This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3178874, IEEE Transactions on Pattern Analysis and Machine Intelligence

Continual Learning for Blind Image Quality Assessment

Weixia Zhang, *Member, IEEE,* Dingquan Li,Chao Ma, *Member, IEEE,* Guangtao Zhai, *Senior Member, IEEE,* Xiaokang Yang, *Fellow, IEEE,* and Kede Ma, *Member, IEEE*

Abstract—The explosive growth of image data facilitates the fast development of image processing and computer vision methods for emerging visual applications, meanwhile introducing novel distortions to the processed images. This poses a grand challenge to existing blind image quality assessment (BIQA) models, which are weak at adapting to subpopulation shift. Recent work suggests training BIQA methods on the combination of all available human-rated IQA datasets. However, this type of approach is not scalable to a large number of datasets, and is cumbersome to incorporate a newly created dataset as well. In this paper, we formulate continual learning for BIQA, where a model learns continually from a stream of IQA datasets, building on what was learned from previously seen data. We first identify five desiderata in the continual setting with three criteria to quantify the prediction accuracy, the plasticity, and the stability, respectively. We then propose a simple yet effective continual learning method for BIQA. Specifically, based on a shared backbone network, we add a prediction head for a new dataset, and enforce a regularizer to allow all prediction heads to evolve with new data while being resistant to catastrophic forgetting of old data. We compute the overall quality score by a weighted summation of predictions from all heads. Extensive experiments demonstrate the promise of the proposed continual learning method in comparison to standard training techniques for BIQA, with and without experience replay.

Index Terms—Blind image quality assessment, continual learning, subpopulation shift

1 Introduction

Aiming to automatically quantify human perception of image quality, blind image quality assessment (BIQA) [1] has experienced an impressive series of successes due in part to the creation of human-rated image quality datasets over the years. For example, the LIVE dataset [2] marks the switch from distortion-specific [3] to general-purpose BIQA [4], [5]. The CSIQ dataset [6] enables cross-dataset comparison. The TID2013 dataset [7] and its successor KADID-10K [8] expose the difficulty of BIQA methods in generalizing to different distortion types. The Waterloo Exploration Database [9] tests model robustness to diverse content variations of natural scenes. The LIVE Challenge Database [10] probes the synthetic-to-real generalization, which is further evaluated by the KonIQ-10K [11] and SPAQ [12] datasets. Assuming that the input domain $\mathcal{X}$ of BIQA is the space of all possible images, each IQA dataset inevitably represents a tiny *subpopulation* of $\mathcal{X}$ (see Fig. 1). That is, BIQA models are bound to encounter subpopulation shift when deployed in the real world. It is thus of enormous value to build robust BIQA models to subpopulation shift.

Previous work [4], [5], [13], [14] on BIQA mainly focuses on boosting performance within subpopulations, while few efforts have been dedicated to testing and improving model robustness to subpopulation shift. Mittal *et al*. [15] aimed ambitiously for *universal* BIQA by measuring a probabilistic distance between patches extracted from natural undistorted images and those from the test "distorted" image. The resulting NIQE only works for a limited set of distortions. Zhang *et al*. [16] modified NIQE by adding more expressive statistical features with marginal improvement.

A straightforward adaptation to subpopulation shift is to fine-tune model parameters with new data, which has been extensively practiced by the BIQA methods based on deep neural networks (DNNs). However, new learning may destroy performance on old data, a phenomenon known as *catastrophic forgetting* [17]. Recently, Zhang *et al*. [18], [19] proposed a dataset combination trick for training BIQA models against catastrophic forgetting. Despite demonstrated robustness to subpopulation shift, this type of method may suffer from three limitations. First, it is not scalable to handle a large number of datasets because of the computation and storage constraints. Second, it is inconvenient to accommodate a new dataset since training samples from all datasets are required for joint fine-tuning. Third, some datasets may not be accessible after a period of time (*e.g.*, due to privacy issues [20]), preventing naïve dataset combination.

In this paper, we take steps towards assessing and improving the robustness of BIQA models to subpopulation shift in a *continual learning* setting. The basic idea is that a BIQA model learns continually from a stream of IQA datasets, integrating new knowledge from the current dataset (*i.e.*, plasticity) while preventing the forgetting of acquired knowledge from previously seen datasets (*i.e.*, stability). To make continual learning for BIQA feasible, nontrivial, and practical, we identify five desiderata: 1) common perceptual scale, 2) robust to subpopulation shift, 3) limited direct access to previous data, 4) no test-time

- *W. Zhang, C. Ma, G. Zhai, and X. Yang are with the MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China. E-mail: {zwx8981, chaoma, zhaiguangtao, xkyang}@sjtu.edu.cn.*
- *D. Li is with Peng Cheng Laboratory, Shenzhen, China. E-mail: lidq01@pcl.ac.cn.*
- *K. Ma is with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong. E-mail: kede.ma@cityu.edu.hk.*

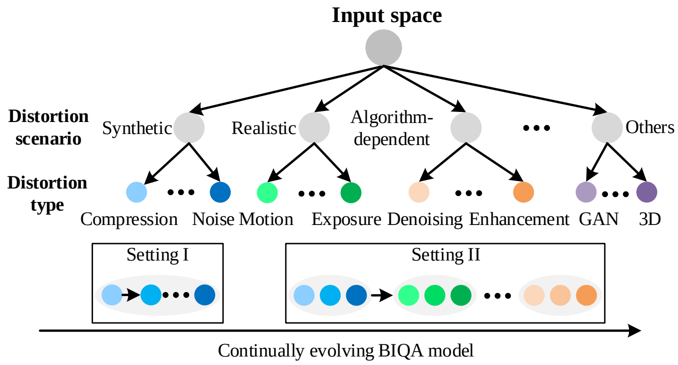


Fig. 1. Illustration of the continual learning paradigm for BIQA. Subpopulation shift exists across distortion types and scenarios. In Setting I, a BIQA model continually evolves from one distortion type to another within the same distortion scenario. In Setting II, a BIQA model continually evolves with varying distortion scenarios.

oracle, and 5) bounded memory footprint. Furthermore, we describe a simple yet effective continual learning method for robust BIQA to subpopulation shift. Specifically, based on a shared and continually-updated backbone network, we add a quality prediction head for each new dataset as a way of promoting plasticity for learning new knowledge. Consolidation of previous knowledge is implemented by stabilizing predictions of previous heads. We summarize the current training dataset using K-means in feature space, and use the learned centroids to compute weights for final quality prediction.

In summary, our main contributions are threefold.

- We establish the continual learning paradigm for BIQA, in which model robustness to subpopulation shift can be evaluated more directly and practically.
- We propose a computational method for continually learning BIQA models, which significantly outperforms standard training techniques for BIQA, with and without experience replay [21].
- We conduct extensive experiments to test various aspects of the proposed method, including plasticity, stability, accuracy, and order-robustness.

2 Related Work

In this section, we give an overview of representative IQA datasets as different subpopulations from the image space $\mathcal{X}$. We then discuss the progress of BIQA driven by the construction of IQA datasets (see Table 1). Finally, we review continual learning in a broader context.

2.1 IQA Datasets

Hamid *et al.* [2] conducted the first "large-scale" subjective user study of perceptual image quality. The resulting LIVE dataset [2] includes 779 distorted images with five synthetic distortion types at five to eight levels. Single stimulus continuous quality rating (SS-CQR) was adopted to collect the mean opinion scores (MOSs). In 2010, Larson *et al.* [6] released the CSIQ dataset, covering 866 images with six synthetic distortions at three to five levels, among which four types are shared by LIVE. A form of the multiple stimulus method was used for subjective testing, where a set of images were linearly displaced according to their perceived quality. The horizontal distance between every pair of images reflected the perceptual difference. In 2011, Ciancio *et al.* [22] built the BID dataset, including mostly blurry images due to camera and/or object motion during acquisition. The same subjective method as in LIVE was adopted to acquire human quality annotations. In 2013, Ponomarenko *et al.* [7] extended the TID2008 dataset to TID2013 with 3,000 images distorted by 25 types at five levels. Paired comparison with a Swiss-system tournament was implemented to reduce subjective cost. In 2016, Ghadiyaram and Bovik [10] created the LIVE Challenge Database with $1,162$ images, undergoing complex realistic distortions. They designed an online crowdsourcing system to gather MOSs using the SS-CQR method. In 2017, Ma *et al.* [9] complied the Waterloo Exploration Database, aiming to probe model generalization to image content variations. No subjective testing was conducted. Instead, the authors proposed three rational tests, namely, the pristine/distorted image discriminability test (D-Test), the listwise ranking consistency test (L-Test), and the pairwise preference consistency test (P-Test) to evaluate IQA methods in a more economic manner. From 2018 to 2019, two large-scale datasets, KADID-10K [8] and KonIQ-10K [11], were made publicly available, which significantly expand the number of synthetically and realistically distorted images, respectively. MOSs of the two datasets were sourced on crowdsourcing platforms using single stimulus absolute category rating. In 2020, Fang *et al.* [12] constructed the SPAQ dataset for perceptual quality assessment of smartphone photography. Apart from MOSs, EXIF data, image attributes, and scene category labels were also recorded to facilitate the development of BIQA models for real-world applications. Concurrently, Ying *et al.* [23] built a large dataset that contains patch quality annotations.

As discussed previously, different datasets may use different subjective procedures, leading to different perceptual scales of the collected MOSs. Even if two datasets happen to use the same subjective method, their MOSs may not be comparable due to differences in the purposes of the studies and the visual stimuli of interest. In Sections 3 and 4, we will give a careful treatment of this subtlety in continual learning for BIQA.

