https://doi.org/10.2352/EI.2024.36.11.HVEI-224

Adapting Pretrained Networks for Image Quality Assessment on High Dynamic Range Displays

Andrei Chubarau[1], Hyunjin Yoo[2], Tara Akhavan[2], and James Clark[1]

[1] Department of Electrical and Computer Engineering, McGill University, Montreal, Canada
[2] Faurecia IRYStec Inc., Montreal, Canada
andrei.chubarau@mail.mcgill.ca, {hyunjin.yoo, tara.akhavan}@forvia.com, james.j.clark@mcgill.ca

Abstract

Conventional image quality metrics (IQMs), such as PSNR and SSIM, are designed for perceptually uniform gamma-encoded pixel values and cannot be directly applied to perceptually non-uniform linear high-dynamic-range (HDR) colors. Similarly, most of the available datasets consist of standard-dynamic-range (SDR) images collected in standard and possibly uncontrolled viewing conditions. Popular pre-trained neural networks are likewise intended for SDR inputs, restricting their direct application to HDR content. On the other hand, training HDR models from scratch is challenging due to limited available HDR data. In this work, we explore more effective approaches for training deep learning-based models for image quality assessment (IQA) on HDR data. We leverage networks pre-trained on SDR data (source domain) and re-target these models to HDR (target domain) with additional fine-tuning and domain adaptation. We validate our methods on the available HDR IQA datasets, demonstrating that models trained with with our combined recipe outperform previous baselines, converge much quicker, and reliably generalize to HDR inputs.

Introduction

Real-world scenes are brighter and more vivid than their digital twin reproductions. While 8-bit gamma-encoded color values drive the common standard-dynamic-range (SDR) displays, high-dynamic-range (HDR) imaging enhances the viewing experience by encoding a significantly wider range of luminance with more precision, allowing to represent a larger portion of the visible color gamut. Despite its advantages, HDR also brings complexity to the imaging pipeline. The vast majority of applications and algorithms operate on SDR content and do not yet extend to HDR.

Image quality is a critical performance metric in all visual applications, be they SDR or HDR. Classical image quality assessment (IQA) relies on hand-crafted mechanisms based on mathematical models of the human visual system (HVS). With deep learning [1], IQA has evolved toward jointly optimizing feature representations and inference directly from image data. Training deep image quality metrics (IQMs) from scratch, however, is a challenging task, because IQA datasets are limited in size [2], especially for HDR [3]. Recent methods address overfitting and handle large images by dividing inputs into smaller patches, computing and aggregating patch-wise metrics. Initially, IQMs based on convolutional neural networks (CNNs) essentially processed patches independently [4, 5]; the current state-of-the-art models take advantage of transformer architecture [6, 7, 8, 9], or combinations of CNNs and transformers [10], to capture more complex global interdependencies between patch-wise inputs.

Only a handful of traditional IQMs, and even fewer deep learning-based IQMs, are natively designed for HDR content. Most notably, HDR-VDP [11, 12], a metric that models contrast detection for a wide range of luminance conditions, has achieved wide recognition. Most SDR algorithms, on the other hand, rely on the perceptual uniformity of gamma-encoded sRGB images, making them unsuitable for accurately processing perceptually non-uniform HDR color values. While perceptually uniform (PU) encoding [13, 14] and perceptual quantizer (PQ) [15] transform linear photometric color values into perceptually uniform units, thereby enabling the application of some SDR metrics, the effectiveness of SDR methods on HDR images is limited. With the scarcity of HDR data, most deep learning methods similarly operate primarily on SDR content. Previous attempts at HDR IQA involve training networks from scratch on PU-encoded data [3, 16], overlooking the clear advantages of transfer learning [17].

In this paper, we investigate more effective strategies for training deep learning-based IQA models for HDR[1]. Instead of training on HDR data from scratch, we leverage networks pre-trained on SDR content and propose special fine-tuning strategies to re-target such networks to HDR. First, we explore several modifications to the training procedure with PU-encoded units to facilitate transfer learning. Second, we train with domain adaptation (DA) to reduce the degradation in performance associated with the domain shift from SDR to HDR. Third, while we focus on IQA for HDR, we aim to provide adequate performance for SDR and HDR data, which allows for flexible usage of the trained models in real-world applications of IQA. We validate our findings by retraining PieAPP [5] and VTAMIQ [9] to outperform previous baselines in HDR IQA on the available datasets (SDR and HDR). Our experiments emphasize the importance of transfer learning, as demonstrated by stronger generalization on both SDR and HDR data.

Related Work

Image Quality Assessment

Conventional full-reference (FR) IQA correlates image quality with the perceptual difference between a reference and a distorted image. The comparison can be based on error visibility [18, 19], structural similarity [20, 21, 22], information content [23, 24], contrast visibility [25, 11], or various other feature similarities inspired by the Human Visual System (HVS) [26, 27, 28, 29, 30] and optionally modulated by visual saliency [31, 32]. More recent work uses deep learning [1] for data-driven IQA. Instead of hand-crafted features, deep FR IQMs typically compare deep layer activations for two images [33, 34, 4, 5]. The training is done

[1] `https://github.com/ch-andrei/HDR-IQA-dom-adapt`

by optimizing mean absolute error (MAE) or mean squared error (MSE) between the predicted and the expected quality values, optionally using additional guidance by pairwise preference [5] or ranking loss [35, 34]. With limited data, the advantages of transfer learning [17, 36] motivate the use of feature extraction networks initially trained for other vision tasks (e.g., classification or segmentation [37]) and subsequently fine-tuned to IQA.

