Balancing Performance and Effort in Deep Learning via the Fusion of

Real and Synthetic Cultural Heritage Photogrammetry Training Sets

Eugene Ch’ng, Pinyuan Feng, Hongtao Yao, Zihao Zeng, Danzhao Cheng and Shengdan Cai

University of Nottingham
UK | CHINA | MALAYSIA

University of Nottingham Ningbo China, 199 Taikang East Road, Ningbo,
315100, China

First published 2021

Ch'ng E., Feng P., Yao H., Cheng D. and Cai S. (2021) Balancing Performance and Effort in Deep Learning via the Fusion of Real and Synthetic Cultural Heritage Photogrammetry Training Sets. Special Session on Artificial Intelligence and Digital Heritage: Challenges and Opportunities - ARTIDIGH, 13th Int' Conference on Agents and Artificial Intelligence - ICAART, 2-4 Feb. 2021

Balancing Performance and Effort in Deep Learning via the Fusion of Real and Synthetic Cultural Heritage Photogrammetry Training Sets

Eugene Ch'ng[1,3], Pinyuan Feng[2], Hongtao Yao[2], Zihao Zeng[2], Danzhao Cheng[3] and Shengdan Cai[3]

[1]NVIDIA Joint-Lab on Mixed Reality

[2]School of Computer Science

[3]Digital Heritage Centre

University of Nottingham Ningbo China

{eugene.chng, scypf1, scyhy1, scyzz3, danzhao.cheng, shengdan.cai}@nottingham.edu.cn

Keywords: digital heritage, deep learning, object detection, data augmentation, photogrammetry, fusion dataset

Abstract: Cultural heritage presents both challenges and opportunities for the adoption and use of deep learning in 3D digitisation and digitalisation endeavours. While unique features in terms of the identity of artefacts are important factors that can contribute to training performance in deep learning algorithms, challenges remain with regards to the laborious efforts in our ability to obtain adequate datasets that would both provide for the diversity of imageries, and across the range of multi-facet images for each object in use. One solution, and perhaps an important step towards the broader applicability of deep learning in the field of digital heritage is the fusion of both real and virtual datasets via the automated creation of diverse datasets that covers multiple views of individual objects over a range of diversified objects in the training pipeline, all facilitated by close-range photogrammetry generated 3D objects. The question is the ratio of the combination of real and synthetic imageries in which an inflection point occurs whereby performance is reduced. In this research, we attempt to reduce the need for manual labour by leveraging the flexibility provided for in automated data generation via close-range photogrammetry models with a view for future deep learning facilitated cultural heritage activities, such as digital identification, sorting, asset management and categorisation.

1 INTRODUCTION

One of the key advancements that led to an increased interest-driven amateur 3D recording of cultural heritage objects can be said to be a critical progress made within the field of Digital Heritage. This critical progress in combined technology and approach is close-range photogrammetry in its many implementations, methods and use (Ch'ng et al., 2019; Luhmann et al., 2006; Mudge et al., 2010; Yilmaz et al., 2007). Its advent has opened up possibilities for both digitisation and digitalisation due to its ease of use and accessibility and thus, is a catalyst for the widespread recording of objects within cultural heritage. The ability to record true appearances of objects can open up many possibilities; it removes barriers such as the effort and cost of 3D laser scanning and closes the gap between visual representations of authentic cultural heritage objects that can populate virtual environments (Cai et al., 2018; Ch'ng et al., 2019; Li et al., 2018). We see further potentials for the use of photogrammetry-based objects, and that is the use of such models for generating complementary data that could be used for augmenting datasets meant for identifying cultural relics. We believe that, for the fact that the 3D recorded objects are 1) digital, and that these objects can be 2) manipulated within a virtual space, and that many facets of each object can be generated as images, and across many collections of objects. Therefore, a dataset composed of sythetic imageries can be used as training datasets that complements existing databases. We have since created such a dataset named *DeepRelic*. What we are unsure of is, if training datasets can be entirely virtual, i.e., photogrammetry-based imageries, or if there is a ratio whereby an inflection point can be found where additional virtual imageries will no longer increase

