Pacific Graphics 2019
C. Theobalt, J. Lee, and G. Wetzstein
(Guest Editors)

Volume 38 (2019), Number 7

Learning to Predict Image-based Rendering Artifacts with Respect to a Hidden Reference Image

Mojtaba Bemana[1] Joachim Keinert[2] Karol Myszkowski[1] Michel Bätz[2] Matthias Ziegler[2] Hans-Peter Seidel[1] Tobias Ritschel[3]

[1]MPI Informatik [2]Fraunhofer IIS [3]University College London

Abstract

Image metrics predict the perceived per-pixel difference between a reference image and its degraded (e. g., re-rendered) version. In several important applications, the reference image is not available and image metrics cannot be applied. We devise a neural network architecture and training procedure that allows predicting the MSE, SSIM or VGG16 image difference from the distorted image alone while the reference is not observed. This is enabled by two insights: The first is to inject sufficiently many un-distorted natural image patches, which can be found in arbitrary amounts and are known to have no perceivable difference to themselves. This avoids false positives. The second is to balance the learning, where it is carefully made sure that all image errors are equally likely, avoiding false negatives. Surprisingly, we observe, that the resulting no-reference metric, subjectively, can even perform better than the reference-based one, as it had to become robust against mis-alignments. We evaluate the effectiveness of our approach in an image-based rendering context, both quantitatively and qualitatively. Finally, we demonstrate two applications which reduce light field capture time and provide guidance for interactive depth adjustment.

1. Introduction

Computer vision or graphics experts easily recognize image artifacts that might be highly domain-specific. An image-based rendering (IBR) specialist will quickly notice where depth estimation failed, where transparency was not handled or where a highlight did not move correctly. Similarly, in computer graphics, artifacts resulting from Monte Carlo noise in image synthesis when producing a feature film, or shadow bias [Wil78] in a computer game are easily spotted by domain experts. The assessment typically is not limited to detection, but importantly includes judging magnitude as well as spatial locality. The importance of interacting with errors can be seen from photographs with spatially annotated over- and under-expose artifacts, as done for instance by Henri Cartier-Bresson [Col12]. Remarkably, all this is not achieved by comparing an image to a reference, but by experience and intuition built from knowing what natural images look like and how images with artifacts differ. Can we enable a machine to also perform such a task?

More formally, we face the challenge illustrated in Fig. 1. Given an image $\mathcal{A}$ that is a distorted version of a reference $\mathcal{B}$ we wish to predict their difference $\mathcal{A} \ominus \mathcal{B}$ without access to $\mathcal{B}$. The lower right image shows the ground truth metric response $\mathcal{A} \ominus \mathcal{B}$. This metric could simply be the mean square error (MSE as used in Fig. 1), a more perceptual metric like SSIM [WBSS04] or even VGG-16 activation differences that are effective as an image metric [SZ14, ZIE*18]. More particularly, we go beyond the typical mean opinion scores [TM18] given to uniform distortions such as noise or JPEG compression, and seek to produce localized distortion visibility maps without accessing the reference.

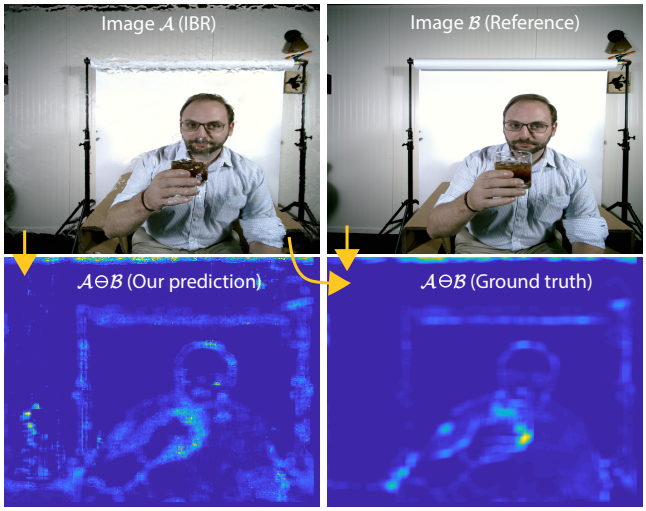

Figure 1: *Given an image* $\mathcal{A}$ (top left) *that is a version of a reference* $\mathcal{B}$ (top right) *distorted by IBR artifacts, we predict their per-pixel difference map* $\mathcal{A} \ominus \mathcal{B}$ (lower left) *without observing* $\mathcal{B}$*. The lower right shows the ground truth difference* $\mathcal{A} \ominus \mathcal{B}$*. We here show MSE, but other metrics such as SSIM or VGG16 are also possible.*

In this paper, we choose to study one specific form of artifacts

that arise in image-based rendering (IBR) [MB95, GGSC96], in particular, when employed for novel-view synthesis from sparse light fields (LFs) [LH96]. It is important in virtual reality and movie production where LFs are used to provide head motion parallax and special effects. Moreover, having a localized error prediction is also important for quality control. In IBR, artifacts are very localized (e. g., around certain depth edges) and creating opinion scoring or even spatio-angular annotated dataset of LF artifacts in a size sufficient for machine learning appears to be a daunting task. Our method proceeds without all of this.

Addressing this challenge, we make use of convolutional neural networks. We will show, how learning this mapping right away will result in many false positives or false negatives. Instead, two important ingredients come together in our approach. First, as the number of images containing artifacts is typically limited, we need to augment the training data with natural images that are free from artifacts. Second, we propose a way to find the right balance between natural and distorted training data.

Not requiring a reference is useful whenever the original is inaccessible (lost, impossible to compute, unavailable, undefined). Furthermore, we demonstrate one application of a non-reference metric in light field capturing. We first capture a sparse light field, followed of by an interpolation of the intermediate views. If our our metric indicate those intermediate views have errors, they views will be recaptured. This allows acquiring higher-quality light field in much shorter time compared to dense LF capturing.

2. Previous Work

In this section, we discuss objective image quality metrics, with special emphasis on those that do not require the undistorted reference image. Then, we briefly characterize IBR-specific artifacts, as well as metrics specialized in their detection, which is the key focus of this work.

Image metrics Some application and functions may require *quality* while others need *visibility* metrics [Cha13].

Image quality metrics (IQMs) evaluate the distortion magnitude and are typically trained on the mean-opinion score (MOS) data [SSB06, PLZ*09] that labels the entire image with as a single quality score. The most commonly used IQMs such as PSNR, SSIM, MS-SSIM [WB06], FSIM [ZZMZ11], and CIELAB [ZW97] are *full-reference* (FR) metrics that take as input the reference and distorted images, and compute local differences that are pooled into a global, single quality score. Recently, it has been demonstrated that CNN-based FR-IQMs achieved best performance in predicting MOS data [APY16, BMM*18]. Zhang et al. [ZIE*18] employed crowdsourcing and created a large scale patch-based dataset in two perceptual experiments: (1) two-alternative forced choice (2AFC) on distortion strength, and (2) “same/not same" near-threshold distortion visibility. They train different network architectures and report in each case a much better performance than traditional FR-IQMs in predicting their data from both experiments.

Visibility metrics (VMs) predict the distortion perceptibility for every pixel in the form of visibility maps. VMs are specifically tuned for detecting near-threshold distortions, which is required in many graphics and vision applications that cannot tolerate any perceivable quality reduction and require local information on the distortion positions. To decide on the visibility of such near-threshold distortions, models of human vision are often employed, where the most prominent FR-VMs examples include: VDM [Lub95], VDP [Dal92], and HDR-VDP-2 [MKRH11]. In the specific task of predicting selected rendering and compression artifacts, best performance has been achieved using machine learning [ČHM*13] and CNN-based techniques [WGY*18, PL18].

