Blind Tone-mapped Image Quality Assessment and Enhancement via Disentangled Representation Learning

Lei Wang, Qingbo Wu* , King Ngi Ngan, Hongliang Li, Fanman Meng, Linfeng Xu
University of Electronic Science and Technology of China, Chengdu, China
E-mail: lwang@std.uestc.edu.cn;{qbwu;knngan;hlli;fmmeng;lfxu} @uestc.edu.cn

***Abstract*—For compatibility with existing low dynamic range (LDR) display devices, the tone mapping operator (TMO) is widely applied to the high dynamic range (HDR) image, which inevitably leads to visual quality degradation. Various blind image quality assessment models have been developed to quantify the distortion degrees across different HDR images. However, these models only extract the quality-aware features and serve as a selector for different TMOs, which excludes visual content information and fails to conduct an end-to-end image enhancement towards a desired quality score. In this paper, we propose to jointly conduct blind tone-mapped image quality assessment and enhancement via disentangled representation learning. An encoder is firstly used to map the input image into the general feature space. Then, two branches are separately developed to extract the quality-aware and content-aware latent representations from the general feature, which are supervised with the quality score and image reconstruction constraints, respectively. Meanwhile, these two branches are also coupled with the adaptive instance normalization, which enables our model to flexibly modify the image towards any desired quality score. Extensive experiments confirm the effectiveness of the proposed method.**

I. Introduction

Recent years have witnessed the rapid growth of high dynamic range (HDR) imaging applications, which enjoy great popularity due to their better visual expression in terms of luminance, contrast and color variations. But the unfriendly price of special display devices also limits the popularization of HDR images in our daily life. Several tone-mapping operators (TMOs) and multi-exposure fusion methods are born to convert HDR images to low dynamic range (LDR) images [1], which are compatible with existing display devices. However, the perceptual quality degradation is inevitable regarding the change of dynamic range during the tone mapping process. Given different HDR images, the degrees of quality degradation produced by each TMO are different as well. To select the optimal TMO, a lot of efforts are devoted to the blind tone-mapped image quality assessment (BTMIQA) to quantify the quality of LDR image without accessing to its original HDR version.

Classical BTMIQA models develop various handcrafted features to capture quality-aware information, such as image sharpness, colorfulness, naturalness and so on. Then, these features are mapped to the quality scores via different regression models. Recently, some deep learning based methods also attepmt to conduct the BTMIQA task by training the feature extractor and quality regressor simultaneously. Although different descriptors or network archtectures are explored in these methods, they all focus on extracting the quality-aware global representation, which is irreversible to the content information in different spatial locations. Therefore, existing BTMIQA models could only passively select the best tone-mapped image from several candidate TMOs instead of directly enhancing the LDR image in an end-to-end manner.

In this paper, we propose a more flexible and interpretable network architecture for joint image quality assessment and enhancement, which is achieved via the disentangled representation learning. More specifically, an encoder is firstly used to map the input image into the general feature space. Then, two branches are separately developed to extract the quality-aware and content-aware latent representations from the general feature, which are supervised with the quality score and image reconstruction constraints, respectively. Meanwhile, these two branches are also coupled with the adaptive instance normalization, which enables our model to flexibly modify the image towards any desired quality score. Extensive experiments on the ESPL-LIVE HDR database verify the effectiveness of the proposed method in both the image quality assessment and enhancement tasks.

II. Related Work

A. Representative BTMIQA Methods

Like natural image quality assessment, most existing BTMIQA methods still assumes that the natural scene statistics (NSS) [2] play a critical role in distinguishing the tone-mapped images with different qualities. Kundu et al. [3] explore the performance of many representative NSS features in the BTMIQA task, such as the statistics of log-derivative, mean subtraction and divisive normalization operators, and gradient information. In [4], Gu et al. extract the TMO specific features by measuring the information volumes of different illuminance ranges, the means and standard deviations of local pathces, and the mean responses of the sobel filter. In [5], Yue et al. further enrich the NSS features with the normalized colorfulness and contrast information. Recently, some deep learning based methods are also explored for BTMIQA task. Kumar et al.

* Corresponding author.

[6] employ the AlexNet to extract the deep features, and then reduce their dimension via principal component analysis. He et al. [7] develop a multiscale deep representation based on the ResNet-50.

