Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Causal Perception Inspired Representation Learning for Trustworthy Image Quality Assessment

Lei Wang, Qingbo Wu, Desen Yuan, Fanman Meng, Zhengning Wang, King Ngi Ngan

University of Electronic Science and Technology of China

ARTICLE INFO

Communicated by

2000 MSC:
41A05
41A10
65D05
65D17

Keywords:
Trustworthy Image Quality Assessment
Adversarial Attack
Causal Perception

ABSTRACT

Despite great success in modeling visual perception, deep neural network based image quality assessment (IQA) still remains untrustworthy in real-world applications due to its vulnerability to adversarial perturbations. In this paper, we propose to build a trustworthy IQA model via Causal Perception inspired Representation Learning (CPRL). More specifically, we assume that each image is composed of Causal Perception Representation (CPR) and non-causal perception representation (N-CPR). CPR serves as the causation of the subjective quality label, which is invariant to the imperceptible adversarial perturbations. Inversely, N-CPR presents spurious associations with the subjective quality label, which may significantly change with the adversarial perturbations. We propose causal intervention to boost CPR and eliminate N-CPR. Specifically, we first generate a series of N-CPR intervention images, and then minimize the causal invariance loss. Then we propose a SortMask module to reduce Lipschitz and improve robustness. SortMask block small changes around the mean to eliminate N-CPR and can be plug-and-play. Experiments on four benchmark databases show that the proposed CPRL method outperforms many state-of-the-art methods and provides explicit model interpretation. To support reproducible scientific research, we release the code at `https://clearlovewl.github.io`.

1. Introduction

The rapid growth of image data on the Internet has made the automatic evaluation of image quality a vital research and application topic. Objective image quality assessment (IQA) methods can be divided into three categories based on the availability of the original undistorted images: full-reference (FR), reduced-reference (RR), and no-reference/blind (NR/B). Among them, BIQA model is a challenging computational vision task, which mimics the human ability to judge the perceptual quality of a test image without requiring the original image content[1, 2, 3, 4, 5]. However, existing BIQA methods are not trustworthy, and minor perceptual attacks can fool the quality evaluator model to produce incorrect output. This vulnerability affects the security issues, resulting in lower reliability of IQA[6]. As shown in Fig 1, after adding minor perturbations, the prediction results of the deep IQA model showed significant errors while the human eye does not perceive the change in quality. In image classification, this is known as adversarial attack (e.g., 'norm constrained attack). A trustworthy IQA model should not be overly sensitive to adversarial attacks, that is, be adversarially robust[7, 8, 9].

The existing SOTA BIQA models use Deep Neural Network (DNN) architectures, which naturally inherits its vulnerability

e-mail: `qbwu@uestc.edu.cn` (Qingbo Wu)

http://dx.doi.org/10.1016/j.displa.xxxx.xx.xxx
Received ; Received in final form ; Accepted ; Available online

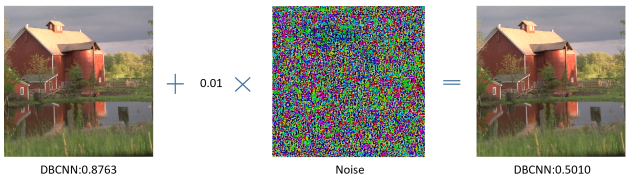

Fig. 1: Demonstration of IQA model adversarial example generation). By adding an imperceptible perturbation, we can drastically change the predicted score of an IQA model for an image.

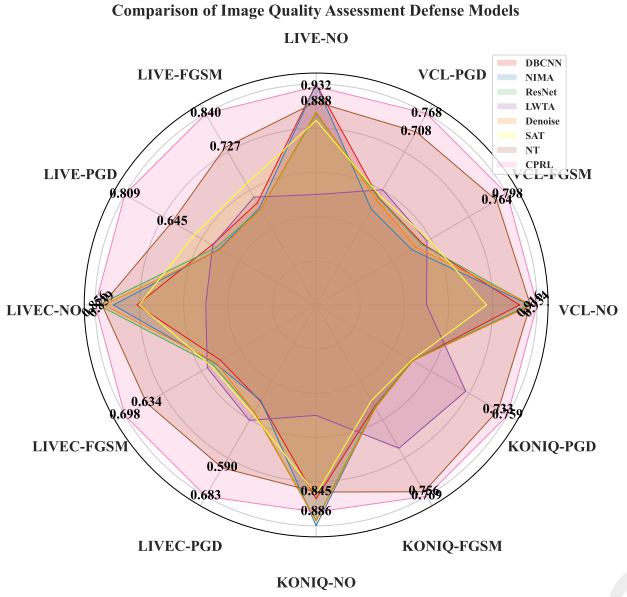

Fig. 2: The CPRL Method in comparison with existing state-of-the-arts, showing exceptional improvements on different attack method and datasets. Metrics are the (SRCC+PLCC)/2.

problem. Slight interference within Just Noticeable Difference (JND) [10] will not affect the perceptual quality of the human eye. The specific reason can be related to the all or nothing principle [11] in Human Visual System (HVS). However, DNN uses a multi-layer structure with high-dimensional dot product operations. This allows small changes in features to accumulate [12], forming a dense mixture and leading to an excessively large Lipschitz constant [13]. This ultimately leads to vulnerabilities different from HVS. The prevalent approach to adversarial robustness remains empirical adversarial training and its variants. However, adversarial training can be seen as a form of passive defense, which is limited to specific attack types and fails to generalize to more diverse and complex attack scenarios. Model-based denoising defense is another important defense strategy. However, in quality assessment tasks, denoising models may backfire and have adverse effects. This is because the denoising operation itself alters image quality without affecting image semantics[14]. Recently, researchers analyzed adversarial examples from the perspective of non-robust features and robust features, and proved that adversarial examples can be viewed as non-robust features associated with labels[7].

Considering the challenges above, We propose a new way to understand the IQA task by using a causal approach. We use a causal framework that shows how the images in the IQA dataset are produced by three factors: reference images, causal perception representation (CPR) and non-causal perception representation (N-CPR). We define N-CPRs as confounders that do not affect image quality but can create false correlations. These confounders can make the model predictions wrong, and they can be any feature that is related to some labels, such as local texture, small edges, and faint shadows within JND. Fig. 3 illustrates the structural causal model (SCM) that we use to describe how the data for the IQA dataset is generated. The normal IQA training and testing can be seen as changing the reference image R while keeping the N-CPR fixed. In this case, the N-CPR are stable, so the model can work well from the training to the test scenario. But in an adversarial scenario, someone can change the predicted image quality by changing the N-CPR without changing how the image looks. Thus IQA models that rely solely on statistical associations are insufficient and untrustworthy[15]. Because they do not use any prior knowledge to remove the confounders and stop the false correlations.

