Depth-NOCTIS: Depth-based Novel Object Cyclic Threshold based Instance Segmentation

Max Gandyra[1, 2]
mgandyra@uni-bremen.de
Alessandro Santonicola[1, 2]
ale_san@uni-bremen.de
Michael Beetz[1]
beetz@uni-bremen.de

[1] AICOR Institute for Artificial Intelligence
University Bremen
28359 Bremen, Germany

[2] equal contribution

Abstract

Zero-shot instance segmentation of novel objects in RGB-D images is a well-known problem in computer vision and significant for many different applications, especially ones in which accurate identification of unseen objects without long retraining is required. In this work we propose Depth-NOCTIS (D-NOCTIS), a unified RGB-D segmentation method that seamlessly integrates depth information to enhance matching accuracy. This pipeline is based upon NOCTIS, leveraging Grounded-SAM 2 for object proposals with precise bounding boxes and corresponding segmentation masks; and DINOv2's zero-shot capabilities for robust *cls* (semantic) and *patch* (appearance) embeddings, while introducing a geometric consistency score. By using RGB-D images instead of RGB-only ones, like NOCTIS, this new score is able to better handle objects that are similar in appearance but differ in size and shape. We empirically show that Depth-NOCTIS through the fusion of RGB and depth based similarity scores, without further training/fine tuning, achieves substantial performance gains, in terms of mean absolute Average Precision (AP); over the best RGB and RGB-D methods on the seven core datasets of the BOP 2023 challenge for the "Model-based 2D segmentation of unseen objects" task.

1 Introduction and related work

The instance segmentation task, in which object instances are identified and located in images via segmentation masks, proves to be a crucial issue in robotics' perception and augmented reality applications, especially when zero-shot adaptability to novel objects without (further) training is required; e.g. a robot wants to identify a specific object instance on a conveyor belt. Historically, classical supervised learning frameworks have exhibited strong performances when target objects are fixed [10, 19, 23, 33, 38], yet their reliance on extensive training and labeled data limits deployment in dynamic and/or industrial environments, where the target objects change constantly.

Large-scale pretrained models have revolutionized zero-shot generalization. Vision transformers (e.g., ViT [5]), contrastive learners (CLIP [31]), and self-supervised backbones

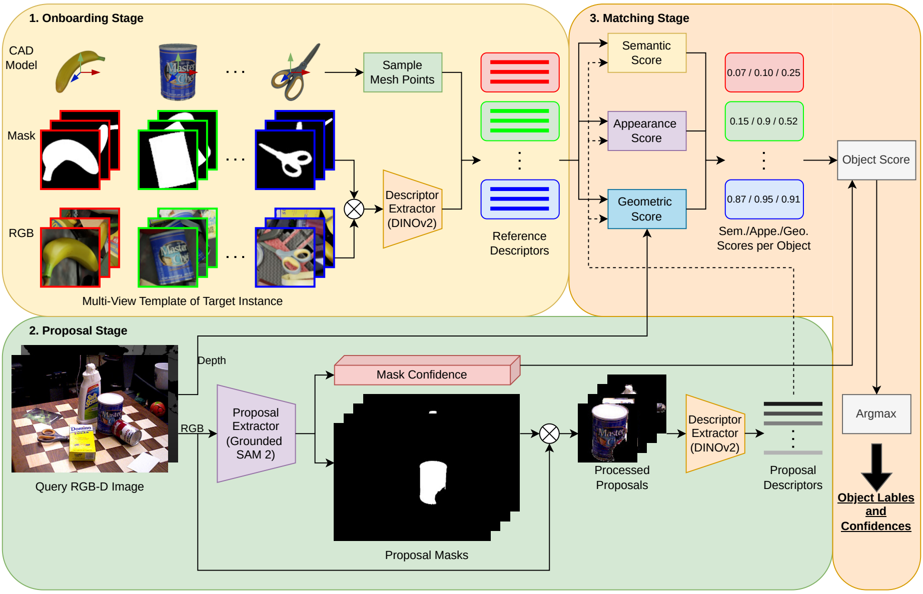

Figure 1: The three D-NOCTIS stages: onboarding stage, represents each object via descriptors from templates and sample points (Section 2.1); proposal stage (Section 2.2), where proposals as masks, and their descriptors, are generated from the query RGB-D image; lastly, in the matching stage, object labels and confidences are assigned to each proposal based on their descriptors (Section 2.3).

