One-Shot Object Detection in Heterogeneous Artwork Datasets

Prathmesh Madhu*†, Anna Meyer†, Mathias Zinnen†, Lara Mührenberg‡, Dirk Suckow¶, Torsten Bendschus§, Corinna Reinhardt§, Peter Bell‖, Ute Verstegen‡, Ronak Kosti†, Andreas Maier†, Vincent Christlein†
*Corresponding Author
†Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg Germany
Email: prathmesh.madhu@fau.edu
‡Chair of Christian Archaeology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
§Chair of Classical Archaeology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
¶Institute of Art History, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
‖German and Art Studies, Philipps University Marburg, Germany

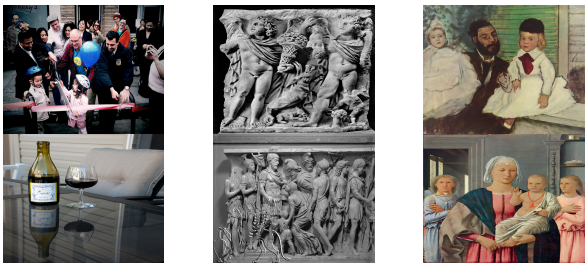
2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA) | 978-1-6654-6964-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/IPTA54936.2022.9784141

***Abstract*—Christian archeologists face many challenges in understanding visual narration through artwork images. This understanding is essential to access underlying semantic information. Therefore, narrative elements (objects) need to be labeled, compared, and contextualized by experts, which takes an enormous amount of time and effort. Our work aims to reduce labeling costs by using one-shot object detection to generate a labeled database from unannotated images. Novel object categories can be defined broadly and annotated using visual examples of narrative elements without training exclusively for such objects. In this work, we propose two ways of using contextual information as data augmentation to improve the detection performance. Furthermore, we introduce a multi-relation detector to our framework, which extracts global, local, and patch-based relations of the image. Additionally, we evaluate the use of contrastive learning. We use data from Christian archeology (CHA) and art history – IconArt-v2 (IA). Our context encoding approach improves the typical fine-tuning approach in terms of mean average precision (mAP) by about 3.5 % (4 %) at 0.25 intersection over union (IoU) for UnSeen categories, and 6 % (1.5 %) for Seen categories in CHA (IA). To the best of our knowledge, our work is the first to explore few shot object detection on heterogeneous artistic data by investigating evaluation methods and data augmentation strategies. We will release the code and models after acceptance of the work.**

***Index Terms*—one-shot, object detection, digital humanities, computational humanities, data augmentation.**

I. Introduction

Recognizing objects has now become a fundamental task in any scene understanding problem irrespective of the data domain. For example, *iconography* is a branch of art history that deals with epistemic image analysis methods for identifying, describing, and interpreting the content of images. Christian archeology, the focus of this work, investigates the material culture of the first Christians in Late Antiquity and Byzantium. A major challenge for computer vision is the rather small amount of preserved genuinely Christian images from these eras. Fortunately, the late antique processes of pictorial invention themselves come to the rescue, because the Christian images did not emerge in a vacuum, but rather recur to the pagan motifs. Well-known elements from non-Christian Roman iconography were recombined, placed in a different context, or simply retained in their message. Many motifs also persist in later periods.



(a) *COCO* (b) *Christian Arch.* (c) *IconArt*

Fig. 1: Examples representing the heterogeneity among real world dataset (COCO [2]) vs. heterogeneous artwork datasets (Christian archeology and IconArt [3]).

In order to approach the analysis of such data, a first step is the detection of the underlying objects within a scene. Afterwards, cultural references and complex relationships within artworks need to be recognized. However, simply recognizing characters with their sophisticated representations in artworks is challenging even for experts [1]. Since there is very little annotated data available in the field of digital humanities (DH), it is also difficult to apply transfer learning on these models to generate satisfactory results.

