Toward a theory of perspective perception in pictures

Aaron Hertzmann Adobe Research, San Francisco, CA, USA

This paper reviews projection models and their perception in realistic pictures, and proposes hypotheses for three-dimensional (3D) shape and space perception in pictures. In these hypotheses, eye fixations, and foveal vision play a central role. Many past theories and experimental studies focus solely on linear perspective. Yet, these theories fail to explain many important perceptual phenomena, including the effectiveness of nonlinear projections. Indeed, few classical paintings strictly obey linear perspective, nor do the best distortion-avoidance techniques for wide-angle computational photography. The hypotheses here employ a two-stage model for 3D human vision. When viewing a picture, the first stage perceives 3D shape for the current gaze. Each fixation has its own perspective projection, but, owing to the nature of foveal and peripheral vision, shape information is obtained primarily for a small region of the picture around the fixation. As a viewer moves their eyes, the second stage continually integrates some of the per-gaze information into an overall interpretation of a picture. The interpretation need not be geometrically stable or consistent over time. It is argued that this framework could explain many disparate pictorial phenomena, including different projection styles throughout art history and computational photography, while being consistent with the constraints of human 3D vision. The paper reviews open questions and suggests new studies to explore these hypotheses.

Introduction

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

How does a viewer interpret shape and space in a photograph or a painting? The information in a picture is ambiguous. Yet, realistic pictures convey to viewers a sense of spatial arrangements of objects and scenes, including the shapes of scene elements, along with their relative sizes, positions, and distances from the viewer.

The term *projection* describes the geometric relationship between the elements of a realistic two-dimensional (2D) picture and an implied three-dimensional (3D) world. Artists throughout history have used many different approaches to projection (Figure 1), some of which can be codified into projection systems (Willats, 1997). Most famously, linear perspective is celebrated and taught as an artistic technique (Kemp, 1990); it is the default option in many consumer cameras and computer graphics systems. Many other systems have been developed, such as the parallel perspectives in ancient Chinese scroll painting, for example, Figure 1b, and various curvilinear perspectives (Kemp, 1990; Zorin & Barr, 1995; Koenderink, van Doorn, Pepperell, & Pinna, 2016a). Artists do not necessarily reason about projection, and many artists use freeform techniques that cannot easily be described by a simple system or formula, such as Figure 1f. Arguably, strict adherence to any set of rules is rare in classical art; as early as the Renaissance, many great artists in history rejected linear perspective, even after achieving mastery of it (Kemp, 1990; Willats, 1997; Koenderink et al., 2016a; Kemp, 2022).

Picture projections—including linear perspective—span a vast literature across the arts, sciences, engineering, history, architecture, and philosophy (Kemp, 1990; Elkins, 1994). The topic has new relevance with modern computational photography, because smartphone camera apps make nonlinear projections widely available. Hence, there is a need for perceptual theories and studies to explain why these methods work, to help guide the design of better tools for artists and photographers, and to inform societal concerns about photographic digital manipulation (Fried, Jacobs, Finkelstein, & Agrawala, 2020). Conversely, computational photography offers new tools for systematic study of projection.

What assumptions about projection, if any, govern viewers' perception of pictures? At present, no paradigm offers a compelling explanation for how viewers perceive 3D shape and space in pictures, accounting for both linear and nonlinear projections. Existing theory and studies of shape and space in pictures focus on linear perspective imagery created with a single center-of-projection (COP) (e.g., Pirenne, 1970; Kubovy, 1986; Hecht, Schwartz, & Atherton, 2003), with rare exceptions. But analysis solely in terms of linear perspective cannot provide understanding of the wide range of projections in art history

Citation: Hertzmann, A. (2024). Toward a theory of perspective perception in pictures. *Journal of Vision*, *24*(4):23, 1–41, https://doi.org/10.1167/jov.24.4.23.

https://doi.org/10.1167/jov.24.4.23 Received November 2, 2023; published April 25, 2024 ISSN 1534-7362

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.

(Gombrich, 1961; Willats, 1997), the ways that "correct" linear perspective creates misperceptions (e.g., Vishwanath, Girshick, & Banks, 2005; Koenderink, van Doorn, de Ridder, & Oomes, 2010; Bryan, Perona, & Adolphs, 2012; Cooper, Piazza, & Banks, 2012), or the perceptual advantages of some nonlinear projection algorithms (Seitz & Kim, 2003; Zelnik-Manor, Peters, & Perona, 2005; Agarwala, Agrawala, Cohen, Salesin, & Szeliski, 2006; Carroll, Agrawala, & Agarwala, 2009; Perona, 2013; Shih, Lai, & Liang, 2019).

This paper proposes hypotheses describing 3D projection perception in pictures. The paper begins by reviewing relevant evidence from art history, distortion phenomena in linear perspective, vantage-point compensation studies, artistic and computational projection techniques, foveal vision, vision-at-a-glance, and 3D vision. Then, informed by this review, the paper proposes several connected hypotheses that could apply, to varying degrees, to any picture depicting a realistic scene, even relatively abstracted ones (such as those in Figure 1).

The first set of proposed hypotheses state that viewers understand the 3D shape and structure within any small region of a picture independent of the rest of the picture, with very specific exceptions. When a viewer fixates on a picture, they interpret shape and space around that fixation point, primarily in a radius related to the size of the fovea. This 3D interpretation does not change after subsequent fixations in a picture, regardless of the content around the region (Figure 2), except after high-level changes in object recognition, such as in bistable imagery. The size of this region may vary with stimuli, without any precise cutoff, but, for sake of discussion, one may consider a 6 degree visual angle, or, equivalently, a 6 cm diameter in a picture viewed at a 60 cm distance. As David Hockney (2006) has pointed out, cropping individual people and objects from classical paintings produces smaller pictures that,

Figure 2. The 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, from Figure 15a; the middle and right images were generated using Adobe Photoshop "Generative Fill.").

Figure 3. Multiperspective collages constructed with multiple vanishing points, illustrating how plausible-looking pictures can be constructed from collages 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.

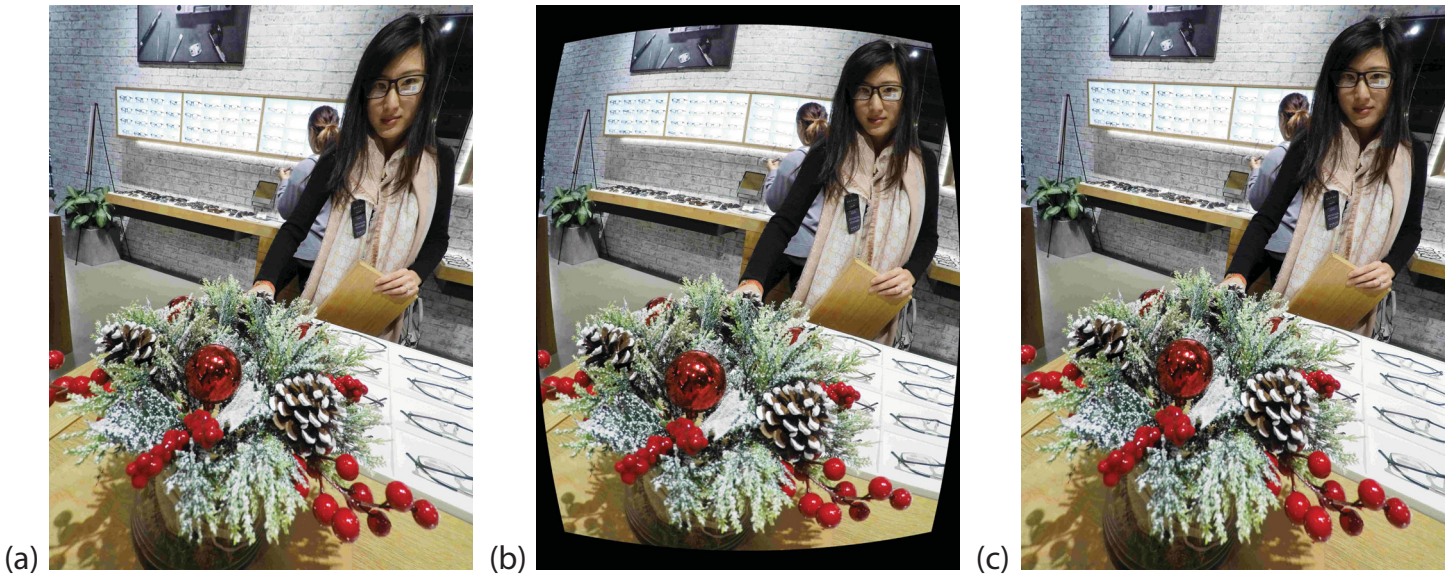


Figure 4. Comparison of photographic perspectives, from Shih et al. (2019). (a) Linear perspective. All straight lines in the world project to straight lines in the image. The spheres and the face in the image margins appear distorted, except when viewing monocularly from the COP. (b) Stereographic perspective. Spheres appear circular, and the face appears undistorted, but straight lines in the world are curved in the image. (c) Content-aware perspective, for correcting face distortion (Shih et al., 2019). The algorithm detects faces and applies a hybrid projection, with stereographic perspective for faces and linear perspective elsewhere. The method preserves straight lines and decreases face distortion, but the spheres in the lower-right corner are still distorted. A version of this method runs on the Google Pixel camera app (see Figure 22).

Figure 5. Raphael's *The School of Athens* uses precise one-point linear perspective for the architecture, but not for the people. None of the faces seem to be distorted as they would be in a true wide-angle linear perspective image, such as in Figure 4a. According to the projection implied by the architecture, the spheres in the lower-right corner should have the aspect of ratio 1.2:1 (Zorin & Barr, 1995), visualized here with a magenta ellipse.

themselves, look like realistic paintings, without any change to the depicted shapes.

What shape assumptions does a viewer bring to the region around a fixation? I argue that shape and space are interpreted according to a linear perspective projection centered on the fixation, which I call *fixation-centered perspective*. Viewers perceive distortion when a shape is depicted inconsistently with this locally linear perspective. This hypothesis fits evidence from compensation studies (Vishwanath et al., 2005), artists' techniques (Kubovy, 1986; Kemp, 1990), and many modern computational photography techniques (Roman, Garg, & Levoy, 2004; Zelnik-Manor et al., 2005; Agarwala et al., 2006; Badki, Gallo, Kautz, &

Sen, 2017) (e.g., Figures 3 and 4). For example, this hypothesis specifies that, to avoid perceived distortion, straight lines in the world should always be straight in pictures, and spheres should be circular. Famously, in *The School of Athens*, Raphael drew spheres as circles, where strict linear perspective would dictate ellipses (Figure 5). Raphael's faces are not skewed either, as they would be in strict linear perspective (Figure 4a).

How does a viewer *globally* combine fixations to interpret 3D in an entire picture? One possibility is that vision aims to infer 3D by "inverse projection" (Juricevic & Kennedy, 2006). However, if human vision aimed to infer a veridical 3D scene from every picture, then many artistic styles might seem incomprehensible when they are ambiguous, inconsistent, or do not follow any apparent strict projection rules. In fact, recent theories of 3D vision suggest that the visual system does not aim to infer coherent 3D shape from the real world (Linton et al., 2022), and so we would not expect picture perception to do so either.

Instead, this paper proposes a two-stage model for 3D human vision. In pictures, the first stage perceives a 3D shape for the current gaze, according to the principles described above. The second stage continually integrates some of the per-gaze information into an overall interpretation of a picture. As a viewer moves their eyes over a picture, they build up an abstracted understanding of the *global* properties of the picture, including, for example, the spatial relationships of the contents of individual picture regions. These relationships may be more specific or more vague, depending on how representational or abstract the visual style is; one even gleans partial relationships from pictures of impossible shapes, such as M. C. Escher's *Belvedere*.

These topics directly relate to the question, "What is a picture?" Some philosophers and art historians have claimed that pictures are a purely cultural phenomenon, using symbolic systems learned like written language (Panofsky, 1927; Goodman, 1968). Conversely, some have treated drawing as a direct recording of mental imagery (Cohn, 2012; Chamberlain & Wagemans, 2016), akin to the commonplace idea that pictures literally show what the artist perceived. Each of these theories fails to explain important pictorial phenomena. Many authors (e.g., Gombrich, 1961; Fan, Bainbridge, Chamberlain, & Wammes, 2023), have instead described pictures as a combination of perceptual factors, on one hand, and stylistic, linguistic, and cultural elements on the other. The hypotheses proposed in this paper provide a possible decomposition of how these elements manifest in pictures: local, per-fixation perception follows real-world 3D shape perception, whereas larger-scale composition is much more flexible. These could be some of the elements of the "language of pictures" (Cohn, 2012; Greenberg, 2021).

Finally, the paper reviews open questions, including those posed by the predictions of the new models here, and suggests new studies to explore these hypotheses.

Linear perspective

Linear perspective has formed the foundation of many theories and studies, including the ones proposed in this paper.

A linear perspective projection is defined by a COP, a view direction, and an image plane (Figure 6b). The view direction is a vector from the COP that intersects the image plane at the *principal point.* Normally, this vector is perpendicular to the image plane, and the principal point is at the image center. The brightness and color of any point on the image plane is determined by the light along the ray through the point toward the COP.

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 COP (Figure 6b), the retinal image should simulate viewing the depicted scene.

This model can be implemented directly with a pinhole camera, in the limit of an infinitely small aperture. Most consumer camera lenses aim to approximate linear perspective (Figure 6a), apart from effects like defocus blur and bokeh. Linear projection is widespread in computer graphics and vision algorithms, often representing the main or only camera option (in addition to orthographic projection).

Filippo Brunelleschi devised one-point perspective around 1413, and Leon Battista Alberti first published it and its derivations in 1435 (Kemp, 1990), possibly inspired by the treatises of ancient Roman architect Vitruvius (Tyler, 2015). Over the centuries, "perspective" grew to comprise a varied collection of geometric constructions for working with straight-edges and brushes to depict complex architectural elements like colonnades and gabled arches. The techniques of one-point and two-point perspective are familiar to many artists and art students today. However, some art historians came to treat it as much more: a rigid set of rules for how to make pictures correctly (Elkins, 1994; Verstegen, 2010), a process that art historian Elkins (1994) called "the fossilization of perspective."

Yet, it is rare that artists strictly follow the "rules" of perspective (Kemp, 1990; Verstegen, 2010; Pepperell & Haertel, 2014; Koenderink et al., 2016a). In one survey of classical paintings, Kemp (2022) found that only a tiny minority followed geometric perspective constructions; instead most used them as "a working tool that delivered convincing results when used in

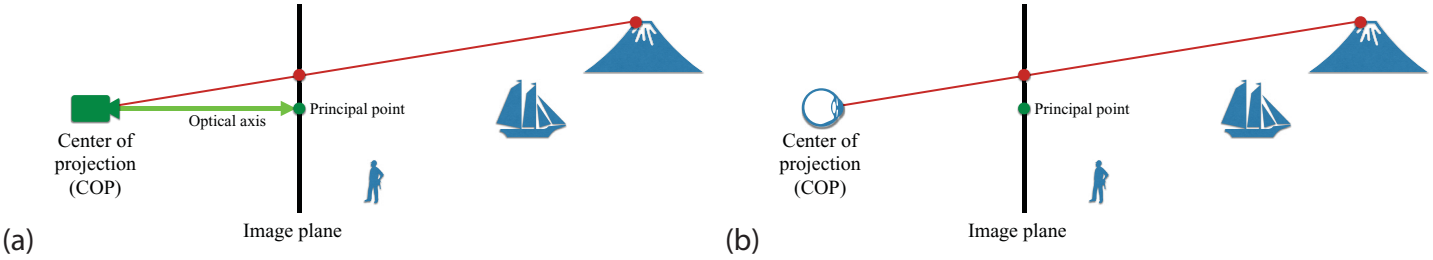


Figure 6. Linear perspective geometry. (a) Linear perspective imaging is defined by an image plane, a COP, and an optical axis (view direction). The principal point is the intersection of the optical axis with the image plane. The color at an image point is determined by the light incoming to the COP along the ray from the image point. (b) When viewing the picture from the 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.

a pragmatic manner, without following the rules slavishly." Many famous painters, including Leonardo da Vinci, J. M. W. Turner, and David Hockney (Kemp, 1990; Kemp, 2022), achieved proficiency in strict linear perspective, and each later wrote about its shortcomings, while exploring more flexible approaches to perspective.

Similarly, discussions of perspective and pictorial space in the perception literature usually assume linear perspective in some way (e.g., Pirenne, 1970; Hecht et al., 2003; Vishwanath et al., 2005; Cooper et al., 2012). Many such studies focus solely on linear perspective stimuli with well-defined COPs, without discussing any other types of projection. Some authors even define pictures themselves in terms that presuppose linear perspective: "A picture acts like a window into a virtual world; it is a frozen cross-section of light to a fixed viewpoint..." (Yang & Kubovy, 1999). Yet, in the words of Koenderink et al. (2016a), "The case for the 'rightness' of linear perspective would not only be stronger if more artists had used it but if fewer great artists had outright rejected it."

Many authors have written about the problems of linear perspective, either as a description of how artists work or as a description of the human visual system's assumptions about pictures. However, no other theory has convincingly offered to replace it in either role.

Vantage point and distortion in wide-angle linear perspective

The derivation of linear perspective assumes that a picture is viewed monocularly from its COP. 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. For example, the reader may consider their own behavior when viewing the pictures in this paper. Hence, the visual system does not *strictly* assume linear perspective with the COP at the viewer's vantage point.

This paper primarily focuses on *wide-angle pictures,* which are very common, yet they cause the most trouble for perceptual theories based purely on conventional linear perspective. A linear perspective picture is wide-angle if it uses a much wider field-of-view (FOV) than the FOV it would normally be viewed with. In other words, the picture's COP is much closer than typical viewing distance (Figure 7a). In photography, a focal length of 35 mm or less (for 35 mm sensor) is considered wide-angle, whereas 50 mm is considered a good focal length for "natural-looking images" (Cooper et al., 2012).

Wide-angle pictures are widespread 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 and 5). Smartphones take wide-angle photos *by default*, despite the fact that these pictures cannot normally be viewed from their COPs on smartphone displays, due to the extremely short viewing distances involved.

Viewing a wide-angle picture from the COP is rare and even uncomfortable (Koenderink et al., 2016a). For example, in Figure 8a, the viewing distance should be approximately 40% of the image width.[1] That is, if the image appears printed or on the screen as 3.5 inches wide, the viewer's eye should be 1.4 inches from the center of the picture. This is very a unusual viewing position, and some peoples' eyes cannot even focus at this distance.

