Wide-Baseline Novel View Synthesis for Urban Navigation

State of the Art and Perspectives

Maxim Quénel Pierre Louis Brun Mame Alé Seye Denis Bereziuc

December 10, 2025

Executive Summary

Traditional immersive "Street View" navigation relies on simplified geometric transitions, often causing noticeable visual artifacts (distortions, tunnel effects). This document examines the State of the Art (SOTA) of technologies required to generate photorealistic transitions between two distant panoramas (approx. 10 meters). We analyze the decomposition of the problem into depth estimation, dense matching, warping-based view synthesis, and generative inpainting, contrasting these methods with volumetric approaches (NeRF/Gaussian Splatting) and classic video interpolation.

1 Introduction and Context

The current approach to navigation in Google Street View relies on proxy geometry. Although Google uses LiDAR data (Velodyne) to project partial ground truth onto panoramas, the transition between two nodes (panos) is performed via projective morphing on simplified planes ("Pancakes" and "Smart Transitions").

- **Critical Limitation:** This method fails to model the complex parallax of near and far objects, creating unnatural stretching at the edges and a "tunnel" effect.
- **Research Objective:** Replace 2D/Proxy interpolation with geometrically consistent view synthesis capable of handling large disocclusions (hidden areas revealed by movement).

2 Geometric Scene Understanding

The first critical step is to recover the 3D structure from a single image (Monocular) or a pair (Stereo).

2.1 Monocular Depth Estimation (MDE)

The recent evolution of Foundation Models has transformed this field.

- **Depth Anything V2 [1]:** Currently the most robust model for relative depth estimation (successor to MiDaS). It excels at fine edge segmentation and "zero-shot" generalization. However, its output is a relative inverse disparity map (non-metric).
- **Metric3D v2 [2]:** (If no GPS coordinates are available). Unlike Depth Anything, this model estimates absolute metric depth.

2.2 Stereo Reconstruction and 3D Correspondence

To overcome monocular limitations, image-pair-based approaches are gaining traction.

MASt3R (Naver) [3] Successor to DUST3R, this model revolutionizes the classic "Feature Extraction + Matching + SfM" pipeline. It achieves direct "Grounding" of image matching in 3D.

- **Advantage:** It jointly provides depth and relative pose without requiring costly global optimization.
- **Constraint:** Even with MASt3R, a post-processing pipeline (Warping + Inpainting + Blending) remains necessary to generate intermediate pixels, as MASt3R produces point clouds, not interpolated dense images.

3 Matching and Pose Estimation (Matching & Flow)

To deform image A towards image B, it is imperative to understand the motion field (Optical Flow) or geometric correspondences.

3.1 Sparse vs. Dense Matching

- **The Classic Approach (Sparse):** SIFT or SuperPoint + LightGlue [4] or XFeat [5], are robust but only provide sparse points, which are insufficient for dense image warping.
- **The Dense Approach (SOTA):**
  - **LoFTR [6]:** Uses Transformers to match textureless areas, outperforming local descriptors.
  - **RoMa [7]:** Represents the current state of the art for Dense Feature Matching. Unlike LoFTR, RoMa estimates dense warps at the pixel level with reliable certainty estimates. It is the ideal candidate for guiding complex non-linear warping.

3.2 Optical Flow

RAFT [8] & SEA-RAFT [9]: Although SEA-RAFT is extremely high-performing for video optical flow, it presents a major limitation for the Street View use case: the distance gap (10m baseline). Optical flow tends to fail when pixel displacements are too large between two frames.

4 Synthesis and Rendering (Warping & Rendering)

Once depth and pose are obtained, the source image must be reprojected into the new virtual viewpoint.

- **Z-buffer Splatting:** A naive method prone to aliasing artifacts (notably used in work on the 3D Ken Burns effect [10]).
- **Softmax Splatting [11]:** This is the reference method. It replaces the rigid Z-buffer with a differentiable operation that handles:
  - Occlusions: Via depth-based weighting (near objects overwrite far ones).
  - Edge Translucency: Avoids tearing and aliasing on object contours.

Comparison with Radiance Fields (NeRF/GS): Although Zip-NeRF [12] (anti-aliasing) and 3D Gaussian Splatting [13] are SOTA for scene representation, they are ill-suited for our configuration.

- **Reason:** They require high view density (dozens of images around an object). With only two panoramic images spaced 10 meters apart, these models cannot converge without creating massive artifacts (floaters).
- **S-NeRF [14]:** Attempts to adapt NeRF to Street View but remains computationally very heavy compared to an image warping-based approach.

5 Handling Disocclusions (Inpainting)

Camera movement inevitably reveals empty areas behind foreground objects. This is where Generative AI comes in.

