TRANSFORMING IMAGES DEMONSTRATES DEGRADATION OF GAZE FIXATION PREDICTION PERFORMANCE

by
James Youngblood

A thesis submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Master of Science

School of Computing
The University of Utah
August 2025

The University of Utah Graduate School

STATEMENT OF THESIS APPROVAL

The thesis of *James Youngblood* has been approved by the following supervisory committee members:

Rogelio Cardona-Rivera,	Chair,	*(Date Approved)*
Paul Rosen,	Member,	*(Date Approved)*
Cem Yuksel,	Member,	*(Date Approved)*

by *Mary Hall*, the Chair of the School of Computing,
and by *Darryl P. Butt*, the Dean of the Graduate School.

ABSTRACT

Using saccadic fixation points collected on reference images and the transformations of those images, we show that common digital image transformations–including cropping, rotation, contrast adjustment, and noise overlay–significantly degrade the performance of state-of-the-art gaze fixation prediction models. We show that there are no reliable heuristics which indicate the degradation of performance for image transformations in general; the collection of real gaze distribution data on transformed images is required.

CONTENTS

LIST OF FIGURES

I. INTRODUCTION

The prediction of human gaze behavior, sometimes referred to as saliency, has many potential use cases. Our study is motivated by the potential for new interactive experiences which operate on knowledge of where the player is looking.

However, for most of these use cases, it is difficult to constrain the image a gaze prediction model will see to a specific style or class of image, such as those we see in datasets that the models train on. The number of studies which test the effects of images which stray from these datasets is sparse, and out of date with the state-of-the-art models.

Gaze prediction models are notably biased towards candid photography: Matthias Kümmerer and Matthias Bethge [1] show that all leading gaze prediction models utilize transfer learning from other problem domains, primarily object recognition. Researchers do this becase of the lack of training data for the gaze recognition task, relative to the object recognition task. The object recognition task prioritizes candid photography over stylistic representations due to the assumption that object recognition applications usually involve unaltered images captured with a camera, but this assumption does not hold true for the gaze prediction task. This assumption is implicitly encoded into most of the training data a gaze prediction model will see, and so the model's performance may suffer when generalizing to stylized images.

Furthermore, a study by Zhaohui Che, Ali Borji, Guangtao Zhai, Xiongkuo Min, Guodong Guo and Patrick Le Callet [2] shows that image transformations influence gaze fixations, and effects of transformations on gaze behavior may be difficult to model. This fact is problematic, because modern visual productions use multiple steps of image processing and rendering for stylistic effect. Each step represents an image transformation, and a possible degradation of gaze prediction performance.

Our work measures the performance of state-of-the-art gaze prediction models on images which have been transformed with common image

transformations,using the data from Che et al., and we find that the performance degrades significantly.

Additionally, we correlate derived metrics gathered from the images and their transformations, in an effort to find heuristics that signal how much a prediction's accuracy has degraded without the need to conduct a human study. Unfortunately, we find that there are only a few weak relationships between the metrics and the performance of the model on the transformed image.

II. RELATED WORK

The effects of image transformations–including cropping, rotation, skewing, and edge detection–on gaze behavior are studied by Che et al. [2]. Their paper is motivated by the possibility of augmenting gaze prediction datasets with transformed images. We extend their work by repeating their experiments using current state-of-the-art models which had not been published at the time. Rather than analyze our results with the intention to augment a training dataset, as Che et al. did, we analyze our results to determine the extent to which prediction performance might degrade when transformations of various kinds are aplpied, and whether there are reliable heuristics which can be used to predict the degradation of performance of a gaze prediction model on a transformed image.

The effects of digital image alterations on gaze behavior have also been studied by Rodrigo Quian Quiroga and Carlos Pedreira [3], who performed an experiment collecting gaze fixations before and after manual Photoshop edits were made to photographs of paintings. Insufficient explanation for the intention behind edits makes it difficult to formally and generally describe the image transformations they studied. Instead, we study computationally-modeled image transformations, such that we can describe the effects they produce on gaze behavior more formally and generally, and so that we can study a greater scale of data which does not require human decisions for each image.

III. BACKGROUND

The foundations we build upon are reviewed comprehensively by Kümmerer et al. [1]. We will briefly describe key terms and concepts.

In the process of studying an image, the human eye will occasionally jump to a new fixation point in what is called a "saccade". Both measured and predicted gaze can be represented by a two-dimensional probability distribution. This is usually represented using a grayscale image called a "saliency map".