2.2 BIQA Models

In the pre-dataset era, the research in BIQA dealt with specific distortion types, such as JPEG compression [3] and JPEG2000 compression [24]. Since the inception of the LIVE dataset, general-purpose BIQA began to be popular. Many early methods relied on natural scene statistics (NSS) extracted from either spatial domain [4], [15] or transform domain [25], [26]. The underlying assumption is that a measure of the destruction of statistical regularities of natural images [27] provides a reasonable approximation to perceived visual quality. Another line of work explored unsupervised feature learning for BIQA [5], [28]. Since the introduction of the LIVE Challenge Database, synthetic-to-real generalization of BIQA models has received much attention. Ghadiyaram and Bovik [29] handcrafted a bag of statistical features specifically for authentic camera distortions. As the number of images in the newly released

TABLE 1
Summary of IQA datasets used in our experiments. CLIVE stands for the LIVE Challenge Database. SS: Single stimulus. DS: Double stimulus. MS: Multiple stimulus. CQR: Continuous quality rating. ACR: Absolute category rating. CS: Crowdsourcing

Dataset	# of Images	# of Training Pairs	# of Test Images	Scenario	# of Types	Testing Methodology	Year
LIVE [2]	779	7,000	163	Synthetic	5	SS-CQR	2006
CSIQ [6]	866	8,000	173	Synthetic	6	MS-CQR	2010
BID [22]	586	10,000	117	Realistic	N.A.	SS-CQR	2011
CLIVE [10]	1,162	20,000	232	Realistic	N.A.	SS-CQR-CS	2016
KonIQ-10K [11]	10,073	95,000	2,015	Realistic	N.A.	SS-ACR-CS	2018
KADID-10K [8]	10,125	95,000	2,000	Synthetic	25	DS-ACR-CS	2019

IQA datasets became larger, deep learning came into play and began to dominate the field of BIQA. Many strategies were proposed to compensate for the lack of human-labeled data, including patchwise training [13], [30], transfer learning [31], and quality-aware pre-training [14], [32], [33], [34], [35]. To confront the synthetic-to-real challenge (and vice versa), Zhang *et al.* [18], [19] proposed a computational method of training BIQA models on multiple datasets. Latest interesting BIQA studies include active learning for improved generalizability [36], meta-learning for fast adaptation [37], patch-to-picture mapping for local quality prediction [23], loss normalization for accelerated convergence [38], and adaptive convolution for content-aware quality prediction [39].

2.3 Continual Learning

Human learning is a complex and incremental process that continues throughout the life span. While humans may forget the learned knowledge, they forget it gradually rather than catastrophically [40]. However, this is not the case for machine learning models such as DNNs, which tend to completely forget old concepts once new learning starts [17]. A plethora of continual learning methods have been proposed, mainly in the field of image classification. Li and Hoiem [41] proposed learning without forgetting (LwF), which uses model predictions of previous tasks as pseudo labels in a knowledge distillation framework [42]. Based on LwF, Rannon *et al.* [43] attempted to alleviate domain shift among tasks in the learned latent space. Aljundi *et al.* [44] introduced a set of gating autoencoders for automatically feeding a sample to the relevant expert network at the inference time. Another family of methods identifies and penalizes changes to important parameters with respect to previous tasks when learning new tasks. Representative work includes elastic weight consolidation [45] and its online variant [46], incremental moment matching [47], variational continual learning [48], synaptic intelligence [49], and memory-aware synapses [50]. Masse *et al.* [51] proposed context-dependent gating as a complementary module to weight consolidation [45], [49].

With the increasing length of task sequence, soft regularization techniques may not suffice to constrain the model parameters in feasible regions. A plausible solution is parameter isolation [52] as a form of hard regularization, which allows growing branches to accommodate new tasks [53] or masking learned parameters for previous tasks [54], [55], [56]. While parameter isolation effectively prevents catastrophic forgetting, it requires the task oracle to activate the corresponding branch or mask during inference.

Experience replay methods are data-level continual learning solutions, which store old samples or generate pseudo-samples with generative models [52]. As a simple baseline, experience replay [57] has been combined with the reservoir sampling [58]. Rebuffi *et al.* [21] developed a class incremental learner, iCaRL, which stores a subset of exemplars per class for representation replay. During inference, iCaRL calculates the mean of each class in the learned feature space, and performs the nearest-mean-of-exemplars classification. Lopez-Paz *et al.* [59] proposed gradient episodic memory (GEM), which was improved by Chaudhry *et al.* [60] in terms of efficiency. Aljundi [61] proposed to replay samples whose predictions will be most negatively impacted by the foreseen parameter updates (*i.e.*, the most interfered). Most recently, Prabhu *et al.* [62] pointed out the caveats in the progress of continual learning for classification. They proposed a naïve method that greedily stores samples in the memory buffer, based on which a model is trained to achieve state-of-the-art performance.

It is important to note that the recent success of continual learning for image classification may not transfer in a straightforward way to BIQA. This motivates us to establish a continual learning paradigm for BIQA, identifying desiderata to make it feasible, nontrivial, and practical. We also contribute to effective and robust continual learning methods for training BIQA models.

3 A Continual Learning Paradigm for BIQA

In this section, we formulate continual learning for BIQA with five desiderata and three evaluation criteria to quantify the prediction accuracy, the plasticity, and the stability, respectively.

3.1 Problem Definition

We define the learning on a new IQA dataset as a new task in our continual learning setting. When training on the t-th dataset $\mathcal{D}_t$, we assume no direct access to $\{\mathcal{D}_k\}_{k=1}^{t-1}$ for the moment, leading to the following training objective:

$$\mathcal{L}(\mathcal{D}_t; w) = \frac{1}{|\mathcal{D}_t|}\sum_{(x,q)\in\mathcal{D}_t}\ell(g_w(x), q), \tag{1}$$

where $x \in \mathbb{R}^N$ and $q \in \mathbb{R}$ denote the N-dimensional "distorted" image and the corresponding MOS, respectively. g_w represents a computational BIQA model parameterized by a vector w. $\ell(\cdot)$ is the objective function, quantifying the quality prediction performance. One may add a regularizer $r(w)$ to Eq. (1) with the goal of gaining resistance to catastrophic forgetting. For evaluation, we may measure the performance of g_w on the hold-out test sets of all tasks seen so far:

$$\sum_{k=1}^{t}\mathcal{L}(\mathcal{V}_k;w)=\sum_{k=1}^{t}\left(\frac{1}{|\mathcal{V}_k|}\sum_{(x,q)\in\mathcal{V}_k}\ell(g_w(x),q)\right), \quad (2)$$

where $\mathcal{V}_k$ is the test set for the k-th task. An ideal BIQA model should perform well on new tasks, and endeavor to mitigate catastrophic forgetting of old tasks, resulting in a low objective value in Eq. (2).

3.2 Five Desiderata

Considering the differences between image classification and BIQA, we argue that careful treatment should be given to make continual learning for BIQA feasible, nontrivial, and practical. Towards this, we list five desiderata.

I **Common Perceptual Scale**. This requires that the MOSs of IQA datasets should admit a common perceptual scale. In other words, there exists a *monotonic* function for each dataset to map its MOSs to this common scale. Otherwise, learning a single g_w on multiple datasets continually is conceptually infeasible. Desideratum I excludes human-rated datasets that measure perceptual quantities closely related to image quality (*e.g.*, visual contrast [64] and scene visibility [63], [65]). To highlight this point, we show some images from the deraining quality assessment dataset [63] in Fig. 2, with a smaller MOS indicating severer rain density. It is clear that rain density is not *monotonically* correlated with visual quality. Therefore, this dataset violates Desideratum I, and should be excluded from the task sequence for BIQA.

II **Robust to Subpopulation Shift**. It is empirically proven that existing BIQA models generalize reasonably to distortions with similar visual appearances (*e.g.*, from additive noise to multiplicative noise; from LIVE [2] to CSIQ [6]). However, when datasets exhibit apparent subpopulation shift (*e.g.*, synthetic to realistic distortions), the generalization of BIQA models remains particularly weak. Although the stream of training data may be distorted in arbitrary form (*e.g.*, with one distortion type only), it is highly desirable to develop continual learning methods that are robust to various levels of (and especially apparent) subpopulation shift.

III **Limited Access to Previous Data**. This is key to the general continual learning setting [52], which makes continual learning continual learning. From the psychology perspective, we humans learn things continually, and thus it is important to develop BIQA models that are consistent with this aspect of human perception and cognition. Experience replay continual learning relies on replaying (at least a small portion of) old data to fight against catastrophic forgetting [21], [59], [60], [66], [67], [68], [69]. In the context of BIQA, Zhang *et al*. [18], [19] proposed to jointly train models on data from all tasks, which can be seen as a performance upper bound [52]. Desideratum III assumes limited access to previous data when training new tasks, meaning that the memory buffer should be carefully controlled within a preset budget. It puts no constraints on the format of old data, which can either be raw images in previous datasets or their feature summaries.

IV **No Test-Time Oracle**. As advocated in [52], [70], a well-designed continual learning method should be independent of the task oracle to make predictions. That is, the method should be unaware of which dataset the test image belongs to. Desideratum IV is imperative in BIQA because if we know in advance the task label, we may be able to train separate and specialized models for each of the datasets, making continual learning for BIQA a trivial task.

V **Bounded Memory Footprint**. The model capacity in the number of model parameters should be relatively fixed, forcing the BIQA method to allocate its capacity wisely to achieve the Pareto optimum between plasticity and stability. Other memory overhead, *e.g.*, to store old data and/or pseudo-labels should also remain bounded, or at least grow very slowly, with respect to the number of tasks seen so far [21].

3.3 Performance Measures

We propose three quantitative criteria to measure 1) the prediction accuracy, 2) the plasticity, and 3) the stability of a BIQA model in the continual learning setting. Conceptually, plasticity and stability refer to the ability of integrating new information and preserving previous knowledge, respectively. Without loss of generality, we use Spearman's rank correlation coefficient (SRCC) between model predictions and MOSs as a measure of prediction monotonicity. Other correlation measures (*e.g.*, Kendall rank correlation coefficient and Pearson linear correlation coefficient) and distance metrics (*e.g.*, mean squared error and mean absolute error) can also be applied.

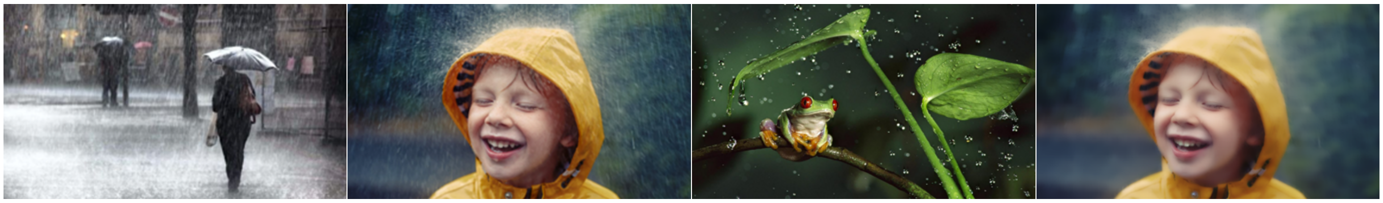
For a BIQA model trained on a sequence of T tasks (denoted by the T-th model), we compute the mean SRCC between model predictions and MOSs of each dataset as a measure of prediction accuracy:

$$\mathrm{mSRCC}=\frac{1}{T}\sum_{k=1}^{T}\mathrm{SRCC}_{Tk}, \quad (3)$$

where SRCC_{tk} is the SRCC result of the t-th model on the k-th dataset ($t,k=1,\cdots,T$). We then define a mean plasticity index (mPI):

$$\mathrm{mPI}=\frac{1}{T}\sum_{t=1}^{T}\mathrm{PI}_t=\frac{1}{T}\sum_{t=1}^{T}\mathrm{SRCC}_{tt}, \quad (4)$$

i.e., the average result of the model on the current dataset along the task sequence. Last, we define a mean stability



(a) MOS = 25 (b) MOS = 39 (c) MOS = 51 (d) MOS = 69

Fig. 2. Images sampled from the deraining quality assessment dataset [63]. A larger MOS in the dataset denotes lighter rain density. It is not hard to observe that rain density is not monotonically correlated with perceived image quality. Therefore, the dataset violates Desideratum I, and should be excluded from the task sequence for BIQA. Images are cropped for improved visibility.

