Visually Imbalanced Stereo Matching

Yicun Liu*[13†] Jimmy Ren*[1] Jiawei Zhang[1] Jianbo Liu[12] Mude Lin[1]
SenseTime Research[1] The Chinese University of Hong Kong[2] Columbia University[3]
{rensijie,zhangjiawei,linmude}@sensetime.com[1] liujianbo@link.cuhk.edu.hk[2] yicun.l@columbia.edu[3]

Abstract

Understanding of human vision system (HVS) has inspired many computer vision algorithms. Stereo matching, which borrows the idea from human stereopsis, has been extensively studied in the existing literature. However, scant attention has been drawn on a typical scenario where binocular inputs are qualitatively different (e.g., high-res master camera and low-res slave camera in a dual-lens module). Recent advances in human optometry reveal the capability of the human visual system to maintain coarse stereopsis under such visually imbalanced conditions. Bionically aroused, it is natural to question that: do stereo machines share the same capability? In this paper, we carry out a systematic comparison to investigate the effect of various imbalanced conditions on current popular stereo matching algorithms. We show that resembling the human visual system, those algorithms can handle limited degrees of monocular downgrading but also prone to collapses beyond a certain threshold. To avoid such collapse, we propose a solution to recover the stereopsis by a joint guided-view-restoration and stereo-reconstruction framework. We show the superiority of our framework on KITTI dataset and its extension on real-world applications.

1. Introduction

There have been remarkable signs of progress in understanding and mimicking the human vision system, and lots of works are focusing on sensing the 3D structures surrounding us. In human's visual brain, depth perception is interceded by a set of scale-variant spatial filters, where low-frequency stimuli establish coarse stereopsis, and then high-frequency visual cues escalate stereo acuity [5]. Early researchers in computer vision define this problem as searching for corresponding pixels [3], edges [1], or patches [2] among different views. Taxonomy and benchmark were later constructed in [35]. With large datasets becoming available, NN-based stereo algorithms exhibited superior performance [45].

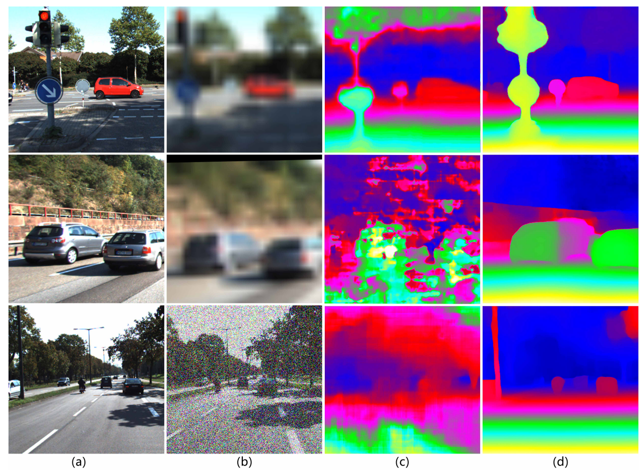

Figure 1. Illustration of visually imbalanced scenarios: (a) input left view (b) Input downgraded right view, from top to bottom: monocular blur, monocular blur with rectification error, monocular noise. (c) Disparity predicted from mainstream monocular depth (only left view as input)/stereo matching (stereo views as input) algorithms: from top to bottom are DORN [9], PSMNet [7] and CRL [31]. (d) Disparity generated from our proposed framework.

However, little attention has been drawn on the imbalanced condition between the stereo views. In many real-world cases, the visual quality of the left view and the right view are not guaranteed to be matched. It is common for human stereopsis to suffer from different degrees of anisometropia and astigmatism in binocular vision [6, 21]; or for computer vision, the master and slave camera in a dual-lens module to have different resolution, lens blur, imaging modality, noise tolerance, and rectification accuracy [43].

Not until recently, discovery in optometry reveals that it is attainable for the human to maintain decent stereo acuity with imbalanced binocular signals. In fact, the monocular downgrading jeopardizes stereo acuity for high spatial frequencies components, but low-frequency targets like structures are merely affected through a natural process called spatial frequency tuning [24]. With this unearthing in mind, we tend to ask: are stereo machines able to handle qualitatively imbalanced inputs?

*Equal contribution. Code will be available at `github.com/DandilionLau/Visually-Imbalanced-Stereo`
†Work was done during internship at SenseTime Research.

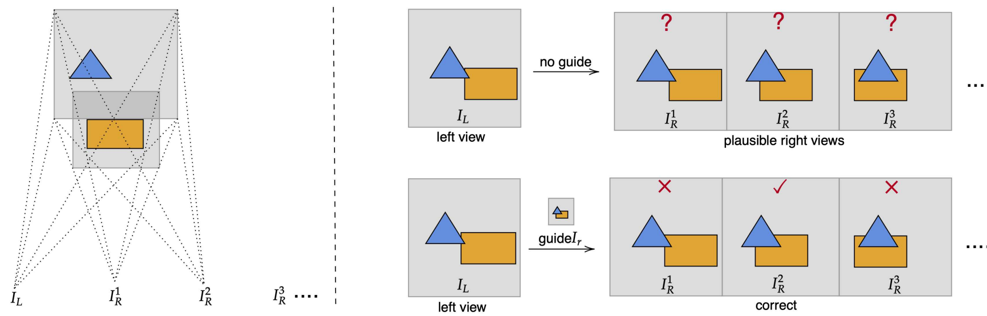

Figure 2. Intuition behind our proposed guided view synthesis framework: predicting the latent view solely based on single view is an ill-posed problem, as there exists a bunch of plausible novel views I_R^i with different disparities. However, with the geometric information in the downgraded right view I_r as a guide, the task is achievable. Even though high-frequency component of I_r is missing, rough object contour can still be inferred, as shown in our toy example. The contour provides a positional hint for the later displacement prediction.