No-reference metrics In this work, we focus on the VMs due to the locality of their prediction, but we are specifically interested in more challenging *no-reference* setup, where the reference image is not available. We discuss the most successful and recent NR-IQMs that rely on machine learning techniques, and we also refer the interested reader to more comprehensive metric surveys in [Cha13, KZG*17]. Early machine learning techniques employed predefined features such as SIFT and HOG [NL10, MB10, SBC12, TJK11], and measured their distortions with respect to natural image statistics [WB06]. Recently, CNN architectures are applied to such feature learning as well as the MOS regression at the same time [BCNS16, KYLD14, BMM*18, TM18]. To compensate for a low number of MOS-labeled images, such solutions typically rely on patches, where they assign the same MOS score for all patches that belong to a given image [KZG*17]. Such practice is justified for specific classes of distortions that affect the whole image uniformly, which might be the case for certain types of image noise or compression artifacts, but might confuse the network in case of localized distortions such as those occurring in IBR.

To compensate for the lack of true local reference images, Bosse et al. [BMM*18] learn the importance of local patches, but their key motivation is not in deriving the localized VM, but rather in estimating relative patch weights in the aggregated MOS rating. Lin and Wang [LW18] employ a quality-aware generative network to hallucinate the reference image, which by employing adversarial learning is further refined by an IQM-discriminator that is trained on ground truth references. Their hallucination-guided quality regression network is fed with the difference between the hallucinated and distorted images, as well as the distorted image itself to predict the MOS value. The quality-aware generative network, hallucination-guided quality regression network, and the IQM-discriminator are jointly optimized in an end-to-end manner. Kim and Lee [KL17] apply state-of-the-art FR-IQMs such as SSIM to generate proxy scores on patches as the ground truth to pre-train the model and then fine-tune their target NR-IQM. At intermediate stages the regression network considers mean values and the standard deviations of per-patch 100-element feature vectors which are then pooled to a per-image quality score.

In this work, we also employ state-of-the-art FR-IQMs to perform an initial per-patch distortion annotation, and strike the required balance between different error magnitudes in the training data, which is essential for meaningful training and shift-invariant properties of our NR-VM.

The research on NR-VMs is extremely sparse, presumably due to limited access to locally labeled images [HČA*12, ČHM*13, WGY*18]. A notable exception is the work of Herzog et al. [HČA*12] who employs support vector machine

(SVM) to predict per-pixel distortions for selected rendering artifacts (they do not consider IBR) and achieve performance comparable to FR-VMs. Here, we demonstrate that time-consuming manual per-pixel distortion labeling is not strictly required.

In cases where training data is both easy to produce–such as uniform distortions like noise, JPEG, etc.–and no perceptual calibration is required, supervised training has been employed to detect aliasing artifacts [PL18]. Our work differs, as we only have very limited training data available, both because only very few ground truth images are available for IBR and we need perceptual calibration. Learning from little data is part of our balancing contribution.

Vogels and colleagues [VRM*18] have proposed a method to denoise path traced images. To steer the amount of denoising, they also trained a neural network to predict distortion in terms of MC variance, which is as unknown as the pixel value to be MC-estimated itself. Interestingly, in both their work and ours, a NR metric is used to steer adaptation: for them it is a denoising algorithm; for us, one application is controlling capture hardware. Their task is different as they predict SSIM error from a pair of images, where one is noisy and the other is denoised. This restricts the distortions to the difference between denoised and reference, which are smaller than IBR artifacts and also does not need to be perceptually calibrated. The fact that images with MC noise can be generated in arbitrary amounts also underlines what is the focus of our work: coping with limited training data.

Image-based rendering for structured or unstructured light fields (LFs) of real-world scenes involves a number of computational steps such as: depth reconstruction, neighboring view-image warping, warped view-image blending, and disocclusion hole in-painting. Each of these steps is prone to inaccuracies that manifest themselves as IBR-specific artifacts such as object shifting (incorrect depth), crumbling, distorted edges (depth discontinuities, e. g., due to compression), popping (fluctuations in depth), ghosting (depth inaccuracy, view blending), stretching, blurry or black regions (in-painting) [TZMD18]. Specialized IBR quality metrics often rely on leaving one view out as the reference [WBF*17, CRM12, SAB11, BPC*11] or searching for matching image blocks after their registration [BBC*15, GJQ*17], and then employing customized FR-IQMs. NR-IQMs typically focus on detecting selected distortion types such as blurring and ghosting [BLL*10], ghosting and popping [GSGC16], blurring, stretching and black holes [TZMD18], and aggregation into one final scalar score. Perceptual experiments have been performed to understand how the observers rate the severity of different artifacts as a function of rendering parameters such as the number of blended views and viewing angles [VCL*11]. A skillful pre-processing of depth (e. g., depth blurring in uncertain regions) and choice of particular algorithmic solutions can substantially suppress artifacts [HRDB16, SKC*19], eventually using a neural network trained to predict blending weights to combine the warped images[HPP*18]. More objectionable distortion types can be traded-off with those that are more visually appealing (e. g., blurry depth that is more consistent but further from the ground truth). Instead of focusing on selected distortion types, Ling et al [LLC18] proposes to learn a dictionary based on manually labeled data. The features extracted from an image allows to predict a MOS value using support vector machine regression. As data labeling can be time consuming, as Ling et al. [LLWC19] create artificial training data that aims to simulate occlusion problems. A Generative Adversarial Network (GAN) discriminator [GPAM*14], targeted to identify in-painted image regions, is used to predict a quality score.

All the discussed work on IBR quality evaluation essentially focuses on providing a single score per-image, which then also serves as a metric for performance evaluation. While some FR-IQMs generate viable per-pixel VMs at intermediate stages [CRM12, SAB11], their accuracy is not formally evaluated. The same holds for the NR-IQM [LLC18]. Our work hence differs from all previous work by pursing the NR-VM setup to detect local IBR distortions using CNN-based techniques.

3. Learning a No-reference Metric

Overview Test-time input to our method is a single distorted RGB image $\mathcal{A}$. While our distortions are always IBR artifacts resulting from a specific depth reconstruction and specific IBR method, the interna of how this image is generated (e. g., the depth map) are transparent, and we only need access to the result. Withheld is the reference RGB image $\mathcal{B}$. In the case of IBR, such a distorted-undistorted pair is typically produced by rendering a known image from other known views.

Output of our proposed method is a single-channel (scalar) image that predicts a given difference metric response $\mathcal{A} \ominus \mathcal{B}$, where the $\ominus$ operator depends on the choice of the specific metric, e. g., MSE, SSIM [WBSS04], or VGG16 [SZ14]. High values are produced where the images are different and small values where they are similar. This output is accurate, if it has little false positives or negatives. False positives correspond to predicting a perceived difference where there are no artifacts and false negatives correspond to visible artifacts the metric fails to report.

Note that two forms of approximations are made here: the first is the error that the metric itself makes when comparing two images relative to human judgment. The second is the error that our method has, with respect to a prediction. Ultimately, our method is a prediction of a prediction, but surprisingly can perform better than one prediction alone.

3.1. Training data

Our training data comprises existing metric responses $\mathcal{A} \ominus \mathcal{B}$ to the distorted image $\mathcal{A}$ and the clean reference image $\mathcal{B}$. Strictly speaking, learning does not even observe the reference image $\mathcal{B}$, but in practice, it is required to compute the metric response $\mathcal{A} \ominus \mathcal{B}$.

For creating our training dataset, we used captured LF images of 42 different scenes, which come from the Stanford LF repository [sta], the Fraunhofer IIS light field dataset [DZD*16], Google Research work [PZ17], and Technicolor [SBV*17] as well as from our own captured images. All 4D LF datasets comprise conventional 2D images in a resolution up to 2k×2k, taken from a range of sparse view points, such as in a 3×3 camera array with known camera positions. For each LF view point, we first estimate the depth using a light field depth estimation technique [DZD*16] and then warp [MMB97] the image into all other views. For each LF, we use the

four corner views to generate novel-view images at the positions of the remaining views. Each warped view corresponds to one original view, and we compute the response of a full-reference metric to this pair. With approx. 9 views per LF and 42 LFs in total, this amounts to only 210 unique images, i. e., a comparatively low number for a training task.

