Improved Zero-Shot Object Localization using Contextualized Prompts and Objects in Context

***Abstract*—Localizing objects is an essential capability for robots to do tasks more autonomously. For instance, finding the door and its handle to open it and navigate to the next room. For scalability to open world settings, it is important to localize objects which have not been seen before (zero-shot). For instance, the door has a push bar, instead of the conventional handle. Pretrained large language-vision object detection models, such as GLIP, can localize a broad variety of object classes reasonably well based on textual prompts and are ideal for zero-shot robotics. We extend GLIP with contextual knowledge to diversify the input prompts for better recall (pre-processing) and to filter the candidate objects using relational information of objects in context for better precision (post-processing). Diversification of prompts is helpful to cover variations of the object (e.g., different types of door handles). Spatial relations of objects are helpful to verify object candidates (e.g., the handle is close to the door). This verification is done by a neuro-symbolic program, endowed with first-order logic to define the spatial relations. We show that recall and precision of GLIP can be improved by leveraging contextual knowledge and without retraining.**

***Index Terms*—object localization, open world robotics, language-vision model, prior knowledge**

I. Introduction

Localizing objects is an essential capability for robots to do tasks more autonomously. An example is a robot that needs to explore a building. The robot should be able to find the door and its handle in order to open it and to navigate to the next room. A robot may have to look for a broad set of objects. In an open world, the robot may even have to localize novel objects, for which it was not trained specifically. It is however not scalable to learn a new or extended model for every new object of interest. Therefore, we consider the problem of localizing novel objects in a zero-shot manner. We do so by vision, i.e., object detection: producing bounding boxes with a label for a given image. We leverage recent pretrained large language-vision models that have a reasonable representation of a broad set of object classes, such that they can also detect novel classes [1]–[4]. We hypothesize that a large model such as the zero-shot object detector GLIP [5] is able to detect such objects in varying settings. GLIP has a very impressive zero-shot performance. Indeed, various types of doors are detected well by GLIP when prompting it for 'door'. This holds even for special cases of doors, such as an elevator door. However, we find that GLIP does not generalize across similar object classes. For instance, when prompting for 'handle', it does not find a door's push bar or knob. Hence the robot will not be able to open the door, when using this simple prompting strategy. To improve this, we use contextual knowledge to extend the prompt to a set of related relevant classes, e.g., 'handle' is extended to 'handle', 'push bar', 'knob'. The aim of this step is to improve the recall of detecting objects, as we want the robot to find as many relevant objects as possible.

The robot should also be efficient. To that end, we aim to improve the precision. We hypothesize that the precision can be increased by taking related objects in the context into account. For instance, a handle is on the door, or very close to it. These spatial relations can be used to further discriminate within the candidate detections, to assess which one is most likely the right one. GLIP can deal with composed prompts [5], such as 'the handle on the door', for which it will output a set of object boxes for both 'the handle' and 'the door'. Surprisingly, we find that both the precision and recall degrade when querying GLIP for composed prompts. Therefore, we take a different route to improve precision. We use contextual knowledge to search for objects in context. We use prior knowledge about spatial relations of objects, e.g., the handle is close to the door. These relations are defined in terms of first-order logic, with predicates about spatial constraints, such as (but not limited to) proximity and relative location. The object(s) of interest are assessed by probabilistic reasoning about the candidate objects produced by GLIP. For probabilistic reasoning, we consider neuro-symbolic programming [6]–[9] because it can deal with uncertainty of object candidates, for which we use the confidences from GLIP. We adopt a framework that can make a trade-off between accuracy and computational scalability [8]. This is necessary, given that there can be many candidate objects in an image. We show that the precision of candidate objects can be improved by this neuro-symbolic reasoning over spatial constraints.

In summary, we provide a method that combines zero-shot GLIP object detection with contextual knowledge, effectively improving the recall by diversifying prompts and the precision by taking spatial relations into account. We demonstrate this on a large set of images, where we improve on object detection metrics. Also, we provide illustrations of improvements and remaining problems, showing its benefits and limitations for object localization in open-world robotics.