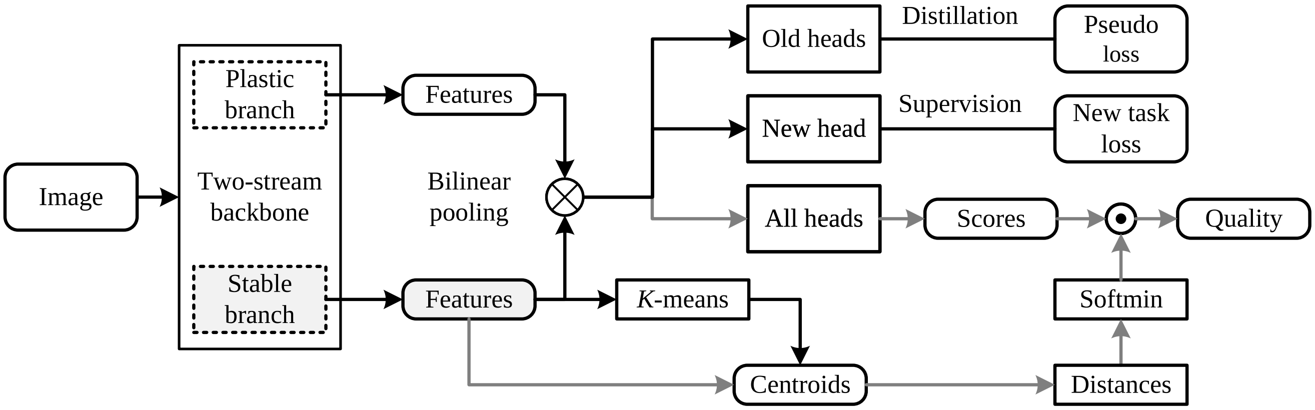


Fig. 3. System diagram of the proposed continual learning method. Black and grey arrows correspond to the training and testing phases, respectively.

index (mSI) by measuring the variability of model performance on old data when new learning has taken place:

$$\mathrm{mSI} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{SI}_t, \tag{5}$$

where

$$\mathrm{SI}_t = \begin{cases} 1 & t = 1 \\ \frac{1}{t-1}\sum_{k=1}^{t-1}\widehat{\mathrm{SRCC}}_{tk} & t > 1 \end{cases}, \tag{6}$$

where $\widehat{\mathrm{SRCC}}_{tk}$ for $k < t$ is computed between the predictions of the t-th and k-th models. It is noteworthy that a higher mSI does not necessarily imply better quality prediction performance on old tasks as no MOS is involved. mSRCC, mPI, and mSI measure different and complementary aspects of a continually learned BIQA model. We may further define a mean plasticity-stability index (mPSI) over a list of T tasks to quantify the plasticity-stability trade-off:

$$\mathrm{mPSI} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{PSI}_t = \frac{1}{2T}\sum_{t=1}^{T}\left(\mathrm{PI}_t + \mathrm{SI}_t\right). \tag{7}$$

4 A Continual Learning Method for BIQA

In this section, we propose a simple yet effective continual learning method for BIQA. The system diagram of our method is shown in Fig. 3.

4.1 Model Estimation

We describe the proposed method with respect to the desiderata stated in Subsection 3.2.

4.1.1 *Learning-to-Rank for BIQA*

According to Desideratum I, it is desirable to work with a common perceptual scale. However, this is difficult because we are only given a stream of T IQA datasets without the monotonic functions to map the associated MOSs. Inspired by [18], [19], we choose to learn a perceptual scale for all tasks by exploiting relative quality information. Specifically, given an image pair (x, y), we compute a binary label:

$$p(x, y) = \begin{cases} 1 & \text{if } q(x) \geq q(y) \\ 0 & \text{otherwise} \end{cases}, \tag{8}$$

where $q(x), q(y) \in \mathbb{R}$ are the MOSs of images x and y, respectively. When learning the t-th task, we transform $\mathcal{D}_t = \{x_t^{(i)}, q_t^{(i)}\}_{i=1}^{|\mathcal{D}_t|}$ to $\mathcal{P}_t = \{(x_t^{(i)}, y_t^{(i)}), p_t^{(i)}\}_{i=1}^{M_t}$, where $M_t \leq \binom{|\mathcal{D}_t|}{2}$.

Our BIQA model consists of a backbone network, $f_\phi : \mathbb{R}^N \mapsto \mathbb{R}^L$, that takes the N-dimensional raw image as input and produces an L-dimensional feature vector. We append a prediction head, $h_{\psi_t}(\cdot)$ parameterized by ψ_t, to compute quality estimates for the t-th task. Under the Thurstone's Case V model, we are able to estimate the probability that x is of higher quality than y by

$$\hat{p}_t(x, y) = \Phi\left(\frac{h_{\psi_t}(f_\phi(x)) - h_{\psi_t}(f_\phi(y))}{\sqrt{2}}\right), \tag{9}$$

where $\Phi(\cdot)$ is the standard Normal cumulative distribution function, and the variance of quality predictions is fixed to one [71]. The full set of parameters, $\{\phi, \psi_1, \psi_2, \ldots, \psi_T\}$, constitute the parameter vector w to be optimized.

For the current t-th task, we measure the statistical distance between the ground-truth and predicted probabilities using the fidelity loss [72], whose advantages over the cross entropy loss have been demonstrated in several BIQA studies [18], [19]:

$$\begin{aligned}\ell_{\text{new}}(x,y;\phi,\psi_t) = 1 &- \sqrt{p(x,y)\hat{p}_t(x,y)} \\ &- \sqrt{(1-p(x,y))(1-\hat{p}_t(x,y))}. \quad (10)\end{aligned}$$

4.1.2 Mitigating Catastrophic Forgetting

Direct optimization of Eq. (10) may cause catastrophic forgetting of old tasks (see Table 4). According to Desideratum III, we consider two cases: 1) no previous data is directly accessed, and the model can only train on new data [52]; 2) there is a preset memory budget for storing a small portion of samples from previous tasks. The preserved old data can be used for experience replay (also called rehearsal) to confront the catastrophic forgetting.

We begin with the simpler replay-free case. Inspired by LwF [41], we add a regularizer to allow forgetting old knowledge gracefully. Before training the t-th task, we use the k-th output head to compute a probability $\bar{p}_{tk}(x,y)$ for each pair of $(x,y) \in \mathcal{P}_t$ according to Eq. (9). This creates $t-1$ datasets $\{\mathcal{P}_{tk}\}_{k=1}^{t-1}$ with pseudo-labels to constrain the updated prediction $\hat{p}_{tk}$ to be close to the recorded prediction p_{tk}. Again, we use the fidelity loss to implement the constraint:

$$\begin{aligned}\ell_{\text{old}}\left(x,y;\phi,\{\psi_k\}_{k=1}^{t-1}\right) = \sum_{k=1}^{t-1}\Bigg(1 &- \sqrt{\bar{p}_{tk}(x,y)\hat{p}_{tk}(x,y)} \\ &-\sqrt{(1-\bar{p}_{tk}(x,y))(1-\hat{p}_{tk}(x,y))}\Bigg). \\ &\qquad (11)\end{aligned}$$

In practice, we randomly sample a mini-batch $\mathcal{B}_t$ from $\mathcal{P}_t$ and use a variant of stochastic gradient descent to minimize the following empirical loss:

$$\begin{aligned}\mathcal{L}\left(\mathcal{B}_t;\phi,\{\psi_k\}_{k=1}^{t}\right) = &\frac{1}{|\mathcal{B}_t|}\sum_{(x,y)\in\mathcal{B}_t}(\ell_{\text{new}}(x,y;\phi,\psi_t) \\ &+\lambda\ell_{\text{old}}\left(x,y;\phi,\{\psi_k\}_{k=1}^{t-1}\right)), \quad (12)\end{aligned}$$

where λ governs the trade-off between the two terms.

There is abundant wisdom in leveraging experience replay to improve continual learning for classification [73]. We will adapt and evaluate representative methods for BIQA in Subsection 5.5.

4.1.3 Network Specification

We adopt a two-stream network as the backbone, which is adapted from DB-CNN [33], a state-of-the-art BIQA model. Our backbone is composed of two branches, a VGG-like CNN and a variant of ResNet-18, with specifications given in Table 2. Following the practice in [33], we pre-train the VGG-like CNN using a large-scale image set with nine synthetic distortion types at two to five degradation levels. This can be formulated as a multiclass classification problem, training the network to discriminate distortion types and levels. It is empirically shown that the learned features through this process are distortion-aware. During continual learning, we fix the VGG-like CNN (denoted as the stable branch in Fig. 3). In contrast, all parameters of the ResNet-18 variant are learnable, denoted as the plastic branch. The input image is fed into both branches, whose outputs are bilinearly pooled to obtain a fixed-length feature vector [74].

TABLE 2
The proposed two-stream network, consisting of a VGG-like CNN [33] and a variant of ResNet-18 [76], for a T-length task sequence. The nonlinear activation, max pooling, and normalization layers are omitted

VGG-like CNN	ResNet-18 Variant
3×3, 48, stride 1	7×7, 64, stride 2
3×3, 48, stride 2	$\begin{bmatrix}3\times3\text{, 64, stride 1}\\3\times3\text{, 64, stride 1}\end{bmatrix}\times2$
3×3, 64, stride 1 3×3, 64, stride 2	$\begin{bmatrix}3\times3\text{, 128, stride 2}\\3\times3\text{, 128, stride 1}\end{bmatrix}\times1$ $\begin{bmatrix}3\times3\text{, 128, stride 1}\\3\times3\text{, 128, stride 1}\end{bmatrix}\times1$
3×3, 64, stride 1 3×3, 64, stride 2	$\begin{bmatrix}3\times3\text{, 256, stride 2}\\3\times3\text{, 256, stride 1}\end{bmatrix}\times1$ $\begin{bmatrix}3\times3\text{, 256, stride 1}\\3\times3\text{, 256, stride 1}\end{bmatrix}\times1$
3×3, 128, stride 1 3×3, 128, stride 1 3×3, 128, stride 2	$\begin{bmatrix}3\times3\text{, 512, stride 2}\\3\times3\text{, 512, stride 1}\end{bmatrix}\times1$ $\begin{bmatrix}3\times3\text{, 512, stride 1}\\3\times3\text{, 512, stride 1}\end{bmatrix}\times1$
Bilinear Pooling	
Full Connection	65,536 $\times T$

$$f_\phi(x) = f_{\phi_p}(x)^T f_{\phi_s}(x), \quad (13)$$

where ϕ_p and ϕ_s are the parameters of the plastic and stable branches, respectively. We append an ℓ_2-normalization layer [75] on top of the backbone network:

$$\tilde{f}_\phi(x) = \frac{f_\phi(x)}{\|f_\phi(x)\|_2} \quad (14)$$

to project the feature representation onto the unit hypersphere. This pushes the predictions of all heads to a similar range, making subsequent computation (*e.g.*, weighted summation of quality scores) more numerically stable.

