Image Quality Assessment using Semi-Supervised Representation Learning

Vishnu Prabhakaran
Amazon, Bangalore, India
visprab@amazon.com

Gokul Swamy
Amazon, Seattle, USA
swagokul@amazon.com

Abstract

In this paper, we propose a framework for learning feature representations for Image Quality Assessment (IQA) using contrastive learning. To account for the absence of large-scale IQA dataset, we pretrain an image encoder to cluster images based on the image quality using synthetically distorted versions of pristine unlabeled images. Images of similar quality are grouped closer in embedding space, while simultaneously pushing apart images of dissimilar quality. In addition we show that, augmenting the contrastive learning task with downstream aware joint supervision results in feature representations that are more suitable and easily transferable for IQA specific tasks. We study the effectiveness of the learnt representations in downstream task of image quality prediction and show that our model achieves superior performance on both synthetically and authentically distorted IQA datasets when compared to other deep feature-based IQA methods.

1. Introduction

With the advent of low-cost and fast image capture mechanisms, images have become ubiquitous and abundant in quantity. However, the quality of images vary drastically depending on the process of image capture, storage and transfer. Image quality is a significant factor that determines the usefulness of captured images for various applications and quality requirements remain application specific. Conventionally, practitioners have resorted to using subjective user ratings as a direct measure of visual quality of images [1]. The quality ratings are obtained through user studies conducted in controlled lab setting with multiple groups of human participants, resulting in a process which is time-consuming, expensive and not real-time applicable.

Image Quality Assessment (IQA) [2] is a computer vision technique to estimate the perceptual quality of images objectively, i.e by predicting the Mean Opinion Scores (MOS) using computational models. Depending on the availability of reference high quality images, objective IQA methods can be broadly classified into: Full-Reference IQA (FR-IQA) and Blind IQA (BIQA). The latter is more challenging and is practically suitable for many applications due to the lack of reference images in general. BIQA is used in applications such as image enhancement [3], image retrieval [4], image ranking [5] and evaluating image capture pipelines.

Due to the wide success of Convolutional Neural Networks (CNNs) in vision domain [6, 7, 8], recent BIQA works have adopted CNN architectures to their solutions resulting in better performance than conventional methods. A common setting in most of these approaches include a CNN encoder for feature extraction, followed by Multi Layer Perceptron (MLP) regressor to predict quality score i.e. MOS. However, the unavailability of large-scale IQA dataset with quality ratings restricts the generalization capability of these CNN based models. The largest IQA dataset till date contains only 10,073 quality scored images [9]. Transfer learning from image classification domain (ImageNet pretrained CNN encoder and fine-tuning) is not suitable as the feature representations of an ImageNet pretrained encoder are invariant to common image transformations such as color-jitter, blur, resize, noise, etc. that degrades image quality.

A common practice in such low-data regimes is to pretrain the network in a task-specific way with unlabeled data using self-supervised objective, followed by supervised fine-tuning with labeled data. Unlike in image classification domain, designing a pretext task to learn features suitable for image quality prediction (a regression task) is non-trivial. Distortion classification i.e predicting the type and severity of synthetic distortion applied to an image is a recently adopted auxiliary task to learn features, but without much success in the downstream IQA regression task. Prior research works [10] [11] have shown that pretraining does not work well when the downstream task is very different. Due to this inherent mismatch between pretrain and downstream tasks, complex transfer learning strategies that does not distort the pretrained features are necessary to achieve good performance on out-of-distribution datasets. In this work, we demonstrate a semi-supervised learning setup that employs joint-training with self-supervised and supervised objectives to help resolve the mismatch between pretrain and downstream tasks. To this end, we train an image encoder on two

sub-tasks jointly: (i) contrastive representation learning task on unlabeled data that aims to group images based on their distortion type and severity (ii) regression task on labeled data to predict image quality score. Our main findings and contributions are summarized as follows:

- A novel self-supervised contrastive learning framework for learning representations for IQA using stochastic synthetic distortions of pristine images.
- Our empirical results suggest that the semi-supervised pretraining scheme learns suitable feature representations that generalizes well to downstream IQA tasks on both in-distribution and out-of-distribution datasets.
- An in-depth analysis of effectiveness of the learnt representations on downstream IQA task under different evaluation modes.

