Anomaly Detection in an Open World by a Neuro-symbolic Program on Zero-shot Symbols

Gertjan J. Burghouts[1], Fieke Hillerström, Erwin Walraven, Michael van Bekkum, Frank Ruis and Joris Sijs

***Abstract*— Mobile inspection robots are typically tasked to find anomalies and report them, such that proper action can be taken. They operate in an open world, in which they encounter previously unseen situations in changing environments. We take the robot's goal, its context, prior knowledge and uncertainties into account to find anomalies that are relevant for the operation. Prior knowledge is expressed by logic formulas. E.g., a tool should not be left on the floor. These symbolic formulas describe anomalous objects in terms of predicates and variables (symbols) that can represent various concepts in the real world, their attributes and their relations. This knowledge can easily be adapted during the robot's operation to a new anomaly that is considered relevant. Reasoning is performed in a probabilistic, multi-hypothesis framework. A neuro-symbolic program evaluates the symbolic formulas against probabilistic, imperfect observations of the symbols. New anomalies require new symbols, which are measured in images by zero-shot language-vision models and their extensions for objects and segments. Starting from the symbolic formulas and predicates that describe the anomaly, our method infers what objects are involved that need to be detected, extracts the probabilistic information from the language-vision models and reasons about that via a neuro-symbolic program in order to find the anomaly of interest. Our contribution is the integration of the neuro-symbolic program and language-vision models. We show the effectiveness of our method to find anomalous situations in a robotic inspection setting.**

I. INTRODUCTION

A foreseen task of mobile robots is to inspect large industrial sites. One such inspection task is to search for abnormal situations that may pose a hazard to the personnel or their environment. Once such an anomaly is found, it can be reported to the operator, such that proper action can be taken. On such sites, there may be many activities. Very likely the robot does not know all possible situations beforehand. Therefore, the robot needs to be able to deal with a (partially) open world, which includes previously unseen actors, objects and situations. We focus on finding anomalies in images, where we have prior knowledge about the anomaly, and need to deal with uncertain observations and previously unseen objects and segments in the scene.

For anomaly detection, a common approach is to use a statistical model of the sensory data [14], [15], [13]. However, the robot's goal, its context and the user's prior knowledge are not taken into account. As a consequence, the detected anomalies are not necessarily relevant. Another drawback of statistical models is that they do not generalize well to new observations in the open world. They cannot be adapted quickly, because it requires a significant amount of training samples to adjust the statistical model.

[1]TNO, The Netherlands. gertjan.burghouts@tno.nl

We take a different approach by leveraging prior knowledge about relevant anomalies. This knowledge can be adapted quickly during operation and via generic definitions it can generalize better to new situations. An example of an anomaly is a tool that is left abandoned on the floor. A tool can be one of many types, such as a hammer, skrewdriver, wrench, and many more. Likewise, floors can be composed of different materials with various appearances. Our goal is to find anomalies in images, based on a high-level definition of the object categories involved and their spatial configuration, without specifying the precise object classes or learning dedicated models for each of them. The rationale is that such a knowledge-based anomaly detection has a broader applicability, because it can generalize better across similar anomalies and is adaptable to new anomalies by formulating a new definition. Furthermore, any detected anomaly has already been interpreted, which allows to connect actions to it.

We search for anomalies by reasoning about spatial relationships between the objects categories involved. We capture such knowledge by means of logic formulas. Since there are many objects in a scene and possibly multiple occurrences of the relevant objects, with a multitude of relations between them, a multi-hypothesis validation is required. Perception is imperfect and probabilistic in nature, therefore this validation needs to be done within a probabilistic framework. A natural choice is to leverage a neuro-symbolic program [7], [6], [3] to test hypotheses about the specified anomaly, since such a program is able to validate symbolic predicates against probabilistic observations of the symbols from the predicates.

