EG UK Computer Graphics & Visual Computing (2018)
G. Tam and F. Vidal (Editors)

A Deep Learning Approach to No-Reference Image Quality Assessment For Monte Carlo Rendered Images

J. Whittle[1] and M. W. Jones[1]

[1]Swansea University, Department of Computer Science, UK

Abstract

In Full-Reference Image Quality Assessment (FR-IQA) images are compared with ground truth images that are known to be of high visual quality. These metrics are utilized in order to rank algorithms under test on their image quality performance. Throughout the progress of Monte Carlo rendering processes we often wish to determine whether images being rendered are of sufficient visual quality, without the availability of a ground truth image. In such cases FR-IQA metrics are not applicable and we instead must utilise No-Reference Image Quality Assessment (NR-IQA) measures to make predictions about the perceived quality of unconverged images. In this work we propose a deep learning approach to NR-IQA, trained specifically on noise from Monte Carlo rendering processes, which significantly outperforms existing NR-IQA methods and can produce quality predictions consistent with FR-IQA measures that have access to ground truth images.

CCS Concepts

•***Computing methodologies*** → ***Machine learning; Neural networks; Computer graphics; Image processing;***

1. Introduction

Monte Carlo (MC) light transport simulations are capable of modelling the complex interactions of light with a wide range of physically based materials, participating media, and camera models to synthesize images that are indistinguishable from photographs. The trade-off for producing images with high visual fidelity is slow computation times as MC based rendering algorithms are only guaranteed to converge in the limit, as the number of samples approaches infinity. Early termination of the rendering process can leave an undesirable amount of visual distortions in the form of spatially varying impulse noise, and missing illumination from complex interactions such as caustic and indirect illumination due to under-sampling when few samples are considered in the image estimate.

When comparing the performance and visual quality of images produced by Monte Carlo rendering processes, a commonly accepted methodology is to render a reference image to a large number of samples to ensure that the reference has high visual quality. This reference image is then assumed to be representative of the ground truth image and can be used experimentally to compare images rendered by competing algorithms under test, either in an equal time or equal quality benchmark. To compare images to the reference image, FR-IQA methods such as MAE, MSE, SSIM [WBSS04], or more recently MS-SSIM [WSB03], HDR-VDP-2 [MKRH11], or SC-QI [BK16] can be used to evaluate the relative quality of test images. Outside of algorithm benchmarking, the ground truth image is often not available as access to the ground truth would preclude the need to perform additional rendering of the image. In such cases NR-IQA measures allow us to predict the perceived quality of images by modelling the distribution of naturally occurring distortions that are present in images before convergence of the rendering process has been achieved.

In this work we present a deep learning approach to NR-IQA, specifically aimed at evaluating distortions found in unconverged MC rendered images. We show that a convolutional neural network architecture is capable of modelling the spatially varying nature of natural image distortions that arise from under-sampling of complex light transport simulations and that predictions made by the model can approach; within a high degree of accuracy, to FR-IQA methods that have access to ground truth images. Our model is designed to estimate image quality blindly by comparing features extracted from the local region around each pixel in an unconverged input image to the distribution of those features which occur naturally in clean images from the target domain. This representation is learned by dense regression from a corpus of images corrupted with varying magnitudes of distortion. The regression target for a given noisy image is a per-pixel quality map computed using a robust FR-IQA measure [WJM17] operating on an unconverged image and its associated ground truth image.

2. Image Quality Assessment

In image processing tasks IQA aims to measure the quality or similarity of images. In FR-IQA, algorithms have access to a potentially distorted test image and a ground truth image which is known to be

correct and distortion free. The algorithm is tasked with evaluating the similarity between the distorted and true images. The meaning of "similarity" can vary between algorithms; ranging from simple definitions of geometric distance, to complex models of the Human Visual System (HVS) and perceived distortion. In contrast, NR-IQA algorithms are tasked with approximating the quality of a test image without outside input on what the true image should look like. In this formulation, image "quality" is predicted by comparing the distorted test image to a learned distribution over the types of distortion we expect to see during inference.

NR-IQA measures are often trained and validated on distorted images sampled from publicly available datasets such as: Live [SWCB14], TID2008 [PLZ*09], TID2013 [PIL*13], Kodak Lossless True Colour [Fra99], MICT [HSKS11], and IRCCyN/IVC [AB09]. These datasets contain natural images (photographs) covering a wide range of scene compositions and subject matters along with synthetically distorted copies, coupled with subjective quality scores pooled over a set of human observers from a controlled user studies.

Measures based on Natural Scene Statistics (NSS) attempt to model the distribution of perceived distortions present within a population of images from a given distortion domain [SBC05, MMB12]. Inference is performed by comparing the statistics extracted from test images to those learned from the population of training images containing the same types of distortion. These measures use hand-crafted features for modelling NSS and are fit against subjective quality scores to minimize the prediction error. Hand-crafted NSS features designed to directly correlate with perceived image quality allow for the creation of "completely blind" NR-IQA measures which do not need to be fit against subjective quality scores [KGBE16].

The synthetic distortions considered attempt to approximate those that can occur naturally during image compression, transmission, and synthesis [KB15]. Examples of the synthetic distortions used are: Additive Gaussian Noise, Gaussian Blur, JPEG-2000 Compression, JPEG Compression, Blockiness, Fast Fading, Sub-band Truncation, and Ringing Artefacts. We find that in many cases synthetic distortions such as these do not provide a representative target for training and evaluating metrics that will be applied to naturally distorted images such as those generated through MC rendering processes. Thus, in this work we propose a model trained directly on domain images containing naturally occurring distortions from unconverged MC rendering processes.

2.1. Image Quality Assessment in Monte Carlo Rendering

In MC light transport simulations each MC sample represents the energy contributed by an individual photon or the flow of energy through a region of path-space [Kaj86]. Such methods approximate an integral to create an image by averaging over many discrete samples and as such no one sample is representative of the integral's final expectation. Only in the limit, as the number of samples goes to infinity will the error of the approximation go to zero, and the average of the samples converge to the expectation of the integral.

