A Modular Deep Learning Framework for Scene Understanding in Augmented Reality Applications

1st Vladislav Li
Dept. of Networks and Digital Media
Kingston University
London, UK
v.li@kingston.ac.uk

2nd Barbara Villarini
School of Computer Science and Engineering
University of Westminster
London, UK
b.villarini@westminster.ac.uk

3rd Jean-Christophe Nebel
Dept. of Computer Science
Kingston University
London, UK
j.nebel@kingston.ac.uk

4th Argyriou Vasileios
Dept. of Networks and Digital Media
Kingston University
London, UK
vasileios.argyriou@kingston.ac.uk

***Abstract*—Taking as input natural images and videos augmented reality (AR) applications aim to enhance the real world with superimposed digital contents enabling interaction between the user and the environment. One important step in this process is automatic scene analysis and understanding that should be performed both in real time and with a good level of object recognition accuracy. In this work an end-to-end framework based on the combination of a Super Resolution network with a detection and recognition deep network has been proposed to increase performance and lower processing time. This novel approach has been evaluated on two different datasets: the popular COCO dataset whose real images are used for benchmarking many different computer vision tasks, and a generated dataset with synthetic images recreating a variety of environmental, lighting and acquisition conditions. The evaluation analysis is focused on small objects, which are more challenging to be correctly detected and recognised. The results show that the Average Precision is higher for smaller and low resolution objects for the proposed end-to-end approach in most of the selected conditions.**

***Index Terms*—Augmented Reality, Object Detection, Scene Analysis, Scene Understanding, Object Recognition, Deep Learning, Super-Resolution, Feature Extraction**

I. Introduction

Augmented Reality (AR) applications enable users to interact with their surrounding environment by overlying digital visuals on top of reality through the camera view. The aim is to achieve an enhancement of the real word through the combination of virtual information, such as text, images, video, or 3D models, with scenes captured by a camera in real time [30]. Furthermore, recent advances of computer system's capabilities, high-speed communication and computer vision technologies has boosted the demand of human-digital interaction through Mixed Reality (XR) headsets, and new three-dimensional interactive displays. The rapid development of AR technologies has fostered its application to different fields such as restoration, education, archaeology, art, tourism, commerce, and healthcare [33].

These immersive technologies rely on the analysis of the surrounding environment to extract content information. For instance, in the field of autonomous vehicles, scene analysis and understanding (e.g., vehicle detection, traffic signs and light recognition, and pedestrian detection) is a key component for decision making tasks and end-to-end control [24] so that the augmented environment can be seamlessly visualised on the car display. In the last decades the advances of computer vision have fostered the design and implementation of object recognition methods increasing computational performance and lowering process time [37]. As a result, current AR technologies based on object recognition use complex computer vision techniques to detect and track objects in the real world. Examples of such technologies include the You Only Look Once (YOLO) model [1], homomorphic filtering and Haar markers [12] and the Single Shot Detector [5]. The use of Convolutional Neural Networks (CNNs) and Deep Learning (DL) led to faster and more accurate detection processes [35]. However, they still deliver poor performance when camera resolution is low, or when the objects to recognise are very small or far away. Thus this can have an impact on the scene understanding and the overall AR experience.

The aim of this study is to provide a novel integrated end-to-end solution that improve performance in such conditions by introducing Super-Resolution (SR) mechanisms. Not only have Generative Adversarial Networks (GANs) been used for new data generation and study adversarial samples and attacks, but also in the recent past they have been investigated to perform SR tasks [15]. Inspired by this, the proposed approach is based on a cascade of two connected networks. The first network is a super resolution network that takes as input transformed images. More specifically, a 3D representation is used where the z-axis represents the colour channel of the image. The second network is based on the YOLO series' architecture that was designed to improve performance at a low computational cost. The key contributions of this work are a) the end-to-end design and training of the two connected networks allowing automatic minimisation of the SR reconstruction error and maximisation of the detection and classification accuracy with

a singe novel optimisation function, b) a complete comparative study under a variety of environmental conditions that are known to affect the overall performance of AR devices and c) a new dataset composed of synthetic objects created under different conditions, which allows unbiased performance evaluation under different sensor and environmental parameters. The paper is organised as follows: Section1 introduces the problem and relevant technologies, sections 2 provides an overview of related work, section 3 describes the proposed end-to-end architecture, section 4 presents results obtained using both a real image dataset (COCO) and a novel synthetic image dataset, and section 4 draws the final conclusions.