From a DH perspective, the labeling costs are high due to the expert requirements and especially for some rare narrative elements only a limited number of examples are available. There is generally much less labeled data available compared to the standard datasets used in state-of-the-art (SOTA) object detection research. For example, the widely used Common Object in Context (COCO) [2] dataset consists of over 2.5 million labeled examples which contains scenes from everyday life (cf. Fig. 1a, where an average viewer can easily to recognize objects. Imagery of artworks, as observed in Fig. 1b is, however,

978-1-6654-6964-7/22/$31.00 ©2022 IEEE

much more complex and a less targeted research area. Examples from a curated dataset from one of these less explored domains namely IconArt [3] can be observed in Fig. 1c. Due to the domain complexity and varying styles across artworks compared to COCO [2], detecting objects is highly challenging even for experts. The average viewer clearly needs practice and instructions and the narrative elements of interest need to be defined by experts in the first place, resulting in a very complex and challenging object detection problem from a computer vision perspective.

This work aims at understanding and creating a framework for identifying objects that an object detection model has *not* seen during training. This being a highly challenging task for computer vision, however, is an important requirement for experts from christian archeology or humanities due to their continual research and data collection. The re-training/finetuning of existing models for new object classes works satisfactorily well, however as soon as the classes change, it requires huge amount of annotated data (often unavailable) and longer re-training time. The goal is to build models that can generalize to novel categories from only a few labeled examples – commonly called *N-way K-shot* learning problem, where *K* is the number of available labeled samples and *N* is the number of categories. When one labeled example per category is given ($K = 1$), the problem is referred as *one-shot learning*. Accordingly, few-shot learning describes the setting in which the novel category is defined by $K > 1$ samples.

Our Contribution: We use the existing state-of-the-art framework titled Co-attention and Co-excitation (CoAE) for one-shot object detection [4] on real-world images as our baseline. In addition, we propose various novel training methodologies and modifications to the architecture that help us answer challenging research questions from the digital humanities perspective. In particular, we propose two data contextualization techniques within our training strategy which help in proposing relevant boxes improving the detection performance significantly for the UnSeen classes. We answer the following research questions: (a) Do SOTA *N-way K-shot* object detection methods generalize well for heterogeneous artworks data? (b) How useful are the transfer learned *N-way K-shot* object detection models in practice? (c) What is the impact of data complexity on the performance of SOTA *N-way K-shot* object detection methods?

II. Related Work

The goal of object detection is to recognize and localize objects in an image. Commonly, this task is solved by generating hypothetical bounding boxes, refining selected boxes and then applying a classifier. Object detection architectures can be then classified into three main categories: single-shot detectors, two-stage detectors and transformers. While Region-proposal Convolutional Neural Networks (R-CNNs) are a family of two-stage detectors [5], [6], single-shot detectors jointly perform detection and classification in one stage, thereby enabling real-time object detection [7]–[9]. All of these methods perform well on real world natural looking images while very little research has been done on object detection in paintings and digital humanities data. The main reason for this is that to train these networks we need a large amount of object-level annotations. However, this still poses a problem when a different set of objects is considered, for which new annotations will be required, limiting the usability of such models without re-training. From a computer vision perspective, one way to solve the problem of learning a method with few annotations and simultaneously enabling the detection of new object classes is *few shot object detection* (FSOD). *One-shot object detection* (OSOD) is a subset of FSOD, where only one query object image is used.

We can formalize the FSOD class setting into base (*Seen*) and novel (*UnSeen*) categories. We have access to a large number of training examples for the *Seen* categories, but only have a few training examples per class for the *UnSeen* categories. Generally, the proposed FSOD methods can be categorized into finetuning-based, metric- or meta-learning methods. Chen *et al.* [10] proposed the first finetuning strategy for FSOD, where they exploit knowledge from the source domain to train an effective detector in the target-domain with very few training examples. A meta learning technique [11] using feature re-weighting can be adapted to the detection of novel objects from a few examples. Very recently, two methods [4], [12] were proposed that can be considered the SOTA for the FSOD task. The former method [4] formalizes co-attention between query and target image to generate efficient region proposals for the one shot setting. In addition to adopting a Squeeze and Co-excitation (SCE) block for finding relevant proposals, they propose a novel margin-based ranking loss that learns a metric to predict regions similar to the query in the target image. Fan *et al.* [12] proposed a generic FSOD network that contrastively learns a matching metric between image pairs based on the Faster R-CNN backbone and a novel multi-relation detector.