We propose causal perception inspired representation learning (CPRL) to build a trustworthy IQA model. Our method uses causal intervention to break false correlations (N-CPR) and find the true cause of image quality (CPR). Based on theoretical analysis, we develop a causal intervention strategy and prior knowledge to improve CPR and eliminate N-CPR. Specifically, to improve CPR, we first generate a series of N-CPR intervention images, and then minimize the causal invariance loss. To eliminate N-CPR, we propose a plug-and-play SortMask module to reduce Lipschitz constant by blocking small changes around the mean.

To further validate our method, we conduct experiments on 8 baseline models and 5 attack methods. The results show the effectiveness of the CPRL strategy in improving the robustness of the baseline model against adversarial attacks. To the best of our knowledge, this is the first time that a defense method for causal interpretation of NR-IQA models is proposed and theoretically supported.

2. Related Work

2.1. Image quality assessment

Image quality assessment aims to obtain an objective model[1] to predict the image quality, making it close to the subjective quality score. Interpretable and trustworthy structural similarity (SSIM)[1] and peak signal-to-noise ratio (PSNR) are favored by researchers as a reference image quality method. However, researchers have been looking for methods similar to the human visual system(HVS) that do not require reference pictures which have been called no-reference(NR) IQA methods[16]. Due to the lack of information on reference pictures, naturally, the researchers used the statistical information of pictures as a reference, called natural scene statistics (NSS)[17], and proposed Natural Image Quality Evaluator (NIQE) [18]. Benefiting from the development of deep learning, some learning-based image quality models [19, 20, 21] have obtained stronger performance than traditional quality evaluation methods[22] due to the powerful feature extraction capabilities of convolution kernels[23]. In order to ex-

plore more feature expression capabilities, researchers proposed many other methods based on Generative adversarial network [5], Variational Auto-Encoder [24], and Transformers [25]. However, powerful feature representation is difficult to explain, so some researchers try to figure out whether the quality evaluation is decoupled from the content[26]. Therefore, this paper focuses on the robustness exploration of deep IQA models, and dedicates to discovering efficient ways to improve the robustness of IQA models.

2.2. Vulnerability of deep neural networks

The vulnerability of Deep Neural Networks is believed to be due to the differences between the features extracted by the model and the human visual system [7]. So small changes can make deep models ineffective [27, 28]. The most classical anti-sample attack method, Fast Gradient Sign Method (FGSM)[29] , which is based on gradient attack, has the following form:

$$\delta = \epsilon \cdot \operatorname{sign}\left(\nabla_x \ell\left(f(x), y\right)\right) \tag{1}$$

where δ is the adversarial perturbation, which equals to the original sample plus a small perturbation, and y is the true label MOS. The direction of the perturbation is determined by the gradient of the predictor output $f(x)$ and the label y with respect to the loss ℓ. The magnitude of the perturbation is given by ϵ. Researchers found that the parameters of the neural network have a dense mixture of terms [12], and adversarial training is doing feature purification which alleviates vulnerability.

2.3. Adversarial robust models

A more robust model is not only more consistent with human perceptual properties[30, 31], but also increases the ability to generalize outside the domain[32]. From the point of view of data preprocessing, Feature Squeezing [33] distinguishes between adversarial and clean samples by reducing the color bit depth of each pixel. From the point of view of feature extraction, Feature denoising [14] reduces noise in latent features. from the empirical results, it can be seen that the feature spectrum is purer than the original deep network. Cisse proposed Parseval networks [34], a layerwise parameters regularization method for reducing the network's sensitivity to small perturbations by carefully controlling its global Lipschitz constant. Qin proposed a Local Linearization regularizer [35] that encourages the loss to behave linearly in the vicinity of the training data. Kannan introduce enhanced defenses using a technique called logit pairing [36], a method that encourages logits for pairs of clean examples and their adversarial counterparts to be similar. All the above methods add additional constraints, such as controlling the Lipschitz constant to suppress the amplification effect of the network, or constraining the local linearity of features, or generating a purpose for feature alignment.

2.4. Causality inference

Various methods have been proposed to achieve causal inference, such as Matching Methods[37], Propensity Score based Methods[38], and Reweighting Methods[39, 40]. However, the balancing methods are not applicable in adversarial scenarios, because they mostly rely on the assumptions that the dataset is sufficient while non-perceptual representation are observed in natural images. Moreover, for high-dimensional images, there is not enough data for statistical analysis of causal estimation. Recently, some researchers have leveraged causal inference tools to optimize models from observable high-dimensional data [15, 41, 42]. Correspondingly, many perspectives have emerged to understand adversarial robustness from causal inference [9, 8, 43, 44, 45]. Furthermore, researchers use causal inference to obtain more trustworthy representations[46, 47, 48].

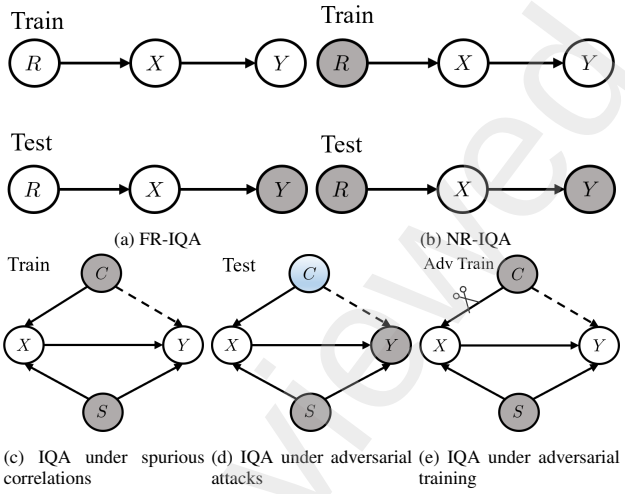

Fig. 3: Causal graphs of IQA learning. Grey nodes represent unobserved variables. (a) Cause graphs of the traditional FR-IQA task during training and testing phases. (b) Cause graphs of the traditional NR-IQA task during training and testing phases. (C) Cause graphs of NR-IQA learning under spurious correlations during the training phase. Due to shortcut learning, the network model learns spurious correlations from the path $X \leftarrow C \rightarrow Y$. (d) Cause graphs of NR-IQA learning during the testing phase. The non-aware variables in blue are not from the same distribution as the training set. During the test phase, the distribution of confounding variables C in blue changes (adversarial attacks). (e) Cause graphs of NR-IQA adversarial learning. Adversarial training obtains new samples X by changing the confounding variable C and adding them to the training. Think of it as a backdoor intervention.