When given a choice, viewers typically choose viewing distance based on picture size, not COP (Cooper et al., 2012). In other cases, such as billboards

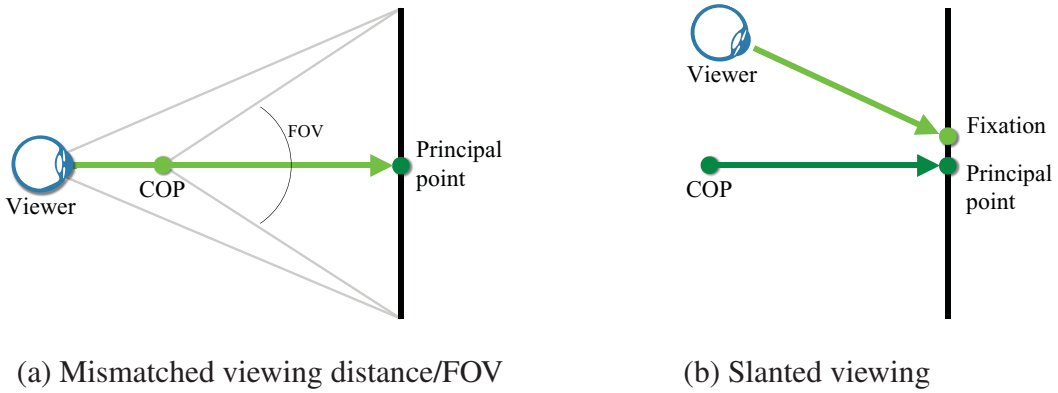


Figure 7. Two ways to view a linear perspective picture when not viewing from the COP. (a) The viewer is behind the COP, but still viewing on the same optical axis. This is typical for wide-angle photography. In this case, the FOV of the picture is wider than the viewer's FOV. (b) Slanted viewing, where the viewer views the picture at an angle, and not necessarily at the COP distance, is the most general case.

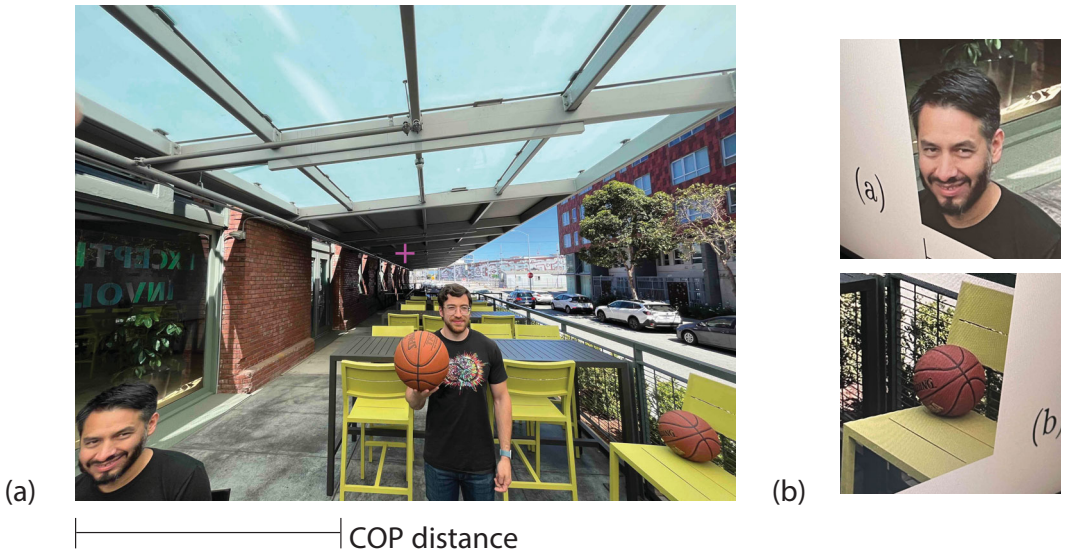


Figure 8. Wide-angle photo for experiencing marginal distortion in COP viewing. (a) Photo taken with an iPhone 13 in ultrawide mode (0.5×, 14 mm). The magenta cross indicates the picture center. COP distance is shown below the picture; it is 40% of the width of the picture. To view from the COP, place 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.

along the road, viewing distance is determined by circumstance.

Distortion: When things "look wrong"

As a clue to visual assumptions, this paper studies *perceived distortions*: when a viewer recognizes that a shape "looks wrong," such as in marginal distortions. In the literature, "distortion" often denotes optical phenomena that deviate from linear projection. This paper does not use the word in this sense, because this paper is concerned solely with how pictures are perceived.

A key concept here is whether objects and faces "look right": does the depiction seem to accurately convey shape, or does the shape appear deformed? People seem to have an intuitive notion of what

Figure 9. 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.

shapes "look right" or "wrong." When Vishwanath et al. (2005) performed experiments asking subjects to judge whether a sphere looked "too wide" or "too narrow" in various viewing conditions, they found it noteworthy that subjects never asked for instructions about how to make these judgements, such as whether they should imagine viewing from a more-canonical position (M. Banks, personal communication, 2022). Likewise, marginal distortions have been discussed as a problem throughout art history; it is understood that they do not "look right."

Hence, perceived distortions provide important clues to the vision system's assumptions about how shapes "should be" depicted. When a normal object's projection looks distorted, the projection does not match viewers' assumptions. Some projections match viewers' assumptions, and some do not. Perceived distortions are considered undesirable in highly-realistic painting and photography, but artists often use them deliberately in other styles, such as, expressionist art and artistic photography (Figure 9).

Perceived distortions give clues to the nature of mental representations. As a metaphor, consider the appearance of a spoon in a glass of water, as in Figure 10. Seen in real life, the spoon looks broken and bent, but we can understand that it is a normal spoon undergoing refraction. Likewise, a picture displays a misleading shape appearance, but, if a viewer knows the normal shape of that object, they can recognize the distortion. An unfamiliar object in the glass would give a confusing or misleading percept. There are several simultaneous mental representations here, corresponding with a glass of water, a deformed spoon, and an undeformed spoon.

Figure 10. 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.

Likewise, perceived distortion suggests the same set of mental shape representations: the apparent shape—which may look "right" or "wrong"—and some knowledge of the actual shape.

Perceived distortions in wide-angle pictures

Two important classes of perceptual distortions in wide-angle linear perspective are widely studied.

Marginal distortions

Wide-angle linear perspective causes objects in the periphery to appear distorted, a phenomenon known as *marginal distortion*. Figures 4a and 8a show examples where spheres and faces seem to be oblong in the corners of photographs. Kubovy (1986) (Ch. 7) reviews experiments aimed at determining the FOVs that produce marginal distortion.

Note that a viewer is not necessarily aware of marginal misperceptions due to linear perspective, as vividly demonstrated by Koenderink et al., 2010 (see Figure 11). The main difference between this case and the marginal distortion of spheres is whether prior

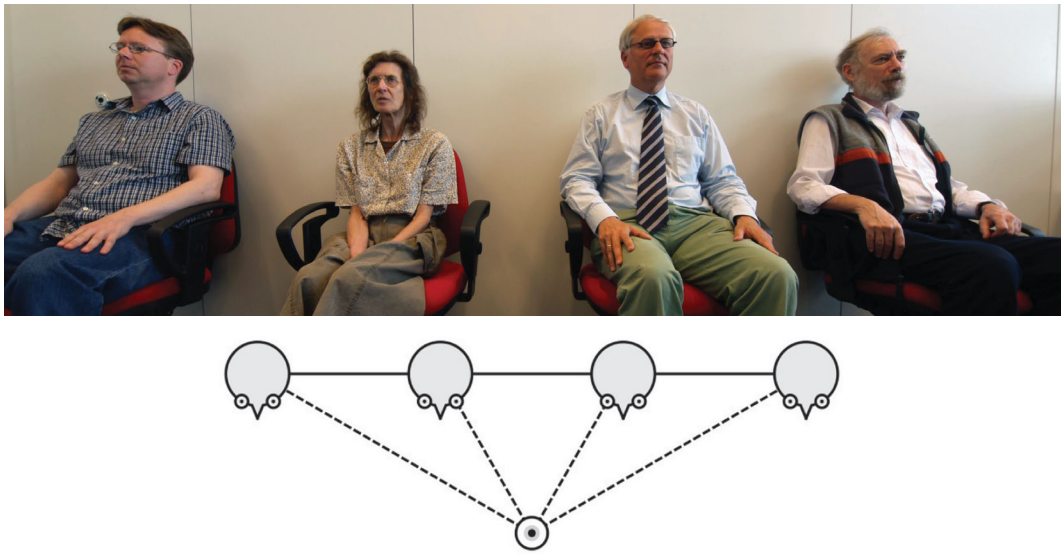


Figure 11. Wide-angle linear perspective photograph, from Koenderink et al. (2010), taken with a 14 mm equivalent lens and cropped. The four individuals face in parallel directions, illustrated in the diagram. Yet, under normal viewing conditions, they seem to be facing in diverging directions in the photograph.

knowledge allows a viewer to recognize a misleading depiction of a familiar object.

Raphael's *The School of Athens* (Figure 5) provides a particularly famous example (Kubovy, 1986) that does not strictly follow linear perspective. Raphael used strict one-point perspective for the architecture. However, 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.

Moreover, none of the faces in *The School of Athens* exhibit marginal distortion, that is, compare the faces with Figure 4a. Large scenes with many faces are common in art, such as in Figures 1 and 5. *But, in the entire history of painting, I am unaware of any face depicted with the marginal distortions dictated by linear perspective.*

This illustrates how rarely artists strictly obey linear perspective for wide-angle depictions with people. Instead, classical painters often used linear perspective to construct architecture like a stage set, and then moved the people around on it freely (M. Kemp, personal communication, 2022).

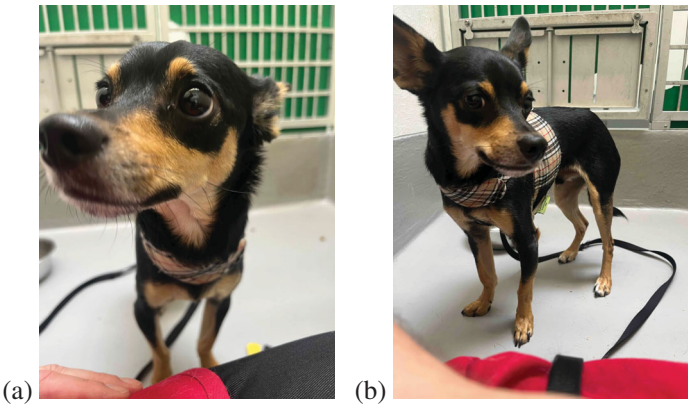


Figure 12. Perspective expansion, illustrated by two wide-angle photos taken moments apart, with the same camera settings (iPhone 13, 1×, 26 mm). (a) In the close-up, the dog's snout appears as big as its body. (b) Increasing the camera's distance to the subject reduces the perceived distortion. For both pictures, COP viewing distance is 75% of image height, e.g., if the picture is displayed 2 inches tall, then COP viewing requires placing the viewer's eye 1.5 inches from the center of the picture.

Perspective compression and expansion

Photos taken with extremely wide or narrow FOVs (equivalently, long or short focal lengths) produce compression or expansion effects, sometimes called "perspective distortion" (Cooper et al., 2012). Figure 12a illustrates a depth-expansion effect of wide-angle close-up photography: the COP distance is much closer than typical viewing distance, exaggerating the size of the dog's snout. This expansion only appears in the close-up view, as seen by comparison with Figure 12b. A viewer can often recognize such shapes as distorted, i.e., that this is a normal dog photographed with an extreme perspective, rather than a dog with a giant snout.

Conversely, telephoto photography (long focal length/narrow-FOV) produce a compression of space that approximates orthographic projection. Pictures

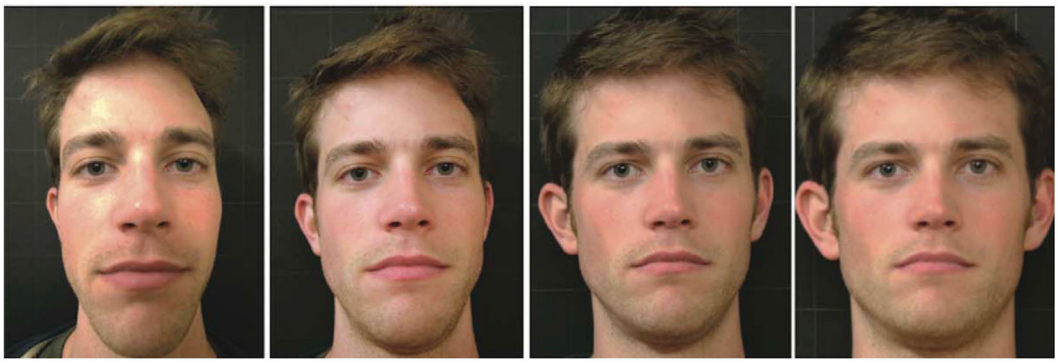


Figure 13. Facial appearance perception depends on COP distance, from Cooper et al. (2012). These four photographs of the same person were taken with four separate focal lengths: 16, 22, 45, and 216 mm (from left to right). Camera distance was kept proportional to focal length, thus keeping constant the subject's interocular distance in the picture.

made this way do not contain the perspective cue that more-distant objects should be smaller in the picture, and can create misleading senses of spaces and distance. For example, early in the COVID pandemic, some news outlets used telephoto pictures to tell misleading stories of people failing to obey social distancing rules (Rachwani, 2020).

Figure 13 shows a vivid demonstration of perspective compression and expansion. Multiple photographs of the same individual were taken, with the camera's distance to the subject and focal length varied together to preserve the size of the face in the picture (Perona, 2007; Cooper et al., 2012). A viewer interprets each picture as having the same scale, but with a different face shape. Bryan et al. (2012) found that these variations affect perception of the subject's personality, for example, with portraits taken at a shorter distance, participants described the subject with more benevolent attributes ("good," "approachable," "trustworthy"), whereas more distant photographs produced attributes like "smart" and "strong" (Perona, 2013).

Vantage-point compensation in linear perspective

The fact that viewers do not normally view wide-angle pictures from the COP has long been a problem for perceptual theories based on linear perspective, since it contradicts the idea that pictures work by simulating the light to the viewer's position. Moreover, if viewers always assumed that they were viewing from the COP, then we would expect perceived shape to continually shift, warp, and shear as one moves one's head in front of a picture.

To resolve this problem, *Vantage-Point Compensation* posits that perception is "robust" to the vantage point. Pirenne (1970), who credited this hypothesis to Albert Einstein, formulated it as:

> When the shape and position of the picture surface can be seen, an unconscious intuitive process of psychological compensation takes place, which restores the correct view when the picture is looked at from the wrong position.

That is, the viewer interprets a picture as though viewing from its COP, regardless of their actual vantage-point.

Two classes of cues could inform compensation. First, Pirenne hypothesized that compensation could depend on the position and shape of the picture frame. As evidence, he pointed out that compensation does not occur when viewing through a peephole. Binocular stereopsis also provides cues to surface slant.

Second, pictorial cues could be used for compensation. For example, a viewer may recover the COP from straight lines in the picture. The appearances of familiar objects could provide scale cues, and thus indirect cues to COP distance (equivalently, focal length).

This section briefly surveys studies that test these theories. These studies have exclusively focused on linear perspective imagery, but the results discussed here have broader implications for realistic pictures. But, first, I describe an informal experiment that the reader may perform, in order to experience some of these phenomena.

Viewing from the COP

It is widely assumed that viewing a linear perspective picture from the COP leads to a realistic percept. Yet, Vishwanath et al. (2005) informally point out that marginal distortion appears when viewing a wide-angle linear perspective picture from the COP,

under *binocular* vision. I am unaware of any formal studies of this particular case.

Here is a simple informal experiment that the reader may undertake to test this observation, and to experience several important compensation phenomena.

(1) First, enlarge Figure 8 to fill a large display, such as a 24-inch monitor, in a normally lit room.
(2) Second, identify the COP location, which is in front of the picture's center, at a distance equal to two-fifths of the displayed image's width. The horizontal line in Figure 8a shows this distance.
(3) Third, open the camera app on a smartphone. Hold the phone with the camera sensor located at the COP, and aimed at a corner of the picture, where a sphere or a face is located. The positioning need not be precise. Notice that the marginal distortion disappears in the smartphone display (Figure 8b). Note how the image continually skews as the camera moves around near the COP.
(4) Next, remove the smartphone, and place one of your eyes at or near the COP location. With your other eye closed, fixate on the ball in the corner of the picture. Does the shape look distorted—does the ball's outline look circular or oblong?
(5) Repeat this test with both eyes open. Compare how the ball or the face changes appearance as you open or close one eye.

You should observe that marginal distortion is visible in binocular viewing, but not monocular.

The final step may not work for stereoblind viewers, who are unable to extract depth from binocular stereopsis. A stereoblind colleague reported to me that the monocular and binocular conditions seem to be the same to him.

One may repeat this experiment using any picture with known focal length, using the formula in Footnote 2.

Discussion

There are several key takeaways from this experiment. First, marginal distortion appears in some viewing conditions, and not others. It does not occur during *monocular* viewing from the COP. *But marginal distortion does occur in binocular viewing from the COP.* Second, monocular viewing from the COP is unusual, difficult, or even impossible for wide-angle pictures under many normal display conditions, owing to physical constraints and visual accommodation. It is the exceptional case, not the norm. However, formal studies are needed to rigorously test these claims.

Vantage-point compensation could explain the disappearance of marginal distortion in binocular viewing. This compensation seems to be automatic, rather the result of conscious reasoning. Another possible explanation is that, in binocular viewing, one of the viewer's eyes cannot be close enough to the COP to cancel distortion, but this explanation would predict only subtle marginal distortions.

Marginal distortion is often explained as occurring "because" the viewer is not at the COP (e.g., Kubovy, 1986). This statement, although accurate, downplays just how unusual monocular COP viewing is. Monocular COP viewing creates a forced perspective illusion, more like an Ames Room (1925) than like normal picture viewing.

The above experiments illustrate a role for binocular vision in picture viewing, but roughly 8% of the population are stereoblind, and a larger fraction have poor stereo vision (Bosten et al., 2015; Levi, 2022). Moreover, many people who are stereoblind do not realize it. This raises the question of whether stereoblindness affects viewers' aesthetic experience and/or space perception in realistic pictures, or whether there is no effect in typical viewing conditions. This is a question for future studies to explore.

