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

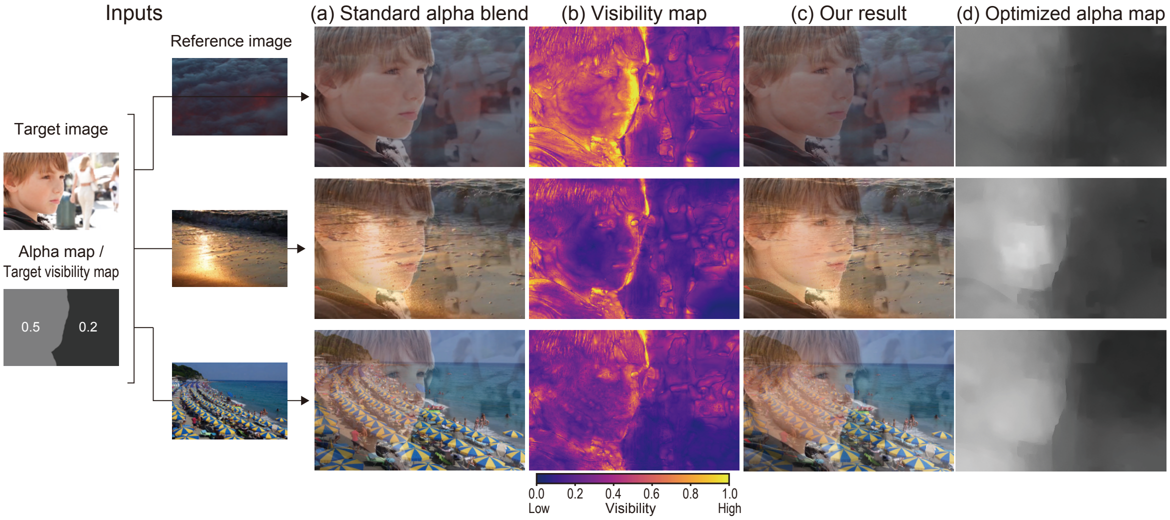

Fig. 1. (a) Results of standard alpha blending. The visibility of the target image is significantly degraded even though the same alpha map is used. (b) Visibility maps of the blend results in (a) predicted by our model. The visibility values are rescaled to [0,1] for visualization as described in Section 6.4. (c) Results optimized by our method. We optimize the alpha map based on the visibility map so that the target image in the blended result achieves the desired visibility level specified in the input target visibility map. (d) The alpha maps obtained by optimization. Input images are by Roxolana / Pixabay, outlinez, and avi_acl / Pixabay.

The visibility of an image semi-transparently overlaid on another image varies significantly depending on the content of the images. This makes it difficult to maintain the desired visibility level when the image content changes. To tackle this problem, we developed a perceptual model to predict the visibility of the blended results of arbitrarily combined images. Conventional visibility models cannot reflect the dependence of the suprathreshold visibility of the blended images on the appearance of the pre-blended image content. Therefore, we have proposed a visibility model with a content-adaptive feature aggregation mechanism, which integrates the visibility for each image feature (i.e., such as spatial frequency and colors) after applying weights that are adaptively determined according to the appearance of the input image. We conducted a large-scale psychophysical experiment to develop the visibility predictor model. Ablation studies revealed the importance

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.

of the adaptive weighting mechanism in accurately predicting the visibility of blended images. We have also proposed a technique for optimizing the image opacity such that users can set the visibility of the target image to an arbitrary level. Our evaluation revealed that the proposed perceptually optimized image blending was effective under practical conditions.

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. A Content-Adaptive Visibility Predictor for Perceptually Optimized Image Blending. *ACM Trans. Appl. Percept.* 1, 1 (October 2022), 30 pages. https://doi.org/10.1145/3565972

1 INTRODUCTION

Alpha blending [41] is a common technique used in rendering semi-transparent objects that is applied in many fields. For example, artists often blend two photos to create impressive artistic effects. Graphical user interfaces also use semi-transparent effects to prevent occlusion of windows. Image blending can also be used to approximate physical effects such as reflection, fire, and smoke, to composite the desired scene.

Although users can adjust the alpha value to achieve the desired level of visibility (i.e., apparent image contrast), this value is not necessarily meaningful. The visibility of a target object or scene, varies significantly depending on the reference scene it is blended with when the same alpha value is used (see the results of standard alpha blending in Fig. 1). The visibility can vary from area to area in an image and adjusting a single alpha value may not achieve the desired visibility over the entire image. Even if one could manage to create an alpha map that achieved the desired visibility, it would be difficult to create one manually for every frame in video editing.

A promising solution to overcome this challenge is to incorporate a perceptual model of visibility to optimize the alpha map. The dependence of visibility on the reference image can be explained by the mechanism of contrast perception in the human visual system (HVS), which has been explored in psychophysical studies [4, 14, 55]. Contrast perception models have also been applied to image quality assessment (IQA) to measure noise visibility in distorted images. In line with this concept, [16] proposed a blending method that optimizes an alpha map to achieve the targeted visibility when rendering virtual objects in an AR display.

However, an IQA model for noise visibility is insufficient for accurately predicting the visibility of blended images. Most existing IQA models weight each image feature (e.g., spatial frequency and colors) in a content non-adaptive manner. They use fixed weights for aggregating feature-wise responses into a single value representing the image quality degradation. These weight values are usually determined by fitting the model to subjectively evaluated quality scores on a diverse image dataset. However, our empirical finding is that the predetermined weights are not effective because the importance of each feature depends on the original features contained in the pre-blended target image (Fig. 2).

We therefore developed a visibility predictor model with a content-adaptive feature aggregation mechanism which integrates the visibility for each image feature after applying weights that are adaptively determined according to the original appearance of the target image. Through ablation studies, we demonstrate that this dynamic weighting mechanism is important for accurately predicting the visibility of arbitrary images. We also conducted a large-scale psychophysical experiment through crowdsourcing to calibrate the model for a wide variety of images. Moreover, we developed a perceptually optimized image blending method using the proposed visibility model (Fig. 15). The method optimizes an alpha map while regularizing the solution such that the original image structure is retained. We demonstrate that our proposed method can control the visibility of alpha-blended images using various types of image content, including photos and videos.

The main contributions of our work can be summarized as follows:

- A visibility model that incorporates a content-adaptive feature aggregation mechanism was proposed to predict the visibility of alpha-blended images.

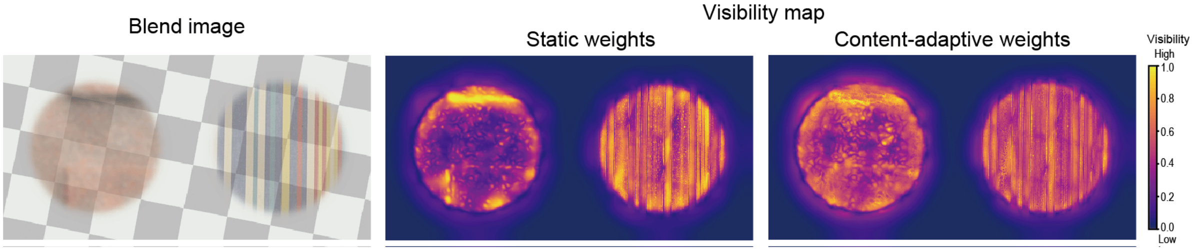

Fig. 2. Comparison of supra-threshold visibility predicted with static feature weights model (middle) and with content-adaptive weights model (right). The two different texture patches are blended with the checker background. In the static model, the visibility of the smooth texture patch is underestimated and that of the stripe texture patch is overestimated. In the content-adaptive model, this problem is alleviated by adaptively weighting each image feature based on the original appearance of the textures. The visibility values of each method are rescaled to $[0, 1]$ for visualization as described in Section 6.4.

- A large-scale psychophysical experiment was conducted to collect subjective data on the visibility of a diverse range of alpha-blended images.
- Perceptually optimized blending techniques were developed based on the proposed visibility model that enable users to intuitively control the visibility of blended images.

In the following section, we first discuss previous work that is related to the content of our paper. Then, in Section 3, we describe the details of the proposed content-adaptive visibility predictor. In Section 4, we provide detailed description of the large-scale psychophysical experiment in which we collected data on the visibility of blended images generated from a variety of image patches, and explain how we optimize the model parameters. In Section 5, we validate our model by comparing it with several previous models. Furthermore, in Section 6, we propose a perceptually-based image blending in which we optimize an alpha map to achieve an arbitrary visibility level based on the proposed model. We demonstrate the effectiveness of the proposed technique through a user study. Finally, in Section 7, we conclude the paper with a discussion of possible avenues for future work.

2 RELATED WORK

Image visibility can be mostly explained from the perspective of our contrast perception. For this reason, we first review the characteristics of contrast perception that have already been identified in vision science literature. We then describe visibility predictor models developed for the IQA problem and briefly explain their relationship to our model. Finally, we discuss other perceptually-motivated image composition techniques and clarify our research position.

2.1 Contrast perception

2.1.1 Contrast sensitivity function. One notable aspect of contrast perception is the spatial frequency dependence of the contrast detection threshold, which has been characterized as the contrast sensitivity function (CSF) [4, 59]. As the retinal signals are known to be mediated by color opponent channels each tuned to the contrast defined by light/dark (i.e., luminance contrast), red/green, or blue/yellow differences, the CSF has also been investigated for each of these opponent color patterns. A typical CSF for luminance-defined patterns is the band-pass, with its peak at approximately 2–5 cycles per degree. The shape of the CSF for the other two color opponent channels (i.e., red/green and blue/yellow) is similar to that of a low-pass filter, and the reduction in sensitivity with increasing spatial frequency is more rapid than that for the luminance channel [36].

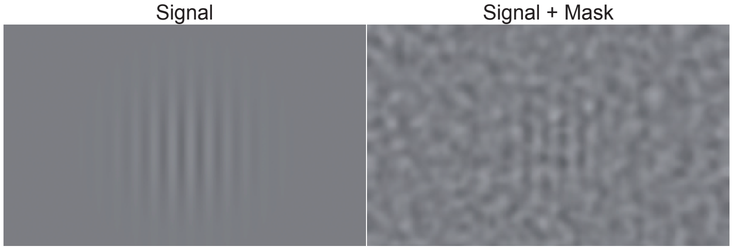

Fig. 3. Degradation of visibility due to a textured background can be explained by "contrast masking," where the perceived contrast of a signal (vertical sinusoidal wave) is degraded when embedded in the mask pattern.

However, it is also known that the role of the CSF is limited in suprathreshold contrast perception. Several studies have demonstrated that the size of perceived contrast is almost constant across spatial frequencies when the stimuli are drawn in high contrast [2, 8, 18]. Therefore, the CSF is not explicitly hard-coded in our visibility model because the main focus of our model is to predict the suprathreshold visibility of blended images. Instead, we allow the weighting scheme for each spatial frequency to be acquired through training on experimental data.

2.1.2 Contrast masking. The characteristics of contrast perception most relevant to our work can be observed in the phenomenon known as contrast masking [14]. A demonstration of contrast masking is displayed in Fig. 3. In the righthand image, the visibility of the sine wave is reduced by the presence of background noise. Qualitatively similar masking phenomena can also be observed when the stimulus is composed of isoluminant color gratings [34]. The mechanism of the nonlinear contrast response function of the HVS is divisive normalization [22], where the response of a visual neuron is divisively normalized by the weighted sum of responses from nearby neurons. This model can explain a vast variety of data, including physiologically measured neural responses and psychophysically measured contrast masking data [22, 46, 48, 55].

We incorporate divisive normalization in our model to account for the dependence of visibility on reference images.

2.1.3 Application of contrast perception models. Contrast perception models have been applied to a wide variety of applications. These include image compression [8], tone mapping [30, 32], adaptive resource allocation for image rendering [42, 50], adaptive radiometric compensation for projection AR [15, 20], watermarking [58], image contrast enhancement [47], and so on. Among these applications, the closest to our work is the watermarking technique, which focuses on the detection threshold of embedded patterns. In contrast though, our work is more concerned with the suprathreshold visibility of blended images.

Image quality assessment (IQA) is another major application domain of the contrast perception models and is discussed in detail in the next subsection.

2.2 Visibility predictor for image quality assessment (IQA)

Our proposed visibility predictor is closely related to the models that have been developed for IQA. In this section, we discuss the relationship between ours and those models.

2.2.1 Image quality metrics. Typical IQA models predict image quality by comparing a corresponding pair of distorted and clean images in the perceptual space [29, 31, 48]. In most cases, the perceptual representation is modeled by band-pass decomposition followed by the divisive normalization process, although some recent works utilized deep features learned by a convolutional neural network [10, 64]. Alternatively, noise components can be extracted earlier in the processing and modulated by simulating the inhibitory mechanisms in the HVS [9, 33, 35]. Another approach to measuring image quality is the structural similarity metric (SSIM [52]) and its multiscale extension (MS-SSIM [54]). Although the SSIM is one of the most widely used image quality metrics due to its algorithmic simplicity, it does not have direct links to physiological mechanisms in the HVS. In this work, we use the blended image as an inhibitory component to modulate the contrast of the target image embedded within it.

2.2.2 Feature aggregation in IQA models. Many IQA models provide a multidimensional feature vector that represents the perceptual contrast difference between the distorted and clean images for each spatial frequency and color channel. Then, this channel-wise visibility information is aggregated into a single value representing the image quality degradation. A popular approach is to integrate it using Minkowski pooling (or L_p norm) [29, 31, 35]. To incorporate the importance of each feature into the IQA task, a channel-wise weighting function can also be applied [35, 54]. We use a weighting function; however, instead of using fixed weights, we dynamically adjust the weights based on the original appearance of the target image content. Some IQA models additionally consider spatially varying weights to aggregate local noise visibility into a single quality score [53, 63]. We, on the other hand, do not incorporate elaborate spatial aggregation mechanisms into our model as our main objective is to predict the image visibility for each location.

