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 images and digital transformations of those images, we show that common transformations–including cropping, rotation, contrast adjustment, and noise–degrade prediction accuracy for state-of-the-art gaze fixation prediction models. We fail to find any generalizable heuristics which indicate the degradation of prediction accuracy for image transformations; the only known way to confirm an arbitrary transformation has caused a degradation in prediction accuracy is to collect real gaze distribution data on transformed images.

CONTENTS

LIST OF FIGURES

I. INTRODUCTION

In the production of visual media, predictions for human gaze behavior provide feedback on the way a scene will be percieved, and can be used to focus production effort on the most important visual aspects of a scene. For robotics, gaze predictions provide a guidance for training an agent to scan its surroundings. For our research, we are motivated by gaze prediction for interactive applications, which will allow unique program logic based on the visual attention of the user.

The field of gaze prediction has seen rapid progress with the emergence of deep learning models over the past decade, but when using deep learning models, users must be careful of possible hidden assumptions the model makes based on its training data. The studies which explore model behavior for application-specific classes of images are sparse, those which exist are out of date with the state-of-the-art models.

Gaze prediction models are notably biased towards a class of images we will refer to as “candid photography”; minimally stylized images captured from a camera for practical purposes. Matthias Kümmerer and Matthias Bethge [1] show that all leading gaze prediction models utilize transfer learning[1] from other problem domains, primarily object recognition. There is a greater volume of object recognition data than gaze recognition data, and so transfer learning from the object recognition domain can improve gaze prediction performance without additional data collection.

The object recognition task deprioritizes visual effects or style, because those do not meaningfully alter outcomes when completing the task, and so prioritizes candid photography. This prioritization does not hold true for our application in gaze prediction, in which we may encounter stylized, illustrated, or computer-generated images and post-processing for aesthetic purposes. The assumption of candid photography, implied by most of the training data a gaze

[1]Transfer learning, generally defined, is a term for retraining a deep learning model designed for one task on another, similar task.

prediction model will see, is a concern for the model's performance when generalizing to stylized images.

A study by Zhaohui Che, Ali Borji, Guangtao Zhai, Xiongkuo Min, Guodong Guo and Patrick Le Callet [2] shows that many digital image transformations influence gaze behavior nontrivially. Visual media produced for aesthetic purposes utilize several post-processing effects with similar characteristics to the digital transformations studied by Che et al. The study supports the concern that deep-learning gaze prediction models may not generalize well to visual media applications.

Using the data the Che et al. collected from human subjects in a new study, our work measures the performance of state-of-the-art gaze prediction models on images which have been transformed with common digital image transformations. We find that the prediction accuracy degrades significantly for almost all transformations, compared to accuracy on untransformed reference images, which strongly indicates that state-of-the-art, deep learning models are biased towards candid photography.

Improving the performance of gaze prediction models for transformations will require more training data, but the space of possible transformations is vast. In order to prioritize training and data gathering efforts, a computational heuristic that indicates a loss in performance after a transformation–without the need for human subjects–would be very valuable for exploring the space of possible transformations. Searching for such a heuristic, we correlate derived metrics gathered from the images against the performance of the model on the transformed images. Unfortunately, we find that there are only a few weak relationships–none that are a strong and general indicator for a loss in performance after a transformation.

II. RELATED WORK

We build on the work of Che et al. [2], who studied digital transformations such as cropping, rotation, skewing, and edge detection on gaze behavior. They were motivated by the possibility of augmenting gaze prediction datasets with transformations that produced a significant difference in the image, but not in the gaze behavior of human subjects, which would allow the gaze fixation data to be reused for a new, transformed image. They analyze the differences in the gaze distributions that they collected. Our study uses the data they collected on image transformations to evaluate performance on transformed images for state-of-the-art gaze prediction models which were not published at the time. We are primarily concerned with analysis of the differences between predicted and real gaze distributions, rather than between the real gaze distributions of various transformations.

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 of the gaze prediction field we build upon are reviewed comprehensively by Kümmerer et al. [1], for which we will provide a brief overview.

The most popular method for representing gaze measurements and predictions for images is with a two-dimensional field spanning the image area. This is usually represented using a grayscale image called a “saliency map”. In the process of studying an image, the human eye will occasionally jump to a new fixation point in what is called a “saccade”. We define a saliency map such that the 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 on. We can further define our saliency map by normalizing it to a probability distribution, such that each pixel has a probability of being the next fixation point.

At this point, our saliency map would be “speckled”, with scattered points of high intensity and all other locations being low intensity. This saliency map 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”. Note that 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.

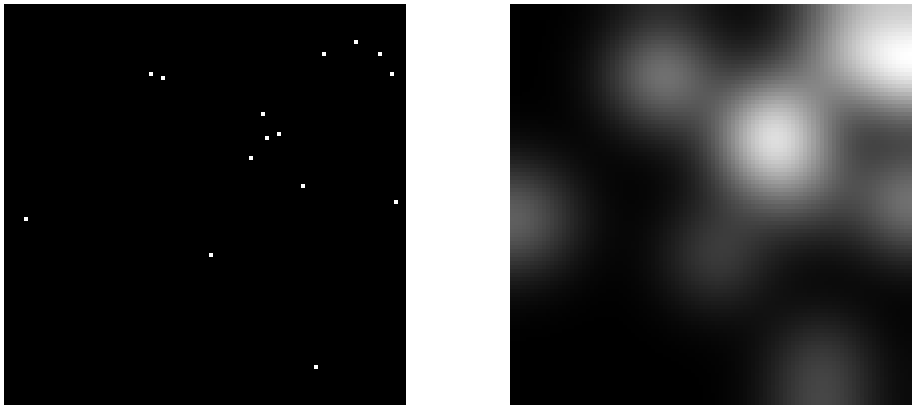

Figure 1: An illustrative example of a 'speckled' saliency map (left) with 15 randomly placed fixation points, and a smoothed saliency map (right) produced from the other saliency map by a regularization step using a gaussian blur.

