VISUAL SALIENCY BASED IMAGE QUALITY ASSESSMENT

Gaurav Verma

Department of Electrical Engineering
Indian Institute of Technology Kanpur
gverma@iitk.ac.in

Paridhi Maheshwari

Department of Electrical Engineering
Indian Institute of Technology Kanpur
paridhi@iitk.ac.in

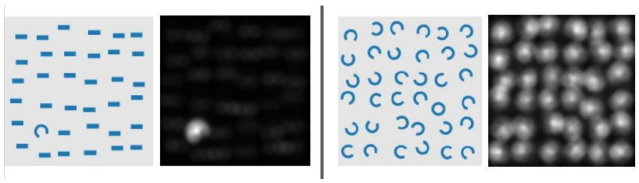
ABSTRACT

We study the use of visual saliency (VS) maps for image quality assessment (IQA). We propose a new Fourier-based metric for comparing saliency maps that quantifies the dissimilarity between them in a more informative manner, compared to conventional metrics like MSE, KL-divergence and NSS. Owing to the ineffectiveness of existing metrics in quantifying distortion types like contrast change and change in color saturation, as reflected in the VS maps, we propose new saliency model that is sensitive to these distortion types, and generates saliency maps that can be used for IQA research as well as for general purposes. [1]

***Index Terms*—** visual saliency, image quality assessment, metrics

1. INTRODUCTION

Visual Saliency (VS) is the distinct subjective perceptual quality which makes parts of an image stand out from their surroundings and immediately grab our attention. This can be best illustrated using the following example. The brighter regions in the map correspond to the visually salient region in the image while the darker regions correspond to the regions that are not of comparable interest. In the left figure, the symbol 'c' stands out, due to its structural difference from the '-' symbol, which is in abundance in the image. However, in the right image, the 'circle' muddles with the 'c' symbols and doesn't stand out, which is evident in the corresponding saliency map.



[1]This work was done as a course project for the course EE698K (Winter Semester'18), under the guidance of Prof. Tanaya Guha, who is with the Department of Electrical Engineering at IIT Kanpur. Email: tanaya@iitk.ac.in

Image Quality Assessment (IQA) is one of the most fundamental and yet challenging problems in image processing and vision research. IQA algorithms are designed to mimic the subjective judgments of humans, which is the ultimate criterion for image quality. Conventional fidelity metrics like peak signal-to-noise ratio (PSNR) or the mean squared error (MSE) do not correlate well with human's subjective fidelity ratings when multiple distortion types are involved [1].Therefore, in the past decade, several sophisticated IQA models, like noise quality measure index (NQM) [2], structural similarity index (SSIM) [3], and visual information fidelity index (VIF) [4], were proposed.

Recent approaches have been trying to incorporate visual saliency information to IQA models to improve their performance. This entire idea is based on the premise that a distortion occurring in a more salient area is more annoying than the one occurring in a less salient area and hence, should we weighted accordingly while pooling the quantified local distortions. One such work by Zhang et al. [5] proposes an IQA metric called visual saliency-based index (VSI). It is motivated by their following observation – with distortion in images, their saliency maps vary significantly.

2. RELATED WORK

2.1. Generation of Saliency Maps

Several methods have been proposed to generate saliency maps from input images. In 2007, Hou et al. [6] proposed a *spectral-residual* approach in which they analyzed the log-spectrum of an input image to extract the spectral-residual of the image in spectral domain. Achanta et al. [7] proposed a method that exploits features of color and luminance to generate saliency maps with well-defined boundaries of salient objects in the image. These boundaries were preserved by retaining substantially more frequency content than erstwhile techniques. Recently, in 2015, Cheng et al. [8] proposed a global contrast based saliency detection method. More recent trends include the use of deep neural networks to learn these saliency maps [9].

