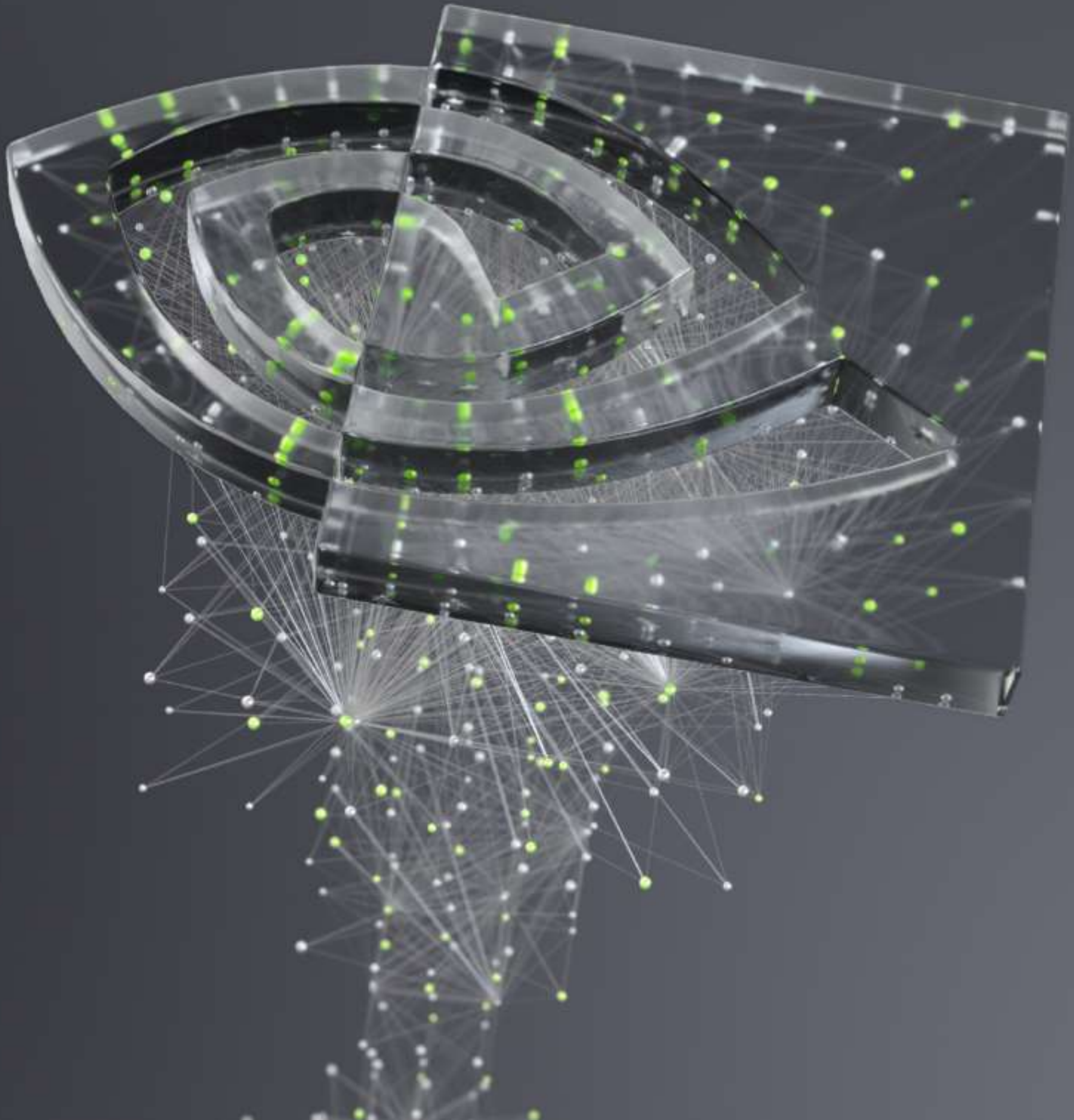
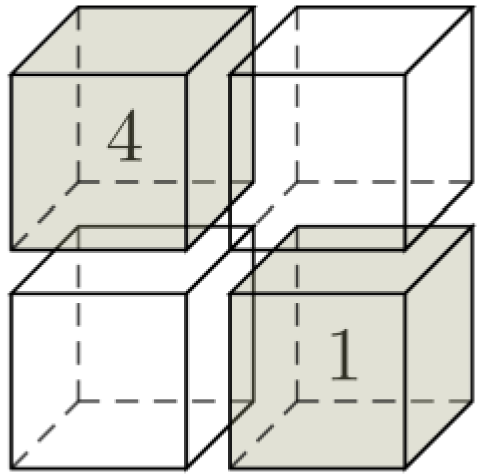


3D PERCEPTION WITH SPARSE TENSORS

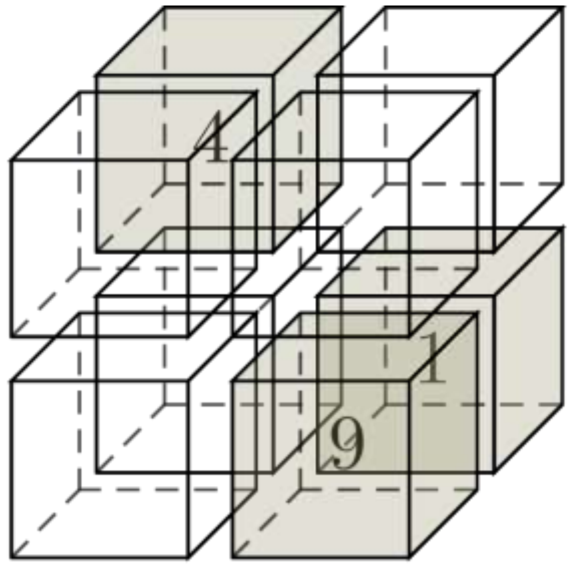
Chris Choy, Nvidia Research

SPARSE TENSOR

Sparse Matrix



Sparse Tensor

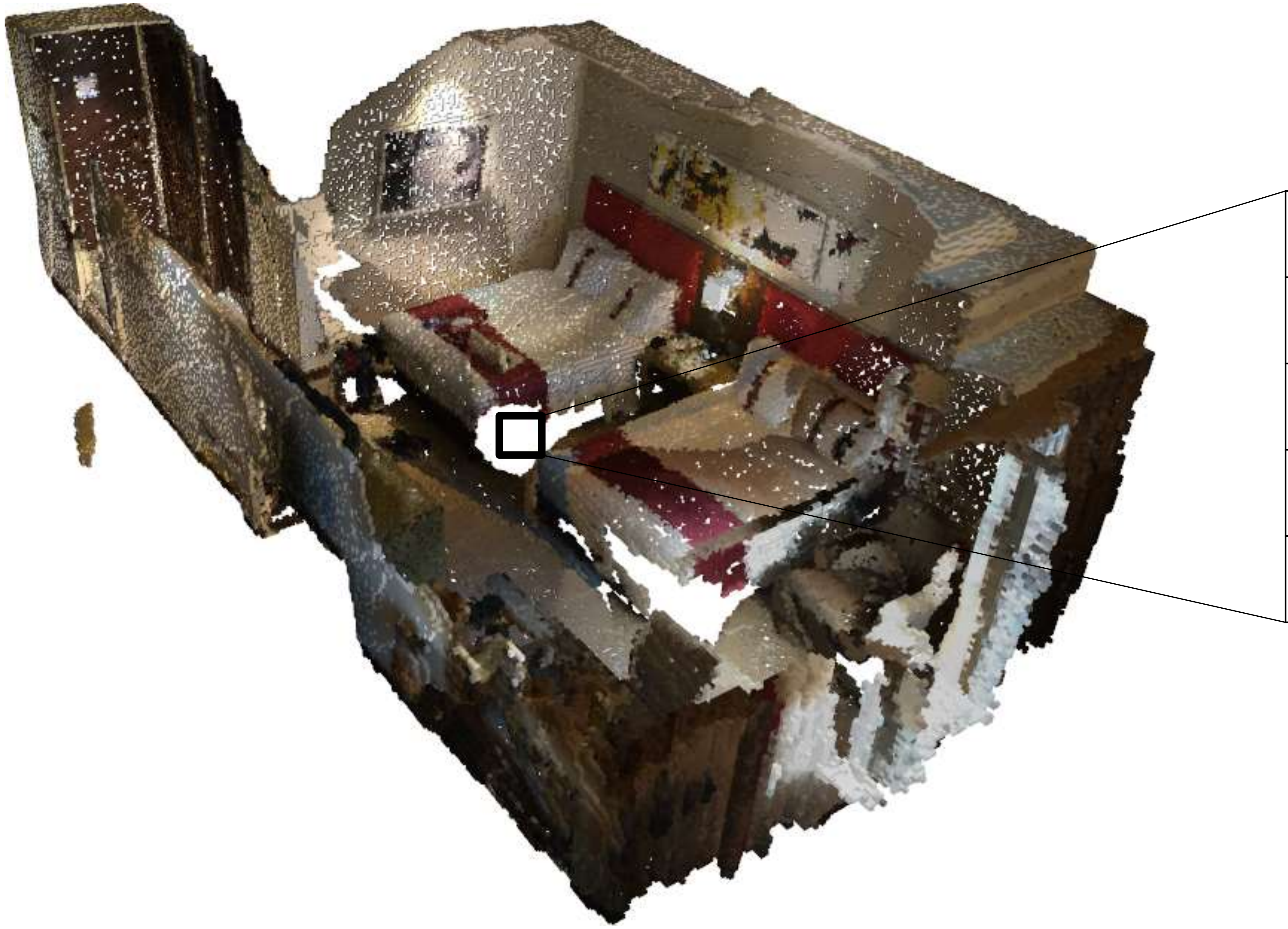


- Sparse Tensor: N-dimensional extension
  - 2x2 matrix
  - 2x2x2 tensor
- COOrdinate (COO) Representation

$$\mathcal{T}[\mathbf{x}_i] = \begin{cases} \mathbf{f}_i & \text{if } \mathbf{x}_i \in \mathcal{C} \\ 0 & \text{otherwise} \end{cases}$$

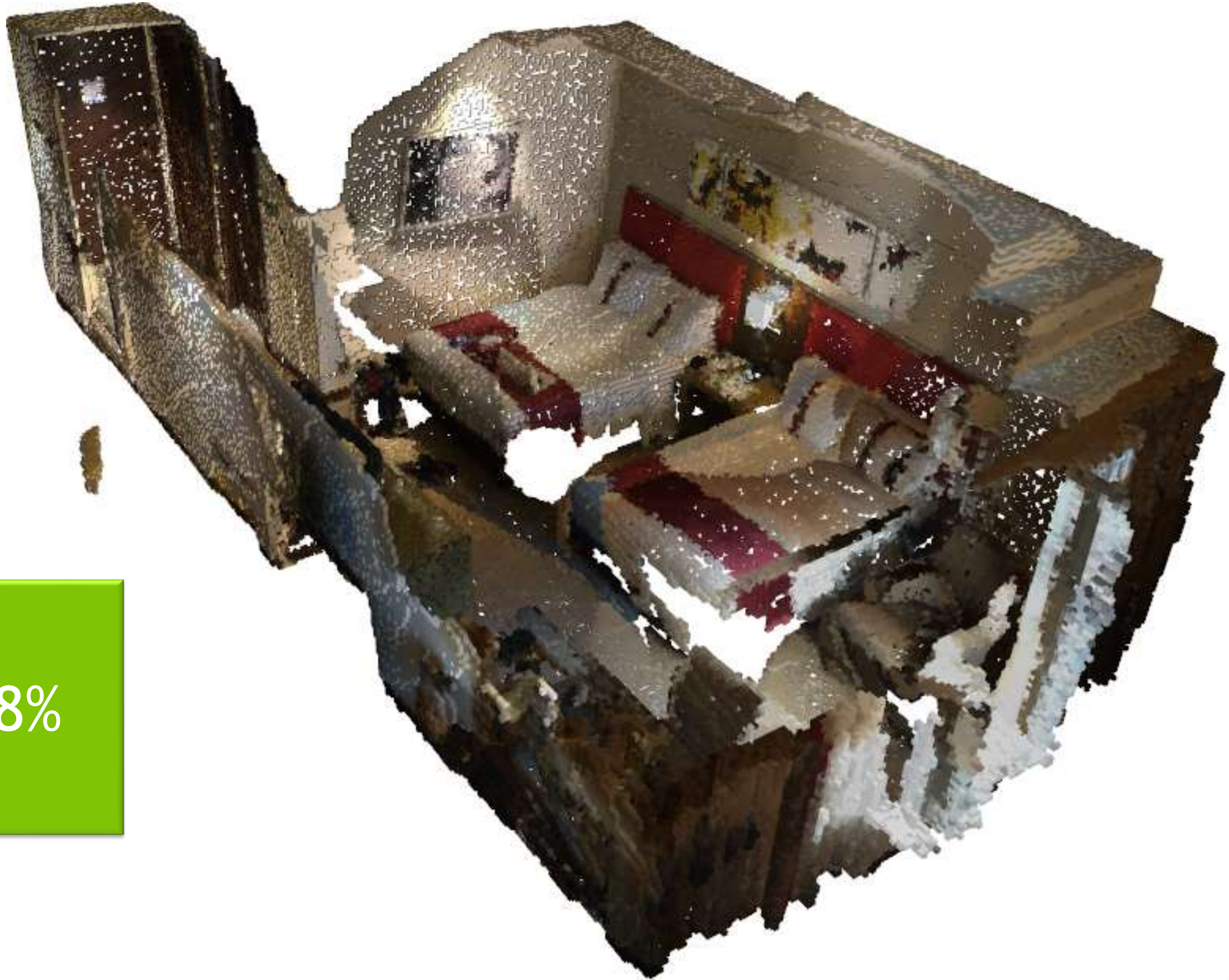
WHY SPARSE TENSOR?

