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, failing to continually adapt to such 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 new setting with a measure to quantify the plasticity-stability trade-off. We then propose a simple yet effective method for learning BIQA models continually. 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 quality score by an adaptive weighted summation of estimates from all prediction heads. Extensive experiments demonstrate the promise of the proposed continual learning method in comparison to standard training techniques for BIQA.

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 during deployment. It is, therefore, 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 synthetic 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], [18]. Recently, Zhang *et al*. [19], [20] 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 [21]), 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 space, 2) apparent subpopulation shift, 3) no test-time oracle, 4) no direct access to previous data, and

- *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.*

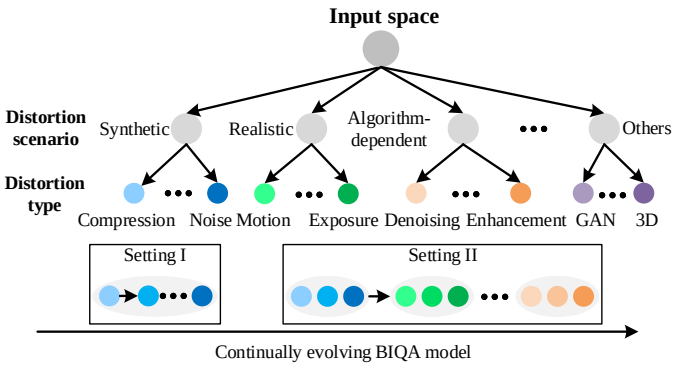


Fig. 1. Illustration of the continual learning paradigm for BIQA. Subpopulation shift exists across distortion types and scenarios. In Setting I, a BIQA model continually evolves from one distortion type to another within the same distortion scenario. In Setting II, a BIQA model continually evolves with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

5) bounded model size. 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 adaptive weightings 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.
- 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 directly 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	7,780	163	Synthetic	5	SS-CQR	2006
CSIQ [6]	866	8,786	173	Synthetic	6	MS-CQR	2010
BID [22]	586	11,204	117	Realistic	N.A.	SS-CQR	2011
CLIVE [10]	1,162	24,604	232	Realistic	N.A.	SS-CQR-CS	2016
KonIQ-10K [11]	10,073	139,274	2,015	Realistic	N.A.	SS-ACR-CS	2018
KADID-10K [8]	10,125	140,071	2,000	Synthetic	25	DS-ACR-CS	2019

IQA datasets becomes 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*. [19], [20] 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. Another family of methods identify and penalize changes to important parameters with respect to previous tasks when learning new tasks. Representative work includes elastic weight consolidation [44] and its online variant [45], incremental moment matching [46], variational continual learning [47], synaptic intelligence [48], and memory-aware synapses [49]. Masse *et al*. [50] proposed context-dependent gating as a complementary module to weight consolidation [44], [48]. Farquhar and Gal [51] noted that soft regularization may not suffice to constrain the model parameters in feasible regions. As a result, parameter isolation [52] as a form of hard regularization has been proposed, 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.

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 a plasticity-stability measure.

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 $\mathcal{D}_t$, no direct access to training images in $\{\mathcal{D}_k\}_{k=1}^{t-1}$ is allowed, leading to the following training objective:

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

where x and q denote the "distorted" image and the corresponding MOS, respectively. f_w represents a 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. During evaluation, we may measure the performance of f_w on the hold-out test sets of all tasks seen so far:

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

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 distinct 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 Space**. This requires that IQA datasets of possibly different perceptual scales should share a common perceptual space. In other words, there exists a *monotonic* function for each dataset to embed its MOSs to this perceptual space. Otherwise, learning a single f_w for multiple datasets continually is conceptually infeasible. Desideratum I

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

Fig. 2. Images sampled from the DQA dataset [57]. A larger MOS in the dataset denotes lower rain density. It is not hard to observe that rain density is not monotonically correlated with perceived image quality. Therefore, DQA violates Desideratum I, and should be excluded to form the task sequence for BIQA. Images are cropped for improved visibility.

excludes human-rated datasets that record only certain aspects of image quality (*e.g.*, contrast perception [58] and scene visibility [57], [59]), and that measure perceptual quantities closely related to image quality (*e.g.*, quality of experience [60] that highly depends on viewing conditions). To highlight this point, we show some images from the deraining quality assessment (DQA) dataset [57] in Fig. 2, with a smaller MOS indicating higher rain density. It is clear that rain density is not *monotonically* correlated with visual quality. Therefore, DQA violates this desideratum, and should be excluded to form the task sequence for BIQA.

II **Apparent Subpopulation Shift**. It is empirically proven that existing BIQA models generalize reasonably to test images with previously seen distortions. Therefore, to make continual learning for BIQA nontrivial, we stipulate that at least two datasets in a task sequence should exhibit apparent subpopulation shift. In other words, part of the distorted images from the two datasets should exhibit noticeably different appearances (see Fig. 1). Desideratum II excludes a series of easy settings, for example, continual learning from additive noise to multiplicative noise and from LIVE [2] to CSIQ [6].

III **No Direct Access to Previous Data**. Some continual learning methods for image classification [52] rely on replaying (at least part of) training data of old tasks to fight against catastrophic forgetting [61], [62], [63], [64]. In the context of BIQA, Zhang *et al*. [19], [20] proposed to jointly train models on data from all tasks, which can be seen as the upper bound of methods with partial access to previous data. To make continual learning for BIQA practical, we assume no direct access to previous data when training new tasks. Notwithstanding, Desideratum III permits summarizing datasets with negligible bits of statistics compared to the dataset sizes.

IV **No Test-Time Oracle**. A well-designed continual learning method should be independent of the task oracle to make prediction. 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 Model Size**. The model capacity in terms of 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. Desideratum V requires the number of learnable parameters introduced by a new task to be negligible compared to that of the current model.

