Uncertainty-Aware Knowledge Distillation for Compact and Efficient 6DoF Pose Estimation

Nassim Ali Ousalah[1], Anis Kacem[1], Enjie Ghorbel[1,2], Emmanuel Koumandakis[3] and Djamila Aouada[1]

Abstract— **Compact and efficient 6DoF object pose estimation is crucial in applications such as robotics, augmented reality, and space autonomous navigation systems, where lightweight models are critical for real-time accurate performance. This paper introduces a novel uncertainty-aware end-to-end Knowledge Distillation (KD) framework focused on keypoint-based 6DoF pose estimation. Keypoints predicted by a large teacher model exhibit varying levels of uncertainty that can be exploited within the distillation process to enhance the accuracy of the student model while ensuring its compactness. To this end, we propose a distillation strategy that aligns the student and teacher predictions by adjusting the knowledge transfer based on the uncertainty associated with each teacher keypoint prediction. Additionally, the proposed KD leverages this uncertainty-aware alignment of keypoints to transfer the knowledge at key locations of their respective feature maps. Experiments on the widely-used LINEMOD benchmark demonstrate the effectiveness of our method, achieving superior 6DoF object pose estimation with lightweight models compared to state-of-the-art approaches. Further validation on the SPEED+ dataset for spacecraft pose estimation highlights the robustness of our approach under diverse 6DoF pose estimation scenarios.**

I. INTRODUCTION

Six Degrees of Freedom (6DoF) object pose estimation is a fundamental area in computer vision, with wide-ranging application fields such as Augmented Reality (AR), robotic manipulation, satellite docking, and space debris tracking [1]–[5]. Its core objective involves estimating both the position and orientation of an object in 3D space relative to the camera coordinate system.

The task of 6DoF pose estimation is typically addressed using a two-stage pipeline: (1) establishing correspondences between the 3D model of the object and input images [4]–[6], and (2) computing the object's pose from these correspondences using a Perspective-n-Point (PnP) algorithm [7]–[10]. While traditional methods address the first step by relying on carefully hand-designed features [11], recent successful approaches have mostly shifted towards deep learning regressors that predict key primitives such as keypoints [4], [6] or local predictions [12], [13], provided as input to the PnP solver.

Despite achieving great accuracy, 6DoF pose estimation methods are practically challenged by real-world constraints. A key limitation lies in their reliance on large deep neural architectures with tens of millions of parameters [14]–[16].

To address this, Guo *et al.* [17] introduced a Knowledge Distillation (KD) strategy for efficient 6DoF pose estimation.
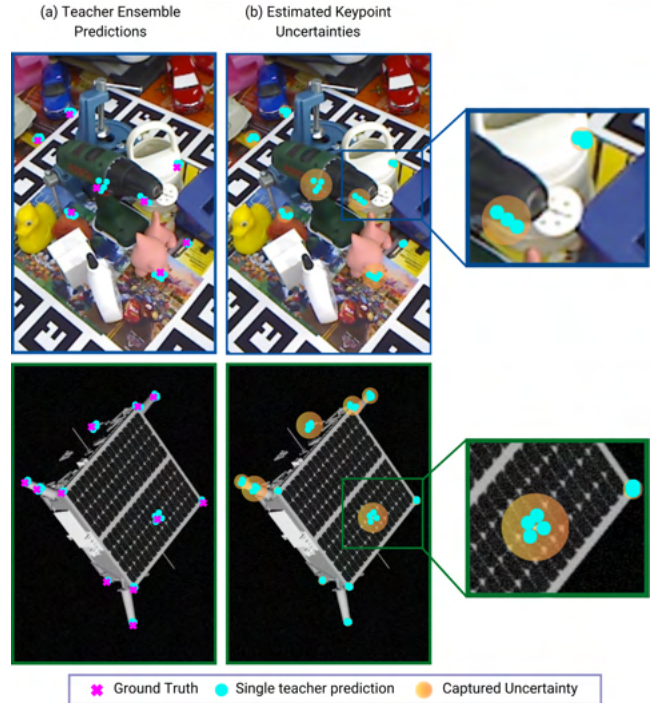
[1]SnT, University of Luxembourg, Luxembourg.
[2]Cristal Lab, ENSI, University of Manouba.
[3]Infinite Orbits, Toulouse.
Corresponding author: `nassim.aliousalah@uni.lu`

Fig. 1: **Uncertainty in Teacher Predictions**. (a) Predictions from multiple teacher models show varied keypoint locations, represented by individual teacher predictions. (b) Keypoint predictions with high uncertainty values are visualized through clustered uncertainty markers.

KD aims at transferring knowledge from a deep teacher model to a more compact student model. While such an approach has been widely explored in the field of computer vision [18]–[20], the work of [17] is the first to consider it in the specific field of 6DoF pose estimation. They observed that compact student networks often struggle to predict accurate keypoints as compared to larger teacher networks. They address this by distilling knowledge from the teacher to the student at the *prediction-level* of the keypoints through an Optimal Transport (OT) approach [21]. Additionally, the *prediction-level* KD was naively combined with an off-the-shelf *feature-level* KD method [22] in an attempt to further enhance the accuracy of the student model.

While KD stands as an appealing approach for achieving efficient 6DoF pose estimation [17], [23], we identify key limitations at two levels in existing methods: (1) *prediction-level* - most works assume that the teacher network predicts all keypoints with equal *confidence*. However, as shown in Figure 1-b, the keypoint predictions from the teacher often exhibit varying levels of *uncertainty*. Current meth-

ods [17], [23] treat all predicted keypoints equally, ignoring this uncertainty. As a result, irrelevant knowledge may be transferred from the teacher, causing the introduction of a bias in the student model towards less reliable teacher predictions. (2) *feature-level* - KD is addressed at the *feature-level* independently from the *prediction-level*, which can lead to inconsistencies when aligning the features and predictions between the teacher and student networks.

In this paper, we propose a novel KD method at both the *prediction-level* and the *feature-level* for efficient keypoint prediction in 6DoF object pose estimation, addressing the identified limitations. Our *prediction-level* KD approach, referred to as Uncertainty-Aware Knowledge Distillation (UAKD), leverages the uncertainties of keypoints predicted by the teacher model to guide the alignment between the teacher and student prediction distributions. The influence of keypoints with higher uncertainty in the teacher model is reduced in the KD process. This is achieved through an uncertainty-aware OT approach, where uncertainty values are integrated into the alignment strategy using an unbalanced Sinkhorn algorithm [24]. As shown in Figure 1-a, we rely on a teacher ensembling method [25], [26] to estimate the keypoint uncertainties, as 6DoF pose estimation techniques do not necessarily provide explicitly this information.

