Image Quality Assessment using Synthetic Images

Pavan C. Madhusudana
UT Austin
pavancm@utexas.edu

Neil Birkbeck, Yilin Wang, Balu Adsumilli
Google Inc.
{birkbeck,yilin,badsumilli}@google.com

Alan C. Bovik
UT Austin
bovik@ece.utexas.edu

Abstract

Training deep models using contrastive learning has achieved impressive performances on various computer vision tasks. Since training is done in a self-supervised manner on unlabeled data, contrastive learning is an attractive candidate for applications for which large labeled datasets are hard/expensive to obtain. In this work we investigate the outcomes of using contrastive learning on synthetically generated images for the Image Quality Assessment (IQA) problem. The training data consists of computer generated images corrupted with predetermined distortion types. Predicting distortion type and degree is used as an auxiliary task to learn image quality features. The learned representations are then used to predict quality in a No-Reference (NR) setting on real-world images. We show through extensive experiments that this model achieves comparable performance to state-of-the-art NR image quality models when evaluated on real images afflicted with synthetic distortions, even without using any real images during training. Our results indicate that training with synthetically generated images can also lead to effective, and perceptually relevant representations.

1. Introduction

Image Quality Assessment (IQA) refers to the problem of objectively quantifying and predicting perceptual judgments of image quality. In No-Reference (NR) or blind IQA, the task is to estimate quality with no additional information about the pristine reference image or the type of distortions present in the degraded image. IQA models are designed with the aim of obtaining accurate quality predictions that have high correlations with subjective judgements. The presence of multiple distortion types, along with the influence of content on perceived image quality, make blind IQA problem a challenging task. Social media applications, image and video-sharing sites such as Facebook, Instagram, YouTube etc. have millions of digital user-generated contents (UGC) uploaded to them everyday, and it is essential to objectively control and govern the quality, before performing additional operations such as compression.

NR-IQA has been extensively studied over the last decade, resulting in a variety of IQA datasets and models. Two types of IQA databases have been proposed in the literature : LIVE-IQA [39], CSIQ-IQA [24] etc. which contain synthetically degraded images, and CLIVE [13], KonIQ [18] etc. which contain images with realistic distortions. Distorted images with synthetic artifacts often contain a single distortion type such as blur, white noise etc. whereas in case of authentic artifacts, a combination of multiple distortion types are present. To objectively capture these artifacts, different approaches have been presented in the literature. Natural Scene Statistics (NSS) based models [33–35, 38] rely on statistical deviations arising due to distortions for obtaining quality-aware features. Recently, deep Convolutional Neural Network (CNN) based, data-driven IQA models [21, 41, 50, 54] have achieved remarkable accuracy in predicting image quality.

Deep CNNs contain millions of trainable parameters, thus require large labeled datasets to achieve better performance. However, currently there is lack of sufficiently large labeled IQA datasets, thus majority of the existing methods use transfer learning methods, where a pretrained model is fine-tuned using image quality labels. In CONTRIQUE [29], an unsupervised training scheme using unlabeled dataset was proposed as an alternative to transfer learning, and the model performance was observed to be comparable to that of current state-of-the-art (SOTA) IQA models. Employing synthetic data for training is another approach that has been explored in the literature for problems like stereo disparity [32], optical flow [6, 10]. For these applications, the models trained on synthetic data was shown to perform well even on real-world datasets [7, 42]. One key drawback of using synthetic data is the presence of domain gap between real and synthetic images, which can be a significant factor for certain applications [4].

Here, we investigate the performance of models trained on synthetic data for IQA problem. The goal is to understand the significance of semantic information in obtaining features which are representative of image quality. We follow the CONTRIQUE [29] framework in our experiments,

where real images are replaced with synthetically generated images, and the training is performed in a self-supervised manner using a contrastive objective. To the best of our knowledge, this is first such work employing synthetic images for the IQA problem. Our contributions in this work can be summarized as

1. We generated synthetic images using the dead leaves (DL) [25] model. DL model is based on statistical properties of natural images. It is a simplistic model for image generation and computationally inexpensive. During testing, model trained on DL data was evaluated on real images with no additional fine-tuning.

2. We introduced an extension to the DL model by including textures, and studied its influence on image quality prediction. We observed that addition of textures always improved model performance.

3. DL images lack semantic information prevalent in natural images. In order to better understand the impact of image semantics, a dataset containing anime images was used for training, and the model performance was observed to be better than that obtained using DL data.

4. Models trained with synthetic data achieved performance comparable to SOTA IQA models on datasets containing synthetic distortions.

