3D MOT with DPE (2022)

- Indoor
- Object detection + reconstruction + recognition
- MOTA Score
- New Dataset (MOTFRONT)

3D MOT with DPE (2022)

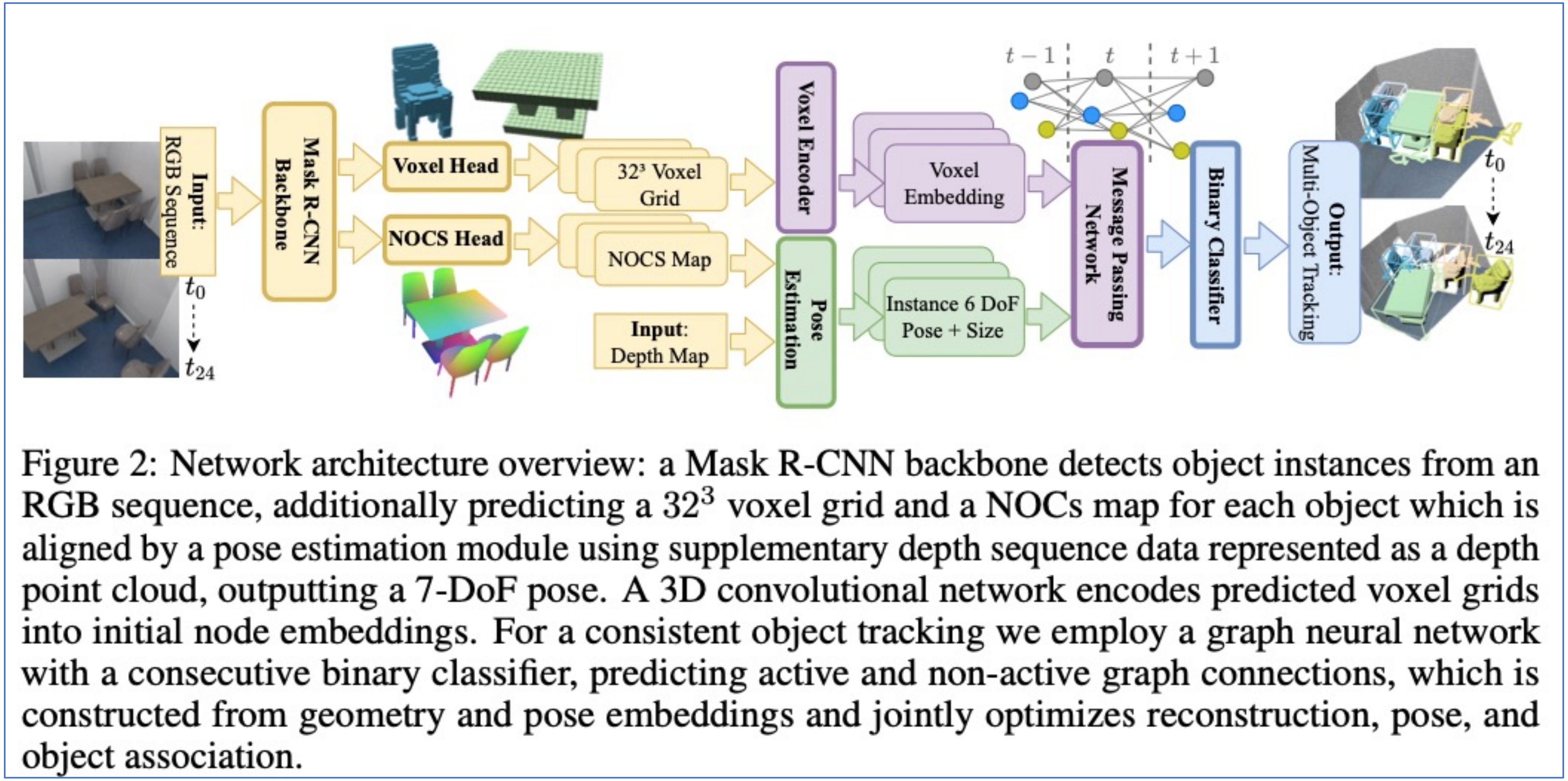

Figure 2: Network architecture overview: a Mask R-CNN backbone detects object instances from an RGB sequence, additionally predicting a 32^3 voxel grid and a NOCs map for each object which is aligned by a pose estimation module using supplementary depth sequence data represented as a depth point cloud, outputting a 7-DoF pose. A 3D convolutional network encodes predicted voxel grids into initial node embeddings. For a consistent object tracking we employ a graph neural network with a consecutive binary classifier, predicting active and non-active graph connections, which is constructed from geometry and pose embeddings and jointly optimizes reconstruction, pose, and object association.