The key challenge is to get observations for the symbols in a manner that generalizes well to the open world setting and new symbols. In previous works, the symbols and their probabilities were produced by specialized, fixed neural networks [3]. In our example of the abandoned tool on the floor, we may have a network to produce symbols for hammers and skrewdrivers, but not for wrenches. Hence, a wrench that was left on the floor will erroneously not be detected as an anomaly. In an open world, we need symbols that generalize to object categories and can recognize unseen object classes. Therefore, we turn to language-vision models [1], [11], [12] that have so-called zero-shot capabilities to recognize novel classes based on a textual description [2], [5], [4]. We use the symbol from the predicate to formulate a textual query (a prompt) to the model, in a fully automated manner. Because the model is based on language, its advantage is that it is likely to have a representation of

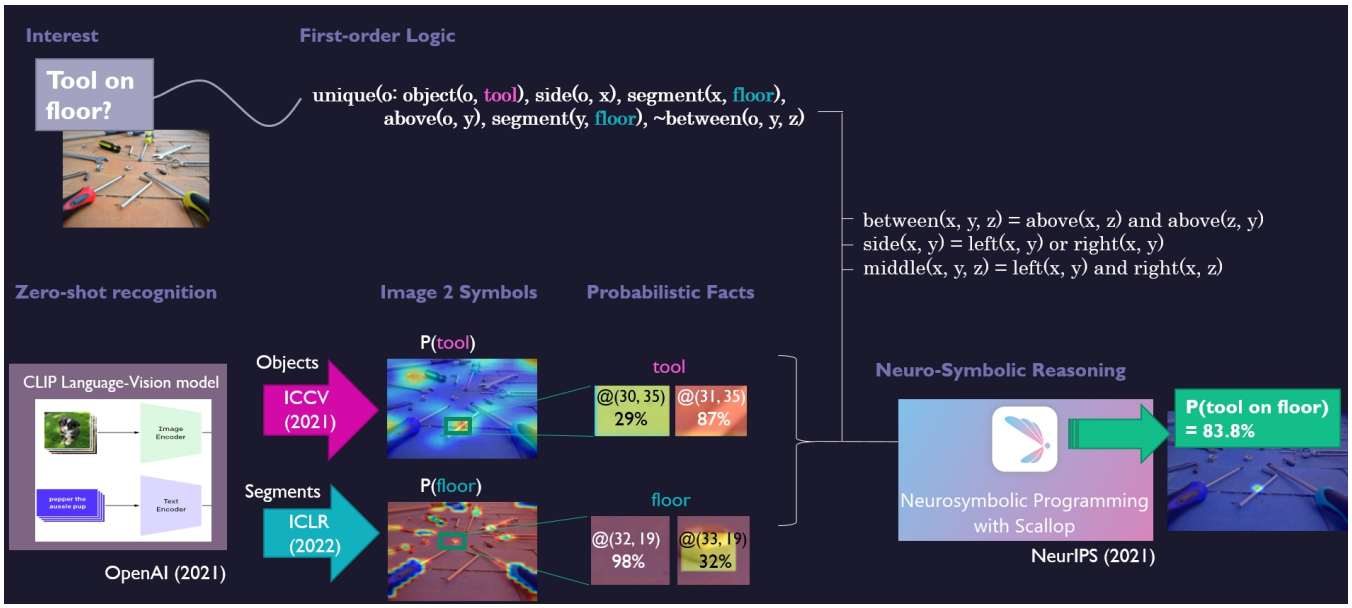


Fig. 1: Knowledge-based anomaly detection via first-order logic, validated by a neuro-symbolic program that operates on language-vision symbols.

the category of an object, since it is similar in terms of the textual description. This enables us to specify anomalies at the category level (e.g., tool). We explore extensions of language-vision models in order to localize both objects and segments in images, in order to extend the range of symbols. With this extension we can localize concepts such as 'floor'. Using these zero-shot capabilities of language-vision models and the segmentation extension, the neuro-symbolic program has broader generalization towards various configurations of the anomaly.

With many symbols in the image, there will be many possible hypotheses. To deal with this, we consider a recent neuro-symbolic programming framework that limits the validation to the top-k hypotheses, while guarantying a minimal performance degradation [3]. The summary of our methodology is as follows. We start with the symbolic predicates that define the anomaly. From the predicates, our method infers the involved symbols. The symbols and their probabilities are measured from images by prompting the language-vision models. The probabilistic symbols and predicates are validated by the neuro-symbolic program. Our method is outlined in Figure 1. The key contribution is the integration of neuro-symbolic programming and language-vision models. We show the effectiveness on real-world images of anomalous situations in a robotic inspection setting.

II. RELATED WORK

To find anomalies based on prior knowledge, we integrate neuro-symbolic programming and language-vision models. Here we discuss related works.

