Our QoE-based fusion method was compared with two state-of-the-art Multi-Exposure Fusion (MEF) methods and two state-of-the-art Tone-Mapping (TM) methods on six standard test sequences (Figures 2 to 7). QBF-1 denotes our algorithm that takes $\alpha_1 = 1, \alpha_2 = 0$. QBF-2 denotes our algorithm that computes $\alpha_1 = 0.6 + \exp(-\bar{L}), \alpha_2 = 0.02 \exp(-\bar{L})$. PF denotes the probabilistic fusion method proposed in our earlier study [1]. EF denotes the exposure fusion method proposed by Mertens et al. [2]. PTR denotes the photographic tone reproduction local TM operator proposed by Reinhard et al. [3]. iCAM06 denotes the TM operator proposed by Kuang et al. [4]. The results by EF, PTR, and iCAM06, were generated by the programs provided by their respective authors. QBF-1, QBF-2, PF, EF, and iCAM06 are Matlab implementations; PTR is C implementation. The same parameter setting as introduced in Section IV in the main article was used in QBF-1 and QBF-2 for all experiments unless otherwise mentioned. The default parameter settings in PF, EF, and iCAM06 were used. The parameters in PTR were generated using the estimation technique in [5]. The High Dynamic Range (HDR) images for TM methods were generated using HDR reconstruction [6] from the corresponding source sequences.

I. OBJECTIVE EVALUATION

A. Evaluation Using $Q^{AB/F}$

The $Q^{AB/F}$ metric [7] was used to measure the correctly transferred edge information from a set of input grayscale images to a fused image. Reproduction of both edge strength and orientation is considered. Traditional fusion quality metrics, including $Q^{AB/F}$, were not designed for cases with more than two source images in MEF. Nevertheless, $Q^{AB/F}$ has been shown to be one of the most robust and consistent metrics [8]. In [8], among twelve compared metrics under four categories, three were recommended for evaluating the resulting image quality of fusing a visible and infrared image pair and are applicable for our study. The four categories are: information theory-based, image feature-based, image structural similarity-based, and human perception-inspired. The three metrics are: the $Q^{AB/F}$ metric, Cvejic’s metric [9], and Yang’s metric [10]. In general, evaluation metrics often give a single global quality score for a fused image. Most evaluation metrics (including the three recommended metrics in [8]) are designed for the case of two source images, which (including Cvejics metric and Yangs metric) largely depend on the calculation and manipulation of covariance (or similar statistics) between the two source images and/or between the two source images and the fused image. Therefore, it is relatively difficult to extend such metrics to cases with multiple source images. The advantage of $Q^{AB/F}$ is that it does not rely on calculating statistical score between two source images, and thus it can be extended to processes involving multiple source images, such as MEF. This metric has also been proven to correspond best with subjective tests among several other popular metrics [11]. Therefore, we adapted $Q^{AB/F}$ in our evaluation.

$Q^{AB/F}$ assumes that stronger edges attract more visual attention, and therefore it also associates each edge with an importance coefficient defined on the edge strength. This metric gives a performance score between $[0, 1]$ for each test image, where a higher score means better performance. Because this metric works on grayscale images, we first converted both source images and resulting images into grayscale before performing the evaluation. The reported metric parameter values from [7] were used in our experiment. The performance scores are reported in Table I and Figure 1. Although PTR and iCAM06 are not MEF methods, they are included in the table for reference purposes only. All of the compared MEF methods successfully transferred most of the edge information, and they have very close performance according to $Q^{AB/F}$. On the average, QBF-
### Evaluation Using the \( Q^{AB/F} \) Metric

