Improving Object Pose Estimation by Fusion with a Multimodal Prior – Utilizing Uncertainty-based CNN Pipelines for Robotics

Jesse Richter-Klug, Patrick Mania, Gayane Kazhoyan, Michael Beetz and Udo Frese*

***Abstract*— Estimating the pose of an object is essential for robot manipulation. In many applications the spatial and geometric relations between the object and the other parts of the world, e.g. the relation between the object and its supporting plane, are a-priori known or can be assumed with a certain accuracy. This information can be leveraged for pose estimation. In this work, we show how this information can be formulated as multimodal prior and probabilistically fused with pose information that a CNN extracts from an image. For this purpose, the CNN pipeline from prior work is utilized. In the cases where the prior fits the ground truth, the approach is able to propel monocular results to binocular / depth data levels. Importantly, in the cases of no fitting priors, the pose estimation does not get negatively affected. The proposed method was evaluated on the T-Less dataset and used in a sample robotic application.**

I. INTRODUCTION

In everyday life as well as in the context of robotics, there is often a lot more known about an object then only its appearance. For example, if one wants to tidy up a dinner table or unload the dishwasher, all the crockery is likely to be located on the table or inside the corresponding dishwasher compartment. In other situations, objects also have relations with other points in space: pedestrians walk on the ground; during grasping, an object is in the hand or the gripper; handles are at specific locations on the doors; doors are fixed at their hinge; and so forth.

This relations can often be expressed as geometric assumptions that can be used to improve the object's pose estimation. This applies especially to algorithms that need to detect objects in an application context, as is the case with a robotic system that is physically interacting with its environment.

Figure 1 illustrates how we utilize such information in this paper. Assume we know the mug on the left side of the figure is standing on a table. We express this information as a prior distribution on the position of the object's center of the bounding box (orange) and fuse it with the distribution from the monocular pose estimation (blue). This pose estimate contains uncertainty information (blue ellipse). Uncertainties in monocular vision are highly anisotropic with the largest

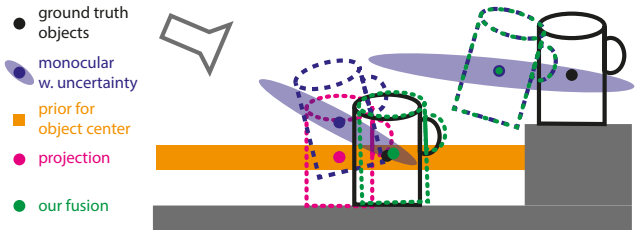

Fig. 1. Pose estimation with multi-modal prior: Two mugs (black) stand on a table and on a book on that table (grey). The monocular pose (blue) has large uncertainty in viewing direction. When the left mug is fused with a support plane prior (orange), the result is much better (green).

error in viewing direction. Taking this into account, probabilistic fusion is much better (green), than simple projection onto the supporting plane (magenta).

In general we only know that "mugs *typically* stand on the table". E.g. the left mug does, the right mug stands on a book on the table instead. To consider this case, we express the prior as a multi-modal distribution with one mode as described previously and one uniform mode. When fusing the prior for the left mug pose estimation, the supporting plane mode fits well, has the largest probability, and is used as described. For the mug that is standing on the book, the supporting plane mode fits badly, the uniform mode has the largest probability and the initial estimate stays as it is. We call this important behavior *non-degrading*.

State of the art pose estimation from RGB images is mostly done via convolutional neural networks (CNNs). Despite their exceptional results, they form nontransparent estimation algorithms where the result's (un)certainty is not apparent. A fusion without uncertainty would lie between the monocular and projected estimate and not be better.

Our previous work [1], [2] proposes a network architecture, where the CNN provides uncertainty estimates for all its (intermediate) results. With these we improved pose estimation, allowing for non-degrading depth data fusion in quasi-no-time as well as binocular and n-ocular integration, and provided the pose estimate with a meaningful uncertainty.

In this work we show how the previously proposed architecture can leverage prior information. By fusing a multimodal prior early during the pose estimation with the RGB-based information, we allow for multiple mutually exclusive assumptions to coexist while improving the monocular RGB result to be comparable or sometimes superior to the RGBD result. This is especially relevant when depth data is missing or corrupt. Our main contributions are:

- Non-degrading integration of a multimodal prior into

This research has been supported by the German Research Foundation DFG, as a part of the Collaborative Research Center (Sonderforschungsbereich) 1320 EASE - Everyday Activity Science and Engineering, University of Bremen (http://www.ease-crc.org/), Subprojects R02 and R04.

*Faculty of Mathematics and Computer Science, University of Bremen, 28359 Bremen, Germany {jesse, pmania, kazhoyan, beetz, ufrese}@uni-bremen.de

the uncertainty-based CNN pipeline from [2].
- Evaluation on the broadly used Dataset T-Less regarding performance and non-degrading.
- Showcasing practical use for robotics by picking plates from a dishwasher (monocular-only with prior).

II. Related Work

It is very common to utilize the fact that objects reside on a supporting plane in one way or another. For instance, [3] extract a ground plane to segment the object. However, this does not improve the precision of the estimate.