If we perform these regularization steps on collected fixations from human subjects for an image, the resulting regularized saliency map is our best proxy for the real gaze distribution. This is referred to as the “gold standard” by Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge [4].

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 a metric for distance between a predicted saliency map and a gold standard, or a metric for the distance of both from another reference point, one can gain a sense for how well the predicted saliency map matches the real gaze distribution.

If the gold standard is an upper reference point, we wish also to find a lower reference point for a more complete picture of the prediction’s accuracy. For a saliency map to be a lower reference point, it should account for 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 any individual image’s contents. This will allow us to see what information the gaze prediction model can inference from an image directly, beyond any general assumptions the model might make based on previous experience with the dataset or class of images.

For datasets exceeding a certain sample size, an adequate lower reference point can be computed by collecting all fixation points for the entire dataset of interest at once and regularizing similar to the gold standard. The process will produce an average of all fixations across all images in the dataset. 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 center bias will account for dataset-wide biases, and as the number of images in the dataset increases, the weight of any individual image’s fixations will trend towards zero.

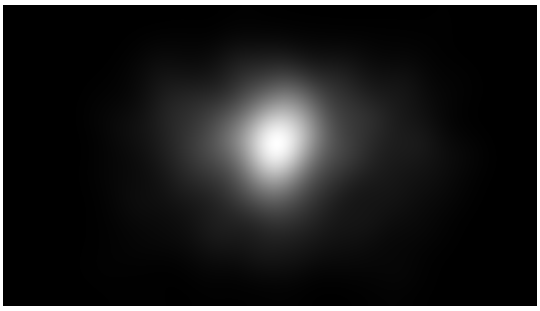

Figure 2: The center bias computed by summing all fixation points for the reference group of images in our dataset and blurring; it is provided as an illustrative example.

Using the collected fixation points, gold standards, center biases, and metrics for comparison compiled by Zoya Bylinskii, Tilke Judd, Aude Olivia, Antonio Torralba and Frédo Durand [5], we can measure the accuracy of a model’s predictions. Each metric is either “location-based” or “distribution-based”, as Bylinskii et al. call them, meaning they either compute a score between a saliency map and fixation points, or between two saliency maps, respectively. We evaluate the qualities and usefulness of each metric as it pertains to our study in the “Method” section. Using these metrics, we can conclude an increase in accuracy of a model’s predictions when they trend towards those measured from the gold standard and exceed those measured from the center bias.

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 (from Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge) [7], and a runner-up with smaller memory footprint and faster inference speed, UNISAL (from Richard Droste, Jianbo Jiao, and J. Alison Noble) [8], as our state-of-the-art models for our study. Unfortunately, we were unable to replicate the performance expected from the

DeepGaze IIE model. We decided to continue our study using only the UNISAL model. More details are described in the “Results” section.

Kümmerer et al. [4] describe the use of cross-validation for enhancing the accuracy of the center bias or the gold standard. As we show in the “Method” the enhancement is marginal and our center bias still performs closely with the center bias on the MIT/Tuebingen Saliency Benchmark, and so we omit the cross-validation step in the interest of simplicity.

They also describe using a leave-one-out policy for the center bias. When using a center bias a lower reference point prediction for an image, they leave out the fixation point data tied to that image when computing the center bias, and use only the fixation points from the rest of the dataset. This will ensure that the center bias does not include any information specific to the image in question, and ensure a better lower reference point.

As we will describe in the “Method” section, we compute center biases using fixations gathered for 100 images. If we omit the leave-one-out policy, the error between our center bias and that of the leave-one-out policy will be within around 1% error of each other, due to the summing and normalizing over 100 images. We decide that this difference is negligible, and so we omit the leave-one-out policy, once again in the interest of simplicity.

IV. METHOD

We aim to measure the accuracy of gaze prediction models on both reference images and their transformed counterparts, and analyze the differences in accuracy between the two.

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. Gaze fixation points are recorded for each image. See figures 3 and 4 for examples of all transformations.

Figure 3: Examples of the Cropping_1 (second) and Cropping_2 (third) transformations applied to the reference image (first).