50	34	67	152
67	79	79	154
72	36	39	160
53	29	46	229
48	120	172	232



0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

Dai et al., **ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes**, CVPR’17



2.5cm voxel : 98%

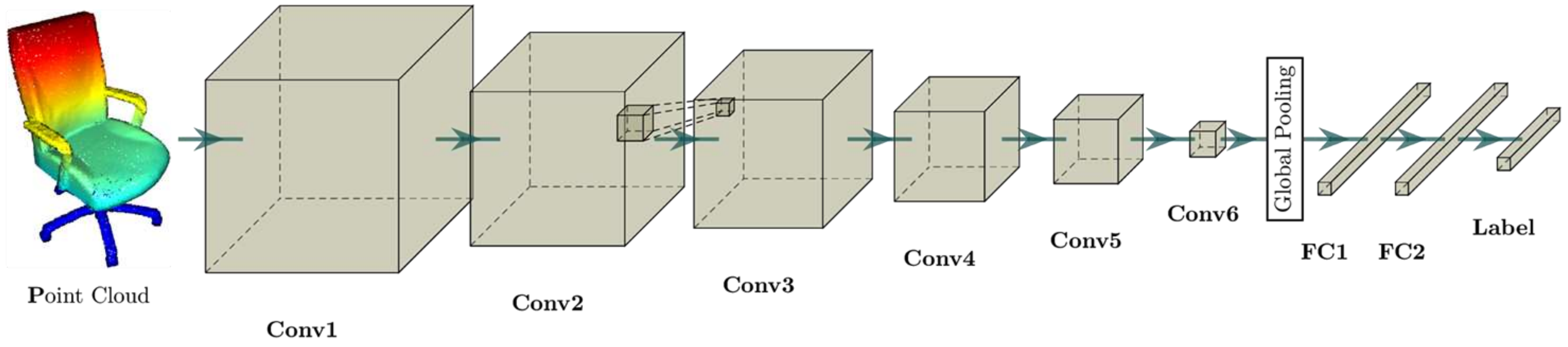
CONTINUOUS VS. DISCRETE

Point Cloud

- No quantization error
- No bound on the number of neighbors
- No random access
- Irregular density
- No hierarchy, or heuristic sampling

Sparse Tensor

- Quantization error
  - Negligible: 1cm for 5m x 5m ScanNet rooms
- Bound on the number of neighbors
- Easy random access
- Hierarchy is deterministic and straight forward

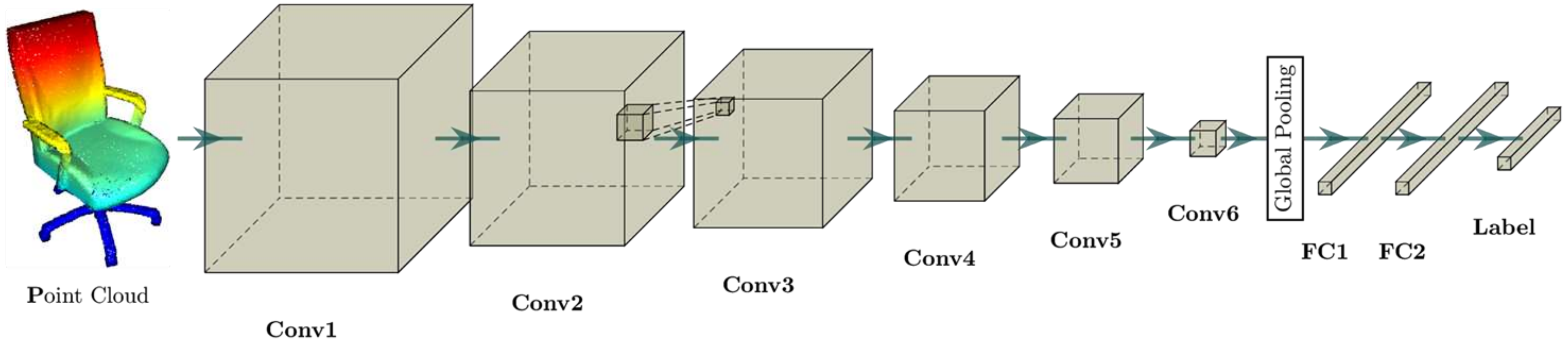


MINKOWSKI ENGINE

Discriminative Networks

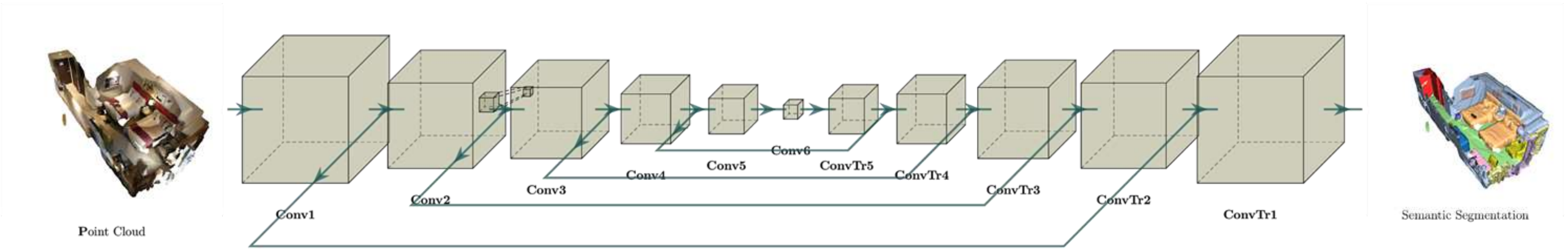
Classification

3D Object to Semantic Label



Semantic Segmentation

3D Scene to Semantic Labels



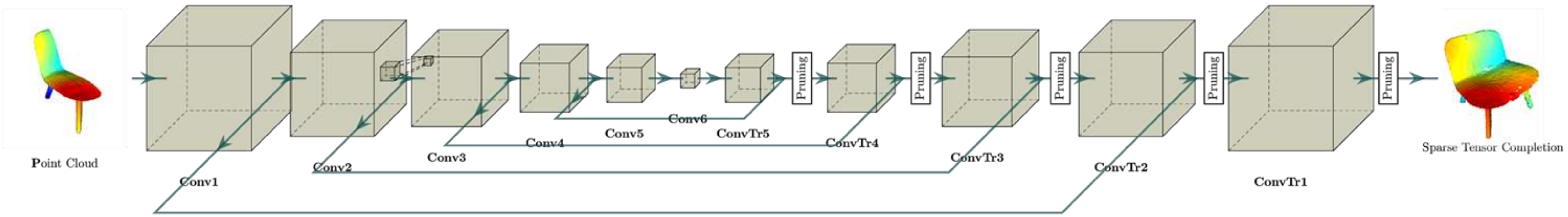
Benjamin Graham, **Sparse 3D convolutional neural networks**, BMVC’15
Chris Choy, JunYoung Gwak, Silvio Savarese, **4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks**, CVPR’19

MINKOWSKI ENGINE

Generation Networks with Generalized Convoluion [Choy et al. CVPR’19]

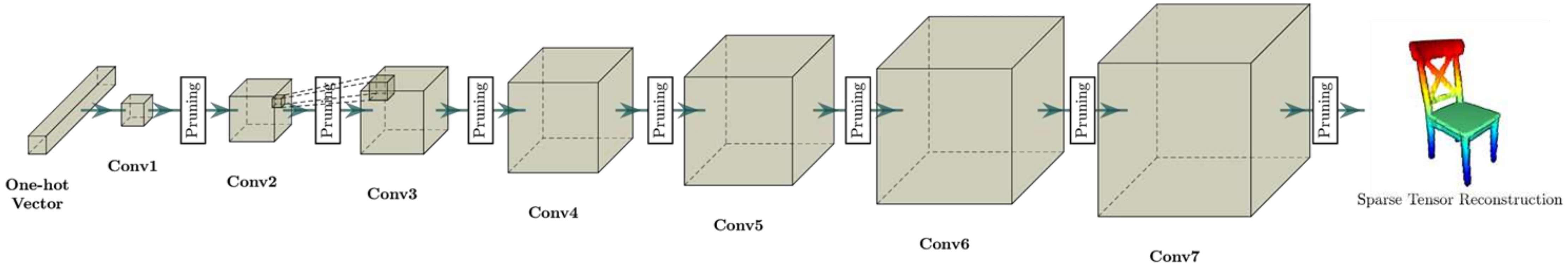
Completion

Partial 3D Object to Complete 3D Object



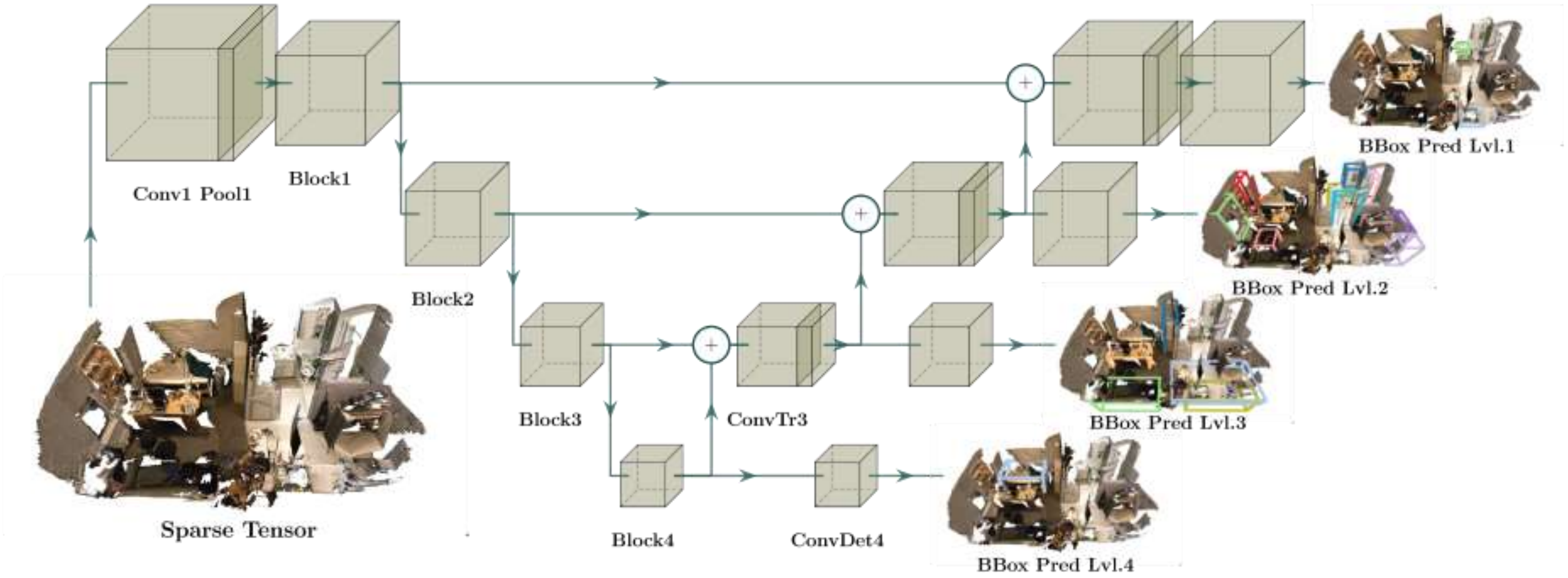
Reconstruction

Feature Vec. to 3D Object



Single-shot Detection

3D Scene to Axis Aligned Bounding Boxes



Chris Choy, JunYoung Gwak, Silvio Savarese, **4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks**, CVPR’19