3.3 A Plasticity-Stability Measure

The plasticity-stability dilemma [65] is pervasive in continual learning of computer algorithms, especially for those implemented by artificial neural networks. Formally, the plasticity and stability refer to the ability of integrating new information and preserving previous knowledge, respectively. Here we propose a quantitative measure to evaluate the plasticity-stability trade-off of a BIQA model during continual learning. Without loss of generality, we use Spearman's rank correlation coefficient (SRCC) to benchmark the performance of a BIQA model. 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. We define a plasticity-stability ratio after the BIQA model has learned the t-th task:

$$\mathrm{PSR}_t = \begin{cases} \mathrm{SRCC}_t & t = 1 \\ \left(\frac{1}{t-1}\sum_{k=1}^{t-1}\frac{\mathrm{SRCC}_{tk}}{\mathrm{SRCC}_k}\right)\cdot \mathrm{SRCC}_t & t > 1, \end{cases} \tag{3}$$

where SRCC_{tk}, for $k \leq t$, is the SRCC result of the model on the k-th dataset when it has just learned on the t-th dataset. We omit a subscript when $t = k$. A larger PSR_t indicates a better plasticity-stability trade-off. Generally, learning a new task will destroy some performance of old tasks, leading to $\mathrm{SRCC}_{tk}/\mathrm{SRCC}_k < 1$. Nevertheless, it also makes sense that new tasks will help improve performance on old ones. In Eq. (3), we give credit to such cases with $\mathrm{SRCC}_{tk}/\mathrm{SRCC}_k > 1$. Finally, we define a mean PSR (MPSR) over a list of T tasks as an overall plasticity-stability measure:

$$\mathrm{MPSR} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{PSR}_t, \tag{4}$$

where we drop the subscript T of MPSR to make the notation uncluttered.

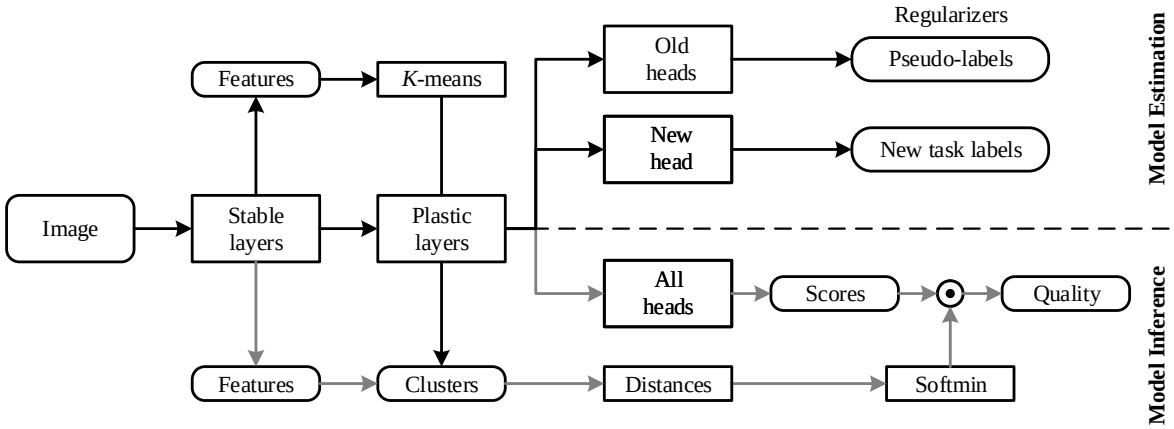


Fig. 3. System diagram of the proposed continual learning method for BIQA. Black and grey arrows correspond to the training and testing phases, respectively.

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 Section 3.2. According to Desiderata I, it is desirable to work in the assumed common perceptual space. However, this is difficult because we are only given a stream of T IQA datasets without the monotonic functions to embed the associated MOSs into this space. Inspired by [19], [20], we want to learn a single perceptual scale for all tasks by exploiting relative quality information. Specifically, under the Thurstone's model [66], the perceptual quality of image x, denoted by q_x, follows a Gaussian distribution with mean μ_x and variance σ_x^2. Assuming the variability of quality between $x \in \mathcal{D}_t$ and $y \in \mathcal{D}_t$ is uncorrelated, the quality difference $q_x - q_y$ is also Gaussian with mean $\mu_x - \mu_y$ and variance $\sigma_x^2 + \sigma_y^2$. We then compute a probability that x is perceived better than y by

$$p(x,y) = \Phi\left(\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2 + \sigma_y^2}}\right) \tag{5}$$

as the ground-truth annotation for pair of images $x, y \in \mathcal{D}_t$, where $\Phi(\cdot)$ is the standard Normal cumulative distribution function. The μ_x and σ_x^2 can be approximated by the collected MOS and the corresponding variance. In summary, when learning the t-th task, we transform $\mathcal{D}_t = \{x_t^{(i)}, \mu_t^{(i)}, \sigma_t^{(i)}\}_{i=1}^{|\mathcal{D}_t|}$ to $\mathcal{P}_t = \{(x_t^{(i)}, y_t^{(i)}), p_t^{(i)}\}_{i=1}^{N_t}$, where $N_t \leq \binom{|\mathcal{D}_t|}{2}$.

Our BIQA model consists of a backbone network, $f_\phi(\cdot)$ parameterized by ϕ, to produce a fixed-length feature vector irrespective of input resolution, and an output 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

$$\hat{p}_t(x,y) = \Phi\left(\frac{h_{\psi_t}(f_\phi(x)) - h_{\psi_t}(f_\phi(y))}{\sqrt{2}}\right), \tag{6}$$

where the variance of quality predictions is fixed to one [66]. The full set of parameters (over a list of T tasks), $\{\phi, \psi_1, \psi_2, \ldots, \psi_T\}$, 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 [67], whose advantages over the cross entropy loss have been demonstrated in several BIQA studies [19], [20]:

$$\ell_{\text{new}}(x,y;\phi,\psi_t) = 1 - \sqrt{p(x,y)\hat{p}_t(x,y)} - \sqrt{(1-p(x,y))(1-\hat{p}_t(x,y))}. \tag{7}$$

