6D Pose Estimation of Unseen Objects for Industrial Augmented Reality

Hugo Durchon, Marius Preda, Titus Zaharia

6D pose estimation of unseen objects for industrial augmented reality

Hugo Durchon
SAMOVAR
Télécom SudParis, IP Paris
Évry-Courcouronnes, France
hugo.durchon@telecom-sudparis.eu

Marius Preda
SAMOVAR
Télécom SudParis, IP Paris
Évry-Courcouronnes, France
marius.preda@telecom-sudparis.eu

Titus Zaharia
SAMOVAR
Télécom SudParis, IP Paris
Évry-Courcouronnes, France
titus.zaharia@telecom-sudparis.eu

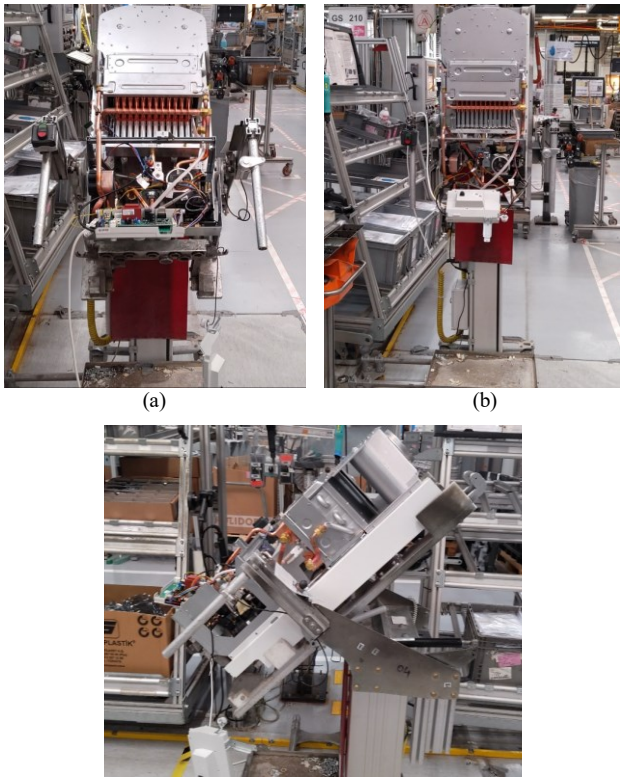
***Abstract*—This paper addresses the challenges of implementing markerless Augmented Reality (AR) in complex manufacturing settings. Making AR systems more intuitive, robust, and adaptable is a required step to make their adoption possible in the industry. Among the hard constraints encountered in uncontrolled, real-world environments, we notably face the dynamic nature of production lines and the evolving appearance of the objects during the assembly process. Emerging deep learning (DL) methods enable 6D object pose estimation for AR registration of moving objects. However, they need a significant amount of 6D object pose ground truth data. In real-world scenarios, such a requirement cannot be fulfilled, because of two factors: the complexity of establishing an accurate 6D pose labeling procedure for large objects in a real production line and the wide variety of object states and appearances encountered along the assembly line. For this reason, it is necessary to develop alternative 6D pose estimation techniques capable of handling unseen objects. To this end, this paper introduces a novel pipeline relying on HoloLens 2 for data capture, Neural Radiance Fields (NeRF) for 3D model generation, and MegaPose for 6D pose estimation. The proposed approach enables 6D pose estimation without object-specific training or laborious pose labeling.**

Keywords— Industrial augmented reality, deep learning, 6D object pose estimation

I. INTRODUCTION

The emergence of Industry 4.0 has generated great interest in applying augmented reality (AR) to various levels and use cases of the supply chain. More specifically, industrial assembly performed by human operators has a lot to gain from AR adoption [1]. AR-based assistants can be used to summarize industrial procedures, check the quality of the assembly, detect unsafe situations, and train novice operators [2]. It has been demonstrated that AR is a relevant medium to guide operators during assembly tasks [3] due to its capacity to interactively display pertinent information in the real world at the right time. Despite the expected benefits, the adoption of AR-based assistants remains limited in practice, and notably in use cases occurring in real industrial environments [4] involving complex industrial objects and uncontrolled environments.

The use case considered in this paper concerns a real-world boiler production line from the elm.leblanc - Bosch manufactories (Fig. 1 [1]), which raises some important challenges. Such production lines lead to cluttered and dynamic scenes. Indeed, the boilers are moved along the assembly line through different workstations and thus their pose is continuously changing. In addition, the shape and visual appearance of the boiler evolve during the assembly process (Fig. 2).



(a) (b) (c)

Fig. 1. Pictures of the boiler on the lifting trolley at one of the same late stages of assembly. (a) front view of the boiler tilted by 45 degrees (b) front view of the boiler with no tilt, but with the max elevation (c) side view of the boiler tilted by 45 degrees. These pictures show the boiler 3D pose changes based on the operator's needs.

A key stage in AR applications concerns the registration step, which aims to align the virtual content with the real world. This step enables the placement of virtual content at meaningful locations so that the user can visualize easily the task to do and interact with the virtual content.

For the time being, the most commonly used registration method is to add a simple visual feature, such as a QR code, to the scene. The marker can be detected and tracked at every frame and acts as an anchor for placing the virtual content. Most of the time, the marker is positioned at a known, fixed location and remains static. As a result, such a method cannot keep virtual content aligned with moving objects. A different possibility would be to set up markers on all moving objects.

[1] Please note that all figures presented in this paper are acquired in real-life conditions at the elm.leblanc - Bosch manufacture in Drancy, France.