2. Related Work

2.1. Blind Image Quality Assessment

The presence of diverse distortion types coupled with the effects of image content on different artifacts makes blind IQA a challenging task. NR models can be broadly categorized based on the design philosophy - traditional/hand-crafted models, and deep CNN based models. Traditional models generally contain a hand-crafted feature extraction framework to obtain quality aware features and a regressor is trained to map these features to quality scores. These include Natural Scene Statistics (NSS) based models such as DIIVINE [35], BLIINDS [38], BRISQUE [33] and NIQE [34], where deviations in the image statistics due to artifacts is used for predicting quality. CORNIA [48] and HOSA [47] employ a codebook based approach, where quality aware representations are obtained using a visual codebook constructed from local patches.

Inspired by the successes of deep learning on various computer vision tasks [16, 17, 42], many CNN-based based models have been applied for NR-IQA achieving impressive performances. The lack of large-scale datasets involving image quality has led to use of transfer learning techniques, whereby a pretrained model is fine-tuned using ground-truth quality scores. Pretrained CNNs extract reliable semantic features, and in [22] it was shown that these features are also excellent indicators of image quality. DB-CNN [54] used two separate CNNs to account for synthetic and realistic artifacts, respectively. In [50], a statistical distribution of subjective scores was used during training yielding superior quality estimates. PaQ-2-PiQ [49] model showed that fine-tuning with both image and patch quality scores can considerably improve model performance. In CONTRIQUE [29], a self-supervised training mechanism using unlabeled dataset was shown to yield robust and accurate image quality representations. All the above models employ real images during training and testing, while here we focus on employing synthetic images for training, and real images for testing.

2.2. Synthetic Data for Training

Certain computer vision problems such as disparity estimation, optical flow etc. have achieved remarkable successes in using synthetic datasets for model training, and fine-tuning on real world data. This is particularly beneficial for those applications for which obtaining large-scale labeled datasets is challenging and expensive. Synthetic datasets such as Sintel [6], Flyingchairs [10], and Scene flow [32] have significantly contributed towards improving stereo disparity [7] and optical flow [42] estimation. Recently Achddou *et al*. [1] employed images generated using dead leaves model for training a deep CNN, and obtained competitive performance on various image restoration tasks. Here we appraoch IQA problem using synthetic data, which has not been explored previously.

3. Method

Here, the goal is to learn effective representations using synthetically generated images which can be used to predict quality of real images. Since artificially created images are used for training, there is no ground-truth quality scores that can be used for training. Hence, a self-supervised training methodology based on CONTRIQUE [29] model is employed for feature learning. The whole training procedure is illustrated in Fig. 1. In the following sections each module present in the training pipeline is presented in detail.

3.1. Synthetic Image Generation

The first step is to create a database of undistorted synthetic images. In the literature, obtaining synthetic images using computer animation, and rendering using 3D graphics software such as Blender[1] has been extensively studied [6, 10, 32]. These images are created with the goal of lending sufficient diversity and realism as observed in natural images. However, this typically requires a careful design of contents in terms of background, color, objects present in the scene, degree of textures etc. Here, we experiment