Many object pose estimation algorithms predict the image position of either 3D bounding box corners or dense object surface points with a CNN and run PnP [4] to obtain a pose. Examples are YOLO6D [5], BB8 [6], [7], [8], DOPE [9] and DPOD [10]. PnP is actually sensor fusion on the 2D camera measurements and usually employs least squares as a final step. However, the mentioned works treat it as a black box. This motivated us in our prior work [1], [2] to make this fusion more explicit by representing every piece of information as a Gaussian in pose space and fusing these Gaussians, *e.g.* fusing estimates from two cameras (binocular) or monocular estimates with depth information, using the same fusion model. Here, we extend this view to prior knowledge in (mixture-of)-Gaussian form.

For autonomous vehicles, [11] obtain a 3D bounding box from its 2D corners predicted by a monocular RGB CNN, by intersecting the ray corresponding to a 2D corner with the support plane. Several support plane candidates are extracted from LIDAR data and the most plausible 3D box is chosen. The central idea of that approach is to make a 2D point 3D by intersecting with the support plane, in contrast to the more general probabilistic fusion view taken in this paper.

The term geometric prior can have different meanings in different contexts. In [12] a latent representation of orientation is learned in an unsupervised manner to avoid having to label poses. A geometric prior stabilizes this process by enforcing that close rotations have close latent code.

The problem of human posture estimation, which has 25-50 degrees of freedom (DoF), is much harder than rigid object pose estimation with 6 DoF. Unobservable or hardly observable DoF are frequent in human posture estimation, *e.g.* due to occlusion, and hence the use of a prior distribution is an established practice in that community [13], [14].

III. Uncertainty-based Pose Estimation (Recap)

The approach from our previous papers [1], [2][1], on which we build here, is that CNNs are best at recognizing something in the image, whereas geometric vision provides a good analytical model for the relation between the 3D world and the 2D image quantities as well as methods for estimating 3D from 2D. With the probabilistic paradigm, we introduce uncertainty estimation by letting the CNN output a distribution and using the established probabilistic

[1] [1] proposes the uncertainty based pose estimation and evaluates the quality of its results. [2] extends the method's output structure to allow symmetric objects. Objects must be known in advance.

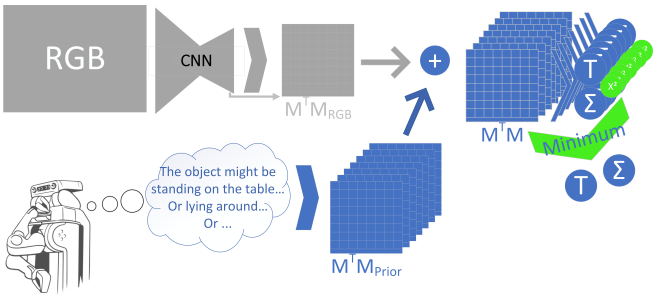

Fig. 2. Approach: A CNN predicts object points with uncertainty in every pixel, which are then aggregated into $M_{\text{RGB}}^T M_{\text{RGB}}$ a Gaussian in pose space. Multi-modal prior information is expressed in the same formalism as $M_{\text{prior}}^{jT} M_{\text{prior}}^j$ and fused by adding. Finally, in every mode the most probable pose is computed and that of the most probable mode is the final output. The blue part is the novel contribution of this paper.

fusion methods in the geometric vision. This paper is about fusion with prior information, which probabilistically means multiplying with the prior.

Thus, our approach is two-staged (see Fig. 2). In the first stage, a CNN extracts the information about what is seen in each pixel with an uncertainty estimate. This comprises the object class, the object pose, and the uncertainty as a 2D Gaussian in image space. The second stage is essentially a generalized perspective n-point solver (gPnP), computing a pose from 2D/3D point correspondences. It aggregates all pixel information into a fixed 13×13-sized matrix ($M_{\text{RGB}}^T M_{\text{RGB}}$), essentially a Gaussian in pose-matrix space ($T \in \mathbb{R}^{4\times4}$) represented in information form:

$$\begin{aligned} p(T_{\text{true}} = T|\text{RGB}) &\propto p(\text{RGB}|T_{\text{true}} = T) \\ &\propto \exp\left(-\tfrac{1}{2}\bar{T}^T (M_{\text{RGB}}^T M_{\text{RGB}})\bar{T}\right) \end{aligned} \tag{1}$$

$$\bar{T} = (\, T_{11}\ T_{12}\ T_{13}\ T_{21}\ T_{22}\ T_{23}\ T_{31}\ T_{32}\ T_{33}\ \ 1\ \ T_{13}\ T_{23}\ T_{33}\,)^T$$

Here, $\bar{T}$ is the to be estimated pose matrix T flattened into a $\mathbb{R}^{13}$ vector. For every object pixel i with coordinates u_i the CNN outputs a 3D point in object coordinates p_i^O and a 2D Gaussian uncertainty as a weighting matrix $W_i^T W_i = \Sigma_i^{-1}$. The information, that p_i^O is observed in pixel u_i with the uncertainty W_i has the mathematical form of (1) [1]:

$$M_{\text{RGB}} = \begin{pmatrix} M_{\text{RGB1}} \\ M_{\text{RGB2}} \\ \vdots \end{pmatrix},\ M_{\text{RGB}}^T M_{\text{RGB}} = \sum_i M_{\text{RGBi}}^T M_{\text{RGBi}} \tag{2}$$

$$\bar{p} = \begin{pmatrix} p_1 & p_2 & p_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & p_1 & p_2 & p_3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p_1 & p_2 & p_3 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_4 & 0 & 0 & 0 \end{pmatrix} \tag{3}$$