3D PERCEPTION WITH SPARSE TENSORS

Papers to present

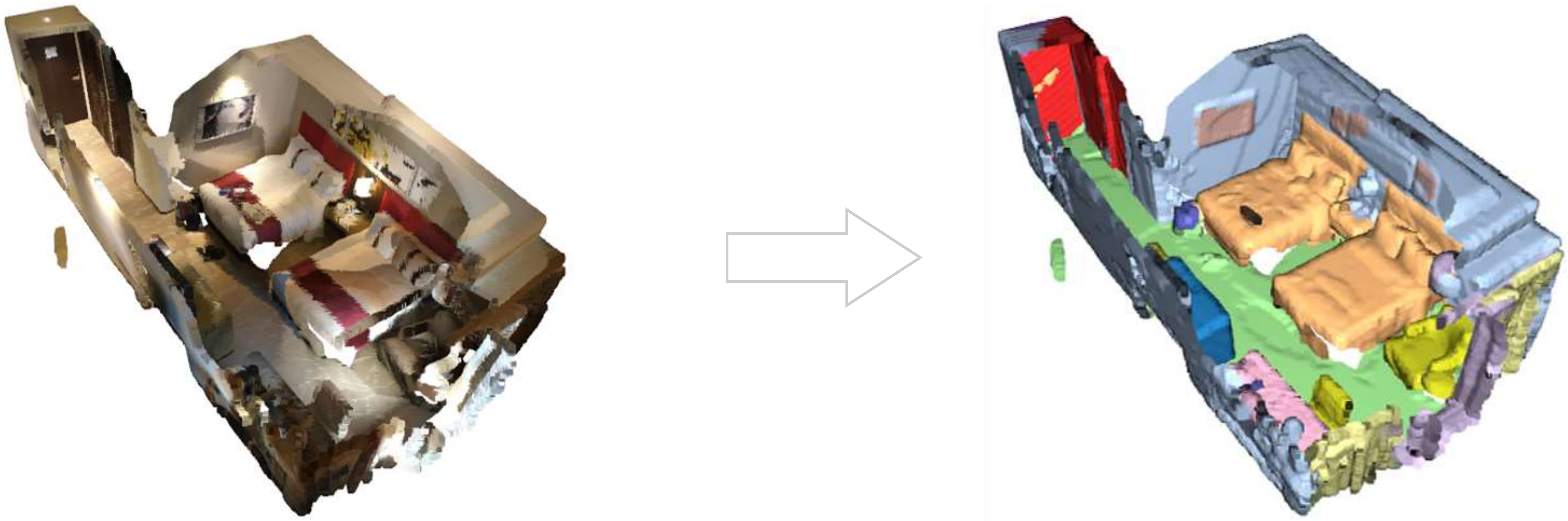
- Choy et al., **4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks**, CVPR’19
- Chris Choy, Jaesik Park, Vladlen Koltun, **Fully Convolutional Geometric Features**, ICCV’19
- Chris Choy, Wei Dong, Vladlen Koltun, **Deep Global Registration**, CVPR’20 Oral
- Choy et al., **High-dimensional Convolutional Networks for Geometric Pattern Recognition**, CVPR’20 Oral
- Gwak et al., **Generative Sparse Detection Networks for 3D Single-shot Object Detection**, preprint 2020

4D SEMANTIC SEGMENTATION

Choy et al., **4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks**, CVPR’19

SEMANTIC SEGMENTATION

- Partition 3D scans or data into semantic parts
- Label each voxel or 3D point as one of semantic labels



Dai et al., **ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes**, CVPR'17

Chris Choy, JunYoung Gwak, Silvio Savarese, **4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks**, CVPR'19

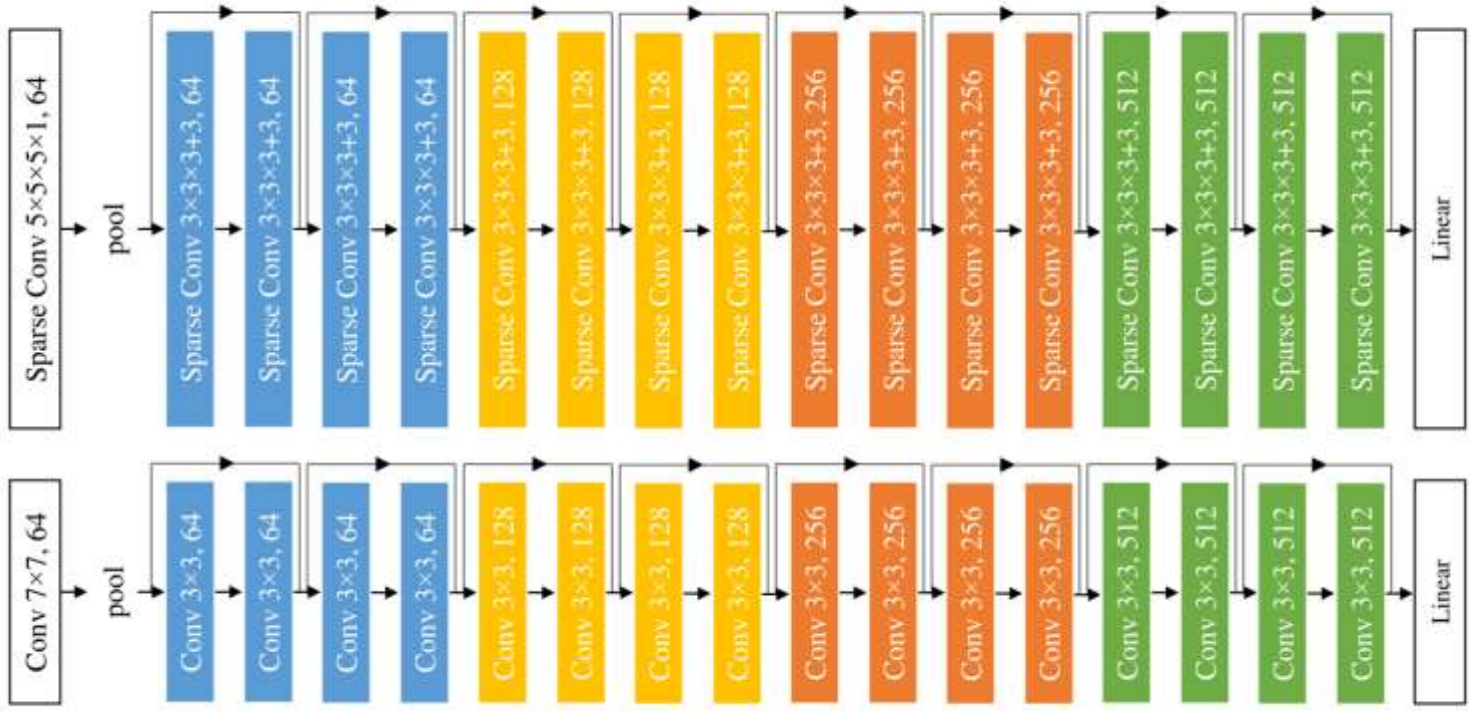
MINKOWSKI NETWORKS

- First very deep convolutional neural networks achieved SOTA on ScanNet (CVPR’19 2018 Nov)
  - 42-layer deep neural networks for semantic segmentation

- Reuse network architectures found from years of research in 2D
  - Residual Network
  - U-Net, or Pyramid Network

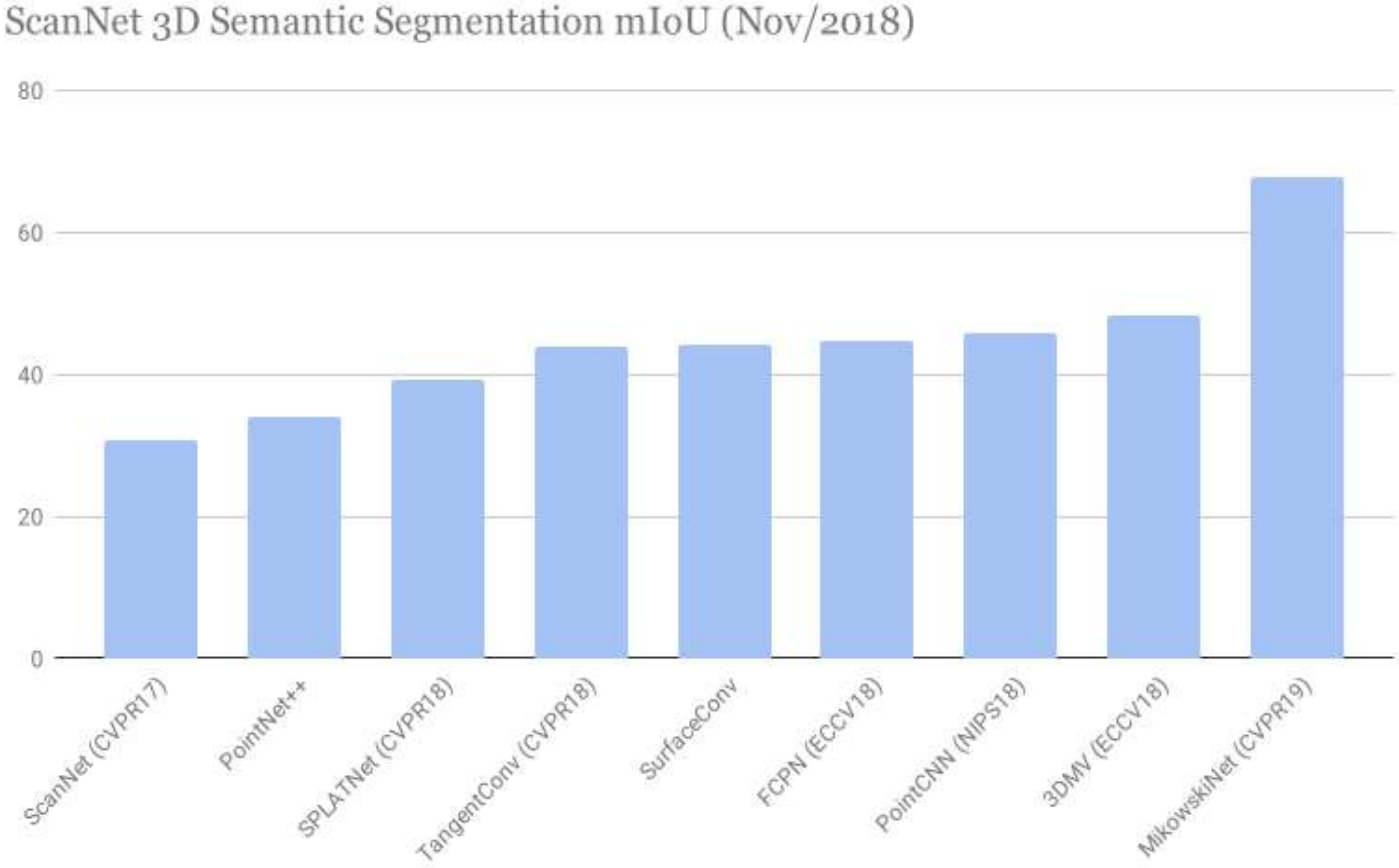
4D MinkNet18

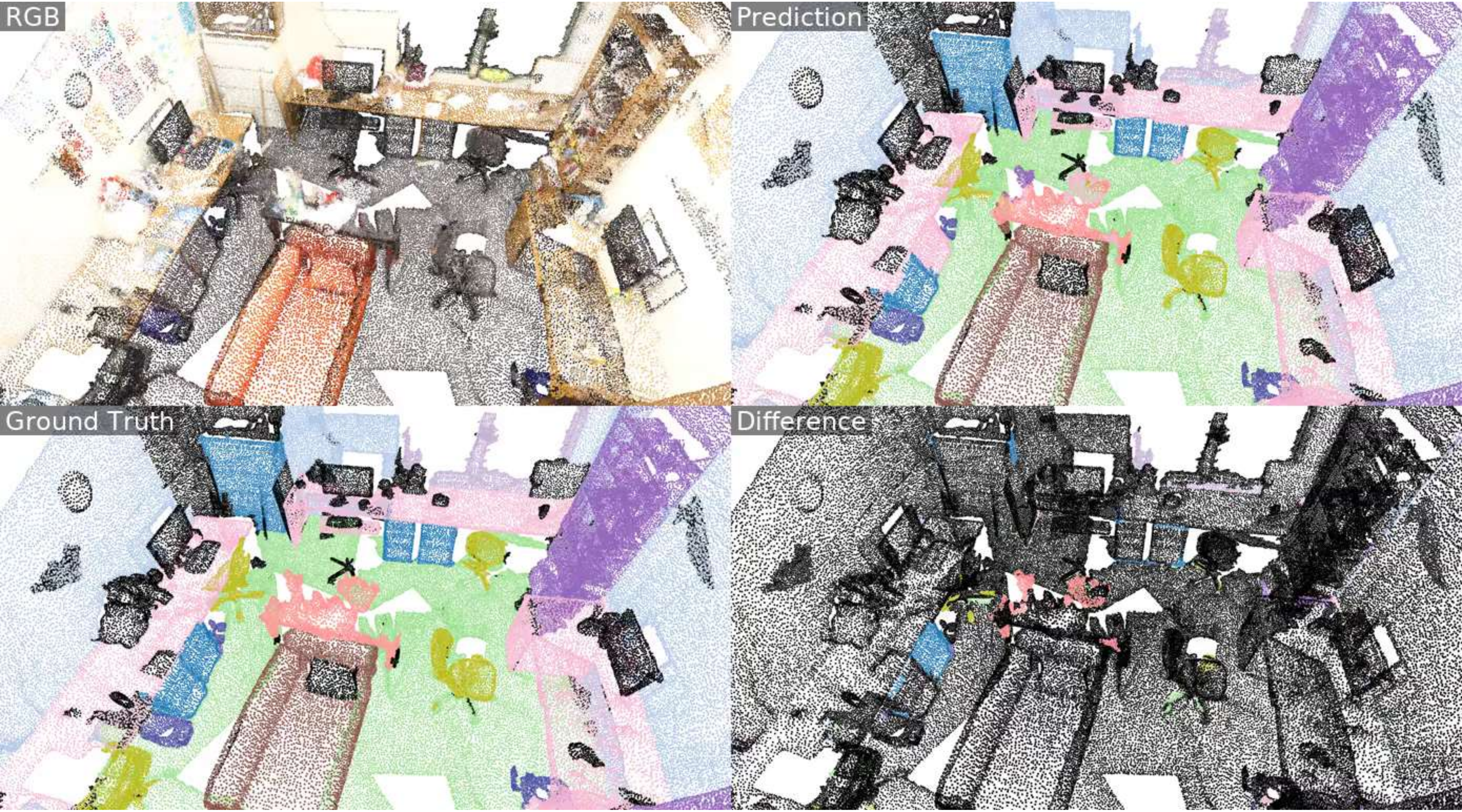
ResNet18



Chris Choy, JunYoung Gwak, Silvio Savarese, **4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks**, CVPR’19