Direct optimization of Eq. (7) may cause catastrophic forgetting of old tasks (see Table 4). Inspired by LwF [41], we add a regularizer to allow forgetting old knowledge gracefully, while respecting Desiderata III. 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. (6). 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 $\bar{p}_{tk}$. Again, we use the fidelity loss to implement the constraint:

$$\ell_{\text{old}}\left(x,y;\phi,\{\psi_k\}_{k=1}^{t-1}\right) = \sum_{k=1}^{t-1}\left(1 - \sqrt{\bar{p}_{tk}(x,y)\hat{p}_{tk}(x,y)} - \sqrt{(1-\bar{p}_{tk}(x,y))(1-\hat{p}_{tk}(x,y))}\right). \tag{8}$$

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:

$$\mathcal{L}\left(\mathcal{B}_t;\phi,\{\psi_k\}_{k=1}^{t}\right) = \frac{1}{|\mathcal{B}_t|}\sum_{(x,y)\in\mathcal{B}_t}\left(\ell_{\text{new}}(x,y;\phi,\psi_t) + \lambda\ell_{\text{old}}\left(x,y;\phi,\{\psi_k\}_{k=1}^{t-1}\right)\right), \tag{9}$$

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

Two delicate design choices are worth elaborating. First, to remind the proposed method of preventing catastrophic

TABLE 2
The network architecture of the proposed method based on ResNet-18 [69] for a T-length task sequence. The nonlinear activation and the normalization layers are omitted for brevity

Layer Name	Layer Specification
Convolution	7×7, 64, stride 2
Max Pooling	3×3, stride 2
Residual Block 1	$\begin{bmatrix} 3\times3, 64, \text{stride } 1 \\ 3\times3, 64, \text{stride } 1 \end{bmatrix}\times2$
Residual Block 2	$\begin{bmatrix} 3\times3, 128, \text{stride } 2 \\ 3\times3, 128, \text{stride } 1 \end{bmatrix}\times1$ $\begin{bmatrix} 3\times3, 128, \text{stride } 1 \\ 3\times3, 128, \text{stride } 1 \end{bmatrix}\times1$
Residual Block 3	$\begin{bmatrix} 3\times3, 256, \text{stride } 2 \\ 3\times3, 256, \text{stride } 1 \end{bmatrix}\times1$ $\begin{bmatrix} 3\times3, 256, \text{stride } 1 \\ 3\times3, 256, \text{stride } 1 \end{bmatrix}\times1$
Residual Block 4	$\begin{bmatrix} 3\times3, 512, \text{stride } 2 \\ 3\times3, 512, \text{stride } 1 \end{bmatrix}\times1$ $\begin{bmatrix} 3\times3, 512, \text{stride } 1 \\ 3\times3, 512, \text{stride } 1 \end{bmatrix}\times1$
Global Average Pooling	-
Full Connection	$512 \times T$

forgetting [64], we treat the backbone network as a composition of two functions $f_\phi = f_{\phi_p} \circ f_{\phi_s}$, where f_{ϕ_s} and f_{ϕ_p} represent the first few and the remaining convolution layers of the DNN. The pre-trained f_{ϕ_s} (for object recognition) is fairly transferable across different vision tasks. We take advantage of this and freeze f_{ϕ_s} to encourage stability during training. The parameters of f_{ϕ_p} are adapted to new tasks, accounting for plasticity. Second, we append an ℓ_2-normalization layer [68] on top of the backbone network:

$$\tilde{f}_\phi(x) = \frac{f_\phi(x)}{\|f_\phi(x)\|_2} \tag{10}$$

to project the feature representation onto the unit hypersphere. This pushes the predictions of all heads to approximately the same 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 directly applicable to BIQA. Instead of relying on the task oracle to precisely activate a task-specific prediction head, we design an adaptive weighting mechanism to compute a weighted summation of quality estimates from all heads as the overall quality score.

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

$$\tilde{f}_{\phi_s}(x) = \frac{\text{pool}(f_{\phi_s}(x))}{\|\text{pool}(f_{\phi_s}(x))\|_2}, \tag{11}$$

TABLE 3
Performance comparison in terms of MPSR and weighted SRCC (with weightings proportional to the sizes of the six IQA test sets). All methods are trained in chronological order

Method	MPSR	Weighted SRCC
Proposed (LwF-AW)	0.8166	0.7886
SL	0.7223	0.6585
SH-CL	0.7698	0.7010
MH-CL	0.7560	0.6704
MH-CL-AW	0.7392	0.6565
LwF	0.7723	0.6996
MH-CL-O	0.7447	0.6711
LwF-O	0.8166	0.8198

where $\text{pool}(\cdot)$ denotes global average pooling over spatial locations. Similar in Eq. (10), we normalize the pooled representations to make them more comparable across different tasks. We then summarize $\mathcal{D}_t$ with K centroids $\{c_t^{(j)}\}_{j=1}^K$ by applying K-means [70] 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 that of the entire training set, Desideratum III is respected. We use f_{ϕ_s} (rather than $f_{\phi_p} \circ f_{\phi_s}$ or $h_{\psi_t} \circ f_{\phi_p} \circ f_{\phi_s}$) as a feature extractor to distill $\mathcal{D}_t$ because it is fixed during model development, which effectively reduces the *task-recency* bias [71].

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 $\mathcal{D}_t$:

$$d_t(x) = \min_{1\le j\le K} \|\tilde{f}_{\phi_s}(x) - c_t^{(j)}\|_2. \tag{12}$$

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

$$a_t(x) = \frac{\exp(-\tau d_t(x))}{\sum_{k=1}^T \exp(-\tau d_k(x))}, \tag{13}$$

where $\tau \ge 0$ is a temperature parameter used to tune the smoothness of the softmin function. Setting a higher value of τ produces a harder weight assignment over T prediction heads. The final quality score is defined as the inner product between two vectors of adaptive weightings and quality predictions:

$$\hat{q}(x) = \sum_{t=1}^T a_t h_{\psi_t}(f_\phi(x)). \tag{14}$$

A final note is that the number of parameters of the T prediction heads is designed to be considerably smaller than that of the backbone network. Thus, our BIQA model meets Desideratum V.

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. 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.