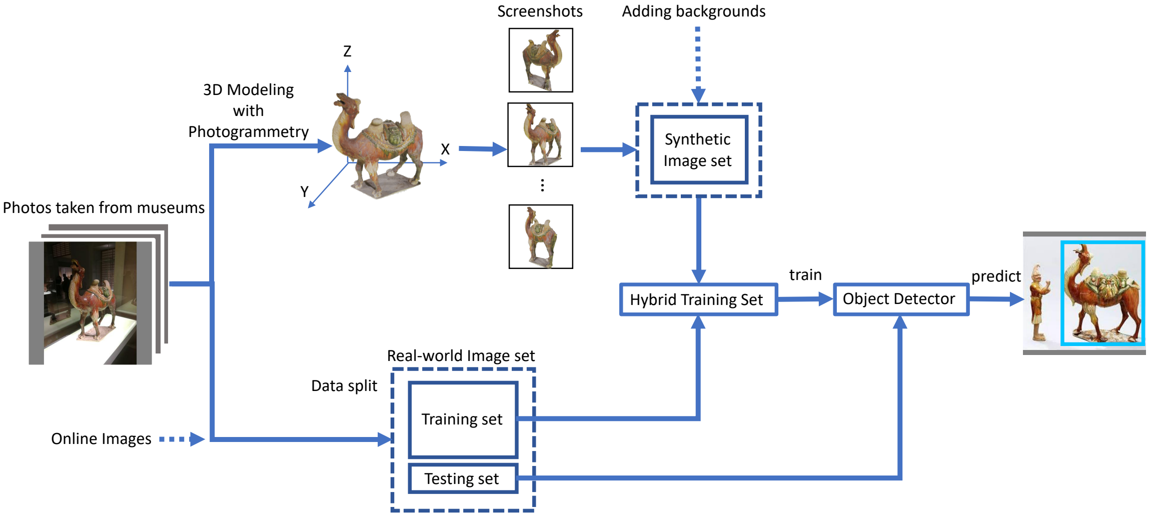

Figure 1 Workflows in our data augmentation pipeline describing the source of the photogrammetry modells, data processing, data fusion, the creation of deep learning training sets, and object detection.

deep learning performances in terms of object identification. We therefore ask the question – can virtual images compliment real images and, if so, what is the right combination as measured by the average ratio of a collection of objects with variable appearances? Our hypotheses are as follows:

H^0: There is no difference in performance between different combinations of ratios of computer-generated images as compared to real images.

H^a: There is a difference in performance if computer-generated images are combined with real images.

Our aim is to test our hypotheses, and if the alternate hypothesis is true, to discover a ratio and, associated with that ratio, further seek for an inflection point where the insertion of virtual imageries will degrade performances using fusion machine learning datasets.

The article is written as follows – we first review literatures related to our core ideas before proceeding to describe our method and approaches in how we manage the fusion of data and design experiments. Within our expriments, we describe the process of creating our *DeepRelic* dataset, through to the training process, and our testing of the performance of deep learning algorithms based on the dataset. The workflow diagram in Figure 1 illustrates our method. The article is then followed by the results of our hypotheses testing, and finally we conclude our study with future work.

2 RELATED WORKS

The genericity of the utility value of AI, and in particularly machine learning and deep learning are pervasive across many different fields. This is true within the many activities of digital heritage, where AI has been employed for various purposes. Deep learning and object detection is one such area for they can be used for assisting, facilitating and complementing human labour for archives and collections, for education and audience engagement within museums. Here, we review relevant approaches that fit our research intention prior to delving into the literatures of related works.

2.1 Object Detection

Object detection, the process of identifying objects in an image along with localisation and classification, has drawn significant benefits from the introduction of deep learning techniques over the past several years. Compared with traditional object detection methods, modern object detection algorithms have made significant improvements in accuracy, speed and memory, and thereby have become pervasive in a wide range of applications. Some examples are malaria image detection (Hung et al., 2017), automobile vision system (Chen et al., 2018) weed detection (Sivakumar et al., 2020), and etc.

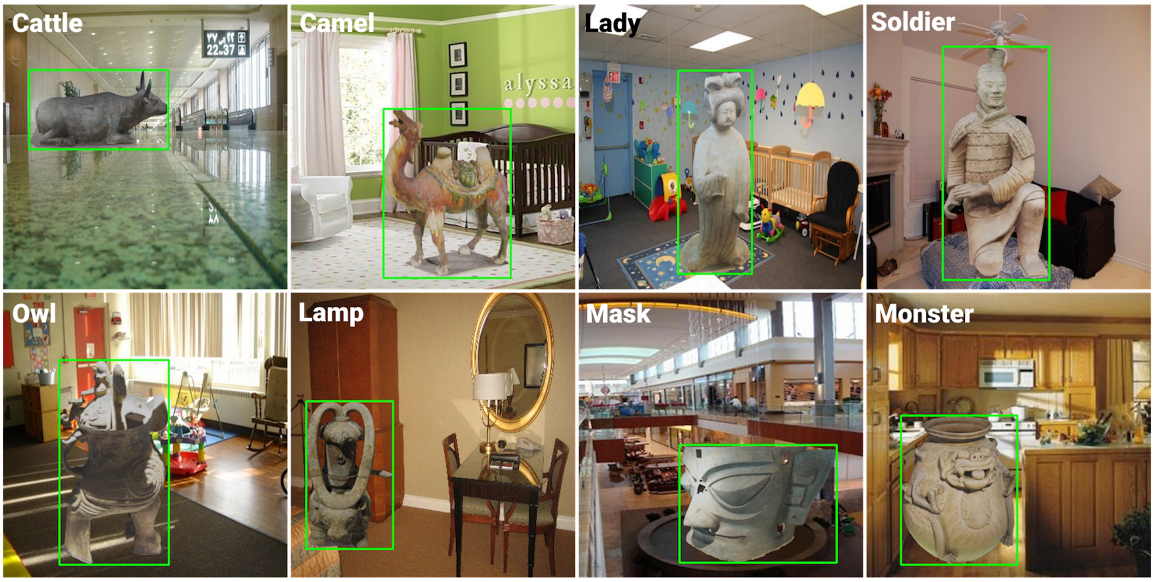

Figure 2 Samples of augmented images

There are two groups of principles behind object detection in deep learning – single-stage and double-stage approach. Single Shot Multi-Box Detector (SSD) (Liu et al., 2016) and YOLO (Redmon et al., 2016) are typical examples of the single-stage approach, which classifies objects as well as their locations within a single step. For double-stage approach, Fast Region-based Convolutional Network method, i.e., Fast R-CNN (Girshick, 2015) is a representative, which creates a set of high-probability regions that surround objects, and then conducts the final localisation and classification steps by taking these regions as input. Fast R-CNN's performance is similar to SSD in comparison with F1 score, recall, precision and IoU, but it might have higher generalisation than the SSD in weed detection (Sivakumar et al., 2020). The SSD algorithm supports low-level computational platforms with ~ 99.3% accuracy and adequate speed when used for detecting vehicle. The YOLO algorithm on the other hand focuses on real-time object detection and performs well on objects.

