ZERO – Detect objects without training examples by knowing their parts.

Gertjan J. Burghouts[a] and Fieke Hillerström[a]

[a] *TNO Intelligent Imaging, Oude Waalsdorperweg 63, 2597 AK, The Hague, The Netherlands*

Abstract

Current object recognition techniques are based on deep learning and require substantial training samples in order to achieve a good performance. Nonetheless, there are many applications in which no (or only a few) training images of the targets are available, whilst they are well-known by domain experts. Zero-shot learning is used in use cases with no training examples. However, current zero-shot learning techniques mostly tackle cases based on simple attributes and offer no solutions for rare, compositional objects such as a new product, or new home-made weapons. In this paper we propose ZERO: a zero-shot learning method which learns to recognize objects by their parts. Knowledge about the object composition is combined with state-of-the-art few-shot detection models, which detects the parts. ZERO is tested on the example use case of bicycle recognition, for which it outperforms few-shot object detection techniques. The object recognition is extended to detection by localizing it, by taking into account knowledge about the object's composition, of which the results are studied qualitatively.

Keywords

Zero-shot learning, Knowledge, Object recognition, Object localization.

1. Introduction

In many computer vision applications, there are no images of the objects of interest. For instance, a new product that has not yet been assembled or photographed, and new variants of home-made weapons. A lack of training images makes it harder to learn to recognize the objects. Standard deep learning offers no solution to recognize such objects, as these models require many labeled images [1]. In zero-shot learning (ZSL) [2], the goal is to learn a new object by leveraging knowledge about that object.

The most common approach is to capture knowledge about the objects by representing their attributes [3]. A new object is modelled as a new combination of known attributes. The state-of-the-art is to learn the relation between attributes (e.g., furry) and appearance [4]. A new object can be predicted if its attributes correspond to the observed appearance. To learn the implicit relations between attributes and appearance, many objects in many different combinations of attributes are needed (e.g., many animals with attributes [5]).

The attribute-based approach does not work for new objects that consist of attributes that are not common in many other objects. For instance, the home-made RC car is composed of wheels, a camera, a battery, some wires, and a small computer (Figure 1, left). Likewise, the home-made explosive (Figure 1, right, i.e., an improvised explosive device, IED) is composed of a mobile phone, tape, wires, bottle. Not many other objects are composed of these specific parts. There are not many other objects that share the IED's parts. Extracting the attributes of these parts and using them for learning, is complex, since the attributes will only be representative for parts of the object. Hence, the implicit relation

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021) - Stanford University, Palo Alto, California, USA, March 22-24, 2021.

EMAIL: gertjan.burghouts@tno.nl (A. 1); fieke.hillerstrom@tno.nl (A. 2)
ORCID: 0000-0001-6265-7276 (A. 1); 0000-0003-1301-3073 (A. 2)

between attributes and appearance cannot be learned, as there is not sufficient training data. For those objects, the attribute-based approach does not fit.

Figure 1: Examples of new, compositional objects for which attribute-based models do not work.

As a novel approach, we leverage the compositionality of new objects, by modelling them explicitly as a combination of reusable parts. For composed new objects, the parts are typically everyday objects. For instance, for the IED, the parts are a phone, tape, wires and bottle, which are all very common. For everyday objects, it is easy to find images and to annotate the region that contains the relevant part. A part model can be learned from these images and annotations. A standard object detector is trained to localize the parts. The new object is modelled by combining the detected parts.

The proposed method is named ZERO. ZERO consists of four steps. Firstly, expert knowledge about the object is captured in terms of its parts. This involves the object parts and the relations between them, i.e., spatial arrangement and relative sizes. Secondly, the parts are learned and detected (localized) in images. Thirdly, the object is learned by combining the parts and their appearance (visual features). Fourthly, the object is localized in the image by assessing the spatial arrangement of parts and their respective sizes relative to the object. ZERO is outlined in Figure 2.

Figure 2: Outline of ZERO and its steps to recognize a new (unseen) object.

The advantages of ZERO are:

a. Recognition of new objects when no training samples nor attribute annotations are available.
b. Taking knowledge of a mid-level abstraction into account, instead of low-level attributes.
c. Using compositional knowledge about the location of properties, which is less feasible for fuzzy attributes.
d. Easier specification of the expert's knowledge, as the parts are mid-level and clear, contrary to fuzzy attributes.
e. To provide predictions that are explainable towards the user. The parts and composition can be expressed more easily to a human than a plain confidence.

Since parts contain a higher level of abstraction than attributes, they encode more information, which makes the added knowledge more valuable. Different parts contain different properties and when they are combined in a composition, these properties are encoded at a location. Attribute-based approaches encode a specific attribute for the whole object. The expert specifies the object composition

in terms of parts, which is related to the our common way of reasoning. The interpretation towards attributes is taken out. The new object and its parts are localized with a confidence for the object and per part. This makes it easier for the user to understand why the algorithm predicted that this object is in the current image. We will show examples in the experimental results of such predictions.

