Synthèse de Vues Inédites à Large Base (Wide-Baseline Novel View Synthesis) pour la Navigation Urbaine

État de l'Art et Perspectives

Maxim Quénel Pierre Louis Brun Mame Alé Seye Denis Bereziuc

10 décembre 2025

Résumé Exécutif

La navigation immersive de type "Street View" repose traditionnellement sur des transitions géométriques simplifiées, provoquant souvent des artefacts visuels notables (distorsions, effet tunnel). Ce document examine l'état de l'art (SOTA) des technologies nécessaires pour générer des transitions photoréalistes entre deux panoramas distants (approx. 10 mètres). Nous analysons la décomposition du problème en estimation de profondeur, correspondance dense, synthèse de vue par *warping*, et *inpainting* génératif, en contrastant ces méthodes avec les approches volumétriques (NeRF/Gaussian Splatting) et l'interpolation vidéo classique.

1 Introduction et Contexte

L'approche actuelle de la navigation dans Google Street View repose sur une géométrie proxy. Bien que Google utilise des données LiDAR (Velodyne) pour projeter une vérité terrain partielle sur les panoramas, la transition entre deux nœuds (panos) s'effectue via un morphing projectif sur des plans simplifiés ("Pancakes" et "Smart Transitions").

- **Limitation critique :** Cette méthode échoue à modéliser la parallaxe complexe des objets proches et lointains, créant des étirements non naturels sur les bords et un effet de "tunnel".
- **Objectif de recherche :** Remplacer l'interpolation 2D/Proxy par une synthèse de vue géométriquement cohérente, capable de gérer de grandes disocclusions (zones cachées révélées par le mouvement).

2 Compréhension Géométrique de la Scène

La première étape critique est de récupérer la structure 3D à partir d'une image unique (Monocular) ou d'une paire (Stereo).

2.1 Estimation de Profondeur Monoculaire (MDE)

L'évolution récente des modèles de fondation (*Foundation Models*) a transformé ce domaine.

- **Depth Anything V2 [1] :** Actuellement le modèle le plus robuste pour l'estimation de profondeur relative (successeur de MiDaS). Il excelle dans la segmentation des bords fins et la généralisation "zero-shot". Cependant, sa sortie est une carte de disparité inverse relative (non métrique).

— **Metric3D v2 [2] :** (Si pas de coordonnée GPS). Contrairement à Depth Anything, ce modèle estime la profondeur métrique absolue.

2.2 Reconstruction Stéréo et Correspondance 3D

Pour dépasser les limites monoculaires, les approches basées sur la paire d'images gagnent en traction.

MASt3R (Naver) [3] Successeur de DUST3R, ce modèle révolutionne le pipeline classique "Feature Extraction + Matching + SfM". Il réalise un "Grounding" direct de l'appariement d'images en 3D.

- **Avantage :** Il fournit conjointement la profondeur et la pose relative sans nécessiter d'optimisation globale coûteuse.
- **Contrainte :** Même avec MASt3R, un pipeline de post-traitement (Warping + Inpainting + Blending) reste nécessaire pour générer les pixels intermédiaires, car MASt3R produit des nuages de points, pas des images denses interpolées.

3 Correspondance et Estimation de Pose (Matching & Flow)

Pour déformer l'image A vers l'image B, il est impératif de comprendre le champ de mouvement (*Optical Flow*) ou les correspondances géométriques.

3.1 Sparse vs Dense Matching

- **L'approche classique (Sparse) :** SIFT ou SuperPoint + LightGlue [4] ou XFeat [5], sont robustes mais ne fournissent que des points épars, insuffisants pour une déformation dense de l'image (*warping*).
- **L'approche Dense (SOTA) :**
  - **LoFTR [6] :** Utilise des Transformers pour matcher des zones sans texture, dépassant les descripteurs locaux.
  - **RoMa [7] :** Représente l'état de l'art actuel pour le *Dense Feature Matching*. Contrairement à LoFTR, RoMa estime des warps denses au niveau du pixel avec des estimations de certitude fiables. C'est le candidat idéal pour guider un warping non linéaire complexe.