We use six scenes for testing and the rest for training. The same split is also applied later for the user study. Our test scenes are totally different from the training scenes, which is important as the number of scenes in the training set is small and generalization across them is an additional challenge.

The natural images used in our training and test dataset are sourced from the Inria Holidays image dataset [JDS08] which have a comparable resolution to our LF images.

Our method is independent of the actual underlying metric $\ominus$ we predict. We will denote this response neutrally as $\mathcal{A} \ominus \mathcal{B}$. We explored three metrics: MSE, SSIM and VGG16. MSE is defined as the average per-pixel RGB difference vector length squared. The SSIM metric is using the original implementation [WBSS04]. VGG16 [ZIE*18] transforms both $\mathcal{A}$ and $\mathcal{B}$ into the VGG16 feature space and picks the activations at layer five, which is 512-dimensional. The L_2 difference of these two vectors is used as the metric response. For each metric, we normalize the 95th percentile of their responses across the training dataset to fall between 0 and 1.

3.2. Architecture

We use a simple encode P [RPB15] that has learnable parameters Θ and predicts the error map $P(\mathcal{A}|\Theta)$ by observing $\mathcal{A}$ (Fig. 2).

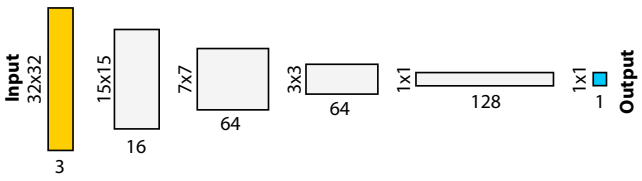

Figure 2: *Our architecture consumes* 32×32 *patches, (yellow left), and applies a cascade of* 3×3 *convolutions, followed by non-linearities (ReLU). Spatial resolution is reduced* (height) *and feature count increases* (width) *before a final prediction of the metric response is produced* (blue, right).

The network comprises 5 layers (32×32 patch size) with the total number of $|\Theta| = 175,537$ learnable parameters and is trained on all patches of the training set in a sliding window fashion.

The loss is the L_1 error of the predicted metric response, so $||P(\mathcal{A}|\Theta) - (\mathcal{A} \ominus \mathcal{B})||_1$. Note that the loss is always L_1, while the metric can be the L-norm-like MSE as well as SSIM or VGG16.

Balancing We have explained why, and will see from the ablation study, that it is important to have natural patches, but the question is how many. If we take an unlimited number, the metric prediction simply always returns zero, because natural patches have no error to themselves.

Our solution is to start with a half-half mix of distorted and clean patches. Regrettably, many of the distorted patches, which make

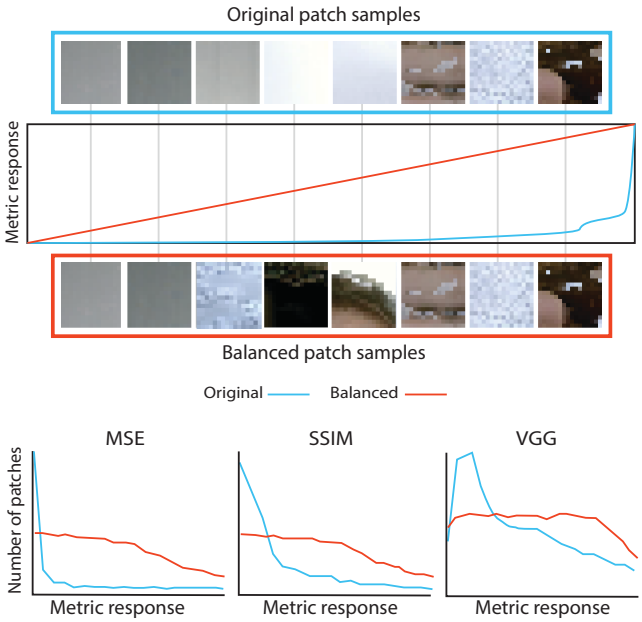

Figure 3: *When sampling uniformly from IBR patches the error distribution is skewed towards low errors* (blue). *Our balancing* (red) *adjusts the samples to have a uniform range of errors. The three lower plots show the actual distribution before and after balancing for different metric responses.*

50 % of the total, also have small errors that are close to zero. These patches are exactly those for which IBR was successful, i. e., did not have any artifacts. Depending on the metric, this imbalance can be very strong, and in particular for MSE, it is extremely heavy-tailed (Fig. 3). To address this, we balance the error distribution for the distorted half when creating the training data as follows: First, we sort all patches by their metric response into a priority queue. Then, we uniformly random-sample the range from zero to the 95th percentile of the metric response distribution. For every sample i with value ξ_i, we find the patch j with the most similar metric response d_i and remove it from the queue and add it to the training dataset. When the minimum difference $\xi_i - d_j$ is larger than a threshold ε, we reject the sample. This is repeated until a target patch count, such as 250 k, is reached.

4. Evaluation

4.1. Methods

Training Strategies We compare three different strategies for training. The first is ours, the other two are ablations. FULL is our complete method involving 50 % natural patches and a balancing of the other 50 % as described in Sec. 3.2. NOBALANCE is realized by a similar 50/50-split, but we train on all distorted patches without the balancing. NONATURAL adapts the balancing to take 100 % of the patches coming from IBR without adding the natural patches as described in Sec. 3.2. All training sets, albeit processed differently, have the same size of ca. .5 M patches.

Error As we predict metric responses, our error is the same as

the loss, the absolute difference between the ground truth metric response and our prediction of that response. As these errors also come in arbitrarily different scales for different metrics, we normalize them per metric by dividing by the global 95th percentile of the GT metric response across the balanced training dataset.

We additionally report errors in metric prediction errors for a split subsets to understand the false/true-positive and false/true-negative tendency. In ALL, we compute the error for the whole test dataset. Additionally, we consider two subsets of the test dataset. The first subset is CLEAN, which includes only natural patches. The second one is DISTORTED that contains only IBR patches, including those that might also come out with very low or even with no error. Please note that this is a partitioning of the test set, and not of the training set.

4.2. Quantitative results

In this section, we discuss both the means and full error distributions of all training strategies for different partitions and different metrics.

Table 1: *Error of the metric predictions on the test data for different variants of our algorithms and different partitions (*ALL/CLEAN/DISTORTED*) of the training data* (columns) *on different metrics* (rows)*. Winners per-partition are marked bold.*

Metric	FULL			NONATURAL			NOBALANCE		
	ALL	CLE.	DIST.	ALL	CLE.	DIST.	ALL	CLE.	DIST.
MSE	**.098**	.006	.189	.137	.092	**.182**	.102	**.003**	.201
SSIM	**.078**	.013	.143	.143	.159	**.127**	.080	**.012**	.149
VGG	**.085**	**.006**	.165	.207	.293	**.121**	.092	.008	.176

Means The means of all methods are compared in Tbl. 1. We see that our method (FULL) has the smallest error across different metrics compared to both other variants (bold in column ALL).

In detail, when we look into the partitioning, we find that for the DISTORTED partition, the NONATURAL strategy performs best. This is expected as training is done with all distorted patches which comprise the maximal variety of distortion. This makes the resulting metric sensitive for all kinds of distortions. As a result, the probability of false negatives, i. e., claiming patches with an error to be fine, becomes low.

We also find, that for the CLEAN partition, the NOBALANCE strategy performs best. This also is expected as in the training, 50 % of data comprises natural (undistorted) patches, and due to the NOBALANCE strategy, small errors dominate in the distorted patches. This makes the resulting metric particularly sensitive for near-threshold distortions. In this case, the probability of false positives, i. e., reporting a high metric response for no-error patches, is low.