(DINOv2 [3, 29]) produce high-quality embeddings that generalize across tasks, i.e. classification, semantic segmentation, depth estimation, and novel instance retrieval. DINOv2 is used in this work due to its robustness against unseen object appearances.
Furthermore, visual foundation models such as Segment Anything (SAM) [17] and its variations/successors FastSAM [45], SAM 2 [32] and others [42, 44] prevail in image segmentation mask generation via enabling open world/class-agnostic scenarios. In recent times, a standard practice is to combine, in a modular way, the strengths of open-set detectors [15, 21, 24, 34] with SAM variants to solve complex problems; here, we employ Grounded-SAM 2 [35] for its efficiency and avoidance of spurious or fragmented proposals.
As already mentioned, classical instance segmentation methods, like Mask R-CNN [10] or similar [23, 33, 38], which demonstrated to be robust in challenging scenarios with heavy occlusions and lighting conditions, always needed to be fine-tuned on specific target objects [13]; making them unable to handle novel objects without retraining. ZeroPose [2] and CNOS [27] were among the first notable models that solved, in a training-free fashion, this task. The core architecture of the latter has also laid the foundation for subsequent models such as SAM-6D [22], NIDS-Net [25] and notably NOCTIS [8]; which combines semantic (*cls* tokens), appearance (*patch* tokens) and mask confidence scores; augmented by a cyclic patch filtering, to match query proposals against multi-view RGB templates.
The proposed pipeline, Depth-NOCTIS (D-NOCTIS), is an extension of NOCTIS that incorporates a newly defined geometric score using depth information into the object matching one, which is useful whenever one has to differentiate between "similar looking" objects with

different sizes or shapes, like e.g. texture-less but differently sized industrial objects.
In our evaluation on the seven core BOP 2023 benchmarks [13] for the “2D instance segmentation of unseen objects” task, D-NOCTIS achieves, without further training, in terms of mean absolute Average Precision (AP) metric, a 1.1% improvement over the best models (NOCTIS and MUSE); moreover, it outperforms the leading RGB-D zero-shot method (LDSeg) by 1.9%. The main contributions of our work can be summarized as follows:

1. We propose Depth-NOCTIS, a zero-shot framework for novel objects instance segmentation that uses vision foundation models and depth information to outperform the current state-of-the-art methods.
2. The introduction of a geometric consistency score exploiting depth for handling similar looking objects with different sizes/shapes.
3. Ablation study is conducted over several object matching score components, demonstrating significant performance gains when using the geometric score.

2 Method

In this section, we explain our approach for performing the instance segmentation, i.e. generating segmentation masks and labeling them, for all novel objects within an RGB-D query image $I \in \mathbb{R}^{4\times\mathrm{W}\times\mathrm{H}}$ with W and H being the width and height in pixels, respectively, and 3 RGB + depth channels; given a set of RGB template images and 3D model sample points of said objects and without any (re-)training.
Our approach, as shown in Figure 1, is carried out in three steps, similarly to [8, 22, 25, 27]. Starting with the onboarding stage in Section 2.1, visual descriptors are extracted from the template images via DINOv2; followed by the proposal stage in Section 2.2, where all possible segmentation masks and their descriptors, from the query RGB-D image, are generated with Grounded-SAM 2 and DINOv2, respectively. Lastly, in Section 2.3, the matching stage, each proposed mask is given an object label and a confidence value, based on the determined object scores using the visual descriptors.

2.1 Onboarding stage

Multiple visual descriptors are generated during the onboarding stage to represent each of the N^O different novel objects $\mathcal{O}$. In the following, in all the descriptions and notations, we will consider just one object $O \in \mathcal{O}$; this is done to keep the notation simple.
In detail, the object is represented by: N^{sample} sample points from its 3D/CAD model; a set of $\mathcal{T}$ template images; and their corresponding ground truth segmentation masks, showing the object from different predefined viewpoints. These templates and masks can either be pre-rendered with renderers like Pyrender [26] or BlenderProc [4] using the 3D model of the object and some fixed viewpoints, or even be extracted out of some selected frames, e.g. annotated videos, where the object is “visible enough” and has a viewpoint close to a predefined one.
In a preprocessing step, the segmentation masks are used to remove the background and to crop the object instance in each template, then, the crop size is unified via resizing and padding. Afterwards, the instance crops are fed into DINOv2, an image foundation network, creating a class embedding/*cls* token and N_T^{crop} patch embeddings/*patch* tokens for each template T in $\mathcal{T}$,

where N_T^{crop} denotes the number of not masked out patches within the cropped template mask ($N_T^{crop} \leq N^{patch}$). The cropped templates are internally divided into $N^{patch} = 256$ patches, on a 16×16 grid, for the *patch* tokens. The *cls* token and *patch* tokens, together, form the visual descriptor of each template.

