Supplemental Material for a Content-Adaptive Visibility Predictor for Perceptually Optimized Image Blending

TAIKI FUKIAGE, NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Japan
TAKESHI OISHI, Institute of Industrial Science, The University of Tokyo, Japan

In Section S1, we first present a preliminary experiment that compared the data quality of the online and laboratory experiments. In Section S2, we show additional model validation results in which we compare model prediction with human psychophysical data on artificial stimuli. Section S3 shows the training curve of our model. Section S4 provides the details of the comparison models used in the model validation as well as their optimized parameters trained on our dataset. Then, in Section S5, we present additional validation results of our blending method to further demonstrate its effectiveness. Finally, in Section S6, we present the effect of each loss function used in our proposed perceptually optimized image blending method.

CCS Concepts: • **Computing methodologies** → **Perception**; **Visibility**.

Additional Key Words and Phrases: alpha blending, image blending, human visual system, contrast perception, visibility

ACM Reference Format:
Taiki Fukiage and Takeshi Oishi. 2022. Supplemental Material for a Content-Adaptive Visibility Predictor for Perceptually Optimized Image Blending. *ACM Trans. Appl. Percept.* 1, 1 (October 2022), 18 pages. https://doi.org/10.1145/3565972

1 COMPARISON OF DATA OBTAINED IN ONLINE AND LABORATORY EXPERIMENTS

Prior to the main experiment described in Section 4 of the main paper, a preliminary experiment was conducted to investigate whether the same quality of data as obtained in a laboratory experiment could be collected in an online experiment. The task and procedure were identical to the *Same* target condition in the main experiment. For the stimuli, we sampled eight image patches from the McGill Calibrated Color Image Database [8] and used all 56 possible image combinations. Six discrete levels of reference alpha values were tested for each image pair. For the laboratory experiment, we recruited nine participants who had normal or corrected-to-normal vision. The stimuli were presented on LCD displays (Eizo ColorEdge CG2420-Z, 24.1 inch, 1920 × 1200 pixel resolution) that were color-calibrated using built-in calibration sensors. The participants observed the stimuli at a distance of 50 cm from the screen. The stimuli were presented in their original resolution (256 × 256 pixels), which corresponded to a 7.9 × 7.9 deg visual angle. For the online experiment, we recruited 11 participants via the same crowdsourcing service (Prolific.ac) used in the main experiment. The other features of the experimental setup were the same as those described in Section 4.1.1 of the main paper. We excluded from the analysis the data of two crowdsourcing participants who had exceptionally low reliability scores (<0.4) (see Appendix B for the computation of the reliability scores). In both the laboratory and online experiments, each participant completed all conditions.

Figure 1 presents a scatter plot comparing the average responses in the laboratory and online experiments. As seen in the plot, the correlation of the responses between the two experimental conditions was remarkably

Authors' addresses: Taiki Fukiage, NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, 3-1, Wakamiya, Morinosato, Atsugi, Kanagawa Pref., Japan, 243-0198, t.fukiage@gmail.com; Takeshi Oishi, Institute of Industrial Science, The University of Tokyo, 4-6-1, Komaba, Meguro-Ku, Tokyo, Japan, oishi@cvl.iis.u-tokyo.ac.jp.

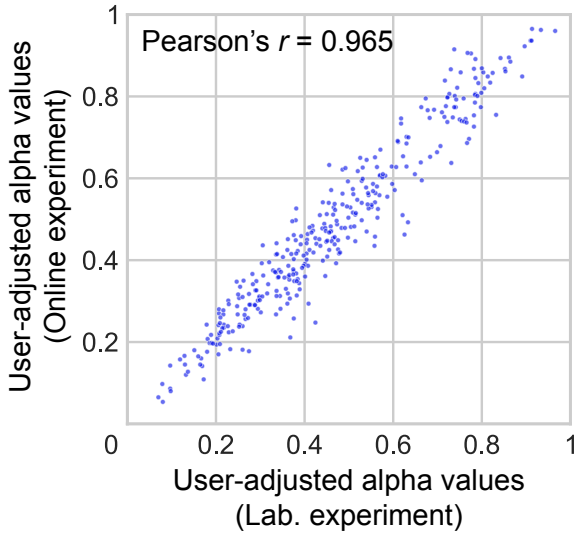

Fig. 1. Comparison of data obtained in online and laboratory experiments.

high (Pearson's $r = 0.965$). Thus, we concluded that for the task used in our study, it was possible to collect data through an online experiment with the same quality as that obtained in a controlled laboratory experiment.

2 MODEL VALIDATION BY COMPARISON WITH HUMAN PSYCHOPHYSICAL DATA ON ARTIFICIAL STIMULI

2.1 Comparison with subjective detectability

Although our visibility model is primarily developed to predict the visibility of blended images, it is also interesting to see whether it can reproduce the known properties of the contrast perception. For this purpose, we tested images taken from the PERCEPTIONPATTERNS dataset [1, 10] with subjective detectability. This dataset provides pairs of a *Reference* stimulus (mask pattern) and *Test* stimulus (mask + target). As our model requires a target-only image for input, we subtracted a *Reference* stimulus from a *Test* stimulus. Then, we replaced the alpha-blending equation used in the model with a linear addition of the target and masker (=reference). For the opaque input, we used a target-only image and a uniform gray image with a mean intensity of the original *Reference* stimulus for an input target image and a reference image, respectively. The image sizes were rescaled to simulate the observation conditions in [10].

The results for these images are presented in Fig. 2. The fourth column displays the visibility maps, while the fifth column displays the spatially-aggregating visibility values obtained by Eq. 10 in the main paper. The third column displays the marking results collected in [10], indicating the subjective detectability of the signal patterns.

It should be noted that the predicted visibility values were not directly comparable to the detectability because the detectability could not represent differences in the suprathreshold visibility. Nevertheless, the variation of the predicted visibility levels exhibited remarkably similar patterns to the marking results. The prediction result for "freq" exhibited a decrease in visibility at both lower and higher frequencies when the stimulus contrast was low, mimicking the typical CSF in the HVS. This was surprising given that our model did not explicitly model the CSF but was derived from our psychophysical experiment.

