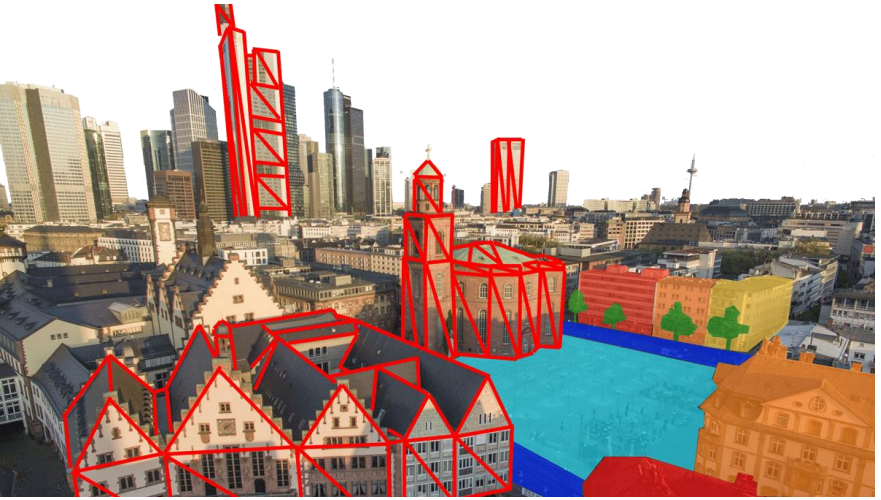
Semantic Scene Understanding from Image Sequences

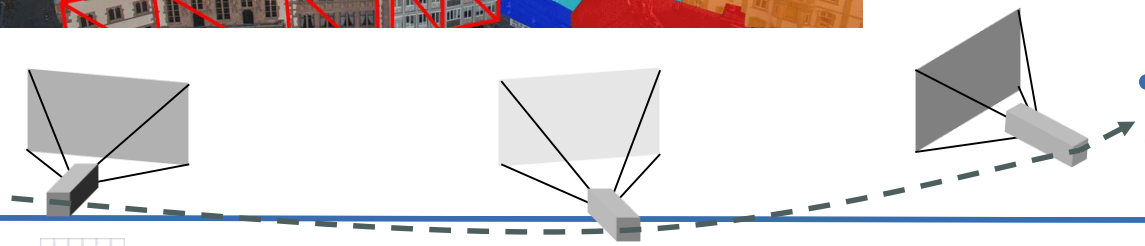
Max Mehltretter

Institute of Photogrammetry and GeoInformation (IPI)
Leibniz University Hannover

Motivation

Goal: Analyse (dynamic) objects w.r.t. 3D spatial, temporal and semantic relationships to its environment and to other objects





Important for:

- Path planning of autonomous systems
- Generation of / interaction with digitial twin
- Can also be transfered to other domains

Use synergies between geometry and semantics

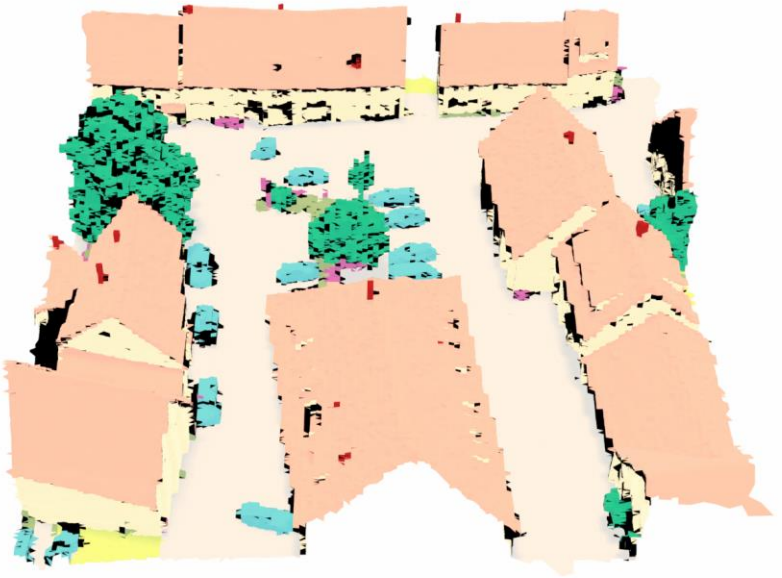
- Joint estimation of both
- Guaranteed consistency between both
- Additional supervision signal via cross-task learning