2.3 Perception-based image composition

In this work, we aim to apply our visibility model to perception-based image composition. Here, we describe several previous works on this topic and discuss their relationship to our own. The saliency model is one of the most successful perceptual models to have been applied to a range of image rendering problems. Grundland et al. [21] proposed using a saliency map to retain the opacity of important regions in an image when blending multiple images. A similar approach has been used for X-ray visualization in AR [24, 44, 45]. These previous approaches are complementary to our work, as their main focus is to analyze the regional importance of the input images while our aim is to control the visibility of a semi-transparently rendered image regardless of the importance of each region within it. Furthermore, the objective of the saliency models is to model visual attention based on the interaction of visual features across spatial areas. Thus, unlike our model, the saliency models cannot properly account for the effect of one image on another image semi-transparently rendered in the same area as we will demonstrate in Section 5.3.

In the context of visualization of more complicated layered structures, techniques have been proposed to optimize rendering parameters such as the layer opacity and color based on the known properties of transparency perception [5, 6, 51, 66]. In contrast to our work, which aims to control the visibility of blended images, the main focus of these techniques is to facilitate segregation and/or ordering of different layers.

Images displayed on typical optical see-through (OST) displays are inevitably blended with incoming light from the background real scene and are thus perceived semi-transparently. Zhang and Murdoch [62] developed a model for the perceived transparency of images presented on OST displays. Gabbard et al. [17] studied the legibility of text presented on an AR display and explored a text drawing style that could mitigate the issue. Zhang et al. [65] proposed a technique to enhance the visibility of a virtual image on OST displays by increasing the chromatic contrast against the background real scene. In the above techniques, however, the properties of the HVS on contrast perception, such as contrast masking, are not fully exploited.

On the other hand, Fukiage et al. [16] optimized an alpha map of a virtual object in AR scenes based on the contrast perception model. They directly applied a model originally developed for the IQA problem [31] and computed the visibility of a semi-transparently rendered object by taking the perceptual distance between the blended image and the background (reference) image. However, this causes unintuitive results when the object is almost opaque: Even though the background content behind the rendered object is invisible, the background

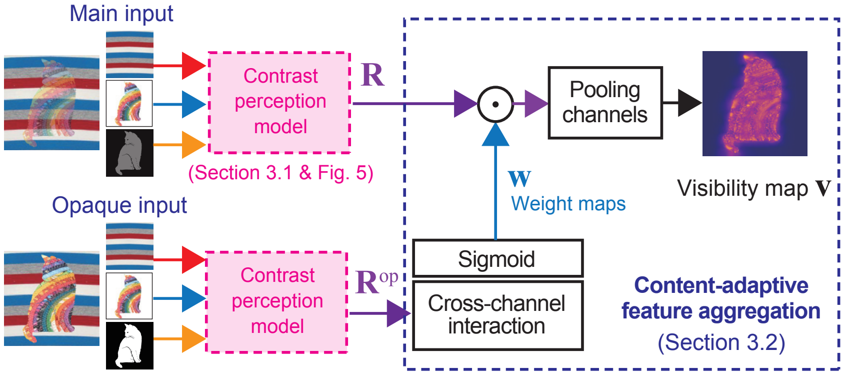

Fig. 4. Overview of proposed visibility predictor model. The weights for the response maps **R** for the target components in the main input are dynamically determined based on the response maps $\mathbf{R}^{\mathrm{op}}$ for the opaque version of the input. For the opaque input, a binary mask image that defines the location of the target image is optionally required if the target image occupies only a portion of the reference image area.

content strongly affects the visibility estimates. In addition, the model was tested using only a limited number of stimuli, and thus its generalizability to a large variety of images remains unknown. In this work, instead of directly applying a model developed for the IQA problem, we propose a custom visibility model specifically designed for blended images to address these problems.

3 CONTENT-ADAPTIVE VISIBILITY PREDICTOR

In this section, we describe our content-adaptive visibility predictor model Ω for blended images. The overview of the entire process is presented in Fig. 4. Our model consists of two main parts: one is based on a psychophysically established model of contrast perception, and the other is a content-adaptive feature aggregation mechanism. As inputs to the model, we assume a target image I^{tg}, a reference image I^{rf}, and an alpha map $A = \{\alpha_i \in [0,1]\}(i = 1, 2, ..., N)$, where N is the number of pixels. A binary mask image that defines the location of the target image is optionally required if the target image occupies only a portion of the reference image area. The objective of the model is to predict the suprathreshold visibility map of the target image $V^{\mathrm{tg}} = \{v_i | v_i > 0\}(i = 1, 2, ..., N)$ blended with the reference image by the alpha-blending equation:

$$V^{\mathrm{tg}} = \Omega(I^{\mathrm{tg}}, I^{\mathrm{rf}}, A), \tag{1}$$

$$I^{\mathrm{bl}}_c = A \odot I^{\mathrm{tg}}_c + (1 - A) \odot I^{\mathrm{rf}}_c, \tag{2}$$

where I^{bl} is the blended image. Blending is performed for each color channel c of the input images.

The input images are first processed using the custom contrast perception model to give a response map **R** for the target image, consisting of feature vector elements **r**: $\mathbf{R} = [\mathbf{r}_1 \mid \mathbf{r}_2 \mid \ldots \mid \mathbf{r}_N] \in \mathbb{R}^{N^{\mathrm{c}} N^{\mathrm{lv}} \times N}$, where N^{c} is the number of color channels, and N^{lv} is the number of Laplacian pyramid levels, as described in Section 3.1.1. (We align the image sizes of the different pyramid levels by upscaling.) The response vector **r** has $N^{c} \times N^{\mathrm{lv}}$ elements: $\mathbf{r} = \{R\} \in \mathbb{R}^{N^{\mathrm{c}} N^{\mathrm{lv}}}$.

The response vector is then scaled by a weight map $\mathbf{W} \in \mathbb{R}^{N^{\mathrm{c}} N^{\mathrm{lv}} \times N}$ and aggregated to give the visibility map V^{tg}. What characterizes the proposed model is that the weights are determined according to the original

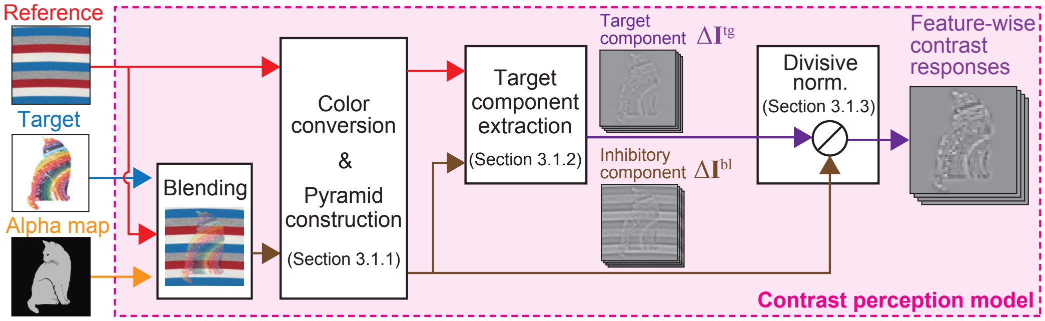
Reference
Target
Alpha map
Blending
Color conversion & Pyramid construction (Section 3.1.1)
Target component extraction (Section 3.1.2)
Target component $\Delta\mathbf{I}^{\mathrm{tg}}$
Inhibitory component $\Delta\mathbf{I}^{\mathrm{bl}}$
Divisive norm. (Section 3.1.3)
Feature-wise contrast responses
Contrast perception model

Fig. 5. Flow chart of proposed model for calculating the perceived contrast of a target image blended with a reference image using an alpha map. The contrast perception model outputs multidimensional contrast response maps, each representing a different spatial frequency and color channel.

appearance of the target image. Therefore, the weight map is determined based on the response map $\mathbf{R}^{\mathrm{op}}$ that is computed by the same contrast perception model using the binary mask image instead of the alpha map.

3.1 Contrast perception model

Here, we describe the process for computing contrast responses for each feature of the target image. The computational flow is presented in Fig. 5. In the first step, we generate a blended image from the two input images and the alpha map. The reference and the blended images then undergo color conversion and Laplacian pyramid decomposition, respectively, to obtain a reference component and a blended image component consisting of multiscale bandpass images for each color channel of the respective images (Section 3.1.1). We then extract the target component, i.e., the image features that belong to the target image in the blended result, as the residual of subtracting the reference image component from the blended image component (Section 3.1.2). Finally, we simulate the contrast masking effect to obtain the contrast responses $\mathbf{R}$ by divisively normalizing the target component using the blended image component as an inhibitory component. By using the blended image rather than the reference image for the inhibitory component, we can incorporate self-masking from the target image in addition to the masking from the reference image.

It is noted that the target component extraction process is a unique element of our model, along with the content-adaptive feature aggregation mechanism described in Section 3.2. Unlike the previous work by Fukiage et al. [16], here we extract the target image component in the blended image early on, before the divisive normalization process. The IQA model used in that work first computes the contrast responses for the blended image and those for the reference image and then obtains the contrast responses for the target image by taking the differences between them. Although the above approach appears straightforward, it presented a problem in that even if the target image is almost opaque (and therefore the reference almost invisible), the content of the reference image significantly influences the visibility of the target. We addressed this issue by explicitly extracting the target component relevant to the visibility estimation in the target component extraction process described in Section 3.1.2.

3.1.1 *Image feature representation.*

We assume that the input images are encoded in the sRGB color space. To simulate the color opponency of the HVS, we convert them into the CIE L*a*b* color space. This color space also takes into account the nonlinearity of the human visual system and is designed to achieve perceptual

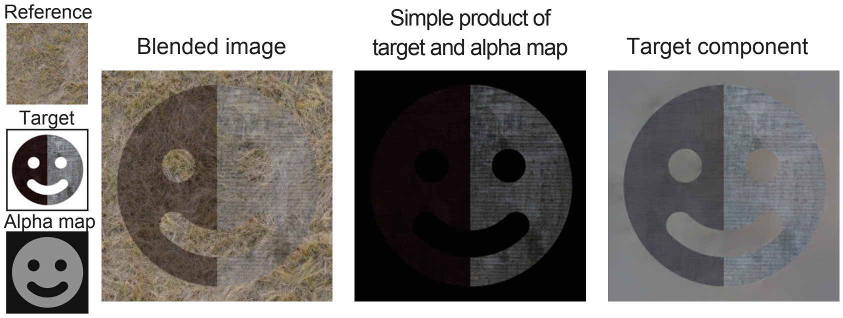

Fig. 6. The importance of the target component extraction process. (Left) In the blended image, image intensity variation for which we want to compute the visibility (target component) is often defined between the target and reference colors due to spatial variation in the alpha map. For example, the right half of the target's face appears to be obscured because the target and reference have similar colors. (Center) The simple product of the input target image and alpha map does not serve as the target component in the blended image because it cannot capture the interaction between the target and reference colors. In this case, the right half of the target's face is high contrast and the left half is low contrast, which is opposite to how the target actually appears in the blended image. (Right) The target component extracted by our target component extraction process more accurately captures the target's appearance in the blended image while removing the reference texture.

uniformity. For this color conversion, we first linearize the RGB values and convert them into the CIE XYZ color space according to the standardized primary values in ITU-Rec.709 [1]. Then, XYZ values are converted to L*a*b* values using the D65 white reference point. To simulate the spatial-frequency-dependent effect of contrast perception, we apply bandpass decomposition using the Laplacian pyramid [3] after the color space conversion. The non-oriented decomposition was used because we did not observe any improvements in prediction accuracy with oriented filters when tested on our dataset obtained in the psychophysical experiment described in the next section. The number of pyramid levels we use is $N^{\mathrm{lv}} = 6$, where the last level represents the low-pass residual component. The low-pass kernel we use is a spatially separable 5-tap filter: $[0.05, 0.25, 0.4, 0.25, 0.05]$. We use the same kernel for all downsampling and upsampling operations throughout the paper.

Below, we denote the Laplacian pyramid representations of the target, reference, and inhibitory component as ΔI^{tg}, ΔI^{rf}, and ΔI^{bl}, respectively. While ΔI^{rf} and ΔI^{bl} are obtained simply by applying the color conversion and pyramid decomposition to the reference and blended image, respectively, ΔI^{tg} is obtained via the target component extraction process described in the following subsection. When necessary, we specify the pixel index i, the color channel $c = \{L*, a*, b*\}$, and the index of the spatial frequency bands $k = \{1, 2, ..., N^{\mathrm{lv}}\}$ of the Laplacian pyramid representation such as $\Delta I^{\mathrm{rf}}_{i,c,k}$. However, in the following equations, we omit some of these indexes for the sake of readability when the same process is performed for all indexes.

3.1.2 Target component extraction. To precisely estimate the visibility of the target image embedded within the blended image, one needs to extract the target component. Here, we refer to the image intensity variations in the blended image for which the visibility should be calculated as "target component." Simply using the product of the alpha map and the target image as the target component is insufficient because the contrast produced by the spatial variation of alpha values is defined by the intensity of both the target and reference images (Fig. 6).

