Small Object Detection in Complex Images: Evaluation of Faster R-CNN and Slicing Aided Hyper Inference

Fatma Mazen Ali Mazen[1]*, Yomna Shaker[2]
Faculty of Engineering-Electrical Engineering Department, Fayoum University, Fayoum, Egypt[1,2]
Engineering Department, University of Science and Technology of Fujairah (USTF), Fujairah, United Arab Emirates[2]

***Abstract*—Small object detection has many applications, including maritime surveillance, underwater computer vision, agriculture, traffic flow analysis, drone surveying, etc. Object detection has made notable improvements in recent years. Despite these advancements, there is a notable disparity in performance between detecting small and large objects. This gap is because small objects have less information and a weaker ability to express features. This paper investigates the performance of Faster Region-Based Convolutional Neural Networks (R-CNN), one of the most popular and user-friendly object detection models for head detection and counts in artworks rather than images of real humans. The impacts of Slicing Aided Hyper Inference (SAHI) on the enhancement of the model's capability to detect small heads in large-size images are also being analyzed. The Kaggle-hosted Artistic Head Detection dataset was used to train and evaluate the proposed model. The effectiveness of the proposed methodology was demonstrated by integrating SAHI into two other object detection models, Cascaded R-CNN and Adaptive Training Sample Selection (ATSS). The experimental results reveal that applying SAHI on top of any object detector enhances its ability to recognize and detect tiny and various scaled heads in large-scale images, which is a significant challenge in numerous applications. At a confidence level of 0.8, the SAHI-enhanced Faster R-CNN achieved the best private Root Mean Square Error (RMSE) score of 5.31337, while the SAHI-enhanced Cascaded R-CNN obtained the highest public RMSE score of 3.47005.**

***Keywords*—Faster R-CNN; Cascaded R-CNN; SAHI; ATSS; artistic head detection; small object detection**

I. Introduction

Recently, there has been a rapid increase in the development of digital fine art collections [1]. The maintenance of digital archives is filled with difficulties, but they have immense potential as a vital resource for documenting studies and stimulating development within museum narratives. This automatic annotation of digital artworks provides content analysis creativity, which helps with the task of protecting and maintaining cultural resources. Moreover, it can enhance virtual reality experiences in museums and access to internet data sources [2]. Deep neural networks beat all prior machine-learning algorithms in computer vision, achieving the best object detection accuracy. Deep learning (DL) is a machine learning technology that enables direct learning of features from data. Unlike traditional machine learning algorithms, which necessitate some human involvement to generate customized features, DL can determine these features on its own. Object detection is commonly achieved with DL utilizing deep CNN, which have made significant contributions [3]. However, including a CNN trained with real-world images in the detection of artistic paintings poses challenges due to the substantial dissimilarities between the two in terms of low-level features, including color histograms and texture statistics. The representation of painting pictures can also vary significantly, as there exist numerous creative approaches through which they can be depicted. In this study, three object detection models, Faster R-CNN [4], which is an extension of Fast R-CNN, Cascaded R-CNN [5], ATSS [6], are trained for the task of head detection in artworks. To train and evaluate the proposed models, the Kaggle-hosted Artistic Head Detection dataset [7] presented by Scale Rapid [8] was utilized. The dataset includes paintings, prints, and drawings from public-domain artwork with different resolutions and various scales. While some images have one head, others include several tiny, medium, and large-scale heads. The high-resolution images are first preprocessed with SAHI [9] to tackle the issue of many tiny heads in high-resolution images during inference time. SAHI was used to segment the images into several overlapping slices, leading tiny objects to occupy more significant pixel regions on the resulting images. As a result, the model's capacity to recognize and detect tiny heads improves.

This research study constitutes the first attempt to address the problem of automatic artistic head detection in artworks using Faster R-CNN, Cascaded R-CNN, and ATSS models. Additionally, this study is the first to experiment with the Kaggle-hosted Artistic Head Detection dataset. The results obtained from this study can provide valuable guidance for future research endeavors in this domain. Furthermore, this paper presents a generic solution for enhancing the accuracy of any object detector, by integrating SAHI into the detection process. The structure of this paper encompasses five distinct sections. A comprehensive overview of the related work is provided in Section. II. Section III outlines the dataset, while Section IV details the Methodology. Section V provides a complete analysis and discussion of the experiment outcomes. Finally, section VI presents the research's conclusion and future scope.

II. Related Work

Many DL methods, like those in [1] and [10], have been proposed to identify the artist, style, or genre in artistic artworks. In [1], a study was conducted to identify the optimal set of visual features that would yield the highest level of accuracy in artist, style, and genre classification. They studied

Fig. 1. Samples of artistic head detection dataset [7].

the application of metric learning methodologies and the performance of various visual features to learn similarities in a collection of fine-art paintings. To test performance for the tasks mentioned above, they performed comparative studies using the most extensive publicly available collection of fine-art paintings. In [10], a large-scale study using CNNs was proposed to classify the genre, style, and artist of fine-art paintings. The key objective of their research was to determine whether the machine can capture "imagination" in paintings. To validate their work, they utilized the large-scale "Wikiart paintings" dataset, which contains over 80,000 paintings. Their approach reached an accuracy of (68%) in overall performance. In another study [11], the authors proposed novel solutions to overcome the shortage of labeled training data for digital fine-art paintings and therefore leverage the promise of deep learning in this application. In their research, they employed artistic style transfer as a means of dataset augmentation on natural images, utilizing specific transformations to enhance the training dataset size. Subsequently, they employed labeled paintings as training images for various classification tasks, including style recognition. Two parallel CNNs were trained, and their output features were combined in a support vector machines (SVM) classifier. The researchers utilized multiple datasets, such as PASCAL VOC 2012, the Painting dataset, and the WikiArt dataset, to train their proposed models. Through a cross-validation test using fine-art painting images, their methodolgy outperformed a competing strategy, demonstrating higher average accuracy. This suggested technique enables real-time object detection on digital paintings, contributing to advancements in cultural heritage preservation, enhancing online resources, and enriching cultural experiences during trips.