3D MOT with DPE (2022)

👍

- Graph approach for differential pose estimation
- Joint reconstruction & pose estimation to achieve robustness
- Algorithm uses 5 frames at time

3D MOT with DPE (2022)

⛔

- Compared with “State-of-the-Art” Seeing Behind Objects (2020)
  - SBO is the precedent paper from the same research team
  - MOTA results do not match, they use the new dataset and get 4.5% difference compared when using DYNSYNTH
  - SBO uses 2 frames, unfair?
- Looks poorly tested, should have also used more an old dataset
- Basically optimized version of SBO that has good performance on the new dataset

OBJECT FUSION (2022)

- Goal: introduce an efficient and performing object reconstruction method
- Focus on efficiency
- Sliding keyframe window (5-10 frames)
- Pose estimation and reconstruction go get a 3D map of the scene
- SceneNet & ScanNet datasets

OBJECT FUSION (2022)

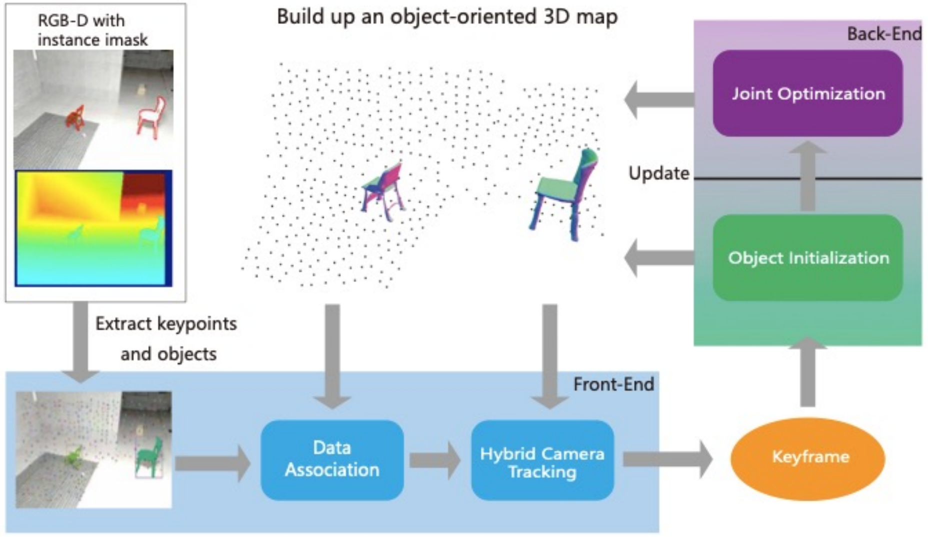

Figure 2. Overview of our ObjectFusion based on deep implicit object representation. ObjectFusion estimates the camera pose of each frame and incrementally builds up 3D surface reconstruction of object instances in the scene.

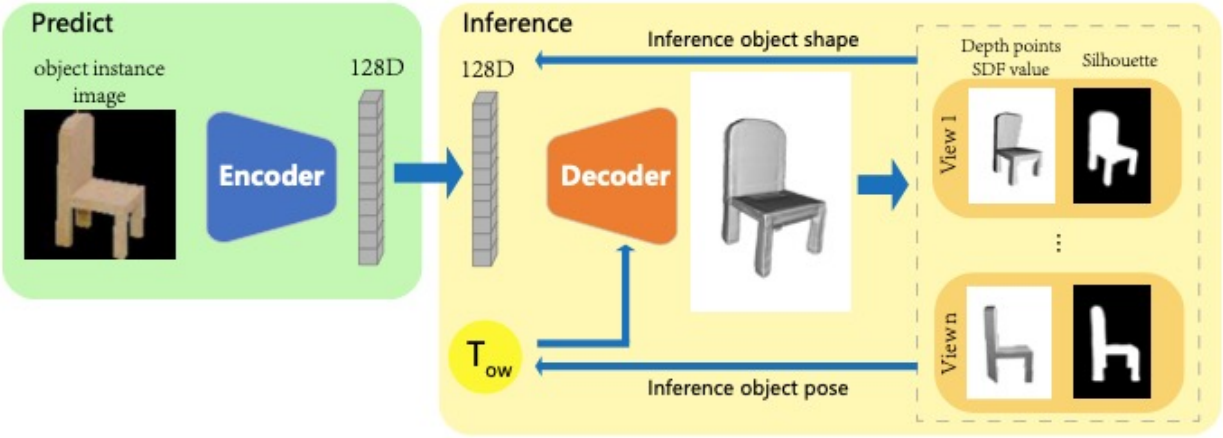

Figure 3. The backbone of our deep implicit object representation. The encoder encodes an object instance image as a latent vector, and then is decoded as a signed distance function of the object. The signed distance value of surface points (depth) and projection silhouette are used object shape and pose inference.

