Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

A unified model for binocular fusion and depth perception

Jian Ding [*], Dennis M. Levi

School of Optometry and the Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, CA 94720-2020, United States

ARTICLE INFO

Keywords:
Disparity threshold
Sensory fusion
Correspondence problem
Phase disparity
Position disparity
Interocular misalignment

ABSTRACT

We describe a new unified model to explain both binocular fusion and depth perception, over a broad range of depths. At each location, the model consists of an array of paired spatial frequency filters, with different relative horizontal shifts (position disparity) and interocular phase disparities of 0, 90, ±180, or −90°. The paired filters with different spatial profiles (non-zero phase disparity) compute interocular misalignment and provide phase-disparity energy (binocular fusion energy) to drive selection of the appropriate filters along the position disparity space until the misalignment is eliminated and sensory fusion is achieved locally. The paired filters with identical spatial profiles (0 phase disparity) compute the position-disparity energy. After sensory fusion, the combination of position and possible residual phase disparity energies is calculated for binocular depth perception. Binocular fusion occurs at multiple scales following a coarse-to-fine process. At a given location, the apparent depth is the weighted sum of fusion shifts combined with residual phase disparity in all spatial-frequency channels, and the weights depend on stimulus spatial frequency and stimulus contrast. To test the theory, we measured disparity minimum and maximum thresholds (Dmin and Dmax) at three spatial frequencies and with different intraocular contrast levels. The stimuli were Random-Gabor-Patch (RGP) stereograms consisting of Gabor patches with random positions and phases, but with a fixed spatial frequency. The two eyes viewed identical arrays of patches except that one eye's array could be shifted horizontally and could differ in contrast. Our experiments and modeling reveal two contrast normalization mechanisms: (1) Energy Normalization (EN): Binocular energy is normalized with monocular energy after the site of binocular combination. This predicts constant Dmin thresholds when varying stimulus contrast in the two eyes; (2) DSKL model Interocular interactions: Monocular contrasts are normalized before the binocular combination site through interocular contrast gain-control and gain-enhancement mechanisms. This predicts contrast dependent Dmax thresholds. We tested a range of models and found that a model consisting of a second-order pathway with DSKL interocular interactions and a first-order pathway with EN at each spatial-frequency band can account for both the Dmin and Dmax data very well. Simulations show that the model makes reasonable predictions of suprathreshold depth perception.

1. Introduction

Although we have two eyes, we seldom perceive two images under normal viewing conditions. Instead, we almost always see a single sharp 3D image. Binocular disparity (the differences in image locations of an object seen by the two eyes, resulting from the eyes' horizontal separation) provides the cue to the brain to align the two eyes' images at each location and to compute depth. However, little is known about how the brain achieves the remarkable feat of fusing the two 2D images to construct a 3D percept. Motor fusion, through vergence eye movements, aligns the two eyes images globally, but this is not sufficient for binocular depth perception (stereopsis). Binocular depth perception requires an interocular matching mechanism to match the two eyes images. Traditional matching mechanisms used features, e.g., zero-crossings (Marr & Poggio, 1979), or maximum interocular correlation (Max operation) (Filippini & Banks, 2009; Fleet, Wagner & Heeger, 1996) to precisely match the two eyes images. However, false matches often occur (Fleet et al., 1996; Qian, 1994) and late-stage operations are required to improve matching performance (Filippini & Banks, 2009; Fleet et al., 1996). Furthermore, false matches occur more frequently with the Max operation than with real V1 neurons (Henriksen, Tanabe & Cumming, 2016).

Recently, more advanced matching strategies have been proposed. Tanabe et al (2011) revealed suppressive mechanisms in monkey V1 that help to solve the stereo correspondence problem. Read and Cumming (2007) found that phase-disparity neurons tend to be more strongly

[*] Corresponding author.
E-mail address: jian.ding@berkeley.edu (J. Ding).

https://doi.org/10.1016/j.visres.2020.11.009
Received 28 March 2020; Received in revised form 13 November 2020; Accepted 16 November 2020

activated by false matches, and may act as ‘lie detectors’, enabling the true correspondence to be deduced by a process of elimination. They noted that matching regions of a real image contain no phase disparity and took advantage of this fact to develop a robust matching strategy for solving the correspondence problem. However, based on optimal information encoding, Goncalves and Welchman (2017) showed that formulating the problem as identifying “correct matches’’ is suboptimal. They proposed an alternative approach that mixes disparity detection with “proscription’’: exploiting dissimilar features to provide evidence against unlikely interpretations, and demonstrated the role of these “what not” responses in a neural network optimized extraction of depth in natural images.

Indeed, these matching mechanisms seem to reverse causation, because “correct matches” might be the result of perfect binocular fusion. Binocular fusion is a process of reducing interocular misalignment of monocular outputs to achieve single binocular vision with misalignment below a threshold. Beyond this threshold, a binocularly combined image might be diplopic (possibly accompanied by suppression of one of the two images) or locally blurred. Although global alignment can be achieved by fusional vergence eye movements (motor fusion), it is still unclear how to achieve local alignments when the two eyes’ images have multiple disparities.

Binocular fusion and depth perception might be accomplished by different mechanisms. Indeed, we can perceive depth even when binocular fusion fails (McKee & Verghese, 2002; Richards, 1971; Schor & Wood, 1983) and persons who are stereo-blind might be able to perform binocular fusion (Richards, 1970). These suggest that binocular fusion might require a separate mechanism that reduces misregistration of the two eyes’ images, resulting in (1) a correct match (a single image) with a perfect alignment (perfect fusion) of the monocular outputs with or without depth perception, or (2) diplopia and a large misalignment (failure of fusion) with or without depth perception. However, based on a matching mechanism, it is unclear how to evaluate depth perception when the two eyes’ images are misregistered.

In the present paper we propose and test a new unified model to explain both binocular fusion and depth perception over a broad range of depths. An earlier version of this model without a fusion mechanism has been previously published in abstract form (Ding & Levi, 2016a). That model assumed a Max operator to select peak energy to compute stimulus disparity and a depth-disparity function to transfer disparity to depth perception. More recently, we used the same model to explain depth perception at suprathreshold levels (Ding & Levi, 2019). However, this model failed to accurately predict the reduced depth perception at large disparities near the disparity maximum threshold (Dmax), when the two eyes’ images are mismatched. To fit experimental data, we assumed an ideal Max operator without any mismatch even at large disparities near Dmax to detect stimulus disparity, and a depth-disparity function (the product of a disparity power function and an exponential decay function) was used to model the reduced depth performance near Dmax threshold. The model provided a reasonable fit to the data, but with an implausible mechanism.

The current study includes rich new psychophysical data using Random Gabor Patch Stereograms (RGP) and an expanded unified model that includes a binocular fusion mechanism as a solution to the correspondence problem. The unified model provides an evaluation of reduced depth perception of mismatched inputs (diplopic images) when fusion fails at large disparities, thus accounting for the full range of binocular single vision from Dmin to Dmax. (In *Discussion* we consider whether the visual system actually needs sensory fusion mechanism for depth perception if it is able to measure a stimulus disparity without considering fusion).

In previous studies (Ding, Klein & Levi, 2013a, 2013b), we proposed a binocular combination model with a binocular fusion mechanism to explain 2D binocular combination of sinewave gratings with different phases. The model assumes that phase disparity energy is calculated as binocular fusion energy for motor/sensory fusion to remap the two eyes’ inputs to realign them until phase disparity is eliminated. The model successfully predicts that the binocular contrast combination is independent of monocular phase differences at high contrast (Baker, Wallis, Georgeson & Meese, 2012; Ding et al., 2013b; Huang, Zhou, Zhou & Lu, 2010), but is dependent on the phase difference at low contrast levels (Baker et al., 2012; Ding et al., 2013b). This is because, at high contrast, the binocular fusion energy is sufficient to realign the two eyes’ images resulting in phase-independent binocular contrast perception, while at low contrast levels, the fusion energy is not sufficient to realign them, resulting in phase-dependent binocular contrast perception. In the current study, we elaborated this binocular fusion mechanism into a unified model for 3D depth perception and integrate it into the depth model to realign the two eyes images under a 3D view. To the best of our knowledge, to date this binocular fusion mechanism has not been addressed directly by physiological studies.

Motor fusion (vergence eye movements) brings the two eyes’ images into global alignment; however, binocular sensory fusion is necessary for local alignment when the images have multiple disparities. Both share the common primary stimulus—binocular disparity. With sensory fusion, small vergence errors (fixation disparity, FD) can occur without diplopia (Fogt & Jones, 1998; Ukwade, 2000), and misaligned (non-corresponding) retinal images are perceived as single as long as they are within Panum’s area (Panum, 1858), i.e., to any given retinal point in one eye there corresponds a small group of points in the other eye. Fixation disparity can be measured objectively using eye movement recording (Fogt & Jones, 1998; Hyson, Julesz & Fender, 1983) or subjectively by aligning nonius lines (Fogt & Jones, 1998; McKee & Levi, 1987; Schor, Wood & Ogawa, 1984; Ukwade, 2000). Hyson et al. (1983) recorded vergence eye movements while their observers viewed a random-dot stereogram and misaligned the stereo images by moving them apart until fusion was lost. They found that the vergence error, the difference between image separation and eye vergence, could be as large as 3°. They postulated that neural remapping occurs during sensory fusion that compensates for the retinal misalignment. Fogt and Jones (1998) compared fixation disparity obtained by objective and subjective methods by measuring FD as a function of forced vergence. They found that the slope of the objective FD curve was significantly greater than the subjective FD curve, indicating an alteration in retinal correspondence. Based on objective measurement of human cyclofusional response, Kertesz and Jones (1970) found that the maximum fused vertical disparity introduced by the cyclofusional stimulus for various peripheral angles of the retinas was close to, but always less than, the disparity threshold for diplopia values obtained by Volkman.

However, beyond these observations, very little is known about sensory fusion or neural remapping. Here, we used a conceptual schema to demonstrate a sensory fusion mechanism that we propose for binocular vision.

2. A conceptual schema for sensory fusion

Poggio and Fischer (1977) classified disparity-selective neurons into four types: a) tuned excitatory (TE) neurons, the most common type, are excited over a narrow range of stimulus disparities around the fixation plane often with inhibitory flanks nearer and farther, and they typically received a balanced binocular input; b) tuned inhibitory (TI) neurons whose responses are suppressed by small disparities of either direction (i.e., for targets fairly close to the fixation plane), but respond weakly to large disparities of either direction, hence the suggestion that they are defined by receptive fields in anti-phase arrangement; c) near neurons (NEAR), which responded well to crossed disparities and are suppressed by uncrossed disparities; d) far neurons (FAR), the opposite of NEAR neurons. However, later studies showed a continuum of tuning types of disparity-selective neurons (Prince, Cumming & Parker, 2002); many neurons have intermediate tuning types. Studies in cat and non-human primate V1 have shown that most disparity-selective neurons are hybrid, with both preferred phase and position disparities (Anzai, Ohzawa &

Freeman, 1997, 1999; Livingstone & Tsao, 1999; Prince et al., 2002; Tsao, Conway & Livingstone, 2003). In fact, each disparity-selective neuron has both preferred phase and position disparities; a pure position disparity neuron has a preferred phase disparity = 0 and a pure phase disparity neuron has a preferred position disparity = 0.

Based on these facts, Fig. 1 illustrates a conceptual schema for sensory fusion and depth perception in one spatial-frequency channel projected to the xz plane. The vertical (y-axis not shown) is orthogonal to the paper plane (xz-plane). The x-axis represents the horizontal dimension and indicates the fixation plane (position disparity u = 0). The z-axis represents position disparity (horizontal relative shift of paired filters). The blue and red horizontal bars represent two vertical Gabor patches presented to the left (LE) and right (RE) eyes respectively. They are not overlaid with each other in the fixation plane (u = 0), but have a stimulus disparity of d. Horizontal space is sampled by LE (blue open boxes) and RE (red open boxes) vertical spatial-frequency filters (only partial filters are shown). At each location, both position and phase disparities are sampled by paired filters. The phase disparity is sampled at 0 (TE), 90 (NEAR), ±180 (TI), and −90 (FAR) phase degree at each location and each position disparity. However, for clarity, we only show paired filters at location × = 0 for each position-disparity plane (inside a thick black box). For example, in the fixation plane, at location × = 0, there are four pairs of filters with TE, NEAR, TI, and FAR tuning curves, preferred at 0°, 90°, ±180^o, and −90° phase disparity respectively. In the position disparity plane u, all four pairs of filters (TE, NEAR, TI and FAR) have a preferred position disparity of u. We define a depth sensor to be an array of paired filters with different position disparities and different phase disparities at one location in the 2D xy plane. At any one time and one location, the system selectively reads out the outputs of the paired filters only in one position disparity plane as the depth sensor's output.

As shown in Fig. 1, when the two eyes are presented with input images with uncrossed disparity d, the FAR (with preferred −90° phase disparity) neuron in the fixation plane at × = 0 detects the uncrossed misalignment. To reduce the misalignment of monocular outputs, the system's readout is shifted from the paired filters in the fixation plane (0 position disparity) to those in an uncrossed position disparity plane. This shift of readout of depth sensor's output from one to another position disparity plane is defined as *sensory readout shift*, or simply *sensory shift*. The shift continues until the misalignment (or phase disparity) of monocular outputs is eliminated (when u = d), at which the pair with 0-phase disparity reaches the maximum correlation, and the pairs with 90, ±180 and −90 phase disparites becomes uncorrelated or anti-correlated. The 2D monocular outputs of the paired filters with 0-phase disparity at u = d plane are perfectly aligned (fused pair in Fig. 1). However, if the phase disparity energy is not sufficient at the fixation plane, e.g., at a larger stimulus disparity and/or low stimulus contrast, the sensory fusion process might stop at $u = u_1 < d$ before reaching the target-depth plane (u = d). This will produce diplopic images with a horizontal separation of d - u_1 and a reduced depth perception given by u_1 .

Our model assumes that there are multiple overlapping depth sensors at different locations and scales across the visual field. They compete with each other, and the one with the minimum misalignment and maximum correlation after the fusion process wins the competition. Typically a depth sensor based on the appropriate corresponding inputs, e.g., the one at x = 0 in Fig. 1, wins the competition. However, false matches might occur especially when phase disparity energy is not sufficient at low stimulus contrast or with a large stimulus disparity.

The blue and red dashed lines in Fig. 1 indicate the local visual directions at x = 0 in the left and right eyes respectively. The z-axis indicates the visual direction for the cyclopean eye (CE). Each depth sensor has its own visual directions in the left, right and cyclopean eyes, which should follow the geometric constraints of stereovision. For example, if a depth target is directly in front of the LE, the LE's local visual direction is orthogonal to the x-axis, i.e., local sensory fusion is

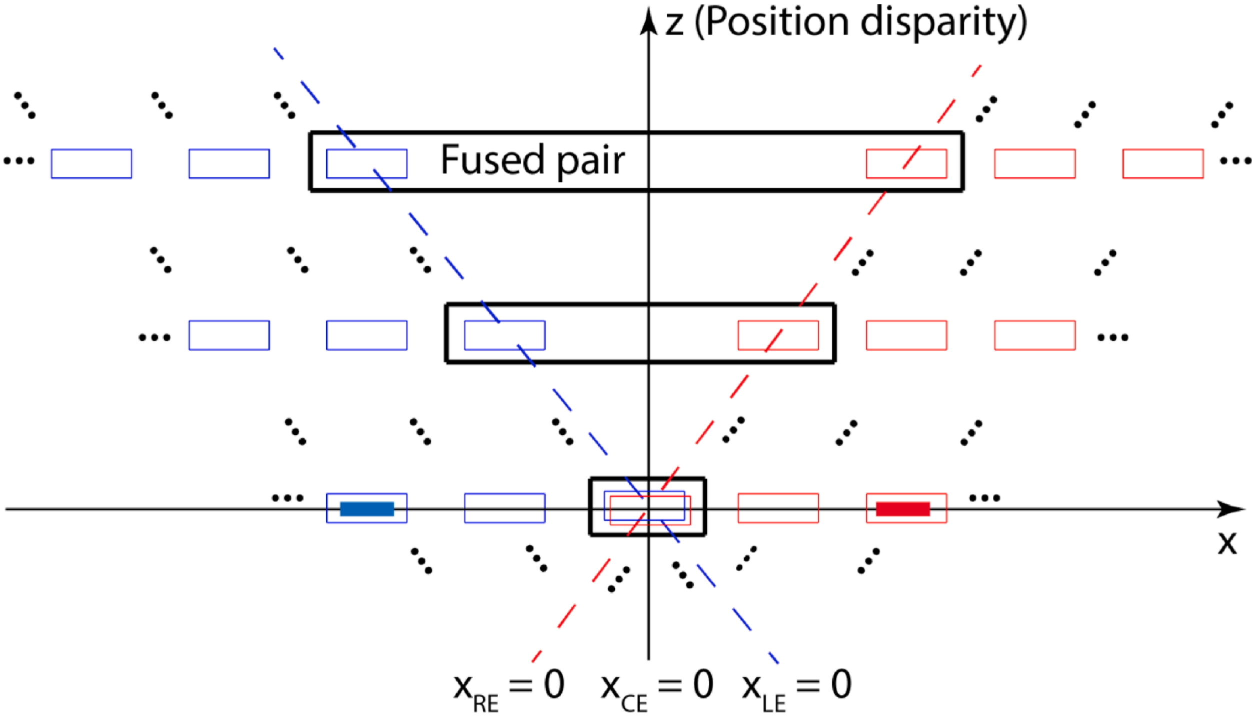

Fig. 1. A schema for sensory fusion in one spatial-frequency pathway. The schema is a projection onto the xz-plane. The vertical y-axis (not shown) is orthorgonal to the paper (xz-plane). The z-axis represents the relative shift of the selected paired filters (position disparity). Blue and red open boxes represent the LE's and RE's vertical spatial frequency filters respectively. The blue and red solid bars represent vertical Gabor patches presented to the LE and RE, respectively, in the fixation plane with stimulus disparity *d*. Paired filters are inside a solid black box with different relative shifts (position disparities). The fused pair has a relative shift of *d* that gives depth perception, and the fused paired filters with 0 phase disparity outputs locally aligned 2D images. The blue and red dashed lines indicate the local visual directions at x = 0 in the LE and RE respectively, and the z-axis indicates the visual direction of the cyclopean eye (CE).

achieved by shifting the readout to the location of the matching filter of the RE, without shifting that of the LE.

Because disparity space may not be sufficiently sampled, even for the best matching pair, phase disparity might not be eliminated completely, i.e., the binocular fusion might not be prefect if the depth samples are not sufficient. To get accurate depth information, after the fusion process, the residual phase disparity is measured to estimate the relative depth of a target to the position disparity plane, i.e., the final perceived depth is the combination of position and phase disparities after sensory fusion.

3. Contrast normalization

Several models have been proposed to explain binocular combination based on interocular contrast gain-control (Ding et al., 2013b; Ding & Levi, 2016b; 2017; Ding & Sperling, 2006, 2007; Georgeson, Wallis, Meese & Baker, 2016; Huang et al., 2010; Meese, Georgeson & Baker, 2006; Yehezkel, Ding, Sterkin, Polat & Levi, 2016), and some of these have been extended to stereovision (Ding & Levi, 2016a; Hou, Huang, Liang, Zhou & Lu, 2013). Previously, we (Ding et al., 2013b) compared multiple normalization mechanisms in binocular combination. We found that all of them can explain binocular contrast combination, but normalization mechanisms with contrast values or responses as inputs to the models, e.g., the two-stage model (Meese, Georgeson et al., 2006), need to be revised by including the spatial domain in order to address phase combination. For the same reason, they need to be revised to address depth perception. Based on interocular contrast gain control (Ding & Sperling, 2006), Hou et al (2013) proposed a multi-pathway contrast gain-control model (MCM) to explain both binocular combination and stereovision. The model simultaneously accounts for Dmin disparity thresholds and cyclopean contrast perception of dynamic random dot stereograms (dRDS). However, the MCM did not address the correspondence problem or Dmax thresholds. Based on studies in anesthetized cats, Ohzawa and Freeman (1994) suggested that a single gain control mechanism is not sufficient to account for the properties exhibited by cortical neurons, and there appear to be at least two mechanisms of contrast gain control either before or after binocular convergence. Later, the same group reported that contrast gain reductions occur primarily at a monocular site, before convergence of information from the two eyes (Truchard, Ohzawa & Freeman, 2000). In this study we address contrast normalization in depth perception based on interocular interactions before the site of binocular combination (Ding & Sperling, 2006; Ding, Klein, & Levi, 2013a; 2013b), and energy normalization after binocular combination (EN: binocular energy is normalized by monocular energy).

In the present study, we measured both minimum and maximum disparity thresholds (Dmin & Dmax) psychophysically using Random Gabor Patch (RGP) Stereograms when stimulus contrast differed in the two eyes. We developed a depth model with a binocular fusion mechanism including two different contrast normalizations to explain both Dmin and Dmax threshold data. We compare a number of different models of depth perception and show that in order to evaluate the reduced depth perception of diplopic images when fusion fails at large disparities, and therefore predict Dmax threshold data, the model needs a fusion mechanism. Without a fusion mechanism, conventional human vision models fail to provide a reasonable explanation of Dmax, even with acceptable model assumptions.

