TranSalNet+: Distortion-aware saliency prediction

Jianxun Lou[a,b], Xinbo Wu[b], Padraig Corcoran[b], Paul L. Rosin[b], Hantao Liu[b]

[a]School of Computer Science, Northeast Electric Power University, Jilin, China
[b]School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom

Abstract

Predicting the saliency of images affected by distortion is a challenging but emerging research problem. Given a distorted image, we wish to accurately predict saliency as perceived by humans. A recent distortion-aware saliency benchmark – the CUDAS database – reveals the inadequacy of existing saliency models in handling distorted images. In this paper, we devise a deep learning Distortion-Aware Saliency Module (DASM) that enables capturing saliency features related to image distortions, and integrates this module into a saliency prediction architecture. To achieve the high expressive capability of DASM using supervised learning, we create a dedicated dataset that draws upon a large-scale saliency dataset and machine-generated image quality assessments. Experimental results demonstrate the superior performance of the proposed model in predicting the saliency of distorted images.

Keywords: Saliency, eye-tracking, distortion, image quality, deep learning

1. Introduction

Visual attention – a crucial function of the human visual system (HVS) – refers to the ability to selectively focus on pertinent information within a visual field [1]. Specifically, foveal vision encompasses a small central portion of the visual field and provides the most detailed and informative visual signals within the HVS [2]. The visual attention mechanism directs foveal vision to prioritise visual stimuli, thereby reducing the cognitive load on the cerebral cortex by selecting the most relevant information within a visual field [1, 3, 4]. Visual saliency that reflects the degree of selective attention, and in turn the detection of the most relevant and meaningful parts of a visual scene, has become an essential component in various applications in multimedia and computer vision [5, 6, 7, 8], human-computer interaction [9], and medical imaging [10].

In machine perception, visual saliency is typically modelled using a saliency map (also known as a fixation density map), providing a quantitative representation of the distribution of visual attention. Significant research has been dedicated to comprehending gaze and simulating visual saliency in computational methods [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. However, these studies focus on pristine images without any distortion, which limits the applicability of the results in many real-world imaging scenarios.

In many cases, digital images are vulnerable to distortions at various stages of the visual processing pipeline including acquisition, processing, compression, storage, transmission, and reproduction [22]. These distortions can alter human attention, thereby changing the visual saliency of the scene [23]. Predicting visual saliency of distorted images has significant potential for various tasks related to human visual perception. For example, a saliency model tailored for distorted images can be used to optimise objective image quality assessment (IQA) metrics [24] and to refine image enhancement algorithms, enabling more nuanced adjustments that better align with human visual experiences [25]. In a recent study [26], a distortion-aware saliency benchmark entitled the CUDAS (Cardiff University Distortion-Aware Saliency) database was created. In this study, a fully controlled eye-tracking experiment was conducted to collect eye movements of 96 human subjects viewing 600 stimuli with differing forms of degradation and varying degrees of perceived quality. Based on an analysis of the behaviour of 20 state-of-the-art saliency models applied to the CUDAS database, the research reveals that these models often fall short in predicting the saliency of distorted images [26]. Therefore it is critical to develop new saliency prediction models that can effectively handle images affected by various types and levels of distortion.

To address this new challenge, this paper proposes a distortion-aware visual saliency prediction model. We first develop a deep learning Distortion-Aware Saliency Module (DASM) to produce a saliency feature representation relating to image distortions. To enable high expressive ability for the DASM during supervised learning, we create a dedicated dataset that draws upon an existing large-scale saliency dataset (i.e., SALICON [27]) and machine-generated image quality assessments. More specifically, we extend the images of the SALICON dataset by generating nine distorted variants for each image, encompassing three different types of distortion at three perceptually distinct levels of degradation, based on the recommendations of previous research [26]; and assigning an image quality judgement using state-of-the-art image quality assessment (IQA) algorithms namely Machine Mean Opinion Score (MMOS). The core innovation involves using MMOS to allow DASM to produce distortion-aware saliency features during learning. Then, we integrate the DASM into a perceptually relevant visual saliency prediction algorithm *TranSal-*

Net [21], resulting in a new distortion-aware saliency model named *TranSalNet+*. Experimental results demonstrate the added value of the proposed DASM in handling distorted images, and the proposed model *TranSalNet+* achieves state-of-the-art performance on the CUDAS distortion-aware saliency benchmark [26].

Our contributions in this paper can be summarised as follows:

- We develop a dedicated dataset, SALICON-MMOS, by leveraging an existing large-scale saliency dataset, namely SALICON, in conjunction with machine-generated image quality assessments. This is achieved through the incorporation of the Machine Mean Opinion Score (MMOS), which is derived from state-of-the-art objective IQA models.

- We introduce the Distortion-Aware Saliency Module (DASM), benefiting from the guidance of the Machine Mean Opinion Score (MMOS) to generate distortion-aware saliency features during the learning process. It allows the saliency prediction model to consider the impact of image quality on visual attention, an aspect often overlooked in conventional saliency models.

- We propose a distortion-aware visual saliency prediction model that achieves state-of-the-art performance on the CUDAS distortion-aware saliency benchmark. It shows the efficacy of incorporating image quality information into the saliency prediction of distorted scenes.

2. Related work

2.1. Saliency prediction models

There is substantial work that has been dedicated to developing computational models for the automatic prediction of visual saliency. Early foundational models, such as IttiKoch [11], GBVS [12], Torralba [28], QSS [29], and CovSal [30], have been instrumental in establishing the field, primarily by leveraging low-level visual features like colour, luminance, texture, and contrast. These features aim to mimic the HVS's innate ability to spotlight salient regions within a scene, offering initial insights into the basic constituents of visual attention.

Despite their contributions, these traditional models face limitations, particularly in harnessing higher-level features—such as semantic context and object recognition—that are critical for a comprehensive understanding of visual saliency [31]. Due to recent advancements in deep learning [32, 33, 34, 35, 36], saliency models have entered a new era [13, 14, 15, 16, 17, 18, 19, 20]. These models transcend the sole reliance on handcrafted features, learning intricate image representations directly from data, which has significantly propelled the field forward.