3. Preliminaries

3.1. Trustworthy IQA in adversarial attack

To form a credible IQA task in adversarial scenarios, we assume that there is a dataset $\mathcal{D}$ consisting of tuples of images x_i, corresponding labels y_i, and reference images R_i. We can represent the data as $\mathcal{D} = \{(R_i, x_i, y_i)_{i=1}^N\}$, where N is the number of samples. The IQA task can be divided into NR and FR, depending on whether a reference image is given or not. A scene is FR if the reference image is observable during both training and testing phases. If the reference image is unobservable, then the scene is NR. The purpose of the adversarial attack defined in this paper is to find a perturbation δ that reduces the network prediction quality score $f(x)$, which can be expressed as:

$$\arg\max_{\delta} \ell(y, f(x+\delta)), \text{ s.t. } d(\delta) \leq \epsilon, \tag{2}$$

where d is a signal fidelity measure, ϵ is the bound of small perturbations, and $\ell(\cdot,\cdot)$ represents the loss criterion for out-

put and label. In practice, we define d as the ℓ_∞-norm measure $\|\cdot\|_\infty$. Following the above formula, we clarify the differences between adversarial scenarios and traditional IQA tasks. Let $P(\cdot)$ represent a marginal distribution. The difference between them is that the image X and its marginal distribution $P(X)$ do not satisfy the assumption of independent and identically distributed (i.i.d.) samples, that is, for traditional IQA tasks, $P(X_{\text{train}}) = P(X_{\text{test}})$. For adversarial scenarios, $P(X_{\text{train}}) \neq P(X_{\text{test}})$.

3.2. Causal graphs in IQA learning

A causal graph is a directed acyclic graph in which nodes represent variables and directed edges represent the causal relationship between two variables. We treat images, labels, reference images, causal perception representation (CPR) and non-causal perception representation (N-CPR) as nodes in a causal graph. Our causal graph reveals the underlying statistics of IQA learning, as we will show in the following sections. As shown in 3, we denote the image as X, the mean opinion score (MOS) as Y, CPR as S, and N-CPR as C. The edges are listed as follows: $X \rightarrow Y$ means that MOS Y is derived from image X; $R \rightarrow X$ means that the reference image R has a causal influence on image X; $C \rightarrow Y$ with dotted line means that the N-CPR variable C has a spurious correlation with MOS Y; $C \rightarrow X$ means that the N-CPR variable C has a causal influence on image X; $S \rightarrow X$ means that the CPR variable S has a causal influence on image X; $S \rightarrow Y$ means that the CPR variable S has a causal influence on MOS Y. For traditional IQA learning in 3(a)(b), its causal graph contains the causal relationship between the image variable X, the label Y, and the reference image R. During the training phase of supervised learning, both the image X and the label Y are observable, and we use them to learn the causal effects in the model. But in the testing phase, only the image X and the causal path are known, and Y is the label to be predicted. For NR-IQA learning, R is unobservable, while for FR-IQA learning, R is observable. For the trustworthy IQA learning scenario, we model an unobserved variable, the non-perceptual variable C, and they are illustrated in 3(c). Their differences in the test phase are shown in 3(d). Due to changes in the non-perceptual variable C, X may have an indirect effect on Y through the causal path $X \leftarrow C \rightarrow Y$, establishing a spurious correlation with C. The purpose of trustworthy IQA is to establish true quality-related correlations based on image quality evaluation, independent of non-perceptual variables.

4. Methodology

In this section, we present a causal framework to analyze the biases induced by non-perceptual features in IQA, a causal intervention method to eliminate them, and the implementation details and pipeline of our approach.

4.1. Non-perceptual-feature induced bias in IQA

According to the causal graph, we can explain why the IQA model fails in adversarial scenarios. We attribute this phenomenon to biases induced by non-perceptual features. In 3(c), the causal path $X \leftarrow C \rightarrow Y$ is a fork structure, and this path is spurious because it is outside the causal path from X to y. Opening this path will create non-perceptual feature-induced biases and produce erroneous outcomes. We give a simple example to illustrate this point. A causal graph A country's per capita chocolate consumption ← Economic conditions → Number of Nobel Prize winners There is a strong correlation between a country's per capita chocolate consumption and its number of Nobel Prize winners. This correlation seems absurd because we cannot imagine winning a Nobel Prize for eating chocolate. A more plausible explanation is that more people eat chocolate in wealthy Western countries, and Nobel laureates are preferentially selected from these countries. But this is a causal explanation, which leads to the observed correlation between chocolate and Nobel Prizes. If we could control for the confounding factor of economic conditions and collect data on the poor economic conditions under which chocolate is rarely consumed, then we could come to the correct conclusion that chocolate consumption is not related to the number of Nobel Prize winners. In 3(e), The fork structural causal path $X \leftarrow C \rightarrow Y$ can be causally intervened through backdoor adjustment to eliminate the bias induced by non-perceptual features in IQA and estimate the causal effect from X to Y. We can get the expression of $P(Y|do(X))$ according to the backdoor criterion[49]:

$$P(Y \mid do(X)) = \sum_c P(Y \mid X, c)P(C) \tag{3}$$

We find that adversarial training can be understood as a backdoor adjustment to a simplified situation:

$$\begin{aligned} P(Y \mid do(X)) &= P(Y \mid X, c_a)P(c_a) \\ &+ P(Y \mid X, c_n)P(c_n) \end{aligned} \tag{4}$$

For a single sample x, we convert the summation term into two samples. Natural samples and corresponding adversarial samples are added to the training to eliminate the dependence on c. In the loss function of adversarial training, the weighted values of adversarial samples and natural samples can be regarded as the prior probability of the above formula. However, adversarial training serves as a passive defense mechanism. In contrast, we aim to develop more advanced causal representations by leveraging network structure and training strategies. The core idea is to utilize causal invariance to boost CPR and incorporate a SortMask module to further eliminate N-CPR.

4.2. Causal Perception Inspired Representation Learning for IQA

In this section, we introduce a strategy designed to learn robust causal perceptual representations (CPR) by leveraging both training strategies and the network structure. Our approach aims to strengthen the model's defense against adversarial perturbations, as illustrated in Figure 4.

4.2.1. Causal Invariant Strategies

Adversarial training can be viewed as a form of causal intervention, but it is usually limited to specific adversarial examples generated to mislead the model, focusing on specific samples. Therefore, we extend traditional adversarial training to a more

