNN [24] + DeepIM [14]	PVNet [20] + PPC (Ours)
ADD(-S)-0.1D	30.40	41.37	**55.50**	55.33
REPROJ-S-5PX	60.86	61.84	56.61	**66.37**
5CM/5°-S	–	33.36	30.93	**41.52**

Table 1: Results on Occlusion LINEMOD. Note that [17] report results according to the REPROJ-5PX metric instead of REPROJ-S-5PX.

We also present results on the Occlusion LINEMOD dataset where we train purely on synthetic data, see Table 2. Our initial pose proposals are obtained from CDPN [15], which was the previous state-of-the-art for this set-up (cf. Benchmark for 6D Object Pose Estimation (BOP) evaluation server [6]). Also note that for these experiments only a subset of 200 test frames is used, in compliance with BOP.

	CDPN-synth [15]	CDPN-synth [15] + PPC-synth (Ours)
ADD(-S)-0.1D	18.76	**23.59**
REPROJ-S-5PX	32.22	**35.99**
5CM/5°-S	16.13	**19.81**

Table 2: Results on Occlusion LINEMOD using only synthetic training data.

Finally, results on the LINEMOD dataset are presented in Table 3 with the purpose of providing a direct comparison with DeepIM [14] with identical initializations. We outperform DeepIM on all metrics using the same proposals from PoseCNN [24]. Note that although no results on LINEMOD for PoseCNN are reported in [24], predictions by PoseCNN

[2] At times we observed negative depth estimates from PVNet, which was corrected for according to Section C.1 in the supplementary material.

are made available by [14]. The results are also good when compared to the state-of-the-art pose estimation methods of Li *et al*. [15] and Peng *et al*. [20] on LINEMOD.

	PoseCNN [24]	PoseCNN [24] + DeepIM [14]	PoseCNN [24] + PPC (Ours)
ADD(-S)-0.1D	62.04	88.33	**88.67**
REPROJ-5PX	64.52	97.53	**97.60**
5CM/5°	18.14	85.21	**89.74**

Table 3: Comparison with the refinement method of DeepIM and ours on LINEMOD with PoseCNN as initialization. Note that [14] reports results according to REPROJ-S-5PX and 5CM/5°-S metric instead of REPROJ-5PX and 5CM/5°.

4.4 Running Time

Experiments were run on a workstation with 64 GB RAM, Intel Core i7-8700K CPU, and Nvidia GTX 1080 Ti GPU. Our pose refinement pipeline takes on average 33 seconds per frame during inference for the 100 iterations to be carried out, meaning 3 iterations / s. One way to improve on this could be by enabling analytical differentiation through differentiable rendering, although care should be taken in order to make sure that $J(\theta)$ behaves smoothly enough, for instance, with a regularization scheme. Furthermore, rather than using iterative gradient-based optimization, gradient-free and sample-efficient approaches such as Bayesian optimization could be worth exploring, but is left as future research.

5 Conclusion

We have presented a novel rendering-based pose refinement method, which shows improved performance compared to previous refinement methods, and is robust to partial occlusions.

On the Occlusion LINEMOD benchmark, we initialize our method with pose proposals from PVNet [20], yielding state-of-the-art results for two out of three metrics on this competitive benchmark, while performing on-par with previous methods for the third metric. Furthermore, additional experiments on Occlusion LINEMOD show that our method works well also when trained purely on synthetic data, improving on the pose estimates of CDPN [15]. Finally, on the LINEMOD benchmark, previous refinement methods are outperformed for all metrics.

6 Acknowledgements

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, the Swedish Research Council (grant no. 2016-04445) and the Swedish Foundation for Strategic Research (Semantic Mapping and Visual Navigation for Smart Robots).

References

[1] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6D object pose estimation using 3D object coordinates. In *The European Conference on Computer Vision (ECCV)*, September 2014.

[2] Thanh-Toan Do, Ming Cai, Trung Pham, and Ian Reid. Deep-6DPose: Recovering 6D object pose from a single RGB image, 2018.

[3] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[4] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. Second Edition.

[5] 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 *Computer Vision – ACCV 2012*, pages 548–562. Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-37331-2_42. URL https://doi.org/10.1007/978-3-642-37331-2_42.

[6] Tomáš Hodaň, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiří Matas, and Carsten Rother. BOP: Benchmark for 6D object pose estimation. *European Conference on Computer Vision (ECCV)*, 2018.

[7] Omid Hosseini Jafari, Siva Karthik Mustikovela, Karl Pertsch, Eric Brachmann, and Carsten Rother. iPose: instance-aware 6D pose estimation of partly occluded objects. In *ACCV*, 2018.

[8] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Segmentation-driven 6D object pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[10] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[11] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[12] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-DoF camera relocalization. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[13] 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. URL http://arxiv.org/abs/1412.6980.

[14] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. DeepIM: Deep iterative matching for 6d pose estimation. *International Journal of Computer Vision*, 128 (3):657–678, November 2019. doi: 10.1007/s11263-019-01250-9. URL https://doi.org/10.1007/s11263-019-01250-9.

[15] Zhigang Li, Gu Wang, and Xiangyang Ji. CDPN: Coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[16] Fabian Manhardt, Wadim Kehl, Nassir Navab, and Federico Tombari. Deep model-based 6D pose refinement in RGB. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[17] Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Making deep heatmaps robust to partial occlusions for 3D object pose estimation. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[18] Kiru Park, Timothy Patten, and Markus Vincze. Pix2Pose: Pixel-wise coordinate regression of objects for 6D pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[19] Georgios Pavlakos, Xiaowei Zhou, Aaron Chan, Konstantinos G. Derpanis, and Kostas Daniilidis. 6-DoF object pose from semantic keypoints. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2017. doi: 10.1109/icra.2017.7989233. URL https://doi.org/10.1109/icra.2017.7989233.

[20] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. PVNet: Pixel-wise voting network for 6DoF pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[21] 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 *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[22] Bugra Tekin, Sudipta N. Sinha, and Pascal Fua. Real-time seamless single shot 6D object pose prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[23] Henning Tjaden, Ulrich Schwanecke, and Elmar Schomer. Real-time monocular pose estimation of 3D objects using temporally consistent local color histograms. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[24] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. *Robotics: Science and Systems (RSS)*, 2018.

[25] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. DPOD: 6D pose object detector and refiner. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.