However, despite these advancements, a critical gap remains conspicuously unaddressed: the majority of existing visual saliency prediction research focuses on pristine, undistorted images, leaving the saliency in distorted scenarios largely unexplored. This oversight is non-trivial, as prior studies have convincingly shown that image quality degradation can influence human visual attention, altering saliency patterns in distorted images [24]. Besides, recent studies have shown that existing visual saliency prediction models often fall short in in accurately predicting saliency within distorted images [26]. This underscores the necessity of integrating an understanding of image quality perception with visual saliency prediction models for images suffering from various distortions.

2.2. Image quality assessment

Since humans are the ultimate receivers of most visual information, subjective evaluation, where participants rate the perceived quality of images in controlled environments, is considered to be the most reliable method for assessing image quality. Previous research has employed extensive subjective experiments on image quality perception to elucidate human perception and evaluation of image quality [37, 38, 39]. Despite the invaluable insights offered by subjective assessments, the inherent limitations of subjective assessments, notably their high cost, time consumption, and limited scalability, significantly curtail their utility in large-scale applications.

In recent years, objective IQA models, such as MANIQA [40], HyperIQA [41], and TReS [42], have been proposed and yielded remarkable progress, achieving results on IQA benchmarks that closely approximate the subjective evaluations of human observers. By virtue of their scalability and efficiency, these IQA models present themselves as feasible alternatives to human observers in IQA applications. Therefore, in order to facilitate the development of distortion-aware saliency prediction, this study employs these three state-of-the-art IQA models as observers to generate Machine Mean Opinion Score (MMOS) for the expanded SALICON dataset [27] with distorted variants.

2.3. Distortion-aware saliency benchmark - CUDAS

The cornerstone of distortion-aware saliency prediction lies in the availability of reliable distortion-aware saliency benchmarks. CUDAS [26] stands as a leading benchmark in this domain, established on the basis of a large-scale eye-tracking study employing rigorous experimental methodologies to investigate visual attention towards distorted images. Specifically, CUDAS comprises a collection of 60 high-quality and high-resolution (1920×1080 pixels) pristine images. Figure 1 shows the diverse stimuli scenes within the CUDAS dataset. Distorted variants of these images are generated by simulating three distinct distortion types: contrast change (CnC), JPEG compression (JPEG), and motion blur (MB). For each distortion type, three distinct levels of perceived image quality (namely, Q1, Q2, and Q3) are created by varying the distortion strength. Consequently, CUDAS yields a total of 600 stimuli. The eye-tracking experiments conducted for CUDAS adhered to the International Telecommunication Union (ITU) standards [43] and were carried out in a standardised office environment. To ensure the reliability of the experimental data, a between-subjects

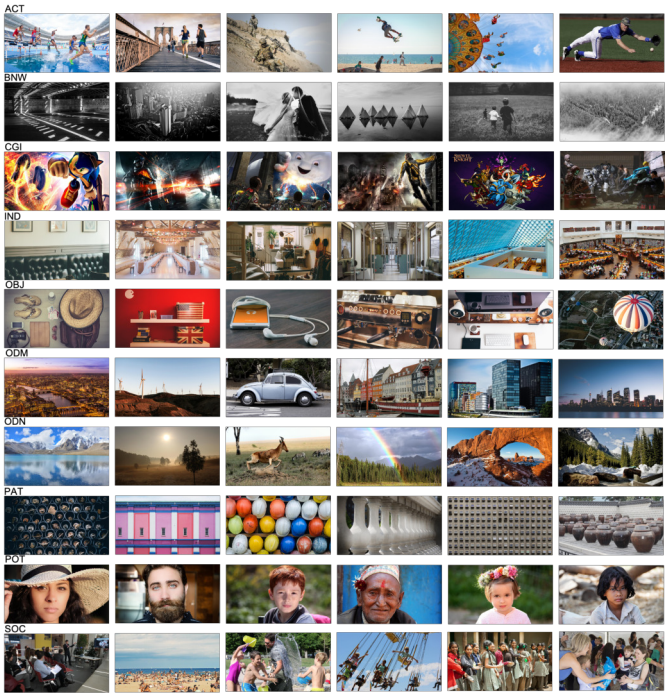

Figure 1: Illustration of the 60 source images (6 pristine images × 10 scene categories) contained in CUDAS [26]. From top row to bottom row, the categories are ACT (Action), BNW (Black and White), CGI (Computer-Generated Imagery), IND (Indoor), OBJ (Object), ODM (Outdoor Manmade), ODN (Outdoor Natural), PAT (Pattern), POT (Portrait), and SOC (Social).

approach [44] was employed during data acquisition, involving 96 participants to mitigate subject bias potentially arising from stimulus repetition [45]. Given that CUDAS provides the largest of its kind distortion-aware saliency dataset, this paper adopts it as a benchmark to explore distortion-aware saliency prediction.

3. Proposed method

3.1. Overall Architecture

The architecture of the proposed distortion-aware visual saliency prediction model, *TranSalNet+*, is illustrated in Figure 2.2. Building upon the previous research [21], which leveraged transformer encoders to capture long-range contextual dependencies for improved saliency prediction, *TranSalNet+* adopts the image encoder from the aforementioned work. This encoder utilises a ResNet-50 [33] backbone architecture, strategically integrating transformer encoders (T_1, T_2, and T_3) into the last three convolutional blocks in a progressive manner (deeper to shallower layers). These transformer modules effectively capture global dependencies within the feature maps, leading to a more comprehensive understanding of the image content. The encoded feature maps are subsequently processed by the Distortion-Aware Saliency Module (DASM), detailed in Section 3.2, to incorporate crucial image distortion information. Finally, the processed features are fed into the CNN decoder to generate the saliency map.

The CNN decoder, detailed in Figure 3, is tasked with efficiently decoding the latent features from the encoder and DASMs, restoring the spatial dimensions, and ultimately generating saliency maps. It achieves this by employing a sequential architecture consisting of five convolutional blocks (Conv_Block_1 to Conv_Block_5), followed by a dedicated Readout block. Each Conv_Block leverages a well-established pipeline of convolution, normalisation, ReLU activation, and upsampling to effectively learn upsampled feature representations. The final Readout block replaces the upsampling operation with a convolutional layer followed by a Sigmoid activation layer to generate the final saliency map. This design choice ensures the output saliency map lies within the range of [0, 1], signifying the relative importance of each pixel within the image.