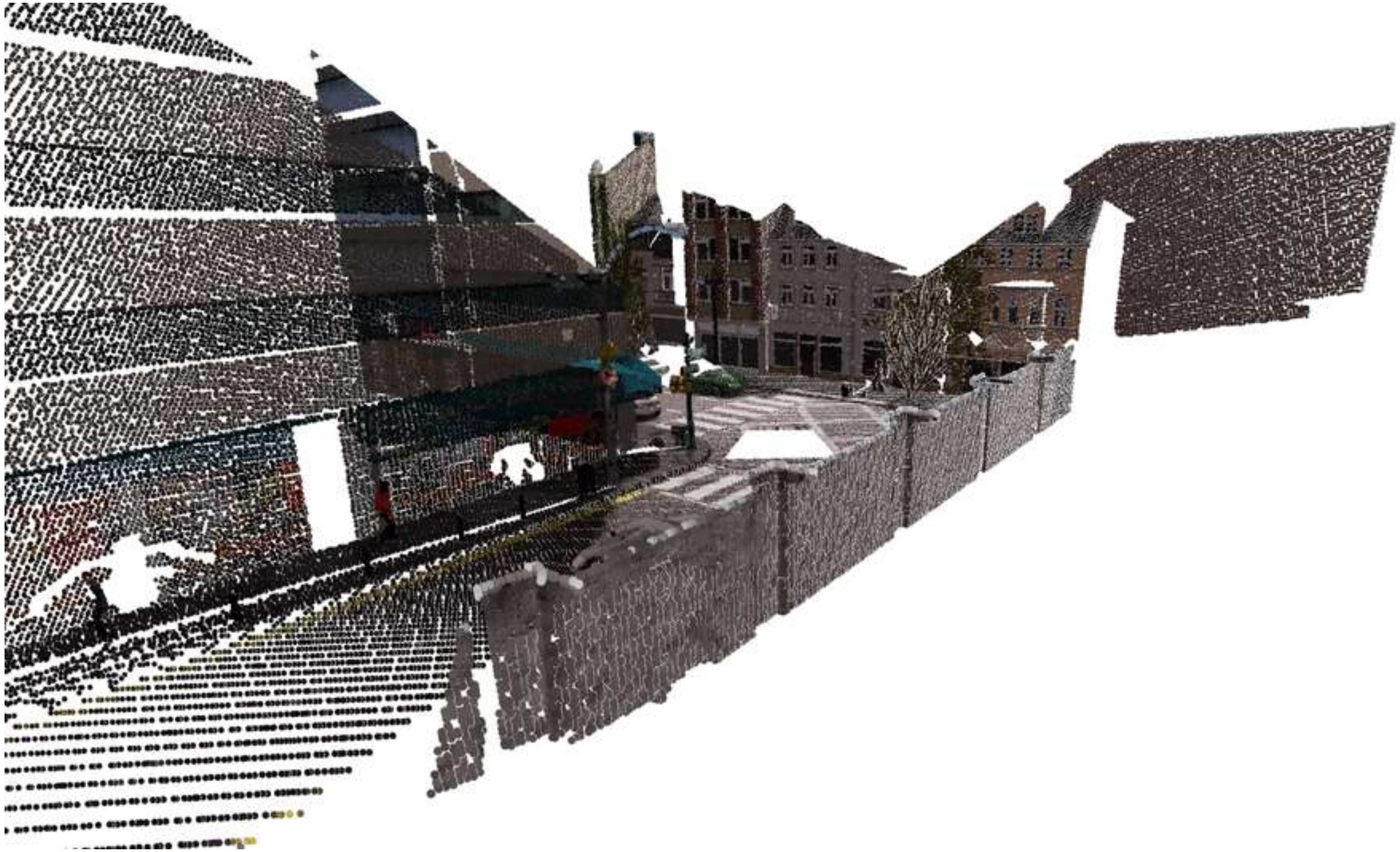
SCANNET 3D SEMANTIC SEGMENTATION BENCHMARK




RGB
Prediction
Ground Truth
Difference

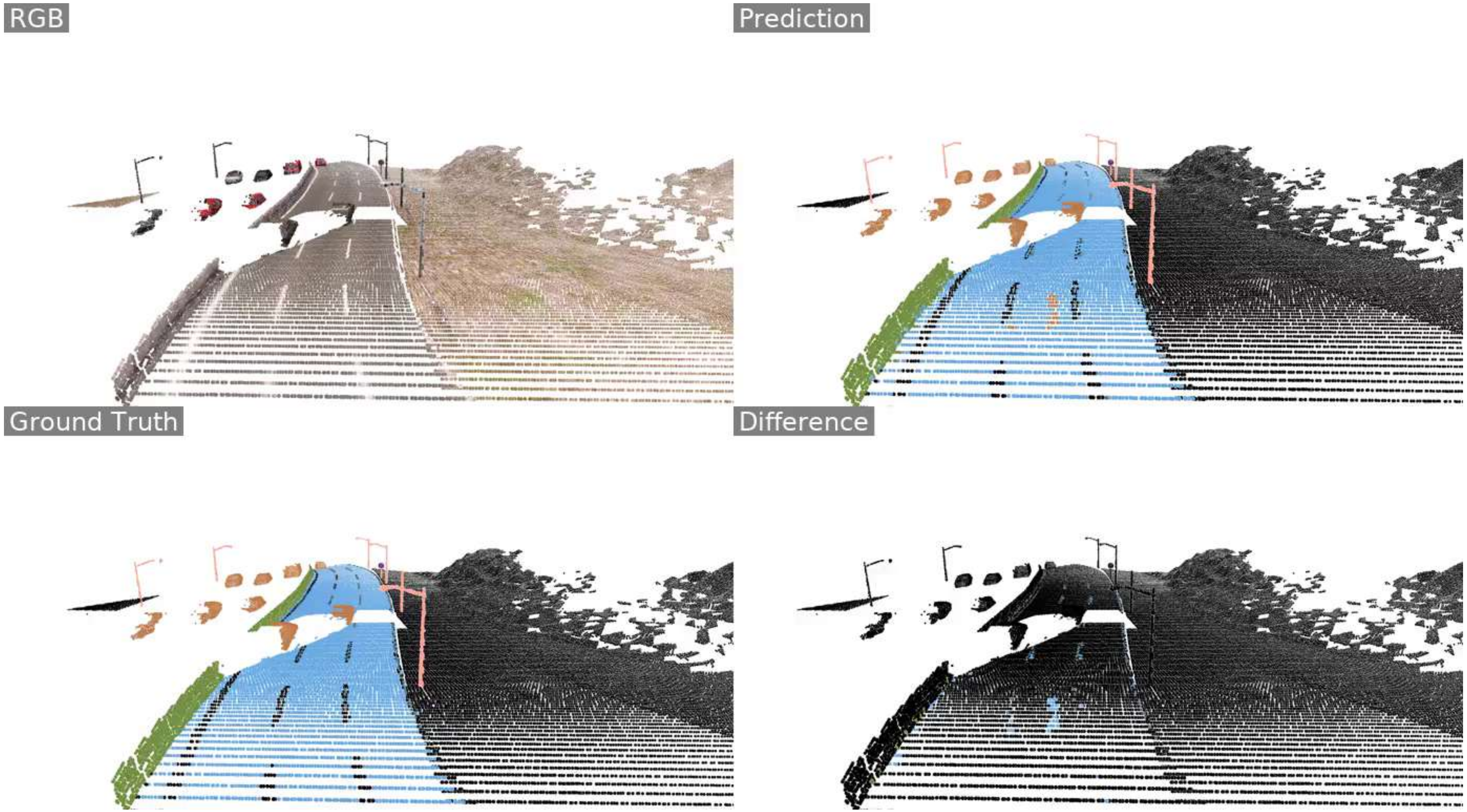
4D SPATIO-TEMPORAL SPACE

3D space + time as a single entity (Minkowski space)



4D CONVNET OVER SPACE AND TIME

3D space + time as a single entity (Minkowski space)

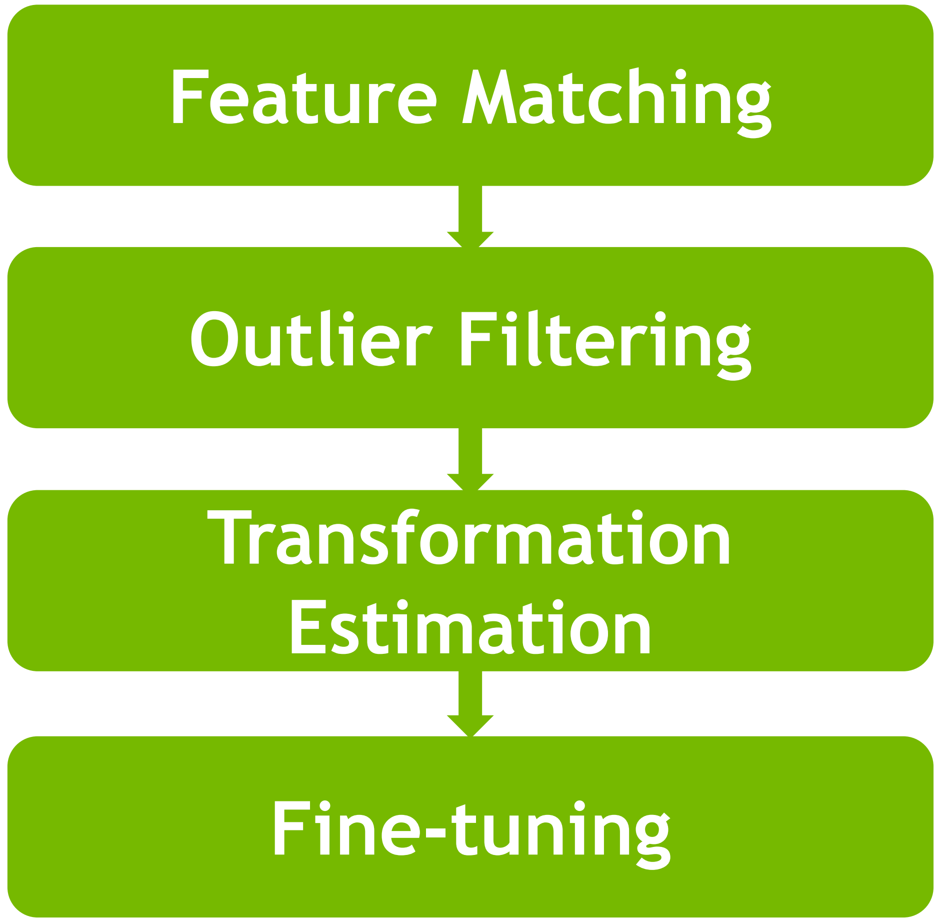


3D FEATURE MATCHING

Chris Choy, Jaesik Park, Vladlen Koltun, **Fully Convolutional Geometric Features**, ICCV'19

