Novelty Assessment Report

Paper: SatDreamer360: Multiview-Consistent Generation of Ground-Level Scenes from Satellite Imagery

PDF URL: https://openreview.net/pdf?id=wmQoigkqUt

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-05

Abstract

Generating multiview-consistent 360° ground-level scenes from satellite imagery is a challenging task with broad applications in simulation, autonomous navigation, and digital twin cities. Existing approaches primarily focus on synthesizing individual ground-view panoramas, often relying on auxiliary inputs like height maps or handcrafted projections, and struggle to produce multiview consistent sequences. In this paper, we propose SatDreamer360, a framework that generates geometrically consistent multi-view ground-level panoramas from a single satellite image, given a predefined pose trajectory. To address the large viewpoint discrepancy between ground and satellite images, we adopt a triplane representation to encode scene features and design a ray-based pixel attention mechanism that retrieves view-specific features from the triplane. To maintain multi-frame consistency, we introduce a panoramic epipolar-constrained attention module that aligns features across frames based on known relative poses. To support the evaluation, we introduce VIGOR++, a large-scale dataset for generating multi-view ground panoramas from a satellite image, by augmenting the original VIGOR dataset with more ground-view images and their pose annotations. Experiments show that SatDreamer360 outperforms existing methods in both satellite-to-ground alignment and multiview consistency.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Generating Multiview-Consistent Ground-Level Scenes from Satellite Imagery**

A total of **34 papers** were analyzed and organized into a taxonomy with **16 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Cross-View Image Synthesis and Generation**
- **3D Scene Reconstruction and Representation**
- **Cross-View Geo-Localization**
- **Multiview Scene Analysis and Understanding**
- **Datasets, Benchmarks, and Survey Studies**

Complete Taxonomy Tree

- Generating Multiview-Consistent Ground-Level Scenes from Satellite Imagery Survey Taxonomy
- Cross-View Image Synthesis and Generation
  - Single-View Ground Image Synthesis
  - Geometry-Guided Synthesis (3 papers)
    - [2] Geospecific view generation geometry-context aware high-resolution ground view inference from satellite views (Ningli Xu, 2024) View paper
    - [7] Geometry-Guided Street-View Panorama Synthesis From Satellite Imagery (Shi YuJiao, 2022) View paper
    - [17] Geometry-aware satellite-to-ground image synthesis for urban areas (Xiaohu Lu, 2020) View paper
  - Learning-Based Synthesis (3 papers)
    - [14] View Synthesis with Scene Recognition for Cross-View Image Localization (Uddom Lee, 2023) View paper
    - [18] RETRACTED CHAPTER: CrossViewDiff: A Cross-View Diffusion Model for Satellite-to-Ground Image Synthesis (Y Chen, 2024) View paper
    - [33] What Is It Like Down There? Generating Dense Ground-Level Views and Image Features From Overhead Imagery Using Conditional Generative Adversarial Networks (Xueqing Deng, 2018) View paper
  - Controllable and Pose-Aligned Synthesis (2 papers)
    - [5] Controllable Satellite-to-Street-View Synthesis with Precise Pose Alignment and Zero-Shot Environmental Control (Song, 2025) View paper
    - [24] From Satellite to Street: A Hybrid Framework Integrating Stable Diffusion and PanoGAN for Consistent Cross-View Synthesis (Anwar Abbas, 2025) View paper
  - Multiview-Consistent Synthesis
  - Panoramic Video Synthesis ★ (3 papers)
    - [0] SatDreamer360: Multiview-Consistent Generation of Ground-Level Scenes from Satellite Imagery (Anon et al., 2026) View paper
    - [11] Sat2Vid: Street-view Panoramic Video Synthesis from a Single Satellite Image (Li, 2021) View paper
    - [27] SatDreamer360: Geometry Consistent Street-View Video Generation from Satellite Imagery (Zhu Bei-Yi, 2025) View paper
  - Large-Scale Consistent View Generation (1 papers)
    - [1] Satellite to GroundScape-Large-scale Consistent Ground View Generation from Satellite Views (Xu, 2025) View paper
  - BEV-Conditioned Street-View Synthesis (2 papers)
  - [3] Street-View Image Generation From a Bird's-Eye View Layout (Alexander Swerdlow, 2024) View paper

