Perceptual real-time 2D-to-3D Conversion using Cue Fusion

Thomas Leimkühler, Petr Kellnhofer, Tobias Ritschel, Karol Myszkowski, and Hans-Peter Seidel

Abstract—We propose a system to infer binocular disparity from a monocular video stream in real-time. Different from classic reconstruction of physical depth in computer vision, we compute perceptually plausible disparity, that is numerically inaccurate, but results in a very similar overall depth impression with plausible overall layout, sharp edges, fine details and agreement between luminance and disparity. We use several simple monocular cues to estimate disparity maps and confidence maps of low spatial and temporal resolution in real-time. These are complemented by spatially-varying, appearance-dependent and class-specific disparity prior maps, learned from example stereo images. Scene classification selects this prior at runtime. Fusion of prior and cues is done by means of robust MAP inference on a dense spatio-temporal conditional random field with high spatial and temporal resolution. Using normal distributions allows this in constant-time, parallel per-pixel work. We compare our approach to previous 2D-to-3D conversion systems in terms of different metrics, as well as a user study and validate our notion of perceptually plausible disparity.

Index Terms—Depth cues, Stereo, Image-based rendering, Perceptual reasoning, Video analysis, Viewing algorithms, Pixel classification, Real-time systems.

1 Introduction

The majority of images and videos available is 2D and automatic conversion to 3D is a long-standing challenge [1]. For applications such as view synthesis, for surveillance, autonomous driving, human body tracking, relighting or fabrication, accurate physical depth is mandatory, and obviously binocular disparity can be computed from such data, resulting in a perfect stereo image pair. However, for 2D-to-3D stereo conversion, such physical depth is not required. Instead, we seek to compute perceptually plausible disparity in this work. It differs from physical depth by three properties. First, the absolute scale of disparity is not relevant, and any reasonable smooth remapping [2], [3] is perceived equally plausible and may even be preferred in terms of viewing comfort and realism. Second, the natural statistics of depth and luminance indicate that depth is typically spatially smooth, except at luminance discontinuities [4], [5]. Therefore, not reproducing disparity details can be acceptable and is often not even perceived, except at luminance edges [6]. Third, the temporal perception of disparity allows for a temporally coarse solution, as fine temporal variations of disparity are not perceivable [6], [7]. Consequently, as long as the error is 2D-motion compensated, depth from one point in time can be used to replace depth at a different, nearby point in time.

Our method is modular (Sec. 3) and based on priors learned in a pre-process (Sec. 3.1) combined with stereo cues extracted from 2D images or videos at runtime (Sec. 3.2). Both priors and cues are represented as normal distributions allowing to fuse a plausible disparity map with high spatial and temporal resolution in real-time (Sec. 3.3). Image-based rendering produces a stereo video stream from this map (Sec. 3.4). In Sec. 4 we validate our notion of perceptually plausible disparity and discuss our results. We find that our system can perform 2D-to-3D conversion at ca. 35 Hz for HD video and compares favorable to off-line methods in terms of different error metrics as well as user ratings. In summary, our contributions are

- a real-time 2D-to-3D conversion system based on the fusion of learned priors and depth cues into a coherent disparity estimate,
- an analysis of the importance of different depth cues in different scenes based on estimated confidence, and
- a perceptual analysis of disparity plausibility, including spatial and temporal sampling requirements for perceptual disparity processing tasks.

2 Previous Work

In this section, we review manual and automatic approaches for 2D-to-3D with an emphasis on real-time conversion, the use of luminance and depth edges in computational stereo, as well as perceptual modeling of binocular and monocular depth cues.

2.1 2D-to-3D conversion

Manual conversion produces high-quality results but requires human intervention, which can result in substantial cost. It is based on painting depth annotations [8] with special user interfaces [9] and propagation in space and time [10]. The semi-supervised method of Assa and Wolf [11] combines cues extracted from an image with user intervention to create depth parallax. User intervention can be included in our approach as an additional depth cue.

Automatic conversion does not induce manual effort, but results in long computation times to produce results

- *T. Leimkühler, K. Myszkowski and H.-P. Seidel are with MPI Informatik. E-Mail: {tleimkueh, karol, hps}@mpi-inf.mpg.de*
- *P. Kellnhofer is with MIT CSAIL. E-Mail: pkellnho@mit.edu*
- *T. Ritschel is with University College London. E-Mail: t.ritschel@ucl.ac.uk*

Manuscript received XX XX, 20XX; revised XX XX, 20XX.

of medium quality. The system of Hoiem et al. [12] infers depth from monocular images by a low number of labels. Make3D [13] is based on learning appearance features to infer depth. This approach shows good results for static street-level scenes with super-pixel resolution but requires substantial computation. Non-parametric approaches rely on a large collection of 3D images [14] or 3D videos [15] that have to contain an exemplar similar to a 2D input. Conceptually, such an approach aligns all 3D images or 3D videos in a large collection (hundreds of exemplars) with a monocular query input image or video and transfers their depth to the query. Aligning to a large collection of images or videos of hundreds of elements contradicts our real-time requirements. We include prior disparity knowledge learned from exemplars into our inference by means of per-category disparity and confidence maps conditioned by image location and appearance. For cel animations, where each frame is drawn manually and therefore usually contains pronounced outlines, T-junctions have been shown to provide sufficient information to add approximate depth [16]. Our approach includes T-junctions in combination with other cues. The work of Tao et al. [17] is conceptually similar to our approach. They estimate depth by fusing information obtained from defocus and correspondences in a confidence-aware fashion using a Markov Random Field. Their offline method requires full light fields, while our real-time approach extracts and fuses an arbitrary number of cues from conventional videos.

Real-time methods to produce disparity from 2D input videos usually come at low visual quality. Individual cues such as color [18], motion [19] or templates [20] are combined in an ad-hoc fashion. A simple and computationally cheap solution is to time-shift the image sequence independently for each eye, such that a space-shift provides a stereo image pair [21]. This requires to identify the camera velocity and only works for horizontal motions. For rigid motions in animations, structure-from-motion (SfM) can directly be used to produce depth maps [22]. Classical SfM makes strong assumptions about the scene content such as a rigid scene with camera motion. More recent work relaxes these assumptions [23], but comes along with high computational costs. In the case of no or very unstructured motion, no stereo is provided by SfM alone, whereas in our fusion-based approach, motion is just one of many cues used when available. Commercial 2D-to-3D solutions [1] based on custom hardware (e. g., JVC's IF-2D3D1 Stereoscopic Image Processor) and software (e. g., DDD's Tri-Def-Player), reveal little about their used techniques, but anecdotal testing shows the room for improvement [15]. We subsume all such approaches in a principled framework that combines an arbitrary selection of cues in a common disparity-plus-confidence representation that can be effectively computed. Finally, most approaches produce low spatial resolution, and lack agreement between depth and luminance edges, as discussed next.

2.2 Depth and luminance edges

Since luminance and depth edges often coincide, e. g., at object silhouettes, full-resolution RGB images have been used to guide depth map upsampling both in the spatial [24] and the spatio-temporal [25] domain. An analysis of a database with range images for natural scenes reveals that depth maps mostly consist of piecewise smooth patches separated by edges at object boundaries [4]. This property is used in depth compression, where depth edge positions are explicitly encoded, e. g., by using piecewise-constant or linearly-varying depth representations between edges [5]. This in turn leads to a significantly better depth-image-based rendering (DIBR) quality than is possible at the same bandwidth of MPEG-style compressed depth, which tends to blur depth edges. In this work, we follow all these guidelines while reconstructing depth maps, as we also use DIBR to secure a high-quality 3D experience.

2.3 Computational models of depth perception

In previous work, perception was taken into account for stereography when disparity is given [3], but it was routinely ignored when inferring disparity from monocular input for 2D-to-3D conversion. In this work, we employ depth perception models to guide the 2D-to-3D reconstruction. Inference of depth from monocular images is based on depth *cues*. In this work, we use monocular cues (Sec. 3.2) to infer the missing binocular cue. A discussion of individual cues is beyond the scope of this article and can be found in Howard and Rogers [7]. The combination of cues into a perception of depth is called *fusion*. If multiple cues are extracted, their computational fusion is considered difficult, and left to the user as in the system of Assa and Wolf [11]. Two main opposing paradigms of fusion exist: the *weak* and the *strong* model [26]. In the weak model, cues act in isolation to produce an estimate of depth which is directly combined in a fixed linear weighting. In a strong model, cues interact in an unspecified and arbitrarily complex way. Our work is based on *modified weak* fusion [26], in which cues are independent, but their combination is not a linear mixture with fixed weights, as it locally adapts to the confidence of each cue. Bayesian fusion [27] using normal distributions is a formal way to achieve modified weak fusion. Here, cues are weighted by their confidence before they are combined. Besides using only the cues of the present stimulus, one strength of Bayesian inference is that it can account for prior experience [27]. We acquire disparity distribution priors for different scene classes using range scanners [4] or by manual annotation. While Bayesian fusion has been considered in perception literature [7, Ch. 30] for weighting specific cues according to their confidence to explain certain observations, we show for the first time a computational model to fuse multiple cues and a prior in order to solve a real-world task such as 2D-to-3D stereo conversion in real-time.

2.4 Spatio-temporal disparity sensitivity

The spatial disparity sensitivity function determines the minimum disparity magnitude required to detect sinusoidal depth corrugations of various spatial frequencies [7, Ch. 18]. The highest resolvable spatial frequency is about 3–4 cpd (cycles per degree), which is almost 20 times below the cut-off frequencies for luminance contrast [28]. Similar investigations in the temporal domain indicate that the highest sinusoidal disparity modulation that can be resolved is about 6–8 Hz [7], which is significantly lower than the 70 Hz measured for luminance [28]. As analyzed by Kane et al. [6],

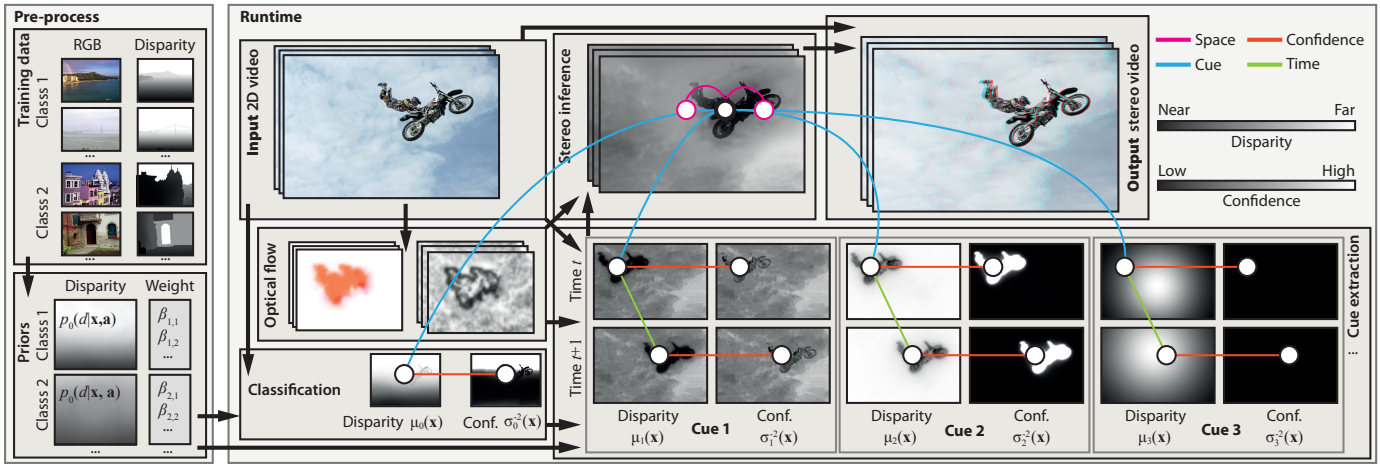

Fig. 1. Overview of our approach *(from left to right)* as described in Sec. 3. The grey coding used is annotated in the top right.

