Toward a Theory of Perspective Perception in Pictures: **Condensed Version**

Aaron Hertzmann*
Adobe Research

May 24, 2024

Abstract

"Projection" describes how painters and cameras depict 3D scenes in 2D. Perceptual theories and studies of projection are dominated by linear perspective, which originated in the Italian Renaissance. Yet, linear perspective fails to fully explain both how we humans understand realistic pictures, and how artists make them. I propose a theory of projection perception based on a two-part model of 3D human vision. In picture viewing, the picture contents around each fixation are interpreted according to a local linear perspective, centered at the fixation. Over multiple fixations, an abstracted global understanding arises from spatial relationships between these fixations. This framework offers new understanding of many pictorial phenomena across many kinds of realistic pictures, and suggests new ways to make and understand pictures.

This is a condensed version of a much-longer paper published in *Journal of Vision,* April 2024. This version focuses on the hypotheses that I propose for picture perception, and omits many important details, as well as most of the literature survey. If you wish to cite this, please cite the full-length version. See http://www.dgp.toronto.edu/~hertzman/perspective **for the full paper and other links.**

I recommend reading this version first, and then, if desired, reading/skimming the long paper for the full story and research survey.

1 Introduction

> *"The eye moves all the time. When my eye moves in one direction, the perspective goes that way." —David Hockney (Gayford, 2022)*

The geometric relationship between 3D points in space and their 2D positions in a picture can be described in terms of a projection. Some projections follow simple parametric formulae, such as linear perspective and orthographic perspective, whereas others may be more free-form. Classic drawing techniques construct projections implicitly (Willats, 1997), whereas modern cameras and computer graphics systems are explicitly designed in terms of mathematical projections.

When interpreting shape and space in a picture, what assumptions do viewers make about the projection? The information in a picture is ambiguous. Yet, there must be some assumptions, since a realistic picture conveys to viewers a sense of the shapes of scene elements, along with their relative sizes, positions, and distances from the viewer. Throughout history, different artists have portrayed objects and spaces in different ways, whether the freeform arrangements in cave paintings and Modern Art, ancient Chinese isometric perspective, or strict linear perspective typical in consumer photography (Figure 1). Some approaches seem

(a) (b) (c)

(d) (e)

Figure 1: Examples of different approaches to projection in art history. (a) Prehistoric painting, using simple arrangements of elements to convey scenes. (b) Chinese scroll painting, using parallel projection. All people are the same size, regardless of distance to the viewer. (c) Russian icon painting, using reverse perspective, in which some objects expand away further from the viewer. (d) Early Renaissance painting before linear perspective. Objects closer to the viewer are larger, and closer to the bottom of the painting. (e) Modern Art painting with a more freeform projection. *Sources:* (a) Prehistoric rock paintings of Tassili N'Ajjer, Algeria, photograph by Dmitry Pichugin. (b) *Eighteen Songs of a Nomad Flute: The Story of Lady Wenji* (detail), unidentified artist, 15th century CE. (c) *The Trinity*, Andrei Rublev, 15th century CE. (d) *Adoration of the Mystic Lamb* from the Ghent Altarpiece, 15th century CE. (e) *The Pink Studio*, Henri Matisse, 1911.

more realistic, and some more abstract or stylized. The diversity of these different approaches and their percepts makes it difficult to articulate what, exactly, viewers' assumptions about projections might be.

This paper proposes hypotheses to explain the projection assumptions in human perception of pictures. Recent developments in vision science indicate that visual awareness is not the sort of 3D reconstruction of the world around us that has often been assumed. Instead, I argue that most of our 3D shape inference occurs per-gaze, in each fixation in the world, and the richest shape detail occurs in foveal vision. Over multiple fixations, much of this 3D information is discarded, as a viewer builds up an abstracted mental representation of the world, and not a detailed 3D reconstruction. Likewise, in picture viewing, shape in each fixation is inferred according to a linear perspective around the fixation point. Over multiple fixations, the viewer constructs an abstracted representation of the 3D contents of the picture. I articulate these ideas in a series of hypotheses that can explain many phenomena of 3D perception across many kinds of pictures, while being consistent with the nature of real-world foveal and 3D vision.

I begin by describing properties of projections that do and do not produce perceived distortions, using examples from art history and computational photography. Perceived distortion is an important part of

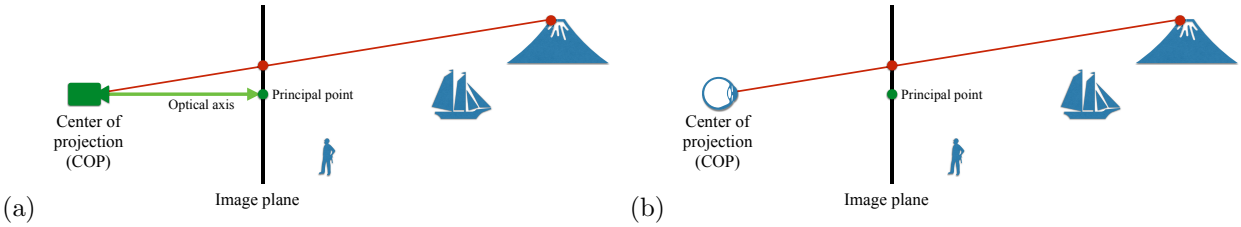


Figure 2: Linear perspective geometry. (a) Linear perspective imaging is defined by an image plane, a center-of-projection (COP), and an optical axis (view direction). The principal point is the intersection of the optical axis with the image plane. The color at an image point is determined by the light incoming to the COP along the ray from the image point. (b) When viewing the picture from the center-of-projection (COP) in ideal conditions, the eye will receive the same light as if it were looking through a window into the real scene, regardless of their gaze direction.

artistic style and technique, and I do not suggest they should be avoided. However, I focus on techniques to avoid perceived distortion since they provide valuable clues toward viewers' projection assumptions.

