The Consistency of Visual Search Models on High Dynamic Range and Tone Mapped Images

Andre Harrison, Michael Green, Chou Hung and Adrienne Raglin

November 7, 2019

Symposium: 24th International Command and Control Research and Technology Symposium

Topic 9: Experimentation, Analysis, Assessment and Metrics

Title: The Consistency of Visual Search Models on High Dynamic Range and Tone-Mapped Images

Authors:

Andre V. Harrison
U.S. Army CCDC - Army Research Lab, FCDD-RLC-IB
2800 Powder Mill Rd, Adelphi, MD 20783
andre.v.harrison2.civ@mail.mil

Michael A. Green
U.S. Army CCDC - Army Research Lab, FCDD-RLC-IB
2800 Powder Mill Rd, Adelphi, MD 20783
michael.a.green85.ctr@mail.mil

Chou P. Hung
U.S. Army CCDC - Army Research Lab, FCDD-RLH-FC
7101 Mulberry Point Rd, Aberdeen PG, MD 21005
chou.p.hung.civ@mail.mil

Adrienne J. Raglin
U.S. Army CCDC - Army Research Lab, FCDD-RLC-IB
2800 Powder Mill Rd, Adelphi, MD 20783
adrienne.raglin2.civ@mail.mil

The Consistency of Visual Search Models on High Dynamic Range and Tone-Mapped Images

Abstract:

Tone mapping operators (TMO) are compression algorithms that compress the bit depth of high dynamic range (HDR) images in such a way that when the tone mapped version of the HDR image is shown on a low dynamic range screen, large portions of the image don't appear over/under-exposed. But the process of reducing the bit-depth of an image often alters the appearance of the image due to the loss of information and the introduction of artifacts, changing the gaze patterns when people look at those tone mapped images. Saliency-based TMOs aim to compress the bit-depth of an HDR image while trying to keep the most salient locations in the HDR image salient after tone mapping. However, these models don't ensure that the saliency model used is actually able to predict eye gaze in HDR environments accurately and assess how eye gaze is influenced by artifacts introduced into the image by the compression process. In this paper, we evaluate a suite of saliency models against 8 well-known TMOs and the original HDR image to see how well each saliency model can predict the change in eye gaze when different TMOs are applied. By doing this, we can establish a firm basis on which to develop a saliency-influenced tone mapping model to both compress HDR images and influence attention within the tone-mapped image. If TMOs can selectively choose what to emphasize or de-emphasize in a tone-mapped image based on saliency results, then saliency-based TMOs can be used to effectively direct attention even in complex environments.

1. Introduction:

Within a battlefield environment, there are many pieces of information coming from different sources within and related to the battlefield environment. All of these potentially relevant pieces of information must compete for the attention of commanders, analysts, and other decision makers. These information sources may have different impacts on the decisions and the decision quality those commanders and analysts make. As such, what information a commander or analysts sees first can have a significant impact on the decisions they make and their final decision-making process. The common operational picture (COP) is one environment where information from different sources is integrated together upon which commanders can make decisions. As the COP is further digitized through the use of Augmented Reality/Mixed Reality systems, more and more information must be curated by software agents rather than people when that information is added to the COP. Also, methods need to be developed to not just organize information based on the software agents estimate of its importance, but when decisions require human judgements the systems need to develop and utilize ways that can convey information prioritization through explicit and implicit means to the commanders and analysts who need to make the decisions.

Visual data (GEOINT, surveillance information, on the ground battlefield imagery, etc.) can often be densely packed with data that commanders can use to inform their decision-making process, however the interpretation of an image into information can be subjective especially for complex imagery. In other situations, the most relevant element within an image can be missed by the human viewer without the appropriate priming. To ensure the right information is taken from a given image, explicit interpretations of this information can be presented to the commander (in the form of audio dialogue, labels, drawings/annotations, or explicit region highlights). However, the use of these communication methods requires the software agent to have a high confidence in its interpretation of this data and the presence of these other cues will limit or prevent other information from being explicitly described so that the final image isn't overly cluttered. For image information where the software agent doesn't have a high enough confidence, if other (higher-priority) information is already explicitly indicated, or if the commander is too cognitively loaded to process more explicit information, other ways of promoting information within an image may be necessary. These less explicit methods aim to draw eye gaze to distinct locations within an image without covering up other information within an image or without being too distracting to other tasks. One well-known class of image processing techniques that has been used to preserve information in some locations while demoting or suppressing information in others are tone mapping operators.

Tone mapping operators (TMOs) are high dynamic range (HDR) compression methods used to convert HDR images into images with a standard dynamic range (SDR) so that when these images are shown on SDR displays, all of the most important information within the image is perceivable [1], [2]. Thus, by choosing what information to ensure is perceivable, they explicitly direct viewers' attention to different locations by ensuring those locations remain visible. However, TMOs have primarily been designed to only suppress the large changes in luminance that are typically considered unimportant in an HDR image while trying to maintain the perceived reflectance of the relevant objects in the image. These methods haven't historically been used to promote some locations within an image over others using a basis other than large luminance changes. The few TMOs that have been developed to compress an HDR image using a measure other than luminance variation have primarily used visual saliency models to select the different regions of an image that should be kept visible or have their visibility enhanced [3]–[5]. Models of visual salience try to estimate the likelihood that a location within an image will attract a person's attention. These models typically try to predict what is salient in an image based on how unique a location or object is within a scene or how relevant that location is relative to the viewer's task [6]–[9].

