# Resolving Scale Ambiguity Via XSlit Aspect Ratio Analysis

Wei Yang[1] Haiting Lin[1] Sing Bing Kang[2] Jingyi Yu[1,3]
[1]University of Delaware [2]Microsoft Research [3]ShanghaiTech University
{wyangcs,haiting}@udel.edu sbkang@microsoft.com yu@eecis.udel.edu

## Abstract

*In perspective cameras, images of a frontal-parallel 3D object preserve its aspect ratio invariant to its depth. Such an invariance is useful in photography but is unique to perspective projection. In this paper, we show that alternative non-perspective cameras such as the crossed-slit or XSlit cameras exhibit a different depth-dependent aspect ratio (DDAR) property that can be used to 3D recovery. We first conduct a comprehensive analysis to characterize DDAR, infer object depth from its AR, and model recoverable depth range, sensitivity, and error. We show that repeated shape patterns in real Manhattan World scenes can be used for 3D reconstruction using a single XSlit image. We also extend our analysis to model slopes of lines. Specifically, parallel 3D lines exhibit depth-dependent slopes (DDS) on their images which can also be used to infer their depths. We validate our analyses using real XSlit cameras, XSlit panoramas, and catadioptric mirrors. Experiments show that DDAR and DDS provide important depth cues and enable effective single-image scene reconstruction.*

## 1. Introduction

A single perspective image exhibits scale ambiguity: 3D objects of difference sizes can have images of an identical size under perspective projection, as shown in Fig. 1. In photography and architecture, the forced perspective technique employs this optical illusion to make an object appear farther away, closer, larger or smaller than its actual size while preserving the aspect ratio. Fig. 2 shows an example in the film "The Lord of the Rings" where characters apparently standing next to each other would be displaced by several feet in depth from the camera. For computer vision, however, such an invariance provides little help, if not harm, to scene reconstruction.

Prior approaches on resolving the scale ambiguity range from imposing shape priors [3, 10], extracting local descriptors [16] to analyzing the vanishing points [13]. In this paper, we approach the problem from a different angle: we analyze aspect ratio changes of an object with respect to its depth. Consider a frontal-parallel rectangle $\mathbf{R}$ of size $s_h \times s_v$ located $d$ away from the sensor and $d > f$ where $f$ is the camera's focal length. Under perspective projection, its image is an rectangle $\mathbf{R}'$ similar to $\mathbf{R}$ of size $[s'_h, s'_v] = \frac{f}{d-f}[s_h, s_v]$. This implies that the aspect ratio $r = s_v/s_h$ of $\mathbf{R}$ and $\mathbf{R}'$ remain the same. The property can termed as aspect-ratio invariance (ARI). ARI is an important property of perspective projection. ARI, however, no longer holds under non-centric projections, exhibiting depth-dependent aspect-ratio (DDAR).

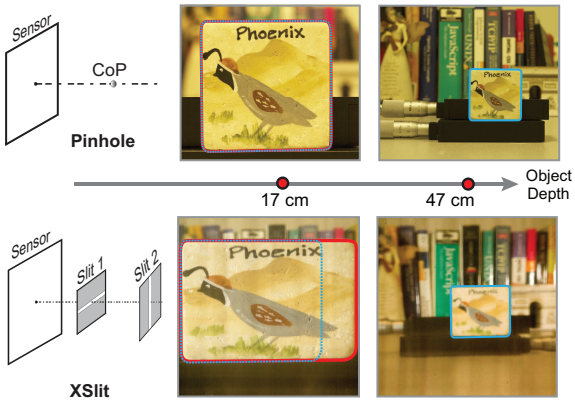



Figure 1: Images of the same object lying at different depths have an identical aspect ratio (AR) in a perspective camera (Top) but have very different ARs in an XSlit image (Bottom).

In this paper, we explore DDAR in a special type of non-centric cameras called the crossed-slit or XSlit camera [29]. Earlier work in XSlit imaging includes the pushbroom camera used in satellite imaging and XSlit panoramas by stitching a sequence of perspective images. The General Linear Camera theory [28] has shown that the XSlit camera is generic enough to describe a broad range of non-centric cameras. In fact, pushbroom, orthographic and perspective cameras can all be viewed as special XSlit entities. Geometrically, an XSlit camera collects rays that simultaneously pass through two oblique (neither parallel nor coplanar) slits in 3D space, in contrast to a pinhole camera whose rays

Figure 2: The perspective trick used in the movie "The Lord of the Rings".

pass through a common 3D point. Ye et al.[27] has further proposed a practical realization by relaying a pair of cylindrical lenses coupled with slit-shaped apertures.

We show that the XSlit camera exhibits DDAR that can help resolve scale ambiguity. Consider two 3D rectangles of an identical size lying at different depth with their images being $\mathbf{R}_1$ and $\mathbf{R}_2$ respectively. Different from the pinhole case, the AR of $\mathbf{R}_1$ and $\mathbf{R}_2$ will be different, as shown in Fig. 1. We first develop a comprehensive analysis to characterize DDAR in the XSlit camera. This derivation leads to a simple but effective graph-cut based scheme to recover object depths from a single XSlit image and an effective formulation to model recoverable depth range, sensitivity, and errors. In particular, we show how to exploit repeated shape patterns exhibiting in real Manhattan World scenes to conduct 3D reconstruction.