In our experiment, SSD algorithm was used within our training framework. SSD is capable of directly finding the bounding boxes of objects and is able to predict classes from feature maps in a single shot. This can enable us to speed up the training process. Its performance is able to achieve a level of accuracy adequate for object identification in our training datasets. Our research attempts to balance the trade-off between performance and usefulness and therefore SSD as our choice of algorithm for hypotheses testing.

2.2 Data Augmentation

Deep learning is dependent on large volumes of data for achieving best performances. However, obtaining adequate amounts of data for acceptable performance is challenging for many domains. This is especially true in the field of cultural heritage considering the limited access to museum collections and the fragility of cultural heritage objects. A viable approach is to use data augmentation. Data augmentation maximises the utilisation of small datasets whilst achieving acceptable performance, and it provides a diverse viewpoint and scale coverage to generate more images with minimal effort (Dwibedi et al., 2017). The most common approach is the transformation of existing datasets by flipping, rotating, scaling, cropping, translation and noise injection (Shorten & Khoshgoftaar, 2019). Dataset transformation are based on adequate real-world images. Cultural heritage objects are often well-preserved in the archives or encased in display glass, which makes it difficult to acquire multi-angled pictures of an object that would become composites of a good training dataset. Cultural heritage objects for image-based resources are limited as such, and also that relics are uncommon, unique and often unpublished and inaccessible in the public domain. Therefore, in order

Table 1. Our choice of cultural heritage objects used in this research and their metadata.

Item Full Name	ID	Sample Image	Year	Museum
Gilt Bronze Bull	Cattle		Western Xia Dynasty (1038-1227)	Ningxia Museum
Tri-coloured Camel	Camel		Tang Dynasty (618–907)	Nanjing Museum
Kneeling Archer	Soldier		Qin Dynasty (221-206 BC)	Emperor Qinshihuang's Mausoleum Site Museum
Ox-shaped Bronze Lamp	Lamp		Eastern Han Dynasty (25-220)	Nanjing Museum
Bronze Mask	Mask		Shu Kingdom (circa 2800-1100BC)	Sanxingdui Museum
Bronze Owl-shaped Zun with Inscription of Fuhao (*Xiaozun*)	Owl		Shang Dynasty (1600-1046BC)	Henan Museum
Pottery Figure of a Standing Lady	Lady		Tang Dynasty (618–907)	National Palace Museum, Taipei
Celadon Glazed Porcelain Zun with Design of a Monster	Monster		Western Jin Dynasty (265-316)	Nanjing Museum

to resolve this issue, our study documents and reconstructs relics as 3D models through close-range photogrammetry technniques, and makes use of the 3D nature of objects for automatically generating imageries that represent the different facets of each object, and across many different objects for data augmentation.

2.3 3D Reproductions of Cultural Heritage

The significance of digital documentation and reconstruction of cultural heritage has been emphasised (Alker & Donaldson, 2018). Digitisation can be said to be the prerequisite towards the full potentials of the digitalisation of cultural heritage. 3D repositories such as Sketchfab have provided a platform for the sharing of cultural heritage. While there are methods that provide scientific and archivable copies (Mudge et al., 2007), methods for capturing directly from museums (Ch'ng et al., 2019), and the more expensive laser scanning and even a combination of close-range photogrammetry and laser scanning techniques (Hess et al., 2015), in reviewing literatures, and in having practically captured over 200 artefacts, we think that any leisurely methods that can capture true appearances (Ch'ng, 2021), i.e., copies that retains the overall identity of the object will be adequate for data augmentation.

3 METHODOLOGY

Our main goal is to augment our cultural heritage image dataset with synthetic imageries so as to test our hypotheses. We began with a dataset of 100% real images and incrementally adds synthetic images consisting of photogrammetry-based 3D objects that is automatically generated via 360° rotations and positioning using Blender Scripts, in combination with different backgrounds. We selected cultural heritage datasets within our photogrammetry object repository, and made the choice of object structure and appearance to increase the variability of our dataset as a simulation for scalability in future work.

The dataset we created consisted of 800 real-world images and 800 synthetic images in 8 categories (Table 1). The size of images is 416 x 416 x 3 (RGB). Some real-world images were patially downloaded from the Web, while others were manually photographed at different museums. Professional annotators have been adopted to ensure that bounding boxes are tightly encapsulating the bounding edges of the target objects. We resized images with the original aspect ratio and then added paddings to avoid distortion to the image contents.

In our experiment, we randomly combined real-world images and synthetic images to form a training set of 800 object images based on different image proportions. We further created a testing set consisting of real-world images for performance evaluation.

3.1 Data Preparation

This section describes our data preparation pipeline, which consists of how 3D objects are managed and how imageries generated from the objects in combination with background images can be used for augmenting our training dataset.