[1] https://www.blender.org/

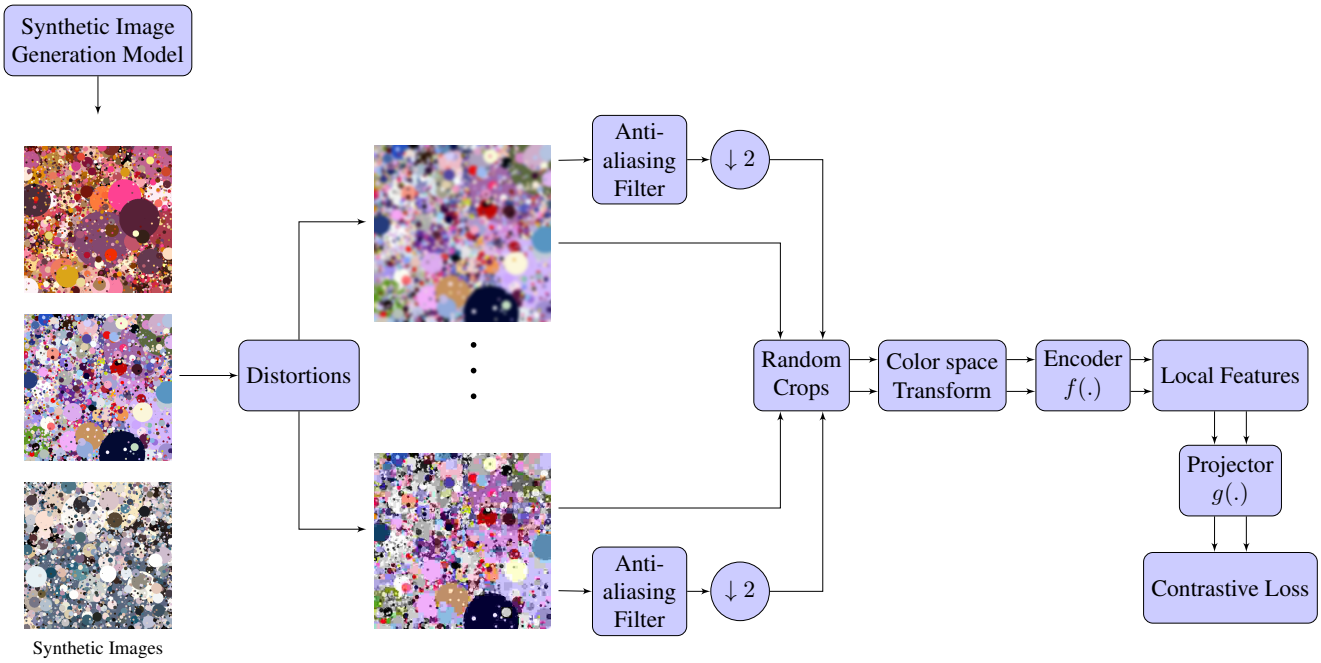


Figure 1. Illustration of training pipeline using synthetic images.

with Dead Leaves (DL) model, a simplistic image generation model based on natural image statistics.

The DL model was originally proposed by Matheron [31] to study morphological properties of materials, and was later observed to exhibit many statistical characteristics commonly seen in natural images [25]. In the DL model it is assumed that the image is formed by a set of template based objects whose locations are sampled from a Poisson process, and are arranged in a layered manner with partial occlusions. The marginal and joint derivative statistics of these DL images was observed to be similar to that of natural images [25]. Additionally, the power spectrum of DL images showed inverse square variation with frequency, commonly seen in natural scenes [12, 40].

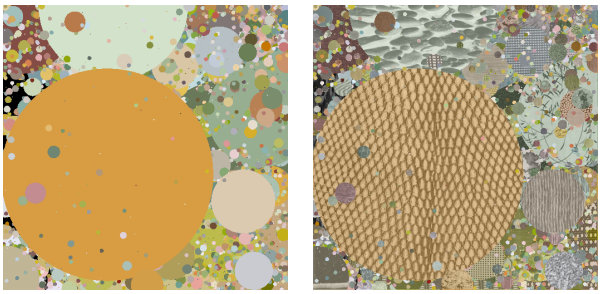
Here, for obtaining DL images we follow the procedure detailed in [1], using circular disks as template objects. The radius r of circular disks is sampled from $f(r) = Kr^{-3}$ distribution, where K is a normalizing constant. In [25] it was shown that this constraint was essential to have statistics similar to natural images. In order to have similar color histograms as natural images, each disk was assigned colors by randomly sampling from natural image color histogram. For this purpose, a natural image was fed as input during image generation (different scenes had different natural images).

3.1.1 Textured Dead Leaves

The expressive power of the original DL model can be enhanced by introducing textures to template objects present in the DL model. From Fig. 2a it can be observed that the images generated from the original DL model can contain significant smooth regions. Since smooth regions have small gradients, marginal gradient distributions are peakier when compared to that of natural images, as can be observed in Fig. 3. Including textures has several advantages: (i) It boosts gradient values, particularly in smooth regions. (ii) The statistics of resulting textured DL images are closer to that of natural images. This is illustrated in Fig. 3 where distributions are visually as well as objectively (using Kullback-Liebler divergence values) compared. Additionally, we also show in Sec. 4.2 that representations learned from textured DL images yield better quality estimates. The textures were applied to each circular disk separately using alpha blending with equal weights on texture and background color. The textures were randomly chosen from Brodatz texture database [5]. In Fig. 2, a sample DL image and corresponding textured version is shown.

3.2. Auxiliary Task

An auxiliary task is an alternate but closely related task for which ground-truth labels are either known or can easily be obtained. Following the CONTRIQUE framework, the auxiliary task is to obtain embeddings that can distinguish



(a) DL Image (b) Textured DL Image

Figure 2. Sample DL image and corresponding textured version.

images based on distortion type as well as the degree of degradations. This can be considered as a classification problem with images afflicted with same distortion type and degradation level categorized under the same class.