Our DDAR analysis can further be extended to model the slopes of lines. Specifically, for parallel 3D lines of a common direction, we show that as far as the direction is different from both slits, their projections will exhibit depth-dependent slopes or DDS, i.e., the projected 2D lines will have different slopes depending on their depths. DDS and DDAR can be combined to further improve 3D reconstruction accuracy. We validate our theories and algorithms on both synthetic and real data. For real scenes, we experiment on different types of XSlit images including the ones captured by the XSlit lens [27] and synthesized as stitched panoramas [21]. In addition, our scheme can be applied to catadioptric mirrors by modeling reflections off the mirrors as XSlit images. Experiments show that DDAR and DDS provide important depth cues and enable effective single-image scene reconstruction.

## 2. Related Work

Our work is most related to Manhattan World reconstruction and non-centric imaging.

A major task of computer vision is to infer 3D geometry of scenes using as fewer images as possible. Tremendous efforts have focused on recovering a special class of scene called the Manhattan World (MW) [4]. MW is composed of repeated planar surfaces and parallel lines aligned with three mutually orthogonal principal axes and fits well to many man-made (interior/exterior) environments. Under the MW assumption, one can simultaneously conduct 3D scene reconstruction [6, 10] and camera calibration [22].

MW generally exhibits repeated line patterns but lacks textures and therefore traditional stereo matching is less suitable for reconstruction. Instead, prior-based modeling is more widely adopted. For example, Furukawa *et al*. [10] assign a plane to each pixel and then apply graph-cut on discretized plane parameters. Other monocular cues such as the vanishing points [5] and the reference planes (*e.g*. the ground) have also been used to better approximate scene geometry. Hoime *et al*. [12, 11] use image attributes (color, edge orientation, *etc*.) to label image regions with different geometric classes (sky, ground, and vertical) and then "pop-up" the vertical regions to generate visually pleasing 3D reconstructions. Similar approaches have been used to handle indoor scenes [6]. Machine learning techniques have also been used to infer depths from image features and the location and orientation of planar regions [19, 20]. Lee *et al*. [14] and Flint *et al*. [9] search for the most feasible combination of line segments for indoor MW understanding.

Our paper explores a different and previously overlooked properties of MW: the scene contains multiple objects with an identical aspect ratio or size (e.g., windows) but lie at different depths. In a perspective view, these patterns will map to 2D images of an identical aspect ratio. In contrast, we show that the aspect ratio changes with respect to depth if one adopts a non-centric or multi-perspective camera. Such imaging models widely exist in nature, e.g., a compound insect eye, reflections and refractions of curved specular surfaces, images seen through volumetric gas such as a mirage, etc. Rays in these cameras generally do not pass through a common CoP and hence do not follow pinhole geometry. Consequently, they lose some nice properties of the perspective camera (e.g., lines no longer project to lines); at the same time they also gain some unique properties such as the coplanar common points [25], special shaped curves [26], etc. In this paper, we focus on the depth-dependent aspect ratio (DDAR) property for inferring 3D geometry.

The special non-centric camera we employ here is the crossed-slit or XSlit camera. An XSlit camera collects rays simultaneously passing through two oblique lines (slits) in 3D space. The projection geometry of an XSlit has been examined in various forms in previous studies, *e.g*., as projection model in [29], as general linear constraints in [28], and as ray regulus in [18]. For long the XSlit camera has been restricted to a theoretical model as it is physically difficult to acquire ray geometry following the slit structure. The only exception is the XSlit panoramas [23, 17] where an XSlit panorama can be stitched from a translational sequence of images or more precisely a 3D light field [15]. Recently, Ye et al.[27] presented a practical XSlit camer-

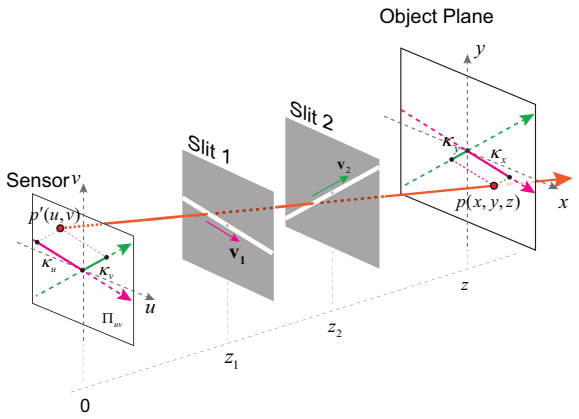



Figure 3: XSlit camera geometry: rays collected by the camera should simultaneously pass through two slits at different depths.

a. Their approach relays two cylindrical lenses with perpendicular axes, each coupled with a slit shaped aperture to achieve in-focus imaging.

## 3. Depth Dependent Aspect Ratio

We first analyze how aspect ratio of an object changes with respect to its depth in an XSlit camera. We call this property *Depth-Dependent Aspect Ratio* or DDAR.

### 3.1. XSlit Camera Geometry

A XSlit camera collects rays that pass through two oblique slits (neither coplanar nor parallel) simultaneously. For simplicity, we align the sensor plane to be parallel to both slits and corresponds to the $x$-$y$ plane. Such a setup is consistent with the real XSlit design [27] and the XSlit panoramas [29]. Further, we assume the origin of the coordinate system corresponds to the intersection of the two slits' orthogonal projections on the sensor plane, as shown in Fig. 3. The two slits lie at depth $z_1$ and $z_2$ and have angle $\theta_1$ and $\theta_2$ w.r.t the $x$-axis, where $z_2 > z_1$ and $\theta_1 \neq \theta_2$. Under this setup, the $z$ components along the two slits are 0. And the $x$-$y$ directions are $\mathbf{v_1}[\cos\theta_1, \sin\theta_1]$ and $\mathbf{v_2}[\cos\theta_2, \sin\theta_2]$ that spans $\mathbb{R}^2$ space.