4. Methods.

Stimuli. Random-Gabor-Patch (RGP) stereograms (Fig. 2), in which vertical Gabor patches with random positions and phases, but with a fixed spatial frequency, were used as stimuli. RGP stereograms provide stereo signals in a narrow spatial frequency-and-orientation channel without monocular depth cues. The two eyes have identical arrays of patches except that one eye's array can be shifted horizontally, and they can differ in contrast. The jth Gabor patch pair is given by

$$I_{jL} = m_L e^{-\frac{\left(x-x_j-\frac{d}{2}\right)^2+\left(y-y_j\right)^2}{2\sigma^2}}\cos\left(\omega\left(x-x_j-\frac{d}{2}\right)+\theta_j\right) \quad (1)$$

$$I_{jR} = m_R e^{-\frac{\left(x-x_j+\frac{d}{2}\right)^2+\left(y-y_j\right)^2}{2\sigma^2}}\cos\left(\omega\left(x-x_j+\frac{d}{2}\right)+\theta_j\right) \quad (2)$$

To produce an RGP stereogram, a large square (14.1x14.1 deg^2) was divided into 18x18 small grids. Each grid contains a Gabor patch with ω

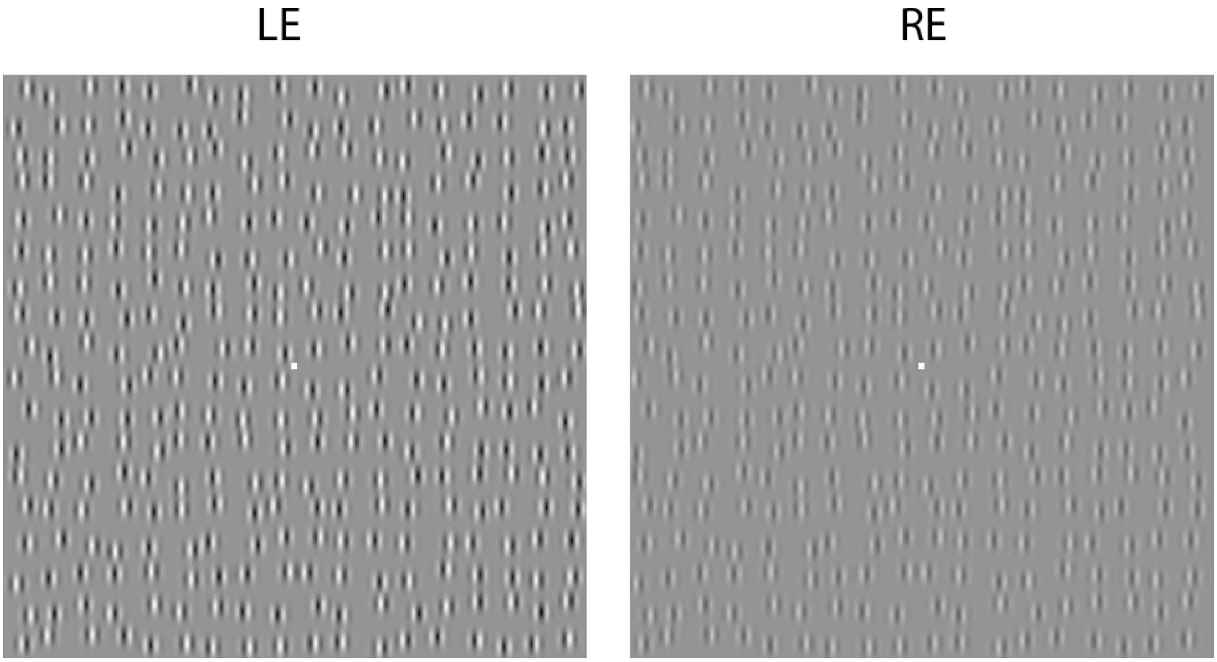

Fig. 2. Random-Gabor-patch stereograms. Gabor patches had random positions and phases, but with a fixed spatial frequency. The two eyes had two identical arrays of patches except that one eye's array could be shifted horizontally, and they could differ in contrast.

= 3 cpd and $\sigma = 0.167$ deg and its position randomly distributed (with equal distribution) inside (gridwidth – σ) × (gridwidth – σ) area. Along the grid border, two patches could be partially overlapped. Stimulus disparity was produced by horizontally shifting the two eyes' images by equal amounts (=half the stimulus disparity) but in reversed directions. A circular shift was made to maintain stereogram size constant. A stereogram given by Eqs. (1) and (2) can produce a sub-pixel disparity accurately because the perceived position of a Gabor patch depends on its centroid of light distribution (Aiba & Morgan, 1985; Georgeson, Freeman & Scott-Samuel, 1996).

Observers were asked to judge the depth of the entire array, either near or far, relative to the central fixation point. For RGP stereograms containing smaller Gabor patches with $\omega = 6$ cpd and $\sigma = 0.083$ deg, the big square (14.1x14.1 deg^2) was divided into 36x36 small grids. For RGP stereograms containing larger Gabor patches with $\omega = 1.5$ cpd and $\sigma = 0.333$ deg, the big square (14.1x14.1 deg^2) was divided into 9x9 small grids.

The stimulus duration was 107 ms. For 1.5 and 3 cpd spatial frequencies, we tested five base contrasts (the higher contrast in the two eyes): 0.96, 0.48, 0.24, 0.12, and 0.06, and for 6 cpd spatial frequency, we tested four base contrasts: 0.96, 0.48, 0.24 and 0.12. Stimuli were presented on a 22-inch NEC MultiSync CRT monitor with a 1920x1440 spatial pixel resolution and 75 Hz vertical refresh rate. The experiments were controlled by a Mac Mini running Matlab (MathWorks, Inc.) with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997). A special circuit (Li, Lu, Xu, Jin & Zhou, 2003) was used to yield 14 bits gray-scale levels, which ensured sub-pixel accuracy in the rending of binocular disparity even at low contrast. Gamma correction was applied and verified by measuring 10 luminance levels using a Minolta LS-110 photometer. The luminance of the monitor with all pixels set to the minimum value was 0.2 cd/m^2; the luminance with all pixels set to the maximum value was 74.2 cd/m^2. Displays were viewed in a mirror stereoscope and positioned optically 68 cm from the observer.

4.1. Psychometric functions

We modeled the psychometric function as the sum of two cumulative Gaussian distribution functions, one rising for Dmin and the other falling for Dmax when disparity increases. Fig. 3 shows sample psychometric functions (40 trials per point: 20 trials for crossed and 20 trials for uncrossed disparity) when the spatial frequency was 3 cpd, the base contrast (the higher of two eyes' contrasts) was 0.96 and the interocular contrast ratios were 0.125, 0.25, 0.5, 0.71 and 1 as labeled on the right side of the figure. Stereo performance was best for both Dmin and Dmax when the two eyes had identical contrast. Please note that Dmax is a measure of the collapse of depth perception, not of the failure of fusion.

4.2. Observers

Three observers with normal or corrected to normal vision signed the written consent forms and participated in the experiment. The data were averaged across the three observers. The experiments were conducted in accordance with the Declaration of Helsinki and the ethical permission for the study was given by Institutional Review Board (IRB) for the University of California, Berkeley.

5. Model

In the following, we first develop simple models for either phase or position disparity with either of two contrast normalization mechanisms: (1) Energy normalization (EN) after the binocular site: binocular energy is normalized by monocular energy; (2) Interocular contrast gain-controls and gain-enhancement before the binocular site (DSKL contrast normalization) (Ding & Sperling 2006; Ding et al., 2013b): the two eyes inputs first mutually suppress and enhance each other and then the binocular energy is calculated. Next, we develop a model with a conventional Max operator to explain both Dmin and Dmax thresholds.

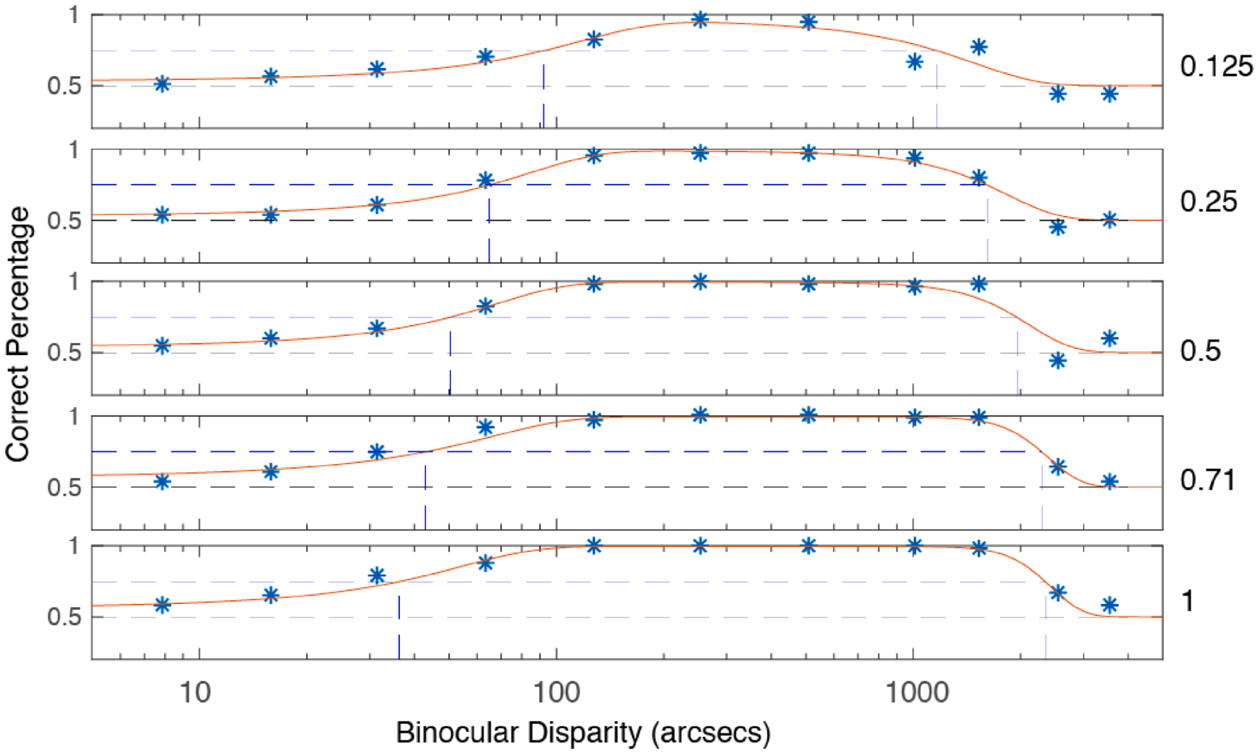

Fig. 3. Examples of psychometric functions. Probability of correct response as a function of binocular disparity. Stimulus spatial frequency was 3 cpd, the base contrast (the higher of two eyes' contrasts) was 0.96. The interocular contrast ratios were 0.125, 0.25, 0.5, 0.71 and 1 as labeled on the right side. The smooth curves are the best fit of two cumulative Gaussian functions, one for Dmin and one for Dmax. The threshold Dmin and Dmax values are the disparities that result in 75% correct responses.

Finally, we propose models for binocular fusion, and develop the unified model with both position and phase disparity detectors and binocular fusion mechanisms. In this section we provide a brief description of each model. Specific details and equations are provided in Appendices A and B.

5.1. A model with EN for phase disparity

As shown in Fig. 4A, the two eyes' images first go through two quadrature pairs of spatial filters. Each pair has filters with symmetric (even) and asymmetric (odd) profiles in the LE and RE, respectively. Linear binocular summation of the outputs of the paired filters is squared to produce binocular energy (BE). The FAR pair with an even filter in the LE and an odd filter in the RE outputs BE for uncrossed disparities, and the NEAR pair with the reversed order of filters outputs BE for crossed disparities. The difference of BE of two pairs (FAR - NEAR) is normalized by the monocular energies (EN: energy normalization) to give normalized binocular energy (NBE) with its baseline removed by the difference operation. NBE is proportional to the stimulus disparity in the range of −90 to 90 phase degrees with a positive response for uncrossed and a negative response for crossed disparity. The NBE is integrated over space and time through late-stage filters to output phase disparity energy for depth perception. This model is similar to the energy model (Adelson & Bergen, 1985) and the Reichardt detector (Reichardt, 1961; Van Santen & Sperling, 1984) for motion perception and the energy model for feature detection (Morrone & Burr, 1988). This model has a disparity-tuning curve similar to the models based on binocular energy neurons that have non-zero phase disparity preference (Fleet et al., 1996; Ohzawa, DeAngelis & Freeman, 1990; Qian & Zhu, 1997), except that the model binocular energy neuron for phase disparity has a positive baseline, and always has a positive output.

The phase disparity energy provides a good estimation of stimulus disparity when its absolute value is sufficiently small. Besides estimating the absolute value of disparity, the phase disparity energy can also detect the direction of disparity, with one quadrature pair (FAR) detecting uncrossed and the other (NEAR) detecting crossed stimulus disparity. The output phase disparity energy is given by Eq. A4,

5.2. A model with EN for position disparity

When stimulus disparity further increases beyond 90 phase degrees, the output of the phase disparity model (Fig. 4A) decreases and reaches zero at 180 phase degrees. A relative horizontal shift of the paired spatial filters is required for depth perception. Fig. 4B shows a depth perception model for position disparity. The two eyes' images first go through two pairs of spatial filters, the paired filters with identical spatial profiles (TE: tuned excitatory) but shifted relatively (u) in horizontal position. One pair of filters has identical symmetric profiles and the other pair has identical asymmetric profiles. Linear binocular summation of the outputs of the paired filters is squared to produce binocular energy (BE). The summation of BE of two TE pairs is normalized by the monocular energies (EN: energy normalization) to output the normalized BE (NBE) for depth perception. This model has a similar structure to models based on binocular energy neurons (Fleet et al., 1996, Ohzawa et al., 1990, Qian & Zhu, 1997).

Unlike the phase disparity detector, the NBE in the position disparity detector is not proportional to stimulus disparity and does not reflect its direction, but reaches the maximum when the relative shift of paired filters matches the stimulus disparity, either crossed or uncrossed. An extra operation is needed to search for the matched pair to estimate both the absolute value and direction of stimulus disparity. At a given location, the NBEs are calculated at multiple relative shifts (position disparity), either crossed or uncrossed, and the Max operation performed across the range of position disparities in both directions to select the peak NBE to estimate both maximum interocular correlation and stimulus disparity (the value and direction). The maximum correlation and matched position disparity are further translated into local position disparity energy by a depth-disparity power function, which is proportional to both the maximum correlation and the estimated disparity, giving a positive response for an uncrossed disparity and a negative response for a crossed disparity. After integration over space and time through late-stage filters, the model outputs the position disparity energy for depth perception. We note that to improve its matching performance, the Max operator typically follows the late-stage filters or operations in the literature (Filippini & Banks, 2009, Fleet et al., 1996). However, in this study, we assumed an ideal Max operator without any false matching and placed it before the late-stage filters. Because we did not have data to test the late-stage filters in this study, for simplicity, we did not put them in the following model diagrams.

5.3. Depth models with DSKL contrast normalizations

Ding et al. (2013) proposed five nested gain-control-gain-enhancement models (DSKL models) to explain binocular contrast and phase combination. They further elaborated their model to explain binocular combination of luminance profiles (Ding & Levi 2017). Here, we expand these models for depth perception. Unlike energy normalization after the site of binocular combination, the DSKL model normalizes monocular contrast before the binocular site (Fig. 4C), which successfully predicts contrast-dependent depth perception (Ding & Levi, 2016a; Hou et al., 2013). As shown in Fig. 4C, in a narrow spatial-frequency band, the LE's signal (Black) is gain-controlled (Blue) and gain-enhanced (Red) by the RE's total contrast energy (TCE), which is a weighted summation over spatial-frequency bands. The RE's gain-control (Blue) and gain-enhancement (Red) of the LE are gain-controlled by the LE's TCE. For simplicity, Fig. 4C shows a half of the DSKL circuit for the LE's normalization; the other half, for the RE's normalization, has a symmetric structure. After normalization, the two eyes' signals are linearly combined and squared to produce BE. Similarly, the BE of the other pair is calculated. The combination of BEs of the two pairs is output for depth perception. Other aspects of the model are similar to Fig. 4A for phase disparity, or to Fig. 4B for position disparity.

5.4. A unified model with Max operator for both Dmin and Dmax

To explain both Dmin and Dmax thresholds, we developed a unified model with a Max operator. The model assumes a depth-disparity function, which is the product of a disparity power function and a disparity exponential decay function, initially increasing with disparity and then decreasing exponentially with further increases in disparity. An earlier version of this model was presented by Ding & Levi (2016a) and was further tested at supra threshold levels over the whole range of stimulus disparities (Ding & Levi 2019).

Because our modeling (see section of Model fitting results) showed that Dmin and Dmax have different contrast normalization mechanisms, we developed a unified model with both first- and second-order pathways, each with a different contrast normalization mechanism, as shown in Fig. 5. In the first-order pathway, the two eyes images pass through two pairs of first-stage spatial-frequency filters, each pair of filters with identical in profile (TE) but shifted in horizontal position (position disparity). The summation of BE of two TE pairs is normalized by monocular energy and goes through a maximum (MAX) operator for estimating the maximum correlation and selecting the matched position disparity, and then goes through a depth disparity function to compute local position disparity energy.

Previous studies (Ding & Levi, 2017, Zhou, Georgeson & Hess, 2014) showed that the second-order signals (contrast modulations) were first normalized based on first-order contrast energy before binocular combination, and that the rectification occurs before the binocular site because the second-order binocular combination is independent of interocular correlation of first-order carries (Ding & Levi, 2017, Zhou

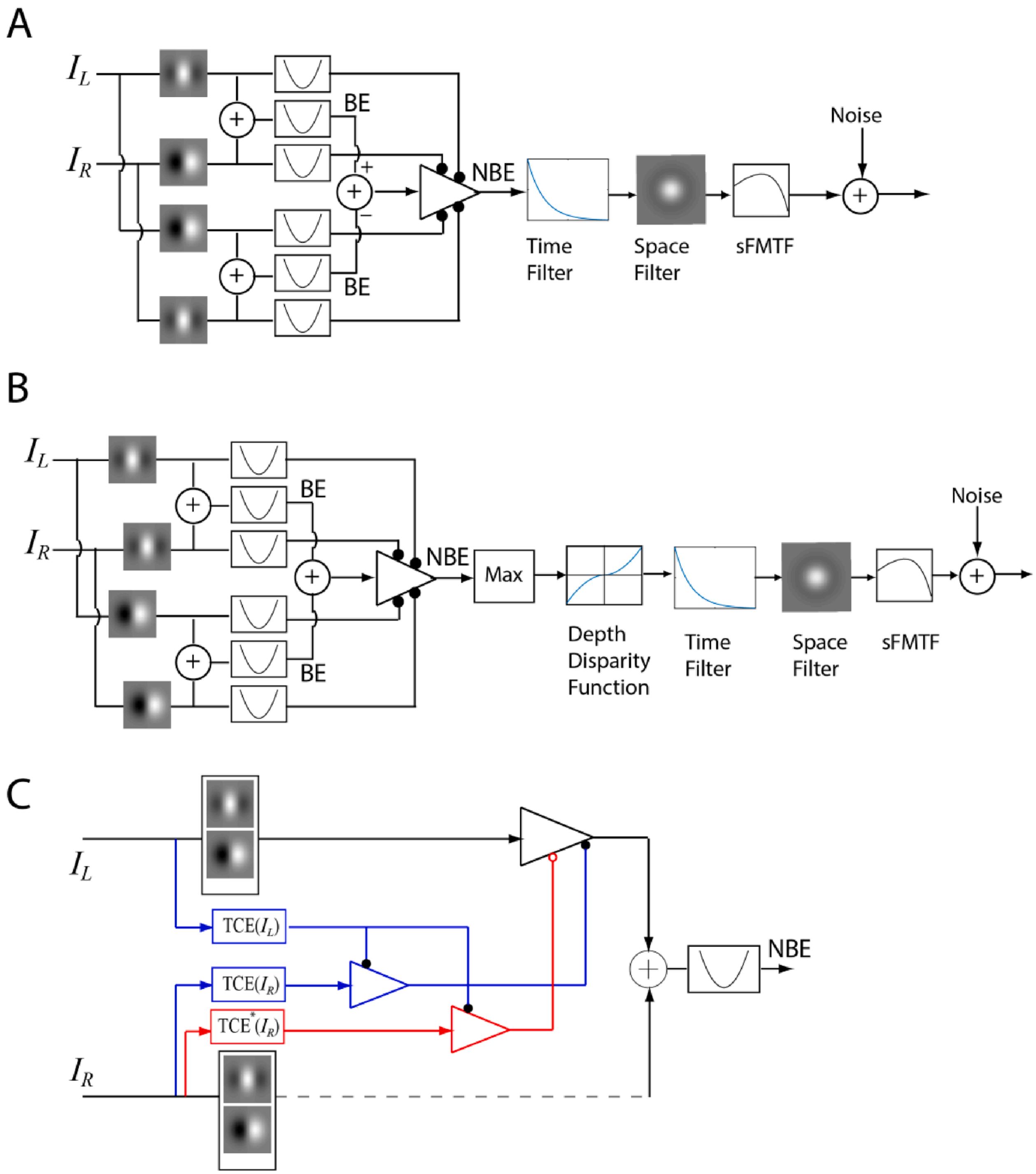

Fig. 4. Models. (A). A depth perception model for phase disparity with energy normalization. In each spatial frequency channel, the two-eyes' images first go through two pairs (FAR and NEAR) of filters. The paired filters are in the same position in each eye (no position disparity) but differ in phase by −90° (FAR) for one pair and 90° (NEAR) for the other pair. Linear binocular summation of the outputs of the paired filters is squared to produce binocular energy (BE). The difference of BE of two pairs (FAR - NEAR) is normalized by the monocular energies and is integrated over space and time through later-stage filters to output phase disparity energy for depth perception. The spatial-frequency modulation transfer function (sFMTF) is included to reflect the fact that the phase disparity energy varies with spatial frequency. (B). A depth perception model for position disparity with energy normalization. The model has a similar structure to the phase-disparity detector in (A) except that the paired filters are in different positions in the two eyes (position disparity) but with identical phase (TE: no phase disparity). The Max operator selects the peak of the normalized binocular energy (NBE) to estimate maximum correlation and stimulus disparity (either crossed or uncrossed), which is further translated into depth information by a depth-disparity power function and is integrated over space and time through late-stage filters. (C) A depth perception model for position disparity with DSKL contrast normalization. The contrast normalization is performed before binocular combination through a DSKL circuit (details see Appendix B and (Ding et al., 2013b)). For simplicity, only partial model with a half of DSKL circuit is shown. The other half of DSKL is symmetric, and the other parts of the model are similar to (B).