Under this definition of a saliency map, pixels with the highest intensity are those most likely to be the next fixation target after a saccade for an arbitrary person under "free-viewing" conditions (which means the person has not been instructed to search for an element of the image).

When collecting gaze distribution data from human subjects, we receive a collection of saccadic fixation points from an eye tracker over the area of the image presented to the subject. Converting these points into a saliency map can be done by brightening the pixels each point falls under and normalizing to a probability distribution, but such a saliency map is "speckled" rather than smooth, and likely diverges from the real gaze distribution because of the limited sample size of fixation points.

When testing greater fixation point sample sizes, we see that on the limit of infinite sample size, the saliency map would converge to a smooth probability distribution rather than a discrete point cloud. Thus, if we wish to obtain a better estimate of the real gaze distribution, we should blur the saliency map with a Gaussian kernel to obtain a smoother probability distribution. This step is referred to as "regularization". (It is convention to set the size of the Gaussian kernel to a pixel value equivalent to one degree of visual angle in length from the human subject's perspective.)

This regularized saliency map is our best proxy for the real gaze distribution. We will refer to it as the "real map" for brevity.

Gaze prediction models will compute a saliency map for an image using only the image's contents. It will not have access to any fixation points

gathered from human subjects. Using performance metrics whose computation we will describe shortly, one can use those metrics computed on the real map as a reference point for the upper limit of performance for those metrics computed on the gaze prediction model's output.

We wish also to find a lower reference point for performance. For a prediction to be a lower reference point, it should predict any features or biases that are common in the dataset or class of images we are studying, but should not predict any features based on the image's contents. This will allow us to see what information the gaze prediction model can inference from an image itself, rather than any assumptions the model might make based on previous experience with the dataset or class of images.

An adequate lower reference point can be computed by collecting all fixation points for the entire dataset of interest, and regularizing similar to the real map. This is usually referred to as the "center bias", due to the prevalent tendency of most human subjects (and therefore datasets of human gaze behavior) to fixate towards the center of an image. The centerbias will account for most dataset-wide biases because of the averaging across all images in the dataset, and it will weight any fixations which were specific to any given image to a lesser degree due to the averaging as well.

(Note that one can compute a more "competitive" lower reference point by using a cross-validation approach to improve upon the centerbias, as shown by Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge [4]. We find that the improvement this step provides is marginal, and so we omit the process. For more details, see the "Method" section.)

We compare saliency maps using metrics compiled by Zoya Bylinskii, Tilke Judd, Aude Olivia, Antonio Torralba and Frédo Durand [5]. Each metric computes a measure of either divergence or similarity between two saliency maps, or measures the saliency values at a set of fixation points. We evaluate the usefulness and qualities of each metric as it pertains to our study in the "Method" section.

Using these metrics, if a saliency map produced by a gaze prediction model is closer to the real map than the center bias, we can conclude that the gaze prediction model is performing well.

The most widely adopted benchmark for gaze prediction model performance is the MIT/Tuebingen Saliency Benchmark [6], which lists

many of the metrics described by Bylinskii et al. [5] to compare the performance of submitted models. We intended to select the current top contender on the benchmark, DeepGaze IIE [7], and a lower accuracy but faster-running and smaller memory-footprint model, UNISAL [8], as our state-of-the-art models for our study. Both of these models perform better than the centerbias on the MIT/Tuebingen Saliency Benchmark for the CAT2000 dataset (which the Che et al. [2] dataset is derived from), while the models Che et al. [2] studied performed worse.

Unfortunately, we were unable to replicate the performance expected from the DeepGaze IIE model, and so we continued our study only using the UNISAL model. More details are described in the “Results” section.

IV. METHOD

We use the dataset provided by Che et al. [2], which includes 100 randomly selected images from the CAT2000 dataset [9], with 18 different transformations applied to each image. This produces a total of 1900 images, including the reference untransformed images. See figure 2 for examples of all 18 transformations. Gaze fixation points are recorded for each image.

We compute the centerbias for the reference (untransformed) set and each transformation set by collecting all fixation points for the images of the transformation and applying a Gaussian blur with a kernel size of 57 pixels, which is one degree of visual angle according to Che et al. As reported in the "Results" section, we find that the centerbias performs closely to the reported performance of the centerbias on the MIT/Tuebingen Saliency Benchmark for the CAT2000 dataset for the reference (untransformed) images. Given this, we decide to omit any cross-validation step as improvements would be marginal.