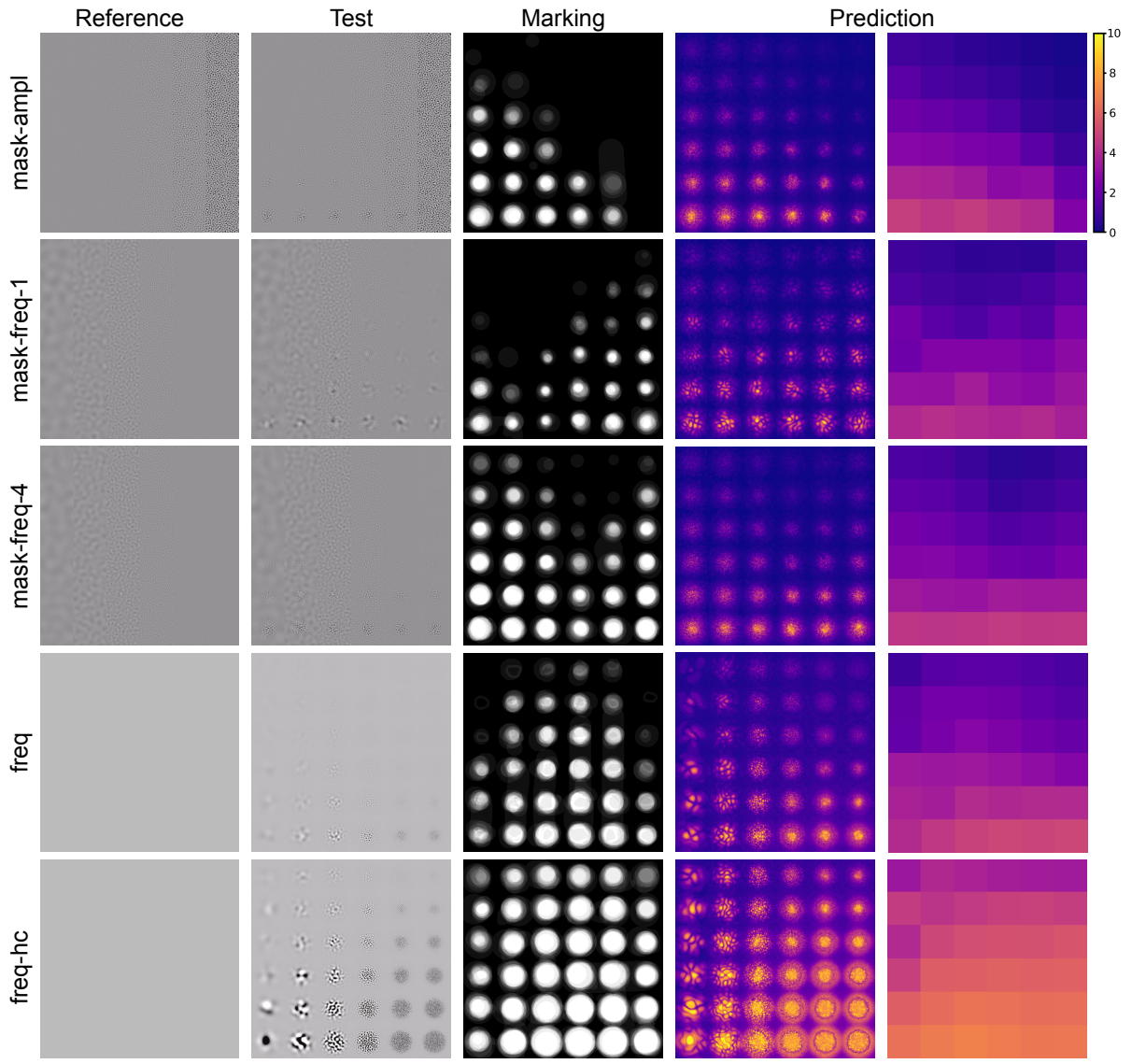

Fig. 2. Validation on PERCEPTIONPATTERNS dataset [10]. The first and second columns show the stimulus images. The third column shows the marking results of human subjects [10]. The fourth and fifth columns respectively show the predicted visibility maps and the spatially-aggregated visibility values.

The prediction result for "freq-hc" exhibited higher sensitivity in the high-frequency range as the contrast increased. Although this behavior is not captured in the marking results, exactly the same trend was observed in the contrast masking study using high contrast stimuli (see "CSF flattening" section in [7]).

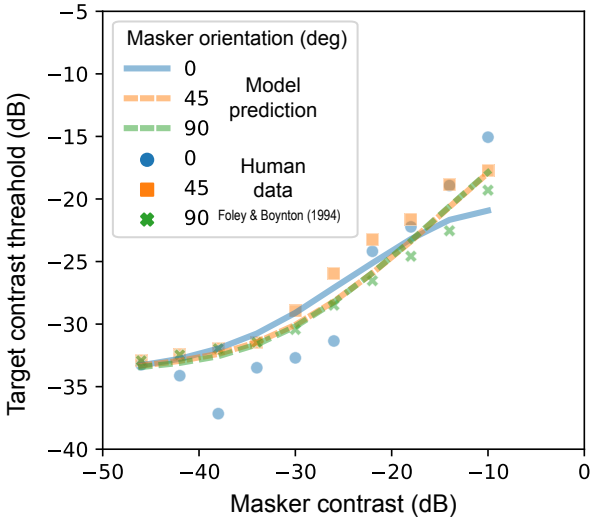

Fig. 3. Model prediction for contrast masking data by [2].

2.2 Comparison with contrast masking data

To validate the non-linearity produced by the divisive normalization process in our model, we compared the model prediction with the psychophysically measured contrast masking data reported by Foley and Boynton [2]. In this experiment, a target Gabor pattern with a 2 cpd vertical grating was embedded in a masker pattern with a grating of the same spatial frequency and the same phase. The contrast thresholds for the target pattern were measured for various masker contrasts and orientations. Minor modifications were required to our model to predict the visibility of contrast-masking stimuli. First, we replaced the alpha-blending equation used in the model with a linear addition of the target and masker (=reference). The stimuli were then encoded in sRGB color space to offset the linearization process in the L*a*b* color conversion used in our model. For the opaque input (the bottom path in Fig. 4 of the main paper), we used a target Gabor pattern with the maximum contrast for the target and a uniform gray image for the reference. (Note that the result did not significantly change with the choice of contrast level used for the opaque input). The target contrast threshold was computed as the target contrast at which our model outputs a threshold visibility level. Only this threshold visibility level was adjusted so as to best fit the contrast masking data with all the other parameters fixed.

Comparison between the model prediction and the human data is shown in Fig. 3. The results show that our model can successfully simulate the characteristic nonlinearity in suprathreshold contrast perception. However, our model cannot explain the slight differences between orientations because it does not analyze the orientation of patterns. In addition, the deviation is larger under the 0-deg condition in which the target and masker had exactly the same pattern (i.e., orientation, spatial frequency, and phase). However, the threshold decrease observed in the low masker contrast condition is known to disappear when the target and masker have a different spatial frequency, orientation, phase, or temporal frequency [2]. Thus, we believe that this effect has little impact in practical situations.

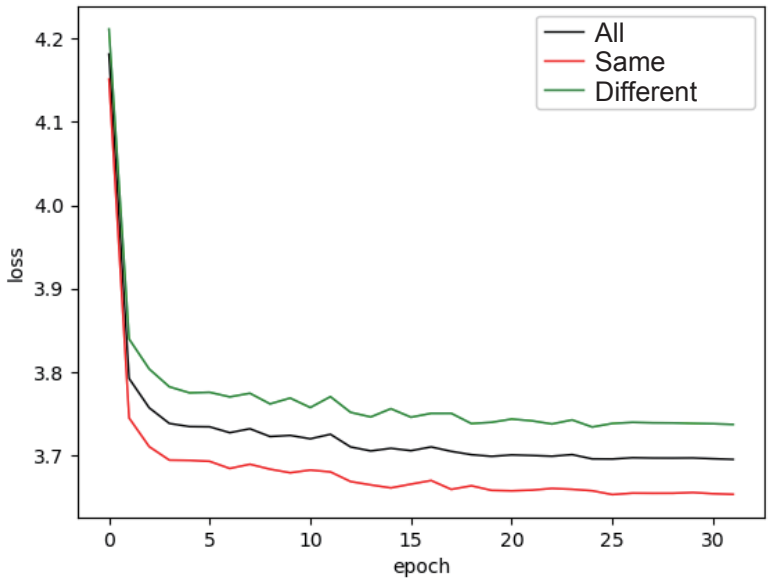

Fig. 4. The training curve when we optimized our model on the entire dataset. The red, green, and black lines indicate the loss (i.e., expected negative log likelihood) computed over the Same condition, Different condition, and both of the conditions, respectively.