2. Related Work

2.1. Blind Image Quality Assessment

Blind image quality assessment have been studied extensively and is a long-standing research problem [12]. Traditional BIQA methods [13, 14] focused on hand crafted features by domain experts that convert raw image into a representative vector and learn a non-linear mapping to quality scores. Alternatively, few works [15, 16, 17] focused on Natural Scene Statistics (NSS) to estimate image quality. Due to absence of large dataset, deep learning based methods [18, 19, 5] use transfer learning of pretrained models on IQA dataset. In [20] [21], the authors train a CNN from scratch by using 32×32 image patches, assuming that the sampled patches have same quality score as the original image. To address the limited IQA dataset, recent works propose using different pretraining strategies [22, 23, 24]. RankIQA [22] used a Siamese Network to rank image pairs that are artificially distorted. DeepFL-IQA [23] proposed a weakly supervised feature learning approach that utilizes FR-IQA scores of artificially distorted images as proxy scores to train on a large unlabeled dataset in supervised manner. MEON [24] is pretrained for distortion identification task. CONTRIQUE [25] discusses a contrastive learning approach to learn representations with distortion classification as auxillary task. Unlike the recent works, our method simultaneously solves a contrastive learning task and a regression task in a semi-supervised setup to learn representations that are more suitable and generalizes well to downstream IQA tasks without complex fine-tuning strategies.

2.2. Contrastive Representation Learning

Self-supervised methods for visual representation learning can be broadly categorized into two types: generative and discriminative. Generative methods tries to model the distribution over pixels in the input space [26], while discriminative methods design pretext tasks using pseudo labels to learn representations from unlabeled dataset. These pretext tasks include predicting image rotations [27], relative position of image patches [28], solving jigsaw puzzles [29], etc that have shown to learn good representations. Contrastive learning is a discriminative approach that learns an embedding space where data points from same class are closer to each other while data points from different classes are far apart. Recent approaches introduce modified frameworks casting the pretext task differently such as cluster assignment [30], dynamic dictionary look-up [31], maximizing mutual information across augmented views [32], etc. In *SimCLR* [33] [34], the authors proposed a simple framework that learns representations by maximizing agreement between differently augmented views of the same sample. Structurally our contrastive learning task is similar to [34], but differs conceptually as we aim to maximize agreement between images of similar quality and learns representations suitable for IQA.

3. Method

Our method learns feature representations for IQA by utilizing a combination of labeled and unlabeled samples in a semi-supervised setting. The setup uses a joint training strategy that comprises of (i) self-supervised contrastive learning task for unlabeled samples and (ii) supervised regression task using labeled samples. Figure 1 illustrates our proposed method.

3.1. Contrastive Representation Learning

The task is to maximize agreement between images of same distortion class. The image distortion module transforms the unlabeled input pristine image (x^u) into a corrupted version (x^d) by randomly choosing a distortion type (d_t) and severity (d_s). We use the *imgaug* [35] library to generate artificial distortions by applying a combination of color and geometric transformations: blur, noise, pixelate, color-jitter, jpeg compression, etc. (for details, refer Figure 5 in Appendix). We employ 25 distortion types with 5 severity degrees, thus resulting in 125 classes. The image distortion types and degradation levels are configured and chosen depending on the downstream task. These distortions capture most of the real world image distortions that occur naturally during image capture, storing and transfer.

An encoder network $f_{\theta_1}(.)$ extracts the feature representations of two augmented views (x_i^d, x_j^d) of the distorted image (x^d). Here, we use transforms that do not affect the image quality to generate two different views of the image. In particular, we use random crop, horizontal and vertical flip augmenters. A shallow MLP network $g_{\theta_2}(.)$ projects the feature representations to a lower dimensional space before