Let an undistorted synthetic image s be distorted by $d^i, i \in \{1, \ldots, D\}$ with degradation degree $l^{ij}, j \in \{1, \ldots, L^i\}$ resulting in a distorted image $\tilde{s}_i^j$. Here D and L^i denote number of distortion types and degradation degrees, respectively. Thus, this is a classification problem with $\sum_{i=1}^{D} L^i + 1$ classes (total number of degradation levels + one undistorted image). To extract features, the images are fed to a deep model consisting of two parts: an encoder and projector. An encoder can be any popular CNN architecture like Resnet [17] (with fully connected terminal layer removed) and the projector is a multi-layer perceptron (MLP) which reduces the dimensionality of the features produced by the encoder. For a given image $s \in \mathbb{R}^{3\times H\times W}$

$$k = f(s), \quad z = g(k) = g(f(s)) \quad k \in \mathbb{R}^B, z \in \mathbb{R}^K \quad (1)$$

where k is the B-dimensional output from the encoder. Similar to [8, 15], the intermediate features k are L_2 normalized before feeding as input to the projector. In the last step, a contrastive loss for image s_i is calculated as

$$\mathcal{L}_i = \frac{1}{|P(i)|} \sum_{j\in P(i)} -\log \frac{\exp(\phi(z_i, z_j)/\tau)}{\sum_{m=1}^{N} \mathbb{1}_{m\neq i} \exp(\phi(z_i, z_m)/\tau)}, \quad (2)$$

where N is the number of images present in the batch, $\mathbb{1}$ is the indicator function, τ is the temperature parameter, $P(i)$ is a set containing image indices belonging to the same class as s_i (but excluding the index i) and $|P(i)|$ is its cardinality. ϕ measures similarity between a pair of representations and is calculated as a dot product $\phi(a, b) = a^T b/||a||_2||b||_2$. The expression (2) is similar to the supervised contrastive loss [20] with class labels derived from prior knowledge about distortions.

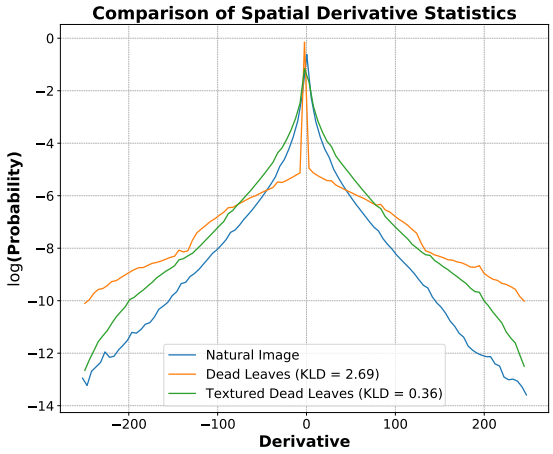


Figure 3. Comparison of distributions of spatial derivatives of natural and DL images. For plotting purposes 100 images from each type were employed. For DL and textured DL distributions, KL-divergence values with respect to natural images is also shown.

3.3. Multiscale Learning and Augmentations

Images, as well as artifacts present in them are multi-scale, and for obtaining better quality estimates it is essential to consider the effects of both local as well as global image characteristics. To obtain more accurate quality estimates, prior IQA methods [33–35, 46] have used feature extraction at multiple scales. In the training pipeline, we employ images at two scales: full resolution, and half-scale resolution obtained by downsampling by a factor of two. An anti-aliasing filter is used before downsampling to avoid aliasing artifacts as shown in Fig. 1.

The images are then randomly cropped to a fixed size $M \times M$. Although the cropped version can have different perceived quality as the original image, we assume that the distortion class remains the same. For each image two cropped versions are obtained, one each at full-scale and half-scale. The cropped versions are then subjected to two augmentations : horizontal flipping and color space conversion. Different color spaces are employed to extract complementary quality information present in them. Four color spaces RGB, LAB. HSV and grayscale along with a band-pass transform obtained using local Mean-Subtraction (MS) coefficients were randomly chosen for each crop of the input image. Prior NSS based models [3, 14, 30, 43] have also demonstrated the perceptual significance of using these transformation for quality prediction. Note that the above augmentations are quality preserving, whereby quality of the input image remains unchanged on application of these transforms.

The last step in the training pipeline involves partitioning

the transformed image into non-overlapping patches of size $P \times P$. This is done with the goal of capturing image quality attributes in a more granular manner. These patches were then fed as input to the encoder to obtain local representations and subsequently used in the loss function (2). Note that we assumed that patches will inherit the same distortion class labels as the original image, as was the case with cropping operation.