It is noted that all of aforementioned methods focus on extracting the quality-aware global representation, which excludes the content information and fails to directly conduct the image enhancement.

B. Disentangled Representation Learning

Disentangled representation learning aims to obtain interpretable and attribute-specific latent codes, which is widely discussed in many conditional image generation task. In [8], Yan et al. achieve this target by optimizing the neural network parameters to maximize the conditional log-likehood, whose variables include the input image, attribute, and the latent code. In [9], Chen et al. introduce the attribute related latent code into the generative adversarial network, which rewards the mutual information between the latent code and the generator distribution. In [10], Lin et al. further increase the distinction of disentangled representations via a contrastive regularizer. In [11], [12], the disentangled representation learning is also explored in different domain or style transfer tasks.

These works inspire us to decompose the image feature into the quality-aware and content-aware latent codes which could serve for image quality assessment and enhancement tasks respectively.

III. Method

In contrast to conventional BTMIQA, the joint image quality assessment and enhancement is more challenging. We need to ensure that the extracted features are correlated with different target attributes. In this section, we first describe the motivation of the proposed model. Then, we introduce the detailed architecture design and loss function. At last, we describe the method of transferring the desired quality to the input tone-mapped image.

A. Model

Let x be the input tone-mapped image. We want to extract a latent z which correlates to image quality. Existing BTMIQA models typically optimize a regressor $f(.)$ to minimize the difference between $f(z)$ and the mean opinion score (MOS) y, while imposing no restrictions on z. As a result, the latent representation z may be used in a highly entangled way, which brings quality irrelevant attributes into z. In the test phase, the distribution of such latent representation z may abruptly change across different image content, thus tampering the generalization capability of the model and limiting its quality assessment. In this context, we want to decompose the image feature into two distinct latent representations, which correspond to the image quality and content respectively. More specifically, quality-aware and content-aware representations could be learned with the supervision of MOS and image reconstruction based self-supervision. By integrating the quality-aware and content-aware representations together, we can achieve a quality-controlled image enhancement.

B. Architecture Design

Fig. 1 shows the detailed network architecture of our joint image quality assessment and enhancement model. We first map the input image x into a general representation t with the encoder $\mathbf{E}$, a popular network architecture, ResNet50 [13]. Inspired by recent work [12], the representation n including little information about quality which we could treat as style of image, we use Instance-Normalization for content feature extraction. Naturally, the discarded mean and variance of the feature layers are treated as quality representations. specifically, the content-aware representation $n = \frac{t-\mu(t)}{\sigma(t)}$ and quality-aware representation $z = \mu(t) \bigoplus \sigma(t)$ where $\bigoplus$ is feature concatenation. Here $\mu(x)$ and $\sigma(x)$ are computed across spatial dimensions independently. Given z, we use a Multi-Layer Perceptron(MLP) which contain 4 fully connected layer to predict the score. Given such content representation n, we require that n guarantees the preservation of content information to reconstruct image. Here we use U-net [14] architecture connect decoder our $\mathbf{D}$, as its skip-connection help to propagate the multilayer spatial information. To cooperate with quality representation and speed up the translation, image are reconstructed by the decoder which is equipped with an Adaptive Instance Normalization (AdaIN) [12]. In order to guarantee skip-connection, decoder $\mathbf{D}$ have same scale architecture as $\mathbf{E}$, which contain 5 maxpool layer and several convolution layers.

C. Loss Function

Inspired by the deep variational information bottleneck [15], We regard the quality representation as a stochastic encoding z of the input source x, defined by a parametric encoder $p(z|x)$. We expect the quality representation to be aligned with the Gaussian distribution while predicting the quality score. This has two advantages, one is to reduce relevance to the content and the other is to generate quality representations from samples taken from gaussian distributions. Loss related to quality representation is established as:

$$\mathbb{E}_{p(z|x)}\left[\log q(y|z)\right] - \beta D_{KL}\left(p(z|x)\|m(z)\right) \tag{1}$$

where $q(y|z)$ is a variable approximation of the decoder to conditional posterior $p(y|z)$, $m(z)$ is a variable approximation to marginal posterior $p(z)$, and $p(\cdot|x)$ is conditional prior of the encoder. Given n we want to be able to reconstruct x. Loss related to content representation can be write as :

$$\mathbb{E}_{p(z,n|x)}\left[\log q(x|z,n)\right] \tag{2}$$

We now propose a method for extract a quality representation z and retain appropriate content information meanwhile. We provide the decoder network with both the quality representation z and the content latent code n. Ideally, the images generated through the generator $q(x|z,n)$ are based on two factors n and z, But in practice, it is easy to converge to a solution satisfying $q(x|z,n)$=$q(x|n)$ that latent code z has little effect on the output and the change of image is determined by x totally. To cope with the problem of trivial codes, we added a little