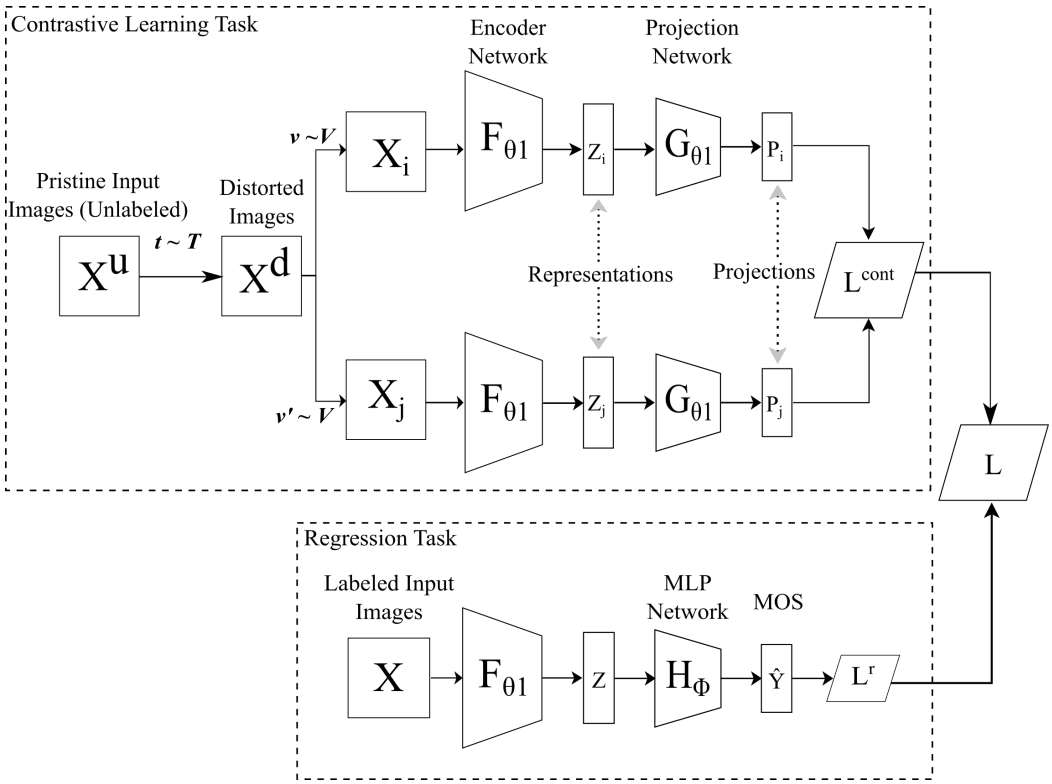


Figure 1: Semi-supervised pretraining pipeline to learn feature representations for Image Quality Assessment (IQA). (*a*) For unlabeled pristine images, we randomly apply distortions ($t \sim T$) with different severity levels and generate two augmented views (v, $v' \sim V$) of resulting distorted images. An encoder transforms the image pairs to latent representations and the contrastive loss is computed over their lower dimensional projections. (*b*) For quality scored images, the shared image encoder F_{θ_1} followed by MLP network predicts quality score and regression loss is computed. The final loss term is a weighted combination of L^{cont} and L^r.

applying the contrastive loss L^{cont}.

$$z_i^d = f_{\theta_1}(x_i^d) \tag{1}$$

$$z_j^d = f_{\theta_1}(x_j^d) \tag{2}$$

$$p_i^d = g_{\theta_2}(z_i^d) \tag{3}$$

$$p_j^d = g_{\theta_2}(z_j^d) \tag{4}$$

In this work, $f_{\theta_1}(.)$ is the standard ResNet-50 architecture [7], $z_i^d, z_j^d \in R^d$ are outputs of its average pooling layer, $g_{\theta_2}(.)$ is a 2-layer MLP network (with 2048) to project the representation to a 128-dimensional latent space and $p_i^d, p_j^d \in R^{d'} (d' < d)$ are the projected vectors. We employ the supervised version of the normalized temperature-scaled cross entropy loss (NT-Xent) [34] as the contrastive loss function. The contrastive loss function $l^{cont}(\theta_1, \theta_2)$ is defined as

$$\frac{1}{|P(i)|} \sum_{j \in P(i)} -\log \frac{\exp(\text{sim}(p_i^d, p_j^d)/\tau)}{\sum_{k=1}^{N_U} \mathbb{1}_{k \neq i} \exp(\text{sim}(p_i^d, p_k^d)/\tau)} \tag{5}$$

where $\text{sim}(a, b)$ is the cosine similarity between two projected vectors, τ denotes the temperature parameter, N_U is the number of unlabeled images in the batch, $\mathbb{1}$ is the indicator function, $P(i)$ is a set of indices of all positive samples belonging to same distortion class as p_i^d in the batch and $|P(i)|$ is its cardinality. The final loss L^{cont} is the average of l^{cont} computed for all classes in the batch.

