SPIQ: A Self-Supervised Pre-Trained Model for Image Quality Assessment

Pengfei Chen, *Graduate Student Member, IEEE*, Leida Li, *Member, IEEE*, Qingbo Wu, *Member, IEEE*, and Jinjian Wu, *Member, IEEE*

***Abstract*—Blind image quality assessment (BIQA) has witnessed a flourishing progress due to the rapid advances in deep learning technique. The vast majority of prior BIQA methods try to leverage models pre-trained on ImageNet to mitigate the data shortage problem. These well-trained models, however, can be sub-optimal when applied to BIQA task that varies considerably from the image classification domain. To address this issue, we make the first attempt to leverage the plentiful unlabeled data to conduct self-supervised pre-training for BIQA task. Based on the distorted images generated from the high-quality samples using the designed distortion augmentation strategy, the proposed pre-training is implemented by a feature representation prediction task. Specifically, patch-wise feature representations corresponding to a certain grid are integrated to make prediction for the representation of the patch below it. The prediction quality is then evaluated using a contrastive loss to capture quality-aware information for BIQA task. Experimental results conducted on KADID-10 k and KonIQ-10 k databases demonstrate that the learned pre-trained model can significantly benefit the existing learning based IQA models.**

***Index Terms*—Blind image quality assessment, self-supervised pre-training, contrastive learning.**

I. Introduction

THANKS to the popularization of smartphones, trillions of digital images have been taken and shared among daily lives. In almost every stage of the visual communication systems, *e.g.*, acquisition, compression, transmission, and display, various types of distortions are introduced. This is the case where image quality assessment (IQA) is needed to ensure the quality of visual contents delivered to the end-users [1]–[5]. Among all IQA paradigms, researchers have made tremendous strides in developing blind image quality assessment (BIQA) methods that can automatically predict perceptual quality without any information of reference images.

Manuscript received November 30, 2021; revised January 11, 2022; accepted January 18, 2022. Date of publication January 25, 2022; date of current version February 11, 2022. This work was supported by the Outstanding Innovation Scholarship for Doctoral Candidate of CUMT under Grant 2019YCBS032. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Moacir A. Ponti. *(Corresponding author: Leida Li.)*

Pengfei Chen is with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China (e-mail: cpf00790079@gmail.com).

Leida Li and Jinjian Wu are with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: ldli@xidian.edu.cn; jinjian.wu@mail.xidian.edu.cn).

Qingbo Wu is with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: qbwu@uestc.edu.cn).

Digital Object Identifier 10.1109/LSP.2022.3145326

Earlier BIQA methods rely on hand-crafted features to build a prediction model [6]–[11]. Benefiting from the strong learning power, the deep learning based methods have demonstrated remarkable progress on standard IQA benchmarks [12]–[17]. Notably, Ma *et al.* [14] proposed a deeper network to learn distortion type and image quality simultaneously. In [15], Zhang *et al.* proposed a two stream network architecture to predict both synthetic and authentic image distortions. Su *et al.* [16] proposed a self-adaptive architecture to aggregate multi-scale content features for quality prediction. In [17], Zhu *et al.* tried to learn meta-knowledge shared between various distortions and then adapted them to unknown distortions. Very recently, some vision Transformer based BIQA methods [18]–[21] have been developed to bypass the inherent constraints in CNN structures such as the fixed input size, and have shown impressive advances. However, the performances of these deep models heavily depend on the size of the training data. Considering data collection and annotation for BIQA task is extremely labor-intensive and time-consuming compared to other computer vision tasks, existing public IQA datasets are of limited scales with respect to both sample numbers and distortion characteristics.

An emerging approach to handle this incoming problem is to make use of the models pre-trained on those large-scale datasets, among which the supervised ImageNet [22] pre-training has been dominant for years. Although these pre-trained models have led to great performance advances, it becomes questionable whether deploying them for pre-training is optimal in BIQA researches. This makes sense since these models are trained to specialize towards solving a single supervised task, such as image classification, which is significantly distinguished from BIQA task. Another alternative to solve the data shortage problem resides in taking advantage of the unlabeled data, where the recent advance focus of self-supervised learning is proved to be one promising methodology that could learning useful representations from massive amount of unlabeled data.