Studies with straight-line cues

Many studies have tested compensation hypotheses using straight-line renderings, with mixed results. Some studies confirm the effect when the picture surface is visible (Perkins, 1973; Hagen, 1976; Bengston, Stergios, Ward, & Jester, 1980; Rosinski, Mulholland, Degelman, & Farber, 1980), and some do not (Adams, 1972; Bengston et al., 1980; Cutting, 1987; Todorović, 2008). Yang and Kubovy (1999) modify the hypothesis to account for this variability by hypothesizing that the degree of compensation increases with the strength of surface slant cues. Some studies find support for pictorial compensation, and others reject it.

It is worth noting that the studies summarized above employ simple wireframe line renderings as stimuli, without any familiar objects. That is, these renderings contain enough information to recover the COP *only by* reasoning about conventional one-point or two-point perspective of scenes containing parallel and perpendicular lines. The studies described next employ richer stimuli and provide much more compelling evidence.

Slant compensation

Vishwanath et al. (2005) provide a compelling study of compensation based on both pictorial cues and surface slant. Viewers were shown either wireframe or shaded 3D renderings of basic shapes against a checkerboard ground plane (Figure 14a), using a bite bar to precisely control viewing position and angle.

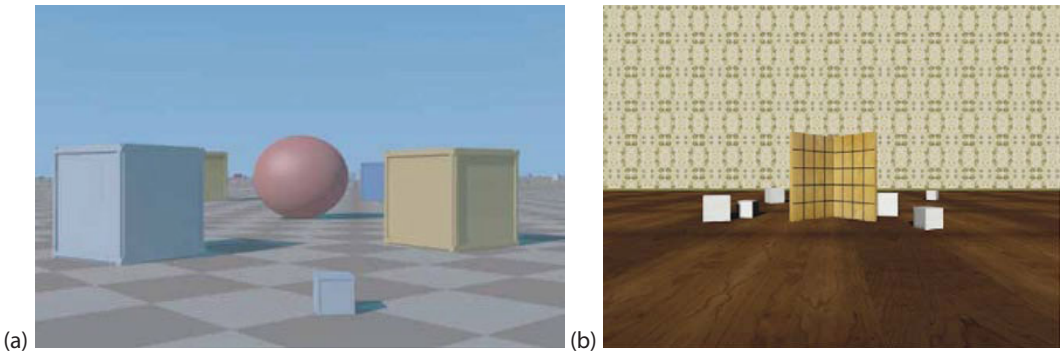


Figure 14. Experimental stimuli. (a) Typical stimulus used by Vishwanath et al. (2005). Under various slanted viewing conditions, participants were asked to judge whether the sphere appears too width or too tall. (b) Typical stimulus used by Cooper et al. (2012). For various picture sizes at a fixed viewing distance, participants were asked to judge whether the angle of the hinge is less than or greater than 90°. In both sets of stimuli, straight lines (checkerboard floor, cube edges) are the primary cue to perspective.

These renderings included sufficient information for estimation of the COP based on parallel and perpendicular lines. Viewers were asked whether ovoids or tilted squares were depicted as too wide or narrow. From the responses, a staircase method was used to determine the aspect ratio that made ovoids appear spherical and tilted rectangles appear square.

In their first experiment, they varied the surface slant and the viewing conditions, providing more or fewer cues to surface slant. In the extreme case with no slant cues—viewing through a peephole without a visible picture frame—viewers did not compensate for slant when judging aspect ratios. However, with visible slant cues, including binocular vision and a visible picture frame, viewers did respond consistently with surface slant compensation. This confirms that viewers compensate for the slant of the display surface when sufficient cues are available.

In their second experiment, they tested whether pictorial cues provide invariance to slant, by rendering pictures with varying off-axis COP locations and slanted view directions, that is, so the picture must be viewed off-axis and slanted (as in Figure 7b) to reproduce the retinal image of the original scene. In an extensive set of experiments removing geometric cues, they found that viewers always responded as though viewing the scene from the central surface normal, rather than from the picture's true COP. From this, they concluded that viewers do not compensate based on pictorial perspective cues.

Finally, they performed a third experiment to determine how slant compensation is performed. Compensation theories would predict that the viewer infers the picture's slant at the center of the picture, and compensates accordingly across the entire image, as though mentally rotating the entire picture to be fronto-parallel. Instead, they found results to be consistent with a local slant hypothesis, which which they phrase as:

> [The] location of the CoP is not recovered. Instead, the observed invariance is due to an adjustment of the retinal-image shape based on measurements of the local slant of the picture surface at the point of interest.

We may call this *local slant compensation*. The authors also point out that this hypothesis predicts that marginal distortion occurs even when viewing binocularly from the COP.

The hypothesis suggests a connection to eye movements, because viewers were, most likely, gazing at the objects when judging their aspect ratios. A natural prediction is, then: if one reproduces the above experiments with an eye tracker, then compensation would be best predicted by the slant at the fixation point relative to eye gaze direction, rather than any pictorial information like object location.

Erkelens (2013b) describes a set of experiments with the opposite conclusion about slant compensation. In these experiments, a skewed wireframe grid is displayed on a slanted screen, and viewers are instructed to align a physical object to the perceived slant of the grid. The perceived slants were predicted well by a virtual stimulus of a grid of parallel lines with the same retinal image as the stimulus. The orientation of the display surface had little effect; that is, viewers responded as though perceiving the grid without any slant compensation, using parallel lines cues to shape. The experimental setup of aligning a physical slant to a perceived slant seems to preclude slant compensation, as does the instruction to the participants to ignore the display's slant. However, Erkelens argues that these principles predict other phenomena, such as the hollow face illusion, and Patrick Hughes' reverspectives (Wade & Hughes, 1999).

Distance compensation

Cooper et al. (2012) performed two experiments to elicit peoples' biases for viewing distance compensation.

First, they displayed computer graphics renderings of hinge shapes on textured backgrounds (Figure 14b). Viewers observed each picture at a fixed distance, controlled with a bite bar. Participants were asked to judge whether each hinge angle was less than or greater than 90°. A staircase method was used to determine, for each viewing condition, the hinge angle that appeared to be perpendicular. This process was repeated for renderings with five different COP distances, to judge the relationship between the picture's COP distance with the viewer's behavior.

The participants behaved as though using their viewing position as the picture's COP. The true COP did not affect their results, that is, the participants did not compensate for viewing distance.

In a second experiment, Cooper et al. (2012) asked participants to select the best viewing distance for various sizes of pictures. The stimuli included photographs and renderings of various familiar types of scenes, at various camera focal lengths. They found that viewers chose their viewing distance based on how much of their FOV was taken up by the picture, and not on pictorial cues. Specifically, viewers preferred that pictures take up 36° of their FOV for all but the smallest prints, for which they preferred 22°. They argued that their experiments together explain common rules of thumb for "normal FOV" photography, and provide more perceptually-based recommendations for focal length.

Likewise, Erkelens (2018) argues against distance compensation for familiar objects. He provides compelling examples for the claim that perceived distance depends only on the size of the object on the display, and not on the size of the picture itself, nor the camera focal length.

Summary: What compensation do viewers perform?

Compelling evidence shows that viewers do not compensate based on estimates of a picture's COP, for example, by estimating vanishing point from straight lines (Vishwanath et al., 2005), by adjusting for viewing distance (Cooper et al., 2012), or by adjusting for picture size and shape (Erkelens, 2018).

Whether or how viewers even perform slant compensation remains more controversial. Slant compensation may occur automatically, or it may be performed by conscious reasoning and mental rotations. The brain may use multiple representations: some behaviors and tasks may depend on uncompensated (retinal) representations, and other may operate on compensated/unslanted ones (Goldstein, 1979; Koenderink, van Doorn, Pinna, & Pepperell, 2016b; Linton, 2017; Morales, Bax, & Firestone, 2020).

Vishwanath et al. (2005) find that viewers perform Local Slant Compensation in normal viewing conditions (Figure 7b): object shape is judged based on the surface slant at the object's location in the picture. Local slant compensation does not require linear perspective; it can apply for any picture. In contrast, Erkelens (2013b) reports experiments in which no slant compensation occurs. One may find other stimuli for which compensation cannot occur, e.g., high-curvature picture surfaces (Cavanagh, von Grünau, & Zimmerman, 2004).

Even in studies that demonstrate slant compensation, whether or not compensation occurs depends on whether sufficient slant cues are available (Perkins, 1973; Yang & Kubovy, 1999; Vishwanath et al., 2005). This, too, is illustrated in the informal COP viewing experiment described earlier (Figure 8). Moreover, the degree of compensation depends on the strength of the cues.

Is Local Slant Compensation automatic? Vishwanath et al. (2005) and Vishwanath (2023) results could have been the result of conscious mental rotations, rather than an automatic, unconscious process. But the viewing-from-the-COP experiment (Figure 8) seems to demonstrate automatic compensation.

A set of following-the-viewer illusions demonstrate uncompensated phenomena. Koenderink et al. (2016b) describe how a frontal facial portrait appears to continually face the viewer as they move around, as though the picture is following the viewer. For example, the famous recruiting posters of Uncle Sam and Lord Kitchener ("I Want You!") seem to point at the viewer, regardless of view angle or knowledge of the unslanted surface. The effect can be very robust and compelling with objects that are not flat pictures, as in the hollow face illusion, and Patrick Hughes' reverspectives (Wade & Hughes, 1999). Nonetheless, these illusions are striking because they are so unusual; slant compensation may still occur in many pictures.

Future experiments with fast presentation times and/or behavioral studies could help to test whether compensation is automatic, although they must somehow deal with the fact that unslanted representations affect task performance even for slanted real-world objects (Morales et al., 2020). There remain several other gaps in the formal experimental studies, including the following. How is shape perceived in other kinds of projections besides linear perspective? How does local slant compensation interact with varying viewing distance for shape inferences? Does stereoblindness or weak stereovision affect compensation, shape perception, marginal distortions, or aesthetic experience?

Nonlinear projections for wide-angle views

The previous four sections surveyed the advantages and limitations of linear perspective as a tool for depiction, as a description of artistic practices, and as a basis for perceptual theories. On its own, linear perspective provides a useful but insufficient model for each of these purposes. Hence, it is necessary to study other projections. This section reviews nonlinear projections that can address some of these shortcomings of linear perspective. I mainly discuss projections of wide-angle views.

The term *projection* here refers to any mapping from 3D space to 2D positions, and *perspective* projections are those that can be defined by rays to a single viewpoint.

Single-viewpoint visual experience

Before discussing different projections, it is worth analyzing visual experience from a single viewpoint in the real world, because many projections aim to capture this experience in some way. Consider the visual experience of standing by the side of a street, as in Figure 15. All conventional projection techniques have significant shortcomings at capturing this experience. Linear perspective with a normal FOV (Figure 15a) does not capture much of the scene. Wide-angle linear perspective preserves straight lines (Figure 15b),

(a) Separate linear perspective photos

(b) Ultrawide linear photo

(c) Curvilinear perspective

(d) Zoom 

Figure 15. Challenges in depicting in a wide-angle scene from a single viewpoint. In real-world experience, straight lines are straight in every view we see, but they change image-space orientation when we rotate our eyes or head. These two facts cannot be captured together by wide-angle perspective, and so any projection system must compromise. Furthermore, foveal vision observes substantially more detail. In real-world viewing, a viewer can easily read distant text by fixating on it, whereas the text is unreadable in these pictures without significant zoom. Photos taken on iPhone 13: (a) Non-wide angle at 52 mm (1.9×), per (Cooper et al., 2012), (b) ultrawide mode (0.5×, 14 mm), (c) panorama mode, (d) linear photo, zoomed (5×, 131 mm).

but produces marginal distortions, which become increasingly extreme as the view widens. Curvilinear perspectives (Figure 15c), can combine different views into a single coherent whole, but do not preserve straight lines.

Why is this experience so difficult to depict?

Some writers have answered that spatial perception itself is naturally curved (Tyler, 2015). For example, Johannes Kepler wrote in 1625 that the natural form of perspective is spherical, not planar. In the seventeenth-century engraving in Figure 17b, Abraham Bosse attributed the challenges of perspective to the distortion inherent in mapping a spherical view to a flat plane. Such answers assume a fixed vantage-point. But, as we have seen, viewers understand pictures well from many vantage-points.

Panofsky (1927) argued that we really see straight lines in the world as curved, in part because the retina is curved. This rationale is easily dismissed, because a vision system could easily correct for retinal curvature.

Eye movements change the projection

Rotating your eyes can create a sensation of curved space (Gombrich, 1974; Tyler, 2015). Consider looking up and down a street, by rotating your eyes or head: each gaze gives a different view into the same scene. In each view, straight lines in the world appear straight. However, rotating one's eyes causes straight lines to change direction in the retinal image, just as rotating a camera causes straight lines to change direction (Figure 15a).

Hence, if we were to depict this experience by concatenating the individual views in Figure 15a, we would get bent lines. This illustrates a fundamental difficulty in representing visual experience: *a viewer sees a different scene projection with each eye movement.* Hence, no *single* projection can fully replicate visual experience when a viewer moves their eyes.

Blending the different projections, as in the curvilinear picture in Figure 15c, removes the discontinuities between different projections, but creates visible curvature.

Defining and categorizing projections

Since the Renaissance, artists and scientists have developed many alternative projections and perspective systems intended to better capture visual experience than linear perspective, and to provide more options for artists and photographers. For example, Leonardo da Vinci, as he became aware of the problems of linear perspective, distinguished between "artificial perspective," that is, linear, and "natural perspective," which would better represent real-world visual experience, such as the relative sizes of objects (Kemp, 1990).

Hence, it is useful to categorize different types of projections to study them. Note that this paper takes the position that no projection is necessarily "correct" nor "wrong" (Gombrich, 1961; Willats, 1997; Koenderink et al., 2016a; Hertzmann, 2022). Each projection produces different percepts and aesthetics, and the preference for one or another depends on the photographer or painter's goals, and their cultural and historical context.

Projections are defined in terms of an underlying 3D scene: either a view of the real world, or a computer graphics model. A projection defines how light rays in the scene project onto points in a picture, thus determining, for a given scene, the colors for each part of the picture. Classical perspective techniques are defined in terms of constructions, but each can be defined by some mathematical projection formula (Willats, 1997).

Projections can be described by whether they achieve specific desired properties or goals (Zorin & Barr, 1995; Carroll et al., 2009; Koenderink et al., 2016a). One class of goals aims to replicate elements of real-world visual experience, including preserving aspects of object appearance (e.g., lines that appear straight in the real world should appear straight in a picture), preserving relative object scale (so that objects' size in a picture is related to their real-world size and distance from the viewer), and preserving orientation (so that the upward direction in the world is upward in the image). Projections may also attempt to satisfy user-specified compositional goals, just as a conventional photographer uses zoom and crop to alter the composition and aesthetics of a linear perspective image.

Taxonomy

In this section, I group projections by the number of COPs required to define them (Figure 16). The main categories I discuss are single-perspective projections (one COP) and multiperspective collage (a set of distinct COPs).

A third category, which I call "infinite multiperspective," includes projections that would require an infinite number of COPs to specify—or, for a real camera, one COP per pixel (or row of pixels). They are usually defined directly in terms of ray directions, rather than in terms of COPs. The most well-known are parallel projections, including orthographic and oblique (Willats, 1997). Parallel projections have an ancient history long predating the Renaissance, and are still used in many contexts, such as maps and video games. Some projections can be defined in terms of simple parameterizations, such as pushbroom panoramas (Seitz & Kim, 2003; Roman et al., 2004), which have

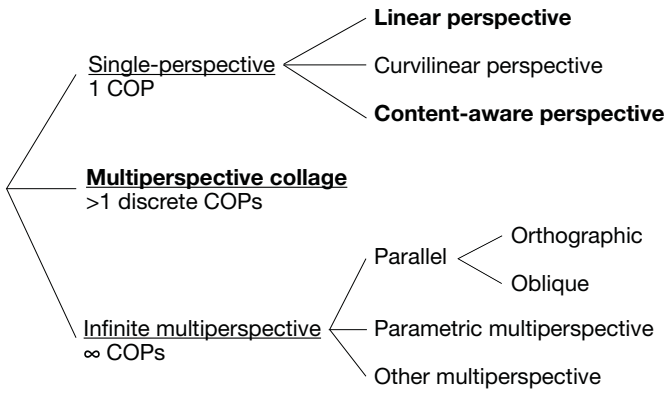


Figure 16. Taxonomy of projection systems used in this paper, based on the number of COPs required to define the projection. The categories shown in bold-face are the ones that are important to this paper. See text for details.

been used for visualizations of long parallel subjects, such as street scenes. Parametric generalizations of pushbroom panoramas allow more-general effects (Yu, McMillan, & Sturm, 2008). User-authored tools allow direct artistic control over freeform projections (Rademacher & Bishop, 1998; Coleman & Singh, 2004; Roman et al., 2004), and may be useful for visualization as well. I do not discuss infinite multiperspectives further.

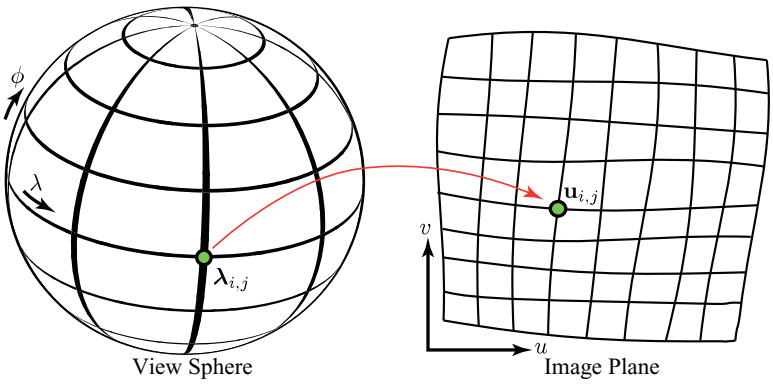
Single-perspective projections

Many projection systems, including linear perspective, can be described as *single-perspective*. A single-perspective projection can be represented as a mathematical function that maps from the sphere of directions around a COP, to points on an image plane (Figure 17a). Figure 9 shows examples of stereographic projections of entire view spheres. In nearly all photography, only a subset of the sphere is captured. Linear perspective is one example of single-perspective projection.