II. Overview of Previous Work

Augmented reality applications rely on machine learning and computer vision techniques to recognise the presence of physical objects in the real world so that virtual objects can be added and rendered in real time. In recent years, the use of Deep CNNs significantly improved performance and accuracy of computer vision for many tasks such as object detection and recognition. In 2014 Girishick et al. proposed the Regions with CNN features (RCNN) for object detection [11]. First, initial object candidate boxes are extracted by a selective search. Then, each box is rescaled to a fixed size image that is fed to a CNN model trained on an AlexNet [31] for feature extraction. Finally, object detection is performed using a linear SVM classifier. Although this approach led to significant improvement of the mean Average Precision when compared with previous approaches, it suffers from slow detection speed. To overcome this issue, He et.al. proposed the Spatial Pyramid Pooling Network (SPPNet) [16]. Its main novelty is a Spatial Pyramid Pooling (SPP) layer, which generates a fixed-length representation regardless of image size and scale allowing to feed images with varying sizes during the training process which improves scale-invariance and reduces overfitting. In the case of object detection, the feature maps are computed from the entire image only once, and then the features are aggregated in sub-regions to generate fixed-length vectors for training the detectors. Evaluation of this method showed that it could detect objects 24 to 102 times faster than RCNN. In 2015, Girshick proposed an improved version over the previous the RCNN architecture called Fast RCNN detector [10]. Although this network allows to train a detector and a bounding box regression simultaneously with the same network configuration, slow speed remained an issue. The same year, Ren et al. proposed the Faster RCNN detector [28], which is considered the first almost real-time deep learning detector using an end-to-end training This architecture introduces a Region Proposal Network (RPN) to speed up the detection process. Numerous variants of this approach have been suggested the following years to decrease any computational redundancy [20].

In particular, Cao et al. (2020) proposed a method called D2Det [3], which is based on the Faster R-CNN framework. Here the Region of Interest (ROI) features are processed through two different stages: a high-density local regression, which replaces the Faster RCNN offset regression, and a discriminant ROI pooling. In contrast to all the methods mentioned above, which are considered as two-stage detectors as they perform a coarse to fine process, in 2016 Joseph et al proposed a one-stage detector called You Only Look Once (YOLO) [26]. The image is divided into regions, and the network predicts bounding boxes for each region at the same time. With such approach the whole process is completed in one step applying a single network to the entire image increasing significantly processing speed. Although YOLO's second and third versions have improved its prediction accuracy [27], they still underperform in terms of localisation accuracy when compared with the two-stage methods. Liu et al. tried to improve this aspect proposing a Single Shot MultiBox Detector (SSD) [23] introducing a multi-reference and multi-resolution detection method detecting objects at different scales on different network layers. As a result, the SSD network gains small improvement outperforming YOLO in PASCAL VOC detection task [6]. Suggesting that extreme foreground-background class imbalance is the cause of the lower accuracy of the one-stage detectors, in 2018, Lin et al. introduced the RetinaNet [21] where a new loss function called "focal loss" was added to improved their approach. Indeed by modifying the standard cross entropy loss, the detector is more attentive on misclassified examples during training. A recent trend in object detection methods is anchor-free techniques where the methods infer the bounding box corners instead of fixed bounding boxes. A notable example is CenterNet proposed by Zhou et el [36]. The CenterNet method is a state-of-the-art Lidar-based 3D detection and tracking framework. It could be viewed as an improvement over CornerNet [18] which is an anchor-free approach to detect bounding box as a pair of keypoints. The keypoints are the top-left and the bottom-right corners retrieved via the corner pooling technique introduced by the same authors [18]. The CenterNet method has introduced a notion of a center keypoint to help associating the corner keypoints with an object in the image. The CenterNet method has outperformed common anchor-based solutions such as Faster RCNN and YOLO by a significant margin. In 2020, Perez-Rua et al [25] introduced OpeN-ended Center nEt (ONCE) which offered a functionality that can detect objects from classes with a small number of examples inside its training dataset. The more recent approaches start to investigate possibilities of transformers covered in the DEtection TRansformer (DETR) method [4] with an advantage being simple yet on par with the rest of the detection techniques used in the field. Later, Zhu et al proposed Deformable DETR as an improvement to address the performance in detecting small objects achieving the state-of-the-art performance.