All statements are true (significant, $p < .01$, t-test after testing for Gaussianity) across all metrics, indicating that the FULL approach is independent of the underlying metric. A positive exception is VGG, where the FULL approach even performs better than NOBALANCE on the CLEAN partition.

Distributions In Fig. 4, we show the distribution of errors for different metric predictions (top) and the correlation of the prediction error and metric response (bottom). In each plot, colors encode the variants of our approach (NONATURAL, NOBALANCE, FULL).

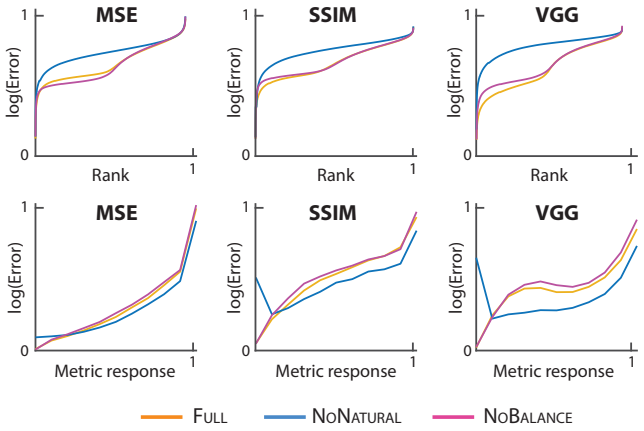

Figure 4: *Analysis of metric prediction error, for different metrics and variants of our method. The top plots show sorted error distributions. The bottom row plots show the correlation of metric response and metric prediction error. All vertical axis are log scale.*

Each plot in the first row of Fig. 4 shows the sorted error of our metric prediction in ascending order. We see that across the entire range, with the exception of MSE prediction for low errors; the FULL approach performs better than other variants. This indicates that the mean is a good characterization of the performance. In all cases, we noticed a sudden increase in the error that occurs around 50 % of the population, i. e., the error for the first half of the population seems to follow a different trend than the second half. We hypothesize that, these are the patches where reference and input are (partially) not aligned, which make up roughly 50 % of the population as well. Unfortunately, there is no way to tell apart a misaligned patch that is judged by FR metrics as different with respect to a displaced reference. Hence, large errors are expected to become undetectable at some error level. The exception is the regime in MSE where the FULL approach is worse on low errors and slightly better on high errors, while it performs best on average in (Tbl. 1). This can be difficult to comprehend due to the log scale of the vertical axis.

Each plot in the second row in Fig. 4 shows the error of our prediction on the vertical axis and the metric response on the horizontal axis as a connected scatter plot. We can see that the plots are in accordance with Tbl. 1: The NONATURAL method which performs best in predicting high metric responses, has a high error on patches with small metric response (false positives). Symmetrically, the NOBALANCE method which is the best at predicting low metric responses, produces high errors on patches with high metric response (false negatives). FULL method is always a bit worse than one other method in one region (except at the unique point where both cross), but on average performs best overall.

4.3. Qualitative results

Example metric outputs Fig. 5 shows an analysis of the response

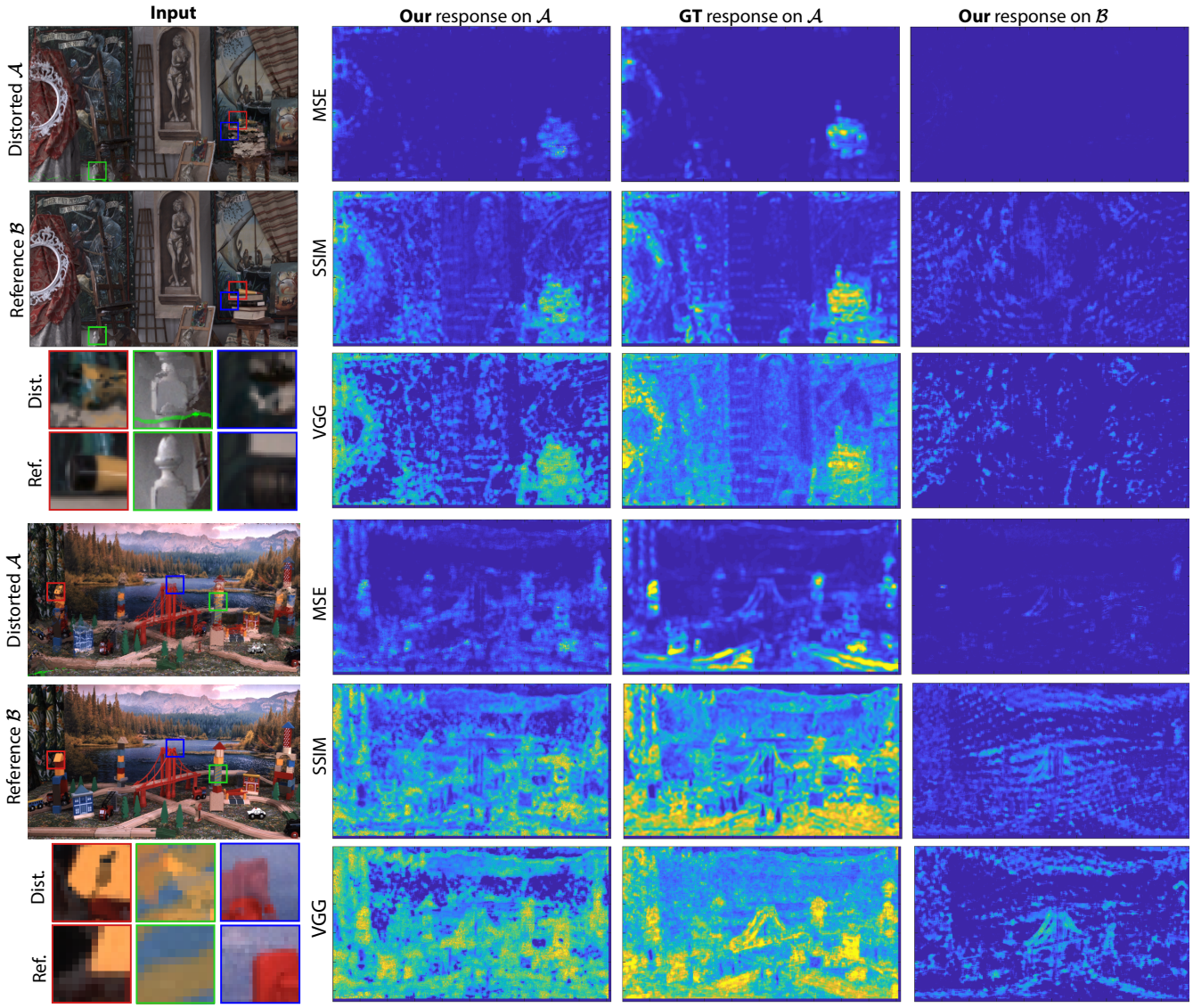

Figure 5: *Comparing the response to a pair of an image $\mathcal{A}$ and its distorted version $\mathcal{B}$* (first column). *Our response* (second column) *is similar to the ground truth* (third column). *When executed on the clean reference* (fourth column), *only very few false positives are reported.*

of all metrics to two different LFs from the test set. The first column shows the distorted input $\mathcal{A}$ in the top, below the hidden reference $\mathcal{B}$ and below this three insets from both. The second column shows our predicted response $\mathcal{A} \ominus \mathcal{B}$ for different metrics: MSE on top, followed by SSIM and VGG. A false color coding, where cold colors indicate a low response and warm colors indicate a high response, is used. The third column shows the GT response for the same. It is evident that there is a similarity between our prediction and the ground truth. We slightly err towards conservative, i. e., miss a few errors. How some of these errors are only false findings, i. e., a limitation of the metrics, becomes apparent from the user study to follow.