Previous approaches study projection using XSlit projection matrix [29], light field parametrization[28], and linear oblique[17]. Since our analysis focuses on aspect ratio, we introduce a simpler projection model analogous to pinhole projection. Consider a 3D point $p$ to $p'$. The process can be described as follows: first decompose the $x$-$y$ components of $p$ into two basis vectors, $\mathbf{v_1}$, $\mathbf{v_2}$ and write it as $[\kappa_x, \kappa_y, z]$. Next project individual component to $[\kappa_u, \kappa_v]$. Each component can be viewed as pinhole projection as they are parallel to either slits. Finally obtain the mapping from $p$ to $p'$.

We first represent $p$ on the basis of $\mathbf{v_1}$ and $\mathbf{v_2}$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \kappa_x \mathbf{v_1} + \kappa_y \mathbf{v_2}$$

We then project $\kappa_x \mathbf{v_1}$ and $\kappa_y \mathbf{v_2}$ independently. Notice the two components are at depth $z$. And $\kappa_x \mathbf{v_1}$ is parallel to slit 1 and $\kappa_y \mathbf{v_2}$ is parallel to slit 2. Their projections imitate the pinhole projection except that the focal lengths are different:

$$\kappa_u = -\frac{z_2}{z - z_2}\kappa_x, \kappa_v = -\frac{z_1}{z - z_1}\kappa_y \tag{1}$$

Notice the XSlit mapping is linear, we can combine $\kappa_u$ and $\kappa_u$ to compute $p'$.

$$p' = \kappa_u \mathbf{v_1} + \kappa_v \mathbf{v_2}$$

$\kappa_u$ and $\kappa_v$ are also the linear representations of $p'$ on basis of $\mathbf{v_1}$ and $\mathbf{v_2}$.

### 3.2. Aspect Ratio Analysis

Equation 1 reveals that $\kappa_x$ and $\kappa_y$ are projected to $\kappa_u$ and $\kappa_v$ with different scale on the two directions parallel to the slits. In other words, with the change of depth, the ratio will be change accordingly. Specifically, we can compute the ratio as:

$$\frac{\kappa_u}{\kappa_v} = \frac{z_2(z - z_1)}{z_1(z - z_2)}\frac{\kappa_x}{\kappa_y} \tag{2}$$

This is fundamentally different from the pinhole/perspective case where the ratio remains static across depth. To understand why it is the case, recall that the pinhole camera can be viewed as a special XSlit camera where the two slits intersect, i.e., they are at the same depth $z_1 = z_2$. In that case, Eqn. degenerates to $\kappa_x/\kappa_y = \kappa_u/\kappa_v$, i.e., the aspect ratio is invariant to depth.

For the rest of the paper, we use $r_o = \frac{\kappa_x}{\kappa_y}$ to represent the base aspect ratio and $r_i = \frac{\kappa_u}{\kappa_v}$ represents the aspect ratio after XSlit projection. From Eqn. 2, we can derive the depth from the aspect ratio as:

$$z = \frac{z_1 z_2 (r_i - r_o)}{z_1 r_i - z_2 r_o} \tag{3}$$

**Monotonicity:** Given a fixed XSlit camera, Eqn. 3 reveals that the AR monotonically decreases with respect to $z$. In fact, we can compute the derivative of $z$ with respect to $r_i$:

$$\frac{\partial z}{\partial r_i} = \frac{z_1 z_2 (z_1 - z_2) r_o}{(z_1 r_i - z_2 r_o)^2} \tag{4}$$

Since $z_1 < z_2$, we have $\frac{\partial z}{\partial r_i} < 0$, i.e., the depth $z$ decrease monotonically with $r_i$. In fact the minimum and the maximum ARs correspond to:

$$r_i^{\min} = r_i|_{z\to\infty} = \frac{z_2}{z_1} r_o, r_i^{\max} = r_i|_{z\to z_2} = \infty \quad (5)$$

**Depth Sensitivity:** Another important property we address here is depth sensitivity. We compute the partial derivative of $r_i$ respect to $z$ for $z$ ranging from $z_2$ to $\infty$ and we have:

$$\frac{\partial r_i}{\partial z} = \frac{z_2(z_1 - z_2)}{z_1(z - z_2)^2} r_o \quad (6)$$

The sensitivity is the absolute value of $\frac{\partial r_i}{\partial z}$ and it decrease monotonically for $z > z_2$. This implies that as objects get further away, the depth accuracy recoverable from the AR also decreases. According to Eqn. 6, the sensitivity is positively related to $\frac{z_2}{z_1}$ and $z_1 - z_2$. Farther separated slits and greater ratio between two slits distances corresponds to higher sensitivity. This phenomenon resembles classical stereo matching using two perspective cameras where the deeper the object, the smaller the disparity and the less accuracy that stereo matching can produce.