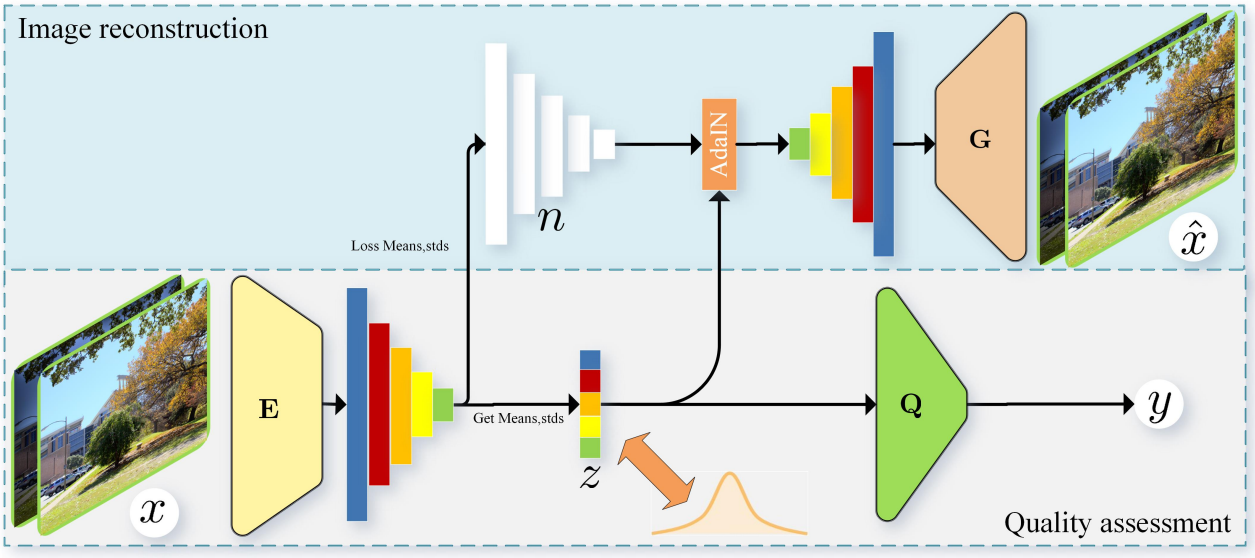

Fig. 1. The architecture of the proposed QD-net. **E**:encoding module, **Q**:a score prediction module, **G**:a decoding module. The bottom pipeline is similar to general image quality method. The top pipeline is the added reconstruction constraint part

trick to the training process. When we optimized reconstruct term, We added a pair term to increase the influence of z:

$$\mathbb{E}_{p(z,n|x)}\left[\log q(x|z,n) + \log q(\tilde{x}|\tilde{z},n)\right] \tag{3}$$

Where $\tilde{z}$ and $\tilde{x}$ are corresponding representation and image with the same content but different quality. Formulation, Replace function $\tilde{x} = F(x)$, $\tilde{z} = E_z(F(x))$.The technique of replacing features in this way to enhance the effect of features is seen in domain translation papers, such as [11].Let us establish a network G which estimates the parameters of the distribution $q(x|z,n)$. We assume further, as it is common practice [16], that the distribution $q(x|z,n)$ has constant standard deviation and the function $G(z,n)$ is a deterministic function in z,n. As a consequence, the network $G(z,n)$ can be considered as an image generator network and can replace Eq.(1) with the reconstruction loss $\mathbb{E}_{p(z,n|x)}\left[\|x - G(z,n)\|_1\right]$. The total loss can be written as :

$$\begin{aligned}\mathcal{L} = &\mathbb{E}_{p(z|x)}\left[\log q(y|z)\right] - \beta D_{KL}\left(p(z|x)\|m(z)\right) \\ &+ \alpha\mathbb{E}_{p(z,n|x)}\left[\|x - G(z,n)\|_1\right] \\ &+ \gamma\mathbb{E}_{p(z,n|x)}\left[\|F(x) - G(E_z(F(x)),n)\|_1\right]\end{aligned} \tag{4}$$