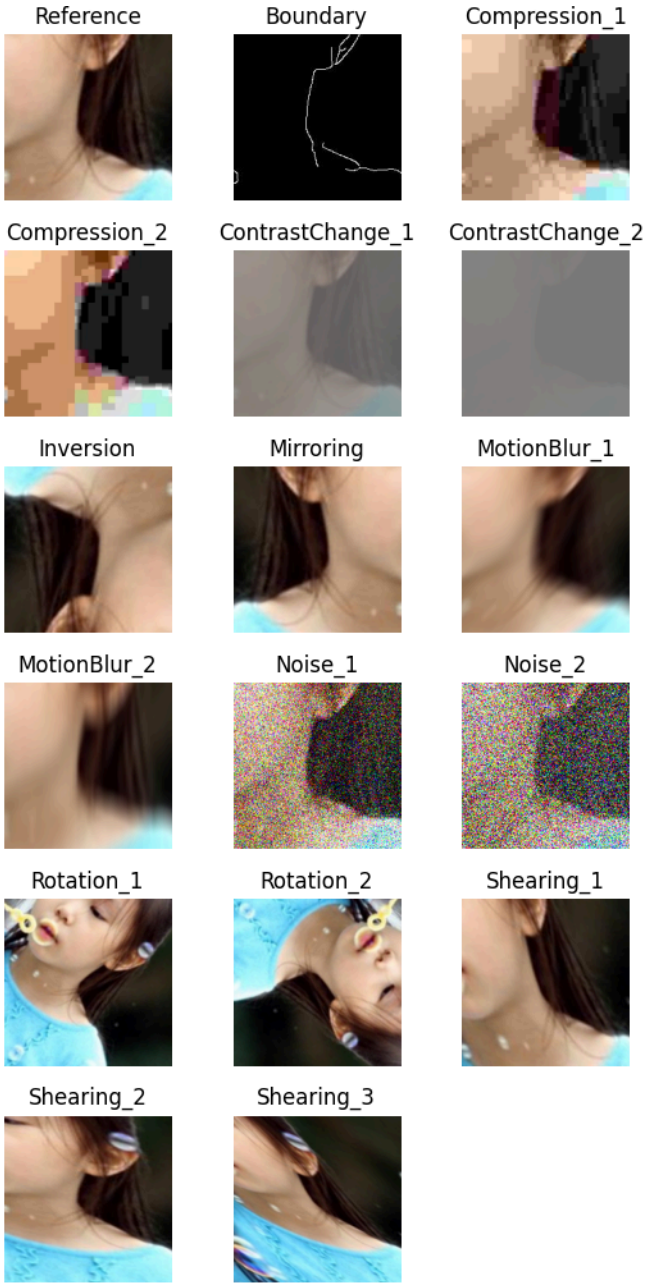

Figure 4: A slice of one of the images in the dataset, along with applications of all transformations except for Cropping_1 and Cropping_2, which are shown in figure 3.

There will be two steps to our study. First, we wish to compare the prediction accuracy of the models between the reference and transformed images. We hypothesize that the model's performance will be degraded as images are transformed.

Some transformations are similar to others in all but intensity, i.e. the `ContrastChange_1` and `ContrastChange_2` transformations. Although we do not have strict measures for the relative intensity between two transformations, we will still plot the prediction accuracy of similar groups of transformations

in ascending order such that we can reveal trends in the effect that a transformation will have as intensity increases.

Second, we wish to find a heuristic for expected prediction accuracy loss for a transformation, without requiring human subject data. We will collect a set of relevant metrics images we can derive from a source gaze distribution dataset and transformations upon that dataset, and compute their correlation to the prediction accuracy of the models on the transformed images. This leads us to examine the metrics compiled by Bylinskii et al. [5] for both the first and the second step.

We will need both of what Bylinskii et al. [5] call “location-based” and “distribution-based” metrics. Location-based metrics compare a saliency map to a set of fixation points, and distribution-based metrics compare two saliency maps. We will select location-based metrics for the first step of our study, where we must evaluate the performance of models given a set of fixation points, because location-based metrics require fewer parameters to configure. For the second step of our study, the comparison between the saliency maps produced for both reference and transformed images may be a valuable heuristic, and so we will select distribution-based metrics as well.

Listing the metrics considered, we see the area-under-the-curve (AUC-Judd) metric [10], the shuffled area-under-the-curve (sAUC) metric [11], the normalized scanpath saliency (NSS) metric [12], the information gain metric (IG) [4], the Earth Mover’s Distance metric (EMD) [13], as well as image-based versions of histogram similarity (SIM), correlation coefficients (CC), and Kullback-Leibler divergence (KL).

We wish to isolate the most relevant metrics for our study. With relevant metrics, we can evaluate prediction accuracy and compute correlations. 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.

We decide against the AUC metric because it is invariant to monotonic transformations, and has been deemed to be relatively saturated and uninformative in benchmarks compared to other metrics by Bylinskii et al. 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. Additionally, it is highly rank-correlated to the NSS and CC metrics, which means that relationships involving the NSS and CC metrics are likely to be present with SIM metric as well.

We decide against the EMD metric because it is computationally expensive, and because it is also highly rank-correlated to the NSS and CC metrics.

We select the remaining metrics: NSS, CC, IG, and KL. NSS and IG are location-based, while CC and KL are distribution-based. NSS and CC have been likened as the discrete and continuous analogs of each other, respectively, as a similarity metric. Meanwhile, the IG and KL metrics utilize similar information-theoretic foundations. IG is favored, because it provides a comparitive measure against a baseline (the center bias), but it also has the limitation that it does not provide a meaningful measure for the center bias itself.

Additionally, we will use the structural similarity index (SSIM) [14] metric to compare the difference not between saliency maps but between images, before and after a transformation. We hypothesize that a measure of difference before and after a transformation may also be a valuable heuristic for expected prediction accuracy loss.