Regarding DL and object recognition in digital fine-art painting, a new methodology was proposed in [12] for performing object retrieval in paintings using CNN and transfer learning. They demonstrated that CNNs features generated from diverse natural picture resources could effectively retrieve paintings containing these specific objects. Moreover, they developed a system that trains object classifiers from Google Photos and then utilizes them to detect a wide range of previously unknown items in a dataset that contains 210,000 paintings.

There are other machine-learning researches on using brushstrokes to recognize artists, like those proposed by [13] and [14]. In [13], various signal processing approaches were utilized such as Wavelet transforms, the Hidden Markov Model (HMM), and geometric characteristics of strokes to visually analyze brushwork in paintings for artist identification. Van Gogh utilized pre-packaged tube colors, thus the rheology of his paints was predominantly influenced by the commercial methodologies employed in their preparation. The surface upon which brushstrokes are placed is another crucial component influencing their appearance. The authors used a dataset of 101 high-resolution grayscale scans of paintings to evaluate the results of the proposed approaches. A computational method was presented in [14] to authenticate artistic works, primarily sketches, and paintings, using high-resolution scans of the originals. This approach utilizes the statistical analysis of first- and higher-order wavelet statistics to construct a model that characterizes an artist based on authenticated artwork scans. This model is subsequently employed to compare and evaluate new works for authentication purposes. Their early findings demonstrated that these approaches, in conjunction with current physical authentication, would play a significant role in art forensics.

In their research [15], the authors introduced a three-stage methodology aimed at improving the detection accuracy of small objects within aerial images. Employing the VisDrone-2019 dataset for both training and evaluating a modified RetinaNet model, they adjusted anchor parameters as part of this process. To address the issue of class imbalance, various augmentation techniques were employed. Their proposed approach demonstrated superior performance compared to other existing object detection models.

To enhance the real-time capabilities of detecting small targets within aerial imagery, the authors of [16] developed the CMF-YOLOv5s model. This included the design of a

novel multi-scale fusion module (MFF) and the construction of a multi-scale detection head with four outputs, aimed at augmenting the network's capacity to perceive small targets. They employed a genetic algorithm to optimize the K-means algorithm, thereby generating more suitable anchor boxes for aerial images. The proposed model was evaluated using the VisDrone-2019 dataset. In comparison to the original YOLOv5s, the detection accuracy metrics, specifically mAP_0.5 and mAP_0.5:0.95 for small targets, were enhanced by 5.5% and 3.6%, respectively. Furthermore, the model demonstrated superior performance over eight lightweight object detection models.

In another related study [17], a novel RetinaNet model was introduced to improve the detection of small drones in infrared imagery. Firstly, the researchers developed a super-resolution texture-enhancement network aimed at improving the texture-related information for small infrared targets. Additionally, they incorporated an asymmetric attention fusion mechanism to enhance semantic and locational detail information. Furthermore, a global average pooling layer was utilized to capture the global spatial information necessary for the classification stage. The proposed model was trained and evaluated using the publicly available infrared image dim-small drone target detection dataset. The experimental results demonstrated that this approach outperformed other existing mainstream methods in terms of detection accuracy and can be applied to any small object detection task.

In the study [18], the ASFF-YOLOv5s model, a real-time algorithm for detecting small targets in unmanned aerial vehicle (UAV) imagery, is presented. The model employs Adaptively Spatial Feature Fusion (ASFF) to enhance the capability of multi-scale information fusion. Furthermore, the quality of anchor frames was improved using the K-means algorithm. The authors also incorporated the Convolutional Block Attention Module (CBAM) to effectively capture significant features while suppressing redundant ones. The SIoU loss function was utilized to achieve a better convergence rate. The proposed model was trained and evaluated using the VisDrone2021 dataset. Compared to the original YOLOv5s model, the proposed model demonstrated significant improvements in precision, F1-score, and mean Average Precision (mAP) values.

Feng, Qihan et al. [19] provided a comprehensive survey on recent approaches based on deep learning for addressing the challenge of small object detection (SOD). They examined the various challenges inherent in SOD and systematically analyzed the methodologies employed to mitigate these challenges, such as data augmentation, scale-aware training, and enhancement of input feature resolution. Furthermore, the study emphasized the prevalent SOD tasks, including the detection of small pedestrians, faces, and objects in aerial imagery. Finally, the authors conducted a detailed evaluation of the performance of SOD models utilizing four well-recognized small object datasets.

IMD-Net [20] is an interpretable multiscale detection network developed to identify dim and small objects in infrared images with complex backgrounds. The network first enhances objects and extracts shallow detail features before acquiring high-level semantic features through a series of multiscale object enhancement modules. Low-level and high-level features are then iteratively fused after computing the global object response, allowing for pixel classification of objects and background noise. The process is finalized by multiple loss joint constraint networks that refine pixel classification to match actual object distributions. Comparative and ablation tests validate the robustness and effectiveness of the network, showcasing its strong object detection and contour description capabilities in challenging infrared conditions and its high reliability.