4.2 Model Inference

During inference, the original LwF for image classification needs the task oracle, which violates Desideratum IV and is not applicable to BIQA. Instead of relying on the task oracle to precisely activate a task-specific prediction head, we design a K-means gating (KG) mechanism to compute a weighted summation of quality estimates from all heads as the overall quality score.

During the t-th task learning, we compute the fixed-length quality representations $\{\tilde{f}_{\phi_s}(x_t^{(i)})\}_{i=1}^{|\mathcal{D}_t|}$ by a feedforward sweep of $\mathcal{D}_t$:

$$\tilde{f}_{\phi_s}(x) = \frac{\text{pool}(f_{\phi_s}(x))}{\|\text{pool}(f_{\phi_s}(x))\|_2}, \quad (15)$$

where $\mathrm{pool}(\cdot)$ denotes global average pooling over spatial locations. Similar to Eq. (14), we normalize the pooled representation to make it more comparable across different tasks. We then summarize $\mathcal{D}_t$ with K centroids $\{c_t^{(j)}\}_{j=1}^{K}$ by applying K-means [77] to $\{\tilde{f}_{\phi_s}(x_t^{(i)})\}_{i=1}^{|\mathcal{D}_t|}$. As the number bits to store K centroids is considerably smaller than storing raw training images, Desideratum III is respected. We use f_{ϕ_s} as the feature extractor to distill $\mathcal{D}_t$ because it is fixed and distortion-aware, which effectively reduces the *task-recency* bias [78]. We measure the perceptual relevance of the test image x to $\mathcal{D}_t$ by computing the minimal Euclidean distance between its feature representation and the K centroids of $\mathcal{D}_t$:

$$d_t(x) = \min_{1\leq j\leq K} \|\tilde{f}_{\phi_s}(x) - c_t^{(j)}\|_2. \tag{16}$$

We then pass $\{d_t(x)\}_{t=1}^{T}$ to a softmin function to compute the weight for the t-th prediction head:

$$a_t(x) = \frac{\exp(-\tau d_t(x))}{\sum_{k=1}^{T}\exp(-\tau d_k(x))}, \tag{17}$$

where $\tau \geq 0$ is a temperature parameter used to tune the smoothness of the softmin function. The final quality score is defined as the inner product between two vectors of relevance weights and quality predictions:

$$\hat{q}(x) = \sum_{t=1}^{T} a_t h_{\psi_t}(f_\phi(x)). \tag{18}$$

A final note is that the number of parameters of the T prediction heads grows slowly compared to that of the backbone network[1]. It is important to note that the proposed KG mechanism is orthogonal to the selection of backbone networks and the number of parameters introduced by a new head can be further optimized on backbone network structure. Meanwhile, the computations of the backbone network are shared across all prediction heads (see Fig. 3), which means that the computational overhead introduced by a new prediction head is negligible. Thus, our BIQA model meets Desideratum V and is rather scalable in terms of the number of training datasets.

5 Experiments

In this section, we describe a realistic and challenging experimental setup for continual learning of BIQA models, which strictly obeys Desiderata I and II. We divide the experiments into two parts according to the direct accessibility to previous data. As the proposed continual learning method is the first of its kind, the performance comparison is done mainly with respect to its variants, some of which can be treated as performance upper bounds.

5.1 Experimental Setup

We select six widely used IQA datasets, including LIVE [2], CSIQ [6], BID [22], LIVE Challenge [10], KonIQ-10K [11], and KADID-10K [8], whose details are summarized in Table 1. We leverage the CORrelation ALignment (CORAL) [79], a widely used non-parametric distance metric to measure the domain similarity, to quantify the dissimilarity between datasets. We define the CORAL loss as the distance between second-order statistics (i.e., covariances) of the VGG-like features of two datasets. As shown in Fig. 4, we find that datasets with synthetic distortions (LIVE, CSIQ, and KADID-10K) are closer to each other compared to those with realistic distortions (and vice versa). We organize these datasets in chronological order, *i.e.*, LIVE → CSIQ → BID → LIVE Challenge → KonIQ-10K → KADID-10K. In Subsection 5.4, we also test on task sequences of different orders to probe the order-robustness of the proposed method. We randomly sample 70% and 10% images from each dataset for training and validation, leaving the remaining for testing. We follow [18], [19] to form image pairs in $\{\mathcal{P}_t\}_{t=1}^{T}$, whose numbers are given in Table 1. To ensure content independence in LIVE, CSIQ, and KADID-10K, we divide the training and test sets according to the reference images. Although in the proposed continual learning setting, test sets of future tasks are assumed to

TABLE 3
Performance comparison in terms of mSRCC, mPI, mSI, and mPSI. All methods are trained in chronological order

Method	mSRCC	mPI	mSI	mPSI
MH-CL-O	0.7672	0.8794	0.9243	0.9019
EWC-O	**0.8529**	0.8754	0.9559	0.9157
SI-O	0.8344	**0.8809**	0.9474	0.9142
MAS-O	0.8354	0.8741	0.9553	0.9147
LwF-O	0.8485	0.8765	**0.9862**	**0.9314**
SL	0.6767	0.8761	0.8045	0.8403
SH-CL	0.7091	0.8778	0.8493	0.8636
MH-CL	0.7079	0.8794	0.8103	0.8449
MH-CL-KG	0.7225	0.8694	0.9086	0.8890
EWC	0.6815	0.8754	0.8132	0.8443
EWC-KG	0.7919	0.8607	0.9418	0.9013
SI	0.7323	**0.8809**	0.8224	0.8517
SI-KG	0.7848	0.8765	0.9315	0.9002
MAS	0.6170	0.8741	0.7974	0.8358
MAS-KG	0.7795	0.8619	0.9433	0.9026
LwF	0.6693	0.8765	0.8331	0.8548
Proposed (LwF-KG)	**0.8150**	0.8563	**0.9796**	**0.9180**

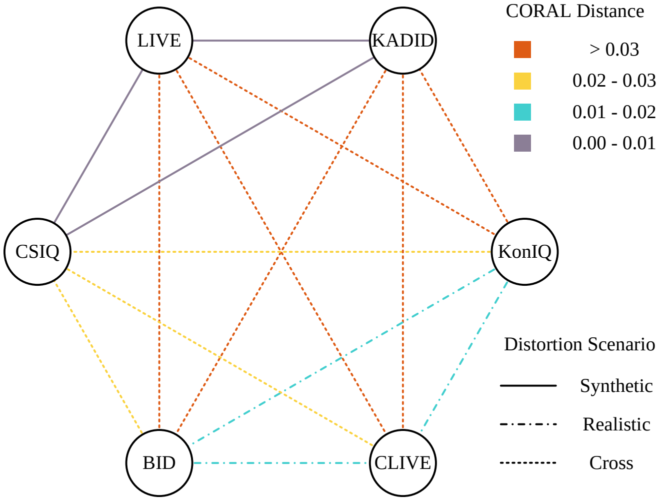


Fig. 4. The pairwise CORAL distances of the six IQA datasets with different distortion scenarios. A larger CORAL value indicates a higher level of dissimilarity between two datasets.

[1] In our implementation, each new prediction head is implemented by a fully connected layer with $65,536$ parameters, accounting for less than 0.6% of the total parameters

TABLE 4
Performance comparison in terms of SRCC between the proposed method and its variants. Best results in each session are highlighted in bold, while results of future tasks are marked in grey