Problem Statement

Jointly estimate geometry and semantics in 3D from image sequence(s)

$$R = h(\{I_t, P_t\}_{t \in \{1,..,T\}})$$

Task: Development of functional relation h



Input:

- Image sequence with images I_t
- Image orientations P_t

Output:

- Semantically annotated 3D reconstruction R

Open Challenges

- Dynamic scenes
- 3D reconstruction of unseen object parts
- Uncertainty in input and output

Real scenes are often not static

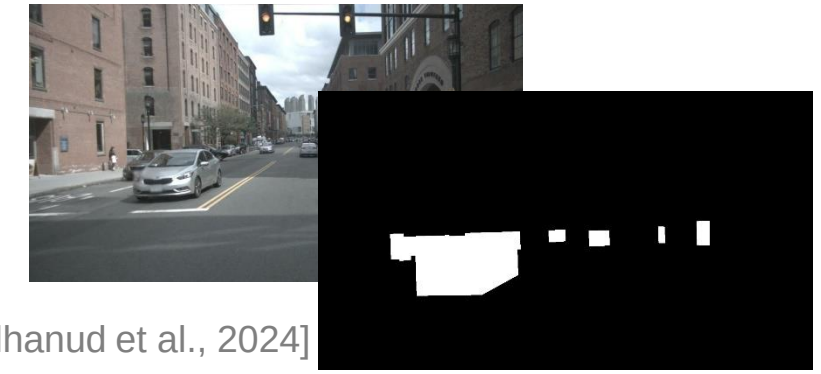
- Most 3D reconstruction methods assume photo-consistency in overlapping images
  - Multi-view stereo, NeRF, Gaussian Splatting, …
  - Requires images to be taken simultaneously (not the case for sequences)
  - or the scene to be static

- Real scenes are typically not static, but contain dynamic objects
  - Violation of photo-consistency assumption
  - Leads to artefacts in reconstructed geometry and semantics
  - Temporal information of sequence allows to investigate dynamic processes

Handling dynamic scenes

Naïve approach

- Filter out observations related to dynamic objects
- Reconstruct static part of scene only
- Often dynamic objects most relevant



[Abualhanud et al., 2024]

[Fridovich-Keil et al., 2023]

3D representation as a function of time

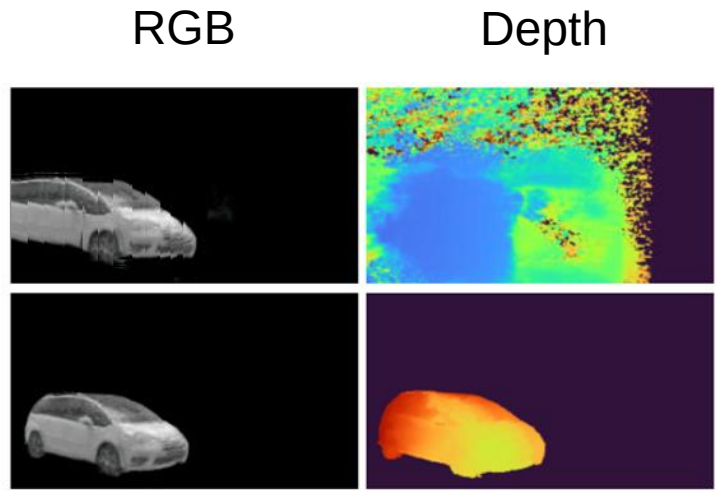
- Time-dependent voxel grid, deformable NeRF, moving Gaussians in Gaussian Splatting
- Artefacts due to naïve local shape interpolation over time
- Computationally expensive (static parts also modelled time-dependent)

Abualhanud et al. (2024): Self-Supervised 3D Semantic Occupancy Prediction from Multi-View 2D Surround Images. PFG Journal.
Fridovich-Keil et al. (2023): K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. CVPR.

Future work

Handle dynamic and static parts separately

- 3D reconstruction if sufficient
- 4D (incl. time) reconstruction if necessary
- Optimise complete reconstruction as a whole



[Abualhanud, 2023]

Instance-based reconstruction of dynamic objects

- Requires to detect and track object instances
- Perform reconstruction in object centric coordinate system
- Avoid naïve shape interpolation
- Enforce instance specific object model instead (shape priors, geometric constraints, …)
- Needs to be extended to deformable objects

[Fridovich-Keil et al., 2023]

Abualhanud, S. (2023): Image-based 3D Reconstruction of Dynamic Objects using NeRF.
Fridovich-Keil et al. (2023): K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. CVPR.