Thus, to remove the reference structure while maintaining the contrast information made by the target image and alpha map, we first linearly fit the bandpass component of the reference image to that of the blended image, and then use the residual component as the target component (Fig. 7). Formally, given the bandpass

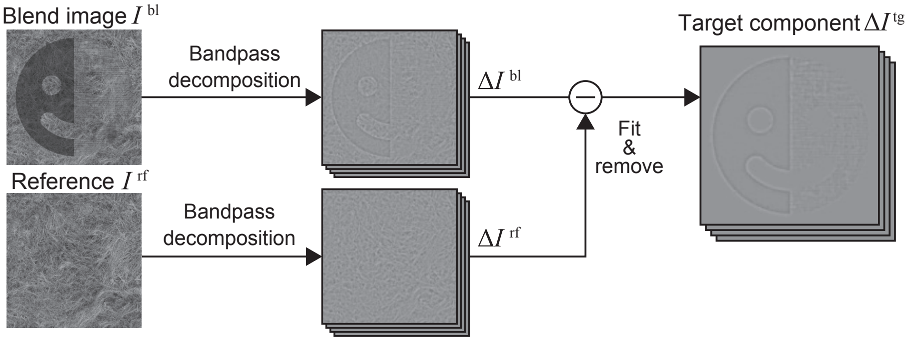

Fig. 7. Process to extract target component.

image of reference image ΔI^{rf} and that of the blended image ΔI^{bl} for each color channel c and for each pyramid level $k = \{1, 2, ..., N^{\text{lv}-1}\}$, we estimate the target component at the i-th pixel by removing the reference image component weighted with spatially varying weights l_i from the blended image component as:

$$\Delta I_i^{\text{tg}} = \Delta I_i^{\text{bl}} - l_i \Delta I_i^{\text{rf}}. \tag{3}$$

A weight l_i is estimated so as to minimize $\|\Delta I_i^{\text{bl}} - l\Delta I_i^{\text{rf}}\|_2$ within a 5×5 window surrounding the i-th pixel, $\mathcal{N}(i)$. The least squares solution for this problem can be obtained as:

$$l_i = \frac{\sum_{j\in\mathcal{N}(i)} \left(\Delta I_j^{\text{bl}} - \overline{\Delta I_i^{\text{bl}}}\right)\left(\Delta I_j^{\text{rf}} - \overline{\Delta I_i^{\text{rf}}}\right)}{\sum_{j\in\mathcal{N}(i)} \left(\Delta I_j^{\text{rf}} - \overline{\Delta I_i^{\text{rf}}}\right)^2 + \epsilon} \approx \frac{\sum_{j\in\mathcal{N}(i)} \Delta I_j^{\text{bl}} \Delta I_j^{\text{rf}}}{\sum_{j\in\mathcal{N}(i)} \left(\Delta I_j^{\text{rf}}\right)^2 + \epsilon}, \tag{4}$$

In the above approximation, we exploit the fact that the local means of the bandpass images, $\overline{\Delta I_i^{\text{bl}}}$ and $\overline{\Delta I_i^{\text{rf}}}$, are nearly zero. $\epsilon (= 10^{-8})$ is a small constant to avoid numerical instability. This also ensures that the solution becomes $I_i = 0$ when the denominator is zero (i.e., when the reference image is uniform and $\Delta I^{\text{rf}} = 0$). In this case, the target component is correctly estimated as $\Delta I_i^{\text{tg}} = \Delta I_i^{\text{bl}}$. We also clip l_i within the range of $[0, 1]$ to minimize the effect of noise in low signal regions. The weight values are computed for each pixel of the input bandpass image in a sliding window fashion.

The process to estimate the target components is repeated for each color channel c and for each pyramid level k except for the lowpass residual level of the Laplacian pyramid.

By collapsing the pyramid, we can obtain the target component image as shown in the right column of Fig. 6. The target component image more accurately captures the contrast information that exists in the blended image (left) than the simple product of the input target image and alpha map (center). In practice, we do not collapse the pyramid representation because we directly use the Laplacian pyramid representation for the following process.

3.1.3 Divisive normalization. The band-pass images of the target component are then divisively normalized by the band-pass images of the inhibitory component to simulate the contrast masking effect. This process is independently performed for each color channel c and for each pyramid level k. We obtain the normalized contrast response for the target component R_i as:

$$R_i = \frac{|\Delta I_i^{\mathrm{tg}}|}{(B^2 + \frac{1}{|\mathcal{N}|}\sum_{j \in \mathcal{N}(i)} |\Delta I_j^{\mathrm{bl}}|^2)^{\gamma}}, \tag{5}$$

where B and γ are variables that control the point at which saturation of the response begins and the slope of the nonlinear response function, respectively. $\mathcal{N}(i)$ represents a group of pixels neighboring the i-th pixel. Similar to several previous works [30, 50, 56], we use the surrounding 5×5 pixels as the definition of the neighborhood (thus, $|\mathcal{N}| = 25$).

We allow B to have different values for each level k and color channel c; however, we regularize them by keeping their relative sizes the same as the ratio of the mean square deviation of the bandpass image in the corresponding channel, as done in [31]. The mean squared deviation in each channel is averaged across the image patches used in the calibration (Section 4.2). Namely, $B_{c,k}$ is given by

$$B_{c,k} = \frac{\beta}{|\mathcal{X}||\mathcal{I}|} \sum_{x \in \mathcal{X}} \sum_{i \in \mathcal{I}} |\Delta I_{c,k,i}^{x}|^2, \tag{6}$$

where ΔI^x denotes the bandpass image of an image patch x. $\mathcal{X}$ is the collection of all the image patches used in the calibration, and $\mathcal{I}$ is a group of pixels contained in the bandpass image. β is a scaling parameter. β and γ are optimized in the model calibration (Section 4.2). For more details about the justification for choosing Eq. 5, please refer to Appendix A.

3.1.4 *Visibility of low-pass residual component.*

To account for the residual information, we take the difference between the blended image and reference image in the lowest level of the Laplacian pyramid, $\Delta I_{N^{\mathrm{lv}}}^{\mathrm{bl}}$ and $\Delta I_{N^{\mathrm{lv}}}^{\mathrm{rf}}$ for each pixel i for each color channel c as

$$R_{k=N^{\mathrm{lv}}} = |\Delta I_{k=N^{\mathrm{lv}}}^{\mathrm{bl}} - \Delta I_{k=N^{\mathrm{lv}}}^{\mathrm{rf}}|. \tag{7}$$

3.1.5 *Concatenation of contrast response maps.*

At this point, we have contrast response maps of the target components contained in the blended image for each pyramid level k for each color channel c. We concatenate these response maps to give $\mathbf{R} = [\mathbf{r}_1 \mid \mathbf{r}_2 \mid \ldots \mid \mathbf{r}_N] \in \mathbb{R}^{N^{\mathrm{c}} N^{\mathrm{lv}} \times N}$. Here, we align the image sizes of the different pyramid levels in an upsampling operation using the same kernel as the Laplacian pyramid decomposition.

3.2 Content-adaptive feature aggregation

The process so far has allowed us to obtain a contrast response map $\mathbf{R}$. In the next step, these responses are weighted and then aggregated to obtain a visibility map (Fig. 4). A key component of this stage is that it dynamically determines the weights according to features originally contained in the pre-blended target image.

For the first step, we compute the responses for the target component when the target image is opaquely overlaid on the reference image: $\mathbf{r}^{\mathrm{op}} = \{R\} \in \mathbb{R}^{N^{\mathrm{c}} N^{\mathrm{lv}}}$. The responses are computed by performing all the above operations with the alpha map set to completely opaque within the target image region. Then, the response vector of each pixel undergoes cross-channel interaction to generate a weight vector $\mathbf{w} \in \mathbb{R}^{N^{\mathrm{c}} N^{\mathrm{lv}}}$ as follows:

$$\mathbf{w} = \mathrm{sigmoid}(\mathbf{X}\mathbf{r}^{\mathrm{op}} + \mathbf{b}), \tag{8}$$

where $\mathrm{sigmoid}(\cdot)$ is a function that applies $1/(1+\exp(-z))$ for each element z in the input vector. $\mathbf{X} \in \mathbb{R}^{N^{\mathrm{c}} N^{\mathrm{lv}} \times N^{\mathrm{c}} N^{\mathrm{lv}}}$ is an interaction kernel that characterizes how each feature interacts with other features to determine the weight values. Furthermore, $\mathbf{b} \in \mathbb{R}^{N^{\mathrm{c}} N^{\mathrm{lv}}}$ is a bias term that represents a fixed baseline component of the weights. We optimize $\mathbf{X}$ and $\mathbf{b}$ as parameters in the calibration step. We constrain $\mathbf{X}$ to be symmetric and positive along the diagonal. We also apply weight decay to parameters in $\mathbf{X}$ when optimizing it. We observed that these regularizations led to better interpretability of our model without harming the validation accuracy on our dataset.

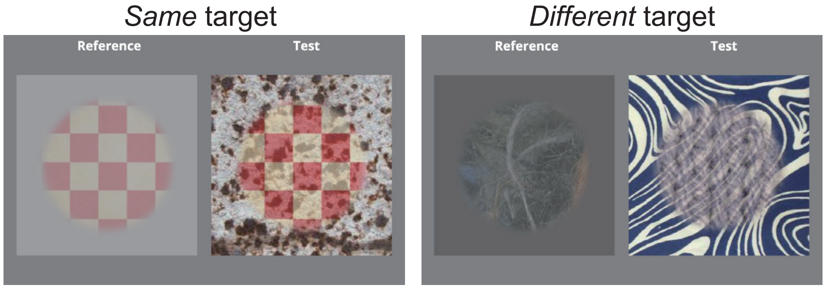

Fig. 8. Examples of experimental stimuli.

Finally, we aggregate the weighted response vector using the L_p norm to obtain the visibility value v_i:

$$v_i = \left(\sum_{n=1}^{N^{\mathrm{c}} N^{\mathrm{lv}}} w_n r_{n,i}^{p} \right)^{\frac{1}{p}}, \tag{9}$$

where w_n and r_n denote n-th components of the weight vector $\mathbf{w}$ and response vector $\mathbf{r}$, respectively. In addition, p is a parameter to be estimated in the calibration process.

Overall, our model has the following free parameters to be optimized in the calibration: $\theta = \{\beta, \gamma, p, \mathbf{X}, \mathbf{b}\}$.

4 MODEL CALIBRATION WITH LARGE-SCALE PSYCHOPHYSICAL EXPERIMENT

In this section, we describe the model calibration method along with a psychophysical experiment that was conducted to collect subjectively evaluated visibility data for alpha-blended images. We employed a visibility matching task in which participants adjusted the alpha value of a test stimulus such that the visibility of the test stimulus matched that of a reference stimulus (Fig. 8). The model parameters were optimized to best predict the matching data. Under the calibration, the model needs to predict a single visibility score for each of the test and reference stimuli. Since the visibility predictor described in Section 3 is a per-pixel model, we introduced a spatially aggregated visibility score $\bar{v}$:

$$\bar{v} = \left(\frac{1}{|\mathcal{N}^{\mathrm{tg}}|} \sum_{i \in \mathcal{N}^{\mathrm{tg}}} v_i^{q} \right)^{\frac{1}{q}}, \tag{10}$$

where $\mathcal{N}^{\mathrm{tg}}$ denotes a group of pixels that belong to the target component, and q is jointly optimized with the model parameters θ (see Section 4.2).

4.1 Psychophysical Experiment

4.1.1 Experimental setup. Participants used their own laptop or desktop computers to run experiment sessions on web browsers. We controlled the displayed sizes of the stimuli in terms of visual angle using a screen calibration task, in which a participant placed a credit card on the screen and adjusted the displayed rectangle to fit the card. We also instructed participants to maintain an observation distance of approximately 50 cm. Finally, we rescaled the displayed sizes of the stimuli according to individual pixel density values (i.e., pixels/deg) to equalize the stimuli sizes.

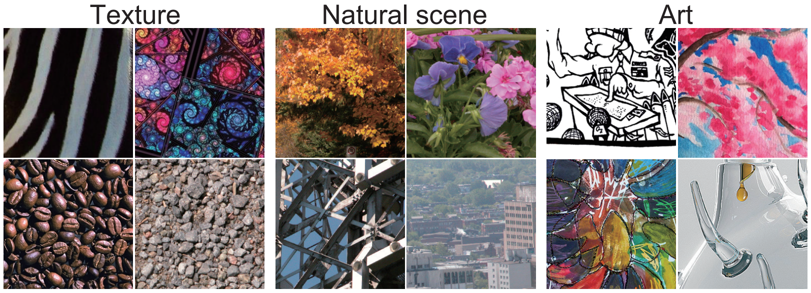

Fig. 9. Examples of image patches used in the psychophysical experiment.

4.1.2 Image dataset. We collected images from four databases, and categorized them into three classes: Texture, Natural scene, and Art.

- Texture: Images taken from the Describable Textures Dataset [7], MIT Vision Texture Database [40], and the "Texture" category in the McGill Calibrated Color Image Database [39].
- Natural scene: Images from the McGill Calibrated Color Image Database except for those in the "Texture" category.
- Art: Images categorized as "3D Graphics," "Pen ink," "Comic," "Vector Art," "Graphite," "Watercolor," and "Oil paint" in the Behance Artistic Media Dataset [57].

The image patches were extracted from the dataset at a size of 256×256 pixels (7.9×7.9 deg in visual angle). To diversify image features included in the stimuli, we first clustered image patches generated from each category by k-means clustering using the mean absolute deviation of the pixels in each level of the Laplacian pyramid (similar to [50]). The number of clusters was 128, 64, and 64, for Texture, Natural scene, and Art, respectively. Each image patch was sampled with equal probability from these clusters. Examples of image patches are presented in Fig. 9.

