GAN Saliency Detection, Depth Based Saliency Comparison

Hamish Sams, *Member, IEEE,* Charith-Abhayaratne, *Fellow, HEA,*

***Abstract*—In this paper, we propose a novel method of combining RGB-D data for saliency prediction based on generative adversarial networks (GANs). Our method allows the same generator neural network to combine depth and RGB data before saliency prediction compared to previous state of the art systems based on depth and RGB combination post saliency prediction. We also attempt to qualify the added benefit of depth in saliency prediction. Our neural network is trained and tested on the Olesova eye tracking dataset and compared with classic and state of the art saliency prediction models along with the NLPR dataset.**

***Index Terms*—Generative Adversarial Networks,/ Saliency detection, RGB-Depth.**

I. Introduction

SALIENCY prediction algorithms attempt to quantify how salient or eye-catching an image or video is similar to how a human would [1]. Saliency is, therefore, a large role both in the computing field but also the psychological and biomedical world [2]. More recently the saliency field has expanded and focused on detecting objects compared to processing large busy scenes [3].

Saliency has always been a topic for research, but as technology has improved, researchers have been able to implement such systems leading to more discoveries historically and recently such as overt and covert attention along with the respective computing methods top-down and bottom-up processing [4]. Overt (top-down) meaning that the eye is selectively chosen to look at an object such as while driving a traffic light would be overtly seen. On the other hand, covert (bottom-up) vision is what is passively eye-catching such as being the passenger in a car traffic lights may not be stimulating but a bright green building may be. Recently, Microsoft released the Xbox Kinect which has allowed researchers access to 3D data for RGB-D saliency [5]. As humans, we see can sense depth due to having two eyes and with this 3D data, we can start to create algorithms that act more human-like [6].

As this technology is growing so are the applications. Saliency models are being used in many image processing tools such as: automatic cropping [7] [8], image thumbnails [9] and image summarizing [10]. On top of image processing tools, saliency is being used to improve computer vision software to improve CNN (Convolutional Neural Networks) speed and accuracy [11], [12], [13].

Saliency as a whole has unlimited applications in computing, as models improve these systems will likely move into our everyday lives. Some likely situations could be inspection robots or cameras designed to detect unnatural items on streets or in buildings. These systems could also be reverse engineered to design naturally eye-catching advertisements or products. These could also be used in health-care to diagnose patients in collaboration with eye-tracking systems to sense unnatural vision [14] [15].

What is holding back these technologies is the huge amounts of data to be processed (around $10^8 - 10^9$ bits per second for the brain [16]). This means neural networks struggle with our current computing power relying on convolution to cut down data [17]. A common problem for quantifying model efficiency is that a lack of challenging datasets exist, generally proposing of a single salient object near the centre of an image with the most common being [18] and [3].

With all of this in mind this paper has the following aims:

1) Propose a new saliency prediction algorithm using generative adversarial networks trained on RGB-D eye-tracking data comparing the effect of training with both RGB and depth available simultaneously.
2) Compare different saliency models using RGB, Depth and RGB-D models to qualify the effect of depth on saliency prediction.

A. Economic, Legal, Social, Ethical and Environmental Context

As this system could be developed and implemented in such a huge amount of areas it is hard to pinpoint exact worries. Saliency is already being developed in AI systems and with technology increasing so rapidly, saliency prediction algorithms could phase out low skilled human jobs that rely solely on vision. As any system that would run a saliency prediction algorithm must use electricity it is, therefore, contributing to global warming given the current methods of generation, however, these standalone systems could be completely green if used in conjunction with solar or wind technology. Currently, this technology does not have the ability to take over from design jobs, but is used to speed up usually time-consuming tasks such as cropping [7] and image summarising [10]. Overall I do not believe developing these systems for generally aiding time-consuming tasks is harmful but implementing these systems in industry where jobs may be taken or in throwaway technologies is not only harmful to

H. Sams is from the Department of Electronic and Electrical Engineering, University of Sheffield, South Yorkshire, UK (e-mail: hcsams1@sheffield.ac.uk).

C.Abhayaratne is from the Department of Electronic and Electrical Engineering, University of Sheffield, South Yorkshire, UK (e-mail: c.abhayaratne@sheffield.ac.uk).

Manuscript received May 3, 2019.

the environment and families but also the local economy by reducing cash flow.