Because CNN-based methods tend to restrict input image resolution, recent work uses patch-wise processing for more flexibility: an image is split into smaller patches (randomly sampled or tiled), and patch-wise quality scores are computed and combined with averaging or weighted pooling [38, 4]. As a notable example, the PieAPP quality metric [5] predicts individual quality scores and the corresponding patch weights for patches of 64 × 64 pixels. The use of patches further allows for data augmentation: a sequence of randomly sampled patches offers a reasonably novel "view" of the same data. However, for real-world applications with HD images, this can quickly become computationally intensive as the number of patches increases to cover more pixels. Lastly, transformers instinctively conform to patch-wise IQA, because the transformer architecture [6, 39] natively uses sequences as inputs. Among recent work on transformer-based IQMs, MUSIQ [8] and VTAMIQ [9] employ multi-scale patch processing to adapt to large resolution inputs common in practical applications of IQA.

Representing Real-World Displays

Although the viewing experience varies widely according to viewing conditions and across different displays [40, 13], many computer vision algorithms operate directly on 8-bit gamma-encoded sRGB color values designed for cathode-ray tube (CRT) displays with around 100 cd/m^2 peak luminance. HDR displays, on the other hand, depict a significantly wider range of visible color with luminance levels that reach 5000 cd/m^2. To describe both SDR and HDR content on a unified scale, it is convenient to represent visual content in physical units of luminance emitted by a display as modeled by the gain-offset-gamma model [41]:

$$L = (L_{max} - L_{blk})F(V) + L_{blk}, \tag{1}$$

where L is the emitted luminance in cd/m^2, L_{max} and L_{blk} are the maximum and the black level luminance of the display in cd/m^2, V is the display-encoded luma in the range 0–1, and F is the EOTF, the inverse of the opto-electronic transfer function (OETF). For SDR, $F(V) = V^\gamma$, where γ is the gamma-correction parameter (typically, $\gamma = 2.2$), or the sRGB non-linearity. For HDR, F can be Hybrid Log Gamma [42], PQ [15], or $F(V)$ can directly encode linear scene luminance.

We can optionally extend Equation 1 to account for ambient light reflected from the display [43] by adding the ambient reflection term L_{amb}, computed as

$$L_{amb} = \frac{k}{\pi} E_{amb}, \tag{2}$$

given the display's reflectivity k (for common displays, $k < 0.01$) and the ambient illumination level E_{amb} in units of lux. The final observed luminance is then equal to $L + L_{amb}$. With this extended model, we include the effect of varying ambient conditions on the viewing experience [44, 45].

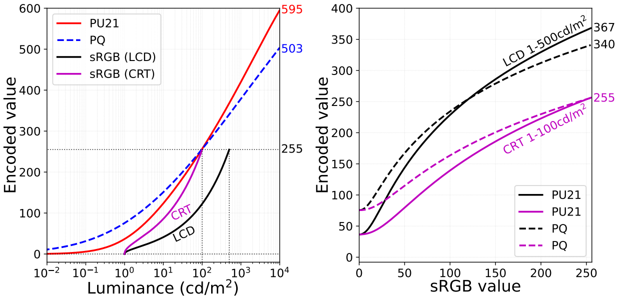

Figure 1. *Perceptually uniform encoding with PU21 [14] (banding + glare variant) and PQ [15] (scaled by 255) contrasted with the approximate mapping between luminance and the sRGB non-linearity for typical CRT and LCD displays (simulated with Equation 1). Left: the full range of encoded luminance. Right: the mapping between sRGB values and PU units.*

Extending IQA to HDR

Most existing SDR metrics are calibrated to perceptually uniform pixel values. On the other hand, trichromatic color values stored in HDR images are not perceptually uniform because they are linearly related to luminance, which humans perceive on a logarithmic scale. By extension, most SDR metrics cannot reliably predict image quality for HDR inputs. To improve the accuracy of SDR metrics on HDR data, perceptually uniform (PU) encoding [13] first transforms luminance into approximately PU values by matching contrast detection thresholds across a wide range of luminance conditions. Existing SDR quality metrics, such as PSNR and SSIM, were shown to produce significantly more accurate predictions for PU-encoded HDR data. As illustrated in Figure 1, PU21 encoding [14] transforms luminance inputs of 0.005–10000 cd/m^2 to PU units. By design, luminance levels of 0.1–100 cd/m^2 (typical SDR display luminance) map to approximately 256 steps in the PU space, ensuring that SDR metrics produce comparable results for PU-encoded SDR data.

Lastly, PQ [15] transforms luminance to a relatively PU space using similar derivations as PU21. While PU21 is concerned with the application of SDR metrics to HDR content, PQ is optimized to reduce visible quantization artifacts in HDR image formats, offering a coding scheme more aligned with human perception.