4.1.3 Stimuli. The stimuli were generated by blending two image patches with Eq. 2. The alpha map was generated by multiplying a single alpha value with a mask pattern I^{mask} that circularly cropped the target patch. The mask pattern was necessary to help observers distinguish the target from the background. To reduce the effect of the edge of the mask on the overall visibility, the boundary of the mask was blurred as $I_i^{\text{mask}} = 1 - \frac{1}{1+\exp(30-d_i/3)}$, where d_i is the distance (in pixels) from the image center.

4.1.4 Conditions. We measured the visibility in two types of situations: visibility matching between images with the same target patch (*Same* condition) and visibility matching between images with different target patches (*Different* condition), as illustrated in Fig. 8.

The target patch was blended with a uniform gray in the reference stimulus while it was blended with another background patch in the test stimulus. A total of 2,000 unique combinations of image patches were sampled for each of the *Same* and *Different* conditions. The intensity of the gray background was randomly chosen between the range [0,1]. For each combination, we changed the alpha value to three different levels as an independent variable. The alpha values were chosen randomly and uniformly on a logarithmic scale in the range $[\alpha_{min}, 0.8]$, where α_{min} was determined for each target patch such that the root-mean-square contrast of the reference stimulus did not fall below 0.006; this value was empirically set such that the reference stimulus was always visible. Both the *Same* and *Different* conditions consisted of 6000 subconditions (12000 subconditions in total).

4.1.5 Participants. Participants aged between 18 and 40 years were recruited via a crowdsourcing service (Prolific.ac). Only those who declared themselves to have normal or corrected-to-normal vision were able to participate in the experiment. The total number of participants was 527.

4.1.6 Procedure. During the visibility matching task, the reference and test stimuli were presented side by side, below which a horizontal slider was displayed that could control the alpha value of the test stimulus. The initial position of the slider was randomly set at either the left or right end of the slider bar for each trial. The participants were instructed to first move the slider to the opposite far side to see the appearance of the entire alpha range, and then adjust the slider such that the visibility of the target patch in the test stimulus matched that in the reference stimulus. The meaning of "visibility" was explained to participants as follows: "Visibility refers to how well you can see features (such as edges, structures, colors and so on) contained in the image."

Each participant performed a training session consisting of 10 trials before proceeding with the main session. The main session contained 150 trials, of which 75 trials used the *Same* condition while the remaining 75 trials used the *Different* condition. The *Same* trials and the *Different* trials were mixed together and presented in a randomized order. Short breaks were inserted every 50 trials. On average, a participant took approximately 30 min to complete the entire experiment, including instructions and training session.

4.1.7 Data validation. After all data were collected, we performed a screening to remove responses from unreliable participants. The details of the screening process are described in App. B. The number of participants who passed the screening was 486. As a result, we obtained 6-8 (mean = 6.05) valid responses for each subcondition. We also confirmed that the responses were consistent across participants; the Pearson correlation between the mean responses of two randomly split participant groups was 0.857 for the *Same* condition and 0.809 for the *Different* condition. Finally, we computed the mean response from all valid participants for each subcondition and used it for the model calibration.

4.2 Optimization of model parameters

In the experiment, we obtained a response alpha α_x averaged over participants for each condition $x \in \mathcal{X}$, where $\mathcal{X}$ is the set of training conditions. For condition x, we had reference stimulus data D_x^{R} (i.e., set of a target patch, background patch, and alpha value), and test stimulus data D_x^{T} (i.e., set of a target patch and background patch), and α_x. The parameters θ were calibrated by minimizing the difference between the predicted visibility values of the test and reference stimuli.

We defined the probability P that α_x was chosen for condition x using the softmax function as follows:

$$P(\alpha_x|D_x^{\mathrm{R}}, D_x^{\mathrm{T}}, \theta) = \frac{\exp(-s^2|\bar{v}(D_x^{\mathrm{R}}|\theta) - \bar{v}(D_x^{\mathrm{T}}, \alpha_x|\theta)|^2)}{\sum_{\alpha^{\mathrm{ref}}} \exp(-s^2|\bar{v}(D_x^{\mathrm{R}}|\theta) - \bar{v}(D_x^{\mathrm{T}}, \alpha^{\mathrm{ref}}|\theta)|^2)}, \tag{11}$$

where $\bar{v}(D_x^{\mathrm{R}}|\theta)$ is the predicted visibility value of the reference stimulus given parameters θ, and $\bar{v}(D_x^{\mathrm{T}}, \alpha|\theta)$ is the predicted visibility value of the test stimulus blended with α. s is a scaling parameter that controlled the uncertainty of the model prediction (width of the probability mass function). α^{ref} was discretely sampled at 0.01 intervals during optimization. We jointly optimized s with the model parameters.

The optimal parameters are obtained by minimizing the following negative log likelihood function:

$$L(\theta, q, s) = -\sum_{x \in \mathcal{X}} \log P(\alpha_x|D_x^{\mathrm{R}}, D_x^{\mathrm{T}}, \theta, q, s). \tag{12}$$

The model parameters were optimized by stochastic gradient descent using the Adam optimizer [28] implemented in PyTorch. We used the default parameters, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, except for the learning rate, which we set to 0.005 at the beginning and halved at the 12th, 18th, and 24th epoch. The training was completed with

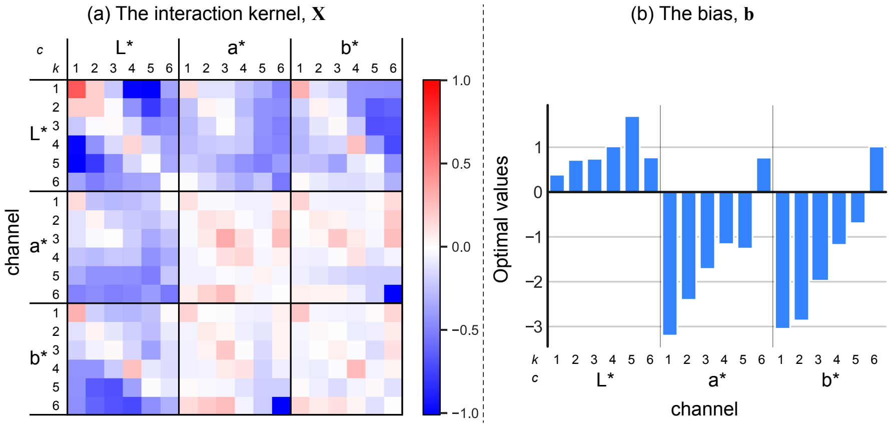

Fig. 10. (a) Optimized interaction kernel **X** and (b) bias **b** in Eq. 8.

Table 1. Optimized parameter values of the visibility predictor model.

β	γ	p	(q)	(s)
1.0121×10^{-3}	0.3973	0.7681	5.4250	4.5903

convergence at the 32nd epoch. We applied weight decay of 10^{-3} to parameters in the interaction kernel **X**. The batch size was set to 12.

The obtained optimal parameters are presented in Fig. 10 and in Table 1. We can observe that positive values emerged around the diagonal line of the interaction kernel, while negative values dominated in the off-diagonal elements, especially within achromatic channels. A similar trend appeared in the achromatic-chromatic interactions. This suggests that our model successfully learned a mechanism that weighted the features originally present in the target image and suppressed the others. Moreover, the optimized bias exhibited a bandpass shape in the achromatic channels and a lowpass shape in the chromatic channels. This is consistent with the known properties of the HVS [36].

5 VALIDATION OF THE PROPOSED MODEL

In this section, we first show how our content-adaptive weights change depending on feature variations in the input target images. Then, we validate our model by comparing its performance with several different models. Finally, we provide visual comparison of the visibility maps obtained by our model with those obtained by different techniques. In the supplementary material, we further demonstrate the validity of our model in terms of its ability to reproduce some of the basic properties of the HVS such as the contrast sensitivity function and contrast masking.

5.1 Analysis of the learned content-dependent weighting mechanism

The content-dependent weighting mechanism is introduced to appropriately weight the contrast of each feature based on its original appearance. Here, we show how the feature weights change depending on the distribution of the target image features.

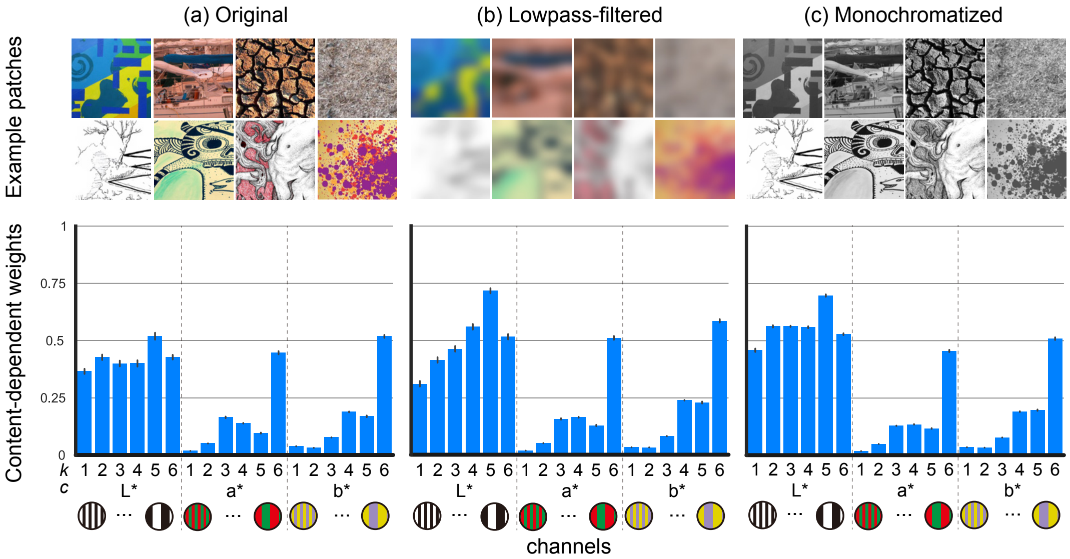

Fig. 11. Weight variation dependent on target content. We sampled 100 natural image patches for a target image and computed the mean content-dependent weight for each channel. (a) Original image patches were used for the target image. (b) Lowpass-filtered version of the image patches were used for the target image. (c) Grayscale version of the image patches were used for the target image. The error bars represent ±95% confidence intervals.

Figure 11 presents the averages of content-dependent weights computed from 100 natural image patches randomly sampled from the training data. When the weights are computed from the original image patch without any modification (Fig. 11a), they exhibit broadband sensitivity in the luminance (L *) channels, while they exhibit lowpass sensitivity in the chromatic (a* and b*) channels. The lack of bandpass shape in CSF under the suprathreshold condition is a known property of the HVS (often referred to as "CSF flattening" or "contrast constancy") [2, 8, 18, 35]. On the other hand, when the weights are computed from the lowpass-filtered version of the same image patches (Fig. 11b), they increase in the lower frequency channels while they decrease in the highest frequency channel (especially in the luminance channels). When the weights are computed from the monochromatized version of the same image patches (Fig. 11c), the weights increase in the luminance channels. These results supported the view that our data-driven approach could successfully realize a mechanism that adaptively weights each feature so that the representative features of the image have larger weights.

It should be noted that the weight for a channel is not necessarily zero even if there is no corresponding component in the input image (i.e., the weights for high frequency and chromatic channels are non-zero for blurred and monochromatic input images, respectively). This is because our calibration process did not constrain the weights to zero for non-activated channels. We leave this issue to be resolved in the future. However, these non-zero weights for non-activated channels do not pose a practical problem because they do not affect the visibility prediction results.

5.2 Comparative model evaluation

Here, we compared the predictive performance of our model to that of several previous models: peak signal-to-noise ratio (PSNR), MS-SSIM [54], High Dynamic Range Visible Difference Predictor (HDR-VDP) (ver. 3.0.6) [35], normalized Laplacian pyramid distance (NLPD) [30], and the IQA model [31] used in [16]. When testing these

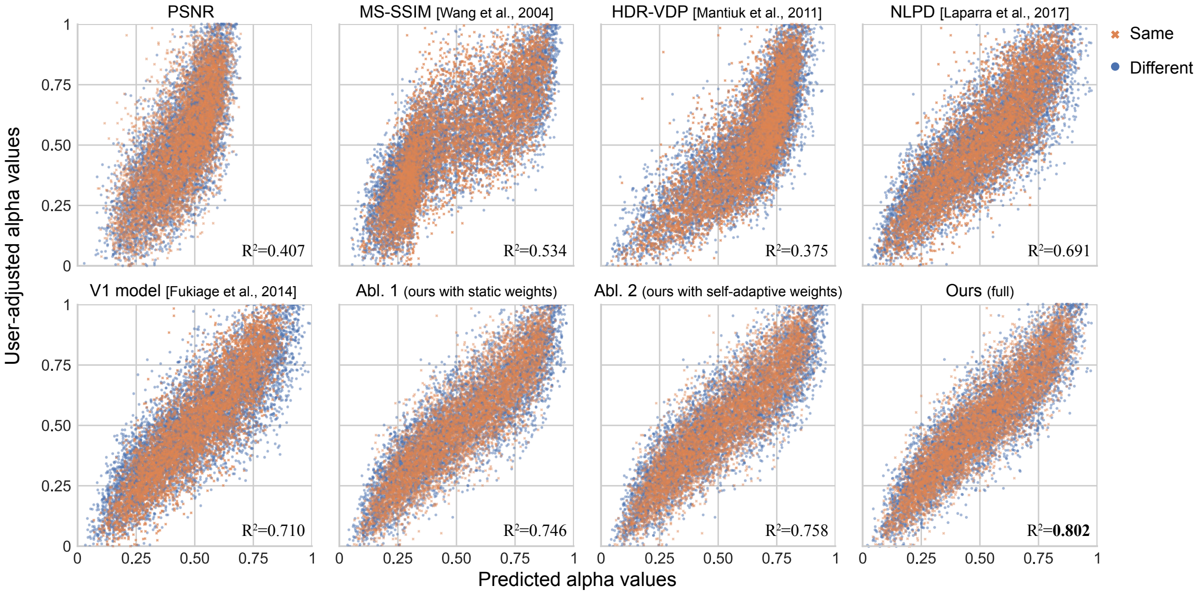

Fig. 12. Predictions of user-adjusted alpha values obtained in the visibility matching task. Each data point represents the mean response across participants for each subcondition. The predictions were obtained from the models trained on the entire dataset.

models, we used a background and blended image pair as a clean and distorted image pair. For the models using the luminance scale (i.e., HDR-VDP and NLPD), we rescaled the pixel values so that their range was [1, 200] (cd/m^2).