3.2. Distortion-Aware Saliency Module

Previous studies have indicated a connection between human visual attention distribution and image distortion [24], suggesting that incorporating distortion-aware features could enhance the accuracy of modelling human visual attention in distorted images. Building upon this insight, we devise the Distortion-Aware Saliency Module (DASM), a novel approach designed specifically to integrate distortion-aware features into the saliency prediction process, as the structure illustrated in Figure 2.2. DASM is based on the principle of the self-attention mechanism, which has been proven highly effective in perception-related tasks, such as visual saliency prediction [21] and image quality assessment [40]. Let $\mathcal{F}_i$ represent the output saliency feature maps from the transformer encoder T_i. The Distortion-Aware Features ($\mathcal{F}_i^{DA}$) are obtained from the DASM$_i$ by processing $\mathcal{F}_i$. More specifically, $\mathcal{F}_i$ is first transformed linearly to generate $\mathcal{F}_i^q$, $\mathcal{F}_i^k$, and $\mathcal{F}_i^v$ for the self-attention mechanism, i.e., $\mathcal{F}_i^q = \mathrm{FC}_q(\mathcal{F}_i)$, $\mathcal{F}_i^k = \mathrm{FC}_k(\mathcal{F}_i)$, and $\mathcal{F}_i^v = \mathrm{FC}_v(\mathcal{F}_i)$. Then $\mathcal{F}_i^{DA}$ is obtained by:

$$\mathcal{F}_i^{DA} = h(h(\mathcal{G}_i)), \tag{1}$$

where:

$$\mathcal{G}_i = \mathrm{Softmax}(\mathcal{F}_i^q \times (\mathcal{F}_i^k)^T) \times \mathcal{F}_i^v + \mathcal{F}_i, \tag{2}$$

$$h(\cdot) = \mathrm{ReLU}(\mathrm{FC}(\cdot)), \tag{3}$$

FC and ReLU represent a fully connected layer (FC) and a ReLU function, respectively. The $\mathcal{F}_i^{DA}$ is fused into the decoder stream by a fusion function [46], which can be expressed as:

$$\mathcal{F}_i^{Fused} = \mathcal{F}_i^* \times \mathcal{F}_i^{DA} + \mathcal{F}_i^*, \tag{4}$$

where $\mathcal{F}_i^*$ denotes the saliency features that are fused with $\mathcal{F}_i^{DA}$ in the CNN decoder. Details regarding the specific layers where fusion takes place are illustrated in Figure 3.

To ensure that DASM generates distortion-aware saliency features, the mechanism based on MMOS (see details in Section 3.3) is introduced to explicitly guide feature expression during its training stage. To facilitate this capability, we make DASM derive a image quality score from the intended

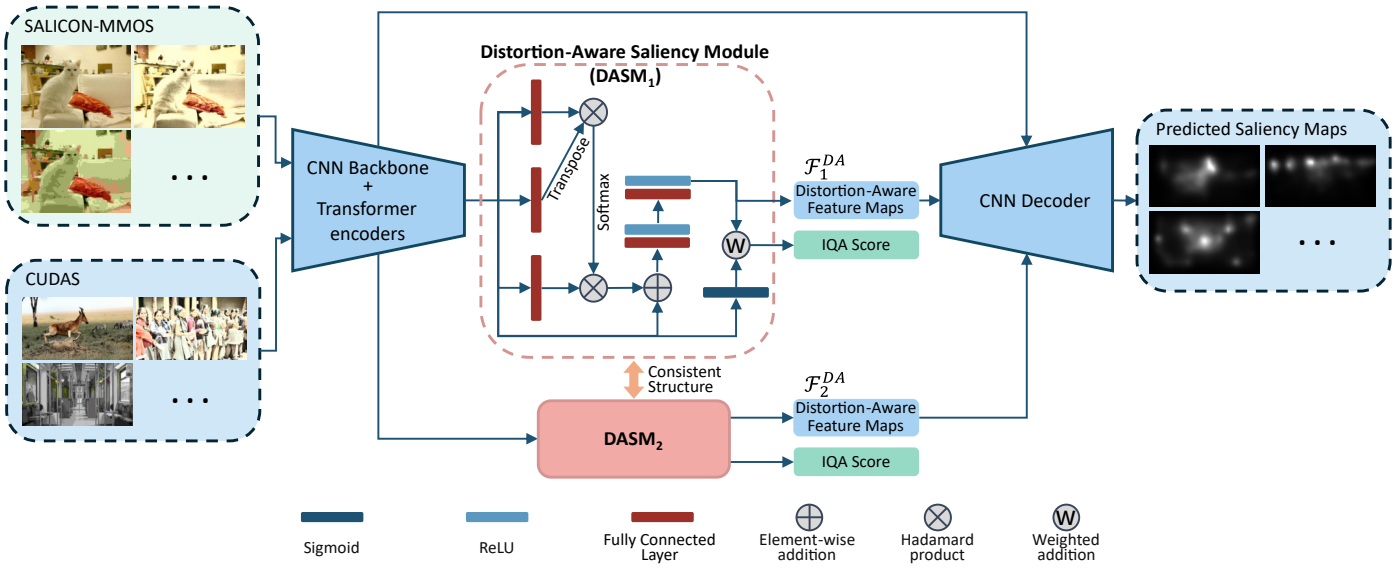

Figure 2: Schematic overview of the proposed architecture (*TranSalNet+*) for distortion-aware visual saliency prediction. The input is encoded by a deep learning encoder consisting of a CNN backbone and transformer encoders. The DASMs yield IQA scores for the distortion-aware training phase using the SALICON-MMOS dataset, and generate the distortion-aware features for learning the intended task of predicting saliency of distorted visual scenes. The predicted saliency maps are generated by the CNN Decoder based on the features from the encoder and DASMs.