II. Related Work

A. Saliency Models

Saliency prediction algorithms can generally be broken down into two sections: Bottom-up and top-down methods [4].

1) bottom up: Bottom-up approaches are based around covert sight and are based around how we naturally look at scenes. The basis of the majority of bottom-up saliency models are three feature maps: intensity, colour and orientation [19] [20] [21] [22] [23]. Derived from psychological behavioural analysis [24] [2]. Bottom-up approaches are generally good at detecting salient points (high precision) but lead to blurry maps with low recall (high false positive rate) [22]. Bottom-up is comprised of two more categories: local and global methods.

- Local methodologies determine region wise saliency based on surrounding region contrast $i.e.$ the edge of an object. Regions may be classed as pixels or features or convolved maps etc. [19] [3] [25]. Local methodologies are still being developed as state of the art [26].
- Global methodologies determine contrast based on difference compared to the rest of the image such as a bright red post box in a town. Some methods include colour histograms to compare contrast [22].

2) top-down: Top-down approaches are designed for overt sight and require pre-determined knowledge for the application such as driving. Top-down methodologies depend on objectness, object proposal [27] [28] and object detection to detect specific items and quantify objects [29] [30] [31]. A top-down methods have been trained for many different applications such as facial recognition [32] and people/car recognition [33].

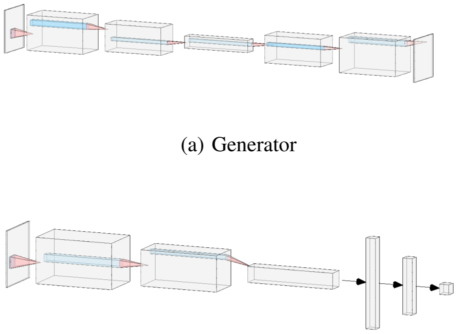
B. Convolutional Neural Networks (CNN)

Convolutional neural networks were first used to classify handwriting [34]. CNNs have been used for a variety of tasks especially image classification [17] and object detection [35]. CNNs are currently outperforming classical models [36] [37] [38] [39] with most being trained on the ImageNet dataset [17] which has been proposed to be applicable to be useful for generic tasks [40].

C. Depth combination models

All current RGB-D saliency methods work by computing RGB saliency and Depth saliency separately and then combining the two, this is split into three methods: Depth-weighting, Depth-Pooling and Learning-Based. Our method proposes a combination method determined by the saliency neural network.

1) Depth-Weighting: Depth-weighted models use feature map fusion to improve the accuracy of saliency prediction [41]. Working by taking into account depth while calculating RGB saliency. The difference between 2D and 3D data has been analysed to determine the saliency improvement using depth data and has been shown to increase saliency by around 6-7% [6]. Methods of weighing depth into saliency models are still being developed to improve accuracy [42].



(a) Generator

(b) Discriminator

Fig. 1. Simplified generator and discriminator neural networks showing convolutional compression, expansion and classification respectively. The generator does not show the U-Net style encoder-decoder with skipped connections [52].

2) Depth-Pooling/Saliency: Depth pooling models work by calculating and combining RGB and depth salience maps together to make a joint RGBD map [43]. One of the most straightforward methods is by multiplying depth saliency maps with RGB saliency maps pixel-wise [44] [45]. Different methods of integrating depth have been proposed based off depth saliency maps taught from synthetic stimuli [46] [47]. Basic RGB-D fusion frameworks have been developed and are widely used in models [48].

3) Learning-Based/mono-vision: Learning based methods are saliency prediction algorithms that calculate the depth map using machine learning from an RGB image. These methods allow for depth saliency without needing stereo-vision equipment or data [49]. Support vector machines (SVM) have been successfully used to classify RGB data into depth maps [50].

III. Proposed approach

In this paper, we propose a new saliency model using RGB and Depth data using a generative adversarial network trained on eye-tracking data in a busy environment using an Xbox Kinect depth camera [51].

A. GAN

The GAN used was that of P. Isola et al [53]. The generator that reads the RGB and depth inputs is a U-Net style network proposed by O. Ronneberger et al where certain feature maps skip some convolutional steps and are passed through the system [52]. A simplified version of the GAN can be seen in Figure: 1a. The Discriminator that classifies the real and fake data is based on the work in C. Li et al [54], a simplified version of this CNN can be seen in Figure: 1b. The GAN was trained on half of the dataset for 100 epochs to train the generator and discriminator. The larger the epochs the more likely the model is to succeed but takes much longer to run, 100 was chosen due to a tradeoff of these two and pushed time limits on my hardware.