Section 2 discusses related work. Section 3 describes the proposed ZERO method. Section 4 details the experimental results and findings. Section 5 concludes the paper.

2. Related work

Zero-shot learning based on attributes, e.g., [4], leverages big datasets with many object classes and many attributes in various combinations, e.g., AWA2 [2], CUB-200 [5] and SUN [6]. AWA2 has 40 object classes for training with 85 attributes [2]. CUB-200 has 200 object classes and 312 attributes [5]. SUN has 717 object classes with 102 attributes [6]. These datasets have more than thousands of training images with many object classes that share attributes, which enables models to learn the relations between attributes and appearance. For many types of new objects, such as the IED (Figure 1), such datasets of shared attributes are not available. In this paper, we are interested in such composed objects.

There are significant differences between this paper and attribute-based ZSL, which are summarized in Table 1. In the attribute-based ZSL, there are many annotations of other objects that share similar attributes, whereas our setup is that there are none. In attribute-based ZSL, the object classes of the abovementioned datasets are closed-world. For instance, the problem is about animals only, which limits the learned models to recognize only new animals and not other objects. In this paper, we aim to recognize a broad set of new objects. The expert knowledge used in attribute-based ZSL only covers the combinations that constitute the object. ZERO uses additional knowledge about the spatial arrangement of parts for localizing the object in the image (localization). In attribute-based ZSL, the importance of each attribute is learnable, because there are so many combinations of attributes and objects and appearance to learn from. Contrary, in this paper, there are no other annotated objects, which requires a different approach. Finally, attribute-based ZSL involves knowledge about attribute composition, where in this paper we leverage more knowledge about the object, i.e., parts, part composition and spatial arrangement of parts.

Table 1
Differences between attribute-based ZSL and ZERO.

	Attribute-based ZSL	ZERO (this paper)
Annotated other objects	Many	Zero
Closed world	Yes	No, objects can be from any category
Importance of elements learnable	Yes	No, lack of labeled data
Expert knowledge	Only about attribute composition	About part composition and spatial arrangement of parts

Our approach is to model a new (unseen) object explicitly as a combination of (reusable) parts. The parts are typically very common, so there is sufficient data to learn part models. Supervised modelling of an object by its parts has earlier been investigated [7], by combining a holistic object and body parts, with the objective to handle large deformations and occlusions of parts. The key of this approach is to automatically decouple the holistic object or body parts from the model when they are hard to detect. The model learns the relative positions and scales of the object, based on many training instances of the object. In contrast, in this paper, we aim to model objects by their parts, but without training samples of the actual object. To that end, we rely on knowledge about the object, and specifically the parts and the relations between them. We leverage this knowledge in a learning scheme. In the absence of training images of the new object, object-parts training samples are synthesized to learn the model for the new object.

Instead of only recognizing images, we also aim for object localization. The combination of object recognition and localization is known as zero-shot detection (ZSD), which aims to detect and localize

instances of unseen object classes. There are roughly three type of methodologies for ZSD. The first type of methods use a region proposal network on top of a feature extraction backbone [8, 9, 10]. In [11] these proposals are improved by explicitly taking the background into account while learning. The features of the proposed regions are used to determine the object classes, using neural layers on top of the features. The region proposals are used to localize the objects. Often a bounding box regression model is trained to fine-tune the locations of the region proposals. The region proposals are trained on common data, or defined as default anchor boxes. These methods are beneficial when no visual samples of the input is available. However, in our case we do have visual samples of the subparts and can take knowledge and features of these parts into account. The second type of methods [12] synthesize training images for the unseen classes, based on semantic information, for example using a GAN. These synthesized images are used to train a state-of-the-art object detector network. In a way this is comparable to our method, since for the detection synthesize part combinations, using knowledge. However, we directly use them for object localization, instead of introducing the additional effort of image generation and detection network training. The third type of methods use attention based models to determine the location of the object in the image [13]. These attention maps learn to differentiate objects from the background, using learned attention weights. Comparable to the region proposal-methods, this is beneficial when no part detections are available. We use the benefits of being able to recognize common-known parts of the objects. In summary, none of methods take explicit knowledge about object parts and configuration into account for object localization.

3. ZERO

The proposed method, ZERO, consists of four steps, starting with knowledge about the new object, up to localizing the new object in a test image. These four steps are detailed in the next subsections.

3.1. Knowledge

Knowledge about the new (unseen) object is captured in terms of its parts. This involves the object parts and the relations between them, i.e., spatial arrangement and relative sizes. An example is a bicycle, which is defined by its parts *wheels*, *saddle*, *chainwheel* and *handlebar*. The arrangement is defined by parts that are not allowed to overlap; only the wheel and the chainwheel are allowed to overlap. And the expected relative sizes of parts are given by the knowledge and are defined by a minimum ratio and maximum ratio, referred to a reference part. This knowledge is summarized in Table 2.

