Causal-IQA: Towards the Generalization of Image Quality Assessment Based on Causal Inference

Yan Zhong [1 2] **Xingyu Wu** [3] **Li Zhang** [4] **Chenxi Yang** [1 2] **Tingting Jiang** [2 5]

Abstract

Due to the high cost of Image Quality Assessment (IQA) datasets, achieving robust generalization remains challenging for prevalent deep learning-based IQA methods. To address this, this paper proposes a novel end-to-end blind IQA method: Causal-IQA. Specifically, we first analyze the causal mechanisms in IQA tasks and construct a causal graph to understand the interplay and confounding effects between distortion types, image contents, and subjective human ratings. Then, through shifting the focus from correlations to causality, Causal-IQA aims to improve the estimation accuracy of image quality scores by mitigating the confounding effects using a causality-based optimization strategy. This optimization strategy is implemented on the sample subsets constructed by a Counterfactual Division process based on the Backdoor Criterion. Extensive experiments illustrate the superiority of Causal-IQA.

1. Introduction

Image quality assessment (IQA) plays a crucial role in optimizing visual experiences across different domains, including image denoising (Tian et al., 2020), restoration (Cui et al., 2023), and generation (Elasri et al., 2022). Its primary objective is to develop algorithms that accurately quantify the perceptual quality of images. Based on the availability of reference information, existing IQA methods can be classified into Full-Reference IQA (FR-IQA), Reduced-Reference IQA (RR-IQA), and Blind IQA (BIQA) (Zhai & Min, 2020), where BIQA exhibits broader applicability due to its independence from reference image (Su et al., 2020; Simeng Sun, 2023; Feng et al., 2021).

Traditional BIQA methods aim to predict image quality scores that align with subjective human ratings, such as Mean Opinion Scores (MOS), by relying on manually extracted statistical features from the provided distorted images (Jiang et al., 2017; Liu et al., 2020; Zhou et al., 2017). Although these methods are flexible and easy to implement, they cannot achieve ideal performance when confronted with complex scenes or diverse distortions. Consequently, Deep Learning (DL)-based IQA methods have received widespread attention due to their strong ability in fusing discriminative features in visual domains. (Ke et al., 2021; Talebi & Milanfar, 2018; Madhusudana et al., 2022; Simeng Sun, 2023). However, the high cost of annotating IQA datasets restricts DL-based BIQA methods to be trained only on small-scale datasets, leading to overfitting and limited generalization on unseen distorted information and images with authentical distortions (Yue et al., 2022).

To address this issue, one intuitive strategy is to initialize BIQA models with the backbones pre-trained on large-scale databases in the field of image classification, such as ImageNet (Deng et al., 2009). Although this enables DL-based models to learn general image features, its effectiveness in IQA tasks is limited due to the lack of specific representation learning for distortions. Another strategy is to design unsupervised IQA method for test time adaptation (Roy et al., 2023), which can effectively mitigate distribution shifts between train and test data, thus improving the performance. However, this strategy has limited applicability as it requires additional unsupervised training to improve the prediction accuracy on different test datasets. Furthermore, these studies often overlook the investigation of the underlying factors that lead to the limited generalization capability in IQA models. Therefore, we focus on exploring an end-to-end BIQA method with good interpretability and generalization capability.

In the standard training paradigm for an IQA network f_θ, the goal is to mine the correlation between distorted image

[1]School of Mathematical Sciences, Peking University, Beijing, China [2]National Engineering Research Center of Visual Technology, National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University, Beijing, China [3]Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China [4]Hefei Institute of Physical Science, Chinese Academy of Sciences, University of Science and Technology of China, Hefei, China [5]National Biomedical Imaging Center, Peking University, Beijing, China. Correspondence to: Tingting Jiang <ttjiang@pku.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

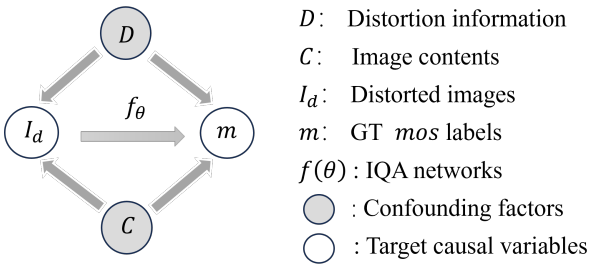


Figure 1. The structural causal graph in IQA tasks, where the arrows denotes the causations between parent nodes to child nodes. We only consider distortion type in D in our method.

I_d and its corresponding MOS label m, which is achieved by maximizing the conditional probability $P(m|I_d)$ through empirical risk minimization (ERM). However, it's important to note that $P(m|I_d)$ is influenced by both the distortion types and the image contents (Li et al., 2018), leading to the detrimental confounding effects and poor generalization in the standard training paradigm (Li et al., 2023). To address this issue, we analyze the causal mechanisms in IQA tasks and construct a causal graph (Figure 1) that sheds light on the underlying causal mechanisms in IQA tasks. This Directed Acyclic Graph (DAG) provides valuable insights into the interplay and potential confounding effects between these variables, allowing us to better understand the challenges associated with the standard training paradigm and the limitations of correlation-based IQA models. According to Figure 1, distortion types and image contents are confounders of cause-effect pair I_d and m. The learning objective of the aforementioned standard training paradigm focuses solely on the correlations between I_d and m within the specific IQA datasets used for training, which is the main reason for their poor generalization since correlation does not imply causation (Pearl et al., 2016).

Based on the above analyses, a generalizable IQA model should be robust to different distortion information and image contents. To achieve this goal, we propose a Causal Learning-based Image Quality Assessment model (Causal-IQA). By shifting the focus from correlations to causality between variables I_d and m, Causal-IQA aims to improve the estimation accuracy of the conditional probability $P(m|I_d)$ through the learning of causal relationships. Specifically, we first construct sample subsets based on confounder sets (including distortion type and image content) from the perspective of causality (Pearl, 2009). To distinguish the content C for each image, the extracted semantic features of these images are partitioned through Gaussian mixture clustering (Do & Batzoglou, 2008). Subsequently, sample subsets[1] with invariant D and C are constructed to achieve distortion-invariant and content-invariant representation learning, which can eliminate the confounding effects (caused by C and D) on the causal relationship between I_d and m to improve the robustness of BIQA model. Ultimately, inspired by Meta-Learning (Finn et al., 2017; Nichol et al., 2018), the causal learning process is instantiated according to a causality-based optimization strategy based on the Backdoor Criterion (Pearl et al., 2016). The contributions of this paper are summarized as follows:

- We provide a comprehensive theoretical analysis from a causal perspective, shedding light on the underlying reasons for the limited generalization of IQA models. We identify that the training process commonly struggles to effectively eliminate the confounding effects caused by distortion and content information, which sets the stage for the introduction of our novel method.
- The proposed Causal-IQA model, as a breakthrough in BIQA training, possesses at least three advantages: (1) Enhanced generalization capacity due to the effective elimination of the confounding effects caused by distorted images and image content; (2) Interpretability for providing valuable insights into the IQA process; (3) Adaptability that can be seamlessly integrated into any BIQA network.
- We conduct extensive experiments on both authentically and synthetically distorted image databases to validate the effectiveness and generalization capabilities of the proposed method. These experiments provide empirical evidence of the superiority of Causal-IQA over existing approaches.

2. Background

2.1. Blind Image Quality Assessment

Blind Image Quality Assessment (BIQA) has gained significant attention recently due to the absence of reference images in realistically distorted image datasets (Zhai & Min, 2020). With the development and wide applications of Deep Learning (DL), there spring up various DL-based methods achieving remarkable progress in BIQA (Ke et al., 2021; Golestaneh et al., 2022). For example, RankIQA (Liu et al., 2017) proposed a Siamese Network to rank image pairs that are synthetically distorted. CONTRIQUE (Madhusudana et al., 2022) considered the problem of obtaining image quality representations in a self-supervised manner, which is trained with contrastive learning. And GraphIQA (Simeng Sun, 2023) integrated graph representation learning into IQA to learn the distortion graph representations. Although these studies can improve the ability of quality perception on training datasets, they lack enough

[1]This process conforms to the calculating of Counterfactuals from the perspective of Causal Learning. See appendix for details.

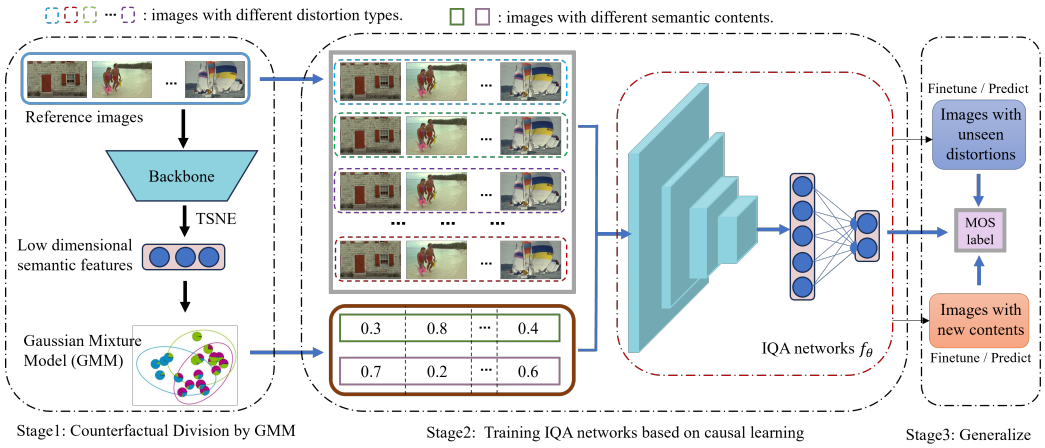


Figure 2. Overview of Causal-IQA. In **Stage 1** (Counterfactual Division), GMM and confounding factors in Figure 1 are used to construct image subsets, where images are distortion-invariant and content-invariant in the same subset. In **Stage 2**, the IQA network f_θ is trained on these sample subsets with a causality-based representation learning method, to improve the generalization of f_θ. In **Stage 3**, the priori IQA model be can used to assess the quality score for new images with unseen information about distortions and contents.

generalization since the scale of the training IQA datasets are limited. In this regard, MetaIQA (Zhu et al., 2020) introduces meta-learning with synthetic distortions, employing a shared quality prior knowledge model for versatile adaptation to diverse distortion types. In order to address different distortions and content variations in images, Su et al. (Su et al., 2020) devise a self-adaptive hypernetwork to assign weights to parameters in the quality prediction module. More recently, one of the latest methods is to integrate domain adaptive and ensemble learning into the IOA task, such as (Roy et al., 2023). Unfortunately, the improvements in generalization in these existing methods come at the expense of training costs. On the contrary, we improve the generalization and robustness of BIQA model from a causality view, which is end-to-end and interpretable.