2.2. Comparison Metrics for Saliency Maps

Mean Square Error (MSE) for comparing saliency maps and in geneal, images, holds well as a conventional fidelity measure but fails to capture the finer aspects involved in comparison. Bylinskii et al. [10] carried out a thorough investigation of several metrics that have been frequently used to compare saliency maps. They provided an analysis of metrics like area under ROC curve (AUC), normalized scanpath saliency (NSS), earth mover's distance (EMD), similarity or histogram intersection (SIM), Kullback-Leibler divergence (KL), information gain (IG) etc., and listed down the properties of the inputs and how they affect these metrics. Some of the metrics have been found to be invariant to linear transformations (contrast) – like AUC, NSS and CC – while some of them have been found to scale with spatial distance – like EMD.

2.3. Saliency Maps for Image Quality Assessment

The close relationship between visual saliency (VS) maps and quality perception has led to a number of approaches that try to integrate VS into IQA metrics. In [11], Zhang et al. have shown that VS values actually vary with the change of image quality. They propose an index that uses VS as a feature to compute the local similarity between the reference image and its distorted version. However, VS maps do not show any significant difference (in terms of mean square error) for the distortion types of contrast reduction (CR) and change of color saturation (CCS). This is evident from Figure 1. Its primary reason is the normalization operation involved in VS computational models. To overcome this shortcoming, they use gradient maps and chrominance features along with the original features and propose a new similarity metric, that accounts for these two distortions.

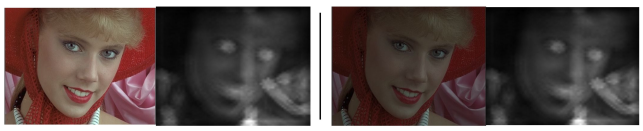


Fig. 1: Example inspired from [5] to illustrate how the current visual saliency computational models do not capture contrast reduction.

3. CONTRIBUTIONS

3.1. Fourier-based Distance Metric

We propose a new Fourier-based metric for comparing saliency maps (which can also be used for comparing heat maps). Given two saliency maps S_1 and S_2 corresponding to the two images I_1 and I_2 (of dimension $N \times M$ each), the distance between these two saliency maps can be evaluated using the following equation:

$$d(S_1, S_2) = \sqrt{\sum_k \frac{(\mathfrak{F}(S_1 - S_2)_k)^2}{1 + (2\pi \mid k \mid)^2}}$$

where, $k = (k_x, k_y) \, \forall \, k_x \in \{0, \cdots, N-1\}$ and $k_y \in \{0, \cdots, M-1\}$. Intuitively, the numerator simply encapsulates the difference in corresponding Fourier coefficients of the two saliency maps. The denominator scales down the contribution of higher frequency coefficients as $\mid k \mid = \sqrt{k_x^2 + k_y^2}$ increases as $k = (k_x, k_y)$ increases. Scaling down of the higher frequency components is desirable as higher frequencies largely correspond to noise while the lower frequencies correspond to more subtle changes.

The following transformation

$$\hat{d}(S_1, S_2) = \frac{d(S_1, S_2)}{1 + d(S_1, S_2)}$$

limits the distance between 0 (exactly similar) and 1 (completely different). The proposed distance metric also follows the essentials of being one – it is (i) symmetric, (ii) non-negative and (iii) follows the triangle inequality.

To assess the quality of the proposed metric, we systematically vary parameters of a saliency map and report effects on the distance scores. Each row in Figure 2 corresponds to varying a single parameter of the prediction: (a) variance, (b,c) location, and (d) relative weight. The variation of the distance score in response to these methodical changes suggests that the metric is able to capture the subtle differences in saliency maps and also align well with human observations.

An interesting observation in this regard, is the last row corresponding to variation in relative weight. It can be noted that as the weights of the two spots are varied in predictions (Row 4; Column 2), the metric gives a lower score to instances when the weights of the two spots are almost similar, than to the instances when the weights are significantly different. A casual experiment carried with the help of ten of our friends, indicates that the third prediction in Row 4 does seem more similar to the ground truth than the other predictions in the same row. Some of the conventional metrics that are used to compare saliency maps (like Information Gain, KL divergence and Area Under the Curve) fail to capture such changes [10].