the picture is different for disparity step-edges in space and time, which are important in real-world images. They found that, for step-edge depth discontinuities, observers might still notice blur due to the removal of spatial frequencies up to 11 cpd, indicating that while overall disparity can be smoothed significantly, this is not the case for depth discontinuities. In this work, we follow this strategy by maintaining high precision in reconstructing sharp depth discontinuities, while otherwise allowing for substantial disparity blurring. Kane et al. could further show that filtering temporal frequencies higher than 3.6 Hz from a step signal remains mostly unnoticed. Their findings indicate that the temporal disparity signal might be sparsely sampled and even more aggressively low-pass filtered, without causing visible depth differences. In this work, we conduct similar experiments for complex scenes.

Surprisingly, depth edges appear sharp, even though human ability to resolve them in space and time is low. One explanation for this is that the perceived depth edge location is determined mostly by the position of the corresponding luminance edge [29]. In this work, we explicitly align imprecisely reconstructed and excessively blurred depth edges with detailed luminance edges. Interestingly, depth discontinuities that are not accompanied by color edges of sufficient contrast poorly contribute to the depth perception and do not require precise reconstruction in stereo 3D rendering [3].

The upper disparity gradient limit determines the maximum disparity for corrugations of a certain frequency the human visual system can fuse [7, Fig. 18.28]. Intuitively, when increasing the disparity gradient (e. g., by slanting a surface), retinal images become dissimilar and fusion becomes impossible [30]. Kane et al. [6] generalize this observation to space-time. In this work we explicitly enforce our disparity maps to obey the upper disparity gradient limit in a post-process.

3 Our Approach

An overview of our approach is shown in Fig. 1. It has two main parts: a pre-process (Sec. 3.1) to extract disparity priors (Fig. 1, left) and a runtime component (Fig. 1, right). While the pre-process uses many example images and requires considerable time, the runtime components execute in real time.

At runtime, first disparity and disparity confidence maps are extracted from monocular images (Sec. 3.2). This is the most computationally intensive part of our pipeline and implemented as parallel algorithms to require only a few milliseconds each. We support a flexible combination of both static cues (defocus, aerial perspective, vanishing points and occlusions) and dynamic cues (depth-from-motion). Each cue alone often has a low confidence in many areas and might contradict other cues. The cue evidence is then fused into plausible disparity maps (Sec. 3.3) using a robust maximum a posteriori (MAP) estimate [27]. This fusion happens again in real-time, producing results that are smooth in time and space, except at luminance edges. Finally, the monocular input image is converted into a stereo image pair obeying the disparity gradient limit (Sec. 3.4).

We will use a simplified disparity space ranging from 0 (close, depicted as black) to 1 (far, shown as white). As our goal is producing plausible disparity, we choose *not* to work in physical units like difference of vergence angles [7] or pixel disparities [31]. Rather, our perceptually plausible disparities arise from a smooth and monotonic remapping of physical disparities and are inspired by the way depth maps for manual stereoscopic conversion are painted. The perceptual effect of monotonic remappings of disparity is analysed in Sec. 4.2. The resulting disparity values will later be remapped to a comfortable range depending on the reproduction device.

3.1 Pre-processing

In a pre-process we learn prior information about disparity for certain classes of images and how to detect those classes.

3.1.1 *Disparity priors*

Priors model what is known about disparity in general without considering any specific image. This information is acquired from example depth images, validated and calibrated, and finally fit to a conditional distribution.

A disparity prior is the probability distribution of disparity $p_0(d)$. For efficient storage and computation, the

probability distribution $p_0(d) = \mathcal{N}(d|\mu_0, \sigma_0)$ is modeled as a normal distribution $\mathcal{N}$ of a certain mean μ_0, standard deviation σ_0, and variance σ_0^2 in this work. Furthermore, our priors $p_0(d|c, \mathbf{x}, \mathbf{a})$ are *conditioned* on three parameters: the scene class c (the depth distribution in "street" is different from "open countries"), the location $\mathbf{x} \in \mathbb{R}^2$ inside the image (the upper areas are more likely to be distant) and the appearance (RGB color) $\mathbf{a} \in \mathbb{R}^3$ (blue in the top of a forest image is more likely distant than green). For final cue fusion, scene class, image location and appearance are known and *unconditioned* priors will be used. Formally, the conditioned prior is defined as two 6D maps containing mean disparity $\bar{\mu}_0(c, \mathbf{x}, \mathbf{a})$ and the confidence of disparity $\bar{\sigma}_0^{-2}(c, \mathbf{x}, \mathbf{a})$. A high-variance value is found for a wide and unreliable distribution, while a high-confidence value $\bar{\sigma}_0^{-2}$ indicates a reliable estimate.

We use 10 representative scene classes consisting of about 40 example images each. Disparity maps were acquired both by sensors and by human annotation. Sensor-acquired classes are "street" and "indoor". For all other classes ("close-up", "coast", "forest", "inside city", "mountain", "open country", "portrait", "tall buildings"), depth maps were painted manually. Annotation was done in parts by 2D-to-3D conversion professionals, and experienced users of image manipulation software. Images have a resolution of ca. 100 k pixels. We provide the annotated database of our hand-painted depth maps and the resulting priors in our supplemental materials.

To compare human annotation performance to physical measurements, additional manual depth map painting was repeated for classes where sensor measurements are available by participants naïve in respect to the purpose of the procedure. A linear fit from painted depth x to physical vergence angles y with $y = .74\,x - .03$ has an error of adjusted $R^2 = .40$, indicating humans do a fair job when painting vergence compared to a sensor (Fig. 2).

Fig. 2. Example disparity maps for the scene class "street". *(a)* Appearance. *(b)* Disparity from sensor. *(c)* Disparity from human annotation.

Priors are extracted from example data independently for each class (see examples in Fig. 3). Each prior is represented as a 5D regular grid where the spatial dimension is discretized into 62×38 and the color dimension into 3×3×3 bins. Normalized image coordinates between 0 and 1 are used for the spatial component and $YCrCb$ color coordinates for the color component. Consequently, our prior contains $n_\mathrm{b} = 63\,612$ bins, with coordinates denoted as $\mathbf{b}_i \in \mathbb{R}^5$. The 2D positions and 3D colors of the n_s input pixels from all input images from that class are concatenated into a set of 5D samples $\mathbf{s}_j \in \mathbb{R}^5$, where each sample is labeled with its disparity d_j. Note that the number of bins is much smaller than the number of samples, $n_\mathrm{b} \ll n_\mathrm{s}$ (Fig. 4). Prior mean and confidence are each computed independently for all grid cells in two consecutive passes. In the first pass, the prior mean is computed as

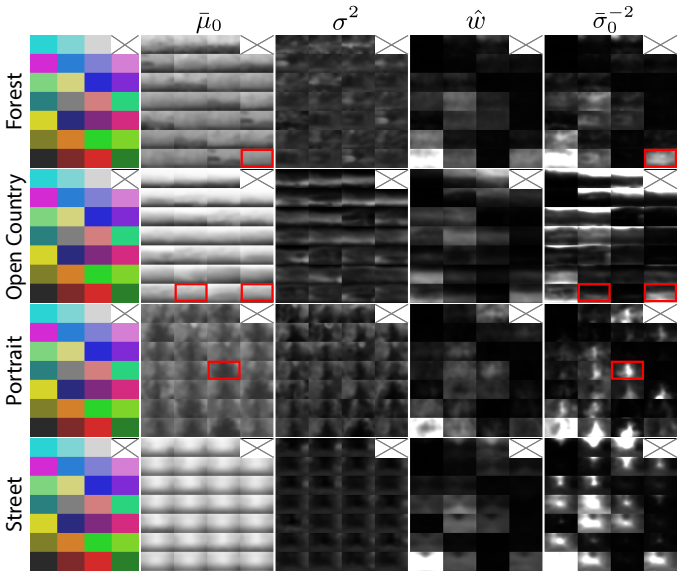

Fig. 3. Mean, variance, weight and confidence *(columns)* at different colors *(tiles)* for priors of different classes *(rows)*. For "forest", green central pixels have a medium depth. For "open country", brown and green lower pixels have a nearby depth. For "portrait", skin-colored central pixels are more nearby.

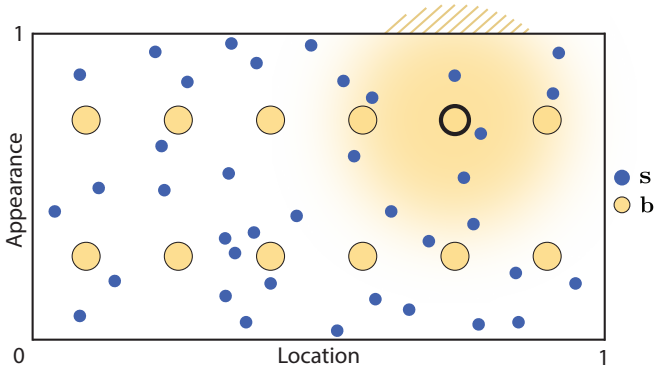

Fig. 4. A schematic of our prior extraction procedure with a 1D location and appearance domain. The samples $\mathbf{s}_j$ (*blue dots*) stem from the example images. The sparse bins $\mathbf{b}_i$ (*orange circles*) constitute the actual prior data ($\bar{\mu}_{0,i}$ and $\bar{\sigma}_{0,i}^{-2}$) to be determined. This is done by calculating the Gaussian weight $\psi(\mathbf{s}, \mathbf{b})$ for each possible sample/bin combination (*shown as a radial gradient for the highlighted bin*). To avoid a boundary bias (*hatched*) a correction α_i has to be determined for each bin.