3.2 Optical Flow

RAFT [8] & SEA-RAFT [9] : Bien que SEA-RAFT soit extrêmement performant pour le flux optique vidéo, il présente une limitation majeure pour le cas d'usage Street View : l'écart de distance (*baseline* de 10m). Le flux optique tend à échouer lorsque les déplacements de pixels sont trop importants entre deux frames.

4 Synthèse et Rendu (Warping & Rendering)

Une fois la profondeur et la pose obtenues, l'image source doit être reprojetée dans le nouveau point de vue virtuel.

- **Z-buffer Splatting :** Méthode naïve sujette aux artefacts crénelés (utilisée notamment dans les travaux sur l'effet Ken Burns 3D [10]).
- **Softmax Splatting [11] :** C'est la méthode de référence. Elle remplace le Z-buffer rigide par une opération différentiable qui gère :

— Les occlusions : Via une pondération basée sur la profondeur (les objets proches écrasent les lointains).
— La translucidité des bords : Évite les déchirements et l'aliasing sur les contours d'objets.

Comparaison avec les champs de radiance (NeRF/GS) : Bien que Zip-NeRF [12] (anti-aliasing) et 3D Gaussian Splatting [13] soient le SOTA pour la représentation de scènes, ils sont inadaptés à notre configuration.

— **Raison :** Ils nécessitent une densité de vues élevée (plusieurs dizaines d'images autour d'un objet). Avec seulement deux images panoramiques espacées de 10 mètres, ces modèles ne peuvent pas converger sans créer d'artefacts massifs (*floaters*).
— **S-NeRF [14] :** Tente d'adapter NeRF au Street View, mais reste très lourd computationnellement par rapport à une approche basée sur le warping d'image.

5 Gestion des Disocclusions (Inpainting)

Le déplacement de la caméra révèle inévitablement des zones vides derrière les objets de premier plan. C'est ici que l'IA générative intervient.

— **Approches "Copy-Paste" (Navier-Stokes, Telea) :** Obsolètes, elles créent du flou.
— **LaMa (Large Mask Inpainting) [15] :** Utilise des Convolutions de Fourier Rapides. Très efficace pour remplir des textures répétitives (ciel, route, murs) rapidement.
— **MAT (Mask-Aware Transformer) [16] :** SOTA pour les grands trous. Il comprend mieux la structure globale que LaMa, mais est plus lourd.
— **Modèles de Diffusion (SDXL [17], PowerPaint [18]) :** Offrent la meilleure qualité "hallucinatoire" (peuvent inventer une porte là où il n'y a rien), mais souffrent d'instabilité temporelle (clignotements) et de lenteur d'inférence. Pour un projet temps-réel ou interactif, un petit CNN ou un GAN comme LaMa/MAT est préférable.

6 Interpolation de Cadres Vidéo (VFI)

Pourquoi ne pas simplement utiliser des interpolateurs vidéo ?

FILM [19], RIFE [20], VFIformer [21] : Ces modèles excellent pour fluidifier une vidéo (ex : 30fps → 60fps).

— **Le problème :** Ils sont conçus pour des mouvements minimes et linéaires. Sur un saut de 10 mètres (Street View), la relation géométrique est brisée. Ces modèles créent des effets de "morphing" fantomatiques (*ghosting*) au lieu d'une véritable parallaxe. Ils ne servent ici que de baseline de comparaison.

7 Analyse des Projets SOTA "Système"

Ces projets représentent des architectures intégrées qui tentent de fusionner la géométrie et la génération.

7.1 Infinite Nature (Google Research)

— **Concept clé :** "Render-Refine-Repeat" (Perpetual View Generation) [22]. Contrairement à l'interpolation entre A et B, Infinite Nature apprend à générer des vues perpétuellement en avançant dans une image unique.