Before we can begin evaluation of the models, we must compute the center biases and gold standards for the dataset. We compute the center bias for the reference and all transformations separately, collecting all fixation points for the 100 images of each and applying a Gaussian blur with a kernel size and sigma value of 57 pixels, which would be one degree of visual angle during the data collection according to Che et al.

At this point, we test to confirm that our center bias performs as expected. We compare against the MIT/Tuebingen Saliency Benchmark reported as 2.0870 on the NSS metric (one of the metrics compiled by Bylinskii et al. [5], which we will cover in more detail shortly) for their center bias on the CAT2000 dataset. Our centerbias achieves a mean NSS of

approximately 2.0665 on the untransformed image set. The error between the two scores is under 1% of the expected value. If we find the range between the gold standard and the center bias listed on the Benchmark (0.6559), which is the range we expect our model prediction scores to fall within, the error between our center bias and the listed centerbias would be 3% of that range. Additionally, our center bias scores lower, and so is more conservative for the purposes of determining whether transformations have significantly degraded the model's prediction accuracy. We decide that this error range is acceptable, and continue without the using the techniques described in the "Background" section to improve our center bias score.

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

Figure 5: Left is the gold standard for the image shown in figures 3 and 4, right is the centerbias for the reference group in the dataset.

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, which matches the aspect ratio of the original 1920x1080 image, 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 [15] resolution), 384x288 (which matches the SALICON dataset [16] resolution), and 384x216 (which preserves the aspect ratio of the original image). We run inference for each model at each resolution, and intend to select the best-performing resolution for each model.

For the first step of our study, we will compute our location-based metrics (NSS and IG) for each transformation set, and compare the average prediction accuracy of the models on the transformed images to the average

prediction accuracy of the models on the reference images. We will also plot the prediction accuracy for transformations that are similar to each other at varying intensities.

For the second step of the study, we will compute the SSIM 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.

We recognize that these metrics are not an exhaustive list of all relevant characteristics of the transformation or predictions, and we task future studies with enumerating metrics with possible relationships more thoroughly.

We collect the five metrics mentioned above as our independent variables. Our dependent variables will be the NSS and IG metrics for the transformed saliency map the model produced, which measure the model's prediction accuracy 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 plot the 10 pairs and compute 10 correlation coefficients for each transformation. We will interpret any correlation coefficient above 0.5 as significant, and proceed to interpret the applicability of each significant relationship on a case-by-case basis.

As we plot the data, we find that some outliers exist. We filter 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 computing a measure of how likely it would be that a relationship for our heuristics arose due to a non-representative sample of images, such as a p-value, would allow greater confidence in our results. In order to compute a p-value for our heuristics, we must determine the likelihood of a given image being representative of a class of images which we wish to study. Defining rigorous distinctions with which to isolate the class of

images relevant to our motivating use cases is an extraordinarliy difficult task, which we will leave for future work.

Instead, we will limit our claims on potential heuristics: any heuristics found only indicate a likely loss in accuracy for images found in the CAT2000 dataset or similar gaze prediction datasets, from which Che et al. have sampled 100 images at random. We argue that our sample size is large enough to provide a reasonable basis for our study.

We take efforts to ensure our study is reproducible. We publish our code at our repository on Codeberg [17].

V. RESULTS

When averaging performance increases for all transformation image sets, and additionally when only considering the untransformed set, we find that inferencing UNISAL at a resolution of 384x224 is optimal, and the same is true for inferencing DeepGaze IIE at a resolution of 1024x576.

For the untransformed image set, we find that the UNISAL model performs similarly to the expectation set by the MIT/Tuebingen Saliency Benchmark for the CAT2000 dataset when considering the IG metric. The Benchmark lists a score of 0.0321, while our inference achieves a score of approximately 0.0381. This results in an error of about 18% of the expected value, but for the error is less than 1% of the range between the gold standard and the center bias (0.8026). This means that it is more likely that the high error percentage relative to expected value is due to a low IG score (and higher variance at smaller scales) than unexpected performance of the model.

For the NSS metric, UNISAL outperforms the expectation set by the MIT/Tuebingen Saliency Benchmark. The Benchmark lists a score of 1.9359, while our inference achieves a score of approximately 2.1563. This results in an error of about 11% of the expected value, or about 33% of the range between the gold standard and the center bias (0.6559). Besides noise, speculative differences in how the data was collected, or errors that elude us in our inference code or measurement process, the reason for this unexpected increase in prediction accuracy is unclear. Nevertheless, we do not believe the error is large enough that we will not obtain meaningful results from continuing our study with the UNISAL model. Additionally, because the UNISAL model is outperforming expectations rather than underperforming, it is more likely to give us a conservative estimate in the case that prediction accuracy will be degraded by a transformation.

We were unable to replicate the prediction accuracy expected from the DeepGaze IIE model, and despite our best efforts to follow the protocol outlined in the DeepGaze IIE paper, the model's performance was significantly worse than expected. The Benchmark lists a score of 0.1893 for

the IG metric, while our inference achieves a score of −0.9402, which is signifcantly worse than the center bias. The error is 597% of the expected value, or 141% of the range between the gold standard and the center bias (0.8026). For the NSS metric, the Benchmark lists a score of 2.1122, while our inference achieves 1.5638, once again worse than the centerbias. The error is 26% of the expected value, and 84% of the range between the gold standard and the center bias (0.6559).