3.2. Semi Supervised Learning

The regression network shares the ResNet-50 encoder $f_{\theta_1}(.)$, followed by a 2-layer MLP network (with 512 hidden units) $h_\phi(.)$ and outputs a single scalar value indicating the perceptual image quality (usually MOS). Let $\{x_i, y_i\}_{i=1}^{N_L}$ denote the labeled samples in the training batch where x_i is the input image and y_i is the ground truth quality score. We use the Mean Squared Error (MSE) as loss function for the

regression sub-task.

$$\hat{y}_i = h_\phi(f_{\theta_1}(x_i)) \tag{6}$$

$$L^r(\theta_1, \phi) = \frac{1}{N_L} \sum_{i=1}^{N_L} (y_i - \hat{y}_i)^2 \tag{7}$$

Let N be the training mini-batch size, including both labeled N_L and unlabeled samples N_U. We define the semi-supervised objective function of the form:

$$L(\theta_1, \theta_2, \phi) = \alpha L^{cont} + (1 - \alpha) L^r \tag{8}$$

where α is a chosen hyperparameter for differential weighting of the two loss terms.

4. Experimental Results

In this section, we enclose the details of the datasets, experimental setups and evaluate the effectiveness of the learnt representations in downstream tasks: BIQA and image quality verification for RLR.

4.1. Datasets

For the semi-supervised pretraining setup, we use KADIS dataset [23] that contains 140,000 unlabeled pristine images (80% train and 20% validation set) and KADID-10k dataset [36] that contains 10,125 artificially distorted images with subjective quality scores (60% train, 20% validation and 20% test set). The raw source images for both datasets exhibit similar distribution of images.

For downstream BIQA, we use KADID-10k [36] (in-distribution) and KonIQ-10k [9] (out-of-distribution) datasets. KonIQ-10k contain 10,073 images with authentic distortions and their Mean Opinion Scores (MOS). MOS are scaled to a standard range of [1, 10], where higher scores corresponds to better quality. For easier comparison to existing BIQA methods, we follow the split strategy used in [23]: training (60%), validation (20%) and test (20%) sets.

4.2. Setup

Representation Learning For the joint training using labeled and unlabeled dataset, we use PyTorch Lightning's [37] `CombinedLoader()` that helps to combine the two dataloaders and allows sampling in parallel for training. We use mini-batch sizes of 1000 for KADIS dataset and 32 for KADID-10k dataset. We train for 100 epochs using LARS optimizer [38] with a initial learning rate of 10^{-4} for regression MLP network and 10^{-2} for the image encoder and projector networks. A weight decay of 10^{-6} is used. Based on hyper parameters search, we found $\tau = 0.5$ and $\alpha = 0.9$ works well. Learning rate is dynamically reduced by a decay factor of 10^{-1} whenever learning stagnates by monitoring the validation loss. We employ early stopping mechanism to avoid over-fitting. Input images are randomly cropped to standard 224×224×3 and subject to random horizontal and vertical flipping augmentation.

BIQA For BIQA, we resort to CNN based image encoder stacked with 2-layer MLP regressor network (with 512 hidden units). To analyze the effectiveness of learnt representations, we use the pretrained image encoder + MLP regressor in three evaluation modes:

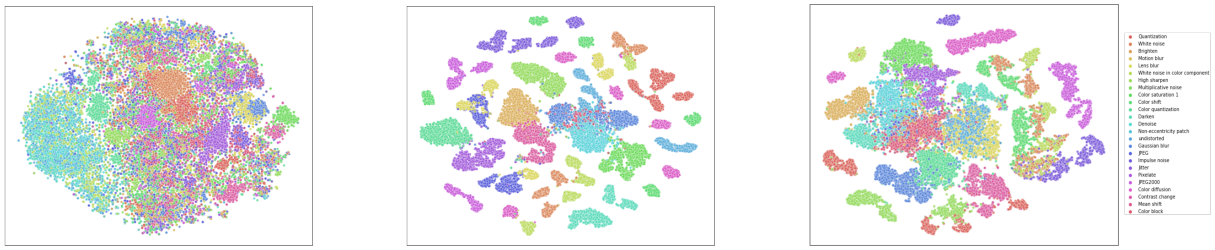
1. Linear Probing (LP) - updating only the MLP network
2. Fine Tuning (FT) - updating both encoder and MLP networks
3. Linear Probing + Fine Tuning (LP-FT) - linear probing followed by fine-tuning

The network is trained for 50 epochs at mini-batch size of 32 using MSE loss function. We use Adam optimizer [39] with learning rate of 10^{-5}. We evaluate the IQA models using two metrics: the Spearman Rank Order Correlation Coefficient (SROCC) and the Pearson Linear Correlation Coefficient (PLCC), computed between predicted and ground truth MOS.