OBJECT FUSION (2022)

- Detect instance segmentation mask of frame
- Encode object to “latent vector”
- Iteratively update object latent vector & pose using hybrid cues
- Joint optimization of object shape, pose and camera pose on a sliding window of frames (5 to 10 frames)

OBJECT FUSION (2022)

👍

- New object representation method, encoded through simple Encoding/Decoding Network
- Great reconstruction quality

OBJECT FUSION (2022)

⛔

- Evaluation strategy is not fine grained, it becomes hard to understand the actual contribution of each optimization
- Unclear what they mean with ”Ours (w/o Obj)”
  - “No object landmarks to evaluate the effect of object term”??
- Still worst results than ORB (2015) in pretty much each scene

Seeing Behind Objects (2020)

- Indoor
- Focus on complete object geometry
- 72h training
- DYNSYNTH & SCANNET datasets

Seeing Behind Objects (2020)

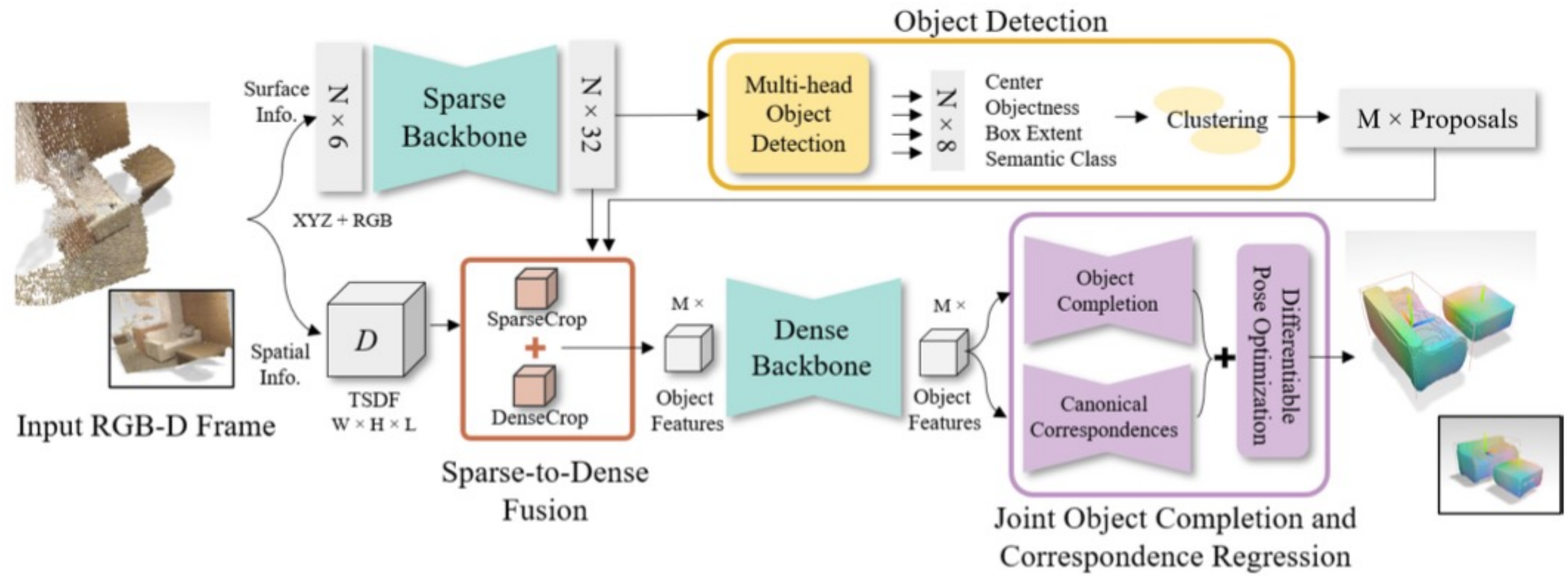

Figure 2. Overview of our network architecture for joint object completion and tracking. From a TSDF representation of an RGB-D frame, we employ a backbone of sparse 3D convolutions to extract features. We then detect objects characterized by 3D bounding boxes, and predict for each object both the complete object geometry beyond the view observation as well as dense correspondences a canonical space; the correspondences on the complete geometry then inform a differentiable pose optimization to produce object pose estimates and within-frame dense correspondences. By predicting correspondences not only in observed regions but also unobserved areas, we can provide strong correspondence overlap under strong object or camera motion, enabling robust dynamic object tracking.

Seeing Behind Objects (2020)

- Object detection via predicting objects boundary boxes
  - Encoder-decoder to extract features
- Sparse to dense fusion to get final object features
- Object completion
  - 3D Convolutions via encoder-decoder -> get dense features
- Object correspondences
- Object tracking based on previous results
  - Hungarian algorithm to find optimal assignment

Seeing Behind Objects (2020)

⛔

- Compared with DETECTRON Mask R-CNN, but this was made for object segmentation and not optimized for object tracking!
  - MASK R-CNN used on Resnet 101 dataset of real world outdoor images, it might overperform with indoor images

DYNA-SLAM (2018)

- Outdoor
- Addresses the issue of having dynamic objects in the scene
- Both monocular camera images and RGB-D
- 5 frames
- TUM and KITTI datasets
- 3D reconstruction

DYNA-SLAM (2018)

👍

- Achieves full dynamic object detection and localization
- Tracks the camera creating a static and reusable map of the scene
- Best in the TUM dataset

DYNA-SLAM (2018)

⛔
- Less accurate then ORB (2015) in KITTI dataset
  - But in scenes with many dynamic objects always performs better