II. Related Work

Large progress has been achieved in language-vision tasks. Language-vision models learn directly from huge datasets of images with textual descriptions, which offers a broad source of supervision [1]–[4]. They have shown great promise to generalize beyond crisp classes and towards semantically related classes. This so-called zero-shot capability is beneficial for classifying images into a broad set of classes, even ones

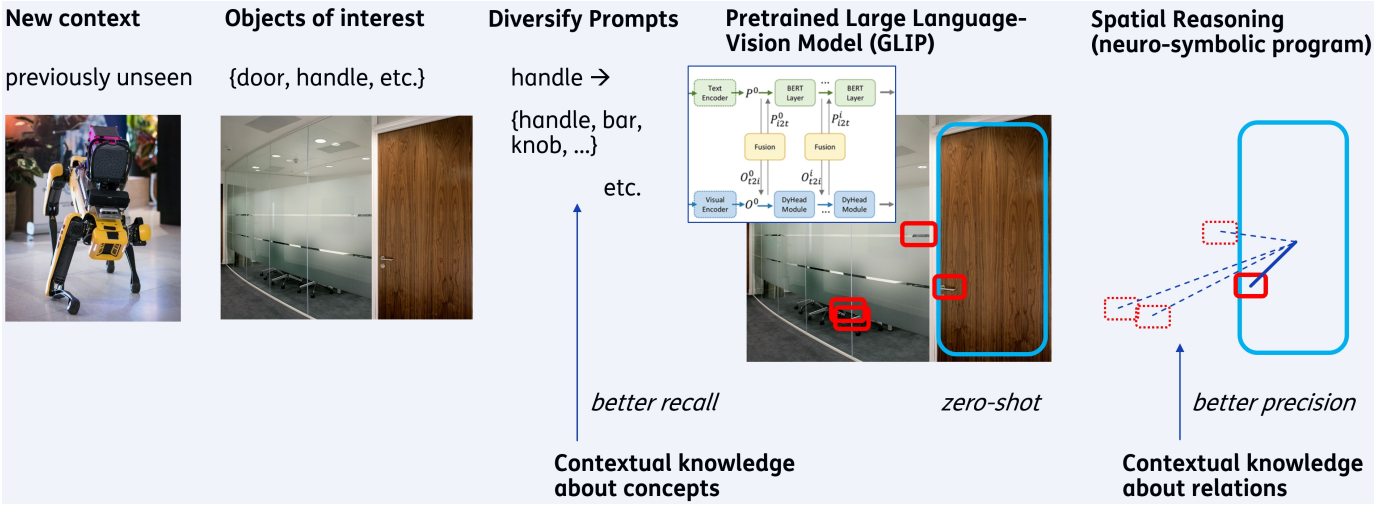


Fig. 1: Zero-shot localization of objects based on a pretrained large language-vision model such as GLIP. We leverage prior contextual knowledge to diversify prompts (better recall) and validate spatial relations (better precision).

that were not seen during training. Recently, these models were extended with capabilities to localize objects in images via co-attentions [10] and to segment parts of the scene based on textual descriptions [11]. Recently a method was proposed called GLIP [5], which for a textual prompt directly provides bounding box estimates. Our interest is in localizing objects. Therefore, we take GLIP as a starting point. Our first contribution is to extend it with contextual knowledge to create diverse prompts with related classes in order to improve the recall of the found objects. To analyze objects in context, knowledge about spatial relations can be leveraged. Connecting knowledge representation [12] and reasoning mechanisms with deep learning models [13] show great promise for reusing knowledge, more efficient learning and higher-level reasoning tasks [14]. Previous reasoning methods based on logic, such as DeepProbLog [6], [9], [15], were limited in terms of scalability when there were many possible hypotheses. A more efficient variant of DeepProbLog was proposed [7]. Recently, a framework was proposed that further improved the efficiency: the neuro-symbolic programming framework called Scallop [8]. Scallop is based on first-order logic and introduces a tunable parameter k to specify the level of reasoning granularity. It restrains the validation of hypotheses by the top-k proofs. This asymptotically reduces the computational cost while providing relative accuracy guarantees. This is beneficial for our purpose, as we expect many possible hypotheses in complex environments with many objects and imperfect observations. Our second contribution is to extend GLIP with the ability to search for objects in context.