3.4. Evaluating Representations

Once the model training is complete the last step involves mapping learned embeddings to quality scores. The correlations of human judgments against predicted quality scores serve as a proxy for evaluating efficiency of the learned representations. During evaluation, the projector network $g(.)$ is discarded and output k from encoder is used as image features. An L_2 regularized linear regressor (ridge regression) is trained on top of the frozen encoder network using ground-truth quality scores from a suitable IQA database for predicting quality. The expression for ridge regression is given by

$$y = Wk, \quad W^* = \underset{W}{\operatorname{argmin}} \sum_{i=1}^{N} (GT_i - y_i)^2 + \lambda \sum_{j=1}^{M} W_j^2, \tag{3}$$

where y denotes predicted scores, GT ground-truth quality scores, λ is the regularization parameter, W is a trainable vector having same dimensions as h, M is number of dimensions of h, and N is the number of images present in the training set. During inference, all the features are computed at the native resolution of the input image, and no data augmentations are performed. Features are extracted at two scales : full-scale and half-scale, and a concatenated version of these two scales is used for regression. Note that no additional *fine-tuning* of encoder using ground-truth quality scores is performed as this can modify encoder weights, and will not be a true indicator of the effectiveness of self-supervised training process as well as the training data employed.

4. Experiments and Results

In this section we perform a series of experiments to investigate the effect of using synthetic data for training. First we will describe experimental settings, evaluation procedure and methods used for comparison. We then compare the performance of models trained using synthetic data against SOTA IQA models. Additionally, we also analyze the outcome of using anime images for training. Lastly, we extend the current model to a Full Reference (FR) setting where features from both reference and distorted images are used for predicting quality.

4.1. Experimental Details

Training Data

We generated 5000 DL images using the method detailed in Sec. 3.1. The generated images were then corrupted with $D = 25$ distortion types with each type having $L^i = 5$ degrees of degradation. The distortion types and degrees were same as those employed in KADIS [27] dataset. Interested readers can refer to [27] for more details about distortion types and degradation levels. Since there are 125 (25×5) possible distortions for each image, the generated dataset contains around $5000 \times 25 = 625,000$ images in total. This results in 126 distortion classes (125 distortion classes + 1 undistorted type) that are used in the contrastive objective (2).

Training Details

Resnet-50 [17] architecture (with fully connected layer removed) was used as the encoder network $g(.)$ and MLP with 2 hidden layers as projector network $g(.)$. Both the hidden layers of MLP contained 2048 neurons. The training was done with a batch size of $N = 512$ and the sampled images were randomly cropped to square blocks of size $M = 256$. For local feature extraction the image crops were further partitioned to patches of size $P = 64$, resulting in 4 patches from each image crop. The temperature parameter used in (2) was fixed at $\tau = 0.1$ and dimension of final feature z was chosen to be $K = 128$. Local representations were calculated using adaptive average pooling layer at the end of encoder module. The models were trained from scratch for 25 epochs using a stochastic gradient descent (SGD) optimizer having initial learning rate of 0.3 and cosine decay schedule without restarts [28] and a linear warmup for first two epochs. The implementations used in this work are available here. [2]

Compared Methods

We compared the performance of models trained on DL and textured DL data against nine SOTA IQA models. The compared methods include traditional models such as BRISQUE [33], NIQE [33], CORNIA [48] and HOSA [47]. The above models (except NIQE) use a support vector regressor (SVR) for predicting quality. The compared methods also contain deep learning based IQA models such as DB-CNN [54], PQR [50], BIECON [21], HyperIQA [41] and CONTRIQUE [29]. For numerical comparison of above IQA models, we copied the numbers provided by the respective authors or as available in the literature. Note that all the above models are trained on real images with no presence of synthetic images in their design framework.

[2] https://github.com/pavancm/CONTRIQUE_syn

Table 1. Performance comparison of NR models on IQA databases containing **synthetic** distortions. In each column, the first and second best models are boldfaced. Entries marked '-' denote that the results are not available.

Method	LIVE-IQA [39]		CSIQ-IQA [24]		TID2013 [36]		KADID [26]	
	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑
BRISQUE [33]	0.939	0.935	0.746	0.829	0.604	0.694	0.528	0.567
NIQE [34]	0.907	0.901	0.627	0.712	0.315	0.393	0.374	0.428
CORNIA [48]	0.947	0.950	0.678	0.776	0.678	0.768	0.516	0.558
HOSA [47]	0.946	0.950	0.741	0.823	0.735	0.815	0.618	0.653
DB-CNN [54]	**0.968**	**0.971**	**0.946**	**0.959**	0.816	**0.865**	0.851	**0.856**
PQR [50]	**0.965**	**0.971**	0.872	0.901	0.740	0.798	-	-
BIECON [21]	0.961	0.962	0.815	0.823	0.717	0.762	-	-
HyperIQA [41]	0.962	0.966	0.923	0.942	**0.840**	**0.858**	**0.852**	0.845
CONTRIQUE [29]	0.960	0.961	**0.942**	**0.955**	**0.843**	0.857	**0.934**	**0.937**
Dead Leaves	0.940	0.941	0.852	0.873	0.703	0.731	0.776	0.774
Textured Dead Leaves	0.950	0.951	0.920	0.930	0.751	0.776	0.820	0.820

Figure 4. Sample images from the Danbooru database

Table 2. SROCC performance comparison of model trained on the Danbooru dataset against models trained on dead leaves trained data. In each column, the top performing model is boldfaced.