4.3. Performance Evaluation

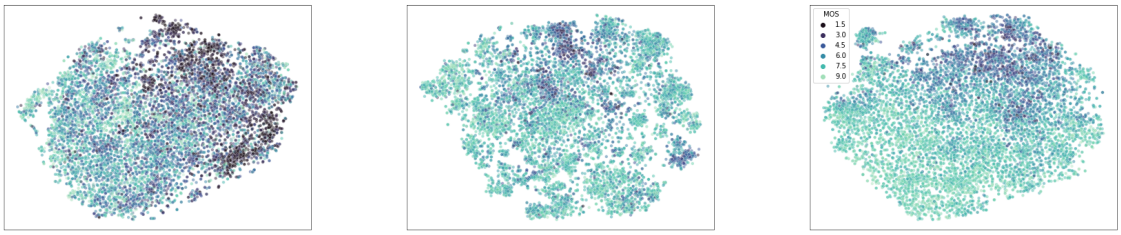
We compare the performance of three image encoders: (i) *SL(ImageNet)* that uses the pretrained encoder trained using supervised learning for ImageNet classification task, (ii) *CL* which uses pretrained weights trained using the contrastive learning approach only (refer Section 3.1) and (iii) *SSL* that uses pretrained weights trained using the semi-supervised learning (refer Section 3.2) approach. The *CL* and *SSL* encoders are trained seperately. Figure 2 shows the t-SNE visualization of the 2048-dimensional embeddings of randomly distorted KADIS validation images from these encoders. Firstly, we observe that the hidden vectors from ImageNet pretrained encoder fails to distinguish images based on their quality. The representations learnt using our contrastive learning setting forms distinct clusters of different distortion classes and shows superior performance. The features from our proposed semi-supervised learning scheme not only form clusters of distortion types but also groups images of similar quality closer. This is very evident in Figure 3 where we visualize the feature vectors on KonIQ dataset color-mapped with ground truth MOS. We notice that the encoder trained with semi-supervised objective is able to group images with similar quality scores (gradually increasing from top to bottom).

In Table 1, we analyze the effectiveness of learnt representations on different transfer learning strategies. We note that LP-FT performs better than LP and FT in most cases. We find that the IQA specific pretraining helps the models to perform better than naive transfer learning from object



(a) SL(ImageNet) (b) CL (c) SSL

Figure 2: t-SNE visualizations of representations learned using different pretraining schemes. The data points comprises of randomly distorted images from KADIS validation set. The learnt representations via contrastive and semi-supervised learning schemes form vivid distinct clusters, while the representations from ImageNet pretrained encoder fails to distinguish different distortion types. Note that each distortion type has 5 different severity levels leading to multiple clusters of the same color.



(a) SL(ImageNet) (b) CL (c) SSL

Figure 3: t-SNE visualizations of latent features from different pretrained encoders on KonIQ dataset color-mapped with their ground truth quality scores. The encoder trained in semi-supervised manner is better at grouping images with similar quality scores.

classification domain. *SSL* shows superior performance on all modes which largely benefits from downstream consistent semi-supervised pretraining. Our model improves the SROCC over the *SL(ImageNet)* baseline by 0.06 and 0.13 on both datasets respectively. Further, we probe the performance of IQA models on predicting the quality scores on four commonly observed real-world distortion types In Table 2.