3.1.1 Photogrammetry Objects

We used a combination of close-range photogrammetry objects, captured from museums and processed via RealityCapture. These processes align images and generate point-cloud and polygon data, and additional effort was needed to remove access data that are not part of the model (e.g., pedestals, visitors, labels, and etc.). Additional processing and editing were conducted within Blender to position and scale the model. Images generated from the 3D models were used for extracting features which will then be used for recovering the positions of surface points. Images of facets of the models are then sorted and categorised together with real-world object data. Details of how we digitise museum objects are available (Ch'ng et al., 2019).

3.1.2 Automated Data Augmentation

Instead of synthesising images through re-scaling, rotating, cropping, and etc., we adopted the "Model-To-Image" methodology to enlarge our dataset. We considered our data augmentation approach to be a further step from traditional approaches in that we have considerably increased the diversity of the dataset.

To speed up the production of our dataset, we exported our model from RealityCapture to Sketchfab and used a standard plugin that imports models directly from Sketchfab into Blender to save time from having to remodel our 3D objects within Blender. We added an object constraint property to the camera in Blender, so that objects are always within the viewport. An additional step was to affix the trajectory of the camera to a spherical object surface and establishing a mathematical correlation of its coordinates in three-dimensional Euclidean space. When the camera is setup, we randomly modified the camera positional coordinates so that we can capture images from all sides of the object. We set the background as transparent in the rendering pipeline of so that we can overlay the object images on background images. Through the steps above, we automated the image generation process and created projections of models with consistent resolution and transparent backgrounds across all objects. This increases the diversity of imageries, and saves considerable effort in the need for manual labour.

The next step is to use Python with image processing package to encase bounding boxes around the target objects by detecting pixel values in each image and the location of object-pixel edges. In the process, we generated the documents that record the coordinates of the bounding box and the corresponding labels that we use for inputs for the machine learning training model. The completion of our pipeline yielded a preliminary dataset containing images of objects surrounded by bounding boxes and corresponding documents containing the coordinates of the boxes (Figure 1).

We employed existing open-source datasets, i.e., "Indoor Scene Recognition" created by MIT

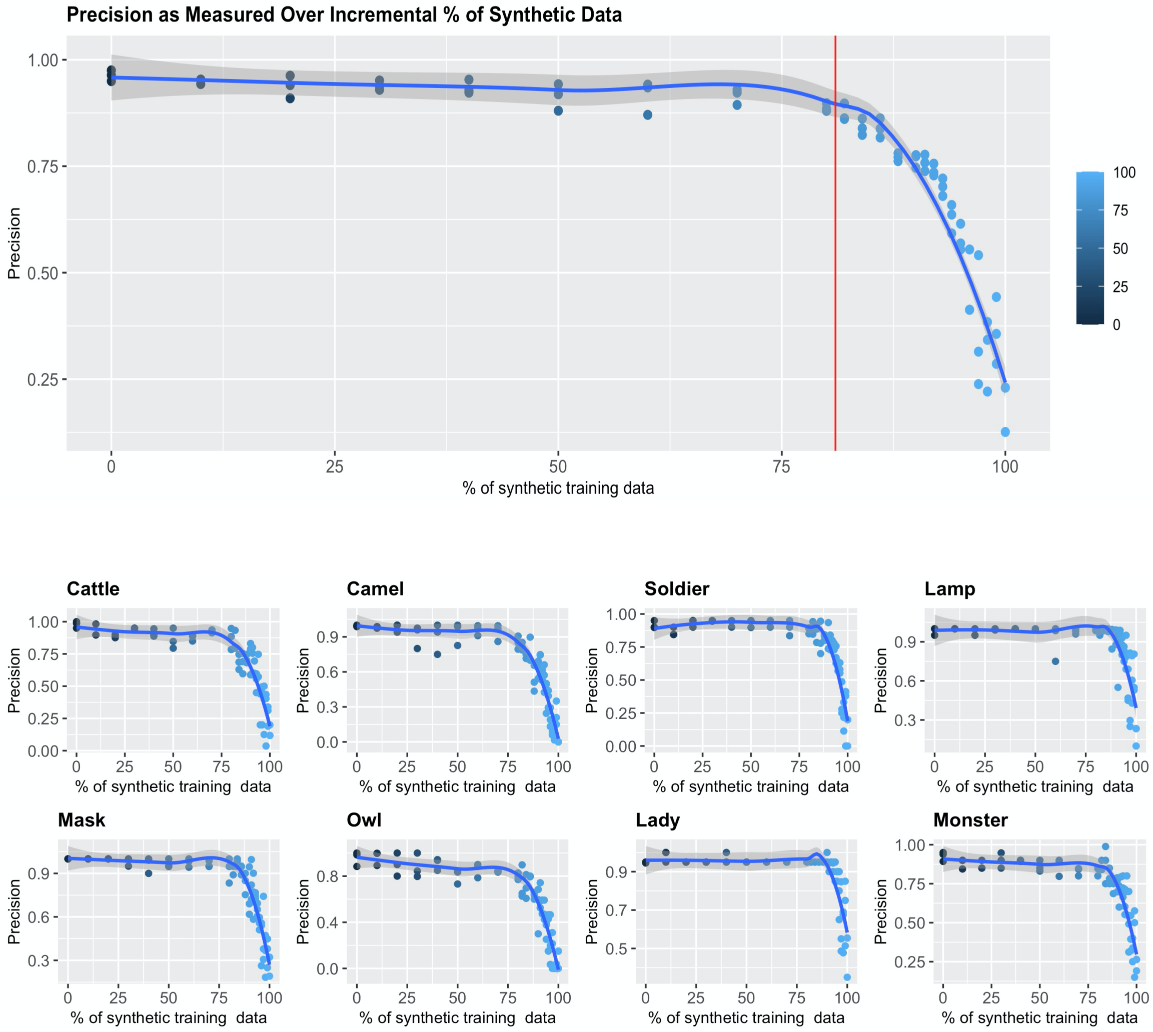

Figure 3. Average precision for all datasets (top), and precision graphs for each relic. The precision (*y axis*) were plotted against the incremental percentage of synthetic images in the dataset.