<table>
<thead>
<tr>
<th>Input</th>
<th>QBF-1</th>
<th>QBF-2</th>
<th>PF</th>
<th>EF</th>
<th>PTR</th>
<th>iCAM06</th>
</tr>
</thead>
<tbody>
<tr>
<td>House</td>
<td>0.715</td>
<td>0.739</td>
<td>0.715</td>
<td>0.732</td>
<td>0.515</td>
<td>0.578</td>
</tr>
<tr>
<td>Chateau</td>
<td>0.774</td>
<td>0.772</td>
<td>0.766</td>
<td>0.769</td>
<td>0.419</td>
<td>0.418</td>
</tr>
<tr>
<td>Memorial Church</td>
<td>0.665</td>
<td>0.658</td>
<td>0.661</td>
<td>0.662</td>
<td>0.628</td>
<td>0.495</td>
</tr>
<tr>
<td>Belgium House</td>
<td>0.609</td>
<td>0.617</td>
<td>0.621</td>
<td>0.609</td>
<td>0.568</td>
<td>0.560</td>
</tr>
<tr>
<td>National Cathedral</td>
<td>0.697</td>
<td>0.703</td>
<td>0.578</td>
<td>0.670</td>
<td>0.286</td>
<td>0.281</td>
</tr>
<tr>
<td>Lamp</td>
<td>0.564</td>
<td>0.583</td>
<td>0.552</td>
<td>0.571</td>
<td>0.365</td>
<td>0.478</td>
</tr>
</tbody>
</table>

Fig. 1. Objective evaluation using \( Q^{AB/F} \). All of the compared MEF methods successfully transferred most of the edge information, and they have very close performance according to \( Q^{AB/F} \). On the average, QBF-2 has slightly better performance than the others.

### B. Evaluation Using DRIVDP

To strengthen the evaluation capability of \( Q^{AB/F} \), we incorporate the Dynamic Range Independent Visible Difference Predictor (DRIVDP) [12] to assess per-pixel fusion quality. DRIVDP was used to assess the visual distortions between a test image and each source image. Three distortions are considered: loss of visible contrast; amplification of invisible contrast; reversal of visible contrast. These distortions are reflected in a distortion map, where green, blue, red, and gray pixels indicate contrast loss, amplification, reversal, and no distortion, respectively. Please note that contrast amplification is normally considered as one of the objectives in image fusion. We assume that the images were viewed on a typical LCD with a maximum luminance equivalent to \( 100 \text{ cd/m}^2 \), a gamma value of 2.2, and a visual resolution of 30 pixels per degree at a viewing distance of 0.5 meter and that the peak contrast sensitivity of the viewer is 0.25%.

We chose two images from each of the six source sequences for this evaluation. The evaluation result using two images from the Memorial Church sequence is given in Figure 8 (please refer to Figure 2 for visual comparison). The source image on the left gives good exposure for the ceiling and wall that appear in the right portion of the image. The source image on the right gives good exposure for the lower left window. For the ceiling and wall, PF, QBF-1, and QBF-2 show relatively less distortion, followed by EF, iCAM06, and PTR. For the lower left window, iCAM06 shows the least distortion, followed by QBF-1, QBF-2, PTR, PF, and EF.

The evaluation result on the National Cathedral sequence is given in Figure 9 (please refer to Figure 3 for visual comparison). The two source images with good exposures respectively for the windows and the wall are given in Figure 9(a). The contrast distortion maps generated using DRIVDP for each method are given in Figure 9(b)-(g). QBF-1 performs best in preventing contrast distortions, followed by QBF-2, PF, EF, PTR, and iCAM06.

The evaluation result using two images from the House sequence is given in Figure 10 (please refer to Figure 4 for visual comparison). The source image at the top gives good exposure for the outdoor scene, and the source image at the bottom gives good exposure for the bookshelf. For these two source images, QBF-2 and EF have close performance, followed by QBF-1, PF, iCAM06, and PTR.
The evaluation result using two images from the Chateau sequence is given in Figure 11 (please refer to Figure 5 for visual comparison). The source image on the left gives good exposure for the balcony and trees. For these regions, QBF-1, QBF-2, and iCAM06 have close performance, followed by PTR, PF, and EF. The source image on the right gives good exposure for the indoor scene. For the indoor scene, QBF-1, QBF-2, PF, and EF have close performance, followed by PTR and iCAM06.

The evaluation result using two images from the Lamp sequence is given in Figure 12 (please refer to Figure 6 for visual comparison). The two source images give good exposures for the bulb and the books, respectively. For the bulb, QBF-2 shows the least distortion, followed by QBF-1, PTR, PF, EF, and iCAM06. For the books, QBF-2 shows the least distortion, followed by EF, PF, QBF-1, iCAM06, and PTR.