B. Testing

The second half of the dataset is used to test the GANs ability of eye-tracking prediction, which makes sure the CNN hasn't simply memorised the dataset as it has never seen these entries. This data is also used in conjunction with a mix of classical and state of the art methods to classify the efficiency [55] [19] [56] [57]. As this dataset provides depth as a grayscale map instead of a matrix the depth data had to be transformed into distances for other models. This was done by calculating the max length of the room shown in the dataset ($8.99m$) and finding the grayscale value at that point (255). This lead to the equation $x * 35.25 = y$ to calculate the distance in mm. The common dataset proposed by H. Peng et al was also used to compare the model's ability to detect salient objects [18]. Multiple methods of classification have been used, the main two being F-Measure and the receiver operating characteristic curve (ROC). The F-Measure is defined as the harmonic mean of precision and recall (where precision and recall are defined in Equation 1 and 2 respectively) as shown in Equation: 3. The receiver operating characteristic (ROC) curve is created by varying the cutoff threshold of the predicted saliency map and comparing with the ground truth map and then plotting the true positive rate vs false positive rate. With a large dataset, all ROC curves are calculated and averaged to get the plotted ROC curves shown.

$$P = \frac{Truepositives}{Truepositives + Falsepositives} \tag{1}$$

$$F = \frac{Truepositives}{Truepositives + Falsenegatives} \tag{2}$$

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

IV. Results & Evaluation

The two main objectives of this paper are to evaluate the performance of our model compared to other state of the art and classical models, but also to research the effect of depth data in saliency prediction and the effect of training a neural network with both RGB and Depth compared with combining RGB and Depth maps post detection.

A. Saliency model comparison

As we have no specific goal (i.e. high precision or recall exclusively) the main measure of an algorithms effectiveness is represented by the F-Measure. All results can be seen in Table: VI and Table: V for the Olesova and NLPR datasets respectively with the best results (based on F-Measure) for each model shown in Table: I and Table: II.

1) Olesova: The Olesova dataset contained over 1000 images from a RGB-D video with corresponding eye-tracking data from participants; this means salient information was directly taken from eye movement and couldn't be easily forged. From Table: I we can see that our model is much higher than the rest in all measures. The precision of our model is around 70% which is higher than the rest but not by a huge amount, as precision is the measure of how much selected is relevant this shows the majority of what our system has defined as salient is correct. However when we look at recall, our system has by far taken the lead, this means our system has selected the majority of what is defined as salient compared to the other models. This means the other models are looking more for small and specific salient areas compared to our model which rates every area on its saliency. The other models have a much higher precision than recall meaning the chosen salient region is generally correct but the majority of the rest of the frame is missed. This is also shown in Figure 3 where all other models seem to plateau at around 30% false positive rate with a true positive rate of 70% whereas ours is at around 95%. Overall it seems fair to say our model has beaten state of the art systems with this style of data with an increase of 21% in AUC and 129% in F-Measure.