2 Linear Perspective

A linear perspective projection is defined by a center-of-projection, a view direction, and an image plane (Figure 2(a)). Most consumer camera lenses aim to approximate linear perspective, and linear projection is widespread in computer graphics and vision algorithms. The techniques of one-point and two-point perspective are familiar to many artists and art students.

Linear perspective has dominated theories of projection, e.g., (Kemp, 1990; Elkins, 1994; Hecht et al., 2003). Yet, it is rare that artists employ strict linear perspective (Kemp, 1990; Verstegen, 2010; Pepperell and Haertel, 2014; Koenderink et al., 2016; Kemp, 2022). Many famous painters, including Leonardo da Vinci, J. M. W. Turner, and David Hockney, achieved mastery of linear perspective, and each later wrote about its shortcomings, while exploring more flexible approaches to perspective (Kemp, 1990, 2022).

Linear perspective arises from the idea that a picture simulates the light the viewer would see if they were looking through a window. That is, when viewing monocularly from the picture's center-of-projection (Figure 2(b)), the retinal image should simulate viewing the depicted scene. At one stage in his investigations, Leonardo da Vinci wrote that a picture will "look wrong, with every false relation and disagreement of proportion that can be imagined in a wretched work, unless the spectator, when he looks at it, has his eye at the very distance and height and direction where the eye ... was placed" (Kubovy, 1986). Yet, pictures can "look right" from many different viewpoints, even pictures that do not have well-defined centers-of-projection.

A linear perspective picture is *wide-angle* if it uses a much wider field-of-view than it would normally be viewed with (Cooper et al., 2012). Wide-angle pictures are common throughout art history and photography. Many historical paintings display large-scale scenes that would have required a wide-angle linear perspective to capture a comparable spatial extent and object scale, e.g., Figures 1, 3. Smartphones take wide-angle photos *by default*.

Viewing a wide-angle picture from the center-of-projection is rare and even uncomfortable (Koenderink et al., 2016). For example, in Figure 4(a), the viewing distance should be approximately 40% of the image width. That is, if the image appears printed or on the screen as 10cm wide, the viewer's eye should be 4cm from the center of the picture. This is a very unusual viewing position, and some peoples' eyes cannot even focus at this distance. Center-of-projection viewing on smartphone displays is often physically impossible for pictures taken with smartphone defaults.

The full-length version of this paper includes instructions for trying COP viewing yourself, by enlarging a paper figure on a large display, in the "Viewing from the COP" section (page 10). This is a worthwhile and instructive exercise.

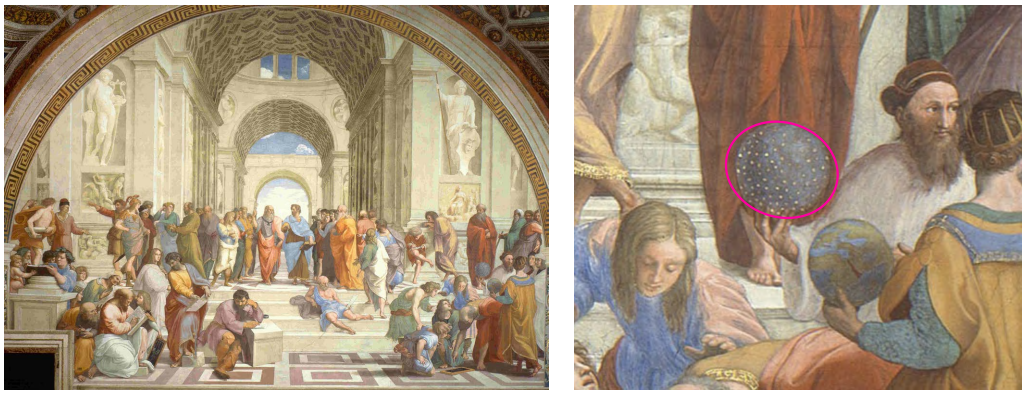


Figure 3: Raphael's *The School of Athens* employs precise one-point linear perspective for the architecture but not for the people. None of the faces appear distorted as they would be in a true wide-angle linear perspective image, such as in Figure 4(a). According to the projection implied by the architecture, the spheres in the lower-right corner should have the aspect ratio 1.2:1 (Zorin and Barr, 1995), visualized here with a magenta ellipse.