However, there are several issues with TMOs that use a more arbitrary method to promote the visibility of different objects in a tone-mapped image. It is not clear how well TMOs actually influence eye gaze and in what conditions do they seem to work. Secondly, most TMOs that incorporate saliency methods in their approach don't use saliency models that have actually been evaluated on HDR data to show how well those selected saliency models actually can predict eye gaze in an HDR image. Typically, they make the assumption that a saliency model that has been developed and tested on standard dynamic range imagery will work just as well on HDR imagery [3], [5]. However, if a saliency model could accurately predict eye gaze using different TMO algorithms, then saliency models could act as a feedback element for a tone mapping-based image information enhancement. The aim of this paper is to identify which models of visual salience are robust enough to predict eye gaze across a large set of TMOs and how well these models can predict eye gaze in HDR images. This represents a first step towards developing a system to iteratively refine methods that can covertly influence attention.

To this end we evaluate the performance of several saliency models on the ETHyma dataset to determine how well different saliency models are able to predict eye gaze in tone-mapped images across a set of tone mapping methods [10], [11]. In the following section, we briefly cover different approaches to modeling attention and influence attention in imagery. In section 3, we discuss the analysis we conducted on the dataset. In section 4, we discuss our results from this analysis. In section 5, we discuss the implications from our results, and in section 6, we provide concluding remarks.

2. Background

2.1 Predicting visual attention (Visual Salience)

Computational models of visual salience have been developed to estimate how likely each object or location within an image will attract a person's attention as evidenced through patterns of eye fixation. Models of visual salience were founded on the seminal work by Treisman and Geladé where they published their concept of Feature-Integration-Theory [12]. Building upon this, Koch and Ulman proposed that attention could be modeled using a winner-take-all calculation [13]. The most influential computational visual saliency model that is still utilized in saliency evaluations today is the Itti model by Itti, Niebur, and Koch, who developed a bio-inspired approach to modeling visual saliency [14]. Since then, there have been a plethora of saliency models, many of which take a bio-inspired approach [6], [8], [15], [16]. However, saliency models have also been developed based on information theory [14], [15], statistical methods [19], or graph-based approaches [20], to name some of the most popular non-deep learning methods.

Over the last few years, the performance of visual saliency models has started to approach the patterns of attention of human eye gaze due in large part to the advancements brought on by deep learning models [9]. These methods have been able to achieve this level of performance because the convolutional neural network (CNN) architectures within deep learning models are able to recognize higher-order visual concepts like objects and faces. Many of the latest deep learning-based visual saliency models are built on top of deep learning-based image classification models like AlexNet, VGG-16, and GoogleLeNet [21]–[23].

2.2 Tone Mapping

For over a decade, cameras have been able to capture imagery and video at a higher dynamic range than most displays can actually show. This was primarily through the development of multi-exposure stitching techniques, which made it very easy to faithfully capture images of scenes with high dynamic ranges (HDR) (>10^3:1) [24]. This disparity led to the development of several research papers on how to compress HDR images so that they could be better displayed on SDR displays; these methods are known as tone mapping operators (TMO) [2]. There are typically two categories of TMOs, global methods and local methods. Global TMOs compress HDR images so that there is a 1-to-1 relationship between the pixel values in the tone-mapped image and the pixel values in the original HDR image. Global TMOs are typically faster than local TMOs and are somewhat reversable if the mapping is preserved, but their ability to ensure the most relevant information is always visible can be limited due to this consistent mapping. Local TMOs are compression methods that use the local spatial patterns within the HDR image to identify what information to compress or preserve without requiring the 1-to-1 mapping be preserved. Due to their higher complexity, they can be more computationally-intensive in comparison to global TMOs, but they can also better adapt to the variation in luminance that may occur in an HDR image. However, due to the loss of the 1-to-1 mapping, local TMOs can also introduce visible artifacts into the tone-mapped image.

What defines an image as better is inherently a subjective concept and could have many different interpretations, though TMOs have typically only used one of three different criteria to evaluate their performance. How perceptually faithful is the tone-mapped version of the image in comparison to the original HDR version? How pretty or aesthetically pleasing is the tone-mapped image? Or, how much detail from the original HDR image is visible in the tone-mapped image? The first two criteria have received a lot of focus and attention during the initial development of TMOs due to their value to the scientific and entertainment communities, respectively. The criteria of detail maximization has received less focus and attention, but it is, however, the most relevant criteria for military tasks [25], though not all details are equally important. Several TMOs have been developed that are able to selectively enhance or maintain the visibility of certain regions within an HDR image, while neglecting other (presumably less important) regions. These TMOs have relied on computational models of visual saliency to select locations to be enhanced. By applying these models to HDR imagery, TMOs can enhance only the regions that are likely to be paid attention to anyway thereby reducing the amount of visual clutter that a person would need to filter. Several TMOs have utilized saliency models to tune their TMOs, however, most of those papers used saliency models that were developed and had only ever been evaluated on SDR imagery [3], [5], [26], [27]. These TMOs simply took existing SDR-developed saliency models and tried to adapt them to handle HDR imagery, but never evaluated how well those models were now able to predict eye gaze in an HDR image. Hence the locations selected by the visual saliency model to be enhanced by using the TMO may have been incorrect or not the only salient locations within the HDR image. Several papers within other areas in computer vision have found that existing and well-known algorithms within computer vision (Canny edge detectors, Blob filters, Harris corner detectors, etc.) perform poorly when applied to raw, linearly encoded, HDR imagery [28]–[30]. Only Lin and Yan [26] used a visual saliency model that was actually developed for and evaluated on HDR imagery, but even that model was only ever evaluated on a single HDR image [31], [32].