MULTI-VIEW 3D RECONSTRUCTION

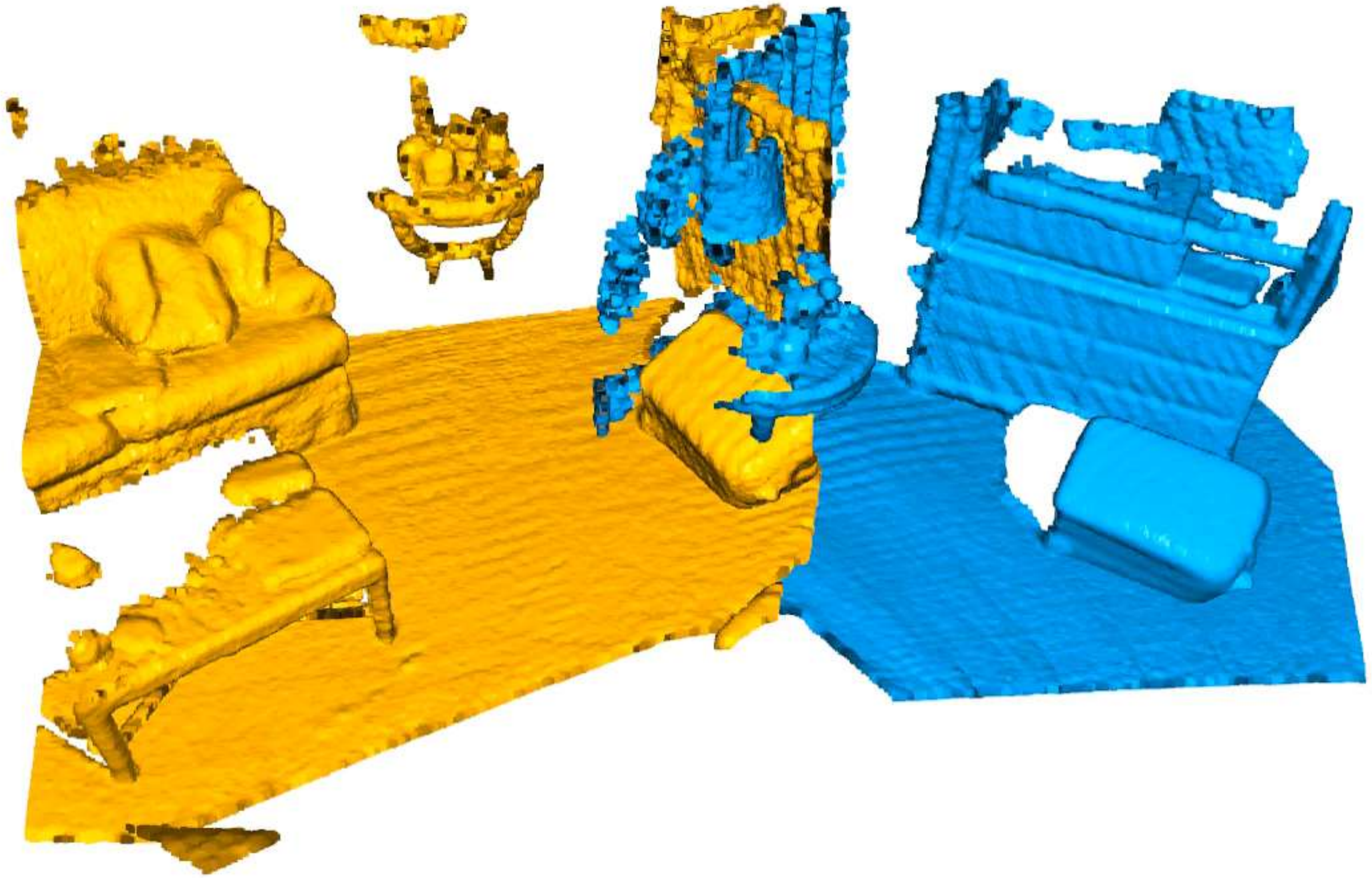
Pipelines when no camera extrinsics are given


Feature Matching

Outlier Filtering

Transformation Estimation

Fine-tuning



PRIOR WORKS IN 3D GEOMETRIC FEATURES

Sliding-window-style (crop and extract) features

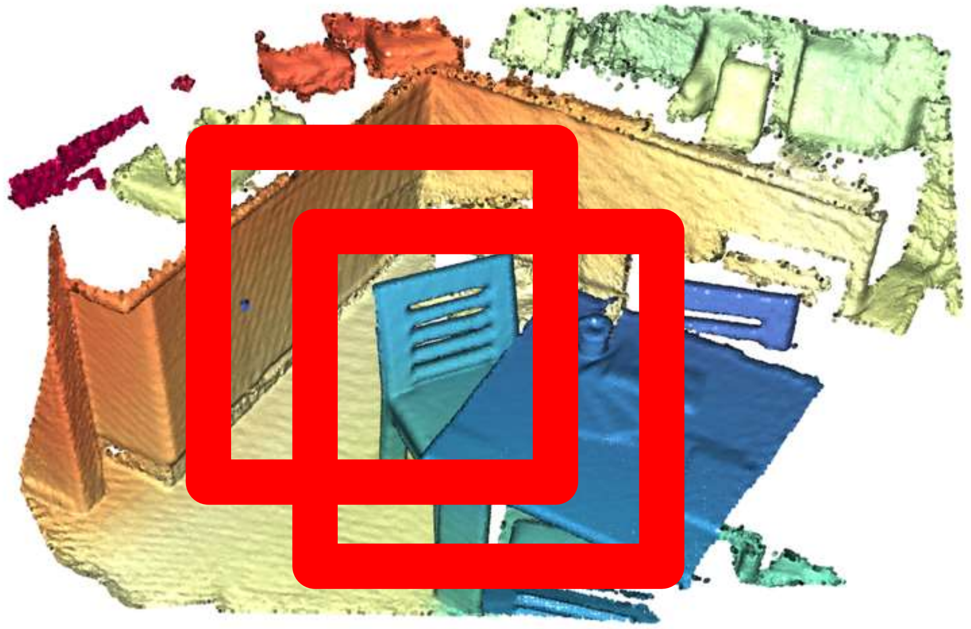
Hand-designed Features

Spin Image, USC, SHOT, PFH, FPFH

Learned Features

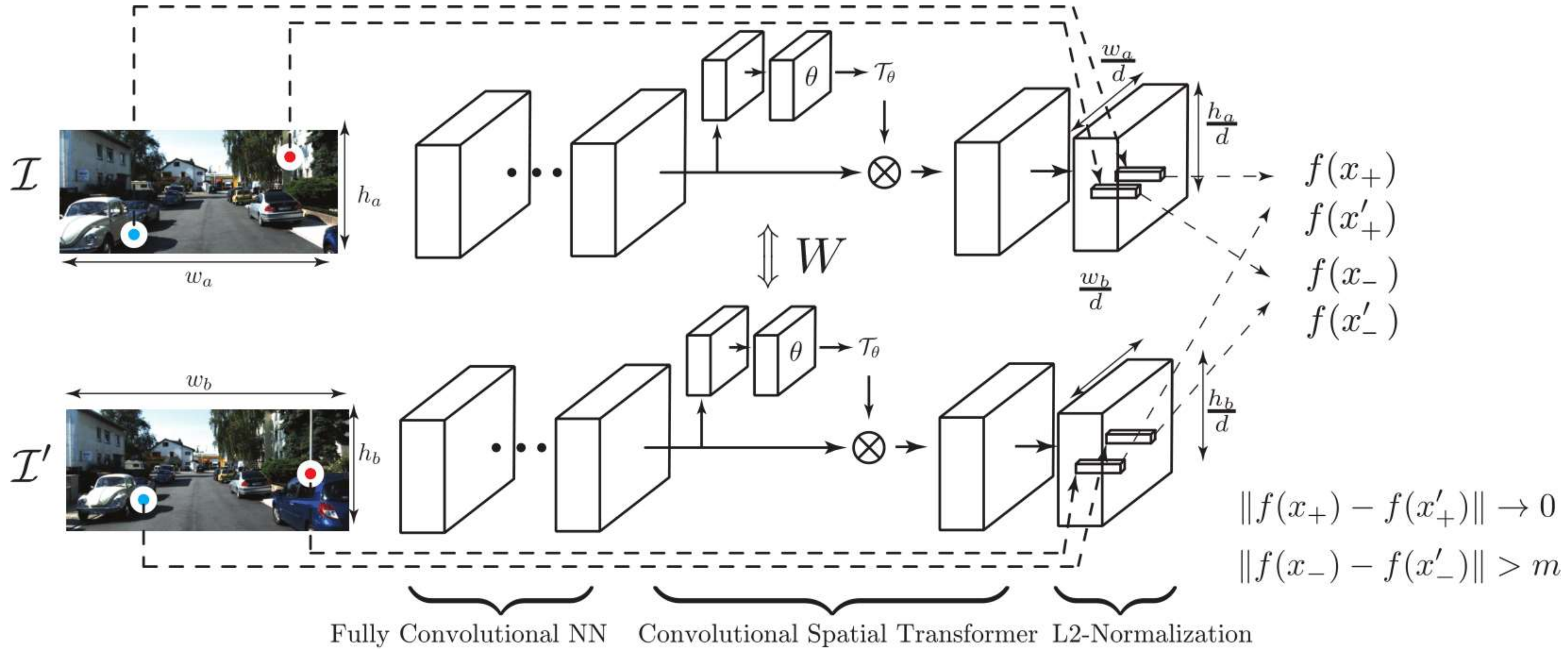
3DMatch, CGF, PointNet, PPF, FoldNet, PPFFold, CapsuleNet, DirectReg, 3DSmoothNet

- Extract a small 3D patch
  - Limits context, receptive field
  - Features extracted separately
- Preprocessing
  - Normal, Signed Distance Function, curvatures

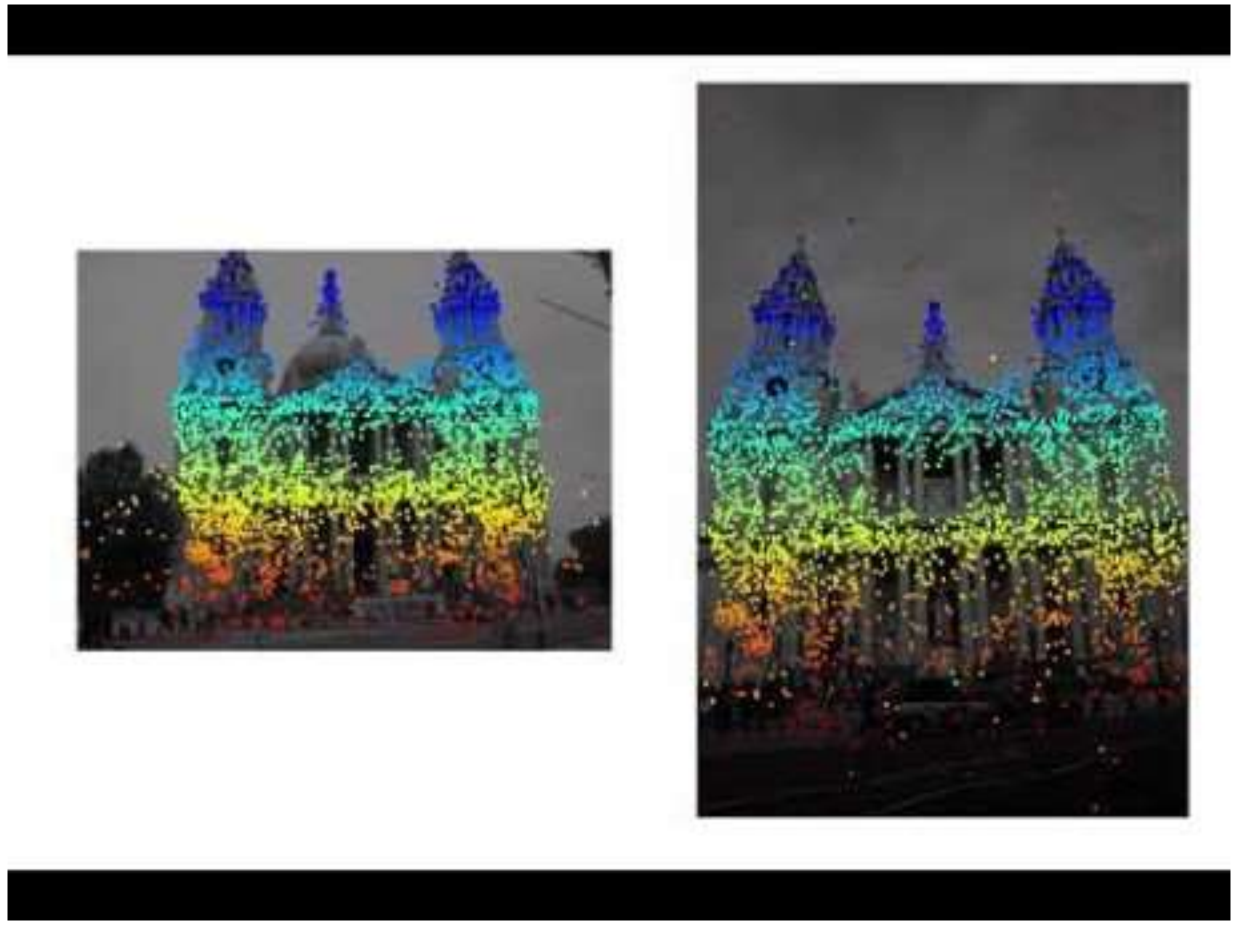


FULLY CONVOLUTIONAL METRIC LEARNING

Dense geometric feature learning with metric-learning loss



- The first fully convolutional metric learning
- Convolutional Spatial Transformer
  - Precursor of deformable convolution

