EUROGRAPHICS 2014 / B. Lévy and J. Kautz
(Guest Editors)

Volume 33 (2014), Number 2

Manipulating refractive and reflective binocular disparity

Ł. Dąbała[1,4] P. Kellnhofer[1] T. Ritschel[1,3] P. Didyk[2] K. Templin[1,2] K. Myszkowski[1] P. Rokita[4] H.-P. Seidel[1]

[1]MPI Informatik [2]MIT CSAIL [3]Saarland University [4]Warsaw University of Technology

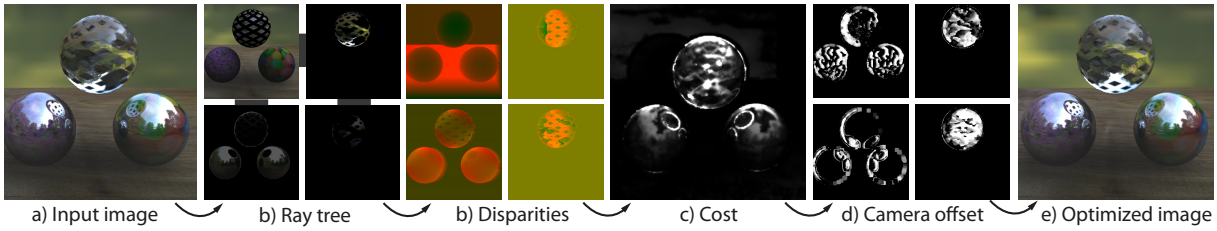

Figure 1: *We split a synthetic stereo image* (a) *into a ray tree* (b)*, and compute its multiple disparity interpretations* (c)*, which are put into a cost function* (d)*, that are used to find new camera setting* (e) *leading to an optimized combined stereo image* (f)*.*

Abstract

Presenting stereoscopic content on 3D displays is a challenging task, usually requiring manual adjustments. A number of techniques have been developed to aid this process, but they account for binocular disparity of surfaces that are diffuse and opaque only. However, combinations of transparent as well as specular materials are common in the real and virtual worlds, and pose a significant problem. For example, excessive disparities can be created which cannot be fused by the observer. Also, multiple stereo interpretations become possible, e. g., for glass, that both reflects and refracts, which may confuse the observer and result in poor 3D experience. In this work, we propose an efficient method for analyzing and controlling disparities in computer-generated images of such scenes where surface positions and a layer decomposition are available. Instead of assuming a single per-pixel disparity value, we estimate all possibly perceived disparities at each image location. Based on this representation, we define an optimization to find the best per-pixel camera parameters, assuring that all disparities can be easily fused by a human. A preliminary perceptual study indicates, that our approach combines comfortable viewing with realistic depiction of typical specular scenes.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Raytracing

1. Introduction

Current 3D display technology does not reproduce all depth cues, limiting the range of binocular disparity that can be presented on a screen. As a result, content needs to be manipulated beforehand, to assure sufficient viewing comfort. The most common manipulation is remapping disparity into the so called "comfort zone", i. e., a disparity range which can be comfortably viewed by a regular observer [LIFH09, SKHB11]. These techniques [JLHE01, LHW*10, DRE*11, etc.] are mostly concerned with disparities produced by diffuse opaque surfaces. Content with transparency and specular shading poses several challenges: First, *excessive absolute disparities*, both horizontal and vertical, are created and there is no way to control them. Second, multiple disparity interpretations at one location e. g., for glass, that reflects and refracts at the same time, result in an *excessive relative disparity*. No known method handles manipulation of multiple stereo interpretations. Additionally, *rivalry* i. e., fusion of different colors for the same scene space locations further degrades viewing experience. In this work, we propose a system to control disparity in computer generated images of such scenes (Fig. 1). First, we decompose the image into a ray-tree, where every bounce of light is represented by a single im-

age node. Next, we use a computer vision-inspired matching on deferred shading buffers to estimate disparities, as well as rivalry and importance for each image node. Finally, this information is used to optimization stereo rig parameters, to provide more comfortable disparity values and smaller binocular rivalry. As there is no definite perceptual model of what reflections are preferred or comfortable to fuse we provide an interactive user interface for intuitive control of specular stereo.

2. Background and Previous Work

In this section, we provide a background on stereo perception as well as an overview of stereoscopic image editing techniques which are related to our work.

Stereopsis In order to achieve a good perception of layout, the human visual system (HVS) combines information from many different cues (i. e., occlusion, perspective, motion parallax, etc.). One of the strongest cues, which also creates the most immersive experience, is stereopsis [Pal99, HR12]. Binocular vision provides two views of the environment from slightly different locations. As a consequence, the two retinal images of the same object are shifted relatively in both eyes. The magnitude of this displacement is used to estimate the position of objects in the 3D space. Although stereopsis is a very compelling cue and became widely used (e. g., in movies, video games, visualizations, TV, etc.), at the same time it introduces many problems. One of them is visual discomfort.

Visual discomfort in stereocopic images The most prominent source of visual discomfort is the interplay between the accommodation and vergence depth cues. While the first acquires depth information from kinesthetic sensations of relaxing and contracting intraocular muscles responsible for correct light focusing, the latter deduces a similar information from extraocular muscles which control the eye movements. Since these two mechanisms are linked to improve their performance, when stereoscopic 3D content is shown on a flat screen, a conflict between these mechanisms arises [LIFH09]. The HVS can usually tolerate this contradiction only in a small region around the screen (the so-called comfort zone) [SKHB11]. Limiting disparities to the comfort zone does not yet guarantee a good 3D perception. In order to perceive a clear three-dimensional object, both the left- and the right-eye images need to be fused. This requires solving a stereo-matching problem between left and right retinal images. The HVS performs the matching only in a small region around the horopter called Panum's fusional area [HR12, Ch. 14]. Beyond this region double vision occurs, which significantly impairs visual comfort. Visual discomfort can also be caused by significant differences between the left- and the right-eye images – so called binocular rivalry [HR12, Ch. 12]. In this case, the perceived image is not stable, but alternates between both views. This phenomenon occurs often in the real world, where it is a consequence of view-dependent light effects (e. g., specular highlights, refraction or glare) [TDR*12]. When differences between the left and right views are moderate, the HVS can fuse the images and tolerate the rivalry.