Concerning SAHI, the authors of [9] conducted experiments with Fully Convolutional One-Stage Object Detection (FCOS) [21], Task-aligned One-stage Object Detection(TOOD) [22], and VFNet [23], models and discussed the results of sliced fine-tuning and slicing-aided hyper inference for their models. They have shown that SAHI enhanced tiny object recognition performance while decreasing big object detection performance in particular circumstances. They also demonstrated that sliced fine-tuning enhances tiny object detection performance. The only drawback to take into consideration is that sliced inference requires a longer model inference time due to the additional quantity of information that the models must process.

In another study [24], the performance of Exceeding You Only Look Once (YOLOX) and YOLOv5 was evaluated for tiny object detection. They used the challenging VisDrone2019Det dataset to train and test the proposed models. This dataset is hard to analyze since most items are tiny compared to the image sizes. They demonstrated the benefits of slicing-aided inference in boosting the Average Precision (AP50) score in all experiments.
The main aim of this study is to build an automated system capable of detecting and counting artistic heads in artworks. To achieve this, three commonly utilized object detection models, known for their effectiveness in addressing this complex task, were employed. Additionally, SAHI, a generic approach for enhancing the accuracy of detecting small objects, was applied. The key parameters that can influence model predictions were then reviewed. Our future directions include the integration of SAHI with cutting-edge object detection models to enhance detection accuracy. Furthermore, the development and deployment of a mobile application specifically designed for museum environments, allowing widespread access to the SAHI model, is also aimed for.

III. The Dataset

The dataset utilized in this study is the Kaggle-hosted Artistic Head Detection dataset [7] created by Scale Rapid, the fastest platform that assists in annotation and obtaining high-quality labels. The key purpose of the challenge is to build a model for identifying and counting heads in works of art instead of images of real people. The Metropolitan Museum of Art in New York provided the original images for this dataset. Each image is a print, painting, or drawing from public domain artwork, as shown in Fig. 1.

Each head is at least 50 pixels wide and 50 pixels tall. The dataset labelers were told to disregard heads with no visible face. The image files are stored in the train/ and test/ directories, with the filename representing the unique id. For example, the train with boxes.csv comprises one entry for each

image in the train/ folder, with three columns: id, num human heads, and boxes.

The filename in the train/ folder corresponds to the id. The num human heads are the number of heads in the image that meet the conditions mentioned above. Finally, the boxes column is a list of bounding boxes, where each bounding box has the format (x min, x max, y min, y max) that specifies the pixel coordinates of the box, measured from the image's upper left-hand corner. It was converted to the Common Objects in Context (COCO) format to facilitate training. Although the data set comprises images with only one head, it also contains images with multiple heads. Fig. 2 depicts some images and their corresponding bounding boxes overlaid on them.

IV. Methods

This section presents an introduction to the fundamental principles of the Faster R-CNN, Cascade R-CNN, and ATSS models.

A. Faster R-CNN

Faster R-CNN is an extension of Fast R-CNN. It is composed of two blocks; the RPN module generates region proposals, while the Fast R-CNN module identifies objects in the suggested regions. As shown in Fig. 3, the first stage involves applying a proposal sub-network ("H0") on the whole image to generate initial detection hypotheses defined as object proposals. These hypotheses are then processed in the second stage by a region-of-interest detection sub-network ("H1"), also known as the detection head. Each hypothesis is given a final classification score ("C1") and a bounding box ("B1").

Cascade R-CNN is a multi-stage version of the well-known two-stage R-CNN object identification method as depicted in Fig. 4.

B. Cascaded R-CNN

It is comprised of a sequence of end-to-end trained detectors with progressively increasing Intersection over Union (IoU) thresholds, making them pickier for near false positives. The output of a prior stage detector is passed on to a subsequent stage detector, and the detection results are enhanced stage by stage.

C. Adaptive Training Sample Selection (ATSS)

Adaptive Training Sample Selection (ATSS) is a technique proposed for automatically selecting positive and negative samples based on the statistical properties of the object. It acts as a bridge between anchor-free and anchor-based detectors. It considerably enhances the performance of state-of-the-art detectors by a wide margin to 50.7% AP without adding any overhead.

V. Results and Discussion

This section presents an analysis of the outcomes obtained from the object detection models proposed in this study, utilizing the competition evaluation metric and other established metrics commonly employed for object detection problems. Furthermore, an examination of the integration of the SAHI method is undertaken, with emphasis placed on its fundamental role in the accurate detection of tiny objects. The experiments were executed using Python programming language on a Kaggle platform, utilizing an NVIDIA TESLA P100 GPU for computational acceleration.

For the sake of simplicity, the evaluation metric for this competition is the root mean square error or RMSE. RMSE is often used in forecasting and regression analysis to validate experimental results. RMSE is given by (1):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{\text{true}} - y_{\text{pred}})^2} \tag{1}$$

where the variables y_{true} and y_{pred} represent the actual and predicted number of artistic heads, respectively, and n is the number of samples in the dataset. The RMSE metric calculates the differences between predicted values and actual values, equally penalizing overestimations and underestimations to evaluate the accuracy and precision of the prediction. The result of this calculation is then subjected to a square root operation to obtain the root-mean-square value.