We compare the performance of our fine-tuned IQA model *SSL* with state-of-the-art BIQA methods in Table 3. For conventional BIQA methods including BIQI [15], BLIINDS-II [17], BRISQUE [40], CORNIA [41], DIIVINE [16] and HOSA [42], we report the experimental results published in [23]. As reported in [23], a Support Vector Regressor (SVR) was trained to predict quality scores from the extracted handcrafted features from the above methods. Among the conventional BIQA methods, CORNIA and HOSA shows good improvement by using local features to predict quality scores in comparison to methods based on global features. For recent deep learning based methods, we reimplemented RankIQA [22], MEON [24] and KonCept224 [9]. To ensure fair comparison, for RankIQA and MEON we follow their pretraining procedure on KADIS dataset using our synthetic distortion module (refer Section 3.1). Overall performance of deep learning based methods are significantly better compared to conventional methods. KonCept512 [9] achieves the best performance on KonIQ-10k dataset suggesting the usage larger resolution images (512×384) helps the IQA models. However, using high resolution input images is computationally expensive for CNN models and leads to poor generalization capabilities. KonCept224 is another model version from [9] that uses downscaled input images (224×224) similar to our setting. The weak supervision using FR-IQA model scores on unlabeled dataset helps DeepFL-IQA [23] method to achieve best performance on KADID-10k dataset. CONTRIQUE (as

Method	KonIQ-10k		KADID-10k	
	SROCC	PLCC	SROCC	PLCC
LP:				
SL(ImageNet)	0.77	0.73	0.70	0.73
CL	0.79	0.80	0.84	0.85
SSL	**0.84**	**0.83**	**0.88**	**0.87**
FT:				
SL(ImageNet)	0.82	0.80	0.75	0.76
CL	0.81	0.82	0.83	0.83
SSL	**0.91**	**0.89**	**0.90**	**0.89**
LP-FT:				
SL(ImageNet)	0.84	0.83	0.78	0.79
CL	0.86	0.86	0.84	0.83
SSL	**0.90**	**0.89**	**0.91**	**0.90**

Table 1: Performance comparison of BIQA models in different evaluation modes

Method	Gaussian Blur	JPEG	JPEG 2000	White Noise
SL(ImageNet)	0.79	0.82	0.79	0.78
CL	0.84	0.83	0.84	0.85
SSL	**0.88**	**0.86**	**0.90**	**0.87**

Table 2: Performance evaluation (SROCC) on KADID-10k dataset on commonly observed distortions types

reported in [25]) shows promising results on both datasets, however the authors use multiple input data transformations (multiscale features, color and patch transforms) increasing the complexity. Moreover, their pre-training uses both sythetically and authentically distorted images comprising of 1.3 million images (9x higher than our training data). For fair comparison, CONTRIQUE-syn [25] that uses only sythetically distorted images in its pretraining regime performs worse than our *CL* baseline model. Our model *SSL* achieves second-best performance on both datasets, showing the effectiveness of joint training strategy for pretraining. Our model exhibits consistently good performance on both artificially and authentically distorted IQA datasets, unlike the other IQA models. Since the performance of batch contrastive learning is highly effective with larger batch sizes, we stick to standard resolution images for computation reasons. Figure 4 shows the MOS model predictions on a few exemplary images, ranked based on the predicted quality scores.

Method	KonIQ-10k		KADID-10k	
	SROCC	PLCC	SROCC	PLCC
BIQI [15]	0.68	0.70	0.42	0.44
BLIINDS-II [17]	0.60	0.61	0.53	0.55
BRISQUE [40]	0.70	0.71	0.52	0.56
CORNIA [41]	0.75	0.77	0.51	0.55
DIIVINE [16]	0.70	0.71	0.47	0.52
HOSA [42]	0.79	0.80	0.61	0.65
RankIQA [22]	0.81	0.83	0.79	0.81
MEON [24]	0.87	0.89	0.87	0.85
DeepFL-IQA [23]	0.87	0.88	**0.93**	**0.93**
KonCept224 [9]	0.86	0.88	0.80	0.81
KonCept512 [9]	**0.91**	**0.92**	0.84	0.85
CONTRIQUE [25]	0.89	0.90	0.93	0.93
CONTRIQUE-syn [25]	0.85	-	-	-
SL(ImageNet)	0.84	0.83	0.78	0.79
CL	0.86	0.86	0.84	0.83
SSL	0.90	0.89	0.91	0.90

Table 3: Performance comparison with existing BIQA models on two public datasets

5. Summary

In this work, we introduced a semi-supervised representation learning framework for IQA to address the limitations due to absence of large-scale IQA dataset. We use a joint training strategy that uses both unlabeled and labeled images. We apply stochastic image distortions over unlabeled pristine images to generate multiple positive and negative examples of different distortion classes in an online fashion and define a contrastive predictive task to group images based on their distortion type and severity. The joint regression sub task over labeled images guides the pretraining to learn downstream aware feature representations. We show that the representations learnt via our pretraining scheme is more suitable and effective for IQA-specific downstream tasks in comparison to other methodologies. Our best IQA model achieves consistent superior performance on both artificially and authentically distorted IQA datasets compared to existing state-of-the-art methods. The broader impact of this research work can be realized by integrating our proposed model to existing/new image capture, upload and verification pipelines where image quality is absolutely critical, thereby reducing manual intervention and process time.