3.2. Comparison of Saliency Maps Metrics

In order to compare the performance of the proposed Fourier-based metric to other metrics and also, quantify different metrics in their ability to perceive quality difference only using saliency maps, we conducted a statistical analysis on the VS maps of images in TID2013 [12]. TID2013 is the most comprehensive dataset available for IQA research, and contains 25 reference images and 24 distortion types. The distortion types cover a wide range of noises such as additive gaussian noise, contrast reduction, gaussian blur and chromatic aberrations. For each image and distortion, there are

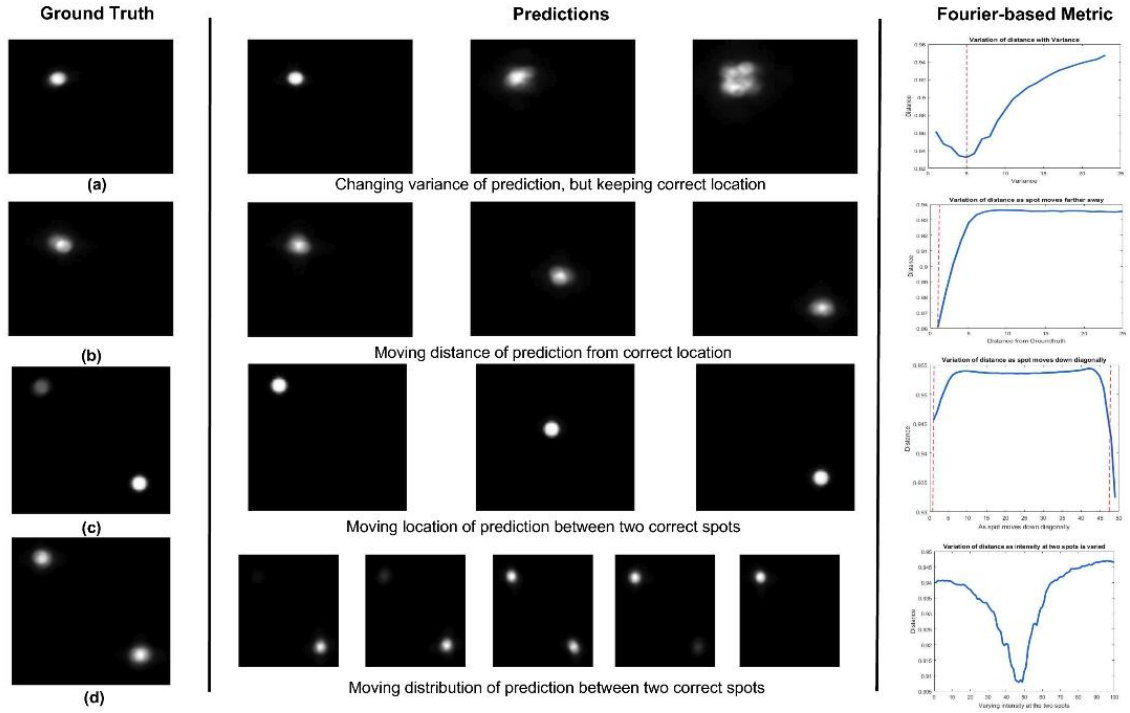


Fig. 2: The x-axis of each graph spans the parameter range, with the dotted red line corresponding to the ground truth parameter setting. A higher y-value indicates a greater dissimilarity between the prediction and the ground truth. The experiment is inspired from Bylinskii et al. [10]. The figure is best viewed on screen with zoom.

5 levels, with an increased level indicating higher distortion. Human subjects were asked to rate the image quality on a scale of 0-10. For all images of the dataset, saliency maps were generated using two techniques: (a) Spectral residual (SR) approach proposed by Hou et al. [6], and (b) LSTM-based saliency attentive model (SAM-VGG) proposed by Cornia et al. [13]. The saliency maps generated using these two techniques have been shown for one sample image in Figure 3.