2.2. Causal Learning

Traditional machine learning approaches often consider the statistical dependencies between inputs and outputs, which can hinder the model's ability to generalize to unknown data (Yao et al., 2021). To address this limitation and identify the true causal impact of inputs on outputs, causal inference has been introduced as a means to discover and leverage causal relationship information embedded in the data. By employing causal graph models and intervention operations, causal inference aims to eliminate confounding effects on the outputs, ultimately yielding causal effects. This aspect is crucial for enhancing both the interpretability and generalization capabilities of models. In recent years, causal inference methods based on Structural Causal Models (SCM) (Pearl, 2010; Pawlowski et al., 2020) have witnessed significant advancements and garnered widespread adoption among computer science researchers due to their ability to offer clearer descriptions of causal relationships between variables (Gresele et al., 2022). Leveraging the structural information between variables, SCM enables causal analysis even in scenarios where interventions on variables are not fully observed, thus presenting a distinct advantage for tackling complex tasks. In the domain of computer vision, various tasks are susceptible to the influence of latent confounding factors. Causal inference techniques provide a valuable framework for identifying and leveraging causal relationships to mitigate the impact of confounding factors, such as scene graph generation (Chen et al., 2019), image recognition (Tang et al., 2020), video analysis (Kanehira et al., 2019), etc. In this paper, we investigate the performance gains afforded by causal relationships within the context of IQA problems.

3. Method

In this section, we initially present the overview of Causal-IQA. Then, we analyze the causal mechanism in BIQA tasks, based on which we propose the IQA formula for robust representation learning, followed by its training paradigm.

3.1. Overview of the Proposed Method

As shown in Figure 2, the basic idea of our method is to learn the distortion-invariant and content-invariant representations based on causal learning. And the whole framework consists of three stages. Given the reference images from synthetically distorted image datasets, we construct sample subsets based on confounding factors first in **Stage 1**: images with the same distortion type are collected into the same sample subset. Due to unavailability of content labels in IQA datasets, we utilize the GMM (Do & Batzoglou, 2008) model for clustering after extracting semantic features by pretrained backbone (ResNet18 (He et al., 2016) is cho-

sen in our method), to prepare for the subsequent training process. In **Stage 2**, we propose a causality-based training paradigm to improve the generalization of IQA, which can be used to eliminate the confounding effects from distortion types and image contents, so that the causal relationship between distorted images and corresponding MOS labels is identifiable (the optimization progress is executed based on a meta-learning strategy using the sample subsets and the cluster probabilities obtained in Stage 1). Finally, in **Stage 3**, the trained IQA model can be used to assess the quality score for new images with unseen distortions and contents.

3.2. The Causal Mechanism of BIQA

BIQA aims to predict the MOS score m of distorted image I_d without reference images, denoted as $m = f_\theta(I_d)$, where θ means the model parameters. From a causal perspective, we model this BIQA progress with a causal structure graph in Figure 1, where I_d and m are regarded as Intervention Object and Intervention Outcome respectively. Suppose there exists an ideal image I_v for each distorted image I_d, the path $D \rightarrow I_d \leftarrow C$ in Figure 1 means the degradation process from I_v to I_d, where D and C denote the variables of distortions and image contents (For distortion information, we only consider distortion type in D in our method, and the discussion about distortion degrees is presented in the Appendix). Besides, the path $D \rightarrow m \leftarrow C$ in Figure 1 denotes the knowledge learned from D and C to m.

In order to improve the generalization of the IQA model, f_θ should be trained to learn the causal representation based on the causation between I_d to m. However, there are two virtual paths connected I_d and m, including $I_d \leftarrow C \rightarrow m$ and $I_d \leftarrow D \rightarrow m$ caused by C and D, which are the confounding factors that bring confounding affects on both the Intervention Object (i.e. I_d) and Intervention Outcome (i.e. m) under study. More precisely, the existing of virtual paths causes the conditional probability $P(m|I_d)$ learned by IQA network f_θ is also conditioned on these two confounding factors, thus the traditional IQA training paradigm is actually fitting $P(m|I_d, C, D)$, leading to the IQA network f_θ not robust to different distortions and contents due to that it is not independent of D and C. To address this issue, we learn the robust IQA representation by Counterfactuals (Pearl, 2022) and Backdoor Adjustment (Pearl, 1995) in the causal learning domain since D is observable in synthetically distorted image datasets.

Backdoor Adjustment: In causal inference, the primary objective of backdoor adjustment is to estimate the intervention distribution, enabling the identification of the causal relationship of the target variables, which is based on Backdoor Criteria (see details in supplementary materials). Specifically speaking, we make a *do-calculus* based on the discretization results of confounding factors C and D, and *do-calculus* transforms the intervention distribution required for causal inference into a probability distribution in statistical learning (Pearl, 1995), which can block the causal relationships from C and D to I_d to eliminate their confounding effects. In this way, we can learn the robust IQA representation independent of distortion types and image contents by instantiating the conditional probability $P\left(m \mid do\left(I_d\right)\right)$. We have the following proposition for Backdoor Adjustment:

Proposition 3.1. *The causal conditional probability in IQA tasks can be formulated as:*

$$P\left(m \mid do\left(I_d\right)\right) = \sum_{(c_i,d_j)\in S} P\left(m \mid I_d, c_i, d_j\right) P\left(c_i, d_j\right) \quad (1)$$

where $P\left(c_i, d_j\right)$ denotes the probability of each distortion type and content cluster.

Proof. See supplementary materials. □

Suppose there exist p distortion types and the cluster number is set as q in Eq. 1, which satisfy uniform distribution in IQA datasets, then we have $P(c_i, d_j) = P(c_i)P(d_j) = \frac{1}{pq}$.

Counterfactual Division: In order to make the trained IQA model independent of different distortion types and image contents with enough generalization ability, we construct the confounders $S = \{(c_i, d_j)\}$ first in Stage 1 of Figure 1, which consists of different (distortion type, image content) pairs. According to (Li et al., 2023), it is better for each reference image to have corresponding distorted images with varying distortion types and image contents. Therefore, we collect images with the same distortion type in the same content cluster into the same sample subset, and each sample subset is corresponding to one element in S. We call this process Counterfactual Division. In fact, the label of distortion type is available in synthetically distorted image datasets, and we obtain the cluster probabilities about image contents by GMM (Do & Batzoglou, 2008) and the lower-dimensional semantic features extracted by ResNet18 and TSNE (Van der Maaten & Hinton, 2008)). Here we have the following proposition to support the rightness of Backdoor Adjustment and the division based on S.

Proposition 3.2. *The construction of above IQA data subsets conforms to the calculating of Counterfactuals. Given S, for $\forall i_d \in I_d$, counterfactual m_{i_d} is conditionally independent of I_d (i.e. $m_{i_d} \perp I_d \mid S$).*

$$P\left(m_{i_d} \mid I_d, S\right) = P\left(m_{i_d} \mid S\right) \quad (2)$$

An easy corollary of Proposition 3.2 is the following:

Corollary 3.3. *The probability of counterfactual m_{i_d} is equal to the calculation of $P\left(m \mid do\left(I_d\right)\right)$. That is:*

$$P\left(m_{i_d} = y\right) = \sum_{s=(c_i,d_j)} P(m = y \mid S = s, I_d = i_d)P(s) \quad (3)$$

Algorithm 1 Causal-IQA-S (CIS)

Input: training dataset $\mathcal{X} = \{(I_d, m) \mid d \in S\}$, confounders $S = \{(c_i, d_j)\}, 1 \le i \le q, 1 \le j \le p$, learning rate α and β for inner and outer loop updating.
Initial: BIQA network f_θ.
repeat
  $\phi_{\bar{s}} \leftarrow \theta$.
  for $s_{i,j}$ **in** S **do**
    Sample training pairs $(I_{s_{i,j}}, m)$ from $\mathcal{X}$
    Task Updating: $\phi_{\bar{s}} = \phi_{\bar{s}} - \alpha \nabla_{\phi_{\bar{s}}} \mathcal{L}\left(f_{\phi_{\bar{s}}}\left(I_{s_{i,j}}\right), m\right)$
  end for
  Meta Updating: $\theta \leftarrow \theta - \beta(\phi_{\bar{s}} - \theta)$
until convergence.
Output: convergent meta parameters θ^*.

Algorithm 2 Causal-IQA-P (CIP)

Input: training dataset $\mathcal{X} = \{(I_d, m) \mid d \in S\}$, confounders $S = \{(c_i, d_j)\}, 1 \le i \le q, 1 \le j \le p$, learning rate α and β for task updating and meta updating.
Initial: BIQA network f_θ.
repeat
  $\phi_{\bar{s}} \leftarrow \theta$.
  Sample training pairs $\left\{(I_{s_{i,j}}, m)\right\}_{i,j}$ from $\mathcal{X}$
  Task Updating:
    $\phi_{\bar{s}} = \phi_{\bar{s}} - \alpha \nabla_{\phi_{\bar{s}}} \frac{1}{pq} \sum_{s_{i,j} \in S} \mathcal{L}\left(f_{\phi_{\bar{s}}}\left(I_{s_{i,j}}\right), m\right)$
  Meta Updating: $\theta \leftarrow \theta - \beta(\phi_{\bar{s}} - \theta)$
until convergence.
Output: convergent meta parameters θ^*.