It is required to forecast the number of human heads larger than 50px by 50px and not look away from the viewer. Compared with the baseline network, the performance of all models is enhanced when using SAHI. With a confidence level of 0.8, the SAHI-enhanced Faster R-CNN achieved the best private RMSE of 5.31337, while the SAHI-enhanced Cascaded R-CNN obtained the highest public RMSE of 3.47005. This study aims to thoroughly assess object detection models and evaluate their ability to identify objects of varying sizes, shapes, and orientations. To evaluate and quantify the performance of these models, various forms of the mean average precision (mAP) metric are typically employed, including mAP_0.5, mAP_0.75, mAP_s, mAP_m, and mAP_0.5:0.95 are shown in Fig. 6. Equations (2), (3), and (4) outline the mathematical procedure for computing Precision (P), Recall (R), and mean Average Precision (mAP) respectively:

$$P = \frac{(TP)}{(TP + FP)} \tag{2}$$

$$R = \frac{(TP)}{(TP + FN)} \tag{3}$$

$$mAP = \frac{1}{n}\sum_{j=1}^{n} AP_j \tag{4}$$

where:

AP= $\int_0^1 P(R)dR$, TP is the True Positive, FP is the False Positive, FN is the False Negative, and n is the number of classes. One commonly used metric is mAP@[.5:.95], which is defined as the average precision of the model at different IoU thresholds ranging from 0.5 to 0.95. Specifically, mAP_0.5 measures the average precision when the IoU threshold is set at 0.5, while mAP_0.75 measures the average precision at an IoU threshold of 0.75. In contrast, mAP_s, mAP_m, and mAP_l utilize the average precision value within the IoU threshold

(a) Single head.

(b) Multiple heads.

Fig. 2. Sample images and corresponding bounding boxes [7].

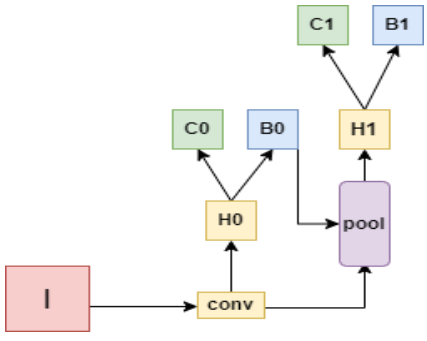


Fig. 3. Faster R-CNN network architecture.

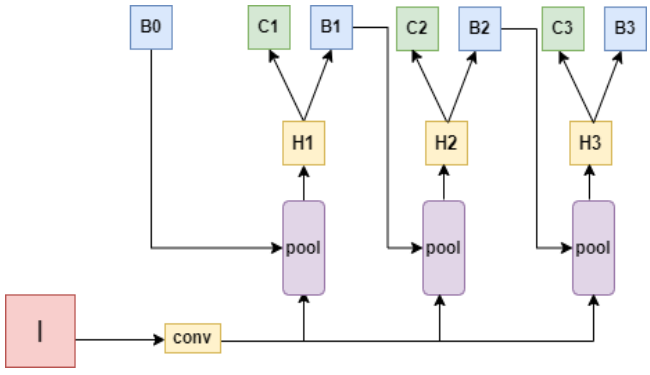


Fig. 4. Cascaded R-CNN network architecture.

range of 0.5 to 0.95 for small, medium, and large objects, respectively.

Table I highlights the public RMSE, private RMSE, AP, and Average Recall (AR) at various IoU values for the baseline and SAHI-enhanced proposed models.

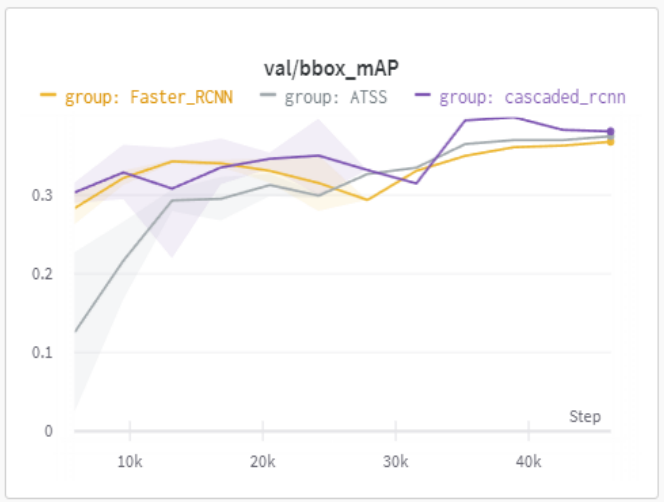
For evaluation purposes, a representative sample image from the test set was chosen. The image included tiny, medium, and large heads to highlight the significant effect of integrating SAHI into object detection models. Each input image has been divided into multiple overlapping slices of size 1024×1024 with overlap height ratio = 0.2 and overlap width ratio = 0.2. The size of the test image is 3753×2698, so it has been divided into 20 overlapping slices, as shown in Fig. 5.