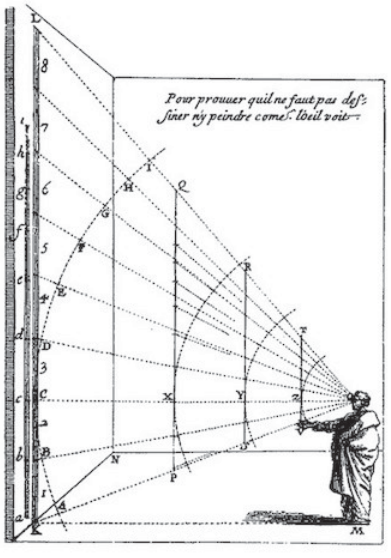
Zorin and Barr (1995) prove that no single-perspective projection can simultaneously guarantee that all possible straight lines project to straight lines, and that all spheres project to circles. Hence, trade-offs are necessary. Likewise, the problem of mapping spheres to 2D pictures has been thoroughly studied in cartography, with hundreds of different projections developed (Snyder, 1993).

We can group single-perspective projections into three categories: linear perspective, curvilinear perspective, and content-aware projection. I have described linear perspective already, and now describe the latter two.

Leonardo's term "natural perspective" is sometimes used to describe techniques designed to capture visual experience better than linear perspective does.



(a)



(b)

Figure 17. Projection as a mapping from the sphere of viewing directions to a plane. (a) Parameterization of single-viewpoint projection, from Carroll et al. (2009). The view sphere parameterizes all light rays reaching the COP at the center of the sphere. Projections map a subset of the sphere to a flat picture plane. Classical projections, like linear and stereographic, are defined algebraically, whereas content-aware projection is computed by a numerical optimization that depends on the image contents. (b) Rays on a sphere, projected to an image plane, by Abraham Bosse, with caption "To prove one can neither draw nor paint as the eye sees." From his 1665 text *Traité Des Pratiques Géométrales Et Perspectives* (Tyler, 2015).

Curvilinear perspective

General-purpose, single-viewpoint alternatives to linear perspective necessarily cause some straight lines to become curved. Some methods derive from the claim that visual space is itself curved (e.g., Barre & Flocon, 1968; Hansen, 1973). Zorin and Barr (1995) define a parametric family of projections that trade-off between straight line preservation and local distortion minimization. Koenderink et al. (2016a) follow Helmholtz in advocating for stereographic projection to capture the effects of eye movements. Stereographic projection does have the desirable property that it projects spheres to circles (Figure 4b), but it does not preserve relative object scales (Carroll et al., 2009). Perhaps because of their perceived distortions, curvilinear projections most often find use for artistic effects, such as in fisheye lenses (Figure 9), and for visualization.

In some cases, moving one's eyes over a curvilinear perspective can simulate the real-world experience of eye movements (Koenderink et al., 2016a). For example, the reader may scan their eyes horizontally over Figures 15a and 15c. The painting in Figure 18 offers an even more subtle example. The actual painting is quite large (122 cm wide), and the reader is encouraged to zoom into the picture on a large monitor. The painting includes a subtly curved wall, but, when viewed up close, the wall seems to be straight in each individual fixation. Hence, the painter has replicated the experience of scanning one's eyes along a straight line in space, while compressing more horizontal space into the same canvas width. The compromise here, though, is that the wall's curvature is visible when viewed from further away. More extreme curved perspectives allow for more horizontal compression of space, such as in Figure 15c and *Arrival of Emperor Charles IV at the Basilica St Denis in 1378* by Jean Fouquet (Koenderink et al., 2016a).

Figure 18. A subtle example of nonlinear projection, in *Mortlake Terrace, Summer's Evening* by J. M. W. Turner (1827). At first glance, the picture appears to be a conventional linear perspective with a straight wall. However, the shape of the wall is subtly curved. The original painting is quite large (122.2 cm wide); the reader is encouraged to view it up close on a large display. As one's eyes scan over the image, each individual fixation seems to contain a normal linear projection. The nonlinearity allows the artist to depict more of the scene horizontally.

Natural perspective for relative scale

Wide-angle linear perspective often makes distant objects seem "too small." You may experience the effect by taking a smartphone photo of a large distant object, like a building or a mountain or the moon, in the default (wide-angle) zoom setting. Then, compare the photo with your visual experience at that location. The landmark will often look "too small" in the picture, as compared with objects nearer to you that frame the landmark. Figure 19 shows a painting by Pepperell (2015) specifically intended to convey full FOV visual experience from a single viewpoint and eye gaze direction, together with corresponding linear perspective images.

Likewise, artists often depict objects larger than they appear in comparable photographs, as illustrated in Figure 20. Sharpless et al. (2010) describe how a class of Baroque painters systematically compress depth. Indeed, my own interest in understanding perspective began when I compared my own drawings to smartphone photos, such as in Figure 32.

To study this phenomenon in artwork, Pepperell and Haertel (2014) compared 18 Cézanne paintings with photographs taken from the approximately the same viewpoints, and found the main subject of the painting to usually be larger than in the corresponding photographs. They then performed an experiment in which they recruited eleven art students without training in linear perspective. The students always drew the central objects of a still life larger than linear perspective would prescribe.

How might we devise projections to better capture these phenomena? Sharpless et al. (2010) describe a projection that compresses depth for wide-angle pictures, inspired by Baroque architectural paintings. Their method projects the scene onto a vertical cylinder,

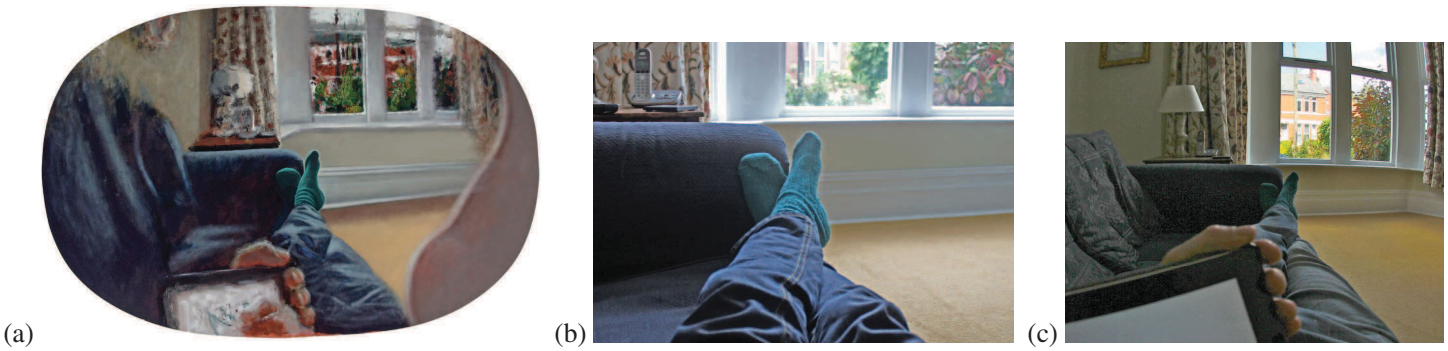


Figure 19. Natural perspective. (a) Robert Pepperell, *Self-portrait (after Mach)*, from Pepperell (2015), painted to visualize the artist's subjective experience of visual space in a single monocular fixation. Much more detail appears in the picture center than in the periphery, as compared to a wide-angle photograph taken from the same view. (b) Normal FOV photograph (50 mm). (c) Wide-angle fisheye photograph (8 mm).

Figure 20. Painters often depict distant objects as larger than than in a comparable wide-angle linear perspective (Pepperell & Haertel, 2014). (a) *High Street, Oxford*, by J. M. W. Turner, painted in 1810. (b) A composite photograph of Oxford High Street, made to replicate Turner's view, depicts distant towers much smaller than Turner did. (c) One of the source photographs for this picture. Two photographs were taken with a tilt-shift lens (17 mm) to keep vertical lines vertical. The two photos were stitched vertically, warped by an affine transformation to better match Turner's viewpoint, touched-up to remove clouds and street signs, and color-adjusted. This picture, from a slightly different viewpoint than the two main source photos, was used to fill holes caused by the street signs. See Fisher (2024) for source photos and process. (Photos by Fisher Studios, 2015.)

and then projects the cylinder to a location behind its center. The method preserves vertical straight lines, and straight lines passing through the image center, but other lines will become curved. They also describe how their projection, hypothetically, could have been used by Baroque artists.

Burleigh et al. (2018) fit a projection to visual space observations made by Pepperell; observers report that this projection better represents visual space than does linear perspective (Burleigh, Pepperell, & Ruta, 2018; Pepperell, Ruta, & Burleigh, 2019). Their "natural rendering" method offers wide FOV with increased object scale at the center of the picture, but does not preserve straight lines.

Working from a single viewpoint limits the options for capturing these phenomena, and necessarily means that they are curvilinear perspectives. Later, I discuss multiperspective techniques that offer more options for adjusting relative object scale (Badki et al., 2017; Liu, Agrawala, DiVerdi, & Hertzmann, 2022).

Content-aware perspective

Content-aware methods adapt the projection to the scene at hand, and can minimize perceived distortion far better than conventional methods. The insight is that one often does not need a single projection formula that works for all scenes. For example, one does not need a projection that preserves all possible straight lines in all possible scenes—it need only preserve the straight lines actually present in a given scene.

Zelnik-Manor and Perona (2005) showed that wide-angle panoramas can be segmented into a few linear projections to decrease marginal distortion artifacts, provided that the boundaries between the different projections can be suitably aligned to the contents of the scene.

Content-aware warping, introduced by Carroll et al. (2009), offers a general formulation that produces high-quality wide-angle pictures (Figure 21). This method computes a freeform mapping from input picture coordinates to output picture coordinates,

Figure 21. 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 22. A taut piece of string in front of a face, photographed in the Google Pixel 5 camera app in ultrawide mode (0.5×). 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.).

thereby producing a new single-view projection of the same scene. They compute the map by a numerical optimization that trades-off several goals, including preserving the appearances of straight lines and faces present in the input photograph, while minimizing a spatial distortion ("conformality") measure. This penalty is applied much more strongly to textured regions than for untextured regions, because it is much harder to perceive distortions in empty regions.

Shih et al. (2019) describe a faster, more-automatic content-aware warp, specialized just for removing face distortion (Figure 4). The camera captures a linear perspective photograph, and automatically detects faces in it. Then, the image is warped by blending stereographic projection for the faces and linear perspective elsewhere. This method runs in real-time on the Google Pixel camera app in ultrawide mode (0.5×); the reader with access to a recent Google Pixel can test this themselves, as in Figure 22.

A key observation from these techniques is that *distortion-avoidance techniques are local.* Local optimization terms to preserve circles, straight lines and faces do not depend on the rest of the input image. (I use the terms "global" and "local" somewhat differently from Carroll et al.).

These content-aware methods have three conceptual shortcomings. First, they require explicit descriptions of the visual properties to preserved (e.g., straight lines), but it is unclear how to define such properties for other shapes and objects. Second, there is no reason to believe that maximiziming conformality (or stereographic projection) genuinely minimizes perceived distortion, for example, it does not preserve straight lines. Third, they are limited to using the light captured with a single COP.

Multiple perspectives in art history

How have classical artists created realistic wide-angle views? Many classical paintings can be described as combining multiple linear projections with different COPs, in a content-aware manner (Kubovy, 1986; Agrawala, Zorin, & Munzner, 2000; Perona, 2013).

As previously discussed, painters depicting large-scale scenes throughout art history depict faces without apparent distortion. Indeed, Pirenne (1970) and Kubovy (1986) observed that Renaissance painters depict spheres and people as though moving the principal point to the object center.

Beyond distortion avoidance, we can observe other uses of multiple perspectives in classical paintings. For example, in the fifteenth century, Paolo Uccello subtly combined multiple viewpoints in his portrait of Niccolò da Tolentino (Kubovy, 1986), as did Andrea del Castagno in his portrait of Dante Aligheri (Perona, 2013). In each case, different elements in the same person or object are depicted as though viewed from different viewpoints, which is hard to notice unless it is pointed out. Hockney (2006) (pp. 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. He also asserts that compositions formed this way provide a more immediate sense of space than conventional linear perspective.

As more recent examples, the painter Richard Estes composes street scenes from multiple source photographs with different viewpoints, making paintings that, as one critic puts it, "emulat[e] the ever changing focus of the restless human eye" (Keats, 2015). Photographer Michael Koller (2004) manually composites photographs into wide-angle panoramas of street scenes.

Inspired by many of these techniques, Perona (2013) composed full-body portraits from multiple aligned COPs (Figure 23), allowing the head and other body portions to imaged from a closer camera distance than needed for single full-body photo. In a survey, seven experts did not notice anything out of the ordinary in the multiperspective portraits. They unanimously agreed that the multiperspective portraits had a different "feel" from the single-perspective versions, often preferring the multiperspective portraits.

Photomontage art offers a useful metaphor for making pictures with local elements: in a photomontage, each individual cut-out picture may look undistorted, despite their differing perspectives. For example, the reader may consider the discrete composition of distinct elements into a coherent scene in Richard Hamilton's 1956 photomontage *Just what is it that makes homes so different, so appealing?* David Hockney's "joiners" compose many distinct views into new compositions, such as his 1986 *Pearblossom Highway* composition, which conveys a painterly perspective composed of ordinary photographs.

These examples illustrate that many *realistic wide-angle paintings can be accurately described as combining multiple linear projections*. For example, the presence of a human face in the margin of any wide-angle realistic picture, without marginal distortion, indicates the use of a separate viewpoint for the face. The same goes for most other distinctly-recognizable object classes in picture margins.

Computational multiperspective collage

This section describes a class of projections that I group under the term *multiperspective collage*. These projections seamlessly combine multiple linear perspective projections into a single picture. These methods can effectively produce large-scale imagery with little or no perceived distortion.

In a multiperspective collage, the picture plane is partitioned into parts, each of which has its own linear perspective projection, with the COP in front. Blending may be used between regions. The partitioning and individual projections depend on the content of the scene being depicted, as well as user goals. Hence, all of these methods are *content-aware*.

A key initial step in this area was the work of Agarwala et al. (2000), who pointed out many uses of multiperspective techniques in the painting and visualization, and categorized their uses for artistic, comprehensibility, and visualization purposes. They described a simple computer graphics technique based on rendering each object with its own linear perspective, with the COP in front of the object, and showed how it can mimic projections in art.

Multiperspective street panoramas (e.g., Figure 3a) can provide more effective visualization for street imagery than linear perspective (Roman et al., 2004; Agarwala et al., 2006). For example, compare Figure 3a with the conventional visualizations in Figure 15. By collaging separate linear projections with separate COPs, these methods can create large-scale panoramas with little apparent distortion. These panoramas have

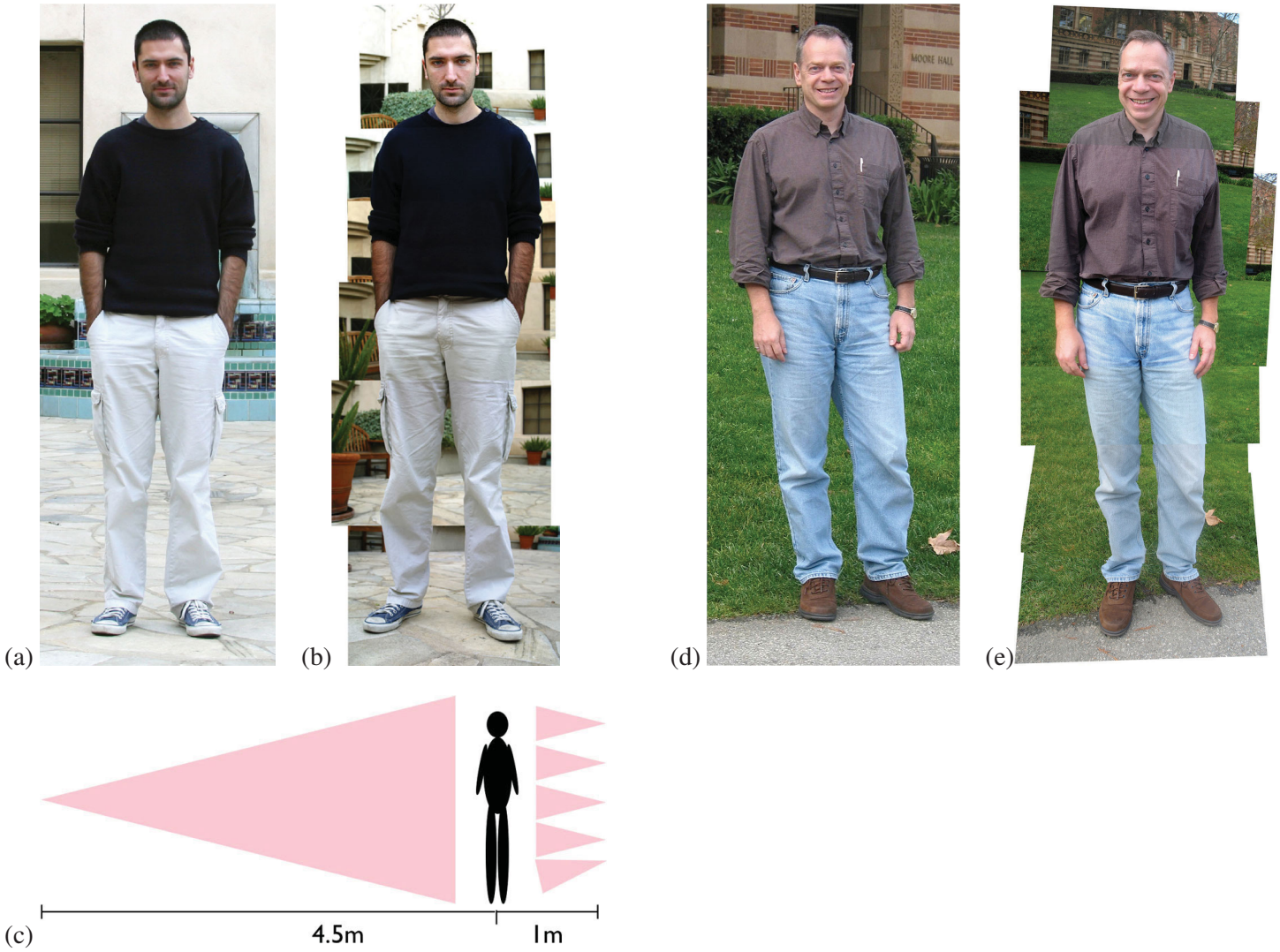


Figure 23. Full-body portraits by Perona (2013). (a) Single-perspective photograph taken from 4.5 m. (b) Composite of multiple photographs of the same individual, taken from 1 m. Note the different appearance of the subject's face; compare with Figure 13. (c) Diagram of camera views for each component photograph in the first two portraits. (d) Single-view portrait taken at 5-m distance. (e) Composite portrait with face taken at 60 cm distance, body from 100 cm. Note the perspective distortion on the face.

unrealistic elements, such as multiple vanishing points and implausible building shapes, but, locally, the images look plausible; identifying many of the unrealistic elements requires cognitive effort over multiple eye fixations.