In light of the above observations, we introduce a novel self-supervised learning framework, termed *Self-supervised Pre-training for Image Quality assessment* (*SPIQ*), that enables discovering rich and descriptive feature representation for BIQA task. In our design, distorted images with diverse categories are first generated based on the designed distortion augmentation strategy. Each of these distorted images is then divided into non-overlapping patches, which are sent to the feature encoder to extract patch-wise feature representations. While different ways of generating self-supervised signals are invented, of particular interest are a family of contrastive learning based methods that self-train deep networks by distinguishing the representation of positive queries from their negative counterparts. By regarding the integrated representation from a grid of patches as the

1070-9908 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on June 07,2022 at 06:31:36 UTC from IEEE Xplore. Restrictions apply.

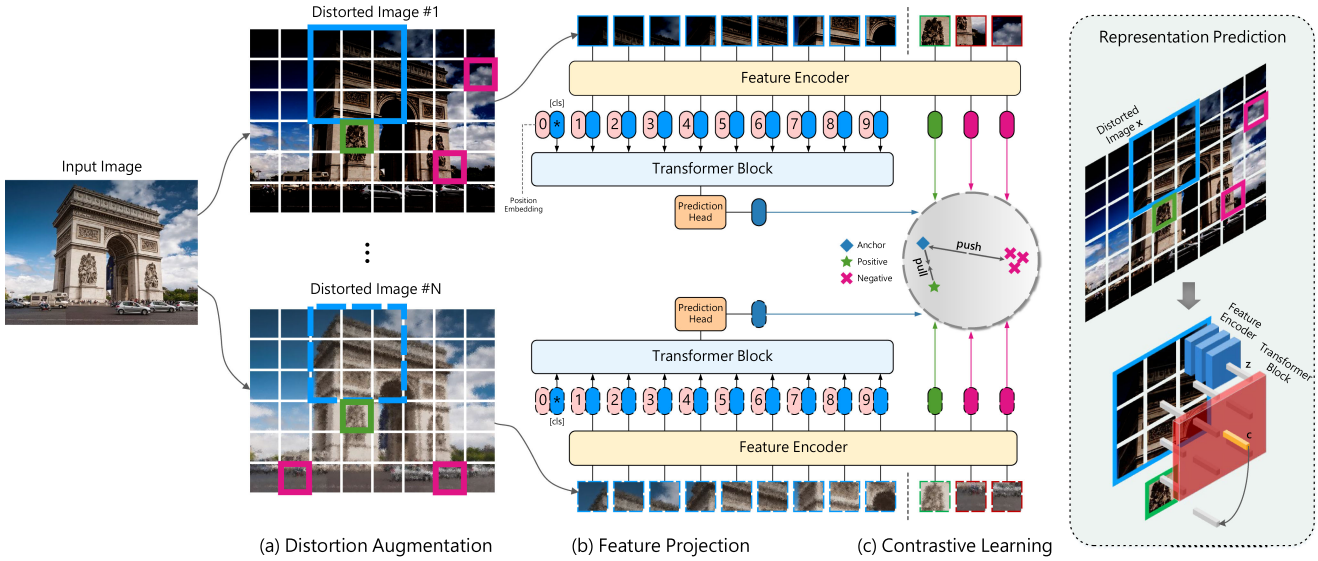

Fig. 1. The overall structure of the proposed method. (a) Based on the distortion augmentation method, training samples with different distortion categories are generated; (b) While the feature encoder is leveraged to extract patch-wise feature representations, the obtained representations that corresponding to patches in a certain grid (blue boxes, anchor) are integrated using the transformer blocks to predict the patch below (green boxes, positive sample); (c) The training procedure is regularized by a contrastive loss, where the the anchor is pulled closer to the positive sample while pushing away from the negative samples (red boxes) in the feature embedding space to learn effective BIQA representations.

anchor in the contrastive learning framework, we formulate the pre-training task by enforcing it to predict the patch beneath the grid in the feature embedding space. The corresponding feature representation is treated as the only positive sample. Enlightened by the observation that the perceptual quality is closely related to both distortion diversity and content variation [23]–[27], the construction of the negative samples is divided into two parts, as *distortion negatives* and *content negatives*. Experimental results suggest that the proposed approach can be well integrated into off-the-shelf IQA metrics as the pre-trained model to achieve better prediction performances.