Proof. See Appendixes for the proof of Eq. 2 and Eq. 3. □

3.3. Causality-based Representation Learning

With the causality-based optimization direction defined in Eq. 1, we can achieve a robust representation learning by instantiating the intervention of different distortion types and image contents. In the traditional ERM-based training paradigm in BIQA tasks, the network f_θ is trained by minimize the loss function $\mathcal{L}\left(f_\theta\left(I_d\right), m\right)$ in the entire training dataset $\mathcal{X} = \{(I_d, m) \mid d \in S\}$ to maximize the condition probability $P(m|I_d)$, which can be denoted as:

$$\theta^* = \arg\min_\theta \mathbb{E}_{(I_d,m)\in\mathcal{X}}\left[\mathcal{L}\left(f_\theta\left(I_d\right), m\right)\right], \tag{4}$$

where $\mathcal{L}(\cdot)$ is usually L_1 loss or EMD loss (Talebi & Milanfar, 2018), and θ^* is the learned parameters. As mentioned above, due to the poor generalization in this training paradigm, we implement the causality-based representation learning by maximizing the causal conditional probability $P\left(m \mid do\left(I_d\right)\right)$ instead of $P\left(m \mid I_d\right)$.

In order to model the intervention from confounders S to instantiate conditional probability $P(m|I_d, c_i, d_j)$ in Eq. 1, we execute the one-step update for each pair $(c_i, d_j) \in S$ to obtain the IQA model f_θ conditioned on specific distortion type and content cluster, which is implemented on the corresponding sample subset obtained in the Counterfactual Division stage in Figure 1, derived as:

$$\phi_{s_{i,j}} = \theta - \alpha \nabla_\theta \mathcal{L}\left(f_\theta\left(I_{s_{i,j}}\right), m\right), s_{i,j} \in S \tag{5}$$

where $I_{s_{i,j}}$ denotes the distorted images I_d in content cluster c_i with distortion type d_j, and $\phi_{s_{i,j}}$ is the corresponding one-step updated model. Note that one image in the wild does not necessarily belong strictly to one content cluster, so the loss $\mathcal{L}\left(f_\theta\left(I_{s_{i,j}}\right), m\right)$ can be computed with the cluster probabilities of each samples:

$$\mathcal{L}(f_\theta\left(I_{s_{i,j}}\right), m) = \sum_r \frac{w_{r,j}}{w_{:,j}} \mathcal{L}\left(f_\theta\left(I_{s_{i,j}}(r)\right), m(r)\right) \tag{6}$$

where $w_{r,j}$ ($\sum_j w_{r,j}$=1) denotes the probability of the r-th sample attaching to j-th content cluster, and $w_{:,j}$=$\sum_r w_{r,j}$. Then we can maximize $P(m|I_d, c_i, d_j)$ by minimize the overall loss $\mathcal{L}(f_{\phi_{s_{i,j}}}(I_d), m)$. Therefore, $P\left(m \mid do\left(I_d\right)\right)$ can be maximized on the whole training set:

$$\theta^* = \arg\min_\theta \mathbb{E}_{(I_d,m)\in\mathcal{X}}[\sum_{s_{i,j}\in S} \mathcal{L}\left(f_{\phi_{s_{i,j}}}\left(I_{s_{i,j}}\right), m\right)] \tag{7}$$

Finally, we can learn the generalizable BIQA networks by the optimization direction in Eq. 7

3.4. The Proposed Training Paradigm

It is evident that the optimization procedure derived in Eq. 5 6 7 is challenging to execute, which is memory-consuming and computationally expensive. We have the following proposition to address this issue.

Proposition 3.4. *The optimization procedure for Causal-IQA in Eq. 5 7 is equivalent to the optimization direction described as follows:*

$$\begin{aligned} &\theta^* = \arg\min_\theta \mathbb{E}_{(I_d,m)\in\mathcal{X}}\left[\mathcal{L}\left(f_{\phi_{\bar{s}}}\left(I_d\right), m\right)\right] \\ &\text{and } \phi_{\bar{s}} = \theta - \alpha\nabla_\theta \frac{1}{pq} \sum_{s_{i,j}\in S} \mathcal{L}\left(f_\theta\left(I_{s_{i,j}}\right), m\right) \end{aligned} \tag{8}$$

where $\mathcal{L}\left(f_\theta\left(I_{s_{i,j}}\right), m\right)$ is defined in Eq. 6.

Proof. See supplementary materials. □

According to Proposition 3.4, $\phi_{\bar{s}}$ is the median model parameters virtually updated with loss function $\mathcal{L}(\cdot)$ on all the image subsets obtained in Counterfactual Division, thus the optimization procedure in Eq. 8 can be implemented by emulating the meta-learning strategy MAML (Finn et al., 2017). However, a second-order gradient needs to be calculated during the iteration in MAML, which is relatively complicated in computation. Inspired by Reptile (Nichol et al.,

Table 1. Attributes of five typical IQA databases in experiments.

Databases	Number	MOS Range	Distortion Type
TID2013	3,000	[0,9]	Synthetic
KADID-10K	10,125	[1,5]	Synthetic
KonIQ-10K	10,073	[1,5]	Authentic
LIVE-C	1,162	[0,100]	Authentic
CID2013	480	[0,100]	Authentic

2018), we use the sequential parameter updating strategy based on the one-order gradient to approximate the second-order gradient, thus the optimization process can be greatly simplified. Eventually, according to the updating mode of the median parameters $\phi_{\bar{s}}$, we propose two versions of the optimization algorithm for Causal-IQA method, which are summarized in Algorithm 1 and Algorithm 2 termed as Causal-IQA-S (CIS for short) and Causal-IQA-P (CIP for short) respectively. Causal-IQA-S Causal-IQA-P compute the task loss for virtual updating in the serial and parallel manner respectively, the relevant analyses are illustrated in supplementary materials.

4. Experiments

4.1. Datasets

In this paper, we perform experiments on the following five representative IQA databases:

- TID2013 (Ponomarenko et al., 2015). This database contains 3000 images, which is obtained from 25 reference images, 24 types of distortions for each reference image,and 5 levels for each type of distortion.
- KADID-10K (Lin et al., 2019). It includes 81 pristine images, where each pristine image was degraded by 25 distortions in 5 levels. For each distorted image, 30 reliable degradation category ratings were obtained by crowdsourcing performed by 2,209 crowd workers.
- KonIQ-10K (Hosu et al., 2020). It consists of 10,073 authentically distorted images chosen from YFCC100M (Thomee et al., 2016) with a wide distribution about brightness, color, contrast and sharpness.
- LIVE-C (Ghadiyaram & Bovik, 2015). LIVE-C consists of 1162 authentically distorted images captured from many diverse mobile devices. Each image was assessed by an average of 175 unique subjects.
- CID2013 (Virtanen et al., 2014). It includes six image sets; on average, 30 subjects have evaluated 12–14 devices depicting eight different scenes for a total of 79 different cameras, 480 images, and 188 subjects.

The details of these datasets are shown in Table 1, where synthetically distorted image datasets are mainly used for training and the authentically distorted datasets are used for verification and fine-tuning. For training and testing uniformity, all the MOS ranges are normalized to $[0, 1]$.

Table 2. Fine-tuning results on authentically distorted image datasets with models trained on TID2013 and KADID-10K. Bold indicates the best performance, and the same convention is applied to other metrics and tables.

Methods	CID2013		LIVE-C		KonIQ-10K	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
BLIINDS-II	0.565	0.487	0.507	0.463	0.615	0.529
BRISQUE	0.648	0.615	0.645	0.607	0.537	0.473
ILNIQE	0.538	0.346	0.589	0.594	0.537	0.501
CORNIA	0.680	0.624	0.662	0.618	0.795	0.780
HOSA	0.685	0.663	0.678	0.659	0.813	0.805
BIECON	0.620	0.606	0.613	0.595	/	/
MEON	0.703	0.701	0.693	0.688	/	/
WaDIQaM-NR	0.729	0.708	0.680	0.671	0.761	0.739
DistNet-Q3	/	/	0.601	0.570	0.710	0.702
DIQA	0.720	0.708	0.704	0.703	/	/
NSSADNN	0.825	0.748	0.813	0.745	/	/
MetaIQA	0.784	0.766	0.835	0.802	0.887	0.850
$CIS(Ours)$	**0.887**	**0.895**	**0.847**	**0.828**	**0.918**	**0.881**
$CIP(Ours)$	0.873	0.894	0.844	0.823	0.896	0.865

4.2. Experimental Settings

Implementation Details. During training with Causal-IQA, we choose the same backbone ResNet18 (He et al., 2016) for a compelling performance comparison with MetaIQA (Zhu et al., 2020), which also designed the generalizable training paradigm based on Meta Learning. In Stage 1 of Counterfactual Division, we set the number c of clustering in GMM as 2, which can be considered as contents-based knowledge guiding for $high\ quality$ and $low\ quality$. Before GMM, the image features extracted by ResNet18 are reduced in dimension to 10. The training network is constructed following MetaIQA, which is trained with Pytorch library (Paszke et al., 2017) on two Intel Xeon E5-2609 v4 CPUs and four NVIDIA RTX 2080Ti GPUs. During training, images are scaled to the size of 256×256 and then randomly cropped to 224×224 before feeding to models. We use Adam optimizer with $\beta_1 = 0.9$ and $\beta_1 = 0.9999$ for both task updating and meta updating, and the learning rates for task updating and meta updating are set as $1e-4$ and $1e-2$ respectively. The learning rate for task updating drops by 0.8 times after every 10 epochs with the total epoch number set as 100. In the process of fine-tuning, the learning rate of Adam optimizer is also set as $1e-4$ with total epoch number 30. The batch sizes B in TID2013 and KADID-10K are set as 32 and 102 for combined training, and $B = 64$ during test process. The backbone is initialized by the pre-training weights obtained by classification task on ImageNet (Deng et al., 2009).