The last column shows a sanity check where we put the hidden reference image $\mathcal{B}$ into our metric. The hidden reference obviously does not contain any error, and consequently reporting one is a false positive. We see, that our image has a responses in areas that are correct but look like IBR artifacts, but in most areas has no response. In summary, this indicates that we localize and scale errors to a hidden reference in images with artifacts, while avoiding to produce a signal when facing clean images. It might appear that MSE has less false positives than SSIM or VGG when inspecting the last column; simply more deep blue, very close to perfect in the first row. However, such a trend is not supported by the numbers in Tbl. 1 or the plots in Fig. 4. The true reason for this impression might be that the SSIM and VGG response simply have a larger receptive field per-se: MSE is per-pixel while VGG is affected by up to 32×32 pixels. Even the ground truth response is more dense (less deep blue). Consequently the metric prediction, in case of error, also makes spatially more extended, more dense, mistakes.

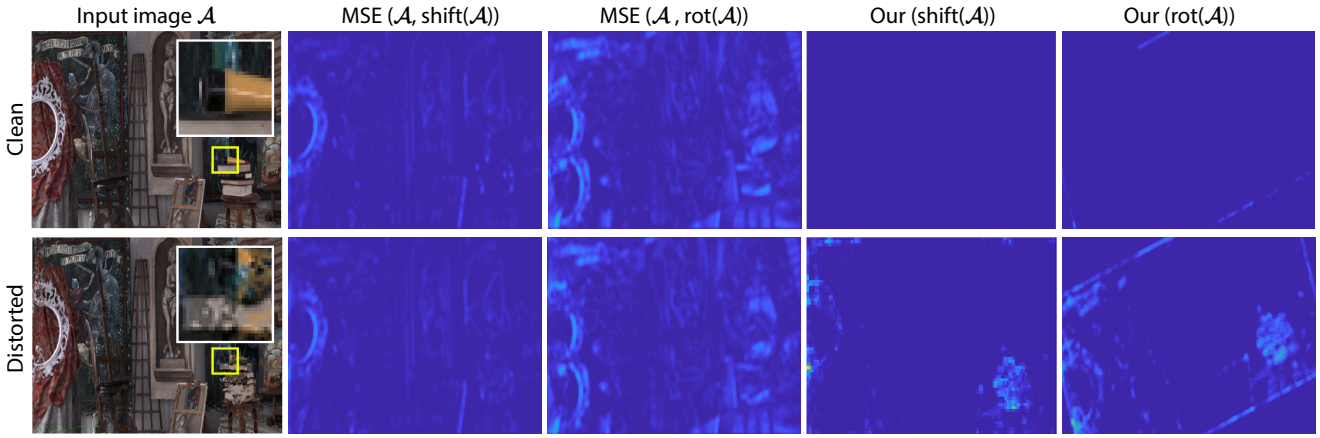

Figure 6: *Transform-invariance of our approach: When computing the distance between a clean input image $\mathcal{A}$* (first column, first row) *and a misaligned reference $\mathcal{B}$ (not shown here, 20-px shifted or 20 degrees rotated copy of $\mathcal{A}$), a common metric such as MSE will show a strong response* (first row, second and third columns). *Such a response is numerically correct, but far from human assessment, which would be more similar to our response* (first row, fourth and fifth columns). *Symmetrically, repeating the experiment on a distorted input, our approach correctly localizes the distortions around the books* (inset) *as if the reference had been aligned.*

Transformation-invariance Surprisingly, results produced by our approach can turn out to be better than their own supervision, as our method is forced to come up with strategies to detect problems without seeing the reference. This makes it immune to a common issue of many image metrics: misalignment [KRMS16]. Even a simple shift in image content will result in many false positives for classic metrics (Fig. 6). An image that has merely been shifted is reported to be very different from a reference by all the metrics used for our supervision; however, it shows less differences in case we add IBR artifacts to it. In contrast, our method does not care about transformation, but when IBR artifacts are added, they are detected. As our proposed method is oblivious to the ground truth, it is not subject to such a misconception. While not quantifiable, the result is arguably more similar to human judgment, as indicated by the user experiment in the next subsection.

4.4. User study

We have conducted a user experiment to validate that our predicted metric responses spatially correlate with the visibility of artifacts to human subjects. We quantify the human responses by means of per-pixel annotations, which are painted on top of images showing IBR artifacts. Note that no user responses was used for training.

Methods Naïve users were asked to use a binary painting interface to mark errors in a rendered image for each of the six LFs of our test dataset in an open-ended session that took 15 minutes on average. We average the binary response into a continuous fraction (percentage) of users that detected the location of the artifacts.

Analysis Asking $N = 10$ users, we find the correlation (Pearson linear correlation R, higher values are better; statements highly significant as the correlation is computed on a high number of image pixels) reported in Fig. 7-b. We see that for many scenes as well as for the average across scenes, our method has a higher correlation with user annotation than the metric it was supervised on. We hypothesize, that this is due to the fact that our network had learned to become independent of a reference, a similar robustness that the HVS employs. There is no clear trend on which of our metric response predictions correlates the most with the user annotations. The differences between scenes, however, seem more pronounced.

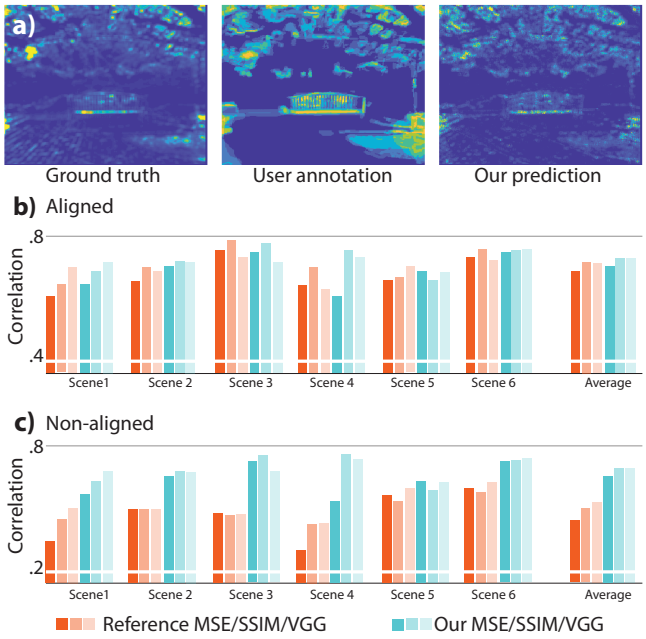

Figure 7: *Exemplary user study result* (a). *Correlation (significant, $p < .001$) of MSE/SSIM/VGG and user responses (red) compared to our predictions of the three metrics (blue) for different scenes and as an average across scenes to the right* (b). *We can see that in the non-aligned conditions, these differences get stronger* (c).

When repeating the experiment with a non-aligned reference (shifted a mere 20 px to the right), we find the correlations reported

in Fig. 7-c. We see that our correlation even improves in this condition(our metric shows higher correlations for all metrics across different scenes), showing we are more robust to alignment issues when predicting user responses.

Perceptualization Finally, we computed a linear correlation R by fitting a model $x_i = a \cdot y_i + b$, where x_i is the user response and y_i is our prediction of the metric response for pixel i. This allows a "perceptualization" of our metrics response. Fitting multiple models a, b in a leave-one-out protocol to 5 of our 6 scenes produces an average error of $.05/.04/.02$ for MSE/SSIM/VGG respectively, indicating that this perceptualization generalizes to some extent.

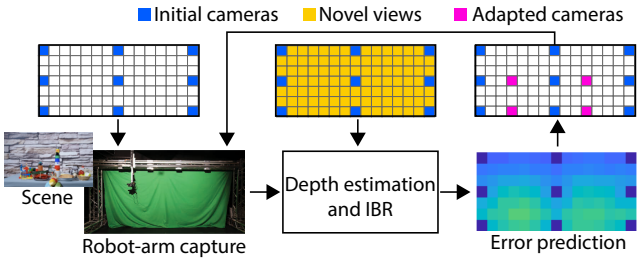

Figure 8: *Proposed pipeline for adaptive LF sampling by bounding the reconstruction error predicted by our no-reference metric.*

4.5. Other architectures

We also explored using other architectures with or without balancing. A simple solution would be to use a supervised image translation network such as Pix2Pix [IZZE17] to map from entire IBR images to the metric response. Unfortunately, training these on our data converges to a flat response of zero, as artifacts are too rare and subtle to be picked without the balancing we suggest. Future work could investigate combining our balancing with other architectures.