We also conducted ablation studies by replacing the core component of our model. 1) Static feature aggregation (Abl. 1): the model statically weighted each channel, and the weights for each channel were optimized in conjunction with all other parameters. 2) Self-adaptive feature aggregation (Abl. 2): the weights for each channel were dynamically determined based on the features output from the previous stage, not on the features of the original target content. The number of parameters of this model was exactly the same as for our full model.

The parameters of all the comparison models were recalibrated to better predict our dataset (see details in the Supplemental Material). We ensured that all of the training was converged using the same procedure as used for our model (Section 4.2).

Figure 12 displays scatter plots of the prediction results for each method. Our model achieved the best goodness of fit as measured by the coefficient of determination (R^2) shown in the bottom right corner. To further quantify the prediction performance of each model in terms of generalization, we performed 5-fold cross-validation and compared the negative log likelihoods on the validation set averaged over each fold (Fig. 13). We split the data such that image patches used in the training set did not appear in the validation set. The results demonstrated that our model performed significantly better than all other methods in predicting the visibility matching data, as demonstrated by the 95% confidence intervals in Fig. 13. Moreover, comparison with the ablation models suggested that both the adaptive weighting mechanism and access to the original features contained in the target image are vital for precise visibility prediction.

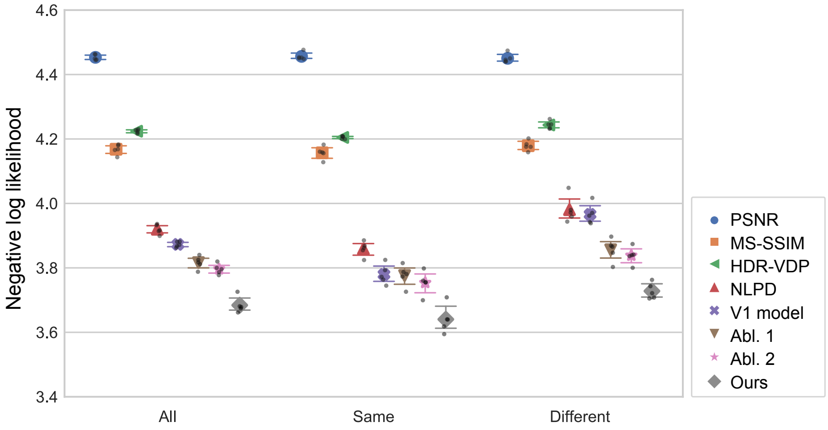

Fig. 13. Negative log likelihoods of the compared models. The color markers represent the mean of the 5-fold cross-validation. The black dots represent the data for each fold. The error bars represent ±95% confidence intervals.

5.3 Visual comparison of visibility maps

Finally, we compared the visibility maps obtained by our model with those obtained by two other techniques — an IQA model [16] and a saliency model [61] — for three opacity levels (Fig. 14). Here, the visibility values of our model and the IQA model are rescaled to $[0, 1]$ for visualization as described in Section 6.4. When the opacity is not high (top and middle rows), our model and the IQA model can correctly estimate the visibility by taking into account the effect of contrast masking due to the reference image component (e.g., the visibility reduction around the traffic sign in the reference image). However, since the saliency model does not have a mechanism for separately analyzing the target component from the reference image, it assigns high visibility to salient objects in the reference image such as the traffic sign and traffic light. When the opacity of the target image is high (bottom row), the IQA model shows the stronger effect of reference image content on the visibility map. For example, the shape of the traffic sign is clearly visible in the visibility map even though it is completely occluded by the target image in the blended result. This happens because the IQA model computes the visibility as the perceptual distance between the blended image and the reference image. Our model does not exhibit the same issue thanks to the target component extraction process (Section 3.1.2). Note that the respective problems of the IQA and saliency models shown here are commonly observed regardless of the specific choice of model.

6 PERCEPTUALLY OPTIMIZED IMAGE BLENDING

In this section, we present an application of the proposed visibility predictor model, where we optimize an alpha map so that the visibility of the target image achieves the target visibility level (Fig. 15). Since our model can account for the alpha map structure when computing visibility, the solution is not constrained to be uniform. Thus, the alpha map is optimized in a spatially adaptive manner. As a result, users can obtain not only the desired visibility across different reference images but also uniform visibility across different areas within a single image. Alternatively, users can use a spatially varying visibility map for input. The technique provides intuitive control over the appearance of the blending results for users who want to achieve the desired level of visibility regardless of the image combination.

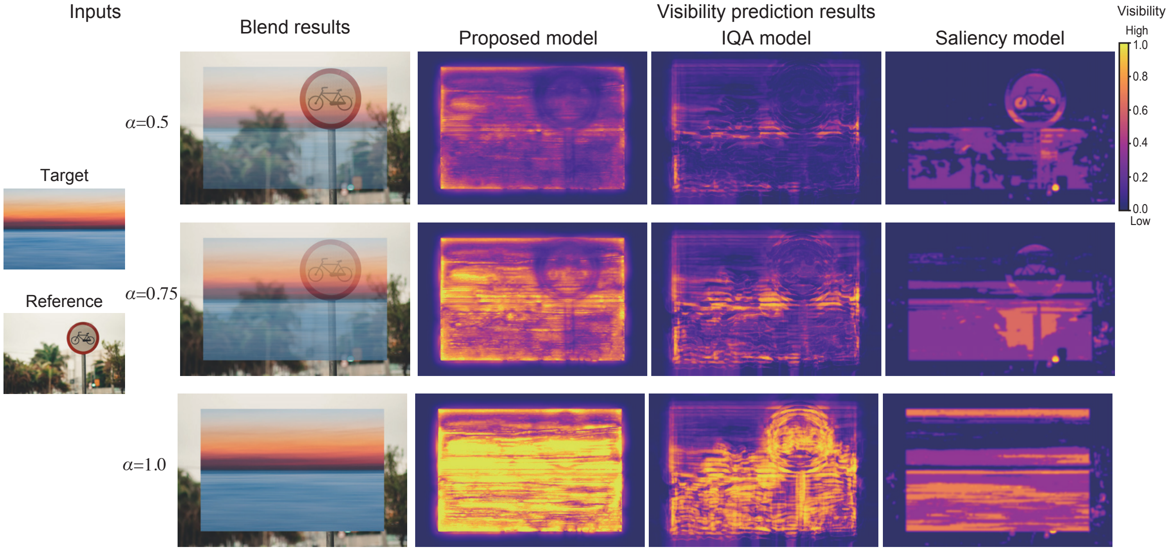

Fig. 14. Comparison of supra-threshold visibility maps. The visibility maps obtained by our model (the third column), IQA model [16] (the fourth column), and the saliency model [61] (the fifth column) are computed for the three blended images (the second column). For visualization, the visibility values of our model and the IQA model are rescaled to [0,1] as described in Section 6.4. For further details, please refer to Section 5.3. The target images and reference image are taken from Abdullah Ghatasheh / Pexels, and Gabriel Santos Fotografia / Pexels, respectively.

6.1 Method

In the image blending procedure, we use an off-the-shelf first-order gradient-based technique as the optimization algorithm [28]. The loss function $\mathcal{L}$ used for the optimization is composed of the following three terms: visibility loss $\mathcal{L}_v$, image fidelity loss $\mathcal{L}_f$, and edge-preserving smoothness loss $\mathcal{L}_s$. The three loss terms are combined as a weighted sum:

$$\mathcal{L} = \lambda_v \mathcal{L}_v + \lambda_f \mathcal{L}_f + \lambda_s \mathcal{L}_s. \tag{13}$$

We empirically set the weights to $(\lambda_v, \lambda_f, \lambda_s) = (0.1, 1, 1)$. The evaluation of each loss function is presented in the Supplemental Material.

6.1.1 Loss functions.

Visibility loss. The visibility loss minimizes the differences between the current visibility level and the user-specified target visibility level. The differences are evaluated for each pixel to achieve spatially consistent visibility levels. The mean absolute difference is used as the visibility loss term:

$$\mathcal{L}_v = \frac{1}{|\mathcal{N}^{\text{tg}}|} \sum_{i \in \mathcal{N}^{\text{tg}}} |\hat{v} - v_i|, \tag{14}$$

where $\hat{v}$ is the target visibility level. $\mathcal{N}^{\text{tg}}$ is the same as Eq. (10). To compute the local visibility levels for the current alpha map, we use the pixel-wise visibility estimate in Eq. (9).

Image fidelity loss. To retain the structure of the target image, we introduce the image fidelity loss. The image fidelity loss maximizes the correlation between the target component and the opaquely rendered target image

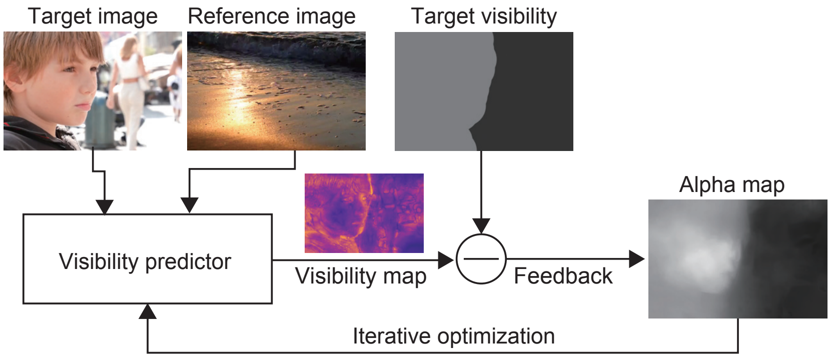

Fig. 15. Schematic diagram of the perceptual optimization procedure. The target and reference images by Roxolana / Pixabay and outlinez, respectively.

I^{op} in the Laplacian pyramid domain as follows:

$$\mathcal{L}_f = \frac{1}{N^{\text{lv}} - 1} \sum_{k=1}^{N^{\text{lv}}-1} \left(1 - \rho\left(\Delta\mathbf{I}_k^{\text{op}}, \Delta\mathbf{I}_k^{\text{tg}}\right)\right), \tag{15}$$

where $\rho(a, b)$ is a function to evaluate the correlation between images a and b. Here, we ignore the low-pass residual component because we only need to constrain the structural components of the image. We use the original RGB color space to compute the Laplacian pyramid because it gives more visually pleasing results than using the L*a*b* color space. Thus, to generate $\Delta\mathbf{I}^{\text{tg}}$, we repeat the target component extraction process in Section 3.1.2 without color conversion.

Correlation function ρ. We first compute correlation coefficients for each pixel by the covariance and standard deviations from the pixel values weighted by a Gaussian window centered at that pixel. The kernel size and standard deviation of the Gaussian window are 15 pixels and 6 pixels, respectively. Then, the local correlation coefficients are averaged over the entire image space as well as the color channels.

Edge-preserving smoothness loss. There are often cases in which optimal alpha values are ambiguous for a given pixel; an example is when both target and reference pixels have the same intensity. This ambiguity can make the optimization unstable and in the worst case can result in visible noise in the blended image. To avoid this problem, we incorporate a regularization term that works in smooth target areas while allowing for large variations in areas where the intensity of the target image changes significantly. The loss function we use is as follows:

$$\mathcal{L}_s = \frac{1}{HW} \sum_{x=1}^{W} \sum_{y=1}^{H} \left(\left| \frac{\partial}{\partial x} \alpha_{x,y} \right| e^{-\left| \frac{\partial}{\partial x} I_{x,y}^{\text{op}} \right|} + \left| \frac{\partial}{\partial y} \alpha_{x,y} \right| e^{-\left| \frac{\partial}{\partial y} I_{x,y}^{\text{op}} \right|} \right), \tag{16}$$

where $\alpha_{x,y}$ and $I_{x,y}^{\text{op}}$ respectively represent the pixel values of the alpha map and the opaquely rendered target image at the x-th column in the y-th row. The luminance channel is used as input for the target image. W and H denote the width and height of the image, respectively. We used the Sobel filter to compute the horizontal and vertical derivatives.