Table 2
Knowledge about the object at hand (bicycle) and its parts.

Object parts	Disallowed overlap of parts	Minimal area ratio	Maximum area ratio
Wheel	Wheel, saddle, handlebar	Reference part	Reference part
Wheel	Wheel, saddle, handlebar	0.5	2
Saddle	Wheel, handlebar, chainwheel	1.5	7
Chainwheel	Saddle, handlebar	1.5	7
Handlebar	Wheel, saddle, chainwheel	1	4

3.2. Object parts

Given the object definition, its parts are learned. Generally, it is possible to obtain annotations for object parts, as most parts are everyday instances and it is easy to find or collect images of them. The annotations are bounding boxes, each with a label of the part. A part model can be learned by leveraging modern object detectors that can be retrained by fine-tuning from the broad MS-COCO dataset to the annotations at hand. We selected Retinanet [14] for this purpose, as it proved to be robust for many

types of images and small objects. The latter is important, as parts are generally smaller. For annotations of parts, we use the dataset in [7]. After learning, a model is acquired that is able to detect (localize) the object's parts in test images.

3.3. Recognition

The object is learned by combining the parts and their appearance. We aim to learn which specific part-based features are discriminative of the full object. The parts and their features are combined in a graph representation such that all features are available to the learning of the object model. To that end, an graph is composed, where each node resembles one part. Each specific node represents one part and has a fixed position in the graph representation. The node contains the features of that part. For the features of a part, we extract the specific region of the image and run it through a standard convolutional neural network, i.e., a Resnet-50, of which the embedding before the final layer is used as a feature vector [15]. On top of the graph, a classifier is learned. We will experiment with various classifiers. The goal is that the classifier learns which features of which parts are most discriminative.

Our contribution is in how the graph is learned, i.e., classifying the combined nodes' features to assess whether the current image contains the new object or not. The challenge is how to train the graph model, with no training images of the object at hand. This is done by synthesizing training samples. A training sample for the object is obtained by leveraging the part definition. For each part, a randomly selected instance of that part is plugged into it. In this way, a huge amount of synthesized training samples can be obtained (in the experiments we set this to 10K), and many variations of part combinations are presented to the model during the learning. The rationale is that this should lead to good generalization.

3.4. Localization

The object is localized in the image by assessing the spatial arrangement of parts and their respective sizes relative to the object. Our localization method tries to answer the question '*Given that the image would contain the object of interest, which combination of parts represents that object the best?*' and assumes that when the convex hull of these selected parts is taken, the location of the object is found. The selection of object-parts is based on predefined knowledge; the object composition, variations of the overlap of parts and variations of the ratios of part areas (see Table 2).

The localization starts with a preprocessing step in which the number of parts is reduced. From all the detected parts in the image, per part-class the 20% with the highest detection confidence is selected (see Figure 3, first step). The selection is done per part-class in order to remain sufficient part-options for every class. The value of 20% is chosen in order to remain sufficient variety of bounding boxes for the object construction, while increasing speed and reducing noise as much as possible. This value is validated on a small set of bicycle images.

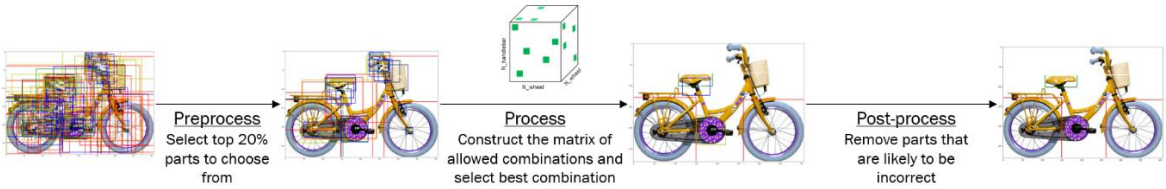

Figure 3: The object localization method starts start with preprocessing to obtain a subset of parts. This subset is checked for allowed combinations, using matrices. Post-processing is applied to remove possible incorrect parts.

After this preprocessing step, all the possible combination of parts are checked whether they are allowed, based on the knowledge. This is done by constructing *N*-dimensional matrices of allowed combinations, where *N* stands for the number of parts that form the object, as defined in the object definition. By implementing the localization method using matrices, the use of multi-variable input (possibility for additional knowledge) and multi-hypothesis output (possibility to return multiple likely answers) is enabled, which makes the methodology very flexible in use. The *N*-dimensional matrices

are combined using the logical-AND operation into one final matrix of allowed part combinations (see Figure 4). From this combined matrix the parts representing the object are selected, based on two scores; the median of the confidence and the median of the confidence when post-processing would be applied.