$$A_i = \begin{pmatrix} -f & 0 & u_{i1}-u_{01} & 0 \\ 0 & -f & u_{i2}-u_{02} & 0 \end{pmatrix} \tag{4}$$

$$0 = W_i(-f p_{i12}^C + (u_i - u_0)p_{i3}) = W_i A_i p_i^C \tag{5}$$

$$= W_i A_i T_R^C \bar{p}_i^O \bar{T} := M_{\text{RGBi}} \bar{T} \tag{6}$$

The matrix $\bar{p}$ expresses the transformation of a point p from object coordinates to reference coordinates as a function of the pose T, i.e. $\bar{p}\bar{T} = Tp$. T_R^C transforms from reference to camera coordinates. This is needed for multi-camera systems, here $T_R^C = I$. The matrix A_i expresses the perspective transform with focal length f and image center u_0. The

M_{RGBi} are then stacked, which is equivalent to adding the $M_{\text{RGBi}}^T M_{\text{RGBi}}$ together.

We will formalize several pieces of information in the form of (6), with a notation as in (5). The $0 = \dots$ means that the following expression is a standard normal distributed residual, linear in the pose T, which is then algebraical transformed into a matrix times $\bar{T}$, the flattened T.

The Gaussian in pose space $M_{\text{RGB}}^T M_{\text{RGB}}$ is valid in the full range of orientations, not only locally around one reference pose, as with a 6D Gaussian in $SE(3)$-tangential space. Hence, $M^T M$ can represent a partially defined pose. Orthonormality, *i.e.* $T \in SE(3)$, needs to be ensured separately.

To fuse different cues, *e.g.* a second camera, the depth channel of an RGBD camera, or the prior distributions considered here, these matrices can be added, effectively multiplying distributions. Here the fixed 1-entry in $\bar{T}_{10}$ allows to represent information that defines the scale.

The pose is then estimated by minimizing over $SE(3)$:

$$\hat{T} = \underset{T \in SE(3)}{\arg\min} \, \|f(T)\|^2 = \underset{T \in SE(3)}{\arg\min} \, \bar{T}^T (M_{\text{RGB}}^T M_{\text{RGB}}) \bar{T} \quad (7)$$

$$= \lim_k T_k, \quad T_{k+1} = T_k \boxplus \overset{\text{linearized}}{\underset{\delta \in \mathbb{R}^6}{\arg\min}} \, \|f(T_k \boxplus \delta)\|^2 \quad (8)$$

Gauss-Newton with $\boxplus$-manifolds [15] are utilized to ensure the orthonormality of the rotation matrix, refer to [1] for details. Afterwards, a 6D-Gaussian in $SE(3)$-tangential space is computed as the more commonly used uncertainty output.

One note on the uncertainties W_i: the CNN is normally overconfident as it learns these on the training data. This is compensated by normalizing the resulting $M_{\text{RGB}}^T M_{\text{RGB}}$ to have the theoretically expected minimum. The theory, however, assumes statistically independent pixels which is not true in reality and the CNN learns a correction factor for that [1]. Here, these steps are applied before the fusion.

The architecture also handles symmetric objects by using a symmetry specific representation for the p_i^O [2] and converting back when forming the $M_{\text{RGB}}^T M_{\text{RGB}}$. This is transparent to the fusion and so not discussed here.

IV. Method

In this paper we fuse prior information, such as a supporting plane, by expressing it in the form of (1) and adding it to the $M_{\text{RGB}}^T M_{\text{RGB}}$ matrix. We extend this approach to multimodal priors by doing this for every mode and taking the overall minimum. For that, $M_{\text{RGB}}^T M_{\text{RGB}}$ is only computed once (Fig. 2).

Any prior information can be fused with the CNN-based perspective measurements by simply adding 13×13-sized matrices. There is one precondition: it should be possible to express the prior information in the form shown in (1), *i.e.* the prior should be a square of something linear in $\bar{T}$.[2]

$$p(T_{\text{true}} = T | \text{prior}) \propto \exp\Big(-\tfrac{1}{2} \bar{T}^T (M_{\text{prior}}^T M_{\text{prior}}) \bar{T}\Big) \quad (9)$$

$$M_{\text{prior}} = \begin{pmatrix} M_{\text{prior}1} \\ M_{\text{prior}2} \\ \vdots \end{pmatrix}, \; M_{\text{prior}}^T M_{\text{prior}} = \sum_i M_{\text{prior}i}^T M_{\text{prior}i}, \quad (10)$$

The posterior is then obtained by multiplying the distributions, respectively adding the matrices:

$$p(T_{\text{true}} = T | \text{RGB}, \text{prior}) \quad (11)$$

$$\propto p(\text{RGB} | T_{\text{true}} = T) p(T_{\text{true}} = T | \text{prior}) \quad (12)$$

$$\propto \exp\Big(-\tfrac{1}{2} \bar{T}^T (M_{\text{RGB}}^T M_{\text{RGB}} + M_{\text{prior}}^T M_{\text{prior}}) \bar{T}\Big) \quad (13)$$

We now derive a toolbox of priors for different situations and then explain how to get the final estimate from the posterior.