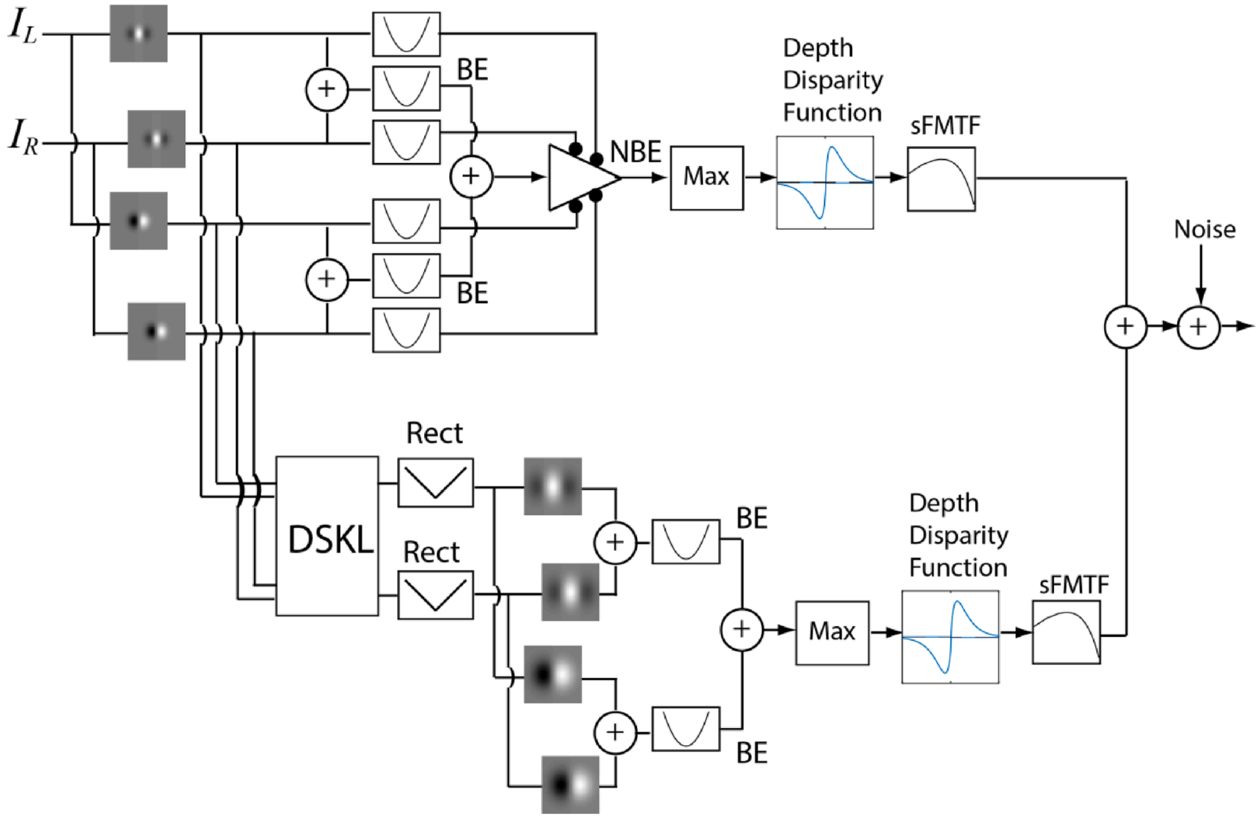

Fig. 5. A unified model with Max operator for both Dmin and Dmax. In the first-order pathway, the two eyes images pass through two TE pairs of first-stage spatial-frequency filters, each pair of filters with identical in profile but shifted in horizontal position (position disparity). The summation of BE of two TE pairs is normalized by monocular energy and goes through a maximum (MAX) operator for estimating the maximum correlation and selecting the matched position disparity, and then goes through a depth disparity function to compute local position disparity energy. In the second-order pathway, the two eyes' outputs of first-stage filters go through the DSKL circuit for contrast normalization before the binocular site. After rectification, the two eyes' second-order contrast modulations go through two TE pairs of second-stage spatial-frequency filters. The summation of BE of two TE pairs goes through the MAX operator and a depth-disparity function to output for second-order position disparity energy. A depth-disparity function is the product of a disparity power function and a disparity exponential decay function, initially increasing with disparity and then decreasing exponentially with further increases in disparity.

et al., 2014). Zhou, Georgeson & Hess (2014) reported linear binocular summation of second-order contrast modulation when the first-order contrast remains constant, which can be explained by the contrast-weighted summation model, a simplified DSKL model (see Appendix B). Ding & Levi (2017) expanded the DSKL model to explain both the first- and second-order binocular combination and tested it measuring second-order binocular combination when first-order contrast varied in the two eyes.

In the present study, we propose DSKL interocular interactions (Fig. 4C) as the contrast normalization mechanism in the second-order pathway for depth perception. As shown in Fig. 5, the two eyes' outputs from first-stage filters go through the DSKL circuit for contrast normalization before the binocular site. After rectification, the two eyes' second-order contrast modulations go through two pairs of second-stage spatial-frequency filters, each pair of filters with identical profiles (TE) but shifted in horizontal position. The summation of BE of the two TE pairs goes through the MAX operator and a depth-disparity function to output second-order position disparity energy. Except for the DSKL contrast normalization and the disparity decay in depth-disparity function, the model's second-order disparity detector is similar to the second-stage convergence model proposed by Tanaka & Ohzawa (2006) to explain their physiology data, and the second-order energy model for binocular disparity in natural images (Hibbard, Goutcher & Hunter, 2016). There are also some close similarities between the present model for binocular combination of second-order modulations (Fig. 5) and the model of Georgeson & Schofield (2016).

For simplicity to fit real data, we assume an ideal Max operator that always selectively reads out a pair of filters with a relative horizontal shift that matches the stimulus disparity for depth perception, and a depth-disparity decay function to interpret decreasing depth performance at large stimulus disparities. This contradicts the idea that the Max operator usually fails to select the correct match at large stimulus disparities, resulting in diplopic images (McKee & Verghese, 2002, Richards, 1971, Schor & Wood, 1983) and poor depth performance. In the following, we propose a sensory fusion mechanism to solve this contradiction.

5.5. 3D binocular fusion mechanism

In previous studies we (Ding, Klein, and Levi, 2013a and 2013b) proposed the DSKL model with a 2D binocular fusion mechanism to explain binocular phase and contrast combination. Here, we elaborate this 2D mechanism to 3D binocular fusion, as a mechanism to solve the correspondence problem in depth perception.

For simplicity, we remove the base line from the binocular energy, which is equivalent to calculation of cross correlation (CC). Fig. 6A shows CC vectors in the complex plane. The positive real part of CC (RC^+) reflects 0° phase disparity energy with TE tuning type, which is calculated by the paired filters with identical spatial profiles. The negative real part of CC (RC^-) reflects ± 180° phase disparity energy with TI tuning type, which is calculated by the paired filters with anti-correlated spatial profiles. The positive imaginary part of CC (IC^+) reflects 90° phase disparity energy with NEAR tuning type, and the negative imaginary part of CC (IC^-) reflects −90° phase disparity energy with FAR tuning type. IC^+ and IC^- are calculated by the paired filters with orthogonal spatial profiles. A depth sensor comprises an array of

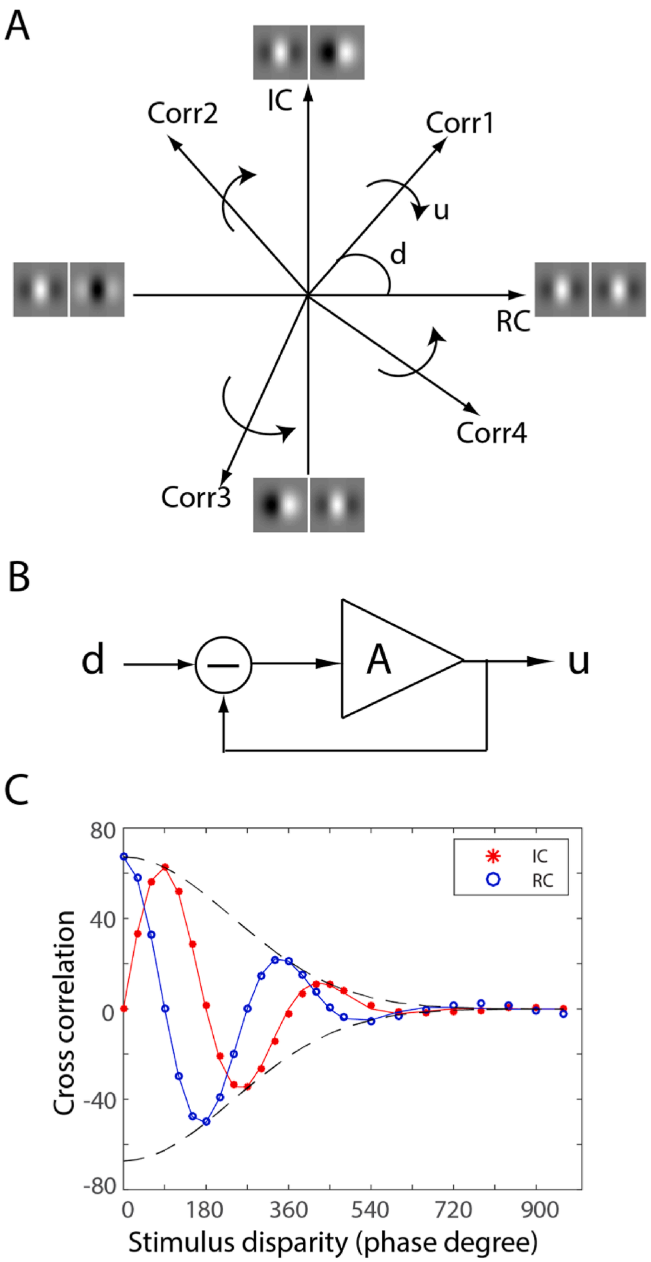

Fig. 6. Binocular fusion mechanism. (A) Cross correlations (CC) represented in the complex plane. In a narrow spatial-frequency band, an interocular cross-correlation can be represented in a complex plane with its real axis (RC: real part of CC) as the correlation output of paired filters with identical spatial profile, and its imaginary axis (IC: imaginary part of CC) as the correlation output of paired filters with orthogonal spatial profiles. The angle of a CC is the stimulus disparity *d* in phase degree unit. (B) A binocular fusion process with a negative feedback loop. A motor/sensory shift *u* is made to reduce the misalignment of monocular outputs until the final misalignment *d* - *u* is less than a threshold. (C) Interocular cross-correlation in a narrow spatial-frequency band when using random-dots stereograms as input stimuli.

pairs, sampling both position and phase disparities. At each position disparity, there are four phase disparity detectors, TE, NEAR, TI and FAR, to calculate RC$^+$, IC$^+$, RC$^-$ and IC$^-$. As shown in Fig. 6A, in the fixation plane, there are four pairs with no relative horizontal shift to calculate RC$^+$, IC$^+$, RC$^-$ and IC$^-$. We assume that the phase disparity energy IC$^+$ (NEAR) drives convergence shift to rotate CC1 and CC2, and the phase disparity energy IC$^-$ (FAR) drives divergence shift to rotate CC3 and CC4 to the positive real axis (0 phase disparity). When the two eyes are presented with images with stimulus disparity d (in phase degrees), which have cross-correlation of CC1 ($0^\circ < d \leq 90^\circ$) in the fixation plane, the IC$^+$ drives motor/sensory fusion to reduce the misalignment of monocular outputs, i.e., shifting the input CC1 vector (of the two eyes' images) to align it with the positive real axis (motor fusion) or by shifting (selecting) the depth sensor's output along the position disparity dimension (see Fig. 1), from the fixation plan (the output of four pairs with no position disparity is the sensor's output), to align it with CC1 where the output of a different set of four pairs with position disparity of d is the sensor's output (sensory fusion). In short, motor fusion rotates a CC vector to align it with the axis, while sensory fusion rotates the axis to align it with a CC vector. This motor/sensory fusion system has a negative feedback loop to reduce the misalignment of monocular outputs, as shown in Fig. 6B. The initial stimulus disparity *d* provides an initial driving force to make a fusion shift *u*, resulting in a reduced misalignment *d-u* with less driving force. Under the steady state, we have $u = A(d\text{-}u)$, i.e.,

$$u = \frac{A}{1+A}d \tag{3}$$

where A is the binocular fusion force, which is proportional to phase disparity energy in the fixation plane. When $A \gg 1$ at high stimulus contrast and long duration, the system achieves perfect fusion with $u \approx d$. When $A \ll 1$ at low stimulus contrast and/or short duration, no fusion shift occurs with $u \approx 0$.

When stimulus disparity $90^\circ < d < 180^\circ$ (CC2 in Fig. 6A), the phase disparity energy RC$^-$ (TI) has a positive output. The combination of RC$^-$ and IC$^+$ drives CC2 first to 90° misalignment to eliminate RC$^-$, and then IC$^+$ drives it to align it with the positive real axis. However, when $d = 180^\circ$, the system reaches an unstable equilibrium without shifting direction because IC$^+$ = IC$^-$ = 0. Whenever deviating from $d = 180^\circ$ to $d < 180^\circ$, e.g., because of noise, the fusion process resumes. A similar procedure is performed to fuse CC3 ($-180^\circ < d < -90^\circ$) and CC4 ($-90^\circ \leq d < 0^\circ$) but in the opposite direction. When stimulus disparity $d > 180^\circ$, binocular fusion fails, driving a CC vector to a wrong direction. In this current version of the fusion mechanism, binocular fusion has a half-cycle limit in a narrow-pass spatial-frequency band. However, under a coarse-to-fine process (Marr & Poggio, 1979), fusion can be achieved beyond the half-cycle limit. For example, when stimulus disparity is 360° phase degrees at a small scale, the same stimulus will have a disparity of only 90° phase degrees at a 4x larger scale. After rotation at the larger scale, the misalignment might be reduced to<180° phase degrees at a smaller scale.

When $u = d + n{*}360^\circ$ ($n = 0, \pm1, \pm2\cdots$), the system reaches a stable equilibrium where the interocular misalignment reaches a local minimum and the interocular correlation reaches a local maximum. The system always goes back to a stable equilibrium when deviating off it because of noise. When $u = d + (2n + 1){*}180^\circ$ ($n = 0, \pm1, \pm2\cdots$), the system reaches an unstable equilibrium, where the two eyes' images become anticorrelated, a local maximum misalignment without fusion shifting direction. The system shifts away from an unstable equilibrium whenever deviating off it. However, the system might remain unstable if the fusion force is not sufficient to shift it to the next stable equilibrium. Only the equilibrium of $u = d$ is a real match, and all other equilibria are false matches.

Binocular difference channels (Cohn & Lasley, 1976, Georgeson et al., 2016, Kingdom, Jennings & Georgeson, 2018, May, Zhaoping & Hibbard, 2012) could provide 180-degree position/phase disparity detectors.

Fig. 6C shows CC simulations, RC (Eq. A7) (blue in Fig. 6C) and IC (Eq. A8) (red in Fig. 6C) as functions of stimulus disparity, after filtering random-dots stereograms with paired filters (Eqs. A1 and A2). The colored markers indicate the simulation points and smooth colored curves are the best fits (Eqs. A9 and A10). The CC disparity decay is indicated by the dashed black curves. When using RGP stereograms as stimuli, the CC simulation is similar as in Fig. 6C.

5.6. A depth model with first-order sensory fusion mechanism.

In the following, we develop models (Figs. 7 and 8) with 0°-phase (TE: position disparity detectors) and 90°-phase disparity detectors (NEAR and FAR) for fitting our experimental data. A depth model including 180-degree phase disparity detectors (TI) is too complex for modeling because both TI and NEAR/FAR detectors drive sensory fusion shifts and their coefficients might be partially dependent on each other for modeling and therefore could not be obtained reliably with only threshold data.

If a pair of 2D images has only a single disparity, perfect alignment can be achieved by motor fusion alone (based on absolute disparity) if the stimulus duration is sufficiently long. However, with multiple disparities, sensory fusion is necessary to align the pair of 2D images locally. Fig. 7 shows the depth model with a sensory fusion mechanism in a first-order pathway. The two eyes' images first go through four pairs of luminance filters, two TE pairs with identical spatial profiles for position disparity, and NEAR and FAR pairs with orthogonal profiles for phase disparity. After binocular combination, binocular energy is normalized by monocular energy. The phase disparity detector (FAR - NEAR) is similar to Fig. 4A including late-stage filters for integration over space and time (not shown in Fig. 7). The position disparity detector (TE), also including late-stage filters (not shown in Fig. 7), has a fusion mechanism to search for a position disparity that matches with stimulus disparity without using the Max operator.

Sensory fusion occurs when the relative shift (u, the position disparity) of the paired filters equals the stimulus disparity (d). At the sensory-fused plane ($u = d$), the outputs of the paired filters with identical profiles (TE) are aligned with each other locally, giving locally-fused 2D images, and their relative shift u gives depth perception (D), while the outputs of the paired filters with orthogonal profiles (NEAR and FAR) are uncorrelated (i.e., no phase disparity). However, when u is close, but not equal to d, the non-zero phase disparity energy shifts all pairs of paired filters simultaneously until $u = d$ to align the two eyes images locally at depth D. At low stimulus contrast and/or short stimulus duration and/or large stimulus disparity, the fusion energy is not sufficient for perfect fusion, and the model uses the phase disparity energy to estimate the difference between stimulus disparity and position disparity, i.e., d – u, after fusion. This estimation is accurate if d – u is sufficient small but has an error at a large difference especially when d – u > 90 phase degree, at which diplopia or local blur might occur, and depth perception might decrease. The disparity decay in the depth-disparity function, assumed in the convention model (Fig. 5), can be explained by weak fusion energy at large stimulus disparities. In the model with the fusion mechanism (Fig. 7), the relative shift of the paired filters is transferred to position disparity energy by a depth-disparity power function, with a positive response for an uncrossed disparity and a negative response for a crossed disparity. Depth is perceived based on the combination of position and phase disparity energies after the fusion process. In general, phase disparity energy might have different coefficients for the fusion mechanism and depth perception. We include a model parameter ϕ to reflect this difference (Fig. 7).

Sensory fusion has quite often been described as a coarse-to-fine process (Marr & Poggio, 1979): it occurs at a large scale first to

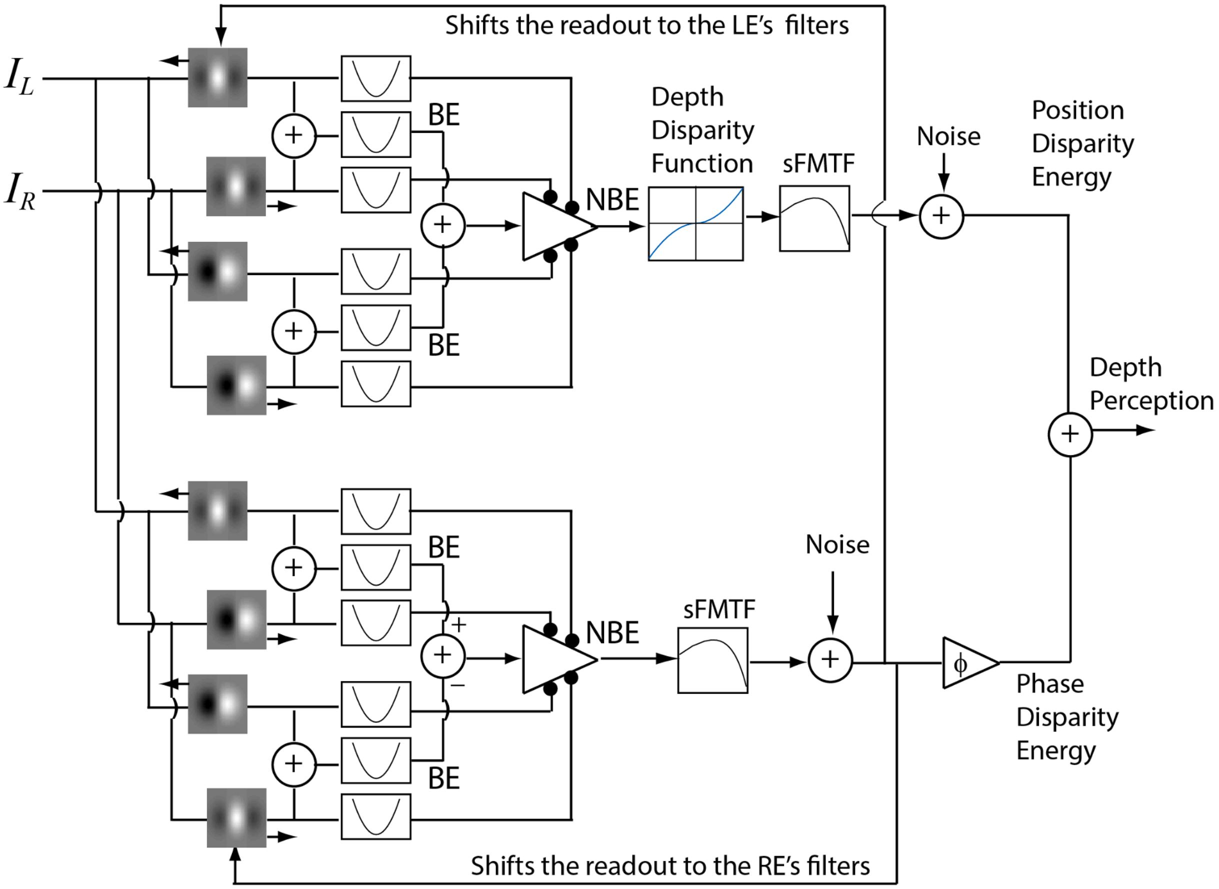

Fig. 7. A depth model with first-order sensory fusion mechanism. The two eyes' images first go through four pairs of luminance filters, two pairs (TE) for position disparity and two pairs (NEAR and FAR) for phase disparity. After binocular combination, the binocular energy is normalized by monocular energy. The phase disparity energy is assumed to shift the two eyes' paired filters relatively until phase disparities is eliminated (sensory fusion). After sensory fusion, the combination of position and possible residual phase disparity energies is calculated for depth perception. The spatial-frequency modulation transfer function (sFMTF) is included to reflect the fact that the disparity energy varies with spatial frequency. The coefficient ϕ is included with phase-disparity energy for depth perception. Late-stage filters are not shown for simplicity.