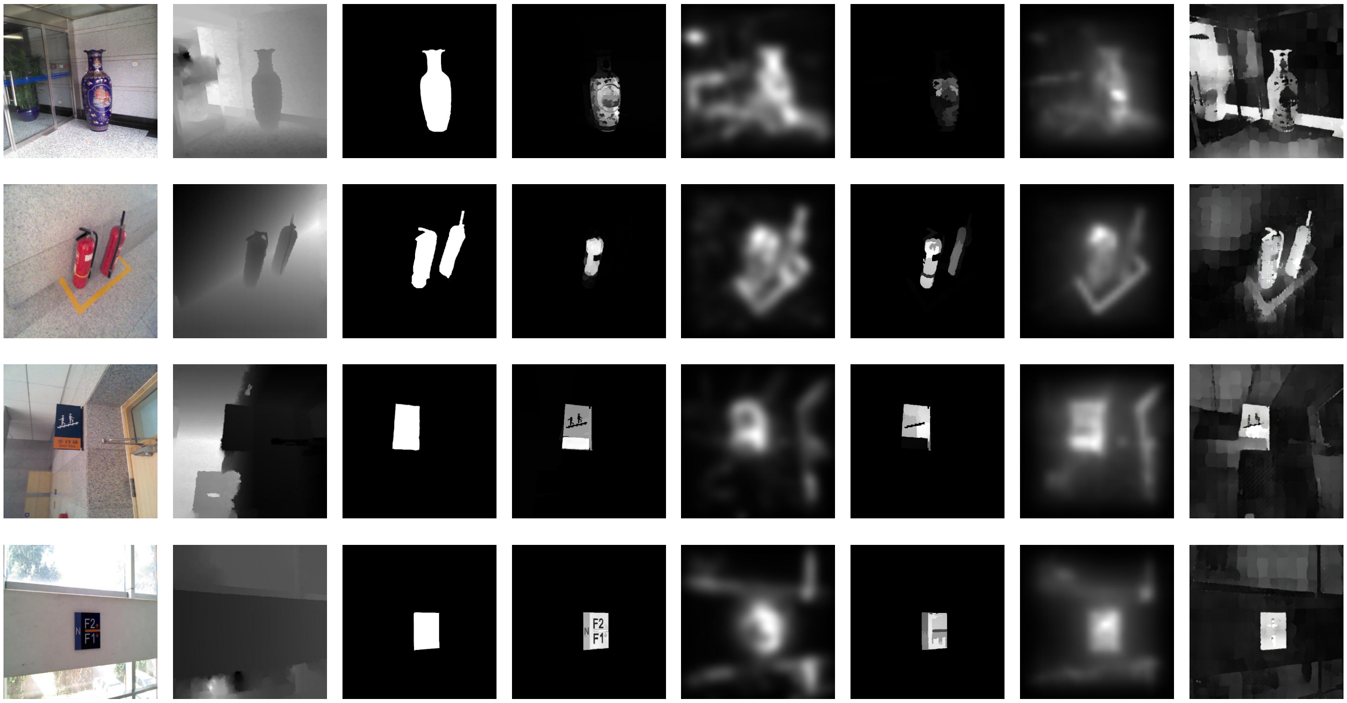
TABLE I
Best Olesova results

Model	F - Measure	Precision	Recall	AUC
Ours(RGB)	0.7248	0.6928	0.7598	0.9227
GBVS(RGB)	0.3163	0.4353	0.2484	0.7052
Itti(RGB)	0.3097	0.4591	0.2336	0.7380
LMH	0.1878	0.6529	0.1096	0.7646

2) NLPR: Unlike the Olesova dataset NLPR [18] contains 1000 images (RGB and depth) with ground truth maps composed of singular salient objects. These points were calculated by humans selecting objects by hand. NLPR is also based on singular images and not constant video compared to Olesova. If we look at Table :II we can see that ours is now around equal to Itti and LMH methods in terms of F-Measure. Looking at the precision and recall we can see that this time our precision is the lowest of all methods meaning that our system has predicted that much more of the image is salient than the ground truth map shows, this is likely due to the fact that our method rates the entire scene and not a single object. This is reinforced by looking at the recall as our system has the highest recall of all, which indicates ours had correctly identified the most relevant elements. We can generally tell that our system has identified much more of the scene as being salient which agrees with all other information. The graph in Figure: 4 shows all models ROC, we can see that most systems are similar in performance, but ours is lowest with a low false positive rate pointing to a system that fails to identify the majority of the main salient point, the true positive rate then rises quickly above other models. This means that our system identifies the whole salient object fully but classes it as a less salient point. Looking at Figure: 2 our method stands out as classifying more of the scene as salient compared to all other methods especially GP and LMH. These images re-enforce the previously made assumptions of over-classifying and others under-classifying (especially LMH).

B. Depth comparison

The Olesova results can be used as a method of comparing how well our neural network has learned the saliency from the eye-tracking data. A table of the combination methods can



(a) RGB (b) Depth (c) Ground (d) GP (e) Itti (f) LMH (g) GBVS (h) Ours

Fig. 2. A comparison of salience prediction algorithms on the NLPR [18] database. (a) RGB image from the dataset. (b) Depth image from the dataset origianally provided as a matrix. (c) Ground truth supplied from the dataset. (d) Saliency prediction using Global Priors [55]. (e) Saliency prediction using Itti and Kochs algorithm [19] from the GBVS matlab library [57]. (f) Saliency prediction using Low, Mid and High-level(LMH) stage saliency [56] (g) Saliency prediction using Graph-Based Visual Saliency (GBVS) [57]. (h) Our predicted saliency map

