Joint Quality Assessment and Example-Guided Image Processing by Disentangling Picture Appearance from Content

Abhinau K. Venkataramanan, Cosmin Stejerean, Ioannis Katsavounidis,
Hassene Tmar, and Alan C. Bovik *Life Fellow, IEEE*

***Abstract*—The deep learning revolution has strongly impacted low-level image processing tasks such as style/domain transfer, enhancement/restoration, and visual quality assessments. Despite often being treated separately, the aforementioned tasks share a common theme of understanding, editing, or enhancing the appearance of input images without modifying the underlying content. We leverage this observation to develop a novel disentangled representation learning method that decomposes inputs into content and appearance features. The model is trained in a self-supervised manner and we use the learned features to develop a new quality prediction model named DisQUE. We demonstrate through extensive evaluations that DisQUE achieves state-of-the-art accuracy across quality prediction tasks and distortion types. Moreover, we demonstrate that the same features may also be used for image processing tasks such as HDR tone mapping, where the desired output characteristics may be tuned using example input-output pairs.**

***Index Terms*—Disentangled Representation Learning, Domain Transfer, Quality Assessment, High Dynamic Range, Example-Guided Image Processing.**

I. Introduction

Recent years have witnessed an explosion in the amount of image and video content being shared over the internet. These images and videos are captured, often by uncertain hands, using cameras of various capabilities that may introduce distortions such as blur, noise, under/overexposure, etc. Following capture, they are commonly subjected to distortions such as compression, scaling, and brightness or contrast distortions during the transmission and display processes. Moreover, images and videos may also be edited by artists to modify their appearance, by making images brighter or darker, changing colors and color saturation, boosting contrast, etc.

At the same time, more sophisticated imaging and display modalities such as high dynamic range (HDR), high frame rate (HFR), and immersive media are rapidly growing. In particular, HDR enables the capture and representation of a wider range of brightnesses and colors, thereby enabling a more realistic reproduction of natural scenes as compared to legacy Standard Dynamic Range (SDR) imaging systems. However, a substantial portion of existing displays are not capable of displaying brightnesses above 1000 nits, which is essential for HDR [1]. So, HDR images and videos must be down-converted to SDR using a process called tone-mapping, so that they may be displayed on legacy displays. Although several algorithms have been proposed to automatically perform tone mapping [2] [3], color grading by human experts remains the gold standard.

Therefore, in both of the aforementioned scenarios, corresponding to the handling of SDR and HDR images/videos, two tasks must be effectively conducted to reliably transmit high-quality videos to consumers. First, objective models are needed that predict subjective opinions regarding the visual quality of images and videos. Such models may be used to control the quality of ingested content on social media websites and identify poor-quality content, which may affect downstream recommendation decisions. Quality models have also been used extensively to optimize processing parameters such as compression and resolution while trading off storage and transmission costs against perceptual quality [4].

Secondly, to control the quality of streamed content, image processing methods are also needed that can enhance specified aspects of images, such as brightness, contrast, color, etc. Typically, such fine-grained editing requires the use of specialized algorithms that provide "tunable knobs" corresponding to various image features. Indeed, this approach has led to the development of many tone-mapping algorithms that contain parameters to control aspects of the appearance of tone-mapped images. For example, the photographic tone reproduction method [5], which we refer to as "Reinhard02" here, uses a "desaturation" parameter to correct oversaturated colors. However, a human expert who is performing tone-mapping or evaluating its quality does not target a "desaturation level." Rather, a colorist tunes image properties manually to achieve a desired "look," which may depend on the colorist's experience and preferences, and their perception of consumers' demands. We posit that such specifications are best described using examples, rather than analytical metrics. This motivates the task of "example-guided" image processing, in general, or tone-mapping in particular. More examples of tone-mapping methods and their tunable parameters are provided in Section IV-A.

Here, we propose a Disentangled Representation Learning (DRL) framework to create a deep neural network model that can be used to tackle both image quality prediction and image processing tasks simultaneously. The general framework for using a common deep model for both quality assessment and image processing tasks is illustrated in Fig. 1.

First, an input image is decomposed into two feature sets,

This research was sponsored by a grant from Meta Video Infrastructure, and by grant number 2019844 for the National Science Foundation AI Institute for Foundations of Machine Learning (IFML).

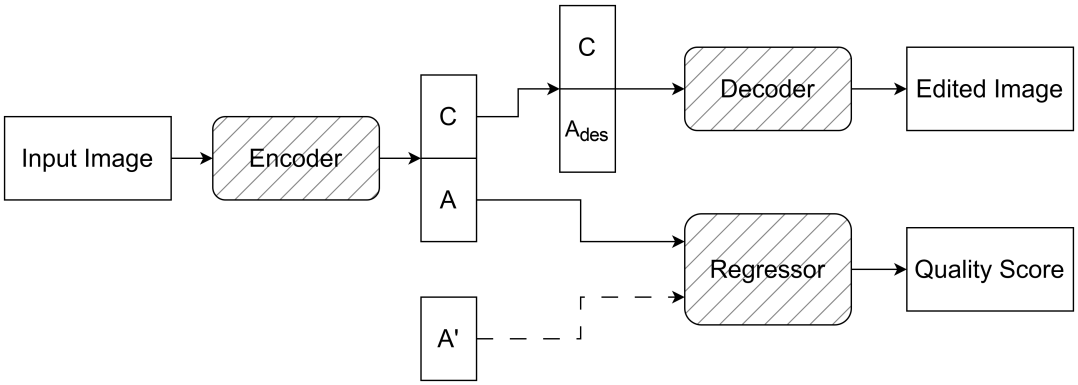

Fig. 1. Performing both quality prediction and image processing using the same disentangled representation learning model.

each describing the "image content" and "image appearance." During quality modeling, the appearance feature is compared against a reference appearance feature, to predict subjective quality. For example, when measuring the visual quality of a tone-mapped HDR video frame, its appearance feature is compared against the appearance feature of the source HDR video frame. We call this model the **Disentangled Quality Evaluator (DisQUE)**.

The input image may then be edited or enhanced by modifying the appearance feature of the input image to a "desired appearance feature," followed by reconstructing it using a decoder. We obtain the desired appearance feature using a pair of input images that are used as an example of (in this case) the desired tone-mapping behavior. We term this **"example-guided tone mapping" (EGTM).**

The remainder of this paper is organized as follows. In Section II, we discuss relevant prior work in the fields of visual quality assessment (VQA) and DRL and explain the novelty of our proposed model. Section III provides a detailed description of our proposed DisQUE model, including the learning objective, deep neural network architecture, and feature extraction protocol for quality prediction. In Section IV, we describe the training and evaluation datasets corresponding to the two domains in which we evaluate DisQUE. Specifically, we describe the datasets used by DisQUE to predict the quality of tone-mapped and compressed HDR videos, and to predict the quality of SDR images. We present training details in Section V-A and the results of quality modeling experiments in Section V-B. Furthermore, we demonstrate the ability of the DRL model to perform example-guided HDR tone-mapping in Section V-C. Finally, we present a summary of our findings and identify avenues for future work in Section VI.

II. Background and Novelty

A. Visual Quality Assessment

Objective models of visual quality may be broadly classified into "classical" (or hand-crafted) or "deep" (data-driven deep networks) methods. Full-reference (FR) quality models compare "distorted" test pictures/videos against their "pristine" reference counterparts to predict their visual quality. Models like SSIM [6], VIF [7], and ST-RRED [8] are examples of general-purpose classical FR quality models. By contrast, models such as DLM [9], VMAF [10], and FUNQUE [11] [12] are task-specific models designed to predict the quality of scaled and compressed videos.

FR models targeting similar applications have also been developed for HDR pictures and videos. Examples of such quality models include HDRMAX-VMAF [13] and HDR-FUNQUE+ [14]. In addition, quality models such as TMQI [15], FSITM [16], and Cut-FUNQUE [17] compare HDR and SDR pictures and videos to assess the quality of HDR tone mapping.

Deep FR quality modeling may be performed using deep networks pre-trained on large datasets such as ImageNet. For example, LPIPS [18], DISTS [19], and DeepWSD [20] all utilize ImageNet-pretrained models.