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

Dataset	Method	LIVE [2]	CSIQ [6]	BID [22]	CLIVE [10]	KonIQ-10K [11]	KADID-10K [8]
LIVE	All	**0.9266**	0.5777	0.6553	0.4827	0.7257	0.5674
CSIQ	SL	0.9193	**0.8449**	0.6357	0.4515	0.6388	0.5729
	SH-CL	0.9360	0.8246	0.6974	0.4916	0.7280	0.6083
	MH-CL	0.9339	0.8189	0.6903	0.4863	0.7281	0.6124
	MH-CL-AW	0.9200	0.8139	0.6960	0.4894	0.7288	0.5928
	LwF	**0.9363**	0.8020	0.7119	0.5117	0.7490	0.6098
	Proposed	0.9038	0.7688	0.7145	0.5346	0.7485	0.5430
BID	SL	0.6509	0.5732	0.8082	0.7308	0.7095	0.3268
	SH-CL	0.8814	0.7764	0.8134	0.7191	0.7463	0.3958
	MH-CL	0.8538	0.7509	0.8117	0.7241	0.7365	0.3912
	MH-CL-AW	0.8674	0.7618	0.8116	0.7232	0.7393	0.4030
	LwF	0.8408	0.7694	**0.8183**	0.7088	0.7365	0.3942
	Proposed	**0.9276**	**0.7901**	0.8150	0.6712	0.7639	0.5142
CLIVE	SL	0.6562	0.5457	0.8327	0.8316	0.7724	0.4138
	SH-CL	0.7430	0.5889	0.8252	0.8387	0.7814	0.3810
	MH-CL	0.6577	0.5911	0.8162	0.8375	0.7681	0.3806
	MH-CL-AW	0.6660	0.5803	0.8170	0.8359	0.7622	0.3549
	LwF	0.7629	0.6441	**0.8407**	**0.8599**	0.7648	0.3047
	Proposed	**0.9230**	**0.7815**	0.8399	0.8034	0.7783	0.5043
KonIQ-10K	SL	0.7651	0.7452	0.7760	0.7043	**0.8811**	0.5437
	SH-CL	0.8104	0.7250	0.7677	0.7024	0.8809	0.5382
	MH-CL	0.8297	0.7413	0.7688	0.6725	**0.8811**	0.5375
	MH-CL-AW	0.7853	0.7169	0.7591	0.6718	0.8783	0.5262
	LwF	0.7992	0.7037	0.7603	0.7107	0.8704	0.5393
	Proposed	**0.9233**	**0.7934**	**0.8156**	**0.7819**	0.8450	0.5584
KADID-10K	SL	0.8206	0.7120	0.5865	0.3822	0.5109	**0.8241**
	SH-CL	0.8496	0.7185	0.6560	0.3777	0.6240	0.8043
	MH-CL	0.8885	0.7233	0.5521	0.3282	0.5531	0.8120
	MH-CL-AW	0.7331	0.5133	0.5169	0.4390	0.5814	0.7706
	LwF	0.8481	0.6958	0.6989	0.3423	0.6104	0.8184
	Proposed	**0.8948**	**0.7836**	**0.7797**	**0.7263**	**0.8117**	0.7655

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 organize these datasets in chronological order, *i.e.*, LIVE $\rightarrow$ CSIQ $\rightarrow$ BID $\rightarrow$ LIVE Challenge $\rightarrow$ KonIQ-10K $\rightarrow$ KADID-10K. In Section 5.4, we also use task sequences of different orders to evaluate the order-robustness of the proposed method. We randomly sample 80% images from each dataset for training, and leave the remaining for testing. We follow [19], [20] 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 be inaccessible, we consider using them for performance evaluation as in the standard cross-dataset setting.

We use a variant of ResNet-18 [69] as the backbone of our BIQA model, which contains more than 10 million trainable parameters. We strip all fully connected layers in ResNet-18, and append a global average pooling layer after the last convolution to produce a 512-dimensional feature vector. Each of the six prediction heads is implemented by a fully connected layer with 512 parameters (and no bias term), accounting for less than 0.03% of the total parameters. As such, the growth of model complexity introduced by each new task is negligible, conforming to Desideratum V. Details of the network is presented in Table 2. We set the first convolution layer and the subsequent three residual blocks as the stable layers by freezing their parameters during continual learning. The last residual block and all full connections are the plastic layers, whose parameters can evolve with the task sequence (see Fig. 3). Different splitting points of the stable and plastic layers will be investigated in Section 5.4.

For each task, stochastic optimization is carried out by Adam [72] with $\lambda = 1$ in Eq. (9). The parameters of the backbone network and the prediction heads are initialized by the weights pre-trained on ImageNet [73] and the He's method [74], respectively. We set the initial learning rate to 3×10^{-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. Empirically, we find that the performance is insensitive to the choice of K. We set the temperature to $\tau = 16$ in Eq. (13). We test on images of original size in all experiments.

5.2 Competing Methods

We present several training techniques that are closely related to our method for comparison.

TABLE 5
Performance comparison in terms of SRCC between the proposed method and its "upper bounds"

Dataset	Method	LIVE [2]	CSIQ [6]	BID [22]	CLIVE [10]	KonIQ-10K [11]	KADID-10K [8]
All	JL	0.9663	0.8691	0.8512	0.8201	0.8971	0.8804
LIVE	All	**0.9266**	–	–	–	–	–
CSIQ	MH-CL-O	**0.9338**	**0.8189**	–	–	–	–
	LwF-O	0.8641	0.8020	–	–	–	–
	Proposed	0.9038	0.7688	–	–	–	–
BID	MH-CL-O	0.8690	0.7594	0.8117	–	–	–
	LwF-O	0.9069	0.7883	**0.8183**	–	–	–
	Proposed	**0.9276**	**0.7901**	0.8150	–	–	–
CLIVE	MH-CL-O	0.6373	0.5951	0.8180	0.8375	–	–
	LwF-O	0.8848	0.7569	0.8176	**0.8599**	–	–
	Proposed	**0.9230**	**0.7815**	**0.8399**	0.8034	–	–
KonIQ-10K	MH-CL-O	0.7987	0.7339	0.7582	0.6880	**0.8811**	–
	LwF-O	0.8721	0.7485	0.8022	**0.8401**	0.8704	–
	Proposed	**0.9233**	**0.7934**	**0.8156**	0.7819	0.8450	–
KADID-10K	MH-CL-O	0.7113	0.4494	0.5268	0.4272	0.5819	0.8120
	LwF-O	0.8604	0.7029	0.7649	**0.8026**	**0.8333**	**0.8184**
	Proposed	**0.8948**	**0.7836**	**0.7797**	0.7263	0.8117	0.7655

- **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. (9) with $\lambda = 0$.
- **Joint Learning (JL)** is a recently proposed method [19], [20] to overcome the cross-distortion-scenario 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. (9) 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. (9) 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. (9) with $\lambda = 0$. It remains to specify one of the heads for final quality prediction. To encourage adaptation to a constantly changing environment, we simply use the latest head to make prediction. Meanwhile, we may incorporate the proposed adaptive weighting during inference, giving rise to **MH-CL-AW**. 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. (9) with $\lambda = 1$. 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-AW)** can be seen as the combination of LwF and adaptive weighting. Moreover, we also explore the task oracle to select the corresponding head for quality prediction, denoted by **LwF-O**.