In parallel, approaches have been developed to enhance the detection of small objects, which is particularly challenging as they have fewer visible details. Super-resolution solutions relying on Generative Adversarial Networks (GAN) [13] have proved particularly successful [19]. Indeed, their competitive process involving two neural networks, i.e., a generator net-

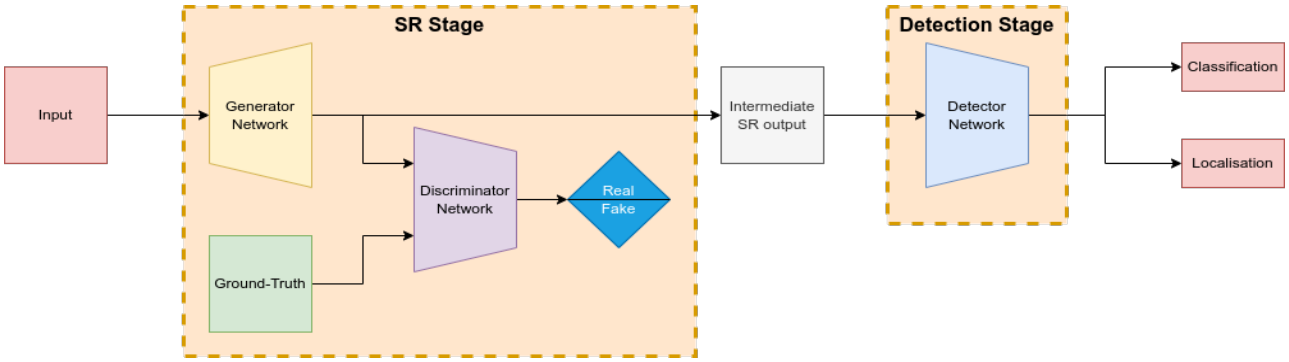


Fig. 1: Overview of the proposed novel framework trained end-to-end. For the SR and detector models any state-of-the-art solutions can be used without affecting the overall pipeline and the proposed modular architecture.

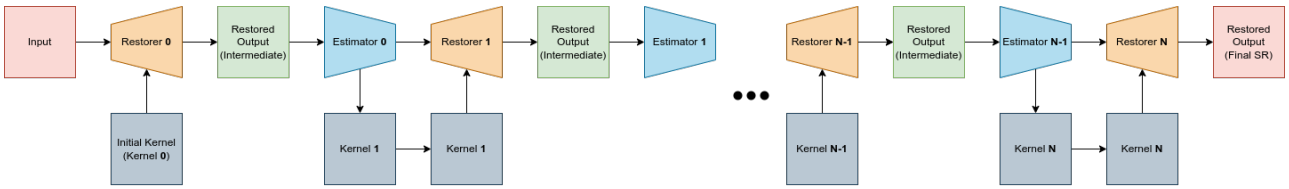


Fig. 2: Training setup for the DAT SR deep network.