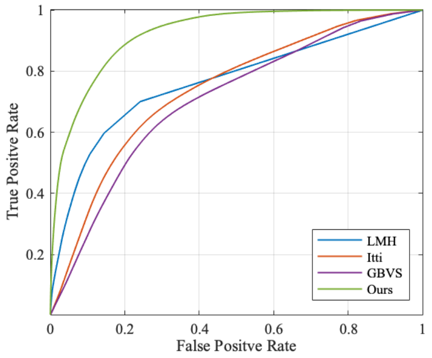


Fig. 3. Olesova Receiver Operating Characteristics

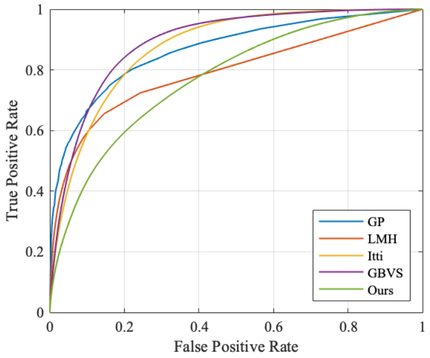


Fig. 4. NLPR Receiver Operating Characteristics

TABLE II
Best NLPR results

Model	F - Measure	Precision	Recall	AUC
GP	0.5607	0.8265	0.4243	0.8668
GBVS (RGB)	0.5113	0.5318	0.4924	0.8927
Itti (RGB)	0.4221	0.4413	0.4045	0.8597
Ours (*)	0.3950	0.3222	0.5103	0.7731
LMH	0.3208	0.6842	0.2095	0.7921

TABLE III
Comparison of depth combination methods on the Olesova dataset

Combinational method	F - Measure	Precision	Recall	AUC
RGB	0.7248	0.6928	0.7598	0.9227
RGBDepth	0.6866	0.6487	0.7292	0.8723
RGB+Depth	0.6121	0.5386	0.7087	0.8674
RGB*Depth	0.4144	0.7816	0.2819	0.8227
Depth	0.4131	0.3566	0.4907	0.6821

be seen in Table:III, the RGBDepth combination proposed is second in terms of F-Measure after plain RGB. Initially, this seems dooming for the model as it has dragged the accuracy down, unfortunately, due to the pre-designed nature of the system both the RGB and Depth maps had to be scaled down in half to fit both into the model. It is had to quantify the loss of saliency by halving the resolution and is certainly an area

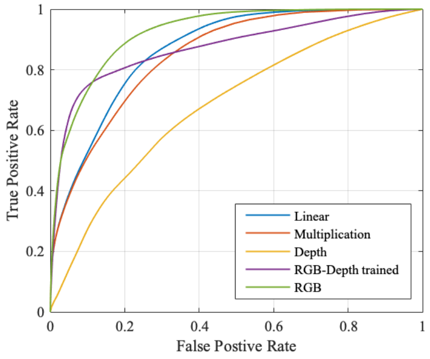


Fig. 5. All methods of combination on our proposed system compared on the Olesova dataset.

TABLE IV
COMPARISON OF DEPTH COMBINATION METHODS ON THE NLPR DATASET

Combinational method	F - Measure	Precision	Recall	AUC
RGB*D	0.3950	0.3222	0.5103	0.7731
RGB	0.3876	0.3205	0.4904	0.7469
Depth	0.2687	0.1858	0.4848	0.6836
RGB+D	0.2587	0.4436	0.1825	0.7684
RGBDepth	0.2460	0.1884	0.3540	0.6312

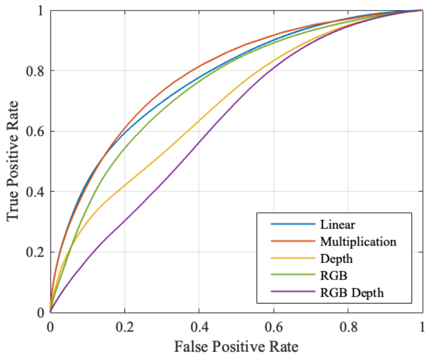


Fig. 6. All methods of combination on our proposed system compared on the NLPR dataset.

for improvement for comparison measures. We can, however, stipulate that depth data is not as important as RGB in saliency detection as both methods had access to the same amount of pixels where the RGB map only had RGB data compared to half and half RGB and depth.