$$\bar{\mu}_{0,i} = \frac{\sum_{j=1}^{n_\mathrm{s}} w_{ij} d_j}{\sum_{j=1}^{n_\mathrm{s}} w_{ij}} \quad \text{where} \quad w_{ij} = \alpha_i \psi(\mathbf{s}_j, \mathbf{b}_i) \qquad (1)$$

and $\psi(\mathbf{s}, \mathbf{b}) = \exp(-(\mathbf{s} - \mathbf{b})^\mathsf{T} \mathsf{A} (\mathbf{s} - \mathbf{b}))$ is a Gaussian kernel with a diagonal precision matrix A. For all results in this paper, the empirically chosen matrix entries are $\mathsf{A}_{11} = \mathsf{A}_{22} = 75$ for the spatial and $\mathsf{A}_{33} = \mathsf{A}_{44} = \mathsf{A}_{55} = 40$ for the appearance term. The normalization α_i for bin i is required because the 5D population can be highly non-uniform, and we use Gaussian filters of infinite support instead of compact (e. g., Epanechnikov) kernels. At the same time, our number of bins introduces a boundary bias for bins closer to the surface of the space-appearance cube which

would receive a lower total weight compared to other pixels. To compensate for this effect, we compute a correction

$$\alpha_i = \left(\int_{(0,1)^5} \psi(\mathbf{s}, \mathbf{b}_i) \mathrm{d}\mathbf{s} \right)^{-1}$$

of each 5D bin using Monte Carlo integration and normalize the result of each bin by this value. In the next pass, prior per-bin variance and weight

$$\sigma_i^2 = \frac{\sum_{j=1}^{n_\mathrm{s}} w_{ij} (\bar{\mu}_{0,i} - d_j)^2}{\sum_{j=1}^{n_\mathrm{s}} w_{ij} - \frac{\sum_{j=1}^{n_\mathrm{s}} w_{ij}^2}{\sum_{j=1}^{n_\mathrm{s}} w_{ij}}} \quad \text{and} \quad \hat{w}_i = \frac{\sum_{j=1}^{n_\mathrm{s}} w_{ij}}{\sum_{i=1}^{n_\mathrm{b}} \sum_{j=1}^{n_\mathrm{s}} w_{ij}}$$

are computed. The final prior confidence is

$$\bar{\sigma}_{0,i}^{-2} = \frac{\hat{w}_i}{\sigma_i^2}. \tag{2}$$

For the confidence $\bar{\sigma}_{0,i}^{-2}$ of the prior to be high, the variance σ_i^2 has to be low (agreement of samples to the mean) and the weight $\hat{w}_i$ has to be high (many samples similar to this bin). This combination prevents bins with a low number of samples to have a high confidence just because their estimate of variance is not stable.

3.1.2 Scene classification

Priors depend on the scene class c which is found from the monocular input RGB image. To this end, an image classifier is trained from example images that were manually labeled by their scene class. To meet our real-time requirements at test time and following ideas from Torralba [32], the image downscaled to 8×8 pixels is used as a feature vector. A linear Support Vector Machine is trained using gradient descent to separate each class from the other classes (one-vs.-one). At test time, we count the number of wins for each class over the other classes and pick the class c with the largest number of wins.

3.2 Depth cues

We model the i-th depth cue as a conditional probability distribution $p_i(d|\mathbf{x})$ of disparity d given a position $\mathbf{x}$. This distribution is described by a spatially-varying map of normal distributions in our approach. We store and process maps of mean disparity $\mu_i(\mathbf{x})$ and their confidence $\beta_{i,c}\sigma_i^{-2}(\mathbf{x})$ at position $\mathbf{x}$. The factor $\beta_{i,c}$ is a global per-cue i and per-category c weight that gives higher weights to cues that have shown to work better for certain scene categories. Actual values were determined empirically and are provided in the supplemental. We now briefly explain the $n_\mathrm{c} = 6$ cues we use. While the input sequence might have an arbitrary spatio-temporal resolution, the typical resolution to store each cue p_i is 300×170 pixels at 3 Hz, which will later be upsampled in space and time by the pairwise fusion (Sec. 3.3.4). We refer to frames of the image sequence holding depth cues as *keyframes*. Their position is not essential to our approach, and we refer to the supplemental material for details and example responses of cues to different input images and videos.

Our cue extraction is conceptually similar to other approaches, but differs with respect to previous work in two ways: First, that all cues can be processed in time linear in the number of pixels and in parallel using common GPU functionality, and second, that they provide an additional measure of per-pixel confidence.

3.2.1 Defocus

Scenes imaged with a finite-size aperture are increasingly blurry at image locations with distances different from the distance of the focal plane. Notably, the defocus only indicates a difference of distance to the focal plane, but not the sign. For the cue to be effective, the image has to contain this depth-of-field, which mostly occurs in images taken with a larger aperture for nearby objects. Depth-from-defocus is computed by measuring the local frequency content around a pixel [33]. Areas with only low-frequency content are considered out of focus. We use a Laplacian pyramid in multiple passes but constant amortized time per pixel (Fig. 5, b). The Laplacian acts as a bandpass while preserving spatial locality and its absolute value can be interpreted as the integral over one octave of the frequency spectrum [34]. On each level of the pyramid, we first soft-threshold the absolute value of the Laplacian up to 0.02 using a sigmoid and then blur the resulting per-level map with a box kernel of size 7×7 (Fig. 5, c). The thresholding is required to avoid interpreting high-contrast features (such as edges) as being more in-focus. We then collapse the pyramid by summing the contributions of all levels for each pixel, leveraging hardware-accelerated texture interpolation.

Out-of-focus regions are assumed to lie behind in-focus regions. This assumption, which is not always valid (Fig. 16, b) but nevertheless common in the absence of additional information [35], [36], corresponds to images with focused objects in front of a defocused background (Fig. 5, a). Consequently, sharp regions map to a disparity of 0 and sufficiently blurred regions to a value of 1 (Fig. 5, d).

Confidence for defocus is inversely proportional to disparity. This is motivated by the fact that high-frequency regions can only stem from scene content close to the focal plane, while there is an intrinsic ambiguity for low-frequency regions: either the depicted object is out of focus or it does not contain any high-frequency details (e. g., a plain-colored wall) [35]. We found this cue to work better when we additionally reduce the overall confidence if no defocus blur is present in the image. In order to determine if in-focus features dominate the image, we simply calculate the mean disparity of this cue by employing a MIP map.

3.2.2 Aerial perspective

Distant objects in images showing a landscape-scale range of depth undergo changes in appearance due to atmospheric scattering. This typically results in a depth-dependent loss of luminance contrast and a color shift towards blue, which can be analyzed to infer depth [37], [38], [39]. As the Cr channel of the $YCbCr$ color space separates low-frequency from high-frequency wavelengths, we use its inverse as the disparity map of this cue [40] in constant time, parallel for all pixels. Pixels with little local contrast in their vicinity (low variance) have higher confidence. Local variance is efficiently estimated using a Laplacian pyramid (cf. Sec. 3.2.1).

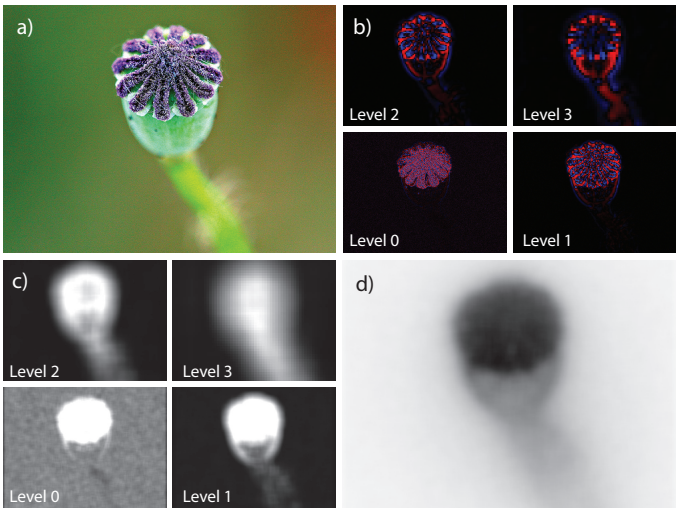

Fig. 5. Cue extraction for defocus builds a Laplacian pyramid (*b*) of the input image (*a*). Here, the red/blue color coding represents positive/negative values. The next step is thresholding and blurring of the absolute Laplacian per level (*c*). The final disparity map (*d*) is created by collapsing the pyramid as described in Sec. 3.2.1.

3.2.3 *Vanishing points*

Perspective projections of parallel 3D lines cross in a 2D vanishing point. If dominant lines are visible in an image, their point of convergence is a strong depth cue we would like to exploit as well. We use an approach based on edge extraction and line accumulation [41]. First, edge orientation is found at multiple image scales [42] and edge strength is measured by counting the number of scales at which the edge is present. Next, all pixels along a line elongating the orientation of every edge pixel are incremented by splatting a line primitive with additive blending (Fig. 6, b). The value of the line increases linearly with the distance to the pixel creating this line. This gradient is required, as vanishing points are more stable if they result in agreement with other lines at an image position far away from the respective pixel causing them. Finally, the pixel in the accumulated line-image that has the highest response to a Harris corner detector is considered the vanishing point pixel. This pixel is found using a parallel reduction.

The drawing area of the accumulated line-image is extended by a factor of 1.5 in both width and height compared to the input frame (red rectangle in Fig. 6, b). This way, vanishing points lying a reasonable distance outside the image boundaries can be detected. We found the recovery of vanishing points with positions further away from the image boundaries to become unstable in practice, while additionally only indicating a diminishing depth gradient.

The vanishing point itself is additionally low-pass filtered in time using a temporal cut-off of 0.5 Hz. Disparity is created according to this vanishing point using a radial gradient that is 1 at the vanishing point and 0 at the pixel farthest away from this point (Fig. 6, c). Confidence is computed by the curvature of the accumulated value: If all lines concentrate on a single pixel, the confidence is high and the vanishing point is reliable. If multiple vanishing points are found or if the accumulated lines do not concentrate in a small region, the cue is considered less confident. While images can contain multiple vanishing points, we found it more stable in practice to only pick the dominant one.

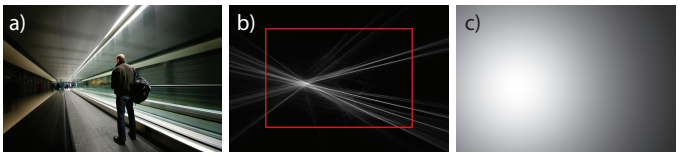

Fig. 6. Vanishing points are determined by splatting a line primitive (*b*) for each multi-scale edge pixel of the input image (*a*). The red rectangle indicates the original image boundary. The final disparity map (*c*) is a radial gradient centered at the estimated vanishing point (cf. Sec. 3.2.3).