Neuro-symbolic programming. An important capability is to reason about an image with external knowledge [9]. This is also the case for anomaly detection based on prior knowledge. Connecting knowledge representation and reasoning mechanisms with deep learning models [7] shows great promise for learning from the environment and at the same time reasoning about what has been learned [6]. Previous reasoning methods were limited in terms of scalability, in case of many possible hypotheses, e.g., industrial inspection with many possible objects and relations. Those methods were ill-suited for real-world applications. A recent neuro-symbolic programming framework is based on first-order logic. It introduces a tunable parameter k to specify the level of reasoning granularity, by restraining the validation of hypotheses by the top-k proofs. This asymptotically reduces the computational cost while providing relative accuracy guarantees [3]. This is beneficial for our purpose, as we expect many possible hypotheses in complex environments with many objects and imperfect observations.

Language-vision modeling. Language-vision models learn directly from large datasets of texts about images which offers a broad source of supervision [1], [2], [11], [12]. They have shown great promise to generalize beyond crisp classes and towards semantically related classes. This so-called zero-shot capability is beneficial for recognizing the object categories that are involved in the anomalies. Recently, these models were extended with capabilities to localize objects in images via co-attentions [5] and to segment parts of the scene based on textual descriptions [4]. We adopt both methods to relate image parts to objects and segments that are of interest for the anomaly at hand. To the best of our knowledge, we are the first to combine neuro-symbolic programming with language-vision models for knowledge-based anomaly detection.

III. METHOD

Our aim is to find an anomaly in an image, based on prior knowledge about the involved objects and their relations.

An overview of our method is shown in Figure 1. At the top, it shows how an anomaly such as 'tool on floor' is translated into symbolic predicates such as $object(o,\, tool)$, $segment(x,\, floor)$ and $above(o, x)$. At the bottom left, the figure shows how the symbols from the predicates, such as 'tool' and 'floor', are measured from images by language-vision models. These measurements are transformed into probabilistic facts, which are fed to the neuro-symbolic program in order to be validated against the logic (bottom right). Each component is detailed in the following paragraphs.

First-order logic. The anomaly is defined by logic formulas and predicates. The symbols in the predicates are about the objects and segments in an image. An example is the anomaly of a tool that is left on the floor:

$$\begin{aligned} \exists o : & object(o, tool) \wedge side(o, x) \wedge \\ & segment(x, floor) \wedge above(o, y) \wedge \\ & segment(y, floor) \wedge \neg between(o, y, z) \end{aligned} \tag{1}$$

This defines the anomaly as a tool that is above and on the side of the floor. This definition is necessary, because the robot's perspective is oblique downward, i.e., the floor will be visible at the bottom of the tool and on the side of the tool. To express that the tool should be on the floor, without anything in between, we define that there should be nothing in between the tool and the floor. Otherwise a tool on a cabinet standing on the floor also fulfils the definition. The definitions of the helper predicates are:

$$\begin{aligned} between(x, y, z) &= above(x, z) \wedge above(z, y) \\ side(x, y) &= left(x, y) \vee right(x, y) \\ middle(x, y, z) &= left(x, y) \wedge right(x, z) \end{aligned} \tag{2}$$

to express that: some z is in between x and y; some x is in the middle of y and z.

Image to symbols. To relate the logic to the image, we search the image for the symbols from the predicates. These involve objects, segments and spatial relations between them. The objects and segments are respectively recognized by the attention-based [5] and segmentation-based [4] methods (see Section II). The output of both methods are probabilistic. We transform them into a heatmap per symbol, in order to acquire probabilities in a spatial grid. The grid is our reference frame to extract spatial relations and to later reason about them. In this way, we acquire $P(object = tool \,|\, image)$ and $P(segment = floor \,|\, image)$.

Inference. From the symbolic heatmaps we derive the probabilistic facts which are used by the neuro-symbolic program for inference. The distance between the robot and the objects may differ from time to time. The heatmaps are finegrained. We add downsampled versions of them to enable both finegrained and coarser inference. To achieve scale invariant reasoning, we select the scale σ (i.e., the image downsampling factor) that maximizes the likelihood for the anomaly A in the given image I given the logic L:

$$P(A \,|\, I,\, L) = \underset{\sigma \in \{1,\, 2,\, 4,\, \ldots\}}{\arg\max} \; P(A \,|\, I,\, L,\, \sigma) \tag{3}$$

IV. ANALYSIS

To analyze the performance of our method, we collected 31 test images in very diverse settings. The goal is to find the anomalous case of an abandoned tool on the floor. To validate how well the method generalizes to various tools, we include images with hammers, skrewdrivers, wrenches, etc. For the same reason, we include various floors, with different materials, textures and colors. Moreover, the viewpoint and zoom are varied significantly. There are 9 images of tools on floors. These are the positives, where we expect anomalies to be detected by our method. To verify true negatives, we include 8 images where there is a both a tool and a floor, but the tool is not on the floor (but on a cabinet, wall, etc.). There are 5 images with only a floor (no tool) and 4 images with only a tool (no floor). To verify true negatives, there are also 5 images where there is no tool and no floor.