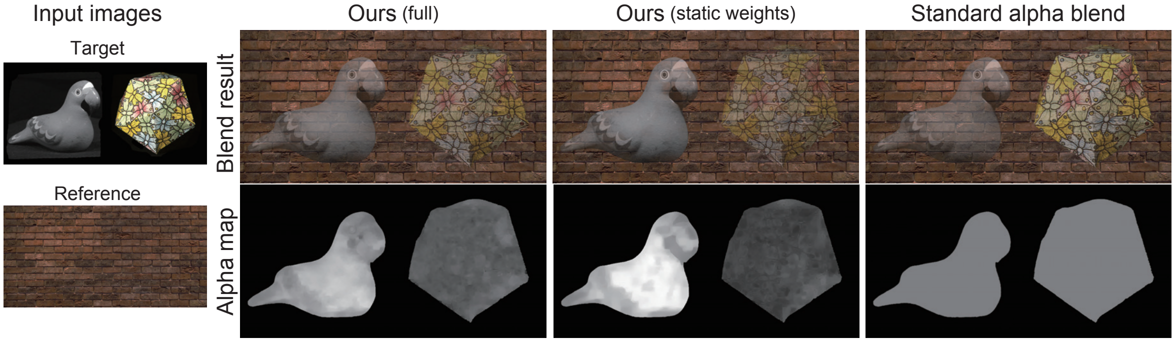

Fig. 16. Effect of content adaptive feature aggregation on blending results. The first column shows a target and reference image pair used as inputs. The result for our method with content-adaptive feature aggregation is shown in the second column. The third column shows the result for our method with static feature aggregation. Both of the results were generated with a normalized target visibility value of 0.5 (see Section 6.4 about the normalization). In the fourth column, we show the results of standard alpha blending using a uniform alpha value of 0.5. The grayscale images on the bottom row display the corresponding alpha maps. The results demonstrate that our full method using content-adaptive feature aggregation is the most successful in achieving consistent visibility. For details, please refer to Section 6.2.1. The object images and background reference image are taken from ALOI database, and Michael_Laut / Pixabay, respectively.

6.1.2 Implementation. Assume that the alpha map is initialized with uniform noise in the range [-0.2,0.2] added to 0.5. We first generate a blended image in which the target image is opaquely overlaid on the reference image using the target mask, and compute the response map. Then, we obtain the weight map for each channel by Eq. 8. Computation of the weight maps is performed only once at the beginning of the optimization process. At each iteration of the optimization, we feed the precomputed weight maps into the visibility model and compute the visibility loss as well as the two other loss terms. The derivative of the loss is then back-propagated to the alpha map, and the alpha map is updated in the direction that reduces the loss. For this iterative optimization, we used the Adam optimizer [28] implemented in PyTorch. We set the learning rate to 0.1 and set the other parameters to $(\beta_1, \beta_2) = (0.9, 0.999)$. We completed the optimization at certain iterations (500 in the experiments). Thereafter, we generated the blended image with the alpha map and the target mask.

When optimized on a single GPU (NVIDIA RTX 3090), the entire process to generate an optimized blending image of 768 × 512 pixels took approximately 20 sec. Thus, our method is currently not applicable to interactive applications such as AR/VR. However, the loss function used to optimize an alpha map can also be used to train a feedforward generative model that generates an optimal alpha map to achieve a desired visibility level given an arbitrary target/ reference image pair. Once such a generative model is successfully trained, it runs in any interactive applications. We have already tested this and obtained promising preliminary results.

6.2 Visual evaluation of blending methods

Here, we first demonstrate the effectiveness of our method in several respects with some example results. Further examples are provided in the supplemental material.

6.2.1 Effectiveness of content-adaptive feature aggregation. To validate the effectiveness of the content-adaptive feature aggregation mechanism in our visibility model, we tested a variant of our blending technique in which we replaced our visibility model by a model with static feature aggregation. The alpha values optimized by this ablation method tended to be too high for a texture-less target image while they tended to be too low for a

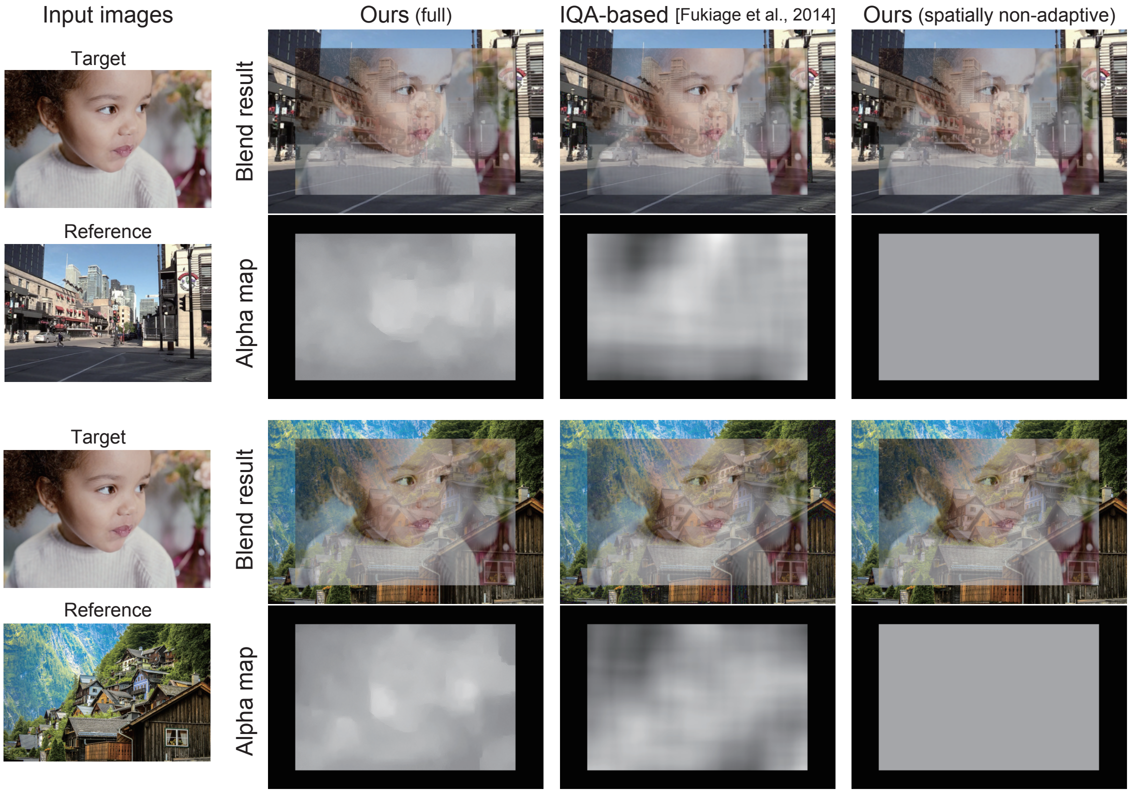

Fig. 17. Comparison of results obtained by three different methods. The first column shows target and reference image pairs used as inputs. The result for our method is shown in the second column. The third column shows the result for the mothod used by Fukiage et al. 2014 [16]. In the fourth column, we show the results for our method without spatial adaption (i.e., the images are blended with a uniform alpha map whose alpha value is optimized by the proposed visibility predictor model). The grayscale images on the bottom row display the corresponding optimized alpha maps. All the images were generated with a normalized target visibility value of 0.5 (see Section 6.4 about the normalization). The results demonstrate that our full method is the most successful in achieving the consistent visibility for the target image across different reference images as well as across different areas within a single image. Input images are by cottonbro / Pexels, German Korb / Pexels, and Rahat Ali / Pexels, respectively.

textured target image, as illustrated in Fig. 16 (the third column). 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 as well as chromatic information while it overestimated the visibility when the image had abundant features. In contrast, our method with content-adaptive feature aggregation achieves more consistent visibility, by appropriately increasing the weights for low-frequency and luminance channels in the smooth gray regions while suppressing overall weights in the textured regions (the second column).

6.2.2 Comparison to IQA-based blending method. The third column in Fig. 17 presents the results of the IQA-based technique proposed by Fukiage et al. [16]. For a fair comparison, we optimized the IQA model used in their method on our dataset. We found that the method shared the same problem as the static feature aggregation method

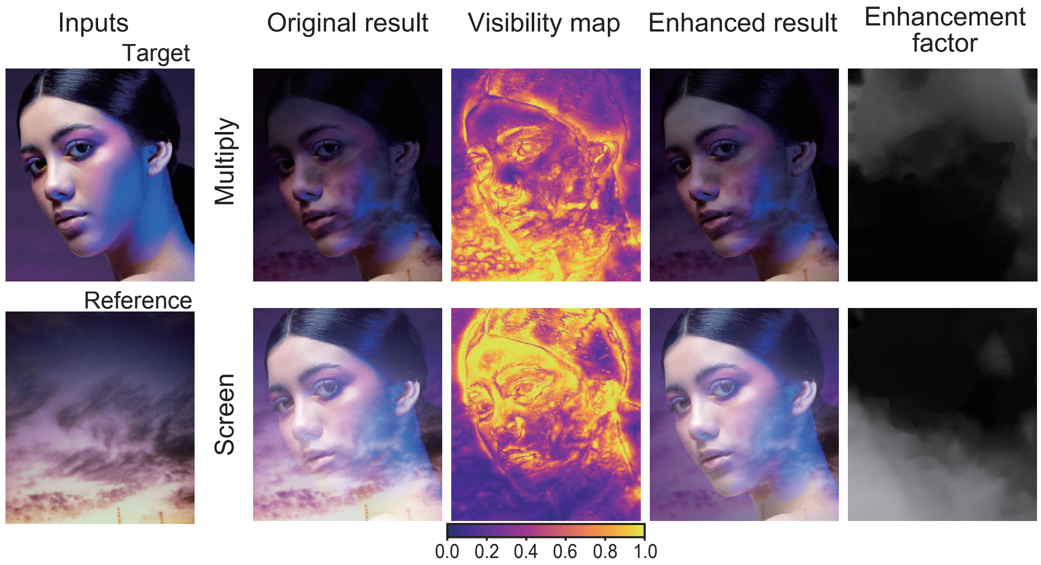

Fig. 18. Application to enhancement for nonlinear blending. The first column shows the input images. In the rest of the columns, the results for multiply blend and screen blend are shown in the top and bottom rows, respectively. The blending equations used for the respective blend modes are: $I^{\mathrm{bl}} = I^{\mathrm{tg}} \odot I^{\mathrm{rf}}$ for multiply blend, and $I^{\mathrm{bl}} = 1 - (1 - I^{\mathrm{tg}}) \odot (1 - I^{\mathrm{rf}})$ for screen blend. Using these blend modes, the visibility of the target image significantly degrades depending on the brightness of the reference image (the second column). Our visibility model is applicable to these nonlinear blending modes by simply replacing the blending equation (Eq. 2) with those given here, and is able to predict this visibility degradation (the third column). Then, we enhance the visibility of the degraded regions by modulating the brightness of the reference image using the following equations: $I'rf = 1 - (1 - A) \odot (1 - I^{\mathrm{rf}})$ for multiply blend and $I'rf = (1 - A) \odot I^{\mathrm{rf}}$ for screen blend, where $A \in \mathbb{R}^N$ is an enhancement factor map. Intuitively, the larger A, the brighter the reference image becomes in multiply blend, thus increasing the visibility of the target image, and the darker the reference image becomes in screen blend, thus increasing the visibility of the target image. We optimized the enhancement factor map using the same procedure as described in Section 6.1, except that we initialized the enhancement factor as uniform noise in the range [0, 0.1] for faster convergence. We set the target visibility value to 0.5. The resulting enhanced blend images and enhancement factor maps are shown in the fourth and fifth columns, respectively. Input images are by Serch Arafat / Pexels and rikka ameboshi / Pexels.

tested above: excessive alpha values in smooth regions (see around the girl's forehead and the corresponding alpha map in the first row). This method also fails to achieve consistent visibility between two results with different reference images even though the target visibility is set to the same value of 0.5.

6.2.3 Comparison to spatially-non adaptive method. To demonstrate the effect of spatially adaptive optimization, we compare our method with the spatially non-adaptive version of our method. In this ablation method, we spatially aggregated the local visibility estimates as in Eq. 10 and identified a single optimal alpha value that achieved the target visibility. The result obtained by the spatially non-adaptive method (the rightmost column of Fig. 17) exhibits inconsistent visibility across image areas; the visibility appears to be significantly reduced in the cluttered regions of the reference image. Our full method achieves more consistent visibility across these image regions by locally adjusting the alpha values (see the alpha maps in the second column of Fig. 17).

6.2.4 Application to spatially varying visibility maps. Our technique can be also applied to a spatially varying target visibility map simply by replacing $\hat{v}$ in Eq. 14 with a pixel-wise visibility map. An example of the result is presented in Fig. 1. We used a hand-drawn target visibility map in which the important region of the target image was filled in with higher visibility. Compared to standard alpha blending, our results exhibited consistent visibility within regions with the same target visibility. Moreover, owing to the image fidelity loss, which preserved the original image structure, a natural transition from high-opacity to low-opacity regions was achieved.

6.2.5 Application to video inputs. We also tested our method using video frames as inputs (Fig. 1, video material). The optimization was performed independently for each frame. For sequential frames in a continuous scene, once the initial frame was completed with 500 iterations, the next round of optimization was performed with fewer iterations (i.,e., 100), using the result of the previous frame as the initial values. It is noted that we did not observe apparent flickers even though we did not explicitly enforce temporal consistency. The reason for this is because changes in the alpha values occurred as changes in the target and reference images, which were unlikely to produce impressions of unnatural variation in the color of the target object. As can be seen in the supplemental videos, our optimization resulted in more consistent visibility of the target scene throughout all video frames than is achieved with standard alpha blending.

6.2.6 Application to nonlinear blending. Although we presented our visibility model only on simple linear blending settings, it is, in theory, applicable to any nonlinear blending by simply replacing the blending equation (Eq. 2). To demonstrate the utility of our optimization framework in nonlinear blending, in Fig. 18, we present an example of our method applied to enhancing the results of two non-linear blend modes: multiply blend and screen blend. When using these blending modes, the visibility of the target image significantly degrades depending on the brightness of the reference image (the second column in the figure above). To improve the visibility of the target image, we optimized the enhancement factor map (instead of the alpha map), which controls the brightness of the reference image. It is observed that the degraded image regions in the original blend results are successfully improved in our enhanced results (the third column). For details, please refer to the caption for Fig. 18.