III. Methodology

Motivated by the research in [4] and [12], we propose a modified OSOD pipeline for the task of FSOD for heterogenous artworks in DH, especially targeting Christian Archaeology. First, we introduce an OSOD baseline network for our work. We then introduce data contextualization methods that we append to OSOD training strategy that enhances the performance of OSOD baseline. Lastly, we explain two adaptations to the OSOD baseline architecture.

A. Baseline-OSOD model

We use the co-attention and co-excitation framework [4] for OSOD, which is presented in Fig. 2. It extends Faster R-CNN by four elements: a siamese network, co-attention (non-local features), co-excitation and a margin-based ranking loss. The backbone extracts features of both, the query patch p and the target image I, denoted as $\phi(p)$ and $\phi(I)$. To enable the Region Proposal Network (RPN) to access additional information about the query patch p, the feature maps $\phi(p)$ and $\phi(I)$ are concatenated with the output of a non-local operation (ψ) [13]. The extended target and query feature map is computed as $F(I) = \phi(I) \oplus \psi(I; p)$ and $F(p) = \phi(p) \oplus \psi(p; I)$, where

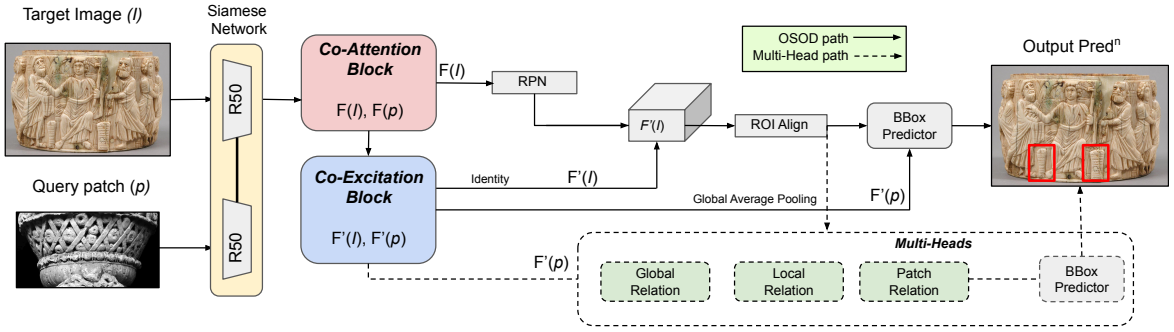


Fig. 2: Proposed Architecture of the FSOD framework. p is the query patch, I represents the target image. $F(p)$ and $F(I)$ are the extended feature maps. $F'(p)$ and $F'(I)$ are the final feature vectors after co-excitation process. BBox Predictor is a combination of BBox and Classifier Head. Architecture adapted from [4].

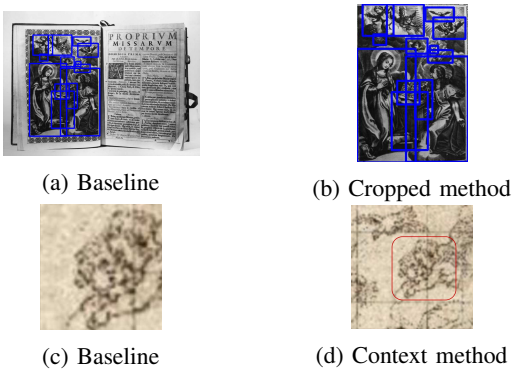
$\oplus$ denotes element-wise addition and $\psi(I;p)$, $\psi(p;I)$ are the results of the non-local operation.

Once we have the extended feature maps $F(p)$ and $F(I)$, we apply an Squeeze and Co-Excitation (SCE) similar to [14]. The processing chain of a SCE block is simple: (i) summarize each feature channel using global average pooling (GAP), (ii) calculate the excitation using two fully connected layers and (iii) apply a sigmoid to obtain scaling values between 0 and 1. The co-excitation is performed with the obtained channel-attention vector $\boldsymbol{w}$ as $F'(I) = \boldsymbol{w} \odot F(I)$ and $F'(p) = \boldsymbol{w} \odot F(p)$. $\odot$ is a dot product of the two vectors. Based on these updated features, the region proposal network (RPN) and region of interest alignment (RoIAlign) [15] are applied on the proposals and the scaled features $F'(I)$. Each proposal output is passed to a convolutional layer and then summarized by GAP resulting in $\boldsymbol{r} \in \mathbb{R}^{T \times 2048}$, where T denotes the batch size. Analogously, the query descriptor $\boldsymbol{q} \in \mathbb{R}^{T \times 2048}$ is computed. The final BBox Predictor block consists of the bounding box regression and classification head. The bounding box regression predicts the offset needed to adjust RPN's box coordinates and the classification head reports the confidence of a box containing p's category.