Stereo content manipulation In order to overcome above limitations, stereo content is often manipulated according to the display conditions [Men09]. For most 3D content, visual comfort is achieved by adjusting the depth range in the scene. Such manipulations are performed either during capturing by manipulating camera parameters [JLHE01, OHB*11, HGG*11], by extraction from a light field [KHH*11], or in a post-processing step [LHW*10, DRE*11]. There is, however, not much work addressing the problem of rivalry and fusion. The rivalry phenomenon has been exploited by Yang [YZWH12] for the purpose of tone mapping. They proposed to use dichoptic presentation to improve the "vividness" of common 2D images. Templin et al. [TDR*12] proposed a method for rendering highlights that facilitates their fusion. No method exists to address stereo content manipulation for more general image formation such as transparency and specular light transport. Any progress in this matter requires solving two challenging problems: the determination of disparities for multiple layers, and estimation of visually comfortable separations between the layers.

Disparity estimation for Lambertian surfaces For the purpose of 3D content manipulations, it is often desired to recover disparity between the left- and right-eye images. An important problem is to predict the HVS performance in stereo-matching, but it is commonly assumed, that some form of correlation between the left and right eye views is exploited [FB09]. Such correlation models proved to be useful in estimating viewing comfort when the left and right images are degraded independently (e. g., compression, blur) [RŚDD11]. Recently, perceptual metrics have been developed to account for the actually perceived disparity [DRE*11, DRE*12]. In contrast to the previous techniques, they assume that a depth map is given, and the perceived disparities are computed based on the HVS sensitivity to different disparity and luminance patterns. Although all above methods work well in many cases, they are limited to opaque diffuse surfaces. In more general scenarios, the problem of finding the correspondence between retinal images is not trivial and may lead to many different interpretations of depth, e. g., as in case of specular reflections or overlapping transparent surfaces.

Disparity of reflective and transparent surfaces There are only few techniques for disparity estimation, that allow for reflections and transparencies in the analyzed scene [TKS06, Siz08, SKG*12]. Here, separate optical flows for scene surfaces and reflected patterns are computed. These techniques are still limited to simple scenarios, e. g., just a single reflection or transparency layer and a piecewise planar image formation. Moreover, overcoming problems with

the accuracy of disparity reconstructed this way requires further research, while high computational costs reduce their practical applicability.

Relatively little is known how the HVS identifies the disparity information resulting from specular reflections. Blake and Bülthoff [BB90] suggest that the brain "knows" the physics of specular reflection and based on it the disparity originating from specular reflection is identified. Muryy et al. [MWBF13] observe that the disparity signals themselves provide the key information that enable to reject potentially unreliable disparity signals, e. g., due to large vertical disparities or horizontal disparity gradients. In the limit the disparity signal is lost completely in infusible regions. The transitions from fusible to infusible regions are not random, but rather exhibit specific binocular properties, which enable the brain to develop robust strategies of detecting abnormal disparity signals and judging their specular origin.

Perception research on stereo-transparency is mostly concerned with random dot stereograms (RDS), where it has been found that the discriminability of transparent layers is affected by increasing the number of layers, layer pooling has been observed for inter-layer disparity below 2 arcmin, and layer discrimination performance dropped for larger disparities, or too high density of dots per-layer [Wei89, TAW08]. In comparison with RDS, there is a number of additional cues in natural scene images, which facilitate the layer discrimination. However, there has been relatively little research on this problem. Akerstrom and Todd [AT88] report that perceptual segregation of overlapping transparent surfaces is facilitated, when depth planes are distinguished by color, but not when they are distinguished by element orientation.

3. Our Approach

Dealing with specular and transparent stereo has four key challenges, addressed in our approach.

First, the non-unique specular and transparent stereo image pair, where every pixel maps to many different depths, needs to be decomposed into a set of perceivable unique stereo image pair layers. While this is hard for real-world images, we demonstrate how such a separation can be obtained for rendered images. To this end we decompose the stereo image pair by its ray tree (Sec. 3.1). A node (or layer) in this tree is a stereo image of a certain indirection, e. g., the refraction, the reflection, the double-reflection, etc.

While disparity of opaque surfaces in computer-generated images is easily computed, it is difficult for transparent and specular light paths, that bend the geometry of rays and alter their disparity. Therefore, in our second step, we estimate disparity of each layer independently (Sec. 3.2). Computer vision-inspired matching is used to emulate human stereo matching, additionally identifying visual importance and rivalry for each layer.

The resulting multiple stereo interpretations suffer from three problems: excessive absolute disparity, excessive relative disparity and rivalry. Absolute disparity is altered, because reflection and refraction change the disparity arbitrarily e. g., beyond what is pleasant to fuse. The absolute disparity also contains a vertical component, which is another source of discomfort. Furthermore, the relative disparity between layers can be excessive for the same reason, consequently preventing fusion of a combination of reflection and refraction, such as found in a glass ball. To this end, we use the data computed before (disparity, rivalry, importance) as an input to four cost functions (Sec. 3.3), that for each layer assess its absolute disparity, its disparity relative to other layers, its rivalry and its similarity to the original disparity.

This cost function is eventually used to optimize per-pixel, per-layer camera parameters, in order to synthesize the optimal stereo image (Sec. 3.4).

3.1. Specular and transparent decomposition

We use the ray tree constructed during Whitted-style [Whi79] ray-tracing, to decompose a non-unique synthetic image with specularities and transparency into multiple unique images with importance weights.

Ray tree Let the image $L(\mathbf{x}) = \sum_{i=1}^{N} L_i(\mathbf{x})$ be the sum of N luminance images $\mathcal{L} = (L_1, \ldots, L_N \in \mathbb{R}^2 \to \mathbb{R}^3)$ which we call the *ray tree* of L (Fig. 2).

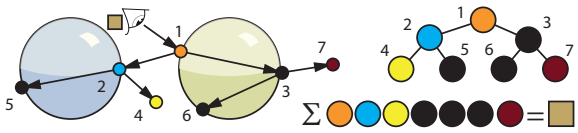

Figure 2: *Light path through a pixel in a scene with two reflecting and refracting spheres* (Left) *and its tree* (Right).

In our image formation model, a ray is either reflected or refracted, allowing to enumerate contributions in a simple scheme: index 1 denotes the diffuse contribution without reflection or refraction, 2 denotes one reflection, 3 is one refraction, 4 denotes one reflection followed by another reflection, and so on. The node L_i contains the diffuse radiance or the background color when there is no local hit point. The particular tree structure and node enumeration is not relevant in the rest of the algorithm and we will refer to the nodes in the tree, which are images, as *layers*. A typical value of N is $7 = 2^3 - 1$, i. e., a diffuse layer and two reflections or refractions at most. For a stereo image pair $(L^{\mathrm{l}}, L^{\mathrm{r}})$ a stereo ray tree is a pair $(\mathcal{L}^{\mathrm{l}}, \mathcal{L}^{\mathrm{r}})$ of two trees, one for each eye. Also, let $\mathcal{P}^{\mathrm{l}} = (p_1, \ldots, p_N \in \mathbb{R}^2 \to \mathbb{R}^3)$ and $\mathcal{P}^{\mathrm{r}}$ be the scene-space location of all surface intersections.