The absence of the analysis of how well different saliency models can actually predict eye gaze for an HDR image weakens the value they add to a TMO as the locations they indicate may not be the only locations of interest. Hopefully, the analysis provided within this paper will clarify how much (if any) work needs to be done to adapt existing SDR and HDR-based visual saliency models to predict eye gaze for both HDR imagery and their tone-mapped versions. We discuss this in the next section.

3. Analysis

The aim of this evaluation was to both assess how well different models of visual salience could predict eye gaze in HDR images, but also to see how well these models could predict how different TMOs influence eye gaze. This would close the loop of influence between TMOs and visual saliency models, as these models could predict what does attract a person's attention in an HDR image, and they can also serve as a method of evaluating how well different TMOs actually influence eye gaze in the final tone-mapped image.

The dataset we conducted our evaluation on was the Eye-Tracking on High dYnamic range iMAges (ETHyma) dataset [10]. The ETHyma dataset is the only publicly-available dataset of eye gaze information for HDR images and tone-mapped versions of each of those HDR scenes. The dataset consists of 11 HDR images and 8 tone-mapped versions of each image using 7 well-known tone mapping methods along with the linear (low dynamic range virtual photograph) version of each image as shown in **Figure 1**. As we have found with other HDR eye gaze datasets, saving an image in the .hdr format does not make the scene that was captured actually HDR. To ensure every HDR image was actually capturing an HDR scene, we use a simple threshold system to label if an image is really of an HDR scene by requiring any HDR image to have a higher luminance dynamic range than the maximum luminance dynamic range of all of the SDR images (tone-mapped images) in the dataset. Based on the authors experiments over several datasets, the maximum luminance dynamic range of an SDR images generally comes to contrast ratios of ($3*10^3$:1).

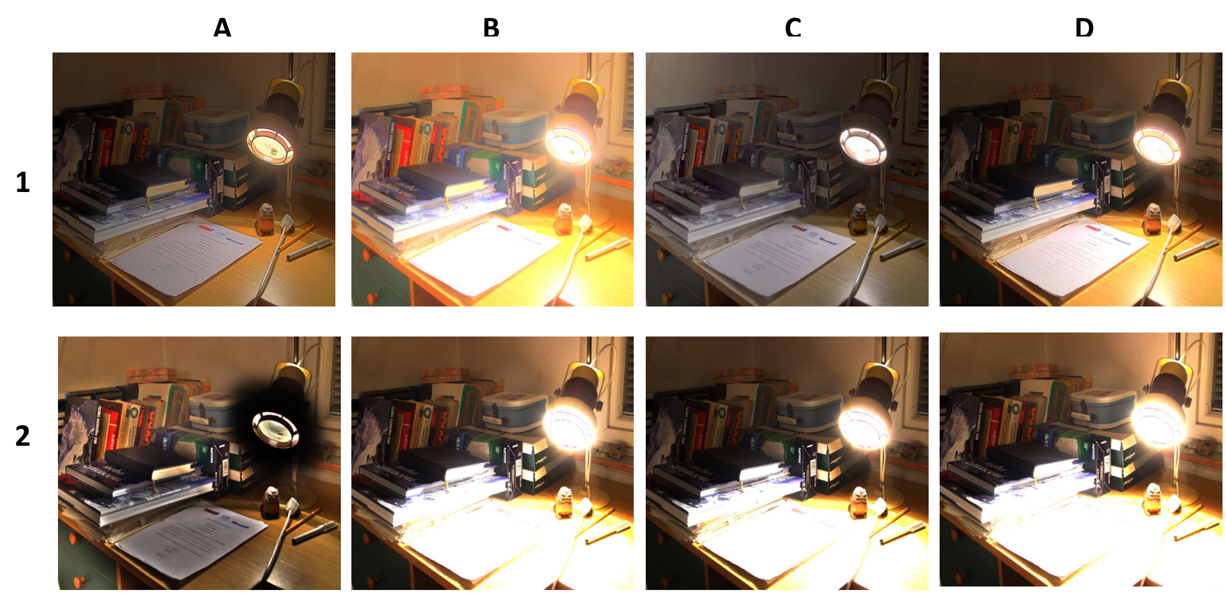

Figure 1. *Sample set of tone mapped images from the ETHyma dataset where an HDR image has been compressed using one of 8 different tone mapping methods. The eight methods are: 1A – Ashikhmin TMO* [33]*. 1B – Adaptive logarithm* [34]*, 1C – Image Color Appearance model (iCAM)* [35]*, 1D – Dodging and Burning TMO* [36]*, 2A – Chiu TMO* [37]*, 2B – Bilateral Filtering TMO* [38]*, 2C – Linearized image, 2D – Tumblin TMO* [39]*.*