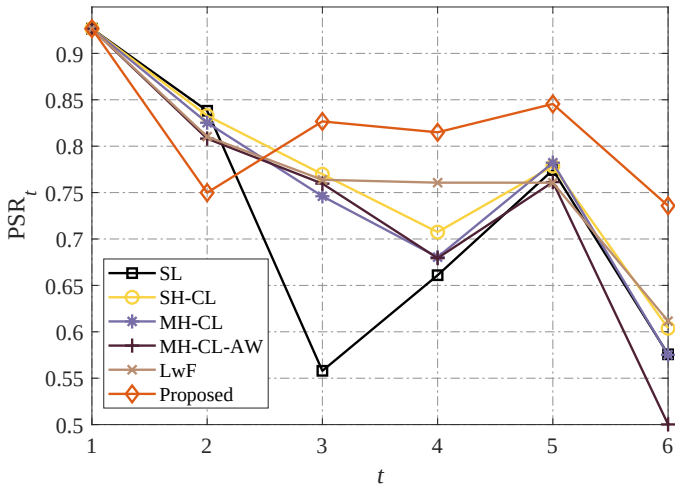
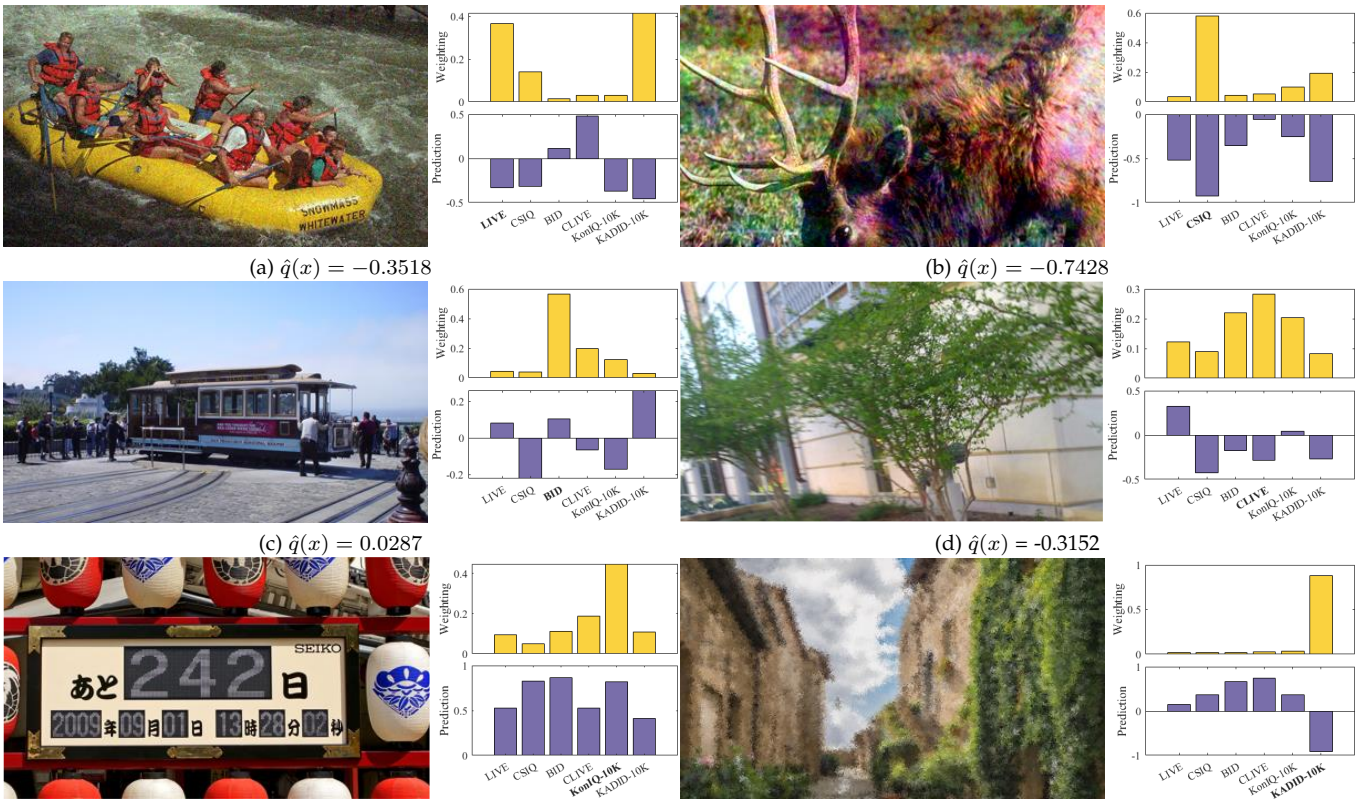


Fig. 4. PSR_t as a function of the task index t.