When pristine reference content is not available, No-reference (NR) quality models are employed. Examples of classical NR models include BRISQUE [21], DIIVINE [22], TLVQM [23], HIGRADE [24], and ChipQA [25]. Deep NR models may be trained either in a supervised or a self-supervised manner. Examples of supervised deep NR models include CNN-based models such as PaQ-2-PiQ [26], Patch-VQ [27], and QFM-IQM [28], and transformer-based models such as MUSIQ [29], RKIQT [30], LoDA [31], and SaTQa.

Recently, a number of high-performing self-supervised NR models have been introduced, including CONTRIQUE [32], Re-IQA [33], and ConViQT [34], and may also be used for NR quality prediction. All three self-supervised methods utilize ResNet-50 backbones and contrastive learning techniques such as SimCLR [35] and MoCo [36] [37] to learn quality-aware representations. The predicted features from test images/videos may be used for NR quality modeling, while the differences in predicted features between the reference and test images may be used for FR quality modeling.

B. Disentangled Representation Learning

Disentangled representation learning (DRL) refers to representation learning techniques that impose a notion of independence between subsets of the learned features. A survey of DRL methods and various taxonomic classifications are provided in [38]. Consider a network learning a vector of features. The disentanglement condition may be applied to each

dimension of the feature vector, which reflects the assumption that each dimension encodes one generative factor of the data distribution being modeled. Examples of dimension-wise disentangling include variational autoencoder (VAE) methods such as FactorVAE [39] and β-TCVAE [40], generative adversarial network (GAN) methods such as InfoGan [41], PS-SC GAN [42] and OroJaR GAN [43], and Barlow Twins [44], which is a self-supervised representation learning method.

Rather than disentangling feature vectors dimension-wise, their subsets may be disentangled to separate specific aspects of the data distribution. For example, DR-GAN [45] separates face and pose information to conduct pose-invariant face retrieval, while MAP-IVR [46] disentangles content and motion information from videos to conduct image-to-video retrieval. Feature subsets may be disentangled using cosine distances [46], minimizing correlations [44], or by minimizing mutual information using techniques such as CLUB [47] or adversarial losses [48].

Two key applications of DRL to image processing that are relevant here are style transfer and domain adaptation. The goal of DRL in such tasks is to decompose an image into a feature set that is common across domains, typically encoding "content," and one that is domain-specific, typically encoding "style." Examples of such methods include DRIT++ [49] and [50].

Efforts have also been made to combine image restoration and image quality assessment tasks, both with and without disentangled representations. QAIRN [51] uses a residual attention mechanism to gate encoder and decoder signals in image restoration networks. Due to the gating effect, the attention maps in QAIRN have been shown to learn local quality-aware feature maps. QD-Net [52] uses disentangled representations to perform no-reference (NR) quality prediction and enhancement of tone-mapped HDR images. QD-Net is trained in a supervised manner using ground-truth subjective ratings, while the enhancement network is trained using predefined enhancement targets. Alhough the "amount of enhancement" may be varied, the exact nature of the enhancement, such as improving color, brightness, or contrast, cannot be controlled. Therefore, if the desired enhancement targets change, for example, due to changes in consumer preferences, the enhancement network must be retrained.

Similar to DQ-Net, DRIQA [53] targets joint supervised quality assessment and restoration of SDR images. In this work, the "content encoder" extracts representations that do not contain distortion information. Therefore, the output of the content encoder is used directly for restoration, while the appearance encoder captures only distortion-related information. The outputs of the distortion encoder are used to augment a Siamese network [54] designed for quality assessment. Once again, the restoration targets are fixed and the decoder must be retrained if the restoration behavior is to be modified.

C. Novelty of the Proposed DRL Method

Our proposed DRL method differs from prior work in the following key aspects. First, prior work on disentangled domain transfer either uses different encoders/decoders for each domain [50] and/or categorical inputs to specify the target domain [49] for multi-domain adaptation. By contrast, we use a **fixed pair of networks to disentangle features**, and the predicted appearance features are used directly to specify the target domain for image processing.

Secondly, both DQ-Net and DRIQA solve restoration tasks using pre-defined restoration targets. We instead adopt an **"example-guided image processing" (EGIP)** framework that learns general appearance-related representations at training time. During inference, the desired processing behavior is expressed using a pair of example images that include a source image and its processed version. The DRL model infers the desired transform from the example and applies it to the input source image to be processed.

Thirdly, we propose a novel method for adapting images across domains called **"appearance mixing,"** as an alternative to "appearance replacement" methods used in prior work [49] [50] [52]. We observed that appearance replacement led to inaccurate adaptation across domains due to the presence of "confounding appearance features" (CAFs) in the source image. We demonstrate the effects of CAFs in Section V-C and show that they may be mitigated by using appearance mixing.

Finally, both DQ-Net and DRIQA-NR were trained in a supervised manner using ground-truth subjective quality scores. By contrast, DisQUE is trained in a **self-supervised** manner without the need for subjectively annotated data.

III. DisQUE

Here, we describe our proposed disentangled representation learning algorithm, which we use to develop DisQUE. The goal of the learning algorithm is to decompose an input image into its "content" and "appearance" components. In prior models such as ReIQA [33], "content" features are extracted to identify semantic content, such as objects in the image. To achieve high object detection accuracy, these models are designed to be robust to small changes in structure and orientation. By contrast, small changes in structure are visible to the human eye and perceptually important, as evidenced by advancements in restoration tasks such as image deblurring.

Here, we interpret the content to be the "high-resolution" intrinsic structure of the image, which acts as a scaffolding that is modulated by "appearance." Appearance, on the other hand, are properties that vary slowly across an image and include aspects such as color, contrast, brightness, and sharpness. Modeling content and appearance under these assumptions is similar to the decomposition of scenes into reflectance and illumination components [55].

A. Learning Objective

Consider a dataset of images $\mathcal{X} = \{x_i\}$ and a bank of image transforms $\mathcal{T} = \{T_j\}$ that alter one or more aspects of the appearance of input images. Examples of such transforms include blurring, brightening, compression, color changes, or tone-mapping operators for HDR. A list of transforms commonly used in HDR and SDR experiments is presented in Section IV. We sample two image patches $x_1, x_2 \sim \mathcal{X}$ and

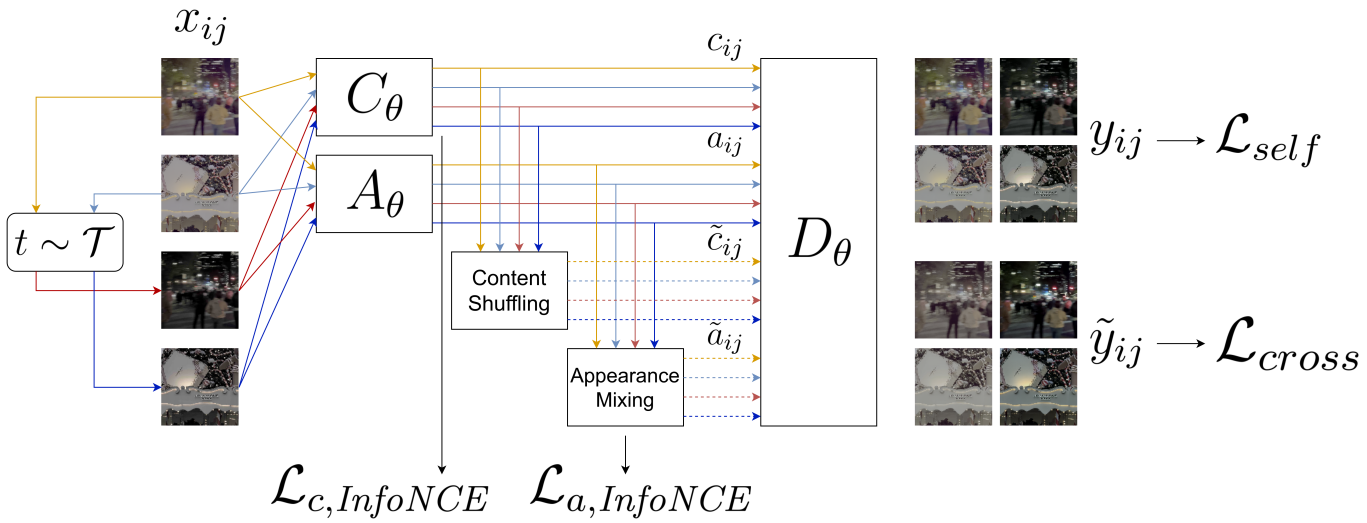

Fig. 2. Visualizing the disentangled representation learning objective.