Evaluation Metrics. We evaluate our Causal-IQA models by two typical metrics (Bosse et al., 2017), including Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank-order Correlation Coefficient (SROCC).

Table 3. SROCC results comparison in leave-one-distortion-out cross validation on the synthetically distorted image database KADID-10K.

	DIST. TYPE	BLIINDS-II	BRISQUE	ILNIQE	CORNIA	HOSA	WADIQAM-NR	METAIQA	CIS	CIP
KADID-10K	GB	0.8799	0.8118	0.8831	0.8655	0.8522	0.8792	0.9461	**0.9524**	0.9349
	LB	0.7810	0.6738	0.8459	0.8109	0.7152	0.7299	0.9168	0.9247	**0.9306**
	MB	0.4816	0.4226	0.7794	0.5323	0.6515	0.7304	**0.9262**	0.9051	0.9135
	CD	0.5719	0.5440	0.6780	0.2432	0.7272	0.8325	**0.9194**	0.9043	0.9026
	CS	-0.1392	0.1821	0.0898	-0.0023	0.0495	0.4209	0.7850	0.8112	**0.8255**
	CQ	0.6695	0.6670	0.6763	0.3226	0.6617	**0.8055**	0.7170	0.7405	0.7231
	CSA1	0.0906	0.0706	0.0266	-0.0194	0.2158	0.1479	0.3039	0.4513	**0.4658**
	CSA2	0.6017	0.3746	0.6771	0.1197	0.8408	0.8358	0.9310	**0.9388**	0.9043
	JP2K	0.6546	0.5159	0.7895	0.3417	0.6078	0.5387	0.9452	0.9379	**0.9508**
	JPEG	0.4140	0.7821	0.8036	0.5561	0.5823	0.5298	0.9115	**0.9348**	0.9001
	WN	0.6277	0.7080	0.7757	0.3574	0.6796	0.8966	**0.9047**	0.8934	0.8877
	WNCC	0.7567	0.7182	0.8409	0.4183	0.7445	0.9247	0.9303	0.9215	**0.9340**
	IN	0.5469	-0.5425	0.8082	0.2188	0.2535	0.8142	0.8673	**0.8810**	0.8767
	MN	0.7017	0.6741	0.6824	0.3060	0.7757	0.8841	0.9247	**0.9316**	0.9089
	DENOISE	0.4566	0.2213	0.8562	0.2293	0.2466	0.7648	0.8985	**0.9428**	0.9041
	BRIGHTEN	0.4583	0.5754	0.3008	0.2272	0.7525	0.6845	0.7827	**0.8267**	0.8087
	DARKEN	0.4391	0.4050	0.4363	0.2060	0.7436	0.2715	0.6219	**0.6851**	0.6424
	MS	0.1119	0.1441	0.3150	0.1215	0.5907	0.3475	0.5555	**0.6429**	0.6199
	JITTER	0.6287	0.6719	0.4412	0.7186	0.3907	0.7781	**0.9278**	0.8909	0.8740
	NEP	0.0832	0.1911	0.2178	0.1206	0.4607	0.3478	0.4184	**0.7012**	0.6723
	PIXELATE	0.1956	0.6477	0.5770	0.5868	0.7021	0.6998	0.8090	**0.8323**	0.8122
	QUANTIZATION	0.7812	0.7135	0.5714	0.2592	0.6811	0.7345	0.8770	**0.8845**	0.8313
	CB	-0.0204	0.0673	0.0029	0.0937	0.3879	0.1602	0.5132	0.6527	**0.7446**
	HS	-0.0151	0.3611	0.6809	0.1142	0.2302	**0.5581**	0.4374	0.5173	0.5091
	CC	0.0616	0.1048	0.0723	0.1253	0.4521	0.4214	0.4377	**0.4891**	0.4589
	AVERAGE	0.4328	0.4136	0.5528	0.3149	0.5598	0.6295	0.7672	**0.8078**	0.7974

4.3. Performances Comparison

We conduct two sets of experiments to verify the superiority of our proposed models. One of them is testing the generalization to authentically distorted images by Causal-IQA that is trained on synthetically distorted datasets, another is testing its generalization to unseen distortion types on synthetically distorted datasets. Twelve typical BIQA methods are used to make comparisons with Causal-IQA, which consist of five traditional BIQA methods (including BLIINDS-II (Saad et al., 2012), BRISQUE (Mittal et al., 2012), ILNIQE (Zhang et al., 2015), CORNIA (Ye et al., 2012) and HOSA (Xu et al., 2016)) and seven DL-based methods (including BIECON (Kim & Lee, 2016), MEON (Ma et al., 2017), WaDIQaMNR (Bosse et al., 2017), DistNet-Q3 (Dendi et al., 2018), DIQA (Kim et al., 2018), NSSADNN (Yan et al., 2019) and MetaIQA (Zhu et al., 2020)), among which MetaIQA is one of the state-of-the-art BIQA training paradigms for generalization enhancement.

Results on Authentical Distortions. In this part, we train Causal-IQA methods on two synthetically distorted datasets and fine-tune the trained model on three authentically distorted datasets, 80% of which are allocated for fine-tuning with 20% for testing. The comparison results between Causal-IQA and other methods are displayed in Table 2.

Compared with these baselines, we can observe that: **(i)**. DL-based methods generally outperform traditional approaches. **(ii)**. Although CIP performs slightly worse than CIS, two versions of Causal-IQA both have better performances than MetaIQA, which illustrates the robustness and effectiveness of the proposed methods.

Results on Synthetical Distortions. To assess how well our Causal-IQA generalizes to unfamiliar distortion types, we evaluate our approach by employing Leave-One-Distortion-Out (Zhu et al., 2020) cross-validation on the TID2013 and KADID-10K databases. In other words, a data subset with the specific distortion type is retained for testing, while all samples with other distortion types are used for training. The SORCC values on KADID-10K of our methods are recorded in Table 3 for comparison with seven SOTA general-purpose BIQA algorithms (See Table 12 in Appendix for the results on TID2013).

From Table 3 and Table 12, we can observe that: **(i)**. Both MetaIQA and CausalIQA show significant superiority over other approaches, demonstrating the generalization of the meta-learning-based training paradigm. **(ii)**. The average SORCC values of our methods are two to four points higher than that of MetaIQA on both TID2013 and KADID-10K, which proves the enough generalization of Causal-IQA on unseen distortion types.

4.4. Ablation Study

To further investigate whether the effectiveness and robustness of our methods are derived from causal relationships, we investigate the impact of different components of the proposed Causal-IQA in this experiment. To be specific, we compare the generalization performances of Baseline, MetaIQA, $CIS\backslash C$, CIS, $CIP\backslash C$ and CIP on KonIQ-10K, CID2013 and LIVE-C. Baseline means training IQA network f_θ with the traditional ERM-based training paradigm in Eq. 4. $CIS\backslash C$ and $CIP\backslash C$ denote training the same IQA network f_θ according to Algorithm 1 and Algo-

Table 4. Ablation study of test results on wild images without fine-tuning using models trained on TID2013 and KADID-10K.

METHODS	CID2013		LIVE-C		KONIQ-10K	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
BASELINE	0.547	0.538	0.562	0.538	0.544	0.553
METAIQA	0.597	0.588	0.606	0.614	0.610	0.629
$CIS\backslash C$	0.673	0.642	0.664	0.652	0.630	0.628
CIS	0.672	0.651	**0.690**	**0.684**	0.662	0.658
$CIP\backslash C$	0.659	0.641	0.652	0.615	0.601	0.613
CIP	**0.712**	**0.683**	0.655	0.624	**0.681**	**0.704**

Table 5. Ablation study of fine-tuning results on authentically distorted images with models trained on TID2013 and KADID-10K.

METHODS	CID2013		LIVE-C		KONIQ-10K	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
BASELINE	0.727	0.712	0.801	0.743	0.832	0.816
METAIQA	0.784	0.766	0.835	0.802	0.887	0.850
$CIS\backslash C$	0.855	0.871	0.839	0.811	0.861	0.864
CIS	**0.887**	**0.895**	**0.847**	**0.828**	**0.918**	**0.881**
$CIP\backslash C$	0.863	0.865	0.826	0.807	0.881	0.832
CIP	0.873	0.894	0.844	0.823	0.896	0.865

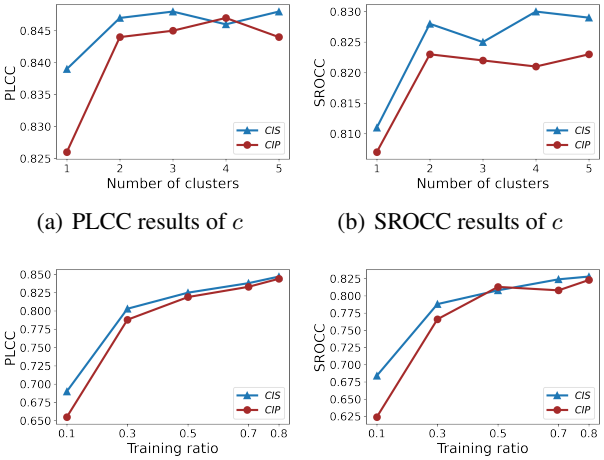
rithm 2 without considering the confounding factor of image content. In other words, there is no path $I_d \leftarrow C \rightarrow m$ in the structural causality graph in Figure 1 and confounders $S = \{d_j\}_{j=1}^{p}$. Same setup as before, these models are trained on TID2013 and KADID-10K first, then Table 4 records the PLCC and SORCC results of directly testing on these authentically distorted image databases without fine-tuning, while Table 5 records that of fine-tuning with an 8:2 training-testing split. Upon closer examination, it becomes evident that: **(i)**. $CIS\backslash C$ and $CIP\backslash C$ perform better than Baseline, yet perform worse than CIS and CIP, which prove the effectiveness of the elimination of confounding effects caused by distortion type and image content. **(ii)**. In Table 4, the PLCC and SROCC of our Causal-IQA is about 7-8 points higher than that of MetaIQA, showing that our method has better zero-shot capability due to its interpretability. **(iii)**. CIS performs better than CIP, which illustrates that the serial manner is more suitable for our causality-based training strategy in BIQA tasks.