2.2 Proposal stage

At this stage, all object proposals from the query image I are acquired. We decided to use Grounded-SAM 2 as proposal generator, since it was shown by the NOCTIS authors's [8] to work better than the original Grounded-SAM [35], SAM or FastSAM; which were used by previous works [1, 20, 22, 25, 27, 37]. Grounded-SAM 2 obtains the bounding boxes of all objects from Grounding-DINO [24], a pretrained zero-shot detector, matching a given text prompt; then, it uses these as a prompt for SAM 2 to create segmentation masks.
Accordingly, Grounded-SAM 2 with the text prompt “objects” is applied on the RGB part of the query image to extract all N^P foreground object proposals $\mathcal{P}$; note that N^P changes according to I. Furthermore, each proposal $p \in \mathcal{P}$ includes a bounding box, a corresponding segmentation mask and a confidence score for both of them. In next step, each proposal with a confidence score lower than a threshold value, or too small relative to the image size, is filtered out. Eventually, the visual descriptor of each proposal p is created using the pipeline from the previous section, where the preprocessing step creates the image crop I_p, which is then used by DINOv2 to generate the *cls* token and *patch* tokens.

2.3 Matching stage

During the matching stage, we determine the matching score for the considered proposal-object pair using the previously gathered descriptors; then, the object label that best suits the proposal is assigned together with its confidence score.
The object matching score s^{obj}, between a proposal p and an object O, represented by its templates and 3D sample points, is the combination of: a semantic score; an appearance score; a proposal confidence; and a newly introduced geometric score, to consider the size of the object.
In the following, a quick overview of the different scores already used/defined in NOCTIS is given. After that, a comprehensive explanation of our new geometric score is provided.

Semantic score The semantic score s^{sem} was established as a robust measure of semantic matching in CNOS and is based on the cosine similarity:

$$cossim(\mathbf{a},\mathbf{b}) = \frac{\langle \mathbf{a},\mathbf{b} \rangle}{||\mathbf{a}|| \cdot ||\mathbf{b}||}, \tag{1}$$

with $\langle , \rangle$ denoting the inner product. The score, is then computed for each proposal as the top-5 average across the cosine similarity values between its *cls* token and all object templates ones.

Appearance score with cyclic threshold The appearance score s^{appe} introduced in NOCTIS, is employed to discriminate between objects which are semantically similar, but with different patch/part-wise appearance. For each proposal-template pair, a sub-appearance score s_T^{appe} is computed as follows:

$$s_T^{appe} = \frac{1}{N_p^{crop}} \sum_{i=1}^{N_p^{crop}} \max_{j=1,\dots,N_T^{crop}} \big(cossim(\mathbf{g}_i, \mathbf{f}_j)\big) \cdot \chi\big(cdist(I_p, T, i) < \delta_{CT}\big), \tag{2}$$

where $\mathbf{g}_i$ and $\mathbf{f}_j$ are, respectively, the corresponding i-th/j-th *patch* token for the image proposal crop I_p and the template T. The function χ is the indicator function, turning the boolean values *True* and *False* to 1 and 0, respectively. Finally, the appearance score, for each proposal is obtained as the maximum sub-appearance score across all object templates.
To increase the quality of the patch-pairs, NOCTIS introduced a patch-pair filtering using patch-wise *cyclic* distance, here *cdist*, since DINOv2 descriptors can assign similar *patch* embeddings/tokens to repetitive textures/similar looking parts (e.g. identical corners or surfaces), leading to many-to-one matches. To filter out some of these unstable matches, a "relaxed" mutual similarity is enforced; thus, only patches whose *cyclic* distance is smaller then the threshold value $\delta_{CT} = 5$ survive.

Bounding box and segmentation mask confidence Proposals might contain a high number of false positives, indeed, background regions and object parts might be misinterpreted as complete objects. To account for this, for each proposal p, the proposal confidence $conf_p$, as the average confidence value of its bounding box and segmentation mask, is included as a weighting factor for the object matching score.