We also compute the real map for each image using the same Gaussian blur kernel on the fixation points for each image separately.

We run the DeepGaze IIE [7] and UNISAL [8] models on all reference and transformed images. The dataset images are at 1920x1080 resolution, and we run inference for both models at this resolution, but we also run inference for downscaled images which better match the expected resolution of the models. DeepGaze IIE [7] expects an image with a width of 1024, so we downscale the images to 1024x576 for another DeepGaze inference. UNISAL [8] expects several resolutions for different datasets it was trained on, so we run an inference for the resolutions of 384x224 (which matches the DHF1K dataset [citation needed] resolution), 384x288 (which matches the SALICON dataset [citation needed] resolution), and 384x216 (which preserves the ratio of the 1920x1080 image, but with a width of 384). We run inference for each model at each resolution, and intend to select the best-performing resolution for each model.

There will be two steps to our study. First, we wish to evaluate the performance of the models on both reference and transformed images. When comparing average performance, we hypothesize that the model's performance will be degraded, but we wish also to get a sense of how much the performance degrades in our analysis, if possible.

Second, we wish to find a correlation between any metrics we can derive from a source gaze distribution dataset and transformations upon that dataset, without having to gather human trials or real gaze distributions, that can be used as a heuristic to estimate the performance of a model's predictions for a type of transformation in general. We will look to the metrics described by Bylinskii et al. [5] for both the first and the second step.

We consider the metrics described by Bylinskii et al. [5] to compare saliency maps produced by models. These metrics include area-under-the-curve (AUC-Judd, citation needed) metric, the shuffled area-under-the-curve (sAUC, citation needed) metric, the histogram similarity metric (SIM, citation needed), the correlation coefficient metric (CC, citation needed), the Kullback-Leibler divergence metric (KL, citation needed), the information gain metric (IG, citation needed), and the Earth Mover's Distance metric (EMD, citation needed).

We wish to isolate the most relevant metrics for our study. With relevant metrics, we can evaluate model performance and conduct statistical and correlation analysis. However, by checking too many metrics, we increase the likelihood of false positives when searching for relationships between metrics due to noise. Thus, we select metrics with the most useful qualities and most significance.

The metrics cluster into rank-correlated groups. Bylinskii et al. [5] show that the AUC-Judd, sAUC, SIM, CC, NSS, and EMD metrics are highly rank-correlated to each other on the MIT/Tuebingen Saliency Benchmark [6]. Additionally, they find that IG and KL metrics are rank-correlated in a separate group. If we find an external correlation exists for a metric in one of these groups, and we assume that the rank correlation between the metrics of the group is close enough to a linear relationship such that it does not excessively weaken or distance any transitive relationships, then the external correlation must also exist for the other metrics of the group. We will make

the above assumption, and so we should select the fewest metrics possible from each group.

We decide against the AUC metric because it is invariant to monotonic transformations, and has saturated benchmarks due to this property. We would like to be sensitive to the relative importance of salient regions, which the AUC metric is not.

We decide against the sAUC metric because it assumes no centerbias is present in the saliency maps that a model produces, and we wish to include centerbias in our study such that we study holistic viewing behavior.

We decide against the SIM metric because it is not symmetrical for false positives and negatives, meaning a false negative will impact the score more than a false positive.

We decide against the EMD metric because it priorities the accuracy of relative saliency of regions, but does not prioritize the accurate placement of those regions. These are opposite priorities to our motivating causes for the study–we want to know where a user is looking as accurately as possible.

Thus, for the first rank-correlated group, we select the NSS and CC metrics. NSS is favored, because it is parameter-free whereas CC requires a decision on the size of a gaussian kernel to blur the saliency map before computation, as Bylinskii et al. [5] recommend for fair comparisons between saliency maps that include various frequencies of information. However, NSS compares a saliency map to a set of fixation points, whereas CC compares two saliency maps. Fixation points are not available when comparing between two saliency maps that where produced by two model inferences, for example.

For the second rank-correlated group, we select the IG and KL metrics. IG is favored, because it provides a comparitive measure against a baseline (the centerbias), and because it is parameter-free, but once again requires fixation points. KL requires the same tuning of the gaussian kernel size as CC, and also does not require fixation points.

Thus, we can apply the metrics we have selected to the two steps of our study. For the first step, we will compute average NSS and IG metrics for each transformation set, and compare the average performance of the models on the transformed images to the average performance of the models on the reference images.