II. Proposed Approach

The proposed SPIQ aims to train a feature encoder that is able to learn effective representations from abundant unlabeled image data. Fig. 1 depicts the overview of our network. Given a high-quality image, we first generate its distorted counterparts with the distortion augmentation strategy. Based on these distorted images, we operate the self-supervised pre-training by predicting the representations of patches below a certain position from those above it, where the quality of the representation learning is optimized in a contrastive manner.

A. Distortion Augmentation

The goal of the distortion augmentation is to derive efficient and promising self-supervision signal to conduct the proposed self-supervised pre-training, where the distorted samples with diverse distortion categories are first brought into being from the collected high-quality images. In specific, for a random input image sample X_i, we recursively apply distortion augmentations to generate its corresponding distorted counterparts $\tilde{X}_{i,1}, \tilde{X}_{i,2}, \ldots, \tilde{X}_{i,K}$, where K denotes the total number of the distortion categories. Afterwards, each distorted sample $\tilde{X}_{i,j}$ is divided into non-overlapping patches $\tilde{\mathbf{x}}_{i,j}^{(p,q)}$, where (p, q) denotes the location of the patch in the image.

To conduct the aforementioned "future observation" prediction task within a distorted sample, for each specific patch $\tilde{\mathbf{x}}_{i,j}^{(p,q)}$ (green boxes), we tend to leverage the feature representations corresponding to the patches in a $d \times d$ grid above it (blue boxes) to predict its representation (Fig. 1, right).

B. Feature Projection

Given the patch $\tilde{\mathbf{x}}_{i,j}^{(p,q)}$ from the distorted image $\tilde{X}_{i,j}$, a non-linear encoder function $f(\cdot)$ is introduced to map the input patch $\tilde{\mathbf{x}}_{i,j}^{(p,q)}$ to its latent representation $\mathbf{z}_{i,j}^{(p,q)}$, as:

$$\mathbf{z}_{i,j}^{(p,q)} = f\left(\tilde{\mathbf{x}}_{i,j}^{(p,q)}\right). \tag{1}$$

Based on the obtained feature representations from a $d \times d$ grid of patches that lie above the specific patch $\tilde{\mathbf{x}}_{i,j}^{(p,q)}$, we consider to leverage an aggregation function $g(\cdot)$ to integrate them into a context representation, denoted as $\mathbf{c}_{i,j}^{(p,q)}$:

$$\mathbf{c}_{i,j}^{(p,q)} = g\left(\{\mathbf{z}_{i,j}^{(u,v)}\}_{(u,v)}\right), \tag{2}$$

where $p-(d-1)/2 \le u \le p+(d-1)/2$, $q-d \le v \le q-1$. While human perception of distorted images is bound up with the combined effect of the salient local distortion and global quality degradation, it is reasonable to take into consideration the attention mechanisms in both local and global aspects when designing the aggregation function $g(\cdot)$. Inspired by recent successes achieved by vision Transformer architecture in many computer vision tasks [28]–[31], where representations from diverse local patches are explicitly correlated, we refer to the vision Transformer to formulate $g(\cdot)$. Specifically, several consecutive Swin Transformer [29] blocks, where the standard multi-head

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on June 07,2022 at 06:31:36 UTC from IEEE Xplore. Restrictions apply.

self attention (MSA) module in an original Transformer block is replaced by a module based on shifted windows with other layers kept the same, are combined to serve as the aggregation function $g(\cdot)$. As the input to the Transformer blocks, the patches in each grid are arranged from top to bottom and from left to right, whose numbers are clearly shown in Figure 1(b).

Unlike some existing IQA methods that rely on the vision Transformer structure to extract features in an end-to-end way [18]–[21], this study proposes a heuristic method to leverage the vision Transformer blocks to achieve a better integration of the feature representations from a certain grid of patches. Afterwards, a predictive function $\phi(\cdot)$ is introduced to take the context representation $\mathbf{c}_{i,j}^{(p,q)}$ as the input to predict the representation of the "future observation" in patch $\tilde{\mathbf{x}}_{i,j}^{(p,q)}$:

$$\hat{\mathbf{z}}_{i,j}^{(p,q)} = \phi\left(\mathbf{c}_{i,j}^{(p,q)}\right) = \phi\left(g\left(\{\mathbf{z}_{i,j}^{(u,v)}\}_{(u,v)}\right)\right). \tag{3}$$

The intuition behind the predictive task is that if the model can predict the representation of future observation from $\mathbf{c}_{i,j}^{(p,q)}$, then the context representation must have encoded strong semantics of the input image sample.