  - [4] BEVControl: Accurately Controlling Street-view Elements with Multi-perspective Consistency via BEV Sketch Layout (Yang, 2023) View paper
- 3D Scene Reconstruction and Representation
  - Neural Radiance Field Based Reconstruction (3 papers)
  - [13] Cross-View Geo-Localization via 3D Gaussian Splatting-Based Novel View Synthesis (Xuanyu Zhang, 2025) View paper
  - [15] Sat2Density: Faithful Density Learning from Satellite-Ground Image Pairs (Ming Qian, 2023) View paper
  - [21] Seeing through Satellite Images at Street Views (Qian Ming, 2025) View paper
  - Explicit 3D Geometry Reconstruction (1 papers)
  - [10] Cross-view SLAM solver: Global pose estimation of monocular ground-level video frames for 3D reconstruction using a reference 3D model from satellite images (Mostafa Elhashash, 2022) View paper
  - Large-Scale Urban 3D Generation (3 papers)
  - [9] Sat2scene: 3d urban scene generation from satellite images with diffusion (Zuoyue Li, 2024) View paper
  - [20] MagicCity: Geometry-Aware 3D City Generation from Satellite Imagery with Multi-View Consistency (X Yao, 2025) View paper
  - [32] Packing Urban Scenes Into Neural Radiance Field: 3D Scene Rendering, Manipulation and Generation (Xiangli, 2023) View paper
- Cross-View Geo-Localization
  - Feature Learning and Matching (3 papers)
  - [8] View from above: Orthogonal-view aware cross-view localization (Shan Wang, 2024) View paper
  - [16] Groundâ□□Satellite coupling for cross-view geolocation combined with multiscale fusion of spatial features (Luying Zhao, 2024) View paper
  - [22] Mutual Relative Position Learning Transformer for Cross-View Geo-Localization (Bo Gu, 2023) View paper
  - Geometry-Aware Localization (3 papers)
  - [19] Accurate 3-DoF Camera Geo-Localization via Ground-to-Satellite Image Matching (Shi YuJiao, 2022) View paper
  - [28] Geometry-Aware Enhancement and Data Augmentation for Street-to-Satellite Geo-localization (Xingbo Wang, 2025) View paper
  - [34] Unleashing Unlabeled Data: A Paradigm for Cross-View Geo-Localization (Supplementary Material) (G Li, n.d.) View paper
- Multiview Scene Analysis and Understanding
  - Cross-View Semantic Analysis (2 papers)
  - [23] Toward Seamless Multiview Scene Analysis From Satellite to Street Level (Lefevre Sebastien, 2017) View paper
  - [29] A Study of Remote Sensing Based Natural and Built Environment Monitoring: From Fully Supervised to Weakly Supervised Learning (Zhang, 2024) View paper
  - Multilevel Map Translation (1 papers)
  - [6] Level-Aware Consistent Multilevel Map Translation From Satellite Imagery (Ying Fu, 2022) View paper
  - Reverse Synthesis and Augmentation (1 papers)
  - [12] Satellite image synthesis from street view with fine-grained spatial textual guidance: A novel framework (Junyan Ye, 2025) View paper
- Datasets, Benchmarks, and Survey Studies
  - Cross-View Task Surveys and Reviews (2 papers)
  - [26] Cross-view Localization and Synthesis--Datasets, Challenges and Opportunities (Xu, 2025) View paper
  - [30] Advances in Neural Radiance Fields for Large-Scale 3D Scene Reconstruction: A Comprehensive Review (Yu Du, 2024) View paper
  - Application-Specific Studies (2 papers)
  - [25] Generative Towns (Simhadri, 2025) View paper
  - [31] Consistent ground-plane mapping: A case study utilizing low-cost sensor measurements and a satellite image (Hang Chu, 2015) View paper

Narrative

Core task: Generating multiview-consistent ground-level scenes from satellite imagery. This field addresses the challenge of synthesizing realistic street-level views conditioned on overhead satellite or bird's-eye-view (BEV) inputs, bridging the substantial geometric and appearance gap between aerial and ground perspectives. The taxonomy organizes research into several main branches: Cross-View Image Synthesis and Generation focuses on methods that translate satellite imagery into ground-level views, often leveraging generative models and geometric priors; 3D Scene Reconstruction and Representation explores volumetric or neural approaches to build consistent 3D models from cross-view data; Cross-View Geo-Localization tackles the retrieval and matching problem between aerial and ground images; Multiview Scene Analysis and Understanding examines broader perception tasks across viewpoints; and Datasets, Benchmarks, and Survey Studies provide the evaluation infrastructure and literature reviews. Within Cross-View Image Synthesis, a dense cluster of works addresses multiview-consistent synthesis, with some methods targeting panoramic or video outputs to ensure temporal and spatial coherence across generated frames.