Several values of the confidence level were investigated in the SAHI technique. Then the results were compared, as shown

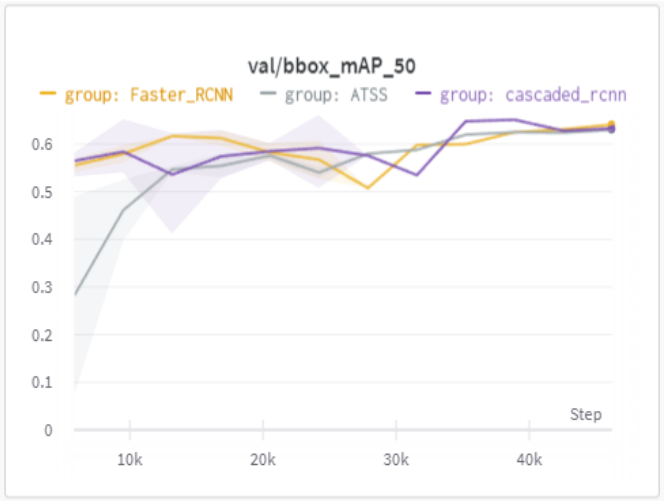
(a) Original image.

(b) Resulting overlapping patches.

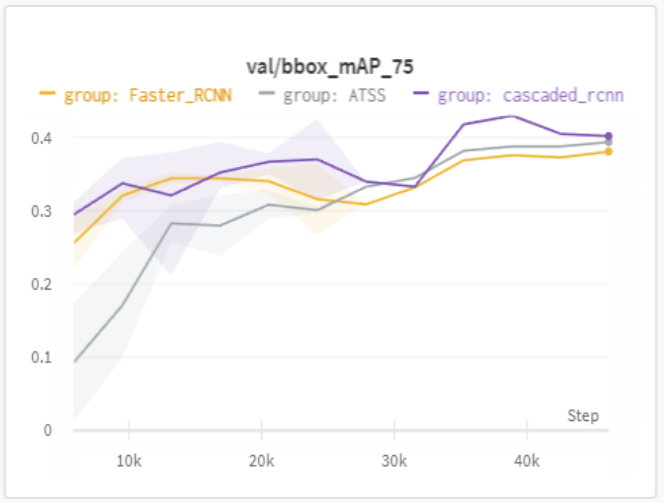
Fig. 5. Cutting the query image into 20 overlapping patches of size 1024×1024 for SAHI inference.



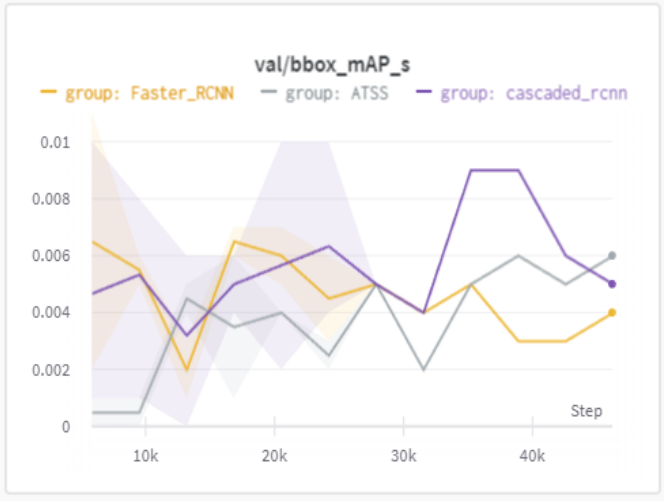
(a) mAP_0.5:0.95



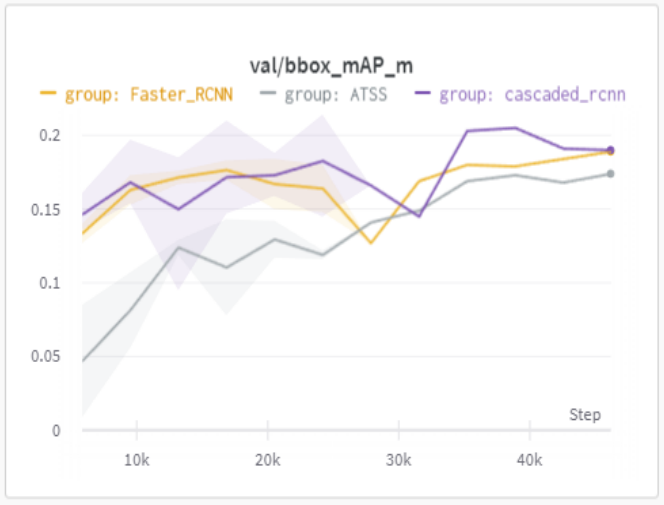
(b) mAP_0.5



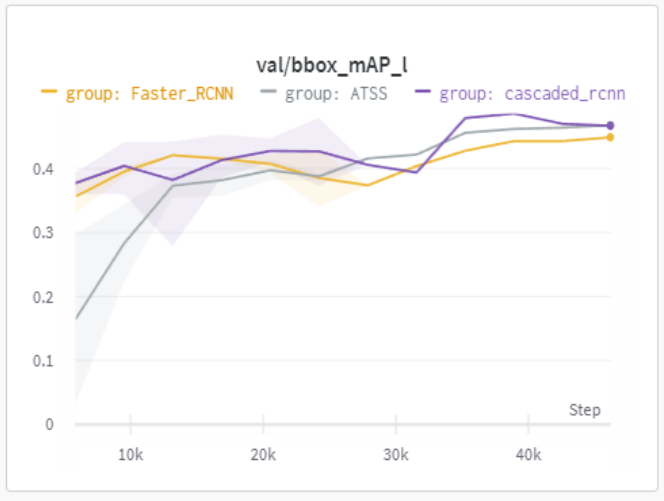
(c) mAP_0.75



(d) mAP_s



(e) mAP_m



(f) mAP_l

Fig. 6. Evaluation metrics for faster R-CNN, Cascaded R-CNN, and ATSS: (a) mAP_0.5:0.95, (b) mAP_0.5, (c) mAP_0.75, (d) mAP_s, (e) mAP_m, (f) mAP_l.