C. Contrastive Learning

The quality of such prediction is then evaluated using a contrastive loss, *i.e.*, the InfoNCE loss [32], [33]. Specifically, we refer to the prediction $\hat{\mathbf{z}}_{i,j}^{(p,q)}$ as the anchor in the contrastive learning, the goal is to correctly recognize its corresponding prediction target $\mathbf{z}_{i,j}^{(p,q)}$ (positive sample) among a set of unrelated negative samples $\{\mathbf{z}_l\}_l$, as:

$$\mathcal{L}_{ctr} = -\sum_{i,j,(p,q)} \log\left[\frac{h\left(\hat{\mathbf{z}}_{i,j}^{(p,q)}, \mathbf{z}_{i,j}^{(p,q)}\right)}{h(\hat{\mathbf{z}}_{i,j}^{(p,q)}, \mathbf{z}_{i,j}^{(p,q)}) + \sum_l h(\hat{\mathbf{z}}_{i,j}^{(p,q)}, \mathbf{z}_l)}\right], \tag{4}$$

where the function $h(\cdot)$ is implemented to measure the similarity between two representations as the following:

$$h(\mathbf{z}_1, \mathbf{z}_2) = exp\left(\frac{\phi(\mathbf{z}_1)\cdot\phi(\mathbf{z}_2)}{\|\phi(\mathbf{z}_1)\|_2 \cdot \|\phi(\mathbf{z}_2)\|_2}\cdot\frac{1}{\tau}\right), \tag{5}$$

where τ is a hyper-parameter that controls the range of the results, and ϕ is a learnable non-linear mapping. The loss function allows the positive pair to attract mutually while repelling the other items in the feature embedding space.

In this paradigm, the negative samples play a critical role since the effectiveness of the feature representations relies on how positive queries can be effectively distinguished from negative examples. To equip the network with the ability to learn feature representations which are both distortion-aware and content-independent, the selection of negative samples $\{\mathbf{z}_l\}_l$ is two-fold:

$$\begin{aligned}\sum_l h\left(\hat{\mathbf{z}}_{i,j}^{(p,q)}, \mathbf{z}_l\right) = &\sum_m \underbrace{h\left(\hat{\mathbf{z}}_{i,j}^{(p,q)}, \mathbf{z}_{i,m}^{(p,q)}\right)}_{\text{distortion}} + \sum_{(p,'q')} \underbrace{h\left(\hat{\mathbf{z}}_{i,j}^{(p,q)}, \mathbf{z}_{i,j}^{(p,'q')}\right)}_{\text{intra-content}} \\ &+ \sum_{i,'(m,n)} \underbrace{h\left(\hat{\mathbf{z}}_{i,j}^{(p,q)}, \mathbf{z}_{i,'j}^{(m,n)}\right)}_{\text{inter-content}},\end{aligned} \tag{6}$$

where we consider the feature representations corresponding to the patches in the same location as the positive from images suffered from different distortion conditions $\mathbf{z}_{i,m}^{(p,q)}(m \neq j)$ as the distortion negatives. On the other hand, patches taken from other locations in the image $\mathbf{z}_{i,j}^{(p,'q')}((p,' q') \neq (p,q))$ are considered as the intra-content negatives while those from other images $\mathbf{z}_{i,'j}^{(m,n)}(i' \neq i)$ are considered as the inter-content negatives. By pulling the anchor closer to the positive while pushing away from the both the distortion and content negatives in the latent space, the learned feature representations are expected to have the capacity to capture the quality-aware information.

III. Experimental Results

A. Experimental Protocols

To demonstrate the superiority of our pre-training models, we rely on two public IQA databases, KADID-10k [25] with synthetically distorted images and KonIQ-10k [34] with authentically distorted images, for their representativeness. In our experiments, the Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank-order Correlation Coefficient (SRCC) are used to evaluate the accuracy and the monotonicity of the predictions, respectively.