Domain Adaptation

In machine learning, when labeled data is scarce (as is the case for HDR IQA), the common solution is to use other available datasets for closely related tasks. Naturally, as such source data may differ from the desired target domain, trained models may have suboptimal performance on the target data due to the problem of domain shift. To address this limitation, various domain adaptation techniques aim to facilitate the transfer of knowledge from a source domain to a target domain, mitigating the degradation in performance caused by domain shift and improving generalization on the target data [46, 47, 48]. In this work, we focus specifically on deep feature activation CORrelation ALignment (CORAL) [49], a DA technique that aligns the statistical properties of source and target distributions. With CORAL, neural networks are trained with an additional loss term defined as the distance between the second-order statistics of the two involved data distributions:

$$\mathcal{L}_{CORAL} = \frac{1}{4d^2} \|C_S - C_T\|_F^2, \tag{3}$$

where C_S and C_T are the covariance matrices of the source and the target d-dimensional feature activations, respectively, and $\|\cdot\|_F^2$ is the Frobenius norm. Optimizing CORAL loss leads to increased statistical similarity between the source and target domains, which in turn allows trained models to learn domain-invariant but task-specific features, consistently improving generalization on the target domain.

Methodology

We investigate various modifications to the training procedure to adapt deep IQMs pre-trained on SDR data to HDR. First, to ensure out-of-the-box performance on PU-encoded SDR data for networks pre-trained with sRGB images, we verify a more intuitive normalization scheme which aligns PU-encoded SDR values to the range of sRGB inputs. Second, to further improve generalization on the target HDR domain, we fine-tune networks on a mix of SDR and HDR data with optional domain adaptation. Lastly, we ensure that the trained models perform well on both SDR and HDR data, making them more practical for real-world applications of IQA.

Training with PU-encoded Data

Similar to prior work on IQA for HDR [13, 14, 3, 45], we represent gamma-encoded SDR and linear HDR color values on a unified scale in photometric units of luminance by computing the display response with Equation 1. Following the specifications of the sRGB color space, SDR displays are modelled with $L_{max} = 100$ cd/m^2 and $L_{blk} = 0.5$ cd/m^2 (effective contrast ratio of 200:1). For HDR image formats, the stored trichromatic color values either directly encode luminance or can be tone-mapped to a given display characteristic using a similar approach. Physical luminance values are then encoded with PU21 [14] for perceptual linearity and used to train neural networks.

To mitigate precision errors and stabilize training, neural networks are trained with values rescaled to a consistent floating-point range (e.g., 0.0–1.0 or similar). PU-encoded values likewise need to be rescaled to some known range when used as input to a neural network. However, the choice of normalization scheme for PU-encoded values is not as intuitive. As illustrated in Figure 1, PU21 (*banding + glare* variant) maps luminance levels of 0.005–10000 cd/m^2 to roughly 0–595 PU units, with SDR luminance levels of 100 cd/m^2 explicitly scaled to 256 PU steps to ensure compatibility with conventional SDR metrics that operate on sRGB. To adhere to the original design of PU encoding, instead of normalizing by the full range of PU-encoded values as was done for training PU-PieAPP [3], we align 255 PU units (100 cd/m^2) with 1.0, while PU-encoded values outside this range exceed 1.0 after normalization. We refer to this normalization scheme as "*255*", because PU values are divided by 255 for normalization instead of the maximum PU encoded value P_{max}. SDR luminance levels of 0–100 cd/m^2 then map to 0–1, while HDR luminance levels of 100-10000 cd/m^2) reach roughly 2.3 after normalization.

The choice of normalization scheme (divide by P_{max} or *255*) has an effect on initial and final performance levels. The benefit of P_{max} normalization is that PU-encoded values (SDR and HDR) match the range of input values used in pre-training; the downside—PU-encoded SDR inputs (originally as 0–1 sRGB values) are arbitrarily compressed to roughly 0–0.4 ($255/595 \approx 0.4$). With *255* normalization, PU-encoded SDR luminance levels are aligned to sRGB range and networks pre-trained on sRGB data consequently produce similar predictions for PU-encoded SDR inputs. On the other hand, PU-encoded HDR luminance levels then map to a range of inputs unseen in training, emphasizing the domain expansion from SDR to HDR. Although a pre-trained network is not expected to produce reliable predictions on PU-encoded HDR data, it has guaranteed performance on SDR. We can then adapt pre-trained networks to the full range of PU-encoded data with additional fine-tuning and domain adaptation, which we hypothesize to be more effective under *255* normalization.