4.6. Supplementary materials

Ground-truth responses of all metrics and our predictions for all input images, for all variants of the algorithm, as well as all user study annotations can be explored in an interactive web application in the supplementary materials.

5. Applications

We will now demonstrate two practical applications of a NR-IQM in light field production. The first is accelerating automated adaptive LF capture (Sec. 5.1), the second employs our NR-IQM as a feedback in an interactive depth manipulation system (Sec. 5.2).

5.1. Adaptive light field capturing

Capturing a dense set of input view images results in a high-quality reconstruction but remains a time-consuming process or may require a bulky setup. Our main observation is that not all input view images contribute equally to the reconstruction of novel-view images. Our metric helps identifying and capturing these.

Images from views dominated by planar diffuse surfaces can reliably be predicted from images taken from other views showing this very same surface. Hence, dense capturing from these views is needed and thus not efficient.

In contrast, occlusions and specularity can be more challenging, because it must be ensured that each scene element is visible in at least two camera views (when using multi-view stereo, as we do) to compute depth. Sparse capturing from these views would sacrifice the reconstruction quality.

To both of these ends, we propose an adaptive capturing mechanism as it illustrated in Fig. 8 to capture an image for a view only if it cannot be extrapolated from other views.

5.1.1. Setup

We study adaptive capturing by means of a large-scale translation stage equipped with a digital camera. The position of the camera can be controlled with a precision of $80\mu m$ in horizontal and $50\mu m$ in vertical direction. This allows for very dense capturing of the scene. While this takes long to capture, it serves as a unique baseline to our study where we can compare our prediction of an error to the actual error present.

5.1.2. Procedure

We first capture a sparse set of images and estimate the depth maps for each view. Then, we use DIBR to render a set of intermediate-views and compute the reconstruction error for each rendered view. All pixels are simply averaged in each view image, producing a single scalar value. The capturing grid is then subdivided into smaller regions where average predicted reconstruction errors is larger than a given threshold. This process is repeated until a desired quality is achieved. By this approach, the number of captured views can be substantially reduced, and we only need to capture images at locations where reconstruction is poor.

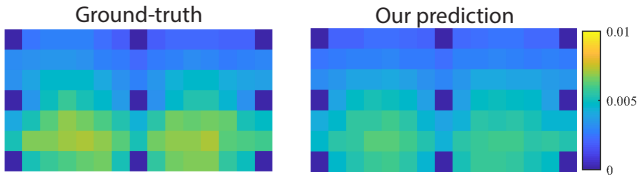

Figure 9: *Reconstruction error of intermediate novel views.* Left*: Ground truth MSE values,* right*: Our network MSE prediction.*

Predicting the reconstruction error of novel view is the key to make such an approach work. Classic full-reference image quality metrics require a dense capture to provide reference images to compute the error, which is not practical as our goal is to reduce the number of captured images in the first place. In contrast, our proposed no-reference metric can measure the error in the novel view images without providing their reference images, resulting in an efficient approach.

5.1.3. Evaluation

To evaluate effectiveness of our metric in this application, we simulate capturing two LFs, adapted according to the MSE metric.

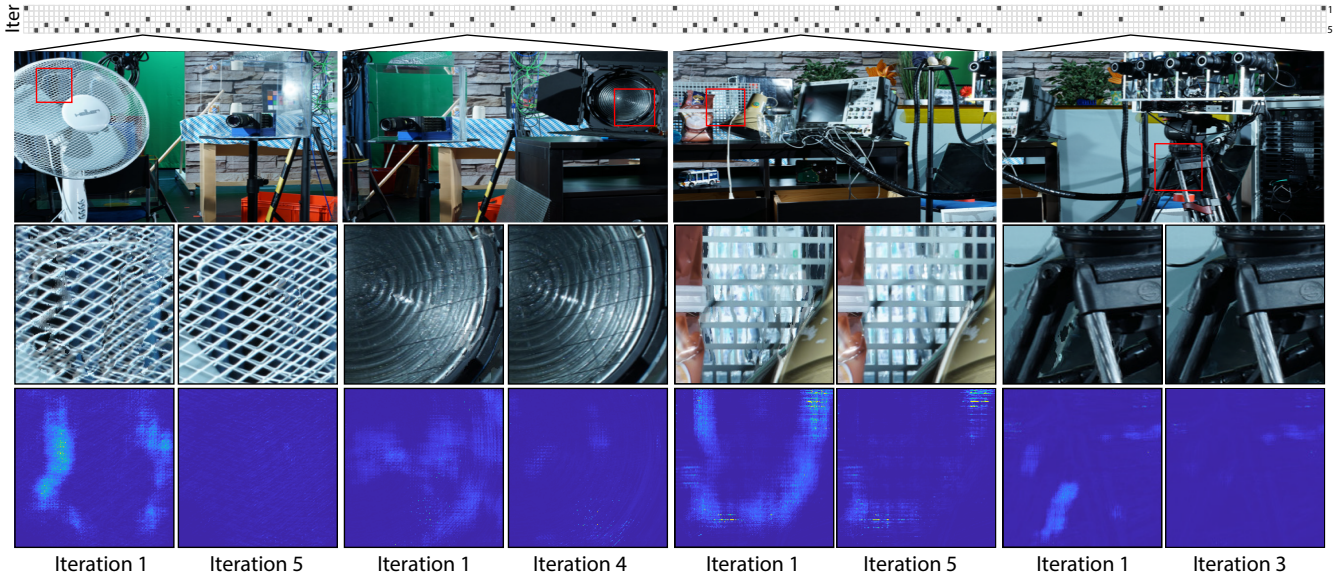

Figure 10: *Adaptive panoramic light field capturing: The top row shows a grid indicating the camera placement at different iterations. The second row shows the selected rendered views based on the key frames that are captured. The insets in the third row show the marked patches from the rendered views in the first iteration and in the iteration that a desired quality is achieved. In the fourth row, we also show our network predictions for the corresponding patches in each iteration.*

Array We captured an array of 7×15 images for the scene shown in Fig. 8 (left). In Fig. 9 we show the ground truth MSE (left) and our network prediction (right), where each grid element denotes a camera position. The dark blue grid elements indicate the camera positions where actual key frames were captured, while rendering has been performed for all remaining intermediate positions.

As we can see, the distribution of reconstruction error as predicted by our metric correlates well with the ground truth. Fig. 8 (right) shows new camera locations that are required to reduce the true average reconstruction error below .004.

Panoramic We also demonstrate the potential benefit of our approach for an efficient panoramic (i. e., one-dimensional, linear) light field capturing. As it is shown in Fig. 10, depending on the scene content, not all regions in the scene require equally dense camera placement. Our metric successfully guides the capturing setup to take more photos in the regions with thin structures, substantial disocclusions or specularites where accurate reconstruction is highly challenging. Overall, capturing 76 instead of 720 images – a sparsity of 10.5 % – reduces the total capture time from 59 minutes to 4.9 minutes, i. e., by 91 %.

5.2. Interactive depth adjustments

Long acquisition times involved in capturing dense light fields make it a tedious and impractical task for some application fields. One of such fields is movie production, where the presence of highly dynamic scenes and time pressure discourages the use of dense light fields, and in such cases, only sparse light field capture using video camera arrays is seen as a convenient solution.