D. From Quality Assessment to Enhancement

Image quality is a continuous variable, so image enhancement should be modeled by a continuous process instead of a discrete domain transfer. Given an image, we want to get a quality score from the IQA model and then form scores to get quality features to transfer an image to another one. In order to achieve the fine-granularity control of the enhancement process, we propose to formulate a IQA guided enhancement equation by:

$$\hat{x} = G(E_c(x), \hat{z}) \tag{5}$$

TABLE I
THE EVALUATION RESULTS OF ALL BIQA MODELS

MODEL	QD-net	He et al.	HIGRADE-2	BIBQA	DESIQUE	GM-LOG
SROCC	**0.841**	0.823	0.730	0.702	0.570	0.556
PLCC	**0.837**	0.827	0.728	0.692	0.568	0.557

where $\hat{z} = Q^{-1}(\hat{y})$. We want to get a quality representation $\hat{z}$ from objective quality $\hat{y}$ through Q^{-1} which is the inverse function of Q. After training, generator G can generate image from $\hat{z}$, which should be sampled from Gaussian distribution. It's easy to training a predicting network Q for mapping $\hat{z}$ to $\hat{y}$, while it is hard to get a higher-dimensional representation $\hat{z}$ from a scalar $\hat{y}$. Ideally, given a image x, the quality representation z is sampled from the Gaussian distribution $p(z|x)$ which parameters are estimated by the encoder network E_z. Aim to get a suitable $\hat{z}$, we use Monte Carlo method [17] repeated sample from $\mathcal{N}(z \mid 0, I)$ to select a quality representation $\bar{z}$ satisfying $Q(\bar{z}) \geq \hat{y}$. Because of the continuity of the neural network, we can search a suitable $\hat{z}$ between $\bar{z}$ and z which quality score is between $\bar{y}$ and y. The objective of searching quality representation can be formulated as:

$$\min_{\epsilon} \|Q(\epsilon\bar{z} + (1-\epsilon)z) - \hat{y}\| \tag{6}$$

where ϵ is the parameter to balance the fusion of $\bar{z}$ and z. Since $\bar{z}$ was already obtained in the previous Monte Carlo sample, We use the SGD [18] optimizer to optimize (6) to get a suitable ϵ. The architecture are show in figure 2