We design a systematic comparison to answer this question. In a controlled-variant setting, we test several major monocular degradation effects on current mainstream stereo matching algorithms, including both NN-based and heuristics-based methods. By selectively increasing the corrupted levels of monocular downgrading factors, we show that existing stereo matching frameworks are resistant to mere degrees of monocular downgrading. Nevertheless, stereo matching accuracy quickly degenerates as the monocular downgrading increases. Similar to human stereopsis, all tested algorithms are observed to 'collapse' beyond specific downgrading threshold, leading to unreasonable disparity predictions.

Ideally, there exist potential cures to alleviate such collapse, but each of them has a certain limit. One intuitive approach is to conduct depth estimation only based on the high-quality monocular view. However, it cannot generalize well for unseen scenarios because it relies on prior knowledge of the object size and other physical attributes. Another approach is to conduct stereo matching on a lower resolution as a compromise to the information loss in the downgraded view, but low-res solutions cannot satisfy the demand of sharp disparity for tasks like portrait defocusing.

Taking a detour in thinking, instead of directly predicting the disparity from imbalanced views, it is easier to first restore the corrupted view using the high-quality textures in the main view and then conduct stereo matching. With the vague object contour observed in the corrupted view, human beings are pretty good at hallucinating the missing textures by 'moving' the objects from the high-quality view to the corresponding position in the corrupted view. The problem of predicting the dense disparity map between imbalanced binocular inputs can be decomposed into two sub-problems: view restoration guided by limited structural information in the corrupted view, and reconstruct stereopsis based on the restored view. For the first sub-problem, we formulate it as a guided view-synthesis process and designed a structure-aware displacement prediction network to achieve that. Our approach achieves unprecedented performance and demonstrates impressive generalization capabilities on a dataset with real-world imbalanced factors.

The main contributions of our work are threefold:

- We discover that constructing stereopsis from imbalanced views is not only feasible for human visual systems but also achievable for computer vision. It is the first work to consider the imbalanced condition for stereo matching.
- We explore the potential of current stereo machines on the task of visually imbalanced stereo matching and examine the threshold of 'stereo collapse' on different models and various imbalanced conditions.
- We exploited a guided view synthesis framework to restore the corrupted view and tackle scenarios beyond 'stereo collapse', which is even out of the capability of human stereopsis.

2. Related Work

Depth Perception There are diverse methods proposed to predict depth from single view [26], stereo views [27], and multiple views [19]. Among these variants, stereo matching is the currently popular way for low-cost depth perception. For the traditional stereo matching setup, a representative taxonomy was proposed in [35]. Many comprehensive testbeds were later introduced for quantitative evaluation of different stereo frameworks [10, 35]. Analysis of subpixel calibration error and radiometric differences emerges in [15, 14]. Significant progress in single view depth estimation emerged in [9], but estimating depth from a single view remains hard considering the ill-posed body of the problem.

Human Stereopsis Early theoretical frameworks of binocular disparity detection mechanisms were proposed in [41, 11]. Neural models of frequency-variant spatial filters in the human's visual brain were then proposed and developed in [5, 4]. To better analyze the neural network subserving stereopsis, a large portion of work has been conducted to characterize the functional attributes for these basic operators in the visual brain [32, 33, 38].

Model	All / Est	1X	2X	3X	5X	8X	10X	15X	20X	SVS
SGBM	D1-bg	14.13%	15.66%	19.53%	24.20%	62.60%	79.49%	83.98%	89.11%	25.18%
	D1-fg	21.99%	22.35%	25.36%	32.32%	58.36%	80.16%	85.48%	90.39%	20.77%
	D1-all	15.88%	16.76%	20.49%	25.88%	61.89%	79.60%	84.23%	89.32%	24.44%
DispNetC	D1-bg	3.09%	3.12%	3.31%	4.69%	11.40%	24.38%	89.51%	98.16%	25.18%
	D1-fg	3.16%	3.21%	3.28%	4.23%	11.08%	24.72%	89.75%	98.35%	20.77%
	D1-all	3.10%	3.13%	3.30%	4.62%	11.35%	24.44%	89.55%	98.18%	24.44%
CRL	D1-bg	3.02%	3.05%	3.25%	4.84%	12.24%	29.16%	94.47%	99.42%	25.18%
	D1-fg	2.89%	3.00%	3.18%	4.41%	12.36%	28.76%	95.27%	99.66%	20.77%
	D1-all	3.00%	3.04%	3.24%	4.77%	12.26%	29.09%	94.60%	99.64%	24.44%
PSMNet	D1-bg	2.36%	2.75%	5.63%	8.23%	20.86%	91.81%	99.32%	99.89%	25.18%
	D1-fg	5.72%	5.78%	8.42%	10.25%	18.85%	100.00%	100.00%	100.00%	20.77%
	D1-all	2.92%	3.25%	6.21%	10.01%	20.52%	92.93%	99.91%	99.97%	24.44%

Table 1. Performance of stereo algorithms under different levels of monocular blur: we mark the 'turning point' observed as red. **D1-bg/D1-fg/D1-all** refers to average percentage of outliers only over **background/foreground/full** regions.

Imbalanced Vision For human vision, it is quite common for patients to have unequal visual acuity and contrast sensitivity in the two eyes [6]. In computer vision, it is usual for slave cameras to equip with cheaper sensors comparing with the master camera, resulting in relatively poor performance in color, resolution, contrast, and noise-tolerance [43]. Stelmach [39] showed that the quality of the binocular perception is dominated by the sharpest image. Seuntiens [36] shows that the perceived-quality of an assymetrically-degraded image pair is about the average of both perceived qualities. Stevenson and Schor [40] demonstrate that human stereo matching does not precisely follow the epipolar lines, where human subjects can still make accurate near/far depth discrimination with 45 arcmin vertical disparity range.