B. Data Contextualization

1) Cropping Method: We crop the target image such that all ground truth bounding boxes are preserved. This removes unlabeled objects in the scene and could be beneficial in avoiding the confusion related to the unlabeled objects. At the same time, the cropping procedure shows the objects at a larger scale. Scale imbalance is an identified reason for the model to have difficulties with under-represented and low-scale objects [16]. Figure 3b depicts an exemplary target image and its cropped version, where the objects are enlarged relative to the target image.

2) Context Method: Providing context for small objects with tight bounding boxes could be a visual cue that can be exploited by the model. In artworks depicting the same or similar scenes, objects often co-occur. Therefore, enlarging the ground truth bounding box for some object instances might provide more useful information during training to learn such contextual



(a) Baseline (b) Cropped method

(c) Baseline (d) Context method

Fig. 3: Data augmentation strategies: (a) & (c) are images used in the baseline method, (b) Cropped method crops the image region around all the annotated bounding boxes, (d) Context method uses larger context when the size of the bounding box is lower than the median size of bounding boxes in all the images.

cues. Experimentally, we see that the models recognize large object instances relatively easily, therefore enlarging context is applied only to small objects. The ground truth bounding box is enlarged for objects with a relative size smaller than the median value of all box sizes. Figure 3d illustrates this concept of larger bounding box giving more contextual information.

C. Modified Architectures

1) Multi-relation detector (MultiHead): The multi-relation heads (MH) provide a detector with strong discriminative ability [12]. The detector R-CNN module performs a class decision for the region proposals and produces an objectness confidence score. Considering the observations on the high number of (background) false positives, we hypothesize that the CoAE framework generates a sufficient number of proposals but struggles to classify their content correctly. The multi-relation heads might therefore benefit the CoAE framework. Figure 2 illustrates the idea of the three heads: a global relation head matches query and image based on a global relationship, a local relation head learns a pixel-wise correlation, and a patch relation head learns an arbitrary non-linear metric for matching.

The three heads jointly generate a matching score and the patch relation head additionally produces bounding box predictions. Hence, MH replaces the head architecture described in Fig. 2 with modifications required for incorporating the multi-relation heads into the existing framework.

2) Contrastive training: The CoAE uses a pair (I, p) during training, which consists of the target image I containing an instance of the target category c and the query p showing an arbitrary instance of an object of category c. Information about other categories is not taken into account. In their experiments, Fan et al. [12] found that only one negative category (n) is sufficient for letting the model learn to distinguish between different categories. Similarly, we incorporate a contrastive training strategy into our framework by using a triplet (I, p_c, $p_{n\neq c}$) for training. The third, additional element $p_{n\neq c}$ is a query of another category chosen randomly (like c). The architecture computes classification scores and bounding box predictions for the two RPN branches in the same way. Because the training is 2-way, the bounding box predictions consist of 8 values instead of 4 for the modified training of CoAE framework. For loss computation, the label values are set to zero for the negative examples. The goal of contrastive training is to let the classifier learn to distinguish between different categories and not to predict bounding boxes for the negative category.

IV. Experimental Setup

A. Datasets

Working in the field of digital humanities is highly challenging as most datasets are not publicly available and are difficult to annotate. Therefore, in our work of iconographical analysis, we present results on our non-public dataset (Christian archeology) and verify them via a public dataset (IconArt V2). The bounding box annotations for these datasets were done by experts. The datasets have an unequal image distribution, leading to a category imbalance. For each dataset, the object categories are separated into different splits for training and testing. In general, with total N object categories, $3N/4$ are used as Seen (train) and the remaining $N/4$ are used as UnSeen (test) categories. During training, input image is the target image, and bounding box of the object (randomly chosen) is considered as the query patch.