979-8-3315-3997-9/24/$31.00 ©2024 IEEE

However, markers must remain visible, which is not trivial for the various objects that are present on the assembly line (boilers, tools, components…), due to inherent occlusions. As a consequence, the AR application will not be able to relocate the virtual content correctly and will lead to significant registration errors. Moreover, precisely positioning virtual content manually at a default location during AR authoring is burdensome. Both issues increase the cognitive load of operators and reduce the assistance quality provided by the AR application [5].

Therefore, our objective is to perform markerless registration of AR content that can automate both the initial placement of the virtual content and the correct update of its location when boilers are moving through the assembly line. Solving this task can help AR applications display dynamic virtual content [6], be more responsive, and provide an overall enhanced user experience.

The markerless registration of virtual content can be formulated as a 6D object pose estimation problem. The objective is thus to perform 6D pose estimation of the boiler at different assembly states.

In recent years, deep learning (DL) methods have revolutionized various computer vision tasks, with spectacular results. Given the highly difficult nature of our task, with uncontrolled environments and boilers represented as complex, evolving objects with reflective or texture-less parts, DL techniques offer a promising axis of research. Let us note that although several DL approaches consider the issues of 2D and 3D object detection for AR scenarios [7], their application to 6D object pose estimation within the framework of industrial AR use cases remains limited.

However, DL-based 6D pose estimation methods require a large amount of labeled training data, which is highly difficult to obtain in real-life industrial settings. To overcome such limitations, we propose a novel pipeline relying on the HoloLens 2 device for acquisition, Neural Radiance Fields (NeRF) [8] for 3D object reconstruction, and the MegaPose 6D pose estimation technique [9]. Trained on millions of synthetic scenes with thousands of different objects, the MegaPose approach offers the advantage of being able to be applied without re-training to estimate the pose of unseen, novel objects. The proposed approach makes it possible to perform 6D pose estimation of several boiler models at different assembly states as required by the considered industrial AR use case.

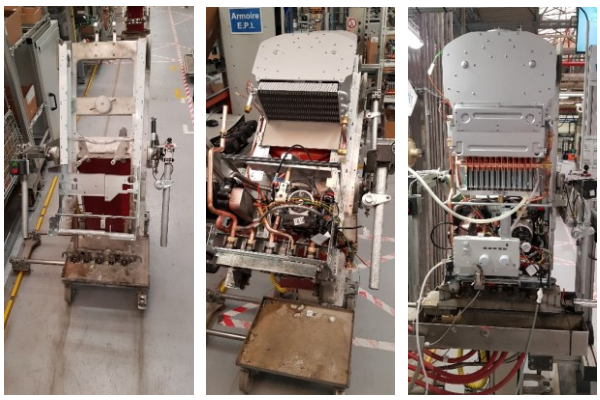


Fig. 2. Changes in the appearance of a boiler during the assembly for a given model.

The rest of the paper is structured as follows. Section II briefly introduces the state of the art in traditional DL-based 6D pose estimation, then analyzes their applications to industrial AR use cases, and lastly details the task of 6D pose estimation on unseen objects. Section III presents the proposed pipeline and the results obtained. Finally, Section IV concludes the paper and draws some perspectives for future work.

II. Related Work

A. 6D pose estimation for industrial AR applications

DL-based 6D object pose estimation is a fast-growing research field. The goal of 6D object pose estimation is to infer the 6D pose of the object of interest, expressed as 3D translation and rotation parameters. Traditionally, DL-based methods are built upon a 3D object detector or feature extractor and use as input either RGB-D data [10] or RGB data only [11]. Such methods are composed of a 6D object pose estimator network supplemented by a pose refinement method which can be DL-based [12] or not, like Iterative Closest Point (ICP) [13]. The 6D object pose estimation methods can be trained to determine the pose of a specific object instance [14, 15, 16] or a whole object category [17, 18, 19].

The application of DL-based 6D object pose estimation methods to industrial AR is still an emerging field due to technological challenges, data limitations, and interdisciplinary nature. This explains the relatively limited number of papers directly addressing this issue.

In [20], the 6D pose estimation of an industrial milling machine has been performed from RGB-D input. The RGB data are captured using a HoloLens 1, and the depth input is captured from an IoT depth camera. The model combines four deep neural networks, each dedicated to a specific task: RGB image segmentation, RGB feature extraction, 3D feature extraction from the reconstructed object point cloud, and finally 6D pose estimation. In order to ensure the related computational requirements, the inference is performed on a remote server. The obtained pose results are finally sent back to the HoloLens 1 device.

MANTRA [21] is an AR system able to align temperature information on industrial objects like water pumps and electronic cardboards. First, the method detects the object and estimates its pose through the association of the Yolo4 model with an altered LINEMOD method [22]. Then, the ICP algorithm [13] is applied to perform a pose refinement. Lastly, a 6D pose tracking module allows updating the pose of the object in real-time. The inputs are RGB-D images, thermal images, and CAD object models. The AR display is screen-based. The method is robust enough to display virtual content on parts of the electronic board.

In [23], the authors introduce a CNN (*Convolutional Neural Network*) approach able to jointly perform object pose and state estimation. The object is a coffee machine with removable parts that can be used to simulate the assembly. The network is based on TridentNet and Faster R-CNN architectures and exploits two distinct CNNs: a first one responsible for object detection and state estimation and a second one for 6d pose estimation. Interestingly, the proposed approach is able to deal with objects being assembled.

The above-cited methods are instance-based and require training a DL model for every new object. They provide accurate pose estimates for learned instance objects but have

inherently weak generalization. This limitation hinders scalability in our evolving industrial environment, where the AR assistant must interact with a large variety of boilers at different assembly stages for which an instance-based pose estimation model has not been trained. Consequently, such 6D instance-based pose estimation methods are impractical in our case.