Looking at other results, however, shows that our method of neural network combination of RGB-D has outperformed linear and multiplication based combination methods despite having access to half of the data available to both other methods. The recall is the highest of all combination methods showing that the majority of salient regions were detected, whereas the precision is not the highest of all methods meaning a few regions were over-estimated. The most interesting entry is the multiplication combination as it has the largest precision yet lowest recall meaning that little regions were selected but these regions were generally correct, theoretically this means that this system will lead to smaller more accurate salient regions like that of the NLPR dataset and can theorise this is why multiplication is best for the NLPR dataset as shown in Table: IV. Comparing this with other models, both Itti and GBVS methods have multiplication based being the least salient due to a tiny recall value meaning tiny amounts of salient regions are detected. This leads to the conclusion that these methods need more complex joining methods to avoid relying too much on depth. Overall in every method (excluding ours due to a poor testing method) depth has reduced saliency accuracy, this is likely due to always treating RGB and Depth equally and not employing any sophisticated weighting methods. To improve this, methods of weighting a scene based on RGB and depth contrast should be used to determine what the driving factor is. This would allow a system to treat images such as those in row 2 and 4 of Figure:2 differently due to no depth contrast in row 4 and one with high depth contrast in row 2, the only models able to do this, shown in this paper, are ours and GP.

Looking at the results from the NLPR dataset in Figure 6 and Table:IV the previously worst of the combination methods is now the most accurate method of all. Methods with the highest ratings in the self-test, showing low loss of the neural data supplied, are getting low scores in the NLPR dataset, This points to the two datasets being very different in types of saliency measured and a generally unfair test.

V. DISCUSSION

The results have shown varying levels of efficiency as testing our system on data similar to that of the training has lead to high precision and low losses meaning a great understanding and learning of the saliency employed. However when testing on generalised data precision dropped drastically. This leads to the thought of an unfair test, and this system needs to be changed. There are multiple methods of change:

- Change the training data for the neural network, the neural network could be trained on the NLPR or similar dataset and then trained on a similar set. This would change the output of the GAN to something more similar to that seen in Figure:2:c which may be desirable or undesirable based on what the system is to be used for.
- Change the testing data, this means changing the norm of testing on the NLPR dataset and instead testing on another eye-tracking dataset, this would keep the output of the neural network in the same style but would require using a non-standard testing set.

Each of these has advantages and disadvantages but ultimately depend on the requirements of implementation and usage which is not in the scope of this document. Due to the nature of our GAN, our neural network has no ability to compare preceding frames for temporal analysis and therefore may not be the greatest method for video analysis, however the neural network may learn to understand blur as movement and can account for this in saliency prediction for video analysis to replace the need for cross-frame analysis. In the future, it may be possible to add frame difference maps as feature maps into the U-Net generator to allow for such movement analysis. Another point raised is the inability to correctly compare the results due to differences in input resolution caused by the GAN system used. Ideally the GAN would have been designed specifically for this purpose with the ability to take singular and multiple images together. This would rely upon a non-square convolution or maxPooling in the neural network which would minimise the difference between models. Alternatively, the GAN could be trained with the RGB and depth downsized as they appear in the joined maps. To improve the models learning accuracy, which doesn't seem proportionally linked to NLPR accuracy, a larger quantity of epochs, around 200, should be employed. This should greatly reduce the loss of the GAN but will require either more computing power or more time. A GPU build of tensorflow should vastly reduce time by around 90% in conjunction with a high performance cluster (HPC) that is GPU based could run the majority of these tests in a day. Depth comparison has unfortunately been hard to achieve in this paper due to floored testing methods, however, it may be concluded that depth data has no ability to hurt saliency detection, as it is simply more data, unless bad quality joining methods are used. Due to this, more future work is key to compare more complex combinational methods Compared to our initial aims and objectives:

1) A new method of saliency has been proposed using generative adversarial networks but also with a new combination strategy of allowing the GAN to combine the RGB and depth data. Future work has been proposed as how to improve this network including re-writing the neural network to accept multiple images/resolutions.
2) A comparison of our method and other depth combination methods has been made with suggestions for improvements in future work. The main of these suggestions is to compare with more complex combination methods that weight depth and RGB separately before combination.

VI. Conclusion

In this paper, we proposed a novel method of 3D saliency prediction that compares both RGB and depth information simultaneously. This is to improve saliency prediction for full scene analysis which is appropriate for applications such as scene compression to reduce file size whilst keeping salient objects at higher resolutions. Our method removes the need for a post saliency combination, beats simple combination methods in GAN accuracy and simplifies the system diagram. Future work has been proposed for more object-based saliency and fairer comparison and testing.