6.3 User study

To further evaluate the effectiveness of the proposed method, we compared our method with several other methods via a visibility rating task. In the supplementary material, we also present the results of an additional user study with pairwise comparisons of visibility consistency.

6.3.1 Method.

Stimuli. Four different classes of images were used as target images: face, text, illustration, and object. We selected four different images for each class. As reference images to be combined with the target images, we selected 10 different photos capturing various indoor/outdoor scenes and a uniform gray image (11 reference images in total) to cover different levels of clutter and contrast. The face images were selected from the Flickr-Faces-HQ (FFHQ) dataset [25], while the object images were selected from the Amsterdam Library of Object Images (ALOI) dataset [19]. The other images were selected from a website for sharing images (Pixabay). All images were resized to 512 × 512 pixels. Examples of these images are presented in Fig. 19.

Comparison methods. We compared our proposed method with the following four methods: 1) standard alpha blending, 2) the IQA-based method proposed by [16], 3) our method with spatially non-adaptive optimization, and 4) our method with static feature aggregation (Section 6.2.1). In the spatially non-adaptive method, we spatially aggregated the local visibility estimates as in Eq. 10 and identified a single optimal alpha value that achieved the target visibility. Then, the blending result was generated by a spatially uniform alpha map.

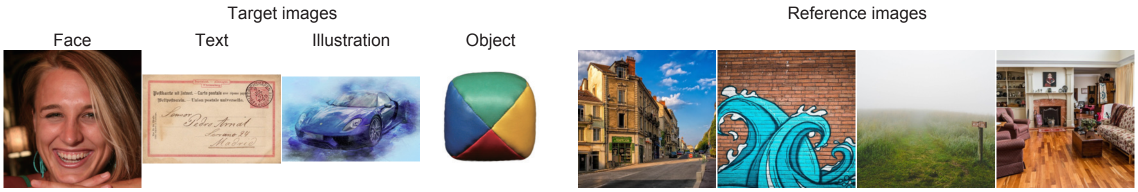

Fig. 19. Examples of input images used to generate stimuli in User study.

Procedure. The participants were instructed to rate the visibility of a target image blended on a reference image. In this experiment, the blended image (test image) and original target image were presented on the right and left sides of the screen, respectively. The participant reported the visibility of the target image using a stepless slider. The slider had five equally spaced scales for reference, each associated with a linguistic description of the visibility level as follows: 1. *Invisible*, 2. *Barely visible*, 3. *Visible*, 4. *Fairly visible*, and 5. *Very clear*. For each method, we generated blended images with eight different target visibility values for each image pair. We set the visibility range of each method such that the range of alpha values of the test images approximately matched across the comparison methods. The proposed method and comparison methods, other than standard alpha blending, were tested. We obtained the responses for the standard alpha blending method from the results of the spatially non-adaptive optimization method by replotting the responses against the raw alpha values. Further details are provided in Appendix C.

6.3.2 Results. Figure 20 presents the visibility scores for each condition averaged across participants against the target visibility levels. We computed the Pearson correlation ρ_P and Spearman correlation ρ_S between the subjectively rated visibility scores and the target visibility levels for each compared method. Our blending method resulted in the highest correlation for both correlation coefficients. We also performed a statistical test on the differences in the correlation coefficients by transforming the coefficients into Fisher's z-scores [37]. The results indicate that the correlation of our method was significantly higher than that of the other methods ($p < 10^{-4}$ for the Pearson correlation and $p < 10^{-3}$ for the Spearman correlation), except for the spatially non-adaptive method ($p = 0.218$ for the Pearson correlation and $p = 0.521$ for the Spearman correlation). We believe that this was because the task of the user study was to rate the global visibility level of an image; thus, the advantage of spatially adaptive optimization was obscured.

In addition, we fitted a nonlinear function to model the relationship between the raw target visibility values x_v and subjective scores y_v and computed the root-mean-square error (RMSE). The generalized logistic function was used as the nonlinear function:

$$y_v = p_1 + \frac{5 - p_1}{(1 + p_2 \exp(-P_3 x_v))^{\frac{1}{p_4}}}, \tag{17}$$

where p_1 - p_4 are parameters. We constrained the solution so that the curve always passed through the origin $(x_v, y_v) = (0, 0)$. The fitted parameters for each comparison method are presented in Appendix D. The gray curves in Fig. 20 represent the fitted functions. The root-mean-square error (RMSE) of each fit was displayed in the top-left corner of each plot. Again, our full method marked the lowest error, meaning that our full method was the best for controlling the visibility of blended images.

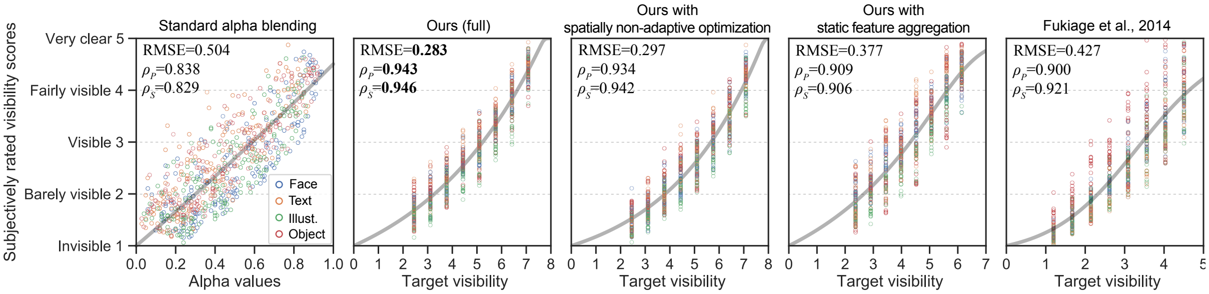

Fig. 20. Subjectively rated visibility scores vs. target visibility for each blending method. The gray curve represents the generalized logistic function fitted to the data. The RMSE value presented in the top left corner denotes the root-mean-square error between the fitted function and the data. ρ_P and ρ_S are the Pearson and Spearman correlation coefficients between the abscissa and ordinate, respectively.

6.4 Normalization of target visibility

Because the raw visibility values provided by the visibility predictor are not intuitive for practical use, we normalized them in the range $[0, 1]$ according to the generalized logistic function (Eq. 17) fitted to the subjective visibility scores in Fig. 20. Namely, we linearly scaled y_v to get the normalized value as: $(y_v - 1)/4$.

7 CONCLUSION AND FUTURE WORK

The visibility of an image semitransparently overlaid on another image significantly varies depending on the content of the images. To overcome this problem, we have proposed a perceptually optimized image blending method in which the opacity of the target image is optimized for each pixel based on a visibility model.

Our visibility model incorporates a content-adaptive feature aggregation mechanism that adaptively weights each image feature depending on the original image content. We conducted a large-scale psychophysical experiment to measure subjective visibility data and calibrate the visibility model. Through several comparative and ablation studies, we demonstrated that our model predicted the visibility of the target component better than other predictor models. The utility of our model was further validated with the proposed image-blending method. Our user study demonstrated that our blending method outperformed previous techniques in achieving the target visibility in the blended results.

While our model has demonstrated state-of-the-art performance in supra-threshold visibility prediction and visibility control of blended images, there are still several remaining issues and additional potential applications to investigate, as described below:

Visual mechanisms underlying the content-adaptive feature aggregation. It is still unclear what mechanisms underlie the content-adaptive feature aggregation found in this work. However, recent studies have suggested that our contrast perception can be modulated by high-level context information as well as visual attention [23, 38, 43, 49]. We also found a suppressive interaction in the learned feature aggregation mechanism (Section 5.1). In line with this finding, a recent study [27] showed that there is a similar suppressive interaction between chromatic and luminance channels in suprathreshold contrast perception. Further investigating the relationships of our model to these previous findings in the vision science literature is an important future direction for our research.

Environmental conditions. Our model does not consider environmental factors, such as the observation distance, display brightness, and color reproducibility. The L*a*b* color space used in our model is also not accurate enough to capture the light adaptation in the HVS. However, while these factors have a significant impact on the near-threshold visibility of an image, their impact is relatively limited in the suprathreshold contrast perception that our model aims to predict. For example, it is known that our suprathreshold contrast perception is nearly scale-invariant [2, 8, 18]. We also ensured in the preliminary test that the correlation between the data obtained in the controlled laboratory experiment using calibrated color monitors and those collected in the online experiment was high (Pearson's $r = 0.965$).

However, it is likely that the display settings more or less affected the individual results and that our model (and thus the proposed blending technique) is calibrated to the average of the display settings of the screens used by the crowd-sourced participants recruited for our experiment. A possible solution for this is to add a display model that can compensate for the observation conditions with additional user data as performed in [60].

Incorporating additional processing in the HVS. As our model relies on spatially local image processing, it does not take into account the interactions between spatially distant areas on image visibility, such as the center-surround interactions and colinear facilitation effects [13]. We also do not explicitly model the layer segregation process in the HVS. Although we extract the target's bandpass component in our model, we presume that the reference component is known. Therefore, our model is not applicable to situations where the target and reference components are hard for human observers to distinguish (e.g., when both images have very similar textures). We believe that incorporating all of these effects will further improve the visibility prediction, including for cases that our model does not currently address, i.e., cases where the target image is enhanced by the addition of a reference image. In addition, our visibility model does not take into account the temporal aspect of the HVS. Although the proposed blending method works in practice for video inputs, incorporating the temporal properties of the HVS, such as the chromatic and luminance adaptation as well as contrast sensitivity to spatio-temporal frequency [26], should improve the results for highly dynamic scenes.

Additional applications. In this paper, we presented only a straightforward application of the proposed visibility model (i.e., optimizing the opacity of a target image to achieve a desired visibility level). However, we think that the applications of our model are not limited to this. For example, our model can be used to control the visibility of multiple images simultaneously when more than two input images are composited together if we properly design the loss function. It will also be possible to enhance the visibility of an image presented on OST displays where images are typically shown semi-transparently (as in the work by Fukiage et al. [16]), although additional mechanisms such as the accommodation and binocular combination processes must be considered to precisely predict the visibility in such cases. Another possible use case is to control the apparent transparency of rendered materials by adjusting the visibility of the transmitted scene, along with other perceptually relevant properties such as distortion [12] and colors [11].

ACKNOWLEDGMENTS

We thank all the anonymous reviewers for their helpful suggestions.

REFERENCES

[1] ITU-R Rec. BT. 709-6. 2015. Parameter values for the HDTV standards for production and international programme exchange. https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.709-6-201506-I!!PDF-E.pdf

[2] N Brady and D J Field. 1995. What's constant in contrast constancy? The effects of scaling on the perceived contrast of bandpass patterns. *Vision Res.* 35, 6 (March 1995), 739–756.

[3] Peter J Burt and Edward H Adelson. 1983. A multiresolution spline with application to image mosaics. *ACM Trans. Graph.* 2, 4 (Oct. 1983), 217–236.

[4] F W Campbell and J G Robson. 1968. Application of Fourier analysis to the visibility of gratings. *J. Physiol.* 197, 3 (Aug. 1968), 551–566.

[5] Ming-Yuen Chan, Yingcai Wu, Wai-Ho Mak, Wei Chen, and Huamin Qu. 2009. Perception-based transparency optimization for direct volume rendering. *IEEE Trans. Vis. Comput. Graph.* 15, 6 (Nov. 2009), 1283–1290.

[6] Johnson Chuang, Daniel Weiskopf, and Torsten Möller. 2009. Hue-preserving color blending. *IEEE Trans. Vis. Comput. Graph.* 15, 6 (Nov. 2009), 1275–1282.

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing Textures in the Wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 3606–3613.

[8] Scott Daly. 1989. Application of a Noise Adaptive Contrast Sensitivity Function to Image Data Compression. In *Human Vision, Visual Processing, and Digital Display*, Vol. 1077. International Society for Optics and Photonics, 217–227.

[9] Scott J Daly. 1992. Visible differences predictor: an algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III*, Vol. 1666. International Society for Optics and Photonics, 2–15.

[10] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. 2022. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 5 (May 2022), 2567–2581.

[11] Franz Faul. 2017. Toward a Perceptually Uniform Parameter Space for Filter Transparency. *ACM Trans. Appl. Percept.* 14, 2 (Jan. 2017), 1–21.

[12] Roland W Fleming, Frank Jäkel, and Laurence T Maloney. 2011. Visual perception of thick transparent materials. *Psychol. Sci.* 22, 6 (June 2011), 812–820.

[13] John M Foley. 2019. Lateral effects in pattern vision. *J. Vis.* 19, 9 (Aug. 2019), 8.

[14] 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.

[15] Taiki Fukiage, Takahiro Kawabe, and Shin ya Nishida. 2019. Perceptually Based Adaptive Motion Retargeting to Animate Real Objects by Light Projection. *IEEE Trans. Vis. Comput. Graph.* (2019), 1–1.

[16] 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.

[17] Joseph L Gabbard, J Edward Swan, Deborah Hix, Si-Jung Kim, and Greg Fitch. 2007. Active Text Drawing Styles for Outdoor Augmented Reality: A User-Based Study and Design Implications. In *2007 IEEE Virtual Reality Conference*. IEEE, Charlotte, NC, USA, 35–42.

[18] M A Georgeson and G D Sullivan. 1975. Contrast constancy: deblurring in human vision by spatial frequency channels. *J. Physiol.* 252, 3 (Nov. 1975), 627–656.

[19] Jan-Mark Geusebroek, Gertjan J Burghouts, and Arnold W M Smeulders. 2005. The Amsterdam Library of Object Images. *Int. J. Comput. Vis.* 61, 1 (Jan. 2005), 103–112.