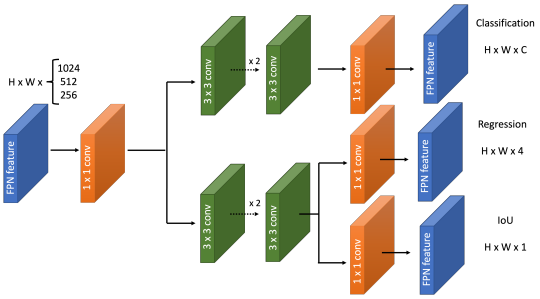


Fig. 3: YOLOX architecture relying on a decoupled head

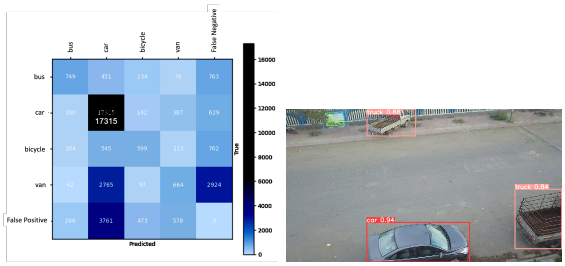


Fig. 4: Confusion Matrix (the white colour of the image was levelled up to see the numbers)

work and a discriminant network, ensures that the generated images are as realistic as possible.

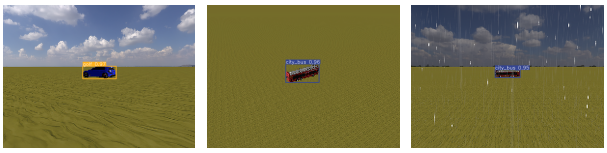
Herein, we address the detection and recognition of visually small objects by proposing a novel approach which is based on a super-resolution network and a second network with a modified YOLO architecture trained end-to-end. This solution aims to provide accurate detection of objects that are very small or very far from the camera sensor of the AR glasses while delivering a fast processing time enabling its usage in real time applications.

III. Proposed Framework

In this paper we propose an end-to-end framework for scene understanding that combines super-resolution, and object detection and classification architectures. Figure 1 shows an overview of the proposed methodology where the two main components take as input an image (or a video) and are trained in an end-to-end manner. Details of these two processing blocks are described in the following sections.

A. Super-Resolution Method

As reviewed in the Section 2, usage of super-resolution in a pre-processing stage has been included in many computer vision pipelines. Typically, the super-resolution model is trained in an unsupervised manner using several independent datasets, while the target classification or detection model is trained only on a single task related dataset. By doing so, some extra information, which is not available in the target labelled dataset, is injected in the SR images [17]. SR models are trained under the assumption that the low-scale images, passed as input, are the results of some low-pass filter, such as a Gaussian blur or a point spread function. The training takes place by first down-sampling high resolution images with such kernel and second optimising the model to reconstruct these high-resolution images. Theoretically, the



(a) Ground - Camera Subcategory - Golf (97%) (b) Ground - Light Subcategory - City Bus (96%) (c) Ground - Weather Subcategory - City Bus (95%)

Fig. 5: Examples of the generated synthetic data. The bottom row represents examples from the Ground category. From the left, the first column shows the Camera sub-category, the second column shows the Light sub-category, the third columns displays the Weather sub-category.

kernel function should match the actual blurring process caused by the camera used in the targeted application. As it is usually unknown, 'standard' kernels have been used. However, as they fail to model the specific optics and sensors of the actual cameras that captured the images of interest, this leads to degraded performance in real-world scenarios. To address this, methods have been proposed to learn the blur kernel. The most accurate approaches, such as the state-of-the-art network Deep Alternating Network (DAT) [17], have relied on deep learning architectures. DAT was selected for our pipeline because its learning is unsupervised and it delivers fast computation, which makes it suitable for mobile devices and low specification desktop computers. Indeed, the authors showed that the average speed is 0.75 seconds per image, more than 500 times faster than its competitors KernelGAN [2], ZSSR [29] and 5 times faster than IKC [14]. The mentioned average speeds are considered to be fast in the domain of SR. Figure 2 shows DAT's training setup. It is composed of two main networks called Restorer and Estimator: the Restorer produces the SR image, while the Estimator provides an estimate of a blur kernel given the restored image. The two networks are used in alternation improving at each Restorer-Estimator step the quality of the SR image and the accuracy of the estimated kernel. The sequence of Restorer-Estimator is optimised end-to-end using a stochastic back-propagation algorithm.