A custom interactive GPU ray-tracer [PBMH02] using BVHs [WBS07] is used to construct each layer following the common Whitted-style [Whi79] recursive ray-tracing, but

instead of point lights, we use image based lighting [Gre86] in combination with pre-computed ambient occlusion [ZIK98] for diffuse shading. Refractive rays are weighted according to Schlick's [Sch94] approximation. Note that this approach excludes all distribution effects such as motion blur and depth of field, but most prominently it excludes all glossy reflections or refractions.

3.2. Disparity estimation

Computing disparity for opaque surfaces from computer-generated images is simple, but intractable for specular, transparent or more general image formations (cf. Fig. 3).

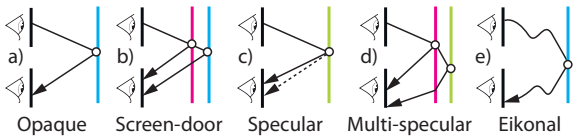

Figure 3: *Different image formation models (see text).*

In the diffuse case (Fig. 3, a), disparity at every pixel can be computed, by just projecting its scene-space position into the other eye's image. This projection depends on the distance to the diffuse object (blue line in Fig. 3, a). The only exception are occlusions, i. e., scene-space positions that are visible to only one eye. For non-CG images, such *stereo matching* i. e., finding this mapping is one of the most classic computer vision problems [SS02]. Still, once the mapping has been found, it is a simple projection due to rays being straight lines.

This principle becomes complicated in the presence of transparency (Fig. 3, b), when multiple scene-space depths (blue and pink line in Fig. 3, b) map to one pixel. Computer vision has addressed this problem, for the case of screen-door transparency, where all scene-space positions that contribute to the pixel happen to lie on a straight line [TKS06, SKG*12]. Using a CG decomposition into layers, this problem is easily addressed. Coincidentally matching different layers is ignored here.

The next complication occurs in the presence of rays that do change direction (Fig. 3, c) if they intersect a specular surface (gren line in Fig. 3, c). In this case, points that do not lie on a line have to be matched. The change of directions leads to disparity (Fig. 3, c, dotted line). For specific conditions, such as planar mirrors [SKG*12], curved reflectors [CKS*05, VAZBS08], water surfaces [Mur92] solutions have been proposed, but no general solution to this problem exists to our knowledge. Even in image synthesis, finding a correct mapping is challenging, even for a moderate number of planar refractors [WZHB09].

Complexity increases, when in addition to direction-changing rays transparency effects are also present (Fig. 3, d). Now, multiple scene points map to multiple image points with changing disparity.

In the most general, "eikonal" case (Fig. 3, e), such as hot gas or mixing fluids, rays do not even follow straight lines [SL96]. Still humans are indeed able to see such phenomena in stereo (Fig. 7). Also stereo matching is possible and computer vision has analyzed optical phenomena such as gas of varying temperature [IKL*10].

We do not account for even more general image formations, such as glossy transport (mapping a ray to a distribution of rays) or chromatic aberration (mapping different colors in a ray to different directions).

To simulate human stereo perception in all conditions using a practical algorithm, we propose to use a specialized stereo matching algorithm on the pair of layered deferred shading buffers $\mathcal{P}$ (possibly created using a non-standard image formation as seen in Fig. 3, b-e) to extract pixel disparity. This disparity serves as an upper bound on human performance, later to be limited by what the HVS actually can perceive, including a model of rivalry and occlusion. Each step is detailed in the following paragraphs.

Stereo matching To our knowledge, no practical method to predict human stereo matching from images, in particular for general image formation as explained before, exists. Classic optical flow [SS02] has difficulties finding the correct flow and requires substantial compute time. Our problem requires to compute flow at interactive rates to put the user into the loop. We therefore propose a specialized matching for stereo CG images.

The key observation is, that we can use any unique feature of a surface position associated with a pixel itself as the key value in stereo matching, because every layer contains only diffuse radiance. We simply use the 3D positions $\mathcal{P}$ stored in a layered deferred shading buffer, which is unique for orientable and self-intersection-free surfaces. We use the approach explained in Fig. 4: For each pixel in each layer, we know its position, i. e., where the ray from one eye intersected the surface. We assume this pixel will match with the pixel in the other eye that is most similar in terms of position i. e., equal up to numerical precision and pixel discretization. This pixel is found using a simple exhaustive search in a 2D neighborhood of pixels. A typical search of size 81×9 pixels ($\approx 24 \times 4$ arc min) can be performed at a rate of 0.6 megapixels per second in parallel over all pixels and layers on an Nvidia Quadro 4000. Note, that there is no other obvious way to invert a general image formation such as in Fig. 3, c–e. Output of this step are N layers of (potentially also vertical) pixel disparity $\mathcal{D} = (d_1, \ldots, d_N \in \mathbb{R}^2 \to \mathbb{R}^2)$, expressed in arcminutes. Since this matching substantially overestimates human performance, we need to limit its output to actually perceivable matches, as described below.

Human limitations The HVS uses luminance patterns to match pairs of corresponding points [FB09]. This process

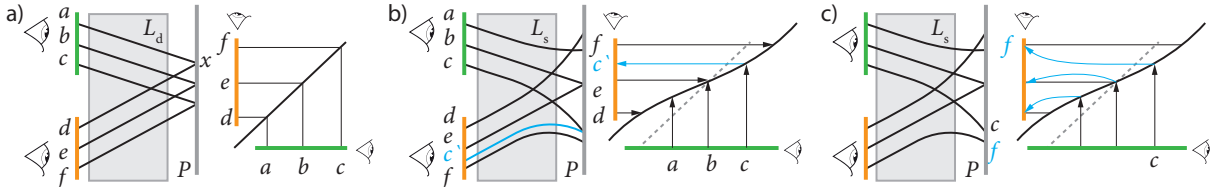

Figure 4: *Stereo vision with non-straight rays in the "flat-land".* (a) *In common stereo image formation L_d, rays go in straight lines. A point a in the left image* (green line) *at scene position x on a diffuse surface P is mapped to a point d in the right image* (yellow line) *by linear perspective. This mapping is simple to invert to establish a matching. The HVS uses luminance patterns to identify matches, but simply assuming a unique parametrization of P is an upper bound on the quality. Binocular stereo perception is solely a deviation from this line (not present in this example).* (b) *In the presence of specular reflections and refraction, rays do not go along straight lines, but change directions, diverge and converge, leading to a complex mapping L_s. Depth perception is created, even for a flat P, simply due to the bending of rays (deviation from the straight dotted line). Analytically inverting the mapping between the eyes, e. g., to map c in one image back to c' in the other image, is impossible in practice.* (c) *Our solution inverts the mapping numerically: c is matched with f which is closest in scene space (distance in P).*

has three main features: matches are proximal, distinctive, and their contrast has to be sufficient. To quantify this, we compute *importance* $\mathcal{W} = (w_1, \ldots, w_N \in \mathbb{R}^2 \to (0,1))$ for each pixel on each layer as a product of the following three factors.

