One Flight Over the Gap: A Survey from Perspective to Panoramic Vision

Xin Lin, Xian Ge, Dizhe Zhang, Zhaoliang Wan, Xianshun Wang, Xiangtai Li, Wenjie Jiang, Bo Du, Dacheng Tao, Ming-Hsuan Yang, Lu Qi

Abstract—Driven by the demand for spatial intelligence and holistic scene perception, omnidirectional images (ODIs), which provide a complete 360° field of view, are receiving growing attention across diverse applications such as virtual reality, autonomous driving, and embodied robotics. Despite their unique characteristics, ODIs exhibit remarkable differences from perspective images in terms of geometric projection, spatial distribution, and boundary continuity, making it challenging for direct domain adaption from existing perspective-based methods. In this survey, we present a comprehensive review of recent techniques in panoramic vision with a particular emphasis on the perspective-to-panorama adaptation problem. At first, we revisit the panoramic imaging pipeline and projection methods to build the prior knowledge required for analyzing the structural disparities between ODIs and perspective ones. Then, we summarize three challenges of domain adaptation including severe geometric distortions near the poles, the non-uniform sampling in Equirectangular Projection (ERP), and the periodic continuity of panoramic boundaries. Based on the discussion above, we cover 20+ representative tasks drawn from more than 300 research papers in two dimensions. On one hand, we present a cross-method analysis of representative mitigation strategies strategies for addressing panoramic distortions across different tasks. On the other hand, we conduct a cross-task comparison and classify panoramic vision into four major categories: visual quality enhancement and assessment, visual understanding, visual generation, and multimodal understanding. In addition, we discuss open challenges and future directions, emphasizing data, models, and applications that will drive the advancement of panoramic vision research. Compared to previous surveys that focused on task-specific pipelines, ours has a more unified and evolving landscape of panoramic visual learning. We hope that our work can provide new insight and forward-looking perspectives to advance the development of panoramic vision technologies. Our project page is https://insta360-research-team.github.io/Survey-of-Panorama/.

Index Terms—Panoramic Vision, Domain Gap, Projection Distortion.

1 Introduction

In recent years, computer vision techniques have made significant progress in understanding 2D perspective images, benefiting a wide range of tasks, including recognition, reconstruction, and generation, in numerous real-world applications. Driven by deep learning, many classic architectures and learning paradigms have been developed under the camera's assumptions of perspective projections, supported by publicly available datasets [1]–[4] and widespread real-world deployment [5], [6]. However, with the growing demand for immersive perception and holistic scene understanding, omnidirectional images (ODIs), which provide a complete 360° field of view, have drawn increasing attention from the research community. Compared to conventional perspective images, ODIs can provide broader spatial coverage and richer contextual information, making them indispensable for emerging applications such as virtual reality (VR) [7], autonomous driving [8], and embodied robotics [9].

Despite their potential, ODIs differ significantly from perspective images in terms of imaging geometry. As illustrated on the right side of Fig. 1, panoramic representations introduce unique challenges, including geometric distortion, boundary continuities, and uneven spatial sampling, which are especially common in standard formats such as Equirectangular Projection (ERP). These differences result in an extreme domain gap, where methods trained in perspective images often fail to generalize effectively to panoramic scenarios. The planar assumptions embedded in conventional deep models hinder their ability to handle spherical geometry and full-scene coverage, thus limiting the adaptability of perspective-based techniques and slowing progress in omnidirectional vision. Then, methods specifically designed for panoramic vision have emerged. Unless otherwise specified, we use the terms omnidirectional and panoramic to represent a 360-degree view, as both terms are widely used.

During the past decade, several surveys have reviewed specific aspects of omnidirectional vision, including 360° video streaming and compression [10], [11], visual quality assessment [12], indoor layout estimation [13], super-resolution [14], optical systems [15], and 3D perception tasks [16]–[19]. More recently, a review [20] has provided a system-level overview of deep learning applications in panoramic vision. In contrast to their structural-paradigm-based categorization, our work begins from the more fundamental perspective–panorama gap, thoroughly examining task-specific differences between perspective and panoramic representations, and systematically analyzing the resulting methodological variations. We aim to provide methodology-level insights for addressing panoramic vision tasks while integrating promising emerging technologies to broaden future research directions.

By this motivation, we investigate various ODI methods for each specific task from a perspective-to-panorama viewpoint, analyzing the strategies and efforts to bridge the domain gap from both vertical (cross-method) and horizontal (cross-task) perspec-

- *Xin Lin, Xian Ge, Dizhe Zhang, Zhaoliang Wan, Xianshun Wang, Wenjie Jiang, and Lu Qi are with Insta360 Research. Xin Lin and Ming-Hsuan Yang are with the University of California, San Diego and the University of California, Merced. Xiangtai Li and Dachenng Tao are with Nanyang Technological University. Lu Qi and Bo Du are with Wuhan University. Work was done during Xin Lin's internship at Insta360 Research. Xin Lin and Xian Ge share equal contributions. Dizhe Zhang is the project leader. Dizhe Zhang and Lu Qi are the corresponding authors.*

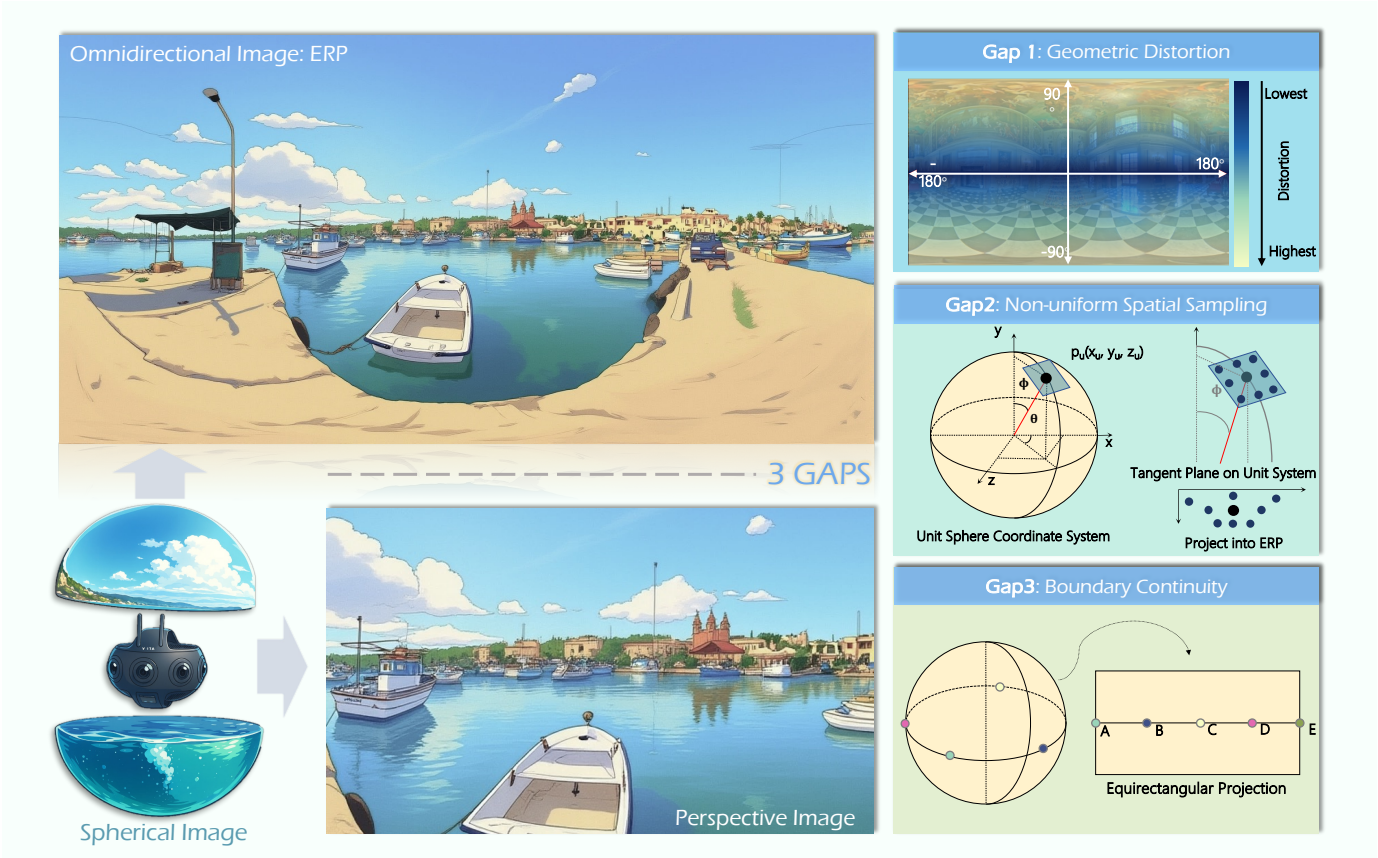

Fig. 1: From spherical image to Panoramic ERP and perspective image, ERP preserves a complete field of view compared to perspective images, but introduces three major domain gaps: (1) geometric distortion, (2) non-uniform spatial sampling, and (3) boundary continuity.

tives. Moreover, we place particular emphasis on two aspects. On one hand, we highlight ODI imaging systems and several emerging and rapidly evolving techniques, emphasizing the potential of diffusion- or auto-regressive- or 3D-reconstruction-based generative paradigms guided by ODI priors. On the other hand, the limitations of existing approaches and promising future directions are discussed. Together, these dimensions bring a holistic understanding of the methodological landscape in omnidirectional vision and uncover opportunities for innovation at the intersection of geometry, semantics, and generation.

To this end, this survey reviews over 20 representative tasks based on over 300 research papers, and is organized into several core sections, each focusing on a key component of panoramic vision. Section 2 revisits the panoramic imaging pipeline, from acquisition to stitching and projection, bringing with a clear foundation for understanding panoramic–perspective differences and supports subsequent methodological analysis. Section 3 presents the three intrinsic characteristics of ODIs which distinguish them from perspective images and reveal the roots of the domain gap, followed by a cross-method analysis of representative mitigation strategies. Section 4 conducts a cross-task comparison that synthesizes common insights and highlights methodological trends. It also identifies several rapidly evolving techniques, such as diffusion models, 3D Gaussian Splatting, and multimodal fusion, which are increasingly emerging but remain systematically unexplored in previous surveys. Last, Section 5 discusses open challenges and promising future directions, with a focus on data, models, and applications that will advance future research on panoramic vision.

As a unique modality that allows spatially comprehensive 360° perception, panoramic vision demonstrates strong potential and practical value in various applications such as spatial intelligence or immersive interaction. Through our comprehensive survey, we identify that bridging existing research gaps by transferring and adapting insights from the conventional perspective-vision domain can substantially benefit omnidirectional computer vision. We hope that our work can provide more insightful and forward-looking guidance for future research in this field.

2 Panoramic Imaging Background

This section presents background knowledge on panoramic imaging, discussing its representative imaging systems, the stitching pipeline, and widely adopted projection formats. Additional details are provided in our supplementary file.

2.1 Imaging Systems

Panoramic imaging systems capture 360° scenes for holistic perception in vision tasks. Unlike perspective cameras, they achieve ultra-wide fields of view through wide-angle refraction, mirror-based reflection, or multi-camera stitching. In this section, we introduce seven representative designs, as illustrated in Fig. 2.
Fisheye Panoramic System is an ultra-wide angle optical system with a field of view (FoV) greater than 180°. In Fig. 2(a), it employs a front group of two to three negative meniscus lenses that compress the wide object-side FoV into a narrower cone, which is then relayed by a subsequent lens group for aberration correction. Such a system has practical advantages, including compact design, simplified image acquisition, lower manufacturing cost, and improved installation stability. Since the optical path is "folded" through multiple front stage elements, fisheye optics inherently produce substantial distortion, often in the range of 15–20%, which renders distortion control a central challenge in ultra–wide-angle lens design.
Catadioptric Panoramic System is shown in Fig. 2(b), which integrates reflective and refractive optical elements to achieve

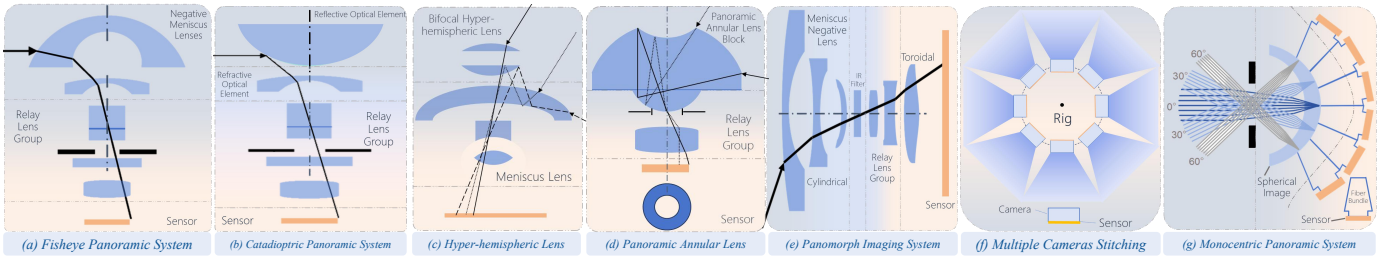

Fig. 2: Illustration of Seven Typical Panoramic Imaging Systems: Optical Designs for Capturing 360° Field-of-View.

single-viewpoint 360° capture. Incoming light is first redirected by a curved mirror and then focused onto the image sensor through a relay lens group, which also contributes to aberration correction. By combining mirror reflection to redirect ambient light into the optical path, catadioptric systems achieve an extended FoV with distortion typically above 10% while also providing broader coverage and higher spatial resolution compared to fisheye optics. However, they often suffer from a central blind spot caused by an occlusion from the mirror structure itself.

Hyper-hemispheric Lens System is illustrated in Fig. 2(c), which employs a series of meniscus lenses to capture light from the entire 360° surroundings. To address the issue of the central blind spot, a forward-view lens group is combined with a mirror to image the blind region, forming two separate optical paths with different focal lengths that include the panoramic and forward channels. By integrating dual optical paths, hyper-hemispheric systems extend the FoV to 360° × 260°, providing nearly complete spherical coverage and comprehensive ultra–wide-angle imaging.

Panoramic Annular Lens, shown in Fig. 2(d), employs a coaxial catadioptric configuration that replaces the fisheye's front lens group with a compact mirror–lens assembly. The light from the full 360° surroundings undergoes two refractions and two reflections, forming a narrow-angle beam that passes through the aperture stop and relay lens group before reaching the sensor, producing a 2D annular panoramic image based on a planar cylindrical projection. This design enables full 360° horizontal FoV with a vertical FoV of β, while significantly reducing system size and complexity compared to fisheye optics. However, the small front mirror inevitably introduces a central blind zone corresponding to a half-FoV angle α and limits its vertical completeness.

Panomorph Imaging System Fig. 2(e) enhances sensor utilization by applying spatially varying anamorphic magnification, realized through cylindrical or toroidal optics, to concentrate pixel density in user-defined regions of interest (ROIs). Compared to the traditional fisheye lenses that uniformly compress angles, Panomorph optics allocate higher resolution to targeted areas while reducing redundancy elsewhere, thus improving data compression, bandwidth efficiency, and semantic scene understanding.

Single Camera Scanning and Multiple Cameras Stitching Fig. 2(f) generate panoramas by stitching images captured from different viewpoints. They can be divided into two categories. The first method utilizes a single static camera that rotates in place to capture the entire field of view, after which the acquired images are stitched together through geometric alignment techniques. The second employs a fixed multi-camera rig, in which cameras simultaneously capture images from different directions, followed by stitching into a seamless panorama. While rotating single-camera systems are low-cost and straightforward, their long scanning process prevents real-time or gaze-based imaging. By contrast, multi-camera rigs enable real-time construction but require precise synchronization and calibration to avoid misalignment artifacts.

Monocentric Panoramic System is presented in Fig. 2(g), which adopts a spherically concentric architecture, where all optical surfaces share a common center of curvature. This design is inspired by the compound eyes of arthropods, which enable curved image sensors to be directly coupled with multiple apertures or optical fibers, resulting in a compact yet high-quality panoramic imaging configuration. By ensuring symmetric light entry, monocentric systems reduce optical aberrations and provide uniform imaging performance across ultra-wide fields of view. They are particularly well-suited for multi-aperture or fiber-coupled setups, combining compact design with high-quality imaging.

2.2 Stitching

Panorama stitching refers to the process of aligning and blending a set of images that cover a 360° view into a seamless panoramic image. In the left side of Fig. 13, we illustrate the typical pipeline, which includes data pre-processing, data association, geometric alignment, and image blending.

Data Preprocessing involves classical image signal processing (ISP) steps such as demosaicing, noise reduction, camera calibration, distortion correction, and exposure/color compensation. Noise is reduced using filters, with its intensity depending on camera parameters [21], while calibration maps 3D world coordinates to 2D pixels, with different camera types (e.g. pinhole, fisheye) requiring distinct models such as polynomial [22] or Zurich [23]. The distortion correction also compensates for the radial and tangential deviations, and motion estimation or common-view extraction may be further applied for more reliable stitching.

Data Association establishes alignment across views, typically categorized into three strategies. First, spatial association matches current and previous frames using visual features via descriptors such as SIFT [24], SURF [25], ORB [26], or LSD [27]. Second, geometric association takes advantage of epipolar constraints to reduce the search space. Finally, photometric association utilizes optical flow to extract motion-consistent feature points between frames for temporal correspondence estimation.

Geometric Alignment ensures robustness and consistency under wide FoV and nonlinear distortions, where planar homographies are often inadequate. Outlier rejection (e.g., RANSAC [28]) is essential for reliable transformation estimation, while local warping (e.g., mesh-based) addresses depth variation and parallax. Seam optimization techniques, such as graph cut or energy-minimizing seams, are then applied to refine transitions in overlapping regions.

Image Blending is the final step to ensure seamless visual transitions, compensating for color and lighting variations between images. Three mainstream strategies are widely used: (1) linear blending (feathering) achieves smooth transitions through weighted averaging in overlaps; (2) multiband blending (Laplacian pyramids) for scale-adaptive fusion and reduced ghosting; and (3)

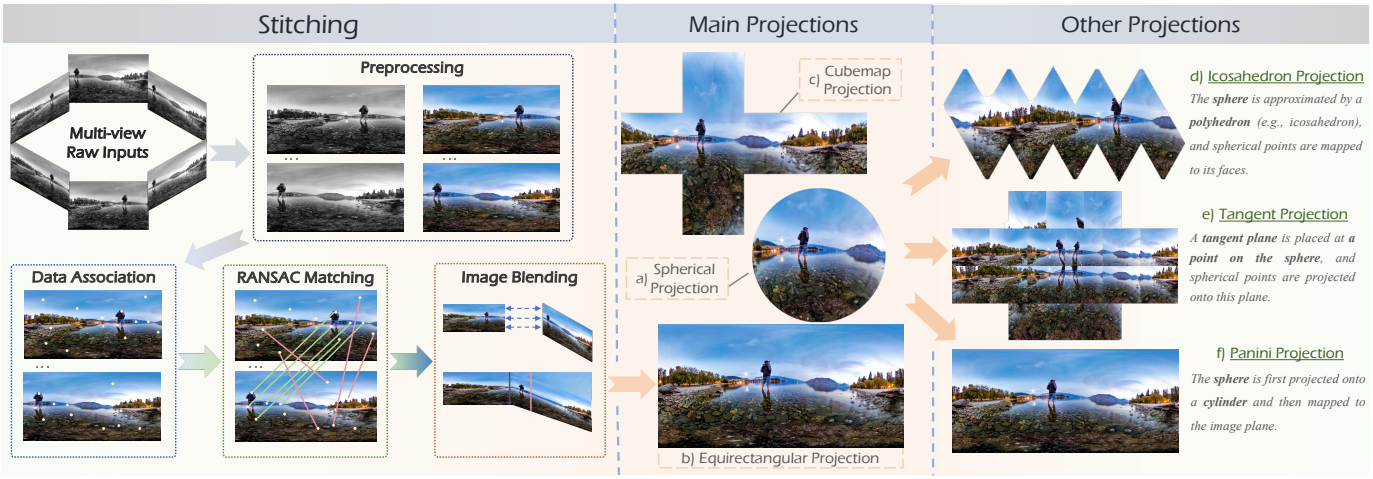

Fig. 3: Comprehensive Pipeline for Panorama Stitching: Preprocessing involves classical image signal processing (ISP) steps, including demosaicing, denoising and correction, Data Association through keypoint detection and matching, Geometric Alignment using RANSAC and homography estimation, and Image Blending with gain compensation and straightening. Representative projection methods for 360° images: (a) spherical projection as the foundational representation mapping content onto the unit sphere, from which several planar projections are derived, including (b) equirectangular projection with longitude–latitude mapping, (c) cubemap projection with six 90° perspective faces, (d) icosahedron projection with near-uniform sampling, (e) tangent projection via local planar mapping, and (f) panini projection preserving vertical lines while compressing horizontal fields.

Poisson-based fusion for global gradient consistency, with efficient variants such as MVC blending to improve speed.

2.3 Projection Formats

Projection is defined as the process of mapping spherical content to 2D formats that serves as the foundation of panoramic vision. Different projection schemes aim to balance distortion, directional continuity, and computational efficiency for specific tasks. As illustrated in the right-hand side of Fig. 13, there are three main and several other projections. On one hand, the spherical projection directly represents directions on the unit sphere, the Equirectangular Projection (ERP) is most widely adopted for its simple bijective mapping, and the Cubemap Projection (CMP) alleviates ERP's severe polar distortion by sampling along cube faces. On the other hand, more advanced designs such as Icosahedron Projection, Tangent Projection, and Polyhedron Projection further improve geometric fidelity and facilitate compatibility with perspective-based vision models. In the following, we briefly introduce those six kinds of projections due to the page limitations. A more detailed description of the projection transformations is provided in the supplementary material.

Spherical Projection. A 360° camera can be modeled as projecting all visible 3D points onto the surface of a unit sphere. This representation provides a unified, distortion-free view of all directions and serves as the foundation for panoramic imaging.

Equirectangular Projection (ERP). As the most widely used format, ERP directly unwraps spherical longitude and latitude onto a 2D plane, similar to a world map. While efficient for storage and rendering, it introduces severe distortions near the poles.

Cubemap Projection (CMP). CMP maps the sphere onto six cube faces, each covering a 90° FoV. This reduces polar distortion compared to ERP and is therefore well-suited for panoramic rendering and processing.

Polyhedron Projection (PP). PP approximates the sphere with a polyhedron (e.g., icosahedron) and maps spherical points onto its polygonal faces. Recursive subdivision of faces yields nearly uniform sampling with reduced distortion, but inevitably increases overall representation complexity.

Tangent Projection (TP). TP projects spherical content onto multiple tangent planes placed around the sphere, producing locally distortion-free patches. This enables the reuse of perspective vision models but requires precise stitching across patches.

Panini Projection. Panini projection reduces distortions of wide-angle rectilinear views ($> 70°$) by preserving vertical and radial lines while compressing the horizontal field. It provides a smooth trade-off between central magnification and edge compression.

