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

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

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. 1, we give an overview of the evaluated datasets and present details on the evaluation setup. In Sec. 2, we qualitatively show appearance ambiguities due to different focal lengths. In Sec. 3, we discuss parameters and strategies used for training. In Sec. 4, we present qualitative examples of our predicted 2D-3D correspondences. In Sec. 5, we show failure cases of our approach. In Sec. 6, 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. 7.

1. 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 [7] (*bed*, *chair*, *sofa*, *table*), Comp [9] (*car*), and Stanford [9] (*car*). These datasets provide category-level 3D pose and focal length annotations 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 [3], T-LESS [4], YCB [1]), or lacked focal length annotations (category-level: Pascal3D+ [11], ObjectNet3D [10]). 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 [9]. 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 [8].

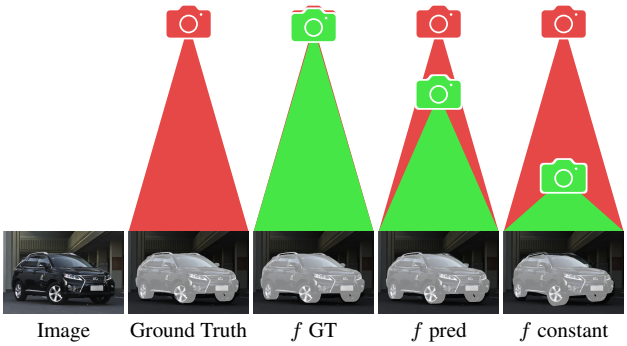


Figure 1: In the case of unknown intrinsics, the 3D pose of an object is ambiguous. Our approach finds a geometric consensus between all projection parameters, which results in a precise 2D-3D alignment for any initial focal length. However, a good initial focal length is required to compute an accurate 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.

2. 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 1. 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 1. Our predicted focal length is significantly more accurate than the best possible constant focal length, *i.e.*, the median of the training dataset.

3. 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 [6]: 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 [2]. 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.

4. Qualitative Predictions

Qualitative examples of our predicted 2D-3D correspondences are presented in Figure 3. The predicted correspondences do not contain single extreme outliers, because they are computed from a low dimensional feature embedding which tends to produce 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 3. 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 [2].

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 3.

5. Failure Cases

Figure 4 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 4. 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.

6. Qualitative Results

Figure 2 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.

7. Ablation Study

Finally, Table 1 presents quantitative results of our approach with and without joint 3D pose and focal length refinement. For this purpose, we compare our initial solu-

Method	Dataset	Class	Rotation		Translation	Pose	Focal	Projection	
			$MedErr_R$ $\cdot 1$	$Acc_{R\frac{\pi}{6}}$	$MedErr_t$ $\cdot 10^1$	$MedErr_{R,t}$ $\cdot 10^1$	$MedErr_f$ $\cdot 10^1$	$MedErr_P$ $\cdot 10^2$	$Acc_{P_{0.1}}$
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 1: 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.

tion obtained by EP*n*P [5] 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 2: 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).

References

[1] 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. 1

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

[3] 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. 1

[4] 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. 1

[5] 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. 3

[6] 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. 2

[7] 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. 1

[8] M. Sundermeyer et.al. Implicit 3D Orientation Learning for 6D Object Detection from RGB Images. In *European Conference on Computer Vision*, 2018. 1