The evaluation result using two images from the Belgium House sequence is given in Figure 13 (please refer to Figure 7 for visual comparison). The two source images give good exposures for part of the outdoor scene and the indoor scene, respectively. For the outdoor scene, iCAM06 shows the least distortion, followed by QBF-2, QBF-1, PF, EF, and PTR. For the indoor scene, QBF-2 and EF show the least distortion, followed by QBF-1, PF, PTR, and iCAM06.

On the average, for the twelve source images considered in this evaluation, the ranking of the six methods are: QBF-2 > QBF-1 > EF ≈ PF > iCAM06 > PTR.

II. Subjective Evaluation

The four criteria in our evaluation, i.e., global contrast, details, colors, and overall appearance, are a subset of those used by Čadík et al. [13]. The two criteria not used are brightness and artifacts. Not using the brightness criterion is because its impact already spreads into the other criteria and therefore reveals itself indirectly [13]. The reason for not using the artifacts criterion is twofold: 1) The subjects were non-experts with no prior experience in MEF or TM. Therefore, their understanding of artifacts can vary from one individual to another. It is difficult for them to identify artifacts solely based on their own judgement. 2) The methods selected for our experiment did not produce major artifacts except some minor color distortion in some results, but this can be covered under the colors criterion.

The subjective evaluation results are summarized in Figure 14. Our QBF-2 received the highest average ranking scores for all four criteria on five scenes, and has similar performance to our QBF-1 on the National Cathedral sequence. Considering all six scenes, the ranking from high to low for the six methods is: QBF-2 > EF > QBF-1 > PF > iCAM06 > PTR for the global contrast criterion; QBF-2 > QBF-1 > EF > iCAM06 ≈ PF > PTR for the details criterion; QBF-2 > EF > QBF-1 > PF > iCAM06 > PTR for the colors criterion; QBF-2 > QBF-1 ≈ EF > PF > iCAM06 > PTR for the overall appearance criterion. On the average, the ranking of the six methods on the six scenes from high to low is: QBF-2 > EF ≈ QBF-1 > PF > iCAM06 > PTR.

The experimental results also suggest that there is no single factor that dominates the user preference, but rather it is the combination of different factors (e.g., global contrast and details) that determines a user’s choice. Below are two examples:

- Although some images may suffer some loss of subtle details due to the enhanced contrast, high global contrast and good color scheme can compensate for that in the visual impression, which results in higher ranking scores. For the Lamp scene (Figure 6), QBF-1 received higher rating than EF under the details criterion, but EF received higher ratings under the global contrast and the colors criteria. Under the overall appearance criterion, EF received higher rating than QBF-1.

- On the contrary, good detail reproduction may also compensate for low global contrast or less vivid color scheme in the overall impression. For the Chateau scene (Figure 5), EF received higher ratings than PF under the global contrast and the colors criteria, but PF received slightly better details rating. Both methods received very close ratings under the overall appearance criterion. Also for the Chateau scene, QBF-1 received very close ratings with EF under the global contrast and the colors criteria, but received much higher rating under the details criterion. Under the overall appearance criterion, QBF-1 has higher rating than EF.

Because user interaction is usually exploited in TM to generate images of different appearances, we also tested different parameter settings of iCAM06. In iCAM06, the parameter p controls the brightness of the tone-mapped image. The default value of p is 0.7 in the implementation of iCAM06 provided by the authors of [4], which is the value that we took in the above experiments. A larger p results in a darker image and a smaller p results in a brighter image. The suggested range of p in [4] is [0.6, 0.85]. We also tested p = 0.6 and p = 0.4 for the six test scenes. The results are given in Figure 15, along with comparisons with our QBF-1 and QBF-2. For each scene, the results by iCAM06
with \( p = 0.6 \), iCAM06 with \( p = 0.4 \), QBF-1, and QBF-2 are given at the left most, left, right, and right most, respectively. Our QBF-1 and QBF-2 produce better details, contrasts, and colors.