When a finite number of samples are used the approximate image will naturally contain a certain amount of error compared to the ground truth image. It is a commonly accepted methodology when benchmarking MC rendering algorithms to produce an image using a large number of samples to ensure it will have high visual fidelity, and to use this image as a ground truth with which to compare images generated by competing algorithms in either an equal time or computation comparison using a well established FR-IQA measure. In this formulation the ground truth image can still potentially contain distortions as it is also the product of an integral approximated using a finite number of samples. Measures such as MS-SSIM [WSB03] and SC-QI [BK16] therefore make good choices for the FR-IQA method as they are robust to distortions in reference images [WJM17].

Without access to the ground truth image NR-IQA techniques can be used to halt the rendering process when an image is predicted to have reached an acceptable quality. Iterative methods have been proposed that evaluate changes between images as samples are computed progressively [Mys98, BM98]. Image change is measured by re-purposing perceptually weighted FR-IQA measures such as Visual Difference Predictor (VDP) [Dal93] and Visual Discrimination Model (VDM) [Lub95]. Treating the current image as the reference and prior images as the those under test, samples are computed progressively until the change in perceptual quality falls below a prescribed threshold. Such methods are limited in their applicability as the underlying FR-IQA are highly susceptible to distortions in reference images, and the computational expense of computing the perceptual measures often outweighs the performance gains from early halting.

Perceptually aware NR-IQA measures can also be used to compute an initial sample density map based on direct illumination and an ambient lighting approximation [RPG99]. This is used to progressively direct a computational budget during the computation of indirect illumination. As the render progresses, the sample density map is updated periodically by comparing the current rendering solution to the one from the previous iteration. Similar schemes have been applied to animated sequences along predefined camera paths, allowing for the sparse computation of high quality key-frames with Image Based Rendering (IBR) [KS00] used to synthesize intermediate frames [MRT00]. A temporally aware extension of VDP, termed Animation Quality Metric (AQM), is used to determine whether the IBR interpolated images are of sufficient visual quality for use in the animated sequence. When they are found not to be of sufficient quality, a new key-frame is recursively computed between the two existing ones and the surrounding frames are re-interpolated through IBR. This process continues until the entire sequence has been generated to a sufficient visual quality. Later methods aim to improve the performance of rendering such animated sequences by using a computationally inexpensive render pass without global illumination to approximate where viewer attention is focused throughout the sequence [YPG01]. This can be used to direct more computation into regions predicted to have high visual importance, yielding improved perceptual quality in the final global illumination pass while using a significantly reduced number of samples.

NR-IQA methods have been designed to model visual sensitivity to domain distortions from biased rendering techniques [HČA*12] such as Virtual Point Lights [Kel97] and shadow map discretisa-

tion. Our method builds on this intuition by developing an NR-IQA which models the effect of domain distortions with specific focus on unbiased MC rendering algorithms.

2.2. Machine Learning in No-Reference Image Quality Assessment

NR-IQA methods often include model parameters that need to be optimized for the image and distortion domains that will be targeted at inference time. Early methods relied heavily on the use of hand-crafted features to extract meaningful statistics about images and Support Vector Regression (SVR) to infer quality scores for image patches. More recently, General Regression Neural Networks (GRNN) were used to predict the subjective quality scores of natural image patches under synthetic distortion from the Live database [LBW11]. The robustness of this method was later improved upon using AdaBoosting [LHZ*16] and further so by the application of Convolutional Neural Networks (CNN) [BMWS16, BMM*16].

The NR-IQA problem is closely related to the problem of blind image denoising, where we wish to recover a clean image from a corrupted one. Intuitively, if we could perfectly identify noise in a distorted image, then subtracting the distortion map would result in a perfectly denoised image. Similarly, if a perfectly denoised image could be extracted from a distorted one, then the difference map would perfectly capture the distortion at each pixel. In this way we can see that there is a shared feature space relevant to both tasks.

Recent work has applied machine learning methods to the task of blind image denoising in the context of MC rendered images [KBS15]. Primary features extracted from the rendering pipeline such as pixel-colour, world-space coordinates, shading normals, texture albedo, and the direct illumination shadow-map are used to compute a set of hand-crafted secondary features based on local neighbourhood statistics. These features are fed independently for each pixel to a Multi-Layer Perceptron (MLP) that predicts a set of variances for each dimension of a cross-bilateral or non-local means filtering of the rendered image. A limitation of such methods is cost of evaluating the MLP independently for the features at each pixel. Later work improves upon this through the use of a CNN architecture by directly regressing the residual error map at each pixel [ZZC*17]. This was demonstrated on natural images distorted by additive Gaussian noise with a constant variance over the entire image, where subtracting the predicted residual from the input gives a denoised image.

CNN have also been applied to image de-raining, a denoising task where natural images contain a naturally occurring distortion, precluding the availability of "clean" reference images for comparison. Such methods can be trained in a supervised fashion through the use of natural images with rain synthetically added to them [FHD*17], or by posing the problem as an unsupervised image to image translation task using a Generative Adversarial training scheme [ZSP17].

3. Our Method

Much of the work on machine learning based NR-IQA is designed to evaluate fixed sized patches extracted from test images [LBW11, LHZ*16, BMWS16, BMM*16]. For a 512×512 image using a stride of 1 pixel, this results in 262144 image patches which need evaluation. Even when batch processing large numbers of patches simultaneously this is prohibitively slow to compute. Proposed methods usually adopt a scheme to mitigate this where a fixed number of random patches are sampled from the image, and their average predicted quality reported as the final NR-IQA score. Due to the translational invariance of convolution, many of the spatial locations between overlapping image patches recompute the same feature activations in offset spatial locations.