(Quattoni & Torralba, 2009) of background images as they are established datasets that meet our criteria for backdrops that contains indoor scenes. We automated the placement and scaling of objects in random locations on the background images. Based on the position and the size of each object's boundary box in the projection image, we calculated the new coordinates of the bounding box in the synthetic picture and record the related training information into annotation files. By repeating the preceding steps, we were able to generate a new dataset containing the synthetic images and the documents from the annotated information.

We now have a series of datasets with the same magnitude but with different stepped proportions of synthetic images (10% incremental). In each dataset, images were evenly divided into diverse parts with each part corresponding to a specific heritage artefact. In selecting the eight sample cultural heritage artefacts, we have considered a range of properties which includes cultural backgrounds, texture and size. We set the magnitude of a single dataset to be 800 images. We have valid reasons for the size as it is difficult to obtain images of rare artefacts, and that every cultural heritage objects are unique. This was also the reason for the size of our dataset in the present, initial research. We made sure that have the same proportion of synthetic images in each dataset. For different datasets, the ratio of synthetic images spans from 0% to 100% set at an incremental interval of 10%. When an inflection point is found in the

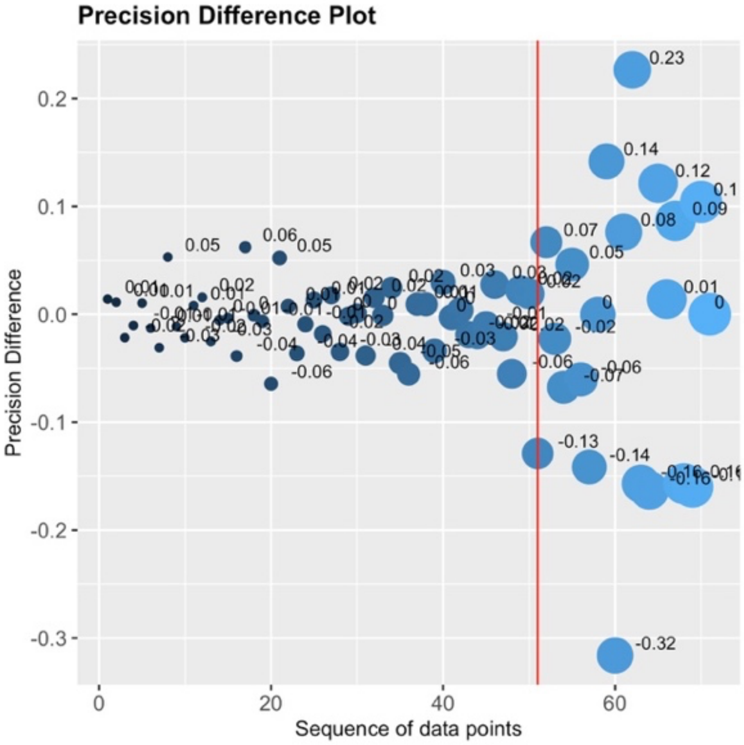

Figure 4. Precision difference plot, showing where precision has diverged. The threshold for divergence is set at T=0.08 (red vertical line).

performance data, we generated smaller incremental intervals of 1% so that we could determine the trend of the performance as it descends in line with the increase of synthetic images.

3.2 Experiment

3.2.1 Hardware and Software

We used a light workstation with GeForce RTX 2080 Super, with 64GB of RAM and Intel(R) Core(TM) i7-9700K CPU @ 3.60GHz (8 CPUs), ~3.6GHz. The training process is conducted using Tensorflow deep learning framework.

3.2.2 Model and Algorithm Selection

Given that the size of the current dataset is small, we selected a light-weight neural network architecture MobileNet to avoid excessive learning on the training set. Rather than training from scratch, we further leveraged transfer learning to fine-tune the model on our dataset to avoid overfitting.

We hope to seek an appropriate training configuration so that the results can reflect the trend with the incremental interval percentage of synthetic images that we gradually added to the dataset. Through repeated training with different object detection algorithms with the increase of ratio of synthetic images, we found that the performance (mean average precision for each object) of SSD ranges from 20% - 95%, while the performance of the other algorithms were generally over 90%. With high variance of the results obtained from SSD, we can easily see the distribution plots and observe the relationship between the change of ratio and how it influences performance. Therefore, the SSD framework satisfies our requirements.