Of the transformations in the dataset, there are a number of them which are similar to each other but applied with different intensities. For example, “ContrastChange_2” is a stronger increase in contrast than “ContrastChange_1”, but otherwise the same transformation. We will group together these transformations in the analysis of our results, and see if we can find a trend in the performance of a model as we increase the intensity of the transformation.

For the second step, we will compute the structural similarity index (SSIM, citation needed) between the reference and transformed images, the CC and KL metrics between the saliency maps produced by the model for the reference and transformed image, and the NSS and IG metrics for the reference saliency map the model produced. These metrics are selected because they can be computed using only a reference dataset with gaze distribution records and any arbitrary image transformation, without the need to measure real gaze distributions for the transformed images.

Furthermore, we select these metrics as the most directly connected to the image transformation process and the saliency maps derived from that process, and thus assumed to be more likely to show relationships when studying the effects of transformations on gaze prediction performance. We recognize that these metrics are not exhaustive, and we task future studies with enumerating metrics with possible relationships more thoroughly. We collect these five metrics as our independent variables. Our dependent variables will be the NSS and IG metrics for the transformed saliency map the model produced–both as measures of the model’s performance on the transformed image.

We wish to find a correlation between any pair of independent and dependent variable. There are 10 possible pairs between these variables, and so we will perform 10 separate correlation tests for each transformation set of images. We will simply graph the independent and dependent variable values for each image in the transformation set to spot patterns, find a line of best fit, and compute the linear correlation coefficient for each pair of variables. From there, we will interpret any correlation with a correlation coefficient above 0.5 as significant, and interpret what meaning those significant relationships might have in the context of our study.

As we graph the data, we find that some outliers exist. We add a filter which will omit any sample which falls beyond three standard deviations from the mean for either the independent or the dependent variable in each graph. These samples are also omitted from the correlation coefficient computation.

We recognize that it would be ideal to compute a measure of how likely it is that a relationship arises due to a non-representative sample of images, such as a p-value. In order to compute a p-value, you must determine how likely it is that a given sample, in this case an image, is representative of the class of images you wish to study. This would be difficult due to the vague nature of image classes and determining whether an image is representative or not. We instead argue that, with our image class defined as “images from the CAT2000 dataset”, and the assumption that the CAT2000 dataset is representative of images found in many other useful tasks we might apply gaze prediction models to, our sample size of 100 randomly selected images from the CAT2000 dataset is large enough to provide a reasonable basis for our study.

We make great effort to ensure our study is reproducible. Find the code and the data at the codeberg repository.

V. RESULTS

We find that our centerbias performs closely to the reported performance of the centerbias on the MIT/Tuebingen Saliency Benchmark for the CAT2000 dataset for the reference (untransformed) images. (More accurate measurements here)

We also find that the UNISAL model performs similarly to the expectation set by the MIT/Tuebingen Saliency Benchmark for the CAT2000 dataset for the reference (untransformed) images. (More accurate measurements here)

However, we were unable to replicate the performance expected from the DeepGaze IIE model, despite our best efforts to follow the protocol outlined in the DeepGaze IIE paper. (More accurate measurements of how bad here) We have reached out to the authors of the paper for comment, but have not heard back yet, and so we continue our study only using the UNISAL model.

Though we found that UNISAL performs as expected for reference images, we find that it performs worse for transformed images. We find that the average NSS and IG metrics for the transformed images are significantly lower than the average NSS and IG metrics for the reference images. If using the centerbias as a binary threshold for whether the prediction is "good enough", then the UNISAL model would be good enough for reference images but not for transformed images. (More intuitive numbers here)

We find that, as we increase the intensity of the transformation, the performance of the model on the transformed images decreases, at varying rates for different transformations. See the following figure.

For the second step of our study, we find that there usually exists a strong correlation (above 0.5 correlation coefficient) between reference image NSS performance and transformed image NSS performance, and the same goes for reference image IG performance and transformed image IG performance. For the ContrastChange and Rotation transformations, the correlation coefficients fall below 0.5. The IG correlation coefficients tend to

be a little weaker than the NSS, and as such they fall below 0.5 for the Boundary, MotionBlur_2, and Shearing_3 transformations as well.

This trend tells us that, at least for the majority of transformations tested, an increase in performance on the reference image leads to an increase in performance on the transformed image. However, upon further analysis we find that this relationship is further qualified.