To address the issue of per-object instance DL retraining, a second family of methods, so-called category-based, has been developed. These methods aim to estimate poses for entire object categories rather than specific instances. Category-based approaches typically rely on large corpora of annotated pose data that are usually produced with dedicated annotation tools. For example, the T-LESS data set [24] uses a sophisticated manual pose labeling process enabled by multiple cameras at known positions to ensure accurate ground truth poses. The development of a pose annotator tool like for the Objectron dataset [25] facilitates the pose labeling process, but human effort is still needed and increases consequently with the number of considered objects. This requirement represents a strong limitation, especially in industrial settings where capturing accurate 6D pose ground truth for large objects is challenging.

The use of synthetic data reduces the need for real-world data acquisition and manual pose labeling. Indeed, synthetic data can be generated from the 3D models of the objects or environments of interest. A model is then trained on these synthetic data and transfer learning is applied to fine-tune the model to real-world data. Such a transfer learning strategy was used for pose estimation [26]. An alternative is to use a self-training mechanism that generates pseudo-pose labels from the synthetic model later used to self-train the model on real unlabeled data. This strategy was effectively applied to AR registration and integrated into a mobile application for industrial robot manipulation [27].

Despite these advancements, category-based methods, like other approaches that rely on training or fine-tuning, need to be adapted for new, unseen objects, which limits their capacity to be used in the wild in our reconfigurable industrial environment. For these reasons, we have not retained category-based techniques in our developments.

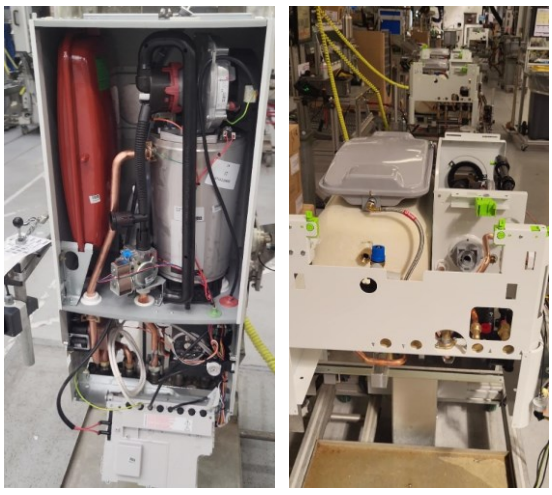


Fig. 3. Examples of two different boiler models in the production line.

B. 6D pose estimation of unseen objects

While instance-based and category-based methods have made significant strides in 6D object pose estimation, they are limited by their reliance on pre-defined object instances or categories. To address this limitation and facilitate the application of DL-based 6D object pose estimation methods to real-world scenarios, recent works have focused on the challenging task of estimating poses for unseen novel objects. This evolution towards more generalizable solutions is crucial for industrial applications where encountering new object variants or objects being assembled is common.

In contrast to the traditional supervised workflow of training and inference used in instance-based and category-based methods, approaches designed for unseen objects introduce an extra step called object onboarding. This step occurs between training and inference and helps the DL model to adapt to entirely new objects. During onboarding, the DL model is provided with a 3D mesh [9] or a few views with pose labels [28] of the novel unseen test object. This approach eliminates the need for retraining for new objects and enables the model to estimate 6D poses of previously unseen objects from input test images during inference.

Two families of DL-based 6D pose estimation methods for unseen objects can be identified: template-based and feature-based matching techniques [29]. Both leverage deep learning for object detection or segmentation, feature extraction, pose estimation, and refinement, but differ in their approach. These modules are often a combination of CNNs or Vision Transformers.

Template-based methods [9, 29] use deep neural networks to extract high-level features from both the cropped region of interest of the test image and a set of rendered image templates from the onboarded 3D mesh. These methods then classify the rendered templates to find the best match with the test image. This classification yields an initial pose hypothesis, which is subsequently refined by another deep neural network. This process can be repeated to generate multiple pose hypotheses and deal with pose ambiguities more easily. To enhance robustness, templates are rendered with diverse lighting conditions, material properties, and background textures. However, these approaches can be computationally intensive due to the extensive rendering requirements and the several forward passes.

Feature-based methods [29, 30] focus on extracting local features from both the segmented test image and the object-onboarding input (3D mesh or labeled views). After feature extraction, these methods use another network to perform the feature matching. They are faster than template-based methods thanks to the absence of template rendering. However, they may still struggle with viewpoints that lack distinctive local features or present pose ambiguities.

DL models capable of performing predictions on unseen objects align with our industrial context. They relieve us from the constraint of capturing 6D pose ground truth and retraining the model for every model and assembly state. Considering that boilers present textureless and reflective parts, and certain background elements closely resemble boiler components, thus potentially generating similar features, template-based methods appear more suitable in our case.

We have used the Benchmark for 6D Object Pose Estimation (BOP) [31] to make the final decision on the pose estimation model to retain. BOP has been organizing since 2017 yearly challenges to track the advancement of 6D pose estimation as part of the R6D workshop. They provide datasets and performance metrics to compare methods and

establish state-of-the-art on several tasks related to 6D pose estimation. In 2023, BOP introduced a task in the challenge for "Model-based 6D localization of unseen objects". At inference, such models require a 3D model as onboarding input.

We have adopted the MegaPose technique [9] to perform the pose estimation because it offers the backbone model of the method that won the best open-source method in the task in 2023 [32]. Furthermore, MegaPose works with rough or low-fidelity 3D models and has already been applied to a practical use case for visual guidance of a robotic arm similar to AR use cases.