Recent efforts reveal contrasting strategies for achieving consistency and realism. Many studies employ diffusion models or GANs with explicit geometric guidance—such as BEVControl[4] and Controllable Sat2Street[5]—to condition generation on layout or depth cues, while others like Sat2Density[15] and Sat2Scene[9] incorporate 3D representations to enforce structural coherence. A smaller handful of works, including Sat2Vid[11] and SatDreamer360 Video[27], extend synthesis to dynamic panoramic video sequences, emphasizing smooth temporal transitions. SatDreamer360[0] sits within this panoramic video synthesis cluster, sharing the goal of producing temporally consistent 360-degree ground-level videos from satellite input. Compared to earlier image-based approaches like BEV to StreetView[3] or Geospecific View Generation[2], SatDreamer360[0] emphasizes full panoramic coverage and video coherence, aligning closely with SatDreamer360 Video[27] in its focus on immersive, multiview-consistent outputs. The main open questions revolve around balancing geometric fidelity with photorealistic detail and scaling these methods to diverse urban environments.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Sat2Vid: Street-view Panoramic Video Synthesis from a Single Satellite Image

Authors: Li, Zuoyue, Zuoyue Li, Li Zhenqiang, Zhenqiang Li, et al. (16 authors total) | **Year/Venue**: 2021 | **URL**: View paper

Abstract

We present a novel method for synthesizing both temporally and geometrically consistent street-view panoramic video from a single satellite image and camera trajectory. Existing cross-view synthesis approaches focus on images, while video synthesis in such a case has not yet received enough attention. For geometrical and temporal consistency, our approach explicitly creates a 3D point cloud representation of the scene and maintains dense 3D-2D correspondences across frames that reflect the geome...

Relationship Analysis

Both papers belong to the Panoramic Video Synthesis category, focusing on generating temporally and geometrically consistent street-view panoramic videos from satellite imagery. They overlap in addressing multiview consistency through explicit 3D geometric reasoning—SatDreamer360 uses a triplane representation with ray-based attention and epipolar constraints, while Sat2Vid employs a 3D point cloud with dense 3D-2D correspondences and cascaded 3D generators. The key difference is that SatDreamer360 operates end-to-end with a diffusion-based framework and panoramic epipolar attention, whereas Sat2Vid uses a GAN-based cascaded architecture (SparseConvNet and RandLA-Net) with point cloud projection and interpolation for video synthesis.

2. SatDreamer360: Geometry Consistent Street-View Video Generation from Satellite Imagery

Authors: Zhu Bei-Yi, Xianghui Ze, Song, Zhenbo, Beiyi Zhu, et al. (11 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

N/A

⚠ Similarity Notice

The candidate paper 'SatDreamer360: Geometry Consistent Street-View Video Generation from Satellite Imagery' appears to be the same work as the original paper 'SatDreamer360: Multiview-Consistent Generation of Ground-Level Scenes from Satellite Imagery', sharing an identical method name and core technical approach. Both papers address panoramic video synthesis from satellite imagery using the same framework, suggesting they are likely the same publication or closely related variants that should be manually verified for potential duplication.

Contributions Analysis

Overall novelty summary. The paper proposes SatDreamer360, a framework for generating geometrically consistent multi-view ground-level panoramas from satellite imagery along predefined trajectories. It resides in the 'Panoramic Video Synthesis' leaf, which contains only three papers total, indicating a relatively sparse research direction within the broader cross-view synthesis field. This leaf focuses specifically on temporally and geometrically consistent street-view panoramic videos, distinguishing it from the more populated single-view synthesis branches that generate individual images without explicit multiview constraints.

The taxonomy reveals that SatDreamer360 sits within 'Multiview-Consistent Synthesis', a subtopic under 'Cross-View Image Synthesis and Generation'. Neighboring leaves include 'Single-View Ground Image Synthesis' (with sub-branches for geometry-guided, learning-based, and controllable methods) and 'BEV-Conditioned Street-View Synthesis'. The scope note for the parent branch emphasizes explicit geometric and temporal consistency, while the exclude note clarifies that single-view methods without consistency constraints belong elsewhere. This positioning suggests the work addresses a more constrained problem—panoramic video coherence—compared to the broader single-image synthesis literature.

Among 15 candidates examined, the analysis identifies mixed novelty signals across contributions. The core SatDreamer360 framework (Contribution 1) examined 3 candidates and found all 3 potentially refutable, suggesting substantial prior work on multiview-consistent generation exists within the limited search scope. The ray-guided triplane representation (Contribution 2) examined 6 candidates with none clearly refutable, indicating this technical approach may be more distinctive. The epipolar-constrained attention module (Contribution 3) examined 6 candidates, with 2 appearing refutable, suggesting partial overlap with existing attention mechanisms for panoramic consistency.