Much like the different evaluation criteria that exist for TMOs, there are multiple measures that can be used to assess the predictive performance of a visual saliency model [40]. To get a fuller sense of how well different saliency models predict attention for different types of imagery, we used a couple of different evaluation measures. The two primary measures we used were the well-known area-under-the-curve-of-the-receiver operator characteristic (ROC-AUC) and the normalized scanpath saliency (NSS) measure. Both of these measures have been at one time the primary measure to assess the performance of a saliency model for visual data; currently, NSS is the preferred measure over the older ROC-AUC measure [40]. In our evaluation, we also include center bias-compensated versions of the ROC and NSS measure (shuffled ROC and shuffled NSS, respectively) [41]. Center bias compensation is often added to saliency analysis to compensate for the strong influence that the center of an image has on a person's gaze, regardless of what is present in the image. There are many factors that contribute to the higher-than- chance likelihood for a person to look at the center of an image, but regardless of the reason, simply assuming that a person will look at the center of an image can generate predictive accuracies much higher than chance (~70% for ROC-AUC), making the evaluation of saliency models more difficult. To reduce the influence of the center bias in the evaluation of saliency models, center bias compensation versions of the ROC-AUC and NSS measures were developed [41]. The result of using these approaches is that only predicting that a person will look at the center of an image is equivalent to chance for these two methods. However, as center bias is still a real effect, reducing it to chance actually over-compensates for center bias, preventing shuffled ROC-AUC and shuffled NSS from replacing their not-center-biased versions. For this reason, the results from using both measures are included and discussed in the following sections.

To better evaluate how well attention can be predicted across different presentations of the same image, we used a testbed of visual saliency models including well-known standard saliency models(the Itti model Itti [14], the Graph Based Visual Saliency [GBVS] model [20], the Judd model [42], the Ideal Observer Model [ioM][18], the Region Covariance Saliency model [COV] [43], the Proto-Object Saliency [Proto] [16], and Fast and Efficient Saliency [FES] model [44]), saliency models developed for data visualization (the Data Visualization model [DVS] [45] and the Visual Importance Model [VisImp] [46]), models developed for HDR imagery (the contrast feature [CF] model [32], and the learning based visual saliency [LBVS] model [47]), and deep learning-based visual saliency models (Ensembles of Deep Networks [eDN] [48], Deep Multi-Level Network [ML-Net] [49], inter-image Similarity and Ensemble of Extreme Learners [iSeel] [50], the Saliency Attentive Model [SAM] [51], and the Saliency with Generative Adversarial Networks [SalGAN] [52]). As in most other areas of computer vision research, deep learning models of visual salience typically have the best performance in predicting eye gaze on SDR imagery, but these models often don't transfer well when applied to data that differs in some way from the training dataset. Thus, deep learning models which have been trained on SDR images, may not generalize well to HDR images with a higher bit depth than 8-bits/color/pixel, making it unclear which class of model is likely to perform better on HDR images.

Within our saliency model test-bed, we have 18 different visual saliency models: 9 shallow models, 2 HDR-developed models, and 7 deep learning-based models. However, not all models of visual salience include a center bias inherently within their model or as a selectable parameter; for models that do not have an intrinsic center bias, we add one by multiplying the saliency output of the original model with a Gaussian kernel the size of the image centered at the center of the image. The x and y standard deviations we used come from [53]. However, not every model is enhanced by the presence of a center bias, and the shuffled variations of the ROC-AUC and NSS measures are likely to penalize models with a center bias. Thus, within our analysis, we have included results from models with and without center bias as indicated by the presence or absence of '_CB' in the results for a given model (i.e. {MODEL_NAME}_CB – for models with center bias.

4. Results

Of the 11 HDR images within the ETHyma dataset, 10 of them were actually of HDR scenes as they had a dynamic range higher than the maximum dynamic range of all of the SDR/tone-mapped images. As such, we applied our testbed of saliency models to only these HDR images within the ETHyma dataset, Figure 2. The saliency map outputs from each of the models reflects the differences in approach taken to predict eye gaze and their impact when applied to these HDR images.