- **"Copy-Paste" Approaches (Navier-Stokes, Telea):** Obsolete, they create blur.
- **LaMa (Large Mask Inpainting) [15]:** Uses Fast Fourier Convolutions. Very efficient for rapidly filling repetitive textures (sky, road, walls).
- **MAT (Mask-Aware Transformer) [16]:** SOTA for large holes. It understands global structure better than LaMa but is heavier.
- **Diffusion Models (SDXL [17], PowerPaint [18]):** Offer the best "hallucinatory" quality (can invent a door where there is nothing), but suffer from temporal instability (flickering) and slow inference. For a real-time or interactive project, a small CNN or GAN like LaMa/MAT is preferable.

6 Video Frame Interpolation (VFI)

Why not simply use video interpolators?

FILM [19], RIFE [20], VFIformer [21]: These models excel at smoothing video (e.g., 30fps $\rightarrow$ 60fps).

- **The problem:** They are designed for minimal and linear movements. Over a 10-meter jump (Street View), the geometric relationship is broken. These models create ghostly "morphing" effects instead of true parallax. They serve only as a baseline for comparison here.

7 Analysis of SOTA "System" Projects

These projects represent integrated architectures that attempt to merge geometry and generation.

7.1 Infinite Nature (Google Research)

- **Key Concept:** "Render-Refine-Repeat" (Perpetual View Generation) [22]. Unlike interpolation between A and B, Infinite Nature learns to generate views perpetually by advancing into a single image.
- **Mechanism:** It uses proxy geometry to warp the current image to a new view, then uses a generative network (SPADE generator) to "repair" artifacts and hallucinate missing details (grass texture, road extension).

- **Relevance to Project:** This is the ultimate solution to the tunnel effect problem. Rather than simply interpolating, it learns to generate plausible content during forward movement (zooming in).
- **Limitation:** Tendency towards "dream-like drift." The geometry can become incoherent over long distances, which is problematic for precise urban navigation.

7.2 Google Immersive View

- **Key Concept:** Large-Scale NeRF & Multi-Modal Fusion [23]. This is Google's current technological showcase.
- **Mechanism:** It involves massive reconstruction using radiance fields (NeRF) combined with classic photogrammetry, trained on thousands of aerial and street images.
- **Critical Difference:** Our project aims for "on-the-fly" (client-side or lightweight) interpolation between two photos. Immersive View is a heavy "offline" approach: the scene is pre-computed and stored in the cloud.
- **Perspective:** Although not directly applicable for a lightweight transition between two panos, the future of this tech lies in 3D Gaussian Splatting, which would allow real-time rendering of these heavy scenes on mobile.

7.3 Learning to Render Novel Views from Wide-Baseline Stereo Pairs (CVPR 2023)

- **Key Concept:** Cross-Attention & Epipolar Constraints [24]. This paper is the most technically relevant for our use case (10m transition).
- **The Solved Problem:** Classic methods (Flow/Warping) fail when the gap (baseline) is large because correspondence search is too difficult.
- **The Solution:** Instead of seeking dense optical flow, they use Transformers with Cross-Attention mechanisms along epipolar lines. The network "constructs" the intermediate image by fetching relevant information from both source images via attention, rather than by rigid geometric warping.
- **Advantage:** Extreme robustness to large perspective changes that classic warping (Softmax splatting) cannot handle without a perfect depth map.

7.4 S-NeRF: Neural Radiance Fields for Street Views

- **Key Concept:** NeRF adapted for unbounded urban scenes [14].
- **Innovation:** Classic NeRFs fail on street scenes (fast near objects + infinite sky). S-NeRF improves vehicle geometry and separates the background (sky) from the foreground.
- **Verdict for Project:** As mentioned in the introduction, S-NeRF remains an optimization method (training-based). To navigate between two panoramas A and B, you would have to "train" a mini-S-NeRF on these two images, which would take several minutes/hours. This is ill-suited for fluid user navigation, unless pre-computed for the entire city (exorbitant cost).

7.5 Disney Frame Interpolation Transformer & Uncertainty Guidance

- **Key Concept:** Explicit uncertainty management for Inpainting [25]. This paper introduces a crucial notion often ignored: uncertainty.

- **Mechanism:** The model does not just predict the intermediate image; it predicts an uncertainty map (where the model "knows" it is failing, often on disocclusions or fast movements).
- **Strategic Application:** In our pipeline, instead of letting the interpolator create blur (ghosting) in difficult areas, you can use this uncertainty map as a mask for our Inpainting model (LaMa).
- **Logic:** "If the interpolator is uncertain here, erase this area and let the Generative AI (LaMa/MAT) completely redraw it."