Model	LIVE-IQA	CSIQ-IQA	TID	KADID
CONTRIQUE	**0.960**	**0.942**	**0.843**	**0.934**
Dead Leaves	0.940	0.852	0.703	0.776
Textured Dead Leaves	0.950	0.920	0.751	0.820
Danbooru	**0.960**	**0.942**	0.790	0.910

Evaluation Criteria

Spearman's rank order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC) were the evaluation metrics used for comparing IQA models. The predicted scores were fed to a four parameter logistic non-linearity [44] before calculating PLCC.

For evaluation we used four IQA databases containing synthetic distortions : LIVE-IQA [39], CSIQ-IQA [24], TID [36] and KADID [26]. These datasets contain images corrupted with synthetic distortions along with corresponding human opinion scores. For calculating weights of the linear regressor, each dataset is randomly divided into 70%,10% and 20% sets corresponding to training, validation and testing, based on reference image to avoid overlap of contents. The above procedure was repeated 10 times with different train-test combinations to avoid any bias on the choice of training contents, and the median performance is reported.

4.2. Correlation Against Human Judgments

In Table 1 we compare the performance of IQA models across four databases. Since models trained on synthetic data are based on CONTRIQUE framework, they are clustered together for ease of comparison. From the Table we can derive two important conclusions. (i) Using textured dead leaves data almost always improves performance. (ii) The performance difference between CONTRIQUE and models trained on synthetic images demonstrate the domain gap between real and dead leaves images. Notably, on LIVE-IQA and CSIQ-IQA datasets this gap relatively low, especially when trained with textured DL data. Also from Table 1 it can be seen that models trained on DL data outperform most traditional IQA models, suggesting that the learned representations from synthetic data more accurately represent perceptual quality than handcrafted features.

4.3. Training with the Danbooru dataset

We analyzed the significance of using synthetic images generated from dead leaves model in Table 1. In this experiment we investigate the effect of using other type of synthetic images obtained by a different generative model. In particular we use Danbooru [2] dataset, which is a large-scale collection of anime images containing in excess of 4 million images. Sample images present in this dataset are shown in Fig. 4. The motivation behind using this dataset is to study the importance of image semantics on IQA as anime images contain more semantic information than dead leaves images. The images present in this database are of high quality with 512×512 resolution and contains images of popular anime characters. We randomly sample 5000 images from this dataset, and artificially degrade them following the same procedure employed for DL images, obtaining a total

Table 3. Full Reference performance comparison across 4 IQA databases. In each column, the first and second best models are boldfaced. Entries marked '-' denote that the results are not available.

Method	LIVE [39]		CSIQ-IQA [24]		TID2013 [36]		KADID [26]	
	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑
PSNR	0.881	0.868	0.820	0.824	0.643	0.675	0.677	0.680
SSIM [45]	0.921	0.911	0.854	0.835	0.642	0.698	0.641	0.633
FSIM [52]	0.964	0.954	0.934	0.919	0.852	0.875	0.854	0.850
VSI [51]	0.951	0.940	0.944	0.929	0.902	0.903	**0.880**	**0.878**
PieAPP [37]	0.915	0.905	0.900	0.881	0.877	0.850	0.869	0.869
LPIPS [53]	0.932	0.936	0.884	0.906	0.673	0.756	0.721	0.713
DISTS [9]	0.953	0.954	0.942	0.942	0.853	0.873	-	-
DRF-IQA [23]	**0.983**	**0.983**	**0.964**	**0.960**	**0.944**	**0.942**	-	-
CONTRIQUE-FR	**0.966**	**0.966**	**0.956**	**0.964**	**0.909**	**0.915**	**0.946**	**0.947**
Dead Leaves	0.948	0.948	0.931	0.932	0.791	0.816	0.869	0.868
Textured Dead Leaves	0.950	0.950	0.945	0.950	0.842	0.853	0.898	0.900
Danbooru	0.962	0.962	0.950	0.957	0.880	0.891	0.935	0.937

of 625,000 distorted images. In Table 2, the performance of the Danbooru trained model is compared against other models, and it can be seen that it outperforms DL trained models. These results indicate that training data containing better semantic information can be beneficial in obtaining more accurate image quality representations.