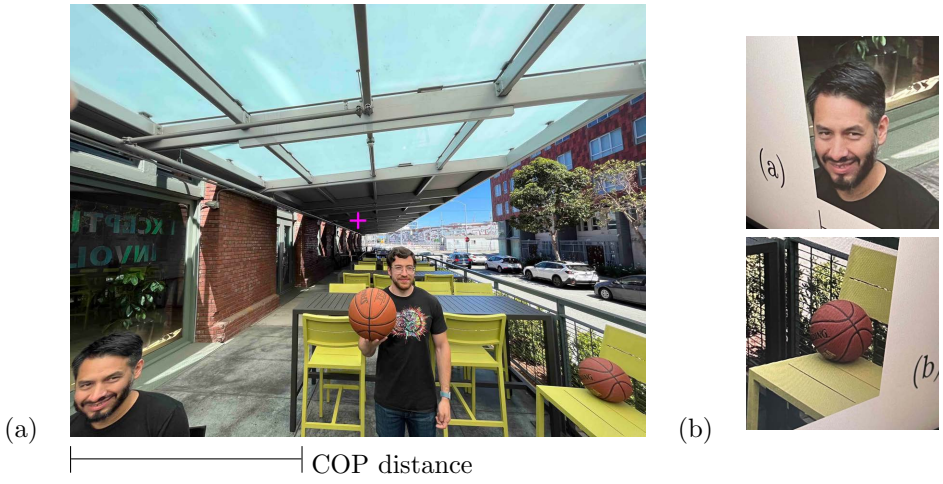


Figure 4: Wide-angle photo for experiencing marginal distortion in COP viewing. (a) Photo taken with iPhone 13 in ultrawide mode (0.5x, 14mm). The magenta cross indicates the picture center. COP distance is shown below the picture; it is 40% of the width of the picture. To view from the COP, place one eye in front of the magenta cross, with distance according to the length of the "COP distance" line. One may need to display the picture on large display or projection in order to be able to. Note that the marginal distortion appears or disappears depending on whether one views monocularly or binocularly. (b) Photos of the same picture displayed on a computer screen, and photographed by a smartphone approximately positioned at the COP location and aimed at the bottom corners.

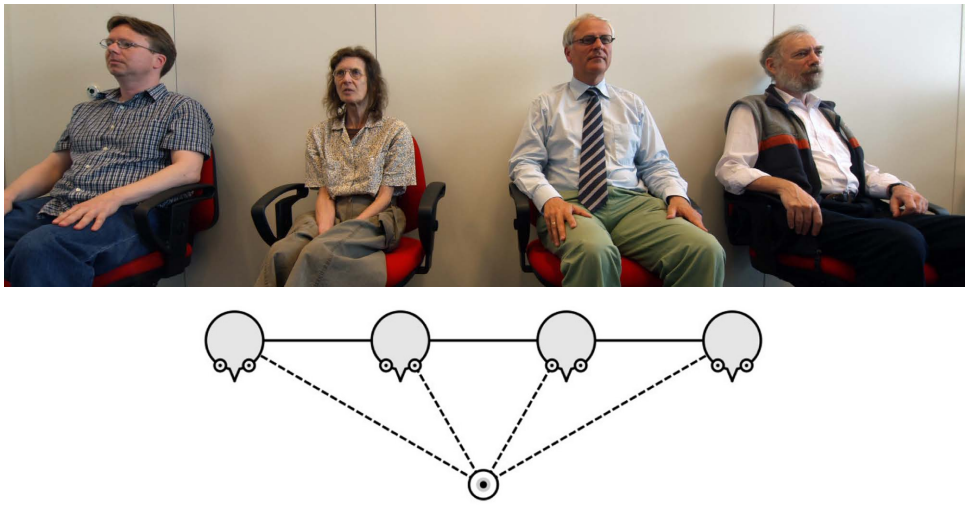


Figure 5: Wide-angle linear perspective photograph, from (Koenderink et al., 2010), taken with a 14 mm equivalent lens and cropped. The four individuals face in parallel directions, illustrated in the diagram. Yet, under normal viewing conditions, they appear to be facing in diverging directions in the photograph.

Typical viewing of wide-angle linear perspective produces several well-known *perceived distortions*: situations where a viewer recognizes that a shape "looks wrong" (Vishwanath et al., 2005; Cooper et al., 2012). These perceived distortions provide important clues to the vision system's assumptions about how shapes "should be" depicted, and give clues to the nature of mental representations.

Wide-angle linear perspective causes objects in the periphery to appear distorted, a phenomenon known as *marginal distortion*. Figure 4(a) shows an example where spheres and faces appear oblong in the corners of photographs. Wide-angle perspective can also make objects appear compressed or expanded in depth, an effect sometimes called "perspective distortion" (Cooper et al., 2012).

Note that a viewer is not necessarily aware of shape misperceptions due to marginal distortion, as demonstrated by Koenderink et al. (2010), see Figure 5. The main difference between this case and the marginal distortion of spheres is whether prior knowledge allows a viewer to recognize a misleading depiction of a familiar object.

3 Multiperspective projections in art and computational photography

Many artistic techniques for avoiding distortion and for conveying wide-angle scenes are *multiperspective* (Kubovy, 1986; Agrawala et al., 2000; Perona, 2013; Koenderink et al., 2016). Moreover, modern computational photography techniques that mimic classical approaches explicitly do so with multiple perspectives.

Many classical paintings appear to combine multiple linear projections with different centers-of-projection. Raphael's *The School of Athens* (Figure 3) provides a particularly famous example (Kubovy, 1986). Raphael employed strict one-point perspective for the architecture. However, he does not paint objects with marginal distortion. For the globes in the right-hand corner of the image, Raphael has painted spheres as circles, whereas linear perspective would dictate that they should be oblong (Kubovy, 1986; Zorin and Barr, 1995).

Moreover, none of the faces in *The School of Athens* exhibit marginal distortion, i.e., compare the faces between Figures 4 and 3. Large scenes with many faces are common in art, e.g., Figures 1. *But, in the entire history of painting, I am unaware of any face depicted with the marginal distortions dictated by linear perspective.* Instead, these depictions can be explained by the use of multiple perspectives.

