WAYLA - Generating Images from Eye Movements

Bingqing Yu  
Centre for Intelligence Machines  
McGill University  
Montreal, Canada  
bingqing.yu@mail.mcgill.ca

James J. Clark  
Centre for Intelligence Machines  
McGill University  
Montreal, Canada  
clark@cim.mcgill.ca

Abstract—We present a method for reconstructing images viewed by observers based only on their eye movements. By exploring the relationships between gaze patterns and image stimuli, the “What Are You Looking At?” (WAYLA) system has the goal of synthesizing photo-realistic images that are similar to the original pictures being viewed. The WAYLA approach is based on the Conditional Generative Adversarial Network (Conditional GAN) image-to-image translation technique of Isola et al. [1]. We consider two specific applications - the first, of reconstructing newspaper images from gaze heat maps, and the second, of detailed reconstruction of images containing only text. The newspaper image reconstruction process is divided into two image-to-image translation operations, the first mapping gaze heat maps into image segmentations, and the second mapping the generated segmentation into a newspaper image. We validate the performance of our approach using various evaluation metrics, along with human visual inspection. All results confirm the ability of our network to perform image generation tasks using eye tracking data.

Keywords—Image Generation; Saliency; GAN; Convolutional Network

I. INTRODUCTION

The main goal of this paper is to propose a novel approach to generate images viewed by observers based only on their eye movements.

Image generation has always been one of the primary topics in the field of computer vision. The existing image generation methods mostly focus on generating an image with higher resolution using a low-resolution image. However, in many instances, one doesn’t even have a lower resolution image. How can we generate an image in this case? Obviously, some additional information is needed.

One approach that is quite popular recently is to train a network to generate an image based on a verbal description. In this paper we look at another situation, wherein we have someone looking at an image, and want to generate an image similar, if not identical, to the one that is being viewed. But we assume that the only information we have available is the viewers’ gaze patterns or eye movement trajectories.

It is not immediately obvious whether there is enough information in the eye tracking data to allow reconstruction of what is being viewed. But, there is a well-examined link between image content and gaze patterns. It has been stated by many studies related to visual attention that viewers scan the interesting locations of a given image by controlling their attention, and the trajectory of the attention can be explored by tracking the eye movements of the viewers. Examples in the literature include the Clark and Ferrier computational model of saliency, which was used in a robotic system to demonstrate the relationships that exist between image saliency and eye movements [2] and the Itti et al. visual attention model [3], which also showed the ability of the eye fixation data to represent viewing behaviors and image salience characteristics. In the context of attention tracking during text reading, researchers have carried out various studies and developed numerous theories in order to describe how the eyes move while reading. For example, O’Regan et al. [4] demonstrated that for a single word recognition task, there exists an initial fixation location which minimizes the probability of refixation and total fixation duration on a word. In the context of continuous reading, this was extended to the notion of Preferred Viewing Location, where the fixation locations depend on the words being read [5]. McConkie and co-workers have discovered that there is an optimal fixation position in sentence reading, which is slightly to the left of a word center. Its duration can be influenced by various factors such as the complexity of a word. Moreover, it has been found that the fixation duration is the longest if the target is at the word center [6]. Grammatical factors also have an impact: functional words such as prepositions often have a higher skipping rate [7]. In sum, reading behaviors and fixation characteristics are strongly influenced by the target locations and various lexical variables of the reading material. This suggests that the particular shape and structure of eye movement trajectories constrain, to some unknown extent, the text being read. Similar conclusions can be drawn with respect to general image viewing - that the eye trajectory information constrain to some extent the contents of the image being viewed.

A. Proposed Method

To investigate the problem of generating images from eye movements, we propose to use a deep learning neural network, which takes in eye fixation data or gaze heat maps and generates, as output, images similar to the original scenes being displayed to the viewers. We call our approach
“WAYLA”, which stands for “What Are You Looking At?”. The feasibility of this approach is enhanced by the presence of numerous datasets of images with corresponding eye tracking annotations that have been published in recent years [8].