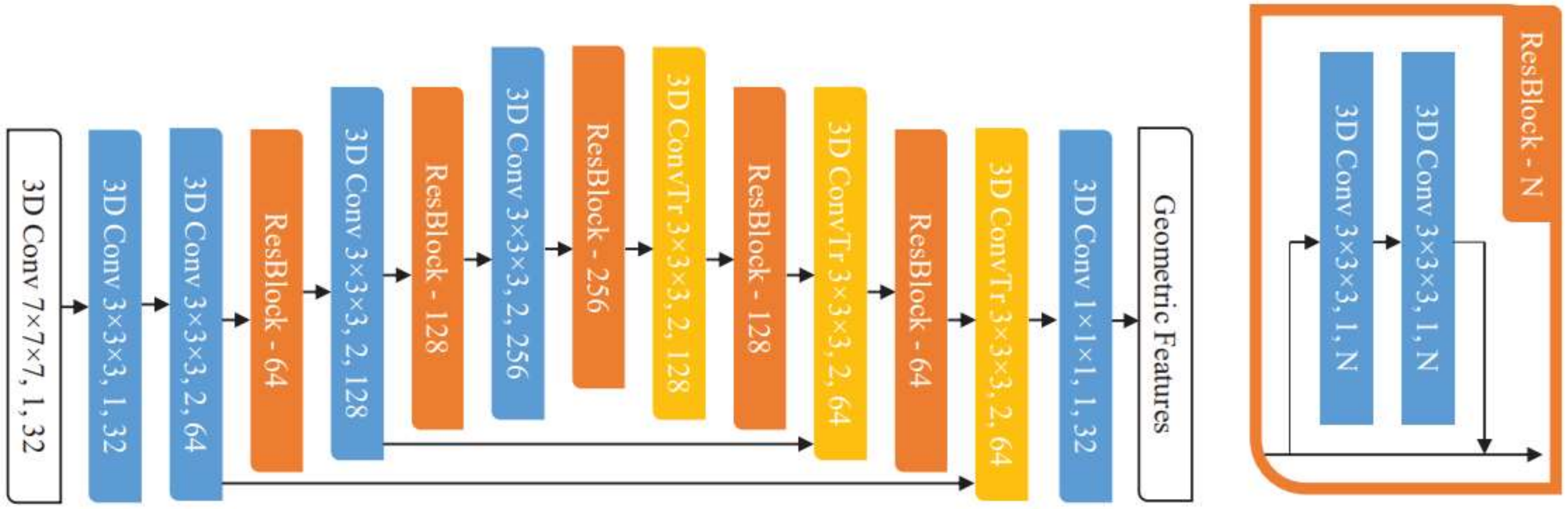


Choy et al., **Universal Correspondence Network**, NIPS'16
Choy et al., **Fully Convolutional Geometric Features**, ICCV'19
Choy and Lee, **Open UCN**, github'20

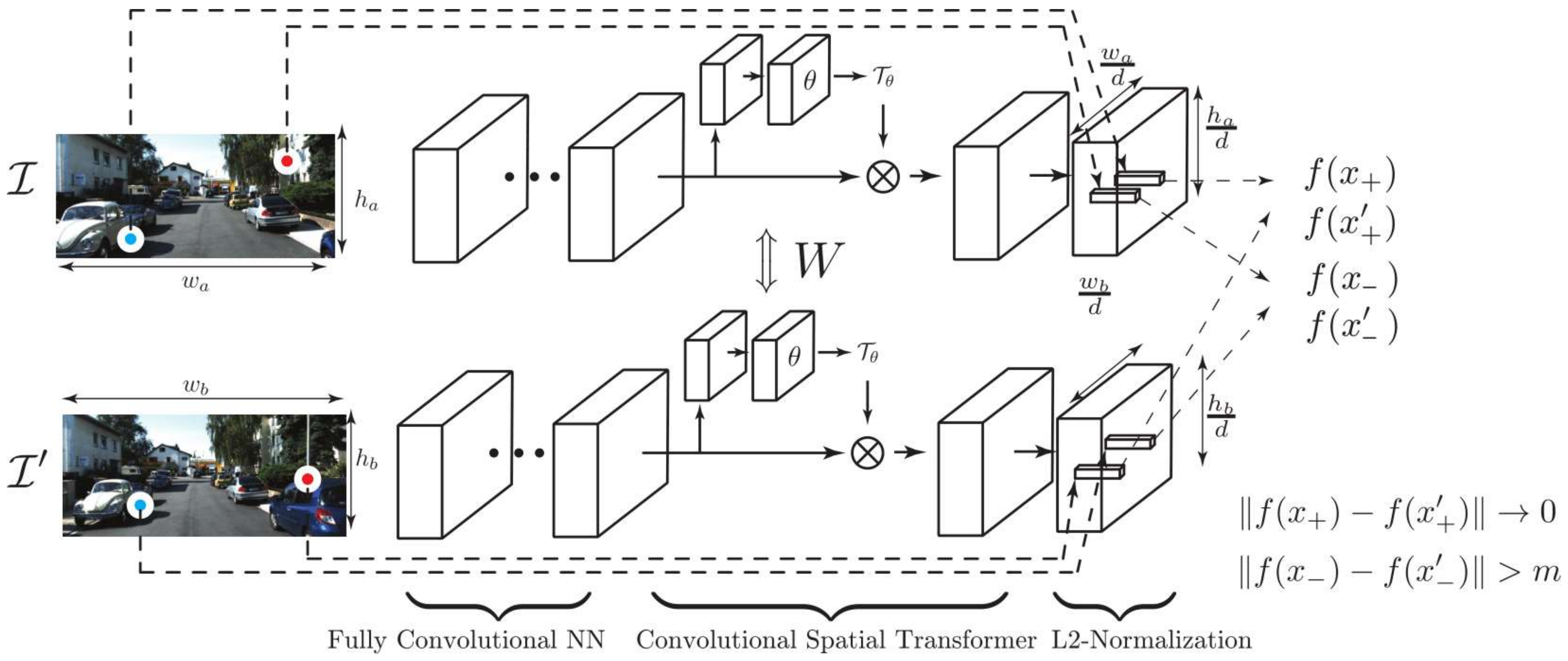
SPARSE FULLY CONVOLUTIONAL METRIC LEARNING

Fully Convolutional Networks on Sparse Tensorized Input

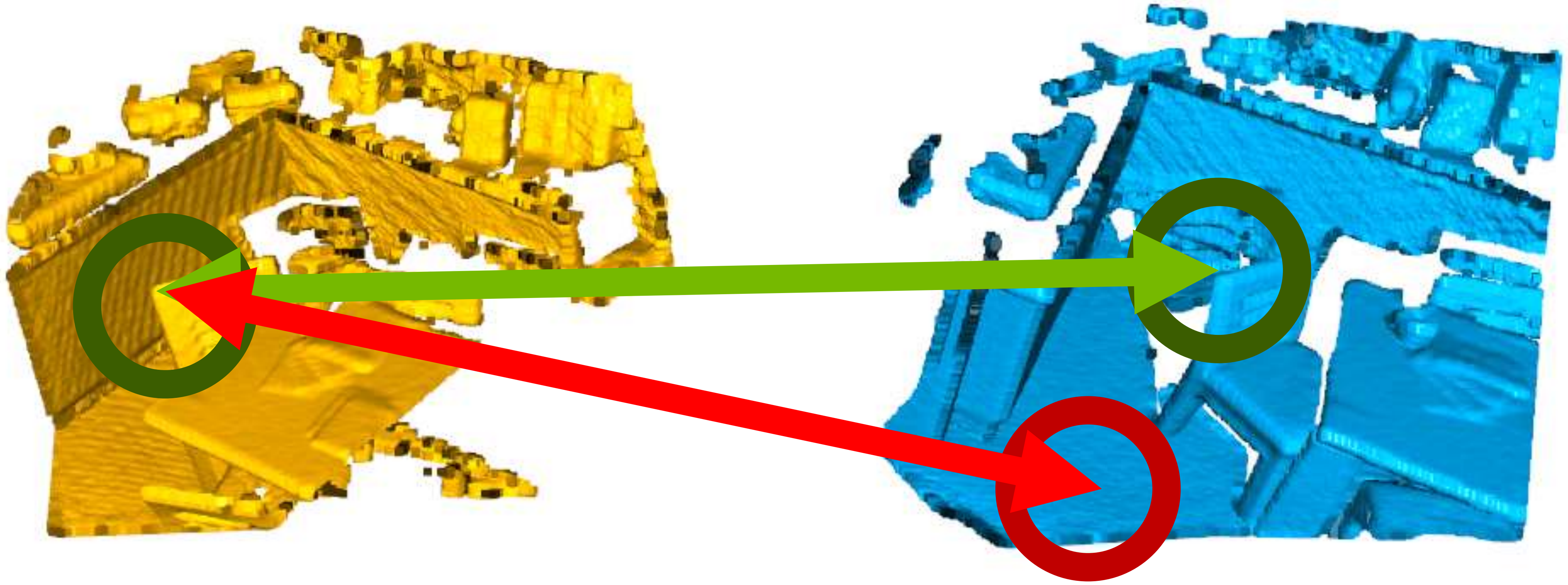
- Dense Image → Spatially Sparse Tensor
- Residual Network + U-Net + Minkowski Engine
  - MinkowskiUNet



Choy et al., **Universal Correspondence Network**, NIPS’16
Choy et al., **Fully Convolutional Geometric Features**, ICCV’19

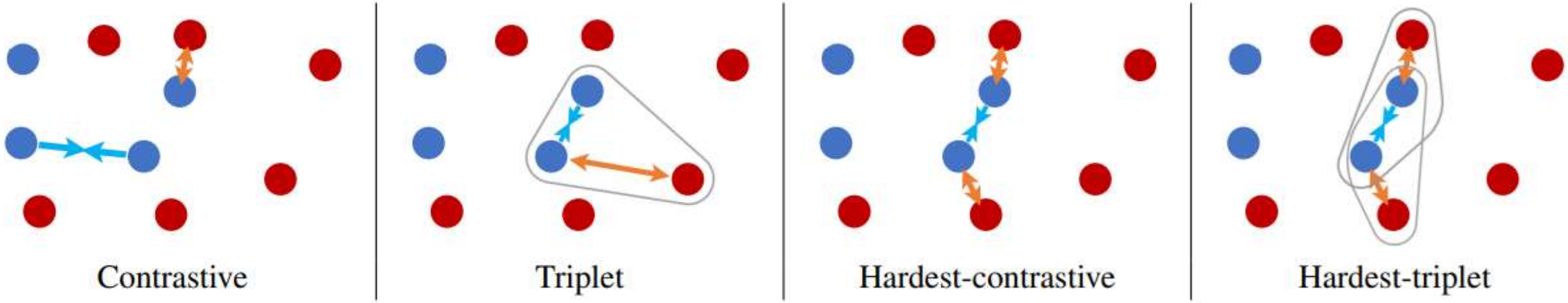


$\mathcal{I}$
$\mathcal{I}'$
h_a
w_a
h_b
w_b
$\mathcal{T}_\theta$
θ
W
$\frac{w_a}{d}$
$\frac{h_a}{d}$
$\frac{w_b}{d}$
$\frac{h_b}{d}$
$f(x_+)$
$f(x'_+)$
$f(x_-)$
$f(x'_-)$
$\|f(x_+) - f(x'_+)\| \to 0$
$\|f(x_-) - f(x'_-)\| > m$
Fully Convolutional NN
Convolutional Spatial Transformer
L2-Normalization