III. Proposed Methodology

Our pipeline (Fig. 4) is composed of three main components: HoloLens 2 for data collection, NeRF for 3D model generation, and MegaPose for 6D pose estimation. The DL part of the pipeline uses Pytorch.

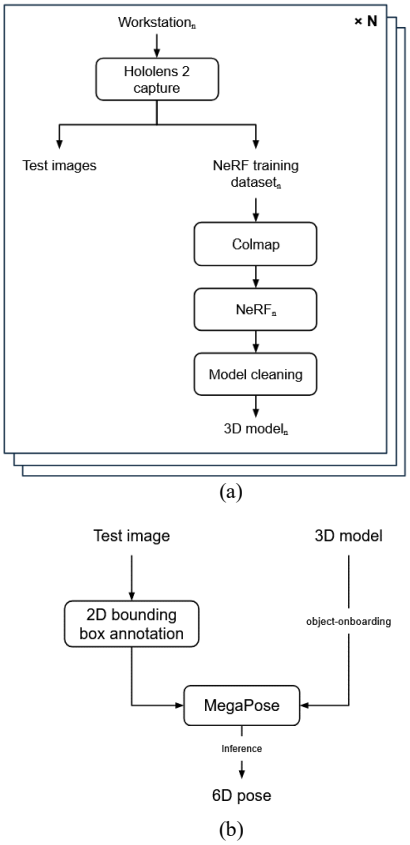


Fig. 4. Proposed pipeline. (a) Data processing including NeRFs models training and 3D models rendering for every workstation. (b) Inference with onboarded 3D models and annotated test images

A. Data collection and 3D model generation

Data collection was conducted using the HoloLens 2 [33] to simulate typical augmented reality use case conditions. Videos are captured around boilers in each workstation of the production lines to capture different viewpoints.

In the assembly industry, 3D models of small components, tools, or final products are often available. In our case, we do not have access to boiler 3D models at every assembly state. Consequently, our approach uses Neural Radiance Fields (NeRF) [8] to generate the required 3D models. Indeed, professional photogrammetry such as Agisoft Metashape [34] is expensive and demands more computational resources.

NeRF is a method capable of generating new views of a given scene or object from an input video. For each image, NeRF casts a set of rays into the image space, and sampling is performed along these rays to select points. Then, two multilayer perceptrons learn the color and density of these points given their position and direction as input. Lastly, a traditional volume renderer is used to generate the color of rays and the loss is computed. At the end of the training, the weights of the NeRF contain all the scene information. Using these trained weights, new RGB images of the scene can be rendered. It is also possible to generate a 3D mesh of the scene by using the marching cube algorithm [35] or the Poisson reconstruction technique [36]. For each workstation, approximately 150 images are acquired, and a dedicated NeRF model is trained. A structure-from-motion pipeline called COLMAP [37] is used to create a sparse model of the workstation scenes containing the transform information (camera calibration and pose of each image with respect to the camera) required to train each NeRF model. We used the Nerfstudio library and the Nerfacto model [38]. This model incorporates several advanced techniques including multi-hash encoding, proposal sampling, and scene contraction. The training takes around 30 minutes per workstation and the corresponding 3D model rendering takes 5 minutes on a GTX 1080 GPU. An example of 3D reconstruction of a boiler model through the assembly line is illustrated in Fig. 5.

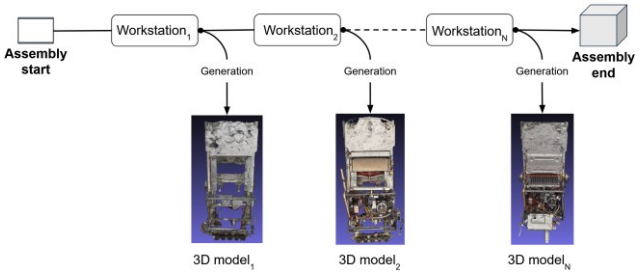
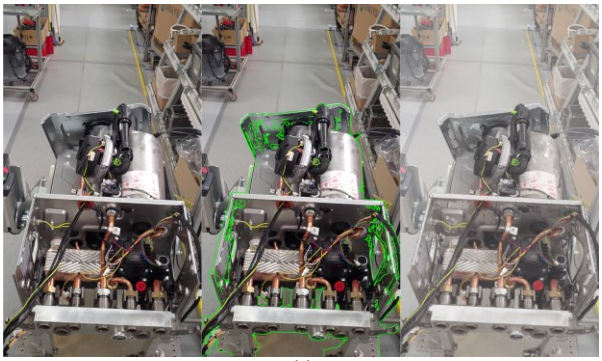


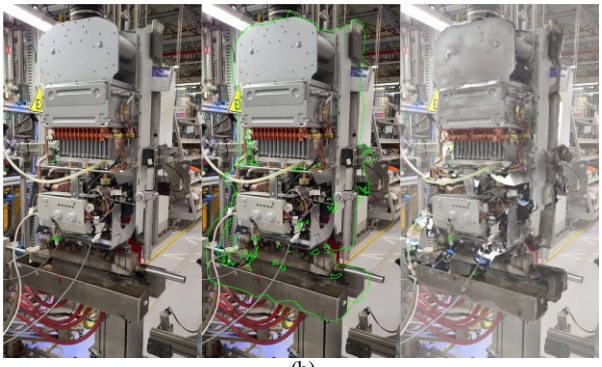
Fig. 5. Example of 3D meshes generated with NeRFs for one boiler model at different workstations through the assembly line. A 3D mesh is generated at the end of each workstation and not for every part added to the boiler.