3.2.4 *Static Occlusions*

Occlusion is a strong depth cue that works on all depth scales: If an object A occludes object B, A is closer. However, occlusion is only a relative cue and furthermore cannot be measured directly, only inferred. Occlusions are found by detecting T-junctions of edges and lines. This is done by convolving the image with a bank of separable filter kernels. 24 kernels are necessary to detect incident edges (Fig. 7, a, top row) and lines (bottom row) with an angular spacing of 30 degrees at a single scale. Note, that the same response can be created by convolving the image with only 12 centered kernels (Fig. 7, b) and then offsetting and/or inverting the resulting responses. We are interested in filter responses at different scales and for this purpose implement filters of increasing size by executing same-sized (15-tap) oriented 1D filters on an image pyramid. The approach of Michaelis and Sommer [43] is used to detect T-configurations based on these responses. As occlusion only indicates ordering, not absolute disparity, it cannot directly produce disparity and confidence, but produces sparse spatial disparity gradients with high confidence. More precisely, if a T-junction is found (Fig. 7, d) at position $\mathbf{x}$ with a vertical bar in direction $\mathbf{d}$ at scale s, a line orthogonal to $\mathbf{d}$ with length $10s$ is drawn with high confidence (we use a constant value of 10 in our implementation) and a positive gradient at $\mathbf{x} + s\mathbf{d}$ and with a negative gradient at $\mathbf{x} - s\mathbf{d}$ (Fig. 7, e).

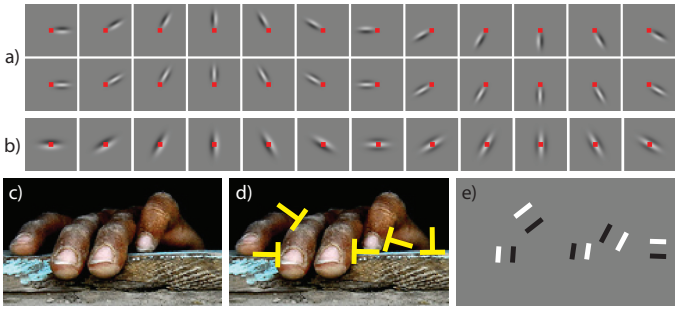

Fig. 7. Occlusions are found by first convolving the input image (*c*) with a filter bank (*a*, *b*; the pixel under consideration is marked red) and then combining the responses to detect T-configurations (*d*), leading to a sparse map of depth gradients (*e*). In *a)*, *b)* and *e)* a grey pixel indicates the value zero. The first row in *a)* and the first six images in *b)* show odd kernels for detecting edges, while the bottom row in *a)* and the last six images in *b)* show even filters for detecting lines. Note that the centered kernels in *b)* can be used to produce responses identical to those of the kernels in *a)*. Individual processing steps are explained in Sec. 3.2.4.

Fig. 8. Cue fusion *(left to right)*. Here, unary fusion combines confident occlusion, aerial perspective and defocus. The prior overrides values in the sky. Inconclusive evidence between prior and other cues is resolved by iterated re-weighting. The pairwise step propagates confident estimates to other locations, preserving space-time luminance discontinuities and eliminating low-confidence noise.

3.2.5 Motion

Several different depth cues are related to motion. Particular observer motions result in typical depth patterns and typical motions in the scene allow predictions about the relative depth of objects. In this work we use the computationally most simple cue that works based on optical flow alone. First, optical flow $\mathbf{f}(\mathbf{x})$ is computed between consecutive frames using a GPU implementation of Lucas-Kanade [44] registration. Although the output of the stereo cues is at low temporal resolution, the flow is computed at the full temporal, but reduced spatial resolution of the input image sequence, as we found flow between consecutive frames to work more reliably than registration of stronger deformations. Flow is augmented by a confidence map $\sigma_{\mathbf{f}}^{-2}(\mathbf{x})$, computed from the local luminance variance of the respective input frame: Flow in featureless regions is considered unreliable. $\mathbf{f}$ is later also used for temporal upsampling and propagation (Sec. 3.3.4).

To determine a disparity and confidence map for each keyframe, the confidence-weighted flow average is removed from the flow, leading to a motion residual

$$\mathbf{f}_r(\mathbf{x}) = \mathbf{f}(\mathbf{x}) - \sum_{\mathbf{x}'} \frac{\sigma_{\mathbf{f}}^{-2}(\mathbf{x}')\mathbf{f}(\mathbf{x}')}{\sigma_{\mathbf{f}}^{-2}(\mathbf{x}')},$$

where the weighted sum over all pixels in the current keyframe is efficiently determined by employing an image pyramid. The residual motion magnitude $\|\mathbf{f}_r(\mathbf{x})\|$ is used as an estimator for motion parallax and finally mapped to disparity, such that fast moving objects are closer. Confidence of this cue is determined by

$$\sigma_{\text{Motion}}^{-2}(\mathbf{x}) = \sigma_{\mathbf{f}}^{-2}(\mathbf{x}) n^{-1} \sum_{\mathbf{x}'} \|\mathbf{f}_r(\mathbf{x}')\|,$$

where n is the number of pixels in the keyframe. Here, the average residual motion magnitude of the keyframe serves as a global indicator that motion parallax is present.

3.2.6 User input

Optionally, user input can be included as another depth cue to augment traditional manual stereo painting with automatic inference in the propagation. A user simply paints a disparity and confidence map and the system includes this additional cue into the inference. No results in this paper were produced using any manual intervention, except for Fig. 12. The supplemental materials demonstrate the stereo improvement achieved by adding a few sparse strokes to the automatic solution.

3.3 Cue fusion

Cue fusion combines evidence from cues over space and time with the scene-specific prior (Fig. 8). Here, we will first explain the use of maximum likelihood estimation (MLE) to fuse evidence from multiple cues in a single pixel. Second, we extend the idea to include priors, yielding a maximum a posteriori (MAP) estimate. Next, we describe an iteratively reweighted variant of the estimate to make it robust to outliers and contradicting cues. Finally, we include interactions over time and space, and compute them using efficient edge-aware filtering.

3.3.1 Unary estimate

The unary estimate predicts the most likely value, given multiple observations with different levels of confidence. For a pixel $\mathbf{x}$, the MLE estimate of disparity $\mu_{\text{MLE}}(\mathbf{x})$ is the confidence-weighted average of disparity means

$$\mu_{\text{MLE}}(\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \sum_{i=1}^{n_c} \mu_i(\mathbf{x}) \beta_{i,c} \sigma_i^{-2}(\mathbf{x}),$$

where Z is the normalizing partition function. Furthermore, the MLE of the confidence simply is

$$\sigma_{\text{MLE}}^{-2}(\mathbf{x}) = \sum_{i=1}^{n_c} \beta_{i,c} \sigma_i^{-2}(\mathbf{x}). \tag{3}$$

This approach was taken in computer vision for measurements in the presence of sensor uncertainty [45] but not for 2D-to-3D conversion.

3.3.2 Prior

Priors are included in the fusion using Bayesian inference, which states that the probability distribution $p(h|e)$ of the hypothesis h given the evidence e is $p(h|e) = p(e|h)p(h)p^{-1}(e)$, where $p(e|h)$ is the probability distribution that the evidence e would be observed when the hypothesis is h, $p(h)$ is the probability distribution of the hypothesis h and $p(e)$ is the probability distribution of the evidence e [27]. A prior is included as an additional observation $\{\mu_0, \sigma_0^{-2}\}$, producing the MAP estimate of disparity

$$\mu_{\text{MAP}}(\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \left(\mu_0(\mathbf{x}) \beta_{0,c} \sigma_0^{-2}(\mathbf{x}) + \sum_{i=1}^{n_c} \mu_i(\mathbf{x}) \beta_{i,c} \sigma_i^{-2}(\mathbf{x}) \right).$$

The MAP estimate of variance $\sigma_{\text{MAP}}^{-2}(\mathbf{x})$ is computed by extending the sum of the MLE confidence (Eq. 3):

$$\sigma_{\text{MAP}}^{-2}(\mathbf{x}) = \beta_{0,c} \sigma_0^{-2}(\mathbf{x}) + \sum_{i=1}^{n_c} \beta_{i,c} \sigma_i^{-2}(\mathbf{x}).$$

In practice, the prior extracted in the pre-process (Sec. 3.1) that expresses information for all possible appearances at a location (conditioned prior $\{\bar{\mu}_{0,i}, \bar{\sigma}_{0,i}^{-2}\}$; Eq. 1, Eq. 2) is used for an image with a specific appearance at a specific location (unconditioned prior $\{\mu_0, \sigma_0^{-2}\}$). Let $L(\mathbf{x}) \in \mathbb{R}^2 \to \mathbb{R}^3$ be this appearance, a simple RGB image. We denote the final unconditioned priors mean and variance as $\mu_0(\mathbf{x}) = \texttt{fetch}(\bar{\mu}_0, (\mathbf{x}|L(\mathbf{x}))$ and $\sigma_0^{-2}(\mathbf{x}) = \texttt{fetch}(\bar{\sigma}_0^{-2}, (\mathbf{x}|L(\mathbf{x}))$. The function $\texttt{fetch}(X, \mathbf{y}) \in \mathbb{R}^5 \to \mathbb{R}$ is the 5D linear filtering of a grid X at position $\mathbf{y}$. For efficiency, we store the prior as a 2D array (spatial domain) of 3D textures (appearance domain). As linear filtering is separable, this texture is read using four 3D linearly-filtered hardware-accelerated interpolations in the appearance domain followed by spatial interpolation.

3.3.3 Robust estimate

If multiple high-confidence cues (including the prior) indicate different disparities, not all can be correct and at least one of them has to be considered an outlier. As MLE and MAP estimates for Gaussian noise models are generalized least-squares fits, they do not perform well in such conditions [46], as a single outlier quadratically skews the entire solution. Consider an example of two cues (e. g., focus and aerial perspective) and the prior that indicate a blurry blue pixel in the top to be far away, and a single cue (e. g., motion) to indicate it is close, all with the same confidence. A least squares-fit would indicate a medium disparity value. A more robust fit would result in a distant disparity and ignore the other cue as an outlier. This can be achieved by an iteratively reweighted MAP estimation. In each step (3 in our implementation) a weighted MAP is computed. In the first iteration, the weight is 1 for all evidence. In later iterations, the weight of evidence not supporting the MAP estimate of the previous iteration is decreased. Evidence does not support the estimate, if it is very different from it. The Cauchy weight function [46] is used to control the reweighting.