Although it is possible to construct an image generation model from scratch, we benefit from an existing deep learning model to perform our eye movements to image prediction task. We propose to use a network based on the Conditional Generative Adversarial Network (Conditional GAN), the details of which will be presented in Section II [9]. Instead of using solely the idea of the traditional GAN network [10] to generate images by minimizing the Euclidean distance between output images and ground truth images, we use the architecture of the Conditional GAN. In this way, our network can learn to generate a synthetic image by conditioning on a corresponding eye fixation heat map. As a result, the goal becomes to minimize the difference between a patch combining the eye fixation heat map with the generated image and a patch combining the eye fixation heat map with the ground truth image. More precisely, the architecture of the proposed WAYLA network uses the Conditional GAN, implemented in the image-to-image translation paper by Isola et al. [1], as a pre-built model. Their study focuses on a specific application, which is transforming an image from one version to another; however, we focused on exploring it as a solution to the eye-fixation-data-to-image generation problem.

### B. Used Datasets

We mainly focus on applying our model to two eye tracking datasets [11], [12]. One contains various scanned images of newspapers with associated gaze heat maps, and the other contains eye tracking data obtained during text reading. More detailed explanation about the motivation behind our choice of these datasets will be presented below.

The dataset provided by Vilkin et al. [11] contains numerous highly detailed newspaper images and their corresponding segmented versions. One main advantage of this dataset is that it only contains one category of images, i.e., a collection of scanned newspaper images. All the images share some common characteristics: they all have a clean layout with text/picture areas and display colorful content in the picture regions. In this way, it becomes feasible for our system to model these characteristics and generate synthetic images similar to the ground truth. Moreover, we chose the Vilkin et al. dataset while considering the applicability of our WAYLA approach. To our knowledge, a great amount of previous work related to eye movement analysis has been done in the context of text reading or web browsing. Therefore, being trained to generate images with text/picture information, the WAYLA system may find its applications in many practical problems involving text reading and attention tracking. Since this database does not have any associated eye track data, we utilized a state-of-the-art saliency model, MLNET [13], to generate all the eye fixation heat maps for the ground truth newspaper images. These heat maps are used to model where people may look at while viewing these newspapers, and serve as the eye fixation data that are fed into our network during training.

The WAYLA network can then learn to take these eye fixation heat maps as input and produce two sorts of output images. The first type of output is a simplified semantic segmentation, while the second type of output is a detailed photo-realistic image. In particular, our network is able to synthesize a segmented version of newspaper appearances. This simplified version of realistic newspapers is consistent with the eye-tracking data and filled with semantically labeled regions representing picture and text. In addition, our network is also effective at then generating images that are at a higher detail level and closely resemble the ground truth scanned newspaper images.

Although the Vilkin et al. dataset allows us to train the network to generate newspaper-like images, the output images are likely to be blurry because they are generated using very limited eye fixation data. As mentioned previously, we applied MLNET, one of the state-of-the-art saliency models, to generate the eye fixation heat maps for the Vilkin et al. dataset. When scientists build this kind of saliency models, the goal is usually to match their saliency predictions with some benchmark datasets such as MIT300 [14]. The main downside of these benchmarks is that all their experiments were conducted by displaying each image stimulus for a short amount of time, i.e., a few seconds. In addition, the datasets usually contain aggregated eye-tracking data of multiple observers. As a consequence, it is impossible to infer the exact text contents from such vague eye fixation data. To address this issue, we also applied our WAYLA approach to GECO, the Ghent Eye-Tracking Corpus [12]. The purpose is to explore whether it is possible to fill text areas with character-like content at a high detail level. The GECO dataset contains eye movements of participants performing detail reading of a novel. It contains the eye trajectories of each participant, and the data were recorded over a long period of reading time. Hence, we attempted to relate eye movements of reading to images of text by making use of GECO, which provided us with more precise gaze data.