a) Christian Archeology (CHA): The individual elements of Christian iconographies were partly taken directly from pagan art and then continued over a long time period. Hence it was possible to expand the rather small image corpus of Christian archeology by non-Christian material. The corpus of Christian archaeology used here includes primarily Western and Byzantine, but also Armenian, Coptic, and Arabic images from the period between Antiquity and the Middle Ages, dating between 500 BCE and 1500 CE. In addition, almost all archeological genres have been included; from painting and sculpture to portable art and textiles. Concretely, the data consists of 16K images, with 16 annotated object categories. Specifically, these chosen objects are the narrative elements like wing, basket, cape, fountain etc. The high class-imbalance is apparent with the most frequent category having 5945 examples, whereas the rarest has only 5. In order to avoid such categories while training, we keep the categories which have at least 10 % of samples of the most frequent category. Each of the remaining 8 categories contain at least 595 samples. We create four random splits of *Seen-UnSeen* categories (like [4]) from this list. *Seen* categories are the ones the model uses for training, whereas the *UnSeen* ones are used to test the model. All the train-test category splits can be seen in Table I.

TABLE I: CHA dataset's object categories for the four train-test splits. Here, w. stands for woman.

	split 1	split 2	split 3	split 4
train	wing basket child trousers phyrg. cap sitting w.	child trousers phyrg. cap sitting w. seat cape	basket child trousers phyrg. cap sitting w. seat	wing basket child sitting w. seat cape
test	seat cape	wing basket	wing cape	trousers phryg. cap

b) IconArt v2 (IA): IconArt dataset (v1 & v2) was introduced by Gonthier *et al.* [3] for classification and detection using multiple instance learning. It consists of 10 classes and 6529 images from the art history domain. Compared to the CHA dataset, IA dataset is simpler w. r. t. objects being clearly recognizable by a human observer, low intra-class variation, and less complex scenes. Also, the category with the lowest number of samples (92) has ≈10 % as many samples as the category with the highest number of samples (1084). Hence, we keep all the categories. As annotations are only available for test-set, we divide it into random splits of *Seen-Unseen* categories for IA.

B. Setup

We consider the OSOD model (cf. Section III-A) pretrained on COCO dataset as our *pretrained* model. This model when finetuned on Split 1 of CHA dataset is defined as *baseline* model. We follow the exact same training procedure as mentioned in [4]. All the models are trained only on the *Seen* categories, but evaluated on both *Seen* and *Unseen* categories. We then compare the *pretrained* and *baseline* methods against the proposed methods (cf. Section III). Motivated from [3], where the authors compare complex artwork datasets with SOTA with a mean average precision (mAP) at intersection over union (IoU) of 0.3, we use a mAP at IoU of 0.25 for all the comparisons.

V. Results

A. *Baseline* vs. *Data Contextualization*

Table IIa shows that the *context* method outperforms all the other methods. Our *context* method improves the performance by 2.16 % mAP and 1.85 % mAP for CHA and IA respectively when compared with the baselines for *UnSeen* classes. When the *cropped* method is added to *context* method, it shows an improvement of 1.38 % mAP and 0.62 % for CHA and IA respectively when compared to the baselines for *UnSeen*

TABLE II: Evaluation of different configurations: (a) shows our Context and Cropped methods while (b) shows results for Contrastive and MultiHead. All methods are compared to the baseline for Christian Archaeology (CHA) - Split 1, and IconArt V2 (IA) - Split 2 (*Seen and UnSeen classes*). All scores represent mAP, taken at IoU=0.25. The *baseline* method is represented as '–'. Best results are highlighted in **bold**.

	Context	*Cropped*	**UnSeen**	**Seen**
CHA	–	–	9.90	25.51
	√		**12.06**	31.66
		√	8.60	27.18
	√	√	11.28	**32.72**
IA	–	–	14.39	41.08
	√		**16.24**	**41.56**
		√	12.78	39.37
	√	√	15.62	40.42

(a)

	Contrastive	*MultiHead*	**UnSeen**	**Seen**
CHA	–	–	9.90	25.51
	√		5.90	15.48
		√	**10.56**	**26.56**
	√	√	3.82	21.79
IA	–	–	**14.39**	**41.08**
	√		13.67	36.92
		√	13.84	39.33
	√	√	8.80	18.83

(b)

TABLE III: Average evaluation of Pretrained and Data Contextualization methods versus baseline for entire Christian Archaeology (CHA, all 4 splits) and IconArt V2 (IA, both splits) datasets (*Seen and UnSeen classes*). All scores represent mAP, taken at IoU=0.25. Best results are highlighted in **bold**.