Geometric score While the previously mentioned scores are already good at handling visual different looking objects, they have problems with differentiating similar looking but differently sized/shaped ones. The introduction of a geometric score, as the one used by SAM-6D, addresses this problem and increases the performance in the area of texture-less but differently sized industrial objects. Following NOCTIS' approach of un-biasing the appearance score, we evaluate the geometric score in a similar fashion by computing it for all templates per object and aggregating results; rather than relying on a single template of the single object with the highest semantic score. Thus, we first define the sub-geometric score s_T^{geo} for each proposal p with template $T \in \mathcal{T}$ as the Intersection-over-Union (IoU or Jaccard index) of the proposal bounding box $\mathcal{B}_p$ and the template based one $\mathcal{B}_{T,i}$:

$$s_T^{geo} = \max_{i=1,\dots,N^{rot}} \left(\frac{|\mathcal{B}_p \cap \mathcal{B}_{T,i}|}{|\mathcal{B}_p \cup \mathcal{B}_{T,i}|} \right). \tag{3}$$

To get $\mathcal{B}_{T,i}$, first, a coarse pose estimation of the object is obtained by combining the object rotation of the template pose/viewpoint with a translation given by the centroid of the reprojected points using the proposal depth image and the camera intrinsic. Afterwards the pose is used to transform the 3D sample points/point cloud of the 3D object mesh, which are then projected onto the image plane, 2D/in-plane rotated and their min/max values are used to create the (axis aligned) bounding box $\mathcal{B}_{T,i}$. Where the index i determines which of the different possible N^{rot} 2D/in-plane rotations is used. The IoU can be increased by adding these in-plane rotations to compensate for the lack of them in the basic viewpoints (see ablation studies 3.3), which limited SAM-6D authors' approach as they relied only on the viewpoint rotations.
As this score is heavily affected by the objects' visibility/occlusion, a weighting factor that

penalizes for scarce visibility has been adopted. Then, we define the sub-visibility vis_T as follows:

$$vis_T = \frac{1}{N_T^{crop}} \sum_{i=1}^{N_T^{crop}} \chi\Big(\max_{j=1,\dots,N_p^{crop}} \big(cossim(\mathbf{g}_i, \mathbf{f}_j)\big) > \delta_{vis} \Big). \tag{4}$$

The value δ_{vis} is a threshold value needed to control the needed minimum visibility of an image patch/part.
Finally, the geometric score s^{geo}, i.e. the best $s_T^{geo} \cdot vis_T$ value across all templates, and its corresponding visual score vis are returned.

Object matching score By combining all these scores and the proposal confidence, we determine the object matching score s_p^{obj} for each proposal p as follows:

$$s_p^{obj} = \frac{s_p^{sem} + 2 \cdot s_p^{appe} + s_p^{geo}}{1 + 1 + vis_p} \cdot conf_p. \tag{5}$$

The object matching scores of all the N^P proposals, over all possible N^O objects, are stored in the $N^P \times N^O$ instance score matrix.

Object label assignment In the final stage, we apply the Argmax function across the objects/rows of the instance score matrix. The object label and its matching score are assigned to each proposal, indicating its corresponding confidence. Eventually, we obtain proposals consisting of: a bounding box of the object instance; its corresponding modal segmentation mask, which encompasses the visible instance part [13]; and an object label with a confidence score. To remove any proposals that may be incorrectly labeled, a confidence threshold δ_{conf} filtering is applied with $\delta_{conf} = 0.2$ as the default value for testing. Moreover, Non-Maximum Suppression is applied to eliminate redundant proposals.

3 Experiments

We begin by presenting our experimental setup (Section 3.1) and then compare it to the state-of-the-art ones for the seven core datasets of the BOP 2023 challenge [13] (Section 3.2). Finally, we perform a short ablation study regarding the score component choices in Section 3.3.

3.1 Experimental setup

Datasets We evaluate our method on the seven core datasets of the BOP 2023 challenge: LineMod Occlusion (LM-O) [1]; T-LESS [14]; TUD-L [12]; IC-BIN [6]; ITODD [7]; HomebrewedDB (HB) [16]; and YCB-Video (YCB-V) [43]. These datasets contain 132 household and industrial objects, which might be textured or not, and symmetric or asymmetric. Furthermore, they are shown in multiple cluttered scenes with varying occlusion and lighting conditions.

Evaluation metric The BOP 2023 challenge's standard protocol [13, Section 2.6] is followed to evaluate the "2D instance segmentation of unseen objects" task, thus we use the Average Precision (AP) as our criterion. The AP metric is computed as the average of precision scores, at different IoU thresholds, in the interval from 0.5 to 0.95 with steps of 0.05.