True positives. There are 9 positives, from which we detect 7 cases. Figure 2 shows 3 out of those 7 cases, one in each column. For each case, the top row shows the results of the neuro-symbolic program. Red indicates a high probability, whereas blue indicates a low probability. The middle and bottom rows show the probabilistic symbols that are used by the program (same color coding). Since the symbols are predicted well (often the tools and floors have a high probability at the respective symbols), the reasoning is able to pinpoint a place in the image where the spatial configuration is fulfilled (red peak).

True negatives. Figure 3a shows true negatives. Although there are both a floor and tools, the reasoner correctly finds that the spatial configuration is not a tool that is on the floor.

False positives. Errors are shown in Figure 3b. The neuro-symbolic program incorrectly reasons that these cases are a tool left on the floor (false positives). This is due to errors in the symbols. There is a wrong association of the symbol tool in both images. On the left, the Gazelle logo is associated with a tool, because Gazelle is a manufacturer of bicycles and many images are about tools. The language-vision model has a bias to associate Gazelle with tools. On the right, the duct tape is considered to be a tool. From a semantic point of view this makes sense. These symbol errors propagate into the reasoner's outputs. Refining the prompts that we pose to the language-vision models, may overcome such errors in the symbols.

False negatives. Figure 3c shows missed cases (false negatives). Again, the source of the errors is in the symbols. On the left, the tool (a grinder) is not recognized as such. This is a flaw in the language-vision model, probably because this tool does not appear often in everyday images and language. On the right, the floor is not recognized as such, because a context is lacking: it could also be a wooden plate. Without the proper evidence for each involved symbol, the reasoner cannot assess these configurations correctly.

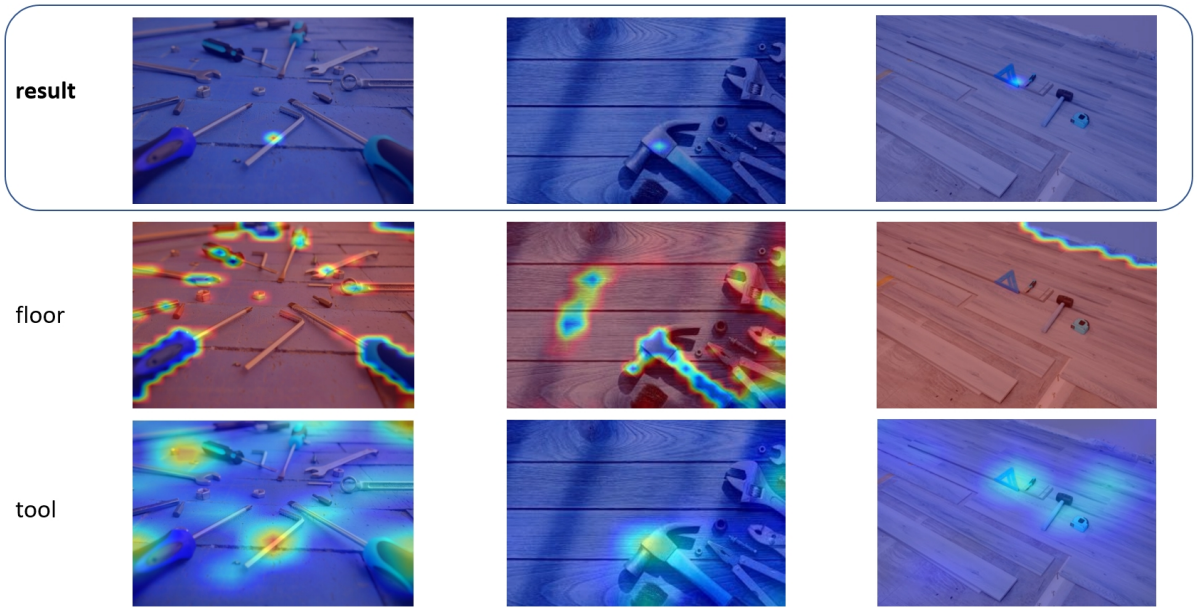
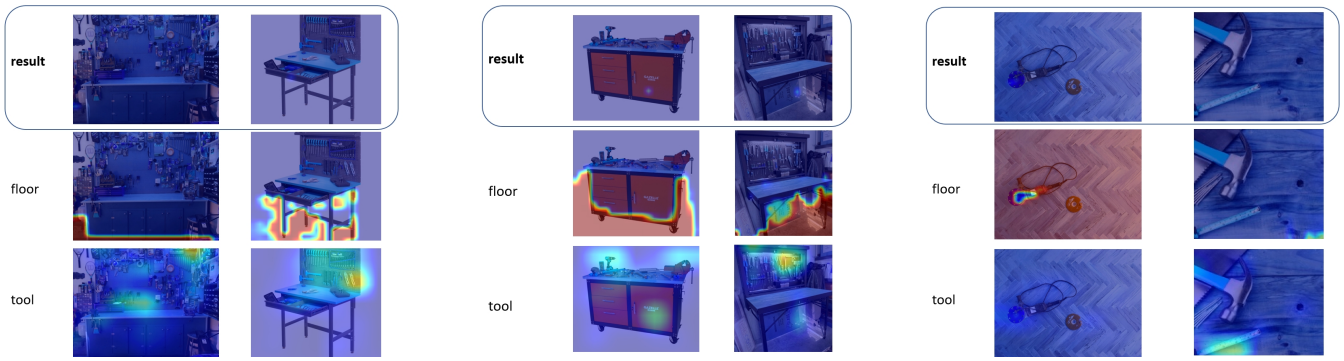