3.3.4 Pairwise estimate

The disparity at one space-time location $\mathbf{x}$ also depends on evidence from other pixels at nearby space-time positions $\mathbf{y}$. This serves both as an additional regularization constraint and as an opportunity to share information between less confident and more confident space-time locations. This dependency is modeled by the *domain* weight (disparity of nearby pixels should be similar) and the *range* weight (pixels with similar luminance values should have similar disparity),

$$v(\mathbf{x}, \mathbf{y}) = \mathcal{N}(\|\mathbf{x} - \mathbf{y}\|_{\mathbf{f}}, \sigma_{\mathrm{d}})\mathcal{N}(I(\mathbf{x}) - I(\mathbf{y}), \sigma_{\mathrm{r}}),$$

where I is the monocular image intensity [24], [25], [47]. Here, we assume the images have been motion-compensated, i. e., $\|\mathbf{x} - \mathbf{y}\|_{\mathbf{f}}$ is the spatial distance of $\mathbf{x}$ and $\mathbf{y}$ moved to the time coordinate of $\mathbf{x}$ along $\mathbf{f}$, or infinite if they are not related by optical flow. Then the final inference that combines spatially-varying cues and priors with confidence maps and interactions of pixels in space and time is

$$\mu(\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \int_{\Omega} v(\mathbf{x}, \mathbf{y}) \sigma_{\mathrm{MAP}}^{-2}(\mathbf{y}) \mu_{\mathrm{MAP}}(\mathbf{y}) \mathrm{d}\mathbf{y} \qquad (4)$$

with confidence

$$\sigma^{-2}(\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \int_{\Omega} v(\mathbf{x}, \mathbf{y}) \sigma_{\mathrm{MAP}}^{-4}(\mathbf{y}) \mathrm{d}\mathbf{y}, \qquad (5)$$

where Ω is the entire space-time domain. This inference is realized in three steps: i) Pixel-wise pre-multiplication of the mean disparity map μ_{MAP} by its confidence map $\sigma_{\mathrm{MAP}}^{-2}$; ii) edge-aware blurring of both the pre-multiplied mean disparity and confidence maps in time and space; iii) per-pixel division of the propagated mean disparity by its confidence [48].

Steps i) and iii) are trivially parallel and equivalent to compositing using pre-multiplied alpha. For propagation in time, the two nearby keyframes are first motion-compensated and then blended [49]. Recall that we compute the flow in full temporal resolution in the depth-from-motion cue component. For motion compensation, we forward-concatenate the flow from the past keyframe and backward-concatenate the flow from the future keyframe and use this flow to warp depth from the respective keyframes into the current frame. Warping disocclusions are filled using push-pull from a Gaussian MIP map. The backward flow is approximated using the negated forward flow, assuming motion is linear on small time scales. The result is then linearly blended using the temporal distance to the future and past keyframe as weights. The output of this step is at full temporal, but still at low spatial resolution. For propagation in space, a two-channel bilateral grid [50] with 8 layers and the full spatial resolution is used. Confidence-weighted disparity and confidence values are inserted into the layers of that grid using the final image intensity I as a guide with a standard deviation of $\sigma_{\mathrm{r}} = 0.1$. This grid is then blurred using a standard deviation of $\sigma_{\mathrm{d}} = 0.5\,\mathrm{deg}$ using a Gaussian MIP map. Next, the bilateral grid is upsampled to the desired high resolution, using the high-resolution luminance as a guide. After this step, the filtered, high-resolution disparity-confidence product is finally divided by the filtered confidence component.

3.4 Stereo image generation

The final step converts the acquired disparity maps into a stereo image pair. This step is a standard 2D-to-3D procedure for which many alternatives exist. We use grid-based image deformation [51] with a cell size of one pixel.

Before converting, however, we assure that the upper disparity gradient limit is maintained. Our disparity is produced by an automatic process and contains disparity with high spatial frequencies (Fig. 9, left column) that is important for the vivid and natural appearance. Consequently, the result may contain areas which are too distorted to be fused or even overlap (red box in Fig. 9). In particular, the result may contain fold-overs, where space runs backwards to create an overlap. We correct for this issue in a post-process as follows. First, a Laplacian pyramid of disparity [3] is produced, which contains gradients of disparity at multiple scales (Fig. 9, top right). Gradient values outside the level-dependent fusible disparity range [6], [7] (dotted lines in Fig. 9, top right) are clamped (red area in Fig. 9, top right). Finally, the resulting pyramid is collapsed into a new disparity map (Fig. 9, center right) that is fusible (green box in Fig. 9, lower right). Note, how the above is not equivalent to global rescaling, nor

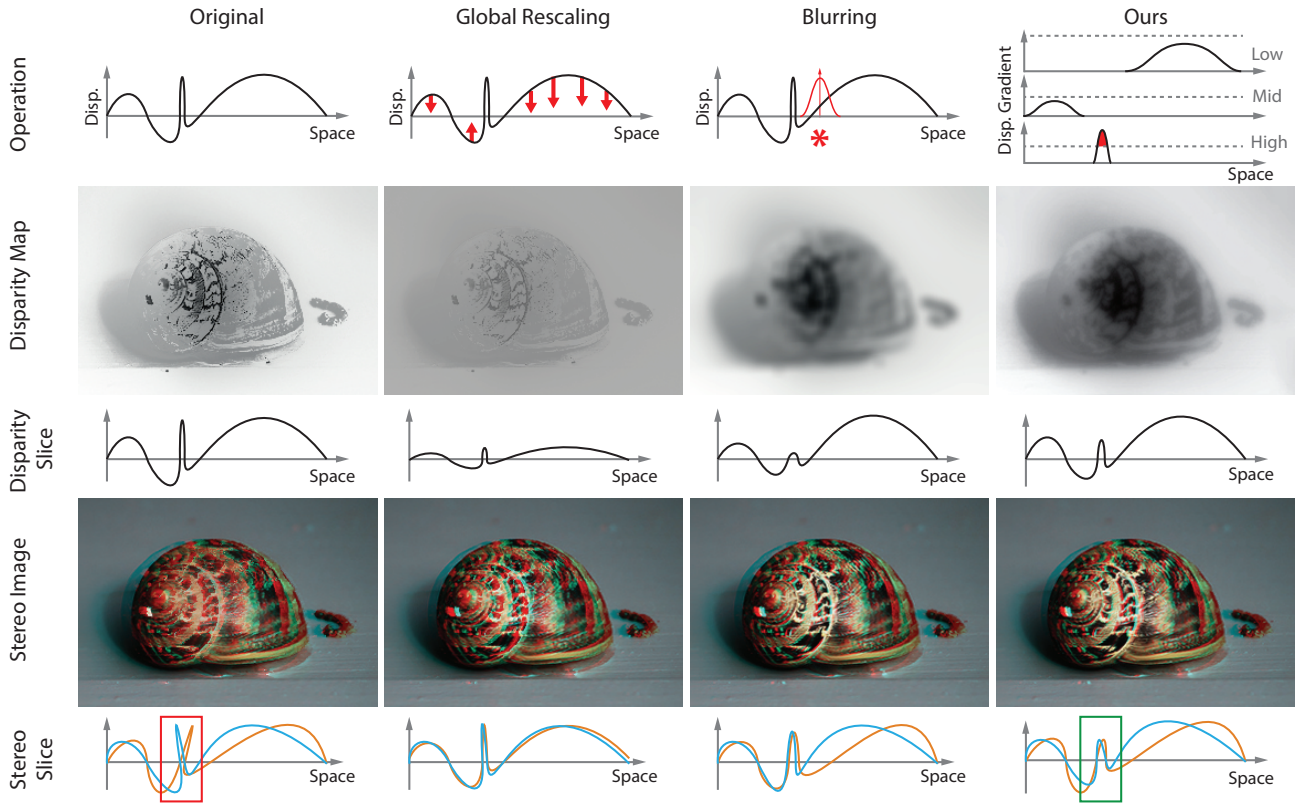

Fig. 9. Disparity maps and stereo images without *(first column)* and with different approaches *(second to fourth column)* to enforce the disparity gradient limit. *First column:* The original disparity map contains gradients exceeding the limits, resulting in fold-overs in the stereo signal (*red box*). *Second column:* A global linear rescaling resolves the issue, but results in a loss of overall depth contrast. *Third column:* Simple low-pass filtering is a natural way to reduce exceeding gradients, but comes at the cost of the loss of fine details. *Fourth column:* Our approach prevents fold-overs while at the same time retaining the global depth range and fine-scale details not exceeding the disparity gradient limit (*green box*).

is it equivalent to blurring. Both are options to fit stereo content into the gradient limit range, but would result in a reduced overall depth impression or in loss of fine details (Fig. 9, second and third column). Instead, our processing only removes disparity variations that are too strong for their spatial extent.

4 Evaluation

Example results of our real-time system are shown in Fig. 13. All are produced at 35 fps on a Geforce GTX 780 with an Intel Xeon E5-1620 CPU. A timing breakdown can be found in Tbl. 1. Results for video are seen in Fig. 10. An example comparison between our cue-guided manual 2D-to-3D conversion and a conventional scribble interface is seen in Fig. 12 and eleven similar results are provided in the supplemental materials. We found that our system works well over a range of scenes, while other approaches are more specific to a certain class, e. g., static street-level outdoor images. While other approaches are specialized to a specific cue (like vanishing points), certain motion (like rigid), a certain shape (like ground plane), or requiring that the image is similar to an image in a database, our technique relies on a greater variety of pictorial depth cues combined with priors based on scene types. Finding a balance between prior information and individual cues is an important component of our system (Fig. 14). To use a prior, the scene needs to be classified, and if classification fails, disparity quality degrades as seen in Fig. 15. Failure cases are discussed in Fig. 16.

TABLE 1
Computation time for a keyframe (every ca. 3 Hz) and for every non-keyframe (more than 30 Hz) at a resolution of 1280×720. Time for the actual computation granularity used is shown in bold.

Part	Step	Time 10 f.	Time 1 f.	Res.
Cue	Aerial per.	**2 ms**	0.2 ms	300×170
	Defocus	**8 ms**	0.8 ms	300×170
	Van. points	**18 ms**	1.8 ms	300×170
	Motion	**9 ms**	0.9 ms	300×170
	Occlusions	**11 ms**	1.1 ms	300×170
Opt. flow		58 ms	**5.8 ms**	300×170
Fusion	Robust MAP	**13 ms**	1.3 ms	300×170
	Temp. prop.	30 ms	**3.0 ms**	300×170
	Spatial prop.	46 ms	**4.6 ms**	1280×720
Warping		80 ms	**8.0 ms**	1280×720
		275 ms	27.5 ms	

4.1 Cue influence analysis

In order to gain insights into the influence of the cues and the prior on the resulting stereoscopic conversion we analyzed the results of our system for 83 videos with 80 keyframes each (a subset of these videos is given in the supplemental materials). Fig. 11, a lists the mean confidence of each cue and the prior. One can observe that the contribution of the prior on the result is about 50 %, while the cues

Fig. 10. Our 2D-to-3D video result with motion-compensated filtering provides temporally stable stereo with fine details.

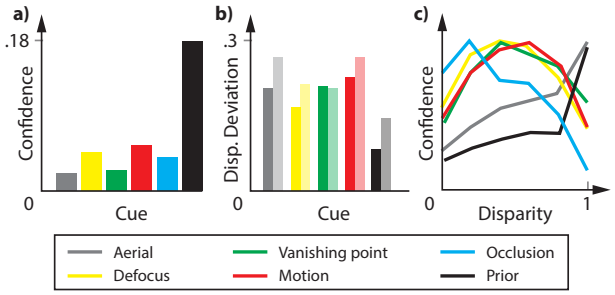

Fig. 11. Cue influence analysis: *a)* Mean confidence. *b)* Mean deviation of each cue's disparity estimate from the robust unary (*dark*) and the final pairwise estimate (*light*). Occlusion is not listed here, as it only provides disparity gradients. *c)* Normalized confidence distribution over disparity.