Appendix

TABLE V
NLPR Results

Model	F-Measure	Precision	Recall	AUC
GP	0.5607	0.8265	0.4243	0.8668
LMH	0.3208	0.6842	0.2095	0.7921
Itti and Koch				
RGB	0.4221	0.4413	0.4045	0.8597
Depth	0.3163	0.3755	0.2732	0.8248
RGB+D	0.3693	0.4632	0.3070	0.8756
RGB*D	0.1098	0.3977	0.0637	0.8630
GBVS				
RGB	0.5113	0.5318	0.4924	0.8927
Depth	0.2519	0.3635	0.1927	0.8212
RGB+D	0.3550	0.4829	0.2806	0.8893
RGB*D	0.0969	0.4028	0.0550	0.8689
Ours				
RGB	0.3876	0.3205	0.4904	0.7469
Depth	0.2687	0.1858	0.4848	0.6836
RGBDepth	0.2460	0.1884	0.3540	0.6312
RGB+D	0.2587	0.4436	0.1825	0.7684
RGB*D	0.3950	0.3222	0.5103	0.7731

TABLE VI
Olesova Results

Model	F - Measure	Precision	Recall	AUC
LMH	0.1878	0.6529	0.1096	0.7646
Itti and Koch				
RGB	0.3097	0.4591	0.2336	0.7380
Depth	0.1233	0.3322	0.0757	0.6914
RGB+D	0.1776	0.3930	0.1147	0.7288
RGB*D	0.0220	0.2762	0.0115	0.7129
GBVS				
RGB	0.3163	0.4353	0.2484	0.7052
Depth	0.1792	0.4127	0.1145	0.6691
RGB+D	0.2362	0.4544	0.1596	0.7013
RGB*D	0.0353	3824	0.0185	0.6864
P2P				
RGB	0.7248	0.6928	0.7598	0.9227
Depth	0.4131	0.3566	0.4907	0.6821
RGBDepth	0.6866	0.6487	0.7292	0.8723
RGB+D	0.6121	0.5386	0.7087	0.8674
RGB*D	0.4144	0.7816	0.2819	0.8227

References

[1] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[2] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[3] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern analysis and machine intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

[4] P. Le Callet and E. Niebur, "Visual attention and applications in multimedia technologies," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2058–2067, 2013.

[5] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[6] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *European conference on computer vision*. Springer, 2012, pp. 101–115.

[7] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen, "Gaze-based interaction for semi-automatic photo cropping," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 2006, pp. 771–780.

[8] C. Rother, L. Bordeaux, Y. Hamadi, and A. Blake, "Autocollage," in *ACM transactions on graphics (TOG)*, vol. 25, no. 3. ACM, 2006, pp. 847–852.

[9] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 2232–2239.

[10] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[11] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, June 2006, pp. 2049–2056.

[12] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2. IEEE, 2004, pp. II–II.

[13] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional selection for object recognitiona gentle way," in *International workshop on biologically motivated computer vision*. Springer, 2002, pp. 472–479.

[14] U. A. Khan, L. Liu, F. A. Provenzano, D. E. Berman, C. P. Profaci, R. Sloan, R. Mayeux, K. E. Duff, and S. A. Small, "Molecular drivers and cortical spread of lateral entorhinal cortex dysfunction in preclinical alzheimer's disease," *Nature neuroscience*, vol. 17, no. 2, p. 304, 2014.

[15] M. L. Balthazar, F. R. Pereira, T. M. Lopes, E. L. da Silva, A. C. Coan, B. M. Campos, N. W. Duncan, F. Stella, G. Northoff, B. P. Damasceno *et al.*, "Neuropsychiatric symptoms in alzheimer's disease are related to functional connectivity alterations in the salience network," *Human brain mapping*, vol. 35, no. 4, pp. 1237–1246, 2014.

[16] K. Koch, J. McLean, R. Segev, M. A. Freed, M. J. Berry II, V. Balasubramanian, and P. Sterling, "How much the eye tells the brain," *Current Biology*, vol. 16, no. 14, pp. 1428–1434, 2006.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[18] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: a benchmark and algorithms," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 92–109.

[19] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision research*, vol. 40, no. 10-12, pp. 1489–1506, 2000.