References

[1] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything v2: A more capable foundation model for monocular depth estimation," in *NeurIPS*, 2024. https://github.com/DepthAnything/Depth-Anything-V2.

[2] W. Yin, C. Zhang, H. Chen, Z. Cai, G. Yu, W. Kaigang, C. Chen, and C. Shen, "Metric3d: Towards zero-shot metric 3d prediction from a single image," in *ECCV*, 2024. Metric3D v2. https://github.com/YvanYin/Metric3D.

[3] E. Vincent *et al.*, "Mast3r: Grounding image matching in 3d with mast3r," *arXiv preprint arXiv:2406.09756*, 2024. Successor to DUST3R. https://github.com/naver/mast3r.

[4] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *ICCV*, 2023. https://github.com/cvg/LightGlue.

[5] G. Potlapalli *et al.*, "Xfeat: Accelerated features for lightweight image matching," in *CVPR*, 2024. https://github.com/verlab/accelerated_features.

[6] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *CVPR*, 2021. https://github.com/zju3dv/LoFTR.

[7] J. Edstedt, Q. Wursthorn, *et al.*, "Roma: Robust dense feature matching," in *CVPR*, 2024. https://github.com/Parskatt/RoMa.

[8] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *ECCV*, 2020. https://github.com/princeton-vl/RAFT.

[9] L. Wan *et al.*, "Sea-raft: Simple, efficient, accurate raft for optical flow," in *ECCV*, 2024. Oral, Best Paper Candidate. https://github.com/princeton-vl/SEA-RAFT.

[10] S. Niklaus, L. Mai, J. Yang, and F. Liu, "3d ken burns effect from a single image," *ACM Transactions on Graphics (TOG)*, 2019. https://github.com/sniklaus/3d-ken-burns.

[11] S. Niklaus and F. Liu, "Softmax splatting for video frame interpolation," in *CVPR*, 2020. https://github.com/sniklaus/softmax-splatting.

[12] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Zip-nerf: Anti-aliased grid-based neural radiance fields," in *ICCV*, 2023. Implementation unofficial: https://github.com/SuLvXiangXin/zipnerf-pytorch.

[13] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," in *SIGGRAPH*, 2023. https://github.com/graphdeco-inria/gaussian-splatting.

[14] Z. Xie, J. Zhang, W. Li, F. Zhang, and L. Zhang, "S-nerf: Neural radiance fields for street views," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025. First appeared in ICLR 2023. https://github.com/fudan-zvg/S-NeRF.

[15] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, *et al.*, "Resolution-robust large mask inpainting with fourier convolutions," in *WACV*, 2022. LaMa. https://github.com/advimman/lama.

[16] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *CVPR*, 2022. https://github.com/fenglinglwb/MAT.

[17] D. Podell, Z. English, K. Lacey, A. Blattmann, *et al.*, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023. https://github.com/Stability-AI/generative-models.

[18] J. Zhuang *et al.*, "Powerpaint: A task is worth one word: Learning to control image inpainting," *arXiv preprint arXiv:2312.03594*, 2023. https://github.com/open-mmlab/PowerPaint.

[19] F. Reda, P. Kontschieder, *et al.*, "Film: Frame interpolation for large motion," in *ECCV*, 2022. https://github.com/google-research/frame-interpolation.

[20] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, "Real-time intermediate flow estimation for video frame interpolation," in *ECCV*, 2022. https://github.com/hzwer/Practical-RIFE.

[21] L. Lu *et al.*, "Video frame interpolation with transformer," in *CVPR*, 2022. https://github.com/dvlab-research/VFIformer.

[22] A. Liu, R. Tucker, V. Jampani, A. Makadia, N. Snavely, and A. Kanazawa, "Infinite nature: Perpetual view generation of natural scenes from a single image," in *ICCV*, 2021. https://github.com/google-research/google-research/tree/master/infinite_nature.

[23] Google, "Google maps updates: Immersive view for routes and more ai features," 2023. https://blog.google/intl/fr-fr/nouveautes-produits/explorez-obtenez-des-reponses/google-maps-immersive-view-itineraires-nouvelles-fonctionnalites-ia/.

[24] Y. Du, C. Smith, A. Tewari, and V. Sitzmann, "Learning to render novel views from wide-baseline stereo pairs," in *CVPR*, 2023. https://github.com/yilundu/cross_attention_renderer.

[25] Disney Research and ETH Zurich, "Frame interpolation transformer and uncertainty guidance," 2023. Technical Report. https://assets.studios.disneyresearch.com/app/uploads/2023/05/Frame-Interpolation-Transformer-and-Uncertainty-Guidance-1.pdf.