Open Challenges

- Dynamic scenes
- 3D reconstruction of unseen object parts
- Uncertainty in input and output

2.5D representation

“Classical” image matching

- Geometry commonly represented as depth map(s)
- Only information for surface points visible in the image(s)

Reference image

Point cloud

[Reis et al., 2015]

NeRF and Gaussian splatting

- Geometry extracted from density
- Photo-realistic for observed parts
- Cloudy artefacts for unseen parts

[Schob et al., 2023]

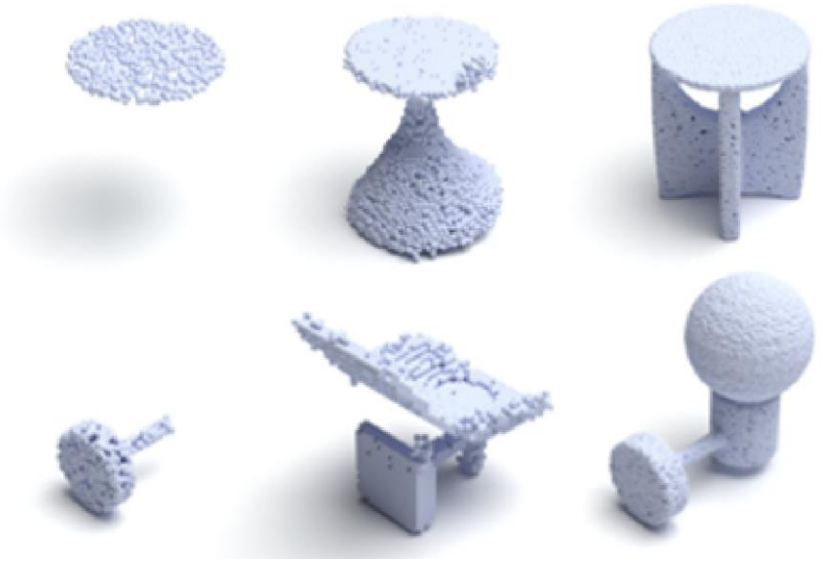
From 2.5D to 3D

Problem:

- 2.5D representation often not sufficient for accurate reasoning on and understanding of observed scene
- Example: measure size / volume of object in reconstruction
- Continuous 3D surface / volumetric representation required

Reconstruct full shape from partial observations:

- Interpolation of small holes (easy)
- Completion of unseen sides / parts of an object
  - Much harder, depending on the object class
  - Often ambiguous



Observation Completion Reference

[Zhang et al., 2021]

Scene completion

Requires strong class specific object model

- Defined based on expert knowledge or learned from training data
- State of the art: hybrid of both (NN-based active shape model [El Amrani et al., 2024], conditional diffusion)
- Open problem for object classes with strong variations in their geometry

Should perform probabilistic reconstruction

- Not just one prediction with maximum likelihood
- But set of plausible solutions / full posterior of shapes
- Posterior can be updated and propagated to down-stream tasks
- Take additional measurements in uncertain regions if needed
- Open problem for complex real-world objects

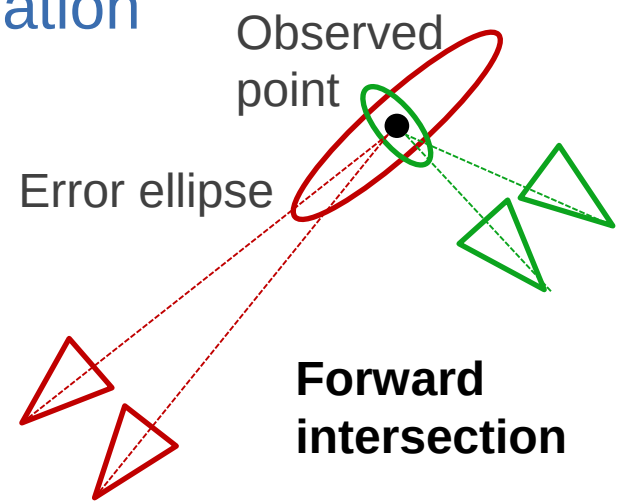
Open Challenges

- Dynamic scenes
- 3D reconstruction of unseen object parts
- Uncertainty in input and output

Uncertain observations

State of the art in DL

- All observations treated as equally important / trustworthy
- Not reasonable, as uncertainty of observations varies, e.g., depending on …
  - distance of 3D points measured via image-based triangulation
  - incidence angle in laser scanner