3.2.3 "Hybrid Training" Setup

In the training process, we leveraged the SSD framework with MobileNet V2 backbone pre-trained on MS COCO dataset. To test our hypotheses and to investigate if synthetic images can influence the performance of cultural heritage object detection, we adopted the "Control Variable" methodologies to

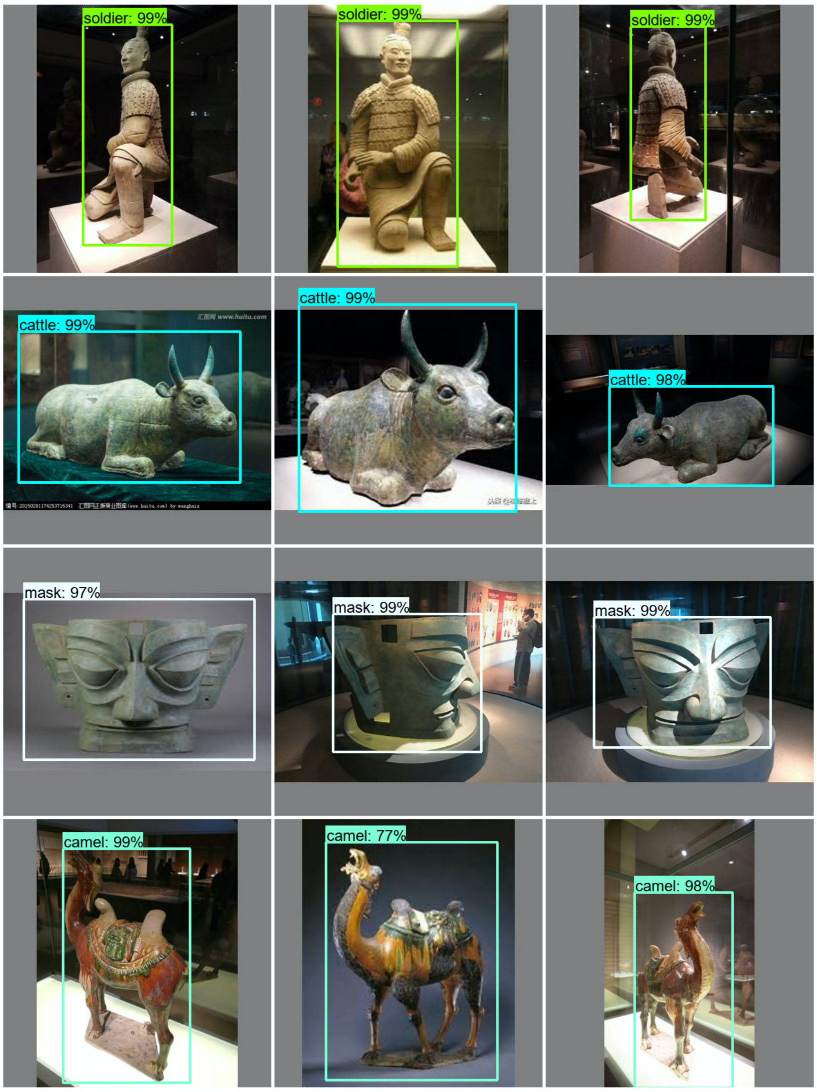

Figure 5. Object detection results of "Soldier", "Cattle", "Mask" and "Camel" based on our fusion dataset.

train different training sets and observe how different combinations of real and synthetic data can affect the learning performance. With the same training configuration, we train models on the datasets, and evaluated the object detection performance based on the same testing dataset. For model hyper-parameters, we fine-tune each deep learning model for 20K iterations using Adam Optimizer with a step-based learning rate schedule. This starts from the learning rate of 0.001, proceeding to the warming up steps until 0.003, and cosine annealing decay were performed, before gradually decreasing the moving step that approaches the minimum. To minimise the probability of outliers in our results for each ratio, we trained the model three consecutive times in order to validate that they are in alignment with each other.

4 RESULTS

Our results reported an Average Precision (AP) at IOU of 0.5 for each dataset. We trained each deep learning models using the fusion datasets we have generated separately, each yielding a result measured as a precision value. Figure 3 shows the result of the incremental percentage of synthetic images added to the dataset in each cycle of training. The *x* axis shows the percentage of synthetic images, and the *y* axis the