A. Synthetic Image Distortions

We use 25 distortion types that simulate real world quality defects in images during image capture, storage and transfer. These include the following: Gaussian blur, Lens blur,

Figure 4: Example images from KonIQ-10k test set with predicted (and ground truth) scores shown below each image.

Motion blur, Color diffusion, Color shift, Color quantization, Color saturation 1, Color saturation 2, JPEG2000, JPEG, White noise, White noise in color component, Impulse noise, Multiplicative noise, Denoise, Brighten, Darken, Mean shift, Jitter, Non-eccentricity patch, Pixelate, Quantization, Color block, High sharpen and Contrast change. Each distortion type is applied at random on input raw pristine image with one of the 5 severity levels.

References

[1] Zhou Wang and Alan C. Bovik. *Modern Image Quality Assessment*. Morgan Claypool Publishers, San Rafael, CA, USA, 2006.

[2] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[3] Hossein Talebi and Peyman Milanfar. Learned perceptual image enhancement. In *2018 IEEE International Conference on Computational Photography (ICCP)*, pages 1–13, 2018.

[4] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. A learning-to-rank approach for image color enhancement. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 2987–2994. IEEE Computer Society, 2014.

[5] Hossein Talebi Esfandarani and Peyman Milanfar. NIMA: neural image assessment. *CoRR*, abs/1709.05424, 2017.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[9] 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*, 29:4041–4056, 2020.

[10] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D. Cubuk, and Quoc V. Le. Rethinking pre-training and self-training, 2020.