Implementation details To generate the proposals, we use Grounded-SAM 2, with an input text prompt "objects", comprised of the Grounding-DINO model with checkpoint "Swin-B" and SAM 2 with checkpoint "sam2.1-L". The corresponding regions of interest (ROIs) are resized to 224 × 224, while using padding to keep the original size ratios. We use the default "ViT-L" model/checkpoint of DINOv2, following the approaches in [8, 22, 25, 27], for extracting the visual descriptors as 1024-dimensional feature vectors, where each *patch* token on the 16 × 16 grid represents 14 × 14 pixels.
We use the "PBR-BlenderProc4BOP" pipeline with the same 42 predefined viewpoints, as described in CNOS [27, Sections 3.1 and 4.1], to select the templates representing every dataset object, because they perform better than the other choices (see the ablation studies of other works [8, 22, 25, 27]). For the 2D rotation $N^{rot} = 8$ equal distributed ones are used and the minimum visibility threshold is set to $\delta_{vis} = 0.35$.
The main code is implemented in Python 3.8 using Numpy [9] and PyTorch[30] (Version 2.2.1 CUDA 11.8). To ensure reproducibility, the seed values of all the (pseudo-) random number generators are set to 2025. The tests were performed on a single Nvidia RTX 4070 12GB graphics card and the average measured time per run, with one run using the same configuration on all the seven datasets, was approximately 130 minutes or 1.324 seconds per image on average.

3.2 Comparison with the state of the art

We compare our method with the best available results from the leaderboard[1] of the BOP challenges, comprising of the top-4 disclosed methods: CNOS [27], SAM-6D [22], NIDS-Net [25] and NOCTIS [8]; and the overall top-3 **undisclosed** ones: "anonymity", LDSeg and MUSE. CNOS uses proposals from SAM or FastSAM and only the semantic score 2.3 for matching. SAM-6D uses the same proposals and semantic score as CNOS, additionally, it also uses simpler versions of the appearance score 2.3 and the geometric one 2.3. NIDS-Net uses proposals from Grounded-SAM and the cosine similarity between the weight adapter refined Foreground Feature Averaging [18] embeddings together with SAM-6D's appearance score. NOCTIS is the model on top of which D-NOCTIS is based; the main difference between the two is that the latter employs depth data for computing a geometric score.
In the Table 1 we show the results for D-NOCTIS and the other methods on all seven datasets and their overall average. We surpass the best established method NIDS-Net by a significant margin, in terms of absolute mean AP, of 4.5% and the best ones (MUSE and NOCTIS) by 1.1%. This comparison clearly shows that fusing depth information with the previous RGB-only pipeline leads to a significant increase over the top performing methodologies. Indeed, while it might look negligible at a first glance, it is actually a significant one; given the nature of the BOP task, being able to improve upon the mean AP score proves to be quite difficult as one methodology might score better on certain datasets and worse on others (given their dissimilarity); leading to minimal gains in performance in most cases. This is

[1] https://bop.felk.cvut.cz/leaderboards/segmentation-unseen-bop23/bop-classic-core/; Accessed: 2025-08-14

		BOP Datasets							
Method	Depth	LMO	TLESS	TUDL	ICBIN	ITODD	HB	YCBV	Mean
CNOS	-	0.397	0.374	0.480	0.270	0.254	0.511	0.599	0.412
SAM-6D	✓	0.460	0.451	0.569	0.357	0.332	0.593	0.605	0.481
NIDS-Net	-	0.439	0.496	0.556	0.328	0.315	0.620	0.650	0.486
LDSeg	✓	0.478	0.488	**0.587**	0.389	0.370	0.622	0.647	0.512
anonymity	-	0.471	0.464	0.569	0.386	0.376	0.628	0.688	0.512
MUSE	-	0.476	0.486	0.550	0.408	0.382	**0.636**	**0.702**	0.520
NOCTIS	-	0.489	0.479	0.583	0.406	0.389	0.607	0.684	0.520
D-NOCTIS	✓	**0.499**	**0.517**	0.542	**0.421**	**0.432**	0.614	0.692	**0.531**

Table 1: Comparison of D-NOCTIS (ours) against different methods on the seven core datasets of the BOP 2023 challenge [13], w.r.t. the AP metric (higher is better). For each dataset, the best result is displayed in bold and the second best is underlined.

	s^{geo}	2D Rot	*vis*	Mean
0	✓	✓	✓	**0.531**
1	-	-	-	0.520
2	✓	-	-	0.525
3	✓	-	✓	0.529
4	✓	✓	-	0.527

Table 2: Ablation studies on the influence of different components on the mean AP metric.

further supported by the fact that the top 4 models, excluding D-NOCTIS, exhibit a mere 0.8% absolute discrepancy in mean AP.