Based on the limited top-15 semantic search, the work appears to occupy a moderately explored niche. The panoramic video synthesis direction itself is sparse (only 3 papers in the leaf), but the underlying techniques—triplane representations, attention mechanisms, and multiview consistency—connect to broader literature. The analysis does not cover exhaustive prior work in neural rendering, diffusion models, or general video synthesis, which may contain additional relevant methods. The novelty assessment reflects what is visible within the examined candidate set, not a comprehensive field survey.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: SatDreamer360 framework for multiview-consistent ground-level scene generation

Description: The authors introduce a unified framework that synthesizes continuous and coherent ground-view sequences from a single satellite image and a target trajectory. The framework addresses the challenge of maintaining both geometric consistency with the satellite image and multiview coherence across generated frames.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Sat2Vid: Street-view Panoramic Video Synthesis from a Single Satellite Image

URL: View paper

Prior Art Analysis

Sat2Vid[11] demonstrates prior work that addresses the same core challenge: generating temporally and geometrically consistent ground-level sequences from a single satellite image and trajectory. Both papers explicitly tackle multiview consistency across generated frames using geometric constraints. Sat2Vid[11] uses 3D point cloud representations with explicit 3D-2D correspondences to maintain consistency, while the original paper uses triplane representations with epipolar constraints. The fundamental contribution of generating multiview-consistent sequences from satellite imagery with pose trajectories was already established by Sat2Vid[11], which predates the original work.

Evidence

Evidence 1 - **Rationale**: Both papers claim to be the first to generate temporally and geometrically consistent ground-level sequences from a single satellite image and trajectory. Sat2Vid[11] explicitly states it addresses video synthesis with temporal and geometric consistency, which directly challenges the novelty claim of the unified framework. - **Original**: we propose satdreamer360, a unified framework that generates continuous and coherent ground-view sequences from a single satellite image and a target trajectory, as shown in figure 1. the key idea is to embed explicit cross-view geometric reasoning between satellite and ground views, as well as acro... - **Candidate**: we present a novel method for synthesizing both temporally and geometrically consistent street-view panoramic video from a single satellite image and camera trajectory. existing cross-view synthesis approaches focus on images, while video synthesis in such a case has not yet received enough attentio...

Evidence 2 - **Rationale**: Both papers identify the same gap in prior work: existing methods focus on single images and lack multiview consistency. Sat2Vid[11] explicitly addresses this limitation by proposing a method for consistent sequences, which refutes the claim that this is a novel contribution. - **Original**: generating multiview-consistent 360◦ ground-level scenes from satellite imagery is a challenging task with broad applications in simulation, autonomous navigation, and digital twin cities. existing approaches primarily focus on synthesizing individual ground-view panoramas, often relying on auxiliar... - **Candidate**: while single street-view image generation from satellite images has recently been investigated [27, 20], these methods are not suitable to create continuous view-point changes around a given location since they built upon random generators and lack constraints on the correspondence between frame pix...

Evidence 3 - **Rationale**: Sat2Vid[11] explicitly claims to be the first work for satellite-to-ground video synthesis from a single satellite image with a trajectory, directly refuting the novelty of the unified framework contribution in the original paper. - **Original**: in this paper, we present satdreamer360, a unified framework that generates continuous and coherent ground-view sequences from a single satellite image and a target trajectory, as shown in figure 1. - **Candidate**: our major contributions can be summarized as follows. (1) we present the first work for satellite-to-ground video synthesis from a single satellite image with a trajectory.

2. Satellite to GroundScape-Large-scale Consistent Ground View Generation from Satellite Views

URL: View paper

Prior Art Analysis

Satellite to GroundScape[1] demonstrates prior work that addresses the same core challenge: generating multiple consistent ground-view images from satellite imagery. Both papers tackle the fundamental problem of maintaining geometric consistency with satellite images and multiview coherence across generated frames. Satellite to GroundScape[1] introduces satellite-guided denoising and satellite-temporal denoising processes to ensure consistency across multiple ground views, which directly overlaps with the original paper's claimed contribution of a unified framework for continuous and coherent ground-view sequences.

Evidence

Evidence 1 - **Rationale**: Both papers identify the same core challenge of generating consistent ground views from satellite imagery and propose frameworks to address multiview consistency, demonstrating that similar prior work exists. - **Original**: we propose satdreamer360, a framework that generates geometrically consistent multi-view ground-level panoramas from a single satellite image, given a predefined pose trajectory. to address the large viewpoint discrepancy between ground and satellite images, we adopt a triplane representation to enc... - **Candidate**: generating consistent ground-view images from satellite imagery is challenging, primarily due to the large discrepancies in viewing angles and resolution between satellite and ground-level domains. previous efforts mainly concentrated on single-view generation, often resulting in inconsistencies acr...