B. 6D boiler pose estimation

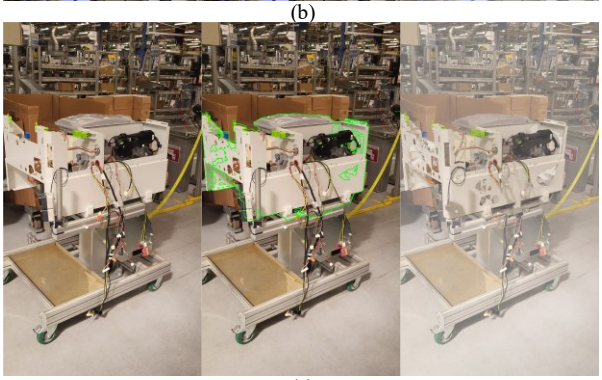
We can directly perform pose prediction on our data using MegaPose without any need for retraining. Indeed, MegaPose is trained on millions of synthetic scenes and a large number of objects. MegaPose takes as input an RGB image annotated with a region of interest, camera intrinsic parameters, and a 3D model to predict the 6D pose of an unseen object. MegaPose is composed of a coarse pose estimator model and a pose refiner model. The objective of the coarse pose estimator is to predict an initial pose estimate. Then, the pose refiner iteratively updates the pose by comparing the input image with the rendered image. The architecture of both models is based on ResNet-34. For the inference, we take as input the test image captured from HoloLens 2 and as onboarding input the 3D model of the object at the current workstation. HoloLens 2 calibration has been performed with OpenCV and a chessboard pattern to obtain the camera's intrinsic parameters. The inference takes around 50 seconds to run on our data on a GTX 1080 GPU.



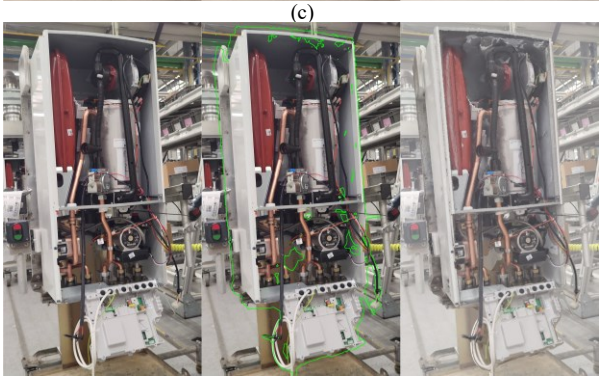
(a)



(b)



(c)

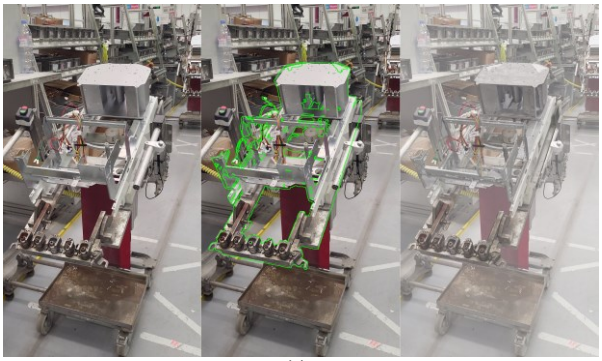


(d)

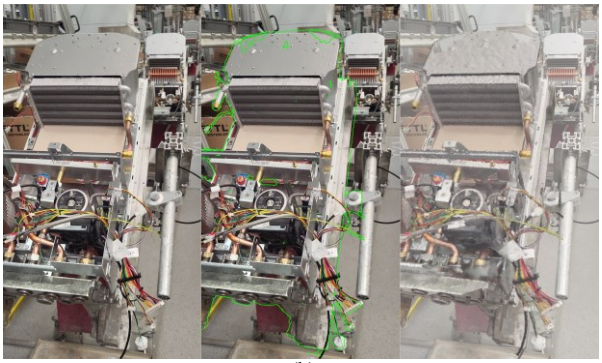
Fig. 6. Results of MegaPose 6D pose estimator for four different boilers models under different viewpoints (a) top (b) front right (c) far back and (d) down. From left to right, the first image is the test RGB image. The second image is the superposition of the test image with the edges of the 3D mesh projected at the predicted pose. The third image is the superposition of the test image with the 3D mesh projected at the predicted pose.

Several 6D pose estimation results are illustrated in Fig. 6 (for successful pose estimations of different boiler models under various viewpoints) and Fig. 7 (which shows successful results at different assembly steps for a given boiler model). From left to right the figures present the test image, the edges of the NERF model superposed on the test image at the predicted pose, and finally, the entire 3D model projected at the predicted pose and superposed on the test image. We can observe that globally, the overall orientation, translation, and scaling factors are correct. We consider that the results obtained are sufficient for display purposes in an AR application.

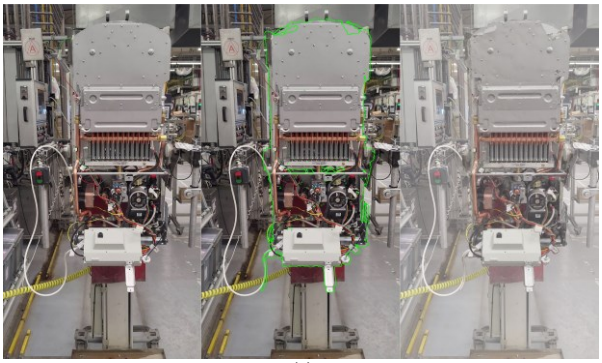
The results illustrated in Fig. 6 and Fig. 7 are obtained using the single hypothesis variant of MegaPose. To improve the results in case of ambiguous pose, it is possible to use the multi-hypothesis variant of MegaPose: the coarse model extracts the top-K hypotheses, runs K refiner iterations for each hypothesis, and then selects the highest refined hypothesis. Fig. 8 shows some examples of improvements obtained by considering the MegaPose multiple hypotheses variant. The trade-off for considering more initial hypotheses and running more refiner iterations is an increased computational cost that needs to be considered for AR applications.