a transform $t \sim \mathcal{T}$. Applying the transform to both images, we obtain two views of each image:

$$x_{11} = x_1, \quad x_{12} = t\left(x_1\right), \quad x_{21} = x_2, \quad x_{22} = t\left(x_2\right). \tag{1}$$

Let C_θ and A_θ denote two parameterized encoders that map input images to their content and appearance features respectively:

$$c_{ij} = C_\theta(x_{ij}), \quad a_{ij} = A_\theta(x_{ij}). \tag{2}$$

Finally, let D_θ denote a parameterized decoder that maps content and appearance features to images. When content and appearance features extracted from an image are reconstructed, we expect to recover the input image. We term this the **self-reconstruction** objective. The reconstruction loss is a weighted sum of pixel-domain and frequency-domain losses, as used in MAXIM [56]. That is, let

$$y_{ij} = D_\theta\left(c_{ij}, a_{ij}\right), \tag{3}$$

$$\mathcal{L}_{self} = \sum_{ij} \mathcal{L}_{char}(x_{ij}, y_{ij}) + \lambda_f \mathcal{L}_{freq}(x_{ij}, y_{ij}), \tag{4}$$

where $\mathcal{L}_{char}$ denotes the Charbonnier loss

$$\mathcal{L}_{char}(x, y) = \sqrt{\|x - y\|^2 + \epsilon^2} \tag{5}$$

and $\mathcal{L}_{freq}$ denotes the frequency loss, which uses the discrete Fourier transform, denoted by $\mathcal{F}$

$$\mathcal{L}_{freq}(x, y) = \left\|\mathcal{F}\left\{x\right\} - \mathcal{F}\left\{y\right\}\right\|_1. \tag{6}$$

The main contribution of our proposed disentangled representation learning algorithm, which enables the separation of content and appearance features, is the **cross-reconstruction objective**. The goal of cross-reconstruction is to predict an image x_{ij} using features from images other than x_{ij}. For example, suppose we wish to predict x_{12}. Since c_{11} encodes the content in image 1 and a_{22} encodes appearance after applying transformation t, one may predict x_{12} as $D_\theta\left(c_{11}, a_{22}\right)$. Such a cross-reconstruction method has been used in prior work, such as DRIT [49], to disentangle content and appearance.

However, the appearance feature a_{22} includes information not only about the effect of t, but also of the source image x_2. Such "confounding" appearance features (CAFs) may be transferred if cross-reconstruction is performed in this manner. For example, if x_1 is a picture of a green field and x_2 is that of a yellow flower, $D_\theta\left(c_{11}, a_{22}\right)$ may yield a field with a yellow hue.

To remove the effect of CAFs, we adopt a novel **"appearance mixing"** method. We first note that the difference between x_{i1} and x_{i2} is only the effect of the transform t. Therefore, $\Delta a_i = a_{i2} - a_{i1}$ captures the effect of t while eliminating CAFs from image x_i. This difference is then "mixed into" other appearance features to add or remove the effect of t and yield cross-reconstructed images. Hence, the crossed appearance features after undergoing appearance mixing are

$$\begin{aligned}
\tilde{a}_{11} &= a_{12} - \Delta a_2, \\
\tilde{a}_{12} &= a_{11} + \Delta a_2, \\
\tilde{a}_{21} &= a_{22} - \Delta a_1, \\
\tilde{a}_{22} &= a_{21} + \Delta a_1.
\end{aligned} \tag{7}$$

To obtain crossed reconstructions, we apply **content shuffling** to replace content features across domains, yielding the following crossed content features

$$\begin{aligned}
\tilde{c}_{11} &= c_{12}, \\
\tilde{c}_{12} &= c_{11}, \\
\tilde{c}_{21} &= c_{22}, \\
\tilde{c}_{22} &= c_{21}.
\end{aligned} \tag{8}$$

The crossed content and appearance features are used to obtain cross-reconstruction predictions

$$\tilde{y}_{ij} = D_\theta\left(\tilde{c}_{ij}, \tilde{a}_{ij}\right), \tag{9}$$

which are evaluated using the cross-reconstruction objective

$$\mathcal{L}_{cross} = \sum_{ij} \mathcal{L}_{char}(x_{ij}, \tilde{y}_{ij}) + \lambda_f \mathcal{L}_{freq}(x_{ij}, \tilde{y}_{ij}). \tag{10}$$

The effect of using appearance mixing to mitigate the effect of CAFs is illustrated in Section V-C.

Since x_{i*} are views of the same image, we expect content representations generated by a good content encoder to satisfy

$$c_{11} \approx c_{12}, \quad c_{21} \approx c_{22}. \tag{11}$$

Moreover, since the two input images were transformed by the same transformation t, we expect appearance representations generated by a good appearance encoder to satisfy

$$\Delta a_1 \approx \Delta a_2. \tag{12}$$

We guide the networks to learn these properties using a symmetrized InfoNCE [57] contrastive loss, as in MoCov3 [58]. InfoNCE loss aims to maximize the similarity between a query-positive key pair (q, k^+) while minimizing the similarity between the query and a set of negative keys K^-:

$$\mathcal{L}_{InfoNCE}(q, k^+, K^-) = -\log \frac{\exp\left(q \cdot k^+/\tau\right)}{\sum\limits_{\{k^+\} \cup K^-} \exp\left(q \cdot k/\tau\right)}. \tag{13}$$

Given a batch of training samples, the "content contrastive loss" $\mathcal{L}_{c,InfoNCE}$ is obtained by using $([c_{11}, c_{21}], [c_{12}, c_{22}])$ from the same sample as positive query-key pairs and those from different samples in the batch as negative pairs. Similarly, the "appearance contrastive loss" $\mathcal{L}_{a,InfoNCE}$ is obtained by using $(\Delta a_1, \Delta a_2)$ from the same sample as positive query-key pairs and those from different samples as negative pairs.

$$\mathcal{L}_{c,InfoNCE} \sim ([c_{11}, c_{21}], [c_{12}, c_{22}]) \tag{14}$$

$$\mathcal{L}_{a,InfoNCE} \sim (\Delta a_1, \Delta a_2) \tag{15}$$

Hence, the parameters of $(C_\theta, A_\theta, D_\theta)$ are trained to minimize the overall learning objective

$$\mathcal{L} = (\mathcal{L}_{self} + \mathcal{L}_{cross}) + \beta\left(\mathcal{L}_{c,InfoNCE} + \mathcal{L}_{a,InfoNCE}\right). \tag{16}$$

A visualization of computation of the training objective is depicted in Fig. 2.

B. Network Architecture

Following prior work [32] [33], we adopted a ResNet-50-based architecture for both the content and appearance encoders, and a reversed ResNet-50 architecture for the decoder, with key modifications. Due to the presence of two encoders and one decoder, we term this architecture a "dual-head" U-Net [59]. To limit the number of features used in downstream picture quality assessment tasks, we adopted a ResNet-50 0.5x architecture, i.e., one that uses half the number of channels at each layer. The other departure from ResNet-50 was the removal of batch normalization layers, since they have been shown to hinder image-to-image translation performance [60] [61].

Moreover, we introduced instance normalization (IN) [62] layers to ResNet blocks in the content encoder to introduce the effect of "appearance normalization." IN layers normalize the statistics of each channel of the input feature map. Consider a feature map $F_{ncij} \in \mathbb{R}^{N \times C \times H \times W}$. Then, the output of the IN layer is

$$\tilde{F}_{nchw} = \frac{F_{nchw} - \mu_{nc}}{\sigma_{nc}}, \tag{17}$$

where

$$\mu_{nc} = \frac{1}{HW} \sum_{ij} F_{ncij} \tag{18}$$

and

$$\sigma_{nc} = \sqrt{\frac{1}{HW} \sum_{ij} \left(F_{ncij} - \mu_{nc}\right)^2}. \tag{19}$$

This follows prior work in style transfer that uses IN layers for style transfer [62] [63], which demonstrated that mean and standard deviations of layer activations may be used to encode "style." As a result, normalizing these statistics using IN layers was found to improve style transfer performance. Hence, we deploy IN layers to normalize appearance and retain only content-related features.