IV. EXPERIMENT

A. Implementation

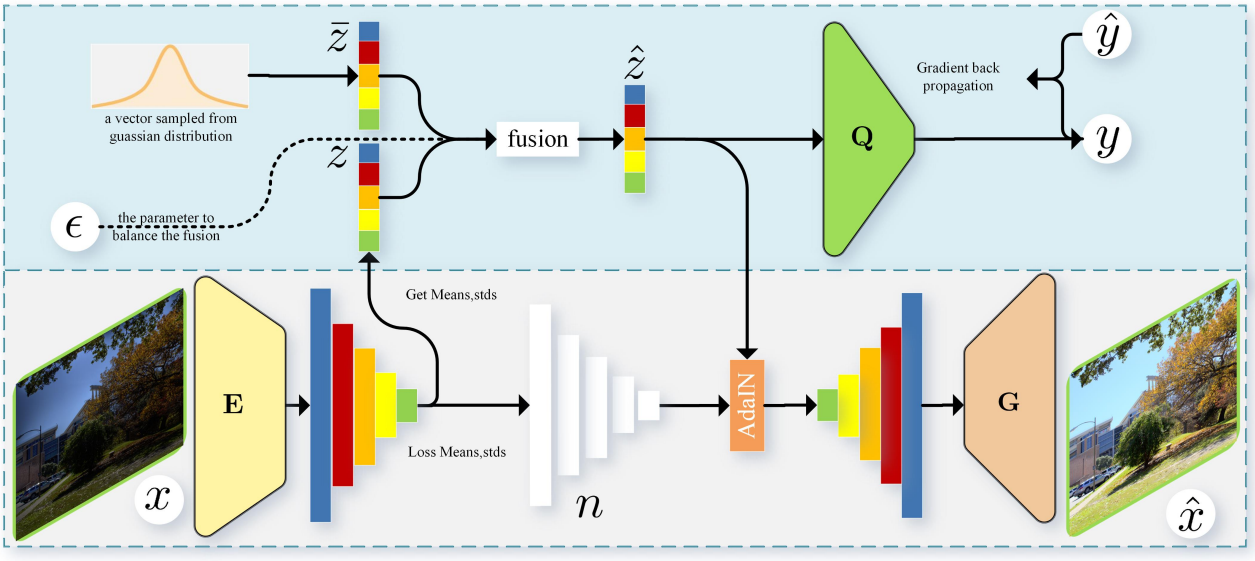

Fig. 2. The architecture of transfer quality to image. A suitable quality representation $\bar{z}$ is generated from a gaussian distribution and it fuses with z of input to get $\hat{z}$ corresponding target score $\hat{y}$. Then $\hat{z}$ is used to generate target images.

1) Datasets: In order to test the performance of our method, the proposed QD-net is trained on ESPL-LIVE HDR image database, which is the largest common tone-mapping image database in TMIQA. It covers a variety of methods for obtaining tone-mapped images: the images in the database were obtained from the HDR illumination map and the SDR image stack through 11 different methods.Therefore, this method is suitable for the experiment of predicting the quality of tone mapping HDR images under the influence of complex factors. There are a total of 1811 tone-mapped HDR images, which are divided into three types according to their acquisition methods: 747 images are generated by four different tone mapping operators; Through five multi-exposure fusion methods, 710 images were directly created from the SDR image stack. Through two different post-processing settings, 354 images were obtained by Photomatix. We randomly split the data into disjoint training and testing sets at a 4:1 ratio and the splits were randomized over 100 trials. Care was taken to ensure that the same source scene did not appear during both training and testing to prevent artificial inflation of the results, following the work in [19].

2) Preprocessing: Images in the training set are randomly cropped to 302*302, because the random crop can augment our dataset amount. Images in the test set are fed to our model without crop. All of this operation is considering that cropped images have the same score as the original one.

3) Detailed Implementation: The proposed QD-net model consists of a encoding module **E**, a score map module **Q**, and a decoding module **G**. We use SGD with learning rate 0.01 for optimization, and learning rate decay 0.8 per 7 epochs, and dropout rate 0.5 for regularization. During training phase, each batch contains 16 pairs of images and can form totally 16 pairs for optimized loss. During testing phase, each batch contains 1 pair of images because images in testing set are fed to our model without crop. During enhancement phase, we input a low-quality image to obtain its quality representation and predicted quality score, and at the same time, we sample to generate another high-quality image, increase the value of the original quality score, and get pictures with different quality scores through (6) We implement the proposed QD-net with the PyTorch library, and perform the experiments in a workstation with Intel Xeon E5-2660 CPU and NVIDIA TI-TAN X GPU.

B. Comparisons

We conduct experiments on ESPL-LIVE HDR Database to compare the performance of our proposed method with some existing TMIQA methods and some state-of-the-art NR-IQA methods which are aimed at general SDR images. Because most of the existing TMIQA methods are full-referenced and the original HDR images are not available in the ESPL-LIVE HDR Database, we compared two NR-TMIQA methods among them. He et al. [7], HIGRADE [3], BIBQA [20], DESIQUE [21], GMLOG [22] are included for comparison.

C. Results

1) TMIQA Result: We evaluate the performance of each BIQA model via two widely used indices: Spearman's rank-order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC). These criteria are calculated for every run, and the medians of the three criteria in 100 runs are reported separately. We repeat the random split experiment 100 times and report the median results in the following. Table I shows the evaluation results of all quality assessment