First, in terms of proximity, matching is performed horizontally in a range of 3 deg [SKHB11] and in a much smaller range of 15 arc min vertically [TLA*12]. We account for this, using a limited search window that is much wider than high (81×9 pixels). If no match smaller than a threshold ε is found, it is considered a (generalized) disocclusion, and the proximity factor is zero. The choice of ε depends on the image resolution and the scene size in scene space; in our experiments ε equals 1/100 of the scene diameter.

Second, the luminance structure must be distinct to be matched. To detect this, we run an SSD metric $\alpha(\mathbf{x}, \mathbf{x}+d_i(\mathbf{x}))$ with a windows size of 3×3 pixels between the CIE LAB luminance of the two images L_i^{l} and L_i^{r} at locations $\mathbf{x}$ and $\mathbf{x}+d_i(\mathbf{x})$ as well as between the location $\mathbf{x}$ and similar alternative disparities $\alpha(\mathbf{x}, \mathbf{x}+d_i(\mathbf{x})+\mathbf{y})$, where $\mathbf{y} \in (-5, \ldots, 5)^2 \subseteq \mathbb{N}^2$. The discrete curvature $\kappa\alpha_i$ of the energy landscape when varying $\mathbf{y}$ is used to rate the match [SS02]: Only if it is high, we found a unique feature the HVS could match.

Third, the HVS needs a minimal luminance contrast to perform matching. As the contrast factor, we use a function similar to Cormack's Q function [CSS91], that equals 0 for a luminance contrast of 0, and smoothly blends into 1 for contrasts above 10 JND. Contrast is computed as the standard deviation of luminance (Peli's contrast [Pel90]) in a 17×17-pixel, Gaussian-weighted window.

Rivalry The main assumptions of our matching approach was that the HVS might fail to match (w low, d correct), but does not create false positive matches (w high, d incorrect). Therefore, we will never have rivalry due to wrong matches: the content of each layer is diffuse, so a scene point has the same color from all points of view, and only points that have the same scene-space locations are matched. However, rivalry may still occur for two other reasons: missing matches / occlusions and because of image structure.

Whenever no match has been detected in the neighborhood – either due to occlusion, or because the match is too distant – we extrapolate matches of nearby pixels, using hole filling. This step is inspired by the filling-in mechanism performed by the HVS [Kom06]. Next, we compare the luminance values of the pixels matched, and report rivalry if differences are too high.

Even if there are no missing matches, rivalry may be still present solely because of the image structure. The HVS does not compare individual points, but finite image patches. Even if pixel colors in the center of the flow are identical, the structure might be different, i. e., rotated, scaled or changed in brightness (Fig. 5). Therefore, we compare the local structure by computing a 7×7 SSD of the luminance of the stereo match. While more advanced models of binocular rivalry exist [YZWH12] they only provide binary values for luminance images in mono. Output of this step is a tree of rivalry images $\mathcal{R} = (r_1, \ldots, r_N \in \mathbb{R}^2 \to \mathbb{R})$.

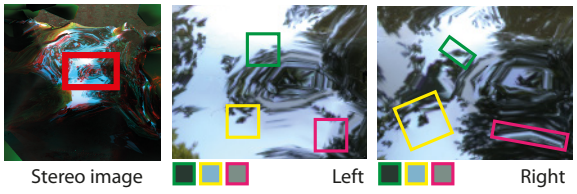

Figure 5: *Rivalry detection: even after our "perfect" matching of a stereo image pair* (left) *and the color difference are zero* (colored rectangles)*, the structure around each matching pair can differ and cause rivalry* (right).

Tone mapping / luminance adaptation Special consideration has to be taken, as the input is an HDR image tree pair. Directly performing stereo matching on the HDR image pairs of each layer, would result in an overestimate of human stereo matching performance, as it would match details that are effectively not perceived, such as in overly bright sky regions (that appear as featureless white) or overly dark regions such as shadows (that appear as featureless black). Also the limitations of the display have to be accounted for: features that will not be reproduced by the screen should also not be used for stereo matching. Consequently, we assume a certain adaptation and display range, and tone-map the HDR image tree pair to LDR using a linear tone mapper filter it by the human CSF [Rob66] before stereo matching. While this works well in practice, and no imperceptible details "pollute" the stereo matching, the underlying mechanics of contrast perception and stereo matching would require further investigation.

3.3. Cost function

The disparity and rivalry computed in the previous step models human perception in a "neutral" way: a certain degree of vertical disparity or rivalry can be good or bad, depending on the scene, display device or the stereographer's objective. We use cost functions to map the neutral physiological information to a cost. The cost is defined on the vector of all disparities $\mathbf{d} \in \mathbb{R}^{2N}$, rivalries $\mathbf{r} \in \mathbb{R}^{N}$ and importances $\mathbf{w} \in \mathbb{R}^{N}$ from all layers and in each pixel.

Disparity The disparity cost combines an *absolute* and a *relative* contribution.

The *absolute* disparity cost function prevents disparity from exceeding the comfort zone, relative to the screen plane [LHW*10]. A typical application is to reduce the disparity of a curvy refracting object, such as a glass ball that magnified disparity. It is defined for horizontal and/or vertical disparity as $f_\mathrm{a}(\mathbf{d},\mathbf{w}) = ||\mathsf{W}\mathbf{d}||^2$, where $\mathsf{W} = \mathrm{diag}(\mathbf{w})$ is an importance matrix.