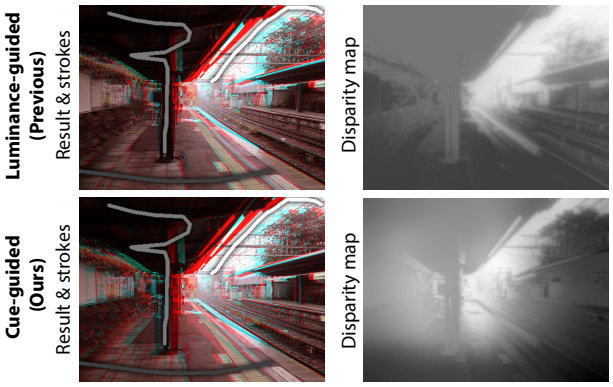

Fig. 12. Manual 2D-to-3D stereo conversion without *(top)* and with *(bottom)* using our cue fusion. Our approach results in a better disparity layout and keeps details, such as the wires.

contribute the other half of the depth information. Fig. 11, b gives each cue's tendency to be an outlier by showing the confidence-weighted deviation of its disparity estimate from the robust unary and the final pairwise estimate. One can observe that the deviation is fairly uniform across the cues, while the pairwise propagation step of our system tends to increase the deviation in order to perform the space-time regularization. Finally, Fig. 11, c shows the normalized confidence distribution of each cue over the disparity range. We can observe that the occlusion cue is mostly covering near distances, while the defocus, vanishing point and motion cues have their strongest influence in the mid-range. The aerial perspective cue as well as the prior mostly cover the larger distances. We conclude that our cues provide a balanced mixture of sources of information. In our versatile test dataset all cues provide important information and tend to complement each other, while our data-driven prior gives strong indications whenever there is not enough evidence from the cues alone. We provide an extended analysis of per-cue and per-scene class influence in the supplemental materials.

4.2 Validating plausible disparity

We would like to know to what extent the three properties of perceptually plausible disparity, which motivate our approach (Sec. 1), are applicable to complex images. To this end, we run perceptual experiments, in which we intentionally reduce physical disparity in these aspects [52].

4.2.1 Experiment

Participants were asked if they consider a physical and a distorted disparity stimulus visually equivalent or not. The physical disparity in our stimuli is distorted by one out of four simple operations: i) remapping by a power curve with a gamma value of $r_1 \in \{0.9, 0.8, 0.6, 0.3, 0\}$, ii) entire removal of a disparity from circles of radius $r_2 \in \{1, 3, 6, 12\}$ visual degrees followed by luminance-based edge-ware inpainting that restores structure but not disparity values, iii) edge-aware spatial blurring with a spatial std. dev. of $r_{3,1} \in \{0.25, 1, 2\}$ visual degrees and range Gaussian std. dev. of $r_{3,2} \in \{0.1, 0.6, \infty\}$ in the intensity range from 0 to 1, as well as iv) temporal blurring with a std. dev. of $r_4 \in \{0.025, 0.25, 1\}$ seconds. The original image or movie in comparison to the reference is used as a control group. 17 participants took part in the experiment, which comprised of 2 repetitions for each of the 4 videos or images being presented with 1 placebo, 5 different remappings, 4 removals, 3×3 spatial blurs and 3 temporal blurs yielding the total of $2 \times 4 \times (1 + 5 + 4 + 3 \times 3 + 3) = 64$ trials. In each trial, participants were shown the reference image and a distorted variant in a randomly shuffled vertical arrangement for 3 seconds and were asked if they provide an equivalent stereo impression or not.

4.2.2 Results and Discussion

We compute sample means and confidence intervals (binomial test, 95% CIs, Clopper-Pearson) for the percentage of trials in which a distorted and an original are considered equivalent (Fig. 17). The control group that is not distorted at all (placebo), is considered equivalent to the reference in 79.0% ± 4.0% of the cases (Fig. 17, a). Consequently, a reduction that is equivalent will at best result in a measure of equivalence of ca. 80%, not 100%. Equivalence is rejected using two-sample t-test (all $p < 0.01$). Additionally, the effect of reduction can be seen from comparing their CIs to the control group, in particular, its lower bound (Fig. 17, dotted line).

Remapping values for $r_1 \leq 0.6$ (stronger deviation from identity) are significantly nonequivalent (Fig. 17, b),

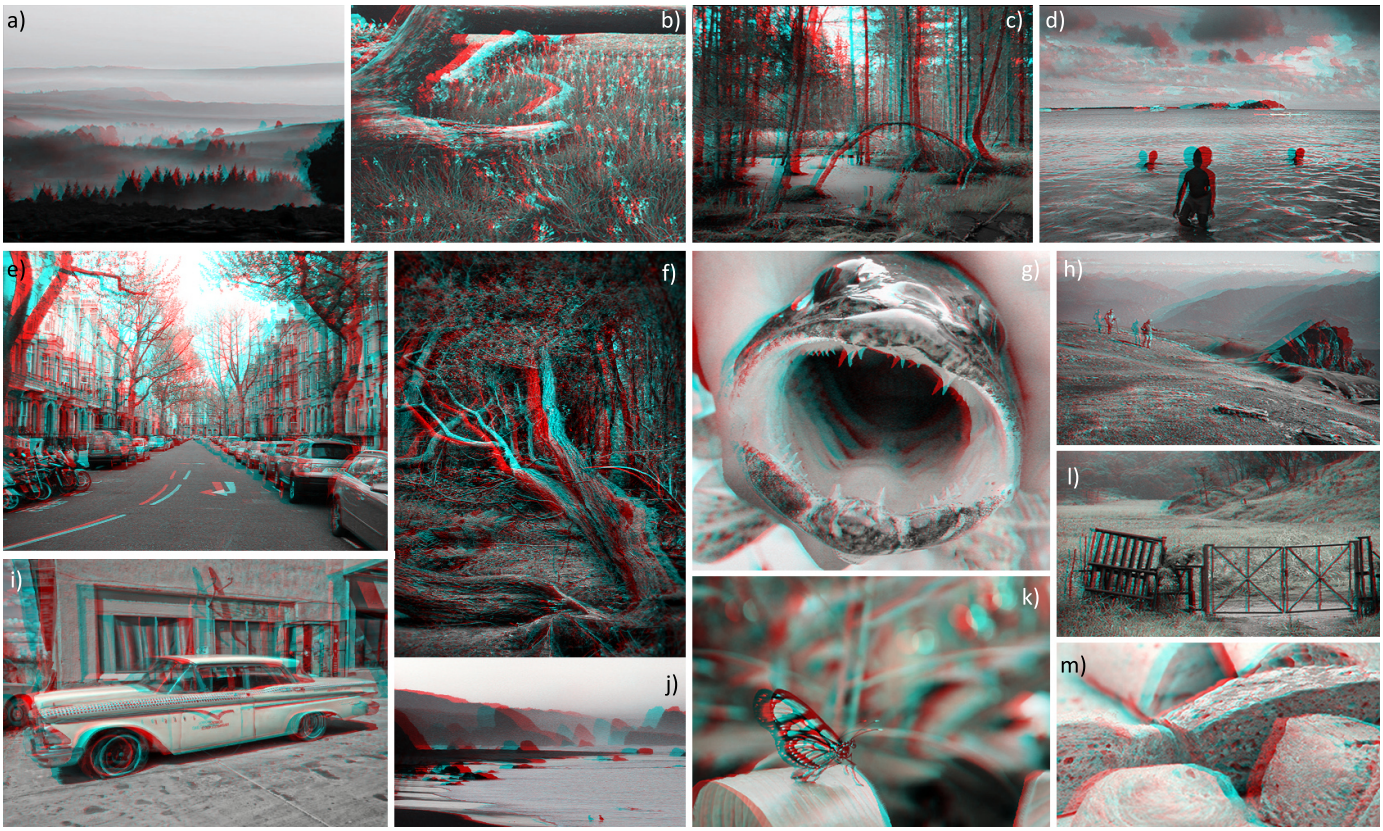

Fig. 13. Results for static images. (a) (Aerial perspective, occlusion) A typically good result as luminance edges give a good indication for depth edges. (b) (Prior) The fine details detach flowers from the ground. The overall ground plane is perceivably non-linear in depth, a typical artefact of our approach. (c) (Prior) The bright sky, the dark trees, the ground plane in the foreground and the forest-typical color-disparity relation in this image allow a plausible, detailed result. (d) (Prior) Classification into coast is easy by the colors. Human shapes are distinct from their context due to the edge-aware pairwise propagation. (e) (Vanishing point, prior) This is a typical street-level scene well covered by other approaches. Our result reproduces the side walls, the sky and the ground plane, but also includes fine details. (f) The twigs indicate occlusions, otherwise this image is dominated by the prior. Disparity is considerably wrong, but the fractal distribution of disparity combined with a correct tendency from the prior produces a consistent stereo look. (g) This image works, because the disparity contrast at the strongest depth discontinuity is correct due to a generic vertical-gradient prior. (h) (Aerial perspective, prior). The color-dependency of the prior correctly places depth edges on the hills horizon lines at all distances. (i) (Occlusion, vanishing point, prior) It can be noted that the ground plane in the front is not a plane in disparity. (j) (Aerial perspective, prior) (k) An image following no prior, where defocus is found as the relevant cue, detaching the butterfly from the backdrop. (m) (Occlusion) (l) An open country prior is fused with occlusion from the fence preserving fine details with good depth discrimination. Disparity mean and confidence maps, the response of all cues and the prior used for more than 60 images and more than 30 videos are found in our supplemental materials.

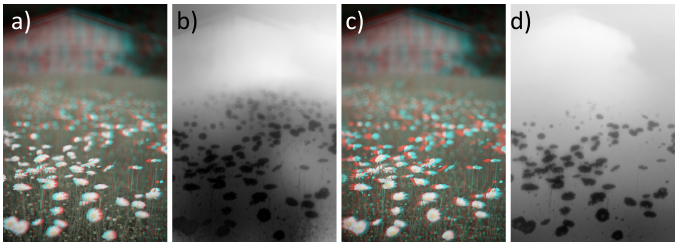

Fig. 14. Result *(a)* and depth map produced by cue fusion without priors *(b)*, and including the prior for open country *(c and d)*. The defocus cue has identified the sharpness gradient complemented by the prior.