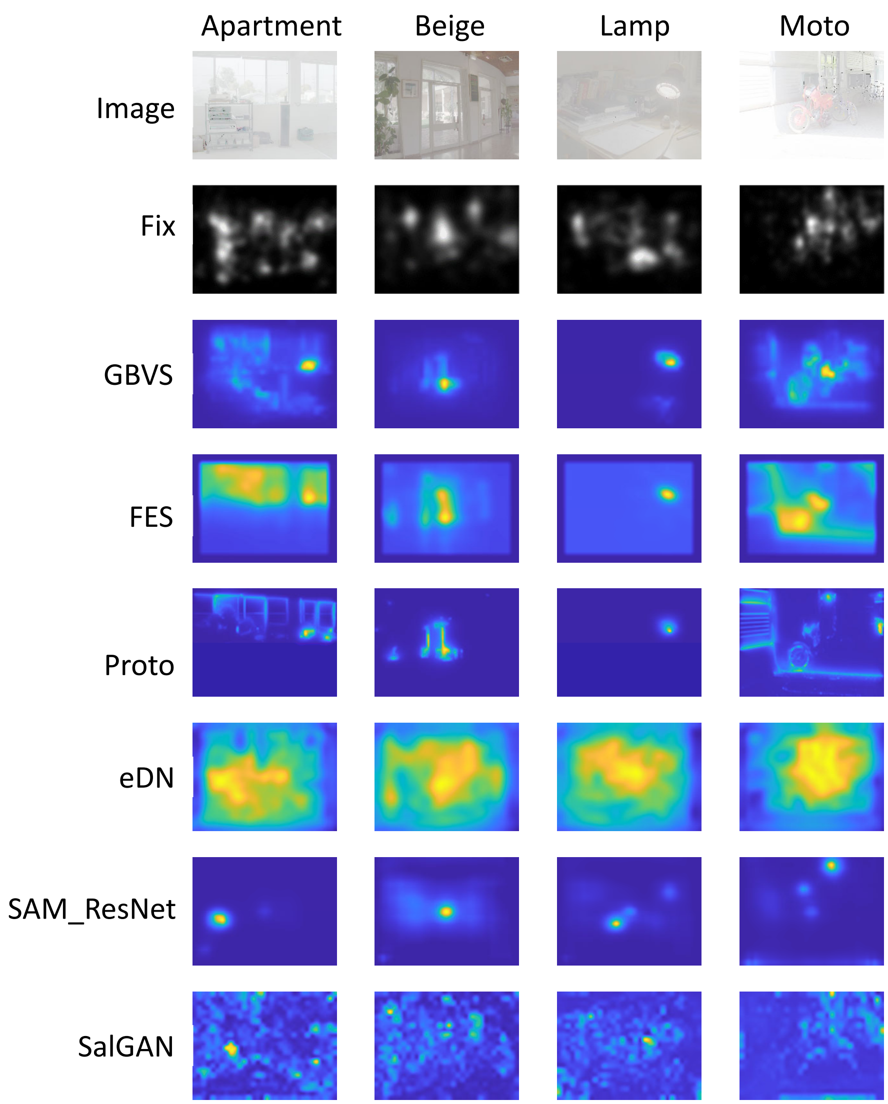

*Figure 2. Qualitative comparison of the saliency maps generated by a sampling of the saliency models used in the testbed. The top row shows the original HDR image. The second row shows maps of the cumulative eye fixation locations for that image over all subjects, where the brightness of an area is proportional to the amount or length of eye fixation. Rows 3 – 5 show the output from some of the top performing shallow visual saliency models. Rows 6 – 8 show the output from some of the top performing deep learning models. For rows 3 – 8 the saliency maps predict that a location is more or less salient as the colors go from yellow -> orange -> green -> blue. *(For display purposes the HDR image is tone mapped by taking the logarithm of the image)*

The outputs from each of the models reflects the differences in approach taken to predict eye gaze. Some of the model outputs are very selective in identifying a given region as salient, often in these maps only a single location with a small spatial extent has a high likelihood of salience, while all other regions are given a low likelihood. Other models have a more smeared saliency map still with 1 or 2 spatial locations of with a high salience value, but the spatial variance of that high likelihood region is wide, often marking more than a third of the image with a moderate to high likelihood of being salient. However, for many saliency models, regardless of whether they have a wide or narrow spread in their salience prediction, the brightest light source in the image is given the highest likelihood, even if the location at the brightest light source is fairly blank. In the fixation maps though, while the brightest locations are consistently a point of fixation within the maps, they do not have the brightest values.

A more quantitative evaluation of the visual saliency models within the testbed is shown in **Table 1** and **Table 2**. The tables contain the predictive performance of the best models within the testbed on the HDR and tone-mapped images within the ETHyma dataset. Due to the number of models within the testbed, and the with and without-center bias variation outputs for each model, only the models that had the best performance in one of the four measures for each model category and for one of the three image categories is shown in **Table 1** and **Table 2**.

		ROC(AUC)				Shuffled-(AUC)		
Best DNN		**TMO**	**HDR**	**log(HDR)**		**TMO**	**HDR**	**log(HDR)**
	Inter-subject	0.777902	0.780041	0.780041	**Inter-subject**	0.748655	0.759617	0.759617
	SALGAN_CB	0.726896	0.714999	0.725769	**SALGAN_CB**	0.740214	0.709718	0.730173
	iSeel	0.724808	0.712156	0.731444	**iSeel**	0.751268	0.709354	0.742304
	eDN	0.702193	0.694773	0.714297	**eDN**	0.737645	0.719698	0.73869
	SAM_ResNet	**0.753034**	0.718182	0.647699	**SAM_ResNet**	**0.776434**	**0.731065**	0.609421
	SalGAN	0.748906	0.716504	**0.741071**	**SalGAN**	0.704066	0.580008	0.614734
Best Shallow								
	GBVS_CB	0.706806	0.690336	0.727528	**GBVS_CB**	0.74358	0.710413	**0.753554**
	Judd_CB	0.70648	0.692976	0.730955	**Judd_CB**	0.728782	0.694992	0.742763
	DVS_CB	0.71843	0.720774	0.730661	**DVS_CB**	0.744969	0.727287	0.739679
	Proto_CB	0.719769	0.631964	0.734962	**Proto_CB**	0.747049	0.628402	0.748433
	FES_CB	0.713014	**0.72253**	0.740432	**FES_CB**	0.663244	0.621972	0.643031
	CF_CB	0.71448	0.71882	0.723998	**CF_CB**	0.739819	0.721561	0.733012
	CF	0.687318	0.688245	0.67211	**CF**	0.74211	0.726052	0.732257
	Center Bias	0.692511	0.706171	0.706171	**Center Bias**	0.699876	0.691302	0.691302