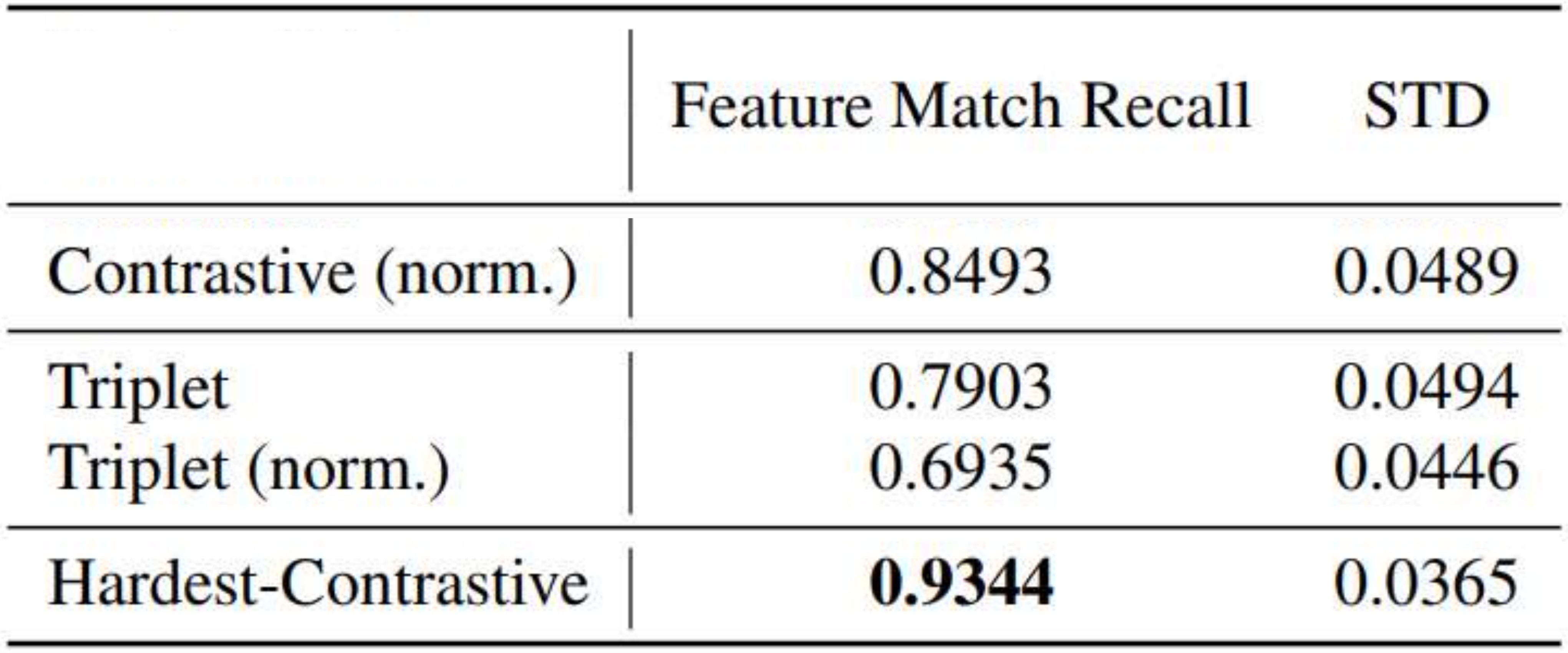
FULLY CONVOLUTIONAL HARDEST CONTRASTIVE LOSS

Fully Convolutional Networks on Sparse Tensorized Input



Choy et al., **Universal Correspondence Network**, NIPS’16
Choy et al., **Fully Convolutional Geometric Features**, ICCV’19

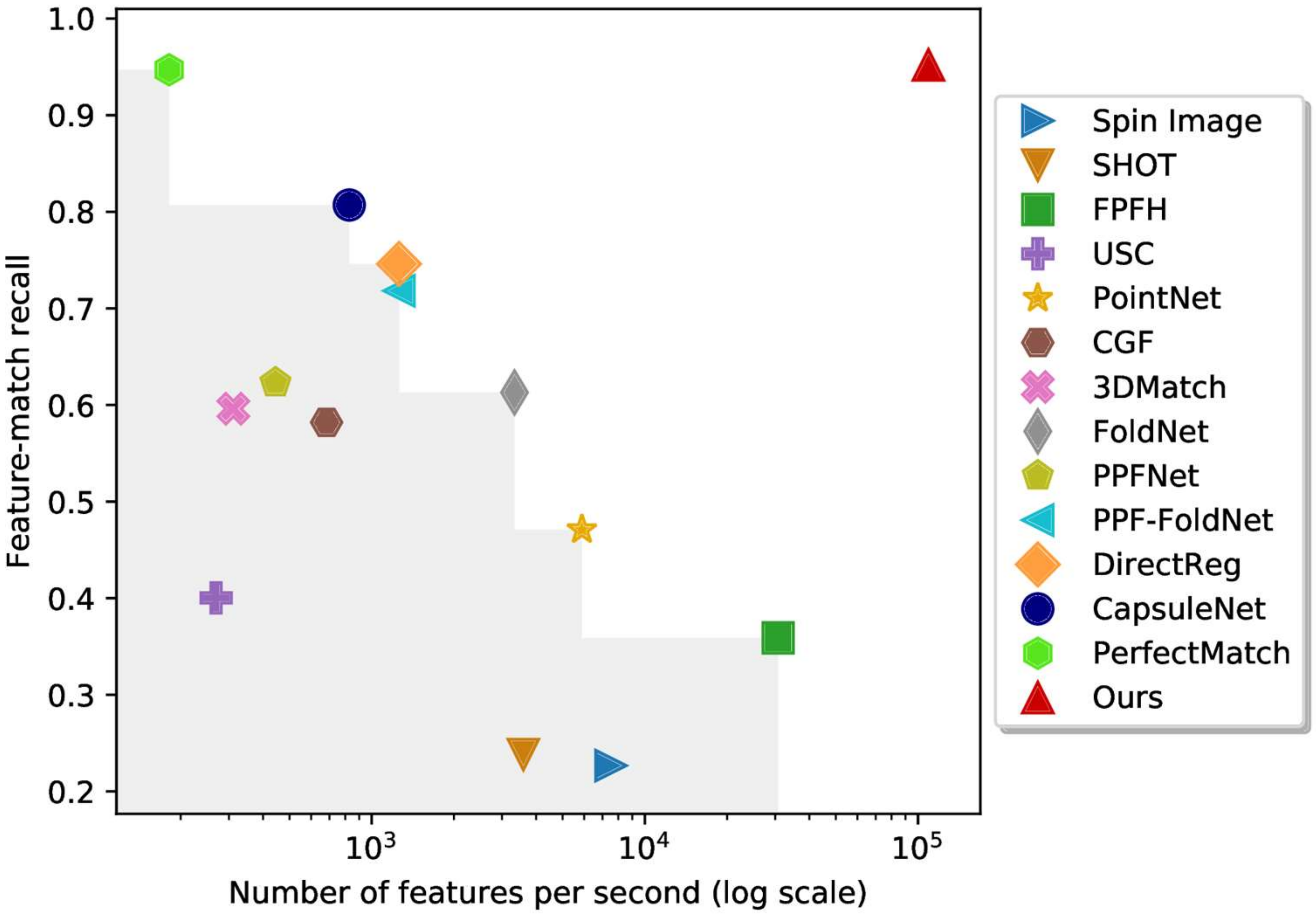
FULLY CONVOLUTIONAL HARDEST CONTRASTIVE LOSS



	Feature Match Recall	STD
Contrastive (norm.)	0.8493	0.0489
Triplet	0.7903	0.0494
Triplet (norm.)	0.6935	0.0446
Hardest-Contrastive	**0.9344**	0.0365

FULLY CONVOLUTIONAL GEOMETRIC FEATURES

Registration Results on the 3D Match Benchmark



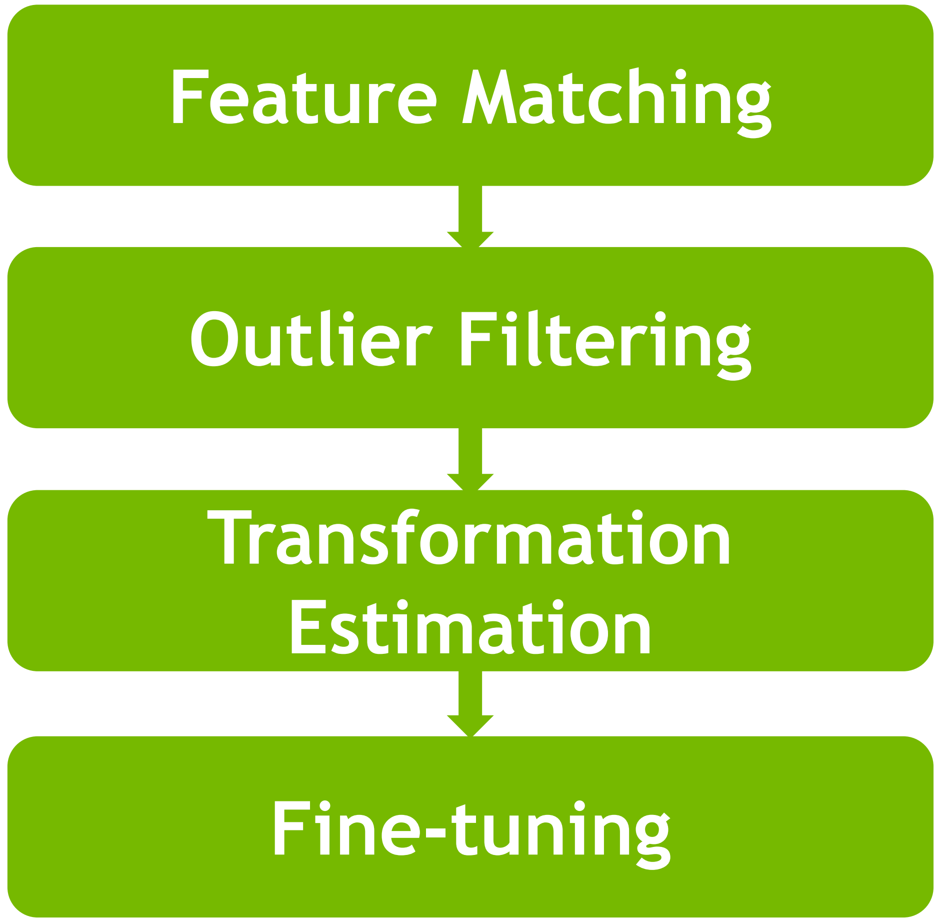
Chris Choy, Jaesik Park, Vladlen Koltun, **Fully Convolutional Geometric Features**, ICCV’19

3D GLOBAL REGISTRATION

Choy et al., **Deep Global Registration**, CVPR'20 Oral
Choy et al., **High-dimensional Convolutional Networks for Geometric Pattern Recognition**, CVPR'20 Oral