One solution to this problem is to use a Fully-Convolutional Neural Network (FCNN) architecture. This class of model was first proposed for the task of dense image segmentation [LSD15] where the goal is to perform multi-class classification on each individual pixel of an input image. In this work the authors took existing pre-trained models from AlexNet [KSH12], VGG [SZ14], and GoogLeNet [SLJ*14] and modified the final layers to make the networks fully-convolutional. Unlike a standard CNN which operate on fixed sized inputs using a combination of pooling, strided and dilated convolution, flattening, and dense layers to down-sample and shape the output size; the output of an FCNN is only dependent on the input size provided. During training, images of a fixed size are grouped into mini-batches to accelerate and stabilize training, while at inference time arbitrarily sized images can be fed to the network individually or in batches of identically sized images. By reusing the common feature map activations between neighbouring spatial locations, such architectures allow for regression to be performed densely at all spatial locations with a significantly reduced computational cost.

For our NR-IQA model we use an FCNN architecture to densely predict the the quality value from a robust FR-IQA measure while only being shown the distorted input image. We chose SSIM [WBSS04] as the target FR-IQA measure for regression as it is efficient to evaluate during model training and because it is known to perform robustly for the types of distortion found in unconverged MC rendered images [WJM17]. By default, SSIM uses an 11×11 pixel neighbourhood to gather local statistics on the means, variances, and covariance between the reference and test images. This yields an SSIM map with the same resolution as the input images which can be averaged to provide a robust scalar valued FR-IQA measure. Given a noisy test image of an arbitrary resolution we aim to predict the value of the SSIM map at each pixel without access to the ground truth image. The predicted SSIM map can then be mean pooled to provide a single value representing predicted image quality or used directly as an indication of the magnitude of distortion present at each spatial location.

We propose a feed-forward FCNN model with two phases. In the first phase we extract image features from the local neighbourhood around each pixel using a series of 3×3 convolutions, expanding the receptive field of the network to a final width of 11×11 pixels over the course of 5 layers, where each layer has 256 feature maps. In the second phase we apply a series of dense layers to each pixel independently with shared weights at each layer. This forms a non-linear recombination of the localized features extracted during the first phase. These dense layers are implemented using a special case of convolution with a kernel size of 1×1 and 128 output fea-

ture maps. In this configuration, the value at each spatial location in each of the output feature maps is formed by a weighted combination of over the features in the same spatial location in each of the feature maps from the preceding layer. For both phases, each convolutional layer is followed immediately by batch normalization [IS15] and ReLU activation. Lastly, we apply a single 1×1 convolutional layer without batch normalization to reduce the 128 features down to a scalar valued quality score at each spatial location. We initialize all convolutional kernels using He initialization [HZRS15] as we use ReLU activations throughout the model. Figure 1 shows our proposed architecture.

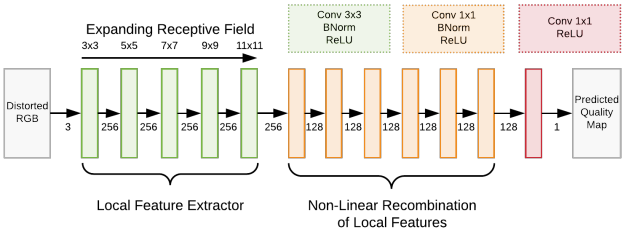

Figure 1: Our fully-convolutional NR-IQA architecture consists of a local feature extraction phase, and a non-linear recombination phase.

3.1. Training

Our model was trained using a dataset of MC rendered images produced with the Mitsuba render [Jak10]. We rendered four scenes (see figure 3) containing a range of lighting conditions, colour pallets, and scene compositions using multiple MC rendering algorithms; namely: PT [Kaj86], BDPT, [LW93, Vea97], PSSMLT [KSKAC02], MLT [VG97], ERPT [CTE05], and Manifold-MLT/ERPT [JM12]. Images were rendered to increasing numbers of independent samples using each of the rendering algorithms considered on a 2^n s.p.p. sequence where $2 \leq n \leq 18$. Images were rendered in High Dynamic Range (HDR), then converted to Low Dynamic Range (LDR) using Mitsuba's Reinhard Tone-Mapping operator [RSSF02]. The resulting LDR images were stored as 24 bits-per-pixel RGB. For each scene we treat a single high quality image as the ground truth used for FR-IQA during training. The ground truth is taken as the highest sample count image rendered with Path Tracing for the Cornell Box and Sponza Atrium scenes, and Bidirectional Path Tracing for the Veach Bidir and Veach Door scenes.

We trained using leave-one-scene-out cross validation whereby the model is trained from random He initialization using each scene in turn as the validation set while training on the remaining scenes. At each training step we sample a random mini-batch of 64×64 distorted RGB patches from the training set and compute the true SSIM map $y_i = SSIM_{map}(x_i, x_i')$ of the i^{th} distorted image patch x_i compared to its ground truth patch x_i'. The noisy patches are fed to the network and the approximate SSIM map of the sampled noisy patch $\hat{y}_i = \mathcal{Q}(x_i, \Theta)$ predicted by the network $\mathcal{Q}$ parametrized by weights Θ. We then optimize all weights in Θ via gradient descent using the Adam optimizer [KB14]. For each cross-validation fold there are ~ 70 million patches in the training set and ~ 24 million patches in the validation set.

For the loss function, we experimented with the MAE and MSE losses before settling on the Charbonnier loss with $\varepsilon = 10^{-3}$ (equation 1). The Charbonnier loss combines benefits of both MAE, as its maximum gradient is 1, and MSE as the derivative transitions smoothly from negative to positive at origin [Bar17]. These properties led to more stable regression training.

$$\mathcal{L}_{\mathcal{C}} = \sqrt{(y - \hat{y})^2 + \varepsilon^2} \tag{1}$$

In the context of an IQA metric a relevant measure of goodness-of-fit is the Pearson's Correlation Coefficient (PCC). Incorporating this into the loss function we jointly minimize the average Charbonnier loss over a batch and the negative absolute batch-wise PCC as shown in equation 2.

$$\mathcal{L}_{\text{Joint}} = \delta(1 - \left|PCC(y, \hat{y})\right|) + \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\mathcal{C}}(y_i, \hat{y}_i) \tag{2}$$

The contribution of PCC to the joint loss is scaled up by δ to balance its importance with the Charbonnier loss. Empirically, we found a value of $\delta = 1$ worked optimally when images were fed to the network in the colour range $[0, 1]$. Incorporating PCC into the loss function had the effect of increasing accuracy and stability during training.