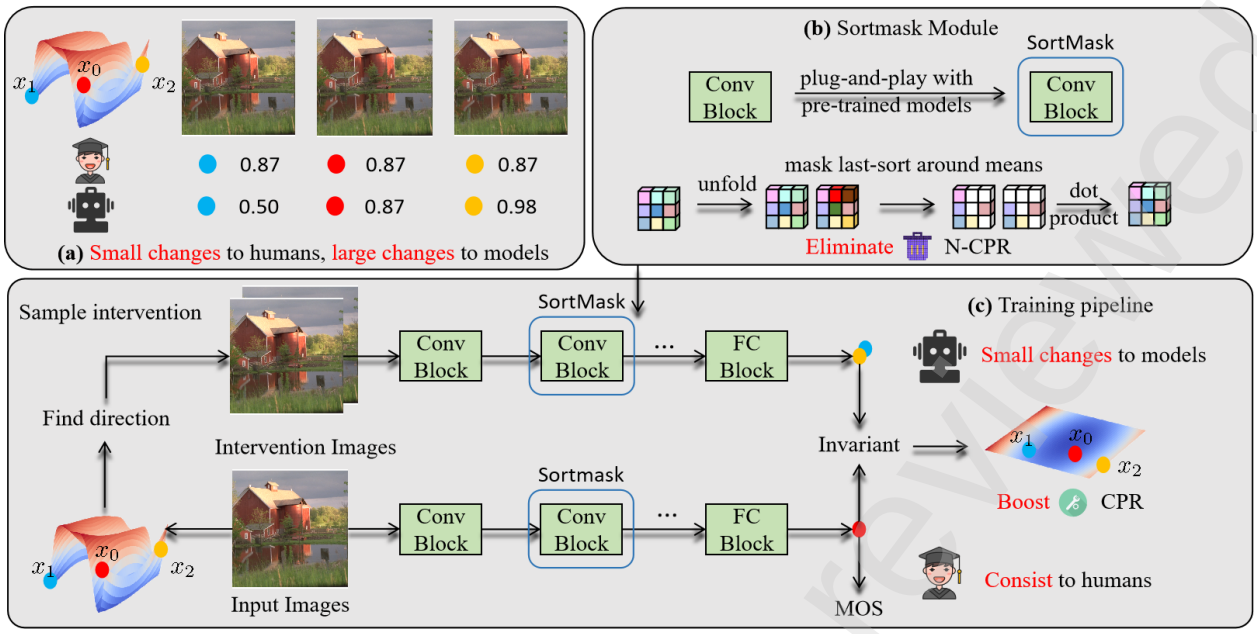

Fig. 4: (a) In the high-dimensional image space, there exist samples that exhibit minor changes imperceptible to humans but cause significant variations to models, reflected in the steep landscape of the predicted scores. (b) SortMask module is integrated into the model to eliminate N-CPR, serving as interventions at the feature level. (c) We sample a series of such intervention samples to perform causal interventions at the data level. Then calculate the invariant loss to enhance CPR and apply the loss on MOS to align the model's outputs with human perception. Through this distillation process, multi-branch parameter updates are achieved.

comprehensive form of causal intervention. This ensures that the model's output remains consistent under perceptually irrelevant perturbations.

Specifically, our proposed intervention involves multiple samples within the ℓ_∞-norm ball. We first define the intervention direction to maximize the output change as follows:

$$\delta_d = \arg\max_{\delta_d \in \mathcal{B}} \|f(x+\delta;\theta) - f(x;\theta)\|,$$

where $f(\cdot)$ represents the model's output function, δ is the perturbation direction, θ denotes the model parameters, and the perturbation $\delta_d \in \mathcal{B}$ belongs to a neighborhood set constrained within an ℓ_∞-norm ball. This maximization objective seeks the direction that produces the greatest difference between the original and perturbed outputs. Consequently, the model aims to minimize this difference across the norm-ball space, focusing on sampling the direction of maximum difference, thereby enforcing the causal invariance constraint:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{\delta\sim\mathcal{P}(\delta)} \|f(x+\delta;\theta) - f(x;\theta)\|,$$

The set $\mathcal{P}(\delta)$ represents a collection of sampled perturbations, referred to as δ_d, constrained within an ℓ_∞-norm ball $\mathcal{B}$ of radius r. Specifically, we use a simple FGSM attack to define $\mathcal{P}$.

$$\mathcal{P}(\delta) = \{\delta \mid \delta = \alpha \mathbf{e}_d,\ \alpha \in [-\epsilon, \epsilon],\ \|\delta\|_\infty \le r\},$$

where $\mathbf{e}_d$ is a constant vector indicating the direction of δ_d, and α represents the perturbation magnitude, sampled within the range $[-\epsilon, \epsilon]$. This ensures that the sampled perturbations remain within the allowable limit defined by the ℓ_∞-norm.

While our proposed causal intervention strategy differs from adversarial training and Norm regularization Training(NT) methods in motivation and implementation, their formulations are similar enough to be expressed in a unified form. Starting from the form of adversarial training, NT can be seen as a variant of adversarial training. CPRL and adversarial training differ in their loss definitions, where the causal invariance loss of CPRL does not depend on the true label. Based on this, the perturbation δ in CPRL is treated as a set, while the perturbations in NT and adversarial training are specific values. NT can thus be viewed as a special case of CPRL, and a detailed comparison of these methods is shown in Table 1. This broader intervention ensures that the model generalizes its causal representation to all samples within the surrounding space, not just adversarial examples. Additionally, our objective updates both branches simultaneously, leading to a more robust understanding of the underlying function f itself.

4.2.2. *SortMask Module*

Causal invariance alone is insufficient, so structural changes to the network are required to enhance the expressive power of Lipschitz networks [13]. To eliminate N-CPR, we introduce a SortMask module into the convolution operation to control sensitivity by leveraging the Lipschitz constant. Specifically, this masking strategy suppresses small changes around the feature mean to alleviate the effects of N-CPR [50]. This structural adjustment limits the model's sensitivity to input perturbations and enhances causal invariance.

Let $x \in \mathbb{R}^{B\times C_i\times H\times W}$ denote the input tensor, where B is the batch size, C_i is the number of input channels, and $H \times W$ represents the spatial dimensions. The convolution operation uses a kernel $\mathbf{W} \in \mathbb{R}^{C_o\times C_i\times k_h\times k_w}$, where C_o is the number of output channels, and $k_h \times k_w$ represents the kernel size. To facilitate computation, the input tensor x is reshaped into an un-

Table 1: Comparison of Formulation.

Method	Net changes	Formulation	δ	ℓ
Adversarial Training	×	$\mathbb{E}_{(x,y)\sim\mathcal{D}}\|\|f_\theta(x+\delta),y\|\|$	$\epsilon\cdot\text{sign}(\nabla_x\ell(f_\theta(x),y))$	$f_\theta(x)\text{sign}(y-y_{\text{mid}})$
Norm regularization Training	×	$\mathbb{E}_{(x,y)\sim\mathcal{D}}\|\|f_\theta(x+\delta),f_\theta(x)\|\|$	$\alpha\cdot\text{sign}(\nabla_x\ell(f_\theta(x)))$	$f_\theta(x)$
CPRL	✓	$\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathbb{E}_{\delta\sim\mathcal{P}(\delta)}\|\|f_\theta(x+\delta),f_\theta(x)\|\|$	$\{\delta \mid \delta=\alpha\cdot\text{sign}(\nabla_x\ell(f_\theta(x))),\ \alpha\in[-\epsilon,\epsilon]\}$	$f_\theta(x)$

folded matrix $x_u \in \mathbb{R}^{B\times(C_i\cdot k_h\cdot k_w)\times P}$, where P is the total number of patches determined by the kernel's receptive field. Consequently, the convolution operation can be expressed as a matrix multiplication $x_o = \mathbf{W}\cdot x_u$, where $\mathbf{W}\in\mathbb{R}^{C_o\times(C_i\cdot k_h\cdot k_w)}$ is the flattened kernel matrix, and x_o is the output.