In addition, we extend our KD framework to operate at the *feature-level*. The proposed KD mechanism is termed Prediction-related Feature Knowledge Distillation (PFKD). In particular, the predicted keypoints are traced back to their corresponding receptive fields in the feature maps of both the teacher and student models. By applying the OT solution from the *prediction-level* to the feature maps, we establish correspondences between relevant regions in the teacher and student feature maps. This enables distillation at key locations in the feature space. By leveraging the *prediction-level* OT solution in the *feature-level* KD, we develop a comprehensive end-to-end uncertainty-aware KD approach that leverages keypoints along with their related feature representations from the teacher model. Experiments conducted on two publicly available datasets, namely LINEMOD [27] for regular object pose estimation and SPEED+ [28] for satellite pose estimation using two different backbones [12], [14], demonstrate that our method outperforms existing KD approaches for efficient 6DoF pose estimation [17], [23]. Our main contributions can be summarized as follows.

- An Uncertainty-Aware *prediction-level* KD (UAKD) for efficient keypoint prediction in 6DoF pose estimation.
- A *feature-level* KD approach (PFKD) that is consistent with the *prediction-level* KD by leveraging the *prediction-level* OT solution to establish correspondences between feature map regions and keypoint predictions in both the teacher and student models.
- Extensive experiments on two well-known datasets with two state-of-the-art pose estimation models [12], [14].

II. RELATED WORK

In this section, we start by presenting the state-of-the-art on efficient 6DoF pose estimation, then review related works on the field of KD.

Efficient 6DoF Pose Estimation. With the growing demand for real-time applications, research on pose estimation has increasingly emphasized the need for accurate yet computationally efficient solutions [17], [23], [29]–[31]. Efforts have been therefore dedicated for introducing pose estimation methods [29], [30], [32] that employ computationally friendly backbones such as SSD [33] and EfficientDet [32]. Nevertheless, these approaches predominantly prioritize the computational cost over other deployment challenges such as memory constraints. As a result, while speed is a crucial aspect, taking into account the number of parameters used in a model is also essential for reaching real-world standards in efficiency. To this end, Guan *et al*. [23] were the first to use knowledge distillation to train an efficient 6DoF pose estimation model based on the HRNetV2-W18 backbone [34]. Additionally, Guo *et al*. [17] proposed a KD approach by aligning keypoint predictions from a highly accurate teacher model to a less performing and compact one using optimal transport [21]. Despite being promising, current methods treat all keypoints equivalently while having different levels of uncertainty; therefore hindering the distillation process. Moreover, they decouple the KD at the feature and the prediction levels, leading to inconsistent knowledge transfer.

Knowledge Distillation (KD). KD has proved to be an effective way to transfer knowledge from a well-performing deeper teacher network to a more compact student one, with the aim of improving the accuracy of the student model [18], [35]–[39]. KD was first proposed by Hinton *et al*. [35] where the authors tried to leverage the last layer outputs of the teacher network. Inspired by this work, many KD variants have been derived by considering different strategies [18], [36], [39]–[42] such as the use of intermediate features from the head [18] or multiple teacher assistants [36]. Nevertheless, most of these approaches do not transfer the knowledge coherently at the *prediction-level* and the *feature-level*, which may cause inconsistencies. Only few methods based on attention [34], [40], [41] tried to address this issue on different computer vision tasks; however, disregarding the specific field of 6DoF pose estimation. Moreover, none of these approaches have considered the uncertainty aspect, which is a core component in this paper. This highlights the need to develop a feature distillation mechanism compatible with an uncertainty-aware *prediction-level* KD.

III. PROBLEM STATEMENT

Estimating the 6DoF pose of a given object from a sample $\mathbf{I} \in \mathcal{I}$ belonging to the space of RGB images $\mathcal{I}$ involves finding a function $\mathbf{\Phi}$ such that,

$$
\begin{aligned}
\mathbf{\Phi} \; : \; \mathcal{I} \; &\longrightarrow \; \mathcal{SO}(3) \times \mathbb{R}^3 \;, \\
\mathbf{I} \; &\longmapsto \; \mathbf{\Phi}(\mathbf{I}) = (\mathbf{R}, \mathbf{t}) \;,
\end{aligned}
\tag{1}
$$

where $(\mathbf{R}, \mathbf{t})$ represents the object pose formed by a rotation matrix $\mathbf{R} \in \mathcal{SO}(3)$ and a translation vector $\mathbf{t} \in \mathbb{R}^3$ w.r.t. the camera coordinate system. $\mathcal{SO}(3)$ denotes the special orthogonal group in dimension 3.

In a two-stage neural 6DoF pose estimation method [4]–[6], the first stage relies on a neural network $\mathbf{\Phi}_k$ that detects from an input $\mathbf{I}$, a total of N 2D keypoints $\mathbf{K} = (\mathbf{k}_i)_{i \in [[1,N]]}$, where $\mathbf{k}_i \in \mathbb{R}^2$ and $\mathbf{K} \in \mathbb{R}^{N \times 2}$. Using the predicted 2D keypoints $\mathbf{K}$ and known keypoints $\mathbf{K}_{3D} \in \mathbb{R}^{N \times 3}$ from a 3D object model, the pose parameters can be estimated using a PnP solver [7], [9], i.e., $(\mathbf{R}, \mathbf{t}) = \mathbf{PnP}(\mathbf{\Phi}_k(\mathbf{I}), \mathbf{K}_{3D})$. The performance of two-stage 6DoF pose estimation methods heavily depends on the accuracy of the predicted keypoints $\mathbf{K}$. Smaller keypoint prediction models tend to have a higher predicted pose error compared to larger ones [17]. Knowledge Distillation (KD) addresses this challenge by considering a trained well-performing large teacher model $\mathbf{\Phi}_k^T$, parametrized by θ_T, and transferring its knowledge when training a smaller student model $\mathbf{\Phi}_k^S$, parametrized by θ_S, for the same task, i.e., keypoint prediction. Here, it is important to note that $|\theta_S| < |\theta_T|$, where $|.|$ refers to the cardinality. The student model is then tasked with two learning objectives at the same time, namely, the minimization of the error between the predictions and the ground-truth, while matching its feature activations and/or predictions with those of the teacher. The parameters of the student are optimized as follows,