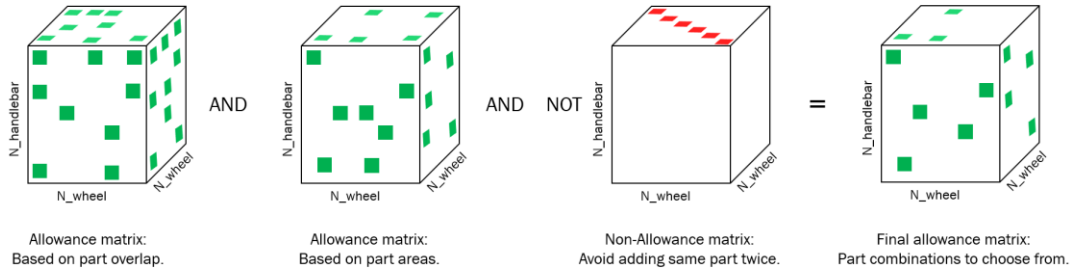

Figure 4: Conceptual visualization of the matrices that capture which combinations are allowed. Possible part combinations are checked for their allowance, based on the knowledge. These sub-matrices are combined into one final matrix to choose part-combinations from. For visualization purposes only 3D matrices are shown. In reality the matrices are *N*-dimensional, with *N* the number of parts that constitute the object.

After the part-combination is selected, post-processing is applied. To take the possibility of missing or occluded parts into account, parts that have a detection confidence lower than 0.6 times the median of all confidence values are probably wrong detections and are removed. When multiple parts of the same class are in the object definition (two wheels for example), parts of this class are removed when their detection confidence is lower than 0.6 times the median of all the confidence values of this class.

4. Experiments

The experiments are performed on the PASCAL VOC dataset [16] for recognition and on downloaded bicycle images for localization. We selected the bicycle as the object of interest, to validate ZERO. To learn the object parts, the annotations from [7] are used.

4.1. Recognition

We compare ZERO's recognition to various baselines. ZERO uses the part models and combines them by a graph. We compare the graph-based approach to simply summing the confidence values of its respective parts. Both are variants with zero examples of the object. We also compare to techniques that require a few examples of the object. To that end, we use the same model [14] as used for the parts, but now for the object. We include these baselines for reference only, because our goal with ZERO is to target the case of zero examples of the object. These baselines cannot deal with that case. For ZERO, we have explored two classifiers on the graph, by concatenating the node features: an SVM (with the radial basis function as kernel) and a deep learning (DL) variant (fully-connected layer with a softmax output).

The ROC curves are shown in Figure 5. Most interestingly, ZERO (see curves for 0 examples) outperforms the baselines that do need several training examples. ZERO also outperforms the naive part combination by summing their confidence values. Note that ZERO's SVM variant performs better than the DL variant, possibly because it is harder to train and optimize the DL variant (more hyper-parameters). For most practical applications, it is essential to have low false positive rates. Therefore, we are especially interested in the left-most part of the ROC curves. In the legend, we report the area under the curve (AUC) at a false positive rate of 5% (0.05). This performance measure is highest for ZERO with the SVM classifier (0.70), outperforming few-shot techniques that required 10 examples (0.64) while ZERO used 0 examples of the object.