TABLE I. EVALUATION METRICS FOR FASTER R-CNN, CASCADED R-CNN, AND ATSS MODELS

Metric	Faster R-CNN	Cascaded R-CNN	ATSS
Inference RMSE (Private) [threshold=0.001]	6.9223	7.79984	56.63364
SAHI-based RMSE (Private) [threshold=0.8]	**5.31337**	5.57163	11.5534
Inference RMSE (Public) [threshold=0.001]	6.29219	6.6753	54.31293
SAHI-based RMSE (Public) [threshold=0.8]	3.80065	**3.47005**	13.36742
Average Precision (AP) @ [IoU=0.50:0.95 — area = all — maxDets = 100]	0.368	0.399	0.375
Average Precision (AP) @ [IoU=0.50 — area = all — maxDets = 1000]	0.641	0.651	0.630
Average Precision (AP) @ [IoU=0.75 — area = all — maxDets = 1000]	0.381	0.430	0.394
Average Precision (AP) @ [IoU=0.50:0.95 — area = small — maxDets = 1000]	0.004	0.009	0.006
Average Precision (AP) @ [IoU=0.50:0.95 — area = medium — maxDets = 1000]	0.189	0.205	0.174
Average Precision (AP) @ [IoU=0.50:0.95 — area = large — maxDets = 1000]	0.449	0.486	0.468
Average Recall (AR) @ [IoU=0.50:0.95 — area = all — maxDets = 100]	0.448	0.480	0.507
Average Recall (AR) @ [IoU=0.50:0.95 — area = all — maxDets = 300]	0.448	0.480	0.507
Average Recall (AR) @ [IoU=0.50:0.95 — area = all — maxDets = 1000]	0.448	0.480	0.507
Average Recall (AR) @ [IoU=0.50:0.95 — area = small — maxDets = 1000]	0.037	0.056	0.037
Average Recall (AR) @ [IoU=0.50:0.95 — area = medium — maxDets = 1000]	0.272	0.293	0.285
Average Recall (AR) @ [IoU=0.50:0.95 — area = large — maxDets = 1000]	0.532	0.569	0.613

TABLE II. SUMMARIZATION OF THE COMPARATIVE ANALYSIS PERFORMED TO FINE-TUNE THE CONFIDENCE LEVEL PARAMETER FOR SAHI INTEGRATED MODELS AND THE CORRESPONDING PUBLIC AND PRIVATE RMSE SCORES

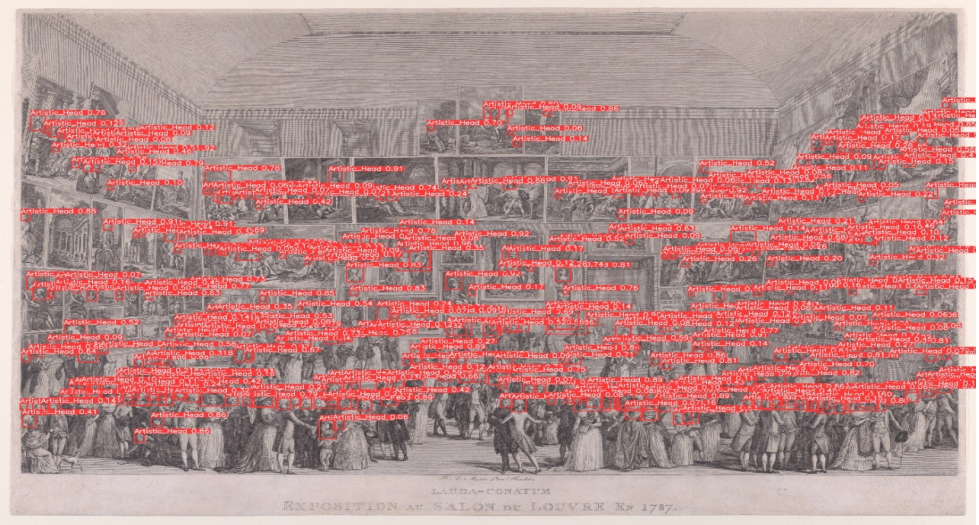
Model	Confidence Level	Number of Detected Heads	Public Score	Private Score
Faster R-CNN	0.001	449 heads	31.28875	27.17679
	0.4	280 heads	7.82784	11.58052
	0.8	131 heads	3.80065	5.31337
Cascaded R-CNN	0.001	443 heads	34.87448	29.35343
	0.4	275 heads	8.89154	12.03526
	0.8	128 heads	3.47005	5.57163
ATSS	0.001	530 heads	12.81895	10.7536
	0.4	100 heads	9.36027	8.09853
	0.8	0 heads	13.36742	11.5534

in Fig. 7. For a confidence level of 0.001, the SAHI-integrated Faster R-CNN discovered 449 heads, the majority of which were fewer than 50 pixels wide and tall, as required by the competition host. The model spotted 280 heads by gradually raising the confidence value to 0.4. When the confidence level is set to 0.8, the model performs best in terms of RMSE. It discovered 131 heads, the majority of which meet the annotation restrictions.