	Method	**UnSeen**	**Seen**
CHA	*pretrained*	8.52	24.14
	baseline	8.64	24.12
	context	**12.26**	**30.14**
IA	*pretrained*	14.98	11.62
	baseline	21.50	29.01
	context	**25.24**	**30.52**

classes. However, the *cropped* method alone is not sufficient for improving the performance. We also observe similar improvements for *Seen* classes.

B. Baseline vs. Modified Architectures

To our surprise, we observe that the *contrastive* method performs worse than the baseline method, cf. Table IIb, whereas MultiHead shows a promising improvement. Compared to the baseline, the *contrastive* method decreases the performance by 4 % and the *MultiHead* improves the performance by 0.66 % for CHA. Similar trend is observed for IA with contrastive method, but the drop in performance is lower. The worst performance is given by the combination of both *contrastive* and *MultiHead*.

C. Pretrained vs. Data Contextualization

Since it gave the best results, we perform an evaluation of *context* methods on all splits and compare it with *pretrained* and *baseline* models in Table III. We observe consistent trend across all the splits. In absolute values, *Context* method achieves at least 3.5 % improvement for CHA and IA in *UnSeen* classes, compared to the baselines. Improvements for *Seen* classes are also similarly observed.

D. Few shot evaluation

We consider the best model from our proposed methods, that is *context* method and compare the few-shot evaluation with the baseline method. From Table IV, we can see an overall

TABLE IV: Few-shot Evaluation of Context method versus baseline for Christian Archaeology (CHA) – Split 1, and IconArt V2 (IA) – Split 2 (*UnSeen classes*). K represents the number of shots. Best results are highlighted in **bold**.

	Method	**K=1**	**K=3**	**K=5**
CHA	*baseline*	9.90	11.90	**12.12**
	context	12.06	12.46	**12.90**
IA	*baseline*	14.39	15.79	**16.04**
	context	16.24	16.96	**17.04**

improvement for 5-shot evaluation when compared to 1-shot or 3-shot evaluation for Christian Archaeology and IconArt datasets. The reason for this is when the network is shown multiple instances of the query patch, the average feature representation of the query class is a better discriminative measure than with a single query, even when it was trained in a one-shot manner.

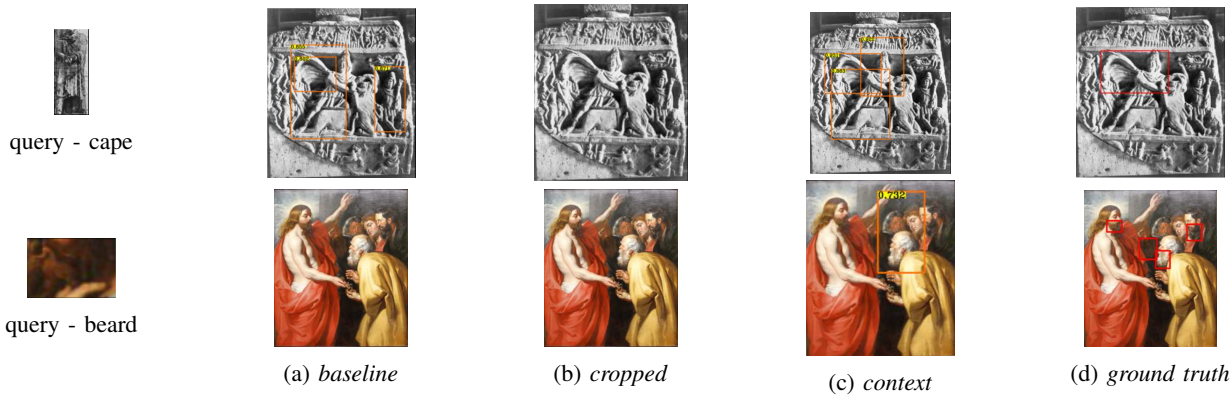
VI. Discussion

A. Qualitative Evaluation

Fig. 4 shows some qualitative results for the baseline method in comparison to our data contextualization methods. The visual results emphasize the challenge of using OSOD for heterogeneous artworks. For CHA, the *context* method is able to predict the correct class with a high confidence ($p = 0.931$) and the baseline with a lower confidence ($p = 0.802$) and lower quality bounding box regression. *Cropped* method fails in both the cases. For IA, the query class is very challenging since its scale is quite low and not easily recognizable as compared to other object classes it was trained on. This makes it even more difficult to detect such classes.