A. Prior that fixes a given object vector

The mathematical most straightforward prior expresses that a given vector p^O in object coordinates equals another given vector p'^C in camera / reference coordinates with uncertainty Σ. One example is a spherical joint, where p is a point vector with a trailing 1. A more common example is that some axis of the object points upwards, where p is a direction vector with a trailing 0.

$$0 = W\left(p^C - p'^C\right) = W\left(\bar{p}^O \bar{T} - p'^C\right) = \quad (14)$$

$$W\left(\bar{p}^O - \begin{pmatrix} 0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,-p_1'^C\;0\;0\;0 \\ 0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,-p_2'^C\;0\;0\;0 \\ 0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,-p_3'^C\;0\;0\;0 \\ 0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,-p_4'^C\;0\;0\;0 \end{pmatrix}\right) \bar{T} =: M_{\text{prior}i} \bar{T}$$

$$W^T W = \Sigma^{-1} \quad (15)$$

Two of these priors can be used to implement a hinge joint.

B. Supporting Plane Prior

The most practically relevant prior expresses that an object rests on a supporting plane. It is not only often applicable and easy to come by but also greatly complements the RGB-based information. The supporting plane is usually at a certain angle with the viewing direction and, therefore, the fusion generates depth information, which is normally uncertain for monocular setups but essential for manipulation.

The rest position of an object can be seen as the distance between an arbitrary object point and the plane in combination with the angles from two object axes and the plane normal (cf. Fig. 3). Note that for rotationally invariant objects the second angle is not applicable. In addition, an object point from inside the rotational axis must be selected as well as the angle between the rotational axis and the plane.

1) Position: The distance d_p of an object point p^O to a plane $\{p | n \cdot p - d = 0\}$ in camera (in general reference) coordinates is defined in Hessian normal form with a given uncertainty σ_p.

$$0 = w\left(n \cdot p^C - d - d_p\right) = w(\,n_1\;n_2\;n_3\;-d-d_p\,)\,\bar{p}^O \bar{T} \quad (16)$$

$$=: M_{\text{prior}_p i} \bar{T}, \; w = \tfrac{1}{\sigma_p} \quad (17)$$

[2] If the information involves additional unknowns, this can be handled by increasing the size of the $M^T M$ matrices and generalizing (8).

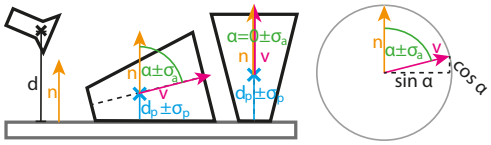

Fig. 3. The object center has a specified distance $d_p \pm \sigma_p$ to the support plane $\{p|n \cdot p - d = 0\}$. For a rolling object, surface normal n and rotation axis v have a specified angle $\alpha \pm \sigma_a$. For a resting object, n and the rest-surface normal v coincide ($\alpha = 0 \pm \sigma_a$).

2) Orientation: The angle α between an object axis (direction vector) v^O and the plane's normal n is defined over their sine and cosine with a given uncertainty σ_a. The cosine based prior is defined as

$$\bar{v} = \begin{pmatrix} v_1 & v_2 & v_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & v_1 & v_2 & v_3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & v_1 & v_2 & v_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \tag{18}$$

$$0 = w\left(n \cdot v^C - \cos\alpha\right) = w\begin{pmatrix} n_1 & n_2 & n_3 & -\cos\alpha \end{pmatrix} \bar{v}^O \bar{T} \tag{19}$$

$$=: M_{\text{prior}_c} \bar{T},\ w = \tfrac{1}{\sigma_a}. \tag{20}$$

Theoretically (20) needs a factor of $\frac{1}{\cos' \alpha} = \frac{1}{\sin \alpha}$ to represent a standard deviation of σ_a in α. However, that would become singular for $\alpha \to 0$. By omitting the factor, we weigh the information implicitly by $\sin^2 \alpha$ in $M_{\text{prior}}^T M_{\text{prior}}$. A complementary sine based prior weighted by $\cos^2 \alpha$ (in $M_{\text{prior}}^T M_{\text{prior}}$) is needed to cover the important range around $\alpha = 0$. However, $0 = |n \times v^C| - \sin\alpha$ is non-linear. Therefore, we express $0 = n \times v^C$ and add $\sin\alpha$ to the uncertainty. Intuitively, this prior expresses $|n \times v^C| \leq \sin\alpha + \sigma_a$ instead of $|n \times v^C| = \sin\alpha \pm \sigma_a$, which is weaker. However, due to the weighting it is only relevant for small α and most important for the $\alpha = 0$ case (Fig. 3).

$$0 = w\left(n \times v^C\right) = w\begin{pmatrix} 0 & -n_3 & n_2 & 0 \\ n_3 & 0 & -n_1 & 0 \\ -n_2 & n_1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \bar{v}^O \bar{T} \tag{21}$$

$$=: M_{\text{prior}_s} \bar{T},\ w = \frac{\cos\alpha}{\sqrt{\sigma_a^2 + \sin^2\alpha}} \tag{22}$$

C. Uniform prior

It is convenient to also describe "no prior" as a prior in the same formalism. This is done by defining

$$M_{\text{prior}}^T M_{\text{prior}} = \theta J^{10,10} \ \Rightarrow\ \bar{T} M_{\text{prior}}^T M_{\text{prior}} \bar{T} = \theta. \tag{23}$$

The only non-zero entry $(M_{\text{prior}}^T M_{\text{prior}})_{10,10}$ binds to the fixed 1-entry in $\bar{T}$, so $M_{\text{prior}}^T M_{\text{prior}}$ behaves like a "uniform distribution" with a constant χ^2-error of θ. θ acts as a threshold and will be discussed later.