(a)

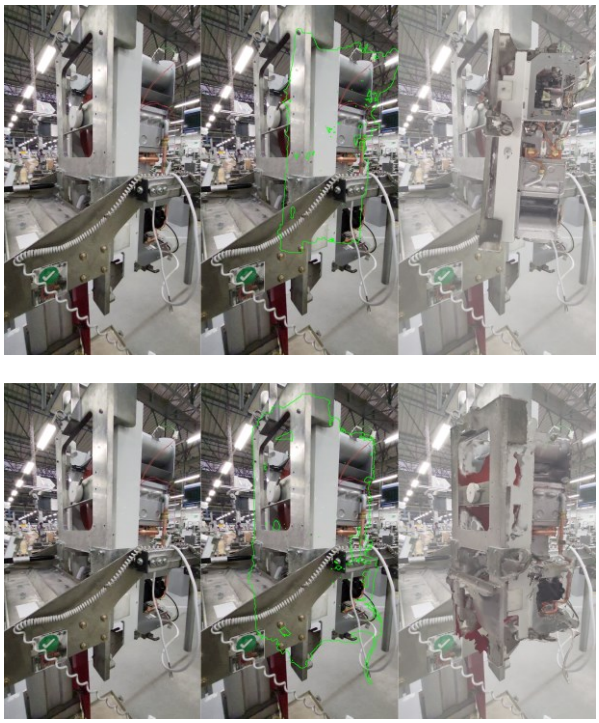


(b)

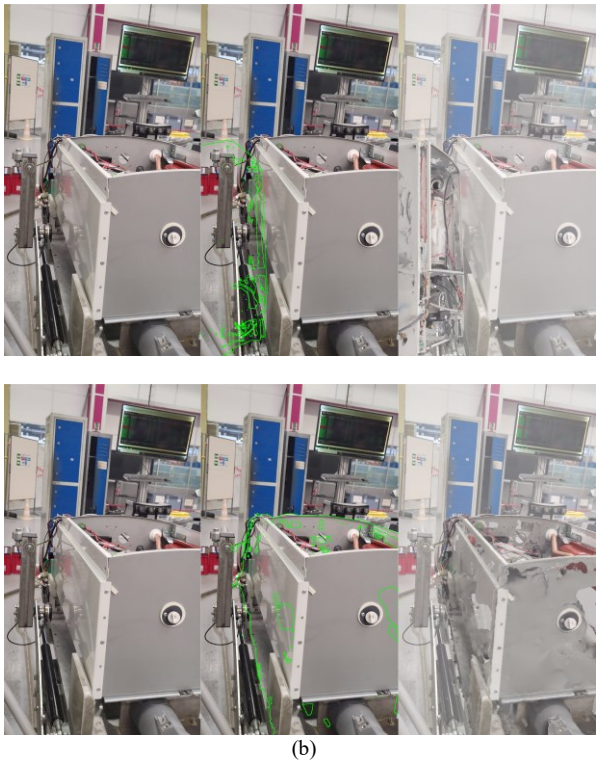


(c)

Fig. 7. Results for one boiler model at different assembly states with different object poses: (a) 60 degrees (b) 45 degrees (c) 90 degrees.



(a)



(b)

Fig. 8. Example of results improvements from single pose hypothesis (images on the top) to multiple (K=5) hypothesis (images on the bottom). (a) The pose was reversed. (b) The predicted pose was outside of the region of interest. In both cases, the multiple hypothesis approach improves the obtained predictions.

Significant failure cases are illustrated in Fig. 9. The large deviations are caused by incorrect initial pose estimates from the coarse model. The pose refiner is trained to correct poses on a restricted interval and cannot correct such consequent pose errors. Even with multiple hypotheses, we were unable to obtain better results on these images. We observe that failures occur in two main scenarios: when only poorly textured parts of the boiler are visible, or when there is significant occlusion of the boiler's fine details. In such cases, the coarse pose estimator proposes a hypothesis that fits the external contours of the scene, but which may be highly different from the correct one.

Table 1. summarizes the results obtained. We have run inferences on five different boiler models at five different assembly states. We have selected ten different viewpoints with respect to the object (front, front right, front left, back, top, down, right, left, top front, and down front). We split the results between a single hypothesis and K=5 multiple hypotheses. The obtained results demonstrate the superiority of the multiple hypotheses approach.

TABLE I. PERCENTAGES OF CORRECT ESTIMATIONS IN SINGLE AND 5-HYPOTHESES MODES

	Correct percentage	***Average inference time***
Single hypothesis	***58.1 %***	***50 s***
5-hypothesis	***77.6 %***	***60 s***