B. Research Questions

From Table III, we can observe that the *pretrained* method does not generalize well for CHA and IA. Even though the transfer learned models mostly improve the performance for heterogeneous artworks, they still seem far from the best method and are hence not very useful in practice. While one of our proposed methods (*context*) outperforms all the methods for



(a) *baseline* (b) *cropped* (c) *context* (d) *ground truth*

Fig. 4: Qualitative result showing the difficulty of one-shot object detection. We compare the (a) baseline with our data contextualization methods (b) cropped and (c) context on *UnSeen* queries for CHA (*cape*, top row) and IA (*beard*, bottom row). Zoom in the figures for more visual details.

CHA and IA, there is still room of improvement to adopt these methods in practice. To study the impact of data complexity on the OSOD performance, we compared CHA and IA datasets and saw that data complexity is directly proportional to the model performance. We also observed that adapting the SOTA OSOD architecture has very little potential to improve the performance on *UnSeen* classes while training it in a contrastive fashion hurts the overall performance. We also observed that the choice of good query patches is more effective than modifying the network architecture for the task of FSOD in heterogenous artworks.

VII. Conclusion

In this work, we analyze one-shot object detection (OSOD) methods for two heterogeneous artwork datasets *viz*. Christian archeology (CHA) and IconArt V2 (IA). Using OSOD, we observed that models pretrained on COCO dataset do not perform well for artworks datasets. We then fine-tuned these models on CHA and IA. Transfer-learning helped in improving the performance substantially. However, we observe that it is quite difficult to adapt state of art methods like contrastive training and multi-relation detector to further improve the results with consistency. Our analysis highlight that there is no assurance about which methods will be useful in practice. However, our data contextualization techniques do help to improve the model performance with fair consistency. Specifically, our *Context* method shows improvements for CHA as well as IA. We validate this observation by evaluating it for all the splits for CHA and IA. It is quite remarkable that adding more scene-context around the object annotations helps to improve general performance of OSOD models.

VIII. Acknowledgment

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the TITAN V GPU used for this research. The paper is partially funded by Odeuropa EU H2020 project under grant agreement No. 101004469.

References

[1] P. Madhu, R. Kosti, L. Mührenberg, P. Bell, A. Maier, and V. Christlein, "Recognizing characters in art history using deep learning," in *Proceedings of the 1st Workshop on Structuring and Understanding of Multimedia HeritAge Contents*, ser. SUMAC '19, 2019, p. 15–22. 1

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., 2014, pp. 740–755. 1, 2

[3] N. Gonthier, Y. Gousseau, S. Ladjal, and O. Bonfait, "Weakly supervised object detection in artworks," in *ECCV Workshops*, L. Leal-Taixé and S. Roth, Eds., 2019, pp. 692–709. 1, 2, 4

[4] T.-I. Hsieh, Y.-C. Lo, H.-T. Chen, and T.-L. Liu, "One-shot object detection with co-attention and co-excitation," in *NeurIPS*, vol. 32, 2019, pp. 2725–2734. 2, 3, 4

[5] R. Girshick, "Fast r-cnn," in *ICCV*, 2015. 2

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*, vol. 28, 2015, pp. 91–99. 2

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016, pp. 779–788. 2

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *ECCV*, 2016, pp. 21–37. 2

[9] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2999–3007. 2

[10] H. Chen, Y. Wang, G. Wang, and Y. Qiao, "Lstd: A low-shot transfer detector for object detection," in *AAAI*, vol. 32, no. 1, 2018. 2

[11] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, "Few-shot object detection via feature reweighting," in *ICCV*, October 2019. 2

[12] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-rpn and multi-relation detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4013–4022. 2, 3, 4

[13] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," *CVPR*, pp. 7794–7803, 2018. 2

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, June 2018. 3

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2980–2988. 3

[16] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, "Imbalance problems in object detection: A review," *IEEE TPAMI*, vol. 43, pp. 3388–3415, 2021. 3