Whereas the above methods collage horizontally, collages in depth can provide control over object scale. In computational zoom (Badki et al., 2017), a scene is partitioned based on object depth from the camera; then, objects for each depth range are rendered as if photographed with separate COPs and focal lengths, selected in a way that preserves continuity across depth (Figure 24). This allows a user to vary depicted scale separately for different depth ranges. One use is to make distant objects larger, in order to create "natural perspective" imagery, as discussed earlier. Computational zoom requires a collection of photographs to be taken at the same time; our ZoomShop method (Liu et al., 2022) works from a single photograph and generalizes computational zoom to allow for smoother transitions between regions (and, thus, amounts to "infinite multiperspective" in some cases).

Multiperspective projections in computer graphics can simulate other nonrealistic artistic techniques, such as de Chirico paintings (Agrawala et al., 2000), Hockney's joiners (Zelnik-Manor & Perona, 2007), cubism (Collomosse & Hall, 2003), and Disney multiperspective background panoramas (Wood, Finkelstein, Hughes, Thayer, & Salesin, 1997). In each of these techniques, the user and/or algorithm chooses multiple COPs specifically for the scenes being depicted.

Figure 24. 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.

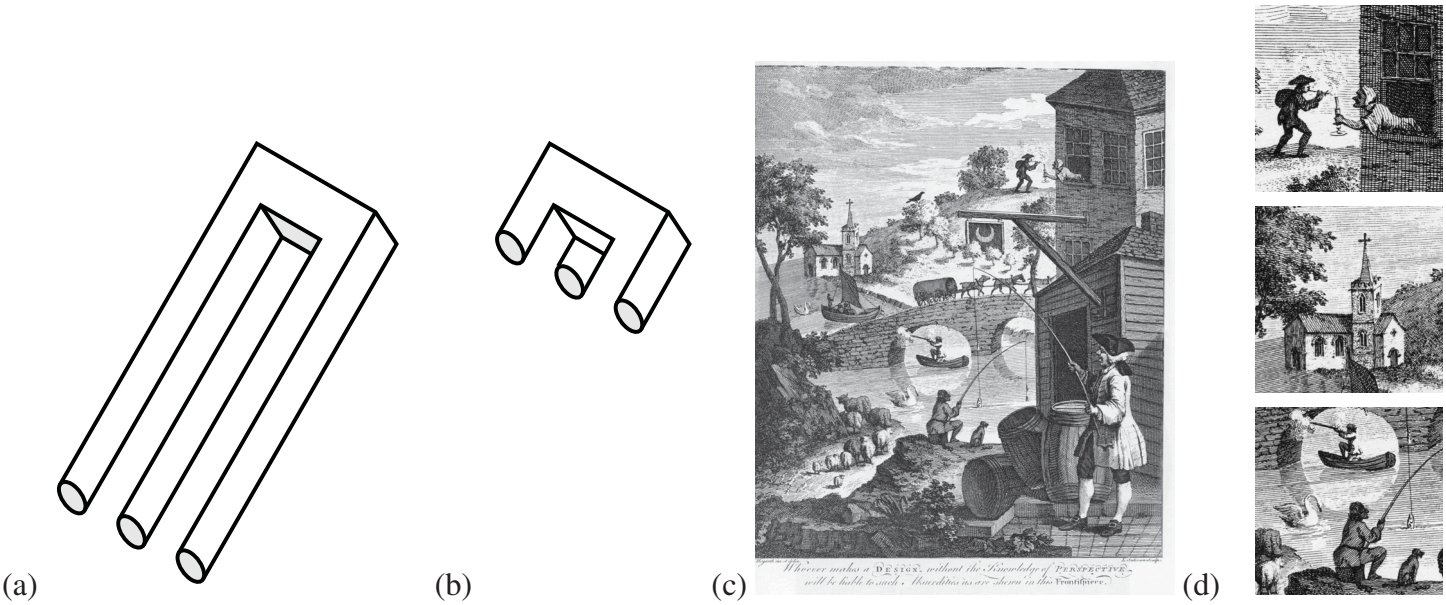


Figure 25. Impossible and indeterminate perspective. (a) Three-Stick Clevis (Schuster, 1964). The spatial contradictions cannot be observed within a single fixation. (b) Short three-stick clevis, viewable in a single fixation. (c) A realistic depiction, with many impossible elements only visible on close inspection. *Satire on False Perspective*, William Hogarth, 1754. (d) On their own, crops of the image appear like geometrically-plausible pictures.

Impossible and indeterminate perspective

Some impossible pictures (Penrose & Penrose, 1958) demonstrate what happens when a picture is plausible *locally* but not *globally*. At first glance, the Three-Stick Clevis (Schuster, 1964; Figure 25a) seems to depict a single coherent object, and each region in the image appears locally consistent. As noted by Schuster (1964), the apparent "illusion" requires multiple fixations in order to detect inconsistencies. Specifically, from the top of the object to the bottom, the roles of the lines (smooth occluding contour or sharp crease) and the position of figure/ground switch, but without any obvious single point of inconsistency. On the other hand, if the contradiction is apparent in a single fixation, as in Figure 25b, the "illusion" goes away. Figure 25c shows a more complex example of a scene that seems to be valid, but contains many small inconsistencies across different regions. These violations do not particularly interfere with one's interpretation of the scene except when fixating on contradictory regions. Several M. C. Escher pieces exhibit more elaborate versions of these phenomena, such as *Belvedere* and *Relativity*. For any of the above pictures, any small region of the picture, considered in isolation, seems to be a plausible linear perspective projection. However, the picture as a whole cannot be described as a projection of

a single plausible scene, without resorting to forced perspective.

Other M. C. Escher pieces explore curved impossible perspectives, such as *House of Stairs* and *High and Low*.

Indeterminate pictures (Pepperell, 2011; Koenderink et al., 2016a) suggest realism, but make it difficult or impossible to determine whether there exists a coherent 3D interpretation. As with impossible perspective, the contents of individual fixations in indeterminate pictures typically appear realistic.

Implications for perception

These projection techniques offer several important implications about perception.

First, there exists a wealth of projection systems with different advantages. Content-aware perspective and multiperspective collage balance multiple goals better than content-independent projections like linear or curvilinear perspectives. Moreover, they can better describe many realistic paintings throughout art history.

Second, computational techniques for minimizing perceived distortion are *local*, and typically simulate multiperspective projection in some way. Multiperspective collages uses separate linear projections for different image regions (Roman et al., 2004; Zelnik-Manor et al., 2005; Agarwala et al., 2006). Content-aware warps use local objective terms to prevent perceived distortion. The specific choice of projection for an image region does not depend on where the region appears in an image, in contrast to conventional projections like linear and stereographic, which project shapes differently in different places in a picture.

Third, many pictures hide impossibility or implausibility across fixations. Inconsistent vanishing points, repeated individuals, and contradictory geometry are all harder to detect when they are far away from each other in the picture. Such pictures appear plausible locally.

Fourth, computational nonlinear projections could provide useful tools for systematic perceptual experiments of picture perception.

Vision at a glance, foveal vision, and visual space

Many conventional theories of vision—as well as common-sense notions of it—assume that we viewers see the visible space in front of our eyes, understand it, and build a mental 3D model of it (e.g., Linton, 2022). However, in the past few decades, many surprising experimental results challenged this view. This section reviews some of these experiments and the counterintuitive new theories of foveal vision, eye movements, and real-world 3D vision that attempt to explain these results. I then argue that these ideas should inform our understanding of picture perception as well.

Vision at a glance

To be most effective at 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, Iyer, Koch, & Perona, 2007; Intraub & Dickinson, 2008; Greene & Oliva, 2009), such as recognizing that a scene comprises a city street. Being able to recognize and interpret space at a single glance is immensely useful for survival: navigating the world would be very challenging if understanding each new environment required moving our eyes all around it first.

Yet, the retinal information at a glance is much more limited than one might think. For example, when photographing the scene in Figure 15, I was able to read the distant street signs in Figure 15d only by fixating directly on them. 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 in Figure 15d and attempt to read a different one. You should find it difficult or impossible to read any text not immediately nearby your fixation. Although it is possible to recognize uncrowded individual letters in peripheral vision (Anstis, 1974), it is extremely unusual to read this way.

This difficulty is due to the limitations of peripheral vision. The retina is often described as comprising the foveal region—the center of the retina where it has greatest acuity—and peripheral vision. We perceive far less detail in peripheral vision than in the fovea. Moreover, the fovea is quite small. One definition of the fovea uses the dimple, which covers 5° of the FOV. Another is the "rod-free fovea," which covers up to 1.7°; peripheral vision covers the remaining 99.9% of the visual field (Rosenholtz, 2016). Using the latter definition, a "rule of thumb" for getting a sense of the size of the fovea is to hold out your thumb at arm's length: at this distance, your thumbnail roughly subtends the angle seen by your fovea. However, note that there is no distinct cutoff between foveal and peripheral vision, and the actual differences between the two are far more gradual and nuanced than usually portrayed (see Rosenholtz, 2016).

To attend to something in the real world, we look at it (O'Regan & Noë, 2001; Wolfe, Kosovicheva, & Wolfe, 2022), and what gets noticed depends on where one's eyes fixate, for how long, and the limitations of peripheral vision (Rosenholtz, 2020).

Fragmentary visual space

If vision at a glance is so effective, then perhaps we do not need to mentally reconstruct a precise 3D model of the world over time (Noë, 2002).

Classical notions of *visual space*—a person's visual experience and internal representation of 3D space—treat it as a coherent geometric representation of the real world (Suppes, 1977; Erkelens, 2015). However, many studies identify inconsistencies and distortions in visual space that cannot be explained by any metric geometry (Linton, 2022): distortions of estimated depth due to irrelevant factors (Johnston, 1991; Todd & Norman, 2003; Vishwanath, 2014; Campagnoli, Hung, & Domini, 2022), inconsistencies between estimates of depth, slant, curvature, and/or shape (Koenderink, 1998; Koenderink, van Doorn, Kappers, & Lappin, 2002; Loomis, Philbeck, & Zahorik, 2002; Di Luca, Domini, & Caudek, 2010), and internal inconsistency of multiple relative judgements (Koenderink, van Doorn, Kappers, Doumen, & Todd, 2008; Svarverud, Gilson, & Glennerster, 2012; Vuong, Fitzgibbon, & Glennerster, 2019). Some recent theories explain inconsistent results by treating visual space as fragmentary in some way, with, for example, separate representations for different surfaces (Koenderink, 1998), different distance ranges (Vishwanath, 2023), or for action versus perception (Goodale & Milner, 1992).

Here I use the idea that 3D vision is fragmentary across fixations (Koenderink et al., 2008). Indeed, change blindness experiments demonstrate inconsistency across fixations: viewers may forget the appearances of individuals before them (Simons & Levin, 1998); in virtual reality experiments, viewers do not notice small rotations of the entire world during saccades and blinks (Langbehn, Steinicke, Lappe, Welch, & Bruder, 2018; Sun et al., 2018), and sometimes forget the existence of entire objects previously within their visual field (Martin, Sun, Gutierrez, & Masia, 2023).

Some 3D information must persist across fixations, but far less than the dense 3D that one might assume.

Many of the above observations are highly counterintuitive. Together they create an *awareness illusion:* we effortlessly perceive a richly detailed, consistent visual experience, yet, when probed, demonstrate a surprising lack of awareness of many details (Dennett, 1991; Noë, 2002; Rosenholtz, 2020).

Implications for picture perception

I claim that these counterintuitive observations directly translate to important features of picture perception.

To understand a picture, a viewer usually must move their eyes over it, because of typical picture viewing distances. For pictures larger than postcards or smartphones, people typically prefer pictures to occupy about 36° of their FOV (Cooper et al., 2012), which is much wider than either definition of the fovea.

However, a viewer need not scan their eyes over an entire picture before interpreting it. Picture interpretation occurs at each glance: a viewer begins interpreting a picture from the first fixation, including recognizing objects and gist, and adding more information with each fixation.

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. This is a direct consequence of the many change blindness studies that have been performed with pictures and videos.

For pictures, the corresponding theory of visual space is *pictorial space:* the notion of a 3D representation that viewers infer for pictures. If visual space is fragmentary across fixations, then we would expect the same for pictorial space. That is, viewers do not reconstruct a 3D pictorial space that is fully coherent across fixations over a picture. Conversely, pictures can create compelling illusions of 3D space without perfect spatial coherence.

In contrast, existing treatments of pictorial space perception seem to imply simultaneous global processing of an entire picture, inconsistent with the nature of eye movements and foveal vision. For example, some vantage-point compensation theories require that a viewer compensates for COP when viewing a picture. However, all known procedures for inferring the COP involve global picture processing, like locating multiple straight lines in the picture, and/or convolutional neural network processing (Kubovy, 1986; Lee et al., 2021; Jin et al., 2023). If viewers compensated according to COP, then we would expect that either a) viewers do not interpret 3D in a picture until after enough fixations to estimate COP, or else b) that viewers' perception of 3D would continually shift as new information arrives from successive fixations. Neither seems to be the case.

The next two sections propose new hypotheses consistent with these observations and those from previous sections.

Local principles of projection perception

No existing theory compellingly explains viewers' projection assumptions and the variety of phenomena surveyed in this paper. Many perceptual theories and studies focus on linear perspective, including the

Figure 26. Cropping an image region does not change the shape percept for that region (except when object recognition is ambiguous, as in Figure 27). Crops are shown from Figure 5.

Vantage-Point Compensation literature surveyed earlier (e.g., Pirenne, 1970; Vishwanath et al., 2005), and many other studies of pictorial space (Hecht et al., 2003). Yet, throughout this paper, I have reviewed the problems of conventional linear perspective with a single COP, either as a foundation for perceptual theory or as a technique for making pictures.

This section and the next formulate new perceptual hypotheses for projection. I begin by describing *local* properties of shape perception, then add increasing specificity about what the local assumptions are, and how shape and scale are related. The following section then discusses entire pictures.

The hypotheses here combine and extend several currently-separate sets of ideas: the roles of fixations, foveal vision, and vision-at-a-glance (Rensink, 2000; Rosenholtz, 2020; Wolfe et al., 2022); effective formulations for removing perspective distortions in computational photography, summarized by the DVC (Zorin & Barr, 1995); the fragmentary nature of 3D vision (Koenderink et al., 2008; Linton, 2022); and local slant compensation (Vishwanath et al., 2005).

Each of the hypotheses proposed here may be tested and refined in future studies.

Shape locality

Vision-at-a-glance and foveation indicate that we interpret scene shape in each fixation, and these interpretations are usually stable. We usually recognize objects at the first glance at a region, and its shape percept does not change after future fixations. Moreover, as we have seen, the most effective methods for reducing perceived distortion are *local*, operating only on specific image regions and not depending on the rest of the picture.

Hence, I propose the following principle, called *shape locality*:

Figure 27. 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. (Pictures from Schwiedrzik, Melloni, & Schurger, 2018, CC-BY.).

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

This applies to unconstrained normal viewing of a picture, over time, across multiple fixations.

Shape locality describes how changing the visual context around an object does not change its appearance, as illustrated in Figure 2. One can generally crop out an object out of a picture without changing its appearance. For example, the crops in Figures 26 have the same apparent shape and distortions (or absence thereof) as they do in the uncropped picture (Figure 5).

Object recognition can change during viewing, as in bistable images and hidden images (Figure 27), at which

Figure 28. A miniature diorama convincingly designed and photographed to look like a large-scale scene, demonstrating how familiar objects provide cues to scale. With the hand covered, the scene looks like a normal-sized room without careful inspection. With the hand visible, scale cues from the hand override other scale cues. (*The Brooklyn Collector* by Dan Ohlmann and Musée Cinéma & Miniature.).

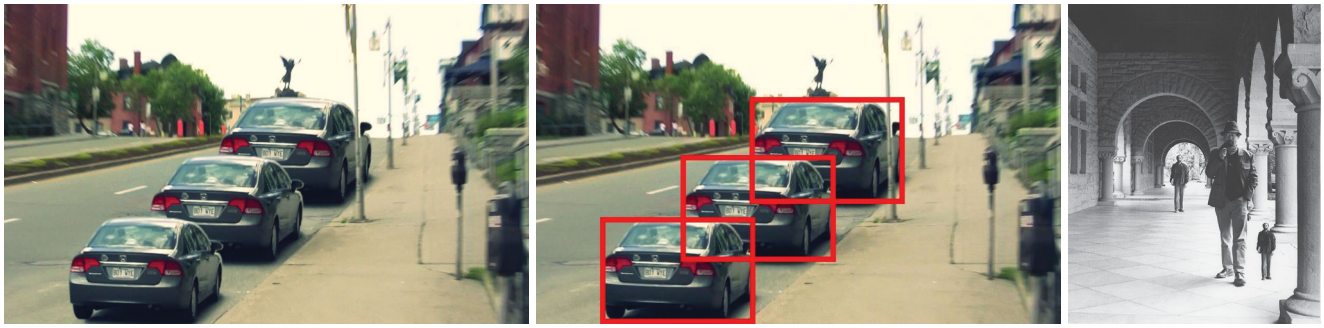


Figure 29. Ponzo illusion, illustrating that object size can depend on context: the three cars seem to have the same shape, but not the same size, even though they comprise identical sets of pixels in the picture. (Left picture by Alex Blouin, with annotation by Paul Linton. Right picture: © The Exploratorium. All rights reserved. Used and adapted with authorization. The Exploratorium is a registered trademark of The Exploratorium, http://www.exploratorium.edu.).

point shape percepts can change, but they remain stable unless recognition changes again.

High-level cognition can recognize shapes as distorted, and one may consciously reason about how a distorted object should look, but this does not change the perceived shape.

The shape locality principle is inconsistent with any perceptual theory that involves estimation of global projection parameters, such as estimating a picture's COP from straight lines.

Further study is required on the minimal size that a picture region can be, but here are a few observations. Any region containing a recognizable object, object part, or surface could constitute a region on its own. Hence, the minimal region size may be stimulus-dependent. Region size may also be related to properties of the retina and viewing conditions, e.g., perhaps locality applies to any region larger than, say, a 6° visual angle, or, equivalently, a circle of diameter 6 cm in a picture viewed at a 60 cm distance.