Evidence 2 - **Rationale**: Both papers propose unified frameworks using latent diffusion models with explicit mechanisms to maintain consistency across multiple generated ground views from satellite imagery, showing substantial overlap in the core technical approach. - **Original**: in this paper, we present satdreamer360, a unified framework that generates continuous and coherent ground-view sequences from a single satellite image and a target trajectory, as shown in figure 1. the key idea is to embed explicit cross-view geometric reasoning between satellite and ground views, ... - **Candidate**: our method, based on a fixed latent diffusion model, introduces two conditioning modules: satellite-guided denoising, which extracts high-level scene layout to guide the denoising process, and satellite-temporal denoising, which captures camera motion to maintain consistency across multiple generate...

Evidence 3 - **Rationale**: Both papers explicitly claim to introduce unified frameworks for generating continuous, consistent ground-view sequences from satellite imagery using similar technical approaches, demonstrating that the novelty claim is challenged by prior work. - **Original**: a unified framework, satdreamer360, for generating continuous and geometrically consistent ground-view sequences from a single satellite image and a target trajectory. - **Candidate**: in this work, we propose a novel approach for satellite-to-ground view synthesis that ensures consistency across the generated ground views, as shown in fig. 1. building on the ldm [31], we introduce a satellite-guided denoising process to bridge the significant domain gap between satellite and grou...

3. Seeing through Satellite Images at Street Views

URL: View paper

Prior Art Analysis

Satellite to Street Views[21] (Sat2Density++) demonstrates prior work that addresses the same core problem: generating multiview-consistent ground-level panoramas from satellite imagery. Both papers tackle the challenge of maintaining geometric consistency with satellite images and multiview coherence across generated frames. Sat2Density++ explicitly learns 3D scene geometry using neural radiance fields and generates continuous street-view videos along trajectories, which directly overlaps with the original paper's claimed contribution of a unified framework for multiview-consistent generation.

Evidence

Evidence 1 - **Rationale**: Both papers identify the same problem: generating multiview-consistent ground-level panoramas from satellite imagery. Sat2Density++ explicitly addresses rendering street-view videos from satellite images with specified trajectories, which is the same goal as the original paper's framework. - **Original**: generating multiview-consistent 360◦ ground-level scenes from satellite imagery is a challenging task with broad applications in simulation, autonomous navigation, and digital twin cities. existing approaches primarily focus on synthesizing individual ground-view panoramas, often relying on auxiliar... - **Candidate**: this paper studies the task of satstreet-view synthesis, which aims to render photorealistic street-view panorama images and videos given any satellite image and specified camera positions or trajectories. we formulate to learn neural radiance field from paired images captured from satellite and str...

Evidence 2 - **Rationale**: Both papers propose frameworks for generating ground-level panoramas from satellite images using 3D representations. The candidate explicitly learns neural fields for this purpose, addressing the same technical challenge of multiview-consistent generation. - **Original**: in this paper, we propose satdreamer360, a framework that generates geometrically consistent multi-view ground-level panoramas from a single satellite image, given a predefined pose trajectory. - **Candidate**: we approach the goal of satstreet-view synthesis by learning the radiance field representation [2] with a feedforward neural network from the satstreet-view image pairs. Along this direction, two main challenges remained to be solved: 1) how to effectively learn the neural fields as the primary 3d r...

Evidence 3 - **Rationale**: Both papers emphasize multiview consistency as a core contribution. The candidate explicitly claims to be the first to synthesize multi-view consistent street-view videos from satellite images, which directly challenges the novelty of the original paper's multiview consistency mechanism. - **Original**: to maintain multi-frame consistency, we introduce a panoramic epipolar-constrained attention module that aligns features across frames based on known relative poses. - **Candidate**: our method is the first that can synthesize multi-view consistent street-view videos from input satellite images without relying on 3d annotations for training. Through extensive qualitative and quantitative experiments, we demonstrate that sat2density++ outperforms existing single-image generation ...

Contribution 2: Ray-guided cross-view feature conditioning with triplane representation

Description: The authors design a mechanism that adopts a triplane representation to encode scene geometry from the satellite image and introduces a ray-based pixel attention module. This module retrieves view-dependent features from the triplane and integrates them into conditional diffusion, enabling geometry-aware and controllable generation without requiring height maps or handcrafted projections.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Epipolar-free 3d gaussian splatting for generalizable novel view synthesis

URL: View paper

Brief Assessment

Epipolar-free Gaussian Splatting[44] focuses on 3D Gaussian splatting for novel view synthesis, not on cross-view satellite-to-ground generation with triplane scene encoding and ray-based attention mechanisms.