3 TRAINING CURVE OF OUR MODEL CALIBRATION

Here, we show the training curve of our model when it was optimized on the entire dataset used in the calibration. We confirmed the convergence of the training by observing that the loss had reached a plateau by the end of the training as shown in Fig. 4.

Table 1. Optimized parameters of proposed model.

Fold	1	2	3	4	5	All
$\beta' \times 10^{-4}$	10.121	8.9496	9.9585	6.8645	12.674	7.6320
γ	0.3973	0.3986	0.3963	0.4018	0.3996	0.3964
p	0.7681	0.7637	0.7615	0.7640	0.7560	0.7658
q	5.4250	5.2820	5.2002	5.4564	5.5425	5.5737
s	4.5903	4.2508	4.2972	4.3239	4.3231	4.2585

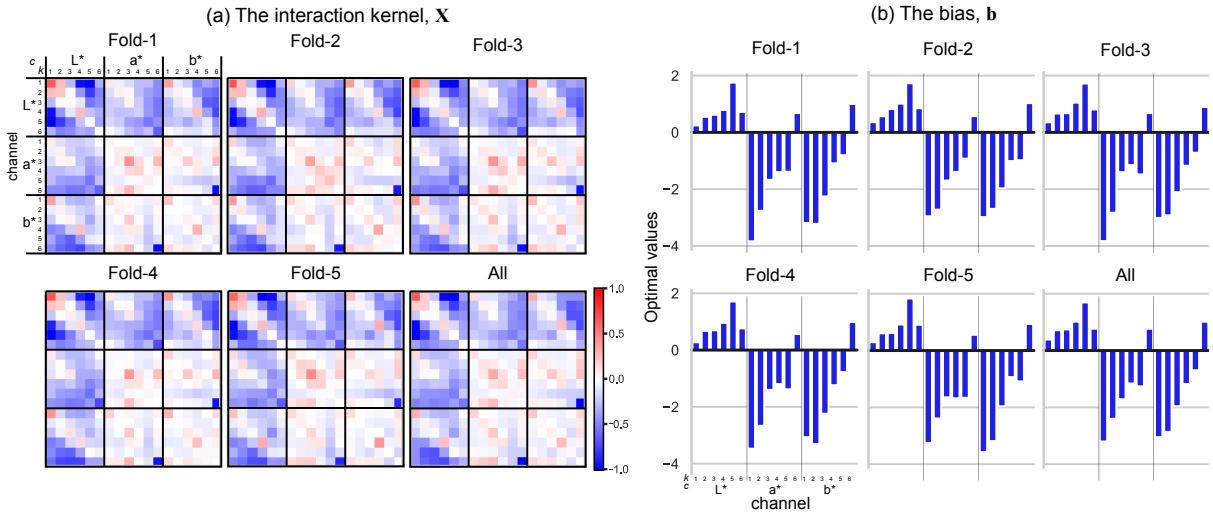

Fig. 5. (a) Optimized interaction kernel $\mathbf{X}$ and (b) bias $\mathbf{b}$ of our proposed model.

4 OPTIMIZED PARAMETERS OF VISIBILITY PREDICTOR MODELS

In this section, we present the optimized parameters of all comparison models (including our proposed model), which were obtained by the 5-fold cross validation. For all models, we used the same training procedure and loss function as those used to train our model. Thus, the parameter s in Eq. 11 of the main paper was also optimized jointly with the parameters of each model.

Ours (full). The optimized parameters of our model obtained using a 5-fold cross validation set as well as the entire dataset are presented in Table 1 and Fig. 5. The parameters obtained from the entire dataset are the same as those in Table 1 and Fig. 10 of the main paper; however, they are presented here for completeness.

PSNR. The optimized parameters obtained with the PSNR model are presented in Table 2. Although the PSNR itself has no free parameters, we fitted the scaling parameter s of Eq. 11 in the main paper to obtain the negative log likelihood of the model.

Table 2. Optimized parameters for PSNR model.

Fold	1	2	3	4	5	All
s	0.0642	0.0664	0.0654	0.0656	0.0644	0.0671

MS-SSIM. The MS-SSIM is the multiscale version of the structural similarity index [9] and contains weighting parameters to control the contribution of each scale. We recalibrated these parameters (i.e., $\{(\alpha_k, \beta_k, \gamma_k) \mid k = 1, 2, \ldots, N\}$) using the data obtained in our experiment. Following [9], we constrained the weights such that $\alpha_k = \beta_k = \gamma_k > 0$ and $\sum_k \gamma^k = 1$. The number of scales used was $N = 5$. Note that this is equivalent to the number of scales in our model ($N = 6$) because the low-pass image similarity measure is included in the last scale of the MS-SSIM, whereas in our model, the low-pass component is counted as an independent scale. The results of the optimization are presented in Table 3.

Table 3. Optimized parameters of MS-SSIM.

Fold	1	2	3	4	5	All
γ_1	0.5186	0.5242	0.5007	0.5002	0.5182	0.5149
γ_2	0.0000	0.0006	0.0000	0.0000	0.0000	0.0000
γ_3	0.4492	0.4423	0.4639	0.4667	0.4465	0.4504
γ_4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
γ_5	0.0322	0.0335	0.0354	0.0331	0.0354	0.0347
s	7.7373	7.7032	7.6581	7.6362	7.6969	7.6856

HDR-VDP. We used the latest version of HDR-VDP (ver. 3.0.6) [7]. We recalibrated the weights $\{w_f\}_{f=1}^{F}$ of the spatial frequency channels for aggregating the band-pass contrast response D into a visibility score Q. The aggregation equation was modified from Eq. (24) in the original paper [7] to introduce extra freedom as follows:

$$Q = \frac{1}{F \cdot O} \sum_{f=1}^{F} \sum_{o=1}^{O} w_f \left(-\epsilon + \log \left(\frac{1}{I} \sum_{i=1}^{I} D^2[f, o](i) + \exp(\epsilon) \right) \right)$$

, where ϵ is an offset parameter that we optimized together with w_f. The weight w_f was constrained to be positive during optimization. The optimized parameters are presented in Table 4. We used a Matlab implementation provided by the authors (HDR-VDP 3.0.6) to compute D. When running the code, we set the task parameter to "quality". The number of spatial frequency channels was $F = 6$, which was determined by setting the pixels per degree (ppd) parameter to 32.4 according to the observation condition in our experiment. The number of orientation channels was by default $O = 1$ in the latest version (3.0.6) of HDR-VDP.

Table 4. Optimized parameters of HDR-VDP.

Fold	1	2	3	4	5	All
w_1	0.1956	0.2028	0.1893	0.1954	0.1955	0.2035
w_2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
w_3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
w_4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
w_5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
w_6	0.1241	0.1219	0.1212	0.1188	0.1194	0.1253
ϵ	-9.3913	-9.4605	-9.4773	-9.4307	-9.5328	-9.8224
s	1.3474	1.3522	1.3429	1.3461	1.3364	1.2965

IQA model used in [3]. Fukiage et al. utilized a variant of the V1 model originally proposed for the IQA task [6] to predict the visibility of alpha-blended images. In this model, the number of scales of the wavelet (including the low-pass residual) was five. For fair comparison, we increased the number of scales to six as in our proposed model and retrained all the parameters using our dataset. In [3], the linear gains ($S_{k,o}$) for each spatial frequency ($k = 1, 2, \ldots, N-1$) and orientation channels ($o \in \{\mathrm{V(vertical)}, \mathrm{D(diagonal)}, \mathrm{H(horizontal)}\}$) were parameterized so that the linear gain always decayed from low to high frequencies. We removed this parameterization and independently optimized S_k to better predict our dataset. However, as in [6], we shared S_k across the horizontal and vertical orientation channels and scaled them by a single parameter d to obtain the linear gains for the diagonal channel. The optimized parameters of the V1 model used in [3] are presented in Table 5.