Unfortunately, automatic error-free light field reconstruction from a sparse capture is still an unsolved problem. To this end, there are ongoing research efforts to address the challenges such as the estimation of disparity in the presence of homogeneous areas, repetitive structures, fine-grained objects, or specularities. In such cases, interactive disparity estimation improvement seems to be the most promising solution to achieve a high-quality view rendering [WFY*11, KK15, LVHDH12, CLD11]. However, this requires detecting possible view rendering artifacts as fast as possible to reduce the post-processing time. As shown in the right-most image of the second row in Fig. 10, spotting an artifact is not a trivial task and sometimes requires carefully scanning the view rendering result. Our quality estimation metric can significantly simplify this process by allowing the automatic analysis of several novel rendered views. By observing the predicted visibility map, which identifies the local distortions, the user can quickly spot the problematic regions. Using a post-production software suite † to perform an interactive view rendering with only a small subset of cameras allows detecting the captured view responsible for the error. The inspection of the corresponding disparity map followed by an approach similar to [WFY*11, KK15] finally allows fixing the view rendering error. This is achieved by manual creation of a geometry proxy in 3D space for objects whose disparity map could not be computed automatically. The proxy is then used to bound the admissible depth values for a subsequent disparity estimation.

(top) and the bottom row shows the corresponding patches after applying our manual disparity refinement.

The results of this procedure are illustrated in Fig. 11. The contained repetitive structures are very challenging for automatic dis-

† https://www.iis.fraunhofer.de/realception

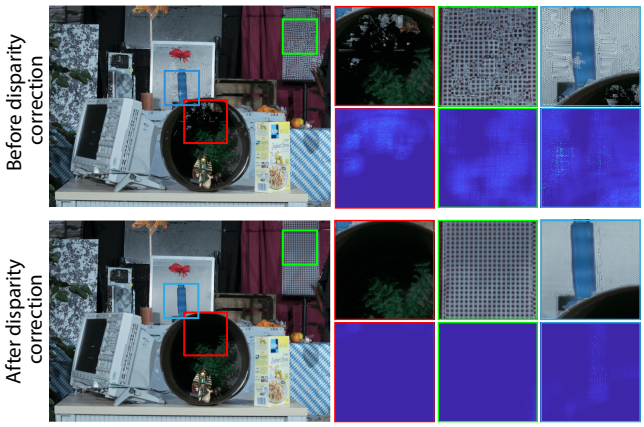

Figure 11: *Interactive depth adjustment. The marked patches are showing the regions in the rendered view where our method predicts the MSE*

parity estimation and consequently lead to many view rendering artifacts as clearly indicated by the depicted error map. For solving these issues, a user has added proxy-based disparity constraints for the waste basket (and the contained figurine), the grid structure behind the flower, and the grid structure in the upper right corner of the image. By these means, a much better view rendering could be achieved as shown in Fig. 11. Our metric has reduced the time required to find those reconstruction errors, leaving more time to a user to correct them.

6. Conclusion

We have demonstrated that with properly adjusted training data (prioritization and natural supervision), a CNN can learn how to predict the difference of an image to a hidden reference. Our approach is independent of the metric used and we have shown MSE, SSIM and VGG prediction. Other metrics such as HDR-VDP-2 [MKRH11] or the CNN-based metric of Wolski et al. [WGY*18] would likely be predictable in a similar fashion.

Such a metric can be applied for several applications. As demonstrated this includes adaptive light field sampling of complex scenes and interactive depth editing. Moreover, since in contrast to any existing non-reference metric, our approach provides a predicted error map, this opens the potential for many novel applications such as interactive or automatic view rendering error correction.

In future work, we would like to overcome the limitations of the paired input, eventually using an adversarial [GPAM*14] design, and learn the prediction only from pairs and without the metric, or only from pairs of undistorted-metric or distorted-metric.

Acknowledgements This work was partly supported by the Fraunhofer-Max Planck cooperation program within the framework of the German pact for research and innovation (PFI) and a Google AR/VR Research Award.

References

[APY16] AMIRSHAHI S. A., PEDERSEN M., YU S. X.: Image quality assessment by comparing CNN features between images. *J Imag. Sci. and Technology 60*, 6 (2016), 60410–1. 2

[BBC*15] BATTISTI F., BOSC E., CARLI M., LE CALLET P., PERUGIA S.: Objective image quality assessment of 3D synthesized views. *Image Commun. 30*, C (2015), 78–88. 3

[BCNS16] BIANCO S., CELONA L., NAPOLETANO P., SCHETTINI R.: On the use of deep learning for blind image quality assessment. *arXiv:1602.05531* (2016). 2

[BLL*10] BERGER K., LIPSKI C., LINZ C., SELLENT A., MAGNOR M.: A ghosting artifact detector for interpolated image quality assessment. In *IEEE Int. Symp. on Consumer Electronics* (2010), pp. 1–6. 3

[BMM*18] BOSSE S., MANIRY D., MÜLLER K. R., WIEGAND T., SAMEK W.: Deep neural networks for no-reference and full-reference image quality assessment. *IEEE TIP 27*, 1 (2018), 206–219. 2

[BPC*11] BOSC E., PEPION R., CALLET P. L., KOPPEL M., NDJIKI-NYA P., PRESSIGOUT M., MORIN L.: Towards a new quality metric for 3-d synthesized view assessment. *IEEE J of Selected Topics in Signal Processing 5*, 7 (2011), 1332–1343. 3

[Cha13] CHANDLER D. M.: Seven challenges in image quality assessment: Past, present, and future research. *ISRN Signal Proc.* (2013). 2

[ČHM*13] ČADÍK M., HERZOG R., MANTIUK R., MANTIUK R., MYSZKOWSKI K., SEIDEL H.-P.: Learning to predict localized distortions in rendered images. In *Comp. Graph. Forum* (2013), vol. 32, pp. 401–10. 2

[CLD11] CAO X., LI Z., DAI Q.: Semi-automatic 2D-to-3D conversion using disparity propagation. *IEEE Trans. Broad. 57*, 2 (2011), 491–9. 9

[Col12] COLEMAN S.: www.theliteratelens.com: Magnum and the dying art of darkroom printing, 2012. 1

[CRM12] CONZE P.-H., ROBERT P., MORIN L.: Objective view synthesis quality assessment. In *Proc. SPIE* (2012). 3

[Dal92] DALY S. J.: Visible differences predictor: an algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III* (1992), vol. 1666, pp. 2–16. 2

[DZD*16] DABAŁA Ł., ZIEGLER M., DIDYK P., ZILLY F., KEINERT J., MYSZKOWSKI K., SEIDEL H.-P., ROKITA P., RITSCHEL T.: Efficient Multi-image Correspondences for On-line Light Field Video Processing. *Comp. Graph. Forum (Proc. Pacific Graphics)* (2016). 3

[GGSC96] GORTLER S. J., GRZESZCZUK R., SZELISKI R., COHEN M. F.: The lumigraph. In *SIGGRAPH* (1996), pp. 43–54. 2

[GJQ*17] GU K., JAKHETIYA V., QIAO J.-F., LI X., LIN W., THALMANN D.: Model-based referenceless quality metric of 3D synthesized images using local image description. *IEEE TIP 27*, 1 (2017), 394–405. 3

[GPAM*14] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial nets. In *NIPS* (2014), pp. 2672–80. 3, 10

[GSGC16] GUTHE S., SCHARDT P., GOESELE M., CUNNINGHAM D.: Ghosting and popping detection for image-based rendering. In *Proc. 3DTV* (2016), pp. 1–4. 3

[HČA*12] HERZOG R., ČADÍK M., AYDIN T. O., KIM K. I., MYSZKOWSKI K., SEIDEL H.-P.: NoRM: No-reference image quality metric for realistic image synthesis. *Comp. Graph. Forum 31*, 2 (2012), 545–54. 2

[HPP*18] HEDMAN P., PHILIP J., PRICE T., FRAHM J.-M., DRETTAKIS G., BROSTOW G. J.: Deep blending for free-viewpoint image-based rendering. *ACM Trans. Graph. (Proc. SIGGRAPH) 37*, 6 (2018). 3