Sequence	Dataset	Method	LIVE [2]	CSIQ [6]	BID [22]	CLIVE [10]	KonIQ-10K [11]	KADID-10K [8]
–	All	JL	0.9602	0.8480	0.8796	0.8305	0.8841	0.8801
1st	LIVE	All	**0.9555**	0.7398	0.5872	0.3988	0.6262	0.5248
2nd	CSIQ	SL	0.9257	0.8801	0.4337	0.3037	0.5951	0.5613
		SH-CL	**0.9560**	0.8783	0.3926	0.2731	0.5482	0.5763
		MH-CL	0.9450	0.8789	0.4141	0.2510	0.5639	0.5460
		MH-CL-KG	0.9473	0.8812	0.4320	0.2729	0.5672	0.5560
		EWC	0.9385	0.8754	0.4705	0.2821	0.5785	0.5578
		EWC-KG	0.9409	0.8773	0.4910	0.3047	0.5818	0.5658
		SI	0.9462	0.8837	0.4326	0.2680	0.5703	0.5498
		SI-KG	0.9497	**0.8866**	0.4615	0.2871	0.5737	0.5589
		MAS	0.9441	0.8808	0.4433	0.2656	0.5704	0.5498
		MAS-KG	0.9456	0.8831	0.4784	0.2880	0.5740	0.5589
		LwF	0.9491	0.8705	0.4668	0.3201	0.5798	0.5551
		Proposed (LwF-KG)	0.9499	0.8848	0.5100	0.3779	0.5904	0.5471
3rd	BID	SL	0.7451	0.7847	0.8433	0.7180	0.6944	0.5158
		SH-CL	0.9088	0.8068	0.8470	0.6165	0.6967	0.5808
		MH-CL	0.8201	0.8186	0.8405	0.6609	0.6958	0.5418
		MH-CL-KG	0.9341	**0.9006**	0.8420	0.5898	0.6687	0.5636
		EWC	0.8416	0.8146	0.8455	0.6593	0.7012	0.5442
		EWC-KG	0.9355	0.8996	0.8455	0.5685	0.6779	0.5690
		SI	0.8464	0.8312	0.8435	0.6496	0.7059	0.5523
		SI-KG	0.9325	0.8968	0.8429	0.5710	0.6803	0.5665
		MAS	0.7830	0.8205	0.8350	0.6625	0.6933	0.5505
		MAS-KG	0.9287	0.9002	0.8351	0.5759	0.6712	0.5781
		LwF	0.8805	0.7931	0.8585	0.6799	0.6973	0.4705
		Proposed (LwF-KG)	**0.9485**	0.8895	**0.8602**	0.4691	0.6426	0.5320
4th	CLIVE	SL	0.7176	0.6055	0.8262	0.8351	0.7573	0.4394
		SH-CL	0.8352	0.6878	0.8356	**0.8517**	0.7758	0.5090
		MH-CL	0.7169	0.6412	0.7735	0.8443	0.7548	0.5005
		MH-CL-KG	0.9504	0.8714	0.7990	0.8194	0.7388	0.5305
		EWC	0.7377	0.6818	0.7905	0.8476	0.7552	0.5079
		EWC-KG	**0.9527**	**0.8886**	0.8261	0.8228	0.7394	0.5414
		SI	0.7133	0.6344	0.7665	0.8470	0.7553	0.5000
		SI-KG	0.9509	0.8702	0.7946	0.8181	0.7387	0.5341
		MAS	0.7450	0.6191	0.8163	0.8464	0.7560	0.5058
		MAS-KG	0.9491	0.8587	0.8349	0.8290	0.7516	0.5379
		LwF	0.7747	0.6025	0.8331	0.8355	0.7460	0.4619
		Proposed (LwF-KG)	0.9337	0.8805	**0.8515**	0.7999	0.7279	0.5185
5th	KonIQ-10K	SL	0.6915	0.6019	0.7434	0.7209	0.8953	0.5578
		SH-CL	0.7294	0.6554	0.7753	0.7155	0.8948	0.5532
		MH-CL	0.7146	0.6013	0.7698	0.7256	0.8936	0.5471
		MH-CL-KG	0.9217	0.7508	0.7720	0.7281	0.8889	0.4929
		EWC	0.6647	0.5914	0.7858	0.7040	0.8905	0.5489
		EWC-KG	**0.9534**	0.8547	0.8374	0.7691	0.8855	0.5231
		SI	0.6933	0.5596	0.7927	0.7412	**0.8960**	0.5447
		SI-KG	0.9357	0.8078	0.8146	0.7618	0.8917	0.5101
		MAS	0.6889	0.5491	0.7760	0.7138	0.8796	0.5634
		MAS-KG	0.9498	**0.8712**	**0.8590**	0.7712	**0.8753**	0.5643
		LwF	0.7324	0.6447	0.7888	0.7055	0.8866	0.5489
		Proposed (LwF-KG)	0.9193	0.8621	0.8485	**0.7851**	0.8752	0.5510
6th	KADID-10K	SL	0.8709	0.7664	0.6554	0.3895	0.5307	0.8473
		SH-CL	0.8577	0.7296	0.7062	0.4548	0.6666	0.8394
		MH-CL	0.8618	0.7402	0.7094	0.4216	0.6455	**0.8634**
		MH-CL-KG	0.8251	0.7333	0.7481	0.5120	0.6871	0.8292
		EWC	0.8584	0.6631	0.6518	0.4528	0.6251	0.8380
		EWC-KG	0.8580	0.7275	**0.8431**	0.6991	0.8463	0.7774
		SI	0.8746	0.7632	0.7418	0.4764	0.6785	0.8594
		SI-KG	0.8575	0.7590	0.7904	0.6588	0.8253	0.8180
		MAS	0.8106	0.6694	0.5300	0.3521	0.4931	0.8469
		MAS-KG	0.8395	0.7334	0.8252	0.6673	0.8185	0.7933
		LwF	0.8548	0.7081	0.5775	0.4379	0.5847	0.8525
		Proposed (LwF-KG)	**0.8996**	**0.7893**	0.8128	**0.7742**	**0.8520**	0.7622

be inaccessible, we consider using them for cross-dataset performance evaluation.

For each task, stochastic optimization is carried out by Adam [80] with $\lambda = 10$ in Eq. (12). The parameters of the tailored ResNet-18 and the prediction heads are initialized by the weights pre-trained on ImageNet [81] and the He's method [82], respectively. We set the initial learning rate to 2×10^{-4} with a decay factor of 10 for every three epochs, and we train our method for nine epochs. A warm-up training strategy is used: only the prediction heads are trained in the first three epochs with a mini-batch size of 128; for the remaining epochs, we fine-tune the entire network with a mini-batch size of 32. During training, we re-scale and crop the images to $384 \times 384 \times 3$, preserving the aspect ratio. During testing, the number of centroids used in K-means is set to $K = 128$ for all tasks. This results in a memory overhead of 64 kilobytes to store the centroids. Empirically, we find that the performance is insensitive to the choice of K. We tune the temperature τ in Eq. (17) on the validation sets and set it to $\tau = 32$. We test on images of original size in all experiments.

5.2 Competing Methods

Separate Learning (SL) is the *de facto* method in BIQA. We train the model with a single prediction head on one of the six training sets by optimizing Eq. (12) with $\lambda = 0$. Although SL is not a continual learning method, we incorporate it as a reference.

Joint Learning (JL) is a recently proposed method [18], [19] to overcome the cross-dataset challenge (as a specific form of subpopulation shift) in BIQA. We train the same model with a single head on the combination of all six training sets by optimizing Eq. (12) with $\lambda = 0$. With full access to all training data, JL serves as the upper bound of all continual learning methods.

Single-Head Continual Learning (SH-CL) is a baseline of the proposed continual learning method, where the same model with a single head is successively trained on $\{\mathcal{P}_t\}_{t=1}^{6}$ by optimizing Eq. (12) with $\lambda = 0$. The difference between SL and SH-CL lies in training the model from scratch for the current task and fine-tuning the model with initialization provided by the previous task.

Multi-Head Continual Learning (MH-CL) is a multi-head extension of SH-CL. MH-CL adds a prediction head for a new task, and optimizes it for Eq. (12) with $\lambda = 0$. It remains to specify the head for final quality prediction. To encourage adaptation to a constantly changing environment, we simply use the latest head to make prediction. MH-CL serves as the baseline for all regularization-based continual learning methods. Meanwhile, we may incorporate the proposed KG mechanism during inference, giving rise to **MH-CL-KG**. Moreover, we leverage the task oracle to precisely activate the corresponding head for prediction, denoted by **MH-CL-O**, which may give the performance upper bound in the multi-head architecture.

Learning without Forgetting (LwF) in BIQA builds upon MH-CL by optimizing Eq. (12) with $\lambda = 10$. In other words, LwF introduces a stability regularizer to preserve the performance of previously seen data. Same as MH-CL, LwF relies on the latest head for quality prediction.

The proposed method (LwF-KG) can be seen as the combination of LwF and KG. We also explore the task oracle to select the corresponding head for quality prediction, denoted by **LwF-O**.

Parameter Importance Regularization follows a similar paradigm that penalizes the changes to the estimated ”important” parameters for previous tasks when learning a new task. Specifically, we implement three such regularizers - elastic weight consolidation (EWC) [45], synaptic intelligence (SI) [49], and memory aware synapses (MAS) [50]. Similar to LwF, all three methods are built upon the MH-CL baseline with the KG mechanism, denoted by **EWC-KG**, **SI-KG**, and **MAS-KG**, respectively. The task oracle can also be leveraged for quality prediction, denoted by **EWC-O**, **SI-O**, and **MAS-O**, respectively.

5.3 Main Results

In this subsection, we present the quantitative and qualitative results in the case that no previous data is directly accessible when learning new tasks.

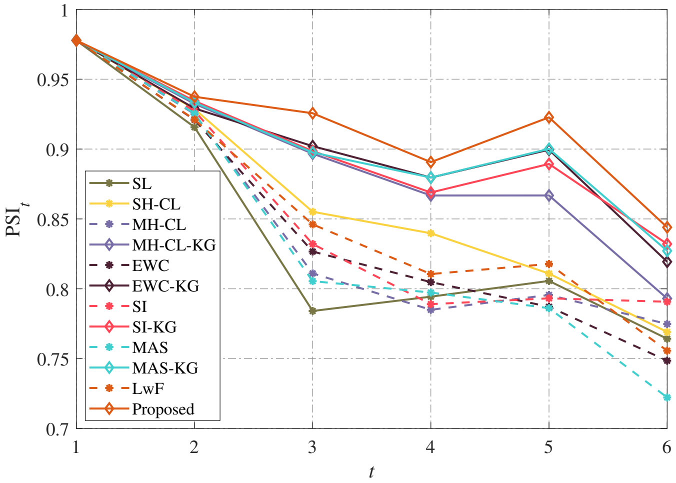


Fig. 5. PSI_t as a function of the task index t.

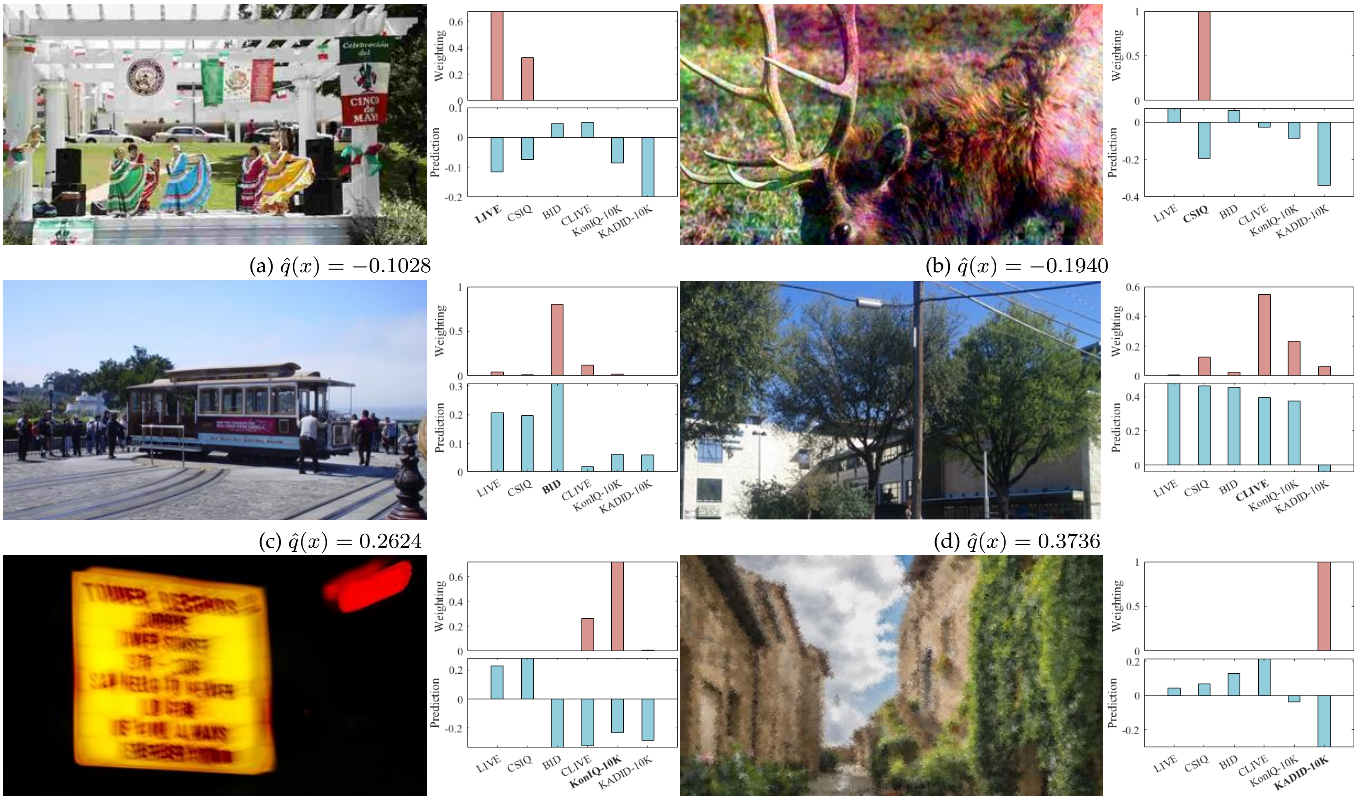
TABLE 5
Performance comparison of different BIQA models. All methods are trained in chronological order