Domain Adaptation for HDR

We incorporate domain adaptation in the training procedure to facilitate the transfer of knowledge between the SDR and HDR domains. To this end, we optimize deep CORAL loss [49] to align the correlations of deep-layer activations between SDR and HDR data. Under such a training regime, at each training iteration, we acquire a batch of SDR data and a batch of HDR data, compute network predictions and feature representations for the inputs, calculate conventional loss functions, e.g., mean average error (MAE), between the expected and predicted values, and finally the deep CORAL loss L_{CORAL} between the SDR and HDR feature vectors. The loss terms are then summed as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{SDR} + \beta\mathcal{L}_{HDR} + \lambda\mathcal{L}_{CORAL}, \tag{4}$$

where L_{SDR} and L_{HDR} are conventional loss functions for IQA (we use MAE) on the SDR and the HDR batches, respectively, and α, β, λ are customizable weight parameters that allow for different trade-offs between the three defined losses. Since CORAL loss specifically adapts feature representations and does not require quality labels (equivalently, $\beta = 0$ or $L_{HDR} = 0$), we can leverage generic unlabeled HDR data which is more widely available than HDR IQA data. The models are then trained with DA to produce statistically similar features for SDR and HDR, but are not explicitly trained for IQA on HDR data. Lastly, the magnitude of $\mathcal{L}_{CORAL}$ varies according to the domains and the feature dimension d, hence λ must be adjusted according to the use case. As in prior work [49], we roughly match the magnitude of L_{CORAL} with other loss terms at the end of the training.

Experiments and Results

To evaluate the effect of our proposed training recipe, we retrain several deep neural network-based IQA models with and without our modifications. In what follows, we describe our training procedure and our performance evaluations, demonstrating the effectiveness of transfer learning and domain adaptation in the context of IQA on HDR images.

Datasets

We test models on the UPIQ dataset [3], which consolidates two popular SDR IQA datasets (LIVE [50] and TID2013 [51]) and two smaller HDR IQA datasets (Korshunov [52] and Narwaria [53]), with quality labels realigned to a common scale (in JOD units) through additional subjective experiments. We train our metrics on IQA data from the larger KADID-10k dataset [54] (SDR). For domain adaptation, we also use HDR images from an HDR image reconstruction dataset (not IQA), SI-HDR [55]. We summarize relevant details on the used datasets in Table 1.

Table 1. Comparison of the used datasets.

Dataset Name	No. Ref. images	No. Dist.	No. Dist. images	Resolution (h×w)	Dynamic range
LIVE [50]	29	5	779	512×768	SDR
TID2013 [51]	25	24	3,000	384×512	SDR
KADID-10k [54]	81	25	10125	384×512	SDR
Narwaria [53]	25	2	140	1080×1920	HDR
Korshunov [52]	25	3	240	1080×944	HDR
SI-HDR [55]	181	N/A	N/A	1080×1888	HDR

Experimental Setup

In previous work on IQA for HDR [3], the CNN-based PieAPP quality metric [5] (model architecture depicted in Figure 2) was trained from scratch on PU-encoded SDR and HDR data normalized to range 0–1 (P_{max}), with the HDR variant termed PU-PieAPP. We re-train PU-PieAPP using our training recipe consisting of pre-training on SDR data and fine-tuning on PU-encoded data with optional domain adaptation, following the optimization criterion defined in Equation 4. Under similar setup, we also re-train VTAMIQ [9], a transformer-based FR IQA model extended for multi-scale patch processing via scale embedding [8] (architecture in Figure 3), with the final model analogously termed PU-VTAMIQ. For both metrics, we use feature extraction networks pre-trained for classification of sRGB images. For PieAPP, we initialize the feature extractor with pre-trained weights for VGG16 [56][2]. For VTAMIQ, we use a pre-trained vision transformer.

Both PieAPP and VTAMIQ use patch inputs, which we obtain by tiling input images with randomized perturbations[3], resulting in a relatively uniform yet randomized view of the full image. For PieAPP, we sample 128 patches of size 64×64 for each image during training and 1024 patches for testing. For VTAMIQ, we use a variant of ViT extended with scale embedding [8]: patches are sampled at five different scales with initial patch sizes $p \in \{16, 32, 64, 128, 256\}$ pixels (downsampled to 16×16 when input to ViT), and with 512 patches for training and 2048 for testing.

For domain adaptation, we apply deep CORAL to the feature vectors produced by the tested models for SDR and HDR data. While PieAPP computes CNN features and the corresponding quality scores for each patch independently, VTAMIQ produces a single compact feature representation and quality score by jointly encoding all patches with the CLS token. For PieAPP, we compute CORAL loss on the features from the last Conv512 layer (y_i in Figure 2) with flattened feature dimension $d = 2048$. We concatenate the feature vectors for all patches for the reference and the distorted images (latter, if available). As an example, for a batch size of 8 image pairs with 64 patches per image, we will compute the CORAL loss between two feature matrices of size 1024×2048 ($8 \times 64 \times 2 = 1024$ and $d = 2048$). Similarly for VTAMIQ, we concatenate the CLS tokens with $d = 768$ for all reference and distorted images in a batch; for batch size and patch count above, CORAL loss is computed between 16×768 feature matrices.

Our implementation is in Pytorch [57], with all training performed on a single NVIDIA GeForce RTX 3090 GPU with 24GB of video memory. We use the AdamW optimizer [58] with the recommended parameters and an initial learning rate of 10^{-4}, exponentially decayed at the end of each epoch with a final goal of 10^{-6}. Practical training times depend on the used dataset, but we generally train for 50 epochs for each run. Since we leverage pre-training, our models converge faster than PU-PieAPP from [3] (trained for 500 epochs). Note that while the original PU-PieAPP [3] was trained on data encoded with PU08 [13], we use the updated PU21 [14], specifically its *banding + glare* variant.