We implemented a Conditional GAN for our system, where the eye fixation data were used to condition the operation of the generator. The discriminator serves to distinguish between “fake” patches and “real” patches. In this way, being trained on the samples described above, the network can be enabled to generate images at different levels of detail using only eye tracking data.
Figure 1. Illustration of our image generation pipeline. Our approach formulates the problem of generating detailed newspaper images from eye fixation data as a two-phase process. The network needs to be fed with the eye fixation heat maps as input and trained with the segmented images as ground truth. Then we can utilize the segmented images to train the network with the highly detailed images as ground truth.

Figure 2. (Left) A ground truth text-embedded image generated from the GECO dataset. (Right) The corresponding eye fixation heat map generated from the fixation data of the GECO dataset.

II. ARCHITECTURE AND TRAINING

A. Data Preparation

To train our network to learn a mapping from eye movement data to newspaper images provided by the Vilkin et al. dataset, we needed to produce eye fixation data and feed them as input for our model. For this purpose we used the state-of-the-art MLNET model [13] to build the eye fixation heat map dataset. To obtain the MLNET saliency predictions, we fed the MLNET network with the original scanned newspapers used by Vilkin et al., and collected the corresponding salience heat map for each input stimulus.

As shown in Figure 1, we broke the end-to-end image generation process into two phases. The first phase has the goal of generating semantic segmentations of the newspaper images, while the second phase is used to generate detailed newspaper images from the segmentations.

We also investigated the effectiveness of our model in reconstructing the text images provided by the GECO dataset. This dataset contains records about the eye fixation positions and durations of each individual for each reading session. Therefore, as shown in Figure 2, we were able to generate eye fixation heat maps that correspond to different parts of the novel that were read by participants. When generating grayscale eye fixation heat maps for the GECO dataset, for each observer and for each time a fixation is made on a specific position of a specific word, we place a bright point in the grayscale heat map according to its recorded fixation location. The brightness of that point is modulated with the recorded percentage of time spent on that specific location out of the total trial time completed by an observer.

As for the ground truth images that serve as training targets, no sophisticated data preparation process was undertaken. The ground truth segmented newspaper images and detailed newspaper images are simply taken from the dataset provided by Vilkin et al. Regarding the GECO dataset, we chose to break the reading material into parts while generating RGB images containing printed text. Each text image is a (256 X 256) size RGB image, with its red channel encoding a constant background and its green channel encoding text content. The blue channel is set to zero everywhere. It was found experimentally that this 3-channel arrangement provides better training stability and allows faster convergence, as compared with using a single channel to encode the text content. For implementation convenience, each image contains 15 words arranged into 3 rows and each row contains 5 words one following another. When generating our eye fixation heat maps for the GECO dataset, the locations of all the saliency points are adapted to the locations of the generated text-embedded images. We ensured that the correspondence is in agreement with the data provided by the GECO dataset.

B. Network Design

In recent years, generative adversarial networks and their extensions led to advances in the field of photo-realistic image synthesis [10], [15], [9], [1]. Given enough training samples, they can produce photo-realistic images that are very close to the ground truth pictures. Therefore, we chose to build our system based on the architecture of Conditional GAN, which has proven to be an efficient way to do image-to-image translation for various computer vision tasks such as image segmentation and grayscale-to-RGB image translation [1]. The overall architecture based on which the WAYLA system is built is illustrated in Figure 3.

Our modifications and some essential features of our network are presented in the remaining part of the section.

C. Two-Phase Training for Newspaper Generation

Since the Vilkin et al. dataset provided us with both segmented and detailed newspaper images, we formulated
the image generation task as a two-phase process. The first phase consists in training the network to do eye-movement-to-segmented-newspaper-image synthesis. The second phase consists in training the network to generate newspaper images with a higher level of detail from the image segmentation.