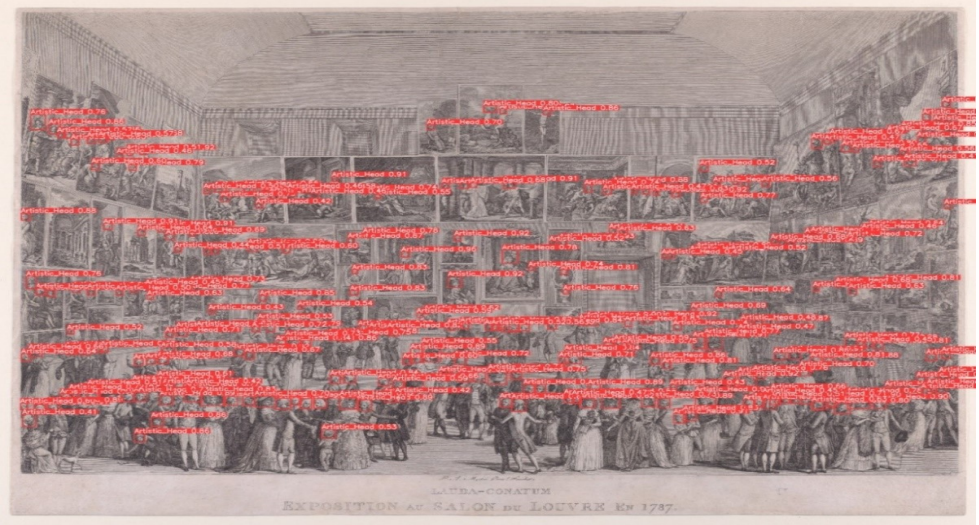
The same approach has been repeated for Cascaded R-CNN and ATSS, and results have been concluded in Table II. The ATSS model, unlike the Faster RCNN and Cascaded RCNN models, could not detect any heads at a confidence level of 0.8. On the contrary, when the confidence level was reduced to 0.4, its performance improved, and it could detect 100 heads. At a confidence level of 0.001, the lowest performance was obtained.

VI. CONCLUSION

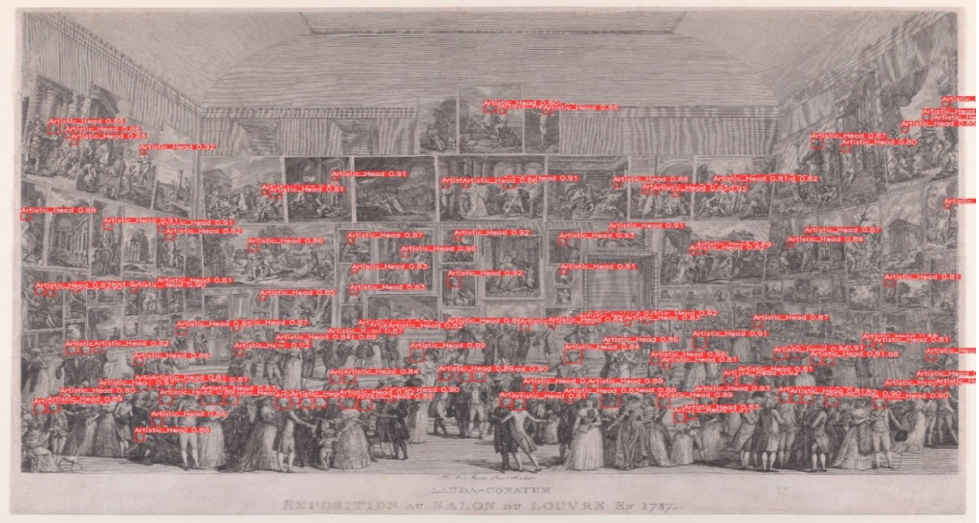
Although deep learning-based object detection architectures have achieved recent breakthroughs in various fields, they struggle to cope with detecting objects in art imagery such as paintings and sketches. In this study, the problem of artistic head detection in artworks was investigated. Three of the simplest and most widely used object detection models were utilized to detect and count heads in artworks instead of photos of natural persons. Finally, the models were extended to the SAHI framework to increase the model's detection performance in detecting small heads in large-size photos. The combined impact of sliced fine-tuning and sliced inference resulted in significant enhancements for all models. The result is a new route forward for training object detection models to interpret artworks. The next step will be to integrate SAHI with cutting-edge object detection models to enhance detection accuracy and release a mobile application specifically designed for museum environments, enabling widespread access to the SAHI model. This approach can be expanded beyond the recognition and detection of heads in artwork to other objects. Head detection in artworks has several real-time applications including virtual museum tours, augmented reality, audience analysis, security and surveillance, and gaming. It can be used to provide personalized content, track head movements, adjust performances, identify unauthorized individuals, and control character movements in games. In Cultural Studies and Anthropology, analyzing the number and characteristics of heads in works of art can contribute to the study of cultural practices, social structures, and historical contexts. It can help researchers gain a deeper understanding of societal norms, power dynamics, and cultural representations of different groups or communities.



(a)



(b)



(c)

Fig. 7. Detection results of Faster R-CNN: (a) at confidence level = 0.001, (b) at confidence level = 0.4, and (c) at confidence level = 0.8.