$$\arg\min_{\theta_S} \gamma_{\text{kpt}} \mathcal{L}_{\text{kpt}}(\mathbf{K}^S, \mathbf{K}_{gt}) + \gamma_{\text{distill}} \mathcal{L}_{\text{distill}}(\mathbf{\Phi}_k^S, \mathbf{\Phi}_k^T) , \quad (2)$$

where $\mathcal{L}_{\text{kpt}}(\mathbf{K}^S, \mathbf{K}_{gt})$ measures the error between the predicted student keypoints $\mathbf{K}^S$ and the corresponding ground-truth keypoints $\mathbf{K}_{gt}$, $\mathcal{L}_{\text{distill}}(\mathbf{\Phi}_k^S, \mathbf{\Phi}_k^T)$ quantifies the discrepancy between the feature activations and/or the predicted keypoints of the student $\mathbf{\Phi}_k^S$ and the teacher $\mathbf{\Phi}_k^T$. The parameters γ_{kpt} and γ_{distill} balance the two terms in (2).

Our goal is to define the knowledge transfer process for keypoint prediction in 6DoF pose estimation $\mathcal{L}_{\text{distill}}$ as an end-to-end distillation process that operates at two levels. Hence, we formulate it as follows,

$$\mathcal{L}_{\text{distill}}(\mathbf{\Phi}_k^S, \mathbf{\Phi}_k^T) = \gamma_p \underbrace{\mathcal{L}_{\text{pred}}(\mathbf{K}^S, \mathbf{K}^T)}_{\text{Prediction-level}} + \gamma_f \underbrace{\mathcal{L}_{\text{feat}}(\mathbf{R}^S, \mathbf{R}^T)}_{\text{Feature-level}} . (3)$$

At the *prediction-level*, we propose to align the student predicted keypoints, $\mathbf{K}^S$, with those of the teacher, $\mathbf{K}^T$, while accounting for potential uncertainties in the teacher predictions. At the *feature-level*, our goal is to ensure the alignment between student and teacher feature activations $\mathbf{F}^S$ and $\mathbf{F}^T$, at key locations $\mathbf{R}^S = f(\mathbf{F}^S, \mathbf{K}^S)$ and $\mathbf{R}^T = f(\mathbf{F}^T, \mathbf{K}^T)$, respectively. These key regions are retrieved through a function $f(.,.)$ driven by keypoint predictions for ensuring a consistent end-to-end process between the *feature-level* and *prediction-level* alignments.

IV. PROPOSED APPROACH

To address the problem outlined in Section III, we propose a Knowledge Distillation (KD) framework for efficient keypoint prediction for 6DoF pose estimation, integrating two complementary KD methods, namely, UAKD which operates at the *prediction-level* and PFKD which acts at the *feature-level*. In particular, UAKD leverages the prediction

uncertainties of the teacher model for performing an Optimal Transport (OT) [21] alignment between the teacher and the student prediction distributions. On the other hand, PFKD relies on the correspondences established between teacher and student keypoints at the *prediction-level* to guide the *feature-level* distillation process, focusing on activated feature regions. Figure 2-(A) illustrates the two-level overall proposed KD framework.

A. Uncertainty-Aware Prediction-Level KD

Let $\mathbf{K}^T = (\hat{\mathbf{k}}_i^T)_{i \in [[1,N]]}$ and $\mathbf{K}^S = (\hat{\mathbf{k}}_i^S)_{i \in [[1,M]]}$ be the teacher and the student predicted keypoints, respectively. To transfer the knowledge from teacher to student keypoints with different cardinalities ($N \neq M$), the keypoint alignment process can be seen as an alignment of two distributions, and therefore formulated as an unbalanced OT problem [17], [43], where each keypoint is weighted. Formally, this is equivalent to finding the optimal transportation plan $\boldsymbol{\pi} \in \mathbb{R}^{N \times M}$ that minimizes the overall alignment defined as,

$$\mathcal{L}_{pred} = \min_{\boldsymbol{\pi}} \sum_{i=1}^{M} \sum_{j=1}^{N} \pi_{ij} \|\hat{\mathbf{k}}_i^S - \hat{\mathbf{k}}_j^T\|_2 ,$$

$$\text{s.t.} \quad \forall i, \sum_{j=1}^{N} \pi_{ij} = \alpha_i^S, \quad \forall j, \sum_{i=1}^{M} \pi_{ij} = \alpha_j^T ,$$

(4)

where $\boldsymbol{\alpha}^T = (\alpha_1^T, \alpha_2^T, \cdots, \alpha_N^T)$ and $\boldsymbol{\alpha}^S = (\alpha_1^S, \alpha_2^S, \cdots, \alpha_M^S)$ denote the weights assigned to the teacher and student keypoint predictions, respectively. The solution to this OT problem can be efficiently found using an unbalanced Sinkhorn algorithm [24], [44].

While Guo *et al.* adopt such an OT-based KD method [17], they only consider the existence probabilities of predicted keypoints that are provided by some keypoint regressors [12]. In this work, we posit that incorporating the uncertainties of the teacher keypoint predictions within the KD process can improve the performance of the student model. To that end, we define the following *confidence* weights for the teacher and the student models,

$$\boldsymbol{\alpha}^{T,c} = \mathbf{1}_N - \mathbf{u}, \quad \boldsymbol{\alpha}^{S,c} = \frac{1}{M} \cdot \mathbf{1}_M , \quad (5)$$

where $\mathbf{1}_N$ and $\mathbf{1}_M$ denote N and M-dimensional all-ones vectors, respectively, and $\mathbf{u} = (u_1, u_2, \cdots, u_N) \in [0,1]^N$ is a vector containing the uncertainty scores for each predicted keypoint $\hat{\mathbf{k}}_i^T$. For models that provide keypoint existence probabilities [12], these confidence weights can be integrated as follows,

$$\boldsymbol{\alpha}^T = \lambda \times \boldsymbol{\alpha}^{T,c} + (1 - \lambda) \times \boldsymbol{\alpha}^{T,e} , \quad (6)$$

where λ denotes a modulating factor and $\boldsymbol{\alpha}^{T,e}$ are the existence probabilities of the teacher's keypoints. We chose summation over multiplication due to superior empirical performances. Figure 2-(B) depicts the proposed UAKD.