In our structural masking approach, sensitivity is controlled by regulating the Lipschitz constant L, which quantifies the impact of input perturbations on the output. By constraining L, we strengthen causal invariance, ensuring the output remains stable under small input changes. The Lipschitz constant L is defined as:

$$L = \sup_{\delta\neq 0}\frac{\|f(x;\theta)-f(x+\delta;\theta)\|}{\|\delta\|},$$

To simplify the estimation of the upper bound, we use the infinity norm of $\mathbf{W}$, which quantifies the maximum impact of input changes on the output [34]:

$$\|\mathbf{W}\|_\infty = \max_{i=1,\ldots,C_o}\sum_{j=1}^{C_i\cdot k_h\cdot k_w}|W_{i,j}|.$$

Thus, the upper bound of the Lipschitz constant can be expressed as $L\le C_i\cdot k_h\cdot k_w\cdot\alpha$, where α is the maximum absolute value of the kernel elements.

To implement structural sensitivity control, we introduce a mask $M\in\{0,1\}^{B\times(C_i\cdot k_h\cdot k_w)\times P}$, where each element is retained with probability $1-p$ and masked with probability p. This mask reduces the number of active features, thereby lowering the model's dependency on non-essential patterns and improving stability under input perturbations. As a result, the masked convolution operation can be formulated as:

$$x_o' = \mathbf{W}(\mathbf{M}\odot x_u),$$

where $\odot$ denotes element-wise multiplication. Let μ represent the mean of the input x_u. The test-time mask M is computed as follows:

$$M_i = \begin{cases}1 & \text{if } |x_u-\mu| > \text{TopK}(|x_u-\mu|,K),\\ 0 & \text{otherwise.}\end{cases}$$

Here, $K=(1-p)\times(C_i\cdot k_h\cdot k_w)\times P$, and TopK returns the K-th largest value. For the unmasked convolution, we denote the Lipschitz constant as $L_0=\|\mathbf{W}\|_\infty$. After masking, the number of active input features is reduced to approximately $1-p$ times, leading to a proportional reduction in sensitivity and improved stability. Consequently, the resulting sensitivity-controlled Lipschitz constant for the masked convolution satisfies:

$$L_m \le L_0.$$

Through this structural sensitivity control, we achieve causal invariance by limiting sensitivity at the structural level, thereby enhancing the model's robustness against adversarial perturbations.

5. Experiments

Table 2: Summary of IQA datasets.

Databases	# of Dist. Images	# of Dist. Types	Distortions Type
LIVE [51]	799	5	synthetic
VCL [52]	575	4	synthetic
LIVEC [53]	1,162	-	authentic
KONIQ [53]	10,073	-	authentic

5.1. Settings

datasets. We evaluate our method on four IQA datasets with different characteristics: LIVE and VCL, which contain artificially distorted images with a small size; and LIVEC and KONIQ, which contain naturally distorted images with a large size. KONIQ is the largest dataset with 10K images. Table 2 summarizes the details of the datasets.

Comparison methods. We evaluate our proposed approach against 3 classic DNN-based IQA models and 4 adversarial robust models including DBCNN [20], NIMA [54], ResNet [23], HyperNet[59], LWTA [55], Denoise [14], SAT [56] and NT [57].

Attack methods. Two classic white-box attack methods FGSM [29] and PGD[60].[1] Two limited white-box attack Target-Free attack[2] [57] and Perceptual Attack [6] . One black-box attacks: UAP[3] [58]. Following Liu et al. [57], To ensure fairness in our evaluations, each attack method uses the same setting (i.e., employing the same hyperparameters in attack) when targeting different models. We set different hyperparameters for different attacks and ensure the majority of attacked images' SSIM was above 0.9 to satisfy the assumption that the MOS is the same for both images before and after the attack. Black-box attacks are indicated with an asterisk and restricted white-box attacks are indicated with Dagger in the tables of this paper.

[1] FGSM and PGD was originally designed for classification tasks, but we modify its loss to for NR-IQA tasks.

[2] Target-Free attack was designed by liu et al. based on FGSM that do not rely on ground-truth target [57].

[3] UAP is proposed as a white-box attack. We employ its perturbation generated on PaQ-2-PiQ model [61], and serve UAP as a black-box attack.

Table 3: Evaluation results (*SRCC, PLCC, RMSE*) of the existing IQA methods under different attack (*NO, FGSM, PGD*). The best performance is marked in **bold** and the second performance is underlined. For each dimension (column) of the result, we used red font to indicate the change of attacked.

Models	Attacks	VCL			LIVE			LIVEC			KONIQ		
		SROCC↑	*PLCC*↑	*RMSE*↓	*SROCC*↑	*PLCC*↑	*RMSE*↓	*SROCC*↑	*PLCC*↑	*RMSE*↓	*SROCC*↑	*PLCC*↑	*RMSE*↓
DBCNN [20]	NO	0.897	0.870	16.643	**0.950**	0.927	**10.100**	0.716	0.753	13.820	0.851	0.868	7.685
	FGSM	0.190	0.168	27.386	0.322	0.350	25.652	-0.321	-0.350	33.287	-0.138	-0.175	20.494
	PGD	0.108	0.082	26.907	0.388	0.336	24.125	-0.363	-0.421	34.467	0.063	0.032	21.142
NIMA [54]	NO	0.938	0.889	22.226	0.940	**0.936**	10.247	0.787	0.829	11.489	**0.901**	**0.922**	**5.385**
	FGSM	-0.387	-0.384	36.810	0.228	0.205	27.203	-0.242	-0.266	31.718	0.081	0.072	17.029
	PGD	-0.403	-0.400	36.194	0.165	0.209	26.230	-0.384	-0.413	33.971	-0.037	-0.058	18.439
ResNet [23]	NO	0.923	0.907	10.724	0.881	0.821	17.263	0.834	**0.863**	**11.000**	0.890	0.912	5.657
	FGSM	-0.089	-0.104	28.125	-0.034	0.070	31.385	-0.180	-0.210	28.443	-0.083	-0.111	19.287
	PGD	-0.087	-0.123	28.302	0.235	0.337	27.276	-0.252	-0.265	28.601	-0.124	-0.151	19.261
LWTA [55]	NO	0.286	0.291	48.436	0.526	0.142	42.285	0.234	0.252	26.230	0.638	0.629	19.365
	FGSM	0.300	0.339	40.731	0.584	0.255	41.701	0.244	0.243	26.851	0.606	0.582	18.111
	PGD	0.354	0.339	47.455	0.589	0.117	42.190	0.257	0.245	25.671	0.629	0.621	18.601
Denoise [14]	NO	0.932	0.909	12.570	0.890	0.824	17.378	0.819	0.848	11.446	0.890	0.915	5.657
	FGSM	-0.329	-0.309	32.894	-0.145	-0.108	33.437	-0.268	-0.288	28.740	0.160	0.152	15.843
	PGD	-0.249	-0.234	29.799	0.040	0.146	28.862	-0.256	-0.262	27.276	0.068	0.058	16.279
SAT [56]	NO	0.802	0.736	17.748	0.841	0.815	14.422	0.701	0.754	13.892	0.832	0.868	7.141
	FGSM	0.377	0.341	25.671	0.510	0.568	22.978	-0.067	-0.056	27.129	-0.318	-0.328	24.104
	PGD	0.244	0.206	28.000	0.496	0.553	23.537	-0.210	-0.198	29.206	-0.151	-0.158	19.209
NT [57]	NO	0.922	0.909	10.354	0.900	0.875	14.634	0.831	0.847	12.391	0.838	0.853	8.078
	FGSM	0.778	0.750	16.013	0.733	0.720	18.897	0.625	0.644	16.488	0.755	0.756	10.202
	PGD	0.725	0.691	17.471	0.652	0.638	19.919	0.580	0.600	17.127	0.735	0.731	10.650
CPRL	NO	**0.943**	**0.926**	**9.108**	0.937	0.927	10.311	**0.851**	0.862	12.741	0.879	0.893	7.638
	FGSM	**0.814**$_{-0.129}$	**0.783**$_{-0.143}$	**15.125**$_{6.017}$	**0.842**$_{-0.095}$	**0.838**$_{-0.089}$	**15.958**$_{5.647}$	**0.700**$_{-0.151}$	**0.697**$_{-0.165}$	**16.205**$_{3.465}$	**0.772**$_{-0.107}$	**0.765**$_{-0.127}$	**10.049**$_{2.411}$
	PGD	**0.783**$_{-0.159}$	**0.754**$_{-0.173}$	**16.050**$_{6.942}$	**0.810**$_{-0.127}$	**0.808**$_{-0.119}$	**16.668**$_{6.357}$	**0.685**$_{-0.166}$	**0.681**$_{-0.181}$	**16.450**$_{3.709}$	**0.763**$_{-0.116}$	**0.754**$_{-0.138}$	**10.253**$_{2.615}$

