GP^{2}C: Geometric Projection Parameter Consensus for Joint 3D Pose and Focal Length Estimation in the Wild

Alexander Grabner[1] Peter M. Roth[1] Vincent Lepetit[2,1]

[1]Institute of Computer Graphics and Vision, Graz University of Technology, Austria

[2]Laboratoire Bordelais de Recherche en Informatique, University of Bordeaux, France

{alexander.grabner,pmroth,lepetit}@icg.tugraz.at

Abstract

We present a joint 3D pose and focal length estimation approach for object categories in the wild. In contrast to previous methods that predict 3D poses independently of the focal length or assume a constant focal length, we explicitly estimate and integrate the focal length into the 3D pose estimation. For this purpose, we combine deep learning techniques and geometric algorithms in a two-stage approach: First, we estimate an initial focal length and establish 2D-3D correspondences from a single RGB image using a deep network. Second, we recover 3D poses and refine the focal length by minimizing the reprojection error of the predicted correspondences. In this way, we exploit the geometric prior given by the focal length for 3D pose estimation. This results in two advantages: First, we achieve significantly improved 3D translation and 3D pose accuracy compared to existing methods. Second, our approach finds a geometric consensus between the individual projection parameters, which is required for precise 2D-3D alignment. We evaluate our proposed approach on three challenging real-world datasets (Pix3D, Comp, and Stanford) with different object categories and significantly outperform the state-of-the-art by up to 20% absolute in multiple different metrics.

1. Introduction

3D object pose estimation aims at predicting the 3D rotation and 3D translation of objects relative to the camera. It is a fundamental yet unsolved computer vision problem with many applications, including augmented reality, robotics, and scene understanding. Recently, there have been great advances in 3D object pose estimation from single RGB images on the category level [9, 31, 37, 45], thanks to the development of deep learning and the creation of large-scale datasets providing 3D annotations for RGB images [50, 51].

While recent approaches achieve high accuracy in terms of 3D rotation, their accuracy in terms of 3D translation is often low [30, 46]. The main reason for this discrepancy is illustrated in Figure 1, where we compare two images of an object captured with cameras having different focal lengths. The appearance of the object is similar in both images, even though the 3D poses are significantly different. In fact, the appearance of an object in an image is not only determined by the 3D pose, but also by the camera intrinsics. While changes in the 3D rotation always significantly effect the appearance, changes in the 3D translation do not if the translation direction and the ratio between the object-to-camera distance and the focal length remain constant. Thus, estimating the 3D translation of objects from RGB images in the case of unknown intrinsics is highly ambiguous.

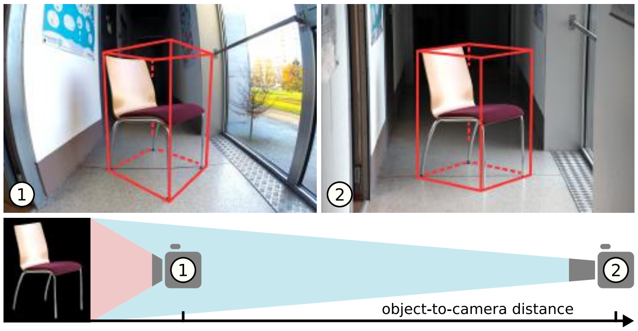


Figure 1: Images captured with two cameras having different focal lengths. The appearance of the chair is similar in both images, but the 3D poses are significantly different due to the distinct focal lengths and object-to-camera distances.

Existing approaches assume that the 3D pose estimation method will implicitly learn the subtle appearance variations caused by different focal lengths from the data and adapt the prediction accordingly [30, 46]. In practice, however, this is not the case, because deep networks do not find the solutions we intend without explicit guidance.

To overcome this limitation, we propose to explicitly estimate and integrate the focal length into the 3D pose estimation. For this purpose, we introduce a two-stage approach that combines deep learning techniques and geometric algorithms. In the first stage, we estimate an initial focal length and establish 2D-3D correspondences from a single RGB image using a deep network. In the second stage, we perform a geometric optimization on the predicted correspondences to recover 3D poses and refine the focal length. In particular, we minimize the reprojection error between predicted 2D locations and 3D points subject to the 3D rotation, 3D translation, and the focal length by solving a P*n*Pf problem [32]. In this way, we exploit the geometric prior given by the focal length for 3D pose estimation.

In contrast to existing approaches, which also predict 3D poses and the focal length but only perform an independent estimation of the individual parameters [46], our approach has two main advantages: First, explicitly modeling the focal length in the 3D pose estimation yields significantly improved 3D translation and 3D pose accuracy. Second, our approach finds a geometric consensus between 3D poses and the focal length. This results in a significantly improved 2D-3D alignment when projecting 3D models of objects back onto the image, which is important for many applications like augmented reality. Therefore, we call our method *Geometric Projection Parameter Consensus* (GP^{2}C).

In addition, we explore two possible methods for establishing 2D-3D correspondences from RGB images, which approach the task from different directions. Our first method predicts 3D points for known 2D locations by estimating a 3D coordinate for each object pixel [1, 2, 20]. Our second method predicts 2D locations for known 3D points by estimating the 2D projections of the object's 3D bounding box corners [9, 36, 41]. Our experiments show that both methods achieve comparable accuracy, but each method has its respective advantages and disadvantages. Thus, we provide a detailed discussion comparing the two methods.

To demonstrate the benefits of our joint 3D pose and focal length estimation approach, we evaluate it on three challenging real-world datasets with different object categories: Pix3D [38] (*bed*, *chair*, *sofa*, *table*), Comp [46] (*car*), and Stanford [46] (*car*). We present quantitative as well as qualitative results and significantly outperform the state-of-the-art. To summarize, our main contributions are:

- We present the first method for joint 3D pose and focal length estimation that enforces a geometric consensus between 3D poses and the focal length.
- We outperform the state-of-the-art by up to 20% absolute in multiple metrics covering different aspects of projective geometry including 3D translation, 3D pose, focal length, and projection accuracy.

2. Related Work

In this section, we discuss previous work on 3D pose estimation for object categories and approaches for estimating the camera intrinsics, in particular, the focal length.

2.1. 3D Pose Estimation

A recent trend in computer vision is to predict pose parameters directly using deep learning. In this context, numerous works predict only the 3D rotation of objects using CNNs. These methods perform rotation classification [37, 44, 45], regression [28, 50], or apply hybrid variants of both [27] using different parametrizations such as Euler angles, quaternions, or exponentials maps.

In this work, however, we focus on the estimation of the full 3D pose, *i.e.*, the 3D rotation and 3D translation of objects. In this case, many approaches combine the 3D rotation estimation techniques described above with 3D translation regression [24, 30, 31]. Because detecting and localizing objects in 2D is often a first step towards estimating the 3D pose, recent approaches integrate 3D pose estimation techniques into object detection pipelines making the entire system end-to-end trainable [21, 22, 46, 52]. However, these methods do not explicitly take the camera intrinsics into account, which results in poor performance on images captured with different focal lengths, for example.

In contrast to these direct approaches, there is a large amount of research on recovering the pose from 2D-3D correspondences, additionally considering a camera model [10]. In this context, recent approaches use CNNs to predict the 2D locations of the projections of 3D keypoints from RGB images [33, 35]. While [35] recovers the 3D pose from the predicted 2D locations and a given 3D model using a P*n*P algorithm, [33] recovers the 3D pose from the predicted 2D locations alone using a trained deformable shape model. However, these approaches rely on category-specific semantic 3D keypoints which need to be selected and annotated manually for each 3D model.

In this work, we also predict 2D-3D correspondences from RGB images, but do not rely on category-specific 3D keypoints. In particular, we explore two different strategies. Our first strategy is to predict 3D points for known 2D locations. A natural choice is to predict a 3D point for each image pixel [1]. In this case, it is important to know which pixels belong to an object and which pixels belong to the background or another object [2]. Recently, it has been shown that deep learning techniques for instance segmentation [11] significantly increase the accuracy on this task [20, 46]. In contrast to our approach, [20] relies on two disjoint networks for instance segmentation and 3D point regression followed by a geometric optimization assuming a constant focal length. Instead, we use a single network to perform both tasks and additionally optimize the focal length. [46] on the other hand also regresses 3D points with a single network, but relies on a second network to estimate the 3D rotation from these points, compared to our approach which uses a geometric optimization on arbitrary 2D-3D correspondences for joint 3D pose and focal length estimation.