An issue that we faced was that the network was prone to memorization from the training set, hindering its ability to generalize to patches from the hidden validation scene. To address this problem we used a data augmentation scheme employing a horizontal flip with 50% probability, 90° rotations with 25% probability each, and lastly a random perturbation in the HSV colour space. The HSV augmentation applies a common hue, saturation, and brightness shift to all pixels within a sampled patch and its associated ground truth patch (equation 3); where the shifts are sampled independently for every training example in the mini-batch. The ground truth SSIM map used as the target for regression is calculated on-the-fly after both the distorted and ground truth patches have been randomly augmented. During validation batches no augmentation is applied.

$$\begin{aligned} H, S, V &= \text{RGBtoHSV}(R, G, B) \\ H' &= (H + \xi_H) \; mod \; 1 \\ S' &= \text{clip}(S + \xi_S, 0, 1) \\ V' &= \text{clip}(V + \xi_V, 0, 1) \\ R', G', B' &= \text{HSVtoRGB}(H', S', V') \\ \text{where} \quad \xi_H &\in U(0,1) \quad \text{and} \quad \xi_S, \xi_V \in U(-0.3, 0.3) \end{aligned} \tag{3}$$

Figure 2 shows the effect of the above rotational and flip augmentation in combination with the proposed HSV perturbations. Image patches are sampled from the Veach Bidir scene of our dataset. The top row shows image patches before augmentation,

Loss	Validation Scene	Training (Rot+Flip+HSV Augmentation) PCC	SROCC	Tau	Validation PCC	SROCC	Tau
$\mathcal{L}_1$	Cornell Box	0.972193	0.973177	0.867047	0.982244	0.860295	0.700377
	Veach Bidir	0.966749	0.968851	0.865146	0.978072	0.975443	0.888591
	Veach Door	0.978891	0.969311	0.859785	0.944208	0.953709	0.835708
	Sponza	0.977722	0.982417	0.897448	0.961138	0.961118	0.838531
	$\mu \pm \sigma$	0.9739 ± 0.0056	0.9734 ± 0.0063	0.8724 ± 0.0170	0.9664 ± 0.0174	0.9376 ± 0.0523	0.8158 ± 0.0807
$\mathcal{L}_2$	Cornell Box	0.975399	0.968849	0.855512	0.988992	0.858504	0.703183
	Veach Bidir	0.974397	0.933385	0.792890	0.982448	0.958809	0.842332
	Veach Door	0.979344	0.951626	0.822898	0.937276	0.933206	0.805445
	Sponza	0.967688	0.947931	0.820128	0.956717	0.960155	0.828935
	$\mu \pm \sigma$	0.9742 ± 0.0048	0.9504 ± 0.0146	0.8229 ± 0.0256	0.9664 ± 0.0239	0.9277 ± 0.0477	0.7950 ± 0.0631
$\mathcal{L}_\mathcal{C}$	Cornell Box	0.965823	0.964170	0.846650	0.984243	0.896299	0.741704
	Veach Bidir	0.964259	0.960643	0.846469	0.977131	0.981226	0.895550
	Veach Door	0.975903	0.979250	0.887376	0.925470	0.931493	0.821929
	Sponza	0.970519	0.962865	0.848396	0.957923	0.956690	0.825556
	$\mu \pm \sigma$	0.9691 ± 0.0052	0.9667 ± 0.0085	0.8572 ± 0.0201	0.9612 ± 0.0263	0.9414 ± 0.0363	0.8212 ± 0.0629
$\mathcal{L}_{\text{Joint}}$	Cornell Box	0.974092	0.975930	0.874233	0.988965	0.954679	0.830864
	Veach Bidir	0.973101	0.961212	0.840273	0.975710	0.972310	0.862354
	Veach Door	0.979507	0.978489	0.883210	0.942559	0.950619	0.848439
	Sponza	0.975069	0.968283	0.859585	0.964857	0.968420	0.850224
	$\mu \pm \sigma$	0.9754 ± 0.0028	0.9710 ± 0.0078	0.8643 ± 0.0188	0.9680 ± 0.0196	0.9615 ± 0.0105	0.8480 ± 0.0130

Table 1: Training and Validation set accuracies for patches after 128 epochs for **leave one scene out** cross validation using rotation, flip, and HSV augmentations. The cross validation folds were computed for each of the four loss functions under test; for each of these configurations, the mean and standard deviation in the reported correlations across the folds are reported in the bottom row of each section. The $\mathcal{L}_\mathcal{C}$ loss shows only a minor improvement over the $\mathcal{L}_2$ and $\mathcal{L}_1$ losses; however, the $\mathcal{L}_{\text{Joint}}$ loss shows significantly more consistent results and has lower standard deviations over the cross validation folds.

with subsequent rows containing the patches after random augmentation has been applied.

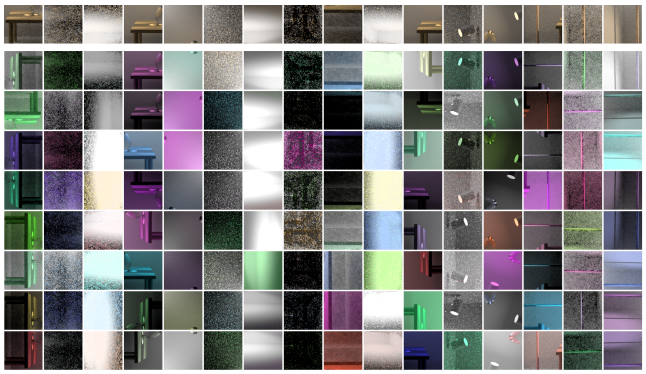

Figure 2: An example of the proposed data augmentation scheme exhibiting rotations, flips, and HSV perturbations. The top row shows image patches before augmentation, with subsequent rows containing the patches after random augmentation has been applied.