B. Object Detection Method

For augmented reality applications, object detection models must deliver high accuracy in real-time. For the proposed framework, the anchor-free model, YOLOX [8], offers the best compromise. Indeed, its simple, powerful and computationally efficient architecture was built upon one of the most used detectors in the industry, YOLOv3 [27], which, in addition to have a limited computational cost, has received excellent software support. However, an important improvement of YOLOX is that, unlike the previous architectures of the YOLO series, it uses a decoupled head which improves convergence speed. Figure 3 provides an overview of the architecture of YOLOX. Following a 1 x 1 convolutional layer used to decrease the number of channels, there are two parallel branches with 3 x 3 convolutional layers. Moreover, compared to the baseline YOLOv3, an Intersection over Union (IoU) aware branch is added in the regression branch.

Another enhancement of YOLOX is, unlike the past versions of YOLO detectors (except for YOLOv1), usage of an anchor-free model. Anchors are candidates bounding boxes with pre-defined dimensions that the detector selects during the detection process and for which it predicts the delta values for their centres and dimensions. Obviously, these additional predictions require extra processing during both the training and inference stages, which impacts the overall computational time. On the other hand, when using an anchor free approach, bounding boxes are predicted directly, which reduces the number of design parameters. As such approach requires advanced data augmentation to match the performance of anchor-based models, state-of-the-art data augmentation approaches, i.e., Mosaic and MixUp, were exploited [8] Indeed, they are known to bring stability and reduce overfitting during the training process. Finally, it is important to specify that YOLOX leverages a high-performance CNN front-end CSPNet [32],which is followed by a feature pyramids network (FPN) [27].

C. End-to-end Framework

The methods described in sub-sections B and C were integrated into an end-to-end framework. Thus, the framework comprises two main components, i.e., Super-Resolution and Detector. Equation (1) illustrates the proposed end-to-end architecture where x is the input low-resolution image, y is the image generated by the super resolution function $S(\cdot)$, and z is the output of the detection function $D(\cdot)$.

$$\begin{cases} \text{y=S(x)} \\ \text{z=D(y)} \end{cases} \rightarrow z = D(S(x)) \tag{1}$$

In this framework, an input image is, first, handled by the SR component which produces a super-resolved output image. Second, this image is passed to the detector component which recognises and locates objects. Through this process, the detector learns from images enhanced by the SR component. The input images are super-resolved using kernels. There are many types of kernels such as common bicubic kernel or linear kernel. They are well studied and don't require an AI network to calculate them. However, in regard to SR task, the real world images don't have information about the kernel therefore it couldn't successfully be restored. Consequently, an estimator is used to infer the kernel during the training process. Then, it is passed to the restorer to generate images. As a result, the restored images contain features which are the product of the kernel. These features could be picked up by the detector during the training process creating a symbiotic relationship between the SR and detector components leading to an improved performance. To monitor and evaluate the training of the framework, several state-of-the-art loss functions were selected. The detector is trained using Varifocal Loss [34] as classification loss function and SIoU [9] (Scylla