Teacher Keypoint Uncertainty Estimation. Here, we describe the estimation of the teacher keypoint uncertainties $\mathbf{u}$ used to define the confidence weights $\boldsymbol{\alpha}^{T,c}$ in (5). We focus on estimating the teacher *epistemic* uncertainty only, which typically arises from the model's parameters themselves [45], preventing the distillation from being influenced by the noise inherent in the data, i.e., *aleatoric*
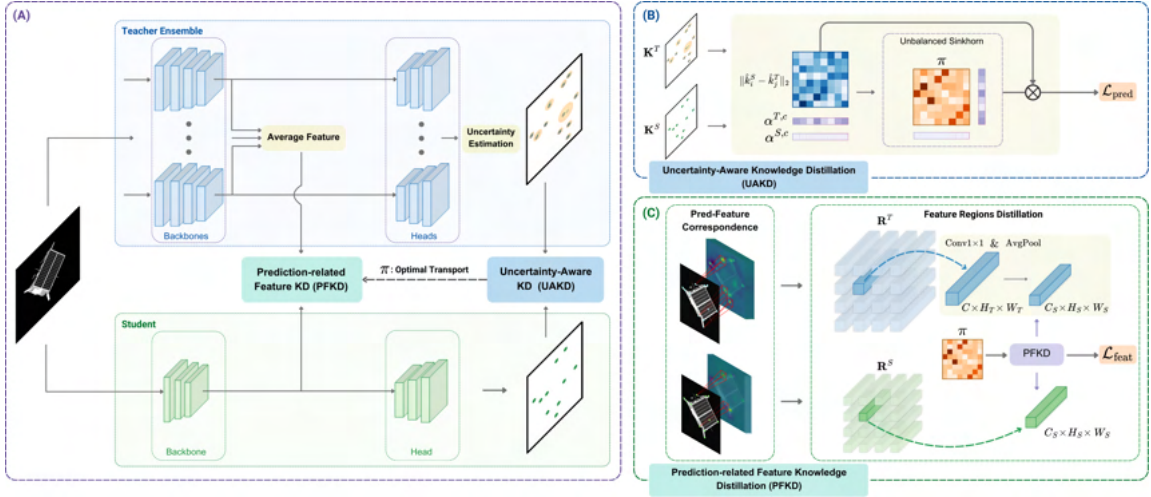
Fig. 2: **Overview of the Proposed Knowledge Distillation (KD) Framework** (Best viewed in color). **(A) General overview**: Keypoints predicted by the ensemble of teachers are used to estimate uncertainties and are subsequently processed by UAKD, while averaged feature maps are directed to PFKD, guided by the OT plan π. **(B) UAKD Module**: Keypoints, along with corresponding confidence scores $\alpha^{T,c}$ and $\alpha^{S,c}$, are aligned using an unbalanced Sinkhorn algorithm [43], where $\otimes$ represent the tensor product. **(C) PFKD Module**: Predicted keypoints are mapped back to their respective regions in the feature maps and are aligned consistently according to the OT plan π.

uncertainty. For this end, we employ the deep ensembling method [25] which approximates the posterior distribution of the teacher model's weights $p(\theta_T|\mathcal{I})$ in a similar way to Monte Carlo sampling [46]. We represent the ensemble of trained teacher weights as $\Theta = \{\theta_{T_j}\}_{j=1}^{E}$. These teacher models are similarly trained but with different weight initializations. At inference, for each input $\mathbf{I} \in \mathcal{I}$, the trained ensemble produces E sets of 2D keypoint predictions, $\mathbf{K}_\Theta^T = \{(\hat{\mathbf{k}}_i^{T_j})_{i\in[[1,N]]}\}_{j\in[[1,E]]}$ where $\hat{\mathbf{k}}_i^{T_j} \in \mathbb{R}^2$ represents the i-th keypoint predicted by the model with the weights θ_{T_j}. We then aggregate the predictions of the teachers for each keypoint i by estimating its per-coordinate mean $(\mu_{x,\hat{k}_i}, \mu_{y,\hat{k}_i})$ and variance $(\sigma_{x,\hat{k}_i}^2, \sigma_{y,\hat{k}_i}^2)$. The collection of mean keypoints are then used as representative of the teacher ensemble predictions $\mathbf{K}^T$. For simplicity, we consider the x- and y- axes independent, Therefore, the total variance of keypoint i is given by $\sigma_{\hat{k}_i}^2 = \sigma_{x,\hat{k}_i}^2 + \sigma_{y,\hat{k}_i}^2$. We then form $\boldsymbol{\sigma}^T = (\sigma_{\hat{k}_i}^2)_{i\in[[1,N]]}$ as the per-keypoint variances of the teacher ensemble. These variance values are further mapped to uncertainty scores in $[0,1]^N$ using a $\mathtt{tanh}$ function to obtain $\mathbf{u} = \mathtt{tanh}(\boldsymbol{\sigma}^T)$. Our experiments demonstrated that an ensemble of just 4 to 6 models is sufficient to closely approximate the true variance, *i.e.*, the epistemic uncertainty, of the teacher model.

Note that in the above, we assume that all teacher models predict the same number of keypoints. However, this may not always hold. In such cases, a majority voting is used to identify the keypoints for which uncertainties will be estimated. The remaining are assigned an uncertainty of 1.

B. Prediction-related Feature-per-Keypoint KD

We now describe the proposed *Prediction-related Feature-per-Keypoint* Knowledge Distillation (PFKD). PFKD leverages the predicted keypoints and the transportation plan

obtained at the *prediction-level* KD (Section IV-A) to distill the knowledge at key locations of the feature space.

Let $\mathbf{F}^T \in \mathbb{R}^{C\times H\times W}$ denote the feature maps extracted from the teacher backbone network, where C, H, and W are the number of channels, height, and width of the feature maps, respectively. We start by retrieving the keypoint coordinates predicted by the teacher model, represented as $\mathbf{K}^T = (\hat{\mathbf{k}}_i^T)_{i\in[[1,N]]}$. A mapping is then established between each keypoint and the feature maps, enabling the identification of keypoint locations within the feature space. Indeed, the predicted keypoints are obtained through a series of convolutions applied on the feature maps that conserve spatial awareness, making it possible to map each keypoint to its corresponding feature location. Thus, we associate each predicted keypoint $\hat{\mathbf{k}}_i^T$ defined by its 2D position $\hat{\mathbf{k}}_i^T = (\hat{x}_i^T, \hat{y}_i^T)$ with a feature region $\mathbf{R}_i^T \in \mathbb{R}^{C\times H_T\times W_T}$ centered at c_i and with spatial dimensions H_T and W_T that are defined as described below,