Developing algorithms and models for automatically identifying and counting heads in works of art can have practical applications in computer vision and artificial intelligence. It can contribute to the development of image recognition systems, object detection algorithms, and crowd analysis tools. These technologies can be used in various domains, such as surveillance, crowd management, and augmented reality. In addition, it enables efficient categorization, identification, and retrieval of artworks based on the number of figures or individuals depicted, facilitating research, exhibition planning, and educational initiatives. Finally, for Art history and analysis, identifying and counting heads in paintings can provide valuable insights into the composition, style, and thematic elements of artworks. It can aid art historians and analysts in understanding the artistic techniques used by the artist, the portrayal of human figures, and the narrative or symbolic significance of the depicted individuals.

ABBREVIATIONS

SAHI: Slicing Aided Hyper Inference
R-CNN: Region-Based Convolutional Neural Networks
ATSS: Adaptive Training Sample Selection
RMSE: Root Mean Square Error
DL: Deep learning
CNN: Convolutional Neural Networks
SVM: support vector machines
HMM: Hidden Markov Model
FCOS: Fully Convolutional One-Stage Object Detection
TOOD: Task-aligned One-stage Object Detection
YOLOX: Exceeding You Only Look Once
MFF: multi-scale fusion module
UAV: unmanned aerial vehicle
ASFF: Adaptively Spatial Feature Fusion
CBAM: Convolutional Block Attention Module
SOD: Small object detection
IMD-Net: interpretable multi-scale infrared small object detection network
AP: Average Precision
COCO: Common Objects in Context
IoU: Intersection over Union
mAP: mean average precision
P: Precision
R: Recall
AR: Average Recall
TP: True Positives
FP: False Positives
FN: False Negatives

REFERENCES

[1] B. Saleh and A. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," *arXiv preprint arXiv:1505.00855*, 2015.

[2] L. Bordoni and F. Mele, *Artificial intelligence for cultural heritage*. Cambridge Scholars Publishing, 2016.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[4] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[5] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[6] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9759–9768.

[7] Kaggle, "Artistic head detection," https://www.kaggle.com/competitions/artistic-head-detection, 2022, [Online; accessed January 31, 2025].

[8] S. Rapid, "Scale rapid," https://scale.com/rapid, 2023, [Online; accessed July 25, 2024].

[9] F. C. Akyon, S. O. Altinuc, and A. Temizel, "Slicing aided hyper inference and fine-tuning for small object detection," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 966–970.

[10] W. R. Tan, C. S. Chan, H. E. Aguirre, and K. Tanaka, "Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3703–3707.

[11] S. Smirnov and A. Eguizabal, "Deep learning for object detection in fine-art paintings," in *2018 Metrology for Archaeology and Cultural Heritage (MetroArchaeo)*. IEEE, 2018, pp. 45–49.

[12] E. J. Crowley and A. Zisserman, "In search of art," in *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13*. Springer, 2015, pp. 54–70.

[13] C. R. Johnson, E. Hendriks, I. J. Berezhnoy, E. Brevdo, S. M. Hughes, I. Daubechies, J. Li, E. Postma, and J. Z. Wang, "Image processing for artist identification," *IEEE Signal Processing Magazine*, vol. 25, no. 4, pp. 37–48, 2008.

[14] S. Lyu, D. Rockmore, and H. Farid, "A digital technique for art authentication," *Proceedings of the National Academy of Sciences*, vol. 101, no. 49, pp. 17 006–17 010, 2004.

[15] V. Pandey, K. Anand, A. Kalra, A. Gupta, P. P. Roy, and B.-G. Kim, "Enhancing object detection in aerial images," *Math. Biosci. Eng*, vol. 19, no. 8, pp. 7920–7932, 2022.

[16] Y. Pan, J. Yang, L. Zhu, L. Yao, and B. Zhang, "Aerial images object detection method based on cross-scale multi-feature fusion." *Mathematical Biosciences and Engineering: MBE*, vol. 20, no. 9, pp. 16 148–16 168, 2023.

[17] Z. Xu, J. Su, and K. Huang, "A-retinanet: A novel retinanet with an asymmetric attention fusion mechanism for dim and small drone detection in infrared images," *Mathematical Biosciences and Engineering*, vol. 20, no. 4, pp. 6630–6651, 2023.

[18] S. Shen, X. Zhang, W. Yan, S. Xie, B. Yu, and S. Wang, "An improved uav target detection algorithm based on asff-yolov5s," *Mathematical biosciences and engineering: MBE*, vol. 20, no. 6, pp. 10 773–10 789, 2023.

[19] Q. Feng, X. Xu, and Z. Wang, "Deep learning-based small object detection: A survey," *Mathematical Biosciences and Engineering*, vol. 20, no. 4, pp. 6551–6590, 2023.

[20] D. Li, S. Lin, X. Lu, X. Zhang, C. Cui, and B. Yang, "Imd-net: Interpretable multi-scale detection network for infrared dim and small objects," *Math. Biosci. Eng*, vol. 21, pp. 1712–1737, 2024.

[21] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[22] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "Tood: Task-aligned one-stage object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2021, pp. 3490–3499.

[23] H. Zhang, Y. Wang, F. Dayoub, and N. Sunderhauf, "Varifocalnet: An iou-aware dense object detector," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8514–8523.

[24] M. C. Keles, B. Salmanoglu, M. S. Guzel, B. Gursoy, and G. E. Bostanci, "Evaluation of yolo models with sliced inference for small object detection," *arXiv preprint arXiv:2203.04799*, 2022.