Our second strategy is to predict the 2D locations of known 3D points. In this case, we choose to predict virtual 3D points which generalize across different objects and categories, *e.g.*, the corners of the 3D bounding box of an object [36, 41], instead of category-specific 3D keypoints. Recently, it has been shown that this approach can be extended to make predictions without the use of 3D models during inference [9]. In contrast to our work, [9] assumes that all objects are already detected and localized in 2D, and uses a constant focal length.

2.2. Focal Length Estimation

Computing the focal length and other camera intrinsics from 2D-3D correspondences has a long tradition in computer vision [8, 10]. In this context, the intrinsic and extrinsic parameters of the camera are often recovered jointly [32, 48]. For this purpose, numerous works explicitly estimate the focal length and the 3D pose of the camera by solving a P*n*Pf problem [34, 54, 55].

In practice, these methods require precise 2D-3D correspondences, which are often selected manually or using calibration grids [43, 53]. Many applications, however, require automatic calibration. In specific cases, it is possible to exploit geometric image elements such as lines [7], vanishing points [40], or circles [5] to compute the intrinsics, but these methods do not generalize to arbitrary natural images.

Thus, recent works estimate the focal length from RGB images without requiring particular geometric structures using deep learning [46, 47]. In this work, we take a similar approach. However, in contrast to existing methods, we propose a different parametrization and additionally use 2D-3D correspondences to refine the predicted focal length.

3. Joint 3D Pose and Focal Length Estimation

Given a single RGB image, we want to predict the focal length and the 3D pose of each object in an image. For this purpose, we introduce a two-stage approach that combines deep learning techniques and geometric algorithms, as shown in Figure 2. In the first stage, we predict an initial focal length and establish 2D-3D correspondences using deep learning (Sec. 3.1). In the second stage, we perform a geometric optimization on the predicted correspondences to recover 3D poses and refine the focal length (Sec. 3.2).

3.1. Stage 1: Deep Focal Length and 2D-3D Correspondence Estimation

To predict the focal length as well as 2D-3D correspondences with a single deep network, we extend the generalized Faster/Mask R-CNN framework [11, 37]. This generic multi-task framework includes a 2D object detection pipeline to perform per-image and per-object computations. In this way, we address multiple different tasks using a single end-to-end trainable network. For our implementation, we use a Feature Pyramid Network [25] on top of a ResNet-101 backbone [12, 13] and finetune a model pretrained for instance segmentation on COCO [26].

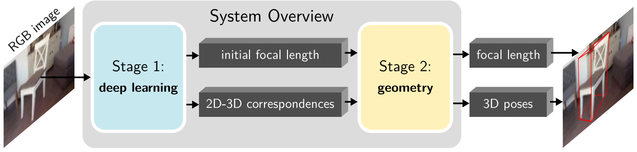


Figure 2: Overview of our proposed two-stage approach. **Stage 1:** We predict an initial focal length and establish 2D-3D correspondences using deep learning. **Stage 2:** We perform a geometric optimization on the predicted correspondences to recover 3D poses and refine the focal length.

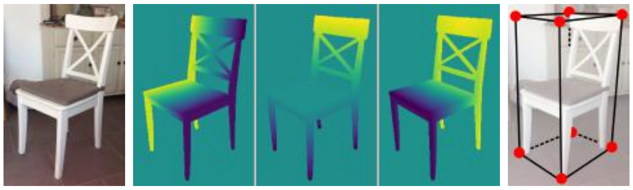
In the context of the generalized Faster/Mask R-CNN framework, an output branch provides one or more subnetworks with different structure and functionality. We introduce two dedicated output branches for estimating the focal length and 2D-3D correspondences alongside the existing object detection branches.

Focal Length. The focal length branch provides one subnetwork which performs a per-image computation. In this case, we regress a scalar for each image from the entire spatial resolution of the shared feature maps computed by the convolutional network backbone. In contrast to previous work, we propose to regress a logarithmic parametrization of the focal length

$$y_f = ln(f), \tag{1}$$

instead of predicting the focal length f directly [46], which has two advantages: First, the logarithmic parametrization reduces the bias towards minimizing the error on long focal lengths during the optimization of the network. This is meaningful because, regarding the estimation of the focal length, the relative error is more important than the absolute error. Second, the logarithmic parametrization achieves a more balanced sensitivity across the entire range of the focal length. Otherwise, the sensitivity is significantly higher for short focal lengths than for long focal lengths. During training, we optimize y_f using the Huber loss [19].

2D-3D correspondences. For establishing 2D-3D correspondences, we explore two distinct methods. Both methods approach the problem from different directions and produce significantly different correspondences and representations, as shown in Figure 3. However, our overall approach works with any kind of 2D-3D correspondences and does not depend on a specific format. Thus, the method for establishing correspondences can be exchanged. This is extremely useful, because different methods have their respective advantages and disadvantages which we discuss in our experiments in Sec. 4.3.



(a) (b) (c)

Figure 3: Visualization of two different forms of 2D-3D correspondences: (a) Image, (b) Location field which encodes XYZ 3D coordinates for each pixel (**LF**), and (c) 2D projections of the object's 3D bounding box corners (**BB**).

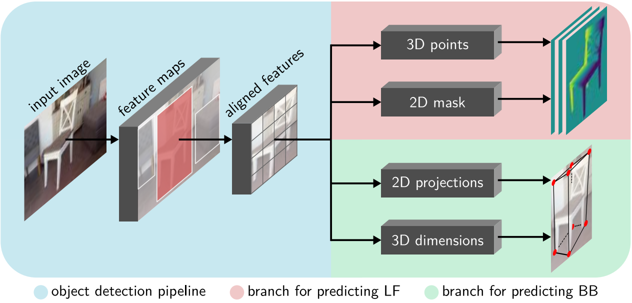


Figure 4: Two alternative branches for predicting 2D-3D correspondences from an RGB image (**LF** and **BB**).

Our first method predicts 3D points for known 2D locations. In particular, we establish correspondences between 2D image pixels which belong to the object and 3D coordinates on the surface of the object. We represent these correspondences in the form of a location field (LF) [46], which provides dense 2D-3D correspondences in an image-like format, as shown in Figure 3b. A location field has the same size and spatial resolution as its reference RGB image, but the three channels encode XYZ 3D coordinates in the object coordinate system instead of RGB colors. Due to its image-like structure, this representation is well-suited for regression with a CNN.

Our second method predicts 2D locations for known 3D points. In this case, we predict the 2D projections of the object's 3D bounding box corners (BB) [36], as shown in Figure 3c. Since the 3D coordinates of the bounding box corners are unknown during inference, we also predict the 3D dimensions of the object along the XYZ axes [9] from which we derive the required 3D points. We represent these sparse 2D-3D correspondences in the form of a 19-dimensional vector, which consists of the 2D locations of the eight bounding box corners (16 values) and the 3D dimensions of the object (3 values).

As shown in Figure 4, we implement a separate 2D-3D correspondences branch for each method. In contrast to the focal length branch, both branches perform region-based per-object computations: For each detected object, an associated spatial region of interest in the feature maps is aligned to a fixed size feature representation with a low spatial resolution, *e.g.*, 14×14. These aligned features serve as an input to one of our two proposed branches. Thus, the chosen 2D-3D correspondences branch is evaluated N times for each image, where N is the number of detected objects. We identify the chosen 2D-3D correspondences method by adding a suffix: Ours-LF or Ours-BB.

For the LF method, the correspondences branch provides two different fully convolutional subnetworks to predict a tensor of 3D points and a 2D object mask at a spatial resolution of 28×28. The 2D mask is then applied to the tensor of 3D points to get a low-resolution location field. We found this approach to produce significantly higher accuracy compared to directly regressing a low-resolution location field which tends to predict over-smoothed 3D coordinates around the object silhouette.