2. GCRayDiffusion: Pose-Free Surface Reconstruction via Geometric Consistent Ray Diffusion

URL: View paper

Brief Assessment

GCRayDiffusion[42] focuses on pose-free 3D surface reconstruction from multi-view images using ray-based diffusion for camera pose estimation, not on satellite-to-ground view synthesis with cross-view feature conditioning for controllable generation.

3. SatDreamer360: Geometry Consistent Street-View Video Generation from Satellite Imagery

URL: View paper

Brief Assessment

No candidate paper text was provided for comparison. The status 'n/a' in the candidate context prevents any assessment of whether prior work exists that could refute this novelty claim.

4. Controllable generation with disentangled representative learning of multiple perspectives in autonomous driving

URL: View paper

Brief Assessment

Disentangled Driving Generation[41] uses triplane representation for autonomous driving scene generation with semantic factor control (weather, speed), not for satellite-to-ground cross-view synthesis. The ray-based sampling mechanism serves different purposes in each work.

5. City-on-web: Real-time neural rendering of large-scale scenes on the web

URL: View paper

Brief Assessment

City on Web[43] focuses on real-time neural rendering of large-scale scenes using triplane representations for 3D scene encoding, but does not address cross-view synthesis between satellite and ground views, which is the core problem in the original paper. The technical contexts are fundamentally different: City on Web[43] deals with multi-view consistency in neural radiance fields for scene reconstruction, while the original work addresses satellite-to-ground view synthesis with geometry-aware diffusion models.

6. Three-dimensional reconstruction and editing from single images with generative models

URL: View paper

Brief Assessment

Single Image Reconstruction[45] focuses on 3D reconstruction from single images using triplane representations for object-level geometry encoding, not on cross-view satellite-to-ground synthesis with ray-based pixel attention for scene-level generation.

Contribution 3: Epipolar-constrained attention module for panoramic images

Description: The authors extend epipolar constraints from pinhole cameras to panoramic images with equirectangular projections. This module aligns features across frames by leveraging known relative camera poses, maintaining multiview consistency while reducing computational complexity compared to full cross-attention.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Era3D: High-Resolution Multiview Diffusion using Efficient Row-wise Attention

URL: View paper

Brief Assessment

Era3D[38] uses row-wise attention for epipolar priors in multiview diffusion for 3D object reconstruction from single-view images, not for panoramic image generation with equirectangular projections as in the original paper's satellite-to-ground synthesis task.

2. CamPVG: Camera-Controlled Panoramic Video Generation with Epipolar-Aware Diffusion

URL: View paper

Prior Art Analysis

CamPVG[35] demonstrates prior work on epipolar-constrained attention for panoramic images with equirectangular projections. Both papers extend epipolar constraints from pinhole cameras to panoramic settings and use these constraints to maintain multiview consistency. CamPVG[35] introduces a 'spherical epipolar module that enforces geometric constraints through adaptive attention masking along epipolar lines' for panoramic video generation, which directly addresses the same technical challenge of applying epipolar geometry to equirectangular projections that the original paper claims as novel.

Evidence

Evidence 1 - **Rationale**: Both papers extend epipolar constraints to panoramic images with equirectangular projections. CamPVG[35] explicitly introduces a spherical epipolar module for enforcing geometric constraints in panoramic settings, demonstrating that this approach existed prior to the original paper's submission. - **Original**: we draw inspiration from the use of epipolar constraints in pinhole cameras (tobin et al. (2019); he et al. (2020); huang et al. (2022; 2024)) and extend the idea to panoramic images with equirectangular projections. we design an epipolar-constrained attention module for panoramic images, which alig... - **Candidate**: we introduce a spherical epipolar module that enforces geometric constraints through adaptive attention masking along epipolar lines. this module enables fine-grained cross-view feature aggregation, substantially enhancing the quality and consistency of generated panoramic videos.

Evidence 2 - **Rationale**: CamPVG[35] addresses cross-view feature aggregation based on spherical projection for panoramic images, which is the same technical domain as the original paper's epipolar-constrained attention for equirectangular projections. - **Original**: to enhance mutiview consistency, we draw inspiration from the use of epipolar constraints in pinhole cameras (tobin et al. (2019); he et al. (2020); huang et al. (2022; 2024)) and extend the idea to panoramic images with equirectangular projections. - **Candidate**: this limitation is primarily due to the inherent complexities in panoramic pose representation and spherical projection. to address this issue, we propose campvg, the first diffusion-based framework for panoramic video generation guided by precise camera poses. we achieve camera position encoding fo...