D. Multi-modal Prior

In reality, most objects have multiple rest positions, each equally likely. In addition, sometimes an object may not adhere to the priors, *e.g.* being on top of another object, which is another uniform mode. These modes form a mixture model, where each mode, is Gaussian or uniform:

$$p(T_{\text{true}} = T|\text{prior}) = \sum_{j=0}^{m} \exp\left(-\tfrac{1}{2}\bar{T}^T (M_{\text{prior}}^{jT} M_{\text{prior}}^j)\bar{T}\right), \tag{24}$$

The superscript ^{j} enumerates m modes with "or-logic", *i.e.* one of them is asserted to be true. Each mode itself can consist of several priors enumerated by subscript $_i$, which have "and-logic", *i.e.* all are asserted to be true in that mode. The uniform prior is $j = 0$.

By distributive law, when multiplying the mixture prior (24) with $p(\text{RGB}|T_{\text{true}} = T)$, the result is again a mixture with each mode multiplied. Therefore, practically, we fuse the $M_{\text{RGB}}^T M_{\text{RGB}}$ from the image CNN with the respective $M_{\text{prior}}^{jT} M_{\text{prior}}^j$ of every mode j by adding.

$$p(T_{\text{true}} = T|\text{Prior}, \text{RGB}) = \tag{25}$$

$$\sum_{j=0}^{m} \exp\left(-\tfrac{1}{2}\bar{T}^T (M_{\text{prior}}^{jT} M_{\text{prior}}^j + M_{\text{RGB}}^T M_{\text{RGB}})\bar{T}\right), \tag{26}$$

E. Computing an estimate from the posterior

We compute the most likely (mode, pose) pair:

$$\widehat{\text{mode}}, \hat{T} = \underset{j, T \in SE(3)}{\arg\max}\ p(\text{mode}_{\text{true}} = j, T_{\text{true}} = T|\text{RGB}, \text{prior})$$

$$= \underset{j, T \in SE(3)}{\arg\max}\ \exp\left(-\tfrac{1}{2}\bar{T}^T (M_{\text{prior}}^{jT} M_{\text{prior}}^j + M_{\text{RGB}}^T M_{\text{RGB}})\bar{T}\right)$$

$$= \underset{j, T \in SE(3)}{\arg\min}\ \bar{T}^T (M_{\text{prior}}^{jT} M_{\text{prior}}^j + M_{\text{RGB}}^T M_{\text{RGB}})\bar{T} \tag{27}$$

Ultimately, this means (see Fig. 2) fusing every $M_{\text{prior}}^{jT} M_{\text{prior}}^j$ with the $M_{\text{RGB}}^T M_{\text{RGB}}$, computing its minimum via the gPnP algorithm [1, eq. 23] and choosing the one with the smallest value, *i.e.* χ^2-error. This value reflects how consistent the prior fits to the RGB information, choosing the most consistent one. For the uniform prior $j = 0$, this value is the original χ^2-error of $M_{\text{RGB}}^T M_{\text{RGB}} + \theta$. This means, if fusing with a prior increases the error by more than θ, the uniform prior is more likely. Thus, effectively θ is a Mahalanobis-threshold gate for the prior [3]. In Fig. 1, this mechanism fuses the left cup pose estimate with the supporting plane prior but keeps the right cup estimate unmodified.

Note that this procedure computes the maximum likelihood T, if (26) is written as $\max_{j=0}^{m}$ instead of $\sum_{j=0}^{m}$, following Olson's max-mixture approach [16]. An computational more expensive alternative would be the maximum of (26) or the mean in pose space. However, we do not think it is practically useful to let all modes contribute to the estimate.

F. Automatic Generation of Priors from CAD models

The priors defined in IV-B have two parameters: a plane defined in the camera / reference coordinate frame and an object-dependent rest position (Fig. 3). This leads to a multimodal prior with one mode per rest position and plane plus the uniform prior. Often, the plane is known in the world frame and is transformed to the camera frame. The error introduced by this transformation is often the largest part of σ_p, σ_a.

Rest positions are object-dependent constants, which we generated as follows: per physics engine we dropped single

[3]This is because the $M^T M$s only represent the Mahalanobis-distance term of the Gaussians. If the normalizer term $-\ln\det(2\pi\Sigma)$ was included in $(M^T M)_{10,10}$, θ would be a density of objects in pose space.

objects multiple times with arbitrary initial rotations and recorded the rest position according to IV-B. For objects with small pedestals we added the intended rest position by hand.

For every rest position, we add a distance prior on the object's origin. If a rotation axis exists, the origin lies on that axis and we select it for the first orientation prior. Else, we arbitrarily select the z-axis. For the second one we select an orthogonal axis to the first (if applicable). The recorded rest positions were clustered by mean shift. The variance inside each cluster (very small in our experiments) and the potential discrepancy between the object model and the real object were approximated and represented in σ_p, σ_a.

G. Further Applications

One can generalize the approach for related pose estimation problems, *i.e.* situations, where the solution is difficult to find with images alone but benefits from additional information (that is linear in T).