**Depth Range:** We can further compute the maximum discernable depth $z^{\max}$. To do so, we first compute $r_i$ when $z \to \infty$ as $r_i^{\infty} = \frac{z_2}{z_1} r_o$. Next we change $r_i^{\infty}$ with $\epsilon$, the smallest ratio change that is discernable in image. We have $r_i^* = \frac{z_2}{z_1} r_o + \epsilon$. The lower bound of $\epsilon$ is $1/L$, $L$ is the image width or height, without considering subpixel accuracy. Sine the depth changes monotonically with $r_i$, the maximum discernable depth is correspond to $r_i^*$. Finally we compute the depth use Eqn. 3:

$$z^{\max} = \frac{z_2}{z_1}[z_1 + (z_2 - z_1)\frac{r_o}{\epsilon}] \quad (7)$$

Eqn. 7 indicates that the larger slit distance ratio $\frac{z_2}{z_1}$ and bigger separating distance of two slits $z_2 - z_1$ correspond to a larger discernable depth range.

## 4. Depth Inference from DDAR

Our analysis reveals that if we know $r_o$ in prior, i.e., the base aspect of the object, we can directly infer the object's depth from its aspect ratio in the XSlit camera. A typical example is using an Parallel-Orthogonal XSlit camera (PO-XSlit) to capture an up-right rectangle. In a PO-XSlit camera, the slits are orthogonal and axis aligned. In this case, $r_o$ directly corresponds to the aspect ratio of the rectangle and $r_i$ corresponds to the observed AR of the project rectangle.

The simplest case is to capture a up-right square whose aspect ratio $r_o = 1$. From the AR change, we can directly infer its depth using Eqn. 3. In practice, we do not know the AR of the object in prior. However, many natural scenes

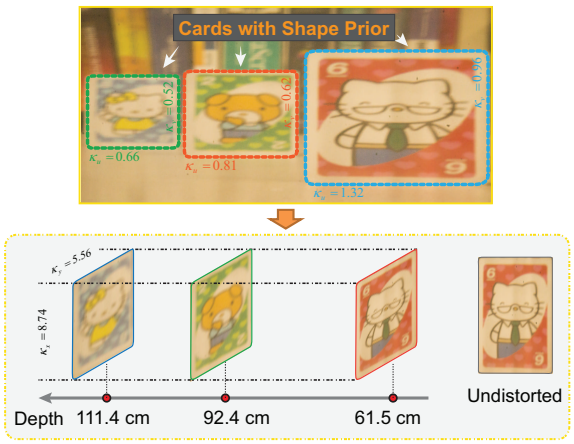



Figure 4: Depth-from-DDAR: Top shows a scene that contains multiple cards of an identical but unknown size. Bottom shows their recovered depths and original size using our scheme from this single image.

contain (rectangular) objects of identical sizes (e.g., windows of buildings) and we can infer their depth even without knowing their ground truth AR.

**Shape Prior** Specifically, consider $K$ rectangles of an identical but unknown sizes and hence ARs. Assume they lie at different depths $z^j$. According to Eqn. 1, we have two equations for each rectangle:

$$\begin{aligned} \kappa_u^j z^j + z_2 \kappa_x &= z_2 \kappa_u^j \\ \kappa_v^j z^j + z_1 \kappa_y &= z_1 \kappa_v^j \end{aligned} \quad (8)$$

Where $j = 1..K$, $z^j$, $\kappa_x$ and $\kappa_y$ are unknowns. And $\kappa_u$ and $\kappa_v$ are computed from the image. For $K$ identical rectangles, we have $K + 2$ unknowns and $2K$ equations. The problem can be solved using SVD when two or more identical rectangles are present. Fig. 4 shows several examples using our technique recovering depth of multiple cards of an identical size. The depth along with the exact scale can be extracted from a single XSlit image under the shape prior.

**Depth Prior** If the objects are of identical aspect ratios but of different sizes, there is still ambiguity. According to Eqn. 2, there are $K$ equations and $K + 1$ unknowns (assume $K$ objects). One useful prior that can be imposed here is the distribution of depth of objects. In real man-made environment, objects are likely to be evenly distributed. We assume that these rectangles are with equal distance along the $z$ direction.

In this scenario/case, we obtain the AR equation for each object:

$$z^j r_o - r_i^j \frac{z_1}{z_2} z^j - z_1 r_o = -z_1 r_i^j, \quad j = 1..K \quad (9)$$

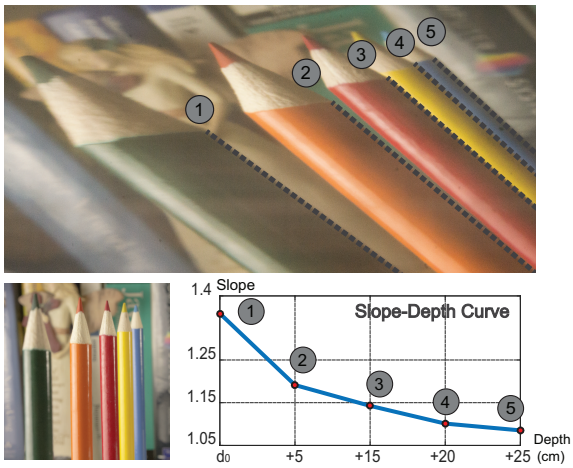



Figure 5: Extending DDAR to DDS. Top: parallel 3D lines map to 2D lines of different slopes in an XSlit image. Bottom: the slopes can be used to infer the depths of the lines.

Furthermore, the equal distance prior gives us the constraint $z^j - z^{j-1} = z^{j+1} - z^j$, for $j = 2...(K-1)$. For $K$ objects in the scene, we have $2K-2$ equations, and $K+1$ unknowns. The problem is determined if we have 3 rectangles in the scene. And it's over-determined if we have more than 3 objects.