In addition, we make comparisons among the fine-tuned performances of Causal-IQA models trained with different backbones on TID2013 and KADID-10K, including ResNet18, ResNet34 and ResNet50 (He et al., 2016). According to Table 6, the generalization performances of CIS get better as the IQA network gets deeper, which is the same as the conclusion of CIP. See Appendix for details.

We also explore the impact of the partition ratio (set t as the training ratio and the test ratio is $1-t$) during fine-tuning and number c of clustering in GMM mentioned in Section 4.2. With the trained CIS and CIP models on TID2013 and KADID10K, we compare the fine-tuning results on LIVE-C with different t and c. Figure 3(a) 3(b) describe the

Table 6. Impact of different backbones on the results of CIS.

BACKBONES		RESNET18	RESNET34	RESNET50
CID2013	PLCC	0.887	0.910	**0.921**
	SROCC	0.895	0.907	**0.923**
LIVE-C	PLCC	0.847	0.872	**0.885**
	SROCC	0.828	0.869	**0.881**
KONIQ-10K	PLCC	0.918	0.923	**0.936**
	SROCC	0.881	0.899	**0.915**



(a) PLCC results of c (b) SROCC results of c

(c) PLCC results of t (d) SROCC results of t

Figure 3. Impact of different cluster numbers c (during training) and training ratios t (during fine-tuning) on database LIVE-C.

variation trends of PLCC and SROCC when c changed in $[1, 5]$ ($c = 1$ means $S = \{d_j\}_{j=1}^{p}$) with $t = 0.8$ fixedly, and Figure 3(c) 3(d) describe the trends when t changed in $\{0.1, 0.3, 0.5, 0.7, 0.8\}$. See Appendix for more results on other datasets, and we can conclude that: **(i)**. Setting c to 2 is optimal, as increasing c leads to larger computational overhead without improving performance. **(ii)**. As training ratio t increases, the performance gradually improves.

5. Conclusion

This paper provides a novel BIQA training paradigm named Causal-IQA, which can improve the generalization on unseen distortion types and image contents. Concretely speaking, we first construct the structural causality graph for BIQA task, which contains two confounding factors hindering the identification of the causal relationship between image and MOS label. Thus we achieve a Counterfactual Division for training datasets to eliminate the confounding effects with Backdoor Adjustment strategy. Finally, we design two versions of causality-based training paradigm CIS and CIP that are proved effective and robust by extensive experiments. It is worth mentioning that Causal-IQA is interpretable and can be applied to any BIQA network. To our knowledge, this is the first work to explore IQA tasks from a causal perspective.

Acknowledgements

This work is partially supported by Sino-German Center (M 0187) and the NSFC under contract 62088102. We also acknowledge High-Performance Computing Platform of Peking University for providing computational resources.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning and Image Quality Assessment. There may exist some potential societal consequences of our work, none of which must be specifically highlighted here.

References

Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T., and Samek, W. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017.

Chen, L., Zhang, H., Xiao, J., He, X., Pu, S., and Chang, S.-F. Counterfactual critic multi-agent training for scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4613–4623, 2019.

Cui, Y., Ren, W., Yang, S., Cao, X., and Knoll, A. Irnext: Rethinking convolutional network design for image restoration. 2023.

Dendi, S. V. R., Dev, C., Kothari, N., and Channappayya, S. S. Generating image distortion maps using convolutional autoencoders with application to no reference image quality assessment. *IEEE Signal Processing Letters*, 26(1):89–93, 2018.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Do, C. B. and Batzoglou, S. What is the expectation maximization algorithm? *Nature biotechnology*, 26(8):897–899, 2008.

Elasri, M., Elharrouss, O., Al-Maadeed, S., and Tairi, H. Image generation: A review. *Neural Processing Letters*, 54(5):4609–4646, 2022.

Feng, Y., Li, S., and Hao, S. An error self-learning semi-supervised method for no-reference image quality assessment. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pp. 1–5. IEEE, 2021.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Ghadiyaram, D. and Bovik, A. C. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015.

Golestaneh, S. A., Dadsetan, S., and Kitani, K. M. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1220–1230, 2022.

Gresele, L., Von Kügelgen, J., Kübler, J., Kirschbaum, E., Schölkopf, B., and Janzing, D. Causal inference through the structural causal marginal problem. In *Proceedings of the International Conference on Machine Learning*, pp. 7793–7824. PMLR, 2022.

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

Hosu, V., Lin, H., Sziranyi, T., and Saupe, D. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

Jayaraman, D., Mittal, A., Moorthy, A. K., and Bovik, A. C. Objective quality assessment of multiply distorted images. In *2012 Conference record of the forty sixth asilomar conference on signals, systems and computers (ASILOMAR)*, pp. 1693–1697. IEEE, 2012.

Jiang, Q., Shao, F., Lin, W., Gu, K., Jiang, G., and Sun, H. Optimizing multistage discriminative dictionaries for blind image quality assessment. *IEEE Transactions on Multimedia*, 20(8):2035–2048, 2017.

Kanehira, A., Takemoto, K., Inayoshi, S., and Harada, T. Multimodal explanations by predicting counterfactuality in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8594–8602, 2019.

Ke, J., Wang, Q., Wang, Y., Milanfar, P., and Yang, F. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021.

Kim, J. and Lee, S. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11 (1):206–220, 2016.

Kim, J., Nguyen, A.-D., and Lee, S. Deep cnn-based blind image quality predictor. *IEEE transactions on neural networks and learning systems*, 30(1):11–24, 2018.

Li, D., Jiang, T., Lin, W., and Jiang, M. Which has better visual quality: The clear blue sky or a blurry animal? *IEEE Transactions on Multimedia*, 21(5):1221–1234, 2018.

Li, X., Li, B., Jin, X., Lan, C., and Chen, Z. Learning distortion invariant representation for image restoration from a causality perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1714–1724, 2023.

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

Liu, X., Van De Weijer, J., and Bagdanov, A. D. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*, pp. 1040–1049, 2017.

Liu, Y., Gu, K., Li, X., and Zhang, Y. Blind image quality assessment by natural scene statistics and perceptual characteristics. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(3):1–91, 2020.

Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., and Zuo, W. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2017.

Madhusudana, P. C., Birkbeck, N., Wang, Y., Adsumilli, B., and Bovik, A. C. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022.

Mittal, A., Moorthy, A. K., and Bovik, A. C. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

Nichol, A., Achiam, J., and Schulman, J. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Pawlowski, N., Coelho de Castro, D., and Glocker, B. Deep structural causal models for tractable counterfactual inference. *Proceedings of the Annual Conference in Neural Information Processing Systems*, 33:857–869, 2020.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Pearl, J. Causal inference in statistics: An overview. 2009.

Pearl, J. Causal inference. *Causality: objectives and assessment*, pp. 39–58, 2010.

Pearl, J. Causal diagrams for empirical research (with discussions). In *Probabilistic and causal inference: The works of Judea Pearl*, pp. 255–316. 2022.

Pearl, J., Glymour, M., and Jewell, N. P. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Pearl, J. et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000.

Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015.

Roy, S., Mitra, S., Biswas, S., and Soundararajan, R. Test time adaptation for blind image quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16742–16751, 2023.

Saad, M. A., Bovik, A. C., and Charrier, C. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE transactions on Image Processing*, 21(8):3339–3352, 2012.

Simeng Sun, Tao Yu, J. X. W. Z. Z. C. Graphiqa: Learning distortion graph representations for blind image quality assessment. *IEEE Transactions on Multimedia*, 2023.

Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., and Zhang, Y. 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*, pp. 3667–3676, 2020.

Talebi, H. and Milanfar, P. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8): 3998–4011, 2018.

Tang, K., Niu, Y., Huang, J., Shi, J., and Zhang, H. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 3716–3725, 2020.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., and Lin, C.-W. Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275, 2020.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Virtanen, T., Nuutinen, M., Vaahteranoksa, M., Oittinen, P., and Häkkinen, J. Cid2013: A database for evaluating no-reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 24(1):390–402, 2014.

Wu, X., Jiang, B., Zhong, Y., and Chen, H. Tolerant markov boundary discovery for feature selection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2261–2264, 2020.

Xu, J., Ye, P., Li, Q., Du, H., Liu, Y., and Doermann, D. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, 25(9):4444–4457, 2016.

Yan, B., Bare, B., and Tan, W. Naturalness-aware deep no-reference image quality assessment. *IEEE Transactions on Multimedia*, 21(10):2603–2615, 2019.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data*, 15(5):1–46, 2021.

Ye, P., Kumar, J., Kang, L., and Doermann, D. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 1098–1105. IEEE, 2012.

Yue, G., Cheng, D., Li, L., Zhou, T., Liu, H., and Wang, T. Semi-supervised authentically distorted image quality assessment with consistency-preserving dual-branch convolutional neural network. *IEEE Transactions on Multimedia*, 2022.

Zhai, G. and Min, X. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63:1–52, 2020.

Zhang, L., Zhang, L., and Bovik, A. C. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015.

Zhou, W., Yu, L., Qiu, W., Zhou, Y., and Wu, M. Local gradient patterns (lgp): An effective local-statistical-feature extraction scheme for no-reference image quality assessment. *Information Sciences*, 397:1–14, 2017.

Zhu, H., Li, L., Wu, J., Dong, W., and Shi, G. Metaiqa: Deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14143–14152, 2020.

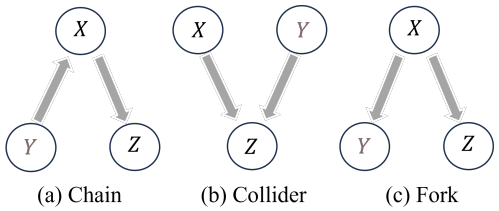


Figure 4. Three typical causal structures in Causal Learning, including Chain structure, Collider structure and For structure.