Formally, we train for 1024 epochs consisting of 256 mini-batches of 16 64×64 RGB image patches in the range $[0, 1]$ randomly sampled from the training scenes with random augmentation applied as described above. We train using the Adam optimizer [KB14] with a base learning rate of 0.001. Training was performed on a single NVidia GTX 1070 GPU and took approximately 4 hours per cross validation fold. Each epoch represents a random sampling with replacement from the training set where the data augmentation scheme ensures that, should the same patch be selected multiple times, it will be shown to the network under different random augmentations.

4. Results

We report the accuracy of our model on patches from the training and validation sets. Accuracy is calculated using the PCC, Spearman's Rank Order Correlation Coefficient (SROCC), and Kendall's Tau between the expected and predicted quality values for each patch in the training and validation sets for each cross validation fold.

To highlight the importance of our joint loss function we initially trained four models for 128 epochs each using MAE, MSE, Charbonnier, and our joint Charbonnier + PCC loss functions. Table 1 shows the result of training using each of the four loss functions. On the validation sets, the joint loss achieves the highest average accuracies of 0.9680 ± 0.0196 in PCC, 0.9615 ± 0.0105 in SROCC, and 0.8480 ± 0.0130 in Tau; and consistently have smaller standard deviations than the other losses between cross validation folds.

While all four loss functions perform similarly when comparing validation PCC values the difference between SROCC and Tau values are more pronounced. For these two measures, we see that the joint loss provides better final accuracy on average by $+0.0201$ to $+0.0338$ for SROCC and $+0.0268$ to $+0.053$ for Tau over the cross validation folds, and has a lower standard deviation between folds. This shows that the regularization properties of incorporating PCC into the loss function are effective in improving training stability and accuracy.

Selecting our joint loss as the best performing loss function we continued training the model to 1024 epochs to improve the accuracy to effective convergence (table 2). Doing so only gave a mod-

Loss	Validation Scene	Training (Rot+Flip+HSV Augmentation) PCC	SROCC	Tau	Validation PCC	SROCC	Tau
$\mathcal{L}_{\text{Joint}}$	Cornell Box	0.974825	0.979942	0.887806	0.988769	0.963563	0.858154
	Veach Bidir	0.978509	0.982749	0.894389	0.988179	0.983878	0.914428
	Veach Door	0.982456	0.986308	0.909126	0.951672	0.957310	0.864841
	Sponza	0.981203	0.986476	0.915918	0.958848	0.963735	0.842833
	$\mu \pm \sigma$	0.9792 ± 0.0034	0.9839 ± 0.0031	0.9018 ± 0.0130	0.9719 ± 0.0194	0.9671 ± 0.0116	0.8701 ± 0.0310

Table 2: Training and Validation set accuracies for patches after 1024 epochs for leave one scene out cross validation using the $\mathcal{L}_{\text{Joint}}$ loss and rotation, flip, and HSV augmentations.

Metric	Cornell Box PCC	SROCC	Tau	Veach Bidir PCC	SROCC	Tau	Veach Door PCC	SROCC	Tau	Sponza PCC	SROCC	Tau
Ours	**0.9996**	**0.9959**	**0.9688**	**0.9982**	**0.9921**	**0.9393**	**0.9928**	**0.9916**	**0.9331**	**0.9989**	**0.9964**	**0.9686**
BLIINDS [SBC10]	0.9840	0.9858	0.9287	0.9619	0.9544	0.8650	0.9640	0.9870	0.9144	0.9066	0.9668	0.8723
BIQI [MB10]	-0.2150	-0.0472	-0.0145	-0.0365	-0.3395	-0.2411	-0.7170	-0.3449	-0.2462	-0.4219	-0.5132	-0.3693
BRISQUE [MMB12]	0.3596	0.6854	0.5414	0.2211	0.3402	0.2172	0.1214	0.4154	0.3546	-0.5622	-0.3427	-0.2879
HIGRADE 1 [KGBE16]	0.2321	-0.1136	-0.0779	-0.7522	-0.6113	-0.4465	0.3415	-0.3348	-0.2901	0.2163	-0.0028	0.0246
HIGRADE 2 [KGBE16]	0.4008	0.4236	0.3167	-0.7526	-0.8591	-0.6761	0.4064	0.1528	0.1009	0.4042	0.2663	0.1867
JP2K-NR [SBC05]	-0.3977	-0.9352	-0.8264	-0.5410	-0.9139	-0.7517	-0.0234	0.2408	0.1091	-0.6635	-0.8156	-0.6524
NIQE [MSB13]	0.6013	0.9490	0.8283	0.6573	0.8651	0.7089	-0.0037	0.4418	0.3581	-0.4974	-0.2841	-0.2371
OG-IQA [LHZ*16]	0.2863	0.4868	0.3278	-0.2013	-0.1947	-0.1759	-0.5076	-0.2150	-0.1673	-0.7798	-0.6177	-0.4556

Table 3: Validation set full image NR-IQA performance for leave one scene out cross validation, compared to prior NR-IQA measures. Each correlation is computed relative to FR-IQA of each 512×512 image in the validation set compared to its ground truth image.

est improvement in accuracy on the validation sets of $+0.0039$ in PCC, $+0.0056$ in SROCC, and $+0.0221$ in Tau, showing that our model can attain high accuracy robustly with only a small amount of training.

Finally, in order to compare our proposed FR-IQA model to existing methods, we feed each 512×512 image in the validation sets to the network and apply average pooling to the resulting prediction maps to give a single scalar valued quality score for each image. Using the predicted values from our method, those from the existing methods, and the ground truth quality scores computed using the SSIM FR-IQA measure; we compute the correlation of the predicted values compared to the FR-IQA values and report these results for each validation scene, using the version of our method trained on each cross validation fold, respectively. The results of this analysis are shown in table 3. Our method consistently achieves the highest correlation across all four validation scenes, achieving average correlations of 0.9974 in PCC, 0.994 in SROCC, and 0.9525 in Tau between cross validation folds. The BRISQUE [MMB12], HIGRADE 1 & 2 [KGBE16], JP2K-NR [SBC05], NIQE [MSB13], and OG-IQA [LHZ*16] NR-IQA measures all exhibit both positive and negative correlations for different validation scenes; indicating that they are not accurately describing the underlying distribution of visible distortions present in the images. Compared to the closest competitor, BLIINDS [SBC10], this represents an increase in average correlation accuracy of $+0.0433$ in PCC, $+0.0205$ in SROCC, and $+0.0574$ in Tau between scenes.