Table 1. *Quantitative results of the best performing saliency models under the ROC(AUC) and the shuffled ROC(AUC) measures. The highest result expected (human performance) is provided by the top row as Inter-subject consistency. The worst performance, considered effectively chance, is the given by Center Bias on the bottom row. Any model measure results with a performance below Center Bias is colored Red. The best deep learning models within each column (image class) are highlighted in green, while the best shallow models are highlighted in yellow. The models that had the highest performance overall (for a given measure and image class) are in bold.*

The strongest pattern evident in the results shown in **Table 1** and **Table 2** is that the deep learning models, SAM_ResNet and SalGAN, are consistently the best predictors of attention for tone-mapped images across all of the different measures. However, the predictive accuracy of saliency models when applied to the raw HDR images is less consistent across the different measures; though, out of the four measures, deep learning models have the highest accuracy, as SAM_ResNet, SalGAN with center bias (SalGAN_CB), and eDN have the highest performance for the shuffled ROC-AUC, NSS, and shuffled NSS measures, respectively. FES with center bias (FES_CB) is the only shallow model that has the highest performance for any of the four measures. However, while FES_CB has the highest overall performance when using the ROC evaluation measure, under the other measures its performance is so low, that it is actually lower than Center Bias prediction, effectively chance, for all of the other measures.

	NSS			Shuffled-NSS		
Best DNN	**TMO**	**HDR**	**log(HDR)**	**TMO**	**HDR**	**log(HDR)**
Inter-subject	1.275972	1.281077	1.281077	1.13023	1.235765	1.235765
SalGAN_CB	0.91316	**0.813933**	0.862809	0.907831	0.575235	0.607325
iSeel	0.870075	0.753285	0.767523	0.906424	0.536665	0.589243
eDN	0.702011	0.676554	0.6907	0.82119	**0.624251**	0.655426
SAM_ResNet	1.053474	0.770949	0.478143	**1.042661**	0.575693	0.315434
SalGAN	**1.054525**	0.754321	**0.886348**	0.884539	0.344578	0.349703
Best Shallow						
GBVS_CB	0.810917	0.570283	0.803664	0.8936	0.434265	0.669686
Judd_CB	0.771501	0.641535	0.791446	0.844576	0.512668	**0.676466**
DVS_CB	0.724439	0.675156	0.792179	0.753481	0.504304	0.581065
Proto_CB	0.766777	0.505446	0.751758	0.794151	0.348691	0.551376
FES_CB	0.583729	0.504395	0.596993	0.480229	0.261768	0.336569
CF_CB	0.696183	0.74153	0.721572	0.74107	0.538229	0.556101
CF	0.640689	0.617829	0.539414	0.800302	0.565947	0.581035
Center Bias	0.646955	0.672464	0.672464	0.686752	0.500476	0.500476

Table 2. *Quantitative results of the best performing saliency models using the NSS and the shuffled NSS measures. The highest result expected (human performance) is provided by the top row as Inter-subject. The worst performance, considered effectively chance is the given by Center Bias on the bottom row. Any model that had a performance below Center Bias is colored Red. The best deep learning models within each column (image class) are highlighted in green, while the best shallow models are highlighted in yellow. The models that had the highest performance overall (for a given measure and image class) are in bold.*

A common method, and probably the simplest method to map an HDR image into a lower dynamic range, is to take the logarithm of the image, which is shown in the third column of each measure in **Table 1** and **Table 2**. For most of the models, across most of the measures, the predictive performance increases when the models are applied to the log version of the HDR image, though often the models still have a lower predictive performance when compared to the tone-mapped images. The only models that don't always benefit from the log-encoded version of the HDR image are the contrast feature (CF) model, using the ROC(AUC) measure, and the SAM_ResNet model, for all measures; in fact, the decrease in the predictive performance of the SAM_ResNet model on the log(HDR) images is so bad, it has a predictive performance below that of the center bias model across all measures.

5. Discussion

Across all of the visual saliency models that we have used in our evaluation, there remains a consistent difference between the predictive performance of all those models, across all measures, and human performance (predicting the eye gaze of one person using the eye gaze of other people). Deep learning models (for SDR/tone-mapped imagery) typically are the closest to matching human performance with some models even exceeding average human performance (under certain measures); however, that difference between model prediction and human performance only increases when those models are applied to HDR imagery, occasionally becoming worse than center bias-based predictions. Simple methods of adaption, like encoding the HDR image using the log values of the HDR image, can help to better adapt the visual salience model to the HDR images with higher bit depths so that it appears more like an SDR image and allowing the assumptions built into those models to not be violated as much. However, the log encoding of an HDR image doesn't work for all models equally as well, and in the cases of the CF and SAM_ResNet models, applying the models to the log-encoded HDR image can actually degrade the predictive performance. This is especially problematic for the SAM_ResNet model, as it is one of the best models for predicting eye gaze in tone-mapped images for most of the saliency measures we used. Thus, while there is a push in the research of visual saliency models to include higher-level processing, to better predict where a person will look in an SDR image, it is clear from the results in this paper that improved lower-level processing is also needed to allow saliency models to better predict eye gaze in HDR environments [54]. If the latest saliency models can perfectly predict how a tone mapping method or some other image manipulation approach will affect attention, but are unable to equally predict where a person's attention would be drawn to in the original HDR image, then the quality of the system to influence attention will likely suffer. A likely way forward would be to utilize approaches from tone mapping. Log compression can be considered one of the simplest tone mapping approaches, to enable saliency models to adapt to the dynamic range of the input image without throwing away potentially useful information about the HDR image.