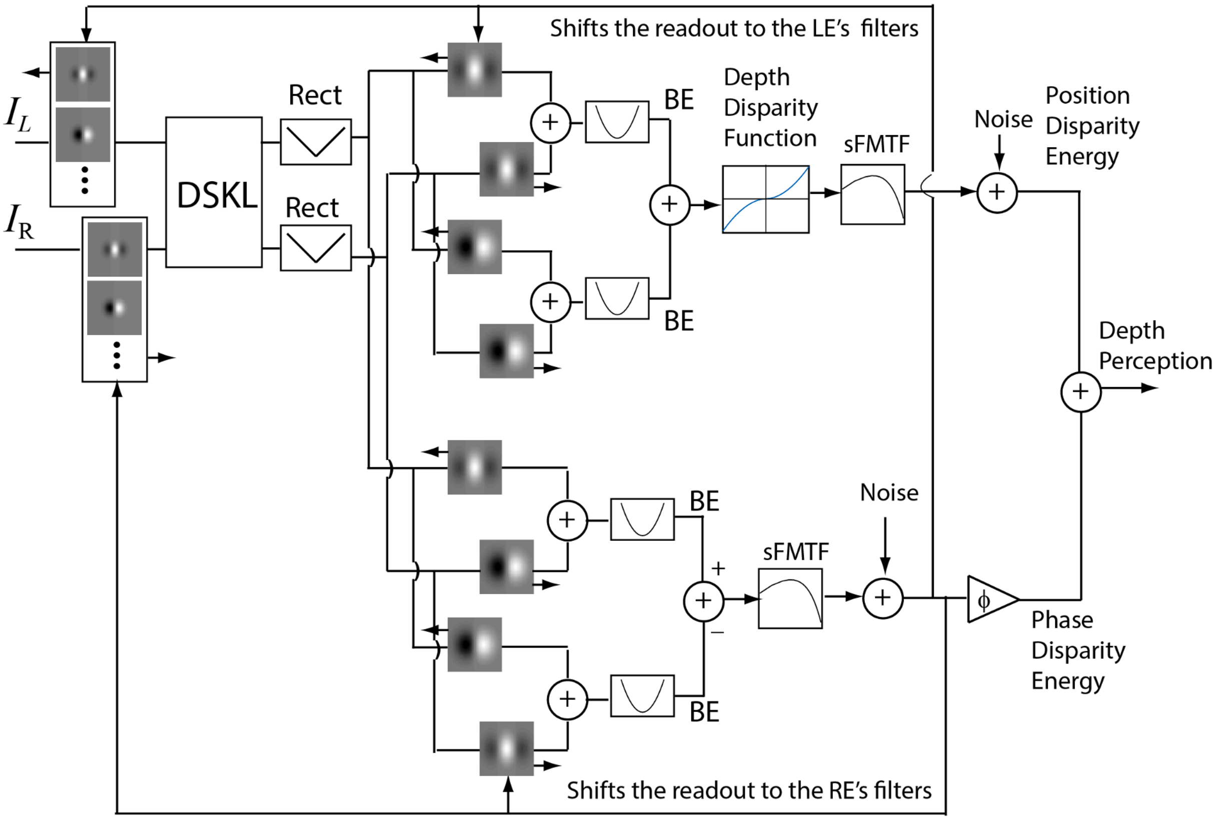

Fig. 8. A second-order sensory fusion mechanism. After going through the first-stage spatial-frequency filters, the two eyes' images go through the DSKL circuit for contrast normalization before the binocular site. After the rectification, the two eyes' second-order signals (contrast modulations) go through four pairs of second-stage filters, two TE pairs for second-order position disparity and both NEAR and FAR pairs for second-order phase disparity. The second-order phase-disparity energy (FAR - NEAR) is assumed to shift both first- and second-stage filters relatively until the second-order phase disparity is eliminated (second-order sensory fusion). After second-order sensory fusion, the combination of second-order position and possible residual phase disparity energies is calculated for second-order depth perception, and the first-order pathway begins to align the two eyes images in a smaller scale and to estimate the relative depth to the second-order depth plane. The spatial-frequency modulation transfer function (sFMTF) is included to reflect the fact that the second-order disparity energy varies with spatial frequency. The coefficient ϕ is included with phase-disparity energy for depth perception.

coarsely align the two images, and then at a smaller scale(s) to achieve fine alignment. Sensory fusion transfers the phase disparity energy in the fixation plane to the position disparity energy in a depth plane at each location for local depth perception. However, we are not clear whether the disparity energy is conserved (the summation of phase and position disparity energies remains constant) during the sensory fusion process.

5.7. A depth model with second-order sensory fusion mechanism

Fig. 8 shows the depth model with a sensory fusion mechanism in a second-order pathway. Except for the DSKL contrast normalization and sensory fusion mechanism, the second-order position and phase disparity detectors are similar to the second-stage convergence model proposed by Tanaka & Ohzawa (2006) for explaining their physiological data, and the second-order energy model for binocular disparity in natural images (Hibbard et al., 2016). After going through the first-stage spatial-frequency filters, the two eyes' images go through the DSKL circuit (Ding et al., 2013b) for contrast normalization before the binocular site. After rectification before the binocular site, the two eyes' second-order signals (contrast modulations) pass through four pairs of second-stage filters, paired filters with identical spatial profiles (TE) for second-order position disparity and paired filters with orthogonal profiles (NEAR and FAR) for second-order phase disparity. After binocular combination, the second-order phase disparity energy (FAR - NEAR) is first integrated over space and time through late-stage filters (Not shown in Fig. 8) and then drives sensory fusion of both first- and second-stage filters until the second-order phase disparity eliminated. In the fused second-order position disparity plane, both first- and second-stage paired filters have the same relative shift to the fixation plane. After the second-order sensory fusion, the combination of second-order position and possible residual phase disparity energies is calculated for second-order depth perception, and the first-order pathway initiates first-order sensory fusion to align two eyes' images at a smaller scale and to estimate the relative depth to the second-order depth plane. This is the typical coarse-to-fine process of sensory fusion.

5.8. A depth model with motor fusion mechanism

A large binocular disparity typically results in vergence eye movement to realign the two images to achieve binocular motor fusion. Generally speaking, multiple spatial-frequency channels are involved in motor fusion. However, for our RGP stereograms, the biggest driving force comes from a second-order spatial-frequency channel whose spatial wavelength is ~ 5–10 times the stimulus spatial wavelength (Ding & Levi 2020). The model structure is similar to the model with second- and first-order sensory fusion (Figs. 7 and 8) except that both second and first-order fusion energies drive vergence eye-movements to align the two eyes' images globally. Unlike sensory fusion which follows a coarse-to-fine process for local fusion at different scales, motor fusion is a global process driven by total global fusion energy summed over space and time across spatial-frequency channels.

Base on efficient coding and task learning of the joint development of stereo disparity perception and vergence eye movements, Zhao et al. (2012) developed a model that, through motor fusion, drives absolute disparity towards zero. Our fusion mechanism also seeks to reduce estimated absolute disparity to zero through motor fusion driven by fusion energy at all spatial-frequency scales. At a large disparity, outside of the half-cycle limit of a small scale, the fusion energy at a large scale is the main force for driving motor fusion. After the disparity is reduced to within the half-cycle limit, the fusion energy at both large and small scales drives motor fusion. After further reducing disparity, the phase disparity at the large scale becomes very small, and the fusion energy at the small scale becomes the main force driving the absolute disparity towards zero.

Our Nonius alignment experiments showed that at least two pathways (one second- and one first-order pathways) are involved with vergence eye movements when using RGP stereograms (Fig. 2) as stimuli (Ding & Levi 2020). In this study, because we did not have sufficient data to test this model (our stimulus duration is too short for motor vergence), we do not elaborate it in detail.

5.9. Full model with sensory fusion mechanism

Fig. 9 shows the full model with multiple scales and with either second- or first-order sensory fusion for each scale. For simplicity, model details (e.g., contrast normalizations) are not shown. The two eyes' images first go through multiple scale spatial-frequency filters and then go through both phase and position disparity detectors on each scale. The large-scale (LS) phase disparity detector detects a LS misalignment (phase-disparity energy), which drives a sensory readout shift (a LS position disparity) of all scales' filters relative to the fixation plane to reduce the LS misalignment. In other words, the output of a depth sensor is selected for readout from the fixation plane to the LS position disparity plane that is sampled by all scales' filters. The medium-scale (MS) phase disparity detector (not shown in Fig. 9) detects a MS misalignment, which drives a sensory shift (a MS position disparity) of all filters with scales not larger than the MS, relative to the LS position-disparity plane to reduce the MS misalignment. In general, there are multiple medium scales, but here we only describe one of them. The small-scale (SS) phase disparity detector detects a SS misalignment, which drives a sensory shift (a SS position disparity) of the SS filters relative to the MS position-disparity plane to reduce the SS misalignment. After sensory fusion, at each scale, the position disparity energy is measured by the position disparity detector and combined with the residual phase disparity

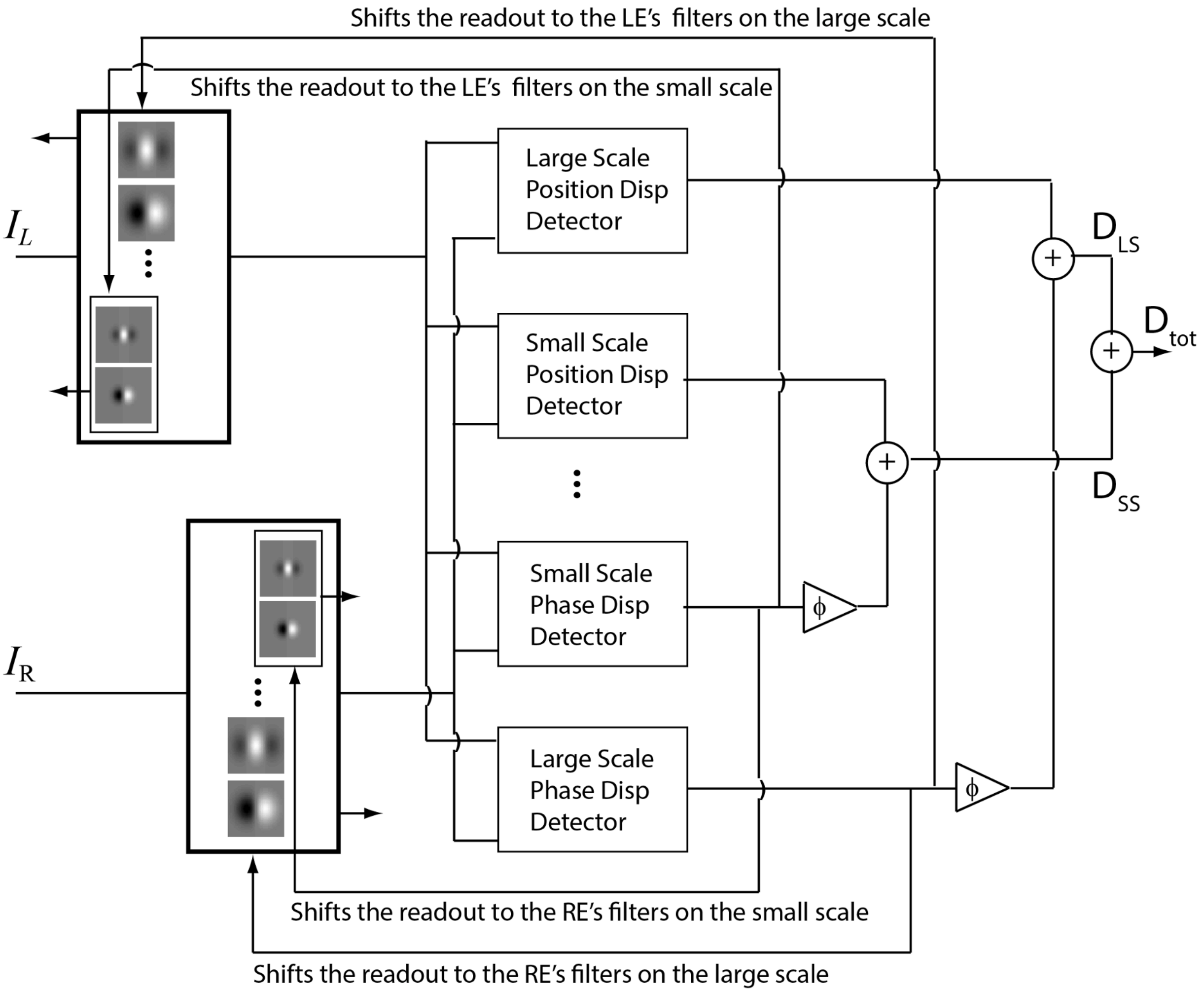
Shifts the readout to the LE's filters on the large scale
Shifts the readout to the LE's filters on the small scale
I_L
Large Scale Position Disp Detector
Small Scale Position Disp Detector
Small Scale Phase Disp Detector
Large Scale Phase Disp Detector
D_{LS}
D_{tot}
D_{SS}
+
ϕ
I_R
Shifts the readout to the RE's filters on the small scale
Shifts the readout to the RE's filters on the large scale

Fig. 9. Full model with sensory fusion mechanism. After going through spatial-frequency filters over multiple spatial scales, the two eyes' images go through phase and position disparity detectors of all scales. The large-scale (LS) phase disparity detector detects a LS misalignment (phase-disparity energy), which drives a sensory shift (a LS position disparity) of all scales' filters, relative to the fixation plane, to reduce the LS misalignment. The medium-scale (MS) phase disparity detector (not shown) detects a MS misalignment, which drives a sensory shift (a MS position disparity) of all filters with scales not larger than the MS, relative to the LS position-disparity plane, to reduce the MS misalignment. The small-scale (SS) phase disparity detector detects a SS misalignment, which drives a sensory shift (a SS position disparity) of the SS filters, relative to the MS position-disparity plane, to reduce the SS misalignment. Then, at each scale, the position disparity energy is measured by the position disparity detector and combined with the residual phase disparity energy for depth perception. The total disparity energy is a weighted summation across all scales.

energy for depth perception. The total disparity energy is a weighted summation across all scales.

Although we described this process stage by stage, all detectors might work simultaneously. When images are presented to the two eyes in the fixation plane, the position disparity detectors at all scales detect zero position disparity and the phase disparity detectors at all scales output phase disparity energy for both depth perception and sensory fusion. At a large stimulus disparity (within half limit of LS sensory fusion, but out of half-limits of MS and SS sensory fusion), phase disparities might have different directions at different scales (i.e., their combination might give ambiguous depth perception), and only LS sensory fusion reduces the misalignment, by selectively shifting the signals for readout from sensors in the fixation plane to those in the LS position disparity plane where the MS and SS detectors continue the process of fusion and depth perception. Therefore, depth perception first appears ambiguous, then clear on a coarse scale and then on a fine scale, following a coarse-to-fine process. However, at a small stimulus disparity, the combined phase disparity energy of all scales might give a reliable depth perception before sensory fusion, and the SS sensory fusion might occur without the LS sensory shift.

5.10. Full model with sensory fusion mechanism for RGP stereograms

For simplicity, we developed the full model for RGP stereograms with a single scale for second-order sensory fusion (~4–8 times the stimulus spatial wavelength) and a single scale for first-order sensory fusion (with the same scale as the stimulus) to explain both Dmin and Dmax thresholds. In the following, we provide some of the formulae that we used in the modeling. Their derivations are described in Appendices A and B. For an RGP stereogram with stimulus disparity d (Eqs. (1) and (2)), based on CC simulation (Fig. 6C), the second-order fusion energy with DSKL contrast normalization mechanism is given by:

$$\widehat{\mathscr{F}}_{\omega_{\rm 2nd}} = h_{\omega_{\rm 2nd}}\widehat{m}_{\rm L}\widehat{m}_{\rm R}\sin\left(\frac{\omega}{a}d\right)\exp\left(-\frac{|d|}{\tau_{\rm 2nd}}\right)+\mathscr{N}(0,\sigma_N) \tag{4}$$

where ω is the stimulus spatial frequency, $\widehat{m}_{\rm L}$ and $\widehat{m}_{\rm R}$ are equivalent contrast (Eqs. B1-B5) after DSKL contrast normalization, and a is the spatial scale factor of first- and second-stage filters, and $\tau_{\rm 2nd}$ is the disparity decay constant. Because of the decay of fusion energy, depth perception decreases at large stimulus disparities. The decay of second-order fusion energy gives a reasonable explanation for Dmax threshold.

The second-order sensory shift $u_{\omega_{\rm 2nd}}$ is given by Eq. (3) with $A = \left|\widehat{\mathscr{F}}_{\omega_{\rm 2nd}}\right|^q$. Here, we note that the fusion energy should go through a temporal filter to be integrated over time to calculate the fusion force A. However, because the stimulus duration was constant in this study, the effect of the temporal filter is assumed to be constant, which was also included in the coefficient $h_{\omega_{\rm 2nd}}$ of the channel.

The position disparity energy produced by the second-order sensory shift is given by:

$$\widehat{D}_{\omega_{\rm 2nd}} = k^p_{\omega_{\rm 2nd}}\,{\rm sign}\left(u_{\omega_{\rm 2nd}}\right)\left|u_{\omega_{\rm 2nd}}\right|^p\widehat{m}_{\rm L}\widehat{m}_{\rm R}\cos\left(\frac{\omega}{a}\left(d-u_{\omega_{\rm 2nd}}\right)\right)\exp\left(-\frac{\left|d-u_{\omega_{\rm 2nd}}\right|}{\tau_{\rm 2nd}}\right) +\mathscr{N}(0,\sigma_N) \tag{5}$$

After second-order sensory fusion, the second-order phase disparity energy for depth perception is given by:

$$\widehat{D}^{\phi}_{\omega_{\rm 2nd}} = \phi_{\omega_{\rm 2nd}}h_{\omega_{\rm 2nd}}\widehat{m}_{\rm L}\widehat{m}_{\rm R}\sin\left(\frac{\omega}{a}\left(d-u_{\omega_{\rm 2nd}}\right)\right)\exp\left(-\frac{\left|d-u_{\omega_{\rm 2nd}}\right|}{\tau_{\rm 2nd}}\right)+\mathscr{N}(0,\sigma_N) \tag{6}$$

and the first-order sensory fusion energy is given by:

$$\widehat{\mathscr{F}}_{\omega_{\rm 1st}} = h_{\omega_{\rm 1st}}\frac{m_{\rm L}m_{\rm R}}{Z^2+m_{\rm L}^2+m_{\rm R}^2}\sin\left(\omega\left(d-u_{\omega_{\rm 2nd}}\right)\right)\exp\left(-\frac{\left|d-u_{\omega_{\rm 2nd}}\right|}{\tau_{\rm 1st}}\right)+\mathscr{N}(0,\sigma_N) \tag{7}$$

and the sensory shift $u_{\omega_{\rm 1st}}$ is given by Eq. (3) with $A = \left|\widehat{\mathscr{F}}_{\omega_{\rm 1st}}\right|^q$. The position disparity energy produced by first-order sensory shift is given by:

$$\widehat{D}_{\omega_{\rm 1st}} = k^p_{\omega_{\rm 1st}}\,{\rm sign}\left(u_{\omega_{\rm 1st}}\right)\left|u_{\omega_{\rm 1st}}\right|^p\frac{m_{\rm L}m_{\rm R}}{Z^2+m_{\rm L}^2+m_{\rm R}^2}\cos\left(\omega\left(d-u_{\omega_{\rm 2nd}}-u_{\omega_{\rm 1st}}\right)\right)\exp\left(-\frac{\left|d-u_{\omega_{\rm 2nd}}-u_{\omega_{\rm 1st}}\right|}{\tau_{\rm 1st}}\right)+\mathscr{N}(0,\sigma_N) \tag{8}$$

After first-order sensory fusion, the first-order phase disparity energy for depth perception is given by:

$$\widehat{D}^{\phi}_{\omega_{\rm 1st}} = \phi_{\omega_{\rm 1st}}h_{\omega_{\rm 1st}}\frac{m_{\rm L}m_{\rm R}}{Z^2+m_{\rm L}^2+m_{\rm R}^2}\sin\left(\omega\left(d-u_{\omega_{\rm 2nd}}-u_{\omega_{\rm 1st}}\right)\right)\exp\left(-\frac{\left|d-u_{\omega_{\rm 2nd}}-u_{\omega_{\rm 1st}}\right|}{\tau_{\rm 1st}}\right)+\mathscr{N}(0,\sigma_N) \tag{9}$$

The total fusion energy is given by:

$$\widehat{\mathscr{F}}_{\omega} = \widehat{\mathscr{F}}_{\omega_{\rm 2nd}}+\widehat{\mathscr{F}}_{\omega_{\rm 1st}}+\mathscr{N}(0,\sigma_N) \tag{10}$$

The total position disparity energy is given by:

$$\widehat{D}_{\omega} = \widehat{D}_{\omega_{\rm 2nd}}+\widehat{D}_{\omega_{\rm 1st}}+\mathscr{N}(0,\sigma_N) \tag{11}$$

The total disparity energy for depth perception is given by:

$$\widehat{D}^{\rm tot}_{\omega} = \widehat{D}_{\omega_{\rm 2nd}}+\widehat{D}^{\phi}_{\omega_{\rm 2nd}}+\widehat{D}_{\omega_{\rm 1st}}+\widehat{D}^{\phi}_{\omega_{\rm 1st}}+\mathscr{N}(0,\sigma_N) \tag{12}$$

For simplicity, we assume $\sigma_N = 1$. The threshold Dmax is defined as the disparity at which ${\rm mean}\left(\widehat{\mathscr{F}}_{\omega}\right) = \sigma_N = 1$. The threshold Dmin is defined as the disparity at which ${\rm mean}\left(\widehat{D}^{\rm tot}_{\omega}\right) = \sigma_N = 1$.