5. Conclusions

In this paper we have presented the results of our experiments applying machine learning to NR-IQA. We have shown that convolutional neural networks acting as localized feature extractors are sufficiently able to detect and extract noise present in MC rendered images and that such features can be used for the purpose of NR-IQA.

When reviewing existing NR-IQA methods we observe that the vast majority of measures in the literature are trained on images from publicly available datasets which contain clean reference images (photographs) and their synthetically distorted test images, coupled with subjective quality scores given by and pooled over a set of human observers. We find that synthetic distortions do not present a representative target for training and evaluating metrics that will be applied to distorted images from MC rendering processes. In our experiments we show that a deep convolutional network NR-IQA model can be trained directly on images containing these naturally occurring distortions.

A limitation of prior methods is that individual patches from images must be extracted and processed separately by the network. This is not efficient and causes many duplicated convolutions between overlapping image patches, resulting the need to sparsely sample patches for performance needs. An FCNN based model can perform a dense regression of the quality score for all pixels in the source image simultaneously, allowing full images to be processed efficiently.

To improve accuracy, we incorporate PCC into the loss function to act as a regularization term, greatly improving training stability and accuracy. The model also had a tendency to over-fit on the training data by focusing on colour palettes, causing it to struggle to correctly predict the quality of images from scenes which were drastically different to those in the training sets. We solved this problem by developing a data augmentation scheme which operates in the HSV colour space. By randomly shifting the hue of input patches, and modulating their contrast and saturation, we discourage the model from using colour information as a feature and encourage it to look for structural information which remains invariant under these distortions.

We compare our proposed method's accuracy on validation sets to values computed using existing state of the art NR-IQA methods. Our model was able to accurately capture the distribution of image

distortion to a much higher degree than existing methods, achieving an average correlation of 0.9974 in PCC, 0.994 in SROCC, and 0.9525 in Tau between cross validation folds.

The code for our experiments are released open source, implemented in Tensorflow [AAB15] and Keras [C*15], and are available on Github under the MIT license at:
https://github.com/JossWhittle/MC-NR-IQA

References

[AAB15] ABADI M., AGARWAL A., BARHAM P.: TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. URL: http://tensorflow.org/. 7

[AB09] AUTRUSSEAU F., BABEL M.: Subjective quality assessment of lar coded art images, 2009. http://www.irccyn.ec-nantes.fr/ autrusse/-Databases/. 2

[Bar17] BARRON J. T.: A more general robust loss function. *CoRR abs/1701.03077* (2017). URL: http://arxiv.org/abs/1701.03077. 4

[BK16] BAE S. H., KIM M.: A novel image quality assessment with globally and locally consilient visual quality perception. *IEEE Transactions on Image Processing 25*, 5 (May 2016), 2392–2406. 1, 2

[BM98] BOLIN M. R., MEYER G. W.: A perceptually based adaptive sampling algorithm. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques* (1998), ACM, pp. 299–309. 2

[BMM*16] BOSSE S., MANIRY D., MÚLLER K., WIEGAND T., SAMEK W.: Deep neural networks for no-reference and full-reference image quality assessment. *CoRR abs/1612.01697* (2016). URL: http://arxiv.org/abs/1612.01697. 3

[BMWS16] BOSSE S., MANIRY D., WIEGAND T., SAMEK W.: A deep neural network for image quality assessment. In *2016 IEEE International Conference on Image Processing (ICIP)* (Sept 2016), pp. 3773–3777. doi:10.1109/ICIP.2016.7533065. 3

[C*15] CHOLLET F., ET AL.: Keras. https://github.com/fchollet/keras, 2015. 7

[CTE05] CLINE D., TALBOT J., EGBERT P.: Energy redistribution path tracing. *ACM Trans. Graph. 24*, 3 (jul 2005), 1186–1195. 4

[Dal93] DALY S.: Digital images and human vision. MIT Press, Cambridge, MA, USA, 1993, ch. The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity, pp. 179–206. URL: http://dl.acm.org/citation.cfm?id=197765.197783. 2

[FHD*17] FU X., HUANG J., DING X., LIAO Y., PAISLEY J.: Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Transactions on Image Processing 26*, 6 (June 2017), 2944–2956. doi:10.1109/TIP.2017.2691802. 3

[Fra99] FRANZEN R.: Kodak lossless true color image suite. *source: http://r0k. us/graphics/kodak 4* (1999). 2

[HČA*12] HERZOG R., ČADÍK M., AYDČIN T. O., KIM K. I., MYSZKOWSKI K., SEIDEL H.-P.: Norm: No-reference image quality metric for realistic image synthesis. In *Computer Graphics Forum* (2012), vol. 31, Wiley Online Library, pp. 545–554. 2

[HSKS11] HORITA Y., SHIBATA K., KAWAYOKE Y., SAZZAD Z. P.: Mict image quality evaluation database. *[Online], http://mict.eng.u-toyama.ac.jp/mictdb.html* (2011). 2

[HZRS15] HE K., ZHANG X., REN S., SUN J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR abs/1502.01852* (2015). URL: http://arxiv.org/abs/1502.01852. 4

[IS15] IOFFE S., SZEGEDY C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (2015), pp. 448–456. 4

[Jak10] JAKOB W.: Mitsuba renderer, 2010. http://www.mitsuba-renderer.org. 4

[JM12] JAKOB W., MARSCHNER S.: Manifold exploration: a markov chain monte carlo technique for rendering scenes with difficult specular transport. *ACM Transactions on Graphics (TOG) 31*, 4 (2012), 58. 4

[Kaj86] KAJIYA J. T.: The rendering equation. In *Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1986), SIGGRAPH '86, ACM, pp. 143–150. 2, 4