6. Conclusion

In this paper, we evaluated a set of visual saliency models to determine how well they could predict how different tone mapping methods influence attention as a proxy for other image manipulation methods. This was aimed at demonstrating the potential for visual saliency methods to serve as a computational feedback method for image manipulation algorithms in order to influence the attentional priority given to different locations within an image. Thus, in a complex command and control environment, visual information can be presented to commanders and analysts in a way to draw their attention to relevant information, potentially without distracting them or adding to the clutter of a potentially already-cluttered image. In our analysis, we showed that deep learning models are consistently able to predict eye gaze in SDR tone-mapped images across multiple measures. However, to properly complete this feedback loop, the visual saliency models need to also predict attention in the original source image. Techniques used within the tone mapping literature may serve as a way forward to adapt existing or to develop new models of visual saliency that can predict eye gaze in images with any dynamic range.

REFERENCES

[1] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 267–276, Jul. 2002.

[2] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec, *High dynamic range imaging*. Morgan Kaufman, 2006.

[3] X. Gao, S. Brooks, and D. V. Arnold, "Saliency-based parameter tuning for tone mapping," in *Proceedings of the 11th European Conference on Visual Media Production - CVMP '14*, 2014, pp. 1–10.

[4] W.-C. Lin and Z.-C. Yan, "Attention-based high dynamic range imaging," *Vis. Comput.*, vol. 27, no. 6–8, pp. 717–727, Jun. 2011.

[5] Z. Li and J. Zheng, "Visual-salience-based tone mapping for high dynamic range images," *IEEE Trans. Ind. Electron.*, vol. 61, no. 12, pp. 7076–7082, 2014.

[6] A. Borji, H. R. Tavakoli, and Z. Bylinskii, “Bottom-Up Attention, Models of,” in *Encyclopedia of Computational Neuroscience*, New York, NY: Springer New York, 2019, pp. 1–19.
[7] L. Itti and A. Borji, “Computational models: Bottom-up and top-down aspects,” pp. 1–30, 2015.
[8] L. Itti and A. Borji, “Computational models of attention,” *Cogn. Neurosci. Biol. Mind (Fifth Ed.*, pp. 1–10, 2015.
[9] A. Borji, “Saliency Prediction in the Deep Learning Era: Successes, Limitations, and Future Challenges,” *arXiv Prepr. arXiv1810.03716v3 [cs.CV]*, pp. 1–43, 2018.
[10] M. Narwaria, M. Perreira Da Silva, P. Le Callet, and R. Pepion, “Effect of tone mapping operators on visual attention deployment,” *Appl. Digit. Image Process. XXXV*, vol. 8499, no. July 2014, p. 84990G, 2012.
[11] M. Narwaria, M. Perreira Da Silva, P. Le Callet, and R. Pepion, “Tone mapping based HDR compression: Does it affect visual experience?,” *Signal Process. Image Commun.*, vol. 29, no. 2, pp. 257–273, 2014.
[12] A. M. Treisman and G. Gelade, “A feature-integration theory of attention,” *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.
[13] C. Koch and S. Ullman, “Shifts in selective visual attention: towards the underlying neural circuitry.,” *Hum. Neurobiol.*, vol. 4, no. 4, pp. 219–27, Jan. 1985.
[14] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.
[15] A. Borji and L. Itti, “State-of-the-Art in Visual Attention Modeling,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.
[16] A. F. Russell, S. Mihala??, R. von der Heydt, E. Niebur, and R. Etienne-Cummings, “A model of proto-object based saliency,” *Vision Res.*, vol. 94, pp. 1–15, 2014.
[17] Tie Liu *et al.*, “Learning to Detect a Salient Object,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.
[18] A. Harrison and R. Etienne-Cummings, “An entropy based ideal observer model for visual saliency,” in *2012 46th Annual Conference on Information Sciences and Systems (CISS)*, 2012, pp. 1–6.
[19] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell, “SUN: Top-down saliency using natural statistics,” *Vis. cogn.*, vol. 17, no. 6–7, pp. 979–1003, 2009.
[20] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in *Advances in neural information processing systems*, 2007, pp. 545–552.
[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Neural Information Processing Systems*, 2012, pp. 1–9.
[22] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in *ICLR 2015*, 2014, pp. 1–14.
[23] C. Szegedy *et al.*, “Going Deeper with Convolutions,” in *CVPR 2015*, 2015, pp. 1–9.
[24] P. E. Debevec and J. Malik, “Recovering high dynamic range radiance maps from photographs,” in *ACM SIGGRAPH 2008 classes on - SIGGRAPH ’08*, 2008, no. 650, p. 1.
[25] A. Harrison, L. L. Mullins, A. Raglin, and R. Etienne-Cummings, “Enhanced visual perception through tone mapping,” in *Next-Generation Analyst IV*, 2016, vol. 9851, p. 98510U.
[26] W. C. Lin and Z. C. Yan, “Attention-based High Dynamic Range imaging,” *Vis. Comput.*, vol. 27, no. 6–8, pp. 717–727, 2011.
[27] S. Afreen, A. Tirmizi, and M. Sarwar, “Pseudo-Multiple-Exposure-based Tone Fusion and Visual-Salience-based Tone Mapping for High Dynamic Range Images: A Review,” *Int. J. Comput. Appl.*, vol. 127, no. 8, pp. 41–46, 2015.
[28] A. Rana, G. Valenzise, and F. Dufaux, “An evaluation of HDR image matching under extreme illumination changes,” *VCIP 2016 - 30th Anniv. Vis. Commun. Image Process.*, 2017.
[29] L. Chermak and N. Aouf, “Enhanced feature detection and matching under extreme illumination conditions with a HDR imaging sensor,” in *2012 IEEE 11th International Conference on Cybernetic Intelligent Systems (CIS)*, 2012, pp. 64–69.
[30] A. Rana, G. Valenzise, and F. Dufaux, “Evaluation of Feature Detection in HDR Based Imaging under Changes in Illumination Conditions,” *Proc. - 2015 IEEE Int. Symp. Multimedia, ISM 2015*, pp. 289–294, 2016.
[31] Y. Dong, M. T. Pourazad, and P. Nasiopoulos, “Human Visual System-Based Saliency Detection for High Dynamic Range Content,” *IEEE Trans. Multimed.*, vol. 18, no. 4, pp. 549–562, 2016.
[32] R. Bremond, J. Petit, and J.-P. Tarel, “Saliency Maps of High Dynamic Range Images,” in *European Conf. on*