[20] Anselm Grundhöfer and Oliver Bimber. 2008. Real-time adaptive radiometric compensation. *IEEE Trans. Vis. Comput. Graph.* 14, 1 (Jan. 2008), 97–108.

[21] Mark Grundland, Rahul Vohra, Gareth P Williams, and Neil A Dodgson. 2006. Cross dissolve without cross fade: Preserving contrast, color and salience in image compositing. *Comput. Graph. Forum* 25, 3 (Sept. 2006), 577–586.

[22] D J Heeger. 1992. Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* 9, 2 (Aug. 1992), 181–197.

[23] Sirawaj Itthipuripat, Kai-Yu Chang, Ashley Bong, and John T Serences. 2019. Stimulus visibility and uncertainty mediate the influence of attention on response bias and visual contrast appearance. *J. Vis.* 19, 14 (Dec. 2019), 8.

[24] Denis Kalkofen, Eduardo Veas, Stefanie Zollmann, Markus Steinberger, and Dieter Schmalstieg. 2013. Adaptive ghosted views for Augmented Reality. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, Adelaide, Australia, 1–9.

[25] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4396–4405.

[26] D H Kelly. 1979. Motion and vision. II. Stabilized spatio-temporal threshold surface. *J. Opt. Soc. Am.* 69, 10 (Oct. 1979), 1340–1349.

[27] Yeon Jin Kim and Kathy T Mullen. 2016. Effect of overlaid luminance contrast on perceived color contrast: Shadows enhance, borders suppress. *J. Vis.* 16, 11 (Sept. 2016), 15.

[28] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

[29] 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.

[30] 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.

[31] 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.

[32] Yuanzhen Li, Lavanya Sharan, and Edward H Adelson. 2005. Compressing and companding high dynamic range images with subband architectures. In *ACM SIGGRAPH 2005 Papers* (Los Angeles, California) *(SIGGRAPH '05)*. Association for Computing Machinery, New York, NY, USA, 836–844.

[33] Yucheng Liu and Jan P Allebach. 2014. A computational texture masking model for natural images based on adjacent visual channel inhibition. In *Image Quality and System Performance XI*, Vol. 9016. International Society for Optics and Photonics, 90160D.

[34] M A Losada and K T Mullen. 1994. The spatial tuning of chromatic mechanisms identified by simultaneous masking. *Vision Res.* 34, 3 (Feb. 1994), 331–341.

[35] 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).

[36] K T Mullen. 1985. The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. *J. Physiol.* 359 (Feb. 1985), 381–400.

[37] Leann Myers and Maria J Sirois. 2006. *Spearman correlation coefficients, differences between.* John Wiley & Sons, Inc., Hoboken, NJ, USA.

[38] Peter Neri. 2017. Object segmentation controls image reconstruction from natural scenes. *PLoS Biol.* 15, 8 (Aug. 2017), e1002611.

[39] 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.

[40] Rosalind Picard, Chris Graczyk, Steve Mann, Josh Wachman, Len Picard, and Lee Campbell. 1995. The MIT Vision Textures database.

[41] Thomas Porter and Tom Duff. 1984. Compositing digital images. *SIGGRAPH Comput. Graph.* 18, 3 (Jan. 1984), 253–259.

[42] Mahesh Ramasubramanian, Sumanta N Pattanaik, and Donald P Greenberg. 1999. A perceptually based physical error metric for realistic image synthesis. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH '99).* ACM Press/Addison-Wesley Publishing Co., USA, 73–82.

[43] Reuben Rideaux, Rebecca K West, Thomas S A Wallis, Peter J Bex, Jason B Mattingley, and William J Harrison. 2022. Spatial structure, phase, and the contrast of natural images. *J. Vis.* 22, 1 (Jan. 2022), 4.

[44] Christian Sandor, Andrew Cunningham, Arindam Dey, and Ville-Veikko Mattila. 2010. An Augmented Reality X-Ray system based on visual saliency. In *2010 IEEE International Symposium on Mixed and Augmented Reality (ISMAR).* Seoul, Korea (South), 27–36.

[45] Marc Ericson C Santos, Igor de Souza Almeida, Goshiro Yamamoto, Takafumi Taketomi, Christian Sandor, and Hirokazu Kato. 2016. Exploring legibility of augmented reality X-ray. *Multimed. Tools Appl.* 75, 16 (Aug. 2016), 9563–9585.

[46] Odelia Schwartz and Eero P Simoncelli. 2001. Natural signal statistics and sensory gain control. *Nat. Neurosci.* 4, 8 (2001), 819–825.

[47] Siwei Lyu and E P Simoncelli. 2008. Nonlinear image representation using divisive normalization. In *2008 IEEE Conference on Computer Vision and Pattern Recognition.* 1–8.

[48] Patrick C Teo and David J Heeger. 1994. Perceptual image distortion. In *Human Vision, Visual Processing, and Digital Display V*, Vol. 2179. International Society for Optics and Photonics, 127–141.

[49] Christoph Teufel, Steven C Dakin, and Paul C Fletcher. 2018. Prior object-knowledge sharpens properties of early visual feature-detectors. *Sci. Rep.* 8, 1 (July 2018), 10853.

[50] Okan Tarhan Tursun, Elena Arabadzhiyska, Marek Wernikowski, Radosław Mantiuk, Hans-Peter Seidel, Karol Myszkowski, and Piotr Didyk. 2019. A Luminance-Contrast-Aware Foveated Rendering. *ACM Trans. Graph.* (2019).

[51] Yunhai Wang, Jian Zhang, Wei Chen, Huai Zhang, and Xuebin Chi. 2011. Efficient opacity specification based on feature visibilities in direct volume rendering. *Comput. Graph. Forum* 30, 7 (Sept. 2011), 2117–2126.

[52] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 4 (April 2004), 600–612.

[53] Zhou Wang and Xinli Shang. 2006. Spatial pooling strategies for perceptual image quality assessment. In *2006 International Conference on Image Processing* (Atlanta, GA), Vol. 8. IEEE, 11.

[54] 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.

[55] A B Watson and J A Solomon. 1997. Model of visual contrast gain control and pattern masking. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* 14, 9 (Sept. 1997), 2379–2391.

[56] Wenjun Zeng, S Daly, and Shawmin Lei. 2000. Point-wise extended visual masking for JPEG-2000 image compression. In *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, Vol. 1. 657–660 vol.1.

[57] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. 2017. BAM! The Behance Artistic Media Dataset for Recognition Beyond Photography. In *2017 IEEE International Conference on Computer Vision (ICCV).* 1211–1220.

[58] 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.

[59] Sophie Wuerger, Maliha Ashraf, Minjung Kim, Jasna Martinovic, María Pérez-Ortiz, and Rafal K Mantiuk. 2020. Spatio-chromatic contrast sensitivity under mesopic and photopic light levels. *J. Vis.* 20, 4 (April 2020), 23.

[60] Nanyang Ye, Krzysztof Wolski, and Rafal K Mantiuk. 2019. Predicting visible image differences under varying display brightness and viewing distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 5434–5442.

[61] Jianming Zhang and Stan Sclaroff. 2013. Saliency detection: A boolean map approach. In *Proceedings of the IEEE international conference on computer vision.* 153–160.

[62] Lili Zhang and Michael J Murdoch. 2021. Perceived Transparency in Optical See-Through Augmented Reality. *Proceedings of 2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (Oct. 2021), 115–120.

[63] Lin Zhang, Ying Shen, and Hongyu Li. 2014. VSI: a visual saliency-induced index for perceptual image quality assessment. *IEEE Trans. Image Process.* 23, 10 (Oct. 2014), 4270–4281.

[64] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 586–595.

[65] Yunjin Zhang, Rui Wang, Evan Yifan Peng, Wei Hua, and Hujun Bao. 2021. Color Contrast Enhanced Rendering for Optical See-through Head-mounted Displays. *IEEE Transactions on Visualization and Computer Graphics* (2021), 1–1. https://doi.org/10.1109/TVCG.2021.3091686

[66] Lin Zheng, Yingcai Wu, and Kwan-Liu Ma. 2013. Perceptually-Based Depth-Ordering Enhancement for Direct Volume Rendering. *IEEE Trans. Vis. Comput. Graph.* 19, 3 (2013), 446–459.

A JUSTIFICATION OF THE EQUATION FOR DIVISIVE NORMALIZATION PROCESS

There were several choices in how we formulated the divisive inhibition process. The equation we used is based on the one proposed by Watson and Solomon [55] because of its generality. Note that our equation takes the same form as that used by Watson and Solomon if we raise their equation to the $1/\gamma$-th power and assume that $p = 1/\gamma$ and $q = 2$ (p and q are the symbols used in their equation). However, we limited the pooling of inhibition components to the spatial dimension only. We ignored the pooling across spatial frequency bands (i.e., different levels of the Laplacian pyramid) because previous studies have shown that the pooling across spatial frequency channels is limited to a small range [31, 55]. We also confirmed that incorporating the spatial-frequency pooling in our model led to a deterioration in the model's performance. We did not consider the orientation because we used non-oriented filters for the image representation. We also tested several different variations in the extent of spatial pooling and found that when the extent of the spatial pooling was larger than 5×5 performance was degraded, while smaller kernel sizes did not affect performance. In the equation, the parameters *gamma* and β (the scaling parameter of B) were the most sensitive to the performance for our dataset. Thus, we left them as free parameters to fit to the data. The parameters obtained by the fitting were validated on the psychophysically measured contrast masking data by Foley and Boynton [14] (See supplementary material).

B SCREENING PROCEDURE FOR THE VISIBILITY MATCHING EXPERIMENT IN SECTION 4

The data submitted by the crowdsourcing participants sometimes contained low-quality data. To exclude unreliable submissions, we designed a reliability measure for our experiment. In each session, there were always at least three conditions in which only the reference alpha value was varied while the combination of the reference/test image patches was the same. Within these conditions, it was expected that a reliable participant would tend to report a higher test alpha value as the reference alpha increased. To quantify this tendency, we computed the Spearman rank-order correlation between the reference alpha and the corresponding test alpha for each set of conditions with common image patches. Then, we averaged the correlation coefficients over all condition sets and used the mean value as a reliability measure for the participant.

Because there is no established way to determine the threshold, we heuristically decided to exclude from further analysis the data of participants whose reliability was lower than the 5th percentile of the entire distribution. The thresholds were 0.569 for the *Same* condition and 0.400 for the *Different* condition. Participants who fell below either of these thresholds were excluded. The number of participants removed by this procedure was 41.

C DETAILS OF THE USER STUDY IN SECTION 6.3

As described in the main paper, we set the visibility range for each method such that the range of alpha values of the test images approximately matched across the comparison methods. For the lower and upper visibility bounds, we used the average of the visibility values of all the test image pairs blended with $\alpha = 0.15$ and $\alpha = 0.85$, respectively. The target visibility values are summarized in Table 2. Four visibility levels selected from either the $\{1, 3, 5, 7\}$- or $\{2, 4, 6, 8\}$-th levels in Table 2 were tested for each image combination.

Table 2. Target visibility values of each comparison method tested in the user study. Raw visibility values before rescaling to 0-1 are presented.

Visibility level	1	2	3	4	5	6	7	8
Ours	2.5	3.1	3.8	4.4	5.1	5.8	6.4	7.1
W/o spatial adaptation	2.5	3.1	3.8	4.4	5.1	5.8	6.4	7.1
Static weights	2.4	2.9	3.4	4.0	4.5	5.1	5.6	6.1
Fukiage et al., 2014	1.2	1.7	2.1	2.6	3.1	3.6	4.0	4.5

Table 3. Parameters of generalized logistic function for each comparison method.

Method	p_1	p_2	p_3	p_4
Ours/full	-0.57836	6.0356×10^{9}	2.8279	17.838
Standard alpha blending	-9.3769	1.1836×10^{44}	95.697	311.26
Ours/spatially non-adaptive	-4.3835×10^{-2}	2.1853×10^{71}	19.901	104.28
Ours/static weights	-0.51609	13264.4	1.3670	7.3501
Fukiage et al., 2014 (IQA-based)	0.88202	13.762	0.81950	0.75775

The entire experiment consisted of 176 image combinations × 4 comparison methods × 4 visibility levels = 2,816 conditions. The conditions were randomly assigned to participants such that each participant was presented with conditions associated with 11 image combinations out of 176 combinations. Each participant rated the visibility for all four visibility levels for all the comparison methods for each image combination, which resulted in 176 conditions in total. All conditions were presented in a randomized order. In the beginning, each participant completed a practice session consisting of 12 trials with stimuli not used in the main test. The experiment took approximately 30 min on average, including the practice session and an instruction phase. A total of 209 participants were recruited via a crowdsourcing service (Prolific.ac).

To remove unreliable participants, we performed a screening process. As in Section B, the reliability was measured by computing the Spearman rank-order correlation between the target visibility levels and the rated visibility scores for the conditions using an identical target-reference pair. Only the conditions from the spatially non-adaptive optimization method were used for screening because in these conditions, the alpha map was always uniform, and the alpha value was thus guaranteed to increase with the target visibility in all areas. We excluded data from participants whose Spearman correlation was lower than the 5th percentile of the entire distribution, which was 0.819. In total, 11 participants were excluded by this procedure. After the screening, an average of 12.3 responses from different participants were collected for each condition.

D PARAMETERS OF LOGISTIC FUNCTION USED TO NORMALIZE TARGET VISIBILITY

As described in Section 6.3 of the main paper, we fit the generalized logistic function (Eq. 17) to model the relationship between the raw target visibility values x_v and subjectively rated visibility scores y_v obtained in User Study. The generalized logistic function was fitted to the data of each comparison method. The optimized parameters for each comparison method are shown in Table 3.