By contrast, the appearance encoder does not include IN layers since its goal is to capture appearance information. This is achieved by average pooling intermediate layer feature activations obtained from each ResNet block. Therefore, the appearance of the input to the network is captured by a single feature vector rather than a spatially-varying feature map. Despite being assumed to be slow-varying over space, appearance is a non-stationary attribute of images. For example, one region of an image may have bright objects while another has dark objects. Hence, the dual-head U-Net is best applied on small image patches, rather than on full images. Here, we use a patch size of 128×128.

The decoder follows a typical U-Net structure, using skip connections to introduce multi-level feature maps from the content encoder. Appearance features are introduced into the encoder using a product-based channel attention mechanism $CA(x, z) = x \otimes a$, similar to that used in residual attention networks [64]. We chose this mechanism since channel attention may be considered an inverse of instance normalization that re-introduces the desired appearance features, as evidenced by Adaptive Instance Normalization [65]. The overall dual-head U-Net structure of the proposed deep network is illustrated in Fig. 3.

C. Visual Quality Assessment

After training, we use the appearance encoder to conduct FR visual quality assessment. Because the encoder was trained to disentangle appearance information from image content, we term our quality predictor the **Disentangled Quality Evaluator (DisQUE)**.

Given a pair of reference and test images I_{ref} and I_{dis}, we obtain feature maps from the output of each of the four ResNet blocks in A_θ:

$$\mathcal{A}_{ref} = A_\theta(I_{ref}), \quad \mathcal{A}_{dis} = A_\theta(I_{dis}). \tag{20}$$

We then characterized feature maps by computing both the mean and standard deviation of each channel

$$a_\mu = E\left[\mathcal{A}\right], \quad a_\sigma = \sqrt{E\left[\left(\mathcal{A} - a_\mu\right)^2\right]}. \tag{21}$$

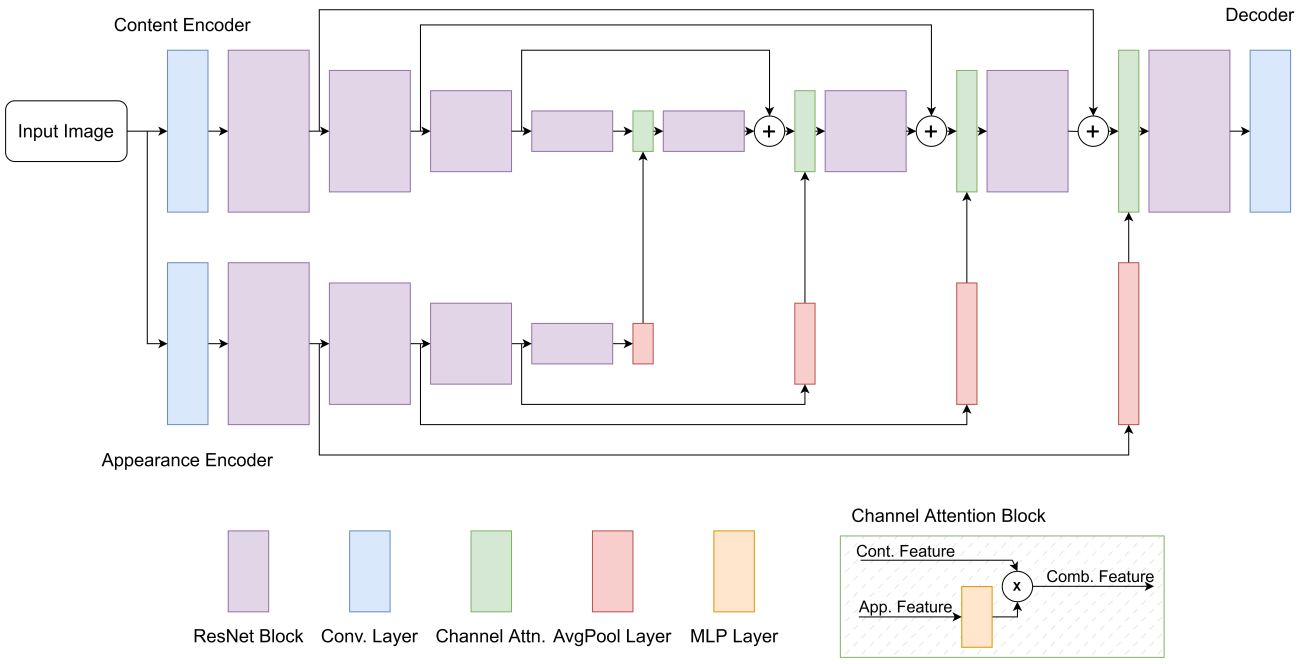

Fig. 3. The dual-head U-Net architecture.

Alhough computing the mean is typical, standard deviations of feature maps have also been used [66], albeit without explicit justification. Here, we justify the use of standard deviation by referring to one of the assumptions behind our disentangled representation model. We posited that appearance varies slowly over space, because of which we used spatially constant appearance vectors to describe image patches. However, since quality assessment is carried out over images, the standard deviation captures variations in appearance over space. Hand-crafted quality models such as ESSIM [67] and GMSD [68] have also benefited from characterizing spatial quality variations using standard deviations.

Following prior work [32] [33] [66], we captured multi-scale appearance features by repeating the process using input images rescaled to half resolution. The final feature vector for the reference and test images (z_{ref}, z_{dis}) was obtained by concatenating mean and standard deviation-pooled features at both scales. Since DisQUE is an FR quality model, the difference in features between the two images $z = |z_{ref} - z_{dis}|$ was used to predict quality.

The use of multi-scale, multi-block features pooled spatially using two methods yields a feature vector of size 8192. Finally, a linear regressor model was used to map the appearance features to subjective quality scores. Note that the self-supervised appearance network was frozen during inference time, and only the linear regressor was recalibrated on each evaluation dataset. Ablation experiments studying the effects of multi-scale features and the use of standard deviation pooling are presented in Section V.

IV. Datasets

A. HDR Datasets

To train DisQUE for HDR quality assessment, we used the recently developed LIVE UGC-HDR database [69], which is the first publicly available large-scale database of HDR videos. The database consists of over 2,153 HLG-encoded [70] videos filmed by amateur iPhone users, containing a diverse collection of scenes, including indoor, outdoor, daytime, nighttime, static, and dynamic scenes containing camera and object motion.

Since the proposed DRL method learns from images, we first sampled video frames from the set of HDR videos. To introduce sufficient content diversity, we sampled video frames at 2-second intervals, yielding a total of 19060 frames at an average of 8.85 frames per video. Since the dataset contains both 1080p and 4K videos, we rescaled all videos to 1080p using Lanczos rescaling. Finally, since the PQ [71] standard can represent a wider range of brightnesses, we re-encoded all sampled frames to 10-bit PQ from 10-bit HLG.

As described earlier, we aimed to train DisQUE to predict the quality of tone-mapped and compressed HDR videos. So, we used a bank of transforms (called $\mathcal{T}$ in Section III-A) consisting of the following ten open-source tone-mapping operators (TMOs), with their parameters varied to generate a diverse set of tone mapping-related distortions.

- **Hable** [3] - A parameter-free pointwise non-linear transform originally designed for use in the video game *Uncharted 2*. A desaturation parameter was varied to control how colorful tone-mapped images would appear.
- **Reinhard02** [5] - A point non-linearity to map luminances from HDR to SDR. A desaturation parameter was varied to control how colorful tone-mapped images would appear.

- **Durand02** [72] - Uses a "fast bilateral filter" to decompose the luminances of HDR frames into "base" and "detail" layers. A contrast parameter was varied to control the degree of global contrast, i.e., the difference between the visibilities of bright and dark regions.
- **Shan12** [73] - Uses an edge-aware stationary wavelet transform (SWT) [74]. The number of wavelet levels was varied, which affected contrast.
- **Reinhard12** [75] - Uses color-appearance models applied in a local manner. The assumed viewing conditions were varied, which introduced color distortions.
- **Eilertsen15** [76] - Applies a "fast detail extraction" method to obtain a base-detail decomposition and applies a dynamic tone-curve. The coarseness of the tone curve was varied, which led to contrast distortions.
- **Oskarsson17** [77] - Uses Dynamic Programming to cluster values in the input image channels. The number of clusters was varied, which introduced quantization artifacts such as banding.
- **Rana19** [78] - Uses a Generative Adversarial Network (GAN) to create a fully-convolutional, parameter-free TMO. A desaturation parameter was varied to control how colorful tone-mapped images appeared.
- **Yang21** [79] - Uses a deep convolutional neural network (CNN) to transform a multi-scale Laplacian pyramid decomposition of each input HDR frame. A desaturation parameter was varied to control how colorful tone-mapped images appeared.
- **ITU21** [2] - A parameter-free TMO proposed by the ITU in Recommendation BT.2446 ("Approach A"). The nominal HDR luminance was varied, which affected the brightness and contrast of tone-mapped images.