Fig. 2: True positives.



(a) True negatives. (b) False positives. (c) False negatives.

Fig. 3: True negatives and errors.

V. CONCLUSIONS

We proposed a method that endows mobile inspection robots with the capability of finding anomalies based on a high-level definition of the involved object categories and their spatial configuration. We expressed the anomalies by first-order logic, about which we reason using a neuro-symbolic program. In the logical definitions, there is no need to specify the precise object classes, so that we can generalize to similar anomalies. To generate the probabilistic observations for the symbols, we leverage zero-shot language-vision models. This extends the scope of the anomalies to previously unseen objects, which is crucial in an open world. Our approach avoids the necessity of learning dedicated models for each of the involved objects, which makes our method flexible and quickly deployable.

ACKNOWLEDGEMENT

This work is sponsored by the TNO Appl.AI program and its SNOW project.

REFERENCES

[1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision.

[2] https://openai.com/blog/clip/

[3] Huang, J., Li, Z., Chen, B., Samel, K., Naik, M., Song, L., Si, X. (2021). Scallop: From probabilistic deductive databases to scalable differentiable reasoning. Advances in Neural Information Processing Systems, 34, 25134-25145.

[4] Li, B., Weinberger, K. Q., Belongie, S., Koltun, V., Ranftl, R. (2021, September). Language-driven Semantic Segmentation. In International Conference on Learning Representations.

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

[6] Garcez, A., Gori, M., Lamb, L. C., Serafini, L., Spranger, M., Tran, S. N. (2019). Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. Journal of Applied Logics, 6(4), 611-632.

[7] De Raedt, L., Dumancic, S., Manhaeve, R., Marra, G. (2020). From Statistical Relational to Neuro-Symbolic Artificial Intelligence. In

Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020 (pp. 4943-4950).

[8] Dai, W. Z., Xu, Q., Yu, Y., Zhou, Z. H. (2019). Bridging machine learning and logical reasoning by abductive learning. Advances in Neural Information Processing Systems, 32.

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

[10] Xu, J., Zhang, Z., Friedman, T., Liang, Y., Broeck, G. (2018, July). A semantic loss function for deep learning with symbolic knowledge. In International conference on machine learning (pp. 5502-5511). PMLR.

[11] Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C. and Liu, C., Florence: A New Foundation Model for Computer Vision.

[12] Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning (pp. 4904-4916). PMLR.

[13] Paschalidis, I. C., Chen, Y. (2010). Statistical anomaly detection with sensor networks. ACM Transactions on Sensor Networks (TOSN), 7(2), 1-23.

[14] Chandola, V., Banerjee, A., Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58.

[15] Pang, G., Shen, C., Cao, L., Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. ACM Computing Surveys (CSUR), 54(2), 1-38.