Future research directions

- Propagation of uncertainty through neural network
- Consideration of temporal dependencies
  - Continuous update of reconstruction with new observations
  - Combination of deep learning and filter-based estimation

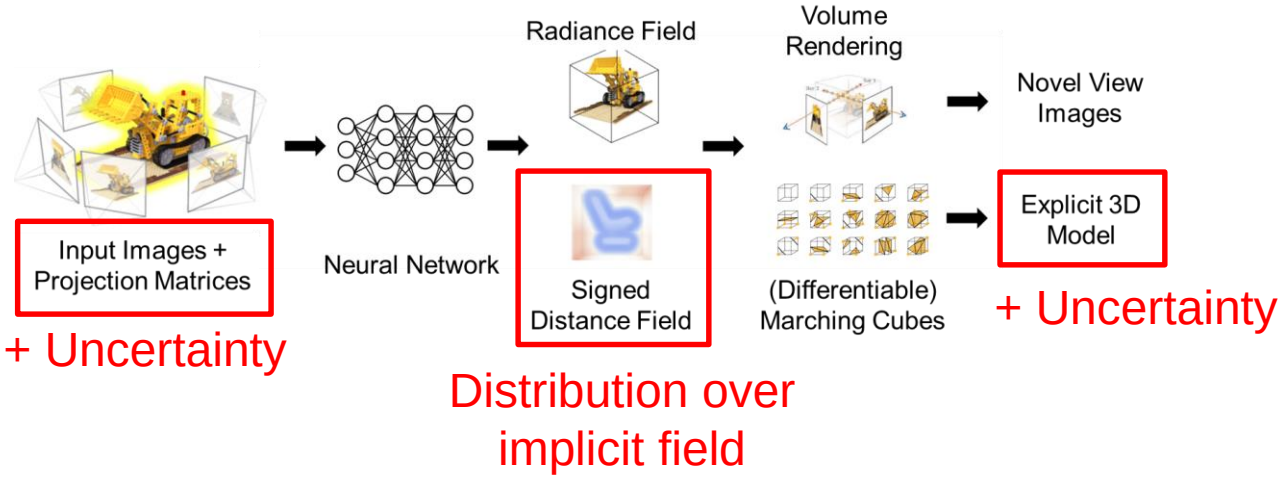
Uncertain estimations

State of the art in DL

- Uncertainty estimation in 2D for discrete points
- Often requires sampling from posterior distribution at test time
- Example: uncertainty-aware depth map via Monte Carlo dropout

Future research directions

- Uncertainty estimation for continuous 3D surface / volumetric representation
  - Basis for probabilistic reconstruction
- Sampling-free uncertainty estimation
  - Evidential deep learning promising direction



Conclusions

- Instance-based 3D reconstruction of dynamic and deformable objects
  - Basis for shape / scene completion
  - Allows integration of object specific constraints and object models
  - Reduces computational effort

- Uncertainty-aware processing
  - Weighting of observations used as input and reference during training
  - Measure of trust for predictions
  - Basis for probabilistic reconstruction

Photogrammetric Computer Vision Group


Christian Grannemann


Max Mehltretter


Maximilian Meyer


Philipp Trusheim


Rasho Ali


Samer Abualhanud


Sara El Amrani Abouelassad


Reza Heidarianbaei


Tianyu Xiu


Dinh Tuan Nguyen

References

- Abualhanud, S. (2023): Image-based 3D Reconstruction of Dynamic Objects using NeRF. Studienarbeit (unpublished).
- Abualhanud, S., Erahan, E., Mehltretter, M. (2024): Self-Supervised 3D Semantic Occupancy Prediction from Multi-View 2D Surround Images. PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science.
- El Amrani Abouelassad, S., Mehltretter, M., Rottensteiner, F. (2024): Monocular Pose and Shape Reconstruction of Vehicles in UAV Imagery using a Multi-task CNN. PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science.
- Fridovich-Keil, S., Meanti, G, Warburg, F.R., Recht, B., Kanazawa, A. (2023): K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 12479-12488.
- Schob, M., Rekittke, J. (2023): Neural Radiance Fields for Landscape Architecture. Journal of Digital Landscape Architecture.
- Zhang, D., Choi, C., Kim, J., & Kim, Y. M. (2021): Learning to generate 3d shapes with generative cellular automata. arXiv preprint arXiv:2103.04130.