Object size perception is non-local

Although I focus on local principles here, it is useful to contrast them with an example of a non-local perception: object scale. An object may look larger or smaller depending on the scene around it, both in absolute terms (e.g., Figure 28), and relative to other scene objects, as in the Ponzo illusion (Moore & Egeth, 1997) (Figure 29). 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 (Held, Cooper, O'Brien, & Banks, 2010)). Hence, perceived object size depends on non-local picture contents, and can change when the rest of the picture changes.

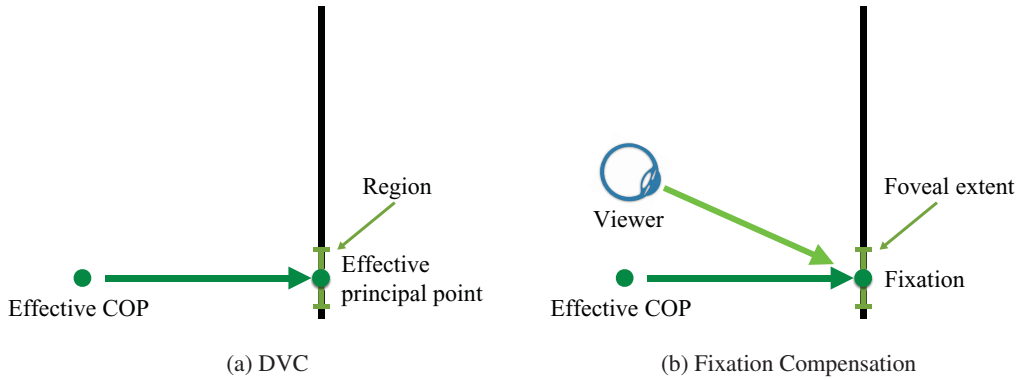


Figure 30. Geometry for two of the hypotheses here. (a) The DVC asserts that the contents of a picture region seems to be undistorted if they look like a linear perspective projection of a plausible 3D scene, with the COP somewhere in front of that region, and, thus, the principal point within the region. (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.

This echoes theories from real-world 3D vision that separate shape and scale, whether in pictures or in real-world 3D vision (Vishwanath, 2014; Linton, 2022), as well as evidence of dissociation between shape and location estimation (Loomis et al., 2002). Vishwanath (2014) points out that internal 3D structure (what I am calling "shape") in real-world vision can be determined solely from relative depth estimates, but absolute depth estimates require additional cues.

More generally, spatial relationships between objects are non-local and typically require multiple fixations to observe.

The DVC

If we interpret shape locally, when do shapes "look right" or look distorted?

Zorin and Barr (1995) observed that "the images of objects in the center of the picture never look distorted, given that the distance to the center of projection is large compared to the size of the object." That is, objects near the principal point of a linear perspective picture normally appear undistorted. Straight lines appear straight, spheres appear circular, and so on. Similarly, Pirenne (1970) and Kubovy (1986) observed that Renaissance painters depict spheres and people as though moving the principal point to the object center.

Based on these observations, Zorin and Barr (1995) proposed a rule for minimizing distortion in pictures, called the direct view condition (DVC): "Objects in the image should look as if they are viewed directly—as they appear in the middle of a photograph."

To make the statement more precise, and to formulate it as a perceptual hypothesis, I propose this formulation of the DVC:

> Under normal viewing conditions, a picture region appears undistorted if and only if it looks like it could appear at the center of a normal-FOV linear perspective picture of a plausible scene.

This statement implies a local, linear perspective with COP in front of the picture region, visualized in Figure 30a.

The normal viewing conditions constraint allows for slant compensation. The normal FOV and normal viewing constraints eliminates the possibility of perspective compression and distortion, e.g., Figures 12, 13 and 23a, that come from wide-angle focal lengths or extremely close or far viewing distances.

Moreover, the degree of deviation from the appearance of a linear perspective determines how distorted the region looks. For example, the visibility of the curve in the wall in Figure 18 may vary depending on viewing. The greater the visible curvature, the more distorted the wall looks.

Figure 23 illustrates how focal length affects perceived shape in individual image regions, and can create local perspective compression/expansion.

The DVC 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 seems to be an undistorted projection of plausible elements.

The DVC is subject to the limitations of vision depending on the viewer's vantage-point, for example, a slightly curved line may seem to be straight when the picture is viewed from up-close (due to peripheral vision limitations) (Figure 18) or from very far enough away (due to limited foveal acuity).

One simple consequence of the DVC is that a normal-FOV linear perspective picture does not appear distorted; distortions are consequences of wide-FOV projections or nonlinear projection. This statement aligns with some of the conventional wisdom (Cooper et al., 2012), but the DVC can also make predictions for wide-angle and nonlinear projections, and recommends ways to reduce perceived distortion in them.

As with shape locality, minimal region sizes are a topic for future study, and may be stimulus-dependent. At a minimum, the DVC should apply to recognizable objects and object parts.

How to make undistorted wide-angle pictures

Distortion-avoidance techniques in existing content-aware warping algorithms (Carroll et al., 2009) can be viewed as special cases of the DVC. Straight lines should be depicted as straight; spheres should be depicted as circles; any projections of an empty, textureless region (such as empty sky) will look undistorted; all of these principles are special cases of the DVC. Texture may change, but it should appear to arise from the same stochastic process (Efros & Leung, 1999; Portilla & Simoncelli, 2000): leaves on a bush can look plausible regardless of the specific placements of the leaves. In this case, the DVC generalizes patch-based texture (Efros & Leung, 1999), because both assess image plausibility in terms of the plausibility of individual image regions.

Hence, based on the DVC, we may predict that a viewer will perceive distortion only when they fixate on regions that violate the DVC. Conversely, a picture that satisfies the DVC everywhere will never appear distorted, including many wide-angle paintings throughout history. For example, *The School of Athens* looks undistorted because anywhere one might fixate will look like it *could have been* in a normal-FOV linear projection of a real scene. Most regions of Figures 3a, 3d and 21c look undistorted, despite the very nonlinear projections used to produce them; they appear much less distorted than a "correct" linear perspective projection would. Even in the impossible perspectives in Figure 25, most regions seem to be undistorted.

Fixation-centered perspective

The DVC does not directly discuss perceptual mechanisms or what shapes are inferred. I now propose a more general hypothesis, fixation-centered perspective:

> In each fixation, a picture is interpreted in terms of a linear perspective projection, with the principal point located at the fixation. The effective COP 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.

There are different possibilities for what the assumed COP is, related to slant compensation. As discussed previously, whether slant compensation occurs automatically remains controversial.

Consequently, I offer two variant hypotheses. The first, direct fixation-centered perspective, simply asserts that the effective COP is always located at the viewer's eye position.

The second, fixation compensation, is as follows:

> In normal viewing conditions, a picture is interpreted in terms of a linear projection with the COP in front of the fixation, at a distance approximately equal to the viewer's distance to the fixation (Figure 30b). When the vision system cannot infer surface slant, such as in peephole viewing and high-curvature regions, the effective COP is simply located at the viewer's eye. As the strength of slant cues increases, the effective COP may be in intermediate positions (Vishwanath et al., 2005).

Higher-level cognition can affect this stage via object identification and recognition. For example, in bistable and hidden images (Figure 27), outside knowledge or cognition may change the identity of observed objects; however, the new interpretation must still obey the fixation compensation perspective.

Fixation-centered perspective involves an effective linear projection to a COP that may differ from the viewer's position. I do not claim that the visual system explicitly reasons about the projection, but, rather, interprets pictures consistently with some effective projection. The effective projection simply describes the possible relationships between the geometry of the picture contents and a viewers' inferred 3D shape interpretation. Similarly, an artificial neural network trained to predict depth from pictures (e.g., as in Ranftl, Bochkovskiy, & Koltun, 2021) need not reason explicitly about perspective, but would be constrained according to the projections in its training set.

Discussion

Here is a possible interpretation of Fixation-Centered Perspective: it is 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. If slant compensation occurs, it occurs because the vision system "knows" slant to be irrelevant to the picture contents. Slant compensation may relate to our normal ability to recognize slanted objects in the world (Vishwanath et al., 2005).

The two versions of the hypothesis offer different explanations of the DVC. Fixation compensation directly predicts the DVC, subsuming it. The argument for how direct fixation-centered perspective predicts the DVC is a little more subtle. If no slant compensation occurs, then a picture region's appearance will subtly skew with every head movement. One might expect that viewers tend to view pictures from different positions in front of a picture. The DVC, then, offers a rule of thumb: depict objects according to zero slant, since this minimizes average distortion across a range of likely viewing positions. This hypothesis may predict different preferred projections for large pictures typically viewed from below.

The two versions disagree in their prediction about whether marginal distortion occurs when viewing from the COP. And, indeed, marginal distortion does occur under binocular viewing from the COP, at least in the informal experiment described previously (Figure 8). This supports the compensation hypothesis. More formal experimentation is needed, however.

Under normal viewing conditions, fixation compensation is a variant of local slant compensation (Vishwanath et al., 2005). Recall that local slant compensation predicts that viewers compensate for the surface slant at the place where an object appears in a picture. To distinguish these hypotheses, we may predict that, when a viewer is required to fixate away from an object, slant compensation for its shape will be centered at the fixation point, not at the object location. Note also that local slant compensation assumes that object location is well defined, and so cannot directly make predictions about very large objects or ambiguously-defined objects.

These fixation-centered hypotheses apply to individual fixations, whereas the shape locality and the DVC hypotheses apply to picture perception over multiple fixations and head movements.

As with the previous hypotheses, I do not specify what, exactly, a region is, as there is insufficient evidence in the literature to do so. For sake of discussion, one may consider the following rule-of-thumb: a region is the picture area viewed by a 6° visual angle, or, equivalently, a circle of diameter 6 cm in a picture viewed at a 60 cm distance. The reality may be far more complicated—different tasks and different stimuli have different "useful fields-of-view" (Wolfe et al., 2022); in the absence of crowding, viewers can recognize shapes in far peripheral vision (Anstis, 1974; Rosenholtz, 2020). There may be no real cutoff at all, but, rather, a fall-off in the precision and detail of information obtained away from a fixation.

All of these hypotheses can account for pictures on curved surfaces, such as in Figure 1a. Fixation compensation, in particular, does not require an entire picture surface to be flat; a region need only be locally, approximately flat.

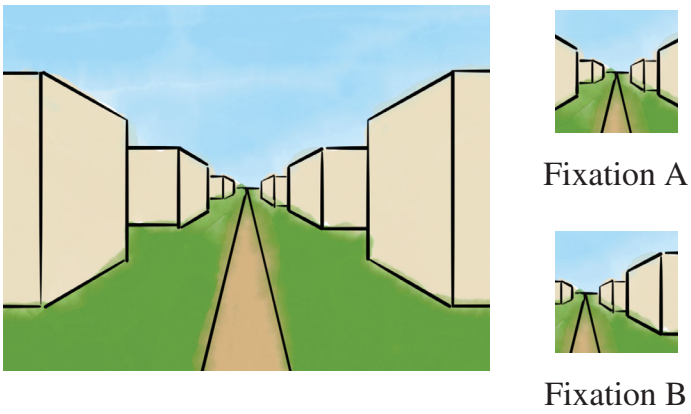


Figure 31. Illustrative example of two fixations in a one-point perspective drawing. Fixation A lies on the vanishing point at the center of the picture, and fixation B is slightly to the right. See text for discussion.

It is interesting to consider the one-point perspective picture in Figure 31. When fixating on the picture's center (fixation A), one sees a conventional one-point perspective picture with the vanishing point at the picture's center; lines perpendicular to the image plane converge to this point. When the viewer fixates a little to the right (fixation B), the vanishing point moves left, so now the lines converging to the vanishing point appear slanted with respect to the image plane, as though the viewer had rotated their eyes to the right in the real world. It is not precisely the same geometry the viewer would have seen if they actually rotated their eyes in the scene (even if the viewer were at the COP, because of compensation). But, nonetheless, this behavior partly mimics the real-world effects of eye rotation, such as in Figure 15a. It also echoes Koenderink et al. (2016a) assertion that "rotations of the eyeball have the effect of translation of the picture plane." And, under compensation, the percept is largely independent of viewer position.

Considerable evidence suggests that 3D vision is not a metric 3D reconstruction; some models of 3D space involve non-Euclid visual spaces (Linton et al., 2022). For example, Erkelens (2021) proposes a projective model of visual space perception that, when applied to pictorial space, predicts how regular structures are spaced in paintings (Erkelens, 2016). In fixation-centered perspective, such models would determine a separate pictorial space in each fixation. Nearby fixations would have very similar pictorial spaces, however.

Summary of evidence and predictions

The twin ideas of shape locality and fixation-centered perspective account for existing evidence about shape and spatial perception in pictures that no existing

theory accounts for. This evidence includes 1) the remarkable successes (and failures) of linear perspective as a projection technique, 2) local slant compensation results (Vishwanath et al., 2005), 3) vision-at-a-glance and foveal vision, which show that viewers must infer perspective for fixations based on limited information, 4) marginal distortion and perspective distortion (compression and expansion), 5) pictures on slightly curved surfaces, 6) the effectiveness of multiperspective and content-aware projections, at least, locally, 7) the fragmentary nature of 3D vision, and 8) the partial 3D perception of impossible shapes.

These hypotheses make predictions that could be tested in the future. Shape locality predicts that shape perception is determined in an initial fixation, and does not change, even if the rest of a picture changes in normal situations. Fixation compensation makes similar predictions to local slant compensation, but differ when objects are not fixated upon. The DVC predicts when objects do and do not seem to be distorted in pictures.

Global pictorial projection perception

The previous section discussed the interpretation of individual fixations in a picture. Now I consider perception of space and shape in a whole picture, across multiple fixations. 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?

This paper hypothesizes an initial answer to these questions by separating per-fixation processing from overall scene awareness and understanding. In each fixation, local perspective follows the relatively rigid rules described in the previous section. As one's eyes move, the visual system identifies objects and infer their spatial relationships, building up an overall awareness of the picture contents. However, this awareness is not a dense 3D representation.

Real-world 3D vision

I begin by outlining a model of some aspects of real-world 3D vision, since 3D perception in pictures surely recruits many of the same processes as in 3D vision.

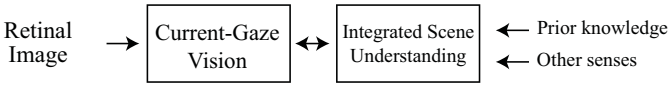
I hypothesize that *all fine-grained 3D vision occurs in per-fixation visual processing*. As previously argued, 3D vision is fragmentary across fixations: only a fraction of 3D information is preserved from moment to moment, and this information is highly-abstracted from both retinal information and detailed shape inferences. Vision does not reconstruct a precisely detailed world model and build upon it across fixations, as demonstrated by change blindness illusions (e.g., Langbehn et al., 2018; Sun et al., 2018; Martin et al., 2023), and by the inconsistencies in viewers' 3D position and shape estimates (e.g., Loomis et al., 2002; Koenderink et al., 2008; Svarverud et al., 2012; Vuong et al., 2019). Instead, to visually attend to something, we look at it. As a result, a viewer can study fine details of shape and color within a fixation but then be unable to recall any details immediately after, other than any that have been consciously committed to working memory.

Per-gaze 3D vision is automatic and unconscious. We cannot choose what we "see," with some rare exceptions, e.g., bistable stimuli. When they conflict, per-gaze 3D vision supersedes knowledge. For example, in the scintillating grid illusion (Schrauf et al., 1997), we see dots appear across different fixations, uninformed by their absence in previous fixations. In the hollow face illusion, each vantage point produces a misleading percept, uninformed by geometric knowledge available from previous vantage-points. Given the choice between a familiar type of convex object undergoing a surprising rotation, and a static but less-likely concave object, vision chooses the former. Similarly, when viewing a picture from its COP, opening or closing an eye changes the perception of marginal distortion, even though we know the picture has not changed.

Hence, the consistency of our 3D perception is explained not by consistency of our representations, but by the consistency of the world.

Two-stage model

To encapsulate these ideas, I suggest separating real-world 3D visual awareness at any instant into two modules, one for the current gaze, and one for scene understanding that persists over time:



The first module, *current-gaze vision* performs all visual interpretation of the current view, including all retinotopic processing in the ventral visual system, combining this information with top-down information to form a percept. This may include 3D information like surface shape and appearance for objects currently in view, primarily in or near the foveal region, as well as the gist of a view, and thus all of the elements of Rensink (2000) model lie in this module. This module includes automatic, unconscious processing, both

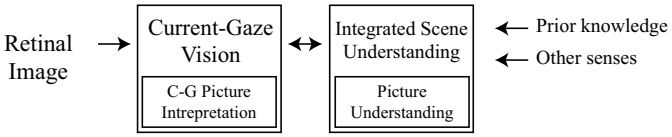
perception and automatic post-perceptual cognition (Linton, 2022).

The second module, *integrated scene understanding* describes overall high-level understanding of the entire world around the viewer—both the elements in view, and also what's below, above, and behind, and incorporates information from prior knowledge and the other senses (sound, proprioception, etc.), and includes working memory. Hence, this is not a vision module per se, though it obtains information from vision and can influence vision with top-down information.

Each representation continuously updates over time, so that one can recognize and interpret objects in a single glance, but also build up richer scene understanding over time.

Pictorial 3D vision

Pictures appear in the world, and so it is reasonable to assume that the same modules process pictures as part of normal vision:



The current-gaze module processes retinal information about a picture, applying fixation-centered perspective in normal viewing to infer shape. It does not see the picture as in full detail, due to foveal vision, though it does receive from integrated scene understanding some high-level information about previously-seen contents of the picture. Integrated scene understanding combines information across fixations and prior knowledge to form an overall high-level interpretation of the contents of the picture and their spatial relationships.

Viewers perceive distortion when the current-gaze interpretation does not match that from integrated scene understanding. For example, current-gaze interprets marginal distortions as sheared objects (e.g., Figure 4a). When the object is familiar, integrated scene understanding recognizes the known shape of the object and the mismatch to the sheared shape. Otherwise the misinterpretation stands, e.g., as in Figure 11.

The different perspectives in fixation-centered perspective imply conflicting 3D space interpretations, but the vision system may not need to resolve these inconsistencies. The interpretation is continually updated, and need not be geometrically stable over time. 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.

How can a flat picture provide an illusion of 3D space?

These models of 3D vision suggest possible answers to some of the puzzles of picture projection. In one telling, pictures use linear perspective to provide the same light to a viewer as does a real scene (Yang & Kubovy, 1999), allowing a viewer to reconstruct the scene as they would in real life (e.g., Juricevic & Kennedy, 2006). As the problems with this traditional view have arisen, it has become puzzling how pictures can provide compelling 3D visual experiences.