3 Structural Challenges and Strategies

Although omnidirectional images (ODI) provide full 360° coverage for immersive perception, their structural differences from perspective images create a domain gap that hinders direct model transfer. In this section, we analyze three key structural characteristics that distinguish ODIs from perspective images: geometric distortion, non-uniform spatial sampling, and boundary continuity, as shown in the right side of Fig. 1. Then, these challenges have motivated a variety of methodological solutions. As summarized in Fig. 4(c), we organize existing approaches through a vertical cross-method analysis into four classes: (1) Distortion-aware, (2) Projection compensation, (3) Physics- or geometry-driven, and (4) Other designs, including diffusion, behavior modeling, and metric learning. Among these, the distortion-aware and projection-compensation methods are the most representative and will be summarized in this section.

3.1 Structural Challenges in ERP-based ODIs

Geometric Distortion. In ERP, unwrapping the sphere onto a 2D plane introduces distortions that increase with latitude and are most severe near the poles (±90°). As shown in Fig. 1, objects near the poles appear significantly stretched and warped, leading to an inaccurate perception of shape and structure. Such distortion limits the effectiveness of standard convolutional neural networks (CNNs), whose translation-invariant filters are ill-suited for spherical geometry. Near the poles, this assumption fails, resulting in a degraded feature extraction.

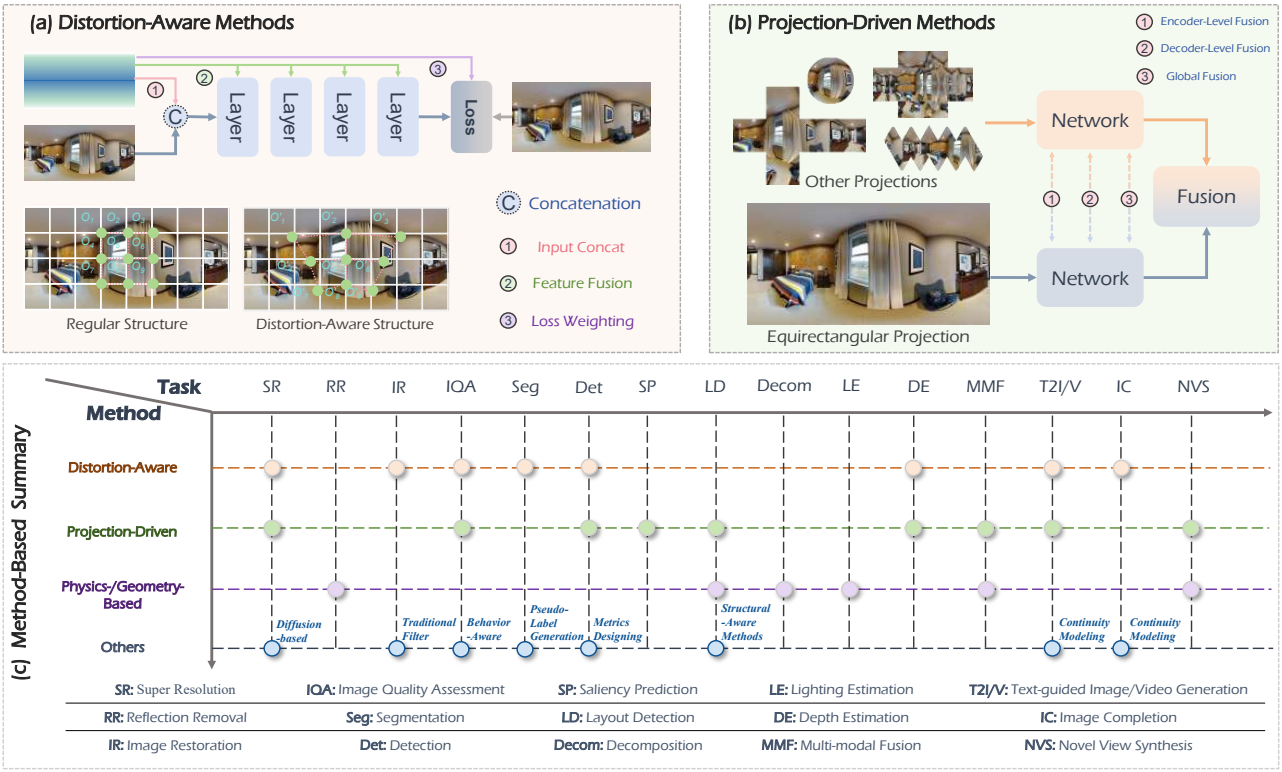

Fig. 4: Overview of representative strategies for mitigating structural challenges and their task-level summary. (a) **Distortion-Aware Methods** either leverage distortion maps (top) with three representative utilizations-input concatenation, feature fusion, and loss weighting, or design distortion-aware architectures (bottom), such as CNNs, Transformers, and diffusion models with adaptive kernels, attention, or noise initialization. (b) **Projection-Driven Methods** alleviate ERP distortions by re-projecting panoramas into alternative views (e.g., cubemap, tangent plane) and fusing multi-projection features. (c) **Method-Based Summary** highlights the task-level applicability across representative panoramic vision tasks, with distortion-aware and projection-driven methods as two core strategies.

Non-uniform Spatial Sampling. Each horizontal line of the ERP corresponds to a constant latitude on the sphere, which leads to the density of pixels to vary across different latitudes. As illustrated in Fig. 1, regions around the equator have dense and fine-grained sampling, whereas the poles exhibit sparse sampling. This leads to significant variation in the pixel-to-surface area ratio across the image. This is an imbalance that disrupts the scale invariance of visual models, as objects of the same size appear at different resolutions depending on their latitude.

Boundary Continuity. In contrast to conventional perspective images, ODIs inherently preserve boundary continuity. Specifically, in ERP, the left and right boundaries correspond to adjacent areas of the sphere, resulting in a seamless horizontal loop. This property preserves spatial continuity across the image boundaries on the spherical domain. However, conventional convolutional neural networks (CNNs), even the position embeddings in the Transformer structure, originally designed for perspective images, often treat ERP images as planar and fail to account for this horizontal continuity. As a result, visual features that span the spherical seam may be incorrectly treated as disjointed, leading to suboptimal performance near the horizontal boundaries.

3.2 Method-Aware Comparison

Building on the structural challenges discussed in subsection 3.1, we further draw a conclusion of current methods to address the inconsistency between data characteristics and task requirements. ODIs suffer from distortion, non-uniform sampling, and boundary continuity, leading different tasks to emphasize either global semantic consistency (e.g., segmentation, restoration) or local geometric precision (e.g., depth, optical flow), which in turn shapes methodological choices.

As summarized in Fig. 4, existing methodologies can be classified into four categories. Distortion-aware methods maintain the ERP format while accounting for distortions. Projection-compensation methods reduce distortion by re-projecting into alternative views. Physics- and geometry-based methods integrate priors such as lighting models or spatial layout constraints. Finally, other designs cover diffusion-based generative approaches, behavior-aware strategies, and metric-oriented frameworks. Among these, distortion-aware and projection compensation methods are the two most widely adopted categories, and the following subsections provide a detailed analysis of their strengths, weaknesses, and task-level applicability.

Distortion-Aware Methods retain the unified ERP representation and embed distortions into the network design, which are shown in Fig. 4(a). Some works [29]–[36] introduce spatially adaptive convolution kernels, where their filter shapes or receptive fields vary with latitude to account for geometric distortion. Some studies [29]–[36] introduce spatially adaptive designs in CNNs, Transformers, and diffusion architectures, where convolution kernels, attention windows, or even the initialization of generative noise are adjusted with latitude to account for geometric distortion and the non-uniform pixel distribution of panoramic images. Others [37]–[39] employ distortion maps—precomputed weight masks indicating the severity of distortion at each pixel—to guide feature learning in multiple ways. As illustrated in Fig.,

distortion maps can be (①) concatenated with the input panorama to provide pixel-wise distortion cues, (②) fused with intermediate feature layers to modulate representation learning adaptively, and (③) incorporated into the loss function as weighted penalties to emphasize errors in highly distorted regions. These strategies collectively compensate for the spatial non-uniformity inherent in ERP images. Their advantages include: (1) preserving global pixel–semantic correspondence without slicing or projection loss; (2) compatibility with CNN/Transformer/diffusion frameworks for end-to-end training; (3) flexible adaptation via deformable convolutions, re-weighted losses, or distortion maps. Limitations are: (1) residual polar distortion in ERP leading to degraded accuracy in high-deformation regions; (2) reduced robustness in geometry-sensitive tasks (e.g., depth, optical flow, keypoint matching) where precise local geometry is critical.

Projection-Driven Methods complement ERP with multiple projections that introduce less distortion, thereby alleviating its adverse effects, which are shown in Fig. 4(b). Representative examples include Cubemap Projection (CP), which reduces polar distortion by splitting the scene into six perspective views, Tangent Projection for locally distortion-free mapping, Polyhedron-based Projections (e.g., icosahedron), and Spherical Projection that directly preserves the underlying geometry. Strengths: (1) effectively suppress distortions, especially at poles and seams; (2) enable direct reuse of perspective models and large pre-trained backbones; (3) achieve stronger performance in geometry-sensitive tasks (e.g., depth, optical flow, NVS); (4) flexibility for task-specific adaptation, as different projections can be selected according to the application. Limitations: (1) fragmented information across projections, requiring additional fusion mechanisms; (2) higher computational and memory overhead from multi-view redundancy; (3) some projections demand bespoke architectures and training.

Applicability Across Tasks. Task-level analysis shows clear preferences for the two strategies: (1) *Distortion-aware methods* suit tasks demanding global semantic consistency and perceptual quality, such as super-resolution, restoration, completion, segmentation, and detection. (2) *Projection compensation methods* excel in geometry-sensitive domains—depth, optical flow, keypoint matching, novel view synthesis (NVS)—and in multi-modal fusion (e.g., LiDAR + panorama, mapping, visual odometry), where alignment with perspective modalities is crucial. (3) Some tasks admit both strategies depending on application goals. For example, super-resolution can prioritize global consistency and perceptual quality (distortion-aware) for video playback or immersive display, or emphasize local geometric fidelity (projection compensation) for architectural preservation or fine-grained reconstruction. Similarly, text-to-image/video generation benefits from distortion-aware designs for holistic semantic alignment, while projection-based schemes provide finer local control via perspective fusion. (4) In physics-driven tasks (e.g., reflection removal, parametric lighting estimation, regression-based layout detection), projection choice plays a secondary role compared to physical priors. (5) For underexplored areas (e.g., tracking, pose estimation, mapping), current evidence is too limited to establish clear preferences, highlighting directions for future study.

4 Panorama Tasks

Recent advances in panoramic vision have catalyzed a wide range of tasks across perception, understanding, and generation. As summarized in Table 5, over 20 representative tasks across four categories have been explored, ranging from low-level image enhancement to high-level scene understanding, multimodal fusion, and immersive content generation. This section gives a comprehensive review of these tasks, conducting a horizontal, cross-task analysis of the methods proposed to address the unique gaps introduced by panoramic scenarios.

4.1 Visual Quality Enhancement and Assessment

The real-world images typically suffer quality degradation due to various disturbances encountered during compression, transmission, and acquisition processes [335]–[337]. The visual quality of images significantly affects human perceptual experiences as well as the performance of subsequent downstream tasks, such as image segmentation, object detection, and 3D reconstruction. To address these issues, in addition to the imaging system-based solutions mentioned in Section 2, image quality enhancement serves as a post-processing approach after image acquisition. It aims to restore high-quality, high-fidelity images from degraded inputs by emphasizing fine details and improving the overall visual clarity. Concurrently, image quality assessment provides quantitative evaluations of both degraded and enhanced images through objective or subjective metrics, guiding the development and optimization of enhancement methodologies.

In particular, perspective images have witnessed remarkable progress in quality enhancement and assessment, benefiting from a range of learning-based techniques. However, panoramic images exhibit distinctive characteristics compared to perspective images, such as severe distortions near the poles and uneven pixel distribution. These features pose substantial barriers to the direct application of existing perspective-based techniques. Therefore, this section systematically reviews recent advancements in panoramic image quality enhancement and assessment, identifying critical technical trends and highlighting open research challenges within this emerging field. Since Section 3.2 provides a detailed analysis of the advantages, limitations, and applicability of the two common strategies, this section presents only a brief discussion for the strategies. For task-specific future directions, please refer to the supplementary material.

4.1.1 Super Resolution

Super-resolution aims to reconstruct a high-resolution image or video from one or more low-resolution inputs by restoring fine details and enhancing visual quality. Some early methods [52], [53] extend the existing perspective-based models [338], [339] with panorama data However, their performance is significantly constrained by model designs that fail to account for the unique characteristics distinguishing panoramic from perspective ones. To address these issues, advanced methods can be categorized into three groups:

Distortion-Aware Methods are designed to address the non-uniform pixel distribution [40] in ODI-SR by introducing distortion-aware priors or adaptive weighting schemes that emphasize perceptually important equatorial regions. The proposed methods with hierarchical modeling (LAUNet [40]), distortion-aware priors (360-SISR [37]), weighted loss designs (OSR-GAN [41]), distortion-aware convolutions and adaptive losses (OSRT [29], An et al. [42], Sun et al. [38]), pixel-wise weighting and distortion-guided attention (FATO [43], GDGT-OSR [44]), and more advanced transformer-based schemes leveraging geometric, semantic, and frequency cues (Cao et al. [45], Shen

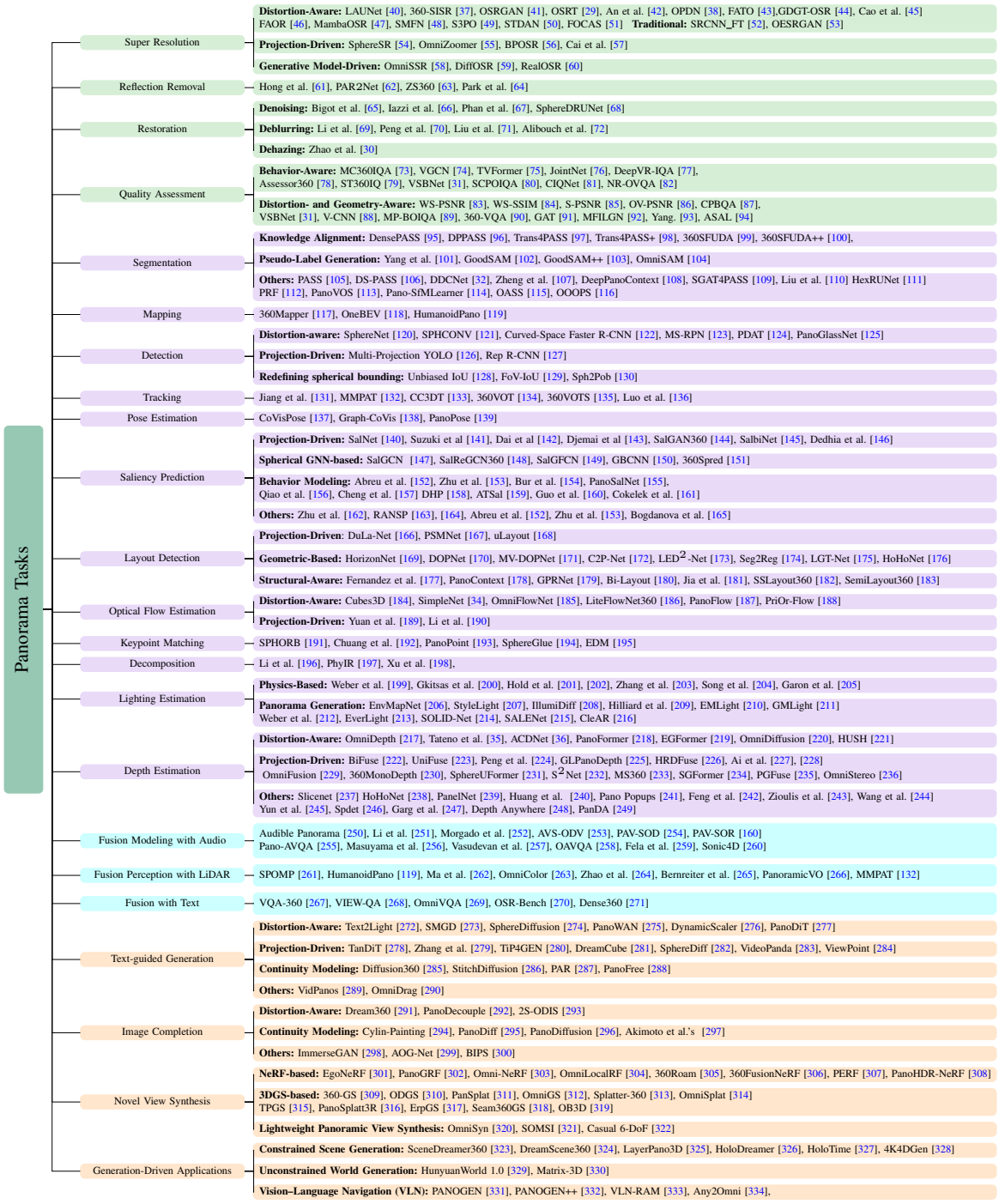

Fig. 5: Summary of Panoramic Vision Tasks and Representative Methods.

et al. [46], Wen et al. [47]). For video SR, works such as SMFN [48], S3PO [49], and STDAN [50] adopt latitude-aware losses to emphasize equatorial regions and temporal-aware modules for spatiotemporal coherence. Central vision-based SR has also been explored in FOCAS [51], leveraging foveated rendering to highlight human central vision.

Projection-Driven Methods leverage alternative projections to reduce ERP distortions, enabling super-resolution models to better pursue local geometric precision and structural fidelity. Some works include continuous spherical modeling with spherical CNNs on icosahedral grids (SphereSR [54]), Möbius projection with spatially adaptive resampling for localized high-precision upsampling (OmniZoomer [55]), dual-branch geometric alignment for structural consistency (BPOSR [56]), and pseudo-cylindrical representations for adaptive latitude sampling compatible with standard 2D SR networks (Cai et al. [57]).

Generative Model-Driven Approaches leverage the strong priors of generative foundation models, particularly diffusion models, to handle unknown and complex degradations and improve generalization in panoramic SR. Some works include ERP-to-TP projection interaction with gradient decomposition correction for detail recovery (OmniSSR [58]), stepwise sampling to approximate high-resolution distributions and reduce texture blurring (DiffOSR [59]), and efficient realistic SR with single-step sampling and unfolding-guided injector for complex degradations (RealOSR [60]).

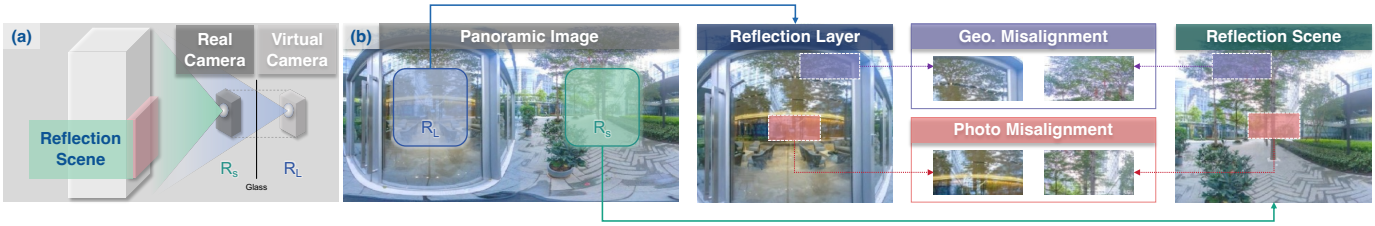

Fig. 6: Reflection removal challenges in panoramic images. The reflection scene (R_s, green) and reflection layer (R_l, blue) present geometric misalignment and photo misalignment, which make transmission-reflection separation more difficult.

Overall, distortion-aware methods emphasize global consistency and perceptual quality, while projection-driven methods prioritize local geometric fidelity. In contrast, generative model-driven approaches remain at an early stage; they still suffer from high computational cost and slow inference.

4.1.2 *Reflection Removal.*

Reflection removal aims to recover the true transmission scene from glass-mixed inputs, but traditional perspective-based methods often fail when applied to panoramic images where reflections can be sharp and complex, as shown in Fig. 6. Most perspective-based methods assume the whole image is glass-mixed with weak and blurred reflections, which often fails in 360 panoramas, making transmission–reflection separation particularly challenging.

Recent panorama-based approaches leverage the unique characteristic that both the reflective scene and the mixture image within the same frame: introducing external priors from visible reflection sources (Hong et al. [61]), unifying correspondence modeling, transmission recovery, and reflection refinement in end-to-end designs (Par^{2}Net [62]), extending to zero-shot reflection removal via iterative geometric matching and disentanglement (ZS360 [63]), and developing fully automated frameworks with long-range dependency modeling and multi-scale alignment for robust performance (Park et al. [64]).

4.1.3 *Omnidirectional Image Restoration*

Omnidirectional image restoration aims to restore high-quality 360° images from their degraded counterparts, which can include noise, blur, and weather distortions. Here, we review methods for main image restoration tasks, including denoising, deblurring, and dehazing. In the following sections, we will highlight how each approach leverages the spatial geometry of omnidirectional data to achieve robust performance.

(1) Image Denoising: Early methods adapt classical filters to spherical geometry, including Wiener filtering, Tikhonov regularization, and Stein block thresholding [65], [66]; panorama-adapted designs introduce space-variant total variation to model ERP-specific distortions [67]; and recent learning-based approaches like SphereDRUNet leverage uniform HEALPix sampling to perform data-driven restoration directly on the sphere [68].

(2) Image deblurring: Only some traditional methods are proposed for omnidirectional images, adapting classical strategies to address geometric and optical challenges. Some approaches include projection-based priors such as Omnigradient, which incorporates cylindrical gradient regularization into deconvolution [69]; hardware-based solutions like coded apertures to capture all-focus information and mitigate defocus blur [70], [71]; and spherical-domain filtering methods such as harmonic-based Wiener filtering on the 2-sphere [72].

(3) Image dehazing: To address panoramic dehazing, Zhao et al. [30] propose a distortion-aware convolution to handle ERP-induced distortion. Their end-to-end framework jointly performs dehazing and depth estimation, establishing a strong baseline for adverse-weather restoration in omnidirectional scenes.

4.1.4 *Visual Quality Assessment.*

Visual Quality Assessment aims to quantitatively evaluate the perceptual quality of panoramic images and videos. Based on whether reference data are used as constraints, these methods can be categorized into full reference (FR) and no-reference (NR) approaches. Unlike conventional IQA and VQA, which often assume uniform visibility and regress a single global score, panoramic quality assessment faces unique challenges arising from ERP distortions near the poles and from user-dependent viewports that expose only localized regions at a time. To address these issues, two representative strategies have recently emerged.

Behavior-Aware Methods aim to reflect human subjective perception by simulating localized viewing on head-mounted displays and integrating behavioral cues such as eye movement, head movement, and saliency. Some works include shared CNNs and graph reasoning for inter-viewport relationships (MC360IQA [73], VGCN [74]); sequence models for temporal scanning and memory effects (TVFormer [75], JointNet [76]); behavioral priors such as viewing coordinates, adversarial learning, and trajectory/saliency-guided weighting (DeepVR-IQA [77], Assessor360 [78], ST360IQ [79]). Recently, fidelity-enhanced designs have been proposed to align distorted content with pseudo-references or to fuse predicted saliency into score aggregation (VSBNet [31], SCPIQA [80]). For VQA, perceptual- and causal-aware models for robust quality estimation (CIQNet [81], NR-OVQA [82]). While effective under uniformly distributed distortions, these methods still struggle to handle the spatially complex and uneven distortion patterns.