Image Enhancement In low-level vision, there exist literatures on restoring blur image by super resolution or image deblur. The current state-of-the-art image enhancement methods [22, 25, 16, 37] generally tackle blur up to 5X, where our framework focuses on more severe monocular blur of 10X or more. Furthermore, although GAN-based enhancement methods [22, 23] generate visually pleasing results, consistency between stereo views is not enforced, which is not very suitable for stereo matching.

View Synthesis The most proximate work of our detailed framework design is the dynamic guided filtering network proposed in [17] and extended in [30, 29]. Instead of predicting the image after spatial transformation, it is more efficient to estimate the transformation matrix itself. There also exist algorithms for novel view synthesis from the single image, such as deep3d [44], appearance flow [48], and stereo magnification [47]. However, none of them operates in a guided manner for imbalanced view restoration.

3. Demystifying Imbalanced Stereopsis

3.1. Motivation and Assumption

In human visual system, disparity signal is generated by responsible cells, by comparing the signals from the spatial receptive field (RF) of the left and right eye [5]. These two spatial RFs can be approximately described by 2D Gabor functions. Correspondingly, there are two types of shift thet formulates the imaging relationship between the two RFs: the positional shift, and the phase shift. The positional shift can be formulated as an overall positional difference between two RFs [4]:

$$RF_L = exp(-x^2/\sigma^2)cos(\omega x) \tag{1}$$

$$RF_R = exp(-(x-d)^2/\sigma^2)cos(\omega(x-d)) \tag{2}$$

where ω is the spatial frequency, σ is the spatial constant, and d is the overall position difference. Also, the phase shift can be expressed as a phase difference between the sinusoidal modulations of the Gabor functions centered at the origin:

$$g(x,y,\phi) = \frac{1}{2\pi\sigma_x\sigma_y} exp(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}) cos(\omega(x+\phi)) \tag{3}$$

where ϕ is the phase parameter of the sinusoidal modulations, and σ_x and σ_y are constants related to the dimensions of the RFs. From the expression of Eq. 2 and Eq. 3, we can see that the positional disparity d can be arbitrarily large, while phase disparity ϕ is limited to a maximum of $\pm\pi/\omega$. In the coarse to fine disparity model proposed in [8], the response of a complex cell with both positional shift d and the phase shift $\Delta\phi$ between the two eyes can be simplified as:

$$r_q^{hybrid} \approx 4A^2 cos^2(\frac{\omega(D-d)-\Delta\phi}{2}) \tag{4}$$

where D is the reference position for RF_L. The construction of the preferred type of disparity under different frequency in Eq. 4 can be approximated as:

$$D^{hybrid} \approx \frac{\Delta\phi}{\omega} + d \tag{5}$$

With higher frequency (refers to smaller search range in phase domain), the response from phase disparity is more robust than the response from position disparity. There exists evidence showing that the search for the optimized d

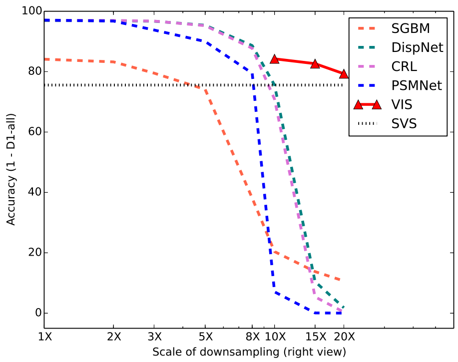

Figure 3. Performance attenuation of stereo machines: the 'turning point' mostly occurs at 5X to 8X, with error rate of 20%-30%, which is close to our baseline [26] of only using the left high-quality image as input. It turns out that rigorous conditions of monocular blur invalidate the multi-scale matching design in stereo machines, leading to unreasonable disparity predictions with error rate $> 80\%$. VIS denotes our proposed guided view synthesis method, which can still maintain decent disparity accuracy regarding severe monocular downgrading.

and $\Delta\phi$ are conducted in a coarse-to-fine iteration [8]. It starts on a large scale and sets $\Delta\phi$ as the full range to optimize positional disparity d. The process will update d iteratively while gradually decreasing the search range. When the iteration approaches a microscopic scale within the range of phase disparity, $\Delta\phi$ will be optimized based on the responses of matching high-frequency textures.

If one view is downgraded with loss of high-frequency components, it mainly jeopardizes the precision of the phase detectors. Nevertheless, position detectors still able to process coarse disparity based on the low-frequency components, and initial iterations at relatively large scale of the coarse-to-fine search are merely affected. Similarly, many stereo machines are embedded with a multi-scale matching mechanism, where the disparity is initially estimated in low resolution and then refined in higher resolution [13, 27]. Until the final refinement step, we expect low and mid-frequency components to be sufficient for coarse stereopsis.

3.2. Benchmark and Comparison

Our next step is to confirm the previous assumption. We consider one of the major factors affecting the stereo accuracy of real-world dual-lens camera modules: monocular blur. For example, budget-limited smartphones are usually equipped with a high-res master camera and low-res, relatively less expensive slave camera[1]. In our test configuration, the right view is first downsampled by a scale factor and then upsampled to its original resolution. Eight different scales are applied to subtilize the degree blur effect: 1X, 2X, 3X, 5X, 8X, 10X, 15X, and 20X. We use the KITTI dataset [28] as our testbed. Results are collected respectively from several stereo matching algorithms (both traditional methods and NN-based methods): SGBM [13], DispNetC [27], CRL [31], and PSMNet [7].