[9] 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

[10] 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

[11] 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

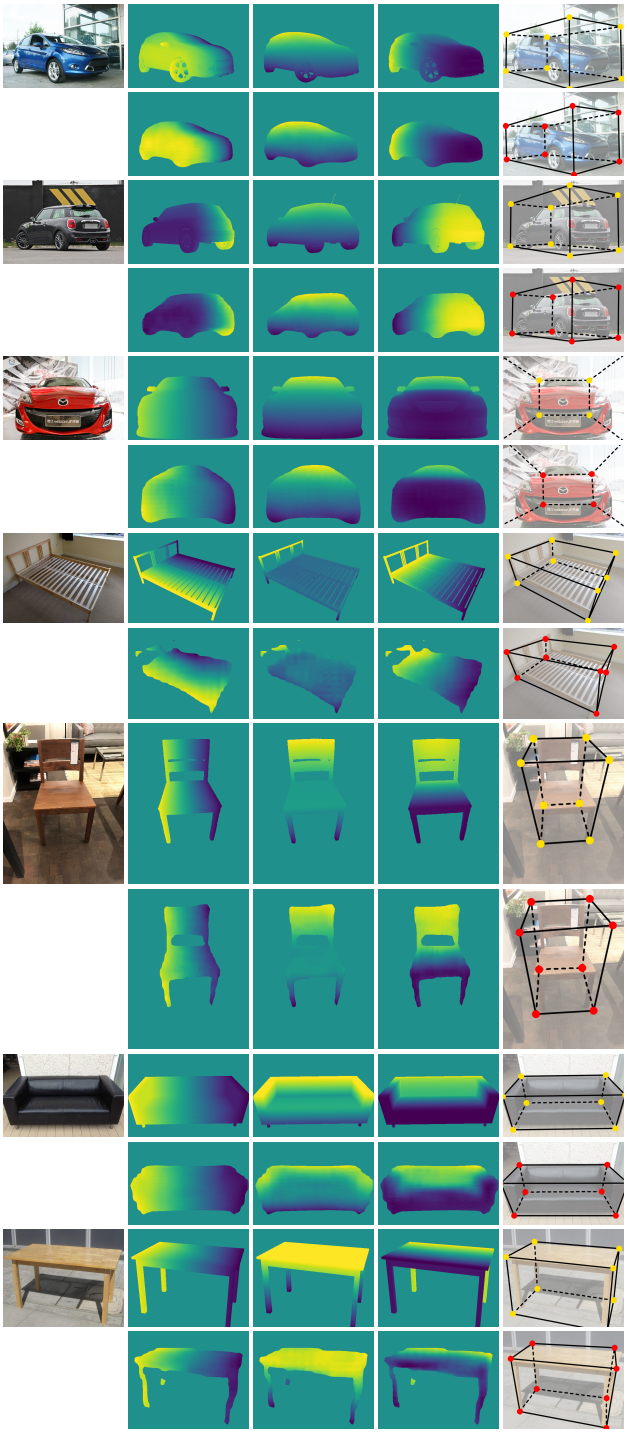
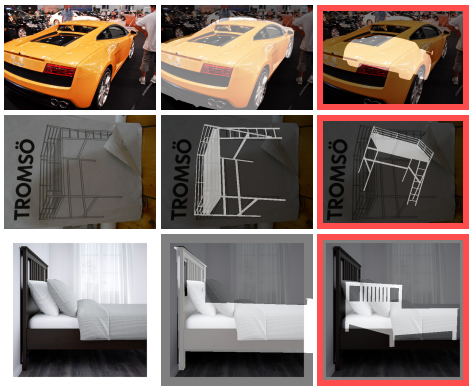
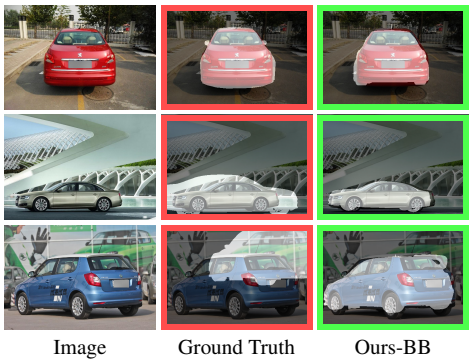


Figure 3: Qualitative examples of our predicted 2D-3D correspondences. For each object, we show two forms of 2D-3D correspondences: the location field (LF) and the projections of the object's 3D bounding box corners (BB). For each example image, the top row shows the ground truth, the bottom row shows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

(a) Failure cases of Ours-LF



(b) Failure cases of Ours-BB



(c) Erroneous ground truth annotations

Figure 4: 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.