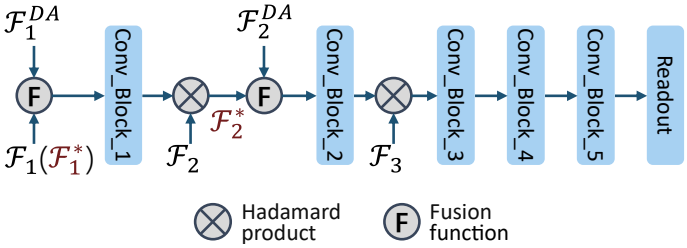

Figure 3: Details of the CNN Decoder, where $\mathcal{F}_i, i = 1, 2, 3$ denotes the saliency features from encoding phase (i.e., T_i, $i = 1, 2, 3$); $\mathcal{F}_i^{DA}$ and $\mathcal{F}_i^*$ denote the distortion-aware features from DASM$_i$ and the features that are fused with $\mathcal{F}_i^{DA}$ in the CNN decoder.

distortion-aware saliency features. According to previous research [40], the overall quality score of an image can be derived by allocating different weight scores to distinct regions of its features. Inspired by this concept, in the DASM, $\mathcal{F}_i$ (i.e., the saliency features) are employed as weights to perform a weighted summation of $\mathcal{F}_i^{DA}$, thereby yielding an image quality score. This process can be expressed as:

$$QS = \frac{\sum \left(\mathcal{F}_i^{DA} \times \sigma(\mathcal{F}_i)\right)}{\sum \sigma(\mathcal{F}_i)}, \tag{5}$$

where QS denotes the predicted image quality score and σ denotes the sigmoid function. Through this process, the quality perception of an image is modelled together with visual saliency, yielding distortion-aware saliency features.

3.3. MMOS-based Mechanism

For deep learning-based methods, employing a large-scale and directly relevant training dataset can often be effective in achieving the intended task [47]. However, there is a lack of large-scale saliency datasets in the context of image quality perception, as existing datasets are typically confined to a few hundred samples [26]. To circumvent this problem and avoid conducting a large-scale perception experiment of eye-tracking and image quality assessment, we provide a practical solution to create a contextually relevant dataset, namely SALICON-MMOS. The intention is to explicitly guide DASM to generate distortion-aware saliency features. We draw upon an existing large-scale saliency dataset (i.e., SALICON) and expand it by including distorted image variants with their quality assessed by a Machine Mean Opinion Score (MMOS). For images of the SALICON dataset, their distorted variants are created using the same protocol as the CUDAS benchmark [26]. Specifically, each image is distorted by three types of distortion including contrast change (CnC), JPEG compression (JPEG), and motion blur (MB); and the magnitude of each distortion type is varied to generate three perceptually distinct quality levels i.e., Q1 (perceptible but not noticeable distortion), Q2 (annoying distortion), and Q3 (very annoying distortion). Consequently, nine distorted variants (three distortion types × three distortion levels) are created for each image in the SALICON dataset, resulting in an expanded set of stimuli named X-SALICON.

The concept of the Machine Mean Opinion Score (MMOS) is informed by the definition of the Mean Opinion Score (MOS) derived from subjective IQA [48]. Given the high correlation between the subjective IQA and objective IQA represented by state-of-the-art IQA models, we use a model to rapidly generate a MMOS for each image contained in the X-SALICON dataset. We choose three highly reliable deep learning-based IQA models; namely MANIQA [40], HyperIQA [41], and TReS [42] to individually assess the quality of an image; and the results

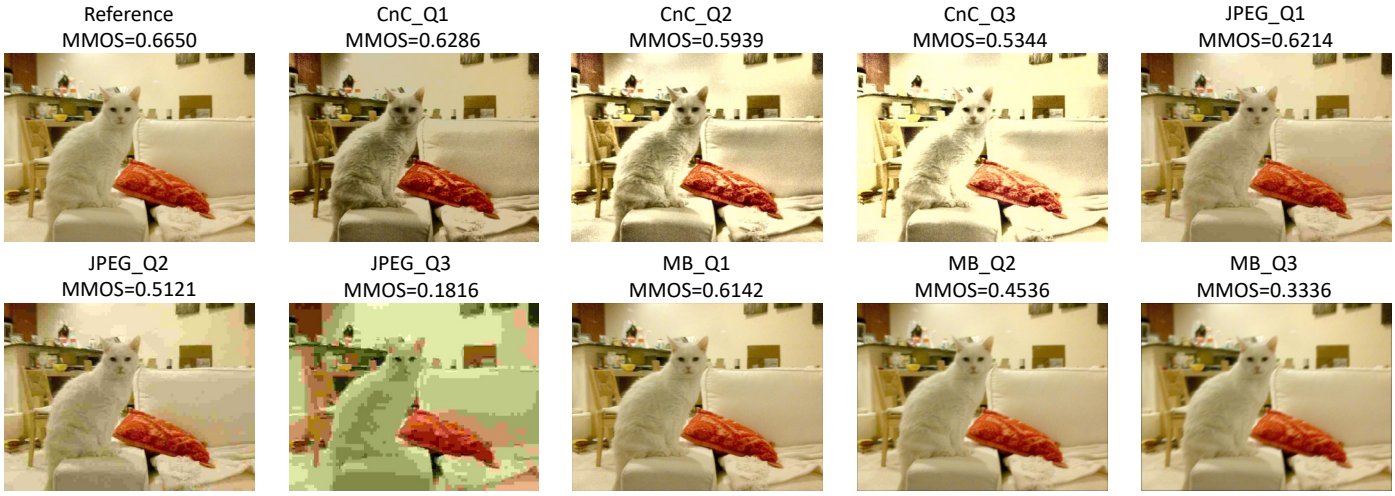

Figure 4: Examples from SALICON-MMOS dataset, including images and their corresponding MMOS values. The top-left image is the original scene from SALICON [27], followed by its nine distorted variants. CnC, JPEG, and MB represent distortions types of contrast change, JPEG compression, and motion blur, respectively; Q1, Q2, and Q3 denote low, medium, and high levels of distortion intensity.

