Supplemental Material for ACM Transactions on Graphics 2017 paper

”Virtual Rephotography:
Novel View Prediction Error for 3D Reconstruction”

Michael Waechter¹, Mate Beljan¹, Simon Fuhrmann¹,
Nils Moehrle¹, Johannes Kopf², and Michael Goesele¹
¹Technische Universität Darmstadt ²Facebook∗

For clarity and due to space considerations we did not include all results for all error metrics, datasets and reconstruction pipelines in the paper. On the following pages we show extended versions of figures found in the paper.

In particular we show errors for all image difference metrics that we investigated. The first metric works on individual pixels:

- Cb+Cr: We use the YCbCr color space and sum up the absolute errors in the two chrominance channels.

All other metrics compare image patches instead of single pixels:

- ZSSD: Zero-mean sum of squared differences handles varying exposure by subtracting the mean luminance within each image patch before comparing two patches.
- DSSIM: The Structural Dissimilarity Index is related to SSIM [3] via DSSIM = \frac{1-SSIM}{2}.
- NCC: Normalized Cross-Correlation normalizes the patch mean to 0 and its standard deviation to 1. We turn NCC into a dissimilarity metric, like the other metrics, by using 1-NCC.
- Census: Census converts each image patch into a binary descriptor by checking for each pixel whether it is lighter or darker than the patch center pixel [4]. Two patches’ difference is computed as their descriptors’ Hamming distance.
- iCID: The improved color image difference metric [2] predicts the perceived difference between two color images by normalizing them to reference viewing conditions (accounting for the viewing distance) and transforming the resulting images into a perceptually uniform color space without perceived cross-contamination of color attribute predictors (lightness, chroma and hue). Within the color attribute channels, local-average, local-contrast and local-structure-based difference features are extracted and merged to a single value reflecting the perceived difference between the images. We analyzed six variants (iCID perceptual, huepreserving, saturating, perceptual csf, huepreserving csf, and saturating csf). For details we refer the reader to Preiss’ paper [2].

∗Part of this work was done while Johannes Kopf was with Microsoft Research.
1 Evaluation with Synthetic Degradation

This section shows the full statistics for the paper’s Section 4.1.

In accordance with the ordering criterion from the paper’s Section 3, all error metrics reflect the increase in noise with a corresponding increase in error. This is even the case for the pixel-based \( C_b + C_r \) metric because there are no luminance differences due to different exposure since all images were exposure-adjusted prior to this experiment.

![Various error metrics](image)

**Figure 1.1:** Various error metrics for (top to bottom) texture noise, geometry noise and mesh simplification applied to Fountain-P11.


2 Evaluation with Multi-View Stereo Data

Here we show the full statistics for the paper’s Section 4.2 with all error metrics and both reconstruction pipelines: CMVS+PSR (1st & 3rd column) and CPC+MS (2nd & 4th column).

To obtain visually similar reconstruction results we use larger image resolutions for CMVS+PSR ($h \in \{375, 750, 1500\}$) than for CPC+MS ($h \in \{93, 187, 375\}$) because CMVS+PSR’s point clouds are sparser than CPC+MS’s depth maps. (Different image resolutions for both pipelines are irrelevant for the key statement of this experiment, because it is not about comparing both pipelines but about showing independently for both pipelines that virtual rephotography assigns higher errors when training set size or training image resolution is reduced.) Rephotos are taken at $h = 750$ regardless of reconstruction image size, and the patch size for the patch-based metrics is $9 \times 9$ pixels.

Figure 2.1: City Wall reconstruction completeness and 11 error metrics for the CMVS+PSR (1st & 3rd column) and the CPC+MS (2nd & 4th column) pipeline. Boxplots show minimum, lower quartile, median, upper quartile, and maximum of 20 cross-validation iterations. $|T|$ is the size of the training image set used for reconstruction and $h$ is the images’ shorter side length.
The plots show that both pipelines exhibit very similar behavior:

1. Completeness: Reconstruction completeness increases with increasing training set size $|T|$.

2. 1-NCC, Census, ZSSD, DSSIM, and six iCID variants: The patch-based metrics can distinguish between reconstructions of different training set size or different image resolution: When comparing boxplots of different $h$ and identical $|T|$ they assign a lower error to larger $h$. The same holds equivalently for different $|T|$ and identical $h$: A lower error is assigned to larger $|T|$. Thus, the patch-based metrics fulfill the ordering criterion from the paper’s Section 3. However, ZSSD, DSSIM and the iCID variants fulfill it less pronounced than 1-NCC and Census do: Boxplots of identical $|T|$ and different $h$ are less clearly separated and in some cases reconstructions with identical $h$ have a lower error for $|T| = 67$ than for $|T| = 133$.