A. Causation VS. Correlation

According to (Pearl et al., 2000), a causal relationship denotes a cause-and-effect connection between two variables, where a change in one variable influences the other. In this scenario, one variable is considered the cause, leading to an observable impact or change in the other variable, known as the effect. Causal relationships imply a directional flow of influence, emphasizing the idea that changes in the cause lead to changes in the effect. Thus causation is interpretable and directed, such as smoking and lung cancer. While correlation refers to a statistical measure that quantifies the degree of association or relationship between two variables, such as yellow fingers and lung cancer. Therefore, correlation alone does not imply causation. Two variables can be correlated without one causing the other (Pearl et al., 2016; Pearl, 2009).

There are three typical causal structures (Pearl et al., 2000), including fork structure, chain structure, and collider structure as shown in Figure 4. For example, in chain structure, X, Y and Z can be rain, slippery roads, and slip, respectively. We cannot conclude that rain is the cause of the slip. Hence, the lack of correlation may not necessarily imply the absence of causation, which may be due to incorrectly controlling for intermediary variables. In collider structure, X, Y and Z can be the blood types of the father, mother, and child, respectively. Therefore, the transformation of X and Y from being uncorrelated to correlated does not necessarily indicate a causal relationship, which could be due to selection bias. In fork structure, X, Y and Z can be cold, fever and runny nose, respectively. So, high correlation between two variables (Y and Z) does not necessarily imply a causal relationship.

In general, the model built based on causality is more robust than that based on correlation (Wu et al., 2020).

B. Backdoor Adjustment in Causal Inference and The Proof of Proposition 3.1

Backdoor Adjustment is designed based on Backdoor Criterion (Pearl et al., 2016; Pearl, 2009), which aims to eliminating the confounding effects for making the causal rela-

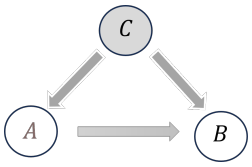


Figure 5. A structural causal graph with is confounding factor C.

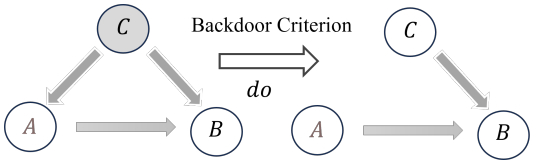


Figure 6. The *do* operation eliminating confounding effects based on Back-door Criterion.

tionship identifiable between two target variables. Here, the concept of confounding is introduced first.

Definition B.1. Confounder: Given three random variables, A, B, and C. C is the confounder of causality between A and B if C is the common cause between A and B, leading to the confounding effect:

$$P(B \mid do(A)) \neq P(B \mid A) \tag{9}$$

In Definition B.1, $do(A)$ denotes the intervention on variable A by setting it to the specific value, $P(B \mid do(A))$ indicates the direct causal relationship from the path $A \rightarrow B$ and $P(B \mid A)$ denotes the correlation between A and B. For example, as shown in Figure 5, there is a spurious correlation between A and B in the fork connection $A \leftarrow C \rightarrow B$, leading to a confounding effect on the identification of the causal relationship from A to B. Therefore, it is necessary to eliminate confounders for estimating the causality from parent node to child node, that is the conditional probability $P(B \mid do(A))$. In other words, as shown in Figure 6, the *do* operation aims to cut off the connection from C to A, which is implemented based on Backdoor Criterion.

Definition B.2. Backdoor Criterion: In a directed acyclic graph $G = (\mathcal{V}, \mathcal{E})$, given a pair of variables $(A, B) \in \mathcal{V}$, the variable set Z satisfies the Backdoor Criterion relative to (A, B) if and only if any element in C is not a successor node of A, and C blocks all paths from variables in A to B that are toward A.

Proposition B.3. *If C is observable and satisfies the Backdoor Criterion with respect to (A, B) in the structural causal graph of Figure 5, the causal effect from A to B can be calculated by:*

$$P(B \mid do(A)) = \sum_c P(B \mid A, C = c)P(C = c) \tag{10}$$

Proof. Without loss of generality, we set $A = a$ and $B = b$, then we have:

$$\begin{aligned} &P\left(B = b \mid do(A = a)\right) \\ &= \sum_c P(B = b|do(A = a), C = c)P(C = c|do(A = a)) \\ &= \sum_c P(B = b|A = a, C = c)P(C = c|do(A = a)) \\ &= \sum_c P(B = b|A = a, C = c)P(C = c) \end{aligned} \tag{11}$$

Eq. 11 is equivalent to Eq. 10, Q.E.D. □

Therefore, we can eliminate confounding effects of C to estimating the causality from A to B by Eq.10 and Eq.11 based on the Backdoor Criterion, which proves Proposition 3.1.

C. Application of Backdoor Criteria in BIQA

As shown in Figure 1, we construct the structural causal graph for BIQA task, where image contents C and distortion type D constitute the confounders $S = \{(c_i, d_j)\}$ that interferes with the estimation of causality between distorted images I_d and MOS labels m consistent with human ratings. Since the confounders in BIQA task are both observable or discretizable variables, the causal relationship from A to B can be estimated by Backdoor Criteria, which is derived as:

$$\begin{aligned} &P\left(m \mid do\left(I_d\right)\right) \\ &= \sum_{(c_i,d_j)\in S} P\left(m \mid I_d, c_i, d_j\right) P\left(c_i, d_j\right) \\ &= \sum_{i=1}^{q}\sum_{j=1}^{p} P\left(m \mid I_d, c_i, d_j\right) P\left(C = c_i\right) P\left(D = d_j\right) \end{aligned} \tag{12}$$

According to (Li et al., 2018), distortion level is another kind of distortion information. However, we have not considered it as a confounding factor in Eq. 12 with two objective reasons: One is that distortion level is strongly correlated with MOS, and the distributions of distortion levels in synthetically distorted image datasets follow a uniform distribution, which are not consistent with the MOS distributions of images in the real-world (i.e. normal distribution). Another reason is that the scale of the sample subset obtained by Counterfactual Division for training will be too small, leading to performance degradation.

To verify the above statements, we compared the experimental results of CIP and CIS with the Causal-IQA methods treating distortion level as an extra confounding factor, the serial and parallel versions of which are named as CIS+L and CIP+L. There are five distortion levels in both TID2013 and KADID-10K, and the results are displayed in Table 7 and Table 8 with the experimental setup same as that in Section 4.4.

Table 7. Ablation study of test results on wild images without fine-tuning using models trained on TID2013 and KADID-10K.

METHODS	CID2013		LIVE-C		KONIQ-10K	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
CIS+L	0.634	0.646	0.653	0.629	0.642	0.637
CIP+L	0.628	0.650	0.638	0.621	0.659	0.655
CIS	0.672	0.651	**0.690**	**0.684**	0.662	0.658
CIP	**0.712**	**0.683**	0.655	0.624	**0.681**	**0.704**

Table 8. Ablation study of fine-tuning results on authentically distorted images with models trained on TID2013 and KADID-10K.

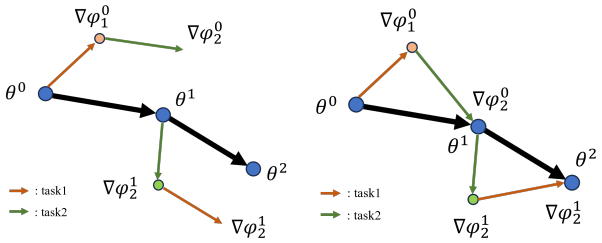
METHODS	CID2013		LIVE-C		KONIQ-10K	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
CIS+L	0.782	0.786	0.801	0.773	0.832	0.816
CIP+L	0.757	0.762	0.789	0.768	0.836	0.805
CIS	**0.887**	**0.895**	**0.847**	**0.828**	**0.918**	**0.881**
CIP	0.873	0.894	0.844	0.823	0.896	0.865

According to Table 7 and Table 8, we can observe that the performances of CIS+L and CIP+L have significantly decreased compared to CIP and CIS, which proves the correctness of our inferences.

D. A Proof of Proposition 3.4

Here, we give the proof of Proposition 3.4:

$$
\begin{aligned}
\theta^* &= \arg\min_{\theta} \mathbb{E}_{(I_d,m)\in\mathcal{X}} \Bigg\{ \frac{1}{pq} \sum_{(c_i,d_j)\in S} \Bigg[\mathcal{L}\big(f_\theta(I_d), m\big) \\
&- \alpha\nabla_\theta \sum_r \frac{w_{r,j}}{w_{:,j}} \mathcal{L}\big(f_\theta(I_{s_{i,j}}(r)), m(r)\big) \nabla_\theta \mathcal{L}\big(f_\theta(I_d), m\big) \\
&+ o\big(\nabla_\theta \mathcal{L}\big(f_\theta(I_d), m\big)\big) \Bigg] \Bigg\} \\
&= \arg\min_{\theta} \mathbb{E}_{(I_d,m)\in\mathcal{X}} \Bigg\{ \mathcal{L}\big(f_\theta(I_d), m\big) \\
&- \frac{1}{pq} \sum_{(c_i,d_i)\in S} \alpha \Bigg[\nabla_\theta \sum_r \frac{w_{r,j}}{w_{:,j}} \mathcal{L}\big(f_\theta(I_{s_{i,j}}(r)), m(r)\big) \Bigg] \\
&\times \nabla_\theta \mathcal{L}\big(f_\theta(I_d), m\big) + o\big(\nabla_\theta \mathcal{L}\big(f_\theta(I_d), m\big)\big) \Bigg\} \\
&= \arg\min_{\theta} \mathbb{E}_{(I_d,m)\in\mathcal{X}} \Bigg\{ \mathcal{L}\big(f_\theta(I_d), m\big) \\
&- \alpha\nabla_\theta \Bigg[\sum_{(c_i,d_i)\in S} \frac{1}{pq} \sum_r \frac{w_{r,j}}{w_{:,j}} \mathcal{L}\big(f_\theta(I_{s_{i,j}}(r)), m(r)\big) \Bigg] \\
&\times \nabla_\theta \mathcal{L}\big(f_\theta(I_d), m\big) + o\big(\nabla_\theta \mathcal{L}\big(f_\theta(I_d), m\big)\big) \Bigg\}
\end{aligned}
\tag{13}
$$



(a) Serial paradigm (CIS) (b) Parallel paradigm (CIP)

Figure 7. The comparison of serial training paradigm and parallel training paradigm.