Table 4: Evaluation results on the different models. The best performance is marked in **bold** and the second performance is underlined.

Model	MOS			Prediction		
	SRCC↑	*PLCC*↑	*RMSE*↓	*SRCC*↑	*PLCC*↑	*RMSE*↓
Target-Free† [57]						
HyperNet	0.237	0.207	20.924	0.250	0.248	15.489
NT	0.824	0.824	13.966	0.923	0.928	6.670
CPRL	**0.839**	**0.835**	**13.398**	**0.963**	**0.965**	**5.191**
Perceptual† [6]						
HyperNet	-0.023	0.023	38.100	-0.070	-0.007	35.516
NT	0.855	0.868	10.976	0.995	0.994	1.818
CPRL	**0.858**	**0.869**	**10.957**	**1.000**	**0.999**	**0.472**
UAP∗ [58]						
HyperNet	0.711	0.736	17.142	0.798	0.832	11.336
NT	0.821	0.817	14.313	0.907	0.914	9.859
CPRL	**0.833**	**0.839**	**12.457**	**0.955**	**0.967**	**6.136**
FGSM						
HyperNet	0.038	0.005	23.591	0.237	0.256	16.424
NT	0.660	0.681	15.767	0.898	0.909	6.580
CPRL	**0.720**	**0.734**	**14.749**	**0.941**	**0.950**	**5.131**

Metrics. We use three widely used metrics to evaluate the performances of the models, i.e., the Pearson's Linear Correlation Coefficient (PLCC) [62], the Spearman's Rank Order Correlation Coefficient (SROCC) [63], the Root Mean Square Error (RMSE).

Implementation. We split the data into training and testing sets at a 4:1 ratio randomly and saved the split-sequence to ensure the same division for all experiments. We also ensure that no source scene appears in both training and testing sets to avoid artificial inflation of the results. In order to operate efficiency, SortMask is used in the middle layer, which is generally more than 128 channels. p is generally set to 0.1. We crop the images to 224×224 randomly. We random crop the same images 25 times in the test. We set the batch size to 16 and use Adam with a learning rate of $2e-5$ and $2e-4$ for optimization. CPRL's backbone is HyperNet[59].

5.2. Evaluations

White-box attack. The existing IQA models and adversarial defense models have poor robustness, as shown by the results in Table 3. Their performance drops significantly with small perturbations to the images under FGSM or PGD attacks. The models' SRCC and PLCC values even become negative, indicating that they provide inconsistent scores for slightly perturbed images. Our CPRL method enhances the robustness of the model and enables it to output more trustworthy scores for perturbed images. Additionally, compared with other methods, our prediction performance on natural samples is also excellent. It can be observed that our method surpasses the NT [57] model. This is because our causal invariance loss is a more generalized representation of the NT loss; that is, the NT loss can be regarded as a special case of the causal invariance loss. Furthermore, our sortmask module provides network structural support for the causal invariance loss to achieve better robustness.

Restricted attacks. In order to further explore the performance differences between our model and the previous SOTA method, we conduct comparative experiments under restricted attacks. In Table 4, columns 2-4 show MOS values of unattacked images and prediction scores on adversarial examples. Columns

Table 5: Performance comparison under different mask rate p on the VCL dataset. The best performance is marked in **bold** and the second performance is underlined.

p	Attack	*SRCC*↑	*PLCC*↑	*RMSE*↓
0.02	NO	**0.943**	**0.926**	**9.108**
	FGSM	0.814	0.783	15.125
	PGD	0.783	0.754	16.050
0.1	NO	0.933	0.912	9.922
	FGSM	**0.841**	**0.801**	**14.632**
	PGD	**0.820**	**0.779**	**15.323**
0.2	NO	0.912	0.892	11.317
	FGSM	0.828	0.780	15.458
	PGD	0.800	0.753	16.180
0.4	NO	0.768	0.673	20.985
	FGSM	0.750	0.650	21.109
	PGD	0.748	0.652	21.099
0.6	NO	0.632	0.641	24.349
	FGSM	0.579	0.590	24.350
	PGD	0.545	0.556	24.373
0.8	NO	0.134	0.102	24.790
	FGSM	-0.063	-0.019	24.802
	PGD	-0.081	-0.073	24.806

5-7 showcase the metrics calculated between prediction scores on unattacked images and scores on adversarial examples. As shown in Table 4, when calculating the RMSE between the prediction scores before and after the attack, the HyperNet model trained by the CPRL method exhibits smaller score changes under nearly all attack scenarios compared with the baseline model HyperNet and the SOTA method NT. When considering the RMSE result between the MOS value and the prediction score, the robustness of CPRL has also been improved compared with the baseline model HyperNet and the SOTA method NT. For restricted attacks—including the Target-Free attack without real labels, perceptual attack, and black-box attack UAP—the attack strength is weaker compared with unrestricted white-box attacks. For example, the metrics for Target-Free, perceptual attack, and UAP on the SOTA model NT can reach above 0.8, but only above 0.6 on FGSM. Compared with the previous SOTA method NT, our method not only guarantees performance under restricted attacks but also significantly improves performance in more difficult white-box attacks.