3. $C_b+C_r$: The pixel-based $C_b+C_r$ metric fails to clearly distinguish different $|T|$ or different $h$ and an error of $\sim 0.02$ is assigned regardless of training set size or image resolution. This is in contrast to the previous experiment with synthetic degradation where the setting was significantly simpler because all images had identical illumination, camera response curves and exposure. The very same would happen with raw RGB difference which Fitzgibbon et al. used to evaluate their IBR method [1, Figure 7d].

4. Leaving out cross-validation (as described in the paper’s Section 4.4): For the two analyzed reconstruction pipelines leaving out cross-validation has no strong influence on the error in the patch-based metrics. In almost all cases the errors of the experiments without cross-validation ($|T| = 561$) are just slightly below those of the largest cross-validation experiments ($|T| = 553$).
3 Lying and Standing Statue

In this section we show two additional datasets which we evaluated analogously to the City Wall in the previous section. For these datasets the properties of our metrics (as discussed in the previous section) start to break, because the reconstructions of these datasets are too flawed in most cases. The reconstructions of the Lying Statue dataset (see Figure 3.1) are often seriously flawed because there are only few (49) input images available. We could not remove many images without completely destroying the reconstruction. Thus, we varied the training set size between $|T| = 23$ and 47 only. In this setting the 1-NCC error is unable to distinguish datasets of different $|T|$. However, it succeeded in separating datasets with different image resolutions $h$, although less pronounced compared to the City Wall dataset.

![Figure 3.1: Lying Statue dataset](image)

$|T|$ is the size of the training image set used for reconstruction and $h$ is the images’ shorter side length.

![Figure 3.2: Lying Statue dataset (49 input images) with rephoto completeness, 1-NCC and $C_C+C_r$ error for the CMVS+PSR (left) and the CPC+MS (right) pipeline. Boxplots show minimum, maximum, lower and upper quartile and median of 20 cross-validation iterations.](image)
Regarding the Standing Statue dataset (see Figure 3.3), the 3D models frequently contained superfluous geometry from the surface reconstruction step due to lots of foreground clutter (mostly plants) in the images. The 1-NCC error correctly ranks datasets with a larger training set size $|T|$ better. Also, it consistently ranks datasets with medium image resolution better than those with low resolution but it fails in distinguishing medium and high resolution. Both datasets demonstrate that virtual rephotography struggles with datasets that do not reconstruct well.

Figure 3.3: Standing Statue dataset

![Standing Statue dataset](image)

Figure 3.4: Standing Statue dataset (334 input images) with rephoto completeness, 1-NCC and $C_b+C_r$ error for the CMVS+PSR (left) and the CPC+MS (right) pipeline. Boxplots show minimum, maximum, lower and upper quartile and median of 20 cross-validation iterations. $|T|$ is the size of the training image set used for reconstruction and $h$ is the images’ shorter side length.
4 Comparison with Geometry-based Benchmark

Here we show the full statistics for the paper’s Section 4.3.

![Plot of rephoto error against geometric error for TempleRing and DinoRing benchmark submissions.](image)

All patch-based metrics have a similar data series layout and seem to be correlated with geometric error, i.e., reconstructions with a high geometric error tend to receive a high rephoto error. Especially the correlation coefficients for 1-NCC and Census in Table 4.1 underline this. ZSSD, DSSIM, and the different iCID variants exhibit a lower correlation to the geometric error.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Correlation of geometric error and rephoto error for all patch-based metrics.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-NCC</td>
</tr>
<tr>
<td>TempleRing</td>
<td>0.63</td>
</tr>
<tr>
<td>DinoRing</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Table 4.1: Correlation of geometric and rephoto error for all patch-based metrics.
5 Different Reconstruction Representations

This section shows the full statistics for the paper’s Section 5.1:

<table>
<thead>
<tr>
<th>Representation</th>
<th>Compl.</th>
<th>C_b+C_r</th>
<th>1-NCC</th>
<th>Census</th>
<th>ZSSD</th>
<th>DSSIM</th>
<th>iCID per. hue.</th>
<th>iCID sat.</th>
<th>iCID hue. csf</th>
<th>iCID sat. csf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mesh with texture</td>
<td>0.66</td>
<td>0.017</td>
<td>0.29</td>
<td>0.21</td>
<td>0.22</td>
<td>0.16</td>
<td>0.48</td>
<td>0.50</td>
<td>0.51</td>
<td>0.46</td>
</tr>
<tr>
<td>View-dep. texturing</td>
<td>0.66</td>
<td>0.017</td>
<td>0.39</td>
<td>0.26</td>
<td>0.24</td>
<td>0.19</td>
<td>0.52</td>
<td>0.54</td>
<td>0.55</td>
<td>0.50</td>
</tr>
<tr>
<td>Mesh with vertex color</td>
<td>0.66</td>
<td>0.016</td>
<td>0.47</td>
<td>0.32</td>
<td>0.26</td>
<td>0.22</td>
<td>0.58</td>
<td>0.59</td>
<td>0.61</td>
<td>0.56</td>
</tr>
<tr>
<td>Point cloud</td>
<td>0.67</td>
<td>0.016</td>
<td>0.51</td>
<td>0.33</td>
<td>0.27</td>
<td>0.23</td>
<td>0.59</td>
<td>0.60</td>
<td>0.62</td>
<td>0.57</td>
</tr>
<tr>
<td>Image-based rendering</td>
<td>0.66</td>
<td>0.022</td>
<td>0.62</td>
<td>0.36</td>
<td>0.46</td>
<td>0.28</td>
<td>0.69</td>
<td>0.70</td>
<td>0.71</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Table 5.1: Rephoto errors for different reconstruction representations of the Castle Ruin shown in the paper’s Figure 1.

Reconstruction completeness is approximately equal for all five representations, because for better comparability we restricted the view-dependent texturing and the image-based rendering to only render those scene parts where the depth maps indicated valid depth.

All patch-based metrics (1-NCC, Census, ZSSD, DSSIM, iCID) consistently rank the scene representations in the following order (from best to worst):

1. mesh with texture
2. view-dependent texturing
3. mesh with vertex colors
4. point cloud
5. image-based rendering

This order corresponds with our visual intuition when manually inspecting the datasets (which is also backed up by the user study in the paper’s Section 5.1.1). Image-based rendering being ranked worst may seem unintuitive but our implementation suffers from strong artifacts caused by imperfectly reconstructed planar depth maps used as geometric proxies. These cause strong error responses. In contrast to the patch-based metrics, the pixel-based $C_b+C_r$ error does not correspond with our intuition and even ranks the point cloud best.
6 Error Localization

In this section we show an extended version of the City Wall from the paper’s Section 5.2.

Figure 6.1: Left to right: Photo, rephoto with multiple color defects on the wall to the left of the door, 1-NCC error projection, and \( C_b + C_r \) error projection.

In Figure 6.1 we show a variant of the paper’s Figure 10 where we added the \( C_b + C_r \) error projection. The 1-NCC projection clearly detects the color defects. In contrast, the pixel-based \( C_b + C_r \) error does not detect it because it cannot distinguish between per-pixel luminance changes due to noise and medium- or large-scale changes due to illumination/exposure differences.

In Figure 6.2’s top row we see the textured City Wall reconstruction and the corresponding 1-NCC error projection. The error projection properly highlights superfluous geometry (at the wall’s boundaries) and hard to reconstruct geometry (the tower’s upper half, which was only photographed from a distance, or grass on the ground). It also strongly highlights mistextured parts, such as tree branches on the tower or the pedestrian on the left. In the bottom row we show a variant of the City Wall where photometric outliers such as the pedestrian and the tree have automatically been removed by the texturing algorithm. The error projection nicely tracks the outlier removal: The error in those regions turns from red to blue.

This scene can also be found in our supplemental video.

Figure 6.2: Textured City Wall reconstruction. Top left: Pedestrian and tree used as texture. Bottom left: A photo-consistency check during texturing removes pedestrian and tree. Right column: Corresponding 1-NCC error projections. The error projection tracks the removal of pedestrian and tree.
References