5.3 Main Results

5.3.1 Quantitative Results

We use the proposed MPSR in Eq. (4) to benchmark the plasticity-stability trade-off once the learning on the task sequence is completed. We also report the weighted SRCC on the six IQA test sets. 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, while SH-CL improves the MPSR result upon SL by a clear margin, it underperforms LwF, indicating that regularizers to stabilize the performance of previous tasks may be necessary. This is also evidenced through the comparison between MH-CL and LwF, where we observe significant performance drops



(a) $\hat{q}(x) = -0.3518$ (b) $\hat{q}(x) = -0.7428$

(c) $\hat{q}(x) = 0.0287$ (d) $\hat{q}(x)$ = -0.3152

(e) $\hat{q}(x)$ = 0.6990 (f) $\hat{q}(x)$ = -0.7566

Fig. 5. Perceptual scaling of images sampled from the six IQA datasets. The bar charts of adaptive weightings and quality predictions of all heads are also presented along with each image. The final quality prediction $\hat{q}(x)$ is shown in the subcaption. Zoom in for better distortion visibility.

of MH-CL in terms of both MPSR and weighted SRCC. Third, adaptive weighting (or the task oracle) alone may hurt the performance of MH-CL, when comparing to MH-CL-AW and MH-CL-O. We believe this arises because the backbone network is constantly evolving with new data, but old prediction heads are not, resulting in a mismatch between the feature extractor and the quality predictors. Fourth, equipped with the adaptive weighting mechanism and the LwF regularizer, our method approaches the upper-bound performance by LwF-O in terms of MPSR. We also plot PSR_t as a function of the task index t in Fig. 4, from which we find that our method is more stable, and performs much better as the length of the task sequence increases.

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 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. Therefore, SH-CL achieves a better plasticity-stability trade-off. Third, both MH-CL and LwF add a prediction head for each new task, which enable learning new quality mapping functions without affecting old ones. However, the new head does not handle old tasks well, which necessitates an effective mechanism to make full use of all learned heads. Fourth, built upon LwF, our method employs adaptive weighting to pool quality estimates, leading to better performance especially on previous tasks.

Table 5 shows the SRCC results of our method against three "upper bounds", which leverage some form of information not allowed by the desiderata in Section 3.2. As expected, JL provides an effective solution to subpopulation shift in BIQA. Despite being unscalable, it serves as the upper bound of all continual learning methods. With access to the task oracle, LwF-O achieves the closest performance to JL. Interestingly, the proposed method is able to deliver better performance on many of the old tasks with strict adherence to Desideratum IV. This may be because the proposed adaptive weighting mechanism effectively implements a BIQA ensemble, which appears to be more resistant to catastrophic forgetting.

5.3.2 *Qualitative Results*

We conduct a qualitative analysis of our BIQA model by sampling test images from the task sequence. Also shown in Fig. 5 are the bar charts of adaptive weightings and quality predictions corresponding to each image. Although the proposed method is not jointly trained on all IQA datasets [19], [20], it successfully learns one perceptual scale for all tasks, well aligning different images in the learned scale. Moreover, visual inspections of the bar charts reveal that the prediction heads may only give accurate quality estimates for the datasets they are exposed to. Fortunately, given a test image, the proposed adaptive weighting is

TABLE 6
Performance comparison in terms of SRCC of the proposed method 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

Order	Dataset	KADID-10K [8]	KonIQ-10K [11]	CLIVE [10]	BID [22]	CSIQ [6]	LIVE [2]
I	KADID-10K	0.8241	0.5109	0.3822	0.5865	0.7120	0.8206
	KonIQ-10K	0.7873	0.8580	0.6783	0.7728	0.7571	0.9252
	CLIVE	0.7695	0.8495	0.8044	0.8299	0.7535	0.8968
	BID	0.7397	0.8261	0.8135	0.8355	0.7554	0.8960
	CSIQ	0.6491	0.8151	0.7647	0.7962	0.8213	0.8897
	LIVE	0.6417	0.8072	0.7247	0.7569	0.7710	0.9356
Order	Dataset	LIVE [2]	BID [22]	CSIQ [6]	CLIVE [10]	KADID-10K [8]	KonIQ-10K [11]
II	LIVE	0.9266	0.6553	0.5777	0.4827	0.5674	0.7257
	BID	0.9431	0.8066	0.6377	0.6838	0.5024	0.7467
	CSIQ	0.9167	0.7792	0.7823	0.6812	0.4477	0.7470
	CLIVE	0.9283	0.8277	0.7710	0.8115	0.4343	0.7673
	KADID-10K	0.9098	0.7593	0.8333	0.7548	0.7703	0.7468
	KonIQ-10K	0.9141	0.7710	0.8510	0.7874	0.7225	0.8114
Order	Dataset	BID [22]	CLIVE [10]	KonIQ-10K [11]	LIVE [2]	CSIQ [6]	KADID-10K [8]
III	BID	0.8082	0.7308	0.7095	0.6509	0.5732	0.3268
	CLIVE	0.8414	0.8390	0.7401	0.6462	0.5557	0.3246
	KonIQ-10K	0.8287	0.8194	0.8438	0.7659	0.6754	0.5313
	LIVE	0.7907	0.7342	0.8129	0.9403	0.6752	0.4774
	CSIQ	0.7690	0.7042	0.7902	0.9394	0.7626	0.4422
	KADID-10K	0.7486	0.6873	0.7649	0.9230	0.8009	0.7219
Order	Dataset	LIVE [2]	CSIQ [6]	KADID-10K [8]	BID [22]	CLIVE [10]	KonIQ-10K [11]
IV	LIVE	0.9266	0.5777	0.5674	0.6553	0.4827	0.7257
	CSIQ	0.9038	0.7688	0.5430	0.7145	0.5346	0.7485
	KADID-10K	0.8859	0.8119	0.7884	0.6988	0.5117	0.7228
	BID	0.9240	0.8352	0.7768	0.7939	0.6753	0.7606
	CLIVE	0.9196	0.8112	0.7503	0.8325	0.7759	0.7754
	KonIQ-10K	0.9172	0.8203	0.7384	0.8087	0.7632	0.8439

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. 5 (a)), the heads trained on CSIQ [6] and KADID-10K [8] of similar synthetic distortions are also assigned relatively high weights. As another example, KADID-10K [8] contains some distinct distortion types (*e.g.*, spatial jitter in Fig. 5 (f)); consequently, the assigned weighting for the head of KADID-10K tends to dominate.