Methods	mSRCC	mPI	mSI	mPSI
NIQE [15]	0.4515	0.4515	1.0000	0.7258
dipIQ [83]	0.3633	0.3633	1.0000	0.6817
Ma19 [34]	0.5664	0.5664	1.0000	0.7832
BRISQUE [4]	0.5235	0.6929	0.6149	0.6539
CORNIA [5]	0.4948	0.6455	0.7069	0.6762
DBCNN [33]	0.6957	0.8814	0.7883	0.8349
MetaIQA [37]	0.6841	0.8535	0.7886	0.8211
KonCept512 [11]	0.6951	0.8680	0.7853	0.8267
Proposed	0.8150	0.8563	0.9796	0.9180

5.3.1 *Quantitative Results*

We use the proposed mSRCC, mPI, mSI, and mPSI in Eq. (3), Eq. (4), Eq. (5), and Eq. (7) to benchmark the performance. From Table 3, we have several interesting observations. First, the unsatisfactory performance of SL calls for continual learning methods to mitigate catastrophic forgetting in BIQA. Second, SH-CL and MH-CL achieve similar mPI result with SL, but improves the mSI and mSRCC results upon SL by clear margins, indicating a vanilla knowledge accumulation process. Third, we see that the proposed KG mechanism leads to consistent performance gains over the baseline regularization methods. Among them, LwF-KG outperforms other regularizers in terms of mSRCC and mPSI, indicating a better plasticity-stability trade-off. Fourth, when the test-time oracle is available, the performance of all methods can be further improved. We also plot PSI_t as a function of the task index t in Fig. 5, from which we find that our method is more stable, and performs much better as the length of the task sequence increases. It is noteworthy that the fluctuation of PSI_t is mainly due to the difficulty of the newly incorporated dataset and the level of subpopulation shift with respect to previous ones.

We take a closer look at the performance variations along the task sequence, and summarize the SRCC results continually in Table 4. Note that all methods begin training on LIVE [2], and their SRCC results are the same before continually learning on any new task. There are several useful findings. First, we observe that subpopulation shift



(a) $\hat{q}(x) = -0.1028$

(b) $\hat{q}(x) = -0.1940$

(c) $\hat{q}(x) = 0.2624$

(d) $\hat{q}(x) = 0.3736$

(e) $\hat{q}(x) = -0.2540$

(f) $\hat{q}(x) = -0.3009$

Fig. 6. Perceptual scaling of images sampled from the six IQA datasets. The bar charts of weights and quality predictions of all heads are also presented alongside each image. The final quality prediction $\hat{q}(x)$ is shown in the subcaption. Zoom in for better distortion visibility.

between different tasks significantly oscillates the results of SL. This is not surprising because it is often challenging for BIQA models trained on datasets of synthetic distortions to perform well on datasets of realistic distortions (and vice versa) [31], [33]. Second, compared with SL, SH-CL generally improves on old tasks with similar performance on new tasks, achieving a better plasticity-stability trade-off. Third, MH-CL and other regularization methods add a prediction head for each new task, which does not handle old tasks well, necessitating an effective mechanism to make full use of all learned heads. Fourth, regularization methods employing the KG module lead to better performance especially on previous tasks. Compared with other regularization methods, LwF-KG delivers more stable performance. Last, all methods are upper bounded by JL as expected.

We find that in the continual learning setting the generalization of BIQA models is heavily influenced by the level of subpopulation shift between the seen and unseen datasets. For example, good generalization in terms of SRCC on the unseen KonIQ-10K (with realistic distortions) is achieved by the proposed method when it has been learned on BID and CLIVE with realistic distortions compared to the one trained on LIVE and CSIQ with synthetic distortions only. We also observe that the proposed method generalizes marginally to KADID-10K when it is continually trained from the 1st to the 5th dataset. This is because KADID-10K [8] contains some distinct distortion types that are not shared by previously trained datasets. In the future, we will improve the generalizability aspect of continually learned BIQA models using advanced machine learning techniques such as domain generalization [84], which is an orthogonal and complementary direction to the current work.

We also evaluate eight representative BIQA methods under the same task sequence in chronological order, including three opinion-unaware methods [15], [34], [83] that do not rely on human-rated IQA datasets for training and five opinion-aware methods [4], [5], [11], [33], [37] that learn from human perceptual scores. As summarized in Table 5, we have several useful observations. First, NIQE [15], dipIQ [83], and Ma19 [34] that do not learn continually deliver the maximal stability, but fail to adapt to changing distortion scenarios, resulting in poor plasticity and overall prediction accuracy. Second, none of the opinion-aware BIQA models is immune to catastrophic forgetting caused by the significant subpopulation shift among tasks. Third, DNN-based models that enjoy joint optimization of feature representation and quality prediction generally exhibit better plasticity than those based on hand-engineered natural scene statistics (NSS).

5.3.2 Qualitative Results

We conduct a qualitative analysis of our LwF-KG model by sampling test images from the task sequence. Also shown in Fig. 6 are the bar charts of weights and quality predictions corresponding to each image. Although the proposed method is not jointly trained on all IQA datasets [18], [19], it successfully learns a common perceptual scale for all tasks, within which images are well perceptually aligned. Moreover, visual inspections of the bar charts reveal that the prediction head may only give accurate quality estimates for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3178874, IEEE Transactions on Pattern Analysis and Machine Intelligence

the dataset it is exposed to. Fortunately, given a test image, the proposed KG mechanism is able to compensate for the prediction inaccuracy, assigning larger weights to the heads trained on images with similar distortions. For example, when evaluating the images sampled from LIVE [2] (see Fig. 6 (a)), the head trained on CSIQ [6] of similar synthetic distortions is also assigned relatively high weighting. As another example, KADID-10K [8] contains some distinct distortion types (*e.g.*, spatial jitter in Fig. 6 (f)); consequently, the assigned weight for the head of KADID-10K tends to dominate. In the open visual world, if a new task is difficult enough to fail the KG mechanism, we may leverage new data to improve the feature representation of the KG mechanism in the future.

5.4 Ablation Study

In this subsection, we conduct a series of ablation experiments to evaluate the robustness of our method to different task orders and alternative design choices.

5.4.1 Order-Robustness

The main experiments are conducted on the task sequence in chronological order. In real-world situations, new distortions may emerge in arbitrary order, and similar distortions may also reappear in the future. Thus a BIQA model is expected to be independent of the task order [85]. To evaluate the order-robustness of the proposed method, we experiment with four extra task orders: (I) reverse chronological order - KADID-10K → KonIQ-10K → LIVE Challenge → BID → CSIQ → LIVE; (II) synthetic and realistic distortions in alternation - LIVE → BID → CSIQ → LIVE Challenge → KADID-10K → KonIQ-10K; (III) synthetic distortions followed by realistic distortions - LIVE → CSIQ → KADID-10K → BID → LIVE Challenge → KonIQ-10K; and (IV) realistic distortions followed by synthetic distortions - BID → LIVE Challenge → KonIQ-10K → LIVE → CSIQ → KADID-10K. We list the mSRCC, mPI, mSI, and mPSI results in Table 6, where we find our method is quite robust to handle task sequences of different orders, providing justifications for its use in real-world applications. Nevertheless the reverse chronological order (Order I) achieves lower mPI, mSI, and mPSI results compared to the other task orders. We believe this is because the harder task (KADID-10K) appears in the beginning of the sequence, making it more difficult for our method to trade off plasticity and stability. Our method offers an SRCC of 0.8474 on KADID-10K [8] when it is first trained on, and fails to stabilize the performance with a final SRCC of 0.7703. The weakest performance in terms of mSRCC is spotted in Order II, which may be due to the frequent switching between tasks of significant subpopulation shifts (*i.e.*, synthetic and realistic distortions).

5.4.2 Gating Mechanism

We compare the proposed KG mechanism with two popular expert gates in classification. The first is the nearest-mean-of-exemplars used in iCaRL [21], which corresponds to setting $K = 1$ of the proposed KG. The only difference is that hard assignment is implemented in [21] for classification, while soft assignment is used in our method for regression. We also implement the expert gate module in [44], which relies on autoencoder-based reconstruction [86]. Our autoencoder consists of two fully-connected layers, one for encoding and the other for decoding, respectively, with ReLU in between. During inference, the weights are computed according to the reconstruction errors through a softmin function. The mSRCC and mPSI results are listed in Table 7, where we see that the proposed KG clearly outperforms the nearest-mean-of-exemplars and the expert gate. We believe this arises because the intra-variance of each dataset in BIQA (containing different distortion types at varying degradation levels) is relatively large, requiring multiple distortion-aware centroids for task summarization.

TABLE 6
Performance comparison for different task orders. I: Reverse chronological order. II: Synthetic and realistic distortions in alternation. III: Synthetic distortions followed by realistic distortions. IV: Realistic distortions followed by synthetic distortions. V: Default chronological order in bold

Order	mSRCC	mPI	mSI	mPSI
I	0.8164	0.8446	0.9521	0.8984
II	0.8049	0.8495	0.9664	0.9080
III	0.8289	0.8617	0.9640	0.9129
IV	0.8078	0.8526	0.9633	0.9080
V	0.8150	0.8563	0.9796	0.9180

TABLE 7
Performance comparison in terms of mSRCC and mPSI of different gating strategies

Gating Strategy	mSRCC	mPSI
Nearest-Mean-of-Exemplars	0.8081	0.9046
Expert Gate	0.8090	0.9061
Without training-time KG (Proposed)	**0.8150**	**0.9180**
With training-time KG	**0.8159**	**0.9175**

The forward computation of the network is different between training and inference time when the KG mechanism is used. We conduct another set of experiments to see whether incorporation of the KG mechanism during training brings additional performance gains. As shown in Table 7, we observe similar performance in terms of mSRCC and mPSI with and without training-time KG mechanism, while keeping other training procedures exactly the same.