The resulting low-resolution location field can be up-scaled and padded to obtain a high-resolution location field with the same spatial resolution as the input image. However, we sample 2D-3D correspondences from the low-resolution location field and only adjust their 2D locations to match the input image resolution. In this way, we avoid generating a large number of 2D-3D correspondences without providing additional information.

For the BB method, the correspondences branch also provides two subnetworks, but this time with fully connected output layers. One subnetwork predicts the 2D locations of the object's 3D bounding box corners, the other subnetwork estimates the 3D dimensions of the object. In this case, we regress the 2D location in normalized coordinates relative to the spatial resolution of the aligned features. Again, we adjust the predicted 2D locations to match the input image resolution.

During training, we optimize the 3D points and 2D mask (Ours-LF), or the 2D projections and 3D dimensions (Ours-BB) using the Huber loss [19]. The final network loss is a combination of our focal length loss, our chosen 2D-3D correspondences loss, and the 2D object detection losses of the generalized Faster/Mask R-CNN framework [11, 37].

3.2. Stage 2: Geometric Optimization

Once we established correspondences between 2D locations and 3D points, we use the same geometric optimization for all methods. In this case, we perform a non-linear optimization of the PnPf problem [32] which finds a geometric consensus between the individual projection parameters. In particular, we minimize the reprojection error

$$e_{\text{reproj}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\|\text{Proj}_{R,t,f}(\boldsymbol{X}_i) - \boldsymbol{x}_i\|_2) \,, \qquad (2)$$

where $\boldsymbol{X}_i$ is a 3D point and $\boldsymbol{x}_i$ its corresponding 2D location. $\text{Proj}_{R,t,f}(\cdot)$ performs the projection from the 3D object coordinate system onto the 2D image plane with respect to the rotation R, translation t, and focal length f. $\mathcal{L}(\cdot)$ is a loss function, such as the standard squared loss $\mathcal{L}(x) = x^2$ or the more robust Cauchy loss [42] $\mathcal{L}(x) = ln(1 + x^2)$, and N denotes the number of correspondences.

We minimize e_{reproj} over both the 3D pose and the focal length. In this case, a minimum of four 2D-3D correspondences is needed to find a unique solution [49], because each correspondence gives two independent equations and we optimize seven parameters: the 3-DoF rotation, the 3-DoF translation, and the 1-DoF focal length. In practice, however, it is important to use more 2D-3D correspondences to compensate for the presence of noise.

Following the strategy of previous P*n*P(f) approaches [14, 23, 34], we compute an initial solution in $O(n)$ time followed by an iterative refinement technique. For our initial solution, we compute the 3D rotation and 3D translation using EP*n*P [23] with our predicted focal length. Providing a good initial focal length is a key factor in achieving high accuracy in terms of 3D translation. In theory, it is also possible to recover the focal length using 2D-3D correspondences from scratch [32, 34], but this requires extremely accurate and clean correspondences. For correspondence estimation on the category level in the wild, however, we are facing fuzzy and noisy predictions. In this case, a low reprojection error is achieved by finding the correct ratio between the object-to-camera distance and the focal length. Thus, we cannot assume that the geometric optimization will find the correct absolute focal length from scratch.

Taking this into account, we jointly optimize the 3D rotation, 3D translation, and focal length during our iterative refinement. For this purpose, we employ a Newton-Step-based optimization [6] depending on the loss function $\mathcal{L}$, *i.e*., Levenberg-Marquardt [29] (squared loss) or Subspace Trust-Region Interior-Reflective [3] (Cauchy loss).

Our approach naturally handles different projection models (egocentric or allocentric) [22]. Additionally, jointly optimizing the 3D poses of multiple objects in an image together with the focal length is straightforward. In this case, we compute the initial solution as before, but perform our iterative refinement for $1 + 6N$ parameters where N is the number of detected objects. We did not evaluate this joint refinement though, because available category level datasets with focal length annotations just provide 3D annotations for one object per image, even if there are multiple objects in the image [38, 46]. In most cases, we are still able to detect the other objects, but do not have ground truth annotations to evaluate them, as shown in our qualitative results in Sec. 4.1. Moreover, our approach can readily be extended to deal with more complex camera models including skew, off-center principal point, asymmetric aspect ratio or lens distortions [32]. However, currently there are no datasets with this kind of annotations.

4. Experimental Results

To demonstrate the benefits of our joint 3D pose and focal length estimation approach (GP^{2}C), we evaluate it on three challenging real-world datasets[1] with different object categories: Pix3D [38] (*bed*, *chair*, *sofa*, *table*), Comp [46] (*car*), and Stanford [46] (*car*). In particular, we provide a quantitative and qualitative evaluation of our approach in comparison to the state-of-the-art in Sec. 4.1, analyze important aspects in Sec. 4.2, and discuss advantages and disadvantages of our two presented methods for establishing 2D-3D correspondences in Sec. 4.3. To cover different aspects of projective geometry in our evaluation, we use the following well-established metrics:

Detection. We report the detection accuracy $Acc_{D_{0.5}}$ which gives the percentage of objects for which the intersection over union between the ground truth 2D bounding box and the predicted 2D bounding box is larger than 50% [51]. This metric is an upper bound for other Acc metrics since we do not make blind predictions.

Rotation. We compute the geodesic distance

$$e_R = \frac{\|\log(R_{\text{gt}}^T R_{\text{pred}})\|_F}{\sqrt{2}} \tag{3}$$

between the ground truth rotation matrix R_{gt} and the predicted rotation matrix R_{pred} which gives the minimal angular distance. We report the median of this distance ($MedErr_R$) and the percentage of objects for which the distance is below the threshold of $\frac{\pi}{6}$ or 30° ($Acc_{R_{\frac{\pi}{6}}}$) [45].

Translation. We report the relative translation distance

$$e_t = \frac{\|t_{\text{gt}} - t_{\text{pred}}\|_2}{\|t_{\text{gt}}\|_2} \tag{4}$$

between the ground truth translation t_{gt} and the predicted translation t_{pred} [18].

Pose. We calculate the average normalized distance of all transformed model points in 3D space

$$e_{R,t} = \underset{\boldsymbol{X} \in \mathcal{M}}{\text{avg}} \frac{d_{\text{bbox}}}{d_{\text{img}}} \cdot \frac{\|\text{Transf}_{\text{gt}}(\boldsymbol{X}) - \text{Transf}_{\text{pred}}(\boldsymbol{X})\|_2}{\|t_{\text{gt}}\|_2} \tag{5}$$

to evaluate 3D pose accuracy [16, 18]. In this case, each 3D point $\boldsymbol{X}$ of the ground truth 3D model $\mathcal{M}$ is transformed using the ground truth 3D pose $\text{Transf}_{\text{gt}}(\cdot)$ and the predicted 3D pose $\text{Transf}_{\text{pred}}(\cdot)$ subject to rotation and translation. We normalize this distance by the relative size of

[1] Details on the datasets and the evaluation setup are provided in the **supplementary material**.