The next problem is how to define 'stereo collapse'. One intuitive method is to search for the 'turning point' of the plot. Additionally, a baseline for evaluating 'stereo collapse' is preferred to be set: we think that before reaching the 'turning point', stereo machines should at least exploit the low-frequency information in the corrupted view, which means they should outperform the result if we only input the high-quality left view. From this perspective, we choose the recent view-synthesis based monocular depth estimation network [26] as the baseline for disparity prediction.

3.3. Implication and Discussion

As shown in Table 1 and Figure 3, stereo machines show the capability to exploit the low-frequency data to establish the coarse disparity. Surprisingly, we observe mere performance decay for deep architectures under low downgrading factors. Only using 1/25 of the original pixels in the right view, stereo algorithms can generate coarse disparity predictions within 30% error. Our tests show that the capability of constructing stereopsis from low-frequency structures is a universal but varying capability of stereo machines. Within the limit of 'stereo collapse', it is possible that coarse disparity can be generated from the remaining structural difference of geometry in both views.

Nevertheless, even in human stereopsis, not all imbalanced binocular conditions lead to reliable disparity results. The core problem is that when vast losses of textural information already exceed the limits of spatial frequency tuning, disparity from imbalanced views should be an unavailable option. In all of the test set-ups, we also observe similar 'stereo collapse' on stereo machines, which is the result of increasing ambiguities and loss of high-frequency data. The turning point for 'stereo collapse' varies from 5X to 8X. Beyond such thresholds, stereo machines tend to predict unreasonable disparities.

4. Stereopsis Beyond Collapse

4.1. Problem Formulation

To reconstruct stereopsis beyond 'stereo collapse', one feasible way is first to restore the corrupted image. Although there is no evidence that the human visual system is capable of doing that, learning-based restoration model has already been widely used in computer vision. In dual-camera modules with relatively small baseline, most object regions appear in both views, and only a few of them are occluded. Admittedly, severe monocular downgrading brings considerable ambiguities, but the rough contour of objects can still be recognized. With such object contour, the next

[1] e.g. Samsung S5KHMX 108MP, S5KGW1 64MP, SONY IMX586 48MP, as master sensor of mobile dual-lens module.

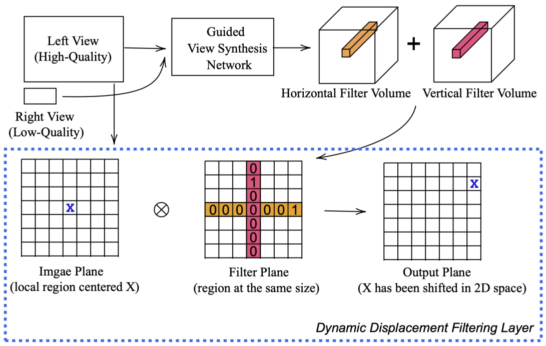

Figure 4. Illustration of the dynamic displacement filters: wrapping is conducted based on the two filter volumes generated by the view synthesis network, of size $H \times W \times L_h$ and $H \times W \times L_v$ respectively. As defined in Equation 6, the hadamard product $\otimes$ of the filter volume and the image plane are calculated to perform pixel-wise 2D displacement. Different from the only '1' in this toy example, in real setting, a pixel in the final output can be the aggregation of multiple pixels from the source image plane.

question is how to properly move and fill the textures from the uncorrupted view to the corrupted view using the contour hints. In many 3D vision tasks, such move-and-fill can be described by various types of 1D or 2D transformation. As shown in Figure 2, our idea is to profile a robust solution by estimating the spatial-variant transformation for each region of pixels in order to move the textures to retore the corrupted view properly.

To simplify the notation, we take the left view I_L as the intact view and right view I_r as the corrupted view. In general, there are two steps of our proposed framework:

1. Guided View Synthesis: Guided by the object contour and low-frequency components in the corrupted right view I_r, estimate spatial-variant transformations to warp the textures from left view I_L to a latent intact view I_R.

2. Stereo Reconstruction: Estimate disparity d based on the restored right view $\hat{I}_R$ and the original left view I_L.

4.2. Dynamic Displacement Filtering Layer

We introduce the dynamic displacement filtering layer inspired by dynamic filtering network (DFN) [17]. As shown in Figure 4, each displacement filter can be thought of as a learned kernel to convolve with a specific local region from the image plane. The job of the dynamic displacement filtering layer is to wrap the high-quality textures in the left view I_L to the appropriate position using the corrupted right view I_r as the positional guide.

The original dynamic filters are designed to be two-dimensional. If we want to estimate the vertical displacement of L_v and horizontal displacement L_h, one 2D dynamic filter requires $O(H \times W \times L_h \times L_v)$ memory. When handling high-res images with large displacements, memory consumption could be arbitrarily large. Instead, inspired by [30], our network predicts two 1D linear filters per pixel to approximate the 2D filters. This design only takes $O(H \times W \times (L_h + L_v))$ memory, which reduces the space complexity by $O(n)$.

Given the pixel $\hat{I}_R$ in the restored right view, and a pair of 1D linear displacement filters $K_h(i,j)$ and $K_v(i,j)$ predicted by the network. We can convolve I_L with the displacement filters to achieve the spatial transformation to the right image plane I_R:

$$\hat{I}_R(i,j) = (K_h(i,j) \times K_v(i,j)^T) \otimes P_L(i,j) \qquad (6)$$

where $K_h(i,j)$ are of size L_h, $K_v(i,j)$ are of size L_v. The cross product of vector $K_h(i,j)$ and vector $K_v(i,j)^T$ approximates the 2D displacement filters. P_L denotes the local image patch of size $L_h \times L_v$, which is the neighborhood region in the left image I_L centering $I_L(i,j)$. Different from general convolution where kenels are applied on the whole image, here $K_h(i,j)$ and $K_v(i,j)$ are only applied on the local patch, which means that each pixel will have the corresponding displacement filters to handle spatial transformation for its corresponding local patch.