It is very important to note that inferring depth under the same setting is not possible in the perspective camera case. In pinhole image $z_1 = z_2$ and $r_i = r_o$, hence Eqn. 8 and Eqn. 9 degenerate. As shown in the introduction, scaling the scene and adjusting the distance from the scene to the pinhole camera accordingly will result in a same projected image as the ground truth scene dose.

### 4.1. Line Slope Analysis

Section 4 reveals that inferring depth from DDAR is that we need to obtain some prior knowledge of either the base AR $r_o$ or the depth distribution of multiple identities. Further, the rectangular shape needs to be in the up-right position to align with the two slits. In this section, we extend the AR analysis to study the slope of lines and we show that this analysis leads to a more effective depth inference scheme.

We treat a line frontal parallel to the XSlit camera as the diagonal of a parallelogram (rectangle in PO-XSlit case), whose sides are along the two slits directions. Given a line with slope $s$ and a point $p_1[x, y, z]$ on it, then we have $p_2[x+1, y+s, z]$ of is on the line. We can map it to a line with slope $s'$ on XSlit image, which $p_1$ and $p_2$ map to points $p_1'(u, v)$ and $p_2'(u+c, v+cs')$ respectively. According to definition of $r_o$, we can decompose the segment $p_1$-$p_2$ onto two slits direction and take the ratio of the two component to get $r_o$:

$$r_o = \frac{\sin\theta_2 - s\sin\theta_1}{s\cos\theta_1 - \cos\theta_2} \tag{10}$$

$r_i$ is computed using Eqn. 10 too, only substitute $s$ with $s'$. Reuse Eqn. 3, we can get the depth.

Eqn. 10 and 3 reveals that we can directly infer the depth of the line from its slope. Similar to the aspect ratio case, such inference cannot be conducted in the pinhole camera since the frontal parallel line slope is invariant to depth.

The analysis above applies only to lines parallel to XSlit camera. For lines unparallel to the camera, previous studies have shown that they map to curves, or more precisely hyperbolas [26]. However, our analysis can still be applied by computing the tangent direction on the hyperbolas, where each tangent direction can be mapped to a unique depth. This can be viewed as approximating a line as piecewise segments frontal-parallel to the camera where each segment's depth can be computed from its projected slope. The complete derivation is included in the supplementary materials.

### 4.2. Scene Reconstruction

Based on our theories, we present a new framework on single-image Manhattan scene reconstruction using the XSlit camera. The main idea here is to integrate depth cues from DDAR (for up-right rectangle objects) and from line slopes (for other lines and rectangles) under a unified depth inference framework. Further, the initial depth estimation scheme can only infer depths on pixels lying on the boundaries of the objects, it is important to propagate the estimation to all pixels in order to obtain the complete depth map.

Our approach is to first infer the depth for the lines or repeat objects from DDAR. Next we cluster pixels into small homogenous patches or superpixels [8]. The use of superpixels not only reduce the computational cost and but also preserves consistency across the regions, i.e the pixels in a homogeneous region such as walls of a building tend to have a similar depth. Finally, we model optimal depth estimation/propagtion as a Markov Random Field (MRF). The initial depth value $V_i$ for superpixel $S_i$ is computed by blending the depths inferred from DDAR according to their geodesic distance to $S_i$. And then we the smooth out $V$ based on distance variations and color consistency. This procedure can be modeled as a Markov Random Field (MRF), where the data term: $E_d(S_i) = U_i - V_i$. And the smoothness term is: $E_s(S_i, S_j) = w_{ij}(U_i - U_j)$, $w_{ij}$ is the weight account for distance variations and color consistency. Finally we estimate the depth map $U$ by optimizing the energy function:

$$E(U) = \sum_{S_i} E_d(S_i) + \lambda \sum_{S_i, S_j \in N} E_s(S_i, S_j) \tag{11}$$

Figure 6: An XSlit image of the arch scene that contains 3D concentric circles (left). Their images correspond to ellipses of different aspect ratios (right).

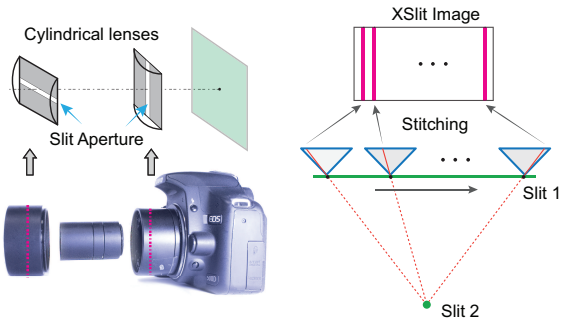



Figure 8: XSlit images can be captured by a real XSlit lens (left) or by stitching linearly varying columns from a 3D light field (right).

$N$ represents the superpixel neighborhood. The problem can be solved using the graph-cut algorithm [2].

## 5. Experiments

We experiment our approach on both synthetic and real scenes. For synthetic scenes, we render images using 3ds Max. For real scenes, we acquire images using the XSlit lens as well as synthesize XSlit panoramas from video sequences.