The *relative* (pairwise) cost considers all pairs of disparity. Typically, it is used, to move a reflection or refraction closer to a reference depth, e. g., make absolute disparity of a diffuse surface and a reflection on it similar. It is defined as $f_\mathrm{p}(\mathbf{d},\mathbf{w}) = ||\mathrm{diag}(\mathsf{Q}\mathbf{w})(\mathsf{P}\mathbf{d} - \mathbf{d}_\mathrm{p})||^2$ where $\mathsf{P} \in \mathbb{R}^{2N\times 4N^2}$ is a matrix constructing differences of all element pairs in a vector, $\mathsf{Q} \in \mathbb{R}^{2N\times 4N^2}$ is a matrix constructing the product of all element pairs in a vector and $\mathbf{d}_\mathrm{p} \in \mathbb{R}^{2N} = (\lambda_\mathrm{x}, \lambda_\mathrm{y}, \lambda_\mathrm{x}, \lambda_\mathrm{y}, \ldots)$. We set $\lambda_\mathrm{x} = 3\,\mathrm{arc\,min}$ and $\lambda_\mathrm{y} = 0\,\mathrm{arc\,min}$ to prefer small horizontal relative disparities.

Rivalry In general rivalry needs to be avoided so the cost $f_\mathrm{r}(\mathbf{r},\mathbf{w}) = ||\mathsf{W}\mathbf{r}||^2$ is a quadratic potential.

Data term Adding a small data term $f_\mathrm{d}(\mathbf{d}) = ||\mathsf{W}(\mathbf{d} - \mathbf{d}_\mathrm{d})||^2$ to the cost function ensures that the optimization prefers the original disparity $\mathbf{d}_\mathrm{d}$, if multiple choices are equally good.

Aggregation All costs are independent between pixels but interdependent between layers. The costs combine as in

$$f(\mathbf{d},\mathbf{r},\mathbf{w}) = \alpha_\mathrm{a} f_\mathrm{a}(\mathbf{d},\mathbf{w}) + \alpha_\mathrm{p} f_\mathrm{p}(\mathbf{d},\mathbf{w}) + \alpha_\mathrm{r} f_\mathrm{r}(\mathbf{r},\mathbf{w}) + \alpha_\mathrm{d} f_\mathrm{d}(\mathbf{d},\mathbf{w}),$$

where $\alpha_{\{\mathrm{a,p,r,d}\}}$ are user-specified weights to control the respective cost. The result of different control settings are shown in Fig. 6.

Our optimization solves for an acceptable combination of multiple conflicting goals (Fig. 6a). Low disparity would provide comfort and avoid rivalry, but provides no depth impression and vice versa. In general α_r is set to the highest value, assuring that rivalry is removed (Fig. 6b). The weight α_d of the data term is contradicting and set to a lower value (Fig. 6c). It is mostly required to pick a solution that is close to the original if multiple solutions are equally good. The weight α_a of absolute disparity reduces rivalry but has to be set to an intermediate value as its term tends to make the scene look flat. Its direct counterpart, the weight of separation α_p, uses the same value to keep a balance (Fig. 6d). Choosing similar weights allows these two terms to produce a distribution of disparity allocating retargeted disparity budget to all layers with a similar inter-layer distributions as in the original content. For typical scenes with mixed light effects such as *Glasses* (Fig. 6a) or *Snowball* (Fig. 8) the weights are $\alpha_{\{\mathrm{a,p,r,d}\}} = 1, 1, 10, 0.1$. For some specific scenes with little rivalry or large disparity difference between layers (e. g. *Skull* or *Chinese Paper* in Fig. 8) we make the data term stronger ($\alpha_\mathrm{d} = 1$) to preserve more of the original depth. Weights can also be spatially refined to strengthen individual terms using a simple painting interface as seen in the supplemental video.

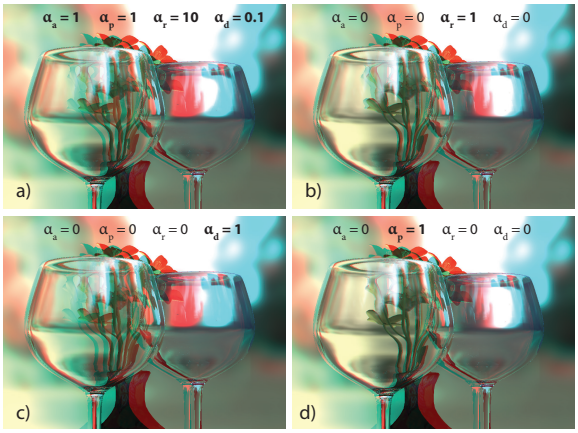

Figure 6: (a) *Our optimization with all costs active, balancing the disparity distortion and residual rivalry.* (b) *Using only the rivalry term, the disparity is locally reduced where rivalry would be unacceptable.* (c) *Using only the data term, rivalry appears in reflections of the right glass and refraction in the left glass.* (d) *When using only separation stereo almost collapses to diffuse disparity.*

3.4. Optimization

The cost defined above, can be used to compare two stereo image pairs created by a general image formation. To produce an image pair that has low cost and therefore high stereo quality, the camera used for shading of the left image is moved along a line, connecting the left and right eye position independently for each pixel and each layer. In particular, if the offset of the camera is reduced to 0, no stereo effects are present (the layer is mono), and if original offset is used, maximal stereo effects occur for that layer.

Let $g : \mathbb{R}^N \to \mathbb{R}^{2N} \times \mathbb{R}^N \times \mathbb{R}^N$ be the mapping from a camera offset for all layers $\mathbf{y} \in \mathbb{R}^N$ to perceived stereo parameters (disparity, rivalry, importance, as described in Secs. 3.1 and 3.2) we like to minimize the cost function $f(g(\mathbf{y}))$.

To this end, we first tabulate g for $k = 32$ inputs in the form $(0/k, 1/k, \ldots, k/k)$. Here, the computation time is dominated by the requirement to raytrace the scene many times and can take up to one minute. The tabulated g has similarities to a light field [KHH*11], but contains disparity instead of radiance and multiple instead of a single layer. The optimization itself can then be performed at interactive rates when a user adjusts a parameter, e. g., by painting weights.

We use gradient descent with restart for optimization. While f is quadratic in its parameters and its derivative can be computed analytically, g involves ray-tracing, matching and our vision model, which do not have an analytic derivative. As a solution in every step of gradient descent we first differentiate g numerically and then apply the analytic gradient of f. We use 32 steps of gradient descent. As g might have several local minima $f(g(\mathbf{y}))$ might have as well. As a solution we restart the optimization four times with different initial camera offsets $\mathbf{y}_i = (1, i/3, \ldots, i/3), i \in \{0, 1, 2, 3\}$. The camera offset for the first diffuse layer is constant and not optimized.

For regularization, the resulting optimal camera offset is blurred using an edge-aware filter. This optimization is performed in parallel, independently for every pixel. The solution is found at lower resolution and later up-sampled [KCLU07]. In this step, the contribution of each pixel is weighted by its importance, i. e., unimportant pixels with mostly meaningless solutions do not affect their neighbors. Once the optimization has finished, we render the scene one last time in RGB using the optimal camera settings. We observed a good temporal stability of the optimization under camera motion as can be seen in the supplemental video.