Figure 2 shows some qualitative segmentation results of our method compared to the publicly available ones, with red arrows indicating any errors in the masks and/or proposals classifications. It is evident that each method has its own strengths and weaknesses. Using SAM/FastSAM as a proposal generator, like SAM-6D does, one has difficulties in distinguishing between objects and some of their parts. While NIDS-Net, due to its usage of Grounded-SAM, is more robust, it can still misclassify by labeling scene objects wrongly or by creating oversized bounding boxes around identified objects, leading to multiple detections. NOCTIS suffers less, on average, from said problems; but it can still misclassify objects that are too similar looking or too close to each other; see columns 1 (left clamp), 4 (lamp holder) and 6 (cylinder hull) for reference. D-NOCTIS with the help of geometric information can avoid some of the previous issues, for some cases, as seen in the YCB-V image; indeed, the large clamp is correctly labeled despite looking very similar to the small one; furthermore, one is able to discriminate objects in industrial (textureless) scenarios, like in the TLESS and ITODD cases. On average, D-NOCTIS shows fewer errors overall than the other methods.

3.3 Ablation studies

In Table 2, we show the influence of the geometric score, the visualization one and the in-plane (2D rotation) on the mean AP metric. Line 0 shows the result attained by the complete

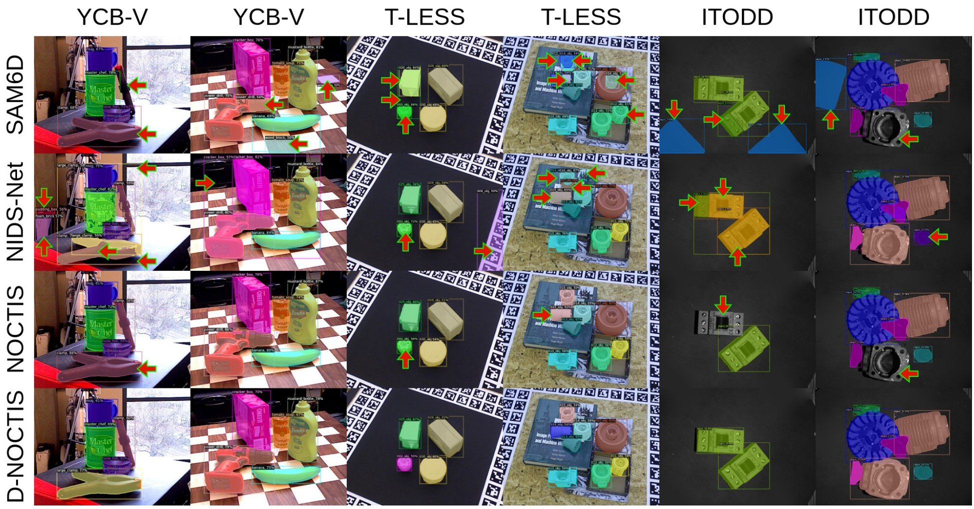

Figure 2: Qualitative assessment of some segmentation results using SAM-6D, NIDS-Net, NOCTIS and D-NOCTIS on YCB-V, T-LESS and ITODD. The image addresses the strengths and limitations of these methods. The red arrows indicate errors in the segmentation masks and/or classifications of the proposals. For better visualization purposes, $\delta_{conf} = 0.5$ was used.

D-NOCTIS model as shown in Table 1 and line 1 shows the original NOCTIS result (no geometric component, pure RGB). The comparison between the two clearly shows that there is a significant increase of 1.1% in absolute mean AP through including the depth information. This increase is built steadily by adding one by one the three components, as shown by lines 2, 3 and 0 in Table 2. The highest incremental gain in performance is obtained when one adopts s^{geo} as shown by line 2; however, the other two increments are still significant 0.4% for the *vis* score and 0.2% for the added in-plane rotation. However, without the geometric components, the other two can not be employed as stated in Section 2.3; thus their total contribution can not be evaluated standalone.

3.4 Discussion and limitations

A limiting factor of our geometric score is most likely the use of a bounding box-based IoU instead of a more detailed one, e.g., between segmentation masks. While the choice of extracting a bounding box from the projected rotated image plane points was due to its ease of implementation (proof of concept), as one just needs to search the min/max values for the xy-coordinates, future work could improve upon this (and with that also on the influence of the 2D rotation) by creating segmentation masks based on convex hulls or alpha-shapes approaches.

Due to a lack of other results for the BOP classic Extra datasets (LM [11]; HOPEv1 [41]; RU-APC [36]; IC-MI [39] and TYO-L [12]), our method is only evaluated on the seven core BOP 2023 datasets; moreover, the BOP 2024 [28] and 2025 [40] challenges are solely focused on a detection task. But, as mentioned in Section 3.1, the chosen datasets include many different scenes, so their evaluation should still be reliable.