Table 5. Optimized parameters of the IQA model used in [3]

Fold	1	2	3	4	5	All
S_1	5.5075	5.2982	4.2423	3.6480	5.5113	4.3217
S_2	0.5263	0.7151	0.8383	1.0610	0.4839	0.8280
S_3	99.030	93.608	93.628	104.31	95.759	104.05
S_4	54.146	58.901	58.934	56.759	56.232	53.698
S_5	9.1132	7.2666	6.8108	8.1499	9.2573	8.3935
d	0.6083	0.6153	0.6076	0.6019	0.6225	0.6004
ω	0.0000	0.0192	0.0270	0.0000	0.0040	0.0173
b	7.478	7.0710	6.7900	6.9753	7.0586	6.9106
γ	0.5172	0.5032	0.5166	0.5196	0.5100	0.5112
σ_e	0.2360	0.2371	0.2346	0.2350	0.2321	0.2108
σ_o	5.0414	5.1642	4.4308	5.3347	5.4218	5.8273
p	1.2296	1.2960	1.2614	1.2759	1.2073	1.2325
q	1.9148	1.9223	1.9621	1.9314	1.9555	1.9712
s	1.2070	1.2605	1.2294	1.2555	1.1487	1.1756

NLPD. The normalized Laplacian pyramid distance (NLPD) was originally proposed as an image quality metric [4] and then successfully applied to image rendering techniques such as tone mapping [5]. We used the implementation employed in [5] and set the number of scales of the Laplacian pyramid to six. Whereas the original model did not have any free parameters to weight each scale of the pyramid, we incorporated these parameters ($\{w_k\}_{k=1}^{N_k}$) to better predict our dataset. Accordingly, we modified the equation to aggregate NLP coefficients (i.e., Eq. 6 in [5]) as follows:

$$D(\mathbf{S}, \mathbf{I}) = \left[\frac{1}{N_k} \sum_{k=1}^{N_k} w_k \left(\frac{1}{N_c^{(k)}} \sum_{i=1}^{N_c} |y_i^{(k)} - \tilde{y}_i^{(k)}|^{\alpha} \right)^{\frac{\beta}{\alpha}} \right]^{\frac{1}{\beta}}$$

, where w_k was constrained to be positive. The optimized parameters of the NLPD are presented in Table 6 and Fig. 6.

Ablation model 1: Ours with static feature aggregation. In this model, the parameters $w_{n_{n=1}^{N^c N^{lv}}}$ in Eq. 9 in the main paper were directly optimized as static weights. These weights were constrained to be positive values during the optimization. The optimized parameters of Ablation model 1 are presented in Table. 7, and the optimized weight values are presented in Fig. 7.

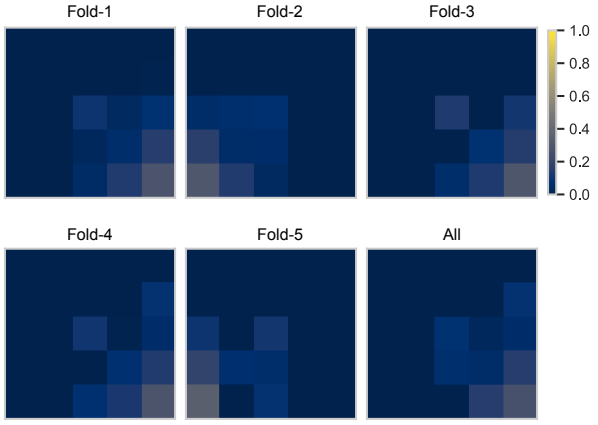

Fig. 6. Optimized local weighting filter $\mathbf{P} \in \mathbb{R}^{5\times5}$ of the normalized Laplacian pyramid distance (NLPD) [5].

Table 6. Optimized parameters of NLPD.

Fold	1	2	3	4	5	All
γ	0.2385	0.2354	0.2408	0.2353	0.2385	0.2367
$\sigma \times 10^{-3}$	8.1565	8.1143	8.4066	8.3360	8.1572	8.4661
σ_{lowpass}	2.1698	2.1754	2.3274	2.3150	2.1572	1.8064
α	2.1540	2.2427	2.1222	2.1799	2.2550	2.2059
β	1.0845	1.0626	1.0738	0.9896	1.0863	1.0251
w_1	0.4391	0.4308	0.4626	0.4092	0.4200	0.4365
w_2	0.4386	0.3517	0.4466	0.4315	0.3907	0.4802
w_3	0.2594	0.2923	0.2467	0.2942	0.3213	0.2621
w_4	0.3504	0.4034	0.3549	0.3546	0.3927	0.3476
w_5	0.4416	0.4935	0.4826	0.4226	0.4500	0.3973
w_6	0.9514	0.8494	0.9013	0.9379	0.9130	0.8867
s	1.7485	1.6755	1.7287	1.6711	1.6885	1.6677

Table 7. Optimized parameters of Ablation model 1.

Fold	1	2	3	4	5	All
$\beta' \times 10^{-4}$	6.1483	7.6589	5.4267	6.9342	5.5746	7.1506
γ	0.3680	0.3626	0.3677	0.3625	0.3610	0.3656
p	1.9714	1.8433	1.7888	1.8879	1.7606	1.7942
q	4.7730	4.7341	4.7646	4.6344	5.0226	4.9396
s	7.1293	6.6361	6.5211	6.7675	6.3947	6.4618

Ablation model 2: Ours with self-adaptive feature aggregation. The optimized parameters of Ablation model 2 are presented in Table. 8, and the interaction kernel $\mathbf{X}$ and bias $\mathbf{b}$ are presented in Fig. 8.

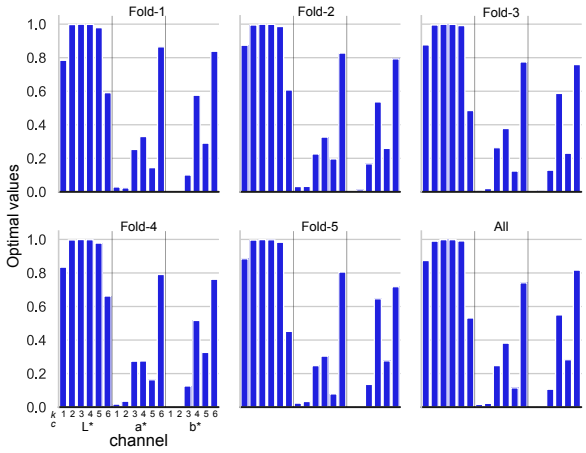

Fig. 7. Optimized weights $\mathbf{w}$ of Ablation model 1 (our model with static weights).