The metrics that we investigated were: Area under the ROC Curve (AUC), Pearson's Correlation Coefficient (CC), Fourier-based Metric (Fourier), Kullback-Leibler divergence (KL), Mean Square Error (MSE), Normalized Scanpath Saliency (NSS), and Similarity of histogram intersection (SIM). KL divergence models the saliency maps as probability distributions and Pearson's correlation is a standard measure in statistics. AUC and NSS have been designed specially for saliency maps whereas, the histogram intersection function operates on image histograms. Similarity/dissimilarity scores between the VS maps of a reference image and its distorted versions were computed for all these metrics. To find the metric that best quantifies these distortions, we computed Pearson's correlation coefficient of the respective metric scores with the subjective scores.

The values of average correlation over all 25 reference images for the two sets of generated saliency maps have been tabulated in Table 1. An important factor is that we desire a high *absolute* value of the average correlation – because some of the metrics quantify dissimilarity (like Fourier, MSE and KL), while other metrics quantify similarity (like SIM and NSS). It is observed that SIM shows a better or comparable correlation with subjective scores than MSE, except when the distortion type is CTC (Contrast change). However, for a metric to be an ideal proxy for a visual image quality, the sign of the correlation values should be consistent throughout all the distortions. None of the metrics that we consider are consistent in this aspect. Although SIM stands best to quantify for most of the distortion types, it can also be observed that the Fourier-based metric outperforms several other metrics by showing a higher correlation with the subjective scores. These experiments suggest that the proposed metric matches the behavior of conventionally used metrics to systemic variation in input images, while providing some additional information.

3.3. Generating Saliency Maps for IQA

We start this section by summarizing the observations made from the experiments conducted as a part of the previous sec-

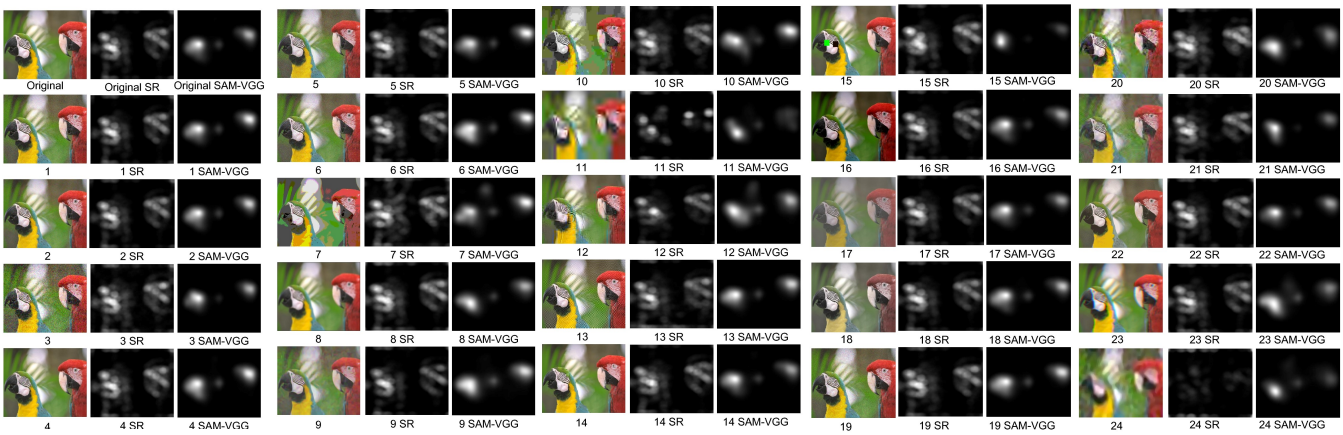


Fig. 3: The 24 distortion types for one sample image from TID2013 along with the saliency maps generated using SR [6] and SAM-VGG [13]. The index indicates the distortion type as per TID2013. The Figure is best viewed on screen with zoom.