Evidence 3 - **Rationale**: Both papers use epipolar constraints in panoramic/spherical settings to align features across frames. CamPVG[35]'s spherical epipolar module with adaptive attention masking along epipolar lines directly parallels the original paper's epipolar-constrained attention module for panoramic images. - **Original**: an interframe attention module that uses panoramic epipolar constraints via equirectangular projections to align features across frames, enhancing multiview consistency. - **Candidate**: we introduce a spherical epipolar module that enforces geometric constraints through adaptive attention masking along epipolar lines. this module enables fine-grained cross-view feature aggregation

3. View synthesis for 360 panoramic spherical images using Multiplane Images

URL: View paper

Brief Assessment

Multiplane Panorama Synthesis[40] focuses on view synthesis for 360 panoramic spherical images using Multiplane Images, not on satellite-to-ground generation with epipolar constraints for multiview consistency in equirectangular projections.

4. DreamCube: RGB-D Panorama Generation via Multi-plane Synchronization

URL: View paper

Brief Assessment

DreamCube[37] focuses on multi-plane synchronization for cubemap generation and does not employ epipolar constraints for panoramic image consistency. The candidate addresses seam inconsistencies through synchronized spatial operators rather than epipolar geometry.

5. Diffpano: Scalable and consistent text to panorama generation with spherical epipolar-aware diffusion

URL: View paper

Prior Art Analysis

DiffPano[36] demonstrates prior work on epipolar constraints for panoramic images with equirectangular projections. The candidate paper explicitly derives spherical epipolar constraints for panoramic images and implements a spherical epipolar-aware attention module to maintain multi-view consistency. The candidate states they 'derived the epipolar line for panoramic images in the equirectangular projection (erp)' and 'extend the principle of epipolar attention to panoramic images to implement the spherical epipolar-aware attention module.' This directly addresses the same technical challenge of applying epipolar constraints to panoramic images with equirectangular projections that the original paper claims as novel.

Evidence

Evidence 1 - **Rationale**: Both papers address extending epipolar constraints from pinhole/perspective cameras to panoramic images with equirectangular projections. The candidate explicitly derives the spherical epipolar line formula for ERP images, demonstrating prior work on this exact technical contribution. - **Original**: we draw inspiration from the use of epipolar constraints in pinhole cameras (tobin et al. (2019); he et al. (2020); huang et al. (2022; 2024)) and extend the idea to panoramic images with equirectangular projections. we design an epipolar-constrained attention module for panoramic images, which alig... - **Candidate**: due to the differences in imaging methods between perspective and panoramic views, existing epipolar attention cannot be directly used for panoramic views. to overcome this challenge, we derived the epipolar line for panoramic images in the equirectangular projection (erp), and the specific proof pr...

Evidence 2 - **Rationale**: Both papers describe implementing an attention module that uses epipolar constraints for panoramic images with equirectangular projections to ensure multi-view consistency. The candidate provides the mathematical formulation and implementation details of this approach. - **Original**: an interframe attention module that uses panoramic epipolar constraints via equirectangular projections to align features across frames, enhancing multiview consistency. - **Candidate**: we extend the principle of epipolar attention to panoramic images to implement the spherical epipolar-aware attention module. given a pixel p in the target view, we calculate its corresponding spherical coordinates psphere based on the spherical projection process

Evidence 3 - **Rationale**: Both papers acknowledge prior work on epipolar constraints for pinhole/perspective cameras and describe extending this concept to panoramic images with equirectangular projections. The candidate explicitly addresses the same technical gap and provides a solution. - **Original**: to enhance mutiview consistency, we draw inspiration from the use of epipolar constraints in pinhole cameras (tobin et al. (2019); he et al. (2020); huang et al. (2022; 2024)) and extend the idea to panoramic images with equirectangular projections. - **Candidate**: epipolar attention was proposed in [20, 51] to ensure consistency between generated multi-view perspective images. however, due to the differences in imaging methods between perspective and panoramic views, existing epipolar attention cannot be directly used for panoramic views. to overcome this ch...

6. Super-resolution reconstruction for stereoscopic omnidirectional display systems via dynamic convolutions and cross-view transformer

URL: View paper

Brief Assessment

Stereoscopic Omnidirectional[39] focuses on super-resolution reconstruction for stereoscopic displays using epipolar lines for binocular parallax attention, not on multiview-consistent generation across camera trajectories with equirectangular projections.