Comp. Vision, 2012, pp. 118–130.
[33] M. Ashikhmin and J. Goyal, “A Reality Check for Tone-Mapping Operators,” *ACM Trans. Appl. Percept.*, vol. 3, no. 4, pp. 399–411, 2006.
[34] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, “Adaptive Logarithmic Mapping For Displaying High Contrast Scenes,” *Comput. Graph. Forum*, vol. 22, no. 3, pp. 419–426, Sep. 2003.
[35] J. Kuang, G. M. Johnson, and M. D. Fairchild, “iCAM06: A refined image appearance model for HDR image rendering,” *J. Vis. Commun. Image Represent.*, vol. 18, no. 5, pp. 406–414, Oct. 2007.
[36] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, “Photographic tone reproduction for digital images,” *ACM Trans. Graph.*, vol. 21, no. 3, pp. 267–276, 2002.
[37] K. Chiu, M. Herf, P. Shirley, S. Swamy, C. Wang, and K. Zimmerman, “Spatially Nonuniform Scaling Functions for High Contrast Images,” 1993.
[38] F. Durand and J. Dorsey, “Fast bilateral filtering for the display of high-dynamic-range images,” in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques - SIGGRAPH ’02*, 2002, p. 257.
[39] J. Tumblin, J. K. Hodgins, and B. K. Guenter, “Two methods for display of high contrast images,” *ACM Trans. Graph.*, vol. 18, no. 1, pp. 56–94, Jan. 1999.
[40] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, “What Do Different Evaluation Metrics Tell Us About Saliency Models?,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, Mar. 2019.
[41] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, “SUN: A Bayesian framework for saliency using natural statistics.,” *J. Vis.*, vol. 8, no. 7, pp. 32.1–20, Jan. 2008.
[42] 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.
[43] E. Erdem and A. Erdem, “Visual saliency estimation by nonlinearly integrating features using region covariances.,” *J. Vis.*, vol. 13, no. 4, pp. 11, 1–20, 2013.
[44] H. Rezazadegan Tavakoli, E. Rahtu, and J. Heikkilä, “Fast and Efficient Saliency Detection Using Sparse Sampling and Kernel Density Estimation,” Springer, Berlin, Heidelberg, 2011, pp. 666–675.
[45] L. Matzen, A. Wilson, K. Divis, M. Armenta, and L. Mcnamara, “Modeling human Comprehension of Data Visualizations ■ Cognition at Sandia,” *Virtual, Augment. Mix. Real.*, no. September, pp. 125–134, 2011.
[46] Z. Bylinskii *et al.*, “Learning Visual Importance for Graphic Designs and Data Visualizations,” in *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology - UIST ’17*, 2017, pp. 57–69.
[47] A. Banitalebi-Dehkordi, Y. Dong, M. T. Pourazad, and P. Nasiopoulos, “A learning-based visual saliency fusion model for High Dynamic Range video (LBVS-HDR),” in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1541–1545.
[48] E. Vig, M. Dorr, and D. Cox, “Large-scale optimization of hierarchical features for saliency prediction in natural images,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 2798–2805, 2014.
[49] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, “A deep multi-level network for saliency prediction,” in *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 3488–3493.
[50] H. R. Tavakoli, A. Borji, J. Laaksonen, and E. Rahtu, “Exploiting inter-image similarity and ensemble of extreme learners for fixation prediction using deep features,” *Neurocomputing*, vol. 244, pp. 10–18, Jun. 2017.
[51] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, “Predicting Human Eye Fixations via an LSTM-Based Saliency Attentive Model,” *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, Oct. 2018.
[52] J. Pan *et al.*, “SalGAN: Visual Saliency Prediction with Generative Adversarial Networks,” *arXiv Prepr. arXiv1701.01081v3 [cs.CV]*, Jan. 2017.
[53] A. D. F. Clarke and B. W. Tatler, “Deriving an appropriate baseline for describing fixation behaviour,” *Vision Res.*, vol. 102, pp. 41–51, 2014.
[54] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand, “Where Should Saliency Models Look Next?,” in *European Conf. on Comp. Vision*, vol. 9909, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Springer, Cham, 2016, pp. 809–824.