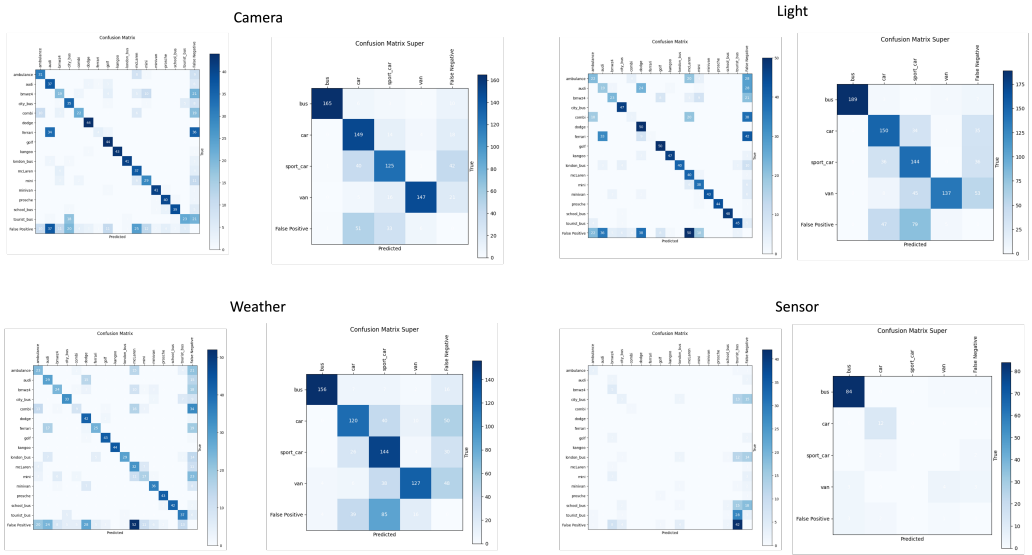


Fig. 6: Confusion Matrices for the four categories

Intersection over Union) as box regression loss function. Moreover, the training process was facilitated with SimOTA - a Simplification of OTA [7] (Optimal Transport Assignment) - for dynamic label assignment [8]. The Varifocal Loss is particularly efficient because it considers both classification and localisation scores when ranking candidates using IoU. Similarly, the SIoU loss function addresses direction mismatch between expected and predicted bounding boxes by exploiting angle, distance, shape, and IoU costs. Finally, the value of SimOTA is to view the task of bounding box assignment as an optimal transport problem where the unit transportation cost between anchor-point and ground truth is expressed as a weighted sum of their classification and regression losses to find the best assignment solution

D. Parameters

The end-to-end framework was fine-tuned by running 10 epochs with a batch size of 3 on both real and synthetic data under four different categories. While the learning rate was set to 0.0001 for the SR component, it was set to 0.0032 with SGD (Stochastic Gradient Descent) optimisation for the detector. Additionally, as mentioned earlier, training was enhanced using Mosaic and MixUp as data augmentation strategies.

IV. Evaluation

The proposed method has been applied to object recognition and scene understanding. Its evaluation was performed using the Common Objects in Context (COCO) dataset [22] and a synthetic dataset where different environmental conditions were applied to affect image quality. COCO is widely used to benchmark computer vision models. It consists of 330K images, with more than 200K labelled images, 1.5 million object instances, 80 object categories, 91 stuff categories,

TABLE I: Model Performance on the COCO Dataset in terms of mAP (%)

RetinaNet	YOLOv3	Faster R-CNN	Proposed
52.61	44.76	40.50	**67.09**

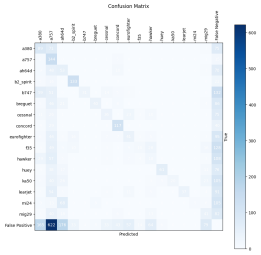
and 5 captions per image. Comparisons with state-of-the-art methods relies on the mean Average Precision (mAP), a standard metric introduced in 2014 to quantify object detection performance based on a user-defined set of criteria [22]. It is defined as the mean value of the average precision of the individual classes:

$$mAP = \frac{1}{n}\sum_{k=1}^{n} AP_k \tag{2}$$

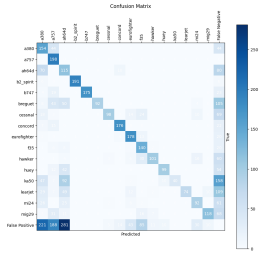
where AP_k is the Average Precision of class k and n is the number of classes.

In this evaluation process, using the COCO dataset, Table I shows performance in terms of mAP of the proposed framework against that of other approaches presented in the literature review. Our framework outperforms significantly all its competitors. Moreover, the added value of the super-resolution component is clearly established as it exceeds YOLO's mAP by over 20%. The confusion matrix in Figure 4 further demonstrates the performance of the model. In particular, it predicts with high accuracy objects belonging to the prevailing "car" category. However, one should highlight that the "van" category is often mistaken for the "car" category, which is due to the visual similarity between images of these two classes.