Furthermore, we introduced compression distortions by applying lossy JPEG compression at four levels to the tone-mapped images. Therefore, each transform in the bank $\mathcal{T}$ consists of tone-mapping using one of the aforementioned TMOs followed by JPEG compression.

We tested the efficacy of DisQUE on the LIVE Tone-Mapped HDR (LIVE-TMHDR) subjective database [80], which is the first public database of subjectively annotated tone-mapped and compressed HDR videos. LIVE-TMHDR consists of 15,000 distorted videos that were generated from 40 source contents (20 each encoded using PQ and HLG) using 13 tone-mapping methods and compressed using libx264 [81] at three quality levels. The 13 tone-mapping methods include the 10 TMOs discussed here, Dolby Vision tone-mapping [82], the Color Space Transform (CST) method used for gamut/tone mapping by colorists, and manual tone-mapping by a human expert colorist. Moreover, the TMOs were applied to videos using three "temporal modes," which varied the degree of temporal distortions.

B. SDR Datasets

To demonstrate the versatility of DisQUE, we also evaluated its performance on SDR FR quality assessment. We followed a similar approach as prior work [33], [32] to create a training dataset of SDR images from the following diverse sources.

- KADIS-700k [83] - ∼140K images
- AVA [84] - ∼ 255K images
- CERTH-Blur [85] - ∼ 2.5K images
- VOC [86] - ∼ 33K images
- COCO [87] - ∼ 330K images
- Places [88] - ∼ 2.2M images

In total, we obtained nearly 3M images from these data resources. As we will describe below, the bank of transforms used for SDR training includes color and contrast distortions. So, we excluded grayscale images and those having significant over/under-exposed regions to create a training dataset of nearly 1.8M training images.

The bank of transforms for SDR training was constructed using the following set of 25 distortions borrowed from [33], which may be applied at five degrees of severity each.

- **NNResize** - Downscale the image and upscale it back to its original resolution using nearest neighbor interpolation.
- **BilinearResize** - Downscale the image and upscale it back to its original resolution using bilinear interpolation.
- **BicubicResize** - Downscale the image and upscale it back to its original resolution using bicubic interpolation.
- **LanczosResize** - Downscale the image and upscale it back to its original resolution using Lanczos interpolation.
- **MotionBlur** - Simulate motion blur by filtering using directional blur kernels.
- **GaussianBlur** - Filter using a Gaussian kernel.
- **LensBlur** - Filter using a circular kernel.
- **MeanShift** - Add a constant value to all pixels.
- **Contrast** - Modify contrast using a sigmoidal non-linear transformation.
- **Compress** - Apply JPEG compression
- **UnsharpMasking** - Increase sharpness of the image.
- **ColorBlock** - Replace regions of images with small randomly colored patches.
- **Jitter** - Apply small random offsets to pixels.
- **PatchJitter** - Apply small random offsets to patches.
- **RGBNoise** - Add white noise to RGB pixels.
- **YUVNoise** - Add white noise in YUV space.
- **ImpulseNoise** - Add salt and pepper noise.
- **SpeckleNoise** - Multiply by speckle noise.
- **Denoise** - Add Gaussian noise and blur to denoise it.
- **Brighten** - Apply non-linear curve to increase brightness
- **Darken** - Apply non-linear curve to decrease brightness
- **ColorDiffuse** - Apply Gaussian blur to the a^* and b^* channels in CIELAB color space.
- **ColorShift** - Offset color channels.
- **HSVSaturate** - Multiply the saturation channel of HSV representation by a factor.
- **LABSaturate** - Multiply the a^* and b^* channels of the CIELAB representation by a factor.

Each of the aforementioned methods typically modifies only one aspect of each image, such as color, brightness, etc. However, the real world may present complex combinations of distortions. To simulate these scenarios, we constructed each transform in $\mathcal{T}$ as a composition of one to three randomly

TABLE I
EVALUATION OF QUALITY PREDICTION MODELS ON LIVE-TMHDR

Model	PCC	SROCC	RMSE
Y-FUNQUE+ [12]	0.4524	0.4343	9.4352
BTMQI [93]	0.4705	0.4663	9.2238
FSITM [16]	0.4813	0.4626	8.9212
NIQE [94]	0.4805	0.4746	9.5563
BRISQUE [21]	0.4811	0.4833	8.9869
DIIVINE [22]	0.4794	0.4925	9.2879
TMQI [15]	0.5062	0.4956	8.6897
FUNQUE [11]	0.5082	0.4949	8.8863
TMVQI [95]	0.5198	0.4969	8.8697
FFTMI [96]	0.5298	0.5315	8.8559
3C-FUNQUE+ [12]	0.5817	0.5661	8.6568
RcNet [97]	0.5985	0.5824	8.2417
HIGRADE [24]	0.6682	0.6698	8.2619
Cut-FUNQUE	0.7783	0.7781	6.4187
CONTRIQUE [32]	0.7360	0.7230	6.8476
ReIQA [33]	0.7583	0.7812	7.2951
MSML [66]	0.7883	0.7740	6.8090
DisQUE	**0.8160**	**0.8215**	**6.3241**

chosen unit distortions, each applied at a randomly chosen level of severity.

We evaluated the SDR DisQUE model on the four FR picture quality assessment datasets. The LIVE-IQA dataset [89] consists of 29 reference images subjected to five distortions - blur, noise, JPEG compression, JPEG2000 compression, and bit errors in JPEG2000 bitstreams. This procedure yielded a total of 982 distorted pictures. The CSIQ dataset [90] consists of 866 test images generated from 30 source contents subjected to six distortions - blur, noise, JPEG compression, JPEG2000 compression, pink Gaussian noise, and global contrast decrements.

TID2013 [91] is a dataset of 3000 test images generated by applying 24 impairments at five levels each to a dataset of 25 images. The set of distortions includes blur, noise, compression, bitstream errors, contrast and color distortions, and spatial distortions such as jitter and color blocking. Finally, KADID-10k [92] is the largest database on the list, containing 10,125 test images generated by subjecting 81 pristine images to 25 distortions at five levels each. The types of distortions in KADID-10k are similar to those in TID2013.

V. EVALUATION

A. Training

In both the HDR and SDR cases, the dual-head U-Net models were trained using a batch size of 36, split across 9 NVIDIA A-100 GPUs. Note that each sample consists of four randomly sampled and transformed (using $t \sim \mathcal{T}$) 128×128 image patches $(x_{11}, x_{12}, x_{21}, x_{22})$, as described in Section III-A. The dual-head U-Net was trained for 400K steps using an Adam optimizer configured with an initial learning rate of 0.0002. The learning rate was decayed by 0.99 every 1,000 steps, and the loss hyperparameters were set to $\lambda_f = 0.1$ and $\beta = 0.5$.

B. Evaluating HDR and SDR Quality Prediction

We evaluated DisQUE's HDR tone-mapping quality predictions on the LIVE-TMHDR video quality dataset, and SDR quality prediction on four datasets - LIVE IQA, CSIQ, TID2013, and KADID-10k. In all cases, as described in Section III-C, DisQUE generated an 8192-dimensional feature vector by applying the appearance encoder to both the reference and test pictures/video frames. For video quality prediction on LIVE-TMHDR, we averaged the feature vector obtained from each frame to yield a video-level feature vector.

We evaluated quality prediction accuracy on each dataset using a 10-fold random cross-validation, where 80% of the dataset was used to train the PLS projector and the Linear SVR predictor, and 20% was used for testing. The hyperparameters of the Linear SVR predictor were selected by performing five-fold cross-validation on the training dataset of each random cross-validation split. In addition to the combination of PLS and Linear SVR, we also experimented with Lasso and Ridge regressors, and the best regression model was chosen. Note that all three approaches yield linear prediction models and so, incur similar computational complexities at inference time. The best regressor models and their hyperparameters were identified by optimizing the average of the Pearson Correlation Coefficient (PCC) and Spearman Rank Order Correlation (SROCC) between the predicted and ground-truth subjective scores on the validation datasets.