Dist. Type	Spectral Residual							SAM-VGG						
	AUC	CC	Fourier	KL	MSE	NSS	SIM	AUC	CC	Fourier	KL	MSE	NSS	SIM
AGN	-0.06	0.93	-0.73	-0.79	-0.69	0.12	**0.96**	-0.04	0.85	-0.53	-0.87	-0.56	0.58	**0.91**
ANC	-0.21	0.95	-0.74	-0.74	-0.63	0.02	**0.95**	0.00	0.91	-0.53	-0.93	-0.68	0.49	**0.94**
SCN	-0.10	0.92	-0.87	-0.89	-0.60	0.43	**0.95**	0.24	0.88	-0.38	-0.86	-0.73	0.77	**0.94**
MN	-0.38	0.94	-0.85	-0.93	-0.79	0.11	**0.97**	-0.30	0.93	-0.67	-0.88	-0.75	0.35	**0.95**
HFN	-0.24	0.92	-0.94	-0.92	-0.94	0.15	**0.96**	0.10	0.92	-0.64	-0.94	-0.89	0.84	**0.97**
IN	-0.02	0.92	-0.78	-0.82	-0.73	0.30	**0.96**	-0.17	0.84	-0.69	-0.85	-0.70	0.61	**0.91**
QN	-0.15	0.83	-0.85	-0.85	-0.78	0.80	**0.92**	0.14	0.85	-0.59	-0.83	-0.57	0.73	**0.91**
GB	0.15	0.77	**-0.88**	-0.82	-0.77	0.65	0.87	-0.03	0.89	-0.75	-0.91	-0.85	0.72	**0.96**
DEN	0.57	0.89	-0.51	-0.91	-0.92	0.68	**0.96**	-0.23	0.89	-0.77	-0.90	-0.67	0.71	**0.94**
JPEG	-0.51	0.91	-0.91	-0.89	-0.88	0.79	**0.97**	-0.15	0.91	-0.77	-0.90	-0.75	0.82	**0.96**
JP2K	0.47	0.84	-0.81	-0.84	-0.91	0.84	**0.91**	0.09	0.89	-0.81	-0.89	-0.76	0.80	**0.96**
JGTE	0.11	0.76	-0.69	-0.77	-0.75	0.70	**0.82**	-0.03	0.82	-0.64	-0.78	-0.60	0.65	**0.86**
J2TE	-0.12	0.81	-0.82	-0.81	-0.78	0.67	**0.87**	0.10	0.78	-0.56	-0.80	-0.59	0.61	**0.86**
NEPN	-0.00	0.89	-0.67	**-0.91**	-0.65	0.73	0.91	0.08	0.80	-0.53	-0.77	-0.53	0.42	**0.85**
Block	0.23	0.33	-0.09	-0.41	-0.14	0.18	**0.55**	-0.00	**-0.29**	0.11	0.20	-0.10	-0.27	-0.20
MS	-0.03	0.64	-0.45	-0.69	**-0.70**	0.26	0.69	-0.05	0.87	-0.56	-0.74	-0.71	0.28	**0.88**
CTC	0.15	-0.53	0.50	0.47	0.32	0.00	**-0.57**	-0.12	0.48	-0.24	**-0.55**	-0.38	0.04	0.54
CCS	0.06	**0.43**	0.03	-0.04	-0.36	-0.01	0.43	-0.15	0.85	-0.83	-0.86	-0.76	0.65	**0.91**
MGN	-0.34	0.92	-0.80	-0.88	-0.62	0.28	**0.95**	-0.24	0.83	-0.63	-0.90	-0.67	0.60	**0.91**
CN	0.13	0.87	-0.79	-0.89	-0.85	0.71	**0.93**	-0.13	0.88	-0.70	-0.88	-0.73	0.52	**0.93**
LCNI	-0.03	0.91	-0.82	-0.92	-0.86	0.67	**0.96**	-0.02	0.90	-0.63	-0.90	-0.75	0.72	**0.95**
ICQD	-0.18	0.88	-0.78	-0.88	-0.90	0.50	**0.95**	0.17	0.89	-0.67	-0.89	-0.81	0.80	**0.95**
CHA	0.37	0.87	-0.79	-0.86	-0.90	0.78	**0.91**	0.17	0.90	-0.73	-0.90	-0.81	0.86	**0.92**
SSR	0.56	0.86	-0.68	-0.88	-0.92	0.86	**0.94**	-0.00	0.92	-0.74	-0.91	-0.79	0.83	**0.96**

Table 1: Correlation of metric scores with the subjective scores, for saliency maps generated using SR [] and SAM-VGG [].

tion:

- The proposed Fourier-based metric provides a better quantification of distance between saliency maps than many of the conventionally used metrics.
- SIM (Similarity of histogram intersection) gives the best measure across all metrics that were considered here.
- No metric meets the desired expectation of being consistent throughout all the 24 distortion types. The problem is prominent for distortion types CTC and CCS.