III. Method

We provide a method that adds contextual knowledge to a language-vision object detection model; in this paper we use GLIP [5] with zero-shot capabilities. Our focus is on providing effective contextual knowledge at the input prompts (pre-processing) and filtering of candidate objects (post-processing) using relational information of objects in context. These are implemented as two respective modules, see Figure 1 for an outline of our method. The proposed modules are described in the next two subsections. We illustrate the effect of each module on the Doors images dataset for robotics [16] (illustrations are shown in the Appendix in Figure 6). A more detailed performance evaluation is presented in Section IV.

A. Diversifying Prompts (Better Recall)

Our first contribution is to diversify the set of prompts to have a better coverage of the object(s) of interest, thereby improving the recall. We use logic production rules based on external knowledge, e.g.:

$$handle \rightarrow \{\, handle, push\text{–}bar, knob \,\} \tag{1}$$

This extends the set of objects that can open a door. This improvement is implemented as a pre-processing step before applying the object model. Figure 2 illustrates the effect of our module. It shows a handle that is localized more accurately, when prompting for ‘bar’ instead of ‘handle’. The improved localization leads to an improved recall.

B. Spatial Relations (Better Precision)

Our second contribution is to take contextual knowledge about spatial relations into account. For instance, the handle is close to the door. For object candidates for door and handle it can be verified whether they fulfill this spatial relation. This verification is performed by a neuro-symbolic program [8], which operates on the (often uncertain) object candidates produced by GLIP. Each candidate object has a confidence associated to its bounding box in the image. The neuro-symbolic program takes first-order logic as contextual input and verifies the candidates accordingly, taking their confidences into account. The spatial relations are defined by symbolic predicates about e.g. proximity and relative location of objects.

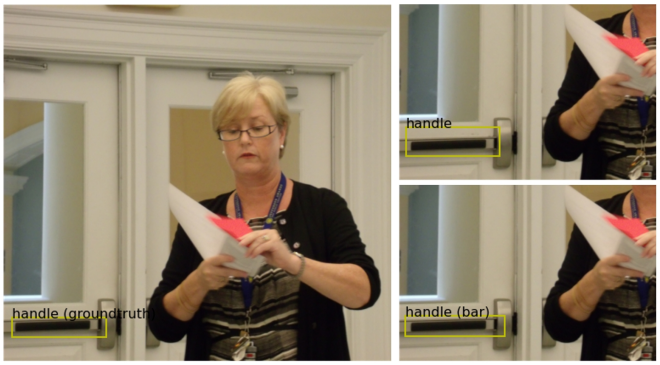


Fig. 2: Diversifying the prompts using contextual knowledge improves the recall. On the left the groundtruth; on the right the detections for searching for 'handle' (top) and searching for handle' with the additional prompt 'bar' (bottom).

$$\begin{aligned} \exists\, d, h : & \, object(d, door) \wedge object(h, handle) \wedge \\ & correct\text{–}size(d, h) \wedge \\ & correct\text{–}range\text{–}horizontal(d, h) \wedge \\ & correct\text{–}range\text{–}vertical(d, h) \end{aligned} \tag{2}$$

This defines that the combination of a door and a handle should be such that the handle has the correct size and at the correct horizontal and vertical range with respect to the door, which in turn are defined as:

$$correct\text{–}size(d, h) = max(1 - \frac{surface(h)}{surface(d)}, 0) \tag{3}$$

$$\begin{aligned} & correct\text{–}range\text{–}horizontal(d, h) = \\ & \qquad max(1 - \frac{hor\text{–}dist\text{–}from\text{–}side(h, d)}{width(d)}, 0) \end{aligned} \tag{4}$$

$$\begin{aligned} & correct\text{–}range\text{–}vertical(d, h) = \\ & \qquad max(1 - \frac{vert\text{–}dist\text{–}from\text{–}middle(h, d)}{height(d)}, 0) \end{aligned} \tag{5}$$

to express respectively that the handle should be small compared to the door, and that it should be close to the horizontal side of the door and vertically close to the middle of the door.

The neuro-symbolic program is implemented as a post-processing step after the object model. It improves the precision by maintaining only the candidates that fulfill the desired spatial relations.

Figure 3 illustrates the effect of this module: the confidence for the handle is increased because it fulfills the spatial relation that a handle should be close to the door. The improved confidence for these objects in context leads to an improved precision.