We have reached out to the authors of DeepGaze IIE so that we may double check our inference code for errors or misconceptions, but have not heard back yet. The performance of the DeepGaze IIE model was so far below expectations that it risked introducing noise into our results in the following steps of the study, and so we decided to continue our study only using the UNISAL model.

Though we found that UNISAL performs as expected for untransformed images, we find that it performs worse for transformed images. All transformations except for `Mirroring` cause the model's prediction accuracy to fall below that of the center bias by a significant percentage of the range between the gold standard and the center bias. In addition, we find an increase the intensity of the transformation leads to a loss in prediction accuracy. See figures 6 and 7.

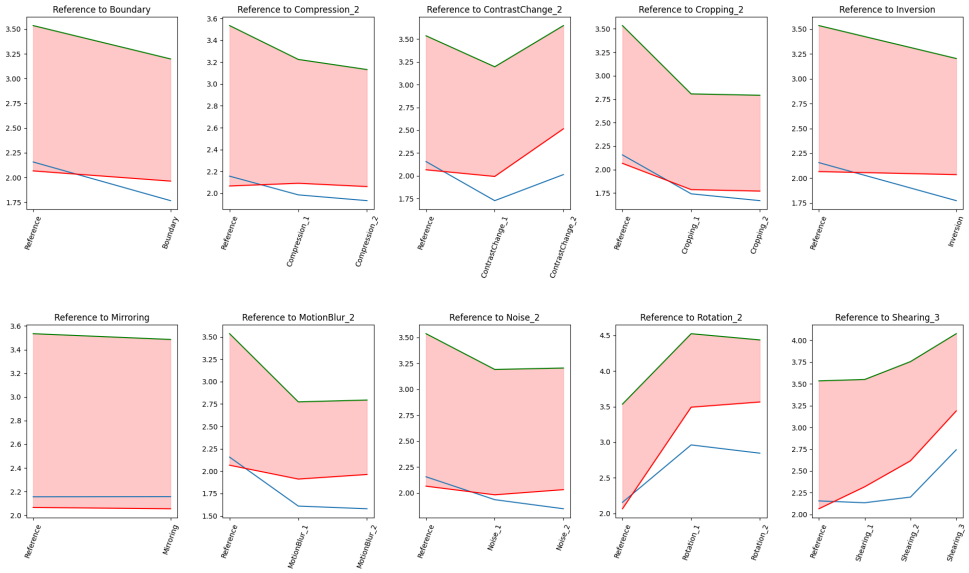

Figure 6: A plot of the NSS metric before and after transformation, as well in order of increasing intensity for those transformations which are similar to each other. Lower red lines are center bias NSS metrics, and upper green lines are gold standard NSS metrics. The red colored region denotes the range between the gold standard and the center bias.

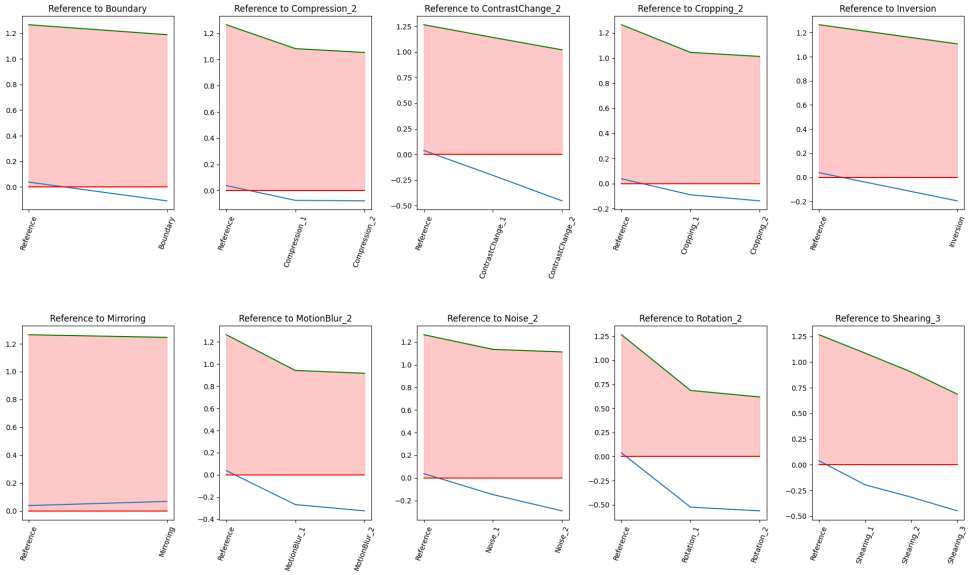

Figure 7: As with figure 6, but plotting the IG metric. We plot before and after transformation, as well in order of increasing intensity of transformations. Lower red lines are center bias IG metrics, and upper green lines are gold standard IG metrics. The red colored region denotes the range between the gold standard and the center bias.

These results confirm our general hypothesis that digital transformations will degrade the model's prediction accuracy. However, for the specific case of the `Mirroring` transformation, the model's prediction accuracy is unaffected,

and may even increase by a marginal amount. We have shown that models will require additional training in order to perform well on transformed images; now we hope to find heuristics that will allow us to quickly explore for transformations which will require additional training.