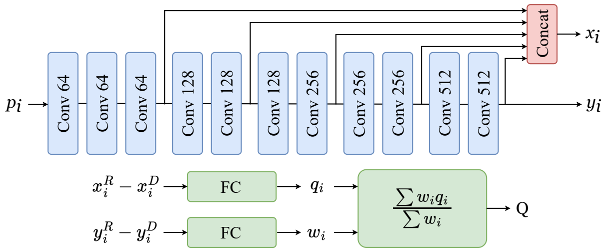

Figure 2. Diagram of PieAPP quality metric [5]. For each 64×64 patch, deep feature activations from 5 convolutional layers are computed and concatenated; a fully-connected layer predicts the quality score given the difference between the reference and the distorted patch features. Features from the last layer are used to predict patch weights. The final image quality is computed as a weighted sum of patch-wise scores. Adapted from [5].

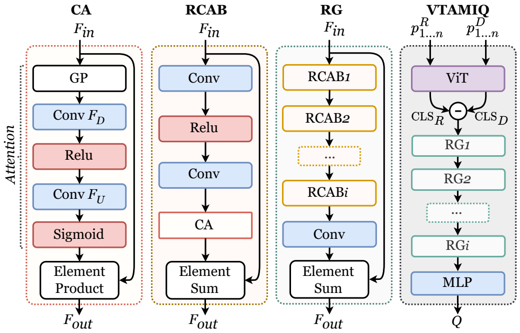

Figure 3. Diagram of VTAMIQ [9]. Patches from the reference and the distorted images are encoded by Vision Transformer (ViT) [39], the corresponding CLS token difference is computed and calibrated by a series of residual groups (RGs) based on channel attention (CA) modules. A fully-connected layer (MLP) predicts the final quality score. Adapted from [9].

Performance Evaluations

We evaluate the tested models on the UPIQ dataset with 5-fold cross-validation, splitting the available data into train, validation, and testing sets with a splitting ratio of 60-20-20 across the reference image dimension. When training on UPIQ, we take extra care to correctly isolate reference images from LIVE and TID2013 subsets of UPIQ (overlapping images are assigned to TID2013) and split the SDR and HDR portions in roughly the same ratio. Following common practice, we assess the performance of the tested models with Spearman rank order correlation coefficient (SROCC) and Pearson linear correlation coefficient (PLCC) between the expected and the predicted quality scores. As in prior work, a logistic fit is applied before computing PLCC [59].

Pre-training and fine-tuning. We test metrics pre-trained on sRGB data from KADID-10k directly on PU-encoded data from UPIQ. As presented in Table 2, without fine-tuning on PU-encoded data, PieAPP and VTAMIQ produce accurate predictions for PU-encoded SDR data (0.87–0.91 SROCC) but are only as reliable on HDR data as the PU variants of conventional IQMs (0.60–0.76 SROCC). Conversely, HDR-VDP [11] and HDR-VQM [60], offer very strong performance on the HDR subsets, but are

[2] PieAPP's CNN feature extractor architecture is similar to VGG16

[3] Sampling details in `https://github.com/ch-andrei/VTAMIQ`

Table 2. Performance (SROCC and PLCC) on subsets of UPIQ. For *Full* set, we report 5-fold cross-validation performance; for *SDR* and *HDR* subsets, we test on the entire subset. We train PieAPP and VTAMIQ on sRGB data, with PU-PieAPP and PU-VTAMIQ further fine-tuned on PU-encoded SDR data. Other metrics are as reported in [3], where PU-PieAPP* is trained on UPIQ (with test set held out). Column *Input* describes input type and normalization scheme. Best scores **bolded**, second best underlined.

Method	**Input**	**Tested subset of UPIQ**					
		Full		**SDR**		**HDR**	
FSIM	Luminance	0.82	0.89	0.54	0.51	0.45	0.34
PU-FSIM	PU08	0.84	0.90	0.77	0.77	0.71	0.66
HDR-VDP	Luminance	0.82	0.84	0.82	0.78	0.81	0.72
HDR-VQM	Luminance	0.78	0.82	0.60	0.62	**0.87**	**0.86**
Trained on KADID-10k (sRGB)							
PieAPP	PU21 (P_{max})	0.85	0.84	0.87	0.87	0.63	0.65
	PU21 (*255*)	0.86	0.85	0.88	0.89	0.60	0.63
VTAMIQ	PU21 (P_{max})	0.87	0.87	0.90	0.91	0.76	0.76
	PU21 (*255*)	0.88	0.89	0.91	0.92	0.71	0.72
Fine-tuned on KADID-10k (PU-encoded)							
PU-PieAPP	PU21 (P_{max})	0.87	0.87	0.89	0.89	0.75	0.75
	PU21 (*255*)	0.87	0.86	0.90	0.90	0.63	0.66
PU-VTAMIQ	PU21 (P_{max})	0.89	0.90	0.91	0.92	0.81	0.81
	PU21 (*255*)	0.90	0.91	**0.93**	**0.94**	0.72	0.73
Trained on UPIQ (PU-encoded)							
PU-PieAPP*	PU08 (P_{max})	**0.94**	**0.96**	0.65	0.67	0.74	0.73

suboptimal for SDR. We then fine-tune VTAMIQ and PieAPP on PU-encoded data from KADID-10k to produce their respective PU variants, PU-VTAMIQ and PU-PieAPP. We note the resulting improvement on both SDR and HDR content. Even only training on PU-encoded SDR data improves performance on the HDR subset. VTAMIQ-based models consistently outperform PieAPP, nearly matching HDR-VDP on the HDR subset (0.81 SROCC).