$$\begin{cases} c_i = (\lfloor \delta \times \hat{x}_i^T \rceil, \lfloor \delta \times \hat{y}_i^T \rceil), \\ H_T = W_T = 1 + \sum_{i=1}^{L}\left((\kappa_i - 1)\prod_{j=1}^{i-1} s_j\right), \end{cases} \tag{7}$$

where L denotes the number of convolutional layers in the keypoint prediction head, δ is the scaling factor relating the feature map size to the output size, $\lfloor \cdot \rceil$ the rounding operation, and κ_i and s_i are the kernel size and stride of the i-th convolutional layer. We then obtain a set $\mathbf{R}^T = \{\mathbf{R}_1^T, \mathbf{R}_2^T, \ldots, \mathbf{R}_N^T\}$ that represents the feature regions that contributed to the prediction of each keypoint. Applying a similar procedure with the student model, we obtain $\mathbf{R}^S = \{\mathbf{R}_1^S, \mathbf{R}_2^S, \ldots, \mathbf{R}_M^S\}$, where $\mathbf{R}_i^S \in \mathbb{R}^{C_S\times H_S\times W_S}$. If necessary, we apply a 1×1 convolution on the feature regions extracted from the teacher model to adjust the channel dimensions and an average pooling

operation to modify the spatial dimensions. This ensures that the elements of $\mathbf{R}^T$ are transformed to $\mathbf{R}_i^T \in \mathbb{R}^{C_S, H_S, W_S}$.

Given $\mathbf{R}^T$ and $\mathbf{R}^S$, the proposed approach is turned into aligning the sets of teacher and student per-keypoint feature regions. To match each region from $\mathbf{R}^T$ to its corresponding one in $\mathbf{R}^S$, we propose to use the optimal transport plan found in (4). Thus, the proposed PFKD loss can be formulated as follows,

$$
\mathcal{L}_{\text{feat}}(\mathbf{R}^T, \mathbf{R}^S, \boldsymbol{\pi}) = \frac{1}{N \cdot M} \sum_{i=1}^{N} \sum_{j=1}^{M} \Bigg[
$$
$$
\pi_{i,j} \sum_{d_1=1}^{C_S} \sum_{d_2=1}^{W_S} \sum_{d_3=1}^{H_S} \frac{\left(\mathbf{R}_i^T[d_1, d_2, d_3] - \mathbf{R}_j^S[d_1, d_2, d_3]\right)^2}{C_S \cdot H_S \cdot W_S} \Bigg]. \tag{8}
$$

It is important to mention that by re-using the transportation plan of the uncertainty-aware *prediction-level* KD, the proposed PFKD also incorporates the uncertainty information of the teacher. Note that the PFKD process explained above considers a single teacher model. However, our approach uses an ensemble of teachers as mentioned in Section IV-A. Thus, we aggregate the different feature regions extracted from different teacher models by averaging them. As a result, the set of the teacher feature regions becomes $\mathbf{R}^T = \{\frac{1}{E} \sum_{j=1}^{E} \mathbf{R}_i^{T_j}\}_{i=1}^{N}$. Figure 2-(C) depicts the proposed PFKD.

Finally, the uncertainty-aware *prediction-level* and *feature-level* KD are together in the overall training. This is achieved by plugging their respective losses defined in (8) and (4) into (3), where γ_p, $\gamma_f \in \mathbb{R}$ are the two modulating factors for each loss. We set $\gamma_p = 5$ and $\gamma_f = 0.1$ in our experiments, which leads to the best results.

V. EXPERIMENTS

In this section, we first describe our experimental setup. Then, we demonstrate the effectiveness of the proposed KD method on a well-known benchmark for generic object pose estimation, namely, the LINEMOD dataset [27]. Furthermore, we consider the specific scenario of spacecraft pose estimation given the relevance of efficient models in the context of space applications. In particular, we validate our approach on the spacecraft pose estimation dataset SPEED+ [28]. Additional experiments are finally conducted to validate the proposed approach.

A. Experimental Setup

Models. We evaluate our approach using two distinct two-stage pose estimation networks, namely, WDRNet [12] and SPNv2 [14]. WDRNet uses a DarkNet-53 [47] backbone coupled with a feature pyramid network [48] to predict 2D keypoint locations across multiple scales. It then employs a sampling strategy that enables feature vectors at each level to contribute probabilistically to the predictions. This results in estimating 2D keypoint clusters for each 3D bounding box corner, thus making it a local prediction-based keypoint detector. On the other hand, SPNv2 is a multi-task CNN that uses a shared multi-scale feature encoder built on an EfficientDet [32] backbone and BiFPN layers. It consists of

multiple prediction heads for various tasks. In our work, we focus on the 2D keypoint prediction head that is trained to predict 2D heatmaps associated with each 2D keypoint on the object of interest. In our experiments, for the teacher models, we use WDRnet with Darknet-53 and the $\phi = 6$ version of SPNv2 based on the EfficientDet-D6 [32] backbone. For the student networks, we use WDRnet with Darknet-Tiny [49] and a lighter Darknet-Tiny-H version with half of the channels of DarkNet-tiny in each layer. Moreover, for the lighter SPNv2 student we train a $\phi = 0$ version which is based on the EfficientDet-D0 [32] backbone. A more detailed comparison of these architectures is presented in Table I.

TABLE I: Comparison of different architectures in terms of number of parameters (in millions), floating point operations per second (in billions), and input resolution.

Model	Backbone	#Param [M]	#FLOPs [B]	Input Res.
	Darknet-53	52.1	36.51	
WDRNet	Darknet-Tiny	8.5	18.84	640×480
	Darknet-Tiny-H	2.3	4.75	
SPNv2	EfficientDet-D6	57.8	288.27	768×512
	EfficientDet-D0	3.8	12.1	

Datasets. LINEMOD [27] is a general object pose estimation benchmark. It contains 16,000 images belonging to 13 different object categories. SPEED+ [28] is a cross-domain dataset for Spacecraft pose estimation. It is formed by 60,000 synthetic images for training and two additional unlabeled subsets containing 6,740 and 2,791 images from two different domains, `lightbox` and `sunlamp`, respectively. The `lightbox` consists of the Tango spacecraft [50] mockup model illuminated to simulate diffuse light in the Earth's orbit, whereas the `sunlamp` images of the same model are illuminated with a different lamp setup to simulate direct sunlight. In our experiments, we follow the same data augmentation setup used in [14].