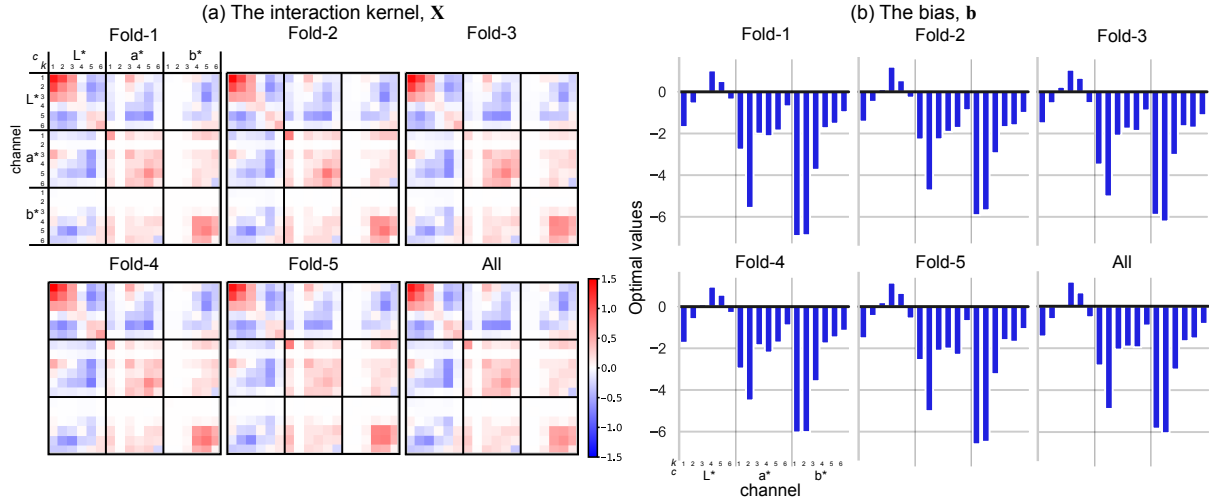

Fig. 8. (a) Optimized interaction kernel $\mathbf{X}$ and (b) bias $\mathbf{b}$ of Ablation model 2.

Table 8. Optimized parameters of Ablation model 2.

Fold	1	2	3	4	5	All
$\beta' \times 10^{-4}$	10.105	7.8414	10.665	6.8795	12.690	9.9986
γ	0.3784	0.3840	0.3781	0.3849	0.3756	0.3834
p	1.3472	1.2689	1.2488	1.2426	1.2592	1.2420
q	4.9178	4.8743	4.8924	4.8571	5.1407	5.0596
s	7.1269	6.4526	6.5327	6.5631	6.5515	6.3389

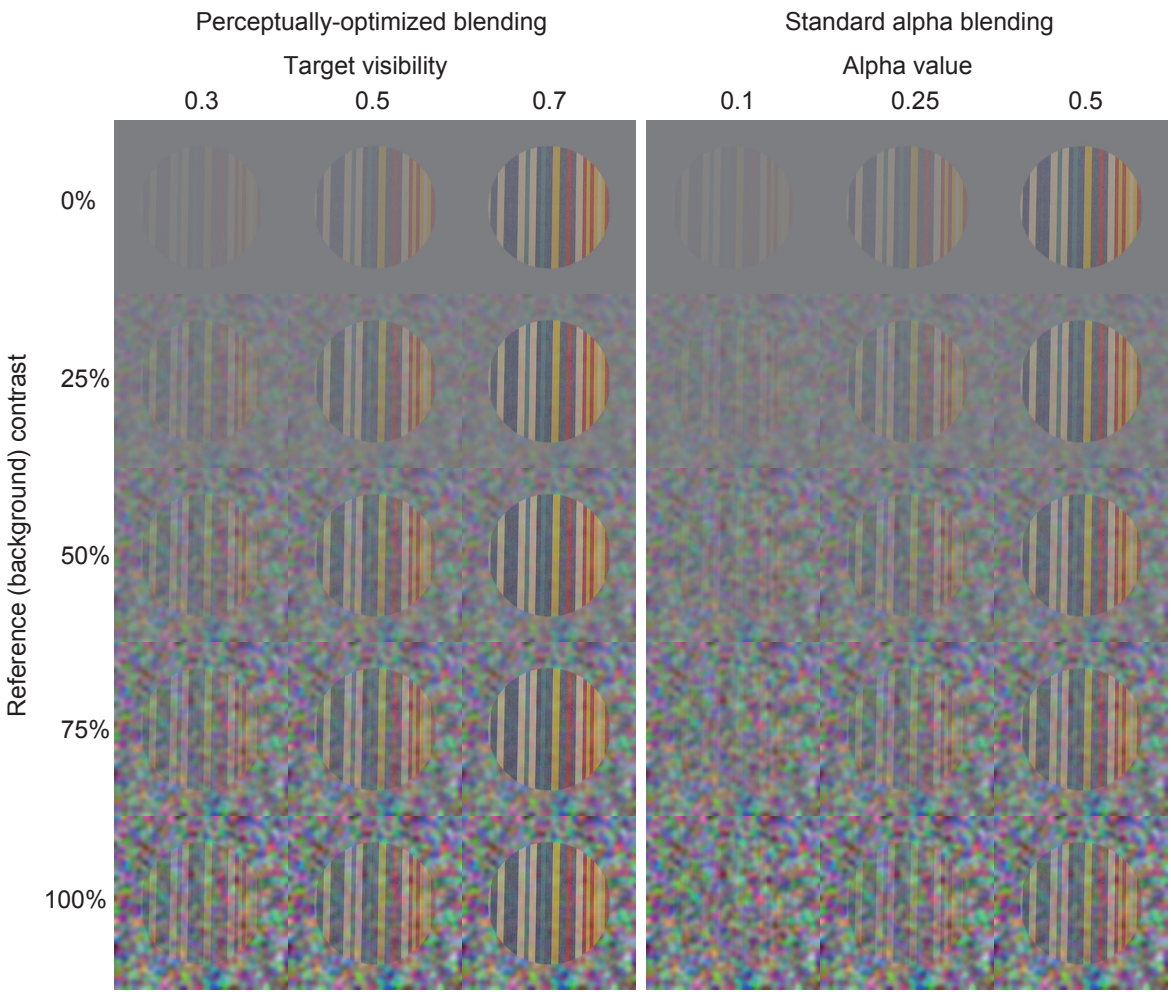

Fig. 9. Blending results with different target visibility levels and different reference (background) contrast levels. The results of our method are shown on the left side and those of the standard alpha blending are on the right side. The alpha values in the standard alpha blending are set to match those of our method under the 0% reference contrast condition (the first row).

5 ADDITIONAL VALIDATION OF THE PROPOSED BLENDING METHOD

In this section, we present additional validation results for our blending method. First, we validate our method using simple artificial stimuli. Then, we show additional results using natural images. Finally, we present an additional user study that complements the first user study presented in the main paper.