We note that for those transformations which suffered more degradation during the first step, the correlation coefficients were weaker in the second step. Upon closer inspection, we find that this is primarily due to a loss in performance for transformed images which performed well before transformation. In other words, if a reference image was hard to predict, and the prediction made was poor, then it would be roughly in line with the performance of the transformed image. However, if an image was easy to predict, and the prediction made was good, then the transformed image's performance would lag behind and would not improve as much as the reference image's performance.

With this, we can hypothesize that many transformations impose a plateau on prediction accuracy which is lower than the reference images, for current models.

For the ContrastChange and Rotation transformations, for which the NSS and IG correlations fell significantly below 0.5, we find that there does exist a weaker relationship (hovering around 0.5) between the CC and KL metrics between the reference prediction and the transformed prediction to the transformed NSS performance. This tells us that, for these transformations which the models perform poorly on, if their prediction for the transformed image looks significantly different from their reference prediction, then it is likely that the transformed prediction will perform poorly. This relationship between CC/KL to transformed NSS does not hold true for all transformations which the reference NSS to transformed NSS relationship was weak. Specifically, it does not hold true for the Boundary, MotionBlur_2, and Shearing_3 transformations.

Further data collection and testing on a greater number of transformations must be done to determine if a weak NSS to transformed NSS relationship is a reliable predictor of a stronger CC/KL to transformed NSS relationship, or if this is a unique property of the ContrastChange and

Rotation transformations. In absence of this data, we can hypothesize that the CC/KL metric is a weak heuristic for predicting the performance of a model on contrast change and rotation transformations.

VI. CONCLUSION

We find that, for all transformations barring the Mirror transformation, the UNISAL model performs worse than the reference set of images. However, there is still a correlation between performance on a reference image and performance on a transformed image, which allows us to hypothesize that a transformation imposes a plateau on prediction accuracy which is lower than the reference images, for current state-of-the-art gaze prediction models.

We find that the CC/KL metric is a weak heuristic for predicting the performance of a model on contrast change and rotation transformations. In this unique case, one can infer some information about a model's performance on transformed images without gathering human trial data.

In general, our work indicates that there are several weakness that current state-of-the-art models have with respect to the image classes and transformations that images might be subject to in common situations. We argue that there is a strong need for much more gaze distribution data for models to be trained and tested on.

For future work, we would like to test a greater number of transformations, including digital distortions, color manipulations, and stylistic filters. We also would like to test whether any emergent effects on the prediction accuracy arise when composing multiple transformations. We would like to test more state-of-the-art models, to see if there is divergence in how the behave on transformed images. We would like to test transformations with more rigorous definitions of "intensity", at a granular level such that we can more accurately elucidate trends in performance as we increase the intensity of the transformation. We would like to test more transformations to determine if the CC/KL to transformed NSS relationship is a reliable predictor of performance on any more general classes of images.

REFERENCES

[1] M. Kümmerer and M. Bethge, "Predicting Visual Fixations," *Annual Review of Vision Science*, vol. 9, no. Volume9, 2023, pp. 269–291, 2023, doi: 10.1146/annurev-vision-120822-072528.

[2] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, and P. L. Callet, "How is Gaze Influenced by Image Transformations? Dataset and Model," *IEEE Transactions on Image Processing*, vol. 29, pp. 2287–2300, 2020, doi: 10.1109/tip.2019.2945857.

[3] R. Quian Quiroga and C. Pedreira, "How Do We See Art: An Eye-Tracker Study," *Frontiers in Human Neuroscience*, vol. 5, 2011, doi: 10.3389/fnhum.2011.00098.

[4] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Information-theoretic model comparison unifies saliency metrics," *Proceedings of the National Academy of Sciences*, vol. 112, no. 52, pp. 16054–16059, 2015, doi: 10.1073/pnas.1510393112.

[5] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *arXiv preprint arXiv:1604.03605*, 2016.

[6] M. Kümmerer *et al.*, "MIT/Tübingen Saliency Benchmark." Accessed: Mar. 15, 2025. [Online]. Available: https://saliency.tuebingen.ai/

[7] A. Linardos, M. Kümmerer, O. Press, and M. Bethge, " DeepGaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling ." [Online]. Available: https://arxiv.org/abs/2105.12441

[8] R. Droste, J. Jiao, and J. A. Noble, "Unified Image and Video Saliency Modeling," in *Computer Vision – ECCV 2020*, Springer International Publishing, 2020, pp. 419–435. doi: 10.1007/978-3-030-58558-7_25.

[9] A. Borji and L. Itti, " CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research ," 2015, [Online]. Available: https://arxiv.org/abs/1505.03581