Distortion- and Geometry-Aware Methods explicitly address spatial non-uniformity and geometric distortions through latitude-aware designings (WS-PSNR [83], WS-SSIM [84], S-PSNR [85], OV-PSNR [86]); saliency- and viewport-weighted assessment with equidistant convolutions for perceptual fidelity (CPBQA [87], VSBNet [31], V-CNN [88]); multi-projection and statistical or dynamic modeling for distortion distribution fitting (MP-BOIQA [89], 360-VQA [90]); hierarchical graph attention module (GAT [91]), weakly supervised frequency domain evaluation via wavelet decomposition and NSS statistics (MFILGN [92]); and large-scale benchmarks with BLIP-2-based modeling to jointly predict perceptual quality and degraded regions for AIGC content (Yang et al. [93]), and continual learning approaches tackling cross-dataset generalization and catastrophic forgetting (ASAL [94]).

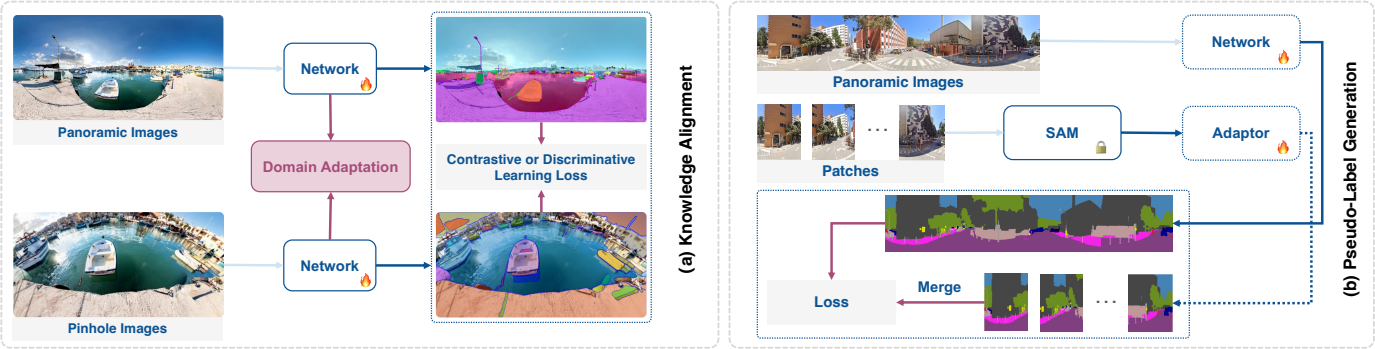

Fig. 7: UDA strategies for panoramic segmentation. (a) Knowledge Alignment: enforces semantic consistency via adversarial learning and prototypical adaptation. (b) Pseudo-Label Generation: derives supervision from ensemble predictions, SAM-based rectification, and fine-tuning with FoV-aware prototypes and dynamic labeling.

4.2 Visual Understanding

Panoramic visual understanding tasks encompass a wide range of objectives, including both high-level semantic interpretation and geometry-centric structural perception. To provide a clear task definition and methodological developments, we categorize existing tasks into two main groups: (1) **Semantic-Level Tasks** (e.g., object segmentation, detection, saliency prediction, visual tracking), which emphasize region- or pixel-level recognition based on semantic priors; and (2) **Structure and Motion-Oriented Tasks** (e.g., layout, optical flow, pose, lighting, and keypoint estimation), which focus on 3D structure, spatial layout, motion, and illumination. Unlike low-level enhancement, these tasks require higher-level perception, where spatial consistency, geometric invariance, and semantic awareness are critical. However, panoramic distortions and view-dependent semantics pose fundamental challenges that drive the need for specialized representations and learning strategies.

4.2.1 Object Segmentation

Object segmentation aims to segment key regions in an image by assigning category labels and, in some cases, instance identities to each pixel. While perspective segmentation [340] has achieved remarkable progress with large-scale datasets and deep neural networks, panoramic images represented in ERP format introduce new challenges due to polar distortions and violations of planar assumptions. Early methods project ERP images into multiple perspective-like patches and segment them individually before fusion (PASS [105], DS-PASS [106]), which suffers from loss of global context, boundary inconsistencies, and redundant computation. To overcome these limitations and reduce reliance on scarce panoramic annotations, recent studies have explored Unsupervised Domain Adaptation (UDA) to transfer knowledge from perspective to panorama. Existing approaches are broadly categorized into two groups, as shown in Fig. 7.

Knowledge Alignment focuses on transferring semantic consistency from the source perspective domain to the target panoramic domain, typically through explicit or implicit alignment mechanisms. Explicit strategies enforce domain invariance via adversarial learning, focusing on global–local consistency (DensePASS [95]) or aligning ERP and perspective branches for cross-domain generalization (DPPASS [96]). Implicit strategies adopt prototypical adaptation, such as self-supervised Mutual Prototypical Adaptation (Trans4PASS [97]) and Segment Anything Model (SAM) [341] enhanced prototype correction for robust supervision (Trans4PASS+ [98]). To further mitigate projection distortion and style inconsistencies, multi-projection fusion approaches (360SFUDA [99], 360SFUDA++ [100]) integrate ERP, FFP, and TP views for multi-level alignment of predictions, prototypes, and features, effectively bridging semantic, geometric, and stylistic gaps.

Pseudo-Label Generation produces supervision signals for unlabeled panoramic images by using pretrained models from other domains. Early approaches generate multiple predictions from transformed panoramic inputs and ensemble them into pseudo-labels (Yang et al. [101]); SAM-based methods leverage zero-shot masks refined with Teacher Assistant guidance and distortion-aware rectification for improved accuracy at lower cost (GoodSAM [102], GoodSAM++ [103]); and recent advances fine-tune SAM2 [342] with LoRA layers, introducing FoV-based prototypical adaptation and dynamic pseudo-labeling to handle distortion, incompleteness, and domain gaps (OmniSAM [104]). Overall, leveraging the strong generalization ability of large pretrained models provides a promising paradigm for more reliable panoramic segmentation under limited supervision.

Others. There are also some more specialized designs: distortion convolution [32], bi-directional learning module [107], multi-task learning [108], spherical deformable embedding [109], dual-branch designing [110], and unfolded icosahedron mesh [111]. In addition, with the development of perspective segmentation, new tasks like panoramic panoptic segmentation [112], video segmentation [113], self-supervised segmentation [114], occlusion-aware seamless segmentation [115], and open-vocabulary panoramic segmentation [116] have also been proposed.

4.2.2 Semantic Mapping

Semantic mapping converts egocentric panoramic inputs into bird's-eye view (BEV) representations, emphasizing spatial localization of objects rather than pixel-level segmentation. Recent works include introducing intermediate projections and distortion-aware indexing (360Mapper [117]); deployment-oriented designs learn direct ERP-to-BEV mappings for efficiency (OneBEV [118]); and fusion-based approaches extend to panoramic–LiDAR integration to mitigate occlusion and FoV limitations in robotics (HumanoidPano [119]).

4.2.3 Object Detection

Object detection locates and classifies object instances by predicting bounding boxes and class scores. Unlike semantic segmentation, detection in panoramas faces unique challenges, as

(a) FoV Boudning Boxes

(b) Planar Bounding Boxes

(c) Comprehensive Comparison between Different Criteria

	Metrics	Sph-IoU	Fov-IoU	Unbiased	Sph2Pob
Consistency ↑	R_{all} / R_{low} / R_{high}	0.7819 / 0.9922 / 0.4274	0.9600 / 0.9974 / 0.8860	1.00 / 1.00 / 1.00	0.9989 / 0.9990 / 0.9988
Time-cost ↓	T_{cpu} / T_{cuda}	0.0364 / 0.0033	0.0372 / 0.0034	46.4417 / -	2.2275 / 0.0096
Detection ↑	AP / AP_{50} / AP_{75}	10.7 / 24.3 / 7.8	10.9 / 25.0 / 7.9	- / - / -	11.5 / 25.7 / 8.2

Fig. 8: Bounding box representations and evaluation criteria for panoramic object detection. (a) Field of view (FoV) bounding boxes and (b) planar bounding boxes. (c) Metric comparison in consistency, computational cost, and detection accuracy.

conventional rectangular boxes fail to represent object shapes and positions under ERP distortions, particularly near the poles, as illustrated in Fig. 8(a)(b). To address this, panoramic detection adopts spherical bounding boxes, angular bounding boxes, or polygon/mask-based annotations that better align with object geometry on the panoramic images. Existing methods can be broadly categorized into three types.

Distortion-Aware Methods explicitly embed spherical priors into the detection model. Approaches include spherical convolutions that define operations on tangent planes or reparameterized kernels to preserve consistency (SphereNet [120], SPHCONV [121]), distortion-aware augmentations with multi-scale kernels or lightweight modules (Curved-Space Faster R-CNN [122], MS-RPN [123]), and task-specific adaptations such as dynamic token partitioning or deformable context modules (PDAT [124], PanoGlassNet [125]).

Projection-Driven Methods mitigate spherical distortion by decomposing panoramas into multiple distortion-free perspective views, enabling reuse of conventional 2D detectors. Some works, such as Multi-Projection YOLO [126] and Rep R-CNN [127] refine detection through bounding box adjustment, soft selection, and reprojection-based RoI alignment.

Overall, the above two strategies achieve end-to-end panoramic detection through architectural design, but they overlook the proper representation of spherical bounding boxes, which is a fundamental distinction of panoramic detection.

Redefining Spherical Bounding Boxes and IoU Metrics builds unbiased spherical representations to better capture object geometry in panoramas. Some metrics directly model overlap on the sphere via spherical rectangles or great-circle angles (Sph-IoU [343], FoV-IoU [129], and Unbiased IoU [128]), eliminating the approximation errors of planar IoU. As shown in Fig. 8(c), these spherical metrics achieve higher consistency and more faithful evaluation but often come with increased computational cost. In contrast, transformation-based approaches such as Sph2Pob [130] map spherical boxes to rotated planar ones, reducing complexity and offering competitive accuracy, though with limited geometric fidelity.

4.2.4 Visual Tracking

Visual tracking aims to continuously track the spatial positions and states of objects in video sequences. The unique panoramic characteristics from spherical projection lead to structural degradation when directly applying perspective-based algorithms, making it difficult to build consistent spatiotemporal continuity. Early efforts enhance object continuity with hardware-driven active vision systems [131]. Cross-modal approaches integrate panoramic images with LiDAR point cloud for geometry-aware detection, association, and bipartite matching (MMPAT [132]). Multi-camera settings are also extended to panoramic space through pre-fused 3D detections for consistent identity assignment (CC3DT [133]). To mitigate distortion and boundary issues, spherical representations and metrics such as BFoV, eBFoV, and Sdual are specially designed to support new 360° benchmarks [134], [135]. Recent frameworks unify tracking-by-detection and end-to-end paradigms under panoramic designs, enhancing association and representation with trajectory feedback and segmentation cues (Luo et al. [136]).

4.2.5 Pose Estimation

Pose estimation aims to estimate the 6-DoF transformation between images. While geometry-based feature matching methods (e.g., SIFT, ORB) have been widely applied in perspective tasks like SLAM and SfM, panoramic images suffer performance drops in feature detection, matching, and modeling due to their unique characteristics distinct from perspective images. Recent solutions explore different paradigms: co-visibility-based approaches enhance matching and extend pairwise estimation to multi-view optimization (CoVisPose [137], Graph-CoVis [138]), while self-supervised strategies leverage photometric losses and rotation-only pretraining to handle large viewpoint differences without ground truth poses (PanoPose [139]).

4.2.6 Saliency Prediction

Saliency prediction aims to simulate the distribution of human visual attention under spherical viewing conditions. Early traditional methods [154], [165] rely on low-level features such as intensity, color opponency, and corner responses. With the success of learning-based saliency models on perspective images, transferring them to omnidirectional scenarios remains challenging due to ERP distortions, the scarcity of large-scale 360° benchmarks, and optimization difficulties. Existing approaches can be broadly categorized into three types:

Projection-Driven Methods directly project panoramas into Cubemap Projections (CP). Some early works [140]–[143] adapted conventional 2D saliency models under this paradigm and fused the cubemap-based results into the original panoramas. These methods effectively avoid ERP-specific distortions with well-trained perspective networks, but fail to bridge the feature-level gap. Subsequent research noticed this challenge and proposed methods with integration of both ERP-based global features and CP-based local features. These approaches [144]–[146] typically employ dual-branch networks or separated pipelines, where

predictions from ERP and CP are fused via weighted averaging or spherical-domain optimization.

Spherical Graph Neural Network-based Methods build graph structures directly on the sphere and apply graph convolutional networks (GCNs) to preserve geometric continuity and spatial adjacency without relying on specific projections [147]–[150]. They typically sample uniformly on the spherical surface (e.g., geodesic grids or superpixels), where saliency reasoning is performed by propagating features across nodes in geodesic space. 360Spred [151] further incorporates spherical optical flow and 3D separable graph convolution to jointly capture spatial–temporal saliency patterns.

Behavior Modeling Methods integrate viewing priors such as equator bias, head–eye motion, or scanpaths, often guided by eye-tracking or viewport trajectories [152], [153]. For video, behavior-aware approaches further incorporate head-movement prediction, FoV dynamics, and viewport biases via sequence models or multi-task learning [155], [156], leverage cube-based spatial–temporal designs for geometric continuity [157], simulate scanpaths with reinforcement learning [158], or fuse global–local cues in dual-stream architectures [159]. More recent multimodal frameworks align audio–visual attention for fine-grained saliency [160], [161].

Others. These approaches operate directly on ERP images and introduce novel modules such as attention mechanisms, spherical-specific strategy. Lightweight solutions improve efficiency and discriminative ability through mobile backbones [162], dynamic convolutions, or ranking-based attention [163], [164]; spherical U-Net [33] defines convolution kernels on spherical crowns to preserve geometric fidelity.

4.2.7 Layout Detection

Layout detection aims to recover the structural boundaries of indoor scenes, including walls, floors, and ceilings from panoramic images. However, severe geometric distortions violate traditional assumptions such as planarity and linearity. To address this, existing methods can be broadly categorized into three types:

Projection-Driven Methods align perspective and panoramic views within a unified spatial framework, reducing inconsistencies across representations. Some approaches include transferring features via dual-branch transformation (DuLa-Net [166]), integrating cross-view projection with a stereo transformer (PSM-Net [167]), and converting perspective inputs into ERP for unified processing (uLayout [168]).

Geometric-Based Methods adapt to panoramic geometry by converting 2D layouts into 1D horizon sequences for efficiency (HorizonNet [169]) and decoupling modeling along orthogonal planes to resolve spatial ambiguities (DOPNet [170], MV-DOPNet [171]). Building on this principle, C2P-Net [172] models components along principal axes and compresses them into 1D representations to enhance spatial reasoning without perspective priors. Differentiable geometric modules further bridge layout and depth (LED^2-Net [173], Seg2Reg [174]), while distortion-aware encodings improve spatial consistency (LGT-Net [175]). Unified frameworks like HoHoNet [176] combine sparse 1D and dense 2D predictions in a shared latent space, facilitating multi-task transfer across layout, depth, and semantics.

Structural-Aware Methods improve the robustness and generalization by integrating geometric priors, ambiguity modeling, and data-efficient strategies. Early works adopt classical cues such as lines and vanishing points [177], [178], while GPR-Net [179] leverages learned geometric tokens and multi-view correspondences for layout registration and pose regression. To address annotation ambiguity in existing datasets, dual-branch frameworks like Bi-Layout [180] jointly predict closed and open layout types with cross-attentive refinement. Beyond Manhattan-world assumptions, normal-aware pipelines [181] adaptively reconstruct 3D structures under mixed constraints. Meanwhile, data-efficient methods further reduce labeling demands through semi-supervised learning [182] or structure-aware perturbations [183], enabling layout estimation in low-resource scenarios.

4.2.8 Optical Flow Estimation

Optical flow estimation aims to compute dense motion fields between panoramic video frames. Conventional assumptions such as brightness constancy and spatial smoothness often fail under severe panoramic distortions. Early work like Cubes3D [184] adapts deep optical flow models with custom projection models and synthetic data, highlighting domain gaps between perspective and panoramic settings. Recent methods can be broadly divided into two categories.

Distortion-Aware Methods adapt convolutional operations to accommodate the geometric distortions for more accurate pixel-wise motion estimation. The proposed methods with deformable convolutions (SimpleNet [34]), distortion-aware kernels aligned with spherical geometry (OmniFlowNet [185]), angular-aware kernels (LiteFlowNet360 [186]), flow-specific augmentations with cyclic estimation (PanoFlow [187]), and ortho-driven distortion compensation (PriOr-Flow [188]).

Projection-Driven Methods integrate motion fields from multiple projection domains. The proposed methods utilize a gnomonic-based cubemap and icosahedron fusion, employing off-the-shelf models (Yuan et al. [189]), and a joint learning approach for equirectangular, cylindrical, and cubemap flows through a fusion network (Li et al. [190]).

4.2.9 Keypoint Matching.

Keypoint matching aims to extract repeatable keypoints, compute descriptors on the sphere, and establish correspondences invariant to spherical rotations and robust to ERP distortions. Traditional methods mitigate distortions by locally approximating the sphere as planar, e.g., SPHORB [191] uses geodesic grids for uniform sampling, and Chuang et al. [192] project local patches onto tangent planes for accurate description. Learning-based approaches embed spherical geometry into models, such as PanoPoint [193], which performs detection/description on ERP images, and SphereGlue [194], which leverages spherical graphs with Chebyshev convolutions. More recently, EDM [195] introduces a spherical spatial alignment module with geodesic refinement, enabling globally coherent yet locally precise dense correspondences.

4.2.10 Decomposition

Decomposition separates 360° imagery into interpretable components such as lighting, reflectance, and geometry, enabling realistic understanding and content editing in immersive settings. The proposed methods with stereo-based full-scene illumination and multi-scale learning (Li et al. [196]), physics-driven decomposition via differentiable rendering (PhyIR [197]), and spherical-constrained optimization with intrinsic priors (Xu et al. [198]).

4.2.11 Lighting Estimation

Lighting estimation aims to recover high-fidelity environmental illumination, supporting tasks such as inverse rendering, relighting, object insertion, and augmented reality. The goal is typically to create an HDR environment map or a compact set of spherical harmonics (SH) coefficients that capture the global scene illumination. Early approaches directly infer illumination from full 360° panoramas [199], [200], but collecting large-scale panoramic data is impractical. Recent studies instead estimate lighting from limited observations, such as perspective images or object-centric RGB-D views, and can be broadly categorized into two main directions.

Physics-Based Methods directly regresses compact parameter sets such as sun position, sky turbidity, or spherical harmonics (SH) coefficients, using physically based sky models, autoencoder-learned codes, or CNN predictors. Outdoor-oriented approaches target natural illumination [201]–[203], while others address complex indoor lighting with spatial variations [204], [205]. Geometry and reflectance priors, often combined with differentiable rendering layers, provide supervision and support photorealistic relighting.

Panorama Generation Methods predict full 360° environment maps as an intermediate or final lighting representation, providing more realistic relighting and greater editing flexibility than parameter-only regression. Generative approaches reconstruct HDR panoramas from partial observations using GANs or diffusion models (StyleLight [207], HDRGAN [206], IllumiDiff [208], Hilliard et al. [209]); physically inspired methods integrate geometric parameterizations or scene cues with editable HDR illumination (EMLight [210], GMLight [211], Weber et al. [212], EverLight [213]); and structured pipelines adopt intrinsic decomposition, hierarchical transformers, or latent-diffusion refinement (SOLID-Net [214], SALENet [215], CleAR [216]). Collectively, these strategies enhance realism, editability, and consistency in panoramic lighting estimation.

4.2.12 Depth Estimation

Depth estimation aims to infer per-pixel scene distances (or disparities) from images, producing dense 3D structural representations. In panoramas, spherical geometry and severe ERP distortions make this task more challenging than in perspective settings. The first learning-based approach, OmniDepth [217], attempts to address this perspective-to-panorama gap by introducing a dedicated dataset and a distortion-aware network, subsequent methods can be broadly categorized into four paradigms.

Distortion-Aware Methods for depth estimation include distortion-aware convolutions that adapt kernel sampling to non-uniform resolution (Tateno et al. [35], ACDNet [36]), Transformer-based architectures with spherical encodings or attention (PanoFormer [218], EGFormer [219]), diffusion frameworks modeling global structure probabilistically (OmniDiffusion [220]), and harmonic-space representations capturing frequency-domain topology (HuSH [221]).

Projection-Driven Methods for panoramic depth estimation include:

(a) Multi-Projection Fusion. Some works include BiFuse [222] and UniFuse [223] with dual-branch ERP–cubemap fusion via learnable masks and distortion-aware padding, perspective-view synthesis for local refinement [224], cubemap vision transformers [225], and feature-alignment modules [226]. More advanced designs such as Elite360D/M [227], [228] integrate ERP and spherical grids (e.g., ICOSAP) with attention mechanisms and multi-task objectives for improved efficiency and generalization.

(b) Projection Transformation. Instead of fusing views, these methods directly replace ERP with distortion-reduced projections. Tangent patches [229], [230], icosahedral meshes [231], and HEALPix-sampled spheres [232] enable uniform spatial reasoning and accurate depth regression. Further improvements include patch-wise fusion with geometry-aware modules [229], [233], spherical transformers [234], frequency-domain designs such as Gabor-based priors [235], and Cassini-based omnidirectional stereo [236] for geometry-preserving multi-view depth.

Other Methods. Beyond structure-aware modeling and projection-based strategies, several complementary directions further advance panoramic depth estimation. Slice-based representations partition ERP into gravity-aligned slices or panels to exploit indoor regularities (SliceNet [237], HoHoNet [238], PanelNet [239]); multi-task learning jointly predicts depth with normals or planar boundaries to enforce geometric consistency (Huang et al. [240], Eder et al. [241], Feng et al. [242]); self- and weakly supervised methods leverage spherical view synthesis or stereo photometric cues for scalable training on unlabeled panoramas [243]–[247]; and pretraining/foundation adaptation transfers large perspective-based models (e.g., Depth Anything) to panoramas via projection conversion, pseudo-label distillation, and Möbius or equator-aware augmentations [248], [249].

4.3 Multi-modal Understanding

These works extend panoramic perception beyond vision-only models by integrating complementary modalities such as audio, LiDAR, and text. With the growing adoption of 360° images and videos in VR/AR, this fusion enables richer semantic understanding and more human-aligned multimodal perception and generation.

4.3.1 Audio–Visual Fusion

These methods leverage spatial audio to complement panoramic vision with orientation cues and event-specific signals, enabling richer 360° scene understanding beyond vision alone. Unlike perspective settings, panoramic scenarios require spatialized audio over the full sphere with continuity and distortion handling. Recent research spans four representative directions.

(a) Spatial Audio Synthesis. Methods generate ambisonic sound aligned with panoramas from static or mono inputs, using depth cues, geometric simulation, or end-to-end audio–visual networks (Audible Panorama [250], Li et al. [251], Morgado et al. [252]).

(b) Audio–Visual Attention Modeling. Large-scale eye-tracking datasets and cross-modal saliency frameworks demonstrate how spatial audio guides visual focus under ERP and cubemap projections (AVS-ODV [253], PAV-SOD [254], PAV-SOR [160]).