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. Consequently, a BIQA model is expected to be independent of the task order it is trained on [75]. 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 detailed SRCC results in Table 6 and the MPSR and the weighted SRCC results in Table 7, respectively, from which we make some useful observations. First, the proposed method is quite robust to handle task sequences of different orders in terms of MPSR, providing justifications for its use in real-world applications. Second, the reverse chronological order (Order I) and the sequence of realistic distortions followed by synthetic distortions (Order III) achieve lower weighted SRCC results compared to the other two task orders. We believe this is because harder tasks appear in the beginning of the sequence, making it difficult for our method to trade off plasticity and stability. Specifically, for Order I, the first task on KADID-10K [8] is considered a much harder one than those on CSIQ [6] and LIVE [2] due to the introduction of more distortion types. Our method offers an SRCC of 0.8241 on KADID-10K [8] when it is first trained on, and

TABLE 7
Performance comparison in terms of MPSR and weighted SRCC 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	MPSR	Weighted SRCC
I	0.8000	0.7338
II	0.8163	0.7924
III	0.7993	0.7488
IV	0.8184	0.7953
V	0.8166	0.7886

TABLE 8
MPSR and weighted SRCC as functions of the splitting point of stable and plastic layers. The default setting is highlighted in bold

Splitting Point	MPSR	Weighted SRCC
None	0.7620	0.6605
Up to First Convolution	0.7937	0.7708
Up to Residual Block 1	0.7895	0.7567
Up to Residual Block 2	0.8069	0.7664
Up to Residual Block 3	**0.8166**	**0.7886**
Up to Residual Block 4	0.6497	0.5859

TABLE 9
Performance comparison in terms of MPSR and weighted SRCC for different design choices

Design Choice	MPSR	Weighted SRCC
LwF-SW with normalization	0.7746	0.7620
LwF-HW with normalization	0.7960	0.7391
LwF-AW w/o normalization	0.7820	0.7321
LwF-AW with normalization (Ours)	**0.8166**	**0.7886**

fails to stabilize the performance with a final SRCC of 0.7219. Similarly, quality prediction of realistically distorted images is a more difficult computational task than evaluating synthetically distorted images. This may help explain the final inferior performance on BID [22], LIVE Challenge [10], and KonIQ-10K [11] as the first three tasks in Order III. The order-robustness experiment reveals that there is still room for improving the model ability to consolidate learned knowledge and acquire new knowledge in the presence of a difficult task order.

5.4.2 Splitting Point of Stable and Plastic Layers

As stated in Section 4.1, the backbone network of the proposed method is composed of a cascade of stable and plastic layers. Recall that we use ResNet-18 [69] as the backbone network, which consists of a first convolution layer followed by four residual blocks (see Table 2). We list the MPSR and the weighted SRCC results as functions of the splitting point in Table 8. We notice a significant performance drop (compared to the default setting) when all parameters are frozen (*i.e.*, up to Residual Block 4). This is not surprising since no plastic layer is reserved for adapting to new tasks. An opposite extreme is when there is no stable layer, namely, all network parameters are updated with new data during continual learning. In this case, we apply Eq. (11) to the responses of the last convolution to compute the feature vector for adaptive weighting. It is clear that this variant is also weak in terms of both MPSR and weighted SRCC. For the remaining cases, our method is relatively insensitive to the choice of the splitting point.

5.4.3 Model Weighting and Feature Normalization

To demonstrate the promise of our adaptive weighting scheme, we compare it with two alternative weighting strategies. The first is to simply average the quality predictions of all heads, which is termed as **LwF-SW**. The second adopts a hard-weighting method, selecting a single head with the highest weight for quality prediction, termed as **LwF-HW**. Mathematically, LwF-SW and LwF-HW correspond to setting $\tau = 0$ and $\tau \to \infty$ in Eq. (13), respectively. We list the MPSR and the weighted SRCC results in Table 9. Detailed SRCC values are shown in Table 10. We find that our method clearly outperforms LwF-SW, suggesting that simple averaging may introduce bias from less reliable prediction heads. LwF-HW shows an MPSR improvement over LwF-SW, indicating that a better plasticity-stability trade-off has been made. However, it is dangerous to rely solely on the prediction head with the highest weight, especially when the weighting function for the given test image is less accurate. This has been reflected by a noticeable reduction in weighted SRCC, where hard weighting is unable to handle the last task on KADID-10K, as shown in Table 10. By contrast, adaptive weighting in the proposed method effectively constructs an ensemble from a set of relatively accurate prediction heads, therefore representing a more reliable means of computing the final quality score.

We then remove the normalization step of the proposed method (computed by Eq. (10)). From Table 9, we observe that our method without normalization gives inferior performance in MPSR and weighted SRCC. We attribute this performance drop to the scale differences of the learned prediction heads. For example, without feature normalization, the prediction head for LIVE produces quality scores in the range of $[-7, 5]$, while the head for CSIQ outputs scores in $[-3, 2]$. As a result, it is less meaningful to combine quality scores linearly. As empirically observed after feature normalization, all prediction heads give scores of similar scales, making adaptive weighting more sensible.

5.5 Further Testing Using Different Continual Learning Regularizers

We have incorporated LwF [41] into BIQA as a regularizer to mitigate catastrophic forgetting. On top of LwF, we have described an adaptive weighting module for the BIQA model with multiple heads, bypassing the task oracle during inference. In this subsection, we show that the proposed method of computing the final quality score is compatible with other regularizers in continual learning. Specifically, we implement three such regularizers - elastic weight consolidation (EWC) [44], synaptic intelligence (SI) [48], and memory aware synapses (MAS) [49]. All the three methods follow a similar paradigm that penalizes the changes to the estimated "important" parameters for previous tasks when learning the new task. Given a mini-batch of samples, $\mathcal{B}_t$, the empirical loss is computed as

$$\begin{aligned}\mathcal{L}\left(\mathcal{B}_t; \phi, \{\psi_k\}_{k=1}^t\right) =& \frac{1}{|\mathcal{B}_t|} \sum_{(x,y)\in\mathcal{B}_t} \left(\ell_{\text{new}}(x, y; \phi, \psi_t)\right. \\ &\left. + \lambda \sum_i \beta_i \left(\phi_i - \phi_i'\right)^2\right), \end{aligned} \tag{15}$$

where β_i refers to the estimated importance of the i-th parameter to previous tasks. ϕ_i' records the value of the i-th parameter before learning the t-th task, and $\phi_i - \phi_i'$ represents the corresponding change. EWC computes β offline as the diagonals of the Fisher information matrix. SI and MAS estimate β online using the accumulated gradients in slightly different ways. We empirically set the λ values for