Fig. 3: Precision improves by verifying contextual knowledge about spatial relations. Left: groundtruth location of the handle. Right top: detection of the handle with highest confidence when no spatial reasoning is applied. Right bottom: detection of the handle with improved confidence after reasoning about its relation relative to the door.

IV. Experiments

We evaluate the performance of our method on a large set of 60 indoor images from the Doors dataset [16]. For examples we refer to the Appendix, Figure 6. We evaluate GLIP as-is [5] and with our proposed extensions. The standard metrics are the mean average precision (mAP) and mean average recall (mAR), for a minimum overlap between the groundtruth and detected boxes. This overlap is measured by intersection-over-union (IoU). Since the annotations of handles are sloppy, we evaluate both at the standard IoU = 0.5 and also at IoU = 0.35 to compensate for misaligned annotations. We refer to the Appendix, Figure 7 for examples that motivate that IoU = 0.35 is indeed sensible. Table I summarizes our findings. GLIP as-is with a single prompt for each object class {door, handle}, cannot find all handles (0.810). A composed prompt such as 'a handle on a door' is able to find almost all handles (0.968) and doors (0.969), but the precision is low for doors (0.528). Our diversified set of prompts improves both the recall (0.969 $\rightarrow$ 0.984) and the precision significantly (0.528 $\rightarrow$ 0.960), but the precision for the handles is still low (0.511). This precision is improved significantly by taking the spatial relations into account (0.511 $\rightarrow$ 0.694), while only losing a small bit of recall (0.968 $\rightarrow$ 0.921). In summary, the proposed extensions on top of GLIP are effective.

Figure 8 (Appendix) shows examples of GLIP detections without our modifications. Figures 4 and 9 (Appendix) show examples where our method achieves a perfect result (mAP = 1). Interestingly, the cases are very different, showing broad applicability. All cases have some clutter, yet the door and

TABLE I: Object Localization on the Door Images Dataset [16]

		IoU $\geq$ 0.5				IoU $\geq$ 0.35			
		Doors		**Handles**		**Doors**		**Handles**	
GLIP prompting	**Spatial relations**	mAP	mAR	mAP	mAR	mAP	mAR	mAP	mAR
single prompt for each class	-	0.960	0.984	0.124	0.810	0.970	0.984	0.444	0.984
composed prompt of classes	-	0.528	0.969	0.511	0.968	0.536	0.984	0.634	1.000
diversified prompt set (ours)	-	0.960	0.984	0.511	0.968	0.970	0.984	0.634	1.000
diversified prompt set (ours)	neuro-symbolic program (ours)	0.960	0.984	0.694	0.921	0.970	0.984	0.943	1.000

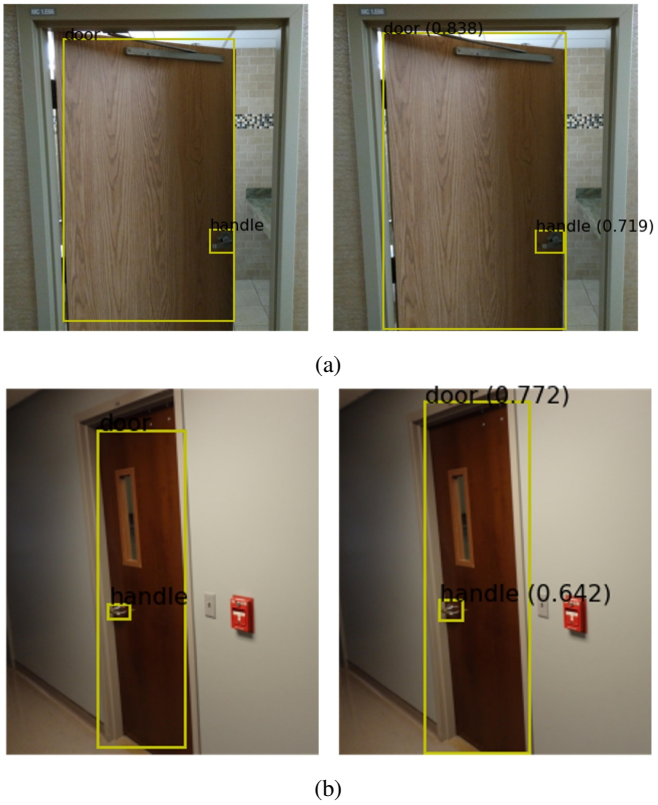


Fig. 4: Results of our method with maximum performance (mAP = 1). Left: the groundtruth. Right: our detections.

handle are correctly found. Figure 5 shows examples where our method performs worst (0.4 < mAP < 0.9). In Figure 5a, a handle is detected on the person, but with a lower confidence. The errors are that two handles are detected on the same location and that the detection of one door is too large. In Figure 5b, the large door on the left is erroneously detected as two separate doors. A similar error is observed in Figure 5c for a zoomed image. In Figure 5d the bar on the side is mistaken for a door handle.