(c) Semantic Reasoning and Sound Source Localization. Panoramic tasks require spherical encodings and long-range reasoning for object-level auditory understanding, with methods addressing AVQA, localization, and semantic prediction via spatial encodings, self-supervision, or multi-task learning (Pano-AVQA [255], Masuyama [256], Vasudevan et al. [257]).

(d) Perceptual Quality and 4D Fusion. Audio-aware models outperform vision-only baselines in panoramic quality assessment [258], [259], while Sonic4D [260] integrates video-to-audio generation, grounding, and room simulation for realistic 4D interactive experiences.

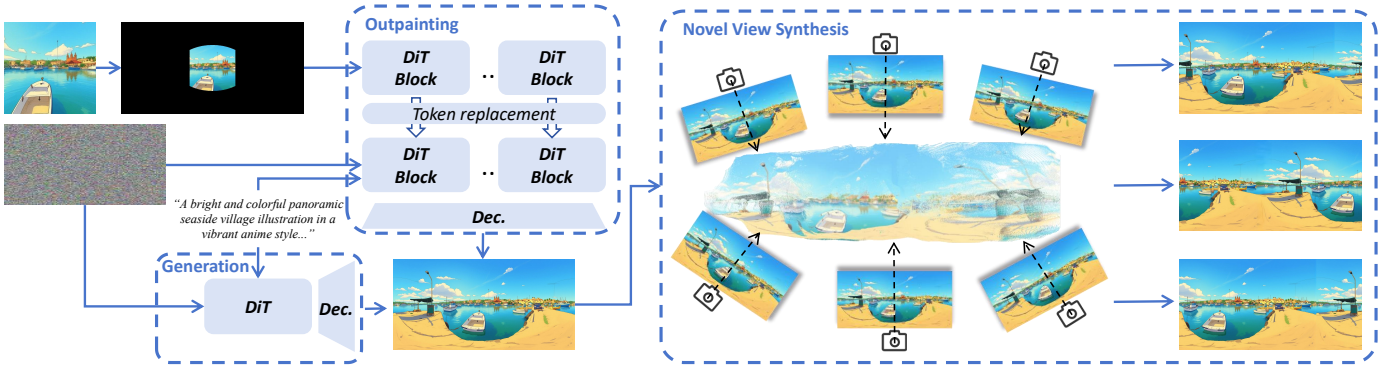

Fig. 9: Overview of a panoramic world modeling pipeline, which integrates text-guided generation, image completion, and novel view synthesis. These stages also correspond to three representative tasks, demonstrating how generative methods collectively advance holistic scene modeling.

4.3.2 Fusion Perception with LiDAR

These approaches integrate panoramic imagery with sparse LiDAR to enhance robustness in driving, robotics, and mapping. Unlike perspective fusion [9], [344] with narrow FoV and rectilinear geometry, panoramic settings require spherical modeling and full-scene alignment. Recent methods explore semantic mapping and planning with panoramic depth cues (SPOMP [261], Humanoid-Pano [119]), indoor reconstruction with fisheye–LiDAR fusion or color-consistent point clouds (Ma et al. [262], OmniColor [263]), cross-modal saliency and viewpoint-robust localization via spherical attention or equivariant networks (Zhao et al. [264], Bernreiter et al. [265]), and trajectory estimation or tracking with LiDAR-enhanced optimization (PanoramicVO [266], MMPAT [132]).

4.3.3 Fusion with Text

Extending vision–language fusion from perspective images to 360° panoramas introduces unique gaps: spherical projections cause severe distortions and wraparound boundaries, while semantics are scattered across the full sphere, making conventional attention and alignment ineffective. Recent methods introduce spatially-aware attention, guided FoV exploration, and spherical encodings to enable effective VQA, captioning, and dense grounding in panoramic contexts, highlighting the importance of geometry-aware fusion for robust panoramic vision–language understanding.

VQA-360 [267] pioneers the task with an observe-then-answer strategy and the first panoramic VQA dataset. VIEW-QA [268] extends this to 360° videos with annotations for assistive scenarios. OmniVQA [269] and OSR-Bench [270] further emphasize spatial reasoning and multimodal grounding, evaluating MLLMs via structured policy learning and cognitive map–based protocols. Most recently, Dense360 [271] introduces dense captioning and grounding for ERP images using an ERP-RoPE encoding and large-scale datasets, enabling spatially consistent language grounding in panoramic contexts.

4.4 Generative-modeling-based Tasks

With the rising interest in world models, an emerging approach leverages the inherent geometric consistency of panoramas as a foundation for generation. Recent advances such as Matrix-3D [330] and HunyuanWorld 1.0 [329] exemplify this trend, pushing panoramic generation beyond single-scene synthesis toward omnidirectional, explorable 3D worlds. Within this paradigm, as illustrated in Fig. 9, text-guided generation introduces semantic controllability through multimodal conditioning, image completion recovers unobserved content in missing or outpainted regions, and novel view synthesis extends NeRF and 3DGS frameworks to panoramic settings, enabling faithful viewpoint expansion under spherical geometry. Together, these components drive the development of world modeling, shifting 360° vision from reconstruction of observations into the generative construction of rich, interactive, and semantically grounded virtual worlds.

4.4.1 Text-guided Generation

Text-guided generation aims to synthesize panoramic images or videos from textual descriptions, providing semantic controllability for 360° content creation. Diffusion models have driven substantial advances in text-to-image/video (T2I/V) synthesis for perspective images and videos, enabling high-quality and controllable generation [345], [346]. Extending these models to 360° panoramas is nontrivial due to spherical topology, which induces geometric distortion and discontinuities across boundaries. Geometry-aware generation is therefore essential, with clear benefits for downstream tasks requiring precise spatial understanding. Early attempts like Text2Light [272] remain widely used benchmarks but do not explicitly enforce spherical consistency or address panorama-specific structural and semantic issues, and thus serve as transitional rather than definitive solutions to the domain gap.

Distortion-Aware Methods explicitly adapt architectures to spherical geometry and ERP distortions for panoramic generation. Representative designs include spherical convolutions (SMConv, SMGD [273]) and SphereDiffusion [274] for native spherical modeling, pixel-wise reweighting (PanoWAN [275]) and projection-specific denoising (DynamicScaler [276]) for distortion-aware optimization, as well as transformer backbones with spherical encodings (PanoDiT [277]) to capture long-range dependencies while preserving geometric fidelity.

Projection-Driven Methods for panoramic generation include tangent-plane decomposition (TanDiT [278]), dual-branch architectures that combine perspective and panoramic contexts (Zhang et al. [279], TiP4GEN [280]), RGB-Depth cube diffusion framework (DreamCube [281]), and hybrid spherical–planar modeling (SphereDiff [282]). Other approaches build features from overlapping perspective views or predefined tangent directions (VideoPanda [283], ViewPoint [284]), achieving localized consistency while maintaining global coherence, and

Continuity Modeling Methods enhances wraparound continuity in 360° panoramas through inference-time strategies without

altering base architectures. Training-free approaches include circular blending (Diffusion360 [285]), dual pre-denoising at borders (Wang et al. [286]), circular padding (Wang et al. [287], 360DVD [347]), and iterative warping with bidirectional guidance (PanoFree [288]), among which PanoFree stands out as a plug-and-play solution balancing efficiency and compatibility.

Other Methods explore complementary directions such as coarse-to-fine temporal modules for panoramic video generation (VidPanos [289]) and user-controllable approaches like OmniDrag [290], which explicitly model spherical motion and enable trajectory-based editing through temporally aware architectures and motion-diverse datasets.

4.4.2 Image Completion

Image completion, including both inpainting (filling missing regions) and outpainting (extending beyond boundaries), has achieved notable success in perspective images using autoencoders, GANs, and diffusion models. However, extending to 360° panoramas remains challenging due to ERP-induced distortions around poles and seams, the difficulty of generating plausible content in large missing regions, and poor handling of edge- and pole-specific structures such as sky, ground, and stitching artifacts. To address these issues, recent research explores two main directions.

Distortion-Aware Methods for 360° completion include Dream360 [291], which builds spherical latent spaces to reduce planar bias; PanoDecouple [292], which separates distortion guidance and content completion with distortion maps and Distort-CLIP loss; and 2S-ODIS [293], which employs a two-stage VQGAN + NFoV refinement pipeline to balance global layout and local detail.

Continuity Modeling Methods enforce circular consistency in panoramic completion. Cylin-Painting [294] preserves circular continuity with cylinder-style convolutions, PanoDiff [295] and PanoDiffusion [296] achieve rotation equivariance and wraparound consistency with distortion–content decoupling and RGB-D cues, while Akimoto et al. [297] enhance texture fidelity and semantic continuity through a dual-network design with circular inference.

Other Methods explore complementary directions beyond distortion correction and continuity modeling, including semantic conditioning (ImmerseGAN [298]), multi-modal autoregressive generation with NFoV, text, and geometry cues (AOG-Net [299]), and structure-aware RGB-D disentanglement with new evaluation metrics (BIPS [300]). These approaches extend panoramic completion toward more controllable, multimodal, and structurally reliable synthesis.

4.4.3 Novel View Synthesis

Novel view synthesis aims to generate unseen panoramic views from limited inputs, enabling consistent scene expansion under spherical geometry. Recent approaches address these issues by incorporating spherical representations, depth priors, and soft occlusion handling into NeRF/3DGS frameworks, and further explore hybrid architectures and AIGC-driven pipelines for efficient, semantically consistent panoramic scene generation.

Novel View Synthesis (NVS) for 360° panoramas aims to generate unseen viewpoints from limited observations. While perspective-based methods such as NeRF [348] and 3D Gaussian Splatting (3DGS) [349] have achieved remarkable progress, directly extending them to panoramic faces faces challenges including spherical distortions, wide-FoV occlusions, and inefficiencies in Cartesian sampling. Recent approaches address these issues by incorporating spherical representations, depth priors, and soft occlusion handling into NeRF/3DGS frameworks, and further explore hybrid architectures and AIGC-driven pipelines for efficient, semantically consistent panoramic scene generation.

NeRF-based Methods adapt neural radiance fields to 360° panoramas by addressing spherical distortion, wide-baseline occlusions, and inefficient Cartesian sampling. Early efforts reformulate ray sampling and feature aggregation in spherical coordinates to improve efficiency and geometric consistency (EgoNeRF [301], PanoGRF [302]); later studies introduce novel camera models (OmniNeRF [303]), local radiance field partitioning (OmniLocalRF [304], 360Roam [305]), and semantic or depth priors for sparse-view or monocular inputs (360FusionNeRF [306], Perf [307]) to enhance scene understanding from sparse or monocular inputs. while latest studies incorporate HDR estimation [308] or inpainting-based augmentation to mitigate data scarcity.

3DGS-based Methods extend 3D Gaussian Splatting to panoramic view synthesis by addressing projection mismatch, sparse inputs, and ERP distortion through geometry-aware projection, sampling, and rasterization. To align with spherical geometry, differentiable splatting on tangent planes or Fibonacci lattices reduces pole distortion and improves ERP rendering (360-GS [309], ODGS [310], PanSplat [311]); OmniGS [312] replaces cubemap approximations with differentiable projection models for stronger generalization. Dual-projection encoders and Yin–Yang decompositions enhance Gaussian parameter prediction (Splatter-360 [313], OmniSplat [314]), while layout priors, boundary optimization, hierarchical cost volumes, RoPE rolling, and distortion-aware losses, dual-Fisheye distortion modeling improve robustness in sparse or low-texture indoor scenes (360-GS [309], TPGS [315], PanoSplatt3R [316], ErpGS [317], Seam360GS [318]). Finally, OB3D [319] introduces a high-fidelity omnidirectional dataset with diverse trajectories, enabling rigorous benchmarking and advancing panoramic 3DGS research.

Other Representations move beyond volumetric NeRFs and point-based 3DGS by adopting layered images (MSI/LDI) or spherical meshes for geometry-aware warping and differentiable compositing. Specifically, SOMSI [321] follows the layered-image paradigm with a soft-occlusion MSI, achieving high quality with few layers and fast feedforward synthesis. OmniSyn [320] follows the spherical-mesh paradigm by pairing 360° depth and a spherical cost volume with a differentiable 360° mesh renderer to handle wide baselines and occlusions. Similarly, Casual 6-DoF [322] enhances the spherical-mesh approach by integrating panoramic depths into a lightweight mesh, allowing for real-time, VR-ready 6-DoF navigation from casually captured panoramas.

4.4.4 Applications.

Applications of panoramic generative models build on the spatial consistency and completeness of 360° vision, extending text-driven control to diverse modalities and downstream tasks. Representative directions include constrained scene generation for bounded single-scene synthesis, unconstrained world generation for open-ended multi-trajectory environments, and vision–language navigation (VLN) for embodied interaction. Together, these applications adapt panoramic generation for spatial reasoning, immersive interaction, and controllable environments.

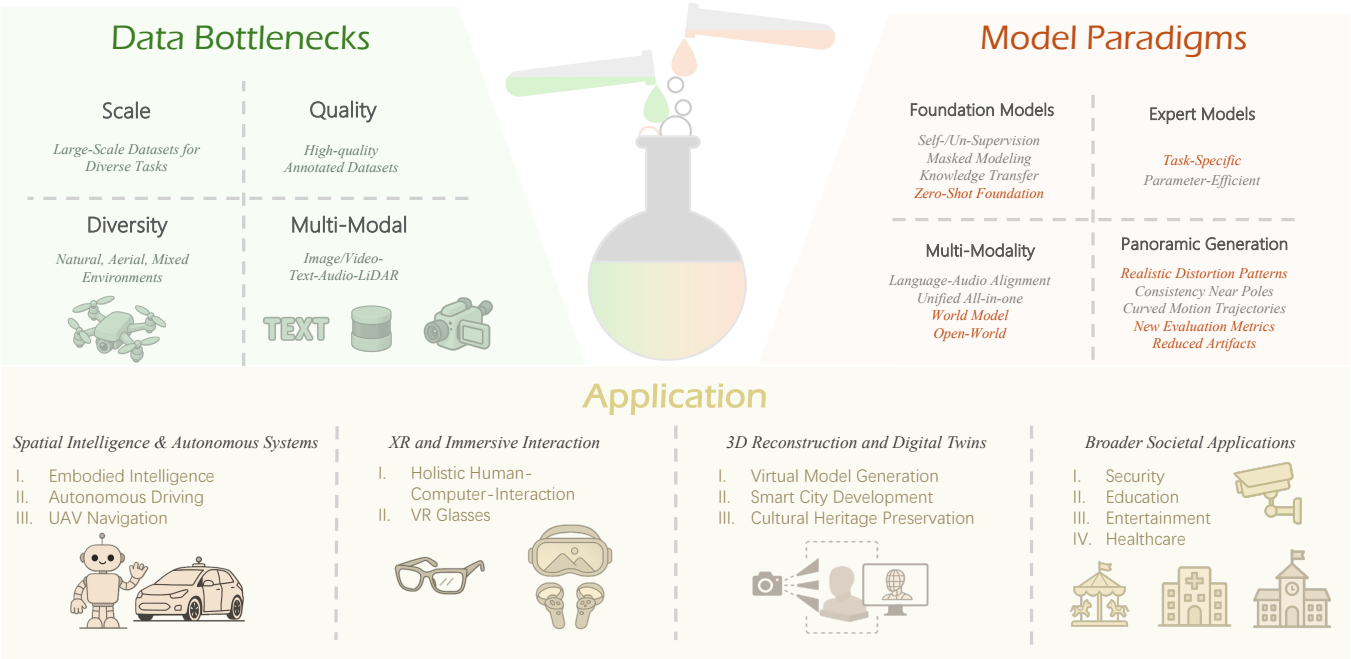

Fig. 10: Summary of future directions in panoramic vision: (1) **Data Bottlenecks**, emphasizing challenges in scale, diversity, quality, and multi-modality; (2) **Model Paradigms**, covering foundation models, expert models, multi-modal integration, and models for panoramic generation; and (3) **Applications**, including spatial intelligence and autonomous systems, XR and immersive interaction, 3D reconstruction and digital twins, and broader societal applications such as security, education, entertainment, and healthcare.

Constrained Scene Generation. leverages text-to-360° panoramas as holistic spatial priors for constructing immersive 3D and 4D environments.

SceneDreamer360 [323] and DreamScene360 [324] synthesize high-quality panoramas with depth estimation to reconstruct 3D Gaussian Splatting scenes, ensuring high-resolution rendering and spatial fidelity. Similarly, LayerPano3D [325] lifts layered decompositions into Gaussian fields for progressive optimization, while HoloDreamer [326] employs a multi-stage depth–geometry pipeline for view-consistent 3D scenes. Extending to dynamics, HoloTime [327] generates ERP videos for re-renderable 4D reconstruction, and 4K4DGen [328] decomposes panoramas into perspective views and recovers spherical geometry with 4D Gaussian fields.

Unconstrained World Generation extends panoramic generation from single-scene synthesis to open-ended, explorable 3D worlds. Matrix-3D [330] employs trajectory-guided panoramic video diffusion conditioned on mesh renders for precise camera control, lifting generated 360° videos into navigable 3D via either fast panorama reconstruction or high-fidelity 3DGS pipelines. Complementarily, HunyuanWorld 1.0 [329] builds interactive worlds from text or a single image through semantic layering, layer-aligned depth, and mesh-based reconstruction, combining foreground asset warping, HDR sky modeling, and world-consistent video diffusion for long-range exploration.

Vision–Language Navigation (VLN) leverages text-to-360° generation to provide panoramic context and semantically consistent observations for instruction following. PANOGEN [331] and PANOGEN++ [332] generate realistic panoramas from room descriptions with semantic layouts and orientation-aware object placement, enriching VLN training. VLN-RAM [333] expands unseen coverage by rewriting scenes and instructions with object-aware augmentation, while Any2Omni [334] introduces a large-scale dataset and a spatially consistent Transformer (Omni^2) for panoramic generation and editing.

5 Challenges and Future Works

Despite rapid progress, panoramic vision still faces fundamental challenges that limit its scalability, robustness, and deployment in real-world scenarios. Based on the analysis of existing methods, we highlight future research opportunities from three complementary perspectives: data, models, and applications.

5.1 Data Bottlenecks

Compared with perspective vision, a major bottleneck for panoramic vision is data scarcity. Existing datasets (a comparative summary with perspective datasets is provided in supplementary material) remain limited in scale, diversity, quality, and modality, thereby constraining model generalization and hindering fair benchmarking.

Scale. Large-scale, standardized 360° image and video datasets across diverse tasks and scenarios remain scarce, constraining model training and reproducible evaluation.

Diversity. Most existing datasets concentrate on indoor or urban settings, with limited coverage of natural, aerial, or mixed environments, thus constraining progress towards open-world generalization.

Quality. High-quality annotated panoramic datasets for tasks such as depth estimation, segmentation, detection, tracking, and mapping remain limited, particularly with fine-grained labels in real-world scenarios.

Multi-Modality. Panoramic image–text and video–text resources remain limited, restricting advances in language-guided generation, VQA, and cross-modal reasoning.

Overall, future progress depends on constructing large-scale, foundational, standardized, multi-modal datasets to enhance generalization, comparability, and performance across various tasks.

5.2 Model Paradigms

Progress in panoramic vision is closely tied to advances in model design, ranging from general foundation frameworks to task-specific expert architectures, and further extending to multimodal and generative paradigms. A major challenge is to develop models that can effectively adapt to panorama-specific representations while progressing toward three key goals: (1) strong ability for generalization and zero-shot transfer, (2) unified all-in-one architectures that jointly support multiple tasks, and (3) world modeling techniques capable of open-world understanding and scene generation.

Foundation Models. Training paradigms such as self-/un-supervision, contrastive learning, and masked modeling require adaptation to 360° data. An important direction is to transfer knowledge from foundational perspective models to panoramic domains, thereby reducing the domain gap and improving efficiency. Future foundation models should emphasize on zero-shot robustness, ensuring reliable performance in novel panoramic environments under limited supervision.

Expert Models. In addition to general-purpose foundation models, task-specific expert models remain essential. On the one hand, incorporating panoramic characteristics into task-specific architecture designs can lead to better efficiency and accuracy. On the other hand, for tasks such as detection, segmentation, depth estimation, and temporal analysis, integrating pre-trained base networks with parameter-efficient expert modules can further boost performance while preserving generalization.

Multi-Modality and Panoramic Generation. Existing multi-modal frameworks still struggle to handle panoramic-specific properties such as spatial continuity and distortion distribution. Future developments may involve incorporating panoramic priors into architectures that better align vision, language, and audio. Detailed promising directions include: (1) panorama–language alignment for improved grounding, (2) unified frameworks that integrate generation and understanding, (3) world models for continuous and interactive panoramic scene synthesis, and (4) open-world adaptation to unseen semantics.

While understanding and generation are inherently coupled in multi-modal panoramic systems, advancing panoramic generation remains equally important. Key challenges include: (i) developing specialized evaluation metrics, (ii) preserving realistic distortion patterns, (iii) ensuring consistency near poles, and (iv) modeling curved motion trajectories distinct from those in perspective video. In addition, panoramic video generation poses further difficulties in maintaining spatiotemporal coherence.

5.3 Application

Spatial Intelligence and Autonomous Systems. 360° panoramic vision provides a complete environmental context, which inherently aligns with the demands of spatial intelligence. By eliminating blind spots and enhancing global perception, it enables robust scene understanding and decision-making, which is particularly crucial for embodied intelligence, autonomous driving, and UAV navigation, where comprehensive situational awareness directly supports safety and reliability.

XR and Immersive Interaction. From panoramic recording to high-resolution content generation, 360° vision forms the cornerstone of extended reality (XR). Future directions include integrating spatial audio, haptic feedback, and other multi-sensory modalities to create a holistic, immersive interaction paradigm. Moreover, supporting rich and perceptually aligned human-computer interaction across the human senses, together with lightweight deployment on portable devices such as VR/AR glasses, will drive practical adoption and daily usage.

3D Reconstruction and Digital Twins. Panoramic imaging captures holistic scene information, enabling the complete reconstruction of 3D environments and the creation of digital twins. Applications range from 3D mapping and spatial digital archiving to virtual model generation, supporting fields such as smart city development and cultural heritage preservation.

Broader Societal Applications. Beyond the above technical directions, panoramic vision also holds broad prospects for practical use in various domains. It can enhance security and surveillance through full-scene monitoring with fewer blind spots, enrich education and training via immersive content delivery, enable new forms of entertainment and media through high-fidelity 360° capture and generation, and support healthcare with XR-assisted telemedicine and rehabilitation. These examples highlight its transformative potential across diverse industries.

6 Conclusion

This survey provides a comprehensive overview of panoramic vision, aiming to bridge the gaps between panoramic and perspective representations. First, we analyze panoramic imaging systems and projection models, which reveal the unique geometric characteristics underlying the fundamental gaps between panoramic and perspective representations: geometric distortion, non-uniform spatial sampling, and boundary continuity. Next, we conduct both cross-method and cross-task analysis across more than 20 representative tasks, synthesizing common strategies while highlighting their advantages, limitations, and applicability. Finally, we outline several future directions, including building larger and more diverse datasets, developing foundational, multimodal, and generation models, and extending to broader downstream applications such as embodied intelligence, autonomous driving, and immersive media. Overall, this survey serves as both a comprehensive reference and a forward-looking guide for the continued development of panoramic vision.

7 Acknowledgment

We would like to thank our colleagues Yunning Peng, Shi Luo, and Haoran Feng for their valuable contributions to improving the quality of this work.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.