are averaged to yield the MMOS. These three models represent distinct deep learning modelling paradigms (i.e., distinct network architectures) for the IQA task, providing diverse capabilities in generating IQA scores. More specifically, MANIQA utilises a multi-dimension attention network, HyperIQA adopts a self-adaptive hyper network architecture, and TReS leverages a hybrid architecture combining CNN and transformers. This approach mitigates potential biases inherent in individual deep models and hence enhances the robustness of the MMOS [49]. In our implementation, first IQA models are each applied to the images of the X-SALICON dataset. Then z-scores are calculated to calibrate the scores of different IQA models towards the same mean and standard deviation. The z-score (z_{ij}) of the i-th IQA model on the j-th image can be computed as:

$$z_{ij} = \frac{r_{ij} - \mu_i}{\sigma_i}, \tag{6}$$

where r, μ, and σ denote the raw IQA score, mean IQA score, and standard deviation, respectively. The outlier detection method described in [50] is applied, and no outliers are found in the z-scores. Finally, the MMOS for each image is calculated as the mean of the z-scores over all IQA models:

$$MMOS_j = \frac{1}{s}\sum_{i=1}^{s} z_{ij}, \tag{7}$$

where $s = 3$ is the number of IQA models. To enhance the interpretability of the final MMOS, these scores are linearly rescaled to be within the [0, 1] range. Consequently, the SALICON-MMOS dataset encompasses 100,000 images in its training set and 50,000 in the validation set, with reference and distorted variants, and their corresponding MMOS. A set of images from the SALICON-MMOS dataset, including the original image and its nine distorted variants and their corresponding MMOS values, is illustrated in Figure 4. It can be seen that for the same distortion type, the MMOS decreases as the distortion intensity increases, which aligns with the subjective image quality perception. The resulting SALICON-MMOS dataset that contains the expanded stimuli with distortions and their MMOS values will be used for the feature guidance of the DASM.

4. Experimental Results

4.1. Datasets

- *SALICON-MMOS*: An extended set of stimuli of the SALICON dataset [27], including both original images and their distorted variants, all assigned with an MMOS value. The details are detailed in Section 3.3.
- *CUDAS* [26]: The distortion-aware saliency benchmark, comprising 600 images of different distortion types of varying perceived quality. Ground truth data of saliency and image quality were collected via eye-tracking and subjective IQA experiments. The details are detailed in Section 2.3.

4.2. Evaluation Metrics

A range of metrics have been established to quantify the agreement between the predicted saliency map and the ground truth. Following previous research [51, 52, 23], we adopt five widely used metrics to thoroughly evaluate the performance of the visual saliency model on distorted images. These metrics include Pearson correlation coefficient (CC), Similarity (SIM), Kullback-Leibler Divergence (KLD), Area Under Curve (AUC), and Normalised Scanpath Saliency (NSS). A brief introduction to these saliency metrics is provided below. Let P, S, and F be the predicted saliency map, ground truth saliency map, and ground truth fixation map, respectively.

- CC is a statistical method to measure the correlation of two variables, which evaluates the accuracy of saliency prediction by:

$$\mathrm{CC}(P, S) = \frac{\mathrm{cov}(P, S)}{\sigma(P) \cdot \sigma(S)}, \tag{8}$$

where $\mathrm{cov}(\cdot)$ is the covariance and $\sigma(\cdot)$ is the standard deviation.

- SIM measures the similarity between the predicted and the ground truth saliency maps by:

$$\mathrm{SIM}(P, S) = \sum_i \min(P_i, S_i), \tag{9}$$

where i indexes the i-th pixel; $\sum_i \mathbf{P}_i = \sum_i \mathbf{S}_i = 1$.

- KLD metric is calculate by:

$$\mathrm{KLD}(P, S) = \sum_i S_i \log\left(\epsilon + \frac{S_i}{\epsilon + P_i}\right), \tag{10}$$

where i indexes the i-th pixel; ϵ is used as a regularisation constant, which is set to 2.2204×10^{-16} as per [51].

- NSS is a metric specifically designed for saliency evaluation, which uses the fixation map (F) as the ground truth reference. It can be calculated by:

$$\begin{aligned} \mathrm{NSS}(P, F) &= \frac{1}{N} \sum_i \bar{P}_i \times F_i \\ \text{where} \quad N &= \sum_i F_i \\ \bar{P} &= \frac{P - \mu(P)}{\sigma(P)}, \end{aligned} \tag{11}$$

where i indexes the i-th pixel; N represents the total number of fixated pixels; $\mu(\cdot)$ is the mean; $\sigma(\cdot)$ is the standard deviation.

- AUC assesses the predicted saliency map as a binary classifier by varying thresholds to determine if pixels are actual fixation points. The ROC (receiver operating characteristic) curve is derived from the true and false positive rates of these threshold-based classifiers.

For KLD, values approaching zero indicate an optimal agreement with the ground truth. For other saliency metrics, higher values indicate a closer alignment with the ground truth.

4.3. Implementation Details

The implementation of our *TranSalNet+* model involves two phases. The first phase is training the DASMs. The backbone and transformer encoders were initialised by the pre-trained weights on SALICON and then frozen. Only the DASMs were trained on the SALICON-MMOS dataset using the mean squared error (MSE) loss to produce IQA scores. Optimal models were obtained through early stopping after 5 epochs of patience, using the AdamW optimiser [53]. The optimisation started with a batch size of 32 and an initial learning rate of 4×10^{-4}, decaying by a factor of 0.1 every epoch.

The second phase is training the proposed architecture for distortion-aware saliency prediction. The saliency prediction components of the model were initialised by the pre-trained weights on SALICON (as per recommendations in [26]), and the DASMs were initialised by the best parameters from the first training phase. We conducted a k-fold cross-validation ($k = 6$) for comprehensive model evaluation on the CUDAS dataset, which was divided into six equal, non-overlapping sets. In each cross-validation cycle, one set was designated for testing, one for validation, and the remaining four sets for training. This approach ensured no overlap or parameter sharing between the cycles was allowed; and the models were tested on unseen samples. The final results represented the mean performance across all six runs. The loss function used for training was a linear combination of CC, SIM, NSS, and KLD, as detailed in [21]. Early stopping and the optimiser were consistent with the first phase, with a batch size of four and an initial learning rate of 8×10^{-5}, decaying by a factor of 0.1 every two epochs.