Method	Depth	Mean AP	Time
SAM-6D (FastSAM)	-	0.428	0.249
SAM-6D (FastSAM)	✓	0.449	0.445
SAM-6D	-	0.450	2.281
SAM-6D	✓	0.481	2.795
NOCTIS	-	0.520	0.990
D-NOCTIS (w/o 2D Rot)	✓	0.529	1.054
D-NOCTIS	✓	0.531	1.324

Table 3: Mean AP and time per image comparison between different versions of SAM-6D and NOCTIS (with and without depth).

In Table 3, we show a comparison of the time per image needed by different versions of SAM-6D and NOCTIS (in seconds). As it can be noticed, the inclusion of depth information increases the overall mean AP score at the expense of a higher processing time. Furthermore, the D-NOCTIS configuration without the additional in-plane rotation (similar to Table 3.3 row 3) seems to have the best balance between the increases in time (6%) and w.r.t. absolute mean AP (0.9%) compared to NOCTIS; making it the preferred model of choice for more time-sensitive scenarios. We also want to note that, while time is usually an important factor for real-world applications, it is not an evaluation criterion for the BOP task. As each participant can use their own hardware for evaluation, e.g. “GeForce RTX 3090 24GB” (SAM-6D), “V100 16GB” (CNOS) or “Nvidia RTX 4070 12GB” (ours), this heavily influences the runtime (“normal” graphic cards vs. server ones), making a direct comparison of the time per image (almost) meaningless.

4 Conclusion

In this paper we presented Depth-NOCTIS, a new framework for zero-shot novel object instance segmentation; which builds upon NOCTIS to increase its performances by adopting a depth-based geometric score for object sizes/shapes. As shown by the experimental results, our method performed better than all other methodologies, in terms of mean absolute AP score, on the seven core datasets of the BOP 2023 benchmark.

Acknowledgment

This paper was done under the “FAME” project, supported by the European Research Council (ERC) via the HORIZON ERC Advanced Grant no. 101098006. While funded by the European Union, views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. *Learning 6D Object Pose Estimation Using 3D Object Coordinates*, page 536–551. Springer International Publishing, 2014. ISBN 9783319106052. doi: 10.1007/978-3-319-10605-2_35.

[2] Jianqiu Chen, Zikun Zhou, Mingshan Sun, Rui Zhao, Liwei Wu, Tianpeng Bao, and Zhenyu He. Zeropose: Cad-prompted zero-shot object 6d pose estimation in cluttered scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2): 1251–1264, February 2025. ISSN 1558-2205. doi: 10.1109/tcsvt.2024.3482439.

[3] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024. URL https://arxiv.org/abs/2309.16588.

[4] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8 (82):4901, 2023. doi: 10.21105/joss.04901.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[6] Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim. Recovering 6d object pose and predicting next-best-view in the crowd. In *2016 IEEE Confeeedingsrence on Computer Vision and Pattern Recognition (CVPR)*, pages 3583–3592, 2016. doi: 10.1109/CVPR.2016.390.

[7] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Härtinger, and Carsten Steger. Introducing mvtec itodd — a dataset for 3d object recognition in industry. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2200–2208, 2017. doi: 10.1109/ICCVW.2017.257.

[8] Max Gandyra, Alessandro Santonicola, and Michael Beetz. Noctis: Novel object cyclic threshold based instance segmentation, 2025. URL https://arxiv.org/abs/2507.01463.

[9] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. doi: 10.1109/ICCV.2017.322.

[11] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In Kyoung Mu Lee, Yasuyuki Matsushita, James M. Rehg, and Zhanyi Hu, editors, *Computer Vision – ACCV 2012*, pages 548–562, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-37331-2. doi: 10.1007/978-3-642-37331-2_42.

[12] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent-Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiri Matas, and Carsten Rother. Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/papers/Tomas_Hodan_PESTO_6D_Object_ECCV_2018_paper.pdf.

[13] Tomas Hodan, Martin Sundermeyer, Yann Labbe, Van Nguyen Nguyen, Gu Wang, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, and Jiri Matas. Bop challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5610–5619, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024W/CV4MR/papers/Hodan_BOP_Challenge_2023_on_Detection_Segmentation_and_Pose_Estimation_of_CVPRW_2024_paper.pdf.

[14] Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jirí Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888, 2017. doi: 10.1109/WACV.2017.103.

[15] Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy, 2024. URL https://arxiv.org/abs/2403.14610.

[16] Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2767–2776, 2019. doi: 10.1109/ICCVW.2019.00338.

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023. doi: 10.1109/ICCV51070.2023.00371.