4.3. Deep Guided Filtering Layer

If we take a slice of the horizontal displacement volume K_h of size $H \times W \times L_h$ with the axis of L_h, we will get L_h slices, corresponding to the displacements from 1 to L_h. Ideally, if one object has displacement d from the left view I_L to the right view I_R, the corresponding d-th slice $K_h[d]$ should reflect the shape of that object. However, the background of that object may be a complex region with sophisticated textures, precisely distinguish the object from the background can be challenging for RGB data. The object shape in the displacement map might be partially complete, with significant notches and surpluses.

One way of refining the edge is to use the object shape knowledge in I_L to conduct guided filtering on the filter volume. The original guided image filtering is proposed by [12]. It is a local linear model between the guide image G and the filter output O. We assume that O is a linear transform of G in a window ω_k centered at pixel k:

$$O_i = a_k G_i + b_k, \forall i \in \omega_k \qquad (7)$$

where i is the index of the pixel, k is the index of local square window ω. The linear coefficients (a_k, b_k) is assumed to be constant in the window w_k. The local linear model ensures that O has an edge only if G has an edge, because $\nabla O = a \cdot \nabla G$. In addition, the filter output O_i should be similar to the input P_i with the constrain:

$$E(a_k, b_k) = \sum_{k:i \in \omega_k} ((O_i - P_i)^2 + \epsilon a_k^2), \qquad (8)$$

where ϵ is a regularization parameter. By minimizing $E(a_k, b_k)$, we can get the filtered output O. In our task,

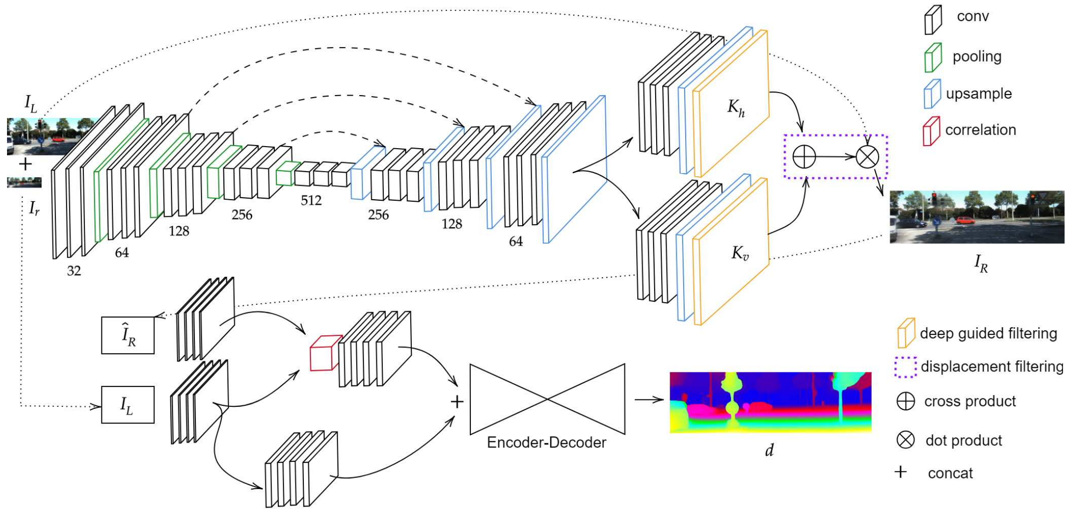

Figure 5. Network architecture for the visually imbalanced stereo matching task: the design contains two sub-networks. The upper one is our guided view synthesis network, which aims to restore the high-quality right view I_R by shifting and deforming the objects in the left view I_L. This shift and deformation operation is handled by the dynamic filters predicted by the first network. The filter volume will be processed by a deep guided filtering layer, which utilizes the shape knowledge from I_L to do edge-aware refinement. The lower one is the stereo reconstruction network, which predicts disparity based on I_L and the restored right view I_R.

we use the left view I_L as the guide G, and the d-th slice of the filter volume $K_h[d]$ as the O. If we consider all windows in the image to be filtered, the linear transformation can be written as:

$$(I_L)_i = \frac{1}{|\omega|} \sum_{k:i\in\omega_k} (a_k \cdot (K_h[d])_i + b_k) = \overline{a}_i (K_h[d])_i + \overline{b}_i \tag{9}$$

where the gradient in the displacement volume slice $\nabla(K_h[d])$ should be consistent with the gradient of the guide image ∇I_L for the optimized $\overline{a}_k$ and $\overline{b}_k$. To embed this design as a differentiable layer in our network, we propose to apply the deep guided filter [42] as the layer after displacement volume in our network, which is an accelerated and fully differentiable version of [12].

4.4. Guided View Synthesis Network

The task of the proposed guided view synthesis network is to take the corrupted right view I_r as the guide and restore the high-quality right view $\hat{I}_R$ by selectively moving the textures from the left view to the proper position in the right view. The guided view synthesis network inputs the uncorrupted left view I_L and the corrupted right view I_r and predicts displacement filter volumes K_h and K_v. The overall network architecture is shown as Figure 5, where the top part refers to our guided view synthesis framework, and the lower part refers to the stereo restoration network.