[4] L. Qi, J. Kuen, W. Guo, T. Shen, J. Gu, J. Jia, Z. Lin, and M.-H. Yang, "High-quality entity segmentation," in *ICCV*, 2023.

[5] C.-J. Qin, R.-Q. Wu, Z. Liu, X. Lin, C.-L. Guo, H. H. Park, and C. Li, "Restore anything with masks: Leveraging mask image modeling for blind all-in-one image restoration," in *ECCV*, 2024.

[6] B. Ren, Y. Li, N. Mehta, R. Timofte, H. Yu, C. Wan, Y. Hong, B. Han, Z. Wu, Y. Zou *et al.*, "The ninth ntire 2024 efficient super-resolution challenge report," in *CVPRW*, 2024.

[7] X. Lin, S. Luo, X. Shan, X. Zhou, C. Ren, L. Qi, M.-H. Yang, and N. Vasconcelos, "Hqgs: High-quality novel view synthesis with gaussian splatting in degraded scenes," in *ICLR*, 2025.

[8] L. Qi, L. Jiang, S. Liu, X. Shen, and J. Jia, "Amodal instance segmentation with kins dataset," in *CVPR*, 2019.

[9] Z. Wan, Z. Bi, Z. Zhou, H. Ren, Y. Zeng, Y. Li, L. Qi, X. Yang, M.-H. Yang, and H. Cheng, "Rapid hand: A robust, affordable, perception-integrated, dexterous manipulation platform for generalist robot autonomy," *arXiv preprint arXiv:2506.07490*, 2025.

[10] M. Zink, R. Sitaraman, and K. Nahrstedt, "Scalable 360 video stream delivery: Challenges, solutions, and opportunities," *Proceedings of the IEEE*, 2019.

[11] A. Yaqoob, T. Bi, and G.-M. Muntean, "A survey on adaptive 360 video streaming: Solutions, challenges and opportunities," *IEEE Communications Surveys & Tutorials*, 2020.

[12] X. Sui, S. Wang, and Y. Fang, "A survey on objective quality assessment of omnidirectional images," in *APSIPA ASC*, 2024.

[13] C. Zou, J.-W. Su, C.-H. Peng, A. Colburn, Q. Shan, P. Wonka, H.-K. Chu, and D. Hoiem, "Manhattan room layout reconstruction from a single 360 image: A comparative study of state-of-the-art methods," *arXiv preprint arXiv:1910.04099*, 2019.

[14] Q. Zhao, C. Guo, T. Zhang, J. Zhang, P. Jia, T. Su, W. Jiang, and C. Li, "A systematic investigation on deep learning-based omnidirectional image and video super-resolution," *arXiv preprint arXiv:2506.06710*, 2025.

[15] S. Gao, K. Yang, H. Shi, K. Wang, and J. Bai, "Review on panoramic imaging and its applications in scene understanding," *TIM*, 2022.

[16] S. Jiang, K. You, Y. Li, D. Weng, and W. Chen, "3d reconstruction of spherical images: a review of techniques, applications, and prospects," *Geo-spatial Information Science*, 2024.

[17] T. L. da Silveira, P. G. Pinto, J. Murrugarra-Llerena, and C. R. Jung, "3d scene geometry estimation from 360 imagery: A survey," *ACM Computing Surveys*, 2022.

[18] J. Yu, A. C. P. Grassi, and G. Hirtz, "Applications of deep learning for top-view omnidirectional imaging: A survey," in *CVPR*, 2023.

[19] M. Meng, Y. Zhu, Y. Zhao, Z. Li, and Z. Zhu, "3d indoor scene geometry estimation from a single omnidirectional image: A comprehensive survey," *Computational Visual Media*, 2025.

[20] H. Ai, Z. Cao, and L. Wang, "A survey of representation learning, optimization strategies, and applications for omnidirectional vision," in *IJCV*, 2025.

[21] H. Faraji and W. J. MacLean, "Ccd noise removal in digital images," *TIP*, 2006.

[22] A. Basu and S. Licardie, "Alternative models for fish-eye lenses," *Pattern recognition letters*, 1995.

[23] D. Scaramuzza, "Omnidirectional vision: from calibration to root motion estimation," Ph.D. dissertation, ETH Zurich, 2007.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.

[25] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *ECCV*, 2006.

[26] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *ICCV*, 2011.

[27] R. G. Von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "Lsd: A line segment detector," *Image Processing On Line*, 2012.

[28] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, 1981.

[29] F. Yu, X. Wang, M. Cao, G. Li, Y. Shan, and C. Dong, "Osrt: Omnidirectional image super-resolution with distortion-aware transformer," in *CVPR*, 2023.

[30] D. Zhao, J. Li, H. Li, and L. Xu, "Stripe sensitive convolution for omnidirectional image dehazing," *TVCG*, 2023.

[31] C. Tian, F. Shao, X. Chai, Q. Jiang, L. Xu, and Y.-S. Ho, "Viewport-sphere-branch network for blind quality assessment of stitched 360 omnidirectional images," *TCSVT*, 2022.

[32] X. Hu, Y. An, C. Shao, and H. Hu, "Distortion convolution module for semantic segmentation of panoramic images based on the image-forming principle," *TIM*, 2022.

[33] Z. Zhang, Y. Xu, J. Yu, and S. Gao, "Saliency detection in 360 videos," in *ECCV*, 2018.

[34] S. Xie, P. K. Lai, R. Laganiere, and J. Lang, "Effective convolutional neural network layers in flow estimation for omni-directional images," in *3DV*, 2019.

[35] K. Tateno, N. Navab, and F. Tombari, "Distortion-aware convolutional filters for dense prediction in panoramic images," in *ECCV*, 2018.

[36] C. Zhuang, Z. Lu, Y. Wang, J. Xiao, and Y. Wang, "Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation," in *AAAI*, 2022.

[37] A. Nishiyama, S. Ikehata, and K. Aizawa, "360 single image super resolution via distortion-aware network and distorted perspective images," in *ICIP*, 2021.

[38] X. Sun, W. Li, Z. Zhang, Q. Ma, X. Sheng, M. Cheng, H. Ma, S. Zhao, J. Zhang, J. Li *et al.*, "Opdn: Omnidirectional position-aware deformable network for omnidirectional image super-resolution," in *CVPRW*, 2023.

[39] R. Liu, Y. Qin, Y. Ying, X. Liu, H. Zhang, W. Sheng, J. Zhang, and S. Chen, "Distortion-aware outdoor panoramic depth estimation via local–global fusion," *TII*, 2025.

[40] X. Deng, H. Wang, M. Xu, Y. Guo, Y. Song, and L. Yang, "Lau-net: Latitude adaptive upscaling network for omnidirectional image super-resolution," in *CVPR*, 2021.

[41] C. Ozcinar, A. Rana, and A. Smolic, "Super-resolution of omnidirectional images using adversarial learning," in *International Workshop on Multimedia Signal Processing*, 2019.

[42] H. An and X. Zhang, "Perception-oriented omnidirectional image super-resolution based on transformer network," in *ICIP*, 2023.

[43] H. An, X. Zhang, S. Zhao, and L. Zhang, "Fato: Frequency attention transformer for omnidirectional image super-resolution," in *ACMMM Asia*, 2024.

[44] C. Yang, R. Dong, J. Xiao, C. Zhang, K.-M. Lam, F. Zhou, and G. Qiu, "Geometric distortion guided transformer for omnidirectional image super-resolution," *TCSVT*, 2025.

[45] J. Cao, Q. Ding, and H. Luo, "Geometric relationship-guided transformer network for omnidirectional image super-resolution," *Signal, Image and Video Processing*, 2025.

[46] X. Shen, Y. Wang, S. Zheng, K. Xiao, W. Yang, and X. Wang, "Fast omni-directional image super-resolution: Adapting the implicit image function with pixel and semantic-wise spherical geometric priors," in *AAAI*, 2025.

[47] W. Wen, Q. Zhao, and X. Shao, "Mambaosr: Leveraging spatial-frequency mamba for distortion-guided omnidirectional image super-resolution," *Entropy*, 2025.

[48] H. Liu, W. Ma, Z. Ruan, C. Fang, F. Shang, Y. Liu, L. Wang, C. Wang, and D. Jiang, "A single frame and multi-frame joint network for 360-degree panorama video super-resolution," *EAAI*, 2024.

[49] A. A. Baniya, T.-K. Lee, P. W. Eklund, and S. Aryal, "Omnidirectional video super-resolution using deep learning," *TMM*, 2023.

[50] H. An, X. Zhang, S. Zhao, L. Zhang, and R. Xiong, "Spatio-temporal distortion aware omnidirectional video super-resolution," *arXiv preprint arXiv:2410.11506*, 2024.

[51] X. Wang, T. D. Nguyen, C. M. Tran, E. Kamioka, and T. X. Phan, "Central vision based super-resolution for 360-degree videos," in *BDIOT*, 2023.

[52] V. Fakour-Sevom, E. Guldogan, and J.-K. Kamarainen, "360 panorama super-resolution using deep convolutional networks," in *VISAPP*, 2018.

[53] Y. Zhang, H. Zhang, D. Li, L. Liu, H. Yi, W. Wang, H. Suitoh, and M. Odamaki, "Toward real-world panoramic image enhancement," in *CVPRW*, 2020.

[54] Y. Yoon, I. Chung, L. Wang, and K.-J. Yoon, "Spheresr: 360deg image super-resolution with arbitrary projection via continuous spherical image representation," in *CVPR*, 2022.

[55] Z. Cao, H. Ai, Y.-P. Cao, Y. Shan, X. Qie, and L. Wang, "Omnizoomer: Learning to move and zoom in on sphere at high-resolution," in *ICCV*, 2023.

[56] J. Wang, Y. Cui, Y. Li, W. Ren, and X. Cao, "Omnidirectional image super-resolution via bi-projection fusion," in *AAAI*, 2024.

[57] Q. Cai, M. Li, D. Ren, J. Lyu, H. Zheng, J. Dong, and Y.-H. Yang, "Spherical pseudo-cylindrical representation for omnidirectional image super-resolution," in *AAAI*, 2024.

[58] R. Li, X. Sheng, W. Li, and J. Zhang, "Omnissr: Zero-shot omnidirectional image super-resolution using stable diffusion model," in *ECCV*, 2024.

[59] L. Liu, T. Luo, G. Jiang, Y. Chen, H. Xu, R. Hu, and Z. He, "Diffosr: Latitude-aware conditional diffusion probabilistic model for omnidirectional image super-resolution," *KBS*, 2025.

[60] X. Sheng, R. Li, B. Chen, W. Li, X. Jiang, and J. Zhang, "Realosr: Latent unfolding boosting diffusion-based real-world omnidirectional image super-resolution," *arXiv preprint arXiv:2412.09646*, 2024.

[61] Y. Hong, Q. Zheng, L. Zhao, X. Jiang, A. C. Kot, and B. Shi, "Panoramic image reflection removal," in *CVPR*, 2021.

[62] ——, "Par2net: End-to-end panoramic image reflection removal," *TPAMI*, 2023.

[63] B.-J. Han and J.-Y. Sim, "Zero-shot learning for reflection removal of single 360-degree image," in *ECCV*, 2022.

[64] J. Park, H. Kim, E. Park, and J.-Y. Sim, "Fully-automatic reflection removal for 360-degree images," in *WACV*, 2024.

[65] S. Bigot, D. Kachi, S. Durand, and E. M. Mouaddib, "Spherical image denoising and its application to omnidirectional imaging." in *VISAPP*, 2007.

[66] A. Iazzi, A. Radgui, M. Rziza *et al.*, "An adapted block thresholding method for omnidirectional image denoising," *Research Journal of Applied Sciences, Engineering and Technology*, 2014.

[67] T. D. K. Phan and T. H. Y. Tran, "A space-variant nonlinear algorithm for denoising omnidirectional images corrupted by poisson noise," *IEEE Signal Processing Letters*, 2020.

[68] R. Fermanian, T. Maugey, and C. Guillemot, "Spheredrunet: A spherical denoiser for omnidirectional images," in *ISMAR-Adjunct*, 2023.

[69] Y. Li and J. Lou, "Omnigradient based total variation minimization for enhanced defocus deblurring of omnidirectional images," *International Journal of Optics*, 2014.

[70] Y. Peng, Y. Liu, Y. Li, and M. Zhang, "Coded aperture techniques for catadioptric omni-directional image defocus deblurring," in *SMC*, 2012.

[71] Y. Liu, H. Li, Y. Li, J. Liu, and M. Zhang, "Coded aperture enhanced catadioptric optical system for omnidirectional image deblurring," *Optik*, 2014.

[72] B. Alibouch and M. Rziza, "Catadioptric omnidirectional images motion deblurring," in *International Symposium on Ubiquitous Networking*, 2021.

[73] W. Sun, X. Min, G. Zhai, K. Gu, H. Duan, and S. Ma, "Mc360iqa: A multi-channel cnn for blind 360-degree image quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[74] J. Xu, W. Zhou, and Z. Chen, "Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks," *TCSVT*, 2020.

[75] L. Yang, M. Xu, T. Liu, L. Huo, and X. Gao, "Tvformer: Trajectory-guided visual quality assessment on 360 images with transformers," in *ACMMM*, 2022.

[76] C. Zhang and S. Liu, "No-reference omnidirectional image quality assessment based on joint network," in *ACMMM*, 2022.

[77] H. G. Kim, H.-T. Lim, and Y. M. Ro, "Deep virtual reality image quality assessment with human perception guider for omnidirectional image," *TCSVT*, 2019.

[78] T. Wu, S. Shi, H. Cai, M. Cao, J. Xiao, Y. Zheng, and Y. Yang, "Assessor360: Multi-sequence network for blind omnidirectional image quality assessment," *NeurIPS*, 2023.

[79] N. J. Tofighi, M. H. Elfkir, N. Imamoglu, C. Ozcinar, E. Erdem, and A. Erdem, "St360iq: No-reference omnidirectional image quality assessment with spherical vision transformers," in *ICASSP*, 2023.

[80] Y. Zhang, L. Wan, D. Liu, X. Zhou, P. An, and C. Shan, "Saliency-guided blind omnidirectional image quality assessment via scene content perceiving," *TIM*, 2024.

[81] Z. Hu, L. Liu, and Q. Sang, "Omnidirectional video quality assessment with causal intervention," *IEEE Transactions on Broadcasting*, 2024.

[82] X. Chai and F. Shao, "Blind quality assessment of omnidirectional videos using spatio-temporal convolutional neural networks," *Optik*, 2021.

[83] Y. Sun, A. Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE signal processing letters*, 2017.

[84] Y. Zhou, M. Yu, H. Ma, H. Shao, and G. Jiang, "Weighted-to-spherically-uniform ssim objective quality evaluation for panoramic video," in *ICSP*, 2018.

[85] H. T. Tran, N. P. Ngoc, C. M. Bui, M. H. Pham, and T. C. Thang, "An evaluation of quality metrics for 360 videos," in *ICUFN*, 2017.

[86] P. Gao, P. Zhang, and A. Smolic, "Quality assessment for omnidirectional video: A spatio-temporal distortion modeling approach," *TMM*, 2020.

[87] H. Jiang, G. Jiang, M. Yu, Y. Zhang, Y. Yang, Z. Peng, F. Chen, and Q. Zhang, "Cubemap-based perception-driven blind quality assessment for 360-degree images," *TIP*, 2021.

[88] C. Li, M. Xu, L. Jiang, S. Zhang, and X. Tao, "Viewport proposal cnn for 360° video quality assessment," in *CVPR*, 2019.

[89] H. Jiang, G. Jiang, M. Yu, T. Luo, and H. Xu, "Multi-angle projection based blind omnidirectional image quality assessment," *TCSVT*, 2021.

[90] R. G. d. A. Azevedo, N. Birkbeck, I. Janatra, B. Adsumilli, and P. Frossard, "A viewport-driven multi-metric fusion approach for 360-degree video quality assessment," in *ICME*, 2020.

[91] H. Yang, X. Zhang, J. Ma, L. Zhu, Y. Zhang, and H. Zhang, "Hierarchical graph attention network for no-reference omnidirectional image quality assessment," *arXiv preprint arXiv:2508.09843*, 2025.

[92] W. Zhou, J. Xu, Q. Jiang, and Z. Chen, "No-reference quality assessment for 360-degree images by analysis of multifrequency information and local-global naturalness," *TCSVT*, 2021.

[93] L. Yang, H. Duan, J. Wang, J. Liu, M. Hu, X. Min, G. Zhai, and P. L. Callet, "Quality assessment and distortion-aware saliency prediction for ai-generated omnidirectional images," *arXiv preprint arXiv:2506.21925*, 2025.

[94] K. Zhou, Z. Hao, L. Wang, and X. Liang, "Adaptive score alignment learning for continual perceptual quality assessment of 360-degree videos in virtual reality," *TVCG*, 2025.

[95] C. Ma, J. Zhang, K. Yang, A. Roitberg, and R. Stiefelhagen, "Densepass: Dense panoramic semantic segmentation via unsupervised domain adaptation with attention-augmented context exchange," in *ITSC*, 2021.

[96] X. Zheng, J. Zhu, Y. Liu, Z. Cao, C. Fu, and L. Wang, "Both style and distortion matter: Dual-path unsupervised domain adaptation for panoramic semantic segmentation," in *CVPR*, 2023.

[97] J. Zhang, K. Yang, C. Ma, S. Reiß, K. Peng, and R. Stiefelhagen, "Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation," in *CVPR*, 2022.

[98] J. Zhang, K. Yang, H. Shi, S. Reiß, K. Peng, C. Ma, H. Fu, P. H. Torr, K. Wang, and R. Stiefelhagen, "Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation," *TPAMI*, 2024.

[99] X. Zheng, P. Zhou, A. V. Vasilakos, and L. Wang, "Semantics distortion and style matter: Towards source-free uda for panoramic segmentation," in *CVPR*, 2024.

[100] X. Zheng, P. Y. Zhou, A. V. Vasilakos, and L. Wang, "360sfuda++: Towards source-free uda for panoramic segmentation by learning reliable category prototypes," *TPAMI*, 2024.

[101] K. Yang, X. Hu, Y. Fang, K. Wang, and R. Stiefelhagen, "Omnisupervised omnidirectional semantic segmentation," *TITS*, 2020.

[102] W. Zhang, Y. Liu, X. Zheng, and L. Wang, "Goodsam: Bridging domain and capacity gaps via segment anything model for distortion-aware panoramic semantic segmentation," in *CVPR*, 2024.

[103] ——, "Goodsam++: Bridging domain and capacity gaps via segment anything model for panoramic semantic segmentation," *arXiv preprint arXiv:2408.09115*, 2024.

[104] D. Zhong, X. Zheng, C. Liao, Y. Lyu, J. Chen, S. Wu, L. Zhang, and X. Hu, "Omnisam: Omnidirectional segment anything model for uda in panoramic semantic segmentation," *arXiv preprint arXiv:2503.07098*, 2025.

[105] K. Yang, X. Hu, L. M. Bergasa, E. Romera, and K. Wang, "Pass: Panoramic annular semantic segmentation," *TITS*, 2019.

[106] K. Yang, X. Hu, H. Chen, K. Xiang, K. Wang, and R. Stiefelhagen, "Dspass: Detail-sensitive panoramic annular semantic segmentation through swaftnet for surrounding sensing," in *IV*, 2020.

[107] Z. Zheng, C. Lin, L. Nie, K. Liao, Z. Shen, and Y. Zhao, "Complementary bi-directional feature compression for indoor 360deg semantic segmentation with self-distillation," in *WACV*, 2023.

[108] C. Zhang, Z. Cui, C. Chen, S. Liu, B. Zeng, H. Bao, and Y. Zhang, "Deeppanocontext: Panoramic 3d scene understanding with holistic scene context graph and relation-based optimization," in *CVPR*, 2021.

[109] X. Li, T. Wu, Z. Qi, G. Wang, Y. Shan, and X. Li, "Sgat4pass: Spherical geometry-aware transformer for panoramic semantic segmentation," in *arXiv preprint arXiv:2306.03403*, 2023.

[110] J. Liu, H. Yu, S. Li, and J. Li, "360-degree full-view image segmentation by spherical convolution compatible with large-scale planar pre-trained models," *arXiv preprint arXiv:2507.09216*, 2025.

[111] C. Zhang, S. Liwicki, W. Smith, and R. Cipolla, "Orientation-aware semantic segmentation on icosahedron spheres," in *ICCV*, 2019.

[112] A. Jaus, K. Yang, and R. Stiefelhagen, "Panoramic panoptic segmentation: Insights into surrounding parsing for mobile agents via unsupervised contrastive learning," *TITS*, 2023.

[113] S. Yan, X. Xu, R. Zhang, L. Hong, W. Chen, W. Zhang, and W. Zhang, "Panovos: Bridging non-panoramic and panoramic views with transformer for video segmentation," in *ECCV*, 2024.

[114] M. Liu, S. Wang, Y. Guo, Y. He, and H. Xue, "Pano-sfmlearner: Self-supervised multi-task learning of depth and semantics in panoramic videos," *IEEE Signal Processing Letters*, 2021.

[115] Y. Cao, J. Zhang, H. Shi, K. Peng, Y. Zhang, H. Zhang, R. Stiefelhagen, and K. Yang, "Occlusion-aware seamless segmentation," in *ECCV*, 2024.

[116] J. Zheng, R. Liu, Y. Chen, K. Peng, C. Wu, K. Yang, J. Zhang, and R. Stiefelhagen, "Open panoramic segmentation," in *ECCV*, 2024.

[117] Z. Teng, J. Zhang, K. Yang, K. Peng, H. Shi, S. Reiß, K. Cao, and R. Stiefelhagen, "360bev: Panoramic semantic mapping for indoor bird's-eye view," in *WACV*, 2024.

[118] J. Wei, J. Zheng, R. Liu, J. Hu, J. Zhang, and R. Stiefelhagen, "Onebev: Using one panoramic image for bird, aos-eye-view semantic mapping," in *ACCV*, 2024.

[119] Q. Zhang, Z. Zhang, W. Cui, J. Sun, J. Cao, Y. Guo, G. Han, W. Zhao, J. Wang, C. Sun *et al.*, "Humanoidpano: Hybrid spherical panoramic-lidar cross-modal perception for humanoid robots," *arXiv preprint arXiv:2503.09010*, 2025.

[120] B. Coors, A. P. Condurache, and A. Geiger, "Spherenet: Learning spherical representations for detection and classification in omnidirectional images," in *ECCV*, 2018.

[121] Y.-C. Su and K. Grauman, "Learning spherical convolution for fast features from 360 imagery," *NeurIPS*, 2017.

[122] K.-H. Wang and S.-H. Lai, "Object detection in curved space for 360-degree camera," in *ICASSP*, 2019.

[123] G. Tong, H. Chen, Y. Li, X. Du, and Q. Zhang, "Object detection for panoramic images based on ms-rpn structure in traffic road scenes," *IET Computer Vision*, 2019.

[124] N. Wakai, S. Sato, Y. Ishii, and T. Yamashita, "Panoramic distortion-aware tokenization for person detection and localization using transformers in overhead fisheye images," *arXiv preprint arXiv:2503.14228*, 2025.

[125] Q. Chang, H. Liao, X. Meng, S. Xu, and Y. Cui, "Panoglassnet: Glass detection with panoramic rgb and intensity images," *TIM*, 2024.