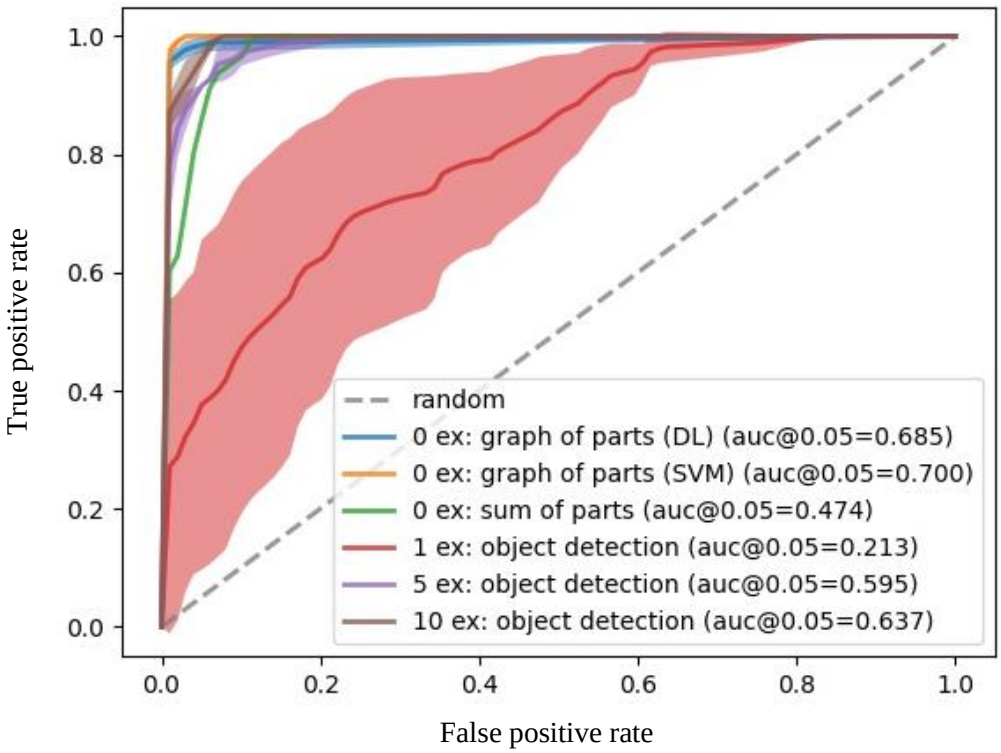

Figure 5: ROC curves of ZERO (zero examples) vs. baselines (few examples).

Four examples of ZERO's predictions are shown in Figure 6. In the upper-left, a positive with a very high confidence (correct). In the lower-left, a negative with a very low confidence (correct). In the upper-right, a negative with a moderate confidence (ideally lower). In the lower-right, a positive with a very low confidence, because of the large occlusion (the bicycle is marginally visible in the back, behind the desk). Obviously, it is hard to recognize a new object, if it is largely occluded.

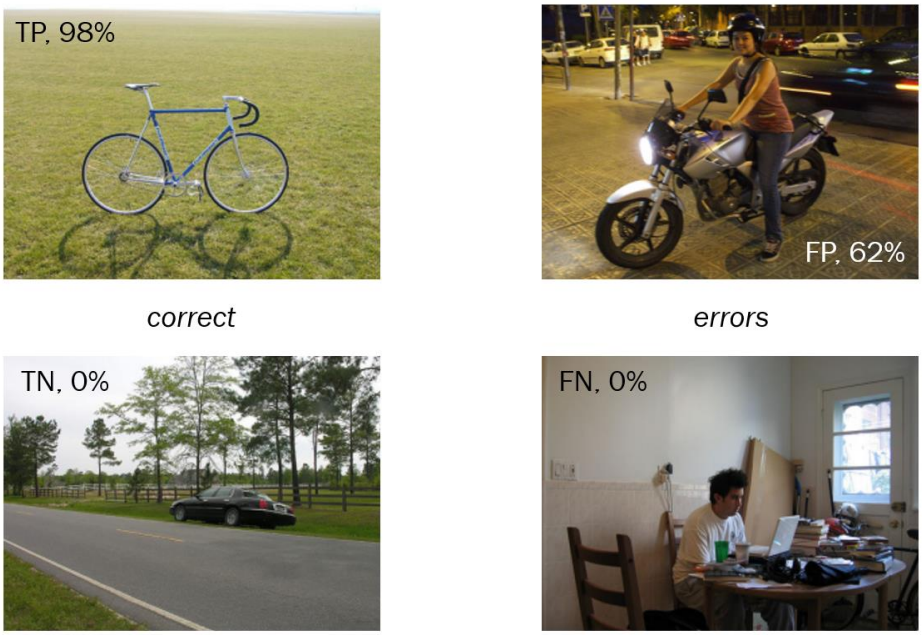

Figure 6: Example predictions by ZERO.

4.2. Generalization

We explore how well ZERO generalizes to new, deviating variants of the object of interest. Our hypothesis is that the training procedure, based on many variations of part combinations, lead to good generalization. We manually selected a set of 25 deviating objects from the internet, as our objects of interest. The background of other objects is the same as in the previous experiment.

Figure 7 shows the ROC curves for ZERO and the baselines, when tested against the deviating objects. ZERO generalizes well to new, deviating variants of the object of interest. Generalization is essential for zero-shot recognition, as not all variants will be known beforehand, and still we want to be able to recognize them well.

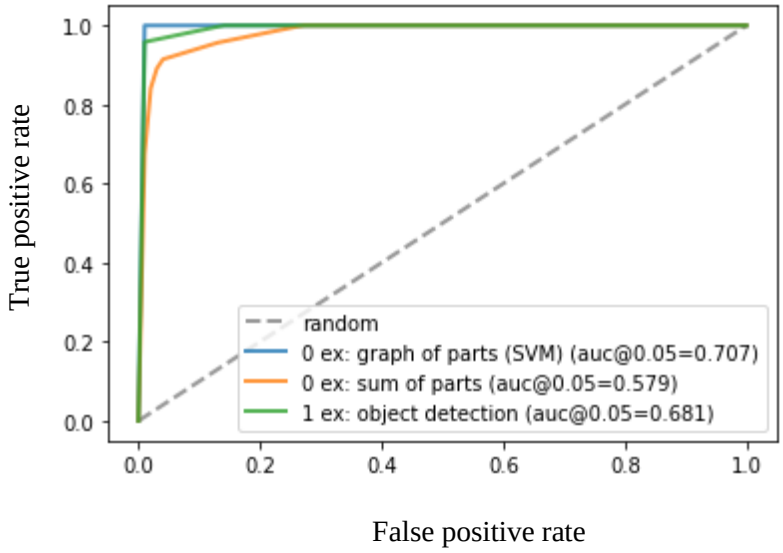

Figure 7: ROC curves on deviating variants of the object of interest.