Metrics. We report our results on LINEMOD using the Average Distance of Model Points (ADD) metric [27], which is defined as the average distance between the transformed 3D model and the predicted pose. More specifically, we use the ADD-0.1d variant, which represents the percentage of images with an average distance lower than 10% of the object diameter. For symmetric objects, the average closest point distance (ADD-S) metric [51] is used. In the experiments on SPEED+, we do not have the 3D model of the object; thus, we use the rotation, translation, and pose errors, denoted as E_R, E_T, and E_{pose}, respectively, as in [14].

Baselines. We evaluate our proposed approaches against several baselines, namely, (1) a fully trained student model without KD, referred to as **Student**, (2) the state-of-the-art KD approach called **ADLP** from [17], which, to the best of our knowledge, is the only KD method specifically designed for 6DoF pose estimation, combined with the feature-based KD approach referred to as **FKD** [22] which has been shown to be orthogonal with ADLP. Our proposed Uncertainty-Aware KD, Prediction-related Feature KD, and their combination are labeled as **UAKD**, **PFKD**, and **UAKD+PFKD**,

TABLE II: Performance on the LINEMOD dataset in terms of ADD-0.1d using DarkNet-Tiny-H and DarkNet-Tiny based students; Objects with * are symmetric and the ADD-S metric is used instead.

Class	Teacher Model	DarkNet-Tiny-H					DarkNet-Tiny				
		Baselines		Ours			Baselines		Ours		
		Student	ADLP [17] + FKD [22]	UAKD	PFKD	UAKD + PFKD	Student	ADLP [17] + FKD [22]	UAKD	PFKD	UAKD + PFKD
Ape	83.0	65.4	69.9	75.6	77.2	**79.1**	73.4	76.2	79.6	**80.3**	80.0
Benchvise	96.5	92.0	93.7	**94.5**	94.1	**94.5**	95.2	96.7	**97.2**	95.3	96.5
Camera	93.8	78.4	84.5	88.0	**89.2**	88.0	91.2	92.0	**93.5**	92.4	93.4
Can	96.7	82.2	83.9	**90.7**	88.9	89.8	92.4	94.0	**97.1**	95.4	95.4
Cat	93.9	81.5	81.6	87.6	85.1	**88.7**	87.2	88.6	91.6	**93.7**	92.8
Driller	95.5	85.5	90.3	93.3	**95.5**	93.4	92.2	94.8	96.9	94.0	**97.0**
Duck	79.0	64.3	68.9	**72.7**	68.7	**72.7**	70.9	74.7	**77.9**	**77.9**	**77.9**
Eggbox*	99.2	95.8	96.4	97.7	**98.3**	97.9	99.3	99.3	**99.5**	98.8	98.9
Glue*	97.9	90.7	93.2	94.8	95.4	**96.4**	97.2	97.7	98.0	98.0	**98.3**
Holepuncher	87.7	73.2	76.3	80.1	**81.4**	**81.4**	78.0	82.2	85.5	**87.1**	85.5
Iron	95.5	86.3	90.5	89.6	**93.2**	92.4	92.1	93.2	94.2	94.6	**95.3**
Lamp	98.1	93.6	94.6	96.4	**96.5**	96.0	96.6	96.8	97.8	97.5	**99.1**
Phone	91.3	76.0	79.2	85.0	83.3	**86.9**	87.5	89.6	89.8	**92.9**	89.6
AVG.	92.9	81.9	84.8 (↑ 2.9)	**88.1** (↑ 6.2)	**88.2** (↑ 6.3)	**89.0** (↑ 7.1)	88.7	90.4 (↑ 1.7)	**92.2** (↑ 3.5)	**92.0** (↑ 3.3)	**92.3** (↑ 3.6)

respectively. For UAKD and UAKD+PFKD, we report only the best average results with an ensemble based on 4 and 6 teacher models.

B. 6DoF Object Pose Estimation with LINEMOD

Results using WDRnet. We report in Table II the results obtained for WDRnet using the DarkNet-Tiny-H and DarkNet-Tiny backbones. Our approach (UAKD + PFKD) achieves the best performance on LINEMOD for both backbones with an ADD-0.1d of 89.0 and 92.3, respectively. Remarkably, even by integrating PFKD or UAKD solely, our method outperforms the state-of-the-art in terms of ADD-0.1d on LINEMOD in both cases. In general, our method reduces the performance gap between the DarkNet53-based teacher model and the student models, namely, DarkNet-Tiny-H (which has 95.5% fewer parameters) and DarkNet-Tiny (which has 83.7% fewer parameters) by 7.1 and 3.6, respectively. It can be noted that DarkNet-Tiny even surpasses the teacher model for some classes such as *Bvise*, *Can*, *Driller*, *Eggbox* and *Phone*.

Results using SPNv2. Table III reports the results obtained when SPNv2 [14] on LINEMOD. As for WDRnet, both UAKD and PFKD, even when integrated alone, outperforms the state-of-the-art. Specifically, it improves the results of the student model by 1.4 and 1.2 points on average LINEMOD, respectively. Moreover, by combining UAKD and PFKD, our approach (UAKD+PFKD) achieves an improvement of 1.8, surpassing the fully-trained teacher model on the *Can* and *Iron* classes while reducing the FLOPs by 96.6%.

C. Spacecraft Pose Estimation with SPEED+

For the SPEED+ [28] dataset, we conduct experiments using SPNv2 only, given the low performance of WDRnet on this specific dataset. Each trained model is evaluated across the three different domains provided by SPEED+. Table IV provides the results obtained on SPEED+ in terms of translation, rotation and pose errors for the three different domains, namely, synthetic, lightbox, and sunlamp. Both

TABLE III: Performance on the LINEMOD dataset in terms of ADD-0.1d and ADD-S using an SPNv2 $\phi = 0$ student.