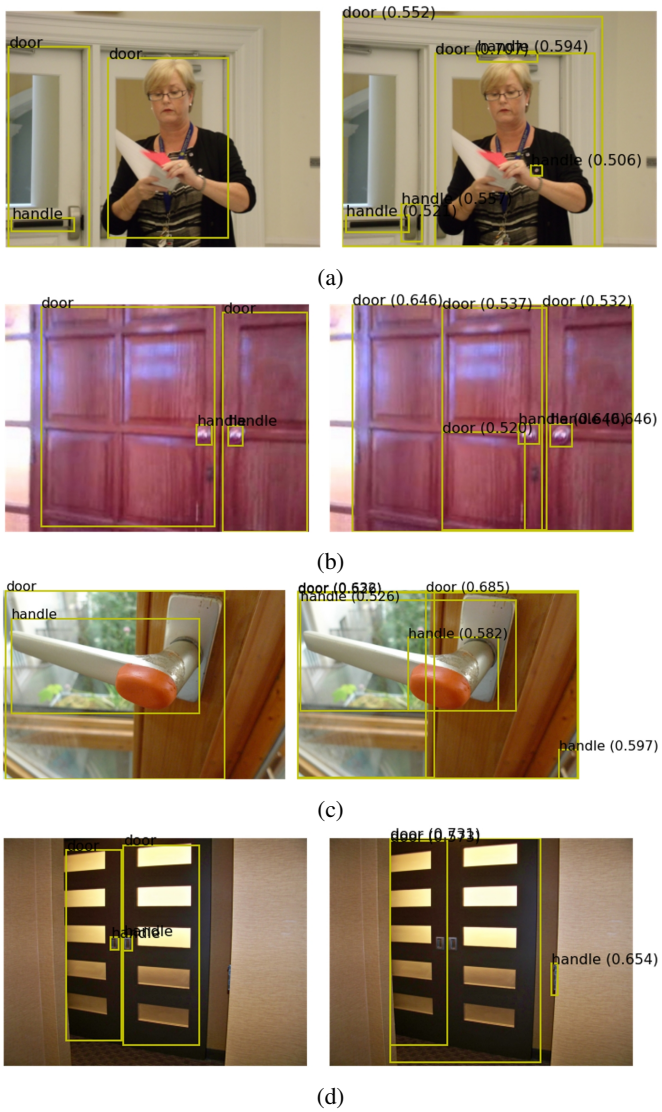


Fig. 5: Results of our method with lowest performance (0.4 < mAP < 0.9).

V. Conclusions

We have proposed a method that adds contextual knowledge to a pretrained large language-vision object detection model, to improve precision and recall. We demonstrate that GLIP [5], a popular and high-performance zero-shot pretrained model, can be improved when adding contextual knowledge at the input prompts (pre-processing) and filtering of candidate objects by known spatial relations (post-processing). The former is implemented as a set of logical production rules and the latter as a neuro-symbolic program. We have demonstrated that the pre-processing improves the recall of objects and the post-processing improves the precision of found objects. Adding this on top of large pretrained models is beneficial for object localization in open-world robotics.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[2] OpenAI, "https://openai.com/blog/clip/," 2021.

[3] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[4] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4904–4916.

[5] H. Zhang, P. Zhang, X. Hu, Y.-C. Chen, L. H. Li, X. Dai, L. Wang, L. Yuan, J.-N. Hwang, and J. Gao, "Glipv2: Unifying localization and vision-language understanding," in *Advances in Neural Information Processing Systems*, 2022.

[6] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, and L. D. De Raedt, "Neural probabilistic logic programming," *Advances in Neural Information Processing Systems*, vol. 31, pp. 3753–3763, 2021.