[18] Klemen Kotar, Stephen Tian, Hong-Xing Yu, Dan Yamins, and Jiajun Wu. Are these the same apple? comparing images based on object intrinsics. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 40853–40871. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/803c6ab3d62346e004ef70211d2d15b8-Paper-Datasets_and_Benchmarks.pdf.

[19] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 574–591, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58520-4.

[20] Bowen Li, Jiashun Wang, Yaoyu Hu, Chen Wang, and Sebastian Scherer. Voxdet: Voxel learning for novel instance detection. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 10604–10621. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/21f1c5bbf2519321c1bee9bfa9edcd46-Paper-Conference.pdf.

[21] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10975, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/papers/Li_Grounded_Language-Image_Pre-Training_CVPR_2022_paper.pdf.

[22] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui 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)*, pages 27906–27916, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lin_SAM-6D_Segment_Anything_Model_Meets_Zero-Shot_6D_Object_Pose_Estimation_CVPR_2024_paper.html.

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, February 2020. ISSN 1939-3539. doi: 10.1109/tpami.2018.2858826.

[24] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. *Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection*, page 38–55. Springer Nature Switzerland, November 2024. ISBN 9783031729706. doi: 10.1007/978-3-031-72970-6_3.

[25] Yangxiao Lu, Jishnu Jaykumar P, Yunhui Guo, Nicholas Ruozzi, and Yu Xiang. Adapting pre-trained vision models for novel instance detection and segmentation, 2025. URL https://arxiv.org/abs/2405.17859.

[26] Matthew Matl. Pyrender, 2021. URL https://github.com/mmatl/pyrender.

[27] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Vincent Lepetit, and Tomas Hodan. Cnos: A strong baseline for cad-based novel object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2134–2140, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023W/R6D/papers/Nguyen_CNOS_A_Strong_Baseline_for_CAD-Based_Novel_Object_Segmentation_ICCVW_2023_paper.pdf.

[28] Van Nguyen Nguyen, Stephen Tyree, Andrew Guo, Mederic Fourmy, Anas Gouda, Taeyeop Lee, Sungphill Moon, Hyeontae Son, Lukas Ranftl, Jonathan Tremblay, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, Stan Birchfield, Jiri Matas, Yann Labbe, Martin Sundermeyer, and Tomas Hodan. BOP challenge 2024 on model-based and model-free 6D object pose estimation. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW, CV4MR Workshop)*, 2025. URL https://arxiv.org/abs/2504.02812.

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[32] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

[34] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, Yuda Xiong, Hao Zhang, Feng Li,

Peijun Tang, Kent Yu, and Lei Zhang. Grounding dino 1.5: Advance the "edge" of open-set object detection, 2024. URL https://arxiv.org/abs/2405.10300.

[35] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded SAM: assembling open-world models for diverse visual tasks. *CoRR*, abs/2401.14159, 2024. doi: 10.48550/ARXIV.2401.14159.

[36] Colin Rennie, Rahul Shome, Kostas E. Bekris, and Alberto F. De Souza. A dataset for improved rgbd-based object detection and pose estimation for warehouse pick-and-place. *IEEE Robotics and Automation Letters*, 1(2):1179–1185, 2016. doi: 10.1109/LRA.2016.2532924.

[37] Qianqian Shen, Yunhan Zhao, Nahyun Kwon, Jeeeun Kim, Yanan Li, and Shu Kong. A high-resolution dataset for instance detection with multi-view object capture. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 42064–42076. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/832ea0ff01bd512aab28bf416db9489c-Paper-Datasets_and_Benchmarks.pdf.

[38] Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6728–6738, 2022. doi: 10.1109/CVPR52688.2022.00662.

[39] Alykhan Tejani, Danhang Tang, Rigas Kouskouridas, and Tae-Kyun Kim. Latent-class hough forests for 3d object detection and pose estimation. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 462–477, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10599-4. doi: 10.1007/978-3-319-10599-4_30.

[40] Antonin Melenovsky Tomas Hodan and Mederic Fourmy. Benchmark for 6d object pose estimation: Bop challenge 2025. https://bop.felk.cvut.cz/challenges/, 4 2025. Accessed: 2025-04-16.

[41] Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13081–13088, 2022. doi: 10.1109/IROS47612.2022.9981838.

[42] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Rep vit: Revisiting mobile cnn from vit perspective. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15909–15920, 2024. doi: 10.1109/CVPR52733.2024.01506.

[43] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018. doi: 10.15607/RSS.2018.XIV.019.

[44] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications, 2023. URL https://arxiv.org/abs/2306.14289.

[45] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything, 2023. URL https://arxiv.org/abs/2306.12156.