The quality prediction accuracies of various hand-crafted and deep quality models on the LIVE-TMHDR database are presented in Table I, while the SDR quality prediction outcomes are presented in Table II. From the Tables, it may be observed that DisQUE outperformed all the compared models compared in Table I on the tone-mapping quality prediction task, and achieved comparable state-of-the-art (SOTA) accuracy among the compared self-supervised models on the SDR quality prediction task.

We further analyzed the effect of feature subsets of DisQUE on quality prediction accuracy in an ablation study. As explained in Section III-C, DisQUE combines four subsets of features generated by applying mean and standard deviation pooling to the appearance encoder's feature maps. The impact of introducing each feature subset into DisQUE is quantified in Table III. From this Table, it may be seen that both multi-

TABLE II
EVALUATION OF QUALITY PREDICTION MODELS ON SDR QUALITY DATABASES

Type	Model	LIVE IQA		CSIQ		TID2013		KADID-10k	
		PCC	SROCC	PCC	SROCC	PCC	SROCC	PCC	SROCC
Hand-crafted	PSNR	0.868	0.881	0.824	0.820	0.675	0.643	0.680	0.677
	BRISQUE [21]	0.935	0.939	0.829	0.746	0.694	0.604	0.567	0.528
	SSIM [6]	0.911	0.921	0.835	0.854	0.698	0.642	0.633	0.641
	FSIM [98]	0.954	0.964	0.919	0.934	0.875	0.852	0.850	0.854
	CORNIA [99]	0.950	0.947	0.776	0.678	0.768	0.678	0.558	0.516
Supervised Deep Nets	DB-CNN [100]	0.971	0.968	0.959	0.946	0.865	0.816	0.856	0.851
	PQR [101]	0.971	0.965	0.901	0.872	0.798	0.740	-	-
	HyperIQA [102]	0.966	0.962	0.942	0.923	0.858	0.840	0.845	0.852
	LoDA [31]	0.979	0.975	0.901	0.869	-	-	0.936	0.931
	DRF-IQA [103]	0.983	0.983	0.960	0.964	0.942	0.944	-	-
	RKIQT [30]	0.986	0.984	0.970	0.958	0.917	0.900	0.911	0.911
	SaTQA [104]	0.983	0.983	0.972	0.965	0.948	0.938	0.949	0.946
	DRIQA [53]	0.989	0.985	0.980	0.978	0.961	0.954	-	-
Self-Supervised Deep Nets	LPIPS [18]	0.936	0.932	0.906	0.884	0.756	0.673	0.713	0.721
	CONTRIQUE [32]	0.966	0.966	**0.964**	0.956	**0.915**	**0.909**	**0.947**	**0.946**
	ReIQA [33]	**0.974**	**0.973**	**0.965**	**0.961**	**0.915**	0.905	0.903	0.901
	DisQUE	**0.972**	**0.970**	0.956	**0.961**	0.909	**0.922**	**0.921**	**0.934**

TABLE III
ABLATION EXPERIMENTS STUDYING THE EFFECT OF MULTI-SCALE AND MULTI-POOLING FEATURES

DisQUE Variant	LIVE-TMHDR		LIVE IQA		CSIQ		TID2013		KADID-10k	
	PCC	SROCC	PCC	SROCC	PCC	SROCC	PCC	SROCC	PCC	SROCC
Single-Scale, Mean Pooling	0.767	0.762	0.961	0.956	0.926	0.940	0.865	0.879	0.889	0.907
Multi-Scale, Mean Pooling	0.803	0.807	0.968	0.961	0.947	0.951	0.891	0.898	0.908	0.923
Single-Scale, Mean+Std Pooling	0.805	0.804	0.967	0.964	0.948	0.956	0.900	0.915	0.910	0.928
Multi-Scale, Mean+Std Pooling	**0.816**	**0.822**	**0.972**	**0.970**	**0.956**	**0.961**	**0.909**	**0.922**	**0.921**	**0.934**

scale features and standard-deviation pooling improved quality prediction accuracy across databases.

C. Example-Guided Tone Mapping

A key element of training DisQUE is the use of appearance mixing to yield crossed predictions $\tilde{y}_{ij}$. In addition to being a representation learning framework, the DisQUE training paradigm may also be used to perform **example-guided image processing (EGIP)**.

We define an EGIP task as consisting of three inputs - the example source, the example target, and the input source image. The EGIP model then processes the input source image to induce a similar effect as shown by the example pair. For example, if the example target is a blurred version of the example source, an ideal EGIP network predicts the target image as a blurred version of the input source. EGIP, as defined here, may be seen as a task this is complementary to guided image filtering [105], where content from a guidance image (analogous to the example pair here) is transferred to the input, rather than appearance.

Here, we demonstrate the EGIP capability of the disentangling network by performing example-guided HDR tone mapping (EGTM). Fig. 4 shows examples of using EGTM to vary visual characteristics such as color, contrast, and brightness of tone-mapped images. We achieved this by varying the corresponding attribute of the example SDR image, and it may be seen that the network was able to characterize the differences between the example HDR and SDR images and transfer those characteristics to the input HDR image to predict the corresponding tone-mapped SDR image.

The EGTM results presented in Fig. 4 all use the appearance mixing method described in Section III-A to improve their robustness to CAFs. To demonstrate the usefulness of appearance mixing, we illustrate the outputs of EGTM when using the naive appearance replacement approach, also described in Section III-A. From Fig. 5, it may be observed that the green color of the wall in the source HDR, which is a CAF, results in a green sky in the predicted SDR when using appearance replacement. However, predicting the SDR image using appearance mixing significantly reduced the effect of the CAF, thereby improving prediction accuracy relative to the "ground-truth" target SDR. Note that in this example, both the example and target SDR images were generated using the Hable TMO [3] with a desaturation parameter of 0.

VI. CONCLUSION AND DISCUSSION

We have developed a novel framework for disentangled representation learning using a new "appearance mixing" framework for adapting images across domains/appearance classes. The DRL network learns to decompose an input image into content and appearance-related features, which we used for two downstream tasks - perceptual quality modeling and example-guided image processing.

We found that our DRL-based quality model DisQUE achieved state-of-the-art accuracy when predicting the quality of tone-mapped and compressed HDR videos, and of synthetically distorted SDR images. In addition, we demonstrated the EGIP capabilities of the DRL model by performing example-guided tone mapping. Specifically, we gave examples of how particular image appearance features can be modulated using appropriately chosen examples.