4. Evaluation

In this section we present results of our approach, which are used as stimuli in a perceptual study.

4.1. Results

A collection of typical results of our approach, which are also used in our perceptual study, are shown in Fig. 8. We compare to a simple solution based on a small constant offset of reflection/refraction layers, which places them near the surface of the objects, in a similar fashion to the solution proposed for reflections by Templin et al. [TDR*12].

The first column (Fig. 8) shows, how our approach can retarget excessive disparity in the presence of transparent surfaces. The dragon behind the curtain is not fusible in the original image. Enforcing on-surface or near-surface disparity removes the stereo impression from all objects behind the curtain. Our approach produces an image where objects behind the curtain have the appropriate amount of disparity.

Similar observations can be made within the second column (Fig. 8), but for refracted and reflected rays: The reflections on the water surface cause rivalry, and the disparity of the riverbed is excessive. Competing approaches can reduce rivalry, but lose the depth difference between the surface and the riverbed (consider the floating leaves). Our approach reduces rivalry of the reflections as well as preserves the contrasting depth of the riverbed while keeping it fusible.

The third column (Fig. 8) shows a similar setting, where the shape of the statue behind multiple layers of reflection and refraction is lost.

In the fourth column (Fig. 8), our approach combines the skull, the dirt and the mirror as well as the envmap into a fusible image, while both on-surface and near-surface reflections make the mirror appear as flat as a diffuse poster. Such an artifact was indicated as a limitation of methods based on a constant shift of layers by Templin et al. [TDR*12].

Fig. 7 shows our approach applied to other non-standard image formation models, such as Eikonal light transport [SL96] and fish-eye lenses, which both lead to excessive and vertical disparity, as well as rivalry on multiple layers. For Eikonal light transport, we defined the "diffuse layer" p_0^1 to be the first intersection where a ray changes direction.

4.2. Study

We conducted a preliminary perceptual study to compare our approach to other alternatives in terms of visual comfort, realistic scene reproduction and overall preference. Seven participants with tested stereo vision took part in the experiment. Prior to the experiment, participants received an instruction sheet stating the purpose of the experiment. This sheet contained an illustrated explanation of the terms "stereoscopic fusion" and "realistic 3D impression" (see the supplemental material for details). Images of six different 3D scenes were used as stimuli covering typical reflection and refraction cases (Figs. 6, 7(left), and 8). Each scene was shown to participants in four different versions corresponding to different methods. The task was to arrange the images in increasing order, first,

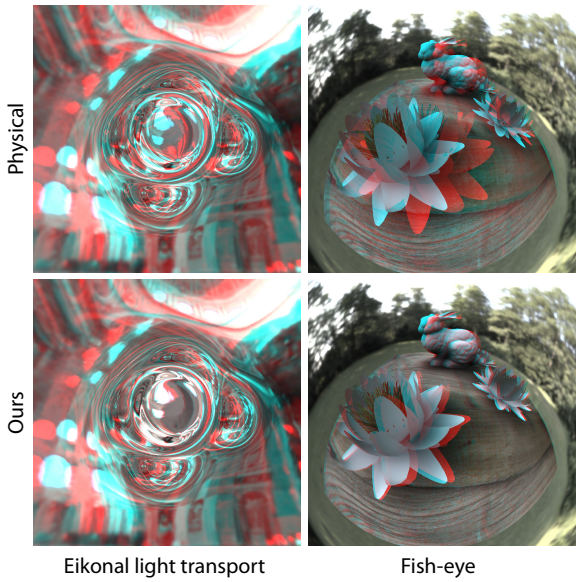

Figure 7: *Results of our approach for Eikonal light transport in a volumetric medium of varying index of refraction (e. g., temperature) and a fish-eye lens rendering.*

according to the ease of stereo fusion, next, to the realism of 3D impression, and eventually, to the overall preference. All images were presented in a randomized horizontal order, at a distance of 60 cm on a Samsung SyncMaster 2233RZ display (1680×1050 pixels, 1000:1 contrast, office lighting conditions), using NVIDIA 3D Vision active shutter glasses. We performed Wilcoxon signed-rank testing to reject the null hypothesis ($p < .05$) that the median difference of ranks is zero and therefore there is no effect (Fig. 9).

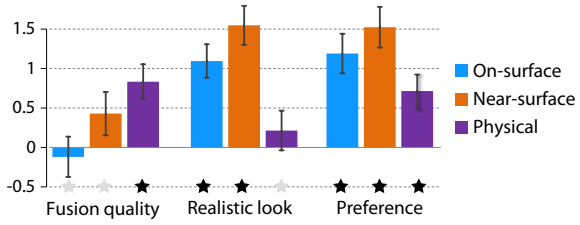

Figure 9: *User study results depicted as mean rank differences between our approach and other approaches for different questions. A high value indicates, that the competing approach ranks behind our approach, a low or negative value indicates a weaker or reverted rank difference. The error bars show standard error of this mean rank difference. Significant ($p < 0.05$) comparisons are marked by a star.*

The study showed that the results produced using our method are easier to fuse than *Physical* rendering. For realistic 3D reproduction, our method outperforms both the *On-surface* and the *Near-surface* method. In terms of overall preference, our technique produces results better than all other methods. While our approach is not always significantly better in all regards than all competitors (but also not significantly worse), we conclude, that our technique can provide a good trade-off between the ease of stereo fusion and realism of depiction.

It may be surprising that the near-surface solution was judged worse than the on-surface and the physical solutions. We hypothesize that this was caused by not accounting for sharp transitions in material properties, which produces numerous artifacts visible particularly in the "River" and "Mirror" scenes. Adding an edge detector on material properties layer would help mitigate this issue.

5. Conclusion

In this work, we approached the problem of manipulating binocular disparity of multiple reflections and refractions, assuming either standard or non-standard image formation model. We adapted computer vision-based stereo matching algorithm to predict what patterns would be fused by the human visual system in the case of multiple disparities present at a single location. The resulting disparity and rivalry were fed into an optimization framework to improve fusibility of synthetic stereoscopic content, while preserving realism of depiction. The approach can react to artist control at interactive rates, and its outcome is significantly preferred when compared to simpler alternatives.

Currently, the main limitation is the lack of support for distribution effects, such as glossy reflections and refraction, but also depth of field and motion blur. Also our optimization is performed on pixel disparity, although the human visual system is sensitive to changes of pixel disparity / vergence of a certain frequency, relative to a reference. Additionally, our optimization ignores cross-talk between layers that might lead to additional false positive stereo percepts. To be applicable to non-CG images and video, further progress in detection of multiple specular flows would be required.