[126] W. Yang, Y. Qian, J.-K. Kämäräinen, F. Cricri, and L. Fan, "Object detection in equirectangular panorama," in *ICPR*, 2018.

[127] P. Zhao, A. You, Y. Zhang, J. Liu, K. Bian, and Y. Tong, "Spherical criteria for fast and accurate 360 object detection," in *AAAI*, 2020.

[128] F. Dai, B. Chen, H. Xu, Y. Ma, X. Li, B. Feng, P. Yuan, C. Yan, and Q. Zhao, "Unbiased iou for spherical image object detection," in *AAAI*, 2022.

[129] M. Cao, S. Ikehata, and K. Aizawa, "Field-of-view iou for object detection in 360° images," *TIP*, 2023.

[130] X. Liu, H. Xu, B. Chen, Q. Zhao, Y. Ma, C. Yan, and F. Dai, "Sph2pob: Boosting object detection on spherical images with planar oriented boxes methods." in *IJCAI*, 2023.

[131] M. Jiang, R. Sogabe, K. Shimasaki, S. Hu, T. Senoo, and I. Ishii, "500-fps omnidirectional visual tracking using three-axis active vision system," *TIM*, 2021.

[132] Y. He, W. Yu, J. Han, X. Wei, X. Hong, and Y. Gong, "Know your surroundings: Panoramic multi-object tracking by multimodality collaboration," in *CVPRW*, 2021.

[133] T. Fischer, Y.-H. Yang, S. Kumar, M. Sun, and F. Yu, "Cc-3dt: Panoramic 3d object tracking via cross-camera fusion," *arXiv preprint arXiv:2212.01247*, 2022.

[134] H. Huang, Y. Xu, Y. Chen, and S.-K. Yeung, "360vot: A new benchmark dataset for omnidirectional visual object tracking," in *ICCV*, 2023.

[135] Y. Xu, H. Huang, Y. Chen, and S.-K. Yeung, "360vots: Visual object tracking and segmentation in omnidirectional videos," *arXiv preprint arXiv:2404.13953*, 2024.

[136] K. Luo, H. Shi, S. Wu, F. Teng, M. Duan, C. Huang, Y. Wang, K. Wang, and K. Yang, "Omnidirectional multi-object tracking," in *CVPR*, 2025.

[137] W. Hutchcroft, Y. Li, I. Boyadzhiev, Z. Wan, H. Wang, and S. B. Kang, "Covispose: Co-visibility pose transformer for wide-baseline relative pose estimation in 360° indoor panoramas," in *ECCV*, 2022.

[138] N. Nejatishahidin, W. Hutchcroft, M. Narayana, I. Boyadzhiev, Y. Li, N. Khosravan, J. Košecká, and S. B. Kang, "Graph-covis: Gnn-based multi-view panorama global pose estimation," in *CVPRW*, 2023.

[139] D. Tu, H. Cui, X. Zheng, and S. Shen, "Panopose: Self-supervised relative pose estimation for panoramic images," in *CVPR*, 2024.

[140] R. Monroy, S. Lutz, T. Chalasani, and A. Smolic, "Salnet360: Saliency maps for omni-directional images with cnn," in *Signal Processing: Image Communication*, 2018.

[141] T. Suzuki and T. Yamanaka, "Saliency map estimation for omni-directional image considering prior distributions," in *SMC*, 2018.

[142] F. Dai, Y. Zhang, Y. Ma, H. Li, and Q. Zhao, "Dilated convolutional neural networks for panoramic image saliency prediction," in *ICASSP*, 2020.

[143] I. Djemai, S. A. Fezza, W. Hamidouche, and O. Déforges, "Extending 2d saliency models for head movement prediction in 360-degree images using cnn-based fusion," in *ISCAS*, 2020.

[144] F.-Y. Chao, L. Zhang, W. Hamidouche, and O. Deforges, "Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks," in *ICMEW*, 2018.

[145] D. Chen, C. Qing, X. Xu, and H. Zhu, "Salbinet360: Saliency prediction on 360 images with local-global bifurcated deep network," in *VR*, 2020.

[146] B. Dedhia, J.-C. Chiang, and Y.-F. Char, "Saliency prediction for omnidirectional images considering optimization on sphere domain," in *ICASSP*, 2019.

[147] H. Lv, Q. Yang, C. Li, W. Dai, J. Zou, and H. Xiong, "Salgcn: Saliency prediction for 360-degree images based on spherical graph convolutional networks," in *ACMMM*, 2020.

[148] D. Chen, C. Qing, X. Lin, M. Ye, X. Xu, and P. Dickinson, "Intra-and inter-reasoning graph convolutional network for saliency prediction on 360° images," *TCSVT*, 2022.

[149] Y. Yang, Y. Zhu, Z. Gao, and G. Zhai, "Salgfcn: Graph based fully convolutional network for panoramic saliency prediction," in *VCIP*, 2021.

[150] Y. Zhu, G. Zhai, Y. Yang, H. Duan, X. Min, and X. Yang, "Viewing behavior supported visual saliency predictor for 360 degree videos," *TCSVT*, 2021.

[151] Q. Yang, W. Gao, C. Li, H. Wang, W. Dai, J. Zou, H. Xiong, and P. Frossard, "360spred: Saliency prediction for 360-degree videos based on 3d separable graph convolutional networks," *TCSVT*, 2024.

[152] A. De Abreu, C. Ozcinar, and A. Smolic, "Look around you: Saliency maps for omnidirectional images in vr applications," in *QoMEX*, 2017.

[153] Y. Zhu, G. Zhai, and X. Min, "The prediction of head and eye movement for 360 degree images," *Signal Processing: Image Communication*, 2018.

[154] A. Bur, A. Tapus, N. Ouerhani, R. Siegwart, and H. Hugli, "Robot navigation by panoramic vision and attention guided fetaures," in *ICPR*, 2006.

[155] A. Nguyen, Z. Yan, and K. Nahrstedt, "Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction," in *ACMMM*, 2018.

[156] M. Qiao, M. Xu, Z. Wang, and A. Borji, "Viewport-dependent saliency prediction in 360 video," *TMM*, 2020.

[157] H.-T. Cheng, C.-H. Chao, J.-D. Dong, H.-K. Wen, T.-L. Liu, and M. Sun, "Cube padding for weakly-supervised saliency prediction in 360 videos," in *CVPR*, 2018.

[158] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang, "Predicting head movement in panoramic video: A deep reinforcement learning approach," *TPAMI*, 2018.

[159] Y. Dahou, M. Tliba, K. McGuinness, and N. O'Connor, "Atsal: an attention based architecture for saliency prediction in 360° videos," in *ICPR*, 2021.

[160] R. Guo, D. Niu, L. Qu, Y. Qi, J. Shi, W. Yue, B. Xing, T. Chen, and X. Ying, "Instance-level panoramic audio-visual saliency detection and ranking," in *ACMMM*, 2024.

[161] M. Cokelek, H. Ozsoy, N. Imamoglu, C. Ozcinar, I. Ayhan, E. Erdem, and A. Erdem, "Spherical vision transformers for audio-visual saliency prediction in 360° videos," *IEEE transactions on pattern analysis and machine intelligence*, 2025.

[162] D. Zhu, Y. Chen, D. Zhao, X. Min, Q. Zhou, S. Yu, G. Zhai, and X. Yang, "A lightweight saliency prediction model for omnidirectional images," in *ICME*, 2021.

[163] D. Zhu, Y. Chen, X. Min, Y. Zhu, G. Zhang, Q. Zhou, G. Zhai, and X. Yang, "Ransp: Ranking attention network for saliency prediction on omnidirectional images," *Neurocomputing*, 2021.

[164] D. Zhu, Y. Chen, X. Min, D. Zhao, Y. Zhu, Q. Zhou, X. Yang, and T. Han, "Saliency prediction on omnidirectional images with brain-like shallow neural network," in *ICPR*, 2021.

[165] I. Bogdanova, A. Bur, and H. Hugli, "Visual attention on the sphere," *TIP*, 2008.

[166] S.-T. Yang, F.-E. Wang, C.-H. Peng, P. Wonka, M. Sun, and H.-K. Chu, "Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama," in *CVPR*, 2019.

[167] H. Wang, W. Hutchcroft, Y. Li, Z. Wan, I. Boyadzhiev, Y. Tian, and S. B. Kang, "Psmnet: Position-aware stereo merging network for room layout estimation," in *CVPR*, 2022.

[168] J. Lee, B. Solarte, C.-H. Wu, J.-C. Jhang, F.-E. Wang, Y.-H. Tsai, and M. Sun, "ulayout: Unified room layout estimation for perspective and panoramic images," in *WACV*, 2025.

[169] C. Sun, C.-W. Hsiao, M. Sun, and H.-T. Chen, "Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation," in *CVPR*, 2019.

[170] Z. Shen, Z. Zheng, C. Lin, L. Nie, K. Liao, S. Zheng, and Y. Zhao, "Disentangling orthogonal planes for indoor panoramic room layout estimation with cross-scale distortion awareness," in *CVPR*, 2023.

[171] Z. Shen, C. Lin, J. Zhang, L. Nie, K. Liao, and Y. Zhao, "360 layout estimation via orthogonal planes disentanglement and multi-view geometric consistency perception," *TPAMI*, 2024.

[172] W. Zhang, M. Zhou, J. Cheng, Y. Liu, and W. Zhang, "C2p-net: Comprehensive depth map to planar depth conversion for room layout estimation," *TPAMI*, 2025.

[173] F.-E. Wang, Y.-H. Yeh, M. Sun, W.-C. Chiu, and Y.-H. Tsai, "Led 2-net: Monocular 360 layout estimation via differentiable depth rendering," in *CVPR*, 2021.

[174] C. Sun, W.-E. Tai, Y.-L. Shih, K.-W. Chen, Y.-J. Syu, Y.-C. F. Wang, H.-T. Chen *et al.*, "Seg2reg: Differentiable 2d segmentation to 1d regression rendering for 360 room layout reconstruction," in *CVPR*, 2024.

[175] Z. Jiang, Z. Xiang, J. Xu, and M. Zhao, "Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network," in *CVPR*, 2022.

[176] C. Sun, M. Sun, and H.-T. Chen, "Hohonet: 360 indoor holistic understanding with latent horizontal features," in *CVPR*, 2021.

[177] C. Fernandez-Labrador, A. Perez-Yus, G. Lopez-Nicolas, and J. J. Guerrero, "Layouts from panoramic images with geometry and deep learning," *RAL*, 2018.

[178] Y. Zhang, S. Song, P. Tan, and J. Xiao, "Panocontext: A whole-room 3d context model for panoramic scene understanding," in *ECCV*, 2014.

[179] J.-W. Su, C.-H. Peng, P. Wonka, and H.-K. Chu, "Gpr-net: Multi-view layout estimation via a geometry-aware panorama registration network," in *CVPRW*, 2023.

[180] Y.-J. Tsai, J.-C. Jhang, J. Zheng, W. Wang, A. Y. Chen, M. Sun, C.-H. Kuo, and M.-H. Yang, "No more ambiguity in 360deg room layout via bi-layout estimation," in *CVPR*, 2024.

[181] H. Jia, H. Yi, H. Fujiki, H. Zhang, W. Wang, and M. Odamaki, "3d room layout recovery generalizing across manhattan and non-manhattan worlds," in *CVPRW*, 2022.

[182] P. V. Tran, "Sslayout360: Semi-supervised indoor layout estimation from 360deg panorama," in *CVPR*, 2021.

[183] J. Zhang, C. Lin, Z. Shen, L. Nie, K. Liao, and Y. Zhao, "Semi-supervised 360 layout estimation with panoramic collaborative perturbations," *arXiv preprint arXiv:2503.01114*, 2025.

[184] A. Apitzsch, R. Seidel, and G. Hirtz, "Cubes3d: Neural network based optical flow in omnidirectional image scenes," *arXiv preprint arXiv:1804.09004*, 2018.

[185] C.-O. Artizzu, H. Zhang, G. Allibert, and C. Demonceaux, "Omniflownet: a perspective neural network adaptation for optical flow estimation in omnidirectional images," in *ICPR*, 2021.

[186] K. Bhandari, Z. Zong, and Y. Yan, "Revisiting optical flow estimation in 360 videos," in *ICPR*, 2021.

[187] H. Shi, Y. Zhou, K. Yang, X. Yin, Z. Wang, Y. Ye, Z. Yin, S. Meng, P. Li, and K. Wang, "Panoflow: Learning 360 optical flow for surrounding temporal understanding," *TITS*, 2023.

[188] L. Liu, M. Feng, J. Cheng, J. Xiang, X. Zhu, and X. Yang, "Prior-flow: Enhancing primitive panoramic optical flow with orthogonal view," *arXiv preprint arXiv:2506.23897*, 2025.

[189] M. Yuan and C. Richardt, "360 optical flow using tangent images," in *BMVC*, 2021.

[190] Y. Li, C. Barnes, K. Huang, and F.-L. Zhang, "Deep 360° optical flow estimation based on multi-projection fusion," in *ECCV*, 2022.

[191] Q. Zhao, W. Feng, L. Wan, and J. Zhang, "Sphorb: A fast and robust binary feature on the sphere," *IJCV*, 2015.

[192] T.-Y. Chuang and N. Perng, "Rectified feature matching for spherical panoramic images," *Photogrammetric Engineering & Remote Sensing*, 2018.

[193] H. Zhang, H. Yi, H. Jia, W. Wang, and M. Odamaki, "Panopoint: Self-supervised feature points detection and description for 360deg panorama," in *CVPRW*, 2023.

[194] C. Gava, V. Mukunda, T. Habtegebrial, F. Raue, S. Palacio, and A. Dengel, "Sphereglue: Learning keypoint matching on high resolution spherical images," in *CVPRW*, 2023.

[195] D. Jung, J. Choi, Y. Lee, S. Jeong, T. Lee, D. Manocha, and S. Yeon, "Edm: Equirectangular projection-oriented dense kernelized feature matching," in *CVPR*, 2025.

[196] J. Li, H. Li, and Y. Matsushita, "Lighting, reflectance and geometry estimation from 360 panoramic stereo," in *CVPR*, 2021.

[197] Z. Li, L. Wang, X. Huang, C. Pan, and J. Yang, "Phyir: Physics-based inverse rendering for panoramic indoor images," in *CVPR*, 2022.

[198] R.-K. Xu, L. Zhang, and F.-L. Zhang, "Intrinsic omnidirectional image decomposition with illumination pre-extraction," *TVCG*, 2024.

[199] H. Weber, D. Prévost, and J.-F. Lalonde, "Learning to estimate indoor lighting from 3d objects," in *3DV*, 2018.

[200] V. Gkitsas, N. Zioulis, F. Alvarez, D. Zarpalas, and P. Daras, "Deep lighting environment map estimation from spherical panoramas," in *CVPRW*, 2020.

[201] Y. Hold-Geoffroy, K. Sunkavalli, S. Hadap, E. Gambaretto, and J.-F. Lalonde, "Deep outdoor illumination estimation," in *CVPR*, 2017.

[202] Y. Hold-Geoffroy, A. Athawale, and J.-F. Lalonde, "Deep sky modeling for single image outdoor lighting estimation," in *CVPR*, 2019.

[203] J. Zhang, K. Sunkavalli, Y. Hold-Geoffroy, S. Hadap, J. Eisenman, and J.-F. Lalonde, "All-weather deep outdoor lighting estimation," in *CVPR*, 2019.

[204] S. Song and T. Funkhouser, "Neural illumination: Lighting prediction for indoor environments," in *CVPR*, 2019.

[205] M. Garon, K. Sunkavalli, S. Hadap, N. Carr, and J.-F. Lalonde, "Fast spatially-varying indoor lighting estimation," in *CVPR*, 2019.

[206] G. Somanath and D. Kurz, "Hdr environment map estimation for real-time augmented reality," in *CVPR*, 2021.

[207] G. Wang, Y. Yang, C. C. Loy, and Z. Liu, "Stylelight: Hdr panorama generation for lighting estimation and editing," in *ECCV*, 2022.

[208] S. Shen, Z. Bao, W. Xu, and C. Xiao, "Illumidiff: indoor illumination estimation from a single image with diffusion model," *TVCG*, 2025.

[209] J. Hilliard, A. Hilton, and J.-Y. Guillemaut, "Hdr environment map estimation with latent diffusion models," *arXiv preprint arXiv:2507.21261*, 2025.

[210] F. Zhan, C. Zhang, Y. Yu, Y. Chang, S. Lu, F. Ma, and X. Xie, "Emlight: Lighting estimation via spherical distribution approximation," in *AAAI*, 2021.

[211] F. Zhan, Y. Yu, C. Zhang, R. Wu, W. Hu, S. Lu, F. Ma, X. Xie, and L. Shao, "Gmlight: Lighting estimation via geometric distribution approximation," *TIP*, 2022.

[212] H. Weber, M. Garon, and J.-F. Lalonde, "Editable indoor lighting estimation," in *ECCV*, 2022.

[213] M. R. K. Dastjerdi, J. Eisenmann, Y. Hold-Geoffroy, and J.-F. Lalonde, "Everlight: Indoor-outdoor editable hdr lighting estimation," in *ICCV*, 2023.

[214] Y. Zhu, Y. Zhang, S. Li, and B. Shi, "Spatially-varying outdoor lighting estimation from intrinsics," in *CVPR*, 2021.

[215] J. Zhao, B. Xue, and M. Zhang, "Salenet: Structure-aware lighting estimations from a single image for indoor environments," *TIP*, 2024.

[216] Y. Zhao, M. Dasari, and T. Guo, "Clear: Robust context-guided generative lighting estimation for mobile augmented reality," *arXiv preprint arXiv:2411.02179*, 2024.

[217] N. Zioulis, A. Karakottas, D. Zarpalas, and P. Daras, "Omnidepth: Dense depth estimation for indoors spherical panoramas," in *ECCV*, 2018.

[218] Z. Shen, C. Lin, K. Liao, L. Nie, Z. Zheng, and Y. Zhao, "Panoformer: panorama transformer for indoor 360° depth estimation," in *ECCV*, 2022.

[219] I. Yun, C. Shin, H. Lee, H.-J. Lee, and C. E. Rhee, "Egformer: Equirectangular geometry-biased transformer for 360 depth estimation," in *ICCV*, 2023.

[220] P. Mohadikar and Y. Duan, "Omnidiffusion: Reformulating 360 monocular depth estimation using semantic and surface normal conditioned diffusion," in *WACV*, 2025.

[221] J. Lee, H. Park, B.-U. Lee, and K. Joo, "Hush: Holistic panoramic 3d scene understanding using spherical harmonics," in *CVPR*, 2025.

[222] F.-E. Wang, Y.-H. Yeh, M. Sun, W.-C. Chiu, and Y.-H. Tsai, "Bifuse: Monocular 360 depth estimation via bi-projection fusion," in *CVPR*, 2020.

[223] H. Jiang, Z. Sheng, S. Zhu, Z. Dong, and R. Huang, "Unifuse: Unidirectional fusion for 360 panorama depth estimation," *RAL*, 2021.

[224] C.-H. Peng and J. Zhang, "High-resolution depth estimation for 360deg panoramas through perspective and panoramic depth images registration," in *WACV*, 2023.

[225] J. Bai, H. Qin, S. Lai, J. Guo, and Y. Guo, "Glpanodepth: Global-to-local panoramic depth estimation," *TIP*, 2024.

[226] H. Ai, Z. Cao, Y.-P. Cao, Y. Shan, and L. Wang, "Hrdfuse: Monocular 360deg depth estimation by collaboratively learning holistic-with-regional depth distributions," in *CVPR*, 2023.

[227] H. Ai and L. Wang, "Elite360d: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion," in *CVPR*, 2024.

[228] ——, "Elite360m: Efficient 360 multi-task learning via bi-projection fusion and cross-task collaboration," *arXiv preprint arXiv:2408.09336*, 2024.

[229] Y. Li, Y. Guo, Z. Yan, X. Huang, Y. Duan, and L. Ren, "Omnifusion: 360 monocular depth estimation via geometry-aware fusion," in *CVPR*, 2022.

[230] M. Rey, M. Y. Area, and C. Richardt, "360monodepth: High-resolution 360 monocular depth estimation," in *CVPR*, 2022.

[231] Y. Benny and L. Wolf, "Sphereuformer: A u-shaped transformer for spherical 360 perception," in *CVPR*, 2025.

[232] M. Li, S. Wang, W. Yuan, W. Shen, Z. Sheng, and Z. Dong, "S2net: Accurate panorama depth estimation on spherical surface," *RAL*, 2023.

[233] P. Mohadikar, C. Fan, and Y. Duan, "Ms360: A multi-scale feature fusion framework for 360 monocular depth estimation," in *Proceedings of the 50th Graphics Interface Conference*, 2024.

[234] J. Zhang, Z. Chen, C. Lin, Z. Shen, L. Nie, K. Liao, and Y. Zhao, "Sgformer: Spherical geometry transformer for 360 depth estimation," *TCSVT*, 2025.

[235] Z. Shen, C. Lin, L. Nie, K. Liao, W. Lin, and Y. Zhao, "Revisiting 360 depth estimation with panogabor: A new fusion perspective," *arXiv preprint arXiv:2408.16227*, 2024.

[236] J. Deng, Y. Wang, H. Meng, Z. Hou, Y. Chang, and G. Chen, "Omnistereo: Real-time omnidireactional depth estimation with multiview fisheye cameras," in *CVPR*, 2025.

[237] G. Pintore, M. Agus, E. Almansa, J. Schneider, and E. Gobbetti, "Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation," in *CVPR*, 2021.

[238] C. Sun, M. Sun, and H.-T. Chen, "Hohonet: 360 indoor holistic understanding with latent horizontal features," in *CVPR*, 2021.

[239] H. Yu, L. He, B. Jian, W. Feng, and S. Liu, "Panelnet: Understanding 360 indoor environment via panel representation," in *CVPR*, 2023.

[240] K. Huang, F.-L. Zhang, F. Zhang, Y.-K. Lai, P. L. Rosin, and N. A. Dodgson, "Multi-task geometric estimation of depth and surface normal from monocular 360° images," *arXiv preprint arXiv:2411.01749*, 2024.

[241] M. Eder, P. Moulon, and L. Guan, "Pano popups: Indoor 3d reconstruction with a plane-aware network," in *3DV*, 2019.

[242] B. Y. Feng, W. Yao, Z. Liu, and A. Varshney, "Deep depth estimation on 360 images with a double quaternion loss," in *3DV*, 2020.

[243] N. Zioulis, A. Karakottas, D. Zarpalas, F. Alvarez, and P. Daras, "Spherical view synthesis for self-supervised 360 depth estimation," in *3DV*, 2019.

[244] X. Wang, W. Kong, Q. Zhang, Y. Yang, T. Zhao, and J. Jiang, "Distortion-aware self-supervised indoor 360° depth estimation via hybrid projection fusion and structural regularities," *TMM*, 2023.

[245] I. Yun, H.-J. Lee, and C. E. Rhee, "Improving 360 monocular depth estimation via non-local dense prediction transformer and joint supervised and self-supervised learning," in *AAAI*, 2022.

[246] C. Zhuang, Z. Lu, Y. Wang, J. Xiao, and Y. Wang, "Spdet: Edge-aware self-supervised panoramic depth estimation transformer with spherical geometry," *TPAMI*, 2023.

[247] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *ECCV*, 2016.