Eq. 13 is derived by Talyor expansion for Eq. 6 at at position θ. Then we conduct the Taylor inverse expansion for Eq. 13, which can be derived as:

$$
\begin{aligned}
\theta^* = \arg\min_{\theta} \mathbb{E}_{(I_d,m)\in\mathcal{X}} \Bigg[\\
\mathcal{L}\Bigg(f_{\left(\theta - \alpha\nabla_\theta \sum_{(c_i,d_j)\in S} \frac{1}{pq} \mathcal{L}\left(f_\theta(I_{s_{i,j}}), m\right)\right)}(I_d), m \Bigg) \Bigg]
\end{aligned}
\tag{14}
$$

where $\mathcal{L}\big(f_\theta(I_{s_{i,j}}), m\big)\big) = \sum_r \frac{w_{r,j}}{w_{:,j}} \mathcal{L}\big(f_\theta(I_{s_{i,j}}(r)), m(r)\big)$. Let $\phi_{\bar{s}} = \theta - \alpha\nabla_\theta \sum_{(c_i,d_j)\in S} \frac{1}{pq} \mathcal{L}\big(f_\theta(I_{s_{i,j}}), m\big)$, Eq. 14 can be rewritten as:

$$
\begin{aligned}
&\theta^* = \arg\min_{\theta} \mathbb{E}_{(I_d,m)\in\mathcal{X}} \left[\mathcal{L}\left(f_{\phi_{\bar{s}}}\left(I_d\right), m \right) \right] \\
&\text{and } \phi_{\bar{s}} = \theta - \alpha\nabla_\theta \sum_{s_{i,j}\in S} \frac{1}{pq} \mathcal{L}\left(f_\theta\left(I_{s_{i,j}}\right), m \right)
\end{aligned}
\tag{15}
$$

which is the same as Eq. 8.

E. Serial Paradigm VS. Parallel Paradigm

The intuitionistic comparison between serial training paradigm and parallel training paradigm are shown in Figure 7, where we suppose there are two sample subsets by Counterfactual Division. For the sake of simplicity, the black arrow means the direction of meta updating, and $\nabla\phi_i^j$ denotes the $(j+1)$-st iteration with i-th subset. Therefore, we can find that the meta parameters are updated through the gradient direction on the last subtask (thus the direction of the black arrow is parallel to the second colored arrow in Figure 7(a)) in the Algorithm 1 (CIS) training with serial paradigm . In addition, the meta parameters are updated through the sum of the gradient direction vectors on all the subtasks (thus the direction of the black arrow is parallel to the vector sum of all the colored arrows in Figure 7(b)) in the Algorithm 2 (CIS) training with parallel paradigm.

Table 9. Impact of different backbones on the results of CIP.

Backbones		ResNet18	ResNet34	ResNet50
CID2013	PLCC	0.873	0.896	**0.914**
	SROCC	0.894	0.903	**0.916**
LIVE-C	PLCC	0.844	0.873	**0.889**
	SROCC	0.823	0.851	0.875
KonIQ-10K	PLCC	0.896	0.909	**0.917**
	SROCC	0.865	0.881	**0.896**

Table 10. Ablation study of test results on wild images without fine-tuning using models trained on TID2013 and KADID-10K.

Methods	CID2013		LIVE-C		KonIQ-10K	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
CIS+M	0.674	**0.686**	**0.693**	**0.689**	0.684	0.687
CIP+M	**0.727**	0.681	0.666	0.643	**0.702**	**0.715**
CIS	0.672	0.651	0.690	0.684	0.662	0.658
CIP	0.712	0.683	0.655	0.624	0.681	0.704

F. A Proof for Proposition 3.2 and Corollary 3.3

In causal learning, counterfactual refers to situations or events that are contrary to the facts. Specifically, it means considering what the results would have been if certain conditions or factors had been different from what actually occurred for events that have already taken place. Therefore, the formal definition of counterfactual can be given as follows:

Definition F.1. Counterfactual: In the context of the causal model M under the environment $U = u$, consider any two variables X and Y. Let M_x denote the modified version of M when X is set to x. The counterfactual $Y_x(u)$ can be formally defined as: $Y_x(u) = Y_{M_x}(u)$. In other words, the counterfactual $Y_x(u)$ in the model M is defined as the solution for Y in the modified submodel M_x.

According to (Pearl et al., 2016; Pearl, 2009), the calculating of counterfactuals follows three steps:

- Abduction (Step 1): Determine the value of U by evidence e .
- Action (Step 2): Remove the structural equations for the variables X to modify the model M. Then, set the X as $X = x$ to obtain the modified M_x.
- Prediction (Step 3): Use the M_x and $U = u$ to compute the value of Y (i.e., the results of the counterfactual).

Thus, conterfactual is actually answering the question: *In the situation U, what $Y_{X=x(U)}$ would be if X is x*. Therefore, in the BIQA task, we implement Counterfactual Division in Section 3.2 to construct our sample subsets by answering the question: *if S is (c_i, d_j), what the distorted image I_d would be with the distribution of distortion and the corresponding ideal image*[2] *invariant?*. Since the image content and distortion type are available in the synthetically distorted image datasets, we call this data partitioning procedure Counterfactual Division.

Table 11. Ablation study of fine-tuning results on authentically distorted images with models trained on TID2013 and KADID-10K.

Methods	CID2013		LIVE-C		KonIQ-10K	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
CIS+M	**0.898**	0.896	**0.871**	**0.857**	**0.922**	**0.891**
CIP+M	0.889	**0.906**	0.868	0.850	0.913	0.879
CIS	0.887	0.895	0.847	0.828	0.918	0.881
CIP	0.873	0.894	0.844	0.823	0.896	0.865

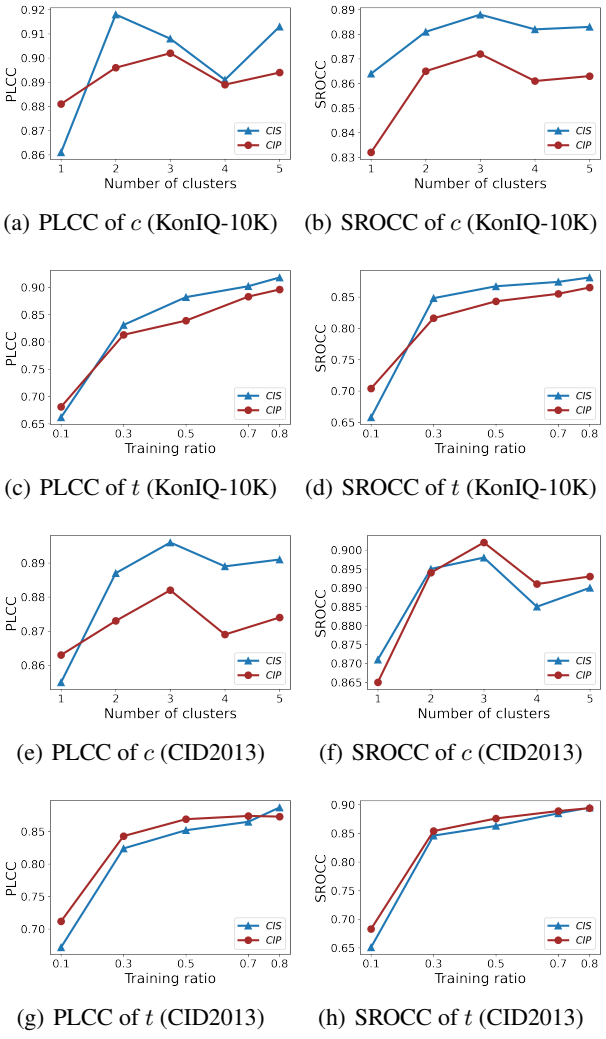
Therefore, we conduct Counterfactual Division through the following process:

- Abduction (Step 1): Use the distorted image I_d to determine the value of I_v, that is $P(I_v|I_d)$.
- Action (Step 2): Modify the degradation model g, so that S is adjusted to the counterfactual value (c_i, d_j).
- Prediction (Step 3): Obtain the consequence $I_{s_{i,j}}$ of the counterfactual based on estimated I_v and modified degradation model $g_{s_{i,j}}$.

In fact, since it is available to high-quality reference images in synthetically distorted image datasets, which can be regarded as the results of degeneration process only based on content, we can ignore the first step and simplify Step 3 as: Obtain the consequence $I_{s_{i,j}}$ of the counterfactual based on corresponding reference images with the i-th content and modified degradation model g_{d_j}.

In Proposition 3.2, counterfactual m_{i_d} means the potential assessment result m predicated on image i_d with fixed S. And node I_d can be regarded as the a result of degradation from ideal image I_v. From a counterfactual perspective, the variable S satisfies the Backdoor Criterion on the original model M in the structural causal graph of BIQA tasks in Figure 1. Hence, it can block the paths from I_d to m_{i_d} on the modified model M_x, as well as the paths from I_d to variables that have influence on m_{i_d} (there is actually no node from I_d to m in Figure 1). Consequently, I_d and m_{i_d} are conditionally independent for any $S = s$. In Causal Inference, the Eq. 2 in Proposition 3.2 is called the Counterfactual Interpretation of Backdoor (Pearl et al., 2016). Then we can use the correction formula Eq. 16 to calculate counterfactuals.

[2]Note that the ideal image (termed as I_v) is virtual and does not represent the reference image. And we denote the virtual degeneration process as $I_d = g(I_v, s)$



(a) PLCC of c (KonIQ-10K) (b) SROCC of c (KonIQ-10K)

(c) PLCC of t (KonIQ-10K) (d) SROCC of t (KonIQ-10K)

(e) PLCC of c (CID2013) (f) SROCC of c (CID2013)

(g) PLCC of t (CID2013) (h) SROCC of t (CID2013)

Figure 8. Impact of different cluster numbers c (during training) and training ratios t (during fine-tuning) on KonIQ-10K (a,b,c,d) and CID2013 (e,f,g,h).