(a) 

(b) 

Figure 6: Photomontage as a metaphor for making pictures from a collection of perspectives. (a) Individual object depictions can be placed independently, without introducing distortion. *Just what is it that makes today's homes so different?* by Richard Hamilton, 1992. (b) An artist creates an overall perspective and sense of a scene by arranging much smaller, undistorted photographs. *Pearblossom Hwy., 11 - 18th April 1986, #2*, David Hockney, 1986

Classical paintings often exhibit many other multiperspective effects (Kubovy, 1986; Agrawala et al., 2000; Perona, 2013; Koenderink et al., 2016). For example, in the 15th Century, Paolo Uccello subtly combined multiple viewpoints in his portrait of Niccolò da Tolentino, as did Andrea del Castagno in his portrait of Dante Aligheri. Hockney (2006) (p. 82–113) characterizes some classical paintings as "multiwindow," in which we see each figure "straight on, regardless of where they are in the scene," which he visualizes by cropping individual elements from the picture. Photocollage art like Hockney's *Pearblossom Highway* and Richard Hamilton's *Just what is it that makes today's homes so different, so appealing?* (Figure 6 illustrate how collaging linear perspective pictures can create varied compositions.

In the past two decades, computer graphics and computational photography research has developed multiperspective techniques that can avoid perceived distortion and create compositions more effectively than strict linear perspective. These methods typically work like collage: they partition the picture plane into regions, each of which has its own linear perspective projection, with its own center-of-projection in front of the region. The partitioning and individual projections depend on the content of the scene being depicted, and the creator's goals. These methods can effectively produce large-scale imagery with little or no perceived distortion.

Here are a few key examples of these computational multiperspective projections. Agrawala et al. (2000) demonstrated that a simple way to avoid distortion is to simply render each object with its own linear perspective, with the center-of-projection in front of the object. Multiperspective street panoramas (e.g., Figure 7(a)) can provide more effective visualization for street imagery than linear perspective (Roman et al., 2004; Agarwala et al., 2006). By collaging separate linear projections with separate centers-of-projection, these methods can create large-scale panoramas with little apparent distortion. Figure 7(b) shows an example of many different collaged linear perspectives. Similarly, collaging in depth can provide control over object scale (Badki et al., 2017), see Figure 7.

The warping method by Carroll et al. (2009) operates from a single perspective (Figure 9), but transforms a picture according to distortion principles from multiperspective. Another warping method by Shih et al. (2019) runs on the Google Pixel's camera app (Figure 10).

I claim that these methods provide better descriptions of artistic practice than does strict linear perspective, since they successfully avoid distortion in ways that mimic classical painting.

Figure 7: Multiperspective collages constructed with multiple vanishing points, illustrating how plausible-looking pictures can be constructed from collage of separate linear perspectives. (a) Computational panorama of a street in Antwerp, from (Agarwala et al., 2006). (b) Six of the 107 fisheye photographs used as input, which were reprojected to a common plane with linear perspective. (c) Visualization of how the panorama was algorithmically composited from individual linear perspective pictures. (d) *Family in a Box* by Frédo Durand (2023). Each compartment was photographed separately and composited, and each compartment has its own vanishing point. (e) Photography stage used for each compartment.

(a)  (b)

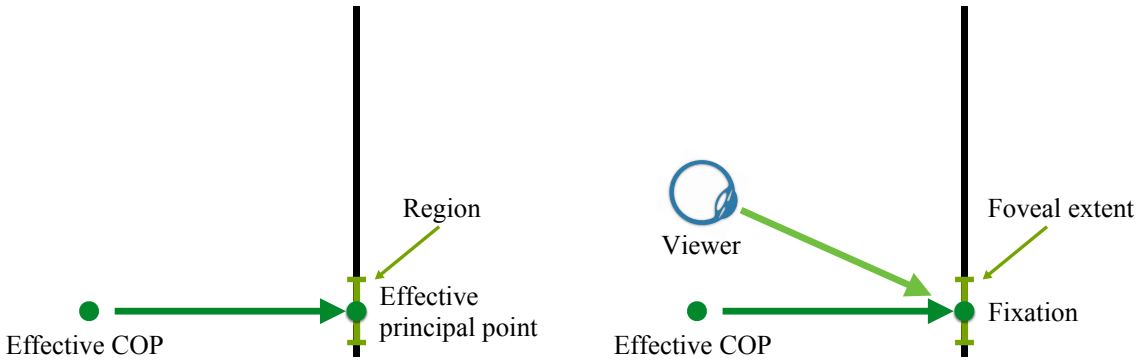
Figure 8: Multiperspective computational zoom, from Badki et al. (2017). (a) Two of the input linear perspective photos, all of which have a dolly-zoom relationship. In the left photo, the building appears very distant; in the right photo, the building appears larger but the people appear distant. (b) Output collage, in which both the people and the building appear larger and more visible, creating a more balanced composition of the people and building.

(a) 

(b) (c)

Figure 9: Content-aware projection of wide-angle photography, from (Carroll et al., 2009). (a) An input wide-angle linear perspective photograph. (b) A stereographic projection computed from the input photo, which creates new distortions. (c) A content-aware projection computed from the input photo, by warping the input photo, in a way that preserves straight lines and other objects, while allowing textureless regions to warp.

Figure 10: A taut piece of string in front of a face, photographed in the Google Pixel 5 camera app in ultrawide mode (0.5x). This app uses a version of Shih et al. (2019), which is content-aware: the face is detected, and the region around it projected with stereographic projection, while the rest of the image uses linear perspective. As a result, the face does not exhibit marginal distortion as it would in linear perspective, but the piece of string is not straight, nor are the lines on the wall near the face. (Photo by Elena Adams.)