5.1 Validation with simple artificial images

Figure 9 (left) presents blending results with three different target visibility levels while varying contrast of the reference (background) noise pattern. The results using the standard alpha blending in which the fixed alpha

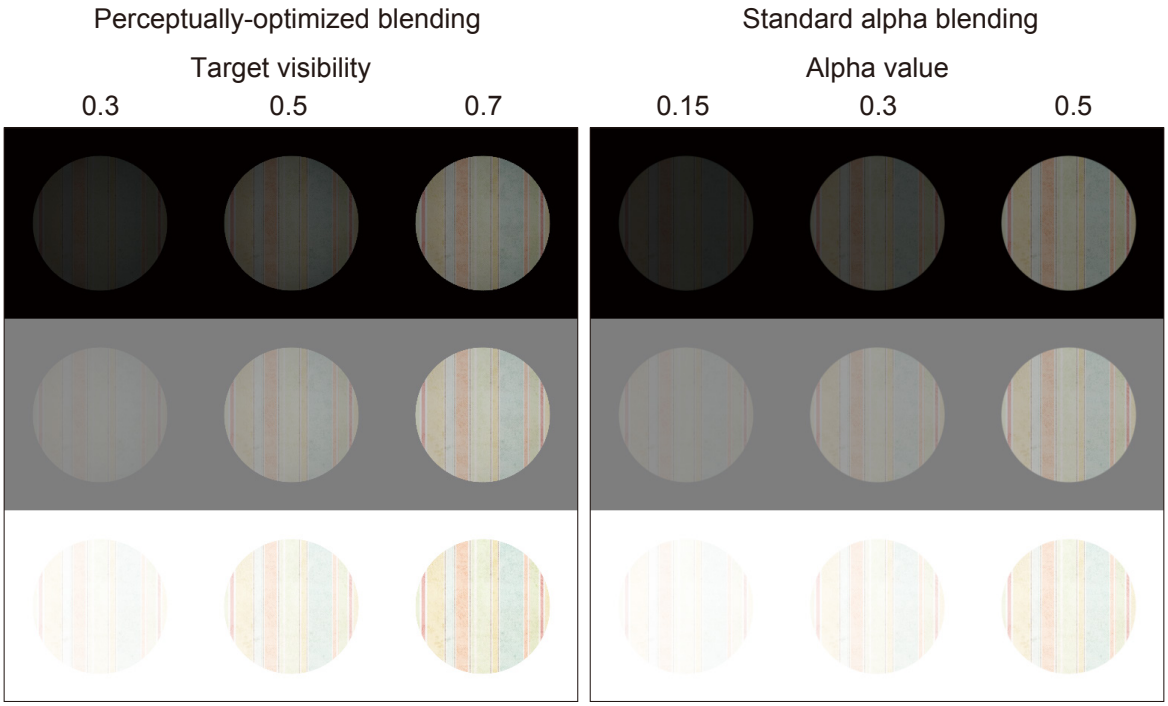

Fig. 10. Blending results with different target visibility levels and different reference (background) luminance levels. The results of our method are shown on the left side and those of the standard alpha blending are on the right side. The alpha values in the standard alpha blending are set to match those of our method under the 0% reference contrast condition (the first row).

value is used across different reference contrast are also shown on the right side of the same figure. The alpha value of the standard alpha blending in each column was set to match those of our method under the 0% reference contrast condition (the first row) in the corresponding column. We can observe that our results maintain an equal level of visibility across different levels of reference contrast while those from the standard alpha blending show significant visibility degradation as the reference contrast increases.

Our method can also compensate for variation in the perceived contrast caused by background luminance changes. Figure 10 presents blending results with three different target visibility levels under three different reference (background) luminance levels. Again, the alpha value of the standard alpha blending in each column was set to match that of our method under the 0% reference contrast condition (the first row) in the corresponding column. The results of standard alpha blending tend to show visibility degradation in the white reference conditions, while our method successfully compensates for this by increasing alpha values.

5.2 Validation with natural images

Here, we show additional blending results using natural images as inputs. Figure 11 presents blending results obtained with four target visibility values for two different reference images (the first and second rows). For comparison, we show blend results obtained by standard alpha blending using a uniform alpha map with four

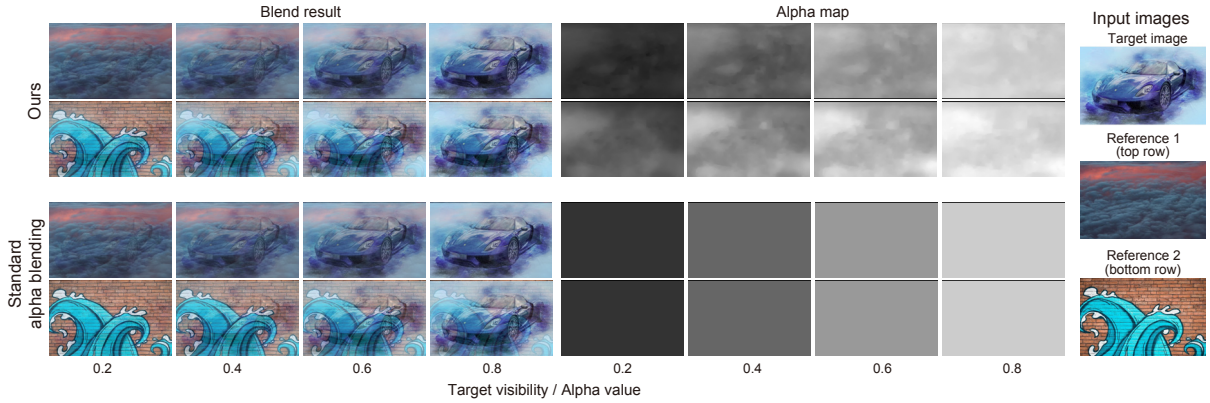

Fig. 11. Comparison of results generated by our method (top) and with standard alpha blending (bottom). The grayscale images on the righthand side show the corresponding optimized alpha map. For each method, the results are generated using the two different reference images and the four different target visibility or alpha values. Input images are by ArtTower / Pixabay, avi_acl / Pixabay, and MichaelGaida / Pixabay in order from top to bottom, respectively.

different alpha values (the third and fourth rows). It can be seen that the visibility of the target image (a sports car) is consistent between the two different reference images when they are generated with the same target visibility. On the other hand, the standard alpha blending results in a significant loss of visibility in the images in the bottom row, even though the same alpha values are used.

To see the dependence of the optimized alpha maps on changes in reference contrast, we also generated results while gradually varying the contrast of the input reference image (Fig. 12). Here, we fixed the target visibility value to 0.5. The alpha values optimized by our method increased as the reference contrast increased (the third column). As a result, the target image (a woman's face) in the blended images showed a consistent visibility regardless of the reference contrast (the second column). On the other hand, the visibility gradually decreased as the reference contrast increased when the fixed alpha value (0.5) was used (the rightmost column).

Figure 13 presents comparisons of the blending results obtained with different methods. Here, all the results are generated to achieve the same target visibility level (a normalized target visibility value of 0.5). The first row presents the results of our proposed method. The second row presents the results of [3] in which the IQA model by [6] is used to optimize alpha values. For fair comparison, we retrained the IQA model on our dataset. In the third row, our model with static feature aggregation (i.e., Ablated model 1) is used to optimize an alpha map. In the fourth row, the standard alpha blend using a uniform alpha map with fixed alpha value is used to obtain the results. When the target content is the same across images, the static weight model can generate results with consistent visibility under the same target visibility setting (see the first and second columns of the third row. However, when compared across images with different target content, the alpha values optimized by this ablation method tended to be too high for the texture-less target image while they tended to be too low for the textured target image, as illustrated in the third and fourth columns of the third row. This can be attributed to the model's inability to adapt feature weights to the original target image: the model always underestimated the visibility when the image lacked high-frequency information while it overestimated the visibility when the image had abundant features. In contrast, our method was able to appropriately increase the weights for low frequencies in the smooth regions while suppressing overall weights in the textured regions (top row).