III. EVALUATION OF THE APPLICABILITY OF OTHER FUSION QUALITY METRICS IN MEF

Aside from \( Q^{AB/F} \), we also investigated the applicability of two other traditional fusion quality metrics (Cvejic’s metric and Yang’s metric) in MEF. These two metrics require two source images, and produce a single global quality score for a fused image. We applied Cvejic’s metric, Yang’s metric, and \( Q^{AB/F} \) metric to the same subsequences used in the DRIVDP-based evaluation. Since such metrics estimate local structural similarity in the luminance channel between source images and the fused image, the details criterion in the subjective study provides some useful information on evaluating their performance. Their quality scores for the six compared algorithms are plotted against the average ranking scores under the details criterion in the subjective evaluation (normalized to \([0, 1]\)) in Figure 16. \( Q^{AB/F} \) metric produced better correspondence with the details criterion in the subjective evaluation than the other two metrics.

IV. ANALYSIS OF THE CONTRAST THRESHOLD

In the local contrast calculation, we apply a threshold \( \theta \) to the physical contrast to suppress the contribution from under-exposed regions. In Figure 17, the local contrast \( C_{i,k}^{n} \) is plotted as a function of the coefficient \( G_{i,k}^{n} \) and its corresponding lowpass-filtered coefficient \([\phi \ast G_{i,k}^{n}]_{i}\) under different contrast thresholds. As shown in the plots, a threshold is necessary for suppressing the contribution from the dark regions, whose lowpass-filtered coefficients are close to zero. As the threshold \( \theta \) increases, the difference in the calculated physical contrasts for different coefficient values becomes less significant.

Visual comparisons of the effects of different thresholds on the fusion results are shown in Figures 18 and 19. When the threshold \( \theta \) is no less than 0.2, the resulting image is brighter and preserves more details. When \( \theta \) is above 0.4, the image shows less vivid colors. The fused images were generated using the perceived contrast measure only. Quantitative comparisons using the \( Q^{AB/F} \) metric are given in Figures 20 and 21. Based on the \( Q^{AB/F} \) score, a threshold value above 0.1 gives relatively better performance. Hence, in practice, we suggest using \( \theta \in [0.2, 0.4] \). Our fusion method, with an appropriate contrast threshold value (0.3 was used in our formal evaluation), produces better results than other methods.

V. EVALUATION OF DIFFERENT TRANSUDER FUNCTIONS

Our fusion method currently employs the transducer function proposed by Foley and Schwarz [14]. Different transducer functions were proposed by other researchers, e.g., [15], [16]. A comparison between different transducer functions is given in Figure 22. Except the transducer function, all other settings in our fusion algorithm were the same when generating the fusion results. The parameters in the individual transducer functions were all set to the reported values from their respective papers. Wilson’s [16] transducer function for threshold and suprathreshold vision produces fusion results with quality very close to Foley’s [14]. García-Pérez’s [15] transducer function produces images with a bit lower local contrasts and less vivid colors. This happens because: the shapes and ranges of Foley’s and Wilson’s transducer functions are similar, and therefore, when combined with the psychometric function, they produce similar detection probabilities for the same normalized physical contrast; compared with Foley’s and Wilson’s combined transducer and psychometric functions, García-Pérez’s combined transducer and psychometric function has a little different shape that produces a bit less differentiation between low contrast levels. These functions are plotted in Figure 23. However, in terms of gray-scale edge information transfer, the three transducer functions have the similar performance (e.g., they have the same \( Q^{AB/F} \) score on the Lamp sequence).

VI. EVALUATION OF DIFFERENT SATURATION MEASURES

Our fusion method currently employs the saturation measure defined in the LHS (luminance, hue, and saturation) color space [17]. Other saturation measures, such as the saturation measure defined in the HSV (hue, saturation, and value) color space and Lübbe’s [18] saturation measure in the CIELAB color space, generate similar results in our fusion scheme. A comparison between different saturation measures is given in Figure 24. Except the saturation calculation scheme, all other settings in our fusion algorithm were the same when generating the fusion results. The saturation measures in the LHS and HSV color spaces and Lübbe’s saturation measure generate very close fusion results when applied in QBF-1. The saturation measure in the HSV color space and Lübbe’s saturation measure show a little overexposure when applied in QBF-2. The fused images
using Mertens’ [2] saturation measure that takes the standard deviation between the R, G, and B components show a lighter color scheme when applied in QBF-1 and a little over-exposure when applied in QBF-2.