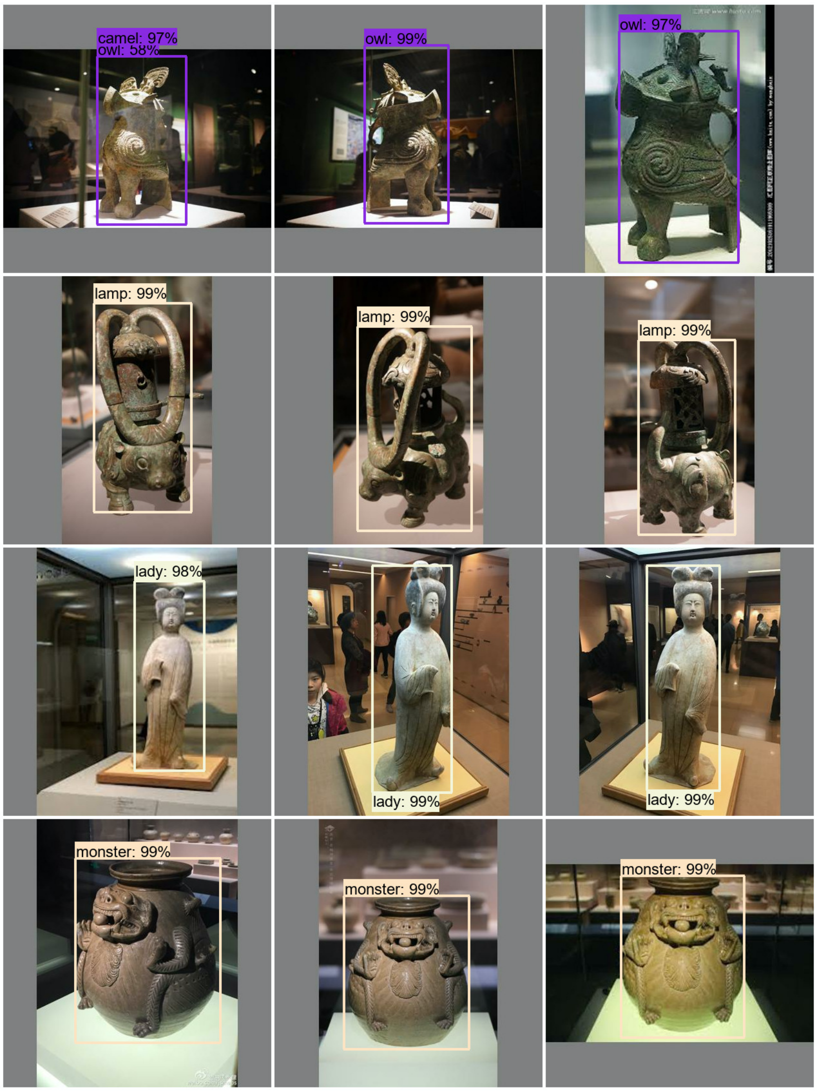

Figure 6. Object detection results of "Owl", "Lamp", "Lady" and "Monster" based on our fusion dataset.

average precision value across all eight artefacts. We can visually identify that between the range of 80% and 100% is a change in precision where the inflection point is (red line). When a visual inflection point is found, and for the subsequent datapoints toward 100% synthetic images, we reduced the granularity of our incremental percentage to 1% so that we are sure that the trend is actually descending. Therefore, there are more datapoints within the range of 80% and 100% in our dataset.

We calculated the difference between each subsequent precision data point (see Figure 4) with a threshold *T=0.08* where the difference begin to matter. The graph with 72 datapoints shows where the precision diverged from stability and begin to degrade over time. The red vertical line shows the calculation of the inflection point. The value at the threshold was used to identify the inflection point for the average precision graph (Figure 4, red vertical line).

We queried the precision of each 3D model dataset for the purpose of investigating if there are general trends. Figure 3 is the precision (*y*) and incremental percentage (*x*) of synthetic images added to the fusion dataset for each artefacts. A similar trend be observed. In Figure 3, we noted that artefact "Owl" has more volatility and lesser precision. In inspecting the 3D model (Figure 2), we realised that the contrast of light and shadow of the environment where the relic was captured was not ideal, thus contributing to darker areas that have no features. This indicated that the quality of capture is important in the dataset, and that care should be taken to ensure appropriate lighting conditions were met before including the data into the training set.

In reviewing our original research questions: 'Can virtual images compliment real images?' and, if so, 'what is the right combination as measured by the average ratio of a collection of objects with variable appearances?', we can confirm that they have been answered via the testing of the hypotheses we formulated at the beginning of the article (*H^0 and H^a*). H^0 '*There is no difference in performance between different combinations of ratios of computer-generated images as compared to real images.'* is false, and therefore the alternate hypothesis H^a *'There is a difference in performance if computer-generated images are combined with real images.'* is true. The inflection point for our present *DeepRelic* dataset is 81%, indicating that up to 80% synthetic images could be used for augmenting the training set without performance tradeoffs.

5 CONCLUSIONS

Our research began from exploring how we may combine deep learning with digitalisation activities, such as the use of deep learning algorithms for object recognition, and for identification and augmentation in mixed reality developments within our range of digital heritage activities. We then looked at a combination of algorithms, datasets and devices for that purpose, and projected that there will be tradeoffs between what types of deep learning algorithms we use, and the devices, i.e., smartphones, AR headsets that are able to support digitalisation endeavours. As with any deep learning developments, one of the greatest challenges in venturing into new areas is the effort needed to acquire data. Within the field of digital heritage, this challenge is amplified by the rarity of cultural heritage object image datasets, and that relics are uncommon, unique and often unpublished and inaccessible in the public domain. This limitation, together with our experience of close-range photogrammetry prompts us to probe if fusion datasets would work, which led to our research enquiry into the possibility of a level of performance in object detection that could be facilitated by an initially limited dataset, that could be augmented by synthetic imageries generated from photogrammetry data. We hypothesised that there would be a difference in performance if computer-generated images are combined with real images, and also asked where the inflection point would be.

We created a workflow to test our hypotheses, resulting in a dataset which we termed *DeepRelic*, and conducted experiments in order to determine where the inflection point if any would be. We discovered based on our fusion dataset that up to 80% of virtual images could be used without significant performance tradeoff and thereby have answered our research questions. Of course, as data increases, and particular collections of cultural heritage objects become diverse, the inflection point and consequently, the ratio may change. Nevertheless, collections of artefacts that are used for digitalisation are often bounded to particular exhibits and therefore, the size of the dataset may not increase that much. It does matter when digitalisation activities are expanded to an entire city.

This research has led to some questions which we will be probing for the future. Presently, we are interested to know if the precision and percentage of synthetic data will be similar in all cases, across collections, since this may be influenced by factors such as the size of the dataset, the content of the images, the quality of synthetic images, and etc. Our aim is to continue our initial exploration and hypotheses testing by expanding our *DeepRelic* dataset with models from different collections that may contain similar forms so as to explore limits for the fusion approach which we have discussed in this article.