[20] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artificial intelligence*, vol. 146, no. 1, pp. 77–123, 2003.

[21] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, no. CONF, 2009, pp. 1597–1604.

[22] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[23] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2007, pp. 1–8.

[24] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature reviews neuroscience*, vol. 5, no. 6, p. 495, 2004.

[25] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2007, pp. 545–552.

[26] C. W. H. Ngau, L.-M. Ang, and K. P. Seng, "Bottom-up visual saliency map using wavelet transform domain," in *2010 3rd International Conference on Computer Science and Information Technology*, vol. 1. IEEE, 2010, pp. 692–695.

[27] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 3241–3248.

[28] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision*. Springer, 2014, pp. 391–405.

[29] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 73–80.

[30] ——, "Measuring the objectness of image windows," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2189–2202, 2012.

[31] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 853–860.

[32] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 2106–2113.

[33] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 438–445.

[34] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[35] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[36] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[37] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[39] W. Ouyang, P. Luo, X. Zeng, S. Qiu, Y. Tian, H. Li, S. Yang, Z. Wang, Y. Xiong, C. Qian *et al.*, "Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection," *arXiv preprint arXiv:1409.3505*, 2014.

[40] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International conference on machine learning*, 2014, pp. 647–655.

[41] Y. Zhang, G. Jiang, M. Yu, and K. Chen, "Stereoscopic visual attention model for 3d video," in *International Conference on Multimedia Modeling*. Springer, 2010, pp. 314–324.

[42] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, and W. Lin, "Saliency detection for stereoscopic images," *IEEE Transactions on Image Processing*, vol. 23, no. 6, pp. 2625–2636, 2014.

[43] K. Desingh, K. M. Krishna, D. Rajan, and C. Jawahar, "Depth really matters: Improving visual salient region detection with depth." in *BMVC*, 2013.

[44] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Ying Yang, "Exploiting global priors for rgb-d saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 25–32.

[45] H. Xue, Y. Gu, Y. Li, and J. Yang, "Rgb-d saliency detection via mutual guided manifold ranking," in *2015 IEEE International Conference on Image Processing (ICIP)*, Sep. 2015, pp. 666–670.

[46] J. Wang, M. P. Da Silva, P. Le Callet, and V. Ricordel, "Computational model of stereoscopic 3d visual saliency," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2151–2165, June 2013.

[47] J. Wang, P. Le Callet, S. Tourancheau, V. Ricordel, and M. Perreira Da Silva, "Study of depth bias of observers in free viewing of still stereoscopic synthetic stimuli," *Journal of Eye Movement Research*, vol. 5, no. 5, pp. pp. 1–11, Sep. 2012. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00730667

[48] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: a benchmark and algorithms," in *European conference on computer vision*. Springer, 2014, pp. 92–109.

[49] C.-Y. Ma and H.-M. Hang, "Learning-based saliency model with depth information," *Journal of Vision*, vol. 15, no. 6, pp. 19–19, 05 2015. [Online]. Available: https://doi.org/10.1167/15.6.19

[50] Y. Fang, W. Lin, Z. Fang, J. Lei, P. Le Callet, and F. Yuan, "Learning visual saliency for stereoscopic images," in *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, July 2014, pp. 1–6.

[51] V. Olesova, W. Benesova, and P. Polatsek, "Visual attention in egocentric field-of-view using rgb-d data," in *Ninth International Conference on Machine Vision (ICMV 2016)*, vol. 10341. International Society for Optics and Photonics, 2017, p. 103410T.

[52] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[53] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[54] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 702–716.

[55] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Ying Yang, "Exploiting global priors for rgb-d saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 25–32.

[56] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: a benchmark and algorithms," in *European conference on computer vision*. Springer, 2014, pp. 92–109.

[57] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2007, pp. 545–552.

Hamish Sams is a M.Eng electronic and electrical student at the University of Sheffield, expecting to receive his degree in 2020, and a member of the IEEE. His research interests are in image processing, machine learning, firmware design and computer vision.

Charith Abhayaratne received his BE degree in electrical and electronic engineering from the University of Adelaide, Australia, in 1998, and his PhD in the same from the University of Bath, UK, in 2002. He is currently a lecturer in the Department of Electronic and Electrical Engineering at the University of Sheffield, UK. His research interests include video and image compression, watermarking, image and video analysis, multidimensional signal processing, graph spectral analysis, and computer vision