The main reason behind this two-phase framework is because the eye data to image generation task can be very hard to achieve, since it is extremely ill-posed and there is not much correlation that we can explore between the eye data and the exact image content. Therefore, it is preferred to divide the task into two sub-tasks. First, we use a network to generate a segmentation image based on the eye data. Then, another network with the same architecture is used to focus on generating a detailed image based on the segmentation image.

When training our network for the first phase, the generator is fed with the eye fixation heat maps that we produced during training. The generator is optimized to produce outputs that resemble the ground truth segmented newspaper images. The upper part of Figure 4 illustrates the input and output setting of this first-phase training.

For the second phase, we trained our network to synthesize detailed newspaper images based on the segmentation images from the dataset. The lower part of Figure 4 illustrates the input and output setting of this second-phase training. During the second phase, the input layer of the generator is fed with the segmented images provided by the Vilkin et al. dataset. Then, the generator is optimized to produce outputs that resemble the ground truth detailed newspaper images.

The training results allowed us to conduct a first set of tests to verify the effectiveness of the first-phase and second-phase networks individually. More precisely, we evaluated the results of the first-phase network by feeding it with the eye fixation heat maps in the testing set. Then, we evaluated the results of the second-phase network by feeding it with the ground truth segmentations in the testing set. Evaluation results will be presented in greater detail in Section III.

D. Integration Testing for the Newspaper Generation Pipeline

The testing procedure described above allowed us to analyze how well the WAYLA method can perform when it has ground truth segmentations at its disposal. Taking a step further, we also tested the ability of the system to generate the final detailed images based on the segmentation images generated from the first-phase training. That is to say, an integrating testing procedure was performed to evaluate how well the system can do without the ground truth segmentation. An illustration of the integrating testing procedure is shown in Figure 5.

For the integration testing, the first-phase network was first utilized to generate its segmentation outputs. Then the generated segmented images were fed to the second-phase network to produce the final images. Evaluation results of the integration testing procedure will be provided in Section III.

1) Training on the GECO Dataset: As mentioned previously, the text parts of the images generated using the Vilkin et al. dataset can have a limited resolution due to the impreciseness of the eye fixation data. Therefore, in attempt to obtain a more concrete representation of what people are looking at during reading, we also applied WAYLA on the GECO dataset. In this case, the network is trained to generate text-only images based on eye fixation data. The generator has as input the eye fixation heat maps that we created from the GECO dataset. The network is expected to produce output images with character-like patterns.

2) Training and Optimization Configuration: The loss functions used for our network are presented in the following, and it is applied to all training phases and all datasets used for our study.

The discriminator, whose task is to classify between real and fake pairs, has the following binary cross entropy objective to maximize:

\[
L_D = E_{x,v,y}[\log D(x,v,y)] + E_{x,v}[\log(1 - D(x,v,G(x,v)))] \quad (1)
\]

In Equation 1, \(x\) is the input of the generator, \(y\) represents all ground truth images that the generator has as target. As for the generator, since it is stated in [16] that mixing the GAN loss with another standard content loss such as
Euclidean loss can improve the training of deep neural networks, we chose to use the $L_1$ distance as the additional loss and combine it with the adversarial loss described above to construct the loss function for our generator. The $L_1$ distance represents the difference between the outputs of the generator and the ground truth images. Thus, the overall loss function of the generator is defined as:

$$L_G = L_D + \lambda L_1(G)$$  \hspace{1cm} (2)

We set the value of $\lambda$ to 0.01, and this decision is taken based on our experiment observation and on the analysis in [1]. When the $L_1$ loss is weighted 100 time larger than the GAN loss, there are fewer artifacts produced in the output of the generator. All layers of the network need to be trained from scratch. Weights are randomly initialized using uniform distribution between $-0.05$ to $0.05$. We always preserve 20 percent of the total samples for testing. The network is trained by updating the generator and the discriminator alternatively. The GAN cross-entropy loss is backpropagated to the discriminator to update its weights. Then, by keeping the discriminator weights constant, we combine the cross-entropy loss with the $L_1$ loss and backpropagate this error to update the generator weights. Mini-batches of 2 samples per batch are used for training. Dropout layers and batch normalization are used in our network to accelerate convergence. The RMSprop optimizer is used to optimize both the generator and the discriminator, with a learning rate of 0.001, a decay rate of 0.9, a momentum of 0 and an $\epsilon$ of $1 \times 10^{-6}$.