**Synthetic Results.** We first render an XSlit images of a scene containing repeated shapes (Fig. 6). The architecture consists of concentric arches of depths ranging from 900cm to 2300cm. We assume that the actual aspect ratio of the arches is 1, i.e., a circle. We position a PO-XSlit camera with $z_1 = -3.2$cm and $z_2 = -346.7$cm frontal parallel to the arches and the images of the arches are ellipses of different aspect ratios. Notice that in the pinhole case, they will be map to circles. We first detect ellipses using Hough transform and then measure their aspect ratios using the major and minor axes. Finally, we use the ratios to recover their depths using Eqn. 3. Our recovered depths for the near and far arches are 906.6cm and 2281.0cm, i.e., the errors are less than 2%.

Next we render two XSlit panoramas, one for the corridor and the second for the facade. Both scenes exhibit strong linear structures with many horizontal and vertical lines. Our analysis shows that for lines to exhibit DDS, they should not align with either slit. Therefore, we rotate the POXSlit, i.e., $\theta_1 = 45°$ and $\theta_2 = 135°$. For the corridor scene, the XSlit camera has a setting of $z_1 = -3.6$cm, $z_2 = -717.9$cm and for the facade scene, $z_1 = -3.1$cm, $z_2 = 4895.9$cm. We first use the LSD scheme[24] to extract 2D lines from the XSlit images and cluster them into groups of horizontal and vertical (in 3D) lines. This is done by thresholding their aspect ratios Eqn. 5. For lines in each group, we compute their depths using Eqn. 10 and 3. This results in a sparse depth map. To recover the full depth map, we apply the MRF (Sec. 4.2) and the final result is shown in Fig. 7. Our technique is able to recover different depth layers while preserving linear structures. For comparison, we render a single perspective image and apply the learning-based scheme Make3D [20]. Make3D can detect several coarse layers but cannot detect fine details as ours since these linear structures appear identical in slope in a perspective image but exhibit different slopes in an XSlit image.

**Real Results.** We explore several approaches to acquire XSlit images of a real scene: by a real XSlit lens and through panorama synthesis. For the former, we use an XSlit lens [26]. The design resembles the original anamorphoser proposed by Ducos du Hauron that replaces the pinhole in the camera with a pair of narrow, perpendicularly crossed slits. Similar to the way of using a spherical thin lens to increase light throughput in a pinhole camera, the XSlit lens relay perpendicular cylindrical lenses, one for each slit. In our experiments, we use two cylindrical lenses with focal lengths 2.5cm (closer to the sensor) and 7.5cm (farther away from the sensor) respectively. The distance between the two slits is adjustable between 5cm and 12cm and the slit apertures have a width of 1mm.

We first capture a checkerboard at known depths and compare the measured AR and our predicted AR using Eqn. 3. We test three different slit configurations, $z_2/z_1 = 1.3$, $z_2/z_1 = 1.59$ and $z_2/z_1 = 2.0$. Fig. 9 shows that the predicted AR curve fits well with the ground truth. In particular, as an object gets farther away from the sensor, its AR also changes slower. Further, the larger the baseline $z_2/z_1$ is, the larger the aspect ratio variations across the same depth range, as predicted by our theory.

Next, we verify our DDS analysis using images captured by the XSlit camera. In Fig. 10, we position a Lego® house model in front of the XSlit camera ($z_1 = 6.12$cm and $z_2 = 11.81$cm). We rotate the XSlit camera by 45 degrees so that the 3D lines on the house will not align with either

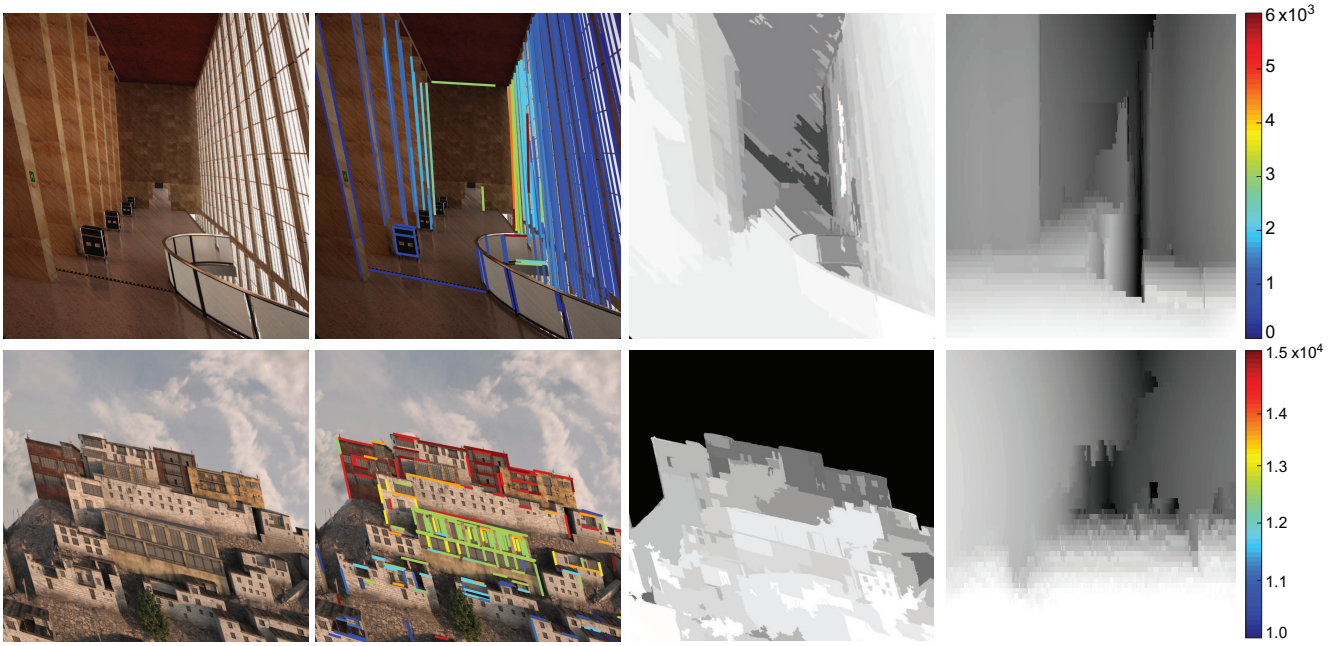



Figure 7: From left to right: An XSlit image of a scene containing parallel 3D lines, the detected lines and their estimated depth using DDS, the depth map recovered using our scheme, and the one recovered using Make3D [20] by using a single perspective image.