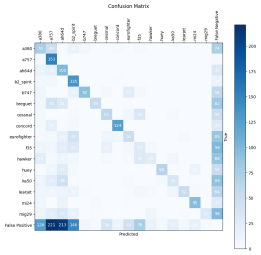
Further evaluation has been carried out using a synthetic dataset that we created using a 3D Rendering Engine. This dataset consists of approximately 3000 low-resolution images



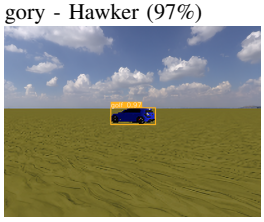
(a) Air - Camera Subcategory - Hawker (97%)



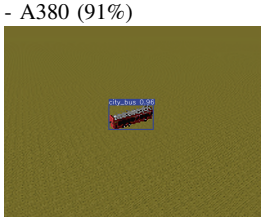
(b) Air - Light Subcategory - A380 (91%)



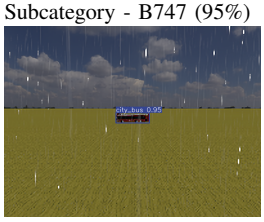
(c) Air Category - Weather Subcategory - B747 (95%)



(d) Ground - Camera Subcategory - Golf (97%)

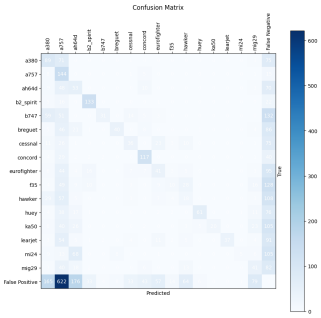


(e) Ground - Light Subcategory - City Bus (96%)

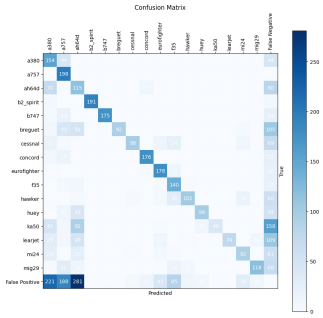


(f) Ground - Weather Subcategory - City Bus (95%)

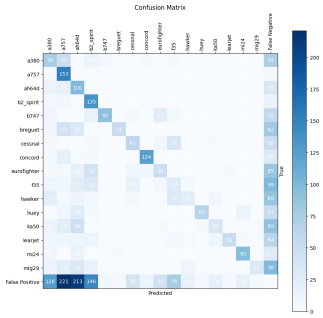
Fig. 7: Examples of the generated synthetic data. The bottom row represents examples from the Ground category. From the left, the first column shows the Camera sub-category, the second column shows the Light sub-category, the third columns displays the Weather sub-category.



(a) Air - Camera Subcategory - Hawker (97%)



(b) Air - Light Subcategory - A380 (91%)



(c) Air - Weather Subcategory - B747 (95%)

Fig. 8: Examples of the generated synthetic data. The bottom row represents examples from the Ground category. From the left, the first column shows the Camera sub-category, the second column shows the Light sub-category, the third columns displays the Weather sub-category.

per category of aerial and ground vehicles in different environments and weather conditions. Its images belong to four different categories, each allowing to assess our model on specific properties: a) Camera, b) Light, c) Weather, and d) Sensor. The "Camera" category contains images of objects seen from different camera angles and distances. In the "Light" category, images are generated under a variety of lighting parameters mimicking different parts of a day such as morning, afternoon, evening, and night. The "Weather" category simulates images captured under diverse weather circumstances, including varying rain and wind conditions. Finally, the "Sensor" category mimics images collected by standard, night vision and thermal cameras. The performance of the proposed framework in terms of mAP is shown in Table II for these three categories.