[248] N.-H. A. Wang and Y.-L. Liu, "Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation," in *NeurIPS*, 2024.

[249] Z. Cao, J. Zhu, W. Zhang, H. Ai, H. Bai, H. Zhao, and L. Wang, "Panda: Towards panoramic depth anything with unlabeled panoramas and mobius spatial augmentation," in *CVPR*, 2025.

[250] H. Huang, M. Solah, D. Li, and L.-F. Yu, "Audible panorama: Automatic spatial audio generation for panorama imagery," in *CHI*, 2019.

[251] D. Li, T. R. Langlois, and C. Zheng, "Scene-aware audio for 360 videos," *TOG*, 2018.

[252] P. Morgado, N. Nvasconcelos, T. Langlois, and O. Wang, "Self-supervised generation of spatial audio for 360 video," *NeurIPS*, 2018.

[253] Y. Zhu, X. Zhu, H. Duan, J. Li, K. Zhang, Y. Zhu, L. Chen, X. Min, and G. Zhai, "Audio-visual saliency for omnidirectional videos," in *ICIG*, 2023.

[254] Y. Zhang, F.-Y. Chao, W. Hamidouche, and O. Deforges, "Pav-sod: A new task towards panoramic audiovisual saliency detection," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023.

[255] H. Yun, Y. Yu, W. Yang, K. Lee, and G. Kim, "Pano-avqa: Grounded audio-visual question answering on 360deg videos," in *ICCV*, 2021.

[256] Y. Masuyama, Y. Bando, K. Yatabe, Y. Sasaki, M. Onishi, and Y. Oikawa, "Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling," in *IROS*, 2020.

[257] A. B. Vasudevan, D. Dai, and L. Van Gool, "Semantic object prediction and spatial sound super-resolution with binaural sounds," in *ECCV*, 2020.

[258] X. Zhu, H. Duan, Y. Cao, Y. Zhu, Y. Zhu, J. Liu, L. Chen, X. Min, and G. Zhai, "Perceptual quality assessment of omnidirectional audio-visual signals," in *CAAI*, 2023.

[259] R. F. Fela, A. Pastor, P. Le Callet, N. Zacharov, T. Vigier, and S. Forchhammer, "Perceptual evaluation on audio-visual dataset of 360 content," in *ICMEW*, 2022.

[260] S. Xie, H. Zhu, T. He, X. Li, and Z. Chen, "Sonic4d: Spatial audio generation for immersive 4d scene exploration," *arXiv preprint arXiv:2506.15759*, 2025.

[261] I. D. Miller, F. Cladera, T. Smith, C. J. Taylor, and V. Kumar, "Air-ground collaboration with spomp: Semantic panoramic online mapping and planning," *TFR*, 2024.

[262] H. Ma, J. Liu, Z. Hu, H. Qiu, D. Xu, Z. Wang, X. Gong, and S. Yang, "A method of generating measurable panoramic image for indoor mobile measurement system," *arXiv preprint arXiv:2010.14270*, 2020.

[263] B. Liu, G. Zhao, J. Jiao, G. Cai, C. Li, H. Yin, Y. Wang, M. Liu, and P. Hui, "Omnicolor: A global camera pose optimization approach of lidar-360camera fusion for colorizing point clouds," in *ICRA*, 2024.

[264] Z. Zhao, H. Yu, C. Lyv, W. Yang, and S. Scherer, "Attention-enhanced cross-modal localization between 360 images and point clouds," *arXiv preprint arXiv:2212.02757*, 2022.

[265] L. Bernreiter, L. Ott, J. Nieto, R. Siegwart, and C. Cadena, "Spherical multi-modal place recognition for heterogeneous sensor systems," in *ICRA*, 2021.

[266] Z. Yuan, T. Xu, X. Wang, J. Geng, and X. Yang, "Panoramic direct lidar-assisted visual odometry," *arXiv preprint arXiv:2409.09287*, 2024.

[267] S.-H. Chou, W.-L. Chao, W.-S. Lai, M. Sun, and M.-H. Yang, "Visual question answering on 360deg images," in *WACV*, 2020.

[268] I. Song, M. Joo, J. Kwon, and J. Lee, "Video question answering for people with visual impairments using an egocentric 360-degree camera," *arXiv preprint arXiv:2405.19794*, 2024.

[269] X. Zhang, Z. Ye, and X. Zheng, "Towards omnidirectional reasoning with 360-r1: A dataset, benchmark, and grpo-based method," *arXiv preprint arXiv:2505.14197*, 2025.

[270] Z. Dongfang, X. Zheng, Z. Weng, Y. Lyu, D. P. Paudel, L. Van Gool, K. Yang, and X. Hu, "Are multimodal large language models ready for omnidirectional spatial reasoning?" *arXiv preprint arXiv:2505.11907*, 2025.

[271] Y. Zhou, T. Zhang, D. Zhang, S. Ji, X. Li, and L. Qi, "Dense360: Dense understanding from omnidirectional panoramas," *arXiv preprint arXiv:2506.14471*, 2025.

[272] Z. Chen, G. Wang, and Z. Liu, "Text2light: Zero-shot text-driven hdr panorama generation," *TOG*, 2022.

[273] X. Sun, M. Xu, S. Li, S. Ma, X. Deng, L. Jiang, and G. Shen, "Spherical manifold guided diffusion model for panoramic image generation," in *CVPR*, 2025.

[274] T. Wu, X. Li, Z. Qi, D. Hu, X. Wang, Y. Shan, and X. Li, "Spherediffusion: Spherical geometry-aware distortion resilient diffusion model," in *AAAI*, 2024.

[275] Y. Xia, S. Weng, S. Yang, J. Liu, C. Zhu, M. Teng, Z. Jia, H. Jiang, and B. Shi, "Panowan: Lifting diffusion video generation models to 360° with latitude/longitude-aware mechanisms," *arXiv preprint arXiv:2505.22016*, 2025.

[276] J. Liu, S. Lin, Y. Li, and M.-H. Yang, "Dynamicscaler: Seamless and scalable video generation for panoramic scenes," in *CVPR*, 2025.

[277] M. Zhang, Y. Chen, R. Xu, C. Wang, J. Yang, W. Meng, J. Guo, H. Zhao, and X. Zhang, "Panodit: Panoramic videos generation with diffusion transformer," in *AAAI*, 2025.

[278] H. Çapuk, A. Bond, M. B. Kızıl, E. Göçen, E. Erdem, and A. Erdem, "Tandit: Tangent-plane diffusion transformer for high-quality 360° panorama generation," *arXiv preprint arXiv:2506.21681*, 2025.

[279] C. Zhang, Q. Wu, C. C. Gambardella, X. Huang, D. Phung, W. Ouyang, and J. Cai, "Taming stable diffusion for text to 360 panorama image generation," in *CVPR*, 2024.

[280] K. Xing, H. Liang, D. Xu, Y. Yin, K. N. Plataniotis, Y. Zhao, and Y. Wei, "Tip4gen: Text to immersive panorama 4d scene generation," *arXiv preprint arXiv:2508.12415*, 2025.

[281] Y. Huang, Y. Zhou, J. Wang, K. Huang, and X. Liu, "Dreamcube: 3d panorama generation via multi-plane synchronization," *arXiv preprint arXiv:2506.17206*, 2025.

[282] M. Park, T. Kang, J. Yun, S. Hwang, and J. Choo, "Spherediff: Tuning-free omnidirectional panoramic image and video generation via spherical latent representation," *arXiv preprint arXiv:2504.14396*, 2025.

[283] K. Xie, A. Sabour, J. Huang, D. Paschalidou, G. Klar, U. Iqbal, S. Fidler, and X. Zeng, "Videopanda: Video panoramic diffusion with multi-view attention," *arXiv preprint arXiv:2504.11389*, 2025.

[284] Z. Fang, K. Zhu, Z. Liu, Y. Liu, W. Zhai, Y. Cao, and Z.-J. Zha, "Panoramic video generation with pretrained diffusion models," *arXiv preprint arXiv:2506.23513*, 2025.

[285] M. Feng, J. Liu, M. Cui, and X. Xie, "Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models," *arXiv preprint arXiv:2311.13141*, 2023.

[286] H. Wang, X. Xiang, Y. Fan, and J.-H. Xue, "Customizing 360-degree panoramas through text-to-image diffusion models," in *WACV*, 2024.

[287] C. Wang, X. Li, L. Qi, X. Lin, J. Bai, Q. Zhou, and Y. Tong, "Conditional panoramic image generation via masked autoregressive modeling," *arXiv preprint arXiv:2505.16862*, 2025.

[288] A. Liu, Z. Li, Z. Chen, N. Li, Y. Xu, and B. A. Plummer, "Panofree: Tuning-free holistic multi-view image generation with cross-view self-guidance," in *ECCV*, 2024.

[289] J. Ma, E. Lu, R. Paiss, S. Zada, A. Holynski, T. Dekel, B. Curless, M. Rubinstein, and F. Cole, "Vidpanos: Generative panoramic videos from casual panning videos," in *SIGGRAPH Asia*, 2024.

[290] W. Li, S. Zhao, C. Mou, X. Sheng, Z. Zhang, Q. Wang, J. Li, L. Zhang, and J. Zhang, "Omnidrag: Enabling motion control for omnidirectional image-to-video generation," *arXiv preprint arXiv:2412.09623*, 2024.

[291] H. Ai, Z. Cao, H. Lu, C. Chen, J. Ma, P. Zhou, T.-K. Kim, P. Hui, and L. Wang, "Dream360: Diverse and immersive outdoor virtual scene creation via transformer-based 360 image outpainting," *TVCG*, 2024.

[292] D. Zheng, C. Zhang, X.-M. Wu, C. Li, C. Lv, J.-F. Hu, and W.-S. Zheng, "Panorama generation from nfov image done right," in *CVPR*, 2025.

[293] A. Nakata and T. Yamanaka, "2s-odis: Two-stage omni-directional image synthesis by geometric distortion correction," in *ECCV*, 2024.

[294] K. Liao, X. Xu, C. Lin, W. Ren, Y. Wei, and Y. Zhao, "Cylin-painting: Seamless 360 panoramic image outpainting and beyond," *TIP*, 2023.

[295] J. Wang, Z. Chen, J. Ling, R. Xie, and L. Song, "360-degree panorama generation from few unregistered nfov images," *arXiv preprint arXiv:2308.14686*, 2023.

[296] T. Wu, C. Zheng, and T.-J. Cham, "Panodiffusion: 360-degree panorama outpainting via diffusion," *arXiv preprint arXiv:2307.03177*, 2023.

[297] N. Akimoto, Y. Matsuo, and Y. Aoki, "Diverse plausible 360-degree image outpainting for efficient 3dcg background creation," in *CVPR*, 2022.

[298] M. R. K. Dastjerdi, Y. Hold-Geoffroy, J. Eisenmann, S. Khodadadeh, and J.-F. Lalonde, "Guided co-modulated gan for 360 field of view extrapolation," in *3DV*, 2022.

[299] Z. Lu, K. Hu, C. Wang, L. Bai, and Z. Wang, "Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation," in *AAAI*, 2024.

[300] C. Oh, W. Cho, Y. Chae, D. Park, L. Wang, and K.-J. Yoon, "Bips: Bi-modal indoor panorama synthesis via residual depth-aided adversarial learning," in *ECCV*, 2022.

[301] C. Choi, S. M. Kim, and Y. M. Kim, "Balanced spherical grid for egocentric view synthesis," in *CVPR*, 2023.

[302] Z. Chen, Y.-P. Cao, Y.-C. Guo, C. Wang, Y. Shan, and S.-H. Zhang, "Panogrf: Generalizable spherical radiance fields for wide-baseline panoramas," in *NeurIPS*, 2023.

[303] K. Gu, T. Maugey, S. Knorr, and C. Guillemot, "Omni-nerf: neural radiance field from 360 image captures," in *ICME*, 2022.

[304] D. Choi, H. Jang, and M. H. Kim, "Omnilocalrf: Omnidirectional local radiance fields from dynamic videos," in *CVPR*, 2024.

[305] H. Huang, Y. Chen, T. Zhang, and S.-K. Yeung, "360roam: Real-time indoor roaming using geometry-aware 360° radiance fields," *arXiv preprint arXiv:2208.02705*, 2022.

[306] S. Kulkarni, P. Yin, and S. Scherer, "360fusionnerf: Panoramic neural radiance fields with joint guidance," in *IROS*, 2023.

[307] G. Wang, P. Wang, Z. Chen, W. Wang, C. C. Loy, and Z. Liu, "Perf: Panoramic neural radiance field from a single panorama," *TPAMI*, 2024.

[308] P. Gera, M. R. K. Dastjerdi, C. Renaud, P. Narayanan, and J.-F. Lalonde, "Casual indoor hdr radiance capture from omnidirectional images," *arXiv preprint arXiv:2208.07903*, 2022.

[309] J. Bai, L. Huang, J. Guo, W. Gong, Y. Li, and Y. Guo, "360-gs: Layout-guided panoramic gaussian splatting for indoor roaming," *arXiv preprint arXiv:2402.00763*, 2024.

[310] S. Lee, J. Chung, J. Huh, and K. M. Lee, "Odgs: 3d scene reconstruction from omnidirectional images with 3d gaussian splattings," in *NeurIPS*, 2024.

[311] C. Zhang, H. Xu, Q. Wu, C. C. Gambardella, D. Phung, and J. Cai, "Pansplat: 4k panorama synthesis with feed-forward gaussian splatting," in *CVPR*, 2025.

[312] L. Li, H. Huang, S.-K. Yeung, and H. Cheng, "Omnigs: Fast radiance field reconstruction using omnidirectional gaussian splatting," *arXiv preprint arXiv:2404.03202*, 2024.

[313] Z. Chen, C. Wu, Z. Shen, C. Zhao, W. Ye, H. Feng, E. Ding, and S.-H. Zhang, "Splatter-360: Generalizable 360 gaussian splatting for wide-baseline panoramic images," in *CVPR*, 2025.

[314] S. Lee, J. Chung, K. Kim, J. Huh, G. Lee, M. Lee, and K. M. Lee, "Omnisplat: Taming feed-forward 3d gaussian splatting for omnidirectional images with editable capabilities," in *CVPR*, 2025.

[315] Z. Shen, C. Lin, S. Huang, L. Nie, K. Liao, and Y. Zhao, "You need a transition plane: Bridging continuous panoramic 3d reconstruction with perspective gaussian splatting," *arXiv preprint arXiv:2504.09062*, 2025.

[316] J. Ren, M. Xiang, J. Zhu, and Y. Dai, "Panosplatt3r: Leveraging perspective pretraining for generalized unposed wide-baseline panorama reconstruction," *arXiv preprint arXiv:2507.21960*, 2025.

[317] S. Ito, N. Takama, K. Ito, H.-T. Chen, and T. Aoki, "Erpgs: Equirectangular image rendering enhanced with 3d gaussian regularization," *arXiv preprint arXiv:2505.19883*, 2025.

[318] C. Shin, W. O. Cho, and S. J. Kim, "Seam360gs: Seamless 360° gaussian splatting from real-world omnidirectional images," *arXiv preprint arXiv:2508.20080*, 2025.

[319] S. Ito, N. Takama, T. Watanabe, K. Ito, H.-T. Chen, and T. Aoki, "Ob3d: A new dataset for benchmarking omnidirectional 3d reconstruction using blender," *arXiv preprint arXiv:2505.20126*, 2025.

[320] D. Li, Y. Zhang, C. Häne, D. Tang, A. Varshney, and R. Du, "Omnisyn: Synthesizing 360 videos with wide-baseline panoramas," in *VRW*, 2022.

[321] T. Habtegebrial, C. Gava, M. Rogge, D. Stricker, and V. Jampani, "Somsi: Spherical novel view synthesis with soft occlusion multi-sphere images," in *CVPR*, 2022.

[322] R. Chen, F.-L. Zhang, S. Finnie, A. Chalmers, and T. Rhee, "Casual 6-dof: free-viewpoint panorama using a handheld 360 camera," *TVCG*, 2022.

[323] W. Li, F. Cai, Y. Mi, Z. Yang, W. Zuo, X. Wang, and X. Fan, "Scenedreamer360: Text-driven 3d-consistent scene generation with panoramic gaussian splatting," *arXiv preprint arXiv:2408.13711*, 2024.

[324] S. Zhou, Z. Fan, D. Xu, H. Chang, P. Chari, T. Bharadwaj, S. You, Z. Wang, and A. Kadambi, "Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting," in *ECCV*, 2024.

[325] S. Yang, J. Tan, M. Zhang, T. Wu, Y. Li, G. Wetzstein, Z. Liu, and D. Lin, "Layerpano3d: Layered 3d panorama for hyper-immersive scene generation," *arXiv preprint arXiv:2408.13252*, 2024.

[326] H. Zhou, X. Cheng, W. Yu, Y. Tian, and L. Yuan, "Holodreamer: Holistic 3d panoramic world generation from text descriptions," *arXiv preprint arXiv:2407.15187*, 2024.

[327] H. Zhou, W. Yu, J. Guan, X. Cheng, Y. Tian, and L. Yuan, "Holotime: Taming video diffusion models for panoramic 4d scene generation," in *ACMMM*, 2025.

[328] R. Li, P. Pan, B. Yang, D. Xu, S. Zhou, X. Zhang, Z. Li, A. Kadambi, Z. Wang, Z. Tu *et al.*, "4k4dgen: Panoramic 4d generation at 4k resolution," in *ICLR*, 2025.

[329] H. Team, Z. Wang, Y. Liu, J. Wu, Z. Gu, H. Wang, X. Zuo, T. Huang, W. Li, S. Zhang *et al.*, "Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels," *arXiv preprint arXiv:2507.21809*, 2025.

[330] Z. Yang, W. Ge, Y. Li, J. Chen, H. Li, M. An, F. Kang, H. Xue, B. Xu, Y. Yin *et al.*, "Matrix-3d: Omnidirectional explorable 3d world generation," *arXiv preprint arXiv:2508.08086*, 2025.

[331] J. Li and M. Bansal, "Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation," in *NeurIPS*, 2023.

[332] S. Wang, D. Zhou, L. Xie, C. Xu, Y. Yan, and E. Yin, "Panogen++: Domain-adapted text-guided panoramic environment generation for vision-and-language navigation," *Neural Networks*, 2025.

[333] Z. Wei, B. Lin, Y. Nie, J. Chen, S. Ma, H. Xu, and X. Liang, "Unseen from seen: Rewriting observation-instruction using foundation models for augmenting vision-language navigation," *arXiv preprint arXiv:2503.18065*, 2025.

[334] L. Yang, H. Duan, Y. Zhu, X. Liu, L. Liu, Z. Xu, G. Ma, X. Min, G. Zhai, and P. L. Callet, "Omni ̋2: Unifying omnidirectional image generation and editing in an omni model," *arXiv preprint arXiv:2504.11379*, 2025.

[335] X. Lin, C. Ren, X. Liu, J. Huang, and Y. Lei, "Unsupervised image denoising in real-world scenarios via self-collaboration parallel generative adversarial branches," in *ICCV*, 2023.

[336] X. Lin, J. Yue, S. Ding, C. Ren, L. Qi, and M.-H. Yang, "Dual degradation representation for joint deraining and low-light enhancement in the dark," *TCSVT*, 2024.

[337] X. Lin, Y. Zhou, J. Yue, C. Ren, K. C. Chan, L. Qi, and M.-H. Yang, "Re-boosting self-collaboration parallel prompt gan for unsupervised image restoration," *TPAMI*, 2025.

[338] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," *TPAMI*, 2016.

[339] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *ECCVW*, 2018.

[340] L. Qi, J. Kuen, Y. Wang, J. Gu, H. Zhao, P. Torr, Z. Lin, and J. Jia, "Open world entity segmentation," in *TPAMI*, 2022.

[341] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *ICCV*, 2023.

[342] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[343] P. Zhao, A. You, Y. Zhang, J. Liu, K. Bian, and Y. Tong, "Spherical criteria for fast and accurate 360 object detection," in *AAAI*, 2020.

[344] J. Yue, Z. Lin, X. Lin, X. Zhou, X. Li, L. Qi, Y. Wang, and M.-H. Yang, "Roburcdet: Enhancing robustness of radar-camera fusion in bird's eye view for 3d object detection," in *ICLR*, 2025.

[345] L. Yang, L. Qi, X. Li, S. Li, V. Jampani, and M.-H. Yang, "Unified dense prediction of video diffusion," in *CVPR*, 2025.

[346] L. Qi, L. Yang, W. Guo, Y. Xu, B. Du, V. Jampani, and M.-H. Yang, "Unigs: Unified representation for image generation and segmentation," in *CVPR*, 2024.

[347] Q. Wang, W. Li, C. Mou, X. Cheng, and J. Zhang, "360dvd: Controllable panorama video generation with 360-degree video diffusion model," in *CVPR*, 2024.

[348] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, 2021.

[349] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *TOG*, 2023.

[350] B. Krolla, M. Diebold, B. Goldlücke, and D. Stricker, "Spherical light fields." in *BMVC*, 2014.

[351] T. L. T. da Silveira, P. G. L. Pinto, J. E. M. Llerena, and C. R. Jung, "3d scene geometry estimation from 360° imagery: A survey," *arXiv preprint arXiv:2401.09252*, 2024.

[352] S. Li and K. Fukumori, "Spherical stereo for the construction of immersive vr environment," in *IEEE Proceedings. VR 2005. Virtual Reality, 2005.*, 2005.

[353] M. Eder, M. Shvets, J. Lim, and J.-M. Frahm, "Tangent images for mitigating spherical distortion," in *CVPR*, 2020.

[354] S. Karigiannis, "Introduction to geometry," in *Lectures and Surveys on G2-manifolds and related topics*. Springer, 2020.

[355] Y. Lee, J. Jeong, J. Yun, W. Cho, and K.-J. Yoon, "Spherephd: Applying cnns on a spherical polyhedron representation of 360deg images," in *CVPR*, 2019.

[356] T. T. H. Uyen, O.-J. Kwon, S. Choi, and I. Hussain, "Subjective assessment of 360 image projection formats," *IEEE Access*, 2020.

[357] Z. Chen, Y. Li, and Y. Zhang, "Recent advances in omnidirectional video coding for virtual reality: Projection and evaluation," *Signal Processing*, 2018.

[358] J. Tremblay, M. Meshry, A. Evans, J. Kautz, A. Keller, S. Khamis, T. Müller, C. Loop, N. Morrical, K. Nagano *et al.*, "Rtmv: A ray-traced multi-view synthetic dataset for novel view synthesis," *arXiv preprint arXiv:2205.07058*, 2022.

[359] S.-H. Chang, C.-Y. Chiu, C.-S. Chang, K.-W. Chen, C.-Y. Yao, R.-R. Lee, and H.-K. Chu, "Generating 360 outdoor panorama dataset with reliable sun position estimation," in *SIGGRAPH Asia*, 2018.

[360] A. Hamdi, B. Ghanem, and M. Nießner, "Sparf: Large-scale learning of 3d sparse radiance fields from few input images," in *ICCV*, 2023.

[361] T. Sugg, K. O'Brien, L. Poudel, A. Dumouchelle, M. Jou, M. Bosch, D. Ramanan, S. Narasimhan, and S. Tulsiani, "Accenture-nvs1: A novel view synthesis dataset," *arXiv preprint arXiv:2503.18711*, 2025.

[362] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J.-F. Lalonde, "Learning to predict indoor illumination from a single image," *TOG*, 2017.