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *CoRR abs/1412.6980* (2014). URL: http://arxiv.org/abs/1412.6980. 4, 5

[KB15] KAMBLE V., BHURCHANDI K.: No-reference image quality assessment algorithms: A survey. *Optik - International Journal for Light and Electron Optics 126*, 11–12 (2015), 1090 – 1097. URL: http://www.sciencedirect.com/science/article/pii/S003040261500145X, doi:https://doi.org/10.1016/j.ijleo.2015.02.093. 2

[KBS15] KALANTARI N. K., BAKO S., SEN P.: A Machine Learning Approach for Filtering Monte Carlo Noise. *ACM Transactions on Graphics (TOG) (Proceedings of SIGGRAPH 2015) 34*, 4 (2015). 3

[Kel97] KELLER A.: Instant radiosity. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques* (1997), ACM Press/Addison-Wesley Publishing Co., pp. 49–56. 2

[KGBE16] KUNDU D., GHADIYARAM D., BOVIK A. C., EVANS B. L.: No-reference image quality assessment for high dynamic range images. In *2016 50th Asilomar Conference on Signals, Systems and Computers* (Nov 2016), pp. 1847–1852. doi:10.1109/ACSSC.2016.7869704. 2, 6

[KS00] KANG S. B., SHUM H.-Y.: A review of image-based rendering techniques. Institute of Electrical and Electronics Engineers, Inc. URL: https://www.microsoft.com/en-us/research/publication/a-review-of-image-based-rendering-techniques/. 2

[KSH12] KRIZHEVSKY A., SUTSKEVER I., HINTON G. E.: Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, Pereira F., Burges C. J. C., Bottou L., Weinberger K. Q., (Eds.). Curran Associates, Inc., 2012, pp. 1097–1105. 3

[KSKAC02] KELEMEN C., SZIRMAY-KALOS L., ANTAL G., CSONKA F.: A simple and robust mutation strategy for the metropolis light transport algorithm. *Computer Graphics Forum 21*, 3 (2002), 531–540. 4

[LBW11] LI C., BOVIK A. C., WU X.: Blind image quality assessment using a general regression neural network. *IEEE Transactions on Neural Networks 22*, 5 (2011), 793–799. 3

[LHZ*16] LIU L., HUA Y., ZHAO Q., HUANG H., BOVIK A. C.: Blind image quality assessment by relative gradient statistics and adaboosting neural network. *Signal Processing: Image Communication 40* (2016), 1 – 15. URL: http://www.sciencedirect.com/science/article/pii/S0923596515001708, doi:https://doi.org/10.1016/j.image.2015.10.005. 3, 6

[LSD15] LONG J., SHELHAMER E., DARRELL T.: Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 3431–3440. 3

[Lub95] LUBIN J.: A visual discrimination model for imaging system design and evaluation. In *Vision Models for Target Detection and Recognition: In Memory of Arthur Menendez*. World Scientific, 1995, pp. 245–283. 2

[LW93] LAFORTUNE E. P., WILLEMS Y. D.: Bi-directional path tracing. In *Proceedings of CompuGraphics* (1993), vol. 93, pp. 145–153. 4

[MB10] MOORTHY A. K., BOVIK A. C.: A two-step framework for constructing blind image quality indices. *IEEE Signal processing letters 17*, 5 (2010), 513–516. 6

[MKRH11] MANTIUK R., KIM K. J., REMPEL A. G., HEIDRICH W.: Hdr-vdp-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on Graphics (TOG) 30*, 4 (2011), 40. 1

[MMB12] MITTAL A., MOORTHY A. K., BOVIK A. C.: No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing 21*, 12 (Dec 2012), 4695–4708. doi:10.1109/TIP.2012.2214050. 2, 6

[MRT00] MYSZKOWSKI K., ROKITA P., TAWARA T.: Perception-based fast rendering and antialiasing of walkthrough sequences. *IEEE Transactions on Visualization and Computer Graphics 6*, 4 (2000), 360–379. 2

[MSB13] MITTAL A., SOUNDARARAJAN R., BOVIK A. C.: Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters 20*, 3 (2013), 209–212. 6

[Mys98] MYSZKOWSKI K.: The visible differences predictor: Applications to global illumination problems. In *Rendering Techniques' 98*. Springer, 1998, pp. 223–236. 2

[PIL*13] PONOMARENKO N., IEREMEIEV O., LUKIN V., EGIAZARIAN K., JIN L., ASTOLA J., VOZEL B., CHEHDI K., CARLI M., BATTISTI F., KUO C. C. J.: Color image database tid2013: Peculiarities and preliminary results. In *European Workshop on Visual Information Processing (EUVIP)* (June 2013), pp. 106–111. 2

[PLZ*09] PONOMARENKO N., LUKIN V., ZELENSKY A., EGIAZARIAN K., CARLI M., BATTISTI F.: Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics 10*, 4 (2009), 30–45. 2

[RPG99] RAMASUBRAMANIAN M., PATTANAIK S. N., GREENBERG D. P.: A perceptually based physical error metric for realistic image synthesis. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (1999), ACM Press/Addison-Wesley Publishing Co., pp. 73–82. 2

[RSSF02] REINHARD E., STARK M., SHIRLEY P., FERWERDA J.: Photographic tone reproduction for digital images. *ACM transactions on graphics (TOG) 21*, 3 (2002), 267–276. 4

[SBC05] SHEIKH H. R., BOVIK A. C., CORMACK L.: No-reference quality assessment using natural scene statistics: Jpeg2000. *IEEE Transactions on Image Processing 14*, 11 (Nov 2005), 1918–1927. 2, 6

[SBC10] SAAD M. A., BOVIK A. C., CHARRIER C.: A dct statistics-based blind image quality index. *IEEE Signal Processing Letters 17*, 6 (2010), 583–586. 6

[SLJ*14] SZEGEDY C., LIU W., JIA Y., SERMANET P., REED S., ANGUELOV D., ERHAN D., VANHOUCKE V., RABINOVICH A.: Going deeper with convolutions, corr abs/1409.4842. *arXiv preprint arXiv:1409.4842* (2014). 3