(a) Direct View Condition (b) Fixation Compensation

Figure 11: Geometry for two of the hypotheses here. (a) The Direct View Condition (DVC) asserts that the contents of a picture region appear undistorted if they look like a linear perspective projection of a plausible 3D scene, with the COP somewhere in front of that region, and, thus, the principal point within the region. (b) Fixation Compensation asserts that, when gazing on a specific fixation point in a picture, a viewer treats the region around the fixation point as a linear perspective picture, with COP in front of the fixation, with COP distance approximately equal to the viewer's distance to the fixation.

4 The Direct View Condition

Furthermore, I claim that the effectiveness of multiperspective projection can be explained by the following hypothesis, the **Direct View Condition**, adapted from a previous computational formulation by Zorin and Barr (1995):

> Under normal viewing conditions, an object (or, more generally, picture region) appears undistorted if and only if it looks like it could appear at the center of a normal-field-of-view linear perspective picture of a plausible scene.

This statement implies a local, linear perspective with the center-of-projection in front of the picture region, visualized in Figure 11(a). The degree of deviation from the appearance of a linear perspective determines how distorted the region looks.

The Direct View Condition dictates that straight lines should always appear straight and spheres should project to circles, since this is how they appear at the center of a linear projection. More generally, to make a picture look undistorted everywhere, every picture region should look plausible according to a center-of-projection in front of the region. Avoiding distortion in wide-angle projection requires simulating multiperspective projection.

The Direct View Condition says nothing about veridical shape inference: an object may appear undistorted but give a misleading shape percept, such as in forced perspective. It does not even require the existence of an underlying scene: the scene depicted in *The School of Athens* never existed, but nonetheless appears to be an undistorted projection of plausible elements.

Figure 12 show examples of stereographic projections that produce perceived distortions, e.g., straight lines are not depicted as straight. Here the distortion is an intentional artistic effect; different artistic styles often depict shape in very different ways.

Even in the presence of distortions, we often have a sense of the underlying shape based on prior knowledge. There seems to be multiple representations in vision: the directly perceived shape, and the understood shape according to prior knowledge. Figure 13 illustrates this concept.

Figure 12: Artistic photography with 360° stereographic projection. Many scene elements are visibly distorted. These images were captured with a Ricoh Theta S camera and then later projected to 2D in an interactive application. The effect on the left is called "little planet," describing the percept it gives. Photos by Rich Radke.

Figure 13: How is a picture like a glass of water? If, in real-life, we view a spoon in a glass of water, the spoon appears on the side of the glass as broken and bent. A viewer sees the spoon as distorted, but can understand that it is a normal spoon. Moreover, the viewer can recognize the difference between the appearance and known shape. I argue that perceived distortion in pictures operates similarly: a picture gives a distorted shape perception, and the viewer infers a more normal shape, and recognizes the mismatch between appearance and known shape. This also illustrates the roles of multiple distinct shape representations in vision.

5 Foveal Vision, Fixations, and 3D Vision

Why would multiperspective projections produce pictures that don't look distorted?

In order to answer this question, I turn to recent developments in foveal and 3D vision. Many conventional theories of vision—as well as common-sense notions of it—assume that we viewers continually see the entire visible space in front of our eyes and build an accurate mental 3D model of it, e.g., see the review in (Linton et al., 2022). However, many surprising experimental results in the past few decades have challenged this view.

First, the retinal information available at a glance is very limited. In real-world viewing, we perceive far more detail in the gaze direction (foveal region) than in peripheral vision (Rosenholtz, 2016). To attend to something, we look at it (O'Regan and Noë, 2001; Wolfe et al., 2022), and what gets noticed depends on where one's eyes fixate, for how long, and the limitations of peripheral vision (Rosenholtz, 2020).

The reader is encouraged to try to read text without fixating directly upon it. For example, fixate on one word on this page, and then see how many other words are readable; or stare at one street sign and attempt to read words on a nearby sign. It is unexpectedly difficult or impossible.

Yet, in order to be most effective for helping us navigate and survive the world, human vision must operate at each glance. Indeed, from a single fixation, a viewer can get a sense of overall scene structure and contents (Fei-Fei et al., 2007; Greene and Oliva, 2009), e.g., recognizing that a scene comprises a city street.

If vision at a glance is so effective, then perhaps we do not need to reconstruct a detailed 3D mental model of the world over time. Indeed, numerous studies identify inconsistencies in viewers' behavior that cannot be explained by any consistent mental 3D reconstruction of the real world (Linton et al., 2022). For example, Koenderink et al. (2008) demonstrate spatial inconsistencies in peoples' behaviors in a real-world pointing experiment. Change blindness experiments demonstrate that viewers may forget the appearances of individuals before them (Simons and Levin, 1998), and, in virtual reality experiments, viewers do not notice small rotations of the entire world during saccades (Sun et al., 2018; Langbehn et al., 2018). Some 3D information must persist across fixations, but far less than the dense 3D that one might assume. For these reasons, I hypothesize that *all fine-grained 3D vision occurs in per-fixation visual processing.*

Many of the above observations are highly counterintuitive, leading to an "awareness illusion:" we experience a richly-detailed visual sense of space, yet, when probed, demonstrate a surprising lack of awareness of many details (Dennett, 1991; Noë, 2002). The consistency of our 3D perception is explained not by consistency of our representations, but by the consistency of the world.

I claim that these counterintuitive observations directly translate to picture perception. The experience of picture viewing creates a "pictorial awareness illusion:" we think we are seeing an entire picture at once, when we are actually moving our gaze to attend to different regions sequentially. Furthermore, viewers do not reconstruct a 3D pictorial space that is fully coherent across fixations over a picture.