For example, prior knowledge can distinguish objects of same appearance but different sizes by providing depth information. Object sizes are modeled as modes in (6) and every combination of them and the prior modes is evaluated.

V. Evaluation

We evaluate our method on the T-Less Dataset [17] according to combined average recall defined in [18]. This broadly used dataset has 30 object classes in 20 test scenes. In 55.3% of all (test set) appearances, the object rests on the main supporting plane. In the remaining 44.7%, objects lean against each other or lie on top of other objects. Therefore, this dataset is perfectly suitable for evaluating the prior integration with and without a fitting prior.

The prior integration was implemented on top of the unchanged CNN from [2], making the RGB information, including its uncertainty value, directly comparable. The χ^2-limit, when a prior is applied was $\theta := 2$. As prior information one support plane was given that is based on the known pose of the robot camera in the world. We guessed the uncertainties as $\sigma_p := 4mm$ and $\sigma_a := 15°$, the systematic study in Sec. V-B was done later.

A. Performance

Table I shows the results according to the evaluation standard from the BOB-challenge [18]. For reference we list the respective best performing entries [19], [20], [21]. We show overall average recall (the challenge's main score) and the average recalls of MSSD (error in space, relevant for grasping) and MSPD (error in image, relevant for monocular systems). All recalls count the percentages with a pose error below different thresholds. These are then averaged over the thresholds.

For object samples that fit a prior (✓), the RGB result is improved greatly and it is slightly better than depth fusion. This confirms our belief that the supporting plane (even if only known approximately) primarily provides depth.

If no prior fits (✗), the results are basically untouched despite the prior fusion. Importantly, the prior information does not deteriorate the RGB information. This indicates that mostly the prior information is correctly discarded if not fitting to the RGB-based uncertainty estimate. When the RGB information is very uncertain, also false priors can fit. This can deteriorate, but also improve the result, because the wrong prior may still be better than the RGB information. This explains the slight improvement on ✗-samples.

TABLE I

Comparison of RGB(+D) data results with and without "standing on the table"-prior with state-of-the-art results for reference. Average Recalls are provided for (All) objects that form the T-Less test set as well as for only those that do (✓) or do not (✗) fit any given prior.

T-Less		RGB [2]	RGB [2] + **Prior**	RGB [19]	RGB[1] [20]	RGB + D [2][2]	RGB + D [2] + **Prior**	RGB-D [21]	RGB[1]+D [20]
AR MSSD	✗	0.421	**0.451**	-	0.652	0.689	0.687	-	0.719
	✓	0.516	**0.772**	-	0.729	0.761	0.770	-	0.772
	All	0.474	**0.628**	0.418	0.695	0.729	0.733	0.689	0.749
AR MSPD	✗	0.727	**0.742**	-	0.792	0.749	0.750	-	0.740
	✓	0.795	**0.831**	-	0.845	0.814	0.818	-	0.788
	All	0.765	**0.791**	0.674	0.821	0.785	0.788	0.696	0.767
AR	✗	0.504	**0.531**	-	0.691	0.684	0.684	-	0.679
	✓	0.590	**0.770**	-	0.759	0.755	0.762	-	0.718
	All	0.552	**0.663**	0.490	0.728	0.723[2]	0.727	0.655	0.701

[1] [20] uses RGB with RGB-based iterative refinement.
[2] The original AR value of 0.651 stated in [2] was flawed due to erroneous depth fusion. After correction, [2] reaches the 0.723 stated here.

With good depth data, the added prior knowledge is redundant, but does not degrade the pose estimates. Therefore, the combination can be used for adding robustness to difficult objects, *e.g.* glasses.

The focus of this work is on the addition of geometric priors, *i.e.* the difference to [2]. Nevertheless, our method outperforms the state-of-the-art except for [20] which utilizes RGB-based iterative refinement to achieve its outstanding result.

We conclude that our proposed prior integration allows for greatly improved pose estimates in the absence of (good) depth data without introducing any drawbacks.

B. Prior Uncertainty Selection

The uncertainty of any supporting plane prior is seldom exactly known, instead the anticipated error must be approximated. In Fig. 4 the change of performance is recorded while varying either σ_p or σ_a with the other σ staying at the manual guess ($\sigma_p = 4$mm, $\sigma_a = 15°$) made before.

One can see that our guess was pretty close to the optimum, but also that there is a large area where one can reach similar results. As it is common for probabilistic fusion, a looser distribution is more forgiving. Within the range of $[2.3mm,\ 38.5mm]$ for σ_p and $[6°,\ 200°]$ for σ_a performance drops by only 1%.

Therefore, we conclude that our prior integration is able to provide a major improvement for RGB-based pose estimates – independent of a perfect uncertainty estimate.