			Detection	Rotation		Translation	Pose	Focal	Projection	
Method	Dataset	Class	$Acc_{D_{0.5}}$	$MedErr_R$ $\cdot 1$	$Acc_{R_{\frac{\pi}{6}}}$	$MedErr_t$ $\cdot 10^1$	$MedErr_{R,t}$ $\cdot 10^1$	$MedErr_f$ $\cdot 10^1$	$MedErr_P$ $\cdot 10^2$	$Acc_{P_{0.1}}$
[46]	Pix3D	bed	98.4%	5.82	95.3%	1.95	1.56	2.22	6.05	74.9%
Ours-LF			99.0%	**5.13**	96.3%	**1.41**	**1.04**	**1.43**	**3.52**	90.6%
Ours-BB			**99.5%**	5.40	**97.9%**	1.66	1.17	1.59	3.55	**93.2%**
[46]	Pix3D	chair	94.9%	7.52	88.0%	2.69	1.58	1.98	6.04	75.3%
Ours-LF			95.2%	7.52	88.8%	1.92	1.21	1.62	3.41	88.2%
Ours-BB			**97.3%**	**6.95**	**91.0%**	**1.68**	**1.08**	**1.58**	**3.24**	**90.9%**
[46]	Pix3D	sofa	96.5%	4.73	94.8%	2.28	1.62	2.42	4.33	82.2%
Ours-LF			96.5%	4.49	95.0%	1.92	1.33	1.79	2.56	93.7%
Ours-BB			**98.3%**	**4.40**	**97.0%**	**1.63**	**1.16**	**1.73**	**2.13**	**95.6%**
[46]	Pix3D	table	94.0%	10.94	72.9%	3.16	2.28	3.03	8.90	53.6%
Ours-LF			94.0%	**10.53**	73.5%	**2.16**	**1.62**	**2.05**	5.92	69.5%
Ours-BB			**95.7%**	10.80	**77.2%**	2.81	1.78	2.10	**5.74**	**72.4%**
[46]	Pix3D	*mean*	96.0%	7.25	87.8%	2.52	1.76	2.41	6.33	71.5%
Ours-LF			96.2%	6.92	88.4%	**1.85**	**1.30**	**1.72**	3.85	85.5%
Ours-BB			**97.7%**	**6.89**	**90.8%**	1.94	**1.30**	1.75	**3.66**	**88.0%**
[46]	Comp	car	**98.9%**	5.24	97.6%	3.30	2.35	3.23	7.85	73.7%
Ours-LF			98.8%	5.23	97.9%	2.61	1.86	2.97	4.21	95.1%
Ours-BB			**98.9%**	**4.87**	**98.1%**	**2.55**	**1.84**	**2.95**	**3.87**	**95.7%**
[46]	Stanford	car	**99.6%**	5.43	98.0%	2.33	1.80	2.34	7.46	76.4%
Ours-LF			**99.6%**	5.38	**98.3%**	1.93	1.51	**2.01**	3.72	96.2%
Ours-BB			**99.6%**	**5.24**	**98.3%**	**1.92**	**1.47**	2.07	**3.25**	**96.5%**

Table 1: Experimental results on the Pix3D, Comp, and Stanford datasets. We significantly outperform the state-of-the-art in the 3D translation, 3D pose, focal length, and projection metrics. We explain the reported numbers in detail in Sec. 4.1.

the object in the image using the ratio between the ground truth 2D bounding box diagonal d_{bbox} and the image diagonal d_{img}, and the L2-norm of the ground truth translation $\|t_{\text{gt}}\|_2$. This normalization provides an unbiased metric for 3D pose evaluation in the case of unknown intrinsics.

Focal Length. We report the relative focal length error

$$e_f = \frac{|f_{\text{gt}} - f_{\text{pred}}|}{f_{\text{gt}}} \tag{6}$$

between the ground truth focal length f_{gt} and the predicted focal length f_{pred} [34, 48].

Projection. To evaluate all projection parameters, we compute the average normalized reprojection distance

$$e_P = \operatorname*{avg}_{\boldsymbol{X} \in \mathcal{M}} \frac{\|\text{Proj}_{\text{gt}}(\boldsymbol{X}) - \text{Proj}_{\text{pred}}(\boldsymbol{X})\|_2}{d_{\text{bbox}}} . \tag{7}$$

In this case, each 3D point $\boldsymbol{X}$ of the ground truth 3D model $\mathcal{M}$ is projected to a 2D location using the ground truth projection parameters $\text{Proj}_{\text{gt}}(\cdot)$ and the predicted projection parameters $\text{Proj}_{\text{pred}}(\cdot)$ subject to rotation, translation, and focal length. d_{bbox} is the ground truth 2D bounding box diagonal. We report the median of this distance ($MedErr_P$) and the percentage of objects for which the distance is below the threshold of 0.1 ($Acc_{P_{0.1}}$) [46].

4.1. Comparison to the State-of-the-Art

We first present quantitative results of our approach using our two different methods for establishing 2D-3D correspondences (Ours-LF and Ours-BB) and compare them to the state-of-the-art. To this end, we reimplemented the approach of [46] and achieve comparable results, even outperforming their reported $MedErr_P$ and $Acc_{P_{0.1}}$ scores due to our improved backbone architecture and initialization. The results are summarized in Table 1. We achieve consistent results across all datasets and categories, thus, we provide a joint discussion based on the evaluated metrics:

Detection. All methods achieve high detection accuracy ($Acc_{D_{0.5}}$). This is not surprising, because we fine-tune a model pre-trained for instance segmentation on COCO [26]. In fact, all evaluated categories are also present in COCO.

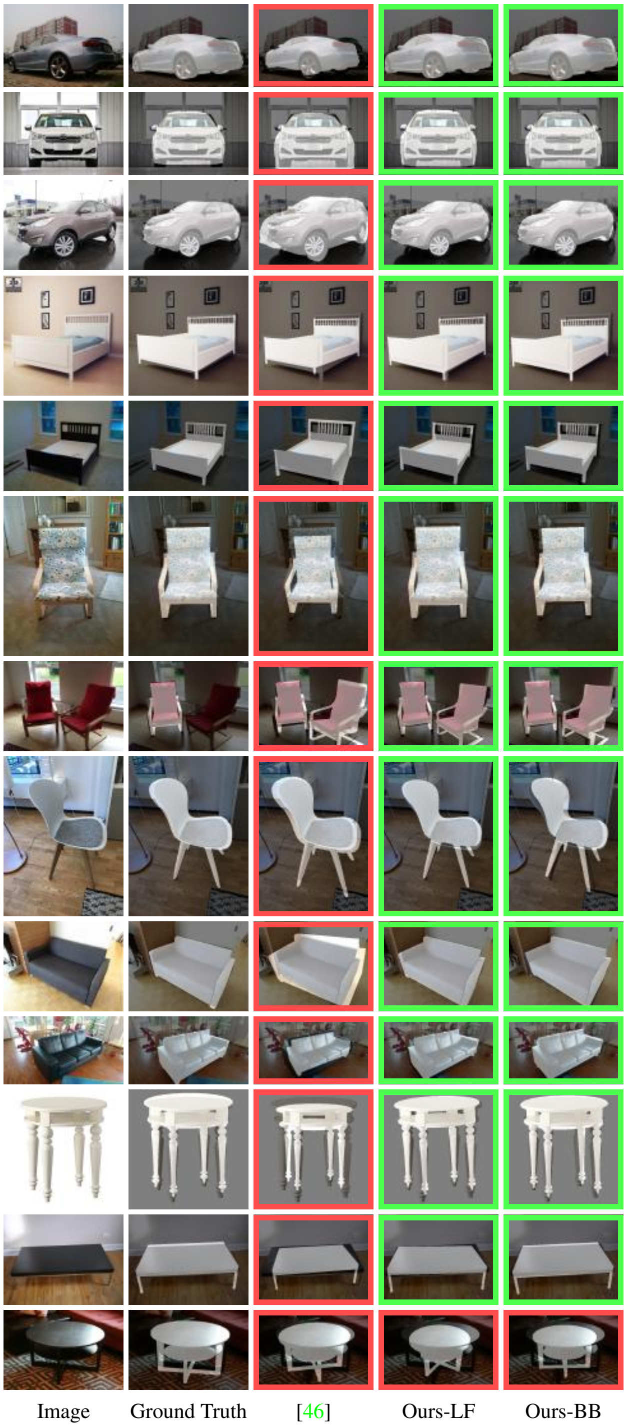


Figure 5: Qualitative 3D pose and focal length estimation results for all evaluated datasets and categories. We project the ground truth 3D model onto the image using the 3D pose and focal length predicted by different approaches. In contrast to [46], our approach finds a geometric consensus between the parameters which results in improved 2D-3D alignment, *e.g.*, the **scale** of the projection. We highlight respective samples with frames. **Best viewed in digital zoom.**

Rotation. Also, all methods achieve high rotation accuracy ($MedErr_R$ and $Acc_{R_{\frac{\pi}{6}}}$). Our reported numbers are in line with the results of previous work on rotation estimation in the wild [9, 45, 46] and confirm that 3D rotation can robustly be recovered from 2D observations up to a certain precision. Only for the category $table$, we observe sub-average accuracy. In fact, almost all tables have symmetries, as can be seen in Figure 5, which sometimes confuse all evaluated methods, because they predict a single 3D pose rather than a distribution (see last $table$ sample).