4.4. Ablation Study

We investigate the contribution of the proposed DASM to distortion-aware saliency prediction. Let DASM_1, DASM_2, and DASM_3 represent the DASMs connected to different encoding levels, i.e., T_1, T_2, and T_3, respectively. To this end, we devised seven model variants, denoted as Variants *A* to *F*. These model variants include the variant without DASM and those utilising a single DASM, i.e., DASM_1, DASM_2, or DASM_3, and a combination of them. Table 1 presents the performance of these model variants on the CUDAS benchmark. The results demonstrate: (1) Variant *A*, without the Distortion-Aware Saliency Module (DASM), and consequently lacking the integration of distortion-aware features informed by the MMOS, consistently exhibits the lowest performance across the majority of metrics. This observation underscores the critical role of distortion-aware features in enhancing saliency prediction for distorted images. (2) In comparing the performance among variants *B*, *C*, and *D*, it is observed that the integration of a single DASM at different encoding levels can improve model performance, suggesting that distortion-aware features generated by DASM are effective for the intended task. However, incorporating DASM into the deeper encoding layers (e.g., T_1 and T_2) of the network yields larger performance improvements, implying that the deeper features are more capable of representing the distortion related features for saliency prediction. (3) In comparing the performance among all variants, the combination of DASM_1 and DASM_2 achieves the highest performance, surpassing the use of DASM_1, DASM_2, and DASM_3 combined together. This suggests that distortion-aware features derived from shallower levels have low representational capability for distortion-aware saliency prediction. Consequently, we adopt the model variant employing DASM_1 and DASM_2 as our definitive model.

4.5. Comparison with the state-of-the-art

To validate the effectiveness of our proposed model against the state-of-the-art for distortion-aware saliency prediction, we conduct a comparative analysis on the CUDAS benchmark. The selection criteria of models are: (1) top-tier performance on general-purpose saliency benchmarks, and (2) the availability

Table 1: Ablation study results on the distortion-aware saliency benchmark (CUDAS) [26]. **Bold** and ***Italicised Bold*** fonts indicate the best and second-best performance, respectively.

Variant	$DASM_1$	$DASM_2$	$DASM_3$	CC ↑	SIM ↑	KLD ↓	AUC ↑	NSS ↑
A	–	–	–	0.7853	0.7070	0.4177	0.8227	1.6814
B	✓	–	–	***0.7935***	***0.7136***	**0.4045**	***0.8241***	1.6916
C	–	✓	–	0.7918	0.7118	0.4136	0.8234	***1.6965***
D	–	–	✓	0.7894	0.7087	0.4170	0.8232	1.6910
E	✓	✓	–	**0.7979**	**0.7156**	***0.4112***	**0.8246**	**1.7057**
F	✓	✓	✓	0.7893	0.7110	0.4258	0.8240	1.6943

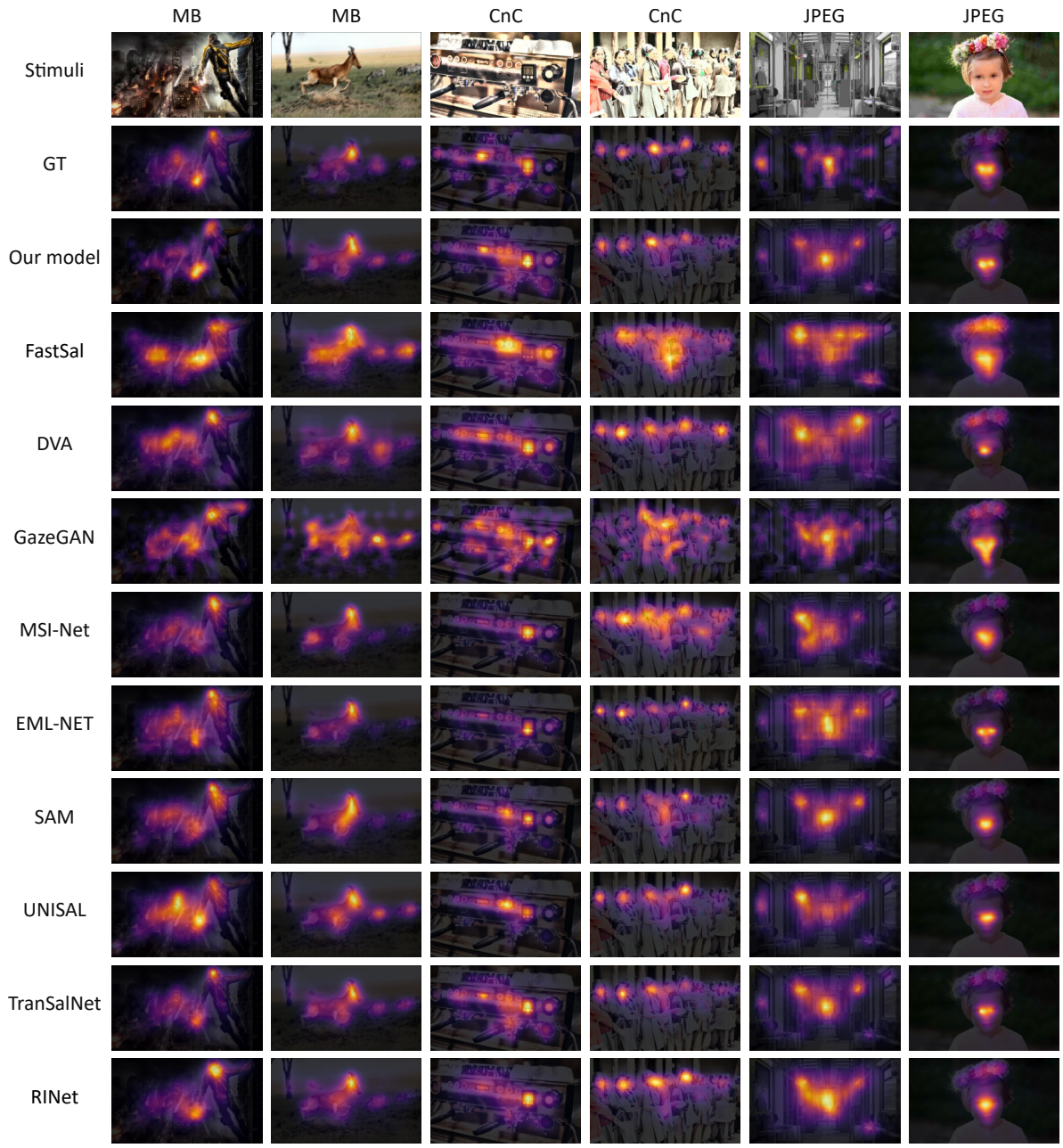

Figure 5: Examples of ground truth (GT) versus predicted saliency of distorted visual scenes on the CUDAS benchmark [26]. From left to right, the first two columns present scenes with motion blur (MB) distortion, the third and fourth columns illustrate scenes with contrast change (CnC) distortion, and the fifth and sixth columns show scenes with JPEG compression (JPEG) distortion. The top two rows represent the distorted stimuli and their ground truth (GT) saliency maps, the other rows show the predicted saliency of our proposed model and state-of-the-art saliency models.