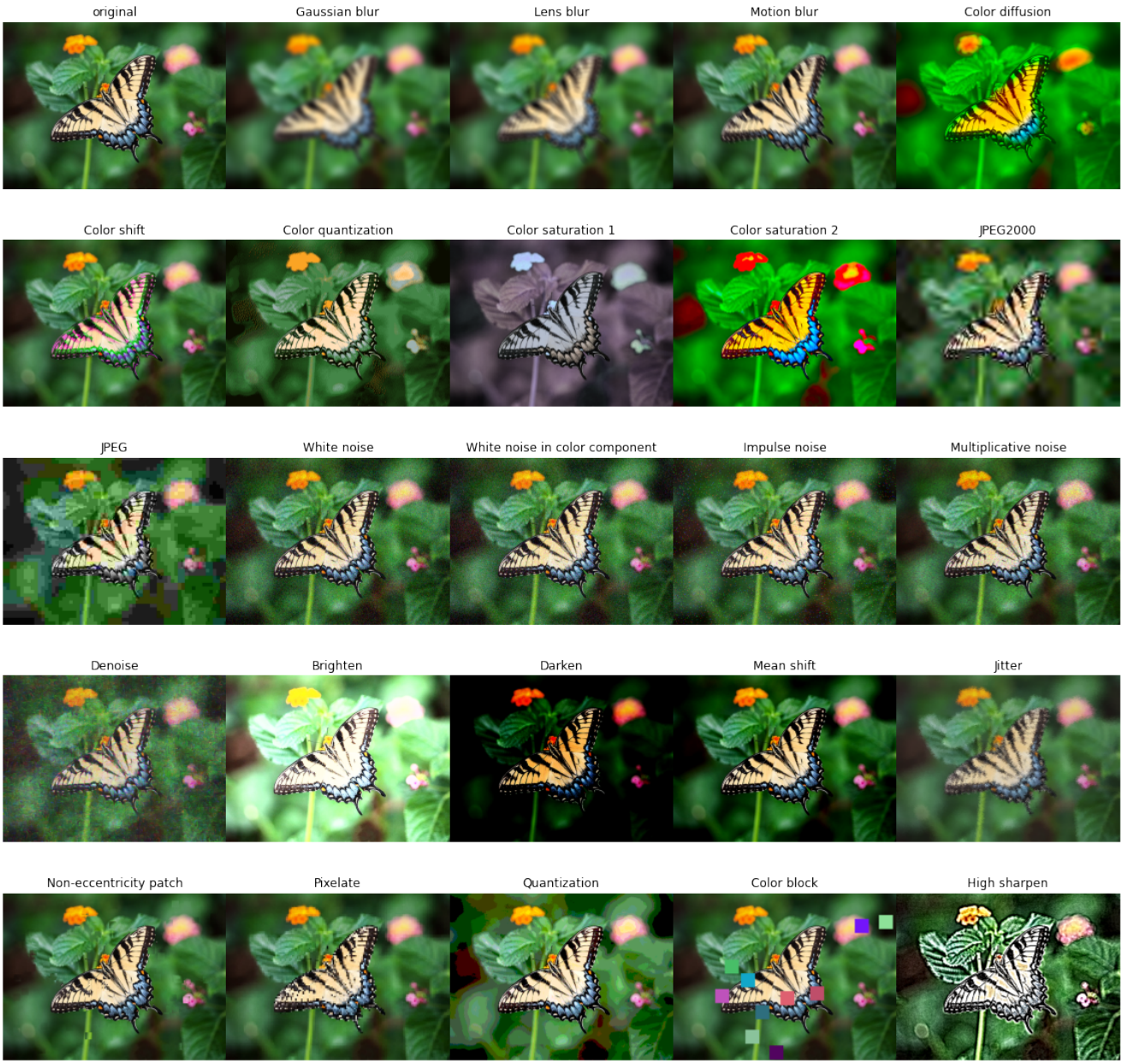


Figure 5: Illustrations of image distortions applied to a clear image. Each distortion operator can randomly transform the image from a range of severity levels. Original image sampled from KADIS dataset [23].

[11] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training, 2018.

[12] Zhou Wang, Alan C. Bovik, and Ligang Lu. Why is image quality assessment so difficult? In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–3313–IV–3316, 2002.

[13] Qiaohong Li, Weisi Lin, Jingtao Xu, and Yuming Fang. Blind image quality assessment using statistical structural and luminance features. *IEEE Transactions on Multimedia*, 18(12):2457–2469, 2016.

[14] Yutao Liu, Ke Gu, Shiqi Wang, Debin Zhao, and Wen Gao. Blind quality assessment of camera images based on low-level and high-level statistical features. *IEEE Transactions on Multimedia*, 21(1):135–146, 2019.

[15] Anush Krishna Moorthy and Alan Conrad Bovik. A two-step framework for constructing blind image qual-

ity indices. *IEEE Signal Processing Letters*, 17(5):513–516, 2010.

[16] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, 2011.

[17] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE Transactions on Image Processing*, 21(8):3339–3352, 2012.

[18] Deepti Ghadiyaram and Alan C. Bovik. Blind image quality assessment on real distorted images using deep belief nets. In *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 946–950, 2014.

[19] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. On the use of deep learning for blind image quality assessment. *CoRR*, abs/1602.05531, 2016.

[20] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014.

[21] Sebastian Bosse, Dominique Maniry, Thomas Wiegand, and Wojciech Samek. A deep neural network for image quality assessment. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3773–3777, 2016.

[22] Xialei Liu, Joost Van De Weijer, and Andrew D. Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1040–1049, 2017.

[23] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Deepfl-iqa: Weak supervision for deep iqa feature learning. *arXiv preprint arXiv:2001.08113*, 2020.

[24] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2018.

[25] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022.

[26] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *CoRR*, abs/1601.06759, 2016.

[27] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *CoRR*, abs/1803.07728, 2018.

[28] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. *CoRR*, abs/1505.05192, 2015.

[29] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. *CoRR*, abs/1603.09246, 2016.

[30] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.

[31] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[32] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *CoRR*, abs/1906.00910, 2019.

[33] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020.

[34] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2020.

[35] Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. imgaug. `https://github.com/aleju/imgaug`, 2020. Online; accessed 01-Feb-2020.

[36] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database.

In *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019.

[37] William Falcon and The PyTorch Lightning team. PyTorch Lightning, 3 2019.

[38] Yang You, Igor Gitman, and Boris Ginsburg. Scaling SGD batch size to 32k for imagenet training. *CoRR*, abs/1708.03888, 2017.

[39] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[40] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[41] David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, page 1098–1105, USA, 2012. IEEE Computer Society.

[42] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, 25(9):4444–4457, 2016.