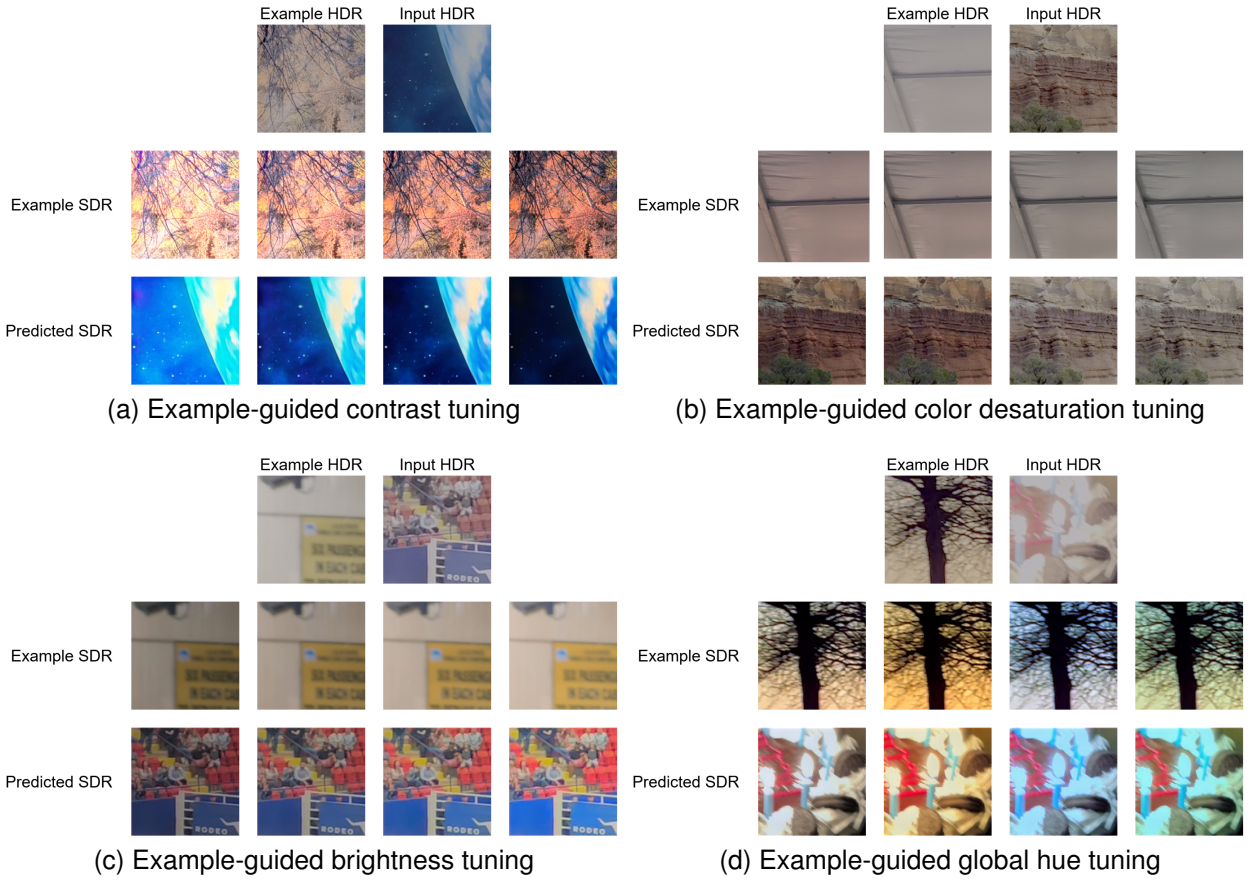

Fig. 4. Examples of using example-driven tone mapping to vary tone-mapping characteristics

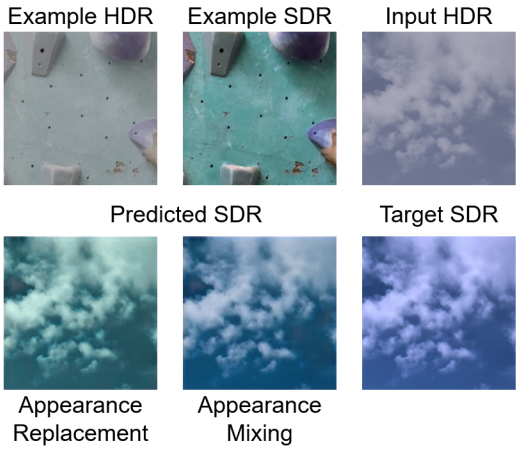

Fig. 5. Effect of confounding appearance features on EGTM.

Despite the promising performance of DisQUE and EGTM, the proposed DRL model has some limitations. First, since appearance mixing relies on differences between source and target domain appearance features, DisQUE is only able to perform full-reference quality modeling. Existing self-supervised methods such as CONTRIQUE and ReIQA can perform NR quality modeling since the appearance features, and not their differences, are used to distinguish between images. Combining these two methods may yield a network suitable for both FR and NR quality modeling.

Moreover, one of the assumptions used to design the DRL method was that appearance does not vary over image patches. However, appearance does vary over larger spatial regions, such as 1080p or 4K images. Therefore, the use of an average-pooled constant appearance vector, rather than a spatially varying feature map, limits the application of EGIP to patches. To enable the example-guided processing of high-resolution inputs, the appearance representation may be modified to include spatial information.

Finally, we note that the EGTM results in Fig. 5 showed that appearance mixing alone may not be sufficient for accurate EGTM. In this example, the predicted SDR was still a different shade of blue compared to the ground-truth target. This may be attributed to the fact that the example and input HDR images were significantly different in their visual characteristics - the example is predominantly a green wall, while the input is predominantly a blue sky. This suggests that a better approach during inference may be to use a bag of example pairs, and dynamically choose the most relevant example for every input image to be processed. More generally, a similar idea may be applied to the processing of high-resolution images by using patch attention methods to exploit "good example patches" from each example image pair.

REFERENCES

[1] CNET. Best TV for 2024: We tested Samsung, LG, TCL, Vizio and more. [Online]. Available: https://www.cnet.com/tech/home-entertainment/best-tv/

[2] ITU-R, "ITU-R BT.2446: Methods for conversion of high dynamic range content to standard dynamic range content and vice-versa," 2021.

[3] J. Hable. Uncharted 2: HDR lighting. [Online]. Available: https://www.gdcvault.com/play/1012351/Uncharted-2-HDR

[4] I. Katsavounidis, "Dynamic optimizer - a perceptual video encoding optimization framework," March 2018. [Online]. Available: https://netflixtechblog.com/dynamic-optimizer-a-perceptual-video-encoding-optimization-framework-e19f1e3a277f

[5] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Transactions on Graphics*, vol. 21, no. 3, p. 267–276, Jul 2002.

[6] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[7] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[8] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 684–694, 2013.

[9] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935–949, 2011.

[10] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, p. 2, 2016.

[11] A. K. Venkataramanan, C. Stejerean, and A. C. Bovik, "FUNQUE: Fusion of unified quality evaluators," in *IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 2147–2151.

[12] A. K. Venkataramanan, C. Stejerean, I. Katsavounidis, and A. C. Bovik, "One transform to compute them all: Efficient fusion-based full-reference video quality assessment," *IEEE Transactions on Image Processing*, vol. 33, pp. 509–524, 2024.

[13] Z. Shang, J. P. Ebenezer, A. K. Venkataramanan, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik, "A study of subjective and objective quality assessment of HDR videos," *IEEE Transactions on Image Processing*, vol. 33, pp. 42–57, 2024.

[14] A. K. Venkataramanan, C. Stejerean, I. Katsavounidis, and A. C. Bovik, "A FUNQUE approach to the quality assessment of compressed HDR videos," *arXiv preprint arXiv:2312.08524*, 2023.

[15] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 657–667, 2013.

[16] H. Ziaei Nafchi, A. Shahkolaei, R. Farrahi Moghaddam, and M. Cheriet, "FSITM: A feature similarity index for tone-mapped images," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1026–1029, 2015.

[17] A. K. Venkataramanan, C. Stejerean, I. Katsavounidis, H. Tmar, and A. C. Bovik, "Cut-FUNQUE: Objective quality assessment of compressed and tone mapped high dynamic range videos," *Manuscript Under Submission*, vol. 1, 2024.

[18] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[19] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2022.

[20] X. Liao, B. Chen, H. Zhu, S. Wang, M. Zhou, and S. Kwong, "DeepWSD: Projecting degradations in perceptual space to wasserstein distance in deep feature space," in *ACM International Conference on Multimedia*, 2022, p. 970–978.

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

[22] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[23] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5923–5938, 2019.

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

[25] J. P. Ebenezer, Z. Shang, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik, "ChipQA: No-reference video quality prediction via space-time chips," *IEEE Transactions on Image Processing*, vol. 30, pp. 8059–8074, 2021.

[26] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3575–3585.

[27] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-VQ: 'Patching up' the video quality problem," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 019–14 029.

[28] X. Li, T. Gao, X. Zheng, R. Hu, J. Zheng, Y. Shen, K. Li, Y. Liu, P. Dai, Y. Zhang *et al.*, "Adaptive feature selection for no-reference image quality assessment using contrastive mitigating semantic noise sensitivity," *arXiv preprint arXiv:2312.06158*, 2023.

[29] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 5148–5157.

[30] X. Li, J. Zheng, X. Zheng, R. Hu, E. Zhang, Y. Gao, Y. Shen, K. Li, Y. Liu, P. Dai *et al.*, "Less is more: Learning reference knowledge using no-reference image quality assessment," *arXiv preprint arXiv:2312.00591*, 2023.

[31] K. Xu, L. Liao, J. Xiao, C. Chen, H. Wu, Q. Yan, and W. Lin, "Local distortion aware efficient transformer adaptation for image quality assessment," *arXiv preprint arXiv:2308.12001*, 2023.