Table 2: Performance comparison of our proposed model *TranSalNet+* and state-of-the-art saliency models on the distortion-aware saliency benchmark (CUDAS) [26]. **Bold** and ***Italicised Bold*** fonts indicate the best and second-best performance, respectively.

Molde Name	CC ↑	SIM ↑	KLD ↓	AUC ↑	NSS ↑
FastSal [19]	0.6940	0.6522	0.5801	0.8004	1.3627
DVA [14]	0.6940	0.6549	0.4269	0.8032	1.4689
GazeGAN [16]	0.7022	0.6535	0.6973	0.8006	1.4386
MSI-Net [17]	0.7379	0.6559	0.4895	0.8194	1.6194
EML-NET [15]	0.7603	0.6950	0.8320	0.8182	1.6759
SAM [13]	0.7672	0.7003	0.5244	0.8193	1.6281
UNISAL [18]	0.7781	0.7017	***0.3857***	0.8194	1.6194
TranSalNet [21]	0.7853	***0.7070***	0.4177	***0.8227***	***1.6814***
RINet [20]	***0.7911***	0.7054	**0.3299**	0.8217	1.6446
Proposed (*TranSalNet+*)	**0.7979**	**0.7156**	0.4112	**0.8246**	**1.7057**

of implementation code for pre-trained models. To ensure a fair comparison, the same implementation protocol as detailed in Section 4.3 for our proposed model including initialisation with SALICON pre-trained weights and 6-fold cross-validation on CUDAS is applied for all other models in the comparative study. Table 2 shows the results of performance comparison, with some examples of visual assessment illustrated in Figure 5. It can be seen that our model achieves the ***best*** scores in CC, SIM, AUC, and NSS, and a competitive score in KLD, demonstrating its overall superiority in predicting visual saliency of distorted images.

4.6. Discussion

Previous studies have demonstrated the importance of carefully selecting appropriate metrics to evaluate the performance of visual saliency models. It is suggested that the choice of saliency metrics should align with the specific application under study [51, 52]. Amongst commonly used saliency metrics, NSS, CC, and SIM are more closely aligned with human perception [51, 52]. Hence these metrics may provide a more suitable evaluation for visual saliency prediction in application scenarios where distortions affect viewers' visual experiences [23]. As shown in Table 2, the proposed model achieves the best results across all metrics except for KLD. Particularly noteworthy is our model's exceptional performance on the NSS, CC, and SIM metrics, demonstrating its superiority in predicting visual saliency in the more demanding conditions, such as distorted images.

The aim of the proposed model is to predict visual saliency for images of varying levels of distortion. In our study, we specifically focus on three common types of distortion, i.e., contrast change, JPEG compression, and motion blur. While these distortions are widely studied, and hence serve as representative examples in our context, it should be noted that they do not encompass all possible distortion types. Other types of distortion, such as white noise and Gaussian blur, can also significantly impact visual attention in various real-world applications. Future research could focus on evaluating the proposed model's performance across a broader spectrum of diverse distortions and improving the robustness of saliency prediction.

5. Conclusion

In this paper, we have presented our work towards predicting saliency of distorted visual scenes. To tackle this challenge, we propose a deep learning Distortion-Aware Saliency Module (DASM), which learns the representational features related to image distortion for the task of saliency prediction. To facilitate the feature expression capability of DASM during training, we create a SALICON-MMOS dataset to encompass images with distortion and their quality assessments derived by a Machine Mean Opinion Score (MMOS). Experimental results have substantiated the efficacy of our proposed model, outperforming state-of-the-art models in predicting saliency of distorted images.

References

[1] J. Jonides, D. Irwin, S. Yantis, Integrating visual information from successive fixations, Science 215 (4529) (1982) 192–194.

[2] E. E. M. Stewart, M. Valsecchi, A. C. Schütz, A review of interactions between peripheral and foveal vision, Journal of Vision 20 (12) (2020) 2–2.

[3] K. Koch, J. McLean, R. Segev, M. A. Freed, M. J. Berry, V. Balasubramanian, P. Sterling, How much the eye tells the brain, Current biology 16 (14) (2006) 1428–1434.

[4] P. Lennie, The cost of cortical computation, Current biology 13 (6) (2003) 493–497.

[5] Y. Han, B. Han, X. Gao, Human scanpath estimation based on semantic segmentation guided by common eye fixation behaviors, Neurocomputing 453 (2021) 705–717.

[6] W. Zhou, G. Yue, R. Zhang, Y. Qin, H. Liu, Reduced-reference quality assessment of point clouds via content-oriented saliency projection, IEEE Signal Processing Letters 30 (2023) 354–358.

[7] Y. Liu, D. Zhang, Q. Zhang, J. Han, Part-object relational visual saliency, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (7) (2022) 3688–3704.

[8] J. Chen, Q. Li, H. Ling, D. Ren, P. Duan, Audiovisual saliency prediction via deep learning, Neurocomputing 428 (2021) 248–258.

[9] Y. Jiang, L. A. Leiva, H. Rezazadegan Tavakoli, P. R. B. Houssel, J. Kylmälä, A. Oulasvirta, UEyes: Understanding visual saliency across user interface types, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1–21.

[10] J. Lou, H. Lin, P. Young, R. White, Z. Yang, S. Shelmerdine, D. Marshall, E. Spezi, M. Palombo, H. Liu, Predicting radiologists' gaze with computational saliency models in mammogram reading, IEEE Transactions on Multimedia 26 (2024) 256–269.

[11] D. Walther, C. Koch, Modeling attention to salient proto-objects, Neural Network 19 (9) (2006) 1395–1407, brain and Attention.