We contrast the performance of our models with PU-PieAPP* trained in [3] from scratch on data from UPIQ encoded with PU08 [13] and under P_{max} normalization. PU-PieAPP* has stronger cross-validation performance on the *Full* set of UPIQ, which is expected, because the model is directly trained on subsets of UPIQ, while our models are trained on KADID-10k. On the other hand, when PU-PieAPP* is trained on the HDR subset of UPIQ and evaluated on its SDR subset, it performs poorly, though its performance on the HDR subset, when trained on SDR, is only slightly lower than ours. Our models thus provide a more optimal balance of performance on both subsets, demonstrating the benefit of transfer learning, where adequate pre-training improves generalization on unseen data for both SDR and HDR inputs. We achieve further performance improvements on HDR inputs with additional fine-tuning and domain adaptation.

Normalization scheme for PU-encoded values. We find that the performance of the trained metrics increases on SDR and decreases on HDR data when *255* normalization is used instead of P_{max}. This is expected because PU-encoded SDR luminance levels appear more similar to sRGB values when normalizing by *255*, while HDR signals exceed the range of data used in pre-training. However, we confirm that P_{max} normalization results in significantly better final performance, discouraging the use of *255* normalization, despite our original motivation of leveraging the similarity between PU-encoded SDR and sRGB. Since our main objective is HDR performance, where *255* normalization is suboptimal, we emphasize our results under P_{max} normalization.

Table 3. Performance (SROCC) on the HDR subset of UPIQ for different training and domain adaptation configurations. All runs use PU21 encoding and P_{max} normalization. Training with CORAL loss indicated in column λ. For $S \rightarrow H_U$, training with no CORAL loss is equivalent to results in Table 2, because only the SDR labels are used. Unlike other DA configurations, for $S \rightarrow H_L$, we apply 5-fold cross-validation on the HDR subset of UPIQ.

Method	λ	**DA Configuration and Target**				
		$S \rightarrow H_U$			$S \rightarrow H_S$	$S \rightarrow H_L$
		SIHDR	KADID	UPIQS	KADID	UPIQH
PU-PieAPP		0.75	0.75	0.75	0.78	0.85
	✓	0.76	0.78	0.76	0.79	0.86
PU-VTAMIQ		0.81	0.81	0.81	0.88	0.89
	✓	0.82	0.85	0.84	0.89	0.91

Training with domain adaptation. We consider three domain adaptation configurations with different training procedures and source-target domains. While we use labeled SDR IQA data from KADID-10k (PU-encoded) as the source domain, the HDR-like target domain can be: (i) unlabeled authentic HDR images ($S \rightarrow H_U$), (ii) labeled synthetic HDR-like images simulated from sRGB images in SDR IQA datasets ($S \rightarrow H_S$), and (iii) labeled authentic HDR images from HDR IQA datasets ($S \rightarrow H_L$). Option $S \rightarrow H_U$ follows the common setting for DA with unlabeled target data, produces subtle improvements in generalization on the target domain, but typically does not outperform training on labeled data. With option $S \rightarrow H_S$, we provide additional optimization guidance with labeled HDR-like data simulated from the more abundant SDR IQA data, granted the distribution of luminance values does not truly come from an HDR source and the reused SDR quality labels are perhaps not as reliable for HDR. Lastly, with $S \rightarrow H_L$, although authentic HDR IQA data is used, its severely limited availability poses practical challenges due to overfitting and noisy evaluations.

For $S \rightarrow H_U$, we experiment with (i) HDR images from the SI-HDR dataset, (ii) SDR images from KADID-10k simulated as HDR, and (iii) SDR images from UPIQ simulated as HDR (UPIQS). For $S \rightarrow H_S$, we train on labeled data from KADID-10k simulated as HDR. To generate synthetic HDR-like IQA data, we reuse the labels and images from SDR IQA datasets (e.g., KADID-10k or the SDR subset of UPIQ), but simulate HDR-like display response by controlling the L_{max} parameter in Equation 1. We sample L_{max} from a normal distribution $\mathcal{N}(100,10)$ for SDR and $\mathcal{N}(5000,500)$ for HDR display response for additional data augmentation, instead of using a constant value as was done in [3]. We apply a similar method to tone-map the linear HDR color values from SI-HDR. Lastly, for $S \rightarrow H_L$, we train with cross-validation on labeled HDR data from UPIQ (UPIQH) and test on a held out set, which makes comparison with other DA configurations problematic, but nevertheless showcases the benefit of CORAL.