5.3. Ablation Studies

In this section, we conduct experiments to study the efect on our approach of 1) Whether to use causal invariant or SortMask loss 2) mask rate

Impact of causal invariant and SortMask . To examine the effect of causal invariance loss and SortMask module, we conduct the ablation study. As shown in table Table 6, causal invariance loss achieves superior results than adversarial training, and SortMask module boost it futher. Both accuracy and robustness are enhanced.

Impact of mask rate. In order to study the impact of the mask ratio p on the performance of the model, we conducted an ablation study varying the rate p. The experimental results are listed in Table 5. We conducted experiments on the VCL dataset using a model based on the HyperNet network with CPRL, evaluating its robustness under different parameter settings. The experimental results show that a smaller parameter p leads to better performance on clean samples but results in lower robustness. Conversely, increasing the parameter p initially improves robustness, but if p becomes too large, it adversely affects the network expression, leading to reduced performance.

Table 6: Performance comparison on the LIVEC dataset. The best performance is marked in **bold** and the second performance is underlined.

p	Attack	*SRCC*↑	*PLCC*↑	*RMSE*↓
Adv	NO	0.681	0.696	16.705
	FGSM	0.601	0.612	17.670
	PGD	0.575	0.579	18.050
intervention	NO	0.818	0.817	14.285
	FGSM	0.595	0.615	17.512
	PGD	0.573	0.597	17.726
intervention+SortMask	NO	**0.851**	**0.862**	**12.741**
	FGSM	**0.700**	**0.697**	**16.205**
	PGD	**0.685**	**0.681**	**16.450**

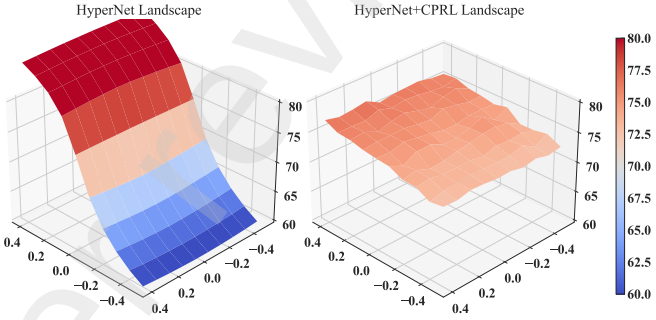

Fig. 5: Output landscape on two-dimensional hyper-plane based on HyperNet. specifically, one direction is the FGSM direction with a length of 1.0 pixels. Another direction is a random choice. Left is the HyperNet. Right is the HyperNet with CPRL. It can be found that the landscape is flat than original one. Regardless of the randomly selected perturbation direction or the perturbation direction of the adversarial attack, our landscape is flat, which empirically proves that CPRL is more robust.

5.4. Case Study

landscape. Fig. 5 depicts the score landscape on a 2D hyperplane based on HyperNet and HyperNet with CPRL. For vanilla HyperNet, the landscape exhibits more variations, while CPRL shows more stability, and experiments confirm that CPRL is more robust.

6. Conclusion

In this paper, we propose to build a trustworthy IQA model through causal-aware inspired representation learning (CPRL), which only requires inserting the proposed SortMask module in the convolutional layer and optimizing it with causal invariance loss. Extensive experiments on popular quality assessment datasets verify that our method can obtain trustworthy IQA models.

References

[1] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing 13 (2004) 600–612.

[2] X. Min, G. Zhai, K. Gu, Y. Zhu, J. Zhou, G. Guo, X. Yang, X. Guan, W. Zhang, Quality evaluation of image dehazing methods using synthetic hazy images, IEEE Transactions on Multimedia 21 (2019) 2319–2333.

[3] G. Zhai, X. Min, Perceptual image quality assessment: a survey, Science China Information Sciences 63 (2020) 1–52.

[4] H. Lin, V. Hosu, D. Saupe, Kadid-10k: A large-scale artificially distorted iqa database, in: 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), IEEE, 2019, pp. 1–3.

[5] J. Ma, J. Wu, L. Li, W. Dong, X. Xie, G. Shi, W. Lin, Blind image quality assessment with active inference, IEEE Transactions on Image Processing 30 (2021) 3650–3663.

[6] W. Zhang, D. Li, X. Min, G. Zhai, G. Guo, X. Yang, K. Ma, Perceptual attacks of no-reference image quality models with human-in-the-loop, Advances in Neural Information Processing Systems 35 (2022) 2916–2929.

[7] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, A. Madry, Adversarial examples are not bugs, they are features, Advances in neural information processing systems 32 (2019).

[8] H. Liu, M. Chaudhary, H. Wang, Towards trustworthy and aligned machine learning: A data-centric survey with causality perspectives, arXiv preprint arXiv:2307.16851 (2023).

[9] J. Kaddour, A. Lynch, Q. Liu, M.J. Kusner, R. Silva, Causal machine learning: A survey and open problems, arXiv preprint arXiv:2206.15475 (2022).

[10] A. Liu, W. Lin, M. Paul, C. Deng, F. Zhang, Just noticeable difference for images with decomposition model for separating edge and textured regions, IEEE Transactions on Circuits and Systems for Video Technology 20 (2010) 1648–1652.

[11] J.W. Kalat, Biological psychology, Cengage Learning, 2015.

[12] Z. Allen-Zhu, Y. Li, Feature purification: How adversarial training performs robust deep learning, in: 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2022, pp. 977–988.

[13] B. Zhang, D. Jiang, D. He, L. Wang, Rethinking lipschitz neural networks and certified robustness: A boolean function perspective, Advances in Neural Information Processing Systems 35 (2022) 19398–19413.

[14] C. Xie, Y. Wu, L.v.d. Maaten, A.L. Yuille, K. He, Feature denoising for improving adversarial robustness, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 501–509.

[15] M. Yang, Z. Fang, Y. Zhang, Y. Du, F. Liu, J.F. Ton, J. Wang, Invariant learning via probability of sufficient and necessary causes, arXiv preprint arXiv:2309.12559 (2023).

[16] Z. Wang, A.C. Bovik, Reduced-and no-reference image quality assessment, IEEE Signal Processing Magazine 28 (2011) 29–40.

[17] Y. Fang, K. Ma, Z. Wang, W. Lin, Z. Fang, G. Zhai, No-reference quality assessment of contrast-distorted images based on natural scene statistics, IEEE Signal Processing Letters 22 (2014) 838–842.

[18] A. Mittal, R. Soundararajan, A.C. Bovik, Making a "completely blind" image quality analyzer, IEEE Signal processing letters 20 (2012) 209–212.

[19] K. Ma, W. Liu, T. Liu, Z. Wang, D. Tao, dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs, IEEE Transactions on Image Processing 26 (2017) 3951–3964.