We use a bottleneck design with skip connections for the guided view synthesis network, which is similar to U-Net proposed in [34]. The intuition behind our guided view synthesis network is illustrated in Figure 2. With I_L and bilinear upsampled I_r concatenated as input, the network uses the two images to estimate the spatial differences and predicts the per-pixel displacement K_h and K_v. As exhibited in Figure 5, our guided view synthesis network has two branches for its last few layers, estimating the horizontal displacement and the vertical displacement, respectively. After the last upsample layer, we obtain the feature map of size $H \times W \times L_h$ or $H \times W \times L_v$ and use them as the horizontal filter volume and vertical filter volume respectively. Then we add a deep guided filter layer in the last to refine the shape in the filter volume. In the last, cross product is applied on the two filter volume, and hadamard product is applied to warp the left view to obatin the latent right view as network output.

4.5. Stereo Reconstruction Network

We select DispNet [27] to conduct stereo reconstruction based on the left view I_L and the restored right view $\hat{I}_R$. We further follow the modification made in [31] to adopt the DispFulNet structure for full resolution disparity output. In this network, I_L and $\hat{I}_R$ are first passed to several convolution layers with shared weights. Then the resulted feature map will be processed by a correlation layer, which embeds the geometry cues into the correlation of different horizontal patches. The feature maps outputted by the correlation layer will be concatenated with higher-level feature maps from the left image I_L. Then followed by an encoder-decoder structure, the network further refines the feature map and outputs the final disparity map d.

4.6. Loss Function

For our guided view synthesis network, although our network aims at learning the displacement volume K_h and K_v, we do not directly supervise dynamic displacement filters. One of the reasons is that while stereo pairs are easy to ob-

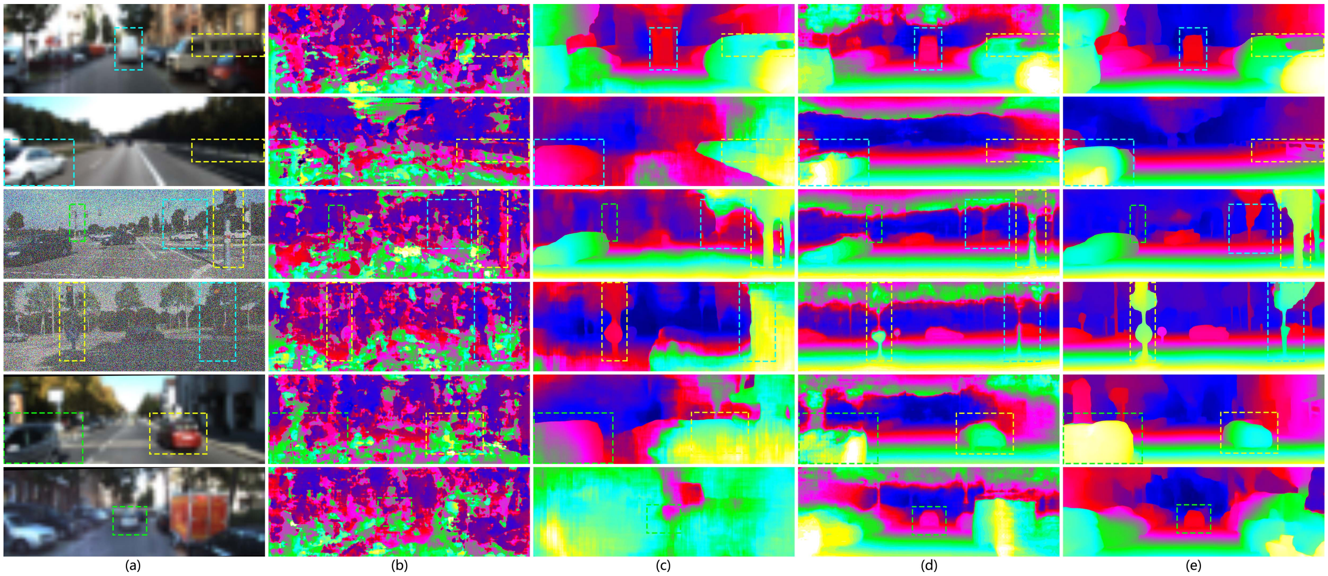

Figure 6. Qualitative evaluation on various downgrading factors: (a) downgraded right view (from top to bottom: monocular blur 10X, monocular blur 15X, gaussian noise with σ=0.5, σ=1, monocular blur 10X and maximum 0.5 degree of rectification error, monocular blur 10X and maximum 1.0 degree of rectification error). From left to right, the rest columns are disparity maps generated by (b) PSMNet [7] (c) CRL [31] (d) DORN [9] (pseudo disparity, visualization is converted from depth) (e) VIS (our framework).

tain by the dual-lens camera, displacement ground truth is hard to gather since it is a many-to-many matching problem. Instead, we consider two types of loss functions that measure the difference of the restored right view $\hat{I}_R$ and its ground truth I_R. The first loss is the photometric loss consisting of the l_1-norm and MS-SSIM:

$$\mathcal{L}_{pixel} = \alpha \cdot \|\hat{I}_R - I_R\|_1 + (1-\alpha) \cdot (\hat{I}_R, I_R)_{MS-SSIM}, \quad (10)$$

where α is the hyper parameter to balance the two term, which is set to be 0.84 in our experiment as suggested in [46]. The other loss function is the perceptual loss proposed in [18]. It is defined as the l_2-norm between feature representations of $\hat{I}_R$ and I_R:

$$\mathcal{L}_{feat} = \frac{1}{\mathcal{C}_j \mathcal{H}_j \mathcal{W}_j} \left\| \psi_j(\hat{I}_R) - \psi_j(I_R) \right\|_2, \quad (11)$$

where $\psi_j()$ denotes the feature map from the j-th VGG-19 convolutional layer and $\mathcal{C}_j$, $\mathcal{H}_j$, $\mathcal{W}_j$ are the number, height and width of the feature maps, respectively. As suggested in [29], we use the feature map generated from the $relu4\_4$ of VGG-19 to calculate the perceptual loss. The overall loss function for the guided view synthesis framework is:

$$\mathcal{L}_{syn} = \beta \cdot \mathcal{L}_{pixel} + (1-\beta) \cdot \mathcal{L}_{feat}, \quad (12)$$

where β is the hyper-parameter balancing the two losses. In our experiment, β is set to be 0.5.