It needs to be mentioned that the RMSprop optimizer is chosen based on our observation from experimentation. We observe that it allows faster convergence and behaves more stably, as opposed to some other optimizers such as SGD. In terms of computational cost, RMSprop is also less expensive than some of its extensions, such as Adam [17], [18]. Moreover, experiments with various activation functions and receptive field sizes for the discriminator were conducted. Based on our observation, Tanh activations are less computationally expensive and allow faster convergence. A (46 x 46) PatchGAN is chosen as our optimal architecture, as it allows the output images to have the highest quality.

### Table I

|                         | SSIM  |
|-------------------------|-------|
| Segmented newspapers    | 0.83  |
| Detailed newspapers (Individual Testing) | 0.54  |
| Detailed newspapers (Integration Testing)  | 0.47  |

**SSIM scores obtained by comparing our synthesized images with the ground truth.**

Figure 6. Example results of WAYLA obtained by testing the two phases separately and independently. The qualitative results of WAYLA on the testing set are compared to ground truth.

### III. Evaluation and Discussion

#### A. Evaluation of the Model for the Newspaper Generation Task

Among the large number of tools available to evaluate the perceptual quality of images, we chose to use the structural similarity (SSIM) index to compute the perceived similarity between our generated images and our ground truth images. In this way, the evaluation results provide a better understanding of the overall quality of generated images, since SSIM not only compares pixel values but also image structures [19], [20]. In Table I, similarity scores are reported for both the segmented and detailed newspaper image generation tasks.

As there is no previous work on generating images from eye movement data, to investigate the significance of our contribution to synthesizing newspaper images, we looked at the SSIM scores reported in various papers on image generation. By comparing our SSIM scores with other studies, we can verify whether our synthesized images are similar enough to the ground truth images. Mihaela R. et al. stated that with their auto-encoding GAN, they can generate images with a SSIM mean value of 0.62 when compared with ground truth images [21]. In another paper, which uses a GAN structure to implement image super-resolution, it is shown that a generated image with SSIM score of 0.6423 is very similar to the original image to human perception [22]. In addition, Zhou et al. introduced the concept of SSIM, they presented an image disturbed with Gaussian noise with a SSIM score of 0.30. The noisy version still preserves most of the texture patterns of the original image. The SSIM score obtained from the integration testing is situated slightly below the other SSIM scores presented in Table I; however, given the fact that the integration testing is done without ground truth segmentation, it is inevitable that the network can only synthesize images of limited quality.

Figures 6 and 7 allow a qualitative evaluation of the
performance of our image generation pipeline. It can be observed that the generated segmented images produce text/picture patterns that are visually similar to the ground truth. It is also worth noticing that when WAYLA is applied on a specific dataset, i.e., the newspaper dataset, the network is able to extract meaningful information from eye movement data and generate images which are compatible with the specific dataset. Taking the first row of Figure 6 as an example, despite the fact that there are numerous eye fixation points in the top part of the eye fixation image, in the generated segmented image, only center part of the eye fixation locations are converted into picture areas. Furthermore, the last column of these figures provides some images generated by directly mapping the gaze data into detailed images. The poor quality of these images justifies the motivation for our two-phase framework.

In sum, the evidence presented above shows that WAYLA is effective at generating images that are highly compatible with the ground truth. It needs to be emphasized that the low resolution of the generated images is mainly due to the limited amount of information provided by the eye tracking data. However, in Section III-C, we will prove the ability of our model to reconstruct text-embedded images given more precise eye tracking data.