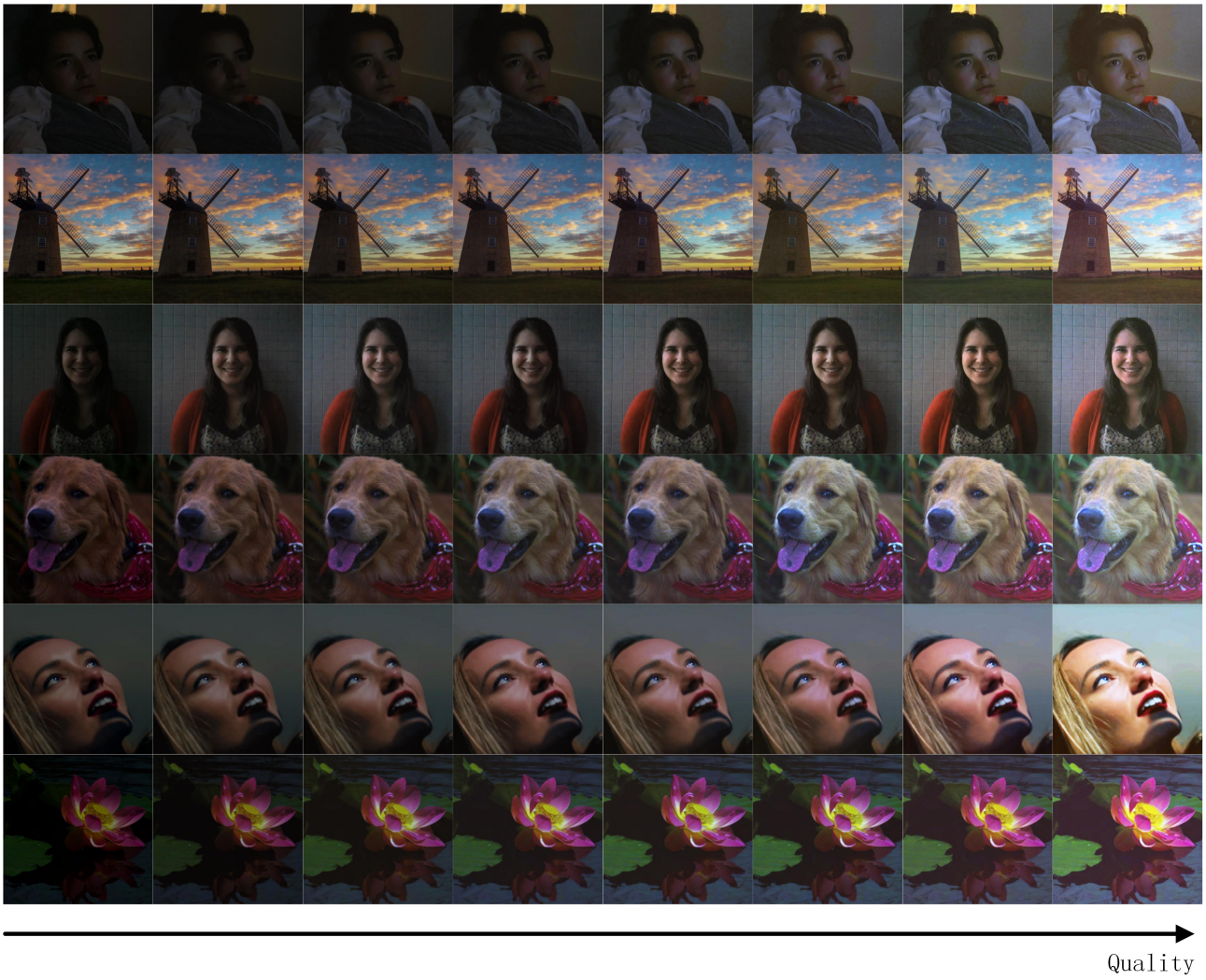

Fig. 3. Quality degree adjustment by controlled quality representation (the leftmost is the original input, from left to right: light to heavy enhancement).

models, where the best results are highlighted by boldface in each column. The results of HIGRADE, DESIQUE, GM-LOG, BRISQUE on the same database are quoted from [3], the results of BIBQA method are quoted from [20], the results of He et al. are quoted from [7].