[363] Z. Ni, S. Du, Z. Hou, C. Wu, and S. Yang, "Para-lane: Multi-lane dataset registering parallel scans for benchmarking novel view synthesis," *arXiv preprint arXiv:2502.15635*, 2025.

[364] L. Zhou, K. Han, N. Ling, W. Wang, and W. Jiang, "Gameir: A large-scale synthesized ground-truth dataset for image restoration over gaming content," in *ACCV*, 2024.

[365] X. Deng, H. Wang, M. Xu, Y. Guo, Y. Song, and L. Yang, "Lau-net: Latitude adaptive upscaling network for omnidirectional image super-resolution," in *CVPR*, 2021.

[366] M. Cao, C. Mou, F. Yu, X. Wang, Y. Zheng, J. Zhang, C. Dong, G. Li, Y. Shan, R. Timofte *et al.*, "Ntire 2023 challenge on 360° omnidirectional image and video super-resolution: Datasets, methods and results," in *CVPRW*, 2023.

[367] H. R. V. Joze, I. Zharkov, K. Powell, C. Ringler, L. Liang, A. Roulston, M. Lutz, and V. Pradeep, "Imagepairs: Realistic super resolution dataset via beam splitter camera rig," in *CVPRW*, 2020.

[368] S.-H. Chou, C. Sun, W.-Y. Chang, W.-T. Hsu, M. Sun, and J. Fu, "360-indoor: Towards learning real-world objects in 360° indoor equirectangular images," in *WACV*, 2020.

[369] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *arXiv preprint arXiv:0909.5206*, 2010.

[370] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *IJCV*, 2020.

[371] D. Yu and S. Ji, "Grid based spherical cnn for object detection from panoramic images," *Sensors*, 2019.

[372] T. Xiao, B. Li, X. Zhang, G. Yu, Z. Shen, Z. Wei, Z. Xiong, Z. Luo, Z. Zhang, and Z. Zhang, "Objects365: A large-scale, high-quality dataset for object detection," in *ICCV*, 2019.

[373] H. Xu, Q. Zhao, Y. Ma, X. Li, P. Yuan, B. Feng, C. Yan, and F. Dai, "Pandora: A panoramic detection dataset for object with orientation," in *ECCV*, 2022.

[374] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *CVPR*, 2020.

[375] A. Gupta, P. Dollar, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in *CVPR*, 2019.

[376] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *IJCV*, 2020.

[377] K. Yang, J. Zhang, S. Reiß, X. Hu, and R. Stiefelhagen, "Capturing omni-range context for omnidirectional segmentation," in *CVPR*, 2021.

[378] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *CVPR*, 2017.

[379] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *arXiv preprint arXiv:1702.01105*, 2017.

[380] J. Li, J. Hu, Y. Huang, Z. Chen, B. Gao, J. Jiang, and Y. Zhang, "A synthetic digital city dataset for robustness and generalisation of depth estimation models," *Scientific Data*, 2024.

[381] M. Li, X. Jin, X. Hu, J. Dai, S. Du, and Y. Li, "Mode: Multi-view omnidirectional depth estimation with 360° cameras," in *ECCV*, 2022.

[382] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *CVPR*, 2015.

[383] Y. Rai, J. Gutierrez, and P. Le Callet, "A dataset of head and eye movements for 360 degree images," in *ACMMM*, 2017.

[384] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," *arXiv preprint arXiv:1505.03581*, 2015.

[385] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang, "Predicting head movement in panoramic video: A deep reinforcement learning approach," *TPAMI*, 2018.

[386] H. J. Lin, S.-W. Huang, S.-H. Lai, and C.-K. Chiang, "Indoor scene layout estimation from a single image," in *ICPR*, 2018.

[387] G. Pintore, E. Almansa, M. Agus, and E. Gobbetti, "Deep3dlayout: 3d reconstruction of an indoor layout from a spherical panoramic image," *TOG*, 2021.

APPENDIX

This supplementary material document provides a detailed introduction of some common projection formats, task-specific future directions, and a comparison of perspective and panoramic datasets. The supplementary material is organized as follows:

1. Projection.
2. Future Work by Tast.
3. Comparison of Perspective and Panoramic Datasets.

PROJECTION

Here, we provide a more detailed summary of representative projections commonly adopted in panoramic vision beyond the main text, with a focus on their characteristics and limitations.

Spherical Projection. A 360° camera can be modeled as a pinhole at the center of a unit sphere, projecting all visible 3D points onto its surface without requiring traditional intrinsic parameters [350]–[352]. A 3D point in the world Cartesian coordinate system $P = [x, y, z]^T$ is first transformed into spherical coordinates (ρ, θ, ϕ), where $\rho = \sqrt{x^2 + y^2 + z^2}$, $\theta = \arccos(z/\rho)$, and $\phi = \arctan 2(y, x)$. By normalizing with respect to ρ, the point is mapped onto the unit sphere, yielding the unit vector $p = (\sin\theta\cos\phi,\ \sin\theta\sin\phi,\ \cos\theta) = (x', y', z')$. The unit sphere is centered at the origin of the world coordinate system. This spherical projection provides a unified, distortion-free representation of all viewing directions, serving as a fundamental basis for analyzing panoramic images and videos.

Equirectangular Projection (ERP). As the most common format for 360-degree panoramas, ERP maps spherical coordinates (ϕ, θ) directly and uniformly to a 2D image plane: ϕ (longitude) to the horizontal axis and θ (latitude) to the vertical axis, with sampling intervals $\Delta\phi = 2\pi/w$ and $\Delta\theta = \pi/h$ for image width w and height h (typically $w = 2h$). A pixel (u, v) corresponds to $(\phi, \theta) = (u \cdot \Delta\phi - \pi,\ v \cdot \Delta\theta)$. Such a simple bijective mapping is analogous to the transformation of the Earth's spherical surface into a world map, making ERP efficient for rendering, editing, and training vision models. Unless specified otherwise, we use ERP as the default projection in this survey.

Cubemap Projection (CMP). The CMP is a widely used alternative to ERP that alleviates geometric distortions, particularly the stretching near the poles. It projects spherical content onto the six faces of a cube, each covering a $90° \times 90°$ field of view. Each cube is centered at the camera origin, with faces oriented toward the front, back, left, right, top, and bottom. For a 3D point $p = [x', y', z']^T$ on the unit sphere, the face index is determined by the dominant coordinate axis, and the point is projected onto that face via perspective projection. For example, if P maps to the front face, the 2D coordinates are

$$u = \frac{x}{|z|}, \quad v = \frac{y}{|z|}.$$

Each face is rendered as a $w \times w$ square, producing a complete 6 unfolded layouts. Overall, the CMP provides a distortion-reduced, face-based representation well-suited for panoramic rendering and processing.

Tangent Projection (TP). To adapt perspective models for high-resolution spherical data, [353] proposes TP, which renders an omnidirectional image as local planar grids tangent to the faces of a subdivided icosahedron. As illustrated in Fig. 2(c), each patch is obtained from a gnomonic projection [354], mapping a point P_s in the sphere to a tangent plane centered at P_c. A point in the unit sphere with spherical coordinates (θ, ϕ) is projected directly onto a tangent plane centered at (θ_c, ϕ_c), yielding plane coordinates (u_t, v_t):

$$u_t = \frac{\cos\phi\ \sin(\theta - \theta_c)}{\cos c},$$
$$v_t = \frac{\cos\phi_c\ \sin\phi - \sin\phi_c\ \cos\phi\ \cos(\theta - \theta_c)}{\cos c},$$

where c is the central angle between (θ, ϕ) and the tangent point (θ_c, ϕ_c). $\cos c = \sin\phi_c \sin\phi + \cos\phi_c \cos\phi \cos(\theta - \theta_c)$. The inverse mapping follows the same geometric principle, enabling one-to-one correspondence between spherical coordinates and tangent patches [229]. Overall, TP preserves local geometric fidelity and reduces distortion compared to ERP. It also allows the reuse of standard perspective vision models on spherical imagery.

Polyhedron Projection (PP). To reduce distortions from mapping spherical images to planar formats while preserving directional continuity, PP [355] approximates the sphere with a subdivided polyhedron, such as an octahedron or icosahedron. Each face of the base polyhedron can be recursively divided into four smaller faces, increasing resolution and reducing distortion [355]. For instance, an icosahedron at level l yields 20×4^l triangular faces, providing a near-uniform sphere sampling.

Panini Projection. The Panini projection mitigates the strong distortions of rectilinear projection at wide fields of view (typically $> 70°$) by preserving vertical and radial lines while non-linearly compressing the horizontal field. It maintains a strong central perspective without exaggerating objects near the periphery. Specifically, the mapping consists of projecting a unit spherical point (ϕ, θ) to a cylinder, followed by a rectilinear projection from a variable center:

$$S = \frac{d+1}{d + \cos\phi}, \quad h = S \cdot \sin\phi, \quad v = S \cdot \tan\theta,$$

where $d \geq 0$ controls horizontal compression. $d = 0$ yields a rectilinear projection, $d = 1$ produces a cylindrical stereographic projection, and $d \to \infty$ approximates the orthographic projection. This flexibility enables a smooth trade-off between central magnification and edge compression.

Other projections. Apart from the aforementioned popular projection formats, there are several other formats supported by the 360Lib software package for coding and processing [356]. They include adjusted equal-area projection (AEP), truncated square pyramid projection (TSP), adjusted cubemap (ACP), rotated sphere projection (RSP), equatorial cylindrical projection (ECP), equiangular cubemap (EAC), and hybrid equiangular cubemap (HEC). Especially, as a map-based projection, AEP adaptively decreases the sampling rate in vertical coordinates and avoids the over-sampling problem in ERP. Equi-Angular Cubemap (EAC) projection [357] maintains spatial sampling rates for different sampling locations on the faces of the cubes to alleviate distortions in CP.

COMPARISON OF PERSPECTIVE AND PANORAMIC DATASETS.

Table 1 shows a striking imbalance phenomenon between perspective and panoramic datasets across a wide range of tasks. For perspective vision, large-scale datasets such as Open Images (9.2M images), Objects365 (600k images), and ScanNet (2.5M frames) have provided abundant data to train powerful foundation

TABLE 1: Comparison of Perspective and Panoramic Datasets across Multiple Tasks

Task	Perspective Dataset	Data Size	Source	Panoramic Dataset	Data Size	Source
Generation	RTMV [358]	300k	Synthetic	360SP [359]	15,730	Real
	SPARF [360]	17M	Synthetic	HDR360-UHD [272]	4,392	Real
	ACC-NVS1 [361]	148k	Real	Laval Indoor HDR [362]	2,100	Real
	Para-Lane [363]	80k	Real	Laval Outdoor HDR [202]	205	Real
Super-Resolution	GameIR-SR [364]	19,200	Synthetic	ODI-SR [365]	1,000	Real
	GameIR-NVS [364]	57,600	Synthetic	Flickr360 [366]	3,150	Real
	ImageParis [367]	11,421	Real	ODV360 [366]	250	Real
Object Detection	MS COCO [2]	328k	Real	360-Indoor [368]	3,000	Real
	Pascal VOC [369]	21k	Real	ERA [126]	903	Real
	Open Images V7 [370]	1.9M	Real	OVS [371]	600	Real
	Objects365 [372]	600k	Real	PANDORA [373]	3,000	Real
	BDD100K [374]	100k	Real	COCO-Men [343]	7,000	Synthetic
Segmentation	LVIS [375]	164k	Real	PASS [105]	1,050	Real
	Open Images V4 [376]	9.2M	Real	WildPASS [377]	2,500	Real
Depth Estimation	ScanNet [378]	5M	Synthetic	Stanford2D3D [379]	1,413	Real
	SDCD [380]	930k	Real	Deep360 [381]	2,100	Real
Saliency Prediction	SALICON [382]	10k	Real	Salient360! [383]	85	Real
	CAT2000 [384]	4k	Real	PVS-HM [385]	76	Real
Room Layout	LSUN Room Layout [386]	5,394	Real	Pano3DLayout [387]	107	Synthetic

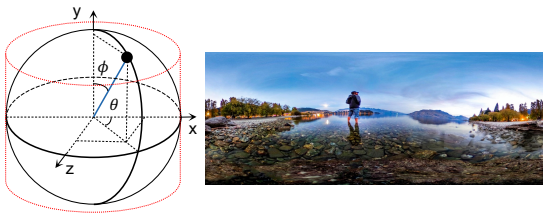

Fig. 11: Equirectangular projection (ERP) mapping spherical coordinates (ϕ, θ) to image pixels (u, v), analogous to flattening the Earth onto a world map.

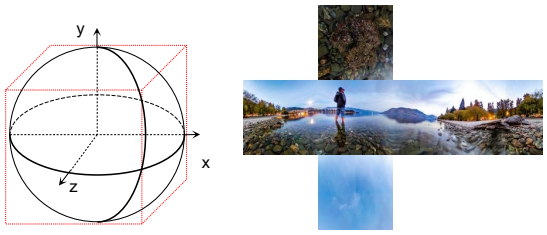

Fig. 12: Cubemap projection (CMP) mapping spherical content onto six cube faces via perspective projection, reducing polar distortions compared with ERP.

models. By contrast, panoramic datasets remain relatively scarce, often orders of magnitude smaller: PASS has only 1,050 annotated panoramas for segmentation, WildPASS 2,500 samples, and Deep360 merely 2k panoramic depth maps. Even in generation and super-resolution tasks, panoramic datasets usually contain only a few thousand samples, compared with hundreds of thousands or even millions on the perspective side.

This imbalance highlights a critical bottleneck: the lack of large-scale, diverse, and richly annotated panoramic datasets hinders the development of generalizable models and fair benchmarking across tasks. While perspective vision has benefited greatly

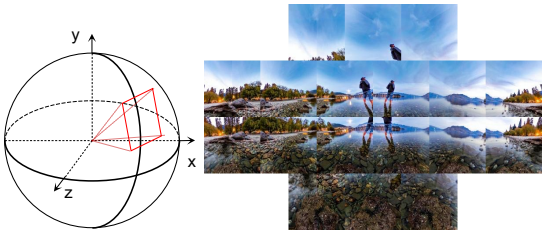

Fig. 13: Tangent projection (TP) maps spherical points onto tangent planes via gnomonic projection, preserving local geometry and enabling perspective model reuse.

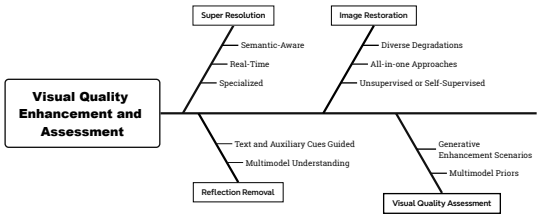

Fig. 14: Summary of Future Directions in Visual Quality Enhancement and Assessment.

from the scale of data, panoramic vision is still constrained by limited data availability, preventing models from fully exploiting the potential of 360° understanding and generation. Bridging this gap is therefore an urgent priority for the community, calling for systematic efforts in panoramic data collection, annotation, and benchmark design.

Future Work by Task

Visual Quality Enhancement and Assessment

As summarized in Fig. 14, future research on visual quality enhancement and assessment for panoramic vision can be explored along several directions.

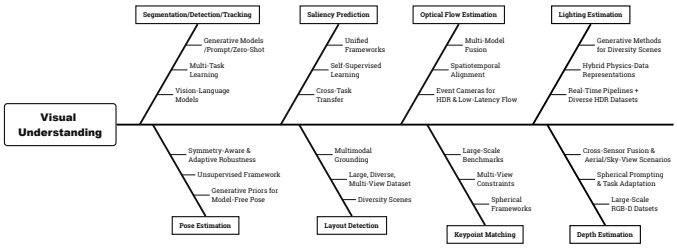

Fig. 15: Summary of Future Directions in Visual Understanding.

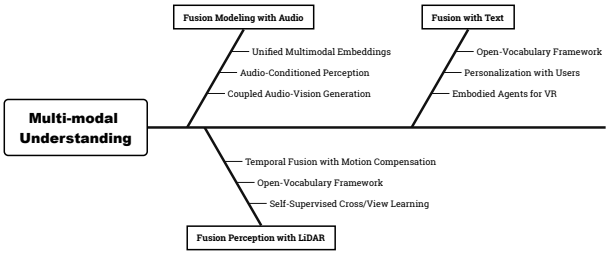

Fig. 16: Summary of Future Directions in Multi-modal Understanding.

Super-Resolution: Leverage diffusion models with textual guidance for semantically aware panoramic restoration, improve real-time efficiency via pruning and architectural optimization, and incorporate modules specialized for key categories such as human faces.

Reflection Removal: Move toward reflection separation by leveraging text or auxiliary cues to disentangle reflection and transmission, while extending to outdoor scenarios with semantic and multimodal understanding to improve robustness and generalization.

Image Restoration: Address diverse degradations (e.g., rain, snow, fog, low-light). Current models are often single-task and trained on synthetic data, highlighting the need for more generalized all-in-one approaches and unsupervised or self-supervised frameworks that adapt to real-world panoramic conditions.

Visual Quality Assessment: FR-OIQA often assumes perfect references, which is unrealistic given sensor noise and stitching artifacts; future work should jointly assess fidelity and perceptual naturalness, particularly in generative enhancement scenarios. Meanwhile, NR-OIQA suffers from limited supervision and weak generalization; integrating vision–language pipelines with LLMs can provide semantic priors and subjective cues to improve distortion reasoning and quality prediction.

Visual Understanding

Future directions in panoramic visual understanding are summarized in Fig. 15.

Segmentation: Use generative models (e.g., diffusion, masked autoencoders) to improve representation learning and zero-shot transfer; adopt multi-task learning with generation tasks (e.g., inpainting, novel view synthesis) to yield richer features; and leverage large vision–language models for open-world segmentation with weakly supervised multimodal alignment.

Detection: Employ prompt-driven detection for flexible zero-shot queries; integrate multimodal inputs (RGB, depth, thermal, text) to improve robustness and generalization; and adopt open-vocabulary detection to foster lifelong learning and uncertainty awareness. Future progress hinges on integrating these capabilities with large-scale vision–language understanding and continual learning for safe, scalable detection.

Tracking: Explore zero-shot paradigms with vision–language models for prompt-based initialization and semantic reasoning, unified frameworks integrating detection, tracking, and segmentation for spatiotemporal consistency, and cross-view or multimodal fusion for robustness under occlusion. Uncertainty-aware designs further enable long-term, real-world deployment, pointing toward more open, semantic-aware, and generalizable tracking systems.

Pose Estimation: Improve generalization and reduce annotation reliance by leveraging generative priors for model-free object pose, enhance robustness in category-level tasks with symmetry-aware and adaptive methods, and advance unsupervised facial/head pose estimation via geometric cues, temporal coherence, and pre-trained representations.

Saliency Prediction: Develop unified frameworks combining distortion-aware architectures, dynamic attention, and multi-modal cues under spherical representations, while leveraging self-supervised panoramic video learning and cross-task transfer (e.g., saliency to scanpath) for robust and human-aligned saliency modeling.

Layout Detection: Generalize beyond Manhattan-aligned scenes to non-Manhattan, multi-level, or dynamic layouts. Dataset scale and diversity remain insufficient, highlighting the need for larger, richly annotated multi-view panoramas. Integrating multimodal cues (e.g., depth, audio, inertial data, language) could further ground layout estimation in embodied perception, enabling more intelligent agents in immersive environments.

Optical Flow Estimation: Advance multi-modal panoramic flow by integrating RGB-D cues for robustness under occlusion or low light, and leverage event cameras for low-latency, HDR motion capture. Networks that fuse event streams with image-based representations via spatiotemporal alignment or neural fusion, supported by dedicated RGB-Event or RGB-D datasets, could enable more consistent and reliable flow estimation in challenging conditions.

Keypoint Matching: Develop spherical frameworks with rotation-equivariant features, incorporate multi-view constraints for robust correspondence, and build large-scale benchmarks with ground-truth panoramic matches.

Decomposition: Develop unified spherical models capable of handling dynamic scenes and complex materials, supported by larger and more diverse panoramic datasets.

Lighting Estimation: Current generative methods remain limited under diverse outdoor and dynamic lighting, motivating the use of temporal cues and multimodal signals. Physically inspired approaches need hybrid physics–data models for greater realism and interpretability. Structured pipelines should be optimized for real-time, lightweight VR/AR deployment, while diverse HDR panoramic datasets with scene–illumination pairs are essential for robust generalization.

Depth Estimation: Tackle the shortage of high-quality RGB-D panoramas by building large-scale datasets for foundation models. Advances may come from spherical prompting and task adaptation for better transfer, cross-sensor fusion with LiDAR or IMU for real-world robustness, and extending beyond ground-level to aerial and sky-view scenarios.

Multi-modal Understanding

As illustrated in Fig. 16, multi-modal understanding provides rich future directions.

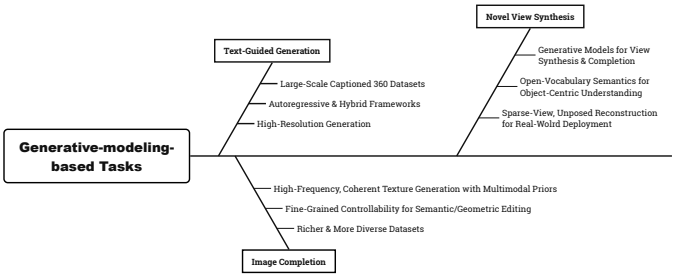

Fig. 17: Summary of Future Directions in Generative-modeling-based Tasks.

Fusion Modeling with Audio: Pursue unified embeddings of spatial, semantic, and temporal cues, audio-conditioned perception for attention and control, and real-time multimodal modeling for embodied AI. Key challenges include reverberation, occlusion-aware propagation, and adaptive feedback, while coupled audio–vision generation (e.g., text-conditioned 4D synthesis with ambient sound) offers another promising path.

Fusion Perception with LiDAR: Explore self-supervised learning with cross-view/modal consistency for robust features, open-vocabulary detection and segmentation via vision–language integration, and temporal fusion with motion compensation to improve dynamic scene perception.

Fusion with Text: Advance open-vocabulary detection and viewpoint-aware grounding, develop lightweight and distortion-aware models for efficient deployment, enhance personalization through user intent and preference alignment, and enable embodied agents for VR instruction following, spatial dialogue, and multimodal 360° interaction.

Generative-modeling-based Tasks

Finally, as summarized in Fig. 17, generative modeling for panoramic data opens several new avenues.

Text-guided Generation: Build large-scale, captioned 360° datasets to improve scalability and semantic grounding; explore autoregressive and hybrid frameworks beyond diffusion for long-range spherical modeling; and advance high-resolution generation by overcoming memory and architectural constraints. Tackling these directions will enable semantically aligned, visually consistent panoramic content for diverse downstream applications.

Image Completion: Expand dataset scale, diversity, and realism to improve generalization; enhance controllability for fine-grained semantic and geometric editing; and improve the generation of high-frequency, coherent textures over large missing regions. Promising directions include building richer datasets, developing interactive and viewpoint-aware frameworks, and integrating external priors with multimodal cues for more realistic and controllable 360° scene completion.

Novel View Synthesis: Leverage generative models for view synthesis and completion, integrate open-vocabulary semantics for object-centric understanding, jointly model motion and geometry for dynamic scenes, and reduce reliance on dense posed inputs to enable sparse-view, unposed reconstruction in real-world settings.