In Corollary 3.3, we have the following derivation:

$$
\begin{aligned}
& P\left(m_{i_d}=y\right) \\
&= \sum_{s=(c_i,d_j)} P\left(m_{i_d}=y \mid S=s\right) P(s) \qquad ① \\
&= \sum_{s=(c_i,d_j)} P\left(m_{i_d}=y \mid S=s, I_d=i_d\right) P(s) \qquad ② \\
&= \sum_{s=(c_i,d_j)} P(m_{i_d}=y \mid S=s, I_d=i_d) P(s) \qquad ③
\end{aligned}
\tag{16}
$$

where ①, ②, ③ are derived from Total Probability Formula, Proposition 3.2 and the principle of consistency, respectively. This verifies that $P\left(m_{i_d}=y\right)$ is another way of saying $P(m_{i_d}=y \mid do(i_d))$ (Pearl et al., 2016).

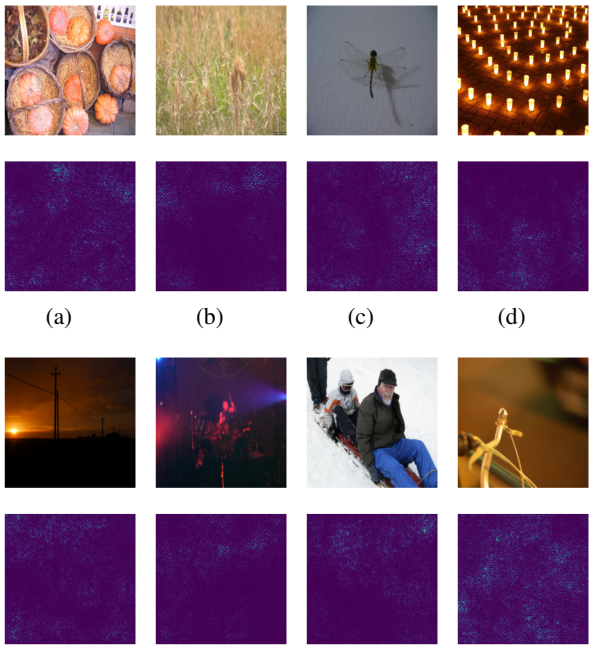


Figure 9. The gradient maps of some authentically distorted images in KonIQ-10K database.

G. Other Experimental Results

G.1. Impact of Different Backbones on CIP

In addition to experiments of exploration on the impact of different backbones on CIS, we make comparisons among the fine-tuned performances of Causal-IQA trained with different backbones on TID2013 and KADID-10K, including ResNet18, ResNet34 and ResNet50 (He et al., 2016). According to Table 9, the generalization performances of CIP get better as the IQA network gets deeper, which is the same as the conclusion of CIS.

G.2. Performances of Training with Distortion-mixed Images

In Section 4, our quality prior Causal-IQA model was trained on the TID2013 and KADID10K datasets. In Table 1, each image sample in both the two synthetically distorted datasets only contains single distortions, while real-world distorted images may involve combinations of multiple distortion types. Therefore, we investigated the performances of training on an additional distortion-mixed image dataset MLIVE[3] (Jayaraman et al., 2012) based on

[3]MLIVE involves 37 subjects and 8880 human judgments on 15 pristine reference images and 405 multiply distorted images of two types, including "Gaussian blur + Gaussian noise" and "Gaussian blur +JPEG compression". DMOS score of MLIVE ranges from 0 to 100.

Table 12. SROCC results comparison in leave-one-distortion-out cross validation on the synthetically distorted image database TID2013.

	DIST. TYPE	BLIINDS-II	BRISQUE	ILNIQE	CORNIA	HOSA	WADIQAM-NR	METAIQA	CIS	CIP
TID2013	AGN	0.7984	0.9356	0.8760	0.4465	0.7582	0.9080	0.9473	**0.9503**	0.9364
	ANC	0.8454	0.8114	0.8159	0.1020	0.4670	0.8700	0.9240	**0.9327**	0.9287
	SCN	0.6477	0.5457	0.9233	0.6697	0.6246	0.8802	0.9534	**0.9677**	0.9594
	MN	0.2045	0.5852	0.5120	0.6096	0.5125	0.8065	0.7277	0.7308	**0.7566**
	HFN	0.7590	0.8965	0.8685	0.8402	0.8285	0.9314	**0.9518**	0.9432	0.9381
	IN	0.5061	0.6559	0.7551	0.3526	0.1889	0.8779	0.8653	**0.9274**	0.8971
	QN	0.3086	0.6555	0.8730	0.3723	0.4145	**0.8541**	0.7454	0.8307	0.8130
	GB	0.9069	0.8656	0.8142	0.8879	0.7823	0.7520	**0.9767**	0.9458	0.9466
	DEN	0.7642	0.6143	0.7500	0.6475	0.5436	0.7680	0.9383	0.9402	**0.9508**
	JPEG	0.7951	0.5186	0.8349	0.8295	0.8318	0.7841	0.9340	**0.9395**	0.9182
	JP2K	0.8221	0.7592	0.8578	0.8611	0.5097	0.8706	0.9586	**0.9593**	0.9647
	JGTE	0.4509	0.5604	0.2827	0.7282	0.4494	0.5191	0.9297	**0.9380**	0.9325
	J2TE	0.7281	0.7003	0.5248	0.4817	0.1405	0.4322	0.9034	0.9112	**0.9133**
	NEPN	0.1219	0.3111	-0.0805	0.3571	0.2163	0.1230	0.7238	**0.8355**	0.7968
	BLOCK	0.2789	0.2659	-0.1357	0.2345	0.3767	0.4059	0.3899	0.4631	**0.5652**
	MS	0.0970	0.1852	0.1845	0.1775	0.0633	0.4596	0.4016	0.5422	**0.6387**
	CTC	0.3125	0.0182	0.0141	0.2122	0.0466	0.5401	0.7637	0.8032	**0.8386**
	CCS	0.0480	0.2142	-0.1628	0.2299	-0.1390	0.5640	0.8294	0.7921	**0.8409**
	MGN	0.7641	0.8777	0.6932	0.4931	0.5491	0.8810	0.9392	**0.9508**	0.9274
	CN	0.0870	0.4706	0.3599	0.5069	0.3740	0.6466	0.9516	**0.9529**	0.9439
	LCNI	0.4480	0.8238	0.8287	0.7191	0.5053	0.6882	**0.9779**	0.9533	0.9386
	ICQD	0.7953	0.4883	0.7487	0.7757	0.8036	0.7965	0.8597	**0.9712**	0.9461
	CHA	0.5417	0.7470	0.6793	0.6937	0.6657	0.7950	0.9269	0.9350	**0.9488**
	SSR	0.7416	0.7727	0.8650	0.8867	0.8273	0.8220	**0.9744**	0.9715	0.9697
	AVERAGE	0.5322	0.5950	0.5701	0.5465	0.4725	0.7073	0.8539	0.8786	**0.8837**

training on TID2013 and KADID10K. The experimental setup is the same as that in Table 4 and Table 5. The models trained under the serial paradigm and parallel paradigm are referred to as CIS+M and CIP+M, respectively. The corresponding results are presented in Table 10 and Table 11, from which we can observe that: CIS+M and CIP+M perform better than CIS and CIP in both scenarios of prediction with fine-tuning and prediction without fine-tuning on unseen image datasets in the wild. This may be due to the addition of more distortion types and sample sets to the training, which to some extent explains the effectiveness of our proposed paradigms in Algorithm 1 and Algorithm 2.

G.3. Impact of Different Cluster Numbers c and Training Ratios t on CID2013 and KonIQ-10K

We compare the experimental results on CID2013 and KonIQ-10K with different partition ratio t and number c of clustering in GMM. The experimental setup is the same as that on LIVE-C: CIS and CIP models are trained on TID2013 and KADID10K, then we compare the fine-tuning results on CID2013 and KonIQ-10K when c changed in $[1, 5]$ and t changed in $\{0.1, 0.3, 0.5, 0.7, 0.8\}$. The trend results are shown in Figure 8, from which we can conclude that: **(i)**. 2 is optimal for c, as increasing c leads to larger computational overhead without improving performance. **(ii)**. As training ratio t increases, the performance improves.

G.4. Visual Analysis for Quality Prior Model

In this part, we conducted a visual experiment to illustrate the efficacy of our Causal-IQA, which is implemented by visualizing MOS specific saliency maps. Specifically, we utilized a CNN visualization code[4] to depict gradient maps at the pixel level under various distortions. The quality prior model was trained on distortion-specific images from the TID2013 and KADID-10K databases. Subsequently, 8 severely distorted images from KonIQ-10K were randomly selected for the visualization experiment. Figure 9 displays both the images and their corresponding gradient maps. The results in Figure 9 are obtained by the model trained based on CIS. We can observe that: the gradient maps successfully capture the precise locations of authentic distortions in the images, such as overexposure in Figure 9(a), blur in Figure 9(b) and underexposure in Figure 9(c).

G.5. Leave-one-distortion-out Cross Validation on TID2013

To further evaluate the generalization of our proposed Causal-IQA to unfamiliar distortion types, we evaluate our approach by employing Leave-One-Distortion-Out (Zhu et al., 2020) cross-validation on TID2013. The experimental setup is the same as that in Table 3. The SORCC values on TID2013 of our methods are recorded in Table 12 for comparison with seven SOTA general-purpose algorithms.

From Table 3 and Table 12, we can observe that: **(i)**. Both MetaIQA and CausalIQA show significant superiority over other approaches, demonstrating the generalization of the meta-learning-based training paradigm. **(ii)**. The average SORCC values of our methods are two to four points higher than that of MetaIQA, which proves the enough generalization of Causal-IQA on unseen distortion types.

[4]The code is available from https://github.com/sar-gupta/convisualize_nb.