For the second step of our study, we find that for most transformations there exists a strong correlation (above 0.5 correlation coefficient) between reference image NSS performance and transformed image NSS performance, as well as reference image IG performance and transformed image IG performance. For the `ContrastChange_1`, `ContrastChange_2`, `Rotation_2`, and `Shearing_3` transformations, the correlation coefficients fall below 0.5. The IG correlation coefficients tend to be a weaker than the NSS, and as such they fall below 0.5 for the `Boundary`, `Rotation_1`, and `MotionBlur_2` transformations as well. See figures 8 and 9.

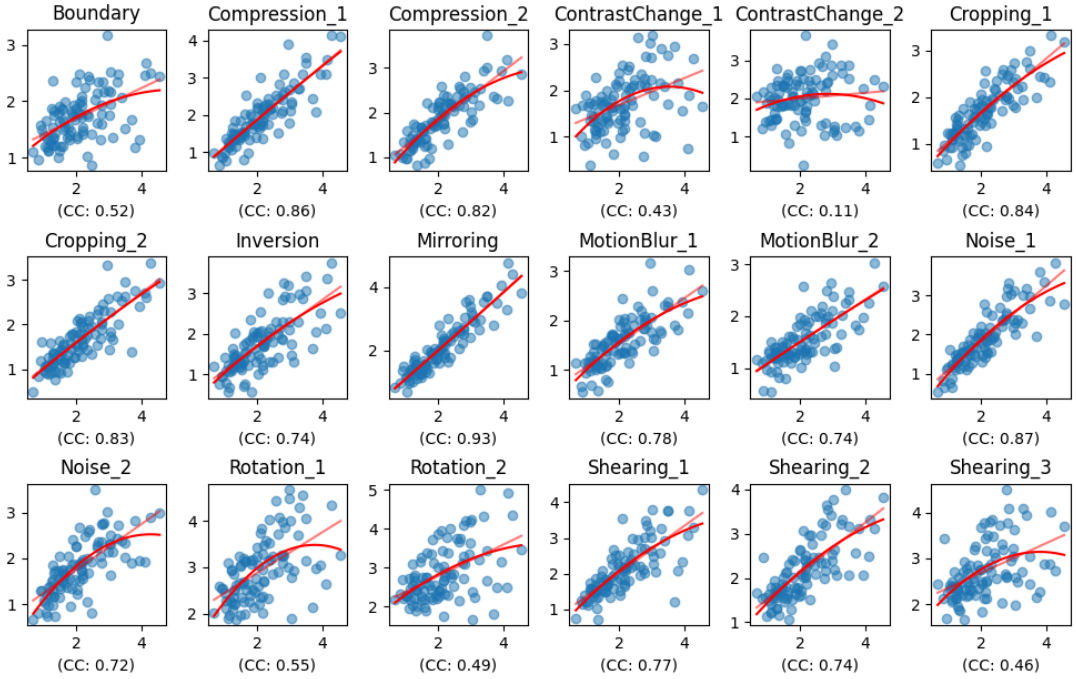

Figure 8: A scatter plot where each point represents an image, with the x-value being the model's NSS metric for its prediction on the untransformed image, and the y-value being the same for the transformed image. We plot lines and paraboles of best fit for each plot, and we compute the correlation coefficient as listed below each plot.

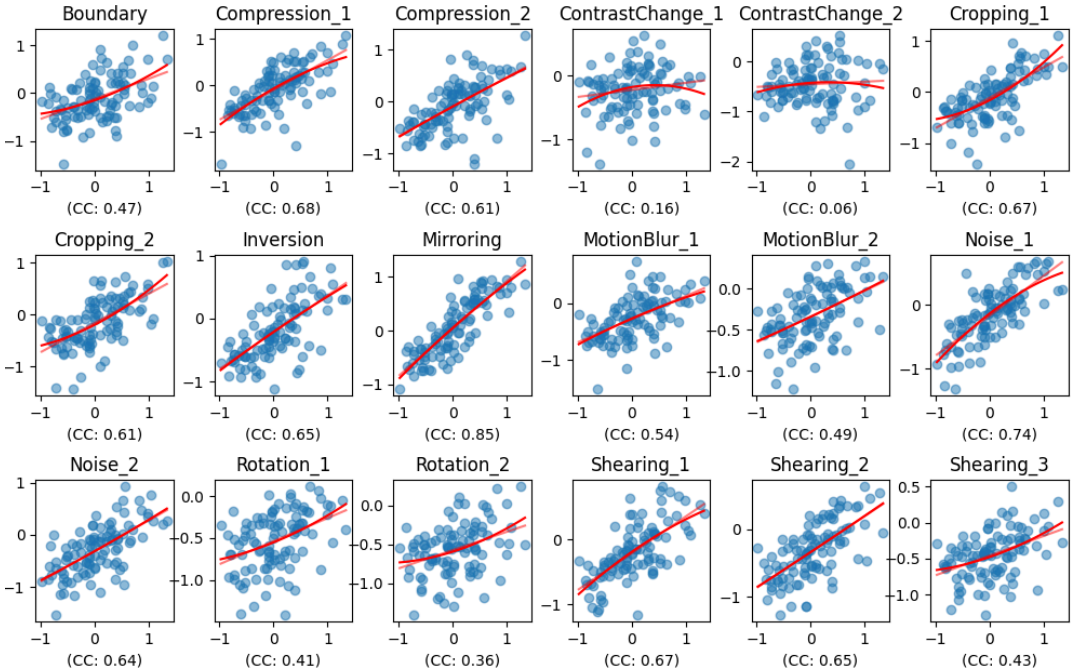

Figure 9: As with figure 8, but plotting the IG metric. We plot points, lines and paraboles of best fit for each plot, with the x-value being the IG metric for the model's prediction on an untransformed image, and the y-value being the same for the transformed image. We compute the correlation coefficient as listed below each plot.