So, then, how can a flat picture convey an illusion of 3D space? My answer is that *real-world 3D vision is far less consistent than it seems, and thus easier to simulate than it seems.* To provide an illusionistic sense of space, a picture merely needs to provide plausible pictorial cues for each fixation. Different fixations need not be perfectly consistent with each other. Thus, when done well, a picture can provide a visual experience that shares much in common with a real-world visual experience.

Moreover, pictures that are abstract, ambiguous, or impossible can provide meaningful 3D percepts, because each individual fixation can provide a 3D shape percept. Even if objects float nearby on an empty background, their spatial relationship still provides some 3D information, and more detailed representations can give more fine-grained information. For example, the cave painting in Figure 1a gives some limited information about shape and spatial relationships, whereas a photograph can provide very detailed information about each. Impossible pictures convey plausible shape within each fixation, and relationships between nearby objects, but these cannot be resolved into a coherent whole. Nonetheless, the visual system extracts local shape information, and whatever relationships between regions it can.

Implications for understanding art

How we understand perspective directly connects to how we understand pictures and representational art, in many disciplines. Elkins (1994) points out that "perspective is not fully at home in any one discipline;" at one university, works on perspective spread across "the mathematics library, the fine arts library, the architecture library, the engineering library, and both general libraries."

This section discusses how the hypotheses in this paper could inform these discussions and offer tools for understanding perspective and composition in representational pictures.

What is a picture?

Perspective relates to a broader question, long pondered by philosophers, psychologists and artists: *what is a picture?* (Gombrich, 1961; Goodman, 1968; Gibson, 1971; Gibson, 1978; Kemp, 1990; Hecht et al., 2003; Greenberg, 2021). That is, how and why does the visual system "understand" pictures, and what do artists do when they make pictures?

Here I review prominent ways to understand the nature of pictures, and discuss how the hypotheses proposed in this paper can inform this understanding.

Depictions as recordings of light

Influential art critic John Ruskin (1857) wrote that the technical power of painting derives from recovering "the innocence of the eye:" recording the colors of the retinal image, "a sort of childish perception of these flat stains of color, merely as such, without consciousness of what they signify—as a blind man would see them if suddenly gifted with sight."

Likewise, linear perspective has often been treated as a rational, scientific approach to making pictures (Elkins, 1994), producing pictures by exactly reproducing the light seen by a viewer of the real scene (Gibson, 1971). Some authors even *define* a picture as a recording of light at a fixed viewpoint (e.g., Yang & Kubovy, 1999), like looking through a window.

Yet, as I have discussed, pictures viewed under normal viewing conditions are not perceived as though looking through a window. Few paintings seem explainable simply as recordings of light—even photography requires many artistic choices, due to its limitations in displaying tone, color, and space (Gombrich, 1961; Wilson, 2021; Hertzmann, 2022).

Pictures as cultural product

In response to such rationalist views, many art historians and philosophers have treated pictures purely as products of culturally-determined "languages" (Panofsky, 1927; Goodman, 1968; Elkins, 1994), without any basis in geometry or perception. Indeed, many common depiction systems cannot be understood by novice viewers (Deregowski, 1989). Moreover (Cohn, 2012; Cohn, 2014) presents a compelling range of evidence that learning to draw is much like learning language.

But the idea of perspective as *solely* a cultural phenomenon is contradicted by extensive studies demonstrating that a person who has never seen a realistic photograph or drawing can understand one (e.g., Jahoda et al., 1977). Moreover, the many visual cues shared by real-world imagery and realistic painting, photography, and computer graphics do not seem coincidental.

Depiction as recording mental representations

Popular accounts of art often *equate* depiction with perception. According to artist Harold Cohen (2014), what we draw is "the internal model of the world inside our head." Some perceptual theories explain drawings as "articulating perception" (Cohn, 2012): an artist looks at an object and draws their mental representation. Chamberlain et al. (2016) express puzzlement that many adults find drawing from life difficult, given how easily novices can trace over a picture, with the implication that drawing could be just an easy task of tracing a mental representation.

Yet, it is unclear that the brain naturally keeps mental representations suitable for tracing full-size pictures. Early vision processes retinal imagery, which is too foveated to be suitable for tracing. One might claim that conscious mental imagery provides these representations (Schwarzkopf, 2024). But Catmull and Wallace (2023) describe examples of highly-talented artists who are unable to conjure mental images (aphantasia), indicating that conscious mental imagery is not strictly necessary for drawing skill.

Pictures are tools for visual communication and aesthetics

In the words of Fan et al. (2023), drawing is a "versatile cognitive tool." I follow Gombrich (1961), Fan et al. (2023), and many others in the belief that people create pictures to communicate information and/or create visual experiences. Pictures can produce aesthetic responses, percepts, interpretations, and/or emotions in a viewer.

An artist or photographer making a realistic picture chooses how to arrange elements in 2D and depict elements in space, whether consciously or not. There is no single right or wrong projection; yet, different choices create different percepts. The local and global elements of pictures provide the elements of a "language" (Greenberg, 2021) of pictures: 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, but many aspects (especially local ones) derive from biological vision.

I next survey how the hypotheses in this paper could help inform the nature of these choices.

Interpreting projection in art and photography

In my own experience, drawing a realistic picture requires trading-off between local and global

Figure 32. A sketch that I drew in 2019, while attempting to make a realistic depiction. This experience, and others like it, inspired the investigations in this paper. (a) A plein air sketch of the Aurora Bridge, Seattle. (b) A wide-angle photograph taken immediately after making the sketch (iPhone Xs, 1×). Some objects in the sketch seem to be larger than in the wide-angle photo, and many others disappear. This may reflect the apparent expansion of objects when fixated upon (see Figure 19); the objects that appear largest were likely those that were fixated upon the most in early stages of the composition.

considerations. One can arrange objects freely on the picture plane, making each element larger or smaller or omitted, as desired, but each arrangement offers different percepts, each conveying the scene in a different ways. Adding more objects visible in a scene makes them more crowded, as does increasing the FOV for the same picture size. When composing, say, a landscape, one would like to convey the sweep of a wide-open space, but squeezing a broad panorama onto a picture plane requires distorting some elements, and removing others entirely. Adding additional elements conveys more of the objects present in the world, at the cost of taking picture space away from other elements (and more time and effort). Some options are more or less realistic, some more or less harmonious. In the push-and-pull of adjusting these elements, one trades off local considerations—making each element look realistic, or look good—versus the overall composition and sense of space.

We can interpret perspective in art history and photography in terms of these perspective choices. In the sequence of pictures in Figure 1, we see individual people and objects depicted with varying degrees of realism (including local distortions), and composed together spatially. Compositional options range from highly rigid to extremely freeform. Renaissance artists using strict linear perspective architecture (e.g., *School of Athens*) constructed virtual spaces and then collaged in individual people and other elements with their own local projections. Artists may roughly follow perspective projection, but still shift and adjust space, such as in Figures 18 and 32, according to compositional goals, for example, to fit a scene into a picture frame. Space may be freeform in the picture plane, for example, curved perspective, or, mixed between 2D and 3D (e.g., Diego Rivera murals), or even impossible (Figure 25). Orthographic projection, often used in various forms in ancient Chinese painting (Figure 1b) and architectural drawings (Willats, 1997) allows very large scenes to be depicted, with individual human-sized elements easily visible in very large-scale scenes.

Photography mirrors many of these choices (Hertzmann, 2022). Conventional photography rigidly creates compositions according to linear perspective rules, so that each image region looks highly realistic, at least within a normal FOV. But linear perspective photography is not the "best" way to compose a realistic image, as illustrated by content-aware and multiperspective projection techniques, which can better satisfy the DVC. Photomontage, nonlinear lenses, and computational photography techniques offer many of the same options as painting, such as curved perspective (Figure 15c), which captures a FOV at the cost of substantial distortion, and multiperspective collage (e.g., Figure 3) (Hockney's "joiners").

Discussion

How do we understand space in pictures? The evidence outlined in this paper suggests that individual fixations and distortion can be understood in terms of local linear perspectives, whereas overall scene understanding is a subtler blend of individual fixations and prior knowledge. The hypotheses here involve many open questions and predictions to be tested.

A key assertion of my hypotheses is the centrality of eye movements and foveal vision to pictorial perception, which leads to several hypotheses about distortion and peripheral vision. These hypotheses could be tested with eye-tracking studies. Some studies do suggest that shape perception in peripheral vision is distorted (Oomes et al., 2009; Baldwin et al., 2016).

The shape locality hypothesis claims that shape perception in a region does not depend on context,

apart from object recognition. This predicts, for example, that perceived shape and distortion in a region is unaffected by changing the rest of the picture, and remains stable from the first glance to later glances, so long as object recognition is unchanged. The DVC and fixation compensation predicts viewers' judgements of whether objects appear distorted, and that these judgments are independent of the rest of the picture. These make predictions for many kinds of objects and pictures.

What is a region?

A "region" is a key concept in these hypotheses, but there is insufficient evidence to specify the exact size of a region. This region may be directly related to a "foveal region," but even the concept of "foveal region" may be a convenient fiction. The distinction between foveal and peripheral vision is much subtler than often portrayed (Rosenholtz, 2016), without a clear cutoff between the two. For search tasks, the "useful FOV" depends on the stimulus (Wolfe et al., 2022). Hence, the concept of a "region" may depend on the contents of the picture itself, as well as depending on the task. There may also be no strict cutoff at all, but, rather, a decay in the precision and detail of information obtained away from a fixation.

The hypotheses formulated here rely on many concepts that could be quantified in experiments. What constitutes "normal FOV" or a normal range of FOVs? How much deviation from linear perspective can one detect in a given amount of clutter and at a given eccentricity? For example, fixation compensation predicts that the curved wall in Figure 18 looks straight when the image is zoomed in enough. How does texture affect perception of distortion? How does scale or viewing distance affect perceived shape and distortion?

Perspective in peripheral vision

What scene and shape information do viewers obtain from peripheral vision, if any? Existing evidence suggest viewers can recognize scene properties at a glance (Fei-Fei et al., 2007; Greene & Oliva, 2009), and some evidence suggests an effect of peripheral vision on foveal shape (Moore & Egeth, 1997).

Are viewers aware of distortion in peripheral vision? If not, can distortion in peripheral vision nonetheless affect task performance or scene perception? At what fields-of-view is distortion no longer detectable? How are these effects influenced by clutter and eccentricity?

What information is preserved across fixations?

There are numerous questions to study about what information is preserved across fixations in a picture, and how. How much does one remember from fixation to fixation, either consciously or not? Recent evidence suggests that at least some information is directly transferred from old to new reference frames between fixations (So & Shadlen, 2022). Are 3D space judgments fused across views? To what extent do the answers to these questions correspond to behavior in real-world 3D vision?

Holistic picture interpretation and pictorial space

If we wish to discuss how pictures as a whole are interpreted, then we also must consider where the viewer's eye fixates, and what determines this, itself a topic of considerable study and discussion. Composition and distortion are most important for salient picture regions, and an artist making a realistic picture need not be so concerned about distorted parts of a picture in places where most viewers do not look; however, composition itself affects where eyes go.

Many previous theories describe 3D scene perception in terms of *pictorial space,* (e.g., Hecht et al., 2003; Erkelens, 2016; Koenderink et al., 2016a). However, many studies have demonstrated the impossibility of a coherent visual space (e.g., Koenderink et al., 2008; Svarverud et al., 2012; Vuong et al., 2019), and so we would not expect there to be a coherent pictorial space either, especially when we consider the varieties of projections and layouts used throughout art history. Instead, there may be some pictorial space during a fixation, and a much looser sense of space that persists across fixations.

Expanding studies beyond linear perspective

Although linear perspective represents a tiny fraction of all historical imagery prior to photography, past studies of pictorial space and perspective have examined only linear perspective, with only a handful of exceptions. Future studies should study broader classes of projections that describe both painting and computational techniques. Nonlinear projections from computational photography and computer graphics could provide tools for systematic study of perspective perception.

What does it mean for a wide-angle picture to "look right?"

Smartphone photos often look very true to life—unless one directly compares them to the scene as while experiencing it (Albert & Efros, 2016), at which point one observes significant differences between the photo and one's direct perceptions of the world.

Suppose we could define an appropriate notion of whether a photo is "perceptually accurate." I support the idea that narrow-FOV pictures could be

perceptually accurate. But narrow-FOV pictures are relatively rare in many scenarios, because smartphones take wide-angle pictures by default, and art history includes many large tableaux. Narrow FOVs in art history may be more limited to close-up portraits and still lives.

For wide-angle pictures, some projections seem more accurate than others, at least in some ways. What can we say about the advantages and disadvantages of different projections, and how accurate they are to visual experience? Although there may be no "right" or "wrong" projection (Gombrich, 1961; Koenderink et al., 2016a; Hertzmann, 2022), different nonlinear projections may have benefits that can be experimentally determined. For example, natural perspective projections seem to better capture relative scale than wide-angle linear perspective (Burleigh et al., 2018), but it is unclear how to formalize them mathematically, in a perceptually grounded way.

Understanding when pictures are "honest" or "deceptive" has many societally-important implications, such as in social media and documentary photography.

Tones, colors, and shadows

The hypotheses here help explain other pictorial phenomena beyond perspective, such as the way pictures can look real despite not reproducing the extreme range of brightnesses of the real-world (Debevec & Malik, 1997; Hertzmann, 2022), and the scale of local-global decompositions (Reinhard et al., 2010; Liba et al., 2019) for tone-mapping may correspond with foveal region sizes. For example, consider Magritte's "Empire of Lights," which harmoniously combines two inconsistent lighting conditions, without the inconsistency being visible within any local region. Similarly, foveal processing could explain some difficulties in detecting inconsistent shadows. For example, Ostrovsky et al. (2005) and Jacobson and Werner (2004) describe cases where shading anomalies do not "pop out," but, rather, require visual search over a picture for a viewer to detect.

The skills in realistic drawing

A better understanding of perspective and foveal vision could provide insight to understanding the skill of drawing (Chamberlain & Wagemans, 2016). Without an internal picture representation, drawing becomes a challenging task of planning pencil movements from eye fixations, relating fixations in a scene to fixations on a page (Perdreau & Cavanagh, 2015), without the benefit of the sort of internal picture representation that a computer would have.

Figure 32 shows a sketch I drew, prior to this research, compared with a photo taken at the same time. In retrospect, this sketch suggests that I fixated on a few main objects, which because the largest elements, whereas others were crowded out to fit the picture. This suggests a multi-fixation version of the "natural perspective" phenomenon (Pepperell & Haertel, 2014; Pepperell, 2015). It could be that, we tend to draw objects very large when fixating on them; when the eye moves, relative proportions invisibly shift. As a result, getting proportions "right" is a bit like trying to stuff a pile of coiled springs into a small suitcase. In contrast, we might expect that the same expansion when viewing the picture might compensate for this effect.

Another factor in projection, evident in this sketch, is that the composition are constrained to fit the composition into the canvas size.

Why is foreshortening so hard? Why a tendency to frontality?

Two phenomena seem significant to the difficulty in drawing, and to each other. First, it is difficult to draw highly foreshortened objects, and, second, artists seem to have a tendency to tilt slanted objects toward the viewer (e.g., Schmidt, Khan, Kurtenbach, & Singh, 2009; Verstegen, 2010; Ward, 1976). I have noticed both phenomena in my own drawing: whether drawing from life or from a photograph, foreshortened objects tend to look "too flat" when I draw freely; "correcting" this flattening with more careful planning and adjustment is difficult. Perhaps these effects are related to Erkelens' finding that viewers perceive linear perspective pictures as compressed relative to the true depths of the scenes depicted (Erkelens, 2013a), or the compression of visual space into pictures (Erkelens, 2016). Eye movements may also be a factor: viewing the entirety of a large foreshortened object would typically require multiple fixations, and relating shape between the fixations in both the world and the picture is a complicated juggling act of working memory and perception, with the outcome highly sensitive to very small changes in the drawing.

New projection systems

Finally, a better understanding of the functioning of distortion and pictorial space ought to inform new projection systems, for example, for more flexible smartphone photography and computer graphics visualizations. Being able to predict when distortion occurs could enable projections to directly minimize distortion; being able to predict viewers' shape interpretation could allow direct optimizing for percepts.

Keywords: picture perception, pictorial space, art, photography, linear perspective, curvilinear perspective, peripheral vision, eye gaze, visual cognition, impossible pictures

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.

Commercial relationships: none.
Corresponding author: Aaron Hertzmann.
Email: hertzman@dgp.toronto.edu.
Address: Adobe Research, 601 Townsend St., San Francisco, CA 94103, USA.

Footnote

[1]This is calculated as follows. The photo was taken using an iPhone 13 using the ultrawide zoom setting (0.5×), which is equivalent to using focal length $f = 14$ mm on film width $w = 35$ mm. When viewing a display of width W, calculating with similar triangles gives the viewing distance $d = Wf/w = W*2/5$. (The iPhone default (1×) zoom is wide-angle: $f = 26$ mm equivalent, for which the COP distance is approximately $W*3/4$.)

References

Adams, K. R. (1972). Perspective and the viewpoint. *Leonardo, 5*(3), 209–217.

Agarwala, A., Agrawala, M., Cohen, M., Salesin, D., & Szeliski, R. (2006). Photographing long scenes with multi-viewpoint panora mas. *ACM Transactions on Graphics, 25*(3), 853–861.

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

Albert, R., & Efros, A. A. (2016). Post-post-modern photography: Capture-time perceptual matching for more faithful photographs. *Technical Report EECS-2016-167*, UC Berkeley.

Ames, A. (1925). The illusion of depth from single pictures. *Journal of the Optical Society of America, 10*(2), 137–148.

Anstis, S. M. (1974). A chart demonstrating variations in acuity with retinal position. *Visual Research, 14*(7), 589–592.

Badki, A., Gallo, O., Kautz, J., & Sen, P. (2017). Computational zoom: A framework for post-capture image composition. *ACM Transactions on Graphics, 36*(4), 1–14.

Baldwin, J., Burleigh, A., Pepperell, R., & Ruta, N. (2016). The perceived size and shape of objects in peripheral vision. *i-Perception, 7*(4), 1–23.

Barre, A., & Flocon, A. (1968). *La perspective curviligne: De l'espace visuel a´ l'image construite*. Flammarion.