For the stereo reconstruction network, In the training step, it is supervised by the ground truth disparity using l_1 loss as shown in Eq. 13. To ensure smooth training, we also adopt the multi-scale disparity loss to supervise the intermediate results.

$$\mathcal{L}_{disp} = \sum_{n=1}^{N} \|\hat{d}_n - d_n\|_1, \quad (13)$$

where $\hat{d}_n$ and d_n are the estimated disparity and ground truth disparity of n-th scale respectively.

5. Experiments

In this section, we present our experiments and corresponding results. In our experiment, we extend the boundary of the imbalanced condition to consider monocular blur, rectification error, and sensor noise in dual-lens cameras. We use KITTI raw [28] for training, which contains in total of 42382 rectified stereo pairs captured from 61 scenes. We benchmark all the models on KITTI 2015 stereo dataset.

5.1. Implementation Details

For both training and testing step in all imbalanced scenarios, only horizontal displacement kernels are enabled, and vertically displacement kernels are disabled. We use Adam [20] to optimize the network with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For our training strategy, the initial learning rate is set to be $1e^{-5}$ and multiplied by 0.9 after every five epochs. We train our framework for 100 epochs.

5.2. Experimental Details

Monocular Blur In this experiment, we think our testing images are well-rectified thus we only use the horizontal dynamic displacement filters and set its maximum size to 201. We test our framework with monocular blur scales larger

Figure 7. Qualitative evaluation of the restored view under monocular blur of 15X. ‘Syn’ denotes the synthesized right view from our network, and ‘GT’ denotes the ground-truth right view.

than those previously discovered ‘stereo collapse’ thresholds. We use the photometric criteria to evaluate the guided view synthesis network, as long as the disparity error to assess the disparity predictions from entire framework. The results are shown in Table 2 and Figure 6. Our method achieves much more resonable disparity predictions with much lower disparity error than all other stereo matching models tested in Section 3. It is worth noticing that even without end-to-end finetuning, our framework can still generate reliable results from severe monocular blur beyond the turning point of ‘stereo collapse’. As reference, other stereo matching performance is shown in Table 1.

	VIS:10X	VIS:15X	VIS:20X	CRL:10X
PSNR	19.066	18.030	17.213	N/A
SSIM	0.8249	0.7891	0.7785	N/A
D1-bg	15.24%	16.72%	18.97%	29.16%
D1-fg	18.61%	20.52%	22.78%	28.76%
D1-all	16.72%	18.32%	21.90%	29.09%

Table 2. Monocular blur as imbalanced factor. The CRL [31] column indicates the performance of direct stereo matching. PSNR and SSIM are calculated based on the restored right view image with respect to the right view groundtruth.

Rectification Error In this experiment, we introduce rectification error as the imbalanced factor, which widely exists in mobile dual-lens modules after accidental drop or a long time of use. To also address the low resolution of the slave camera, we maintain the monocular blur of 10X for the downgraded right view input. We set the horizontal filter size to 201. We simulate the rectification error by rotating the right view (slave lens) by a maximum degree in its X, Y, and Z-axis. Note that such rotation violates the formulation of the 1D search problem in all other stereo matching algorithms, thus even rotation of 0.5 degree leads to disparity error larger than 70%.

However, as shown in Table 3, our proposed framework is still able to fix this rectification error and produce a rectified high-quality right view image. Additionally, we test our framework on the real-world dual-lens dataset. Example portrait stereo pair is provided in Figure 8.

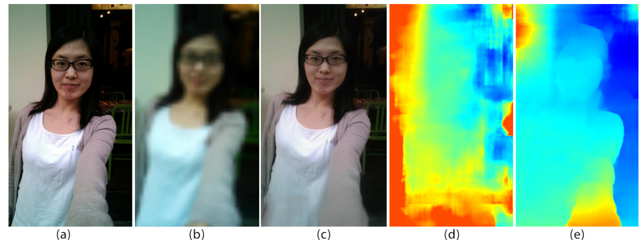

Figure 8. Qualitative evaluation on mobile dataset[2]: (a) high-res left master camera view, (b) low-res right slave camera view. Slave image has severe monocular blur, and some rectification error. CRL [31] leads to failure result in (d). Our framework first estimates the restored view (c) and reconstruct the stereopsis in (e).

	VIS:0.5D	VIS:1.0D	CRL:0.5D
PSNR	17.916	17.732	N/A
SSIM	0.7729	0.7663	N/A
D1-bg	18.63%	20.14%	99.10%
D1-fg	21.35%	24.20%	99.43%
D1-all	19.47%	21.94%	99.24%

Table 3. Rectification error as imbalanced factor. Aside from rectification error, test cases have 10X monocular blur. The CRL [31] column indicates the performance of direct stereo matching.

Sensor Noise In this experiment, we use gaussian noise to synthesize the monocular downgrading due to noise in the low-light environment. We set the horizontal filter size to 201. We control the variance σ in Gaussian distribution and use two of σ values 0.5 and 1.0.

	VIS:σ=0.5	VIS:σ=1.0	DispNet:σ=0.5
PSNR	20.217	19.142	N/A
SSIM	0.8405	0.8261	N/A
D1-bg	13.97%	15.82%	21.19%
D1-fg	16.64%	20.10%	22.36%
D1-all	15.39%	18.91%	21.78%

Table 4. Monocular noise as the imbalanced factor. The DispNet [27] column indicates performance of direct stereo matching.