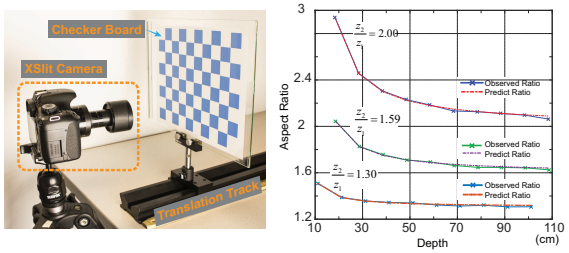


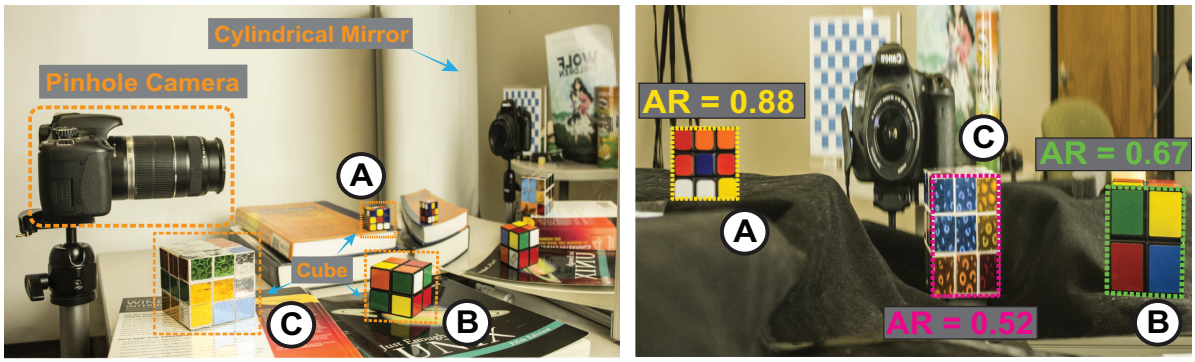

Figure 9: Experimental validations of our analysis. We place a checker board in front of the XSlit camera and move it away(Left). The comparisons of measured AR and predict AR with different silts configurations(Right).


Figure 10: Real result on a Lego® house scene. (a) an XSlit image of the scene captured by the XSlit camera. Detected lines are highlighted in the image. (b) the recovered depth map using our slope and aspect ratio based scheme.

slit. Fig. 10(a) shows the acquired image. Next, we conduct line fitting and slope estimation similar to the synthetic case for estimating the depths of the detected lines. Fig. 10(a) highlights the detected lines and their depths (using color) and Fig. 10(b) shows the complete depth map using the MRF solution. The results shows that major depth layers are effectively recovered. The error on the top-right corner is caused by the lacking of line structures.

A major limitation using the XSlit camera is its small baseline (between the two slits). Our analysis shows that the maximum recoverable depth range depends on this baseline. Further, since images captured by the XSlit camera exhibits noise and strong defocus blurs, the actual recoverable depth range is even smaller. For example, our analysis shows that with baseline $z_2/z_1 = 2$, two cards are placed at $30m$ and $35m$ will have undistinguishable ARs. Their ratio difference reach the lower bound that determined by pixel size. For outdoor scenes, we resort to XSlit panorama synthesis.

To produce XSlit panoramas, Zomet et al. [29] capture a sequence of images by translating a pinhole camera along a linear trajectory at a constant velocity. In a similar vein, Seitz and Adams et al. acquire the image sequence by mounting the camera on a car facing towards the street. Next, linearly varying columns across the images are selected and stitched together. Fig. 8 shows the procedure of generating a XSlit image using a regular camera.




(a) Scene Setup (b) Captured Image

Figure 12: Results on catadioptric mirrors. Left: we capture the scene using a cylindrical catadioptric mirror. Right: the aspect ratios of cubes change with respect to their depthes.

Figure 11: The XSlit image of an outdoor scene. Left: An XSlit panorama and the detected lines. Right: The recovered depth map.

Fig. 11 shows the XSlit panorama synthesized from an image sequence captured by a moving camera. We linearly increase the column index in terms of frame number and stitch these columns to form an XSlit image. The moving path of the camera is 55cm long. And the camera is tilt with 20° angle. The resulting two slits are at -1.8cm and 41cm respectively.

Recent ray geometry studies [7] show that reflections of certain types of catadioptric mirror can be approximated as an XSlit image. In Fig. 12, we position a perspective camera facing towards a cylindrical mirror and Fig. 12(b) shows that DDAR can both be observed on the acquired image. In particular, we put multiple cubes of an identical size at different depths and their aspect ratios change dramatically. This is because two virtual slits of the catadioptric mirror are separated far away where DDAR is more significant than the XSlit camera case.