2) Enhancement Result: To prove the performance of enhancement, we conduct experiments on the real photo from the internet. In Fig. 3, the images in the left column are the inputs, and we gradually enhance their qualities in ascending order from the left to the right columns. Similarly, in Fig. 4, all inputs are listed in the leftmost column, and we gradually degrade them with the proposed method in descending order from the left to the right side. It is seen that the proposed method could flexibly and effectively change the image towards a desired quality. In addition, past work [16] has shown that the encoder network is able to learn to cluster high-dimensional data, so we conjecture that posterior z outputted from the encoder network should cluster the representation into meaningful groups. Figure 5 visualizes the posterior z in the test dataset in 2D space using t-SNE [23]. It is seen that the learned latent space is highly correlated with the quality score, which confirms our assumption.

V. Conclusion

In this paper, we propose a QD-net to jointly conduct the blind tone-mapped image quality assessment and enhancement. Instead of serving as TMO selector like most existing BTMIQA models, the proposed method could learn quality-aware and content-aware latent codes with the disentangled representation learning simultaneously, which could support a quality-controlled image translation. Extensive experiments verify the effectiveness of the proposed method in both the quality assessment and enhancement tasks.

Fig. 4. Quality degree adjustment by controlled quality representation (the leftmost is the original input, from left to right: good to bad quality).

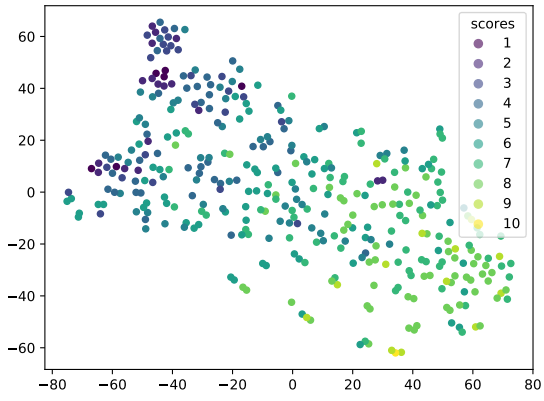

Fig. 5. t-SNE visualization of the quality representation z for test dataset

Acknowledgment

This work was supported in part by National Natural Science Foundation of China under Grant 61971095, Grant 61871078,Grant 61831005 and Grant 61871087.

References

[1] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski, *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010.

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

[3] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference quality assessment of tone-mapped hdr pictures," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2957–2971, 2017.

[4] K. Gu, S. Wang, G. Zhai, S. Ma, X. Yang, W. Lin, W. Zhang, and W. Gao, "Blind quality assessment of tone-mapped images via analysis of information, naturalness, and structure," *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 432–443, 2016.

[5] G. Yue, C. Hou, and T. Zhou, "Blind quality assessment of tone-mapped images considering colorfulness, naturalness, and structure," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 5, pp. 3784–3793, 2019.

[6] V. A. Kumar, S. Gupta, S. S. Chandra, S. Raman, and S. S. Channappayya, "No-reference quality assessment of tone mapped high dynamic range (hdr) images using transfer learning," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2017, pp. 1–3.

[7] Q. He, D. Li, T. Jiang, and M. Jiang, "Quality assessment for tone-mapped hdr images using multi-scale and multi-layer information," in *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2018, pp. 1–6.

[8] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2image: Conditional image generation from visual attributes," in *European Conference on Computer Vision*. Springer, 2016, pp. 776–791.
[9] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *arXiv: Learning*, 2016.
[10] Z. Lin, K. K. Thekumparampil, G. Fanti, and S. Oh, "Infogan-cr: Disentangling generative adversarial networks with contrastive regularizers." *arXiv: Learning*, 2019.
[11] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–189.
[12] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.
[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[15] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.
[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv: Machine Learning*, 2013.
[17] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo method*. John Wiley & Sons, 2016, vol. 10.
[18] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International conference on machine learning*, 2013, pp. 1139–1147.
[19] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "Large-scale crowdsourced study for tone-mapped hdr pictures," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4725–4740, 2017.
[20] G. Yue, C. Hou, K. Gu, S. Mao, and W. Zhang, "Biologically inspired blind quality assessment of tone-mapped images," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 3, pp. 2525–2536, 2018.
[21] Y. Zhang and D. M. Chandler, "No-reference image quality assessment based on log-derivative statistics of natural scenes," *Journal of Electronic Imaging*, vol. 22, no. 4, pp. 043 025–043 025, 2013.
[22] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4850–4862, 2014.
[23] L. V. Der Maaten and G. E. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.