C. Computation Time

Our gPnP-call takes on average $\approx 0.74\,ms$ to compute, which has to be called once per prior included in the

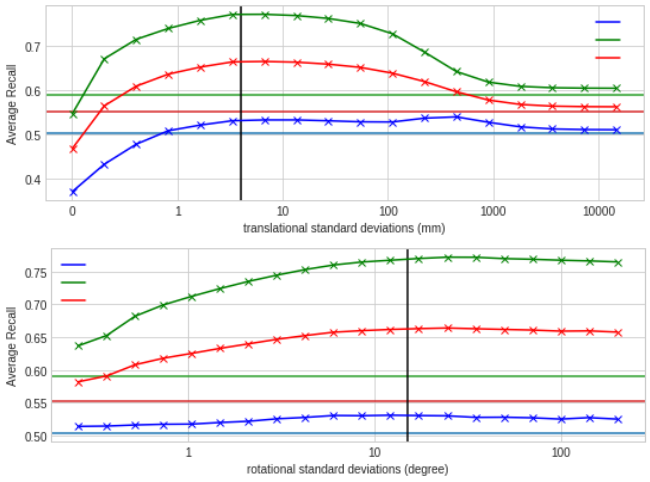

Fig. 4. Performance on T-Less of RGB+Prior dependent on given prior covariance. Isolated variation of σ_p (upper) and σ_a (lower). Average recall for all objects (All/red), only objects with (✓/green) or only objects without (✗/blue) available prior with their respective RGB-only baseline as horizontal. Our manual guess is Indicated by the vertical.

Fig. 5. Qualitative evaluation (left): Robot empties the dishwasher, based on RGB data fused with the knowledge of the dishwasher compartments (as support plane); Detecting object in Hand with gripper pose as prior knowledge (right) - quantitative evaluation against RGB and RGB-D data. The whole task can be seen in the accompanying video.

respective mixture model. For T-Less there are 3 and 17 modes per object. In total, the added cost for our prior integration is between $\approx 2.22\,ms$ and $\approx 10.36\,ms$ per object. In comparison, the pose estimation CNN call costs $\approx 86\,ms$ per object without the bounding box generation.

VI. Robotic Application

We argue that our method of combining prior knowledge with the results from a CNN greatly benefits robotic manipulation. In this section, we take a closer look at two example applications: picking plates from dishwasher compartments and object in robot hand detection. The common ceramic plate as an example object was present in two different sizes to accommodate our proposal of IV-G (Fig. 5). All demonstrations were performed by a PR2 robot within the EASE kitchen environment.

A. Learning Procedure and Implementation Details

1) Training Images: An object model was hand-crafted for both plates and used to generate photorealistic training images. Following [18], the plates were sampled inside a cube with random CC0 textures as background. Instead of dropping the objects, we randomly placed them somehow vertically on the surface – like plates inside dishwashers.

2) CNN Usage: Two CNNs were then learned solely based on these simulated images. The first is a variant for segmentation and bounding box generation. The second is a pose estimation network following the approach of [2]. Because of the very similar appearance of the two plates, they were learned indistinguishable from one another.

Since the bigger plate is now indistinguishable from the smaller plate, during inference we generate two pose estimates per detected object as described in Section IV-G.

3) Robot Demonstrator: Our demonstrator is based on the robot control framework described in [22] with three key parts: CRAM [23] is a high-level planning system, tasked with the parametrization and execution of abstract action descriptions. The system contains geometric environmental features and semantics to infer the relevant entities for manipulation tasks. Motion planning is done by Giskard [22], which is transforming a goal-based motion description from CRAM into a quadratic programming constraint optimization problem to calculate the desired joint velocities. The approach presented in this paper is integrated into RoboSherlock [24], which is a knowledge-based robot perception framework, featuring a multi-expert strategy to analyse sensor data. Perception tasks in RoboSherlock are generally formulated as queries, which are generated by CRAM during task execution. The selection and application of suitable experts is then automatically inferred from the incoming task descriptions.

B. Picking Plates from the Dishwasher

The picking plates from dishwasher scenario was selected because it provides a challenging use case for robot manipulation and perception. Tightly packed plates in a dishwasher require precise pose estimations due to the small space between the plates which the robot gripper has to enter while achieving a firm grasp on the slippery, rigid plate. This is especially challenging for perception, because views of the dishwasher feature lots of occlusions and the compartments are not rigid and not continuous supporting surfaces (which would be otherwise easily detected).

1) Prior Knowledge: In the selected dishwasher, there are two rows of natural plate placing positions: the left side in the top compartment and the front side of the bottom compartment (here plates can be orientated in two opposite ways). In each case, we defined a (horizontal) virtual support plane, in which all plate origins approximately lie, relative to our model of the kitchen. Transformations between this model and the camera frame are calculated by applying forward kinematics combined with the robot pose from the localization. We guessed the prior's uncertainty as $3cm$ since the load slightly pushes down the compartment.

2) Procedure: Starting from random viewpoints around the dishwasher, we take one image and estimate the pose of all seen objects. Without taking another image, the robot will

TABLE II

STD. DEVIATION (MM) OF ESTIMATE WHILE ROTATING GRASPED PLATE

Position	1	2	3	4
RGB	17.2	47.7	20.0	9.7
RGBD	18.0	62.5	25.3	151.4
RGB+Prior	14.6	41.4	16.5	8.6

then attempt to grasp one randomly selected object. Grasp points are pre-defined relative to the object frame. Motion planning is done by a projection-based coarse reachability calculation in CRAM, taking into account the collision bodies in the scene as well as the geometry of the plates. The selected approach direction for grasping is then used for the calculation of the complete grasping motion trajectory in Giskard. In our experiments we evaluated if the robot was able to grasp plates based on the pose estimations given by our prior-based method. We count a grasp as successful if the PR2 could grasp and lift a plate out of the dishwasher. Misgrasped or slipped plates count as failed attempts.