When performing distortion augmentation, we refer to images from the MS-COCO dataset [35] targeting on image segmentation as the high-quality samples, and rely on diverse distortion categories such as motion blur, Gaussian blur, overexposure or underexposure, white noise and JPEG compression to generate distorted images for pre-training the proposed method. For optimization, all models are trained end-to-end using the Adam optimizer to minimize the widely-used ℓ_2 loss. An initial learning rate of $5e-4$ decayed by a factor of 0.2 every 20 epochs is used during the training process. We consider the widely-used ResNet-50 network as the feature encoder f and empirically use four Swin Transformer blocks to achieve the function $g(\cdot)$.

B. Performance Evaluation

To evaluate the effectiveness of the proposed pre-trained model for IQA, we take two widely-used evaluation protocols including fine-tuning and linear evaluation.

Finetuning: We evaluate the effectiveness of the proposed SPIQ as the pre-trained model with two representative supervised IQA baselines, *i.e.*, MetaIQA [17] and HyperIQA [16]. Specifically, the backbone network pre-trained on ImageNet is first replaced with our SPIQ, then the model is finetuned for another 30 epochs on the labeled training set. Table I exhibits the prediction improvements over both test datasets. It can be clearly observed that, equipped with our proposed pre-trained model, both the baselines could achieve more than 0.0313/0.0317 and 0.0269/0.0393 improvements in terms of PLCC/PLCC (marked in green) on two test datasets, respectively. We also provide several other top-performing IQA methods (BRISQUE [8], CORNIA [9], HOSA [10], BIECON [13], DBCNN [15], TRIQ [20] and MUSIQ [18]) to highlight the performance gains with our SPIQ pre-trained model. These results indicate that the feature representation obtained from our method is more effective compared with that pre-trained on ImageNet.

Linear evaluation: We then benchmark the learned representations following the linear evaluation protocol defined in [41], to validate the effectiveness of the pre-trained models without further finetuning. After pre-training on the unlabeled training set for 100 epochs, the feature encoder is freezed and

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on June 07,2022 at 06:31:36 UTC from IEEE Xplore. Restrictions apply.

TABLE I
PERFORMANCE EVALUATION OF THE PROPOSED PRE-TRAINED MODEL ON TWO TEST DATASETS. IN THE BRACKETS ARE THE GAPS TO THE IMAGENET SUPERVISED PRE-TRAINING COUNTERPART. (†) WE REPLACE THE BACKBONE NETWORK IN THE ORIGINAL IMPLEMENTATION FROM RESNET-18 TO RESNET-50. (*) WE DIRECTLY TAKE THE NUMBERS FROM THEIR PAPERS FOR COMPARISON

Database	KADID-10k [25]		KonIQ-10k [34]	
Method \ Criterion	PLCC↑	SRCC↑	PLCC↑	SRCC↑
BRISQUE [8]	0.5733	0.5302	0.6714	0.6262
CORNIA [9]	0.5971	0.5842	0.7361	0.7145
HOSA [10]	0.6384	0.6451	0.6895	0.6728
BIECON [13]	0.6532	0.6360	0.6339	0.6108
DBCNN [15]	0.8126	0.8165	0.8603	0.8487
MetaIQA† [17]	0.8461	0.8533	0.8877	0.8654
HyperIQA [16]	0.8517	0.8590	0.8842	0.8726
TRIQ [20]	–	–	0.925*	0.907*
MUSIQ-single [18]	–	–	0.919*	0.905*
MUSIQ-multi [18]	–	–	0.928*	0.916*
MetaIQA + *SPIQ*	0.8774 (+0.0313)	0.8864 (+0.0331)	0.9146 (+0.0269)	0.9060 (+0.0406)
HyperIQA + *SPIQ*	0.8882 (+0.0365)	0.8907 (+0.0317)	0.9208 (+0.0366)	0.9119 (+0.0393)

TABLE II
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART SELF-SUPERVISED METHODS UNDER THE LINEAR EVALUATION PROTOCOL