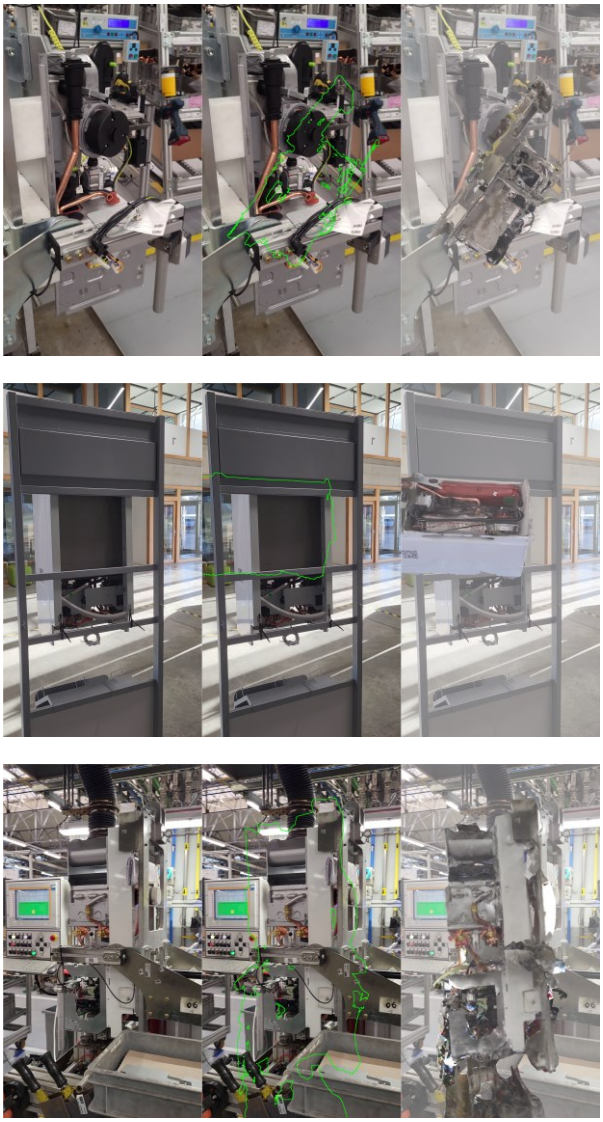


Fig. 9. Significant failure cases.

C. Perspectives of improvement

Despite the slight errors and the failure cases shown above, the pipeline is promising enough to be further enhanced and adapted to run within a HoloLens 2 application. Indeed, the 6D object pose estimation is performed without training and the need for manual pose annotation. The pipeline works for our different models of boilers at different assembly steps and can be directly used for other objects of interest in the production line such as assembly components or auxiliary equipment.

To improve the pipeline's accuracy, speed, and practical applicability in industrial AR settings, we propose several areas of enhancements:

1. Capture RGBD data with HoloLens 2 sensors. The additional depth channel would help solve the pose errors along the depth dimension.
2. Replace NeRF with Gaussian Splatting [39], which offers faster training and view rendering. Gaussian Splatting has been successfully applied in AR for real-time rendering of virtual content. However, it generates a set of 3D Gaussian primitives rather than a 3D mesh. This approach would require either an extra step to generate a 3D model from the Gaussian primitives or adjustments to pose estimation methods to accept Gaussian primitives as object-onboarding input.
3. Incorporate zero-shot segmentation methods, such as those used in [40]. Segmentation would allow for more precise feature extraction from the exact region of interest in the image.
4. Adapt and optimize the pose estimation model for HoloLens 2 deployment. This simplification process may be more straightforward using the latest state-of-the-art methods such as [40, 41] addressing some limitations of MegaPose and offering improved accuracy and significantly faster performance.
5. Incorporate a tracking module, such as in [42], as an essential component for our final industrial AR application.

IV. Conclusion

This research aims to make AR systems more intuitive, robust, and adaptable, thereby advancing the adoption of AR in the industry. We addressed two real-world constraints in our industrial AR use case: dynamic environments and objects with evolving appearances. These constraints emphasize the need for 6D pose estimation and methods capable of handling unseen objects. Our proposed approach, relying on HoloLens 2, NeRF, and MegaPose, eliminates the need for object-specific training and laborious pose labeling, making it adaptable to various boiler models and assembly states. Furthermore, this pipeline can be efficiently applied to any new objects relevant to industrial AR use cases, such as assembly tools, indicating its broader versatility. While our results show that the pipeline achieves reasonable pose estimates in many cases, there are opportunities for enhancement, particularly in handling challenging viewpoints and improving overall accuracy. We discussed promising improvements for future work, including the use of RGBD data, replacement of 2D bounding box annotation by zero-shot segmentation, incorporation of a tracking module, and simplification of the DL model for execution on HoloLens 2. Additionally, conducting a quantitative evaluation and a user study would further validate and improve the system's performance.

References

[1] E. Ras, F. Wild, C. Stahl, and A. Baudet, "Bridging the skills gap of workers in Industry 4.0 by human performance augmentation tools: Challenges and roadmap". In Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments, 2017, p. 428-432.

[2] S. Webel, U. Bockholt, T. Engelke, N. Gavish, M. Olbrich and C. Preusche, "An augmented reality training platform for assembly and maintenance skills". Robotics and autonomous systems, 2013, vol. 61, no 4, p. 398-403.

[3] A. Tang, C. Owen, F. Biocca, Frank, and W. Mou, "Comparative effectiveness of augmented reality in object assembly". In Proceedings of the SIGCHI conference on Human factors in computing systems. 2003. p. 73-80.

[4] T. Masood, and J. Egger, "Augmented reality in support of Industry 4.0—Implementation challenges and success factors". Robotics and Computer-Integrated Manufacturing, 2019, vol. 58, p. 181-195.

[5] D. K. Baroroh, C. H. Chu and L. Wang, "Systematic literature review on augmented reality in smart manufacturing: Collaboration between human and computational intelligence". Journal of Manufacturing Systems, 2021, vol. 61, p. 696-711.

[6] G. Zhao, Ganlin, P. Feng, J. Zhang, C. Yu, and J. Wang, "Rapid offline detection and 3D annotation of assembly elements in the augmented assembly". Expert Systems with Applications, 2023, vol. 222, p. 119839.

[7] Y. Ghasemi, H. Jeong, S. H. CHOI, K. B. Park, and J. Y. Lee ,"Deep learning-based object detection in augmented reality: A systematic review". Computers in Industry, 2022, vol. 139, p. 103661.