TABLE 10
Performance comparison in terms of SRCC for different weighting strategies

Dataset	Method	LIVE [2]	CSIQ [6]	BID [22]	CLIVE [10]	KonIQ-10K [11]	KADID-10K [8]
LIVE	All	**0.9266**	0.5777	0.6553	0.4827	0.7257	0.5674
CSIQ	LwF-SW	**0.9264**	0.7416	0.7011	0.5301	0.7525	0.5693
	LwF-HW	0.8674	**0.8000**	0.6850	0.5207	0.7287	0.5107
	Proposed	0.9038	0.7688	0.7145	0.5346	0.7485	0.5430
BID	LwF-SW	0.9165	0.7452	0.7629	0.6266	0.7723	0.5428
	LwF-HW	0.9031	**0.8248**	**0.8194**	0.6951	0.7332	0.4396
	Proposed	**0.9276**	0.7901	0.8150	0.6712	0.7639	0.5142
CLIVE	LwF-SW	0.8639	0.7175	0.8098	0.7053	0.7872	0.5371
	LwF-HW	0.8728	**0.7960**	0.8235	**0.8606**	0.7472	0.4326
	Proposed	**0.9230**	0.7815	**0.8399**	0.8034	0.7783	0.5043
KonIQ-10K	LwF-SW	0.8564	0.7432	0.7909	0.7104	0.8209	0.5736
	LwF-HW	0.8675	**0.8049**	0.8153	**0.8092**	**0.8625**	0.5151
	Proposed	**0.9233**	0.7934	**0.8156**	0.7819	0.8450	0.5584
KADID-10K	LwF-SW	0.8546	0.7496	0.7624	0.6447	0.7970	0.7345
	LwF-HW	0.8429	**0.7885**	**0.7862**	**0.7861**	**0.8246**	0.6336
	Proposed	**0.8948**	0.7836	0.7797	0.7263	0.8117	**0.7655**

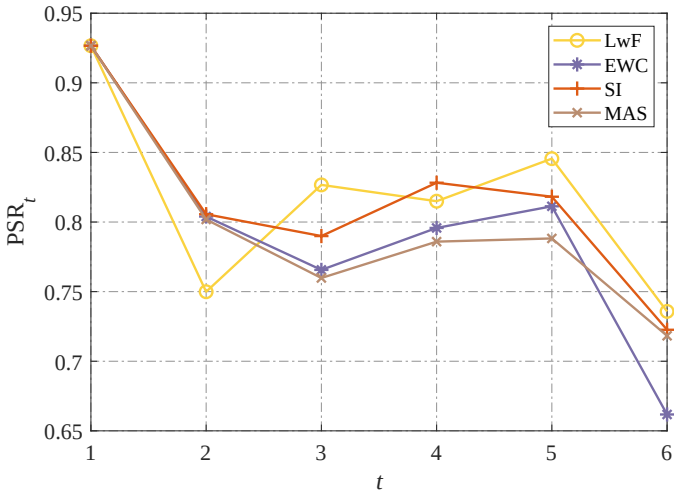


Fig. 6. PSR_t as a function of the task index t for different continual learning regularizers.

EWC, SI, and MAS to $10,000$, 100, and 10, respectively, balancing the magnitude of β for different methods. We compare PSR_t as a function of the task index t in Fig. 6, where we find that the proposed method works equally well with different continual learning regularizers. We also show the performance comparison in terms of MPSR and weighted SRCC in Table 11, where the baselines use the latest head to predict image quality. We see that the proposed method leads to consistent performance gains over the baselines. Therefore, we may conclude that the improvement by the proposed adaptive weighting is orthogonal to the adopted continual learning regularizers.

6 Conclusion

We have formulated continual learning for BIQA with five desiderata and a plasticity-stability measure. We also contributed a continual learning method 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

TABLE 11
Performance improvements upon different baseline continual learning regularizers in terms of MPSR and weighted SRCC. All baselines compute quality scores using the latest prediction head

Method	Baseline	Baseline+AW	Improvement (%)
MPSR			
EWC	0.7587	0.7942	4.68
SI	0.7746	0.8166	5.42
MAS	0.7563	0.7968	5.36
Weighted SRCC			
EWC	0.6513	0.7481	14.86
SI	0.6558	0.7633	16.39
MAS	0.6584	0.7665	16.42

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, 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 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] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, "Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory." *Psychological Review*, vol. 102, no. 3, p. 419, Jul. 1995.

[19] 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.

[20] ——, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *CoRR*, vol. abs/2005.13983, 2020.

[21] R. Aljundi, "Continual learning in neural networks," *CoRR*, vol. abs/1910.02718, 2019.

[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," *CoRR*, vol. abs/2003.03849, 2020.

[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," *CoRR*, vol. abs/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] 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.

[45] 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.

[46] 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.

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

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

[49] 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.

[50] 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.

[51] S. Farquhar and Y. Gal, "Towards robust evaluations of continual learning," *CoRR*, vol. abs/1805.09733, 2018.

[52] M. D. Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *CoRR*, vol. abs/1909.08383, 2019.

[53] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," *CoRR*, vol. abs/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," *CoRR*, vol. abs/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] 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.

[58] 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. 8.1–8.13, Sep. 2008.

[59] 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.

[60] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang, "Developing a predictive model of quality of experience for internet video," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4, pp. 339–350, Aug. 2013.

[61] 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.

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

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

[64] 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.

[65] S. T. Grossberg, *Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control*. Springer Science & Business Media, 2012.

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

[67] 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.

[68] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "Normface: L2 hypersphere embedding for face verification," in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 1041–1049.

[69] 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.

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

[71] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. van de Weijer, "Class-incremental learning: survey and performance evaluation," *CoRR*, vol. abs/2010.15277, 2020.

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

[73] 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.

[74] 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.

[75] 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.