4.4. Full-Reference IQA

Similar to CONTRIQUE, the learned representations can be employed in an FR setting with no additional training of the encoder module. This is accomplished by incorporating reference features in the regressor as

$$y = W|k_{ref} - k_{dist}|,$$
$$W^* = \underset{W}{\operatorname{argmin}} \sum_{i=1}^{N} (GT_i - y_i)^2 + \lambda \sum_{j=1}^{M} W_j^2, \quad (4)$$

where absolute difference between the features of reference and distorted images is used for predicting quality. The performance of FR-IQA models is compared in Table 3. Similar evaluation protocol of dividing datasets into 70%,10%,20% as training/validation/testing sets, respectively based on content, was followed. The train-test division was repeated 10 times and median correlation values are reported. Nine SOTA FR-IQA models were included for performance comparison : PSNR, SSIM [45], FSIM [52], VSI [51], PieAPP [37], LPIPS [53], DISTS [9], DRF-IQA [23] and CONTRIQUE-FR [29]. From the Table 3 we can make similar observations as seen in the No-Reference case, where training with Danbooru data performed better than that using DL data, emphasizing the importance of semantic information for training.

5. Drawbacks of Using Synthetic Images

In the previous sections we analyzed the model performances on synthetically distorted IQA datasets. Every distorted image in these databases was corrupted by a 'single' distortion type. However, if we consider images with realistic distortions such as User Generated Content (UGC) images, a combination of multiple distortions is involved. Models which are trained on synthetic data often under-perform when evaluated on UGC datasets as can be seen in Table 4, where performances across 4 IQA datasets KonIQ [18], CLIVE [13], FLIVE [49] and SPAQ [11] containing authentic distortions are compared. We hypothesize that two factors might have contributed to this performance gap. Firstly, the training data contains distorted images which only have 'single' types of corruption. Thus there exists a significant domain gap in terms of synthetic and authentic distortions resulting in lower performance. Artificially replicating authentic distortions is hard, as they often contain diverse mixtures of unknown distortions. Secondly, image semantics play a major role in quantifying realistic distortions [22, 41]. Thus, the lack of sufficient semantic information might also be a contributing factor of the performance degradation.

6. Conclusion and Future Work

In this work we investigated the effect of using synthetic data in an unsupervised training framework for learning effective image quality representations. A synthetic image dataset from dead leaves model was generated, and discriminating distortion type and degree was used as an auxiliary task to train a deep CNN model. We conducted holistic evaluation on multiple IQA databases and analyzed the significance of texture and semantic information in predicting image quality. We also highlighted the drawbacks of employing models trained with synthetic data on images corrupted with realistic distortions. As part of future work we plan to explore adding multiple distortions to the training data similar to LIVE multiply distorted dataset [19], and analyze its effect on quantifying realistic distortions.

Table 4. Performance comparison of NR models on IQA databases containing **authentic** distortions. In each column, the first and second best models are boldfaced. Entries marked '-' denote that the results are not available.

Method	KonIQ [18]		CLIVE [13]		FLIVE [49]		SPAQ [11]	
	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑
BRISQUE [33]	0.665	0.681	0.608	0.629	0.288	0.373	0.809	0.817
NIQE [33]	0.531	0.538	0.455	0.483	0.211	0.288	0.700	0.709
CORNIA [48]	0.780	0.795	0.629	0.671	-	-	0.709	0.725
HOSA [47]	0.805	0.813	0.640	0.678	-	-	0.846	0.852
DB-CNN [54]	0.875	0.884	0.851	0.869	0.554	**0.652**	0.911	0.915
PQR [50]	0.880	0.884	**0.857**	**0.882**	-	-	-	-
HyperIQA [41]	**0.906**	**0.917**	**0.859**	**0.882**	0.535	0.623	**0.916**	**0.919**
CONTRIQUE	**0.894**	**0.906**	0.845	0.857	**0.580**	0.641	**0.914**	**0.919**
Dead Leaves	0.812	0.826	0.671	0.700	0.460	0.500	0.870	0.877
Textured Dead Leaves	0.820	0.835	0.677	0.700	0.485	0.528	0.872	0.879
Danbooru	0.841	0.851	0.715	0.717	0.520	0.540	0.886	0.893

7. Acknowledgment

This research was sponsored by a grant from YouTube and by grant number 2019844 for the National Science Foundation AI Institute for Foundations of Machine Learning (IFML). The authors would also like to thank the Texas Advanced Computing Center (TACC) for providing computational resources that contributed to this work.

References

[1] Raphaël Achddou, Yann Gousseau, and Saïd Ladjal. Synthetic images as a regularity prior for image restoration neural networks. In *Eighth International Conference on Scale Space and Variational Methods in Computer Vision (SSVM)*., 2021.