As mentioned, the idea could be extended to distribution effects, or to actively change the scene such that depth perception is optimized.

References

[AT88] AKERSTROM R. A., TODD J. T.: The perception of stereoscopic transparency. *Attention, Perception, and Psychophysics 44*, 5 (1988), 421–432. 3

[BB90] BLAKE A., BÜLTHOFF H.: Does the brain know the physics of specular reflection? *Nature 343*, 6254 (1990), 165–168. 3

[CKS*05] CRIMINISI A., KANG S. B., SWAMINATHAN R., SZELISKI R., ANANDAN P.: Extracting layers and analyzing their specular properties using epipolar-plane-image analysis. *Comp. Vis. and Im. Underst. 97*, 1 (2005), 51–85. 4

[CSS91] CORMACK L. K., STEVENSON S. B., SCHOR C. M.: Interocular correlation, luminance contrast and cyclopean processing. *Vis. Res. 31*, 12 (1991), 2195–07. 5

[DRE*11] DIDYK P., RITSCHEL T., EISEMANN E., MYSZKOWSKI K., SEIDEL H.: A perceptual model for disparity. *ACM Trans. Graph. (Proc. SIGGRAPH) 30*, 4 (2011), 96:1–96:10. 1, 2

[DRE*12] DIDYK P., RITSCHEL T., EISEMANN E., MYSZKOWSKI K., SEIDEL H.-P., MATUSIK W.: A luminance-contrast-aware disparity model and applications. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 31*, 6 (2012), 184:1–184:10. 2

[FB09] FILIPPINI H. R., BANKS M. S.: Limits of stereopsis explained by local cross-correlation. *J Vis. 9*, 1 (2009). 2, 4

[Gre86] GREENE N.: Environment mapping and other applications of world projections. *IEEE Comp. Graph. and App. 6*, 11 (1986), 21–29. 4

[HGG*11] HEINZLE S., GREISEN P., GALLUP D., CHEN C., SANER D., SMOLIC A., BURG A., MATUSIK W., GROSS M.: Computational stereo camera system with programmable control loop. *ACM Trans. Graph. (Proc. SIGGRAPH) 30*, 2 (2011), 94:1–94:10. 2

[HR12] HOWARD I., ROGERS B.: *Perceiving in Depth, Volume 2: Stereoscopic Vision*. OUP USA, 2012. 2

[IKL*10] IHRKE I., KUTULAKOS K. N., LENSCH H., MAGNOR M., HEIDRICH W.: Transparent and specular object reconstruction. *Comp. Graph. Forum 29*, 8 (2010), 2400–26. 4

[JLHE01] JONES G., LEE D., HOLLIMAN N., EZRA D.: Controlling perceived depth in stereoscopic images. In *SPIE* (2001), vol. 4297, pp. 42–53. 1, 2

[KCLU07] KOPF J., COHEN M., LISCHINSKI D., UYTTENDAELE M.: Joint bilateral upsampling. *ACM Trans. Graph. (Proc. SIGGRAPH) 26*, 3 (2007), 96:1–96:6. 7

[KHH*11] KIM C., HORNUNG A., HEINZLE S., MATUSIK W., GROSS M.: Multi-perspective stereoscopy from light fields. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 30*, 6 (2011), 190:1–190:10. 2, 7

[Kom06] KOMATSU H.: The neural mechanisms of perceptual filling-in. *Nature reviews neuroscience 7*, 3 (2006), 220–231. 5

[LHW*10] LANG M., HORNUNG A., WANG O., POULAKOS S., SMOLIC A., GROSS M.: Nonlinear disparity mapping for stereoscopic 3D. *ACM Trans. Graph. (Proc. SIGGRAPH) 29*, 4 (2010), 75. 1, 2, 6

[LIFH09] LAMBOOIJ M., IJSSELSTEIJN W., FORTUIN M., HEYNDERICKX I.: Visual discomfort and visual fatigue of stereoscopic displays: A review. *J Imag. Sci. and Tech. 53*, 3 (2009), 1–12. 1, 2

[Men09] MENDIBURU B.: *3D movie making: stereoscopic digital cinema from script to screen*. Focal Press, 2009. 2

[Mur92] MURASE H.: Surface shape reconstruction of a nonrigid transparent object using refraction and motion. *IEEE PAMI 14*, 10 (1992), 1045–52. 4

[MWBF13] MURYY A. A., WELCHMAN A. E., BLAKE A., FLEMING R. W.: Specular reflections and the estimation of shape from binocular disparity. *Proc. of the National Academy of Sciences 110*, 6 (2013), 2413–2418. 3

[OHB*11] OSKAM T., HORNUNG A., BOWLES H., MITCHELL K., GROSS M.: OSCAM-optimized stereoscopic camera control for interactive 3D. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 30*, 6 (2011), 189:1–189:8. 2

[Pal99] PALMER S. E.: *Vision Science: Photons to Phenomenology*. The MIT Press, 1999. 2

[PBMH02] PURCELL T. J., BUCK I., MARK W. R., HANRAHAN P.: Ray tracing on programmable graphics hardware. *ACM Trans. Graph. (Proc. SIGGRAPH) 21*, 3 (2002), 703–712. 3

[Pel90] PELI E.: Contrast in complex images. *JOSA A 7*, 10 (1990), 2032–40. 5

[Rob66] ROBSON J.: Spatial and temporal contrast-sensitivity functions of the visual system. *JOSA 56*, 8 (1966), 1141–2. 6

[RŚDD11] RICHARDT C., ŚWIRSKI L., DAVIES I. P., DODGSON N. A.: Predicting stereoscopic viewing comfort using a coherence-based computational model. In *Computational Aesthetics* (2011), pp. 97–104. 2

[Sch94] SCHLICK C.: An inexpensive brdf model for physically-based rendering. *Comp. Graph. Forum 13*, 3 (1994), 233–46. 4

[Siz08] SIZINTSEV M.: Hierarchical stereo with thin structures and transparency. In *CVPR* (2008), pp. 97–104. 2

[SKG*12] SINHA S. N., KOPF J., GOESELE M., SCHARSTEIN D., SZELISKI R.: Image-based rendering for scenes with reflections. *ACM Trans. Graph. (Proc. SIGGRAPH) 31*, 4 (2012), 100. 2, 4

[SKHB11] SHIBATA T., KIM J., HOFFMAN D., BANKS M.: The zone of comfort: Predicting visual discomfort with stereo displays. *J Vis. 11*, 8 (2011). 1, 2, 5