[SWCB14] SHEIKH H. R., WANG Z., CORMACK L., BOVIK A. C.: LIVE Image Quality Assessment Database Release 2, apr 2014. 2

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014). 3

[Vea97] VEACH E.: *Robust Monte Carlo methods for light transport simulation*. PhD thesis, Stanford University, 1997. 4

[VG97] VEACH E., GUIBAS L. J.: Metropolis light transport. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1997), SIGGRAPH '97, ACM Press/Addison-Wesley Publishing Co., pp. 65–76. 4

[WBSS04] WANG Z., BOVIK A. C., SHEIKH H. R., SIMONCELLI E. P.: Image quality assessment: From error visibility to structural similarity. *Trans. Img. Proc. 13*, 4 (apr 2004), 600–612. doi:10.1109/TIP.2003.819861. 1, 3

[WJM17] WHITTLE J., JONES M. W., MANTIUK R.: Analysis of reported error in monte carlo rendered images. *The Visual Computer 33*, 6 (2017), 705–713. URL: http://dx.doi.org/10.1007/s00371-017-1384-7, doi:10.1007/s00371-017-1384-7. 1, 2, 3

[WSB03] WANG Z., SIMONCELLI E. P., BOVIK A. C.: Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on* (2003), vol. 2, Ieee, pp. 1398–1402. 1, 2

[YPG01] YEE H., PATTANAIK S., GREENBERG D. P.: Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Transactions on Graphics (TOG) 20*, 1 (2001), 39–65. 2

[ZSP17] ZHANG H., SINDAGI V., PATEL V. M.: Image de-raining using a conditional generative adversarial network. *CoRR abs/1701.05957* (2017). URL: http://arxiv.org/abs/1701.05957. 3

[ZZC*17] ZHANG K., ZUO W., CHEN Y., MENG D., ZHANG L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing 26*, 7 (July 2017), 3142–3155. doi:10.1109/TIP.2017.2662206. 3

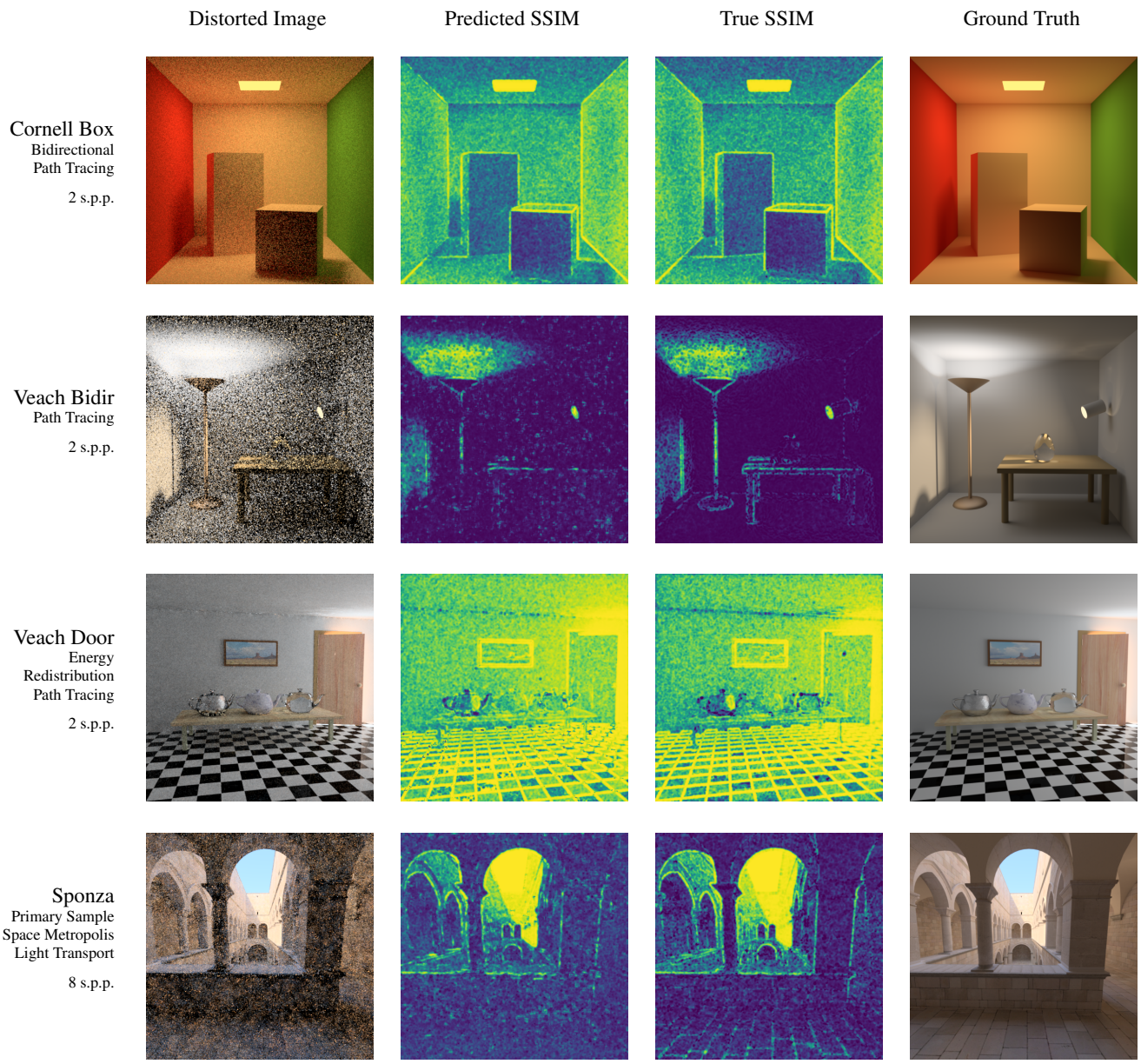

Figure 3: Example predicted quality maps from our model on images from the validation sets. From left to right: The input distorted image to evaluate. The predicted SSIM map of the noisy image. The true SSIM map of the noisy image compared to the ground truth image. The ground truth image.