Appendix: Text Similarity Detection

Textual similarity detection checked 15 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. SatDreamer360: Geometry Consistent Street-View Video Generation from Satellite Imagery

Detected in: Core Task (sibling), Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] SatDreamer360: Multiview-Consistent Generation of Ground-Level Scenes from Satellite Imagery View paper
- [1] Satellite to GroundScape-Large-scale Consistent Ground View Generation from Satellite Views View paper
- [2] Geospecific view generation geometry-context aware high-resolution ground view inference from satellite views View paper
- [3] Street-View Image Generation From a Bird's-Eye View Layout View paper
- [4] BEVControl: Accurately Controlling Street-view Elements with Multi-perspective Consistency via BEV Sketch Layout View paper
- [5] Controllable Satellite-to-Street-View Synthesis with Precise Pose Alignment and Zero-Shot Environmental Control View paper
- [6] Level-Aware Consistent Multilevel Map Translation From Satellite Imagery View paper
- [7] Geometry-Guided Street-View Panorama Synthesis From Satellite Imagery View paper
- [8] View from above: Orthogonal-view aware cross-view localization View paper
- [9] Sat2scene: 3d urban scene generation from satellite images with diffusion View paper
- [10] Cross-view SLAM solver: Global pose estimation of monocular ground-level video frames for 3D reconstruction using a reference 3D model from satellite images View paper

- [11] Sat2Vid: Street-view Panoramic Video Synthesis from a Single Satellite Image View paper
- [12] Satellite image synthesis from street view with fine-grained spatial textual guidance: A novel framework View paper
- [13] Cross-View Geo-Localization via 3D Gaussian Splatting-Based Novel View Synthesis View paper
- [14] View Synthesis with Scene Recognition for Cross-View Image Localization View paper
- [15] Sat2Density: Faithful Density Learning from Satellite-Ground Image Pairs View paper
- [16] Groundâ□□Satellite coupling for cross-view geolocation combined with multiscale fusion of spatial features View paper
- [17] Geometry-aware satellite-to-ground image synthesis for urban areas View paper
- [18] RETRACTED CHAPTER: CrossViewDiff: A Cross-View Diffusion Model for Satellite-to-Ground Image Synthesis View paper
- [19] Accurate 3-DoF Camera Geo-Localization via Ground-to-Satellite Image Matching View paper
- [20] MagicCity: Geometry-Aware 3D City Generation from Satellite Imagery with Multi-View Consistency View paper
- [21] Seeing through Satellite Images at Street Views View paper
- [22] Mutual Relative Position Learning Transformer for Cross-View Geo-Localization View paper
- [23] Toward Seamless Multiview Scene Analysis From Satellite to Street Level View paper
- [24] From Satellite to Street: A Hybrid Framework Integrating Stable Diffusion and PanoGAN for Consistent Cross-View Synthesis View paper
- [25] Generative Towns View paper
- [26] Cross-view Localization and Synthesis--Datasets, Challenges and Opportunities View paper
- [27] SatDreamer360: Geometry Consistent Street-View Video Generation from Satellite Imagery View paper
- [28] Geometry-Aware Enhancement and Data Augmentation for Street-to-Satellite Geo-localization View paper
- [29] A Study of Remote Sensing Based Natural and Built Environment Monitoring: From Fully Supervised to Weakly Supervised Learning View paper
- [30] Advances in Neural Radiance Fields for Large-Scale 3D Scene Reconstruction: A Comprehensive Review View paper
- [31] Consistent ground-plane mapping: A case study utilizing low-cost sensor measurements and a satellite image View paper
- [32] Packing Urban Scenes Into Neural Radiance Field: 3D Scene Rendering, Manipulation and Generation View paper
- [33] What Is It Like Down There? Generating Dense Ground-Level Views and Image Features From Overhead Imagery Using Conditional Generative Adversarial Networks View paper
- [34] Unleashing Unlabeled Data: A Paradigm for Cross-View Geo-Localization (Supplementary Material) View paper
- [35] CamPVG: Camera-Controlled Panoramic Video Generation with Epipolar-Aware Diffusion View paper
- [36] Diffpano: Scalable and consistent text to panorama generation with spherical epipolar-aware diffusion View paper
- [37] DreamCube: RGB-D Panorama Generation via Multi-plane Synchronization View paper
- [38] Era3D: High-Resolution Multiview Diffusion using Efficient Row-wise Attention View paper
- [39] Super-resolution reconstruction for stereoscopic omnidirectional display systems via dynamic convolutions and cross-view transformer View paper
- [40] View synthesis for 360 panoramic spherical images using Multiplane Images View paper
- [41] Controllable generation with disentangled representative learning of multiple perspectives in autonomous driving View paper
- [42] GCRayDiffusion: Pose-Free Surface Reconstruction via Geometric Consistent Ray Diffusion View paper
- [43] City-on-web: Real-time neural rendering of large-scale scenes on the web View paper
- [44] Epipolar-free 3d gaussian splatting for generalizable novel view synthesis View paper
- [45] Three-dimensional reconstruction and editing from single images with generative models View paper