Database	KADID-10k [25]		KonIQ-10k [34]	
Method \ Criterion	PLCC↑	SRCC↑	PLCC↑	SRCC↑
ImageNet Pre-train	0.5437	0.5589	0.5906	0.6064
CPC v2 [36]	0.6244	0.6093	0.6515	0.6561
MoCo v2 [37]	0.6030	0.6178	0.6439	0.6218
SimCLR [38]	0.5871	0.5545	0.6106	0.6097
BYOL [39]	0.6282	0.6314	0.6638	0.6306
SwAV [40]	0.6078	0.5970	0.6414	0.6392
SPIQ (ours)	**0.7434**	**0.7519**	**0.7606**	**0.7493**

a linear regressor consisting of two fully-connected layers is further trained for 30 epochs. We also include other famous self-supervised baselines (CPC v2 [36], MoCo v2 [37], SimCLR [38], BYOL [39] and SwAV [40]) for comparison. As shown in Table II, our method provides a significant boost of 0.1152/0.1205 and 0.0968/0.1187 in terms of PLCC/SRCC over the best competitor (BYOL) on two test datasets, respectively. We stress that there are few algorithms in the literature that work well on both synthetic and authentic distortion settings, and our SPIQ is very competitive in that sense.

In Figure 2, we try to understand what SPIQ learns by visualizing the feature activation maps [42] of some representative samples from the KonIQ-10 k dataset. These maps indicate where the learned feature representations focus on under the linear evaluation setting. Compared with the ImageNet pre-trained model that is more sensitive to foreground objects, the activation maps of SPIQ pre-trained representations contain higher attention values at the distorted regions, which are of prime importance for visual quality perception.

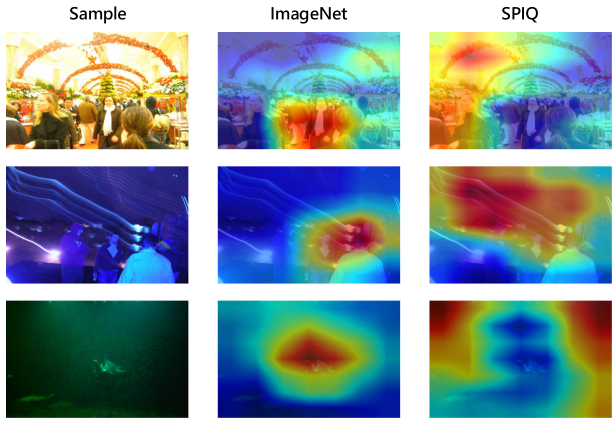

Fig. 2. Visualization of activation maps using the selected samples from KonIQ-10 k dataset.

TABLE III
ALATION ANALYSIS ON DIFFERENT PATCH AND GRID SIZES

Setup		Datasets	
Patch Size	Grid Size	KADID-10k	KonIQ-10k
8	{5×5}	0.6961	0.7244
16	{5×5}	0.7310	0.7525
32	{5×5}	**0.7434**	**0.7606**
64	{5×5}	0.7367	0.7578
32	{1×1}	0.5738	0.6312
32	{3×3}	0.7185	0.7492
32	{5×5}	**0.7434**	**0.7606**
32	{7×7}	0.7406	0.7417
32	{9×9}	0.7243	0.7338

C. Ablation Study

In this experiment, we conduct ablation studies to analyze the performances of the proposed SPIQ under different patch and grid sizes. As shown in Table III, when the grid size is first fixed to 5×5 the performance gradually increases from 0.6961/0.7244 of patch size 8 to 0.7434/0.7606 of patch size 32, indicating that larger patch size could enlarge the receptive field which improves the expressivity of the learned feature representations. However, when it further increases to patch size 64, the content contained in such a patch is too much to be well perceived, making the model confuse to learn discriminative representations. The same trend can be observed when we fix the patch size and change the grid size. Therefore, we adopt a combination of patch size 32 and grid size 5×5 in other experiments.

IV. CONCLUSION

In this paper, we have proposed to learn a general pre-trained BIQA model from the abundant unlabeled image data. On the basis of each distorted image generated with the designed distortion augmentation strategy, the proposed pre-training is operated with a patch representation prediction task. This task is regularized by a contrastive loss, aiming to learn useful representations that are both distortion-aware and content-independent. Experiments demonstrated that the proposed SPIQ could significantly benefit the supervised IQA approaches serving as the pre-trained model.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on June 07,2022 at 06:31:36 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[2] J. Wu, W. Lin, and G. Shi, "Image quality assessment with degradation on spatial structure," *IEEE Signal Process. Lett.*, vol. 21, no. 4, pp. 437–440, Apr. 2014.