6 Local Principles of Shape Perception

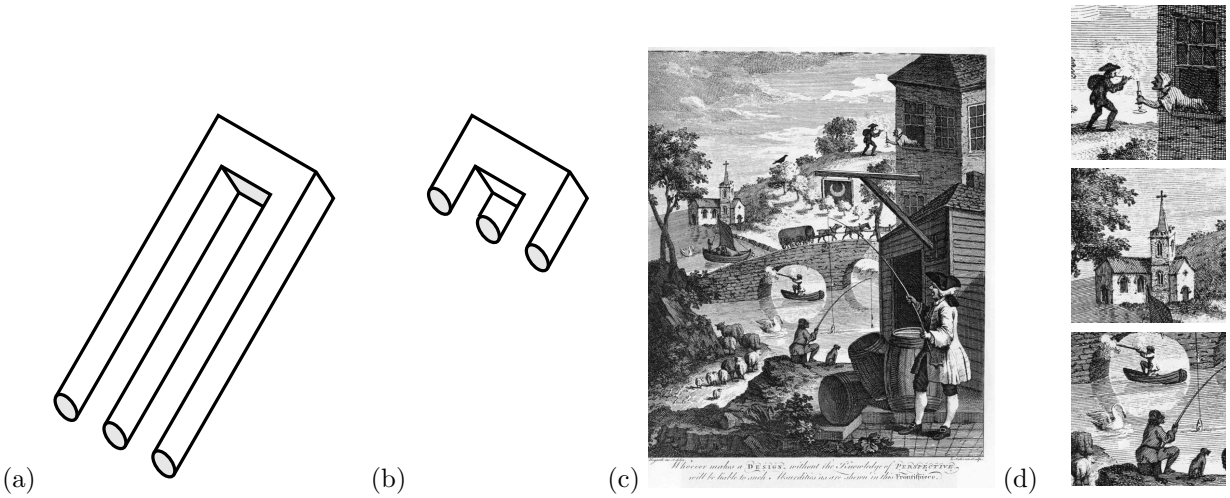
Based on these observations, I propose two hypotheses of how shape perception in pictures depends on eye movements and fixations.

First, I note that shapes are often stable over time, in unconstrained normal viewing of a picture, a principle that I call **Shape Locality**:

> Once objects are recognized, perceived object shape within a small picture region does not depend on the rest of the picture.

Shape Locality describes how changing the visual context around an object does not change its appearance, as illustrated in Figure 14. Likewise, in classical paintings and realistic photographs, one can generally crop out individual portions of a picture without changing object appearance. For example, the contents of the crop in Figure 3 has the same apparent shape as they do in the uncropped picture. Even the impossible

Figure 14: Changing the context around the cars in the middle of the picture does not change their apparent shapes, despite very different visual contexts. (Left image is original photo; the middle and right images were generated using Adobe Photoshop "Generative Fill.")



(a) (b) (c) (d)

Figure 15: Impossible and indeterminate perspective. (a) Three-Stick Clevis (Schuster, 1964). The spatial contradictions cannot be observed within a single fixation. (b) Short three-stick clevis, viewable in a single fixation. (c) A realistic depiction, with many impossible elements only visible on close inspection. *Satire on False Perspective*, William Hogarth, 1754. (d) On their own, crops of the image appear like geometrically-plausible pictures.

perspectives in Figure 15 can be cropped into smaller, realistic pictures, each with valid local shape interpretations, e.g., Figure 15(d). Object recognition can change during viewing, as in bistable images and hidden images (Figure 16), at which point shape percepts can change.

It is useful to contrast Shape Locality with an example of a non-local perception. An object may look larger or smaller depending on the scene around it, both in absolute terms, and relative to other scene objects, as in the Ponzo illusion, Figure 17. In the Ponzo illusion, multiple objects are perceived as having identical shapes but different scales. Scale cues for an object can include local properties (familiar objects) and global properties (the object's spatial relationship to other objects, and defocus blur).

Combining Shape Locality, the Direct View Condition, and the nature of vision-at-a-glance leads a more general hypothesis, which I call **Fixation-Centered Perspective** (Figure 11(b)):

> In each fixation, a picture is interpreted in terms of a linear perspective projection, with the principal point located at the fixation. The effective center-of-projection of this projection may depend on the viewing conditions. When the interpreted shape is inconsistent with prior knowledge of the shape or shape class, the shape is perceived as distorted.

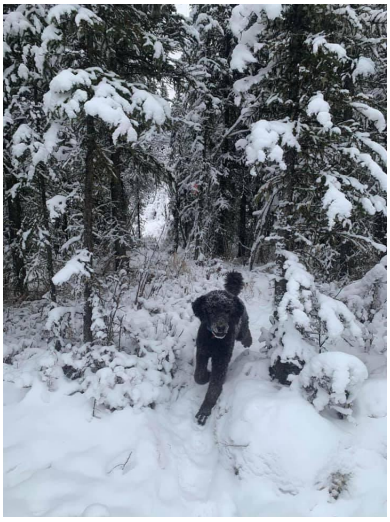


Figure 16: Exceptions to the local perspective principles occur only when object recognition shifts, such as in bistable and hidden imagery. One may not immediately recognize the shapes in these pictures; when they are recognized as faces, shape interpretation changes.

There are different possibilities for what the effective center-of-projection is. It may be that the viewer always interprets the picture region using their own location as the effective center-of-projection. Or, they may perform local slant compensation (Vishwanath et al., 2005), and interpret the region contents as though viewing them from in front of the fixation. The topic of compensation has been controversial in the vision science literature, and is itself a complex topic.