Although the `ContrastChange_1`, `ContrastChange_2`, `Rotation_1`, `Rotation_2`, `MotionBlur_2`, and `Shearing_3` transformations, which saw weak correlation coefficients, also performed noteably poorly on average in the first part of the study (as seen in figures 6 and 7), we cannot find strong connection between the performance results of the first step and the correlation coefficients of this second step of our study. For example, the `Boundary` transformation did not perform particularly poorly in the first part compared to other transformations, and yet saw a weak correlation coefficient.

Figures 8 and 9 also plot paraboles of best fit for the data. We find that, aside from the the `Boundary`, `Cropping_1`, and `Cropping_2`, `Rotation_1`, `Rotation_2`, and `Shearing_3` transformations for the IG metric, which have weak positive coefficients for the quadratic term, all paraboles have a negative coefficient for the quadratic term. For both metrics, we see particularly high negative coefficients for those transformations which have the weakest correlation coefficients.

These results tell us that, for most transformations tested, an increase in prediction accuracy on the reference image leads to an increase in prediction

accuracy on the transformed image. It does not appear to be at a linear rate, however, and there is varying steepness in curves of diminishing returns. The contrast change transformations seem not to follow this trend, and have very little relationship between the reference and transformed image prediction accuracy. The boundary, rotation, motion blur, and shearing transformations a weaker relationship, but fall off the curve of diminishing returns quickly.

For the `ContrastChange_1`, `ContrastChange_2`, `Rotation_1`, and `Rotation_2` transformations, we find that there does exist a weaker relationship (hovering around 0.5) between the CC and KL metrics between the untransformed prediction and the transformed prediction to the NSS metric for the transformed prediction. This relationship is weakened to the point of insignificance when considering the IG metric. See figures 10 and 11.

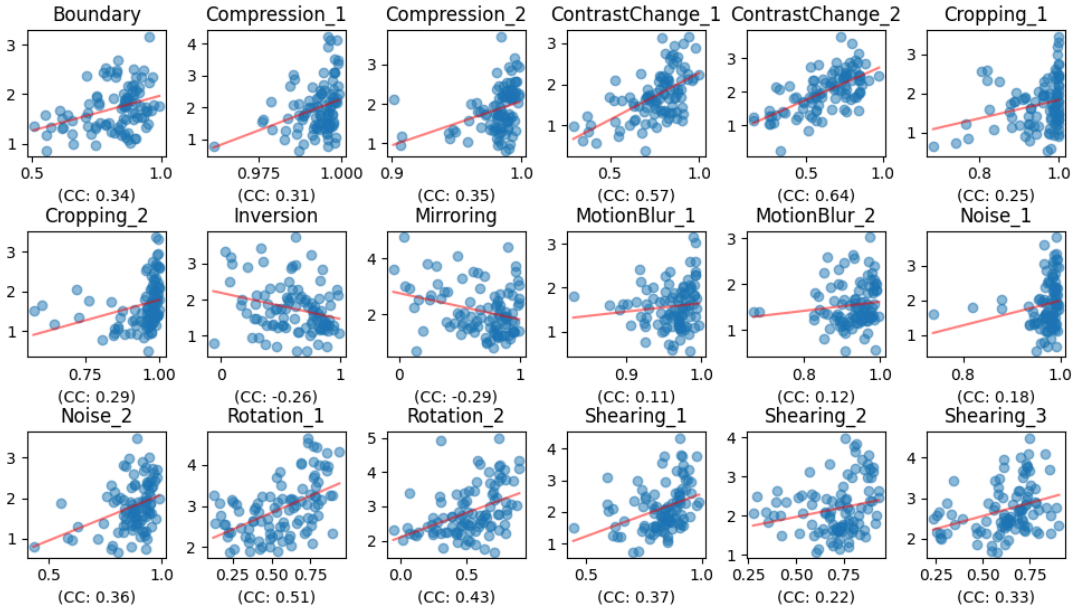

Figure 10: We plot points and lines of best fit for each plot, with the x-value being the CC metric between the predictions for the untransformed and transformed images, and the y-value being the NSS metric for the transformed image. We compute the correlation coefficient as listed below each plot.