Our performance evaluations on the HDR subset of UPIQ for models trained with DA are presented in Table 3, which we contrast with Table 2, where the training is done without DA. We determine that training with CORAL yields subtle but noticeable performance improvements on the target HDR data for all tested DA configurations with both unlabeled and labeled target data. For $S \rightarrow H_U$, training without CORAL loss is equivalent to training on KADID-10k (see Table 2). Our best performance is achieved with option $S \rightarrow H_S$ with for KADID-10k as target, where we simulate synthetic HDR-like data. PU-VTAMIQ trained with $S \rightarrow H_S$ has 0.89 SROCC on the HDR subset of UPIQ, outperforming HDR-

VQM which has 0.87 SROCC; PU-PieAPP, generally, reacts less favorably to DA with CORAL. To isolate the effect of optimizing deep CORAL loss, we also train without the CORAL loss term, i.e., only using the labeled SDR and HDR data but without DA between them. While the main improvement comes from having labeled HDR-like training data, CORAL loss leads to additional albeit subtle gains. Finally, while we experiment with DA using P_{max} and *255* normalization schemes, we find that P_{max} offers stronger performance for all setups, unless we train directly on HDR data from UPIQ ($S \rightarrow H_L$), where the final performance is comparable. For brevity, we omit our DA results for *255* normalization.

Discussion

With transfer learning, a model is trained on one task and repurposed for a different but related task. Domain adaptation complements pre-training by facilitating transfer of knowledge from source to target domain. We train networks on sRGB data and adapt them to PU-encoded HDR data with additional fine-tuning and domain adaptation between SDR and HDR. Fine-tuning on PU-encoded data produces a considerable improvement to generalization on HDR content. DA contributes to an additional incremental gain in IQA performance on the target HDR domain. We explored DA with unlabeled and labeled target data, achieving higher performance in both scenarios. For unlabeled DA for IQA ($S \rightarrow H_U$), we note the importance of distorted images in the target dataset. Although SI-HDR contains authentic HDR images, it produce meager improvement when used as DA target, which we hypothesize is due to its relatively limited size and lack of distorted images. Conversely, SDR images from KADID-10k contain image quality distortions and more variation, resulting in improved performance when used as DA target. Granted, SI-HDR only contains 181 images, which may be insufficient for our application—we leave DA on a larger unlabeled HDR dataset to future work. While we focus on IQA, we expect our findings to also apply to other HDR tasks, where limited labeled task-specific data is available.

We note, however, that DA adds complexity to the training procedure and presents certain challenges. First, with deep CORAL, the weight of the CORAL loss must be tuned to a given use-case: despite recommendations in the original work, this potentially requires extensive empirical experimentation. Secondly, CORAL loss assumes that the second-order statistics of two data distributions can be aligned without limiting domain-specific representation, instead leading to task-specific but domain-invariant learning, but there is no guarantee that this holds for all tasks and data distributions. Moreover, while CORAL of deep feature representations is originally intended for other vision tasks, in IQA, the distortions arguably matter as much as natural image statistics. Using unlabeled and undistorted images for DA is perhaps not enough to transfer knowledge of relevant features for IQA. In that regard, it is then difficult to apply DA in its original form, hence why we find that CORAL is useful as an additional optimization criterion along regular IQA losses on both the source and target domains. Lastly, training with DA requires certain implementation details and data schedules to be changed in order to accommodate DA. This makes comparison against previous work problematic, as potential performance difference may be due to different training procedures. While we addressed this by training specifically with and without CORAL loss for similar data splits and other loss functions, there may be other details that affected the final performance.

Conclusion

We investigate more effective training strategies to adapt networks pre-trained on SDR data to HDR applications. First, we verify the effect of normalization schemes when training on data encoded with perceptually uniform transforms such as PU21. When using networks pre-trained on sRGB data, we find that the benefit of aligning the full range of PU values to the range of values used in pre-training outweighs the benefit of similarity between PU-encoded SDR values and sRGB. Moreover, networks pre-trained on sRGB data produce excellent baselines for fine-tuning on PU-encoded data. With additional training and optional domain adaptation, we further consolidate IQA for SDR and HDR, leveraging the more widely available SDR data to transfer task-specific but domain-invariant knowledge from SDR (source domain) to HDR (target domain). Our results demonstrate that our combined training recipe offers much quicker convergence and stronger generalization on both SDR and HDR data. Lastly, although we focus on IQA, our strategies likely extend to other tasks in HDR imaging.

References

[1] A. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–44, 05 2015.

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

[3] A. Mikhailiuk, M. Pérez-Ortiz, D. Yue, W. Suen, and R. Mantiuk. Consolidated dataset and metrics for high-dynamic-range image quality. *IEEE Transactions on Multimedia*, PP:1–1, 04 2021.

[4] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, 2018.

[5] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen. PieAPP: Perceptual image-error assessment through pairwise preference. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1808–1817, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NeurIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[7] M. Cheon, S.-J. Yoon, B. Kang, and J. Lee. Perceptual image quality assessment with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 433–442, 2021.

[8] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang. MUSIQ: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5128–5137, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society.

[9] A. Chubarau and J. Clark. VTAMIQ: Transformers for attention modulated image quality assessment. *CoRR*, abs/2110.01655, 2021.

[10] S. Lao, Y. Gong, S. Shi, S. Yang, T. Wu, J. Wang, W. Xia, and Y. Yang. Attentions help CNNs see better: Attention-based hybrid image quality assessment network. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1139–1148, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society.