[SL96] STAM J., LANGUÉNOU E.: Ray tracing in non-constant media. In *Proc. EGSR*. 1996, pp. 225–34. 4, 7

[SS02] SCHARSTEIN D., SZELISKI R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *J Comp. Vis. 47*, 1-3 (2002), 7–42. 4, 5

[TAW08] TSIRLIN I., ALLISON R. S., WILCOX L. M.: Stereoscopic transparency: Constraints on the perception of multiple surfaces. *J Vis. 8*, 5 (2008). 3

[TDR*12] TEMPLIN K., DIDYK P., RITSCHEL T., MYSZKOWSKI K., SEIDEL H.-P.: Highlight microdisparity for improved gloss depiction. *ACM Trans. Graph. (Proc. SIGGRAPH) 31*, 4 (2012), 92. 2, 7

[TKS06] TSIN Y., KANG S. B., SZELISKI R.: Stereo matching with linear superposition of layers. *IEEE PAMI 28*, 2 (2006), 290–301. 2, 4

[TLA*12] TYLER C. W., LIKOVA L. T., ATANASSOV K., RAMACHANDRA V., GOMA S.: 3D discomfort from vertical and torsional disparities in natural images. In *Proc. SPIE* (2012). 5

[VAZBS08] VASILYEV Y., ADATO Y., ZICKLER T., BEN-SHAHAR O.: Dense specular shape from multiple specular flows. In *CVPR* (2008), pp. 1–8. 4

[WBS07] WALD I., BOULOS S., SHIRLEY P.: Ray tracing deformable scenes using dynamic bounding volume hierarchies. *ACM Trans. Graph. 26*, 1 (2007), 6. 3

[Wei89] WEINSHALL D.: Perception of multiple transparent planes in stereo vision. *Nature 341*, 6244 (1989), 737–739. 3

[Whi79] WHITTED T.: An improved illumination model for shaded display. *ACM SIGGRAPH Computer Graphics 13* (1979), 14. 3

[WZHB09] WALTER B., ZHAO S., HOLZSCHUCH N., BALA K.: Single scattering in refractive media with triangle mesh boundaries. *ACM Trans. Graph. (Proc. SIGGRAPH) 28*, 3 (2009), 92. 4

[YZWH12] YANG X., ZHANG L., WONG T.-T., HENG P.-A.: Binocular tone mapping. *ACM Trans. Graph. (Proc. SIGGRAPH) 31*, 4 (2012), 93:1–93:10. 2, 5

[ZIK98] ZHUKOV S., IONES A., KRONIN G.: An ambient light illumination model. In *Proc. EGSR*. 1998, pp. 45–55. 4

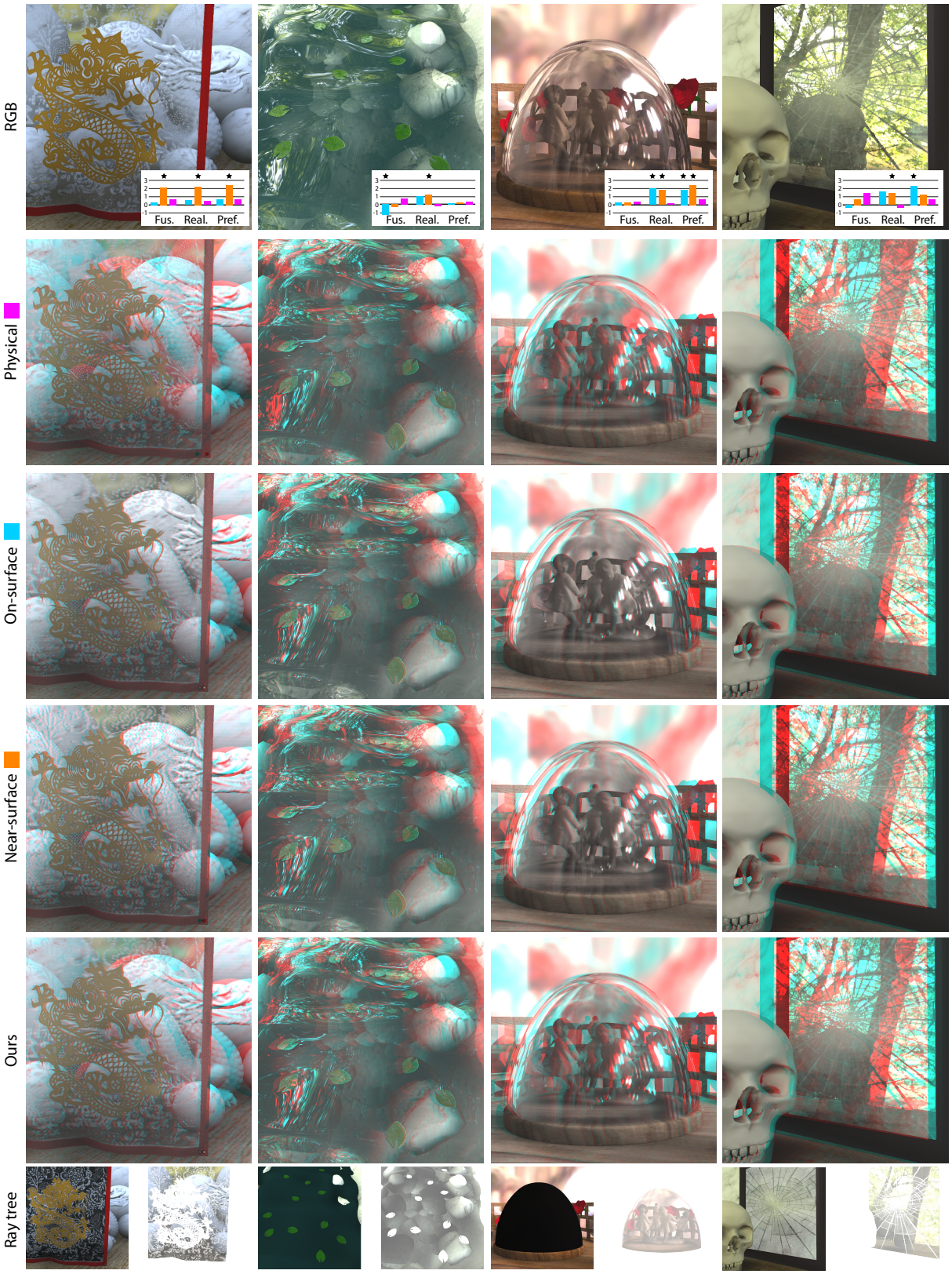

Figure 8: *Results of different approaches* (rows) *in scenes* (columns) *used in the study and discussed in Sec. 4.*