Translation. In terms of translation accuracy ($MedErr_t$), our approach significantly outperforms the state-of-the-art. Directly predicting the 3D translation from a local image window of an object is highly ambiguous in the case of unknown intrinsics. By explicitly estimating and integrating the focal length into the 3D pose estimation, we exploit a geometric prior and achieve a relative improvement of 20%.

Pose. In the case of unknown intrinsics, the 3D pose accuracy ($MedErr_{R,t}$) is primarily governed by the translation accuracy. Therefore, we also observe a relative improvement of 20% compared to the state-of-the-art.

Focal Length. Considering the focal length accuracy ($MedErr_f$), our approach outperforms the state-of-the-art by a relative improvement of 10% due to our logarithmic parametrization and refinement.

Projection. Finally, we report the projection metrics ($MedErr_P$ and $Acc_{P_{0.1}}$), which evaluate all predicted parameters. In these metrics, we achieve the largest improvement compared to the state-of-the-art: **20% absolute** in $Acc_{P_{0.1}}$ and **40% relative** in $MedErr_P$ across all datasets. In contrast to an independent estimation of the individual projection parameters, our approach finds a geometric consensus which results in improved 2D-3D alignment and reprojection error. This quantitative improvement is also reflected in our qualitative results shown in Figure 5. In this experiment, our approach consistently produces a higher quality 2D-3D alignment compared to the state-of-the-art for objects of different categories. This significant improvement can be accounted to the fact that we minimize the reprojection error during inference. However, we want to emphasize that the 3D model is only used for the evaluation. The 3D poses and focal length are solely computed from a single RGB image in our approach.

4.2. Analysis

Next, we analyze two important aspects of our approach: (a) the robustness of our predicted 2D-3D correspondences and (b) the importance of the focal length for estimating 3D poses from these correspondences. For this purpose, we perform experiments on Pix3D, which is the most challenging dataset, because it provides multiple object categories and has the largest variation in object scale.

		Projection	
Method	PnP	$MedErr_P \cdot 10^2$	$Acc_{P_{0.1}}$
Ours-LF	Standard	3.88	85.3%
	RANSAC	3.87	85.4%
	Cauchy	**3.85**	**85.5%**
Ours-BB	Standard	3.68	87.5%
	RANSAC	3.68	87.6%
	Cauchy	**3.66**	**88.0%**

Table 2: Evaluation of different PnP strategies. The results show that our predicted 2D-3D correspondences are reliable and do not contain single extreme outliers.

First, we run our approaches using different PnP strategies and compare the obtained results using the projection metrics ($MedErr_P$ and $Acc_{P_{0.1}}$) in Table 2. In particular, we compare the standard approach, which is sensitive to outliers due to the squared loss $\mathcal{L}(x) = x^2$, to the more robust RANSAC scheme and Cauchy loss [42] $\mathcal{L}(x) = ln(1 + x^2)$.

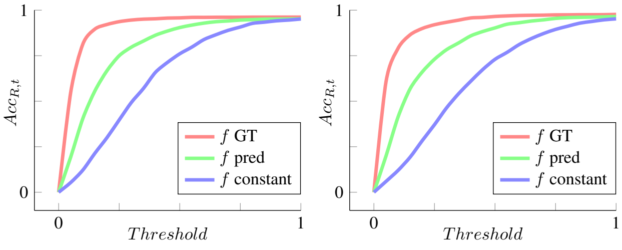
All three PnP strategies achieve similar performance for both Ours-LF and Ours-BB. This experiment shows that our predicted 2D-3D correspondences do not contain single extreme outliers which are often present in traditional interest-point-based approaches. This is due to the fact that all 2D-3D correspondences are computed from a low dimensional feature embedding which produces consistent predictions[2].

Second, to demonstrate the importance of the focal length for estimating 3D poses from 2D-3D correspondences, we initialize the geometric optimization with three different focal lengths and compare the results using the 3D pose distance in Figure 6. In this experiment, we plot the percentage of objects for which the 3D pose distance is below a threshold varying in the range [0,1] ($Acc_{R,t}$).

As expected, if we initialize the geometric optimization with the ground truth focal length, we achieve the highest 3D pose accuracy. However, for 3D pose estimation in the wild, the focal length is unknown during inference. In this case, we can use a constant or a predicted focal length for initialization. Even if we use the best possible constant focal length, which is the median focal length of the training dataset, the accuracy drops significantly. Instead, if we initialize using our predicted focal length, we achieve improved 3D pose accuracy. However, there is still a gap in the accuracy compared to using the ground truth focal length.

4.3. Discussion

So far, our results show that both presented 2D-3D correspondence estimation methods (LF and BB) achieve a similar level of accuracy. However, each method has specific characteristics advantageous for different tasks.



(a) Ours-LF

(b) Ours-BB

Figure 6: Evaluation of different initial focal lengths. The results show that a good initial estimate of the focal length is a key factor for achieving high 3D pose accuracy.

For example, LF implicitly handles truncations and occlusions, because it estimates 3D points for visible object parts and resolves occlusions using the 2D mask. Moreover, the predicted dense 2D-3D correspondences might also be useful for other tasks like dense depth estimation or shape reconstruction. However, this method requires detailed 3D models for training.

In contrast, BB only requires accurate 3D bounding boxes for training. The overall design of this method is simpler and more lightweight, which makes it easier to implement and train. This is also reflected in our reported numbers, which show a slight advantage compared to LF. Additionally, BB always gives a fixed number of sparse 2D-3D correspondences. This results in fast inference, which is beneficial for real-time applications, for example. However, while this method is well-suited for dealing with box-shaped objects like cars, other approaches might perform better on highly non-box-shaped objects.

5. Conclusion

Estimating the 3D poses of objects in the wild is an important but challenging task. In particular, predicting the 3D translation is difficult due to ambiguous appearances resulting from different focal lengths. For this purpose, we present the first joint 3D pose and focal length estimation approach that enforces a geometric consensus between 3D poses and the focal length. Our approach combines deep learning techniques and geometric algorithms to explicitly estimate and integrate the focal length into the 3D pose estimation. We evaluate our approach on three challenging real-world datasets (Pix3D, Comp, and Stanford) and significantly outperform the state-of-the-art by up to 20%.

Acknowledgement This work was supported by the Christian Doppler Laboratory for Semantic 3D Computer Vision, funded in part by Qualcomm Inc. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

[2] Qualitative examples of our predicted 2D-3D correspondences are provided in the **supplementary material**.

References

[1] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6D Object Pose Estimation using 3D Object Coordinates. In *European Conference on Computer Vision*, pages 536–551, 2014. 2

[2] E. Brachmann, F. Michel, A. Krull, M. Ying Yang, S. Gumhold, and C. Rother. Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image. In *Conference on Computer Vision and Pattern Recognition*, pages 3364–3372, 2016. 2

[3] M. A. Branch, T. F. Coleman, and Y. Li. A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems. *SIAM Journal on Scientific Computing*, 21(1):1–23, 1999. 5

[4] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. Dollar. The YCB Object and Model Set: Towards Common Benchmarks for Manipulation Research. pages 510–517, 2015. 11

[5] Q. Chen, H. Wu, and T. Wada. Camera Calibration with Two Arbitrary Coplanar Circles. In *European Conference on Computer Vision*, pages 521–532, 2004. 3

[6] A. R. Conn, N. I. Gould, and P. L. Toint. *Trust Region Methods*. SIAM, 2000. 5

[7] M. Dubská, A. Herout, R. Juránek, and J. Sochor. Fully Automatic Roadside Camera Calibration for Traffic Surveillance. *IEEE Transactions on Intelligent Transportation Systems*, 16(3):1162–1171, 2015. 3

[8] O. Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, 1993. 3

[9] A. Grabner, P. M. Roth, and V. Lepetit. 3D Pose Estimation and 3D Model Retrieval for Objects in the Wild. In *Conference on Computer Vision and Pattern Recognition*, pages 3022–3031, 2018. 1, 2, 3, 4, 7