6. Conclusion

Our paper defines the problem of visually imbalanced stereo matching. We start by discussing the human visual system, illustrating the mechanism behind the stereo capability under imbalanced binocular input. With such evidence at hand, we question the existence of a similar phenomenon in stereo machines, and carry out a systematic comparison to confirm whether and when ‘stereo collapse’ generally occurs on current stereo matching algorithms. Moreover, we present a practical solution to reconstruct stereopsis to support computer vision stereo systems to operate beyond such thresholds. The experiments show that our framework can effectively avoid stereo collapse, which in some sense, outperforms human stereopsis.

[2]Since images are shot horizontally in *landscape mode* and displayed vertically in *portrait mode*, objects in the right view are positionally higher.

References

[1] H Harlyn Baker and Thomas O Binford. Depth from edge and intensity based stereo. In *IJCAI*, 1981.

[2] Simon Baker, Richard Szeliski, and P Anandan. A layered approach to stereo reconstruction. In *CVPR*, 1998.

[3] Stan Birchfield and Carlo Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *TPAMI*, 1998.

[4] Randolph Blake and Hugh Wilson. Binocular vision. *Vision Research*, 2011.

[5] Randolph Blake and Hugh R Wilson. Neural models of stereoscopic vision. *Trends in Neurosciences*, 1991.

[6] Brian Brown and Maurice KH Yap. Differences in visual acuity between the eyes: Determination of normal limits in a clinical population. *Ophthalmic and Physiological Optics*, 1995.

[7] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, 2018.

[8] Yuzhi Chen and Ning Qian. A coarse-to-fine disparity energy model with both phase-shift and position-shift receptive field mechanisms. *Neural Computation*, 2004.

[9] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018.

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[11] F Gonzalez and R Perez. Neural mechanisms underlying stereoscopic vision. *Progress in Neurobiology*, 1998.

[12] K. He, J. Sun, and X. Tang. Guided image filtering. *TPAMI*, 2013.

[13] Heiko Hirschmueller. Stereo processing by semiglobal matching and mutual information. *TPAMI*, 2008.

[14] Heiko Hirschmuller and Stefan Gehrig. Stereo matching in the presence of sub-pixel calibration errors. In *CVPR*, 2009.

[15] Heiko Hirschmuller and Daniel Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE transactions on pattern analysis and machine intelligence*, 2008.

[16] Daniel S Jeon, Seung-Hwan Baek, Inchang Choi, and Min H Kim. Enhancing the spatial resolution of stereo images using a parallax prior. In *CVPR*, 2018.

[17] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *NIPS*, 2016.

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[19] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *NIPS*, 2017.

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[21] Andrew KC Lam, Apries SY Chau, WY Lam, Gloria YO Leung, and Becky SH Man. Effect of naturally occurring visual acuity differences between two eyes in stereoacuity. *Ophthalmic and Physiological Optics*, 1996.

[22] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

[23] Lerenhan Li, Jinshan Pan, Wei-Sheng Lai, Changxin Gao, Nong Sang, and Ming-Hsuan Yang. Learning a discriminative prior for blind image deblurring. In *CVPR*, 2018.

[24] Roger W Li, Kayee So, Thomas H Wu, Ashley P Craven, Truyet T Tran, Kevin M Gustafson, and Dennis M Levi. Monocular blur alters the tuning characteristics of stereopsis for spatial frequency and size. *Royal Society Open Science*, 2016.

[25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR Workshops*, 2017.

[26] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *CVPR*, 2018.

[27] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016.

[28] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015.

[29] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *CVPR*, 2017.

[30] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *ICCV*, 2017.

[31] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *ICCVW*, 2017.

[32] Saumil S Patel, Michael T Ukwade, Scott B Stevenson, Harold E Bedell, Vanitha Sampath, and Haluk Ogmen. Stereoscopic depth perception from oblique phase disparities. *Vision Research*, 2003.

[33] Simon JD Prince and Richard A Eagle. Stereo correspondence in one-dimensional gabor stimuli. *Vision Research*, 2000.

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015.

[35] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002.

[36] Pieter Seuntiens, Lydia Meesters, and Wijnand Ijsselsteijn. Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric jpeg coding and camera separation. *ACM Transactions on Applied Perception (TAP)*, 2006.

[37] Vivek Sharma, Ali Diba, Davy Neven, Michael S Brown, Luc Van Gool, and Rainer Stiefelhagen. Classification-driven dynamic image enhancement. In *CVPR*, 2018.

[38] Harvey S Smallman and Donald IA MacLeod. Size-disparity correlation in stereopsis at contrast threshold. *JOSA A*, 1994.

[39] Lew B Stelmach, Wa James Tam, Daniel V Meegan, André Vincent, and Philip Corriveau. Human perception of mismatched stereoscopic 3d inputs. In *Proceedings 2000*

International Conference on Image Processing (Cat. No. 00CH37101), 2000.

[40] Scott B. Stevenson and Clifton M. Schor. Human stereo matching is not restricted to epipolar lines. *Vision Research*, 37:2717–2723, 1997.

[41] Gerald Westheimer. The ferrier lecture, 1992. seeing depth with two eyes: Stereopsis. *Proc. R. Soc. Lond. B*, 1994.

[42] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Fast end-to-end trainable guided filter. In *CVPR*, 2018.

[43] Ruichao Xiao, Wenxiu Sun, Jiahao Pang, Qiong Yan, and Jimmy Ren. Dsr: Direct self-rectification for uncalibrated dual-lens cameras. In *3DV*, 2018.

[44] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *ECCV*, 2016.

[45] Jure Zbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 2016.

[46] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *TCI*, 3:47–57, 2017.

[47] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018.

[48] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A. Efros. View synthesis by appearance flow. In *ECCV*, 2016.