To this end, we analyze the reasons for the observed anomalies that occur while quantifying the (dis)similarity of saliency maps, when the distortion type is contrast change (CTC) or change in color saturation (CCS). Following the analyses, we propose a variant of pix2pix [14] to generate saliency maps that are good for image quality analysis, as well as for other general purposes.

Einhuser et al. [15] found that when the contrast is varied in a given image, the perceived saliency patterns vary only for the first four seconds (after presenting the stimulus), and become same as that for the original image thereafter. The TID2013 [12] dataset was not collected under such a setting.

Moreover, the trends of subjective scores for the distortion type CTC and CCS are very unpredictable (when compared to those for the other distortion types).

Isola et al. [14] came up with an image-to-image translation model that uses conditional generative adversarial networks (cGANs) to learn pixel-wise mappings to transform images. To generate saliency maps that can capture contrast change, and hence can be used for IQA research, we based our work on this model, more commonly known as the pix2pix model. The network, as proposed by Isola et al. consists of a generator and a discriminator. The job of the generator is to produce outputs that cannot be distinguished by a discriminator that is adversarially trained to detect generator's fakes. The objective that the model tries to minimize is given by:

$$\mathfrak{L}_{cGAN} = \mathbb{E}_{x,y}[\log D(x,y) + \mathbb{E}_{x,z}[\log(1 - D(x, G(x,z)))]$$

G, the generator, tries to minimize this objective against an adversarial D, the discriminator, that tries to maximize it. It has been found that mixing the GAN objective with more tradition L1 or L2 loss, gives an additional responsibility to the generator to generate images that are close the ground truths. Isola et al. use the L1 loss instead of L2, as it encourages less blurring.

$$G^* = \min_G \operatorname{argmax}_D \mathfrak{L}_{cGAN}(G,D) + \lambda \mathfrak{L}_{L1}(G) \text{ , where}$$

$$\mathfrak{L}_{L1}(G) = \mathbb{E}[||y - G(x,z)||_1]$$

We wanted to ensure that we didn't face the problem that Zhang et al. [5] faced, i.e., we wanted our saliency maps to register the change in contrast of original images. Therefore, we decided to add an edge preserving loss to the optimization function. After re-weighting the additional term with an hyper-parameter λ_2, we added it to the original loss function to get the final optimization function as:

$$G^* = \min_G \operatorname{argmax}_D \mathfrak{L}_{cGAN}(G,D) + \lambda_1 \mathfrak{L}_{L1}(G) + \lambda_2 \mathfrak{L}_{EPL}(G)$$

Here, $\lambda \mathfrak{L}_{EPL}(G)$ denotes the binary cross-entropy loss between gradient maps of the original image and its saliency map. The gradient-pixels were treated as 1 or 0, based on a threshold. For our experiments, we had the value of the threshold to be set at 0.5. The value of λ_1 was chosen to be 0.8, and the value of λ_2 was chosen to be 0.2. This optimization with an additional edge preserving loss encourages the model to change the visual saliency maps as edges in the original image change. Following this, we carried out certain statistical experiments to ensure that the generated saliency maps are good for IQA. Table 2, shown in the next page, lists the average correlation of the metric scores with the subjective scores, for the distortion typed CTC and CCS. As it can observed, the sign consistency of the correlation has improved. The values in blue denote that the saliency maps generated using our proposed model have done better than those generated using SAM-VGG or SR, in terms of sign consistency. The values in red denote that the problem still persists, and the values in black denote that there was no problem to begin with.