[10] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003. 2, 3

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *International Conference on Computer Vision*, pages 2980–2988, 2017. 2, 3, 4, 12

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[13] K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. In *European Conference on Computer Vision*, pages 630–645, 2016. 3

[14] J. A. Hesch and S. I. Roumeliotis. A Direct Least-Squares (DLS) Method for P*n*P. In *International Conference on Computer Vision*, pages 383–390, 2011. 5

[15] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lepetit. Gradient Response Maps for Real-Time Detection of Textureless Objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):876–888, 2011. 11

[16] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab. Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. In *Asian Conference on Computer Vision*, pages 548–562, 2012. 5

[17] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis. T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-Less Objects. In *IEEE Winter Conference on Applications of Computer Vision*, pages 880–888, 2017. 11

[18] T. Hodaň, J. Matas, and Š. Obdržálek. On Evaluation of 6D Object Pose Estimation. In *European Conference on Computer Vision*, pages 606–619, 2016. 5

[19] P. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. 3, 4

[20] O. H. Jafari, S. K. Mustikovela, K. Pertsch, E. Brachmann, and C. Rother. iPose: Instance-Aware 6D Pose Estimation of Partly Occluded Objects. In *Asian Conference on Computer Vision*, 2018. 2

[21] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab. SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again. In *International Conference on Computer Vision*, pages 1530–1538, 2017. 2

[22] A. Kundu, Y. Li, and J. M. Rehg. 3D-RCNN: Instance-Level 3D Object Reconstruction via Render-and-Compare. In *Conference on Computer Vision and Pattern Recognition*, pages 3559–3568, 2018. 2, 5

[23] V. Lepetit, F. Moreno-Noguer, and P. Fua. EP*n*P: An Accurate $O(n)$ Solution to the P*n*P Problem. *International Journal of Computer Vision*, 81(2):155, 2009. 5, 13

[24] C. Li, J. Bai, and G. D. Hager. A Unified Framework for Multi-View Multi-Class Object Pose Estimation. In *European Conference on Computer Vision*, pages 1–16, 2018. 2

[25] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature Pyramid Networks for Object Detection. In *Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 3

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, pages 740–755, 2014. 3, 6, 12

[27] S. Mahendran, H. Ali, and R. Vidal. A Mixed Classification-Regression Framework for 3D Pose Estimation from 2D Images. In *British Machine Vision Conference*, pages 238.1–238.12, 2018. 2

[28] F. Massa, R. Marlet, and M. Aubry. Crafting a Multi-Task CNN for Viewpoint Estimation. In *British Machine Vision Conference*, pages 91.1–91.12, 2016. 2

[29] J. J. Moré. The Levenberg-Marquardt Algorithm: Implementation and Theory. In *Numerical Analysis*, pages 105–116, 1978. 5

[30] R. Mottaghi, Y. Xiang, and S. Savarese. A Coarse-To-Fine Model for 3D Pose Estimation and Sub-Category Recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 418–426, 2015. 1, 2

[31] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka. 3D Bounding Box Estimation Using Deep Learning and Geometry. In *Conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017. 1, 2

[32] G. Nakano. A Versatile Approach for Solving P*n*P, P*n*Pf, and P*n*Pfr Problems. In *European Conference on Computer Vision*, pages 338–352, 2016. 2, 3, 4, 5

[33] G. Pavlakos, X. Zhou, A. Chan, K. Derpanis, and K. Daniilidis. 6-DoF Object Pose from Semantic Keypoints. In *International Conference on Robotics and Automation*, pages 2011–2018, 2017. 2

[34] A. Penate-Sanchez, J. Andrade-Cetto, and F. Moreno-Noguer. Exhaustive Linearization for Robust Camera Pose and Focal Length Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2387–2400, 2013. 3, 5, 6

[35] B. Pepik, M. Stark, P. Gehler, T. Ritschel, and B. Schiele. 3D Object Class Detection in the Wild. In *Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–10, 2015. 2

[36] M. Rad and V. Lepetit. BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects Without Using Depth. In *International Conference on Computer Vision*, pages 3828–3836, 2017. 2, 3, 4

[37] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 1, 2, 3, 4

[38] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. Tenenbaum, and W. Freeman. Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling. In *Conference on Computer Vision and Pattern Recognition*, pages 2974–2983, 2018. 2, 5, 11

[39] M. Sundermeyer et.al. Implicit 3D Orientation Learning for 6D Object Detection from RGB Images. In *European Conference on Computer Vision*, 2018. 11

[40] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer Science & Business Media, 2010. 3

[41] B. Tekin, S. N. Sinha, and P. Fua. Real-Time Seamless Single Shot 6D Object Pose Prediction. In *Conference on Computer Vision and Pattern Recognition*, pages 292–301, 2018. 2, 3

[42] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle Adjustment – A Modern Synthesis. In *International Workshop on Vision Algorithms*, pages 298–372, 1999. 5, 8

[43] R. Tsai. A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-The-Shelf TV Cameras and Lenses. *Journal of Robotics and Automation*, 3(4):323–344, 1987. 3

[44] S. Tulsiani, J. Carreira, and J. Malik. Pose Induction for Novel Object Categories. In *International Conference on Computer Vision*, pages 64–72, 2015. 2

[45] S. Tulsiani and J. Malik. Viewpoints and Keypoints. In *Conference on Computer Vision and Pattern Recognition*, pages 1510–1519, 2015. 1, 2, 5, 7

[46] Y. Wang, X. Tan, Y. Yang, X. Liu, E. Ding, F. Zhou, and L. S. Davis. 3D Pose Estimation for Fine-Grained Object Categories. In *European Conference on Computer Vision Workshops*, 2018. 1, 2, 3, 4, 5, 6, 7, 11

[47] S. Workman, C. Greenwell, M. Zhai, R. Baltenberger, and N. Jacobs. DEEPFOCAL: A Method for Direct Focal Length Estimation. In *International Conference on Image Processing*, pages 1369–1373, 2015. 3

[48] C. Wu. P3.5P: Pose Estimation with Unknown Focal Length. In *Conference on Computer Vision and Pattern Recognition*, pages 2440–2448, 2015. 3, 6

[49] Y. Wu and Z. Hu. P*n*P Problem Revisited. *Journal of Mathematical Imaging and Vision*, 24(1):131–141, 2006. 5

[50] Y. Xiang, W. Kim, W. Chen, J. Ji, C. Choy, H. Su, R. Mottaghi, L. Guibas, and S. Savarese. ObjectNet3D: A Large Scale Database for 3D Object Recognition. In *European Conference on Computer Vision*, pages 160–176, 2016. 1, 2, 11

[51] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond Pascal: A Benchmark for 3D Object Detection in the Wild. In *IEEE Winter Conference on Applications of Computer Vision*, pages 75–82, 2014. 1, 5, 11

[52] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In *Robotics: Science and Systems Conference*, pages 1–10, 2018. 2

[53] Z. Zhang. A Flexible New Technique for Camera Calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. 3

[54] Y. Zheng and L. Kneip. A Direct Least-Squares Solution to the P*n*P Problem with Unknown Focal Length. In *Conference on Computer Vision and Pattern Recognition*, pages 1790–1798, 2016. 3

[55] Y. Zheng, S. Sugimoto, I. Sato, and M. Okutomi. A General and Simple Method for Camera Pose and Focal Length Determination. In *Conference on Computer Vision and Pattern Recognition*, pages 430–437, 2014. 3

GP^{2}C: Geometric Projection Parameter Consensus for Joint 3D Pose and Focal Length Estimation in the Wild Supplementary Material

In the following, we provide additional details and qualitative results of our joint 3D pose and focal length estimation approach called *Geometric Projection Parameter Consensus* (GP^{2}C). In Sec. 6, we give an overview of the evaluated datasets and present details on the evaluation setup. In Sec. 7, we qualitatively show appearance ambiguities due to different focal lengths. In Sec. 8, we discuss parameters and strategies used for training. In Sec. 10, we present qualitative examples of our predicted 2D-3D correspondences. In Sec. 11, we show failure cases of our approach. In Sec. 9, we provide additional qualitative 3D pose and focal length estimation results of our approach. Finally, we conduct an ablation study on joint refinement in Sec. 12.