[32] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Image quality assessment using contrastive learning," *IEEE Transactions on Image Processing*, vol. 31, pp. 4149–4161, 2022.

[33] A. Saha, S. Mishra, and A. C. Bovik, "Re-IQA: Unsupervised learning for image quality assessment in the wild," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5846–5855.

[34] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "CONVIQT: Contrastive video quality estimator," *IEEE Transactions on Image Processing*, vol. 32, pp. 5138–5152, 2023.

[35] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, 2020, pp. 1597–1607.

[36] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

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

[38] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu, "Disentangled representation learning," *arXiv preprint arXiv:2211.11695*, 2022.

[39] H. Kim and A. Mnih, "Disentangling by factorising," in *International Conference on Machine Learning*, Jul 2018, pp. 2649–2658.

[40] R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[41] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[42] X. Zhu, C. Xu, and D. Tao, "Where and what? examining interpretable disentangled representations," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 5861–5870.

[43] Y. Wei, Y. Shi, X. Liu, Z. Ji, Y. Gao, Z. Wu, and W. Zuo, "Orthogonal jacobian regularization for unsupervised disentanglement in image generation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 6721–6730.

[44] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*, 2021, pp. 12 310–12 320.

[45] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *IEEE/CVF International*

Conference on Computer Vision and Pattern Recognition (CVPR), Jul 2017.

[46] L. Liu, J. Li, L. Niu, R. Xu, and L. Zhang, "Activity image-to-video retrieval by disentangling appearance and motion," in *AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2145–2153.

[47] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "CLUB: A contrastive log-ratio upper bound of mutual information," in *International Conference on Machine Learning*, vol. 119, 13–18 Jul 2020, pp. 1779–1788.

[48] E. H. Sanchez, M. Serrurier, and M. Ortner, "Learning disentangled representations via mutual information estimation," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 205–221.

[49] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *European Conference on Computer Vision (ECCV)*, September 2018.

[50] V.-H. Tran and C.-C. Huang, "Domain adaptation meets disentangled representation learning and style transfer," in *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019, pp. 2998–3005.

[51] H. E. Gedik, A. K. Venkataramanan, and A. C. Bovik, "Joint deep image restoration and unsupervised quality assessment," *arXiv preprint arXiv:2311.16372*, 2023.

[52] L. Wang, Q. Wu, K. N. Ngan, H. Li, F. Meng, and L. Xu, "Blind tone-mapped image quality assessment and enhancement via disentangled representation learning," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 1096–1102.

[53] Z. Ye, Y. Wu, D. Liao, T. Yu, J. Yang, and J. Hu, "DRIQA-NR: No-reference image quality assessment based on disentangled representation," *Signal, Image and Video Processing*, vol. 17, no. 3, pp. 661–669, 2023.

[54] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," in *Advances in Neural Information Processing Systems*, vol. 6, 1993.

[55] H. Barrow, J. Tenenbaum, A. Hanson, and E. Riseman, "Recovering intrinsic scene characteristics," *Computer Vision Systems*, vol. 2, no. 3-26, p. 2, 1978.

[56] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. C. Bovik, and Y. Li, "MAXIM: Multi-axis MLP for image processing," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 5769–5780.

[57] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[58] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9640–9649.

[59] 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, Munich, Germany, October 5-9*, 2015, pp. 234–241.

[60] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jul 2017.

[61] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017, pp. 3883–3891.

[62] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4105–4113.

[63] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *arXiv preprint arXiv:1610.07629*, 2016.

[64] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017, pp. 3156–3164.

[65] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 1501–1510.

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

[67] A. K. Venkataramanan, C. Wu, A. C. Bovik, I. Katsavounidis, and Z. Shahid, "A hitchhiker's guide to structural similarity," *IEEE Access*, vol. 9, pp. 28 872–28 896, 2021.

[68] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2013.

[69] S. Saini, A. K. Venkataramanan, and A. C. Bovik, "The LIVE user-generated HDR video dataset," 2024.

[70] T. Borer and A. Cotton, "A display-independent high dynamic range television system," *SMPTE Motion Imaging Journal*, vol. 125, no. 4, pp. 50–56, 2016.

[71] SMPTE, "High dynamic range electro-optical transfer function of mastering reference displays," *SMPTE Standard*, vol. 2084, p. 11, 2014.

[72] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," in *ACM Annual Conference on Computer Graphics and Interactive Techniques*, 2002, p. 257–266.

[73] Q. Shan, T. DeRose, and J. Anderson, "Tone mapping high dynamic range videos using wavelets," *Pixar Technical Memo*, 2012.

[74] G. P. Nason and B. W. Silverman, *The Stationary Wavelet Transform and some Statistical Applications*. Springer New York, 1995, pp. 281–299.

[75] E. Reinhard, T. Pouli, T. Kunkel, B. Long, A. Ballestad, and G. Damberg, "Calibrated image appearance reproduction," *ACM Trans. Graph.*, vol. 31, no. 6, Nov 2012.

[76] G. Eilertsen, R. K. Mantiuk, and J. Unger, "Real-time noise-aware tone mapping," *ACM Trans. Graph.*, vol. 34, no. 6, Nov 2015.

[77] M. Oskarsson, "Temporally consistent tone mapping of images and video using optimal k-means clustering," *Journal of Mathematical Imaging and Vision*, vol. 57, no. 2, pp. 225–238, Feb 2017.

[78] A. Rana, P. Singh, G. Valenzise, F. Dufaux, N. Komodakis, and A. Smolic, "Deep tone mapping operator for high dynamic range images," *IEEE Transactions on Image Processing*, vol. 29, pp. 1285–1298, 2020.

[79] J. Yang, Z. Liu, M. Lin, S. Yanushkevich, and O. Yadid-Pecht, "Deep reformulated laplacian tone mapping," *arXiv preprint arXiv:2102.00348*, 2021.

[80] A. K. Venkataramanan and A. C. Bovik, "Subjective quality assessment of compressed tone-mapped high dynamic range videos," *Manuscript Under Preparation*, vol. 1, 2024.

[81] VideoLAN, "x264." [Online]. Available: https://code.videolan.org/videolan/x264.git

[82] (2016) An introduction to Dolby Vision. [Online]. Available: https://professional.dolby.com/siteassets/pdfs/dolby-vision-whitepaper_an-introduction-to-dolby-vision_0916.pdf

[83] H. Lin, V. Hosu, and D. Saupe, "DeepFL-IQA: Weak supervision for deep IQA feature learning," *arXiv preprint arXiv:2001.08113*, 2020.

[84] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2408–2415.

[85] E. Mavridaki and V. Mezaris, "No-reference blur assessment in natural images using Fourier transform and spatial pyramids," in *IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 566–570.

[86] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010.

[87] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*, 2014, pp. 740–755.

[88] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[89] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[90] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011 006–011 006, 2010.

[91] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Color image database TID2013: Peculiarities and preliminary results," in *European Workshop on Visual Information Processing (EUVIP)*, 2013, pp. 106–111.

[92] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *International Conference on Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1–3.

[93] 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.
[94] 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.
[95] H. Yeganeh, S. Wang, K. Zeng, M. Eisapour, and Z. Wang, "Objective quality assessment of tone-mapped videos," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 899–903.
[96] L. Krasula, K. Fliegel, and P. Le Callet, "FFTMI: Features fusion for natural tone-mapped images quality evaluation," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 2038–2047, 2020.
[97] C. S. Ravuri, R. Sureddi, S. V. R. Dendi, S. Raman, and S. S. Channappayya, "Deep no-reference tone mapped image quality assessment," in *Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 1906–1910.
[98] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.
[99] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.
[100] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020.
[101] H. Zeng, L. Zhang, and A. C. Bovik, "A probabilistic quality representation approach to deep blind image quality prediction," *arXiv preprint arXiv:1708.08190*, 2017.
[102] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 3667–3676.
[103] W. Kim, A.-D. Nguyen, S. Lee, and A. C. Bovik, "Dynamic receptive field generation for full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4219–4231, 2020.
[104] J. Shi, P. Gao, and J. Qin, "Transformer-based no-reference image quality assessment via supervised contrastive learning," *arXiv preprint arXiv:2312.06995*, 2023.
[105] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.