Fixation-Centered Perspective could be a consequence of how the vision system adapts real-world vision-at-a-glance to pictures. When fixating on a new picture, human vision "knows" only the contents of a fixation, and must make an interpretation at each glance. It treats the region around the fixation like a picture that simulates real-world appearances with linear perspective. As a viewer's eye moves over a picture, each fixation has its own perspective, hence the effectiveness of multiperspective techniques in art and photography.

The main hypotheses that I've presented—Shape Locality, the Direct View Condition, and Fixation-Centered Perspective—are compatible with the existing evidence about shape and spatial perception in pictures. This evidence includes: 1. the remarkable successes of linear perspective as a projection technique, 2. the perceived distortions produced by linear perspective, 3. vision-at-a-glance and foveal vision, which show that viewers must infer perspective from the limited information in each fixation, 4. the inconsistent nature of real-world 3D vision, 5. the effectiveness of multiperspective and content-aware projections, at least, locally, 6. pictures on slightly curved surfaces, since picture regions need only be locally flat, and 7. the partial 3D perception when viewing inconsistent or impossible perspectives. No other theory that I'm aware of attempts to describe these disparate phenomena.

7 Global Pictorial Projection Perception

Pictures as a whole may depict space in many different ways (Figure 1)—whether strict linear perspective, a more freeform arrangement, ambiguous semi-abstract imagery, or even impossible perspective—and the visual system can extract some spatial information from each. How does picture perception work for so many different types of projection?

The different perspectives as one moves one's eyes over a picture imply conflicting 3D space interpretations. But the vision system may not need to resolve these inconsistencies.

As previously discussed, real-world vision does not maintain a detailed, consistent 3D model of the visual

(a)

(b)

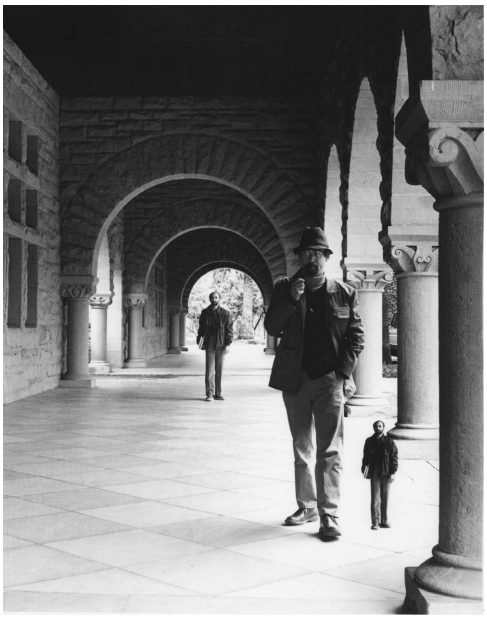


Figure 17: Ponzo illusion, illustrating that object size can depend on context: the three cars appear to have the same shape, but not the same size, even though they comprise identical sets of pixels in the picture. ((a) by Alex Blouin, with annotation by Paul Linton. (b) © The Exploratorium. All rights reserved. Used and adapted with authorization. The Exploratorium is a registered trademark of The Exploratorium, http://www.exploratorium.edu)

world. Hence, pictorial perception need not either. Instead, each view may present its own sense of space, informed by high-level information from adjacent views, consistent enough to appear part of a coherent scene. Only an abstracted 3D interpretation would be maintained over time, representing concepts like general object position and shape, and relationships between objects, rather than a detailed 3D reconstruction.

This idea suggests a possible answer to the question "How can a flat picture provide an illusion of 3D space?" I answer that pictures do not convey consistent 3D information like the real world, but, instead, that real-world 3D vision is far less consistent than it seems. If vision does not aim to reconstruct 3D from the real-world that is consistent across fixations, then there is no reason for it to do so from pictures. In order to provide some illusion of space, a picture merely needs to provide plausible pictorial cues for each fixation.

8 Conclusion

How do viewers understand shape and space in pictures? I have argued that each fixation in a picture produces a 3D percept based on a local perspective. Over multiple fixations, a viewer builds up a 3D scene interpretation from each fixation, but the representation is abstracted, rather than representing detailed 3D shape. The same process, generally speaking, applies both to real-world 3D vision and to viewing pictures, since pictures exist in the 3D world and are processed by the same visual system.

These hypotheses suggest new ways to think about related questions, such as what happens when an artist draws a picture, since local perception and working memory must play significant roles in this process. They also suggest a new way to think about the question of "what is a picture?" The local and global elements of pictures provide the elements of a "language" of pictures (Greenberg, 2021): the rules of perspective for local regions, how artists may distort shape locally, and how they may arrange objects and regions spatially. Some aspects of this language vary in different cultures and styles, and some derive from biological vision. These ideas suggest many avenues for further experimentation and discussion.

Acknowledgments

The author is indebted to Ruth Rosenholtz for invaluable discussions and extensive feedback on paper drafts that provided much wisdom and guidance. Thanks to Robert Pepperell for many inspiring and useful discussions, feedback, pointers, and encouragement. Thanks to Daniel Martin for thorough proofreading and discussions, and to Pietro Perona for detailed comments on a draft. Thanks to Andrew Adams, Elena Adams, David Fisher, Paul Linton, Daniel Martin, Stijn Oomes, Victoria von Ehrenkrook, and Bryan Russell for help with figures, and to Martin Banks, Stephen DiVerdi, Alyosha Efros, Casper Erkelens, Hany Farid, Roland Fleming, Gabriel Greenberg, Martin Kemp, Jitendra Malik, Rich Radke, and Maarten Wijntjes for discussions.