6. Datasets and Evaluation Setup

We evaluate our proposed approach for joint 3D pose and focal length estimation in the wild on three challenging real-world dataset with different object categories: Pix3D [38] (*bed*, *chair*, *sofa*, *table*), Comp [46] (*car*), and Stanford [46] (*car*). These datasets provide category-level 3D pose and focal length annotations for RGB images taken in the wild and have only been available recently.

Previous datasets were either captured using a single camera with constant focal length (category-level: KITTI or instance-level: LineMOD [15], T-LESS [17], YCB [4]), or lacked focal length annotations (category-level: Pascal3D+ [51], ObjectNet3D [50]). Due to the lack of focal length annotations, Pascal3D+ and ObjectNet3D are only meaningful for coarse 3D rotation estimation but not for fine-grained 3D pose estimation because they assume an almost orthographic camera for all images.

As a consequence of this previous lack of datasets, there is little research on 3D pose and focal length estimation in the wild [46]. Existing 3D pose estimation methods either assume the focal length to be given or evaluate on datasets which were captured using a single camera with constant focal length. However, in the wild, images are captured with multiple cameras having different focal lengths and the focal length is unknown during inference. Moreover, approaches for instance-level 3D pose estimation cannot be applied to category-level 3D pose estimation, as they assume that objects encountered during testing have already been seen during training [39].

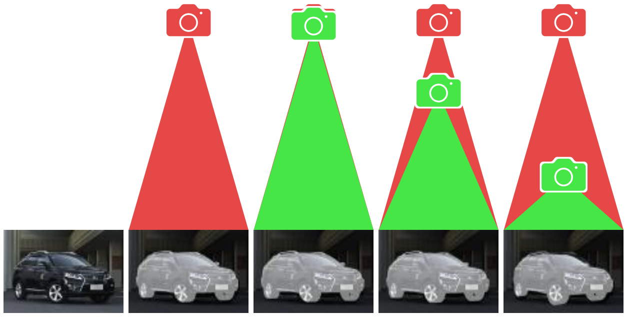


Figure 7: In the case of unknown intrinsics, the 3D pose of an object is ambiguous. Our approach finds a geometric consensus between all projection parameters, which results in a precise 2D-3D alignment for any initial focal length. However, a good initial focal length is required to compute an accurate 3D pose, as illustrated by the visualization of the object-to-camera distance.

The Pix3D dataset provides multiple categories, however, we only train and evaluate on categories which have more than 300 non-occluded and non-truncated samples (*bed*, *chair*, *sofa*, *table*). Further, we restrict the training and evaluation to samples marked as non-occluded and non-truncated, because we do not know which objects parts are occluded nor the extent of the occlusion, and many objects are heavily truncated. For each category, we select 50% of the samples for training and the other 50% for testing. To the best of our knowledge, we are the first to report results for 3D pose and focal length estimation on Pix3D.

The Comp and Stanford datasets only provide one category (*car*). Most images show one prominent car which is non-occluded and non-truncated. The two datasets already provide a train-test split. Thus, we use all available samples from Comp and Stanford for training and evaluation.

7. Appearance Ambiguities

In the main paper, we discuss appearance ambiguities resulting from different focal lengths and show the importance of the focal length for estimating 3D poses from 2D-3D correspondences quantitatively. This is also emphasized by the qualitative example shown in Figure 7. In this experiment, we initialize our geometric optimization with three different focal lengths (ground truth, predicted, and constant). We use the predicted 3D pose and focal length to project the ground truth 3D model onto the image and additionally visualize the object-to-camera distance.

Our geometric optimization finds a consensus between the individual projection parameters, which results in a precise 2D-3D alignment for any initial focal length, because we optimize the reprojection error during inference. However, the 3D pose of an object is ambiguous in the case of unknown intrinsics. Thus, a good initial focal length is a key factor in achieving high accuracy in terms of 3D translation, as can be seen from the visualization of the object-to-camera distance in Figure 7. Our predicted focal length is significantly more accurate than the best possible constant focal length, *i.e.*, the median of the training dataset.

8. Training Details

For our implementation, we resize and pad images to a spatial resolution of 512×512 maintaining the aspect ratio. In this way, we are able to use a batch size of 6 on a 12GB GPU. We train our networks for 200 epochs and employ a staged training strategy for fine-tuning a model pre-trained on COCO [26]: First, we freeze all pre-trained weights and only train our focal length and 2D-3D correspondences branches using a learning rate of $1e^{-3}$. During training, we gradually unfreeze all network layers and finally train the entire model using a learning rate of $1e^{-4}$.

We employ different forms of data augmentation commonly used in object detection [11]. In this case, some techniques like mirroring or jittering of location, scale, and rotation require adjusting the training target accordingly, while independent pixel augmentations like additive noise do not.

Balancing individual loss terms is crucial for training a multi-task network. We weight the focal loss with 0.1, the 2D-3D correspondences loss with 10.0, and the object detection loss with 1.0, however, the specific numbers are highly dependent on the implementation.

9. Qualitative Results

Figure 8 shows additional qualitative 3D pose and focal length estimation results for multiple objects in a single image. We predict 3D poses for multiple objects, however, all evaluated datasets only provide 3D pose annotations for one instance per image.

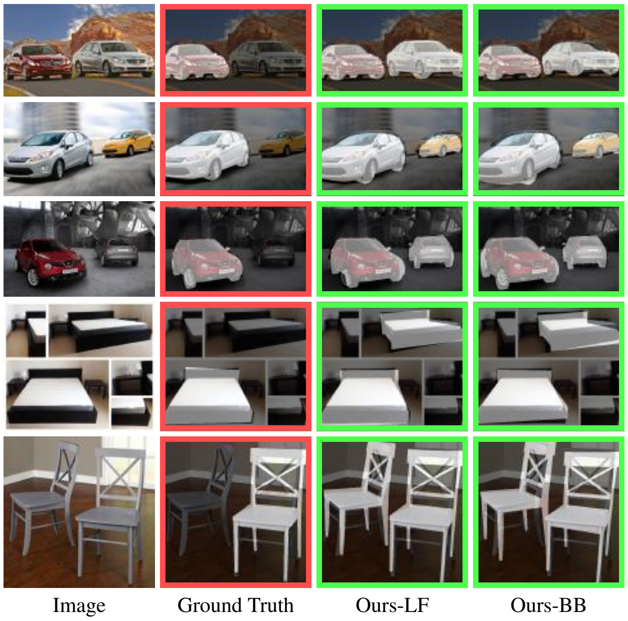


Figure 8: Additional qualitative 3D pose and focal length estimation results for multiple objects in a single image. We predict 3D poses for multiple objects (green frames), however, all evaluated datasets only provide 3D pose annotations for one instance per image (red frames).

10. Qualitative Predictions

Qualitative examples of our predicted 2D-3D correspondences are presented in Figure 9. The predicted correspondences do not contain single extreme outliers, because they are computed from a low dimensional feature embedding which produces consistent predictions. If our prediction fails entire regions of 2D-3D correspondences are corrupt. In such cases, we cannot estimate the pose correctly, not even with robust methods.

Considering our predicted location fields, we observe that the overall shape of the object is recovered very accurately. In specific cases, thin object parts and details are not detected, *e.g.*, the skinny legs of a table as shown in Figure 9. To address this issue, the spatial resolution of the predicted location field can be increased. In this work, we follow the architecture of Mask R-CNN and use a spatial resolution of 28×28 [11].

Considering our 3D bounding box corner projections, we observe that the predicted 2D locations are close to the ground truth 2D locations. Also, the perspective box-shape is well recovered and there is a consensus between the individual points. The predictions are even accurate for corners which project outside the image area, as shown in Figure 9.