6. Modeling

6.1. F-test for comparison of nested models

If Model a is nested within Model b, the F-test that tests whether Model b significantly improves data fitting is given by,

$$\mathrm{F}_{\rm a,b} = \frac{\frac{\chi^2({\rm a})-\chi^2({\rm b})}{\nu({\rm a})-\nu({\rm b})}}{\frac{\chi^2({\rm b})}{\nu({\rm b})}} \tag{13}$$

where χ^2 is the residual sum of square in the least squares fitting and ν is the number of degrees of freedom. Eq. (13) compares the variance between models a and b with the variance inside model b and has F distribution with [ν(a) −ν(b), ν(b)] degree of freedom. When the F value is large enough, Model a can be rejected at a small false-rejection probability p(F).

6.2. The AIC for comparison of different models

We used the Akaike Information Criterion (AIC), a measure of the relative goodness of fit of a statistical model developed by Akaike (1974), to compare different models. Let K be the number of estimated parameters in the model and L_{Max} be the maximized value of the likelihood function for the model, AIC is defined as AIC $= 2K - 2\ln L_{Max}$. Assuming that the errors are normally distributed and independent, after ignoring the constant term, AIC is given by

$$\mathrm{AIC} = \mathrm{N}\ln\left(\frac{\chi^2}{\mathrm{N}}\right) + 2\mathrm{K} \tag{14}$$

where χ^2 is the residual sum of square in the least squares fitting and N is the number of observed data points. To give a greater penalty for additional parameters, Burnham and Anderson (2002) recommended the AIC with a correction for finite sample sizes (AICc), which is given by,

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2\mathrm{K}(\mathrm{K}+1)}{\mathrm{N}-\mathrm{K}-1} \tag{15}$$

For the set of R models, the one with the lowest AICc score is most likely to be the best model of those considered. The relative likelihood of model i is proportional to $\exp(-0.5\Delta_i)$, where Δ_i is the AICc difference between model i and the best model (with the lowest AICc). Given the data and the set of R models, the relative likelihood or Akaike weight is given by (Burnham and Anderson, 2002):

$$w_i = \frac{\exp(-0.5\Delta_i)}{\sum_{r=1}^{R}\exp(-0.5\Delta_r)} \tag{16}$$

The AIC allows one to decide which model, of those considered in the analysis, is most likely to be the 'best' one - meaning closest in information-theoretic terms to an unknown 'true' model that is not (and could not be) in the set of models considered. Putting it another way, if none of the models is any good, picking the model with the lowest AICc will not identify a good model, and certainly not a 'correct' one, just the least-worst model (thanks to an anonymous Reviewer). In summary, the chosen model is the one that minimizes the Kullback-Leibler distance between the model and the truth over the set of models considered (Burnham and Anderson, 2002).

7. Experimental results

The data in the subsequent figures were averaged across the two eyes and across the three observers. Fig. 10 illustrates the data in two different ways: (1) the Dmin threshold as a function of interocular contrast ratio at each base contrast (the higher contrast in the two eyes) (Fig. 10A), and (2) the Dmin threshold as a function of monocular contrast in one eye when the other eye's contrast is fixed (colored curves in Fig. 10B), or as a function of binocular contrast when the two eyes have identical contrast (the thick black curve in Fig. 10B). The Dmin thresholds are specified either in phase degrees (Left y-axis) or in arcsecs (Right y-axis). A high degree of scale invariance can be observed across spatial-frequency channels.

We found that, at a given base contrast (the higher contrast in the two eyes) in Fig. 10A, the best performance (the lowest threshold) occurs when two eyes have identical contrast, i.e., at contrast ratio = 1. Performance decreases when one eye's contrast decreases and the other eye's contrast is fixed (colored curves in Fig. 10B). More interestingly, performance also decreases when only one eye's contrast increases, the stereo contrast paradox (Cormack, Stevenson & Landers, 1997, Halpern & Blake, 1988, Legge & Gu, 1989, Schor & Heckmann, 1989), as indicated by the black vertical dashed line in the top panel of Fig. 10B, where the best performance occurs when the two eyes' contrasts are identical at 0.12 (the thick black circle), and decreases when only one eye's contrast increases to 0.24 (blue square), and further decreases when that eye's contrast further increases to 0.48 (cyan diamond), and to 0.96 (red cross). In Fig. 10B, all colored points (asymmetric contrast in the two eyes) were above (poorer performance) the thick black curve (identical contrast in the two eyes). The more asymmetric the two eye's contrast levels, the poorer the performance. However, when stimulus contrast decreased binocularly (the thick black circles of Fig. 10B), performance first *increased* (threshold decreased) and then decreased slightly (top two panels at spatial frequencies of 1.5 and 3.0 cpd) or remained constant (bottom panel at spatial frequency of 6.0 cpd). This can be explained by the model with both first- and second-order fusion mechanisms (the thick black curves)(see model simulation Fig. 13D), although the EN contrast normalization in the first-order pathway predicts constant performance and the DSKL contrast normalization in the second-order pathway predicts monotonic decreasing performance when stimulus contrast decreased binocularly.

Like the Dmin threshold, the Dmax thresholds are specified either in phase degrees (Left y-axis) or in arcsecs (Right y-axis) in Fig. 11. A high degree of scale invariance can be observed across spatial-frequency channels. At a given base contrast (Fig. 11A), the best Dmax performance (the highest Dmax threshold) occurs when two eyes' contrast is identical, i.e., at contrast ratio = 1. Performance decreases (Dmax threshold decreases) when one eye's contrast decreases, and the other eye's contrast is fixed (colored curves in Fig. 11B). However, unlike the Dmin threshold, the Dmax performance monotonically decreases when the contrast decreases binocularly. This cannot be explained by the model with EN normalization mechanism (see Model fitting results). In Fig. 11B, colored points (asymmetric contrast in the two eyes) are either on both sides of, or overlapped with the thick black curve (identical contrast in the two eyes), i.e., the performance for asymmetric contrast in the two eyes is not always poorer than that for symmetric contrast in the two eyes; the stereo contrast paradox was not observed in Dmax thresholds. The model with DSKL interocular interactions before the binocular site provides the best fit to the Dmax threshold data (see Model fitting results).

Please note that, when the spatial frequency was 6 cpd and the base contrast was 12% (Left bottom plot in Figs. 10 and 11), the data was reduced to only one point when the two eyes' contrasts were identical, because the task was impossible for our participants at unequal-contrast conditions. The model prediction is indicated by a short horizontal green line.

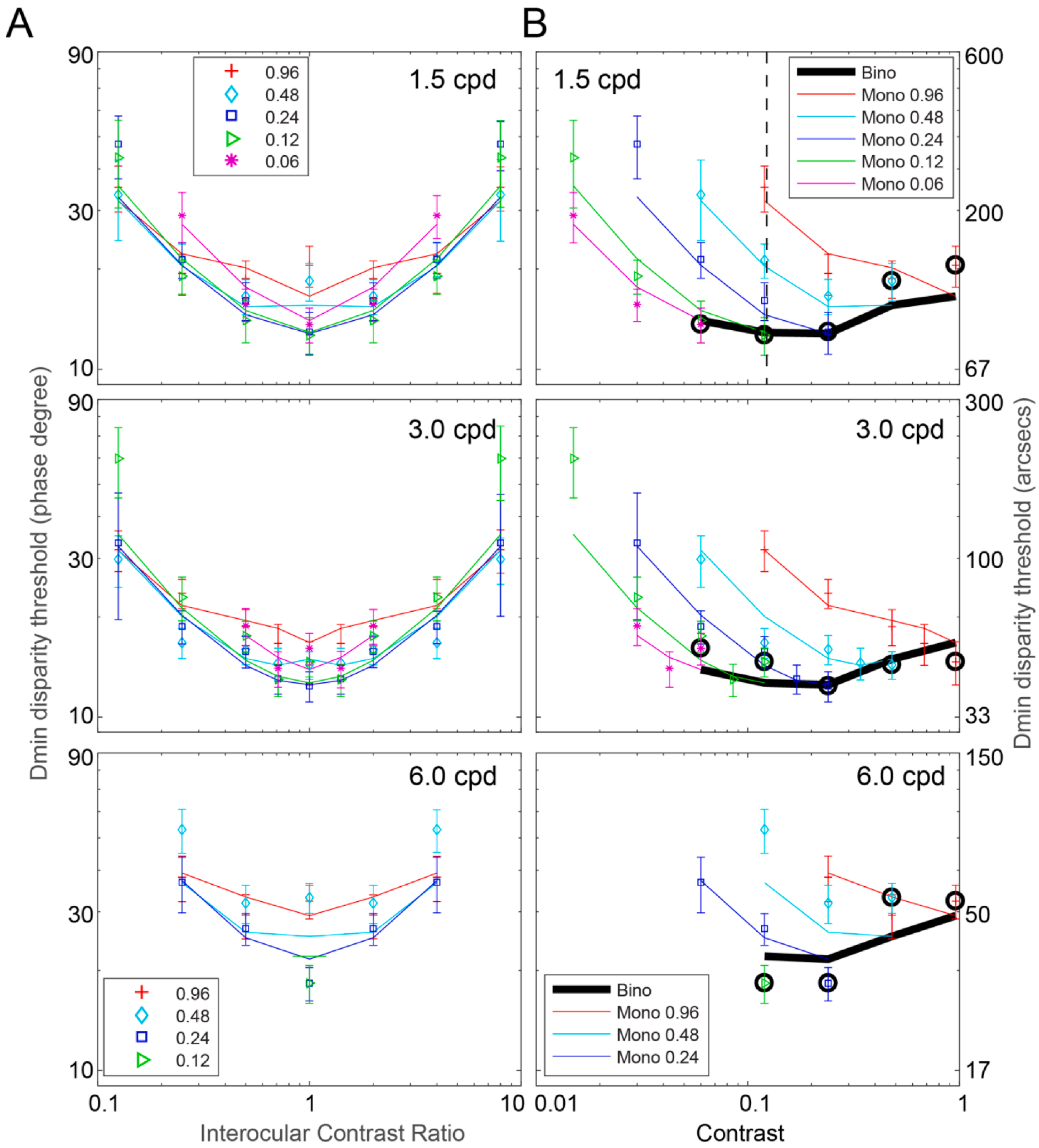

Fig. 10. Results of the minimum disparity threshold (Dmin). A. The Dmin disparity threshold as a function of interocular contrast ratio when the base contrast was 0.96 (red), 0.48 (cyan), 0.24 (blue), 0.12 (green), or 0.06 (magenta). B. The Dmin threshold as a function of one eye contrast when the other eye's contrast was fixed at 96% (red), 48% (cyan), 24% (blue), 12% (green), or 6% (magenta). The thick black circles indicate the performance when the two eyes contrast was identical. The smooth curves are the best fit of the full model with both first- (Fig. 7) and second-order sensory (Fig. 8) fusion mechanisms. The thick black curves are the model prediction when the contrast was identical in the two eyes. The left y-axis is specified in phase-degrees, and the right y-axis in arcsecs. Error bars represent ± 1 standard error. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

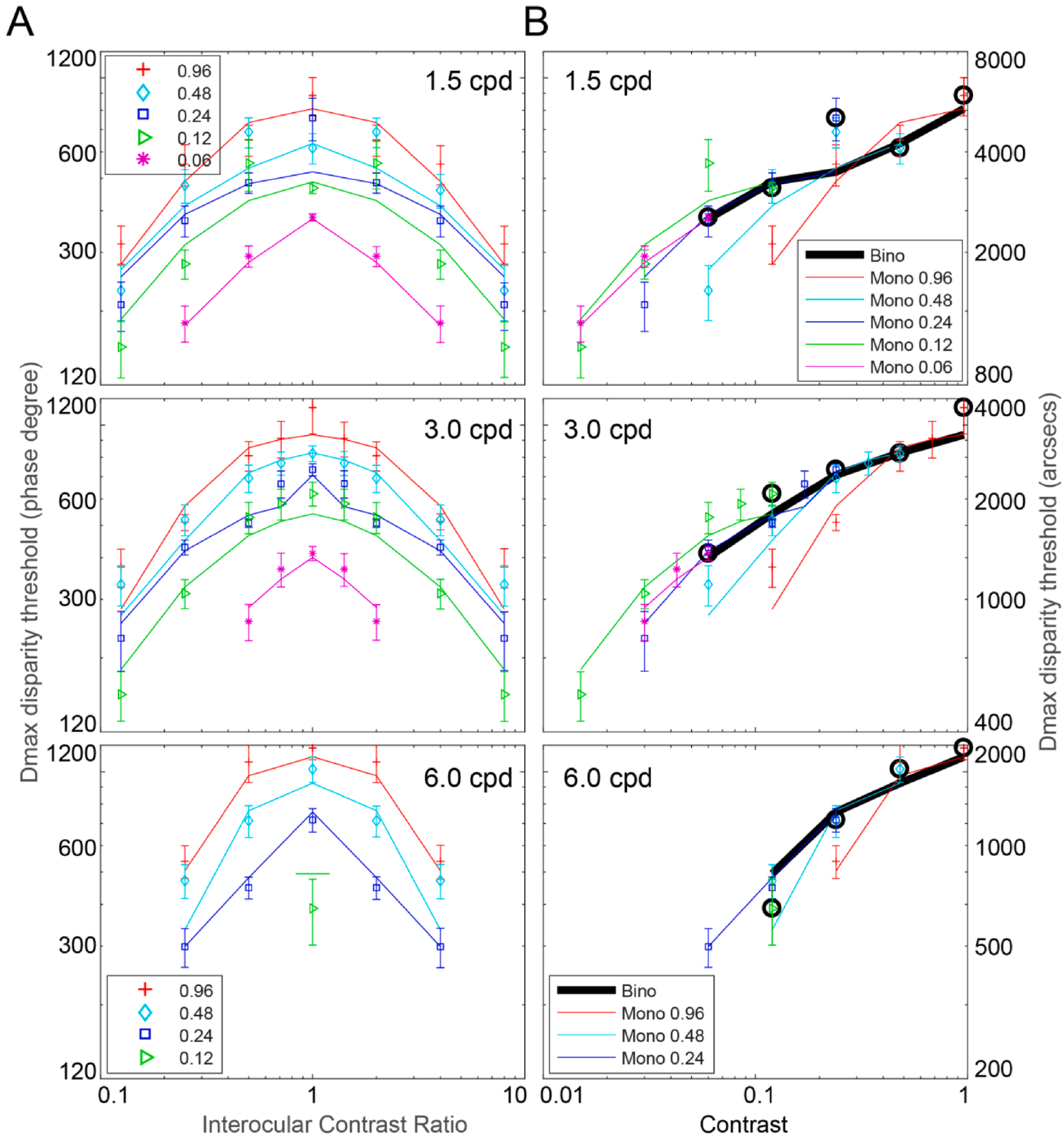

Fig. 11. Results of the maximum disparity threshold (Dmax). A. The Dmax threshold as a function of interocular contrast ratio when the base contrast was 0.96 (red), 0.48 (cyan), 0.24 (blue), 0.12 (green), or 0.06 (magenta). B. The Dmax threshold as a function of one eye contrast when the other eye's contrast was fixed at 96% (red), 48% (cyan), 24% (blue), 12% (green), or 6% (magenta). The thick black circles indicate the performance when the two eyes contrast was identical. The smooth curves are the best fit of the full model with both first- (Fig. 7) and second-order sensory (Fig. 8) fusion mechanisms. The thick black curves are the model prediction when the contrast was identical in the two eyes. The left y-axis is specified in phase-degrees, and the right y-axis in arcsecs. Error bars represent ± 1 standard error. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

8. Model fitting results

8.1. Phase- vs. Position disparity models for Dmin threshold

The phase-disparity model with an EN mechanism (PhaEN) as shown in Eq. A11 has five parameters for three spatial-frequency channels: three coefficients h_ω of sFMTF, one energy-normalization threshold Z, and one disparity-decay constant τ. The full PhaEN model (PhaEN 3) can be simplified to be three nested models PhaEN 1–3. When the stimulus disparity is much smaller than the scale of the filter profile, the disparity exponential decay can be ignored and the model can be simplified to be one with four parameters (PhaEN 2): three coefficients h_ω of sFMTF and one energy-normalization threshold Z. The simplest one (PhaEN 1) only has three coefficients h_ω with $Z = 0$.

Table 1 shows chi square values for model fitting and statistical comparisons of three nested phase-disparity models (PhaEN 1–3), in which a previous model is nested within its successor. The comparison of two neighboring models was made through an F-test with the F value given in the row of the second and subsequent models (F-test and its p-value are only shown for the average data). Adding an energy-normalization threshold Z in the PhaEN 2 model achieved a significant improvement in data fitting; the PhaEN 1 model could be rejected with a very small (<0.001) probability of false rejection. However, because Dmin thresholds are much smaller than the filters' profile scales, adding a disparity exponential decay (PhaEN 3) did not improve model fitting. AICc scores are also shown for the average data and the 'best' model is the one with the lowest AICc score. The Akaike weight (Aw), the relative likelihood of a model being the 'best' one in a set of models considered, is given in the last column. The PhaEN 2 model is the best one with 61.6% Akaike weight, and the Akaike weights of other two models are <22%.

The position-disparity model with an energy normalization mechanism (PosEN) as shown in Eq. A13 has five parameters for three spatial-frequency channels: three coefficients k_ω of sFMTF, one energy-normalization threshold Z, and one power parameter p in depth-disparity power function. The full PosEN model (PosEN 3) can be simplified to be three nested models PosEN 1–3. Assuming a linear depth-disparity function, i.e., $p = 1$, the full model is simplified to be the PosEN 2 model. The simplest one (PosEN 1) has only three coefficients, k_ω with $Z = 0$ and $p = 1$. When stimulus disparity is sufficiently small, the phase-disparity model is almost equivalent to the position-disparity model with a linear depth-disparity function ($p = 1$), i.e., PhaEN 1 ≈ PosEN 1 and PhaEN 2 ≈ PosEN 2. However, as shown in Tables 1 and 2, the PosEN models provide much better fits than the PhaEN models, respectively. With a depth-disparity power function ($p \approx 2$ for the best fit), the PosEN 3 model further significantly improves data fitting.

Based on the modeling statistics, the position disparity model provides a much better account for Dmin threshold than the phase disparity model; however in biological visual systems, phase disparity detectors play an important role in estimating stimulus disparity (DeAngelis, Ohzawa & Freeman, 1991, Ohzawa et al., 1990, Qian, 1994, Qian & Zhu, 1997, Sanger, 1988, Tsai & Victor, 2003). In the position disparity model (Fig. 4B), an ideal Max operator without any false matches is assumed, and the model has sufficient samples in position disparity space, while in a biological visual system, false matches often occur, and samples of position disparity are limited. A system with only position disparity detectors might have no paired filters with position disparity exactly matching a stimulus disparity, an interpolation has to be performed to estimate the stimulus disparity (Fleet et al., 1996). However, a system with only phase disparity detectors will make large errors or even fail to estimate large stimulus disparities (Fleet et al., 1996). A combination of position and phase disparity detectors provide a continuous measurement of stimulus disparity, as proposed in Figs. 7 and 8.

8.2. DSKL contrast normalization fails to predict Dmin threshold

To test whether interocular interactions before the site of binocular combination can predict Dmin thresholds, we developed the position-disparity model with five nested DSKL contrast normalizations (see Appendix B), PosDSKL 1–5, as shown in Eq. A16. The modeling statistics are also shown in Table 2 comparing with PosEN 1–3. Because a depth model with DSKL predicts that the depth performance increases with stimulus contrast, it fails to predict Dmin thresholds that are basically constant while varying stimulus contrast. Based on AIC analysis, the PosEN 3 model is the best one with 100% Akaike weight.

8.3. Testing DSKL contrast normalization for Dmax threshold

Dmax thresholds depend on stimulus contrast, which can be explained by DSKL contrast normalization (Ding & Levi 2016a). Table 3 shows statistical comparisons of five nested DSKL normalizations (see Appendix B) in the model with second-order phase-disparity energy for sensory fusion (Eq. (4): PhaDSKL 1–5), in which a previous model is nested within its successor. Without gain-control of gain-enhancement, i.e., $\beta = 0$, the PhaDSKL model 5 (the equivalent contrast in Eq. (4) is given by Eq. B5) is simplified to be the PhaDSKL model 4 (the equivalent contrast is given by Eq. B4), and without gain-enhancement, i.e., the gain-enhancement threshold $g_e = \infty$, the PhaDSKL model 4 is further simplified to be the PhaDSKL model 3 (the equivalent contrast is given by Eq. B3). When the double gain-controls are symmetric, i.e., $\alpha = 1$, the PhaDSKL model 3 is simplified to be the PhaDSKL model 2 (the equivalent contrast is given by Eq. B2), and when the gain-control threshold $g_c = 0$, the PhaDSKL model 2 is further simplified to be the PhaDSKL model 1 (the equivalent contrast is given by Eq. B1). The comparison of two neighboring models was made through an F-test with the F value given in the row of the second and subsequent models (F-test and its p-value are only shown for the average data). Except for the PhaDSKL 1 model, the other four models fit the Dmax threshold very well. Based on AIC analysis, the best one is the PhaDSKL 3 model without interocular enhancement. However, based on chi-square analysis, adding interocular enhancement in the PhaDSKL 5 model significantly improved model fitting.