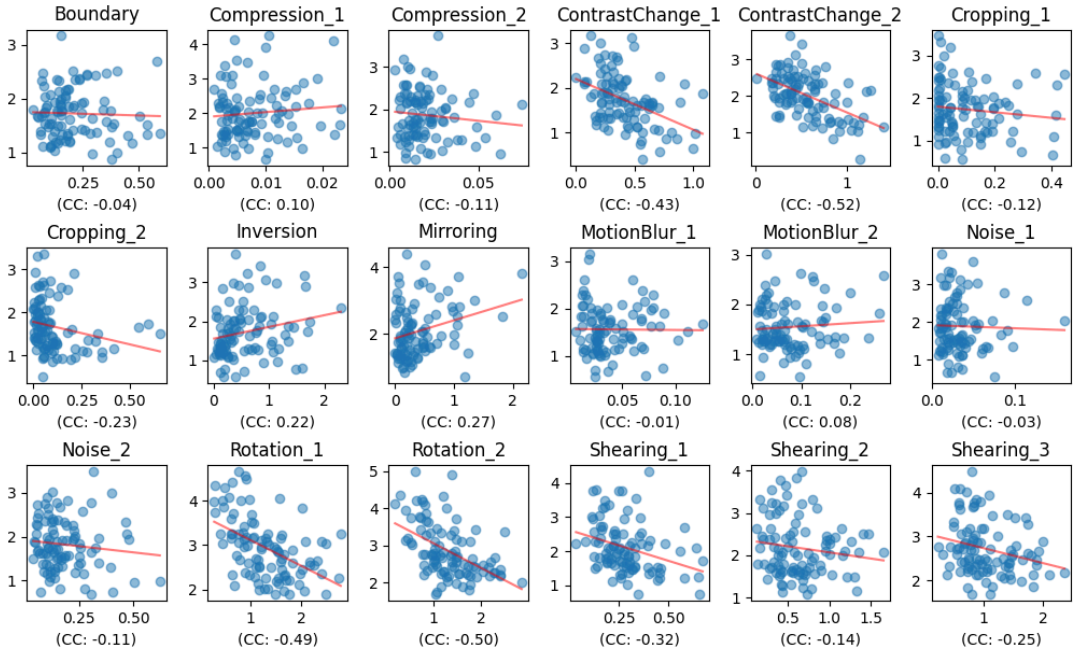

Figure 11: As with figure 10, but plotting the KL metric. We plot points and lines of best fit for each plot, with the x-value being the KL metric between the predictions for the untransformed and transformed images, and the y-value being the NSS metric for the transformed image. We compute the correlation coefficient as listed below each plot.

These results tell us that, for the contrast change and rotation transformations specifically, if their prediction for the transformed image looks significantly different from their reference prediction, then it is slightly more likely that the transformed prediction will perform poorly.

For the contrast change transformations, we might intuitively explain the effect described above by the fact that the salient image features have not changed locations after the transformation, nor has their emphasis relative to other features changed, and so the salient regions will remain the same. For the rotation transformations, it is more difficult to explain why this effect is present, but it may be due to the center bias present in images, which does not change location under rotation.

Further experiments are required for determining what properties of the contrast change and rotation transformations lead to this effect, and whether this effect is a characteristic of transformations in general. In absence of these experiments, 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 only.

We have saved all metrics computed in our study in the `results` directory of our repository on Codeberg [17], along with the code used to compute them and produce visualizations in the same repository.

VI. CONCLUSION

For all transformations except for the `Mirroring` transformation, the UNISAL model performs worse than the reference set of images. Increasing the intensity of the transformation leads to further loss in prediction accuracy. Even so, there is still a correlation between prediction accuracy on a reference image and prediction accuracy on a transformed image, except for the contrast change transformations. We find that for those transformations with a weaker correlation, the data is better modeled by a curve of diminishing returns, where improvements in prediction accuracy for transformed images do not keep up with improvements in prediction accuracy for reference images.

We find that the image-based correlation coefficient or the KL-divergence between the predictions for an untransformed image and its transformation is a weak heuristic for predicting the performance of a model only for 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.

Our work indicates that current state-of-the-art gaze prediction models are likely to have biases present in their training data towards candid photography. Extrapolating from the fact that models struggle with some common digital transformations, and that digital post-processing is common for digital visual media, we hypothesize that models will perform poorly on stylized images, or images which have been altered for aesthetic purposes.

For future work, we would like to test a greater number of transformations, including digital distortions, color manipulations, and stylistic filters, or compositions of all of the above, which are other common digital transformations used in visual media. We might also test compilations of stylized images, rather than pairs of images and their transformations.

We would like to test more state-of-the-art models, including DeepGaze IIE if expected performance can be replicated, to see if there is divergence in how they behave when transforming images.

Finally, 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.

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

[10] T. Judd, K. Ehinger, F. Durand, and A. Torralba, “Learning to predict where humans look,” in *2009 IEEE 12th International Conference on*

Computer Vision, 2009, pp. 2106–2113. doi: 10.1109/ICCV.2009.5459462.

[11] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision Research*, vol. 45, no. 5, pp. 643–659, 2005, doi: https://doi.org/10.1016/j.visres.2004.09.017.

[12] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision Research*, vol. 45, no. 18, pp. 2397–2416, 2005, doi: https://doi.org/10.1016/j.visres.2005.03.019.

[13] Y. Rubner, C. Tomasi, and L. Guibas, "The Earth Mover's Distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, pp. 99–121, 2000.

[14] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004, doi: 10.1109/TIP.2003.819861.

[15] W. Wang, J. Shen, F. Guo, and A. Borji, "Revisiting Video Saliency: A Large-Scale Benchmark and a New Model," 2018, pp. 4894–4903. doi: 10.1109/CVPR.2018.00514.

[16] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in Context," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015.

[17] J. Youngblood, "Transformations and Gaze Prediction." Accessed: Oct. 06, 2025. [Online]. Available: https://codeberg.org/soundeffects/transformations_and_gaze_prediction