			Rotation		Translation	Pose	Focal	Projection	
Method	Dataset	Class	$MedErr_R$ $\cdot 1$	$Acc_{R_{\frac{\pi}{6}}}$	$MedErr_t$ $\cdot 10^1$	$MedErr_{R,t}$ $\cdot 10^1$	$MedErr_f$ $\cdot 10^1$	$MedErr_P$ $\cdot 10^2$	$Acc_{P_{0.1}}$
Ours-LF *initial*	Pix3D	$mean$	7.10	87.9%	1.89	1.32	1.73	3.98	84.7%
Ours-LF *refined*			**6.92**	**88.4%**	**1.85**	**1.30**	**1.72**	**3.85**	**85.5%**
Ours-BB *initial*	Pix3D	$mean$	7.04	90.1%	1.98	1.33	1.77	3.87	86.8%
Ours-BB *refined*			**6.89**	**90.8%**	**1.94**	**1.30**	**1.75**	**3.66**	**88.0%**

Table 3: Ablation study on joint 3D pose and focal length refinement. We compare our initial solution to the final solution obtained by our joint refinement. Jointly optimizing all parameters results in an improvement across all metrics.

11. Failure Cases

Figure 10 shows failure cases of our approach using our two different methods for establishing 2D-3D correspondences (Ours-LF and Ours-BB). Most failure cases relate to strong truncations, heavy occlusions, or poses which are far from the poses seen during training. Naturally, the annotations are not perfect and some occluded or truncated samples are marked as non-occluded and non-truncated, or the 3D pose annotation is incorrect. In some cases, our approach makes a correct prediction, but this prediction is considered wrong because of an erroneous ground truth 3D pose annotation, as shown in Figure 10. Interestingly, there is a large overlap between the failure cases of both methods, which indicates that the respective samples are significantly different from the samples seen during training.

12. Ablation Study

Finally, Table 3 presents quantitative results of our approach with and without joint 3D pose and focal length refinement. For this purpose, we compare our initial solution obtained by EP*n*P [23] with our predicted focal length to the final solution computed by our joint 3D pose and focal length refinement. Jointly optimizing all parameters results in an improvement across all metrics. In fact, the initial solution already outperforms the state-of-the-art by a large margin.

Our geometric optimization is fast and efficient. In our implementation, the geometric optimization with joint refinement (Stage 2) takes only 5 ms, while the CNN forward pass (Stage 1) takes 60 ms per image on average.

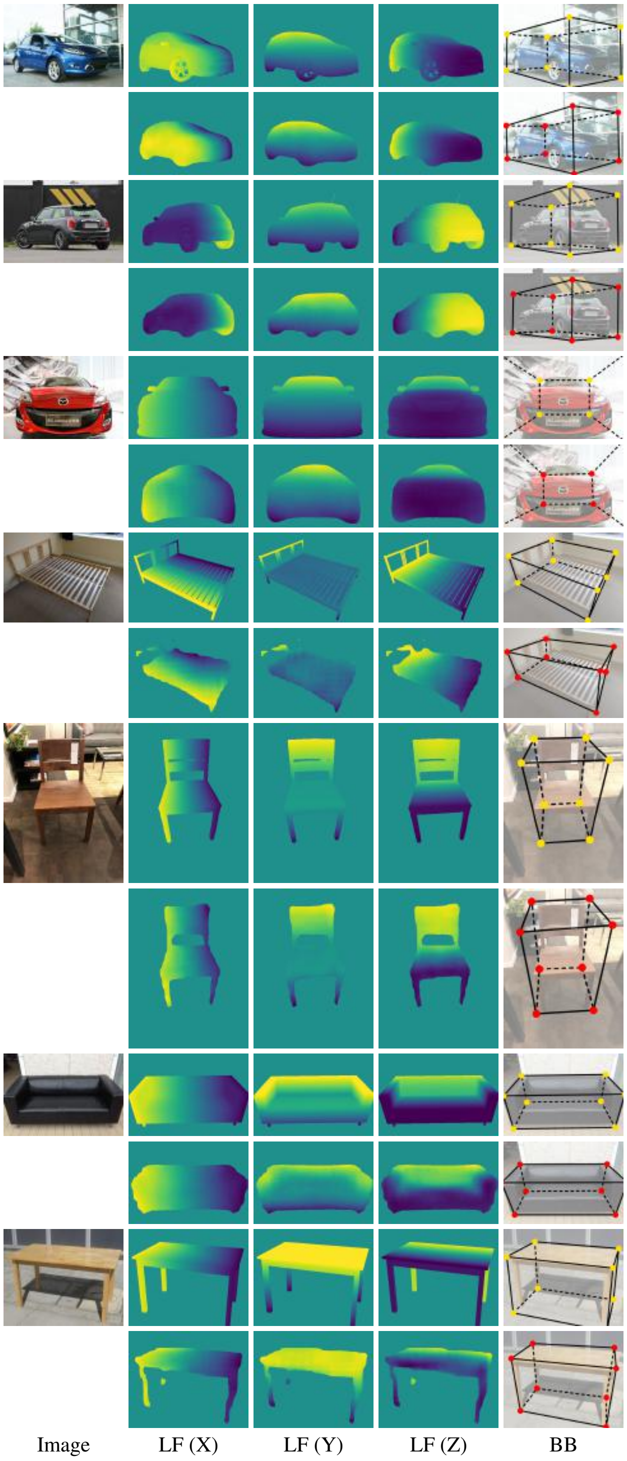
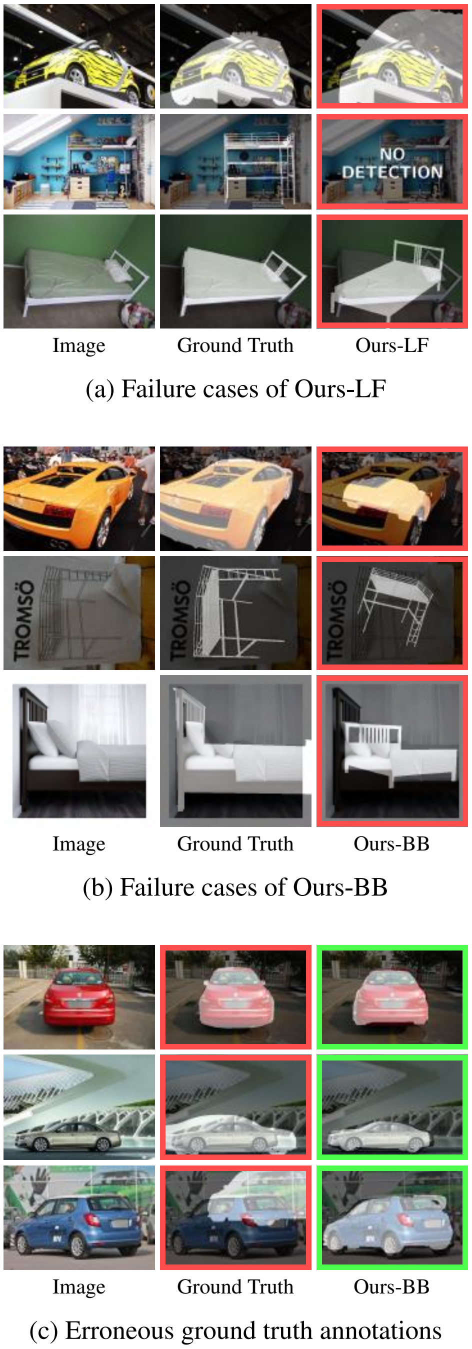


Figure 9: Qualitative examples of our predicted 2D-3D correspondences. For each object, we show two forms of 2D-3D correspondences: the location field (LF) and the projections of the object's 3D bounding box corners (BB). For each example image, the top row shows the ground truth, the bottom row shows our predictions.



(a) Failure cases of Ours-LF

(b) Failure cases of Ours-BB

(c) Erroneous ground truth annotations

Figure 10: Example failure cases of our approach for (a) Ours-LF and (b) Ours-BB. Most failure cases relate to strong truncations, heavy occlusions, or poses which are far from the poses seen during training. (c) In some cases, our approach makes a correct prediction, but the ground truth 3D pose annotation is corrupt, *e.g.*, the annotator confused the back and the front of a car or mislabeled the location of the object in the image. We highlight samples showing incorrect predictions or erroneous ground truth annotations with red frames.