References

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

Agrawala, M., Zorin, D., and Munzner, T. (2000). Artistic multiprojection rendering. In *Eurographics Workshop on Rendering Techniques*, pages 125–136. Springer.

Badki, A., Gallo, O., Kautz, J., and Sen, P. (2017). Computational zoom: A framework for post-capture image composition. *ACM Trans. Graph.*, 36(4).

Carroll, R., Agrawala, M., and Agarwala, A. (2009). Optimizing content-preserving projections for wide-angle images. *ACM Trans. Graph.*, 28(3).

Cooper, E. A., Piazza, E. A., and Banks, M. S. (2012). The perceptual basis of common photographic practice. *Journal of vision*, 12(5):8–8.

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Co.

Durand, F. (2023). Family in a box. https://www.thecomputationalphotographer.net/2023/01/family-in-a-box/.

Elkins, J. (1994). *The Poetics of Perspective*. Cornell University Press.

Fei-Fei, L., Iyer, A., Koch, C., and Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1):10–10.

Gayford, M. (2022). David hockney: space explorer. In *Hockney's Eye: The Art and Technology of Depiction*. Paul Holberton.

Greenberg, G. (2021). Semantics of pictorial space. *Rev.Phil.Psych.*

Greene, M. R. and Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, 20(4):464–472.

Hecht, H., Schwartz, R., and Atherton, M., editors (2003). *Looking Into Pictures: An Interdisciplinary Approach to Pictorial Space*. MIT Press.

Hockney, D. (2006). *Secret Knowledge: Rediscovering the Lost Techniques of the Old Masters*. Viking Studio, second edition.

Kemp, M. (1990). *The Science of Art: Optical themes in western art from Brunelleschi to Seurat*. Yale University Press.

Kemp, M. (2022). Seeing through perspective. In Gayford, M., Kemp, M., and Munro, J., editors, *Hockney's Eye: The Art and Technology of Depiction*. Paul Holberton.

Koenderink, J., van Doorn, A., de Ridder, H., and Oomes, S. (2010). Visual rays are parallel. *Perception*, 39(9):1163–1171.

Koenderink, J., van Doorn, A., Pinna, B., and Pepperell, R. (2016). On right and wrong drawings. *Art & Perception*, 4:1–38.

Koenderink, J. J., van Doorn, A. J., Kappers, A. M., Doumen, M. J., and Todd, J. T. (2008). Exocentric pointing in depth. *Vision Research*, 48(5):716–723.

Kubovy, M. (1986). *The Psychology of Perspective and Renaissance Art*. Cambridge University Press.

Langbehn, E., Steinicke, F., Lappe, M., Welch, G. F., and Bruder, G. (2018). In the blink of an eye: Leveraging blink-induced suppression for imperceptible position and orientation redirection in virtual reality. *ACM Trans. Graph.*, 37(4).

Linton, P., Morgan, M. J., Read, J. C. A., Vishwanath, D., Creem-Regehr, S. H., and Fulvio, D. (2022). New approaches to 3d vision. *Phil. Trans. R. Soc. B*, 378.

Noë, A. (2002). Is the visual world a grand illusion? *Journal of consciousness studies*, 9(5-6):1–12.

O'Regan, J. K. and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24(5):939–973.

Pepperell, R. and Haertel, M. (2014). Do artists use linear perspective to depict visual space? *Perception*, 43(5):395–416.

Perona, P. (2013). Far and yet close: Multiple viewpoints for the perfect portrait. *Art & Perception*, 1(1-2):105–120.

Roman, A., Garg, G., and Levoy, M. (2004). Interactive design of multi-perspective images for visualizing urban landscapes. In *IEEE visualization 2004*, pages 537–544. IEEE.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annu. Rev. Vis. Sci.*

Rosenholtz, R. (2020). Demystifying visual awareness: Peripheral encoding plus limited decision complexity resolve the paradox of rich visual experience and curious perceptual failures. *Atten Percept Psychophys.*

Schuster, D. H. (1964). A new ambiguous figure: A three-stick clevis. *The American Journal of Psychology*, 77(4):673–673.

Shih, Y., Lai, W.-S., and Liang, C.-K. (2019). Distortion-free wide-angle portraits on camera phones. *ACM Trans. Graph.*, 38(4).

Simons, D. J. and Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, 5(4).

Sun, Q., Patney, A., Wei, L.-Y., Shapira, O., Lu, J., Asente, P., Zhu, S., Mcguire, M., Luebke, D., and Kaufman, A. (2018). Towards virtual reality infinite walking: Dynamic saccadic redirection. *ACM Trans. Graph.*, 37(4).

Verstegen, I. (2010). A classification of perceptual corrections of perspective distortions in renaissance painting. *Perception*, 39(5):677–694.

Vishwanath, D., Girshick, A. R., and Banks, M. S. (2005). Why pictures look right when viewed from the wrong place. *Nature neuroscience*, 8(10):1401–1410.

Willats, J. (1997). *Art and Representation: New Principles in the Analysis of Pictures*. Princeton University Press.

Wolfe, J. M., Kosovicheva, A., and Wolfe, B. (2022). Normal blindness: when we look but fail to see. *Trends in Cognitive Science*, 26(6).

Zorin, D. and Barr, A. H. (1995). Correction of geometric perceptual distortions in pictures. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, pages 257–264.