8.4. Energy normalization fails to predict Dmax threshold

To test whether the energy normalization (EN) can predict Dmax thresholds, we developed the model for the second-order phase-disparity energy with energy normalization. As shown in Table 3, energy normalization fails to explain Dmax thresholds with 0.00% Akaike weight; it does not predict contrast dependent Dmax thresholds.

8.5. Testing the full model for both Dmin and Dmax thresholds

Based on the statistical analysis of modeling of Dmin and Dmax thresholds separately (above), we developed the full model (Eqs. 10–12) with one second- and one first-order pathway for each spatial-frequency band. In the second-order pathway with DSKL (Fig. 8), the second-order fusion energy drives second-order sensory fusion and the depth perception is based on the combination of second-order position and phase disparity energies. In the first-order pathway with EN (Fig. 7), the first-order fusion energy drives first-order sensory fusion at each location where the combination of first-order position and phase disparity energies is calculated for its relative local depth to the second-order depth plane. The total fusion energy (Eq. (10)) determines the maximum disparity threshold Dmax, and the total disparity energy (Eq. (12)), the combination of position and phase disparity energies, determines the minimum disparity threshold Dmin. However, to simplify modeling, we assumed unlimited samples in the position disparity space, i.e., there exist paired filters with position disparities exactly matching the sensory fusion shift. With almost-perfect sensory fusion at small disparities (near Dmin threshold), the model's depth estimate depends mainly on position disparity. Indeed, our modeling shows that including phase disparity energy for Dmin thresholds did not improve fitting performance significantly. Therefore, we used the total position disparity energy (Eq. (11)) in the full model to explain Dmin thresholds. We note that, in a biological visual system with limited disparity samples, phase disparity energy might play an important role in depth perception.

Tables 4a and 4b shows the best parameters of the full model for three spatial-frequency bands (1.5, 3, and 6 cpd). For each spatial frequency band, the full model has a total of 14 parameters. The second-order phase-disparity detector for second-order sensory fusion has 9 parameters: 5 DSKL parameters (g_c, α, γ, g_e/g_c, β) + 1 second-to-first-order scale factor (a) + 1 coefficient of sFMTF (h_{2nd}) + 1 disparity decay constant (τ) + 1 power parameter for sensory fusion (q). The second-order position-disparity detector has 9 parameters: 7 parameters shared with the second-order phase-disparity detector (5 DSKL parameters + 1 second-to-first-order scale factor + 1 disparity decay constant) + 1 coefficient of sFMTF (k_{2nd}) + 1 disparity power parameter (p). The first-order phase-disparity detector for first-order sensory fusion has 4 parameters: 2 parameter shared with the second-order pathway (disparity decay constant τ and power parameter for sensory fusion q) +

Table 1
Phase-disparity models with EN for Dmin.

	Np	N	KD		LJ		JW		Average					
			χ^2	χ^2/ν	χ^2	χ^2/ν	χ^2	χ^2/ν	χ^2	χ^2/ν	F-test	p(F)	AICc	Aw
PhaEN 1	3	49	97.1	1.87	187.1	3.40	183.4	3.53	200.2	4.09			78.9	21.6%
PhaEN 2	4	48	84.1	1.65	182.7	3.38	172.6	3.38	183.3	3.82	4.43	0.000	76.8	61.6%
PhaEN 3	5	47	84.1	1.68	182.7	3.45	172.6	3.45	183.3	3.90	0	1	79.4	16.8%

AICc: Akaike Information Criterion with a correction; Aw: Akaike weight

Table 2
Position-disparity models with EN or DSKL for Dmin.

	Np	N	KD		LJ		JW		Average					
			χ^2	χ^2/ν	χ^2	χ^2/ν	χ^2	χ^2/ν	χ^2	χ^2/ν	F-test	p(F)	AICc	Aw
PosEN 1	3	49	81.1	1.56	140.7	2.56	172.7	3.32	163.2	3.33			68.3	0.00%
PosEN 2	4	48	72.1	1.41	137.6	2.55	165.0	3.24	153.1	3.19	3.17	0.000	67.5	0.00%
PosEN 3	5	47	47.3	0.95	71.6	1.35	121.4	2.43	74.3	1.58	49.8	0.000	32.4	100%
PosDSKL 1	**5**	47	176.5	3.53	309.1	5.83	281.9	5.64	538.5	11.5			135.4	0.00%
PosDSKL 2	**6**	46	87.6	1.79	136.8	2.63	183.8	3.75	257.2	5.59	50.3	0.000	99.7	0.00%
PosDSKL 3	**7**	45	55.1	1.15	89.0	1.74	89.1	1.86	102.5	2.28	67.9	0.000	54.6	0.00%
PosDSKL 4	**8**	44	55.1	1.17	89.0	1.78	88.4	1.88	102.5	2.33	0	1	57.6	0.00%
PosDSKL 5	**9**	43	52.5	1.14	89.0	1.82	88.4	1.92	102.5	2.38	0	1	60.7	0.00%

AICc: Akaike Information Criterion with a correction; Aw: Akaike weight

1 energy-normalization threshold (Z) + 1 coefficient of sFMTF (h_{1st}). The first-order position-disparity detector has 4 parameters: 2 parameters shared with the second-order pathway (disparity decay constant τ and disparity power parameter p) + 1 parameter shared with the first-order phase detector (energy-normalization threshold Z) + 1 coefficient of sFMTF (k_{1st}). Adding one more spatial-frequency band, the model only needs extra coefficients for sFMTF and a second-to-first-order scale factor (a) for that band. All other parameters are shared across the spatial-frequency bands. For three spatial-frequency bands, the full model has 20 parameters.

Similar to Table 3, Table 5 shows the model fitting statistics for five-nested DSKL contrast normalizations in the second-order pathway of the full model (FulMod 1–5). The first-order pathway always uses EN2 contrast normalization. Again, we found, based on AIC analysis, the best fitting model is the full model with DSKL 3 (FulMod 3) without interocular enhancement. However, based on chi-square analysis, adding interocular enhancement in FulMod 5 significantly improved the model fitting.

In the full model, we used fusion energy to predict Dmax, and position disparity energy to predict Dmin. However, the other combinations cannot be excluded without further tests. Therefore, we tested all four combinations with DSKL 3: (1) fusion energy for both Dmin and Dmax (fusDmin-fusDmax); (2) position disparity energy for both Dmin and Dmax (posDmin-posDmax); (3) fusion energy for Dmin and position disparity for Dmax (fusDmin-posDmax); (4) fusion energy for Dmax and position disparity energy for Dmin (FulMod 3: posDmin-fusDmax). To test phase disparity energy for model depth perception, we also tested (5) total disparity energy (Eq. (12)) for both Dmin and Dmax (totDmin-totDmax); and (6) fusion energy for Dmax and total disparity energy (Eq. (12)) for Dmin (totDmin-fusDmax). Table 6 also shows modeling statistics for a conventional model (ModMax) with a Max operator (Fig. 5). The best fit is given by the combination of posDmin-fusDmax that is included in the full model (72.1% Akaike weight). Including phase disparity energy for depth perception, the model totDmin-fusDmax failed to improve fitting significantly (27.9% Akaike weight). However, including phase disparity energy, the model totDmin-totDmax significantly improves fitting if comparing with the model posDmin-posDmax. The other combinations and the conventional model of Fig. 5 have a 0.0% Akaike weight. Most likely, the Dmin threshold reflects the disparity detection limitation while the Dmax threshold is due to a binocular fusion limit.

9. Discussion

We have formulated and tested a new model that provides a unified explanation of binocular fusion and depth perception. In this model, a binocular fusion mechanism reduces intereocular misalignment by shifting the input images to depth sensors (motor fusion) and/or by selecting to read out the output of depth sensors that are shifted along the position disparity dimension (Fig. 1) toward the images (sensory fusion), while a depth perception mechanism evaluates depth perception based on position disparity and possibly residual phase disparity in a fused plane for a single image or in an unfused plane for a diplopic image.

9.1. Postion and phase disparities

In the physiological literature, there is clear evidence that both phase and position disparity neurons are present in cortical area V1 (DeAngelis et al., 1991, Ohzawa et al., 1990). Read & Cumming (2007) asked: “Why does the brain devote computational resources to encoding disparity twice over, once through position and once through phase?” Indeed, a

Table 3
Phase-disparity models with DSKL or EN for Dmax.

	Np	ν	KD		LJ		JW		Average					
			χ^2	χ^2/ν	χ^2	χ^2/ν	χ^2	χ^2/ν	χ^2	χ^2/ν	F-test	p(F)	AICc	Aw
PhaDSKL 1	**8**	44	114.8	2.44	975.2	19.5	866.9	18.4	690.8	15.7			156.8	0.00%
PhaDSKL 2	**9**	43	87.5	1.90	148.8	3.04	99.6	2.16	78.9	1.84	333.5	0.000	47.1	34.3%
PhaDSKL 3	**10**	42	79.6	1.77	147.0	3.06	99.4	2.21	73.7	1.75	2.96	0.000	46.7	41.9%
PhaDSKL 4	**11**	41	78.9	1.79	145.7	3.10	99.4	2.26	73.7	1.80	0	1	50.1	7.6%
PhaDSKL 5	**12**	40	78.3	1.82	141.0	3.06	93.6	2.18	66.8	1.67	0	0.01	48.6	16.2%
PhaEN 1	**7**	45	602.0	12.5	718.0	14.1	658.3	13.7	581.2	12.9			144.7	0.00%
PhaEN 2	**8**	44	315.1	6.70	352.5	7.05	313.2	6.66	326.3	7.42	33.6	0.000	117.8	0.00%

pure phase detector can estimate stimulus disparity within a half-cycle limit in one spatial-frequency band (Qian, 1994, Qian & Zhu, 1997, Sanger, 1988), and combined pure phase detectors at multiple scales can estimate disparity over a large disparity range (Sanger, 1988, Tsai & Victor, 2003). However, because the outputs of paired filters with phase disparity are not aligned with each other, it is not clear how to align 2D images for the system with pure phase disparity detectors. On the other hand, pure position disparity detectors can also estimate stimulus disparities over a large disparity range (e.g., Fig. 5) (Fleet et al., 1996) *if the correspondence problem is already solved*. Obviously, the brain uses two coding systems for 3D perception (Fleet et al., 1996), but it is unclear how they work together. Based on optimal information encoding, Goncalves and Welchman (2017) found that hybrid encoding of combined phase and position shifts conveys more information than either pure phase or position encoding. Although their individual model units are not specialized to identify the same feature in the two images, the aggregate readout activity classifies depth with high accuracy. However, it is not clear how their model evaluates the reduced depth perception systematically and predicts the Dmax threshold when the stimulus disparity increases beyond the point where their model loses its ability to give an accurate measurement. Although this does not imply that their model could not be modified to account for Dmax threshold, currently, there is no unified theory to account for both Dmin and Dmax thresholds for the present study. Unlike exploiting dissimilar features to provide evidence against unlikely interpretations (Goncalves & Welchman, 2017), we used a different strategy with a sensory fusion mechanism to reduce interocular misalignment (or dissimilar features), which systematically evaluates the reduced depth perception of diplopic images (see Figs. 13 and 14) and successfully predicts our Dmax threshold data. We proposed two separate mechanisms based on two coding systems, one measuring misalignment, an offset of monocular outputs, for binocular fusion and the other measuring disparity, an offset of monocular inputs, for depth perception.

In order to achieve a sharp 3D view, we proposed a model with hybrid neurons (Fleet et al., 1996) with both preferred position and phase disparities (Fig. 1). In a position disparity plane (TE, NEAR, FAR, and TI neurons with the same position disparity), if phase-disparity detectors (NEAR, FAR, or TI neurons) have positive outputs (misalignment), the system selects local depth sensors shifted along position disparity space to reduce the misalignment until it becomes less than a threshold, which transfers the phase disparity to position disparity. In a fused plane with a fused position disparity, the 2D perception is based on the outputs of paired filters with 0 phase disparity (TE) and the depth perception is based on the fused position disparity and possible residual phase disparity in the fused plane. If fusion energy is not sufficient at low stimulus contrast and/or at large stimulus disparity, fusion might not be completely accomplished and halts in an unfused plane, in which the 2D outputs of any paired filters are misaligned resulting in a local blurred

Table 4a
Full model parameters.

EN	DSKL					Scale factor			Decay	Fusion power	Disp power
Z	g_c	α	γ	g_e/g_c	β	a_1	a_2	a_3	τ/λ	q	p
0.053 ± 0.007	0.123 ± 0.007	0.575 ± 0.070	2.443 ± 0.148	3.550 ± 1.229	1.816 ± 0.742	4.729 ± 0.367	5.473 ± 0.426	6.524 ± 0.613	0.221 ± 0.046	0.745 ± 0.243	1.881 ± 0.242

λ: wavelength of Gabor patches (=600, 1200, or 2400 arcsecs).

Table 4b
Coefficients of sFMTF in the full model.

First-order sensory fusion energy			First-order position-disparity energy			Second-order pathway	
h_{1st1}	h_{1st2}	h_{1st3}	k_{1st1}	k_{1st2}	k_{1st3}	h_{2nd}/h_{1st}	k_{2nd}/k_{1st}
26.13 ± 12.29	27.20 ± 12.53	18.74 ± 8.90	0.040 ± 0.012	0.079 ± 0.025	0.090 ± 0.028	42.77 ± 14.74	1.54 ± 0.74

sFMTF: spatial-frequency modulation transfer function. Standard errors of model parameters are also given.

Table 5
Test five nested DSKLs in the second-order pathway of the full model.

	Np	ν	KD		LJ		JW		Average					
			χ^2	χ^2/ν	χ^2	χ^2/ν	χ^2	χ^2/ν	χ^2	χ^2/ν	F-test	p(F)	AICc	Aw
FulMod 1	16	88	271.9	2.89	174.8	1.75	305.9	3.25	178.2	2.02			97.1	0.0%
FulMod 2	17	87	165.9	1.78	174.8	1.77	208.5	2.24	146.5	1.68	18.8	0.000	79.7	0.0%
FulMod 3	18	86	115.7	1.26	174.8	1.78	208.5	2.27	110.7	1.29	27.8	0.000	53.5	57.3%
FulMod 4	19	85	114.0	1.25	174.8	1.80	208.5	2.29	110.6	1.30	0.077	1	56.5	12.8%
FulMod 5	20	84	93.1	1.03	174.6	1.82	191.1	2.12	105.5	1.26	2.07	0.000	54.8	29.9%

Table 6
Test four combinations of phase and position disparity energies for Dmin and Dmax.

	Np	ν	KD		LJ		JW		Average					
			χ^2	χ^2/ν	χ^2	χ^2/ν	χ^2	χ^2/ν	χ^2	χ^2/ν	F-test	p(F)	AICc	Aw
ModMax	13	91	266.3	2.75	253.8	2.46	216.4	2.23	198.1	2.18			99.8	0.0%
fusDmin-fusDmax	13	91	273.4	2.82	386.5	3.75	400.1	4.13	365.0	4.01			163.3	0.0%
fusDmin-posDmax	18	86	746.8	8.12	925.2	9.44	353.4	3.84	777.9	9.05			256.3	0.0%
posDmin-posDmax	18	86	363.4	3.95	168.5	1.72	208.8	2.27	158.4	1.84			90.8	0.0%
totDmin-totDmax	19	85	150.6	1.65	166.7	1.72	172.0	1.89	129.9	1.53	18.6	0.000	73.2	0.0%
posDmin-fusDmax (FulMod 3)	**18**	**86**	**115.7**	**1.26**	**174.8**	**1.78**	**208.5**	**2.27**	**110.7**	**1.29**			**53.5**	**72.1%**
totDmin-fusDmax	19	85	104.7	1.15	172.7	1.78	164.8	1.81	109.4	1.29	1.01	0.48	55.4	27.9%

binocular view or even diplopia, and depth perception is based on the unfused position disparity and the phase disparity in the unfused plane.

9.2. Possible roles for binocular difference channels in depth perception

Although binocular difference channels have been studied previously (Cohn & Lasley, 1976, Georgeson et al., 2016, Kingdom et al., 2018, May et al., 2012), their role in binocular vision is still unclear. Li and Atick (1994) proposed a depth perception theory that encoded both the sum and the difference of the two eyes images for stereovision, in order to increase coding efficiency. May et al (2012) provided convincing evidence for the existence of binocular difference channels in human vison to support the theory. Here, we argue that binocular difference channels might also play a role in binocular fusion, acting as 180-degree phase disparity detectors to compute interocular misalignment and provide fusion energy when phase disparity > 90 or $< -90°$. The combination of three phase disparity detectors of 90, -90 and $180°$is able to detect any misalignment and make binocular fusion in the range of $-180 \sim 180$ phase degrees at a given spatial scale.

9.3. Local cross-correlation model

The local cross-correlation model has been proposed to explain both the disparity-gradient limit and the stereo-resolution limit (Allenmark & Read, 2010, Banks, Gepshtein & Landy, 2004, Filippini & Banks, 2009). The model is closely based on the known physiology and is able to explain important aspects of human stereo depth perception. The model consists of a local Gaussian window for cross-correlation, with its central position shifted in the visual space. The stimulus disparity is estimated either by a Max operator, searching for the maximum correlation across all horizontal window positions, or by a template matching process (Allenmark & Read, 2010). Because V1 receptive fields appear to prefer uniform disparity (Nienborg, Bridge, Parker & Cumming, 2004), the model assumes constant disparity measurement within the window, which makes the window size a limit for both disparity-gradient and stereo-resolution (Allenmark & Read, 2010, Filippini & Banks, 2009). However, although the model predicts important differences in the ability to detect disparity gratings with square-wave vs. sine-wave profiles, there seems to be little or no difference between the detectability of square- and sine-wave disparity gratings for human subjects (Allenmark & Read, 2010). In particular, the model can detect square-wave gratings up to much higher disparity amplitudes than sine-wave gratings, which is not consistent with human data (Allenmark & Read, 2010).

This local cross-correlation model is similar to our model in Fig. 5, except without disparity-decay at large disparities. If one considers the RGP stereogram (Fig. 2) used in our experiment to be a square-wave disparity grating with 0 spatial frequency, installing a disparity-decay function (Fig. 5) prevents the system from detecting much higher disparity amplitudes than our experimental data, and therefore provides a reasonable explanation for Dmax thresholds. Our sensory fusion mechanism provides a satisfactory explanation of this disparity-decay at large stimulus disparities (See Fig. 13).

9.4. Binocular sensory fusion is a solution for the correspondence problem

To solve the binocular correspondence problem, traditional models typically sample multiple points in disparity space, and select the one with the maximum binocular energy to estimate stimulus disparity for depth perception (Filippini & Banks, 2009, Fleet et al., 1996). However, the stimulus disparity may not agree exactly with the preferred disparity of any one neuron in the population. Interpolation between the samples must be performed to find the peak in the binocular energy function (Fleet et al., 1996), which requires sufficient disparity samples over a large range. However, studies of awake, behaving monkeys found that disparities near zero were most densely represented, and seldom found preferred disparities$>12'$ (crossed or uncrossed) for eccentricities within 2 degrees of the fovea (Poggio & Fischer, 1977).

In this study, we proposed binocular fusion mechanisms to solve the correspondence problem. Unlike a traditional matching process (Filippini & Banks, 2009, Fleet et al., 1996, Julesz, 1971, Marr & Poggio, 1979), which uses feature matches or maximum correlation to search for corresponding inputs, our fusion mechanism uses a negative feedback loop to reduce the two eyes' misalignments. Correct matches are the outcome of perfect fusion. When the matching process fails, it is difficult to evaluate the reduced depth perception of mismatched inputs, and we assumed a depth disparity-decay function (Fig. 5) to explain the reduced depth perception. With a fusion mechanism, however, even when fusion fails, the reduced depth perception of unfused inputs (diplopic images) can be evaluated by the model (e.g., Figs. 7 and 8).

Fig. 12 shows a possible diagramatic conceptual scheme (not generated by a model) for binocular fusion and depth perception under a coarse-to-fine process (Marr & Poggio, 1979), using a depth surface with multiple fronto-parrallel depth planes (0th-order depth planes), i.e., local stimulus disparity remains constant in the surface. The X-axis represents horizontal binocular disparity and the Y-axis represents the vertical dimension. A target depth surface (indicated by bold black lines) is presented in front of a fixation point (black cross) (Fig. 12A). Around the fixation point, the disparity space is sampled at large (magenta box), middle (blue box) and small (red bars) scales (not all samples are shown). At each scale, one position (0 phase disparity) and three phase disparity energies of 90, -90 and $\pm 180°$are calculated at each position disparity. The central magenta box represents four pairs of vertical frequency filters with 0, 90, ±180, and -90 phase disparities (overlaid with each other). At a large scale (LS), one pair of filters with identical phases (idLS paired filters) computes position disparity energy and the other three, with misaligned phases (misLS paired filters), calculate phase disparity energy. Similarly, one blue box represents four pairs of filters at middle scale (MS), and one vertical red bar represents four pairs of filters at a small scale (SS).