We also take the index of the maximum weight in KG as the prediction of the dataset the test image belongs to. We compare the confusion matrices of the pre-trained ResNet-18 and the VGG-like CNN in Table 8. It is clear that the VGG-like distortion-aware representation achieves much higher gating accuracies, which is directly reflected in quality prediction improvements.

5.4.3 Feature Representation

We compare the adopted two-stream network with two variants: a single-stream ResNet-18 network with 1) global average pooling and 2) bilinear pooling as feature extractors, respectively, denoted by **Single Stream-GAP** and **Single Stream-BP**. Note that both variants also use the VGG-like CNN to perform KG during inference, yet its features are ablated from the quality representation. From Table 9, we observe that Single Stream-BP consistently outperforms Single Stream-GAP for all metrics. The two-stream back-

TABLE 8
Confusion matrices of the KG mechanism with different feature representations

Feature	Pre-trained ResNet-18					
Dataset	LIVE	CSIQ	BID	CLIVE	KonIQ	KADID
LIVE	0.5313	0.0813	0.0000	0.2438	0.0750	0.0688
CSIQ	0.0523	0.2791	0.0349	0.1453	0.3953	0.0930
BID	0.0000	0.0085	0.6496	0.1795	0.1624	0.0000
CLIVE	0.0043	0.0089	0.1202	0.7210	0.1459	0.0000
KonIQ	0.0020	0.0010	0.0581	0.0407	0.8918	0.0065
KADID	0.0520	0.0355	0.0280	0.2750	0.2175	0.3920
Feature	**VGG-like CNN**					
Dataset	LIVE	CSIQ	BID	CLIVE	KonIQ	KADID
LIVE	**0.7563**	0.1438	0.0063	0.0125	0.0063	0.0750
CSIQ	0.1919	**0.5988**	0.0000	0.0814	0.0116	0.1163
BID	0.0085	0.0000	**0.9231**	0.0256	0.0427	0.0000
CLIVE	0.0215	0.0258	0.0410	**0.8884**	0.0300	0.0129
KonIQ	0.0015	0.0074	0.0134	0.0362	**0.9236**	0.0179
KADID	0.0545	0.0700	0.0165	0.0160	0.0365	**0.8065**

TABLE 9
Performance comparison of different feature representations

Feature Representation	mSRCC	mPI	mSI	mPSI
Single Stream-GAP	0.7804	0.7776	0.9514	0.8645
Single Stream-BP	0.8092	0.8267	**0.9825**	0.9046
Two Stream-BP	**0.8150**	**0.8563**	0.9796	**0.9180**

bone further boosts the model plasticity, leading to higher mSRCC and mPSI results.

5.5 Further Results based on Experience Replay

In this subsection, we evaluate different continual learning methods based on experience replay for BIQA, where previous data is partially accessible.

5.5.1 Memory Management

The very first step in experience replay is to manage the memory budget. Inspired by [62], we assume all tasks and all data in each task to be equally important. We greedily update a memory buffer $\mathcal{M}$ to accommodate new data while keeping a balanced task distribution, as presented in Algorithm 1.

Algorithm 1 Memory Management

```
Input: ${\mathcal{D}_t}_{t=1}^T = \{\{x_t^{(i)}, \mu_t^{(i)}\}_{i=1}^{|\mathcal{D}_t|}\}_{t=1}^T$
Require: Memory buffer: $\mathcal{M}$, memory budget: $B$, task length: $T$, and $\mathrm{RandSample}(\mathcal{D}, b)$: randomly sample $b$ images from $\mathcal{D}$
$\mathcal{M} \leftarrow \emptyset$
$b = \min(B, |\mathcal{D}_1|)$
$\mathcal{M}_1 = \mathrm{RandSample}(\mathcal{D}_1, b)$
$\mathcal{M} = \mathcal{M} \bigcup \mathcal{M}_1$
for $t = 2, \ldots, T$ do
    $\mathcal{M} \leftarrow \emptyset$
    $m = \lfloor \frac{B}{t} \rfloor$    ▷ Divide the memory budget
    for $k = 1, \ldots, t-1$ do
        $b = \min(m, |\mathcal{M}_k|)$
        $\mathcal{M}_k \leftarrow \mathcal{M}_k \setminus \mathrm{RandSample}(\mathcal{M}_k, |\mathcal{M}_k| - b)$
    $b = \min(m, |\mathcal{D}_t|)$
    $\mathcal{M}_t \leftarrow \mathrm{RandSample}(\mathcal{D}_t, b)$
    $\mathcal{M} = \bigcup_{k=1}^t \mathcal{M}_k$
end for
```

5.5.2 Competing Methods with Experience Replay

Under the pairwise learning-to-rank framework, we need to transform $\mathcal{M}$ to $\mathcal{P}_{\mathcal{M}}$ following the procedure described in Subsubsection 4.1.1. We consider five experience replay methods.

SH-CL with Experience Replay (**SH-CL-ER**) is built on SH-CL that trains a single-head model using mini-batches of images sampled from both $\mathcal{P}_t$ and $\mathcal{P}_{\mathcal{M}}$.

MH-CL with Experience Replay (**MH-CL-ER**) updates each old head using the corresponding images from $\mathcal{P}_{\mathcal{M}}$ and the current head using $\mathcal{P}_t$, respectively. KG is used for quality prediction.

iCaRL-v1 is a direct adaptation of iCaRL [21] in image classification. Driven by knowledge distillation, iCaRL-v1 uses the predictions of the k-th head to compute a probability $\bar{p}_{tk}(x, y)$ as the pseudo label for $(x, y) \in \mathcal{M}_k$, and trains the model using the fidelity loss. The main difference between iCaRL-v1 and LwF is that they distill knowledge using image samples from the memory buffer $\mathcal{M}$ and the current dataset $\mathcal{D}_t$, respectively.

iCaRL-v2 is an advanced adaptation of iCaRL to BIQA. During learning from the t-th task, it performs the same as iCaRL-v1 for all previous heads, and optimizes the current head using the combination of $\mathcal{P}_t$ and $\mathcal{P}_{\mathcal{M}}$ (as in SH-CL-ER). By doing so, the current head is also exposed to a portion of data from previous tasks, aiming for more accurate predictions. This motivates us to compute the final quality score by first aggregating predictions from old tasks with KG and then averaging it with the prediction of the current head.

GDumb is a simple baseline [62] that trains a model with the memory buffer $\mathcal{M}$ only. Although GDumb is not specifically designed for continual learning, we include it in our experiments because GDumb performs competitively against many continual learning methods in image classification [73].

5.5.3 Results and Analysis

We compare the performance of different experience replay schemes in Table 10, where we organize the tasks in chronological order. The key observation is that maintaining a memory buffer with a reasonable budget (*e.g.*, B=1,000) leads to noticeable performance improvements over the reference model LwF-KG, especially under mSRCC. Nevertheless, small memory budgets (*e.g.*, B=100) may hurt the prediction accuracy and the plasticity-stability trade-off due to the danger of over-fitting $\mathcal{M}$. In other words, if the memory budget is extremely limited, replay-free continual learning methods (*e.g.*, the proposed LwF-KG) may be more preferable. Incorporating the learning strategy of SH-CL-ER, iCaRL-v2 shows stronger plasticity than iCaRL-v1, while the latter exhibits stronger stability. For B =1,000, iCaRL-v2 is the best performer in terms of both mSRCC and mPSI. Despite the remarkable performance in image classification, GDumb appears much more data-hungry.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3178874, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 10
Performance comparison in terms of mSRCC, mPI, mSI, and mPSI of different experience replay methods with varying memory budgets. Results of the proposed LwF-KG are listed as reference

Memory Budget	100	500	1,000	2,000	5,000	100	500	1,000	2,000	5,000
Measure			mSRCC					mPI		
LwF-KG			0.8150					0.8563		
SH-CL-ER	**0.8249**	**0.8601**	0.8687	0.8734	0.8736	**0.8795**	**0.8844**	**0.8801**	**0.8802**	**0.8793**
MH-CL-ER	0.7970	0.8256	0.8571	0.8572	0.8658	0.8713	0.8717	0.8730	0.8744	0.8759
iCaRL-v1	0.8021	0.8369	0.8534	0.8603	0.8704	0.8662	0.8543	0.8640	0.8529	0.8721
iCaRL-v2	0.7998	0.8452	**0.8712**	**0.8746**	**0.8844**	0.8729	0.8560	0.8672	0.8660	0.8775
GDumb	0.5669	0.7823	0.7978	0.8188	0.8445	0.6971	0.7883	0.8226	0.8507	0.8676
Memory Budget	100	500	1,000	2,000	5,000	100	500	1,000	2,000	5,000
Measure			mSI					mPSI		
LwF-KG			0.9796					0.9180		
SH-CL-ER	0.9282	0.9487	0.9445	0.9454	0.9459	0.9036	0.9166	0.9123	0.9128	0.9125
MH-CL-ER	0.9420	0.9648	0.9686	0.9742	**0.9753**	0.9067	**0.9183**	0.9208	**0.9243**	**0.9256**
iCaRL-v1	**0.9548**	**0.9743**	**0.9810**	**0.9789**	0.9735	**0.9105**	0.9143	0.9225	0.9159	0.9228
iCaRL-v2	0.9451	0.9697	0.9747	0.9783	0.9638	0.9090	0.9129	**0.9230**	0.9222	0.9207
GDumb	0.9226	0.9586	0.9538	0.9474	0.9517	0.8099	0.8735	0.8882	0.8991	0.9097

6 Conclusion

We have formulated continual learning for BIQA with five desiderata and three performance measures. We also contributed continual learning methods to train BIQA models robust to subpopulation shift in this new setting.

This work establishes a new research direction in BIQA with many important topics left unexplored. First, it remains wide open whether we need to add or remove several desiderata to make continual learning for BIQA more practical. For example, it may be useful to add the online learning desideratum, where learning happens instantaneously with no distinct boundaries between tasks (or datasets). Second, better continual learning methods for BIQA are desirable to bridge the performance gap between the current method and the upper bound by joint learning. Third, with access to partial data of previous tasks, better experience replay strategies would be valuable directions to pursue. Fourth, the current work only considers two distortion scenarios, *i.e.*, synthetic and realistic distortions, to construct the task sequence. In the future, it would be interesting to incorporate multiple distortion scenarios, representing more subpopulation shift during training and testing. Last, the current work only explores small-length task sequences with a limited number of task orders. It is necessary to test the current method on task sequences with arbitrary length and in arbitrary order. It is also important to develop more order-robust and length-robust continual learning methods for BIQA.