Class	Teacher Model	SPNv2 $\phi = 0$				
		Baselines		Ours		
		Student	ADLP [17] + FKD [22]	UAKD	PFKD	UAKD + PFKD
Ape	88.9	67.6	70.8	**70.8**	68.8	70.6
Benchvise	98.5	91.9	92.3	**93.5**	92.9	92.8
Camera	96.9	90.3	92.4	92.5	92.8	**93.0**
Can	89.7	88.6	88.9	89.2	88.2	**89.8**
Cat	93.7	90.3	90.9	89.6	**91.0**	**91.0**
Driller	97.1	82.2	82.9	84.7	83.2	**84.9**
Duck	77.7	68.9	69.7	69.0	69.1	**70.0**
Eggbox*	95.9	93.5	93.6	93.8	**94.3**	94.2
Glue*	96.4	92.4	92.9	92.9	**93.0**	**93.0**
Holepuncher	87.4	75.2	78.4	79.3	79.8	**80.0**
Iron	86.4	87.0	88.0	88.3	88.2	**88.7**
Lamp	95.7	92.4	92.7	92.8	**93.2**	92.9
Phone	91.8	82.2	83.2	83.8	83.6	**84.0**
AVG.	92.8	84.8	85.9 (↑ 1.1)	**86.2** (↑ 1.4)	**86.0** (↑ 1.2)	**86.6** (↑ 1.8)

the proposed UAKD and PFKD outperform the state-of-the-art, i.e., ADLP+FKD. Specifically, Our UAKD method achieves a pose error of 0.024 in the synthetic domain and 0.288 in the lightbox domain, competing with the fully trained SPNv2 model ($\phi = 6$) on the synthetic domain and surpassing the student model by 0.080 in the lightbox domain. In the sunlamp domain, UAKD reduces the pose error by 0.039, corresponding to a decrease of 2.327° in terms of rotation error compared to the student and 1.438° compared to ADLP+FKD. Additionally, our PFKD enhances the performance, slightly improving the pose error by 0.001, 0.079, and 0.024, in the synthetic, lightbox, and sunlamp domains, respectively as compared to ADLP+FKD. With our UAKD+PFKD variant, the performance of the student model aligns with the teacher in the synthetic domain, offering a 0.003° advantage in terms of rotation error, while significantly improving the performance in the lightbox and sunlamp domains by reducing the pose error by 0.120 and 0.041, respectively, corresponding to a reduction of 5.385° and 2.340° in terms of rotation error.

TABLE IV: Performance on the SPEED+ dataset in terms of the rotation error E_R, translation error E_T, and pose error E_{pose} using an SPNv2 $\phi = 0$ student under the `synthetic`, `lightbox` and `sunlamp` domains; E^*_{pose} refer to the HIL pose error [14].

Model			synthetic			lightbox			sunlamp		
			E_T [m]	E_R [°]	E_{pose} [-]	E_T [m]	E_R [°]	E^*_{pose} [-]	E_T [m]	E_R [°]	E^*_{pose} [-]
		Teacher Model	0.042	1.010	0.024	0.272	8.968	0.202	0.259	12.580	0.263
SPNv2 $\phi = 0$	Base.	Student	0.050	1.441	0.033	0.447	16.804	0.368	0.372	19.366	0.401
		ADLP [17]+ FKD [22]	0.045	1.157	0.027	0.482	14.596	0.336	0.387	18.477	0.388
	Ours	UAKD	**0.041**	1.018	0.024	0.346	13.195	0.288	0.383	17.039	0.362
		PFKD	0.048	1.197	0.028	0.288	11.973	0.257	0.388	17.093	0.364
		UAKD + PFKD	0.042	**1.007**	**0.024**	**0.288**	**11.419**	**0.248**	0.373	**17.026**	**0.360**

D. Additional Analysis

Effect of the Ensemble. We examine the impact of using the ensemble strategy in our UAKD, as illustrated in Figure 3. We compare the results obtained for a student backbone based on an ensemble of teachers without considering uncertainty ($\lambda = 0$) to the same backbone without relying on an ensemble. We find that the ensemble contributes with an improvement of 0.2 on DarkNet-Tiny-H and 0.5 on DarkNet-Tiny. This indicates that the increase in performance obtained by our approach returns mostly to the uncertainty-aware knowledge distillation rather than the ensembing approach. The slight improvement coming from the ensemble is attributed to its tendency to produce more accurate predictions [52].

Choice of λ. As presented in (6), we introduce a weight factor λ to balance between the WDRNet's existence score per keypoint and the uncertainty value generated by our ensemble. Figure 3 illustrates the ADD-0.1d values across various LINEMOD training sessions for different λ values. A 50%-50% combination ($\lambda = 0.5$) of uncertainty and existence scores yields the best results on LINEMOD.
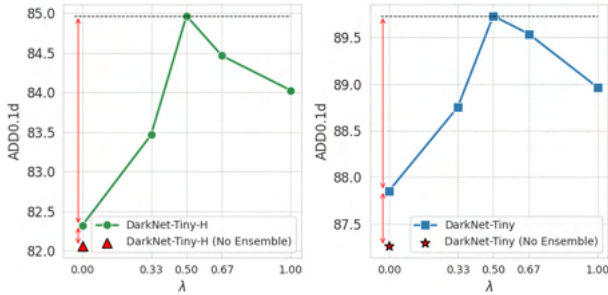

Fig. 3: Impact of λ on the ADD-0.1d Metric using WDRnet on LINEMOD.

VI. Conclusion

In this paper, we present an uncertainty-aware end-to-end knowledge distillation (KD) framework for efficient two-stage 6DoF pose estimation. We argue that the varying uncertainty in teacher predictions is a valuable asset for KD and propose an uncertainty-driven selective weighted distillation method, UAKD, to optimize the alignment between teacher and student keypoints. Leveraging a deep ensemble strategy, we quantify uncertainties for *prediction-level* distillation and extend this to *feature-level* distillation (PFKD) by tracing keypoint predictions to associated feature map regions, enabling a unified uncertainty-aware distillation. Experimental results demonstrate the effectiveness of the proposed method in terms of both accuracy and compactness, highlighting its potential for deployment in pose estimation applications. Future directions include integrating an uncertainty-guided PnP approach [4] and exploring other compression techniques to further improve efficiency and accuracy.

References

[1] J. Liu, W. Sun, C. Liu, X. Zhang, and Q. Fu, "Robotic continuous grasping system by shape transformer-guided multiobject category-level 6-d pose estimation," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 11, pp. 11 171–11 181, 2023.

[2] J. Lin, L. Liu, D. Lu, and K. Jia, "Sam-6d: Segment anything model meets zero-shot 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 27 906–27 916.

[3] V. N. Nguyen, T. Groueix, G. Ponimatkin, Y. Hu, R. Marlet, M. Salzmann, and V. Lepetit, "Nope: Novel object pose estimation from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 17 923–17 932.

[4] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixelwise voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[5] Y. Di, F. Manhardt, G. Wang, X. Ji, N. Navab, and F. Tombari, "So-pose: Exploiting self-occlusion for direct 6d pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 12 396–12 405.