[HRDB16] HEDMAN P., RITSCHEL T., DRETTAKIS G., BROSTOW G.: Scalable inside-out image-based rendering. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 35*, 6 (2016). 3

[IZZE17] ISOLA P., ZHU J., ZHOU T., EFROS A. A.: Image-to-image translation with conditional adversarial networks. *CVPR* (2017). 8

[JDS08] JEGOU H., DOUZE M., SCHMID C.: Hamming embedding and weak geometric consistency for large scale image search. In *ECCV* (2008), pp. 304–317. 4

[KK15] KAP-KEE K.: Apparatus and method for correcting disparity map, 2015. US Patent 9,208,541. 9

[KL17] KIM J., LEE S.: Fully deep blind image quality predictor. *IEEE J Sel. Topics in Signal Processing 11*, 1 (2017), 206–220. 2

[KRMS16] KELLNHOFER P., RITSCHEL T., MYSZKOWSKI K., SEIDEL H.-P.: Transformation-aware perceptual image metric. *J Electronic Imaging 25*, 5 (2016), 053014. 7

[KYLD14] KANG L., YE P., LI Y., DOERMANN D.: Convolutional neural networks for no-reference image quality assessment. In *CVPR* (2014), pp. 1733–40. 2

[KZG*17] KIM J., ZENG H., GHADIYARAM D., LEE S., ZHANG L., BOVIK A. C.: Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal Processing Magazine 34*, 6 (2017), 130–141. 2

[LH96] LEVOY M., HANRAHAN P.: Light field rendering. In *SIGGRAPH* (1996), pp. 31–42. 2

[LLC18] LING S., LE CALLET P.: How to learn the effect of non-uniform distortion on perceived visual quality? Case study using convolutional sparse coding for quality assessment of synthesized views. In *ICIP* (2018), pp. 286–290. 3

[LLWC19] LING S., LI J., WANG J., CALLET P. L.: GANs-NQM: A generative adversarial networks based no reference quality assessment metric for RGB-D synthesized views. *arXiv:1903.12088* (2019). 3

[Lub95] LUBIN J.: *Vision Models for Target Detection and Recognition*. World Scientific, 1995, ch. A Visual Discrimination Model for Imaging System Design and Evaluation, pp. 245–283. 2

[LVHDH12] LIN C., VAREKAMP C., HINNEN K., DE HAAN G.: Interactive disparity map post-processing. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission* (2012), pp. 448–455. 9

[LW18] LIN K. Y., WANG G.: Hallucinated-IQA: No-reference image quality assessment via adversarial learning. *CVPR* (2018). 2

[MB95] MCMILLAN L., BISHOP G.: Plenoptic modeling: An image-based rendering system. In *SIGGRAPH* (1995), pp. 39–46. 2

[MB10] MOORTHY A., BOVIK A.: A two-step framework for constructing blind image quality indices. *IEEE Signal Proc. Letters 17*, 5 (2010), 513–16. 2

[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 Trans. Graph. (Proc. SIGGRAPH)* (2011). 2, 10

[MMB97] MARK W. R., MCMILLAN L., BISHOP G.: Post-rendering 3D warping. In *Proc. i3D* (1997). 3

[NL10] NARWARIA M., LIN W.: Objective image quality assessment based on support vector regression. *IEEE Trans. Neural Networks 21*, 3 (2010), 515–9. 2

[PL18] PATNEY A., LEFOHN A.: Detecting aliasing artifacts in image sequences using deep neural networks. In *Proc. HPG* (2018). 2, 3

[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* (2009), 30–45. 2

[PZ17] PENNER E., ZHANG L.: Soft 3D reconstruction for view synthesis. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 36*, 6 (2017). 3

[RPB15] RONNEBERGER O., P.FISCHER, BROX T.: U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI* (2015), pp. 234–241. 4

[SAB11] SOLH M., ALREGIB G., BAUZA J. M.: 3VQM: a vision-based quality measure for dibr-based 3D videos. In *2011 IEEE Int. Conf. on Multimedia and Expo* (2011), pp. 1–6. 3

[SBC12] SAAD M., BOVIK A., CHARRIER C.: Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE TIP 21*, 8 (2012), 3339–3352. 2

[SBV*17] SABATER N., BOISSON G., VANDAME B., KERBIRIOU P., BABON F., HOG M., LANGLOIS T., GENDROT R., BURELLER O., SCHUBERT A., ALLIE V.: Dataset and pipeline for multi-view light-field video. In *CVPR Workshops* (2017). 3

[SKC*19] SERRANO A., KIM I., CHEN Z., DIVERDI S., GUTIERREZ D., HERTZMANN A., MASIA B.: Motion parallax for 360 RGBD video. *IEEE Trans. Vis. & Comp. Graph. (Proc. IEEE VR) 25* (2019). 3

[SSB06] SHEIKH H., SABIR M., BOVIK A.: A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE TIP 15*, 11 (2006), 3440–3451. 2

[sta] http://lightfield.stanford.edu/lfs.html. 3

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556* (2014). 1, 3

[TJK11] TANG H., JOSHI N., KAPOOR A.: Learning a blind measure of perceptual image quality. *CVPR* (2011), 305–12. 2

[TM18] TALEBI H., MILANFAR P.: Nima: Neural image assessment. *IEEE TIP 27*, 8 (2018), 3998–4011. 1, 2

[TZMD18] TIAN S., ZHANG L., MORIN L., DĂL'FORGES O.: NIQSV+: A no-reference synthesized view quality assessment metric. *IEEE TIP 27*, 4 (2018), 1652–64. 3

[VCL*11] VANGORP P., CHAURASIA G., LAFFONT P.-Y., FLEMING R. W., DRETTAKIS G.: Perception of visual artifacts in image-based rendering of façades. In *Comp. Graph. Forum* (2011), vol. 30, pp. 1241–50. 3

[VRM*18] VOGELS T., ROUSSELLE F., MCWILLIAMS B., RÖTHLIN G., HARVILL A., ADLER D., MEYER M., NOVÁK J.: Denoising with kernel prediction and asymmetric loss functions. *ACM Trans. Graph. (Proc. SIGGRAPH) 37*, 4 (2018), 124:1–124:15. 3

[WB06] WANG Z., BOVIK A. C.: *Modern Image Quality Assessment*. Morgan & Claypool Publishers, 2006. 2

[WBF*17] WAECHTER M., BELJAN M., FUHRMANN S., MOEHRLE N., KOPF J., GOESELE M.: Virtual rephotography: Novel view prediction error for 3D reconstruction. *ACM Trans. Graph. 36*, 1 (2017). 3

[WBSS04] WANG Z., BOVIK A. C., SHEIKH H. R., SIMONCELLI E. P.: Image quality assessment: from error visibility to structural similarity. *IEEE TIP 13*, 4 (2004), 600–12. 1, 3, 4

[WFY*11] WILDEBOER M. O., FUKUSHIMA N., YENDO T., TEHRANI M. P., FUJII T., TANIMOTO M.: A semi-automatic depth estimation method for FTV. *Information and Media Technologies 6*, 2 (2011), 501–507. 9

[WGY*18] WOLSKI K., GIUNCHI D., YE N., DIDYK P., MYSZKOWSKI K., MANTIUK R., SEIDEL H.-P., STEED A., MANTIUK R. K.: Dataset and metrics for predicting local visible differences. *ACM Trans. Graph.* (2018). 2, 10

[Wil78] WILLIAMS L.: Casting curved shadows on curved surfaces. *SIGGRAPH Comput. Graph. 12*, 3 (1978), 270–4. 1

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. *CVPR* (2018). 1, 2, 4

[ZW97] ZHANG X., WANDELL B. A.: A spatial extension of CIELAB for digital color-image reproduction. *J ISD 5*, 1 (1997), 61. 2

[ZZMZ11] ZHANG L., ZHANG L., MOU X., ZHANG D.: FSIM: A feature similarity index for image quality assessment. *IEEE TIP 20*, 8 (2011), 2378–2386. 2