We found that the IQA-based method (the second row) shared the same problem as the static feature aggregation method: excessive alpha values in smooth foreground regions (see the third and fourth columns and corresponding

Fig. 12. Dependence of the optimized alpha map on reference image contrast. A target image (the woman's face) was blended on a reference image while varying the contrast of input reference images (the leftmost column). The second and third columns show the blend results and optimized alpha maps obtained with our method using a target visibility value of 0.5, respectively. The rightmost column shows the results of standard alpha blending using an alpha value of 0.5. The target and reference images are by Cyber Shaman / FFHQ dataset and Free-Photos / Pixabay and respectively.

alpha maps). Moreover, the smoothing step in the postprocessing could not handle the local variation of the alpha map; this step often led to over- or undersmoothing. This is most notable in the result of the woman's face

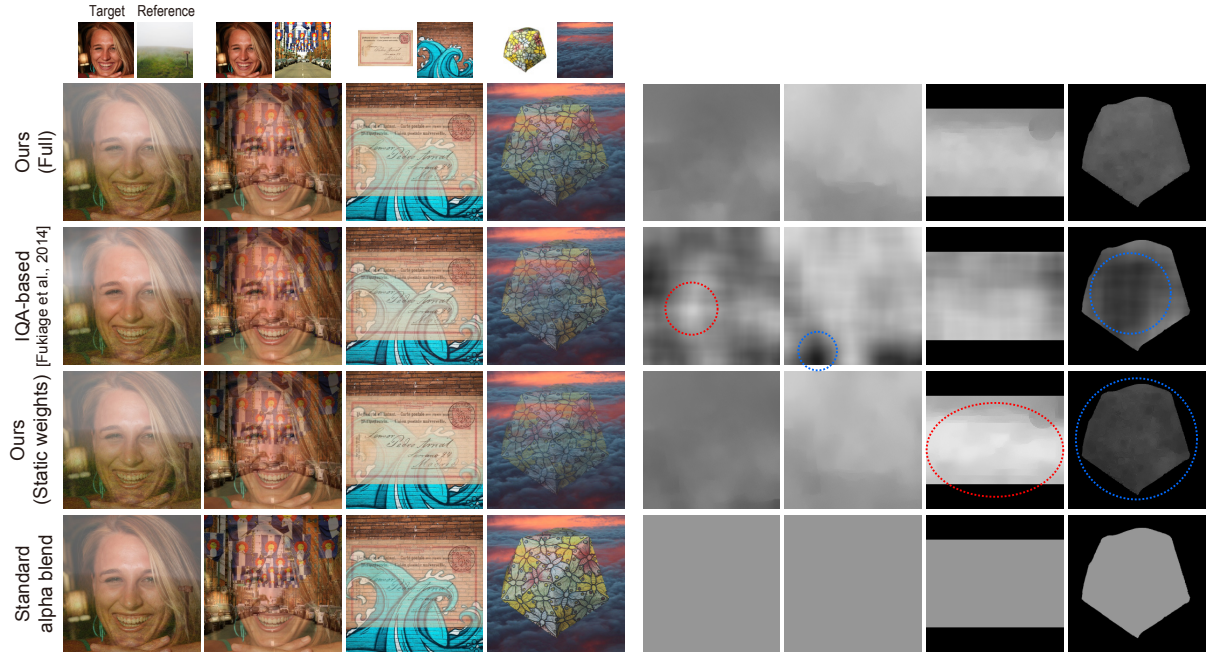

Fig. 13. Results obtained by different methods. For each method, we present the results generated using four different input image pairs displayed in the top. The grayscale images on the right side display the corresponding optimized alpha map. All the images were generated with a normalized target visibility value of 0.5 (see Section 6.4 in the main paper about the normalization). The image areas where alpha values are deemed overemphasized or underemphasized due to the lack of the content-adaptive mechanism are enclosed by red or blue broken lines, respectively. Input images are by Cyber Shaman / FFHQ dataset, Free-Photos / Pixabay, islandworks / Pixabay, 2211438 / Pixabay, MichaelGaida / Pixabay, ALOI database, and avi_acl / Pixabay in order from left to right, respectively.

in the first column. In contrast, our method addressed this problem by considering both the visibility and image structures of the target image in the optimization process.

5.3 Additional user study: Pairwise comparison

The user study presented in the main paper employed the visibility rating task in which the participants evaluated the absolute visibility level for each comparison method and demonstrated that our method could achieve the best performance in terms of visibility control. To further show the superiority of our method over the others more directly, we conducted an additional user study using a pairwise comparison task. In this task, the participants directly compared pairs of blended images generated by two different methods and chose the pair that exhibited more consistent visibility.

Method. The participants were instructed to select an image pair that exhibited more consistent visibility from two blended image pairs generated by different methods. The participants could switch between the image pairs by pressing a button. A short blank period (0.5 s) was inserted between the switching. The original target images were presented for reference above the test stimuli. Each image pair was generated by blending (1) the Same Target image on Different Reference images (ST-DR condition), (2) Different Target images on the Same Reference image (DT-SR condition), or (3) Different Target images on Different Reference images (DT-DR condition). For

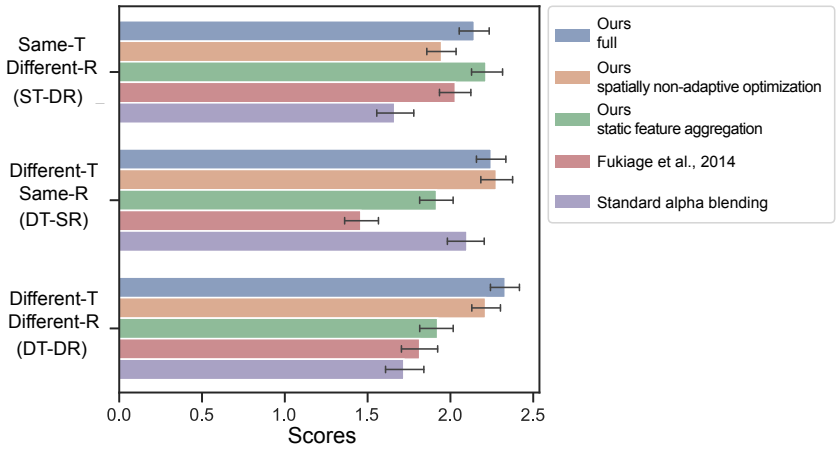

Fig. 14. Average scores for each method obtained in user study 2. The error bars represent ±95% confidence intervals.

each condition, 100 different target-reference combinations were randomly sampled. For each image combination, all possible pairs of comparison methods were compared. The visibility levels of the stimuli were randomly sampled; however, they were approximately matched between compared pairs because the participants would tend to choose the more visible pair as "consistent." This was achieved by choosing the target visibility values of compared methods to give equal visibility scores according to the nonlinear function fitted to the data of the user study in the main paper (gray curves in Fig. 18). The visibility level of each of the comparison pairs was randomly sampled from 2 (Barely visible) to 4 (Fairly visible) with a step size of 0.5. The number of trials was 3 conditions × 100 image combinations × 10 pairs of comparison methods = 3,000. In total, 174 participants were recruited via a crowdsourcing service (Prolific.ac). Each participant performed 100 comparisons in addition to 10 practice trials and 10 sentinel trials used for screening.