Bengston, J. K., Stergios, J. C., Ward, J. L., & Jester, R. E. (1980). Optic array determinants of apparent distance and size in pictures. *Journal of Experimental Psychology: Human Perception and Performance, 6*(4), 751.

Bosten, J., Goodbourn, P., Lawrance-Owen, A., Bargary, G., Hogg, R., & Mollon, J. (2015). A population study of binocular function. *Vision Research, 110*, 34–50.

Bryan, R., Perona, P., & Adolphs, R. (2012). Perspective distortion from interpersonal distance is an implicit visual cue for social judgments of faces. *PLoS One, 7*(9), e45301.

Burleigh, A., Pepperell, R., & Ruta, N. (2018). Natural perspective: Mapping visual space with art and science. *Visionary, 2*(2), 21.

Campagnoli, C., Hung, B., & Domini, F. (2022). Explicit and implicit depth-cue integration: Evidence of systematic biases with real objects. *Vision Research, 190*, 107961.

Carroll, R., Agrawala, M., & Agarwala, A. (2009). Optimizing content-preserving projections for wide-angle images. *ACM Transactions on Graphics, 28*(3), 1–9.

Catmull, E., & Wallace, A. (2023). *Creativity, Inc. (The Expanded Edition): Overcoming the Unseen Forces That Stand in the Way of True Inspiration*. Penguin Random House.

Cavanagh, P., von Grünau, M., & Zimmerman, L. (2004). View dependence of 3d recovery from folded pictures and warped 3D faces. In Proc. 3DV.

Chamberlain, R., & Wagemans, J. (2016). The genesis of errors in drawing. *Neuroscience & Biobehavioral Reviews, 65*, 195–207.

Cohen, H. (2014). Reflections on designing and building AARON, https://www.youtube.com/watch?v=Xlhd8iP1hXo.

Cohn, N. (2012). Explaining 'i can't draw': Parallels between the structure and development of language and drawing. *Human Development, 55*(4), 167–192.

Cohn, N. (2014). Framing "i can't draw": The influence of cultural frames on the development of drawing. *Culture & Psychology, 20*(1), 102–117.

Coleman, P., & Singh, K. (2004). Ryan: Rendering your animation nonlinearly projected. In *Proceedings of the 3rd International Symposium on Non-photorealistic Animation and Rendering* (pp. 129–156).

Collomosse, J. P., & Hall, P. M. (2003). Cubist style rendering from photographs. *IEEE Transactions on Visualization and Computer Graphics, 9*(4), 443–453.

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

Cutting, J. E. (1987). Rigidity in cinema seen from the front row, side aisle. *Journal of Experimental Psychology: Human Perception and Performance, 13*(3), 323.

Debevec, P. E., & Malik, J. (1997). Recovering high dynamic range radiance maps from photographs. In *Proceedings of SIGGRAPH* (pp. 1–10).

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

Deregowski, J. B. (1989). Real space and represented space: Cross-cultural perspectives. *Behavioral and Brain Sciences, 12*(1), 51–74.

Di Luca, M., Domini, F., & Caudek, C. (2010). Inconsistency of perceived 3d shape. *Vision Research, 50*(16), 1519–1531.

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

Efros, A. A., & Leung, T. (1999). Texture synthesis by non-parametric sampling. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*. Kerkyra, Greece: IEEE.

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

Erkelens, C. J. (2013a). Computation and measurement of slant specified by linear perspective. *Journal of Vision, 13*(13), 16–16.

Erkelens, C. J. (2013b). Virtual slant explains perceived slant, distortion, and motion in pictorial scenes. *Perception, 42*(3), 253–270.

Erkelens, C. J. (2015). The perspective structure of visual space. *i-Perception, 6*(5), 2041669515613672.

Erkelens, C. J. (2016). Equidistant intervals in perspective photographs and paintings. *i-Perception, 7*(4), 2041669516662666.

Erkelens, C. J. (2018). Multiple photographs of a perspective scene reveal the principles of picture perception. *Visionary, 2*(3), 26.

Erkelens, C. J. (2021). Geometric constraints of visual space. *i-Perception, 12*(6), 20416695211055212.

Fan, J. E., Bainbridge, W. A., Chamberlain, R., & Wammes, J. D. (2023). Drawing as a versatile cognitive tool. *Nature Reviews Psychology, 2*, 556–568.

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

Fisher, D. (2024). J. M. W. Turner's Oxford High Street. https://fisherstudios.co.uk/photography-blog/j-m-w-turners-oxford-high-street/.

Fried, O., Jacobs, J., Finkelstein, A., & Agrawala, M. (2020). Editing self-image. *Communications of the ACM, 63*(3), 70–79.

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

Gibson, J. J. (1971). The information available in pictures. *Leonardo, 4*(1), 27–35.

Gibson, J. J. (1978). The ecological approach to the visual perception of pictures. *Leonardo, 11*(3), 227–235.

Goldstein, E. B. (1979). Rotation of objects in pictures viewed at an angle: Evidence for different properties of two types of pictorial space. *Journal of Experimental Psychology: Human Perception and Performance, 5*(1), 78.

Gombrich, E. H. (1961). *Art and Illusion: A Study in the Psychology of Pictorial Representation*, 2nd ed. Princeton University Press.

Gombrich, E. H. (1974). The sky is the limit. In R. B. Macleod & L. P. H., Jr., (Eds.), *The Vault of Perception and Pictorial Vision, Perception: Essays in Honor of J.J. Gibson*. Cornell University Press.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15*(1), 20–25.

Goodman, N. (1968). *Languages of Art: An Approach to a Theory of Symbols*. The Bobbs-Merrill Company, Inc.

Greenberg, G. (2021). Semantics of pictorial space. *Review of Philosophy and Psychology*.

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

Hagen, M. A. (1976). Influence of picture surface and station point on the ability to compensate for oblique view in pictorial perception. *Developmental Psychology, 12*(1), 57.

Hansen, R. (1973). This curving world: Hyperbolic linear perspective. *The Journal of Aesthetics and Art Criticism, 32*(2).

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

Held, R. T., Cooper, E. A., O'Brien, J. F., & Banks, M. S. (2010). Using blur to affect perceived distance and size. *ACM Transactions on Graphics, 29*(2), 19.

Hertzmann, A. (2022). The choices hidden in photography. *Journal of Vision, 22*(11), 10, https://doi.org/10.1167/jov.22.11.10.

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

Intraub, H., & Dickinson, C. A. (2008). False memory 1/20th of a second later: What the early onset of boundary extension reveals about perception. *Psychological Science, 19*(10), 1007–1014.

Jacobson, J., & Werner, S. (2004). Why cast shadows are expendable: Insensitivity of human observers and the inherent ambiguity of cast shadows in pictorial art. *Perception, 33*, 1369–1383.

Jahoda, G., Deregowski, J. B., Ampene, E., & Williams, N. (1977). Pictorial recognition as an unlearned ability: A replication with children from pictorially deprived environments. In G. Butterworth (Ed.), *The Child's Representation of the World* (pp. 203–217). USA: Springer.

Jin, L., Zhang, J., Hold-Geoffroy, Y., Wang, O., Matzen, K., Sticha, M., . . . Fouhey, D. F. (2023). Perspective fields for single image camera calibration. In *Proceedings of CVPR*.

Johnston, E. B. (1991). Systematic distortions of shape from stereopsis. *Vision Research, 31*(7), 1351–1360.

Juricevic, I., & Kennedy, J. (2006). Looking at perspective pictures from too far, too close, and just right. *Journal of Experimental Psychology. General, 135*, 448–461.

Keats, J. (2015). How Richard Estes makes his paintings of New York more accurate than photographs. Forbes, https://www.forbes.com/sites/jonathonkeats/2015/03/04/see-how-richard-estes-makes-his-paintings-of-new-york-more-accurate-than-photographs/.

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 M. Gayford, M. Kemp & J. Munro (Eds.), *Hockney's Eye: The Art and Technology of Depiction*. Paul Holberton.

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

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

Koenderink, J., van Doorn, A., Pinna, B., & Pepperell, R. (2016b). Facing the spectator. *i-Perception, 7*(6), 2041669516675181.

Koenderink, J. J. (1998). Pictorial relief. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 356*(1740), 1071–1086.

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

Koenderink, J. J., van Doorn, A. J., Kappers, A. M. L., & Lappin, J. S. (2002). Large-scale visual frontoparallels under full-cue conditions. *Perception, 31*(12), 1467–1475. PMID: 12916671.

Koller, M. (2004). Seamless city. http://www.seamlesscity.com/artwork.html.

Kubovy, M. (1986). *The psychology of perspective and Renaissance art*. Cambridge University Press.

Langbehn, E., Steinicke, F., Lappe, M., Welch, G. F., & Bruder, G. (2018). In the blink of an eye: Leveraging blink-induced suppression for imperceptible position and orientation redirection in virtual reality. *ACM Transactions on Graphics, 37*(4), 1–11.

Lee, J., Go, H., Lee, H., Cho, S., Sung, M., & Kim, J. (2021). Ctrl-c: Camera calibration transformer with line-classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 16228–16237, https://doi.org/10.1109/ICCV48922.2021.01592.

Levi, D. M. (2022). Learning to see in depth. *Vision Research, 200*, 108082.

Liba, O., Murthy, K., Tsai, Y.-T., Brooks, T., Xue, T., Karnad, N., . . . Marc, L. (2019). Handheld mobile photography in very low light. *ACM Transactions on Graphics (TOG), 38*(6), 1–16.

Linton, P. (2017). *The perception and cognition of visual space*. Springer.

Linton, P. (2022). Minimal theory of 3D vision: New approach to visual scale and visual shape. *Philosophical Transactions of the Royal Society B*, https://doi.org/10.1098/rstb.2021.0455.

Linton, P., Morgan, M. J., Read, J. C. A., Vishwanath, D., Creem-Regehr, S. H., & Fulvio, D. (2022). New approaches to 3D vision. *Philosophical Transactions of the Royal Society B,* 378.

Liu, S. J., Agrawala, M., DiVerdi, S., & Hertzmann, A. (2022). Zoomshop: Depth-aware editing of photographic composition. *Computer Graphics Forum, 41*(2), 57–70.

Loomis, J.M., Philbeck, J.W., & Zahorik, P. (2002). Dissociation between location and shape in visual space. *Journal of Experimental Psychology: Human Perception and Performance, 28*(5), 1202–1212.

Martin, D., Sun, X., Gutierrez, D., & Masia, B. (2023). A study of change blindness in immersive environments. *IEEE Transactions on Visualization and Computer Graphics, 29*(5), 2446–2455.

Moore, C. M., & Egeth, H. (1997). Perception without attention: Evidence of grouping under conditions of inattention. *Journal of Experimental Psychology: Human Perception and Performance, 23*(2), 339–352.

Morales, J., Bax, A., & Firestone, C. (2020). Sustained representation of perspectival shape. *Proceedings of the National Academy of Sciences of the United States of America, 117*(26), 14873–14882.

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

Oomes, A. H. J., Koenderink, J. J., van Doorn, A. J., & de Ridder, H. (2009). What are the uncurved lines in our visual field? A fresh look at Helmholtz's checkerboard. *Perception, 38*(9), 1284–1294. PMID: 19911627.

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

Ostrovsky, Y., Cavanagh, P., & Sinha, P. (2005). Perceiving illumination inconsistencies in scenes. *Perception, 34*(11), 1301—1314.

Panofsky, E. (1927). *Perspective as symbolic form*. Zone Books.

Penrose, L. S., & Penrose, R. (1958). Impossible objects: A special type of visual illusion. *British Journal of Psychology, 49*(1), 31–33.

Pepperell, R. (2011). Connecting art and the brain: An artist's perspective on visual indeterminacy. *Frontiers in Human Neuroscience,* 5.

Pepperell, R. (2015). Egocentric perspective: Depicting the body from its own point of view. *Leonardo, 48*(5), 424–429.

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

Pepperell, R., Ruta, N., & Burleigh, A. (2019). Egocentric vision in a 3D game using linear perspective and natural rendering. In Proceedings of EGOAPP.

Perdreau, F., & Cavanagh, P. (2015). Drawing experts have better visual memory while drawing. *Journal of Vision, 15*(5), 5–5.

Perkins, D. N. (1973). Compensating for distortion in viewing pictures obliquely. *Perception & Psychophysics, 14*, 13–18.

Perona, P. (2007). A new perspective on portraiture. *Journal of Vision, 7*(9), 992, https://doi.org/10.1167/7.9.992.

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

Pirenne, M. H. (1970). *Optics, painting & photography*. Cambridge University Press.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40*(1), 49–71.

Rachwani, M. (2020). Picture imperfect: Why photos of 'crowded' beaches may not be what they seem. https://www.theguardian.com/australia-news/2020/sep/13/picture-imperfect-why-photos-of-crowded-beaches-may-not-be-what-they-seem.

Rademacher, P., & Bishop, G. (1998). Multiple-center-of-projection images. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 199–206).

Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 12179–12188).

Reinhard, E., Heidrich, W., Debevec, P., Pattanaik, S., Ward, G., & andMyszkowski, K. (2010). *High dynamic range imaging: Acquisition, display, and image-based lighting*. Morgan Kaufmann, 2nd edition.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition,* 7.

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

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Review of Vision Science, 2*, 437–457.

Rosenholtz, R. (2020). Demystifying visual awareness: Peripheral encoding plus limited decision complexity resolve the paradox of rich visual experience and curious perceptual failures. *Attention, Perception, & Psychophysics, 82*(3), 901–925.

Rosinski, R. R., Mulholland, T., Degelman, D., & Farber, J. (1980). Picture perception: An analysis of visual compensation. *Perception & Psychophysics, 28*(6), 521–526.

Ruskin, J. (1857). The elements of drawing. *Three Letters to Beginners*. Wiley.

Schmidt, R., Khan, A., Kurtenbach, G., & Singh, K. (2009). On expert performance in 3D curve-drawing tasks. In *Proceedings of the 6th Eurographics Symposium on Sketch-Based Interfaces and Modeling, SBIM '09* (pp. 133–140). Association for Computing Machinery.

Schrauf, M., Lingelbach, B., & Wist, E. R. (1997). The scintillating grid illusion. *Vision Research, 37*(8), 1033–1038.

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

Schwarzkopf, D. S. (2024). What is the true range of mental imagery? *Cortex; A Journal Devoted to the Study of the Nervous System and Behavior, 170*, 21–25.

Schwiedrzik, C. M., Melloni, L., & Schurger, A. (2018). Mooney face stimuli for visual perception research. *PLoS One, 13*(7), e0200106.

Seitz, S. M., & Kim, J. (2003). Multiperspective imaging. *IEEE CG&A, 23*(6), 16–19.

Sharpless, T. K., Postle, B., & German, D. M. (2010). Pannini: A new projection for renderingwide angle perspective images. *Computational Aesthetics,* 9–16.

Shih, Y., Lai, W.-S., & Liang, C.-K. (2019). Distortion-free wide-angle portraits on camera phones. *ACM Transactions on Graphics, 38*(4), 1–12.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review, 5*(4), 644–649.

Snyder, J. P. (1993). *Flattening the earth: Two thousand years of map projections*. University of Chicago Press.

So, N., & Shadlen, M. N. (2022). Decision formation in parietal cortex transcends a fixed frame of reference. *Neuron, 110*(19), 3206–3215.e5.

Sun, Q., Patney, A., Wei, L.-Y., Shapira, O., Lu, J., Asente, P., . . . Kaufman, A. (2018). Towards virtual reality infinite walking: Dynamic saccadic redirection. *ACM Transactions on Graphics, 37*(4), 1–13.

Suppes, P. (1977). Is visual space euclidean? *Synthese, 35*(4), 397–421.

Svarverud, E., Gilson, S., & Glennerster, A. (2012). A demonstration of 'broken' visual space. *PLoS One, 7*(3), 1–9.

Todd, J. T., & Norman, J. F. (2003). The visual perception of 3D shape from multiple cues: Are observers capable of perceiving metric structure? *Perception & Psychophysics, 65*(1), 31–47.

Todorović, D. (2008). Is pictorial perception robust? The effect of the observer vantage point on the perceived depth structure of linear-perspective images. *Perception, 37*(1), 106–125.

Tyler, C. W. (2015). The vault of perception: Are straight lines seen as curved? *Art & Perception, 3*(1), 117–137.

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

Vishwanath, D. (2014). Toward a new theory of stereopsis. *Psychological Review, 121*(2), 151–178.

Vishwanath, D. (2023). From pictures to reality: Modelling the phenomenology and psychophysics of 3d perception. *Philosophical Transactions of the Royal Society B, 378*(1869), 20210454.

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

Vuong, J., Fitzgibbon, A. W., & Glennerster, A. (2019). No single, stable 3D representation can explain pointing biases in a spatial updating task. *Scientific Reports, 9*(1), 12578.

Wade, N. J., & Hughes, P. (1999). Fooling the eyes: Trompe l'oeil and reverse perspective. *Perception, 28*(9), 1115–1119.

Ward, J. L. (1976). *The perception of pictorial space in perspective pictures*. Leonardo.

Willats, J. (1997). *Art and representation: New principles in the analysis of pictures*. Princeton University Press.

Wilson, D. M. (2021). Invisible images and indeterminacy: Why we need a multi-stage account of photography. *Journal of Aesthetics and Art Criticism, 79*(2), 161–174.

Wolfe, J. M., Kosovicheva, A., & Wolfe, B. (2022). Normal blindness: When we look but fail to see. *Trends in Cognitive Science, 26*(9), 809–819.

Wood, D. N., Finkelstein, A., Hughes, J. F., Thayer, C. E., & Salesin, D. H. (1997). Multiperspective panoramas for cel animation. *Proceedings of SIGGRAPH 97,* 243–250.

Yang, T., & Kubovy, M. (1999). Weakining the robustness of perspective: Evidence for a modified theory of compensation in picture perception. *Perception & Psychophysics, 61*(3), 456–467.

Yu, J., McMillan, L., & Sturm, P. (2008). Multiperspective Modeling, Rendering, & Imaging. In T. Theoharis & P. Dutre (Eds.), *Eurographics 2008 - State of the art reports*. The Eurographics Association.

Zelnik-Manor, L., & Perona, P. (2007). Automating joiners. In *Proceedings of the 5th International Symposium on Non-Photorealistic Animation and Rendering, NPAR '07* (pp. 121–131). Association for Computing Machinery.

Zelnik-Manor, L., Peters, G., & Perona, P. (2005). Squaring the circle in panoramas. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volumes 1–2* (pp. 1292–1299). IEEE.

Zorin, D., & 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)*, pp. 257–264.