TABLE II: Model Performance on the Synthetic Dataset according to the Four Image Categories

Category	mAP(%)
Camera	80.14
Light	77.82
Weather	76.36
Sensor	23.72

V. Conclusion

The work presented in this paper offers an end-to-end solution for object detection and recognition on AR devises. The modular architecture allows the integration of different SR

and detection models under the same pipeline. An overview of existing solutions and approaches is provided both for super resolution and scene analysis methods with applications in immersive applications. The proposed architecture was tested both in real and synthetic datasets in a comparative study including other state of the art approaches. The obtained results demonstrate a significant improvement especially for low resolution or distant objects. Also, the proposed framework was tested and analysed for different environmental conditions and a variety of camera sensors. Additionally, a new balanced synthetic dataset was produced with annotated data covering multiple objects and environments.

REFERENCES

[1] Ryan Anderson, Juan Toledo, and Hala ElAarag. Feasibility study on the utilization of microsoft hololens to increase driving conditions awareness. In *2019 SoutheastCon*, pages 1–8. IEEE, 2019.

[2] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *Advances in NIPS*, 32, 2019.

[3] Jiale Cao, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. D2det: Towards high quality object detection and instance segmentation. In *Proceedings of the IEEE/CVF CVPR*, pages 11485–11494, 2020.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[5] Nikos Dimitropoulos, Theodoros Togias, George Michalos, and Sotiris Makris. Operator support in human–robot collaborative environments using ai enhanced wearable devices. *Procedia Cirp*, 97:464–469, 2021.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012).

[7] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *IEEE CVPR*, pages 303–312, 2021.

[8] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[9] Zhora Gevorgyan. Siou loss: More powerful learning for bounding box regression. *arXiv preprint arXiv:2205.12740*, 2022.

[10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE CVPR*, pages 580–587, 2014.

[12] Daniel Lima Gomes Jr, Anselmo Cardoso de Paiva, Aristófanes Corrêa Silva, Geraldo Braz Jr, João Dallyson Sousa de Almeida, Antônio Sérgio de Araújo, and Marcelo Gattas. Augmented visualization using homomorphic filtering and haar-based natural markers for power systems substations. *Computers in Industry*, 97:67–75, 2018.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[14] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019.

[15] Rohit Gupta, Anurag Sharma, and Anupam Kumar. Super-resolution using gans for medical imaging. *Procedia Computer Science*, 173:28–35, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans PAMI*, 37(9):1904–1916, 2015.

[17] Yan Huang, Shang Li, Liang Wang, Tieniu Tan, et al. Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems*, 33:5632–5643, 2020.

[18] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, pages 734–750, 2018.

[19] Vladislav Li, George Amponis, Jean-Christophe Nebel, Vasileios Argyriou, Thomas Lagkas, Savvas Ouzounidis, and Panagiotis Sarigiannidis. Super resolution for augmented reality applications. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6. IEEE, 2022.

[20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE CVPR*, pages 2117–2125, 2017.

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE ICCV*, pages 2980–2988, 2017.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[23] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[24] Monirul Islam Pavel, Siok Yee Tan, and Azizi Abdullah. Vision-based autonomous vehicle systems based on deep learning: A systematic literature review. *Applied Sciences*, 12(14):6831, 2022.

[25] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M. Hospedales, and Tao Xiang. Incremental few-shot object detection. In *IEEE CVPR*, 2020.

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE CVPR*, pages 779–788, 2016.

[27] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[29] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE CVPR*, pages 3118–3126, 2018.

[30] Hu Tianyu, Zhang Quanfu, and Dong Huiyuan ShenYongjie. Overview of augmented reality technology. *Computer Knowledge and Technology*, 34:194–196, 2017.

[31] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE CVPR*, volume 1, pages I–I, 2001.

[32] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on CVPRW*, pages 390–391, 2020.

[33] Jianghao Xiong, En-Lin Hsiang, Ziqian He, Tao Zhan, and Shin-Tson Wu. Augmented reality and virtual reality displays: emerging technologies and future perspectives. *Light: Science & Applications*, 10(1):216, 2021.

[34] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. Varifocalnet: An iou-aware dense object detector. In *IEEE/CVF CVPR*, pages 8514–8523, 2021.

[35] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[36] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[37] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023.