Two examples of deviating objects are shown in Figure 8. ZERO is confident that these test images contain the new, unseen object of interest.

Figure 8: Example predictions by ZERO on deviations of the object of interest.

We conclude that the hard cases are not the deviating objects (there is good generalization), but when the object is largely occluded (as in Figure 6).

4.3. Localization

We evaluated our localization method quantitively by showing the good (reasonably good), the bad (understandable mistakes) and the ugly (utterly wrong) localization results on a test set of bicycle images (see Figure 9). The test set contains images downloaded from the internet with different

compositions; bicycles seen from the side or from more difficult angles, sometimes partly occluded. The added value of the different knowledge sources is inspected by comparing the localization results when no other knowledge than the object composition is used with the results when knowledge about part overlap and areas is used, which is shown in Figure 10.

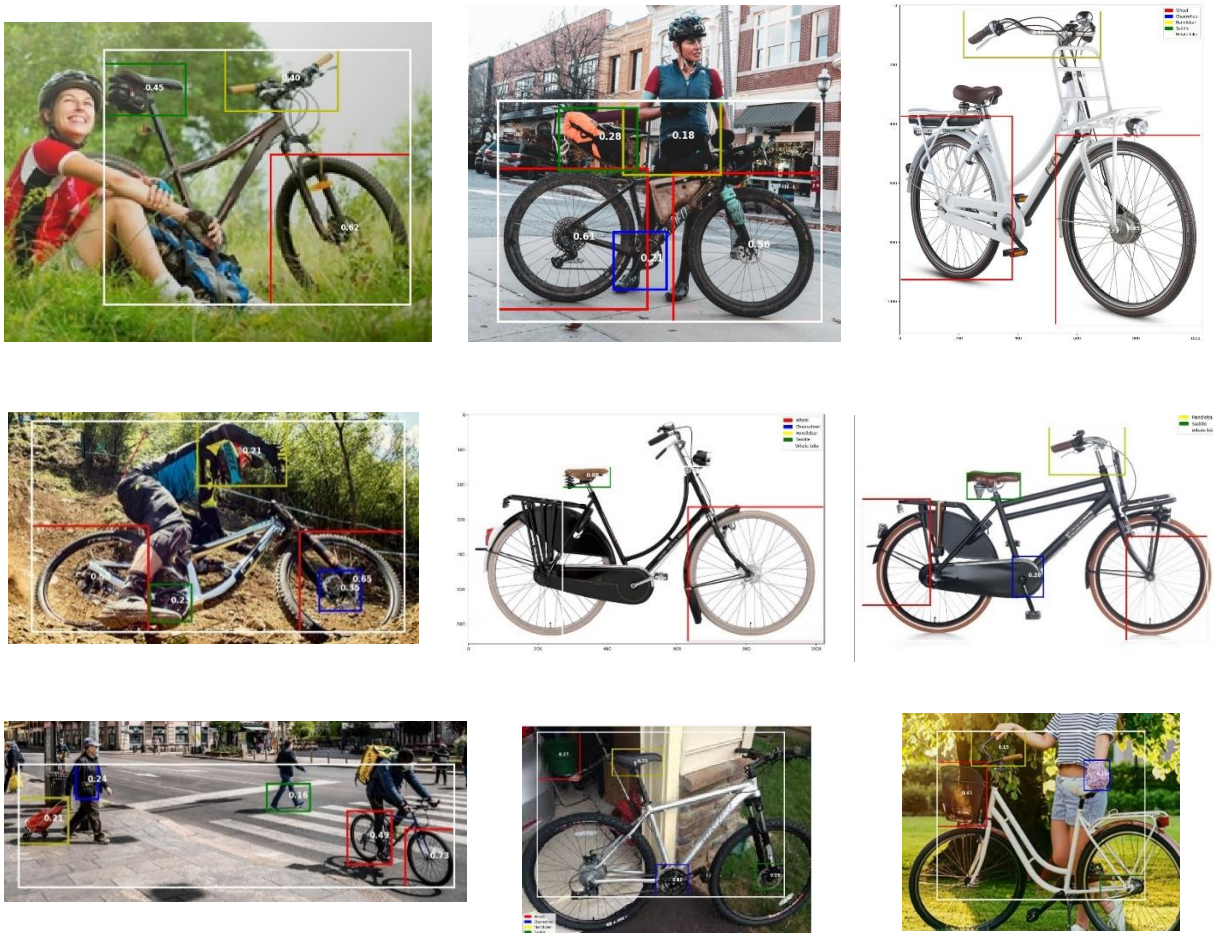

Figure 9: Localization results for different qualitative performance. Upper: The good; reasonably good localization results. Middle: The bad; understandable wrong predictions. Bottom: The ugly; utterly wrong predictions. Red – wheel, blue – chainwheel, yellow – handlebar, green – saddle, white – whole bike, by taking the convex hull of the parts.