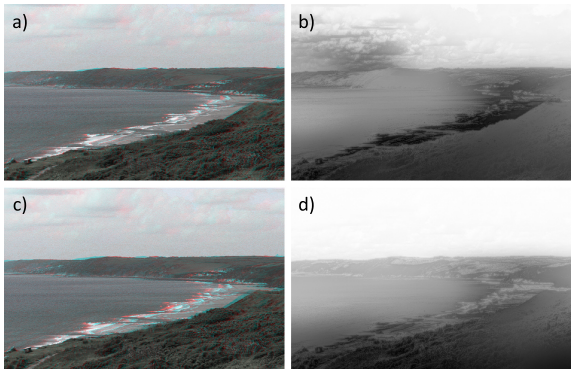

Fig. 15. An image *(a)* was classified to show mountains, resulting in a disparity map *(b)* that is more vertical as seen from the low vertical contrast and the light-grey beach is mapped to a near depth. With correct classification as coast *(c)*, the beach will be placed at medium depth *(d)*.

indicating (but not proving) that more subtle remappings might be equivalent. Our approach does only reproduce disparity up to such a smooth remapping. Not reproducing objects as large as $r_2 = 6$ vis. deg. or larger are significantly nonequivalent (Fig. 17, c), indicating that removal of smaller objects might not be objectionable. In our approach, some objects do not get resolved because neither a cue nor a prior provides evidence for its depth. As long as such objects are consistently embedded into the environment, which typically happens due to our luminance-based edge-aware upsampling, the proper values of depth are not mandatory.

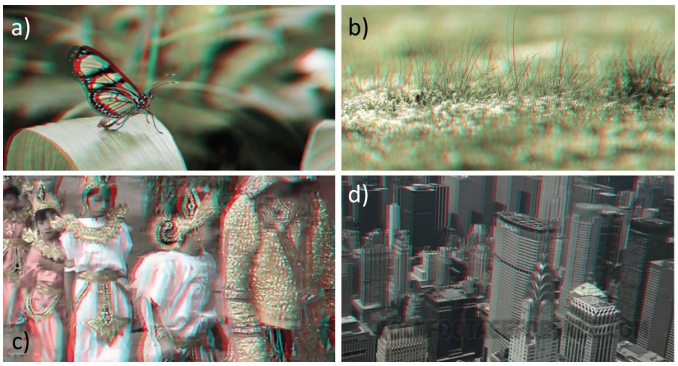

Fig. 16. Failure cases: *a)* High-contrast textures can cause problems in the cue extraction as well as the cue fusion phase. Here, the occlusion module detected several T-junctions in the butterfly wing and hallucinated depth gradients. This misinterpretation cannot be compensated by the pairwise fusion, since it does not distribute the available depth information across the whole object, but rather stops at the luminance edges. This leads to false-positive depth edges in the disparity map. *b)* If an assumption made in a cue extraction module is violated, the module may produce wrong disparity values with high confidence. If only a small number of cues is present in the input video, there is not much reliable information to compensate for that. In this case, the assumption of the defocus cue, that blurred regions are distant, is violated. Since there is no other strong cue present, this leads to a large disparity in the foreground. *c)* The motion cue fails, because the walking subjects cover the image in large part. This leads to residual motion, whose magnitude is low for the subjects and high for the background, hence turning the latter into foreground. *d)* For a camera rotating around an object, both close and far points with high velocity get classified as close.

For blurring (Fig. 17, d), not respecting edges ($r_{3,2} = \infty$), or edge-stopping blurring ($r_{3,2} = 0.6$ and $r_{3,2} = 0.1$) with a spatial Gaussian of std. dev. $r_{3,1} \geq 1$, resp. $r_{3,1} \geq 2$ vis. deg. is not equivalent. This indicates that the slightly larger spatial extent and similar range support used in our approach produces a functionally equivalent result.

For temporal blurring (Fig. 17, e) all reductions with a temporal Gaussian of std. dev. $r_4 \geq 1$s have been found visually nonequivalent. This indicates that temporal disparity sampling can be surprisingly sparse if it is motion-compensated, as in our approach, where disparity is computed only for keyframes at ca. 3Hz which likely is faster than the value required for equivalence. This outcome indicates that in natural images, even more edge-aware spatial blurring and temporal filtering is tolerated than what was reported for disparity-only stimuli by Kane et al. [6]. While the reductions in our experiment (and application) might introduce conflicts between disparity and pictorial cues, the latter seem to play the dominant role in depth perception, and tolerance for disparity reduction is higher. Edges at larger depth discontinuities must be preserved (Fig. 17, d), and in the temporal domain (Fig. 17, e) disparity should follow the image flow, while the temporal update of specific disparity values can be sparse.

4.3 Perceptual comparison study

We would like to know if the results produced in real time by our method are preferred over other approaches. Therefore, image pairs produced by our method and one of three previous methods were presented using Nvidia 3D Vision active shutter glasses on a 27" Asus VG278HE display with a resolution of 1920×1080 pixels at a viewing distance of 60 cm under normal office lighting. 10 participants (all male, 23 to 30 years old) took part in the study. All of them had normal or corrected-to-normal vision and passed a stereo-blindness test. The subjects were naïve to the purpose of the experiment. Overall, 77 image pairs were used. Each pair was presented as a random horizontal arrangement and participants were asked which image provides a better 3D impression. The images have been produced using methods proposed by Saxena et al. [13], Cheng et al. [18], and Karsch et al. [15]. In our study, we include results on our images for the method of Cheng et al., images and depths provided by the original publication for the method of Saxena et al. and a mixture of both for the method of Karsh et al. To produce results for our images the method of Karsh et al. was trained using 400 outdoor images from the Make3D dataset [13] as done in the original paper. Our main goal in this study was to maximize the participants' performance in seeing differences between the methods. Therefore we chose to use static images instead of videos, since human disparity sensitivity decreases with motion [6], [53] and participants were less likely to overlook artifacts. Our method is preferred over the method of Cheng et al. in $69.6\,\% \pm 3.3\,\%$ (0.95 confidence intervals, binomial) of the cases, over the one of Saxena et al. in $64.4\,\% \pm 7.0\,\%$ and over the approach of Karsch et al. in $54.5\,\% \pm 3.6\,\%$ of the cases. All comparisons are statistically significant ($p < 0.02$). Comparing our result and the method of Karsch et al. on a subset containing their images leads to a significant preference for their results ($31.3\,\% \pm 8.1\,\%$ prefer ours) while comparing on a subset only containing our images provides significant preference for our method ($60.0\,\% \pm 3.9\,\%$). This can be attributed to a non-optimal training set for certain images used in the study. We conclude that we can outperform real-time and offline 2D-to-3D conversion methods for general imagery, while the performance of data-driven offline methods highly depends on the training data used.

4.4 Quantitative evaluation

The final quality of a stereo image is due to the complex interaction of monocular and binocular stereo cues, for which no computational model is available. The perceived error of a 2D-to-3D stereo conversion consequently correlates only very little with the predictions of classic image quality metrics such as the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index when they are applied to the disparity maps [52]. Merkle et al. [5] show that more meaningful quality predictions can be obtained when the reconstructed disparity is actually applied to generate stereo-image pairs and those are compared to the ground-truth images. Tbl. 2 nonetheless lists the numerical error with respect to the ground truth NYU (Kinect sensor; well-aligned key luminance and depth edges) and Make3D (laser scanning; low-resolution depth maps) data sets for the approaches of Cheng et al. [18], Karsch et al. [15] as well as ours and a baseline that uses low-frequency fractal noise as a disparity map. We see that according to the PSNR (which is poor in detecting localized disparity distortions and rather assumes their spatially uniform distribution), the approach of Karsch performs best and that most approaches perform better than random, but not on all datasets and according

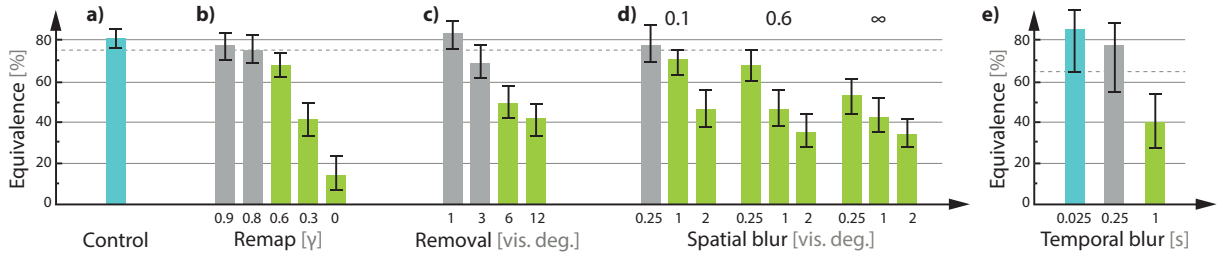

Fig. 17. Perceptual experiment analysis (Sec. 4.2): The horizontal axis shows different bars for different distortions. The vertical axis is equivalence in percentage. A high value means that the distortion is more equivalent to a reference. A green bar has a significantly different equivalence compared to how equivalent the reference is to itself, which is only ca. 80%, not 100%. Bars are grouped by distortions. Inside each group the distortion is the same, just more or less strong in one (*b*, *c* and *e*) or two (*d*) respects.

to all metrics. Overall, in terms of SSIM, the margin starts to get smaller. Finally, when using the most recommended metric by Merkle et al. [5], the difference between all three methods is marginalized.

We conclude that we can achieve similar quality in terms of error numbers as the competitors that either take much longer to compute and / or have a lower user preference. Interestingly, although the visual quality of the fractal baseline stereo-image pairs is clearly not acceptable, the metric predictions (Tbl. 2) do not show them as clear outliers in all cases. The fact that, on the one hand, we do not intend to reproduce ground truth depth but rather perceptually plausible disparity, and, on the other hand, the given quantitative evaluation clearly does not reflect stereoscopic conversion quality, indicates that our perceptual comparison study (Sec. 4.3) provides the most meaningful evaluation results.

TABLE 2
Numerical comparison (larger is better).

	NYU Range				Make3D			
	Disparity		Image pair		Disparity		Image pair	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Cheng [18]	9.96	0.72	21.84	0.80	10.30	0.56	16.91	0.42
Karsch [15]	10.77	0.76	21.77	0.80	11.60	0.77	18.40	0.49
Ours	10.18	0.74	21.03	0.78	10.03	0.66	17.02	0.43
Baseline	10.11	0.75	18.20	0.68	8.69	0.69	16.29	0.37

5 Conclusion

We proposed a system to infer perceptually plausible binocular disparity from a monocular video stream in real time. Several monocular cues estimate disparity and confidence maps of low spatial and temporal resolution. These are complemented by spatially varying, class-specific disparity priors. Robust MAP fusion produces stereo image streams with high spatial and temporal resolution. Perceptual experiments favorably compared our approach to existing techniques. Our method reconstructs perceptually plausible disparity and not physical depth. Instead, we rather draw inspiration from how humans proceed when manually annotating disparity in 2D-to-3D conversion. If physical accuracy is required, e. g., for viewpoint changes larger than inter-ocular distance or for refocusing, it is not advised to use our method. We found our method to produce images that might have physically incorrect depth, yet, they almost always provide a 3D look due to the agreement to high-frequency luminance features and overall plausible layout. Our approach seems to be less sensitive to the variety of scenes and works on priors created by painting. Depending on the problem at hand, working with sensor data can be more or less efficient than our pragmatic approach.