In summary, the creation of new datasets are challenging on many fronts, but our experiments have demonstrated the potentials of fusion datasets that have composites of both real and synthetic images. We strongly believe that imageries of 3D models generated by close-range photogrammetry can complement real-world datasets, and thus can tremendously increase the magitute of datasets, but also considerably reduce human effort.

ACKNOWLEDGEMENTS

This work is part of an annual summer internships carried out at the NVIDIA Joint-Lab on Mixed Reality, at the University of Nottingham's China campus. The authors acknowledge the financial support from the Ningbo Science and Technology Bureau (2017D10035) that has made this research work possible. We would like to especially thank

RealityCapture for the tremendous support which has made this and other exciting, investigative research possible.

REFERENCES

Alker, Z., & Donaldson, C. (2018). Digital Heritage. *Journal of Victorian Culture*, *23*(2), 220–221. https://doi.org/10.1093/jvcult/vcy019

Cai, S., Ch'ng, E., & Li, Y. (2018). A Comparison of the Capacities of VR and 360-Degree Video for Coordinating Memory in the Experience of Cultural Heritage. In *Digital Heritage 2018*. San Francisco, USA: IEEE.

Ch'ng, E. (2021). Asking the Right Questions when Digitising Cultural Heritage. In *"Convergence of Digital Humanities", International Conference on Digital Heritage: Convergence of Digital Humanities*. Gyeongju, Republic of Korea, 20 September 2019.

Ch'ng, E., Cai, S., Zhang, T. E., & Leow, F.-T. (2019). Crowdsourcing 3D cultural heritage: best practice for mass photogrammetry. *Journal of Cultural Heritage Management and Sustainable Development*, *9*(1), 24–42. https://doi.org/10.1108/JCHMSD-03-2018-0018

Ch'ng, E., Cai, S., Zhang, T. E., Leow, F. T., Ch'ng, E., Cai, S., … Leow, F. T. (2019). Crowdsourcing 3D cultural heritage : best practice for mass photogrammetry. *Emerald Publishing Limited 2019*, *9*(1), 24–42. https://doi.org/10.1108/JCHMSD-03-2018-0018

Ch'ng, E., Li, Y., Cai, S., & Leow, F.-T. (2019). The Effects of VR Environments on the Acceptance, Experience and Expectations of Cultural Heritage Learning. *Journal of Computing and Cultural Heritage*.

Chen, L., Zhang, Z., & Peng, L. (2018). Fast single shot multibox detector and its application on vehicle counting system. *IET Intelligent Transport Systems*, *12*(10), 1406–1413. https://doi.org/10.1049/iet-its.2018.5005

Dwibedi, D., Misra, I., & Hebert, M. (2017). Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 1310–1319. https://doi.org/10.1109/ICCV.2017.146

Girshick, R. (2015). Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, *2015 Inter*, 1440–1448. https://doi.org/10.1109/ICCV.2015.169

Hess, M., Petrovic, V., Meyer, D., Rissolo, D., & Kuester, F. (2015). Fusion of multimodal three-dimensional data for comprehensive digital documentation of cultural heritage sites. *2015 Digital Heritage International Congress, Digital Heritage 2015*, 595–602. https://doi.org/10.1109/DigitalHeritage.2015.7419578

Hung, J., Rangel, G., Chan, H. T. H., Leônidas, I., Paulo, U. D. S., Ferreira, M. U., … Carpenter, A. E. (2017). Applying Faster R-CNN for Object Detection on Malaria Images. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 56–61.

Li, Y., Ch'ng, E., Cai, S., & See, S. (2018). Multiuser Interaction with Hybrid VR and AR for Cultural Heritage Objects. In *Digital Heritage 2018*. San Francisco, USA: IEEE. Retrieved from file:///Users/yueli/Library/Application Support/Mendeley Desktop/Downloaded/Li et al. - 2018 - Multiuser Interaction with Hybrid VR and AR for Cultural Heritage Objects.pdf

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9905 LNCS*, 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

Luhmann, T., Robson, S., Kyle, S. A., & Harley, I. A. (2006). *Close range photogrammetry: principles, techniques and applications*. Whittles.

Mudge, M., Ashley, M., & Schroer, C. (2007). A digital future for cultural heritage. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *36*(5/C53).

Mudge, M., Schroer, C., Earl, G., Martinez, K., Pagi, H., Toler-Franklin, C., … Ashley, M. (2010). Principles and practices of robust photography-based digital imaging techniques for museums. In *VAST 2010: The 11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage*.

Quattoni, A., & Torralba, A. (2009). Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 413–420). IEEE.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 779–788. https://doi.org/10.1109/CVPR.2016.91

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0197-0

Sivakumar, A. N. V., Li, J., Scott, S., Psota, E., Jhala, A. J., Luck, J. D., & Shi, Y. (2020). Comparison of object detection and patch-based classification deep learning models on mid-to late-season weed detection in UAV imagery. *Remote Sensing*, *12*(13). https://doi.org/10.3390/rs12132136

Yilmaz, H. M., Yakar, M., Gulec, S. A., & Dulgerler, O. N. (2007). Importance of digital close-range photogrammetry in documentation of cultural heritage. *Journal of Cultural Heritage*, *8*(4), 428–433. https://doi.org/10.1016/j.culher.2007.07.004