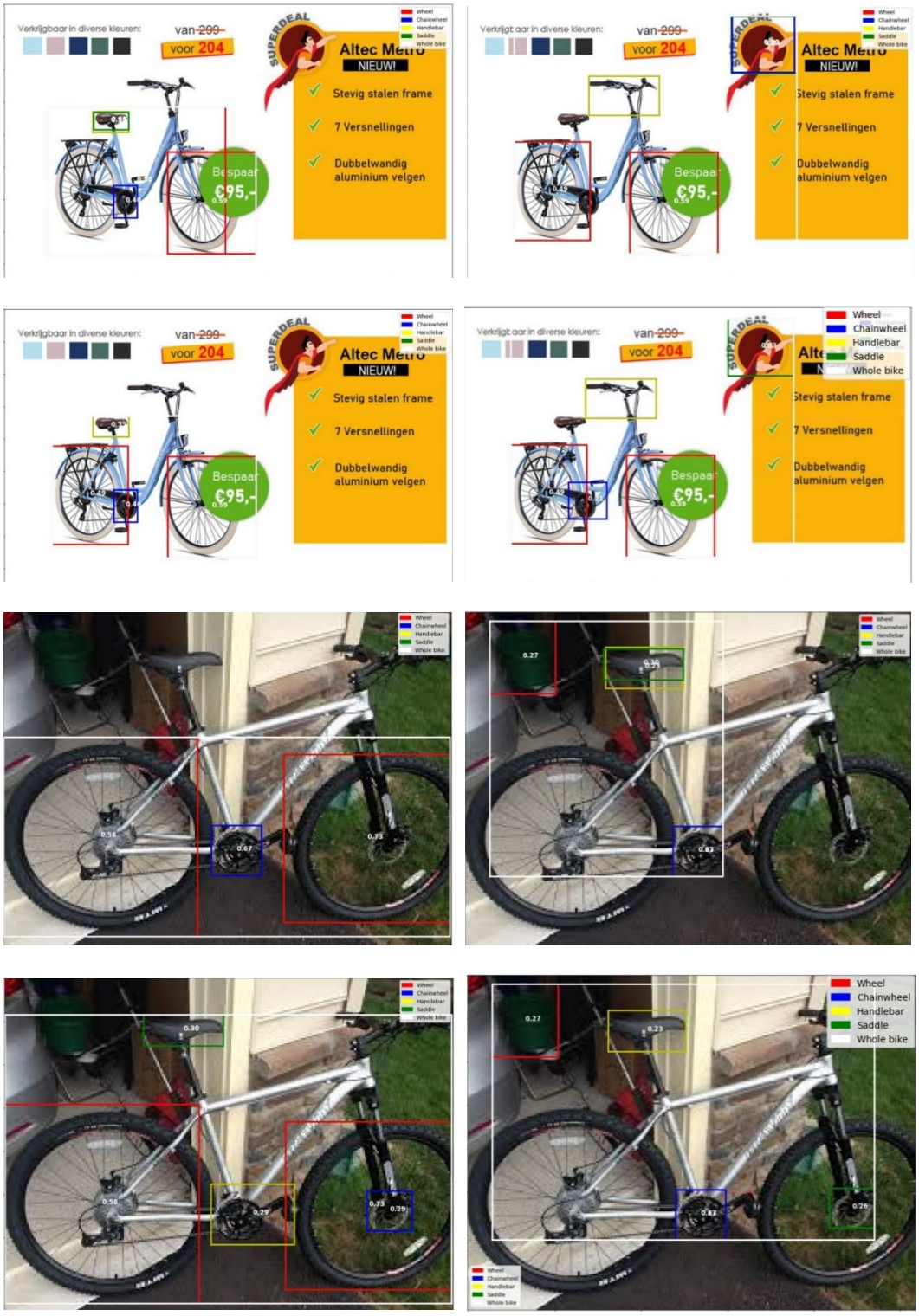

Figure 10: Localization results for two test images, when no other knowledge than the object composition is used (top left), when only area knowledge is used (top right), when only knowledge about the overlap is used (bottom left) and when both additional knowledge sources are used (bottom right).