In future work we would like to integrate more sophisticated cues into our method. Structure-from-motion could be introduced into our system as a cue itself. More elaborate priors conditioned on texture and flow could add to the inference without imposing additional complexity and compute cost. We also would like to model cue fusion with the goal of improving the quality of stereoscopic experience when binocular disparity is given, instead of producing it from monocular images. Finally, our fusion is not limited to inference of depth, but could include other modalities such as observer motion, multiple images or real-time sensor data.

References

[1] L. Zhang, C. Vazquez, and S. Knorr, "3D-TV content creation: Automatic 2D-to-3D video conversion," *IEEE Trans. Broadcasting*, vol. 57, no. 2, pp. 372–83, 2011.
[2] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross, "Nonlinear disparity mapping for stereoscopic 3D," *ACM Trans. Graph. (Proc. SIGGRAPH)*, vol. 29, no. 4, 2010.
[3] P. Didyk, T. Ritschel, E. Eisemann, K. Myszkowski, H.-P. Seidel, and W. Matusik, "A luminance-contrast-aware disparity model and applications," *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, vol. 31, no. 6, 2012.
[4] Z. Yang and D. Purves, "A statistical explanation of visual space," *Nature Neuroscience*, vol. 6, no. 6, pp. 632–640, 2003.
[5] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Müller, P. H. N. de With, and T. Wiegand, "The effects of multiview depth video compression on multiview rendering," *Signal Processing: Im. Commun.*, vol. 24, no. 1-2, 2009.
[6] D. Kane, P. Guan, and M. Banks, "The limits of human stereopsis in space and time," *J Neurosc.*, vol. 34, no. 4, pp. 1397–408, 2014.
[7] I. Howard and B. Rogers, *Perceiving in Depth*. Oxford Psychology Series, 2012.
[8] M. Guttmann, L. Wolf, and D. Cohen-Or, "Semi-automatic stereo extraction from video footage," in *Proc. ICCV*, Sept 2009.
[9] B. Ward, S. B. Kang, and E. Bennett, "Depth director: A system for adding depth to movies," *IEEE Comp. Graph. and App.*, vol. 31, no. 1, 2011.
[10] M. Lang, O. Wang, T. Aydin, A. Smolic, and M. Gross, "Practical temporal consistency for image-based graphics applications," *ACM Trans. Graph. (Proc. SIGGRAPH)*, vol. 31, no. 4, 2012.
[11] J. Assa and L. Wolf, "Diorama construction from single images," *Comp. Graph. Forum (Proc. EG)*, vol. 26, no. 3, pp. 599–608, 2007.
[12] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic photo pop-up," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 577–584, 2005.

[13] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D scene structure from a single still image," *PAMI*, vol. 31, no. 5, pp. 824–40, 2009.
[14] J. Konrad, M. Wang, and P. Ishwar, "2D-to-3D image conversion by learning depth from examples," in *CVPR*, 2012, pp. 16–22.
[15] K. Karsch, C. Liu, and S. B. Kang, "Depth transfer: Depth extraction from video using non-parametric sampling," *IEEE PAMI*, vol. 36, no. 11, 2014.
[16] X. Liu, X. Mao, X. Yang, L. Zhang, and T.-T. Wong, "Stereoscopizing cel animations," *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, vol. 32, no. 6, p. 223, 2013.
[17] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *ICCV*, 2013, pp. 673–680.
[18] C.-C. Cheng, C.-T. Li, and L.-G. Chen, "An ultra-low-cost 2D-to-3D video conversion system," *SID*, vol. 41, no. 1, pp. 766–9, 2010.
[19] X. Huang, L. Wang, J. Huang, D. Li, and M. Zhang, "A depth extraction method based on motion and geometry for 2D to 3D conversion," in *Proc. IITA*, 2009, pp. 294–298.
[20] K. Yamada and Y. Suzuki, "Real-time 2D-to-3D conversion at full HD 1080p resolution," in *IEEE ISCE*, 2009, pp. 103–6.
[21] H. Murata, Y. Mori, S. Yamashita, A. Maenaka, S. Okada, K. Oyamada, and S. Kishimoto, "A real-time 2-D to 3-D image conversion technique using computed image depth," *SID*, vol. 29, no. 1, pp. 919–23, 1998.
[22] G. Zhang, W. Hua, X. Qin, T.-T. Wong, and H. Bao, "Stereoscopic video synthesis from a monocular video," *IEEE TVCG*, vol. 13, no. 4, 2007.
[23] C. Vogel, K. Schindler, and S. Roth, "3d scene flow estimation with a piecewise rigid scene model," *Int. J Comp. Vis.*, vol. 115, no. 1, pp. 1–28, 2015.
[24] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph. (Proc. SIGGRAPH)*, vol. 26, no. 3, 2007.
[25] C. Richardt, C. Stoll, N. Dodgson, H.-P. Seidel, and C. Theobalt, "Coherent spatiotemporal filterung, upsampling and rendering of RGBZ videos," *Comp. Graph. Forum*, vol. 31, no. 2, 2012.
[26] M. S. Landy, L. T. Maloney, E. B. Johnston, and M. Young, "Measurement and modeling of depth cue combination: In defense of weak fusion," *Vis. Res.*, vol. 35, no. 3, pp. 389–412, 1995.
[27] D. C. Knill and W. Richards, *Perception as Bayesian inference*. Cambridge University Press, 1996.
[28] B. A. Wandell, *Foundations of vision*. Sinauer Associates, 1995.
[29] A. E. Robinson and D. I. A. MacLeod, "Depth and luminance edges attract," *Journal of Vision*, vol. 13, no. 11, 2013.
[30] H. R. Filippini and M. S. Banks, "Limits of stereopsis explained by local cross-correlation," *J Vis.*, vol. 9, no. 1, 2009.
[31] R. Szeliski, *Computer vision: Algorithms and applications*. Springer, 2010.
[32] A. Torralba, "How many pixels make an image?" *Visual Neuroscience*, vol. 26, pp. 123–131, 1 2009.
[33] A. P. Pentland, "A new sense for depth of field," *IEEE PAMI*, no. 4, 1987.
[34] T. Darrell and K. Wohn, "Pyramid based depth from focus," in *CVPR*, Jun 1988, pp. 504–509.
[35] J. Lin, X. Ji, W. Xu, and Q. Dai, "Absolute depth estimation from a single defocused image," *IEEE Trans. Image Processing*, vol. 22, no. 11, pp. 4545–4550, Nov 2013.
[36] S. A. Valencia and R. M. Rodriguez-Dagnino, "Synthesizing stereo 3D views from focus cues in monoscopic 2D images," in *Electronic Imaging 2003*, 2003, pp. 377–388.
[37] F. Cozman and E. Krotkov, "Depth from scattering," in *CVPR*, Jun 1997, pp. 801–806.
[38] R. Fattal, "Single image dehazing," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 72:1–72:9, Aug. 2008.
[39] K. B. Gibson, S. J. Belongie, and T. Q. Nguyen, "Example based depth from fog," in *Proc. ICIP*, 2013, pp. 728–32.
[40] W. J. Tam, C. Vázquez, and F. Speranza, "Three-dimensional TV: A novel method for generating surrogate depth maps using colour information," in *Proc. SPIE*, 2009.
[41] S. T. Barnard, "Interpreting perspective images," *Artificial Intelligence*, vol. 21, no. 4, pp. 435–462, 1983.
[42] W.-Y. Ma and B. Manjunath, "Edgeflow: A technique for boundary detection and image segmentation," *IEEE Trans. Image Processing*, vol. 9, no. 8, pp. 1375–1388, Aug 2000.
[43] M. Michaelis and G. Sommer, "Junction classification by multiple orientation detection," in *Proc. ECCV*, 1994, pp. 101–8.
[44] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision." *IJCAI*, vol. 81, pp. 74–679, 1981.
[45] R. Szeliski, "Bayesian modeling of uncertainty in low-level vision," *IJCV*, vol. 5, no. 3, pp. 271–301, 1990.
[46] P. J. Green, "Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives," *J Royal Stat. Soc. B*, pp. 149–92, 1984.
[47] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Int Conf. Comp. Vis.*, 1998, pp. 839–846.
[48] H. Knutsson and C.-F. Westin, "Normalized and differential convolution," in *CVPR*, 1993, pp. 515–23.
[49] O. Miksik, D. Munoz, J. A. Bagnell, and M. Hebert, "Efficient temporal consistency for streaming video scene analysis," in *Proc. ICRA*, 2013, pp. 133–9.
[50] J. Chen, S. Paris, and F. Durand, "Real-time edge-aware image processing with the bilateral grid," *ACM Trans. Graph. (Proc. SIGGRAPH)*, vol. 26, no. 3, p. 103, 2007.
[51] W. R. Mark, L. McMillan, and G. Bishop, "Post-rendering 3D warping," in *Proc. I3D*, 1997, pp. 7–16.
[52] P. Kellnhofer, T. Leimkühler, T. Ritschel, K. Myszkowski, and H.-P. Seidel, "What makes 2D-to-3D conversion perceptually plausible?" in *Proc. Symp. Applied Perception*, 2015.
[53] P. Kellnhofer, P. Didyk, T. Ritschel, B. Masia, K. Myszkowski, and H.-P. Seidel, "Motion parallax in stereo 3D: Model and applications," *ACM Transactions on Graphics (Proc. SIGGRAPH Asia 2016)*, vol. 35, no. 6, 2016.

Thomas Leimkühler is a PhD student at the Max Planck Institute for Informatics and Saarland University in Saarbrücken, Germany. His research interests include efficient filtering and rendering.

Petr Kellnhofer has obtained his PhD in computer graphics at the Max Planck Institute for Informatics and Saarland University in Saarbrücken, Germany in 2016 under supervision of Karol Myszkowski and Hans-Peter Seidel. His thesis focuses on perceptual issues in computer graphics with a special interest in stereoscopic 3D. Currently, he is a postdoctoral associate at MIT CSAIL in the group of Professor Wojciech Matusik.

Tobias Ritschel is a Senior Lecturer at University College London. His research interests include interactive and non-photorealistic rendering, human perception and data-driven graphics. He received the Eurographics PhD dissertation award in 2011 and the Eurographics Young Researcher Award 2014.

Karol Myszkowski is a senior researcher at the Max Planck Institute for Informatics, Saarbrücken, Germany. He received his PhD (1991) and habilitation (2001) degrees in computer science from Warsaw University of Technology (Poland). In 2011 he was awarded with a lifetime professor title by the President of Poland. His research interests include global illumination and rendering, perception issues in graphics, high dynamic range imaging, and stereo 3D.

Hans-Peter Seidel is the scientific director and chair of the computer graphics group at the Max Planck Institute for Informatics and a professor of computer science at Saarland University, Saarbrücken, Germany. He has been on the program committee of all major international graphics conferences, and chaired several of these events. His publication list includes more than 60 papers in ACM SIGGRAPH/ACM TOG and more than 100 papers in Eurographics/CGF. Seidel has received numerous awards for his work, including the DFG Leibniz Prize (2003), and the Eurographics Distinguished Career Award (2012).