REFERENCES


Fig. 2. Comparison of QBF-1 and QBF-2 with PF, EF, PTR, and iCAM06 on the Memorial Church sequence. QBF-1 and QBF-2 preserve more details than the others, especially in the window regions. QBF-2 produces higher global contrast than QBF-1, but may cause loss of subtle details, e.g., in the window regions. QBF-2 has the highest average ranking scores under all four criteria on this sequence in the subjective evaluation.
Fig. 3. Comparison of QBF-1 and QBF-2 with PF, EF, PTR, and iCAM06 on the National Cathedral sequence. QBF-1 and QBF-2 generate the best results, where details in all regions are preserved with high local contrasts and saturated colors. iCAM06 preserves as many details as ours in the window regions but with less vivid colors. Our QBF-1 has the highest average ranking scores under all four criteria on this sequence in the subjective evaluation, followed by our QBF-2.
Fig. 4. Comparison of QBF-1 and QBF-2 with PF, EF, PTR, and iCAM06 on the House sequence. Our QBF-2 has the highest average ranking scores under all four criteria on this sequence in the subjective evaluation.
Fig. 5. Comparison of QBF-1 and QBF-2 with PF, EF, PTR, and iCAM06 on the Chateau sequence. Our QBF-2 has the highest average ranking scores under all four criteria on this sequence in the subjective evaluation.
Fig. 6. Comparison of QBF-1 and QBF-2 with PF, EF, PTR, and iCAM06 on the Lamp sequence. Our QBF-2 has the highest average ranking scores under all four criteria on this sequence in the subjective evaluation.
Fig. 7. Comparison of QBF-1 and QBF-2 with PF, EF, PTR, and iCAM06 on the Belgium House sequence. Our QBF-2 has the highest average ranking scores under all four criteria on this sequence in the subjective evaluation.
Fig. 8. Comparison of QBF-1 and QBF-2 with PF, EF, PTR, and iCAM06 on the Memorial Church sequence using DRIVDP. The source image on the left gives good exposure for the ceiling and wall that appear in the right portion of the image. The source image on the right gives good exposure for the lower left window. In a distortion map, green, blue, red, and gray pixels indicate contrast loss, amplification, reversal, and no distortion, respectively. For the ceiling and wall, PF, QBF-1, and QBF-2 show relatively less distortion, followed by EF, iCAM06, and PTR. For the lower left window, iCAM06 shows the least distortion, followed by QBF-1, QBF-2, PTR, PF, and EF.
Fig. 9. Comparison of QBF-1 and QBF-2 with PF, EF, PTR, and iCAM06 on the National Cathedral sequence using DRIVDP. The two source images give good exposures for the windows and the wall, respectively. QBF-1 performs best in preventing contrast distortions, followed by QBF-2, PF, EF, PTR, and iCAM06.
Fig. 10. Comparison of QBF-1 and QBF-2 with PF, EF, PTR, and iCAM06 on the House sequence using DRIVDP. The two source images give good exposures for the outdoor scene and the bookshelf, respectively. For these two source images, QBF-2 and EF have close performance, followed by QBF-1, PF, iCAM06, and PTR.
Fig. 11. Comparison of QBF-1 and QBF-2 with PF, EF, PTR, and iCAM06 on the Chateau sequence using DRIVDP. The source image on the left gives good exposure for the balcony and trees. For these regions, QBF-1, QBF-2, and iCAM06 have close performance, followed by PTR, PF, and EF. The source image on the right gives good exposure for the indoor scene. For the indoor scene, QBF-1, QBF-2, PF, and EF have close performance, followed by PTR and iCAM06.
Fig. 12. Comparison of QBF-1 and QBF-2 with PF, EF, PTR, and iCAM06 on the Lamp sequence using DRIVDP. The two source images give good exposures for the bulb and the books, respectively. For the bulb, QBF-2 shows the least distortion, followed by QBF-1, PTR, PF, EF, and iCAM06. For the books, QBF-2 shows the least distortion, followed by EF, PF, QBF-1, iCAM06, and PTR.
Fig. 13. Comparison of QBF-1 and QBF-2 with PF, EF, PTR, and iCAM06 on the Belgium House sequence using DRIVDP. The two source images give good exposures for part of the outdoor scene and the indoor scene, respectively. For the outdoor scene, iCAM06 shows the least distortion, followed by QBF-2, QBF-1, PF, EF, and PTR. For the indoor scene, QBF-2 and EF show the least distortion, followed by QBF-1, PF, PTR, and iCAM06.
Fig. 14. Average ranking scores of different algorithms in the subjective evaluation. Our QBF-2 performs the best under all four criteria for 5 out of 6 scenes, and shows similar performance to our QBF-1 on the other scene. QBF-1 and EF have similar performance on the average though QBF-1 gives better detail reproduction, followed by PF, iCAM06, and PTR.
Fig. 15. Comparison of iCAM06 with $p = 0.6$, iCAM06 with $p = 0.4$, our QBF-1, and our QBF-2. For each scene, the results by iCAM06 with $p = 0.6$, iCAM06 with $p = 0.4$, QBF-1, and QBF-2 are given at the left most, left, right, and right most, respectively. QBF-1 and QBF-2 produce better details, contrasts, and colors.
Fig. 16. Comparison of the performance of three fusion quality metrics ($Q^{AB/F}$ metric, Cvejic’s metric, and Yang’s metric) using the same subsequences as in the DRIVDP-based evaluation. $Q^{AB/F}$ metric produced better correspondence with the details criterion in the subjective evaluation than the other two metrics.
Fig. 17. Plots of the local contrast $C_{i,k}^n$ as a function of the coefficient $G_{i,k}^n$ and its corresponding lowpass-filtered coefficient $[\phi * G_{i,k}^n]$, under different contrast thresholds. As the threshold $\theta$ increases, the difference in the calculated contrast for different coefficient values becomes less significant. As shown in the plots, a threshold is necessary for suppressing the contribution from the dark regions, whose lowpass-filtered coefficients are close to zero.
Fig. 18. Comparison of different contrast thresholds on the Flower sequence. When the threshold $\theta$ is no less than 0.2, the resulting image is brighter and preserves more details. When $\theta$ is above 0.4, the image shows less vivid colors.
Fig. 19. Comparison of different contrast thresholds on the Optimus sequence. When the threshold $\theta$ is no less than 0.2, the resulting image is brighter and preserves more details. When $\theta$ is above 0.4, the image shows less vivid colors.
Fig. 20. Comparison of different contrast thresholds on the Flower sequence using the $Q^{AB/F}$ metric. Based on the $Q^{AB/F}$ score, a threshold value above 0.1 gives relatively better performance.