[12] J. Harel, C. Koch, P. Perona, Graph-Based Visual Saliency, in: Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS), NIPS'06, MIT Press, Cambridge, MA, USA, 2006, p. 545–552.

[13] M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, Predicting human eye fixations via an LSTM-based saliency attentive model, IEEE Transactions on Image Processing 27 (10) (2018) 5142–5154.

[14] W. Wang, J. Shen, Deep Visual Attention Prediction, IEEE Transactions on Image Processing 27 (5) (2018) 2368–2378.

[15] S. Jia, N. D. Bruce, EML-NET: An expandable multi-layer network for saliency prediction, Image and Vision Computing 95 (2020) 103887.

[16] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, P. Le Callet, How is Gaze Influenced by Image Transformations? Dataset and Model, IEEE Transactions on Image Processing 29 (2020) 2287–2300.

[17] A. Kroner, M. Senden, K. Driessens, R. Goebel, Contextual encoder–decoder network for visual saliency prediction, Neural Network 129 (2020) 261–270.
[18] R. Droste, J. Jiao, J. A. Noble, Unified Image and Video Saliency Modeling, in: Proceedings of 16th European Conference on Computer Vision, 2020, pp. 419–435.
[19] F. Hu, K. McGuinness, FastSal: a computationally efficient network for visual saliency prediction, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 9054–9061.
[20] Y. Song, Z. Liu, G. Li, D. Zeng, T. Zhang, L. Xu, J. Wang, RINet: Relative importance-aware network for fixation prediction, IEEE Transactions on Multimedia 25 (2023) 9263–9277.
[21] J. Lou, H. Lin, D. Marshall, D. Saupe, H. Liu, TranSalNet: Towards perceptually relevant visual saliency prediction, Neurocomputing 494 (2022) 455–467.
[22] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612.
[23] X. Yang, F. Li, H. Liu, A measurement for distortion induced saliency variation in natural images, IEEE Transactions on Instrumentation and Measurement 70 (2021) 1–14.
[24] H. Liu, I. Heynderickx, Visual attention in objective image quality assessment: Based on eye-tracking data, IEEE Transactions on Circuits and Systems for Video Technology 21 (7) (2011) 971–982.
[25] S. M. H. Miangoleh, Z. Bylinskii, E. Kee, E. Shechtman, Y. Aksoy, Realistic saliency guided image enhancement, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 186–194.
[26] X. Zhao, J. Lou, X. Wu, Y. Wu, L. Lévéque, X. Liu, P. Guo, Y. Qin, H. Lin, D. Saupe, H. Liu, CUDAS: Distortion-aware saliency benchmark, IEEE Access 11 (2023) 58025–58036.
[27] M. Jiang, S. Huang, J. Duan, Q. Zhao, SALICON: Saliency in context, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1072–1080.
[28] A. Torralba, A. Oliva, M. Castelhano, J. Henderson, Contextual Guidance of Eye Movements and Attention in Real-World Scenes: The Role of Global Features in Object Search., Psychological review 113 (2006) 766–786. doi:10.1037/0033-295X.113.4.766.
[29] B. Schauerte, R. Stiefelhagen, Quaternion-Based Spectral Saliency Detection for Eye Fixation Prediction, in: A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, C. Schmid (Eds.), Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II 12, 2012, pp. 116–129.
[30] E. Erdem, A. Erdem, Visual saliency estimation by nonlinearly integrating features using region covariances, Journal of vision 13 (4) (2013) 11.
[31] A. Borji, Saliency prediction in the deep learning era: Successes and limitations, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (2) (2021) 679—700.
[32] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.
[33] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[34] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2261–2269.
[35] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11966–11976.
[36] Y. Liu, D. Cheng, D. Zhang, S. Xu, J. Han, Capsule networks with residual pose routing, IEEE Transactions on Neural Networks and Learning Systems (2024) 1–14.
[37] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612.
[38] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, et al., Image database tid2013: Peculiarities, results and perspectives, Signal processing: Image communication 30 (2015) 57–77.
[39] V. Hosu, H. Lin, T. Sziranyi, D.Saupe, Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment, IEEE Transactions on Image Processing 29 (2020) 4041–4056.
[40] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, Y. Yang, MANIQA: Multi-dimension attention network for no-reference image quality assessment, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2022, pp. 1191–1200.
[41] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, Y. Zhang, Blindly assess image quality in the wild guided by a self-adaptive hyper network, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3667–3676.
[42] S. A. Golestaneh, S. Dadsetan, K. M. Kitani, No-reference image quality assessment via transformers, relative ranking, and self-consistency, in: Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 3209–3218.
[43] RECOMMENDATION ITU-R BT.500-11 Methodology for the subjective assessment of the quality of television pictures, Tech. rep. (1974).
[44] G. Keren, C. Lewis, A Handbook for Data Analysis in the Behavioral Sciences: Volume 1: Methodological Issues Volume 2: Statistical Issues, 1st Edition, Psychology Press, 1993.
[45] D. Shakow, Within-subjects designs: To use or not to use?, Psychological Bulletin 83 (2) (1976) 314–320.
[46] T. Lüddecke, A. Ecker, Image segmentation using text and image prompts, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7086–7096.
[47] C. Chen, J. Han, K. Debattista, Virtual category learning: A semi-supervised learning method for dense prediction with extremely limited labels, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024) 1–17.
[48] V. Hosu, H. Lin, T. Sziranyi, D. Saupe, KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment, IEEE Transactions on Image Processing 29 (2020) 4041–4056.
[49] G. Vardi, On the implicit bias in deep-learning algorithms, Communications of the ACM 66 (6) (2023) 86–93.
[50] H. Liu, N. Klomp, I. Heynderickx, A no-reference metric for perceived ringing artifacts in images, IEEE Transactions on Circuits and Systems for Video Technology 20 (4) (2010) 529–539.
[51] Z. Bylinskii, et al., What Do Different Evaluation Metrics Tell Us About Saliency Models?, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (3) (2019) 740–757.
[52] J. Li, C. Xia, Y. Song, S. Fang, X. Chen, A data-driven metric for comprehensive evaluation of saliency models, in: International Conference on Computer Vision, 2015, pp. 190–198.
[53] I. Loshchilov, et al., Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).