[11] R. K. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4):40:1–40:14, July 2011.

[12] R. K. Mantiuk, D. Hammou, and P. Hanji. HDR-VDP-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content, 2023.

[13] H.-P. Seidel T. O. Aydın, R. Mantiuk. Extending quality metrics to full luminance range images. In *Human Vision and Electronic Imaging*, pages 68060B–10. Spie, 2008.

[14] R. K. Mantiuk and M. Azimi. PU21: A novel perceptually uniform encoding for adapting existing quality metrics for hdr. In *2021 Picture Coding Symposium (PCS)*, pages 1–5, 2021.

[15] S. Miller, M. Nezamabadi, and S. Daly. Perceptual signal coding for more efficient usage of bit codes. In *The 2012 Annual Technical Conference Exhibition*, pages 1–9, 2012.

[16] S. Jia, Y. Zhang, D. Agrafiotis, and D. Bull. Blind high dynamic range image quality assessment using deep learning. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 765–769, 2017.

[17] Y. Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop - Volume 27*, UTLW'11, page 17–37. JMLR.org, 2011.

[18] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4):636–650, April 2000.

[19] D. M. Chandler; S. S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, 16(9):2284–2298, Sep. 2007.

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

[21] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, Nov 2003.

[22] Z. Wang and Q. Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20(5):1185–1198, May 2011.

[23] H. R. Sheikh, A. C. Bovik, and G. de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, Dec 2005.

[24] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, Feb 2006.

[25] R. Mantiuk, S. J. Daly, K. Myszkowski, and H.-P. Seidel. Predicting visible differences in high dynamic range images: model and its calibration, 2005.

[26] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, Aug 2011.

[27] 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*, 23(2):684–695, 2014.

[28] H. Z. Nafchi, A. Shahkolaei, R. Hedjam, and M. Cheriet. Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. *IEEE Access*, 4:5579–5590, 2016.

[29] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand. A Haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication*, 61:33–43, 2018.

[30] S.C. Pei and L.-H. Chen. Image quality assessment using human visual dog model fused with random forest. *IEEE Transactions on Image Processing*, 24(11):3282–3292, 2015.

[31] L. Zhang, Y. Shen, and H. Li. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, Oct 2014.

[32] M. Kummerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge. Understanding low- and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[33] S. A. Amirshahi, M. Pedersen, and S. Yu. Image quality assessment by comparing CNN features between images. *Journal of Imaging Science and Technology*, 60:604101–6041010, 11 2016.

[34] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, June 2018.

[35] X. Liu, J. van de Weijer, and A. D. Bagdanov. RankIQA: Learning from rankings for no-reference image quality assessment. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[36] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 270–279, Cham, 2018. Springer International Publishing.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[38] J. Kim and S. Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1969–1977, 2017.

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[40] R. Wanat and R. K. Mantiuk. Simulating and compensating changes in appearance between day and night vision. *ACM Trans. Graph.*, 33(4):147:1–147:12, July 2014.

[41] R. S. Berns. Methods for characterizing CRT displays. *Displays*, 16(4):173 – 182, 1996. To Achieve WYSIWYG Colour.

[42] T. Borer and A. Cotton. A display-independent high dynamic range television system. *SMPTE Motion Imaging Journal*, 125(4):50–56, 2016.

[43] R. K. Mantiuk, K. Myszkowski, and H.-P. Seidel. *High Dynamic Range Imaging*, pages 1–42. American Cancer Society, 2015.

[44] T. Akhavan, H. Yoo, and A. Chubarau. Solving challenges and improving the performance of automotive displays. *Information Display*, 35(1):13–27, 2019.

[45] A. Chubarau, T. Akhavan, H. Yoo, R. Mantiuk, and J. Clark. Perceptual image quality assessment for various viewing conditions and display systems. *Electronic Imaging, Image Quality and System Performance XVII, pp. 67-1-67-9(9)*, 2020.

[46] H. Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[47] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 513–520, Madison, WI, USA, 2011. Omnipress.

[48] M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[49] B. Sun and K. Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *Computer Vision – ECCV 2016 Workshops*, pages 443–450, Cham, 2016. Springer International Publishing.

[50] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. LIVE image quality assessment database release 2. http://live.ece.utexas.edu/research/quality, 2005.

[51] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015.

[52] P. Korshunov, P. Hanhart, T. Richter, A. Artusi, R. Mantiuk, and T. Ebrahimi. Subjective quality assessment database of HDR images compressed with JPEG XT. In *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2015.

[53] M. Narwaria, M. P. Da Silva, P. Le Callet, and R. Pepion. Tone mapping-based high-dynamic-range image compression: study of optimization criterion and perceptual quality. *Optical Engineering*, 52(10):102008, 2013.

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

[55] Param H., Rafal M., Gabriel E., Saghi H., and Jonas U. Comparison of single image HDR reconstruction methods — the caveats of quality assessment. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery.

[56] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[57] A. et. al. Paszke. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[58] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[59] 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*, 15(11):3440–3451, Nov 2006.

[60] M. Narwaria, M. Perreira Da Silva, and P. Le Callet. HDR-VQM: An objective quality measure for high dynamic range video. *Signal Processing: Image Communication*, 35:46–60, 2015.