5. Discussion

ZERO can be extended to other objects by defining them in terms of their respective parts, collecting images of parts, annotating them, and applying the method described in this paper. Better discrimination between similar but different objects can be achieved by including hard negatives. They can be taken into account explicitly by the data generator in the training procedure, or by hard-negative mining, or by weighting them in the training optimization. If the objects are better described by their properties instead of their parts, attribute-based approaches are more appropriate.

Currently, ZERO's localization method is limited to one object per image. This could be extended to multiple objects per image by anchor boxes (e.g., [14]), for which the object presence is evaluated. This generates multiple hypotheses of where the new object may be located in the image. All hypotheses should be validated one by one, by applying ZERO's recognition. Each hypothesis will result in a confidence, after which the maximum confidence can be determined and the associated localization.

There is more expert knowledge available about localization, for instance, spatial information of how parts relate to each other. This positional encoding is expected to add important cues for the part selection. Another improvement for the world knowledge would be a co-learning setting in which updates to the knowledge can be made during the deployment phase, since it is difficult to select the exact right knowledge.

Note that the parts were extracted from the Pascal VOC part-dataset. As such, the parts are cut out from images of largely visible objects. Hence the parts are not truly isolated, as a small bit of the context is visible (e.g., a small part of the bicycle where the wheel is connected to) and the parts could contain some specific bicycle-part features. This is in contrast to the real envisioned application where no images of the object are available, and the parts and the ZERO model are to be learned from images of truly isolated parts without object specific context and with more general part features. This will be addressed in our near-future research.

It would be interesting to explore the benefits of ZERO's part-based technique for robustness against adversarial attacks. In adversarial attacks, pixels of an image are weakly adjusted to force another prediction from the deep learning model. When using our part-based model, multiple predictions have to be misled in order to change the prediction of the whole image.

In ZERO's part-based recognition method, constructing additional training samples with a new type of part is relatively easy. Therefore our method would allow for fine-grained identification, using knowledge of important recognition cues. Possibly combined with attributes, to answer queries like 'Find the person with the pink bag'. We would like to explore these type of use cases in future work.

6. Conclusion

In this paper we have proposed a zero-shot object detection method based on known parts and world knowledge. Since for our zero-shot learning use case no test-images are available, we tested our method on bicycles and their parts. Our localization method allows for multi-variable input and multi-hypothesis output. For the object recognition, we outperform few shot baselines that require labeled training data. The results of localization show the potential of the method and the multi-variable input allows for updating and extending the used world knowledge.

Acknowledgements

We would like to thank the RVO research program at TNO for financial support.

References

[1] H. Touvron, A. Vedaldi, M. Douze, H. Jegou. Fixing the train-test resolution discrepancy, NeurIPS 2019.

[2] Y. Xian, C. H. Lampert, B. Schiele, Z. Akata. Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. IEEE TPAMI 2018.

[3] Y. Liu, J. Guo, D. Cai, X. He. Attribute Attention for Semantic Disambiguation in Zero-Shot Learning. IEEE ICCV 2019.
[4] V. Khare, D. Mahajan, H. Bharadhwaj, V. Verma, P. Rai. A Generative Framework for Zero-Shot Learning with Adversarial Domain Adaptation. IEEE WACV 2020.
[5] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, P. Perona. Caltech-UCSD Birds 200. 2010.
[6] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. IEEE CVPR 2010.
[7] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, A. Yuille. Detect What You Can: Detecting and Representing Objects using Holistic Models and Body Parts. IEEE CVPR 2014.
[8] Rahman, Shafin, Salman H. Khan, and Fatih Porikli. "Zero-shot object detection: Joint recognition and localization of novel concepts." *International Journal of Computer Vision* 128.12 (2020): 2979-2999.
[9] Rahman, Shafin, Salman Khan, and Fatih Porikli. "Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts." *Asian Conference on Computer Vision*. Springer, Cham, 2018.
[10] Yan, Caixia, et al. "Semantics-Preserving Graph Propagation for Zero-Shot Object Detection." *IEEE Transactions on Image Processing* 29 (2020): 8163-8176.
[11] Bansal, Ankan, et al. "Zero-shot object detection." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
[12] Hayat, Nasir, et al. "Synthesizing the Unseen for Zero-shot Object Detection." *arXiv preprint arXiv:2010.09425* (2020).
[13] Zhu, Yizhe, et al. "Semantic-guided multi-attention localization for zero-shot learning." *Advances in Neural Information Processing Systems*. 2019.
[14] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár. Focal loss for dense object detection. IEEE ICCV 2017.
[15] K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. IEEE CVPR 2016.
[16] M. Everingham, S. Eslami, L. Van Gool, C. Williams, J. Winn, A. Zisserman. The PASCAL Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision, 111(1), 98-136, 2015.