In the sentinel trials, the visibility inconsistency was conspicuously large in one of the comparison image pairs. Participants who selected the inconsistent pair in sentinel trials more than twice were excluded from analysis. 30 participants were removed by this procedure. After the screening, an average of 4.8 responses was collected for each of the 3,000 comparisons. We computed the average number of times each method was preferred, and used the number as a score.

Results. Figure 14 presents the scores for each method. A two-way ANOVA (five comparison methods × three conditions) revealed a significant main effect of the comparison methods ($F(4, 7185) = 47.84, p < 10^{-39}$). We also observed a significant interaction between the methods and conditions ($F(8, 7185) = 20.11, p < 10^{-29}$). Thus, we performed a multiple comparison test (Tukey's HSD) over the scores of five methods separately for each condition. For the *ST-DR* condition, all the methods showed significantly better scores than the standard alpha blend ($p \leq 0.001$). The score of our full method was significantly higher than that for ours with the spatially non-adaptive method ($p \leq 0.05$). However, the differences between our full method and the other two methods were not significant. This is because the advantage of the content-adaptive mechanism should disappear when the content is the same.

When the target's content was different, our full method and ours with the spatially-non adaptive method (both methods utilized the content-adaptive mechanism) showed significantly higher scores than ours with static weights and the IQA-based method [3] ($p \leq 0.001$) as seen in the *DT-SR* and *DT-DR* conditions. On the other hand, the difference between our method and the standard blend was not significant in *DT-SR*. We think that this is

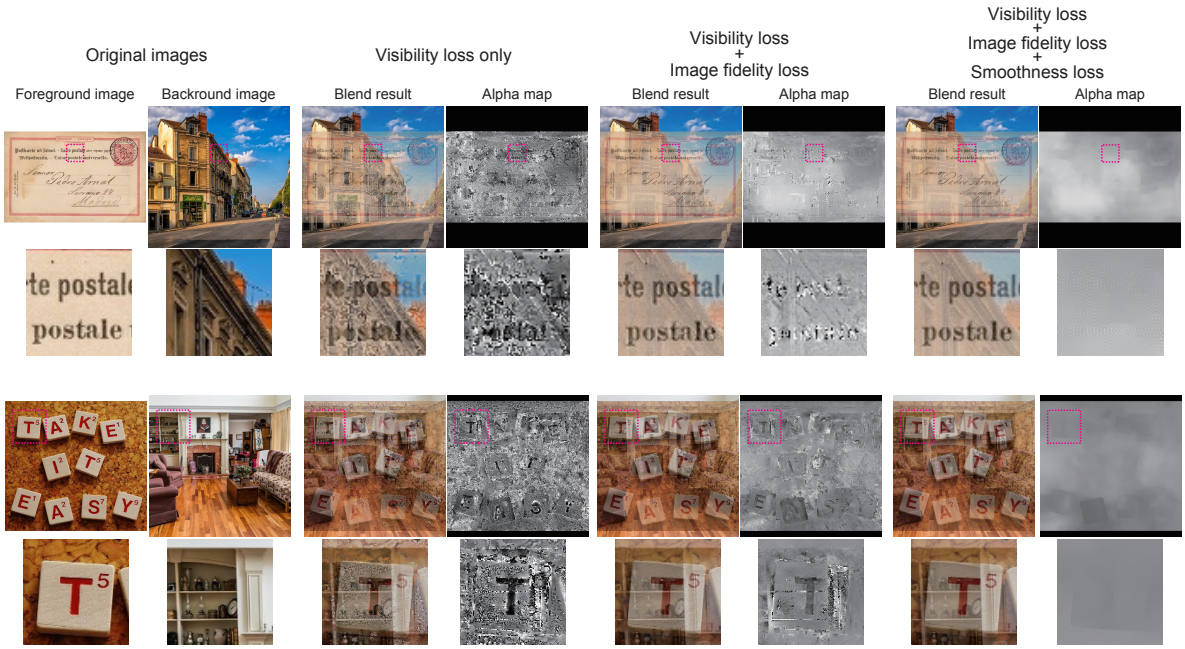

Fig. 15. Results of the ablation study of loss functions. The second and fourth rows show enlarged images of the areas enclosed by dashed lines in the first and third rows, respectively. (Second column) Results obtained using the visibility loss only. Without the image fidelity loss, patterns that do not exist in the original image are produced to achieve the target visibility. As a result, the visual quality of the generated images significantly degrades. (Third column) Results obtained using the visibility loss and image fidelity loss. The image fidelity loss ensures that the results retain the original image structure. However, in some cases, the gradient-based optimization can still be trapped in suboptimal local minima. (Fourth column) When the edge-preserving smoothness loss is combined, the aforementioned problems are resolved, and visually pleasing results are obtained. Input images are by 2211438 / Pixabay, Tama66 / Pixabay, Alexas_Fotos / Pixabay, and Mariamichelle / Pixabay

because the effect of contrast masking was almost equivalent within the image pair when the reference image was the same.

6 EFFECT OF LOSSES USED IN PERCEPTUALLY OPTIMIZED IMAGE BLENDING

To evaluate the effect of the loss functions on our proposed perceptually optimized image blending method, we disabled the edge-preserving smoothness loss and both the edge-preserving smoothness loss and image fidelity loss to generate the blending results. The results were then compared with those obtained by our original method using all loss functions. Details are provided in the caption of Fig. 15).

REFERENCES

[1] Martin Cadík, Robert Herzog, Rafał Mantiuk, Radosław Mantiuk, Karol Myszkowski, and Hans-Peter Seidel. 2013. Learning to Predict Localized Distortions in Rendered Images. *Computer Graphics Forum* 32, 7 (2013), 401–410.

[2] John M Foley and Geoffrey M Boynton. 1994. New model of human luminance pattern vision mechanisms: analysis of the effects of pattern orientation, spatial phase and temporal frequency. In *Computational Vision Based on Neurobiology*, Vol. 2054. International Society for Optics and Photonics, 32–42.

[3] Taiki Fukiage, Takeshi Oishi, and Katsushi Ikeuchi. 2014. Visibility-based blending for real-time applications. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 63–72.
[4] Valero Laparra, Johannes Ballé, Alexander Berardino, and Eero P Simoncelli. 2016. Perceptual image quality assessment using a normalized Laplacian pyramid. *Electronic Imaging* 2016, 16 (Feb. 2016), 1–6.
[5] Valero Laparra, Alexander Berardino, Johannes Ballé, and Eero P Simoncelli. 2017. Perceptually optimized image rendering. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* 34, 9 (Sept. 2017), 1511–1525.
[6] Valero Laparra, Jordi Muñoz-Marí, and Jesús Malo. 2010. Divisive normalization image quality metric revisited. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* 27, 4 (April 2010), 852–864.
[7] Rafał Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. 2011. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on Graphics* 30, 4 (July 2011).
[8] Andriana Olmos and Frederick A A Kingdom. 2004. A biologically inspired algorithm for the recovery of shading and reflectance images. *Perception* 33, 12 (2004), 1463–1473.
[9] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2004. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003.* 1398–1402.
[10] Krzysztof Wolski, Daniele Giunchi, Nanyang Ye, Piotr Didyk, Karol Myszkowski, Radosław Mantiuk, Hans-Peter Seidel, Anthony Steed, and Rafał K Mantiuk. 2018. Dataset and metrics for predicting local visible differences. *ACM Trans. Graph.* 36, 4 (2018), 14.