[2] Anonymous, Danbooru community, and Gwern Branwen. Danbooru2020: A large-scale crowdsourced and tagged anime illustration dataset. `https://www.gwern.net/Danbooru2020`, January 2021. Accessed: September 2021.

[3] Christos G Bampis, Praful Gupta, Rajiv Soundararajan, and Alan C Bovik. SpEED-QA: Spatial efficient entropic differencing for image and video quality. *IEEE Signal Process. Lett.*, 24(9):1333–1337, Sep. 2017.

[4] Manel Baradad, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. *arXiv preprint arXiv:2106.05963*, 2021.

[5] Phil Brodatz. *Textures: a photographic album for artists and designers, by Phil Brodatz*. Dover publications, 1966.

[6] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer, 2012.

[7] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *Int'l Conf. Machine Learning*, 2020.

[9] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–1, 2020.

[10] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[11] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pages 3677–3686, 2020.

[12] David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Josa a*, 4(12):2379–2394, 1987.

[13] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Trans. Image Process.*, 25(1):372–387, 2015.

[14] Deepti Ghadiyaram and Alan C Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of vision*, 17(1):32–32, 2017.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pages 9729–9738, 2020.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pages 2961–2969, 2017.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pages 770–778, 2016.

[18] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Trans. Image Process.*, 29:4041–4056, 2020.

[19] Dinesh Jayaraman, Anish Mittal, Anush K Moorthy, and Alan C Bovik. Objective quality assessment of multiply

distorted images. In *Asilomar Conf. Signals Syst. Comput.*, pages 1693–1697. IEEE, 2012.

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 33, pages 18661–18673, 2020.

[21] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE J. Sel. Topics Signal Process.*, 11(1):206–220, 2016.

[22] Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram, Sanghoon Lee, Lei Zhang, and Alan C Bovik. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Sig. Proc. Magazine*, 34(6):130–141, 2017.

[23] Woojae Kim, Anh-Duc Nguyen, Sanghoon Lee, and Alan Conrad Bovik. Dynamic receptive field generation for full-reference image quality assessment. *IEEE Trans. Image Process.*, 29:4219–4231, 2020.

[24] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *J. Electron. Imag*, 19(1):011006, 2010.

[25] Ann B Lee, David Mumford, and Jinggang Huang. Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. *International Journal of Computer Vision*, 41(1):35–59, 2001.

[26] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10k: A large-scale artificially distorted IQA database. In *IEEE Int'l Conf. on Quality of Multimedia Experience*, pages 1–3, 2019.

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

[28] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *Int'l Conf. Learning Representations*, 2017.

[29] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Image quality assessment using contrastive learning. *arXiv preprint arXiv:2110.13266*, 2021.

[30] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. ST-GREED: Space-time generalized entropic differences for frame rate dependent video quality prediction. *IEEE Trans. Image Process.*, 30:7446–7457, 2021.

[31] G. Matheron. Mod‘ele s´equentiel de partition al´eatoire. *Centre de Morphologie Mathématique*, 1968.

[32] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.

[33] Anish Mittal, Anush K. Moorthy, and Alan C Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, Dec. 2012.

[34] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Sig. Process. Lett.*, 20(3):209–212, Mar. 2013.

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

[36] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database TID2013: Peculiarities, results and perspectives. *Signal Process.: Image Commun.*, 30:57–77, 2015.

[37] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pages 1808–1817, 2018.

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

[39] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.*, 15(11):3440–3451, Nov 2006.

[40] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

[41] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pages 3667–3676, 2020.

[42] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pages 8934–8943, 2018.

[43] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. RAPIQUE: Rapid and accurate video quality prediction of user generated content. *IEEE Open J. Signal Process.*, 2:425–440, 2021.

[44] VQEG. Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment, 2000.

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, April 2004.

[46] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conf. Signals Syst. Comput.*, volume 2, pages 1398–1402 Vol.2, Nov 2003.

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

[48] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference

image quality assessment. In *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pages 1098–1105. IEEE, 2012.

[49] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pages 3575–3585, 2020.

[50] Hui Zeng, Lei Zhang, and Alan C Bovik. A probabilistic quality representation approach to deep blind image quality prediction. *arXiv preprint arXiv:1708.08190*, 2017.

[51] Lin Zhang, Ying Shen, and Hongyu Li. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Trans. Image Process.*, 23(10):4270–4281, 2014.

[52] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.*, 20(8):2378–2386, 2011.

[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pages 586–595, 2018.

[54] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Trans. Circuits Syst. Video Technol.*, 30(1):36–47, 2018.