[20] W. Zhang, K. Ma, J. Yan, D. Deng, Z. Wang, Blind image quality assessment using a deep bilinear convolutional neural network, IEEE Transactions on Circuits and Systems for Video Technology 30 (2018) 36–47.

[21] S. Bosse, D. Maniry, K.R. Müller, T. Wiegand, W. Samek, Deep neural networks for no-reference and full-reference image quality assessment, IEEE Transactions on image processing 27 (2017) 206–219.

[22] X. Min, G. Zhai, K. Gu, Y. Liu, X. Yang, Blind image quality estimation via distortion aggravation, IEEE Transactions on Broadcasting 64 (2018) 508–517.

[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[24] L. Wang, Q. Wu, K.N. Ngan, H. Li, F. Meng, L. Xu, Blind tone-mapped image quality assessment and enhancement via disentangled representation learning, in: 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2020, pp. 1096–1102.

[25] J. You, J. Korhonen, Transformer for image quality assessment, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 1389–1393.

[26] D. Li, T. Jiang, W. Lin, M. Jiang, Which has better visual quality: The clear blue sky or a blurry animal?, IEEE Transactions on Multimedia 21 (2018) 1221–1234.

[27] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, 2014. doi:10.48550/arXiv.1312.6199. arXiv:1312.6199.

[28] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, 2015. arXiv:1412.6572.

[29] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, arXiv:1412.6572 [cs, stat] (2015).

[30] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, A. Madry, Robustness may be at odds with accuracy, arXiv preprint arXiv:1805.12152 (2018).

[31] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, B. Tran, A. Madry, Adversarial robustness as a prior for learned representations, arXiv preprint arXiv:1906.00945 (2019).

[32] B. Shi, D. Zhang, Q. Dai, Z. Zhu, Y. Mu, J. Wang, Informative dropout for robust representation learning: A shape-bias perspective, in: International Conference on Machine Learning, PMLR, 2020, pp. 8828–8839.

[33] W. Xu, D. Evans, Y. Qi, Feature squeezing: Detecting adversarial examples in deep neural networks, arXiv preprint arXiv:1704.01155 (2017).

[34] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, N. Usunier, Parseval networks: Improving robustness to adversarial examples, in: International Conference on Machine Learning, PMLR, 2017, pp. 854–863.

[35] C. Qin, J. Martens, S. Gowal, D. Krishnan, K. Dvijotham, A. Fawzi, S. De, R. Stanforth, P. Kohli, Adversarial robustness through local linearization, Advances in Neural Information Processing Systems 32 (2019).

[36] H. Kannan, A. Kurakin, I. Goodfellow, Adversarial logit pairing, arXiv preprint arXiv:1803.06373 (2018).

[37] E.A. Stuart, Matching methods for causal inference: A review and a look forward, Statistical science: a review journal of the Institute of Mathematical Statistics 25 (2010) 1.

[38] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, Z. Shen, Deep stable learning for out-of-distribution generalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5372–5382.

[39] K. Kuang, P. Cui, H. Zou, B. Li, J. Tao, F. Wu, S. Yang, Data-driven variable decomposition for treatment effect estimation, IEEE Transactions on Knowledge and Data Engineering 34 (2020) 2120–2134.

[40] Z. Shen, P. Cui, K. Kuang, B. Li, P. Chen, Causally regularized learning with agnostic data selection bias, in: Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 411–419.

[41] Y. Wang, M.I. Jordan, Desiderata for representation learning: A causal perspective, arXiv preprint arXiv:2109.03795 (2021).

[42] W. Zhang, T. Wu, Y. Wang, Y. Cai, H. Cai, Towards trustworthy explanation: On causal rationalization, arXiv preprint arXiv:2306.14115 (2023).

[43] J. Kim, B.K. Lee, Y.M. Ro, Demystifying causal features on adversarial examples and causal inoculation for robust network by adversarial instrumental variable regression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12302–12312.

[44] K. Tang, M. Tao, H. Zhang, Adversarial visual robustness by causal intervention, arXiv preprint arXiv:2106.09534 (2021).

[45] Y. Zhang, M. Gong, T. Liu, G. Niu, X. Tian, B. Han, B. Schölkopf, K. Zhang, Causaladv: Adversarial robustness through the lens of causality, arXiv preprint arXiv:2106.06196 (2021).

[46] F. Lv, J. Liang, S. Li, B. Zang, C.H. Liu, Z. Wang, D. Liu, Causality inspired representation learning for domain generalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8046–8056.

[47] H. Cai, Y. Wang, M. Jordan, R. Song, On learning necessary and sufficient causal graphs, arXiv preprint arXiv:2301.12389 (2023).

[48] M. Yang, X. Cai, F. Liu, W. Zhang, J. Wang, Specify robust causal representation from mixed observations, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023, pp. 2978–2987.

[49] J. Pearl, Causality, Cambridge university press, 2009.

[50] S. Amini, M. Teymoorianfard, S. Ma, A. Houmansadr, Meansparse: Post-

training robustness enhancement through mean-centered feature sparsification, arXiv preprint arXiv:2406.05927 (2024).
[51] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE TIP 15 (2006) 3440–3451.
[52] A. Zaric, N. Tatalovic, N. Brajkovic, H. Hlevnjak, M. Loncaric, E. Dumic, S. Grgic, Vcl@ fer image quality assessment database, in: Proceedings ELMAR-2011, IEEE, 2011, pp. 105–110.
[53] D. Ghadiyaram, A.C. Bovik, Massive online crowdsourced study of subjective and objective picture quality, IEEE Transactions on Image Processing 25 (2015).
[54] H. Talebi, P. Milanfar, Nima: Neural image assessment, IEEE transactions on image processing 27 (2018) 3998–4011.
[55] K.P. Panousis, S. Chatzis, S. Theodoridis, Stochastic local winner-takes-all networks enable profound adversarial robustness, arXiv preprint arXiv:2112.02671 (2021).
[56] C. Xie, M. Tan, B. Gong, A. Yuille, Q.V. Le, Smooth adversarial training, arXiv preprint arXiv:2006.14536 (2020).
[57] Y. Liu, C. Yang, D. Li, J. Ding, T. Jiang, Defense against adversarial attacks on no-reference image quality models with gradient norm regularization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 25554–25563.
[58] E. Shumitskaya, A. Antsiferova, D. Vatolin, Universal perturbation attack on differentiable no-reference image-and video-quality metrics, arXiv preprint arXiv:2211.00366 (2022).
[59] 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3667–3676.
[60] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv:1706.06083 (2017).
[61] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, A.C. Bovik, From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality, in: CVPR, 2020, pp. 3575–3585.
[62] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson correlation coefficient, in: Noise reduction in speech processing, Springer, 2009, pp. 1–4.
[63] J.H. Zar, Spearman rank correlation, Encyclopedia of biostatistics 7 (2005).