B. Application of the Training Results on New Datasets

A common practice for demonstrating the applicability of a proposed algorithm is to analyze its performance across different datasets. Thus for the WAYLA model, we also conducted a set of experiments to verify its cross-dataset generalization ability. That is to say, instead of testing the model on the Vilkin et al. testing sets, we applied the training results of WAYLA to two other datasets. The new datasets should contain image stimuli along with eye fixation annotations. In addition, these image stimuli need to have a newspaper-like pattern.

One one hand, we investigated if the WAYLA model performs well when applied to the eye trajectories of one single individual. More specifically, we use the gaze data of one individual, instead of the gaze data from multiple viewers, to feed the newspaper generation pipeline. A dataset called Webtask provides us with such kind of individual data [23]. The saliency maps are generated by using the eye trajectories of one randomly selected participant of the Webtask experiment.

On the other hand, we also evaluated the performance of WAYLA by using the multi-viewer gaze data of another dataset, called Webpage Saliency [24], whose image stimuli are similar to that in the Vilkin et al. dataset.

The same integrating testing procedure, described in Section II-D, was applied to these new data.

It is common practice to generate saliency maps by using a Gaussian kernel, with an optimal sigma, to blur all the eye fixation points. Through our experiments, the sigma
which suits our pre-trained model best is found to be 1.5 degree of visual angle for the Webtask saliency maps, and 0.7 degree of visual angle for the Webpage Saliency saliency maps. The evaluation results give a SSIM score of 0.56 with the Webtask dataset, and 0.52 with the Webpage Saliency dataset. Figure 11 allows a qualitative evaluation by displaying some examples of the generated images. By analyzing the results, it is clear that there are similarities between the texture patterns of the generated detailed images and those of the ground truth images. Although the blurred eye fixation data constrained the quality of the output images, the testing results can still confirm the effectiveness of the model when applied to other datasets.

Finally, it is worth emphasizing that the sigma value of the Gaussian kernel needs to be well tuned for each dataset. A sigma larger than one degree of visual angle is required to intensify the individual saliency maps of the Webtask dataset. However, a smaller sigma value is found to be the most suitable for adjusting the saliency maps of the Webpage Saliency dataset.

C. Evaluation of the Model for the GECO Dataset

As for evaluating the performance of our approach when it is applied to the GECO dataset, we first provide some example results of the generated text-embedded images in Figure 8, from which a qualitative evaluation can be done.

It is important to remember that, although the network is designed to perform image generation without using any natural language processing, the training has proven to be extremely successful in the sense that all the generated images display text-like contents. Furthermore, numerous valid English words can be observed in the generated images. We also utilized various text analysis metrics, in combination with human inspection, in order to evaluate the quality of our outputs. Optical Character Recognition (OCR) was used to extract all the text-like content from the synthetic images generated by our network [25]. In total, 13270 alphabet characters are retrieved from 241 synthetic images. There are only 72 occurrences for which OCR engine is not able to convert a scanned character into a valid alphabet letter. In addition, as presented in Figure 10, we also used text segmentation in attempt to divide the text content in our synthesized images into segments with various lengths. Each segment can be considered as an individual word that the system generates based on the input eye fixation data.

It is worth mentioning that for illustration purpose, we display the qualitative results of the text-embedded images by using a black text color with white background color. Although we used a red and green encoding method to facilitate the training of our network, the generated results can be easily converted to black and white afterwards in order to be more similar to the real image content that is viewed by the readers.

Figure 9 compares the histogram that summarizes the length distribution of all segments obtained from our synthetic images versus the histogram generated using the ground truth images. It can be observed that the shape of the two distributions are similar. Most segments have lengths ranging from 23 to 26 pixels. This means, in both the ground truth and generated images, most words are constructed using 3 to 4 characters. Thus we observe a close similarity between the generated images and the ground truth images. This serves as a demonstration of our network’s ability to generate images filled with character-like content.