Positive phase-disparity energy drives either motor or sensory shifts (see Fig. 1) to reduce its amount until it is eliminated. Perfect fusion is achieved when no positive phase-disparity energy is detected. In Fig. 12, the stimulus disparity is labled with MS phase degrees. The positive LS phase disparity energies (motor-fusion energy) drive vergence eye movements to align the two eyes images globally. After motor fusion (Fig. 12B), the target surface is shifted by vergence to the place without LS phase disparity, i.e, achieving global alignment of the target surface, at which the idLS paired filters have identical outputs and reach the maximum interocular correlation. The vergence motor shift u_{M} determines the global depth of the target surface relative to the fixation point. However, at middle and small scales, the target surface is still misaligned locally. We postulated a sensory fusion mechanism (see Fig. 1) to selectively read out depth sensors in the two eyes' that are locally aligned with the input images. In Fig. 12B, the MS phase disparty energies (calculated by the central misMS paired filters) shifts the readout to MS paired filters that are relatively aligned with the input depth surface. After MS sensory fusion (Fig. 12C), the readout of central MS paired filters is shifted to the positions where the idMS paired filters have identical outputs and reach the maximum interocular correlation. The MS sensory readout shifts calculate the MS local depth relative to the global depth. After MS sensory fusion at the middle scale (Fig. 12C), some locations of the depth surface are already aligned in the two eyes at a small scale (no positive SS phase-disparity energy is detected at these locations) (indicated by bold red bars), and others are still mis-aligned (indicated by thin red bars) and need further sensory fusion at a small scale. After SS sensory fusion, the two eyes images are perfectly aligned at all three spatial scales (Fig. 12D). The SS sensory readout shifts compute the SS depth relative to the MS depth surface. At each location, the apparent depth is a weighted summation of a motor shift, sensory readout shifts and possible residual phase disparity if fusion is not perfect. The weights depend on stimulus contrast and spatial frequency.

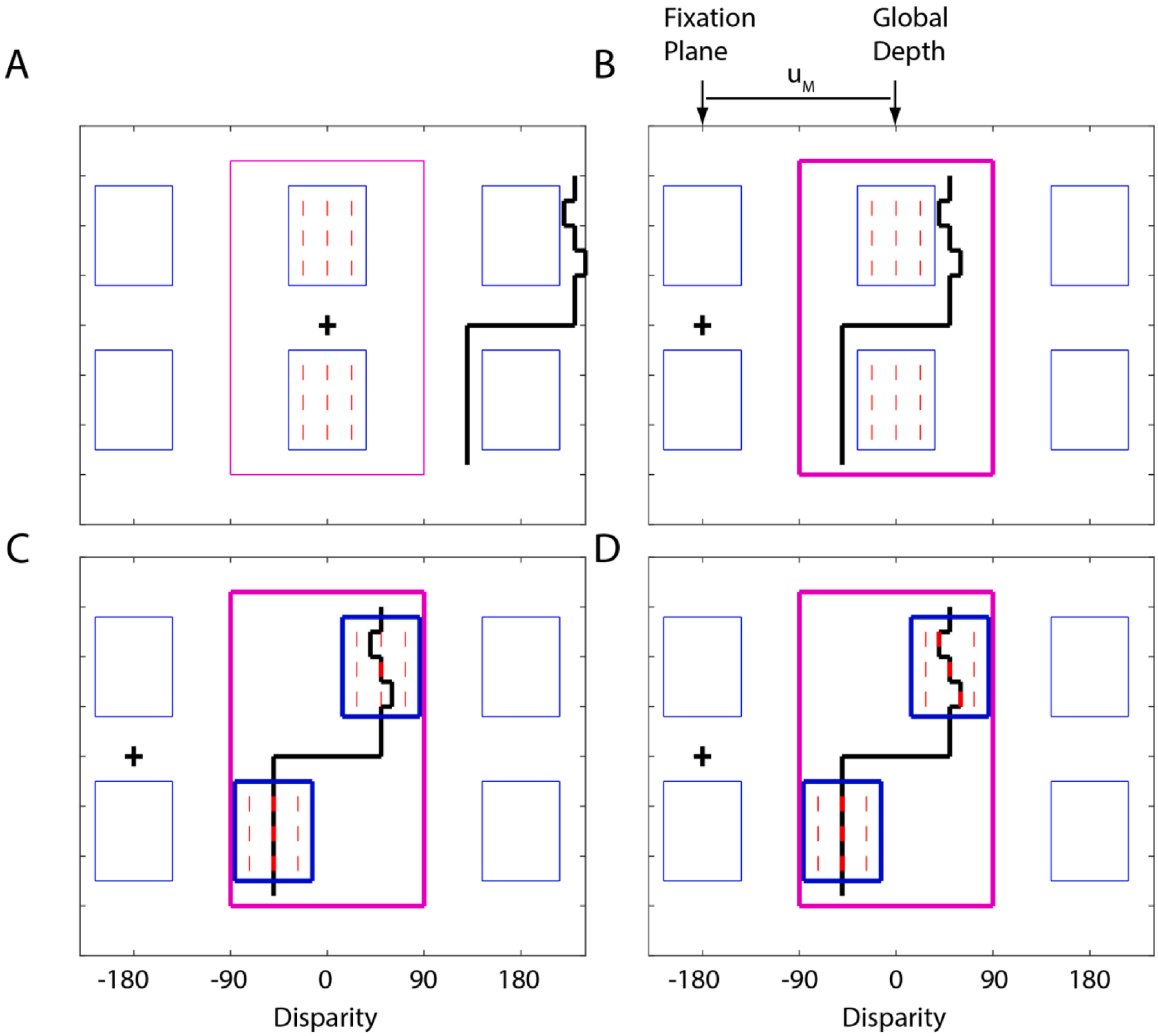

Fig. 12. A diagramatic schema for binocular fusion mechanisms with a multiple fronto-parralel depth surface (0th-order depth plane). (A) A depth surface (indicated by bold black lines) is presented in the front of a fixation point (black cross). The X-axis represents depth or horizontal binocular disparity, and the Y-axis represents the vertical dimension. Around the fixation point, the disparity is sampled at large (magenta box), middle (blue box) and small (red bars) scales (not all samples are shown). The X-axis is labled in phase degrees at the middle scale. (B) The LS (large-scale) phase disparity energy drives vergence eye movements to shift the depth surface to the position without global disparity, i.e., to align the two eyes images globally. (C) The MS (middle-scale) phase disparity energy drives MS sensory fusion to shift the readout to MS paired filters (bold blue boxes) that are relatively aligned with the depth surface at the middle scale. (D) The SS (small-scale) phase disparity energy drives SS sensory fusion for any possible SS misalignments. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

9.5. Does false matching play a role in Dmax thresholds?

If Dmax depends on the probability of false matches, the limit would be near half of the mean distance between pattern elements (e.g., Gabor patches or dots) (Eagle & Rogers, 1996, Morgan, 1992), at which the false-matching probability reaches 50%. However, most of our Dmax data (Fig. 11) are above this half mean distance limit (=420 Gabor phase degrees in a RGP stereogram as shown in Fig. 2). At high contrast, the Dmax thresholds are even greater than the mean distance (=840 Gabor phase degrees). Obviously, in the present study, the false matching of neighboring patches does not play a role in Dmax thresholds, because the fine detail spacing is removed by the second-order spatial filter with a larger scale (wavelength = 1703 ~ 2347 Gabor phase degrees, i.e., the scale factor a = 4.73 ~ 6.52, see Tables 4a and b) than the mean distance. In previous studies on the maximum spatial displacement detectable (Dmax) for random-dot kinematograms (Eagle & Rogers, 1996, Eagle & Rogers, 1997, Morgan, 1992), the Dmax remains relatively constant when the mean distance increases up to the scale of a spatial-frequency filter, and further increasing the mean distance beyond the filter's spatial scale, the Dmax increases with the mean distance – suggesting possible false matching of neighbor elements. Based on the fusion mechanism, the present study revealed that the weighted summation of the half-cycles of second- and first-order filters is the limit for depth perception, and the weights ($\leq$1) depends on both stimulus spatial frequency and contrast. It may be informative to test the model at a mean distance larger than the scale of the second-order filter to see if false matching might play a role in Dmax thresholds at such large mean distances.

9.6. Does the brain need a sensory fusion mechanism for depth perception?

One might ask why a visual system needs a sensory fusion mechanism for depth perception if it is able to measure a stimulus disparity without considering fusion. Indeed, as far as we know, computer vision models of disparity estimation are seldom concerned with the issue of fusion - disparity is measured directly and assigned to the image. However, a model in which disparity is measured and assigned to the image may not be able to predict the systematic reduction in depth perception when the disparity increases beyond the point where the model loses its ability to estimate disparity accurately, and therefore cannot predict the Dmax threshold. While a computer model might not need such a mechanism because one can always increase its measurable range of disparity by increasing its computational resource, given the - limited computational and energic resources of the brain, humans need such a mechanism to provide continuous depth perception when the disparity increases from Dmin to Dmax.

The present study proposes a depth model with sensory fusion mechanism for the brain to evaluate the reduced depth perception of diplopic images systematically, which successfully predicts our Dmax threshold data. This model provides a unified explanation of depth perception, either perfect or reduced, over the entire range of disparities from Dmin to Dmax (see Figs. 13 and 14), no matter whether the image appears single (aligned) or diplopic (misaligned).

9.7. Disparity upper limit for perfect sensory fusion

Please note that the sensory fusion mechanism proposed in the present study operates over the entire range of stimulus disparities from Dmin to Dmax, to reduce or eliminate misalignment, even when sensory fusion fails at the disparity upper limit (<Dmax) and diplopia appears. In one spatial frequency channel, when disparity $\leq$ 90 phase degree, both fusion energy (phase disparity energy) and misalignment increase with disparity; perfect fusion is able to maintain. However, when disparity $> 90^{\circ}$, further increasing disparity, the fusion energy decreases while misalignment increases; the reduced fusion energy is not sufficient to realign the increased misalignment, resulting in failure of fusion (diplopia). This is a correct prediction of the upper limit for sensory fusion that has a constant phase disparity limit of 90° (Schor et al., 1984). Although the fusion fails with diplopia when disparity $> 90^{\circ}$, depth perception and partial sensory fusion persist until disparity $= 180^{\circ}$, which determines the upper depth limit, Dmax, in a single spatial frequency channel. Although the fusion mechanism can explain both Dmax and the upper limit for fusion, they represent two different measures. Dmax is the disparity at which depth perception collapses. At Dmax the fusion energy is just beyond the noise level required to give a fusion direction, no matter whether fusion is actually achieved. The fusion upper limit is determined at a suprathreshold level, well beyond the noise level, at which the fusion energy is insufficient to maintain perfect fusion, but still provides a clear fusion direction and possibly achieves a partial fusion. Typically, the second-order pathway can further increase the Dmax as shown in the present study, but we are not clear if it can also increase the fusion upper limit. Previous work also showed that depth thresholds and the upper limit for fusion have different contrast dependences (Schor & Heckmann, 1989). Currently, we have only tested the unified model with depth thresholds. Testing the model with fusion upper limits is beyond the scope of the current study.

9.8. Depth gradients

As demonstrated in Fig. 12, our model provides a unified account of binocular fusion and depth perception of 0th depth order. Although further development is needed to explain depth gradients (first depth order) and depth curvatures (second depth order), the current model might be able to provide an approximate explanation of depth gradients or depth curvatures with a small number of steps in 0th depth order, for example, with two steps for a gradient and three steps for a curvature (Orban, Janssen & Vogels, 2006). Two steps of 0th depth order can explain the maximum limit of depth gradient, as follows: In one spatial-frequency band, because a sensory shift between two adjacent depth planes has a half-cycle limit (see Fig. 6A), the maximum depth gradient depends on the sampling rate in the 2D xy-plane. Taking two samples per cycle in the 2D xy-plane, i.e., the minimum sampling rate (Nyquist rate) that satisfies the Nyquist sampling criterion, the model predicts the maximum depth gradient is 1, consistent with previous studies using dots (Burt & Julesz, 1980) or vertical lines (Tyler, 1975) as stimuli. The model's interpretation of the maximum depth gradient is similar to the original explanation in Tyler (1975); it reflects a size-disparity correlation (Smallman & MacLeod, 1994) within a sensory fusion mechanism. Based on the explanation by McKee and Verghese (2002), here is our version with a fusion mechanism. For two adjacent dots differing greatly in disparity, a fine scale disparity mechanism can resolve the pair in the 2D xy-plane, but sensory fusion fails to reduce the misalignment for a large disparity that lies beyond its half-cycle limit. A coarse scale can bring about sensory fusion to achieve perfect alignment for the large disparities of widely separated dots but may fail to resolve the pair in the 2D xy-plane if they are too close together. Thus, there exists a suitable scale that can resolve the pair in the both 2D xy-plane and the third disparity dimension. However, McKee and Verghese (2002) used psychophysical measurements of stereo transparency to show that human stereo matching is not constrained by a gradient of 1. They used transparent surfaces composed of many pairs of dots, in which each member of a pair was assigned a disparity equal and opposite to the disparity of the other member, and they found that these opponent–disparity dot pairs produced a striking appearance of two transparent surfaces for disparity gradients ranging between 0.5 and 3. Although diplopia still occurred when gradients were>1, the depth separation of two surfaces could still be measured reliably.

The model with both second- and first-order sensory fusion mechanisms can explain their observation. As shown in Fig. 9, the sensory fusion at a large scale of the second-order pathway reduces the interocular misalignment by shifting the readout of a depth sensor's output from the fixation plane to a large-scale position disparity plane (both planes are also sampled by small-scale position and phase disparity detectors), where the misalignment might be reduced to be within the half-cycle limit of a small scale of the first-order pathway. However, at a large gradient, when the total sensory shift of second- and first-order pathways is less than the stimulus disparity, diplopia occurs but the depth can still be estimated reliably by the combination of sensory shifts and phase disparities of the two pathways.

9.9. Model simulations

In the following, we simulate the model with the best fit of our Dmin and Dmax threshold data (Tables 4a and b). Our model fitting shows that fusion energy (phase disparity energy before fusion) determines the Dmax threshold, and disparity energy (position plus phase disparity energies after fusion) determines the Dmin threshold (See Tabel 6). Fig. 13 demonstrates model simulations of fusion (Fig. 13A) and disparity (Fig. 13B) energies, and second- and first-order sensory shifts (Fig. 13C) as a function of stimulus disparity. We note that both fusion and disparity energies should be above the noise levels in order to perceive depth veridically. When the disparity energy is below the noise level, the system fails to detect the depth. When the fusion energy is below the noise level, the system fails to detect the depth direction, even though it can detect the depth but with uncertain direction. When the stimulus disparity increases below the Dmin threshold, the fusion energy increases across the fusion energy threshold (the horizontal red dashed line at d' = 1 in Fig. 13A). At Dmin threshold (blue dashed lines, at which, the disparity energy = threshold, i.e., d-prime = 1 in Fig. 13B), the fusion energy reaches ~ 3.7 d-prime units, large enough to fuse the

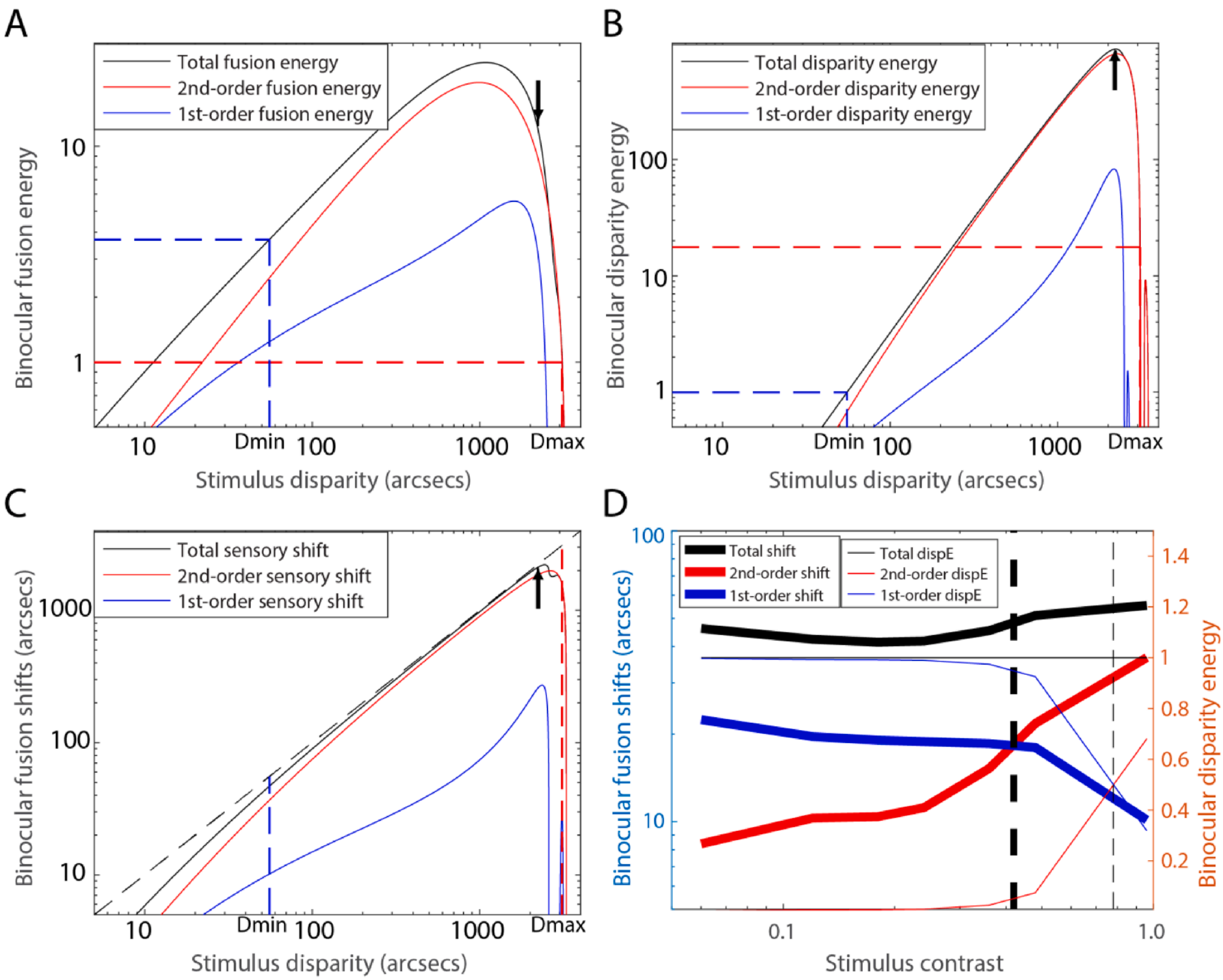

Fig. 13. Simulation of the full model using RGP stereograms with Gabor contrast = 0.96 and Gabor spatial frequency = 3 cpd. Model parameters come from Tables 4a and b. (A) Binocular fusion energy (balck, in d-prime units), summation of second- (red) and first-order (blue) sensory fusion energies, as a function of stimulus disparity. The noise level is indicated by the horizontal red dashed line at d-prime = 1, which determines the Dmax threshold. (B) Binocular disparity energy (in d-prime units) as a function of stimulus disparity. The total disparity energy (black) is the summation of second- (red) and first-order (blue) disparity energies, which determines Dmin threshold (horizontal blue dashed line at d-prime = 1). The arrow indicates the stimulus disparity where the disparity energy reaches the maximum. (C) Binocular fusion shifts as functions of stimulus disparity. (D) Binocular fusion shifts (left y-axis) and binocular disparity energy (right y-axis) as functions of binocular stimulus contrast at Dmin threshold. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

two eyes images into a single, sharp 3D image even at the Dmin threshold (see Fig. 13C: the binocular fusion shift is very close to the stimulus disparity at Dmin threshold). When further increasing stimulus disparity, the disparity energy first increases and then decreases (Fig. 13B). At Dmax threshold (red dashed lines, at which, the fusion energy = threshold , i.e., d-prime = 1 in Fig. 13A), the disparity energy is still at a high level (d-prime = 17.6, Fig. 13B). However, beyond the Dmax threshold, fusion energy cannot overcome the noise in the fusion system (i.e., the direction of sensory fusion becomes uncertain), and the brain fails to detect the direction of depth even with high disparity energy. This is consistent with our observations, at Dmax threshold, although the depth direction is not reliable, the apparent depth is much larger than the minimal threshold.

Fig. 13D shows model simulations of Dmin threshold or the total sensory shift at the threshold level (thick black curve) as a function of binocular contrast when spatial frequency is 3 cpd (the middle panel in Fig. 10B) to predict why Dmin threshold decreased (performance increased) when binocular contrast decreased from 0.96 to 0.24. Based on the model simulations, near the Dmin threshold, the disparity sensitivity of the first-order pathway is much higher than that of the second-order pathway. As indicated by the thick dashed vertical line around 0.4 contrast, the same sensory shift (=0.5 Dmin) produces>10x the disparity energy in the first-order pathway than in the second-order pathway, and as indicated by the thin dashed vertical line around 0.8 contrast, the same disparity energy (=0.5) is produced by a 2x sensory shift in the second-order pathway than in the first-order pathway. The possible reasons for less disparity sensitivity in the second-order pathway are (1) interocular gain-controls before binocular combination; (2) less depth sensitivity at a larger scale near Dmin threshold; (3) possible false matches of Gabor patches. The simulations also show that the first-order pathway dominates depth perception at low contrast; the performance increases slightly with contrast as predicted by EN contrast normalization in the first-order pathway. However, when further increasing contrast, the less sensitive second-order pathway becomes dominant, resulting in decreased total performance of the two pathways, although the performance in the second-order pathway gets a great increase (thick and thin red curves) as predicted by the DSKL normalization. Because the total disparity energy is fixed (=1) at the threshold level, the first-order disparity energy (thin blue curve) decreases when the second-order disparity energy increases with increasing contrast.