[3] S. Wang, K. Ma, H. Yeganeh, Z. Wang, and W. Lin, "A patch-structure representation method for quality assessment of contrast changed images," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2387–2390, Dec. 2015.

[4] Y. Huang, L. Li, H. Zhu, and B. Hu, "Blind quality index of depth images based on structural statistics for view synthesis," *IEEE Signal Process. Lett.*, vol. 27, pp. 685–689, 2020.

[5] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "SpEED-QA: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1333–1337, Sep. 2017.

[6] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[7] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[8] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[9] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 1098–1105.

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

[11] P. Chen, L. Li, X. Zhang, S. Wang, and A. Tan, "Blind quality index for tone-mapped images based on luminance partition," *Pattern Recognit.*, vol. 89, pp. 108–118, 2019.

[12] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[13] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206–220, Feb. 2017.

[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 Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

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

[16] S. Su *et al.*, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 3667–3676.

[17] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 14 143–14 152.

[18] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *Proc. IEEE Int. Conf. Comput. Vision*, 2021, pp. 5148–5157.

[19] M. Cheon, S.-J. Yoon, B. Kang, and J. Lee, "Perceptual image quality assessment with transformers," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshop*, 2021, pp. 433–442.

[20] J. You and J. Korhonen, "Transformer for image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 1389–1393.

[21] L. Li, T. Song, J. Wu, W. Dong, J. Qian, and G. Shi, "Blind image quality index for authentic distortions with local and global deep feature aggregation," *IEEE Trans. Circuits Syst. Video Technol*., early access, 13 Sep. 2021, doi: 10.1109/TCSVT.2021.3112197.

[22] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.

[23] P. Chen, L. Li, L. Ma, J. Wu, and G. Shi, "RIRNet: Recurrent-in-recurrent network for video quality assessment," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 834–842.

[24] H. Liu, U. Engelke, J. Wang, P. Le Callet, and I. Heynderickx, "How does image content affect the added value of visual attention in objective image quality assessment?," *IEEE Signal Process. Lett.*, vol. 20, no. 4, pp. 355–358, Apr. 2013.

[25] H. Lin, V. Hosu, and D. Saupe, "KADID-10 k: A large-scale artificially distorted IQA database," in *Proc. Int. Conf. Quality Multimedia Exper.*, 2019, pp. 1–3.

[26] Y. Fang, J. Yan, L. Li, J. Wu, and W. Lin, "No reference quality assessment for screen content images with both local and global feature representation," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1600–1610, Apr. 2018.

[27] P. Chen, L. Li, J. Wu, W. Dong, and G. Shi, "Unsupervised curriculum domain adaptation for no-reference video quality assessment," in *Proc. IEEE Int. Conf. Comput. Vision*, 2021, pp. 5178–5187.

[28] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[29] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[30] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vision*, Springer, 2020, pp. 213–229.

[31] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2021, pp. 6881–6890.

[32] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[33] P. Chen, L. Li, J. Wu, W. Dong, and G. Shi, "Contrastive self-supervised pre-training for video quality assessment," *IEEE Trans. Image Process.*, vol. 31, pp. 458–471, 2022.

[34] H. Lin, V. Hosu, and D. Saupe, "KonIQ-10 k: Towards an ecologically valid and large-scale IQA database," 2018, *arXiv:1803.08489*.

[35] T.-Y. Lin *et al.*, "Microsoft Coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vision*, Springer, 2014, pp. 740–755.

[36] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *Proc. Int. Conf. Mach. Learn., PMLR*, 2020, pp. 4182–4192.

[37] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[38] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn., PMLR*, 2020, pp. 1597–1607.

[39] J.-B. Grill *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," 2020, *arXiv:2006.07733*.

[40] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," 2020, *arXiv:2006.09882*.

[41] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit*, 2019, pp. 1920–1929.

[42] N. Komodakis and S. Zagoruyko, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Int. Conf. Learn. Represent.*, 2017, p. 1.

Authorized licensed use limited to: University Town Library of Shenzhen. Downloaded on June 07,2022 at 06:31:36 UTC from IEEE Xplore. Restrictions apply.