Acknowledgment

The authors would like to sincerely thank the three anonymous reviewers in the first round for their highest-quality comments, which significantly improve the manuscript. The authors would also like to thank Xuelin Liu for helping illustrate the idea of subpopulation shift in Fig. 1.

References

[1] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. Morgan & Claypool, 2006.

[2] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, “A statistical evaluation of recent full reference image quality assessment algorithms,” *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[3] Z. Wang, H. R. Sheikh, and A. C. Bovik, “No-reference perceptual quality assessment of JPEG compressed images,” in *IEEE International Conference on Image Processing*, vol. 1, 2002, pp. 477–480.

[4] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[5] P. Ye, J. Kumar, L. Kang, and D. Doermann, “Unsupervised feature learning framework for no-reference image quality assessment,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.

[6] E. C. Larson and D. M. Chandler, “Most apparent distortion: Full-reference image quality assessment and the role of strategy,” *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 1–21, Jan. 2010.

[7] P. Nikolay, J. Lina, I. Oleg, L. Vladimir, E. Karen, A. Jaakko, V. Benoit, C. Kacem, C. Marco, B. Federica, and C.-C. J. Kuo, “Image database TID2013: Peculiarities, results and perspectives,” *Signal Processing: Image Communication*, vol. 30, pp. 57–77, Jan. 2015.

[8] H. Lin, V. Hosu, and D. Saupe, “KADID-10k: A large-scale artificially distorted IQA database,” in *International Conference on Quality of Multimedia Experience*, 2019, pp. 1–3.

[9] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, “Waterloo Exploration Database: New challenges for image quality assessment models,” *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 1004–1016, Feb. 2017.

[10] D. Ghadiyaram and A. C. Bovik, “Massive online crowdsourced study of subjective and objective picture quality,” *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, Jan. 2016.

[11] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, “KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment,” *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, Jan. 2020.

[12] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, “Perceptual quality assessment of smartphone photography,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3677–3686.

[13] S. Bosse, D. Maniry, K. R. Müller, T. Wiegand, and W. Samek, “Deep neural networks for no-reference and full-reference image quality assessment,” *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[14] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, “End-to-end blind image quality assessment using deep neural networks,” *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[15] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a ‘completely blind’ image quality analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[16] L. Zhang, L. Zhang, and A. C. Bovik, “A feature-enriched completely blind image quality evaluator,” *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[17] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of Learning and Motivation*, 1989, vol. 24, pp. 109–165.

[18] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Learning to blindly assess image quality in the laboratory and wild," in *IEEE International Conference on Image Processing*, 2020, pp. 111–115.

[19] ——, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, Mar. 2021.

[20] R. Aljundi, "Continual learning in neural networks," *arXiv preprint arXiv:1910.02718*, 2019.

[21] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[22] A. Ciancio, A. L. N. T. Targino da Costa, E. A. B. da Silva, A. Said, R. Samadani, and P. Obrador, "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Transactions on Image Processing*, vol. 20, no. 1, pp. 64–75, Jan. 2011.

[23] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3575–3585.

[24] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: application to JPEG2000," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 163–172, Feb. 2004.

[25] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[26] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[27] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, no. 1, pp. 1193–1216, Aug. 2001.

[28] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, Sep. 2016.

[29] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17, no. 1, pp. 32–32, Jan. 2017.

[30] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[31] H. Zeng, L. Zhang, and A. C. Bovik, "Blind image quality assessment with a probabilistic quality representation," in *IEEE International Conference on Image Processing*, 2018, pp. 609–613.

[32] X. Liu, J. v. d. Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *IEEE International Conference on Computer Vision*, 2017, pp. 1040–1049.

[33] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, Jan. 2020.

[34] K. Ma, X. Liu, Y. Fang, and E. P. Simoncelli, "Blind image quality assessment by learning from multiple annotators," in *IEEE International Conference on Imaging Processing*, 2019, pp. 2344–2348.

[35] J. Wu, J. Ma, F. Liang, W. Dong, G. Shi, and W. Lin, "End-to-end blind image quality prediction with cascaded deep neural network," *IEEE Transactions on Image Processing*, vol. 29, pp. 7414–7426, Jun. 2020.

[36] Z. Wang and K. Ma, "Active fine-tuning from gMAD examples improves blind image quality assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear, 2021.

[37] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 131–14 140.

[38] D. Li, T. Jiang, and M. Jiang, "Norm-in-norm loss with faster convergence and better performance for image quality assessment," in *ACM International Conference on Multimedia*, 2020, pp. 789–797.

[39] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3664–3673.

[40] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in Cognitive Sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[41] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, Dec. 2017.

[42] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[43] A. Rannen, R. Aljundi, M. B. Blaschko, and T. Tuytelaars, "Encoder based lifelong learning," in *IEEE International Conference on Computer Vision*, 2017, pp. 1320–1328.

[44] R. Aljundi, P. Chakravarty, and T. Tuytelaars, "Expert gate: Lifelong learning with a network of experts," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3366–3375.

[45] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, Q. John, T. Ramalho, A. Grabska-Barwinska, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017.

[46] J. Schwarz, W. Czarnecki, J. Luketina, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell, "Progress & compress: A scalable framework for continual learning," in *International Conference on Machine Learning*, 2018, pp. 4528–4537.

[47] S.-W. Lee, J.-H. Kim, J. Jun, J.-W. Ha, and B.-T. Zhang, "Overcoming catastrophic forgetting by incremental moment matching," in *Advances in Neural Information Processing Systems*, 2017, pp. 4652–4662.

[48] C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner, "Variational continual learning," in *International Conference on Learning Representations*, 2018, pp. 1–18.

[49] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *International Conference on Machine Learning*, 2017, pp. 3987–3995.

[50] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *European Conference on Computer Vision*, 2018, pp. 139–154.

[51] N. Y. Masse, G. D. Grant, and D. J. Freedman, "Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization," *Proceedings of the National Academy of Sciences*, vol. 115, no. 44, pp. E10 467–E10 475, Oct. 2018.

[52] M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear, 2021.

[53] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.

[54] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra, "PathNet: Evolution channels gradient descent in super neural networks," *arXiv preprint arXiv:1701.08734*, 2017.

[55] A. Mallya and S. Lazebnik, "PackNet: Adding multiple tasks to a single network by iterative pruning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7765–7773.

[56] A. Mallya, D. Davis, and S. Lazebnik, "Piggyback: Adapting a single network to multiple tasks by learning to mask weights," in *European Conference on Computer Vision*, 2018, pp. 67–82.

[57] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne, "Experience replay for continual learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 350–360.

[58] J. S. Vitter, "Random sampling with a reservoir," *ACM Transactions on Mathematical Software*, vol. 11, no. 1, pp. 37–57, Mar. 1985.

[59] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6467–6476.

[60] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with A-GEM," in *International Conference on Learning Representations*, 2019, pp. 1–20.

[61] R. Aljundi, E. Belilovsky, T. Tuytelaars, L. Charlin, M. Caccia, M. Lin, and L. Page-Caccia, "Online continual learning with maximal interfered retrieval," *Advances in Neural Information Processing Systems*, vol. 32, pp. 11 849–11 860, 2019.

[62] A. Prabhu, P. H. Torr, and P. K. Dokania, "Gdumb: A simple approach that questions our progress in continual learning," in *European Conference on Computer Vision*, 2020, pp. 524–540.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3178874, IEEE Transactions on Pattern Analysis and Machine Intelligence

[63] Q. Wu, L. Wang, K. N. Ngan, H. Li, F. Meng, and L. Xu, "Subjective and objective de-raining quality assessment towards authentic rain image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3883–3897, Nov. 2020.

[64] Z. Wang and E. P. Simoncelli, "Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities," *Journal of Vision*, vol. 8, no. 12, pp. 1–13, Sep. 2008.

[65] L. K. Choi, J. You, and A. C. Bovik, "Referenceless prediction of perceptual fog density and perceptual image defogging," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3888–3901, Nov. 2015.

[66] T. L. Hayes, K. Kafle, R. Shrestha, M. Acharya, and C. Kanan, "Remind your neural network to prevent catastrophic forgetting," in *European Conference on Computer Vision*, 2020, pp. 466–483.

[67] L. Pellegrini, G. Graffieti, V. Lomonaco, and D. Maltoni, "Latent replay for real-time continual learning," in *IEEE International Conference on Intelligent Robots and Systems*, 2020, pp. 10 203–10 209.

[68] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 831–839.

[69] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara, "Dark experience for general continual learning: A strong, simple baseline," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 15 920–15 930.

[70] S. Farquhar and Y. Gal, "Towards robust evaluations of continual learning," *arXiv preprint arXiv:1805.09733*, 2018.

[71] L. L. Thurstone, "A law of comparative judgment," *Psychological Review*, vol. 34, pp. 273–286, Jul. 1927.

[72] M.-F. Tsai, T.-Y. Liu, T. Qin, H.-H. Chen, and W.-Y. Ma, "FRank: A ranking method with fidelity loss," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 383–390.

[73] Z. Mai, R. Li, J. Jeong, D. Quispe, H. Kim, and S. Sanner, "Online continual learning in image classification: An empirical survey," *Neurocomputing*, vol. 469, pp. 28–51, Jan. 2022.

[74] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *IEEE International Conference on Computer Vision*, 2015, pp. 1449–1457.

[75] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "NormFace: L_2 hypersphere embedding for face verification," in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 1041–1049.

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[77] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[78] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. van de Weijer, "Class-incremental learning: Survey and performance evaluation on image classification," *arXiv preprint arXiv:2010.15277*, 2020.

[79] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *European Conference on Computer Vision*, 2016, pp. 443–450.

[80] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015, pp. 1–15.

[81] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[82] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[83] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.

[84] J. Wang, C. Lan, C. Liu, Y. Ouyang, and T. Qin, "Generalizing to unseen domains: A survey on domain generalization," in *International Joint Conference on Artificial Intelligence*, Z. Zhou, Ed., 2021, pp. 4627–4635.

[85] J. Yoon, S. Kim, E. Yang, and S. J. Hwang, "Scalable and order-robust continual learning with additive parameter decomposition," in *International Conference on Learning Representations*, 2020, pp. 1–15.

[86] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biological Cybernetics*, vol. 59, pp. 291–294, Sept. 2004.