Having established that our saliency maps are good for IQA tasks (even for distortions type that are conventionally difficult to deal with, like contrast change (CTC) and changee in color saturation (CCS)), now we try to answer an equally important question: Are the saliency maps that our proposed model generates are good enough for general purposes. To test this, we evaluate our model on CAT2000 Saliency Dataset [16]. As per their evaluation criteria on this dataset, a saliency model that can perfectly mimics the human saleincy model would achieve an NSS score of 3.29 and KL-divergence score of 0.0. For our proposed model, the value of NSS score was 2.31 and KL was 0.67. As it can be inferred from the Table 3 below, these scores attained by our model are very close to the existing scores of the top performing models. This reinforces our belief that our saliency maps are not only good enough for image quality assessment, but are also good for general purposes.

4. CONCLUSION

To summarize our work, we proposed a new Fourier-based metric that shows remarkable ability in quantifying the dissimilarity between saliency maps (over several conventionally used metrics like MSE, KL and NSS). We carried out extensive experiments to assess the effectiveness of several similarity/dissimilarity metrics in quantifying the difference between saliency maps for image quality assessment research. As a result of these experiments, we were able to show the ineffectiveness of existing metrics (as well as that of the proposed metric) in quantifying the dissimilarity when the distortion type is contrast change (CTC) and change in color saturation (CCS). To deal with this ineffectiveness, we proposed a new model for generating saliency maps that effectively register the change in contrast of the original images, and hence can be used for IQA research. This variant of pix2pix model, which was obtained my adding an edge-preserving loss to the optimization function, generates saliency maps that are also good for general purposes and attain NSS anc KL divergence scores that are close to the top performing models on the CAT2000 Saliency Dataset.

5. REFERENCES

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

[2] Niranjan Damera-Venkata, Thomas D Kite, Wilson S Geisler, Brian L Evans, and Alan C Bovik, "Image qual-

	AUC	CC	Fourier	KL	MSE	NSS	SIM
CTC	-0.4289	0.6827	-0.2191	-0.3642	-0.3768	0.4571	0.3061
CCS	-0.2293	0.5321	0.0468	-0.1279	-0.2987	0.2842	0.6821

Table 2: Correlation values with the subjective scores, for distortion type CTC anc CCS, using our proposed model

Model	**NSS**	**KL**
Humans	3.29	0
DPNSal	**2.41**	0.91
DSCLRCN [17]	2.35	0.95
SAM-ResNet [13]	2.34	1.27
pix2pix + EPL	2.31	**0.65**
SAM-VGG [13]	2.30	1.13

Table 3: Comparison of our proposed model (pix2pix + EPL) with top performing models on the CAT2000 Saliency Dataset [16]. The data for other models was taken from [here]

ity assessment based on a degradation model," *IEEE transactions on image processing*, vol. 9, no. 4, pp. 636–650, 2000.

[3] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[4] Hamid R Sheikh and Alan C Bovik, "Image information and visual quality," *IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006.

[5] Lin Zhang, Ying Shen, and Hongyu Li, "Vsi: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[6] Xiaodi Hou and Liqing Zhang, "Saliency detection: A spectral residual approach," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[7] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk, "Frequency-tuned salient region detection," in *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*. IEEE, 2009, pp. 1597–1604.

[8] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[9] Guanbin Li and Yizhou Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5455–5463.

[10] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand, "What do different evaluation metrics tell us about saliency models?," *arXiv preprint arXiv:1604.03605*, 2016.

[11] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[12] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti, "Tid2008-a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[13] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *arXiv preprint arXiv:1611.09571*, 2016.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[15] Wolfgang Einhäuser and Peter König, "Does luminance-contrast contribute to a saliency map for overt visual attention?," *European Journal of Neuroscience*, vol. 17, no. 5, pp. 1089–1097, 2003.

[16] Ali Borji and Laurent Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," *arXiv preprint arXiv:1505.03581*, 2015.

[17] Nian Liu and Junwei Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3264–3274, 2018.