[6] B. Chen, J. Cao, A. Parra, and T.-J. Chin, "Satellite pose estimation with deep landmark regression and nonlinear pose refinement," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[7] H. Chen, P. Wang, F. Wang, W. Tian, L. Xiong, and H. Li, "Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2781–2790.

[8] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o(n) solution to the pnp problem," p. 155–166, 2009. [Online]. Available: https://infoscience.epfl.ch/handle/20.500.14299/60530

[9] S. Li, C. Xu, and M. Xie, "A robust o (n) solution to the perspective-n-point problem," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 7, pp. 1444–1450, 2012.

[10] G. Terzakis and M. Lourakis, "A consistently fast and globally optimal solution to the perspective-n-point problem," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 478–494.

[11] J. Vidal, C.-Y. Lin, and R. Martí, "6d pose estimation using an improved method based on point pair features," in *2018 4th international conference on control, automation and robotics (iccar)*. IEEE, 2018, pp. 405–409.

[12] Y. Hu, S. Speierer, W. Jakob, P. Fua, and M. Salzmann, "Wide-depth-range 6d object pose estimation in space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 870–15 879.

[13] Y. Hu, J. Hugonot, P. Fua, and M. Salzmann, "Segmentation-driven 6d object pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3385–3394.

[14] T. H. Park and S. D'Amico, "Robust multi-task learning and online refinement for spacecraft pose estimation across domain gap," *Advances in Space Research*, 2023.

[15] K. Kajak, C. Maddock, H. Frei, and K. Schwenk, "Segmentation-driven spacecraft pose estimation for vision-based relative navigation in space," in *Proceedings of the International Conference on Computer Vision Systems (ICVS)*, October 2021.

[16] A. Legrand, R. Detry, and C. De Vleeschouwer, "End-to-end neural estimation of spacecraft pose with intermediate detection of keypoints," in *Computer Vision – ECCV 2022 Workshops*, L. Karlinsky, T. Michaeli, and K. Nishino, Eds. Cham: Springer Nature Switzerland, 2023, pp. 154–169.

[17] S. Guo, Y. Hu, J. M. Alvarez, and M. Salzmann, "Knowledge distillation for 6d pose estimation by aligning distributions of local predictions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 18 633–18 642.

[18] J. Wang, Y. Chen, Z. Zheng, X. Li, M.-M. Cheng, and Q. Hou, "Crosskd: Cross-head knowledge distillation for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 16 520–16 530.

[19] H. Zhang, L. Liu, Y. Huang, Z. Yang, X. Lei, and B. Wen, "Cakdp: Category-aware knowledge distillation and pruning framework for lightweight 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 15 331–15 341.

[20] Q. Huangpu and H. Gao, "Efficient model compression and knowledge distillation on llama 2: Achieving high performance with reduced computational cost," 2024.

[21] C. Villani *et al.*, *Optimal transport: old and new*. Springer, 2009, vol. 338.

[22] L. Zhang and K. Ma, "Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=uKhGRvM8QNH

[23] Q. Guan, Z. Sheng, and S. Xue, "Hrpose: Real-time high-resolution 6d pose estimation network using knowledge distillation," *arXiv preprint arXiv:2204.09429*, 2022.

[24] T. Séjourné, J. Feydy, F.-X. Vialard, A. Trouvé, and G. Peyré, "Sinkhorn divergences for unbalanced optimal transport," *arXiv preprint arXiv:1910.12958*, 2019.

[25] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[26] K. Wursthorn, M. Hillemann, and M. Ulrich, "Uncertainty quantification with deep ensembles for 6d object pose estimation," *arXiv preprint arXiv:2403.07741*, 2024.

[27] 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 *Computer Vision – ACCV 2012*, K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 548–562.

[28] T. H. Park, M. Märtens, G. Lecuyer, D. Izzo, and S. D'Amico, "Speed+: Next-generation dataset for spacecraft pose estimation across domain gap," in *2022 IEEE Aerospace Conference (AERO)*. IEEE, 2022, pp. 1–15.

[29] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[30] 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 (CVPR)*, June 2018.

[31] Y. Bukschat and M. Vetter, "Efficientpose: An efficient, accurate and scalable end-to-end 6d multi object pose estimation approach," *arXiv preprint arXiv:2011.04307*, 2020.

[32] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[33] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision– ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[34] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[35] G. Hinton, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[36] W. Son, J. Na, J. Choi, and W. Hwang, "Densely guided knowledge distillation using multiple teacher assistants," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 9395–9404.

[37] Y. Jin, J. Wang, and D. Lin, "Multi-level logit distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 24 276–24 285.

[38] L. Wang, Y. Liu, P. Du, Z. Ding, Y. Liao, Q. Qi, B. Chen, and S. Liu, "Object-aware distillation pyramid for open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 11 186–11 196.

[39] J. Zhang, Y. Gao, M. Zhou, R. Liu, X. Cheng, S. Nikolić, and S. Chen, "Svd-kd: Svd-based hidden layer feature extraction for knowledge distillation," 01 2024.

[40] M. Ji, B. Heo, and S. Park, "Show, attend and distill: Knowledge distillation via attention-based feature matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 7945–7952.

[41] P. Passban, Y. Wu, M. Rezagholizadeh, and Q. Liu, "Alp-kd: Attention-based layer projection for knowledge distillation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, pp. 13 657–13 665, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17610

[42] Z. Yang, Z. Li, X. Jiang, Y. Gong, Z. Yuan, D. Zhao, and C. Yuan, "Focal and global knowledge distillation for detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 4643–4652.

[43] T. Séjourné, G. Peyré, and F.-X. Vialard, "Unbalanced optimal transport, from theory to numerics," *Handbook of Numerical Analysis*, vol. 24, pp. 407–471, 2023.

[44] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013.

[45] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine learning*, vol. 110, no. 3, pp. 457–506, 2021.

[46] A. Shapiro, "Monte carlo sampling methods," in *Stochastic Programming*, ser. Handbooks in Operations Research and Management Science. Elsevier, 2003, vol. 10, pp. 353–425.

[47] J. Redmon, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[48] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[49] J. Redmon, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[50] S. D'Amico, M. Benn, and J. Jørgensen, "Pose estimation of an uncooperative spacecraft from actual space imagery," *Int. J. of Space Science and Engineering*, vol. 2, pp. 171 – 189, 01 2014.

[51] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[52] S. Fort, H. Hu, and B. Lakshminarayanan, "Deep ensembles: A loss landscape perspective," *arXiv preprint arXiv:1912.02757*, 2019.