## 6. Conclusion and Further Work

We have comprehensively studied the aspect ratio (AR) distortion in XSlit cameras and exploited its unique depth-dependent property for 3D inference. Our studies have shown that unlike perspective camera that preserves AR under depth variations, AR changes monotonically with respect to depth in an XSlit camera. This has led to new depth-from-AR schemes using a single XSlit image. We have further shown that similar to AR variations, the slope of projected 3D lines will also vary with respect to depth, and we have developed theories to characterize such variations based on AR analysis. Finally, AR and line slope analysis can be integrated for 3D reconstruction and we have experimented on real XSlit images captured by an XSlit camera, synthesized from panorama stitching, and captured using a catadioptric mirror to validate our framework. We admit that the proposed depth-from-AR technique is not yet comparable to state-of-the-art multi-images (stereo or SfM) or active illumination (Kinect) based techniques. However, a simpler setup lends itself to more cost-effective depth sensing systems. We also hope our work would stimulate more work in the less explored space of single XSlit imaging.

There are a number of future directions we plan to explore. Our cylindrical lens based XSlit has a small baseline (i.e., the distance between the two slits) and therefore can only acquire AR changes within a short range. Constructing a large baseline XSlit camera will be costly as it is difficult to fabricate large form cylindrical lens. A more feasible solution would be adopt a cylindrical catadioptric mirror where the reflection image can be approximated as an XSlit image. In the future, we will explore effective schemes for correcting both geometric distortion and blurs due to imperfect mirror geometry. We will also investigate integrating our AR based solution into prior based frameworks to enhance reconstruction quality. For example, a hybrid XSlit-perspective camera pair can be constructed. Finally, since AR distortions commonly exhibit in synthesized panoramas as shown in the paper, we plan to study effective image-based distortion correction techniques to produce perspectively sound panoramas analogous to [1].

## Acknowledgements

This project was supported by the National Science Foundation under grants IIS-1513031 and IIS-1422477.

## References

[1] A. Agarwala, M. Agrawala, M. F. Cohen, D. Salesin, and R. Szeliski. Photographing long scenes with multi-viewpoint panoramas. *ACM Trans. Graph.*, 25(3):853–861, 2006.

[2] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(9):1124–1137, 2004.

[3] R. Cabral and Y. Furukawa. Piecewise planar and compact floorplan reconstruction from images. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 628–635. IEEE, 2014.

[4] J. M. Coughlan and A. L. Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 941–947. IEEE, 1999.

[5] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *International Journal of Computer Vision*, 40(2):123–148, 2000.

[6] E. Delage, H. Lee, and A. Y. Ng. A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2418–2428. IEEE, 2006.

[7] Y. Ding, J. Yu, and P. Sturm. Recovering specular surfaces using curved line images. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2326–2333. IEEE, 2009.

[8] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[9] A. Flint, C. Mei, D. Murray, and I. Reid. A dynamic programming approach to reconstructing building interiors. In *Computer Vision–ECCV 2010*, pages 394–407. Springer, 2010.

[10] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing building interiors from images. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 80–87. IEEE, 2009.

[11] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. *ACM Transactions on Graphics (TOG)*, 24(3):577–584, 2005.

[12] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 654–661. IEEE, 2005.

[13] J. Košecká and W. Zhang. Video compass. In *ECCV 2002*, pages 476–490. Springer, 2002.

[14] D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2136–2143. IEEE, 2009.

[15] M. Levoy and P. Hanrahan. Light field rendering. In *ACM SIGGRAPH*, pages 31–42, 1996.

[16] J. Novatnack and K. Nishino. Scale-dependent 3d geometric features. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[17] T. Pajdla. Geometry of two-slit camera. *Rapport Technique CTU-CMP-2002-02, Center for Machine Perception, Czech Technical University, Prague*, 2002.

[18] J. Ponce. What is a camera? In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1526–1533. IEEE, 2009.

[19] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *Advances in Neural Information Processing Systems*, pages 1161–1168, 2005.

[20] A. Saxena, M. Sun, and A. Y. Ng. Learning 3-d scene structure from a single still image. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[21] Y. Y. Schechner and S. K. Nayar. Generalized mosaicing. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 17–24. IEEE, 2001.

[22] G. Schindler and F. Dellaert. Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. IEEE Conference on*, volume 1, pages I–203. IEEE, 2004.

[23] S. M. Seitz and J. Kim. The space of all stereo images. *International Journal of Computer Vision*, 48(1):21–38, 2002.

[24] R. G. Von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall. Lsd: a line segment detector. *Image Processing On Line*, 2(3):5, 2012.

[25] W. Yang, Y. Ji, J. Ye, S. S. Young, and J. Yu. Coplanar common points in non-centric cameras. In *Computer Vision–ECCV 2014*, pages 220–233. Springer, 2014.

[26] J. Ye, Y. Ji, and J. Yu. Manhattan scene understanding via xslit imaging. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 81–88. IEEE, 2013.

[27] J. Ye, Y. Ji, and J. Yu. A rotational stereo model based on xslit imaging. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 489–496, Dec 2013.

[28] J. Yu and L. McMillan. General linear cameras. In *ECCV*, 2004.

[29] A. Zomet, D. Feldman, S. Peleg, and D. Weinshall. Mosaicing new views: the crossed-slits projection. *IEEE TPAMI*, 25(6):741–754, June 2003.