3) Results: From 56 tries, the robot was successful in 45 (80%). Failure cases were: missing (4 cases) or false (3 cases) segmentation, orientation error (3 cases) or false prior use (1 case, a small plate from the top compartment was perceived as a big plate in the bottom compartment).

C. Object in Hand

With an object grasped the robot can move without changing the object-in-hand pose. This allows evaluating the consistency of the estimated poses, by comparing the pose variances in the RGB-only, RGBD and RGB+prior pipelines.

1) Prior Knowledge: For this experiment an imaginary support plane was added between the gripper fingers. Priors were formed as if the plate has a rest position based on the approximated grasping position. Since the gripper joints have a π-symmetry, the prior was duplicated along this symmetry.

2) Procedure: After grasping a plate tightly, the hand was set to an arbitrary position in front of the camera. The robot then rotated its wrist step-wise for a total of π. After each step the pose was detected by all three pipelines.

3) Results: Table II lists the measured standard deviations per position. The variances differ, but RGB+Prior is slightly more consistent than RGB only. The RGBD pipeline is prone to false depth data, in particular at position 4, which was below the minimum depth range.

VII. CONCLUSIONS

In this work, we show how to utilize the uncertainty-based CNN pipeline proposed in [1] to integrate multimodal priors. We show that by defining as much as a support plane prior, RGB-only results can be propelled to reach or be slightly above RGBD level if the prior is applicable. At the same time, we also show that out method does not suffer drawbacks in case the prior is inapplicable – the performance merely drops back to RGB-only level.

In addition, we showcase practical performance in a robotic application, namely, removing plates from dishwasher compartments.

REFERENCES

[1] J. Richter-Klug and U. Frese, "Towards meaningful uncertainty information for CNN based 6D pose estimates," in *International Conference on Computer Vision Systems*. Springer, 2019, pp. 408–422.

[2] ——, "Handling object symmetries in CNN-based pose estimation," in *International Conference on Robotics and Automation*. IEEE, 2021.

[3] M. Schwarz, H. Schulz, and S. Behnke, "Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features," in *ICRA*. IEEE, 2015, pp. 1329–1335.

[4] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate O(n) solution to the pnp problem," *IJCV*, vol. 81, no. 2, p. 155, 2009.

[5] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6d object pose prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 292–301.

[6] 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 *ICCV*, 2017, pp. 3828–3836.

[7] A. Crivellaro, M. Rad, Y. Verdie, K. Moo Yi, P. Fua, and V. Lepetit, "A novel representation of parts for accurate 3d object detection and tracking in monocular images," in *ICCV*, 2015, pp. 4391–4399.

[8] M. Oberweger, M. Rad, and V. Lepetit, "Making deep heatmaps robust to partial occlusions for 3d object pose estimation," in *ECCV*, 2018, pp. 119–134.

[9] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," *arXiv preprint arXiv:1809.10790*, 2018.

[10] S. Zakharov, I. Shugurov, and S. Ilic, "DPOD: Dense 6D pose object detector in RGB images," *arXiv preprint arXiv:1902.11020*, 2019.

[11] A. Rangesh and M. M. Trivedi, "Ground plane polling for 6dof pose estimation of objects on the road," *IEEE Transactions on Intelligent Vehicles*, vol. 5, no. 3, pp. 449–460, 2020.

[12] Y. Wen, H. Pan, L. Yang, and W. Wang, "Edge enhanced implicit orientation learning with geometric prior for 6d pose estimation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4931–4938, 2020.

[13] J. Chen, S. Nie, and Q. Ji, "Data-free prior model for upper body pose estimation and tracking," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4627–4639, 2013.

[14] A. M. Lehrmann, P. V. Gehler, and S. Nowozin, "A non-parametric bayesian network prior of human pose," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1281–1288.

[15] C. Hertzberg, R. Wagner, U. Frese, and L. Schröder, "Integrating generic sensor fusion algorithms with sound state representations through encapsulation of manifolds," *Information Fusion*, vol. 14, no. 1, pp. 57–77, 2013.

[16] E. Olson and P. Agarwal, "Inference on networks of mixtures for robust robot mapping," *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 826–840, 2013.

[17] T. Hodaň, 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," *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017.

[18] T. Hodaň, M. Sundermeyer, B. Drost, Y. Labbé, E. Brachmann, F. Michel, C. Rother, and J. Matas, "BOP challenge 2020 on 6D object localization," *European Conference on Computer Vision Workshops (ECCVW)*, 2020.

[19] Z. Li, G. Wang, and X. Ji, "Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7678–7687.

[20] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, "Cosypose: Consistent multi-view multi-object 6d pose estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 574–591.

[21] R. König and B. Drost, "A hybrid approach for 6dof pose estimation," in *European Conference on Computer Vision*. Springer, 2020.

[22] G. Kazhoyan, S. Stelter, F. K. Kenfack, S. Koralewski, and M. Beetz, "The robot household marathon experiment," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[23] M. Beetz, L. Mösenlechner, and M. Tenorth, "Cram — a cognitive robot abstract machine for everyday manipulation in human environments," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 1012–1017.

[24] M. Beetz, F. Bálint-Benczédi, N. Blodow, D. Nyga, T. Wiedemeyer, and Z.-C. Márton, "Robosherlock: Unstructured information processing for robot perception," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 1549–1556.