Fig. 21. Comparison of different contrast thresholds on the Optimus sequence using the $Q^{AB/F}$ metric. Based on the $Q^{AB/F}$ score, a threshold value above 0.1 gives relatively better performance.
Fig. 22. Comparison of different transducer functions on the Lamp sequence. Wilson’s transducer function for threshold and suprathreshold vision produces fusion results with quality very close to Foley’s. García-Pérez’s transducer function produces images with a bit lower local contrasts and less vivid colors.
Fig. 23. Plots of different transducer functions with and without the psychometric function. The shapes and ranges of Foley’s and Wilson’s transducer functions are similar, and therefore, when combined with the psychometric function, they produce similar detection probabilities for the same normalized physical contrast. Compared with Foley’s and Wilson’s combined transducer and psychometric functions, García-Pérez’s combined transducer and psychometric function has a little different shape that produces a bit less differentiation between low contrast levels.
Fig. 24. Comparison of different saturation measures on the Belgium House sequence. The saturation measures in the LHS and HSV color spaces and Lübbe’s saturation measure generate very close fusion results when applied in QBF-1. The saturation measure in the HSV color space and Lübbe’s saturation measure show a little over-exposure when applied in QBF-2. The fused images using Mertens’ saturation measure show a lighter color scheme when applied in QBF-1 and a little over-exposure when applied in QBF-2.