9.10. Depth perception at suprathreshold levels

Although the model was developed based on Dmin and Dmax threshold data, it might be able to predict depth perception at suprathreshold levels. If the disparity energy reflects apparent depth, the model predicts that when stimulus disparity increases, the apparent depth initially increases, reaches a maximum, and then decreases (Fig. 13B). This prediction is consistent with previous studies on apparent depth (Richards, 1971, Schor & Wood, 1983). Schor and Wood (1983) reported that, before reaching the Dmax threshold, usually the stimulus appeared diplopic, and that it always reached a depth maximum beyond which the apparent depth decreased as disparity increased to the Dmax. Richards (1971) measured the apparent depth over a large range of stimulus disparities using a depth matching task. He found that, for a normal stereo-observer (WR), matched depth first increased as disparity increased, and near the depth maximum, the stimulus appeared diplopic, and then it dropped back toward zero depth as the stimulus disparity further increased. Our simulation shows that the fusion energy reaches the maximum earlier than the disparity energy. At the maximum disparity energy (as indicated by arrows in Fig. 13), fusion energy already drops off to a lower level that might not be sufficient to maintain the two eyes images in perfect alignment. As shown in Fig. 13C, the fusion shift first follows stimulus disparity perfectly, correctly informing depth perception at small stimulus disparities. However, near the maximum apparent depth, the fusion shift drops below the stimulus disparity, resulting in diplopia. Further increasing stimulust disparity, the fusion shift decreases, correctly predicting the reduced depth perception (see **), before reaching Dmax threshold.

In our previous studies (Ding & Levi, 2016a; 2019), we proposed a depth perception model without binocular fusion mechanisms (Fig. 5). Instead, we included a disparity window, the product of a disparity power function and an exponential decay function, in the model. This previous model successfully predicted the depth thresholds, Dmin and Dmax (Ding & Levi 2016a), and suprathreshold depth perception (Ding & Levi 2019). The output of a disparity window is very similar to those of phase/position disparity energies (Fig. 13A and 13B) – binocular fusion is the mechanism behind the disparity window.

At small stimulus disparities with perfect binocular fusion, the position disparity energy is very close to a disparity square function (Eqs. (5) and (7) with $p \approx 1.88$). This correctly predicts the accelarating behavior (dipper effect) of increment disparity thresholds, which decrease when pedestal disparity increases in the range of small disparities (Farell, Li & McKee, 2004). However, using disparity modulation (corrugated stereo surfaces), Lunn & Morgan (1997) did not find this acceleration with small pedestal disparities (no dipper effect). Georgeson, Yates & Schofield (2008) showed that the presence of acceleration could be attributed to a procedural effect: trial-by-trial uncertainty about the direction of disparity. Adding a pedestal would reduce this uncertainty and therefore improve the performance. Because our present experiment randomized the disparity direction from trial to trial, uncertainty could be one source of the nonlinearity of depth-disparity power function.

9.11. Apparent depth

To test whether the model can quantitatively predict apparent depth, we re-plotted Richards' data in Fig. 14. Each datum represents the average of crossed and uncrossed apparent depth replotted from the top panel (WR) of Fig. 2 of Richards (1971). Because his stimulus was a single vertical bar with a short duration (distance = 250 cm, duration = 80 ms, stimulus bar = 6′ × 75′), we ignore motor fusion (which requires longer durations). We used a model with two sensory fusion channels, one for stimulus edge energy with higher spatial frequency and the other for stimulus bar energy with a lower spatial frequency. Because the stimulus contrast was fixed and equal in the two eyes in Richards (1971), contrast normalization was not included. The noise terms were also ignored for the matching task.

The model with two sensory-fusion channels has 10 free parameters, but we only have 10 data points. Therefore, for modeling, we reduced the number of free parameters to 6 (for details see Appendix A). The black curve in Fig. 14A shows the best fit, which gives a good fit to all the data. In Fig. 14A, the red curve indicates the depth estimated from the bar sensory shift and the blue curve indicates the depth estimated from the edge sensory shift, with the scale factor $a = 7.17$ ($\omega_{\text{bar}} = 3.17$ cpd and $\omega_{\text{edge}} = 22.7$ cpd). However, with only one sensory fusion channel, the model failed to fit the data (Fig. 14B).

Although our model was developed based on Dmin and Dmax thresholds, it also provides reasonable predictions for suprathreshold depth perception, either perfect or reduced, as discussed above.

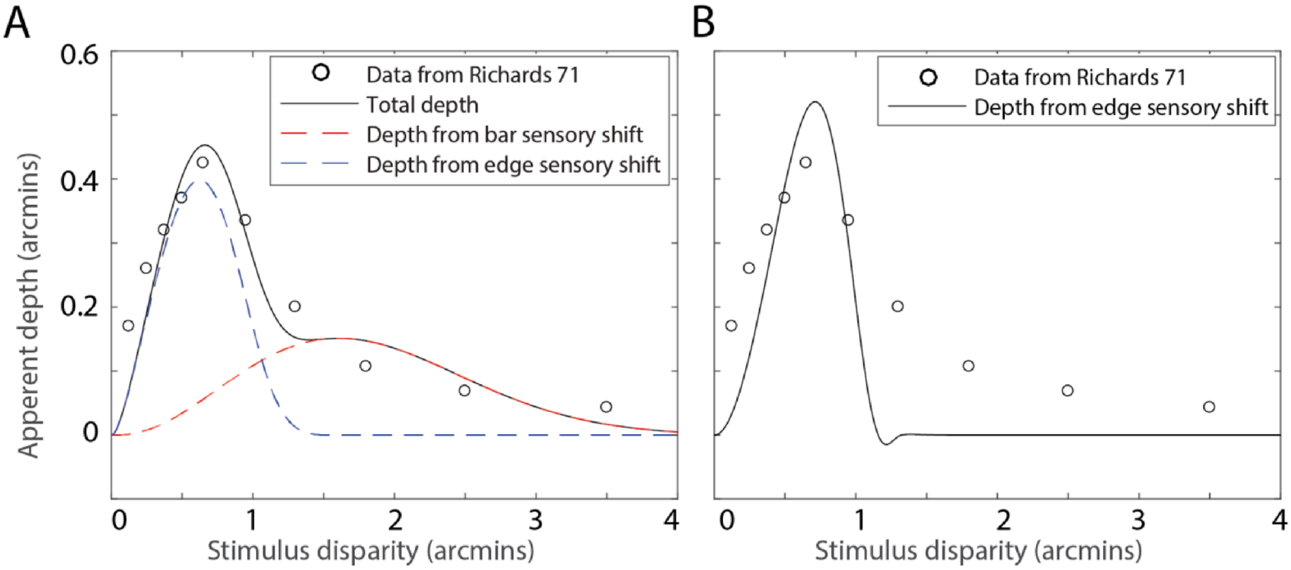

Fig. 14. Model simulations. Apparent depth as a function of stimulus disparity. The data (black circles) are replotted from the top panel of Fig. 2 of Richards (1971), and the crossed and uncrossed apparent depths were averaged. We fit his apparent depth data using a model with (A) two sensory fusion pathways, one for bar (dashed red line) and the other for edge (dashed blue line), and (B) one sensory fusion pathway for edge only. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

However, depth perception has seldom been studied at suprathreshold levels because of a lack of reliable measurement techniques. Typically, in previous studies (Richards, 1971, Schor & Howarth, 1986, Schor & Wood, 1983), a matching task was used to measure the apparent depth at suprathreshold levels: A reference depth was manipulated to match the target depth, which made the measurements highly dependent on the reference, and limited the possible stimulus conditions. In a previous study (Ding & Levi 2019), we performed a rating-scale experiment to study suprathreshold depth perception over a large range of stimulus disparities. We plan to use these suprathreshold data to test our model directly.

CRediT authorship contribution statement

Jian Ding: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing - original draft, Writing - review & editing. **Dennis M. Levi:** Writing - review & editing.

Acknowledgments

This work was supported by National Eye Institute grant R01EY030544 to JD and R01EY020976 to DML. The authors thank Drs. Martin Banks, Emily Cooper, and Clifton Schor for useful discussions. The authors also thank two anonymous Reviewers for their critical comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.visres.2020.11.009.

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A: Optics, Image Science, and Vision, 2*(2), 284–299.

Aiba, T., & Morgan, M. (1985). Vernier acuity predicted from changes in the light distribution of the retinal image. *Spatial Vision, 1*(2), 151–161.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on, 19*(6), 716–723.

Allenmark, F., & Read, J. C. (2010). Detectability of sine- versus square-wave disparity gratings: A challenge for current models of depth perception. *Journal of Vision, 10*(8), 17.

Anzai, A., Ohzawa, I., & Freeman, R.D. (1997). Neural mechanisms underlying binocular fusion and stereopsis: position vs. phase. Proceedings of the National Academy of Sciences, 94 (10), 5438-5443.

Anzai, A., Ohzawa, I., & Freeman, R. D. (1999). Neural mechanisms for encoding binocular disparity: Receptive field position versus phase. *Journal of Neurophysiology, 82*(2), 874–890.

Baker, D. H., Wallis, S. A., Georgeson, M. A., & Meese, T. S. (2012). The Effect of Interocular Phase Difference on Perceived Contrast. *PLoS ONE, 7*(4), Article e34696.

Banks, M. S., Gepshtein, S., & Landy, M. S. (2004). Why is spatial stereoresolution so low? *Journal of Neuroscience, 24*(9), 2077–2089.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436.

Burnham, K. P., & Anderson, D. (2002). *Model selection and multi-model inference: A practical information-theoretic approach* (2nd). Springer-Verlag.

Burt, P., & Julesz, B. (1980). A disparity gradient limit for binocular fusion. *Science, 208*(4444), 615–617.

Cohn, T., & Lasley, D. (1976). Binocular vision: Two possible central interactions between signals from two eyes. *Science, 192*(4239), 561–563.

Cormack, L.K., Stevenson, S.B., & Landers, D.D. (1997). Interactions of spatial frequency and unequal monocular contrasts in stereopsis. PERCEPTION-LONDON-, 26, 1121-1136.

DeAngelis, G. C., Ohzawa, I., & Freeman, R. D. (1991). Depth is encoded in the visual cortex by a specialized receptive field structure. *Nature, 352*(6331), 156.

Ding, J., Klein, S.A., & Levi, D.M. (2013a). Binocular combination in abnormal binocular vision. Journal of Vision 13 (2), 14 11-31.

Ding, J., Klein, S.A., & Levi, D.M. (2013b). Binocular combination of phase and contrast explained by a gain-control and gain-enhancement model. Journal of Vision 13 (2), 13 11-37.

Ding, J., & Levi, D. (2016a). Disparity thresholds Dmin and Dmax both depend on interocular contrast difference. Journal of Vision, 16 (12), 830-830.

Ding, J., & Levi, D. M. (2016b). Binocular contrast discrimination needs monocular multiplicative noise. *Journal of Vision, 16*(5), 1–21.

Ding, J., & Levi, D.M. (2017). Binocular combination of luminance profiles. Journal of vision, 17 (13), 4-4.

Ding, J., & Levi, D.M. (2019). A comprehensive depth perception model with filter/cross-correlation/filter (F-CC-F) structure. Journal of Vision 19 (10), 263a-263a.

Ding, J., & Levi, D. M. (2020). A phase-disparity model for vergence eye-movements. *J Vis, 20*(11), 593–593.

Ding, J., & Sperling, G. (2006). A gain-control theory of binocular combination. *Proceedings of the National Academy of Sciences of the United States of America, 103*(4), 1141–1146.

Ding, J., & Sperling, G. (2007). Binocular combination: Measurements and a model. In L. Harris, & M. Jenkin (Eds.), *Computational Vision In Neural And Machine Systems* (pp. 257–305). Cambridge, UK: Cambridge Unversity Press.

Eagle, R. A., & Rogers, B. J. (1996). Motion detection is limited by element density not spatial frequency. *Vision Research, 36*(4), 545–558.

Eagle, R. A., & Rogers, B. J. (1997). Effects of dot density, patch size and contrast on the upper spatial limit for direction discrimination in Random-dot Kinematograms. *Vision Research, 37*(15), 2091–2102.

Farell, B., Li, S., & McKee, S. P. (2004). Disparity increment thresholds for gratings. *J Vis, 4*(3), 156–168.

Filippini, H.R., & Banks, M.S. (2009). Limits of stereopsis explained by local cross-correlation. Journal of Vision, 9 (1), 8-8.

Fleet, D. J., Wagner, H., & Heeger, D. J. (1996). Neural encoding of binocular disparity: Energy models, position shifts and phase shifts. *Vision Research, 36*(12), 1839–1857.

Fogt, N., & Jones, R. (1998). Comparison of fixation disparities obtained by objective and subjective methods. *Vision Research, 38*(3), 411–421.

Georgeson, M. A., Freeman, T. C. A., & Scott-Samuel, N. E. (1996). Sub-pixel accuracy: Psychophysical validation of an algorithm for fine positioning and movement of dots on visual displays. *Vision Research, 36*(4), 605–612.

Georgeson, M. A., & Schofield, A. J. (2016). Binocular functional architecture for detection of contrast-modulated gratings. *Vision Research, 128*, 68–82.

Georgeson, M. A., Wallis, S. A., Meese, T. S., & Baker, D. H. (2016). Contrast and lustre: A model that accounts for eleven different forms of contrast discrimination in binocular vision. *Vision Research, 129*, 98–118.

Georgeson, M. A., Yates, T. A., & Schofield, A. J. (2008). Discriminating depth in corrugated stereo surfaces: Facilitation by a pedestal is explained by removal of uncertainty. *Vision Research, 48*(21), 2321–2328.

Goncalves, N. R., & Welchman, A. E. (2017). "What Not" Detectors Help the Brain See in Depth. *Current Biology, 27*(10), 1403–1412.e1408.

Halpern, D. L., & Blake, R. R. (1988). How contrast affects stereoacuity. *Perception, 17*(4), 483–495.

Henriksen, S., Tanabe, S., & Cumming, B. (2016). Disparity processing in primary visual cortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences, 371*(1697).

Hibbard, P. B., Goutcher, R., & Hunter, D. W. (2016). Encoding and estimation of first- and second-order binocular disparity in natural images. *Vision Research, 120*, 108–120.

Hou, F., Huang, C.-B., Liang, J., Zhou, Y., & Lu, Z.-L. (2013). Contrast gain-control in stereo depth and cyclopean contrast perception. *Journal of Vision, 13*(8).

Huang, C. B., Zhou, J., Zhou, Y., & Lu, Z. L. (2010). Contrast and phase combination in binocular vision. *PLoS ONE, 5*(12), Article e15075.

Hyson, M. T., Julesz, B., & Fender, D. H. (1983). Eye movements and neural remapping during fusion of misaligned random-dot stereograms. *J. Opt. Soc. Am., 73*(12), 1665–1673.

Julesz, B. (1971). *Foundations of cyclopean perception*. The University of Chicago Press.

Kertesz, A. E., & Jones, R. W. (1970). Human cyclofusional response. *Vision Research, 10*(9), 891–896.

Kingdom, F. A. A., Jennings, B. J., & Georgeson, M. A. (2018). Adaptation to interocular difference. *Journal of Vision, 18*(5), 9.

Legge, G. E., & Gu, Y. C. (1989). Stereopsis and contrast. *Vision Research, 29*(8), 989–1004.

Li, X., Lu, Z.-L., Xu, P., Jin, J., & Zhou, Y. (2003). Generating high gray-level resolution monochrome displays with conventional computer graphics cards and color monitors. *Journal of Neuroscience Methods, 130*(1), 9–18.

Livingstone, M. S., & Tsao, D. Y. (1999). Receptive fields of disparity-selective neurons in macaque striate cortex. *Nature Neuroscience, 2*(9), 825–832.

Lunn, P. D., & Morgan, M. J. (1997). Discrimination of the spatial derivatives of horizontal binocular disparity. *JOSA A, 14*(2), 360–371.

Marr, D., & Poggio, T. (1979). A computational theory of human stereo vision. Proceedings of the Royal Society of London. Series B, Biological Sciences, 204 (1156), 301-328.

May, K. A., Zhaoping, L., & Hibbard, P. B. (2012). Perceived Direction of Motion Determined by Adaptation to Static Binocular Images. *Current Biology*.

McKee, S. P., & Levi, D. M. (1987). Dichoptic hyperacuity: The precision of nonius alignment. *Journal of the Optical Society of America A, 4*(6), 1104–1108.

McKee, S. P., & Verghese, P. (2002). Stereo transparency and the disparity gradient limit. *Vision Research, 42*(16), 1963–1977.

Meese, T. S., Georgeson, M. A., & Baker, D. H. (2006). Binocular contrast vision at and above threshold. *J Vis, 6*(11), 1224–1243.

Morgan, M. (1992). Spatial filtering precedes motion detection. *Nature, 355*(6358), 344–346.

Morrone, M., & Burr, D. (1988). Feature detection in human vision: A phase-dependent energy model. Proceedings of the Royal Society of London. Series B. Biological Sciences, 235 (1280), 221-245.

Nienborg, H., Bridge, H., Parker, A. J., & Cumming, B. G. (2004). Receptive field size in V1 neurons limits acuity for perceiving disparity modulation. *The Journal of Neuroscience, 24*(9), 2065–2076.

Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1990). Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science, 249*(4972), 1037–1041.

Ohzawa, I., & Freeman, R. D. (1994). *Monocular and binocular mechanisms of contrast gain control. Computational vision based on neurobiology* (pp. 43–51). International Society for Optics and Photonics.

Orban, G. A., Janssen, P., & Vogels, R. (2006). Extracting 3D structure from disparity. *Trends in Neurosciences, 29*(8), 466–473.

Panum, P.L. (1858). Physiologische Untersuchungen über das Sehen mit zwei Augen. (Schwerssche Buchandlung Kiel.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.

Poggio, G., & Fischer, B. (1977). Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *Journal of Neurophysiology, 40*(6), 1392–1405.

Prince, S., Cumming, B., & Parker, A. (2002). Range and mechanism of encoding of horizontal disparity in macaque V1. *Journal of Neurophysiology, 87*(1), 209–221.

Qian, N. (1994). Computing stereo disparity and motion with known binocular cell properties. *Neural Computation, 6*(3), 390–404.

Qian, N., & Zhu, Y. (1997). Physiological computation of binocular disparity. *Vision Research, 37*(13), 1811–1827.

Read, J. C., & Cumming, B. G. (2007). Sensors for impossible stimuli may solve the stereo correspondence problem. *Nature Neuroscience, 10*(10), 1322–1328.

Reichardt, W. (1961). Autocorrelation, a principle for evaluation of sensory information by the central nervous system. Symposium on Principles of Sensory Communication 1959 (pp. 303-317): MIT press.

Richards, W. (1970). Stereopsis and stereoblindness. *Experimental Brain Research, 10*(4), 380–388.

Richards, W. (1971). Anomalous stereoscopic depth perception. *JOSA, 61*(3), 410–414.

Sanger, T. D. (1988). Stereo disparity computation using Gabor filters. *Biological Cybernetics, 59*(6), 405–418.

Schor, C., & Heckmann, T. (1989). Interocular differences in contrast and spatial frequency: Effects on stereopsis and fusion. *Vision Research, 29*(7), 837–847.

Schor, C., Wood, I., & Ogawa, J. (1984). Binocular sensory fusion is limited by spatial resolution. *Vision Research, 24*(7), 661–665.

Schor, C. M., & Howarth, P. A. (1986). Suprathreshold stereo-depth matches as a function of contrast and spatial frequency. *Perception, 15*(3), 249–258.

Schor, C. M., & Wood, I. (1983). Disparity range for local stereopsis as a function of luminance spatial frequency. *Vision Research, 23*(12), 1649–1654.

Smallman, H. S., & MacLeod, D. I. (1994). Size–disparity correlation in stereopsis at contrast threshold. *JOSA A, 11*(8), 2169–2183.

Tanabe, S., Haefner, R. M., & Cumming, B. G. (2011). Suppressive mechanisms in monkey V1 help to solve the stereo correspondence problem. *Journal of Neuroscience, 31*(22), 8295–8305.

Tanaka, H., & Ohzawa, I. (2006). Neural basis for stereopsis from second-order contrast cues. *Journal of Neuroscience, 26*(16), 4370–4382.

Truchard, A. M., Ohzawa, I., & Freeman, R. D. (2000). Contrast gain control in the visual cortex: Monocular versus binocular mechanisms. *The Journal of Neuroscience, 20*(8), 3017–3032.

Tsai, J. J., & Victor, J. D. (2003). Reading a population code: A multi-scale neural model for representing binocular disparity. *Vision Research, 43*(4), 445–466.

Tsao, D. Y., Conway, B. R., & Livingstone, M. S. (2003). Receptive Fields of Disparity-Tuned Simple Cells in Macaque V1. *Neuron, 38*(1), 103–114.

Tyler, C. W. (1975). Spatial organization of binocular disparity sensitivity. *Vision Research, 15*(5), 583–590.

Ukwade, M. T. (2000). Effects of nonius line and fusion lock parameters on fixation disparity. *Optometry and Vision Science, 77*(6), 309–320.

Van Santen, J. P., & Sperling, G. (1984). Temporal covariance model of human motion perception. *JOSA A, 1*(5), 451–473.

Yehezkel, O., Ding, J., Sterkin, A., Polat, U., & Levi, D. (2016). Binocular combination of stimulus orientation. *Royal Society Open Science, 3*(11), Article 160534.

Zhao, Y., Rothkopf, C.A., Triesch, J., & Shi, B.E. (2012). A unified model of the joint development of disparity selectivity and vergence control. 2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL) (pp. 1-6): IEEE.

Zhou, J., Georgeson, M.A., & Hess, R.F. (2014). Linear binocular combination of responses to contrast modulation: Contrast-weighted summation in first-and second-order vision. Journal of Vision, 14 (13), 24-24.