IV. CONCLUSION

In this paper, we explored the possibility of inverting the relationships between image stimuli and eye movements and successfully developed an approach to synthesize the content of viewed images based on eye tracking information. We presented WAYLA, a deep-learning technique which utilizes Conditional GAN to generate newspaper images with different detail levels from eye fixation heat maps. Our work proved that the network is able to generate images with text/picture patterns based on the eye movements of people glancing at images. Moreover, when precise eye trajectories are available, the system is able to reconstruct text-embedded images that are similar to the ground truth. It is worth noting that our network is a preliminary attempt to
infer images viewed by observers from their eye movements. Although further improvements could be made to enhance the quality of the generated images, the idea of inverting the path starting from viewed content and ending with eye tracking information can be applied in various settings. By introducing the possibility of inferring image content based on gaze data, the WAYLA approach could open doors for more diversified image generation models in the future.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the staff at the Centre for Intelligent Machines (CIM) at McGill University for their technical support.

REFERENCES

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” arXiv preprint arXiv:1611.07004, 2016.

[2] J. J. Clark and N. J. Ferrier, “Modal control of an attentive vision system.” in ICCV, 1988, pp. 514–523.

[3] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Transactions on pattern analysis and machine intelligence, vol. 20, no. 11, pp. 1254–1259, 1998.

[4] J. K. O’Regan and A. Lévy-Schoen, “Eye-movement strategy and tactics in word recognition and reading.” 1987.

[5] G. W. McConkie, P. W. Kerr, M. D. Reddix, and D. Zola, “Eye movement control during reading: I. the location of initial eye fixations on words,” Vision research, vol. 28, no. 10, pp. 1107–1118, 1988.

[6] F. Vitu, G. W. McConkie, P. Kerr, and J. K. O’Regan, “Fixation location effects on fixation durations during reading: An inverted optimal viewing position effect,” Vision Research, vol. 41, no. 25, pp. 3515–3533, 2001.

[7] C. Wotschack, Eye movements in reading strategies: How reading strategies modulate effects of distributed processing and oculomotor control. Verlag Potsdam, 2009, vol. 1.

[8] S. Winkler and R. Subramanian, “Overview of eye tracking datasets,” in Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on. IEEE, 2013, pp. 212–217.

[9] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in neural information processing systems, 2014, pp. 2672–2680.

[11] A. Vilkin, I. Safo nov, and M. Egorova, “Algorithm for segmentation of documents based on texture features,” Pattern recognition and image analysis, vol. 23, no. 1, pp. 153–159, 2013.

[12] U. Cop, N. Dirix, D. Drieghe, and W. Duyck, “Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading,” Behavior research methods, vol. 49, no. 2, pp. 602–615, 2017.

[13] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, “A deep multi-level network for saliency prediction,” in Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE, 2016, pp. 3488–3493.

[14] T. Judd, F. Durand, and A. Torralba, “A benchmark of computational models of saliency to predict human fixations,” 2012.

[15] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv preprint arXiv:1511.06434, 2015.

[16] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544.

[17] T. Tieleman and G. Hinton, “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude,” COURSERA: Neural networks for machine learning, vol. 4, no. 2, pp. 26–31, 2012.

[18] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[19] Z. Wang and A. C. Bovik, “Mean squared error: Love it or leave it? a new look at signal fidelity measures,” IEEE signal processing magazine, vol. 26, no. 1, pp. 98–117, 2009.

[20] F. Boulogne, J. D. Warner, and E. Neil Yager, “scikit-image: Image processing in python,” 2014.

[21] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed, “Variational approaches for auto-encoding generative adversarial networks,” arXiv preprint arXiv:1706.04987, 2017.

[22] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” arXiv preprint arXiv:1609.04802, 2016.

[23] T. Betz, T. C. Kietzmann, N. Wilming, and P. K, “In-geco: An eyetracking corpus of monolingual and bilingual sentence reading,” Behavior research methods, vol. 23, no. 1, pp. 153–159, 2013.