[7] R. Manhaeve, G. Marra, and L. De Raedt, "Approximate inference for neural probabilistic logic programming," in *Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning*. IJCAI Organization, 2021, pp. 475–486.

[8] J. Huang, Z. Li, B. Chen, K. Samel, M. Naik, L. Song, and X. Si, "Scallop: From probabilistic deductive databases to scalable differentiable reasoning," vol. 34, 2021, pp. 25 134–25 145.

[9] A. d. Garcez and L. C. Lamb, "Neurosymbolic ai: The 3rd wave," *Artificial Intelligence Review*, pp. 1–20, 2023.

[10] H. Chefer, S. Gur, and L. Wolf, "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 397–406.

[11] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *International Conference on Learning Representations*, 2022.

[12] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2019, pp. 3195–3204.

[13] L. De Raedt, S. Dumančić, R. Manhaeve, and G. Marra, "From statistical relational to neuro-symbolic artificial intelligence," *International Joint Conference on Artificial Intelligence*, pp. 4943–4950, 2020.

[14] A. d. Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, and S. N. Tran, "Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning," *Journal of Applied Logics*, vol. 6, pp. 611–632, 2019.

[15] R. Manhaeve, S. Dumančić, A. Kimmig, T. Demeester, and L. De Raedt, "Neural probabilistic logic programming in deepproblog," *Artificial Intelligence*, vol. 298, p. 103504, 2021.

[16] M. Arduengo, C. Torras, and L. Sentis, "Robust and adaptive door operation with a mobile robot," *Intelligent Service Robotics*, 2021.

Appendix

Figure 6 shows examples and their object annotations (groundtruth) from the Doors dataset [16]. There is an interesting variety of scenes with a range of doors and types of handles.

Figure 7 shows examples from the Doors dataset [16] and their overlap with the groundtruth. The examples have an overlap measure of Intersection-over-Union (IoU) ≥ 0.35 but < 0.5. It appears that the groundtruths are not always well positioned. This motivates our choice to evaluate both on the standard setting of IoU ≥ 0.5 as well as IoU ≥ 0.35.

Figure 8 shows examples of GLIP detections without our refinements.

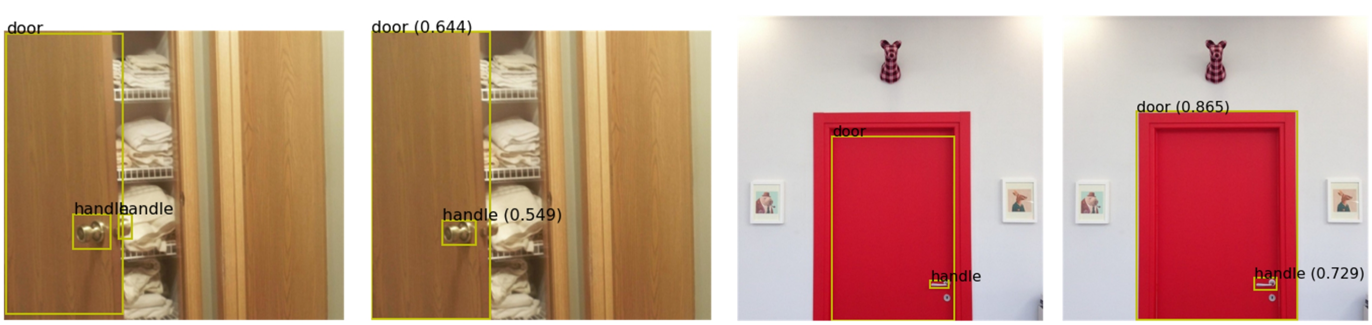
Figure 9 shows more examples where our method achieves a perfect result (mAP = 1). Interestingly, the cases are very different, showing broad applicability. All cases have some clutter, yet the door and handle are correctly found.

Fig. 6: Illustrations of the doors dataset [16].

(a) (b)

Fig. 7: Overlap with the groundtruth: cases that have $0.35 \leq \text{IoU} < 0.5$.

Fig. 8: Examples of original GLIP detections with a confidence > 0.4.



(a) (b)

Fig. 9: Results of our method with maximum performance (mAP = 1). Left: the groundtruth. Right: our detections.