[8] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis". Communications of the ACM, 2021, vol. 65, no 1, p. 99-106.

[9] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, et al. "Megapose: 6d pose estimation of novel objects via render and compare". arXiv preprint arXiv:2212.06870, 2022.

[10] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, p. 3343-3352.

[11] S. Zakharov, I. Shugurov, and S. Ilic, "Dpod: 6d pose object detector and refiner". In Proceedings of the IEEE/CVF international conference on computer vision, 2019, p. 1941-1950.

[12] F. Manhardt, W. Kehl, N. Navab, and F. Tombari, "Deep model-based 6d pose refinement in rgb". In Proceedings of the European Conference on Computer Vision (ECCV), 2018, p. 800-815.

[13] J. Zhang, Y. YAO, and B. Deng. "Fast and robust iterative closest point". IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, vol. 44, no 7, p. 3450-3466.

[14] 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.

[15] 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.

[16] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, p. 16611-16621.

[17] M. Tian, M. H. Ang, and G. H. Lee, "Shape prior deformation for categorical 6d object pose and size estimation". In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. Springer International Publishing, 2020, p. 530-546.

[18] X. Li, H. Wang, L. Yi, L. J. Guibas, A. L. Abbott, and S. Song, "Category-level articulated object pose estimation". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, p. 3706-3715.

[19] M. Z. Irshad, T. Kollar, M. Laskey, K. Stone, and Z. Kira, "Centersnap: Single-shot multi-object 3d shape reconstruction and

categorical 6d pose and size estimation". In 2022 International Conference on Robotics and Automation (ICRA), IEEE, 2022, p. 10632-10640.

[20] Y. Sun, S. N. R Kantareddy, J. Siegel, A. Armengol-Urpi, X. Wu, H. Wang, and S. Sarma, "Towards industrial iot-ar systems using deep learning-based object pose estimation". In 2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC), IEEE, 2019, p. 1-8.

[21] M. Ortega, E. Ivorra, A. Juan, P. Venegas, J. Martínez, and M.Alcañiz, "Mantra: An effective system based on augmented reality and infrared thermography for industrial maintenance". Applied Sciences, 2021, vol. 11, no 1, p. 385.

[22] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, et al. "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes". In : Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11. Springer Berlin Heidelberg, 2013. p. 548-562.

[23] Y. Su, J. Rambach, N. Minaskan, P. Lesur, A. Pagani, and D. Stricker, "Deep multi-state object pose estimation for augmented reality assembly". In 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), IEEE, 2019, p. 222-227.

[24] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects". In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2017, p. 880-888.

[25] A. Ahmadyan, L. Zhang, A. Ablavatski, J. Wei, and M. Grundmann, "Objectron: A large scale dataset of object-centric videos in the wild with pose annotations". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, p. 7822-7831.

[26] J. L. Charco, A. D. Sappa, B. X. Vintimilla, and H. O. Velesaca, "Transfer Learning from Synthetic Data in the Camera Pose Estimation Problem", in VISIGRAPP (4: VISAPP), 2020, pp. 498–505.

[27] K. B. Park, S. H., Choi, and J. Y. Lee, "Self-training based augmented reality for robust 3D object registration and task assistance". Expert Systems with Applications, 2024, vol. 238, p. 122331.

[28] Y. He, Y. Wang, H. Fan, J. Sun, and Q. Chen, "Fs6d: Few-shot 6d pose estimation of novel objects". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, p. 6814-6824.

[29] J. Liu, W. Sun, H. Yang, Z. Zeng, C. Liu, J. Zheng et al. "Deep Learning-Based Object Pose Estimation: A Comprehensive Survey". arXiv preprint arXiv:2405.07801, 2024.

[30] M. Gou, H. Pan, HS. Fang, Z. Liu, C. Lu, and P. Tan. "Unseen object 6D pose estimation: a benchmark and baselines". *arXiv preprint arXiv:2206.11808*, 2022.

[31] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. GlentBuch, D. Kraft et al. "Bop: Benchmark for 6d object pose estimation". In Proceedings of the European conference on computer vision (ECCV), 2018, p. 19-34

[32] T. Hodan, M. Sundermeyer, Y. Labbe, V. N. Nguyen, G. Wang, E. Brachmann et al. "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, 2024, p. 5610-5619.

[33] Microsoft: HoloLens 2, AR headset. 2021. https://www.microsoft.com/en-us/hololens/hardware/

[34] Agisoft Metashape - https://www.agisoft.com/

[35] W. E. Lorensen ORENSEN, and H. E. Cline. "Marching cubes: A high resolution 3D surface construction algorithm". In : Seminal graphics: pioneering efforts that shaped the field. 1998. p. 347-353.

[36] M. Kazhdan, M. Bolitho, and H. Hoppe. "Poisson surface reconstruction". In : Proceedings of the fourth Eurographics symposium on Geometry processing. 2006.

[37] J. L. Schonberger, and J. M. Frahm, "Structure-from-motion revisited". In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, p. 4104-4113.

[38] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, T. Wang et al. "Nerfstudio: A modular framework for neural radiance field development". In ACM SIGGRAPH 2023 Conference Proceedings, 2023, p. 1-12.

[39] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering". ACM Trans. Graph., 2023, vol. 42, no 4, p. 139:1-139:14.

[40] 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, 2024, p. 27906-27916.

[41] V. N. Nguyen, T. Groueix, M. Salzmann, and V. Lepetit, "Gigapose: Fast and robust novel object pose estimation via one correspondence". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, p. 9903-9913.

[42] B. Wen, W. Yang, j., Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, p. 17868-17879.