— **Mécanisme :** Il utilise une géométrie proxy pour warper l'image actuelle vers une nouvelle vue, puis utilise un réseau génératif (SPADE generator) pour "réparer" les artefacts et halluciner les détails manquants (la texture de l'herbe, le prolongement de la route).
— **Intérêt pour le projet :** C'est la solution ultime au problème de l'effet tunnel. Plutôt que de simplement interpoler, il apprend à générer du contenu plausible lors du mouvement avant (*zooming in*).
— **Limitation :** Tendance à la dérive onirique ("dream-like drift"). La géométrie peut devenir incohérente sur de longues distances, ce qui est problématique pour une navigation urbaine précise.

7.2 Google Immersive View

— **Concept clé :** NeRF à grande échelle & Fusion Multi-Modale [23]. C'est la vitrine technologique actuelle de Google.
— **Mécanisme :** Il s'agit d'une reconstruction massive utilisant des champs de radiance (NeRF) combinés à de la photogrammétrie classique, entraînés sur des milliers d'images aériennes et de rue.
— **Différence critique :** Notre projet vise une interpolation "on-the-fly" (client-side ou légère) entre deux photos. Immersive View est une approche "offline" lourde : la scène est pré-calculée et stockée sur le cloud.
— **Perspective :** Bien que non applicable directement pour une transition légère entre deux panos, l'avenir de cette tech réside dans le 3D Gaussian Splatting, qui permettrait un rendu temps réel de ces scènes lourdes sur mobile.

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

— **Concept clé :** Cross-Attention & Epipolar Constraints [24]. Ce papier est le plus pertinent techniquement pour notre cas d'usage (transition 10m).
— **Le problème résolu :** Les méthodes classiques (Flow/Warping) échouent quand l'écart (*baseline*) est grand car la recherche de correspondance est trop difficile.
— **La solution :** Au lieu de chercher un flux optique dense, ils utilisent des Transformers avec des mécanismes d'attention croisée (*Cross-Attention*) le long des lignes épipolaires. Le réseau "construit" l'image intermédiaire en allant chercher l'information pertinente dans les deux images sources via l'attention, plutôt que par un warping géométrique rigide.
— **Avantage :** Robustesse extrême aux grands changements de perspective que le warping classique (Softmax splatting) ne peut pas gérer sans une carte de profondeur parfaite.

7.4 S-NeRF : Neural Radiance Fields for Street Views

— **Concept clé :** NeRF adapté aux scènes urbaines non bornées [14].
— **Innovation :** Les NeRFs classiques échouent sur les scènes de rue (objets proches rapides + ciel infini). S-NeRF améliore la géométrie des véhicules et sépare l'arrière-plan (ciel) du premier plan.
— **Verdict pour le projet :** Comme mentionné dans l'introduction, S-NeRF reste une méthode d'optimisation (*training-based*). Pour naviguer entre deux panoramas A et B, vous devriez "entraîner" un mini-S-NeRF sur ces deux images, ce qui prendrait plusieurs minutes/heures. C'est inadapté pour une navigation utilisateur fluide, sauf si pré-calculé pour toute la ville (coût exorbitant).

7.5 Disney Frame Interpolation Transformer & Uncertainty Guidance

— **Concept clé :** Gestion explicite de l'incertitude pour l'Inpainting [25]. Ce papier introduit une notion cruciale souvent ignorée : l'incertitude.

— **Mécanisme :** Le modèle ne se contente pas de prédire l'image intermédiaire ; il prédit une carte d'incertitude (où le modèle "sait" qu'il échoue, souvent sur les disocclusions ou les mouvements rapides).

— **Application stratégique :** Dans notre pipeline, au lieu de laisser l'interpolateur créer du flou (*ghosting*) dans les zones difficiles, vous pouvez utiliser cette carte d'incertitude comme un masque pour notre modèle d'Inpainting (LaMa).

— **La logique :** "Si l'interpolateur est incertain ici, effacez cette zone et laissez l'IA générative (LaMa/MAT) la redessiner complètement."

Références

[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. Successeur de 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.