MULTI-VIEW 3D RECONSTRUCTION

Pipelines when no camera extrinsics are given

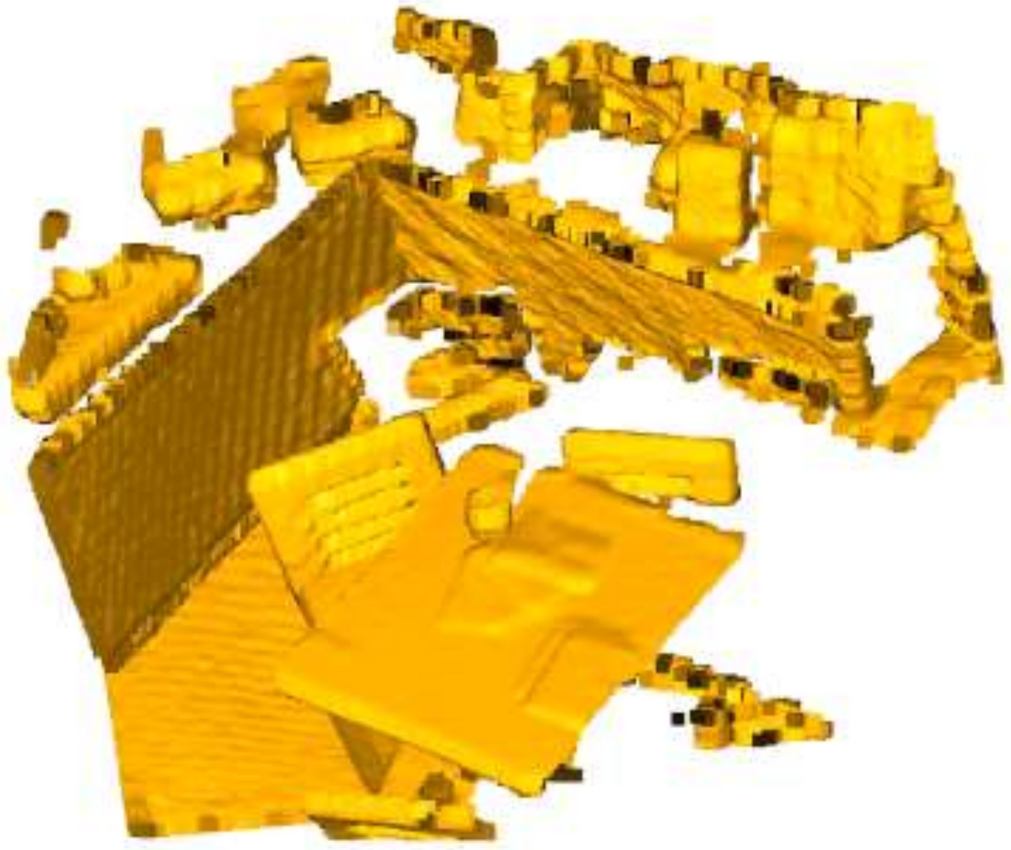


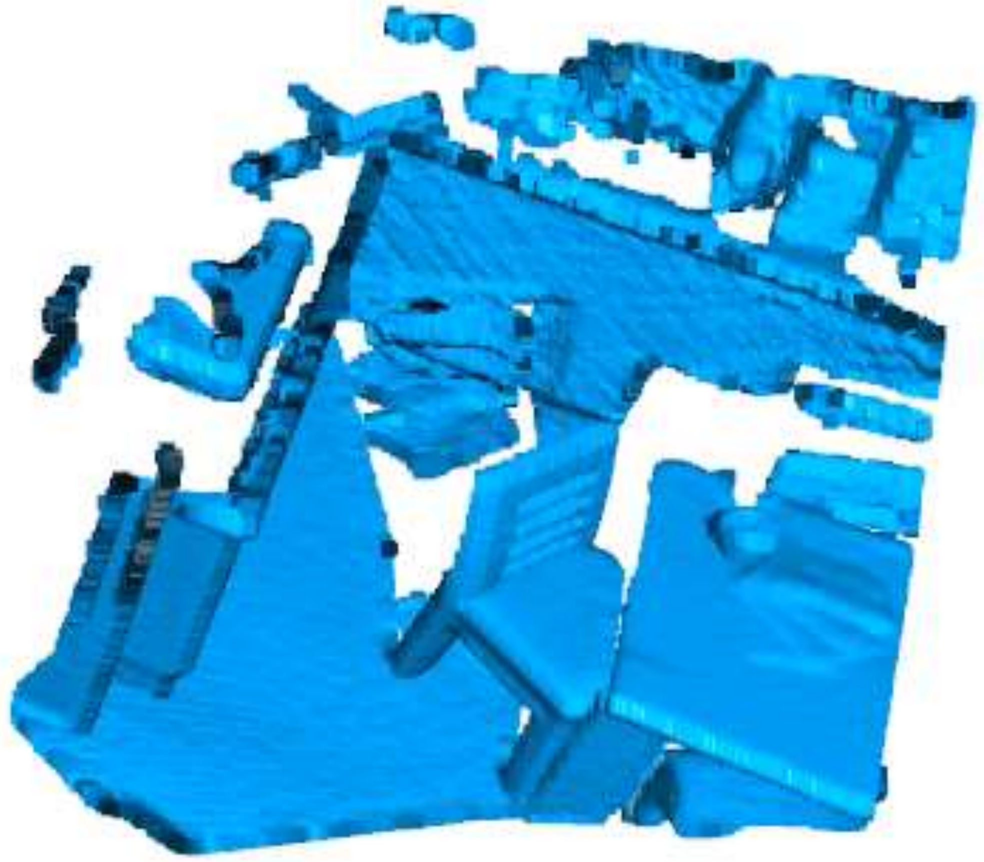
Feature Matching

Outlier Filtering

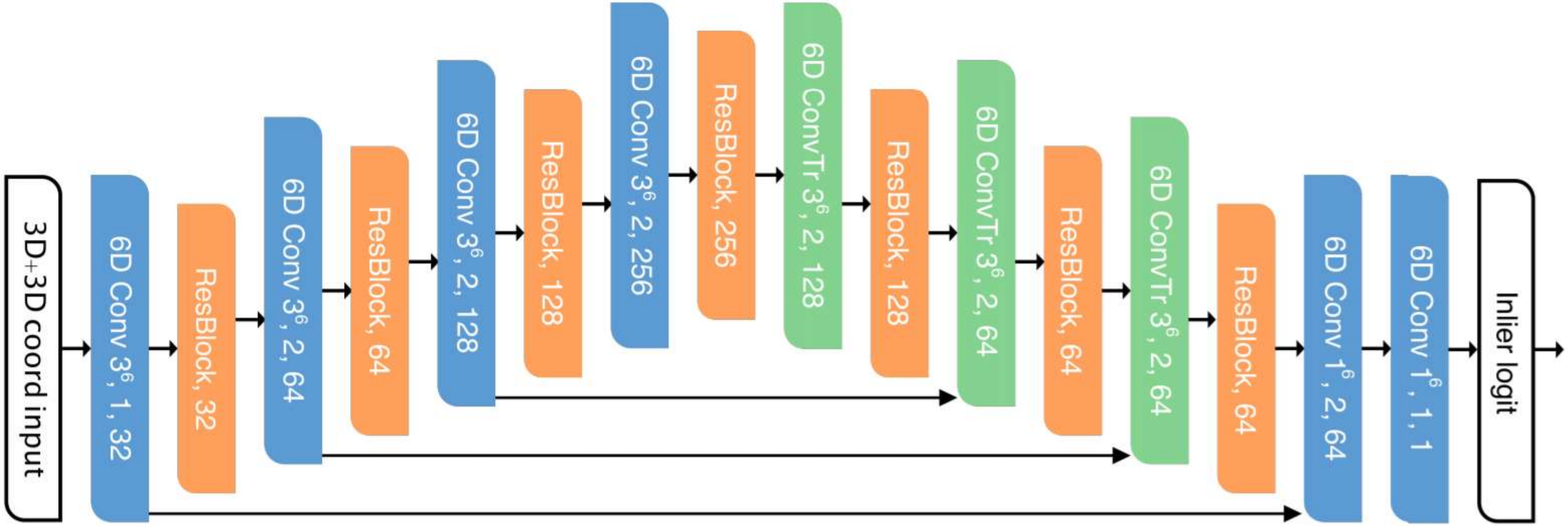
Transformation Estimation

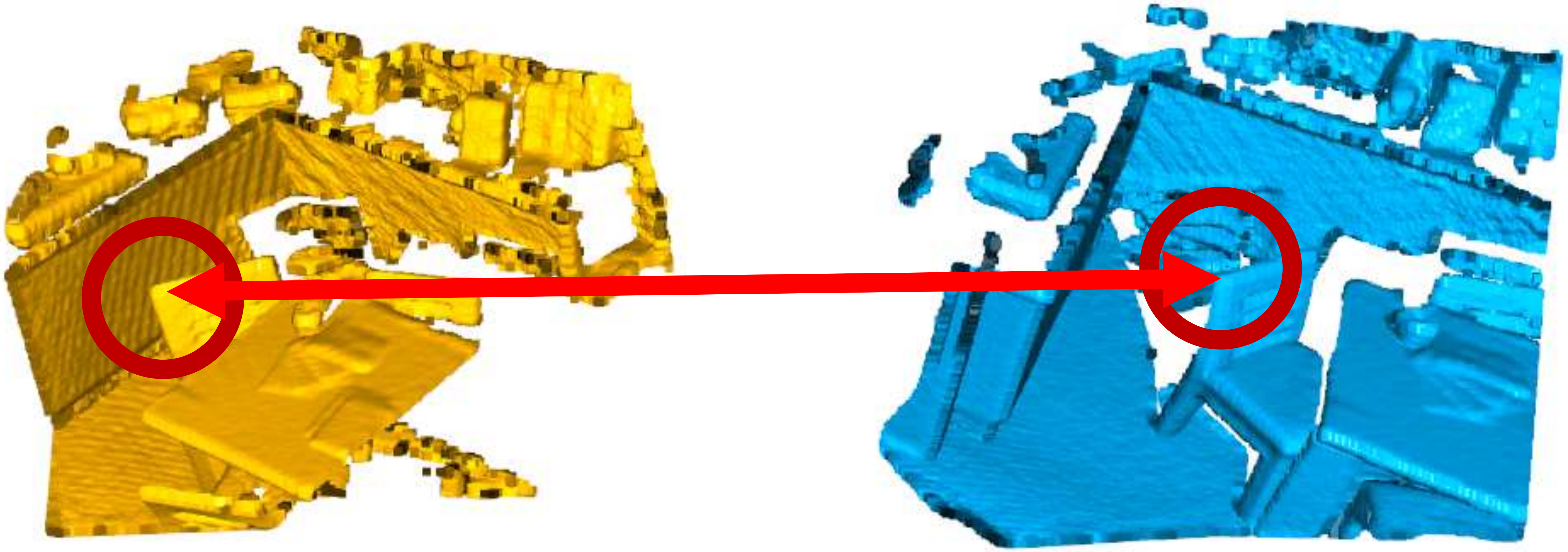
Fine-tuning



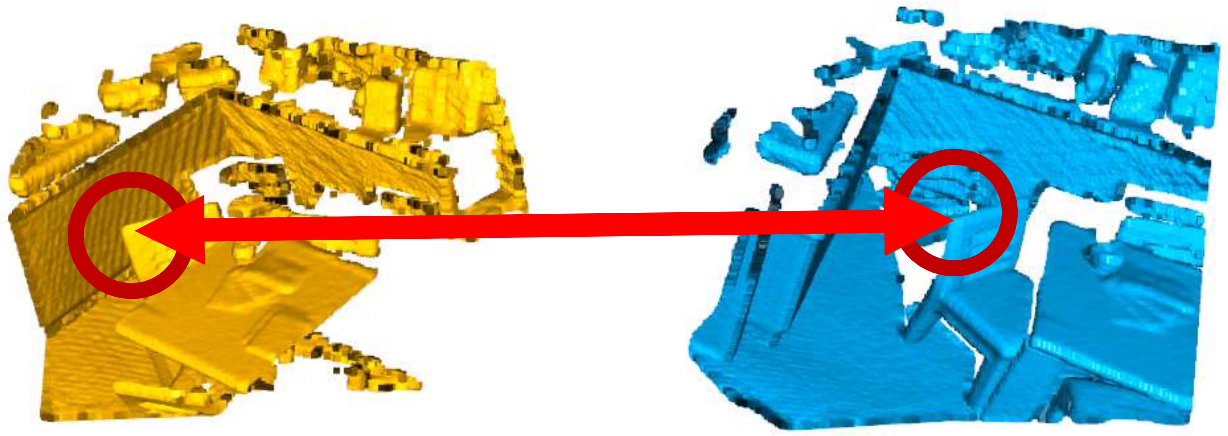


6D CONVOLUTIONAL NETWORK



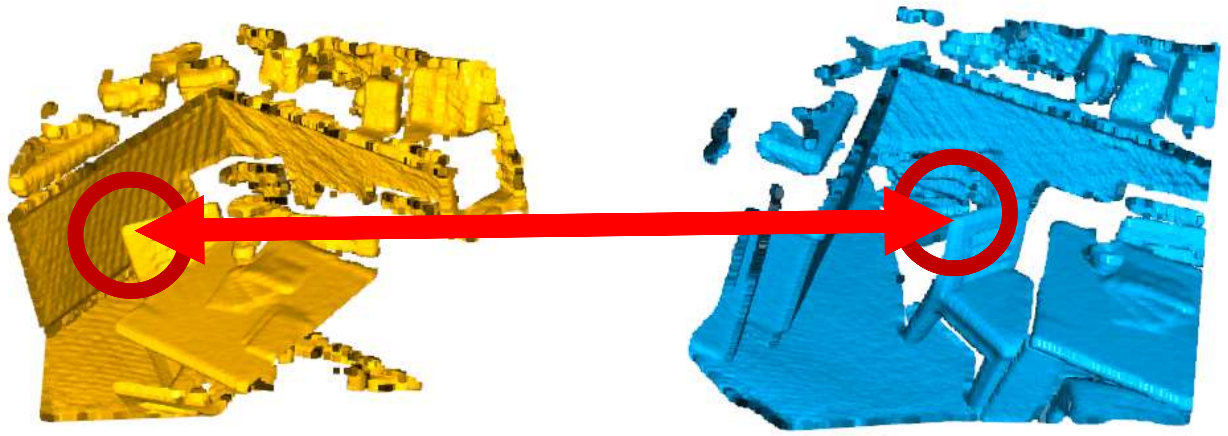


(x_1, y_1, z_1) (x_2, y_2, z_2)



(x_1, y_1, z_1) (x_2, y_2, z_2)

$$R \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} + \mathbf{t} - \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix} \approx \mathbf{0}$$



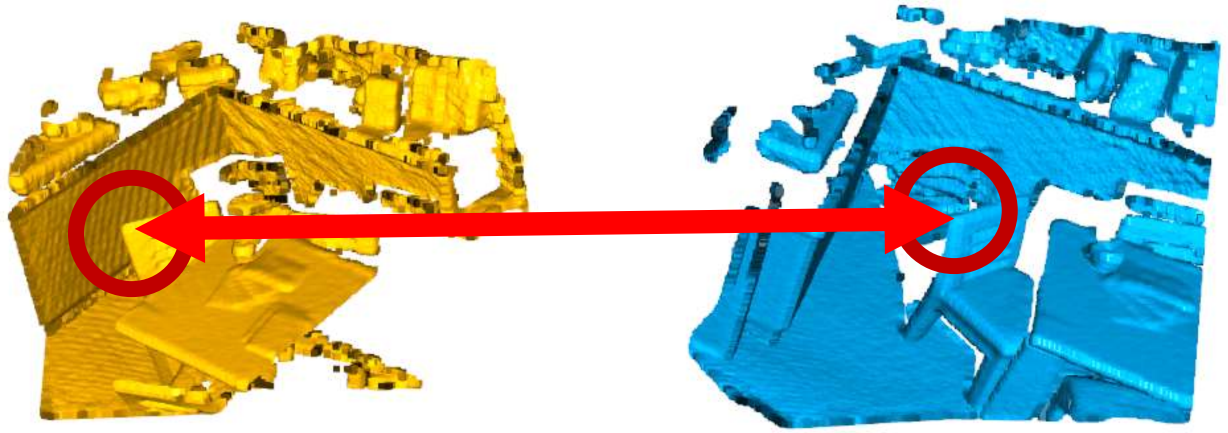
(x_1, y_1, z_1) (x_2, y_2, z_2)

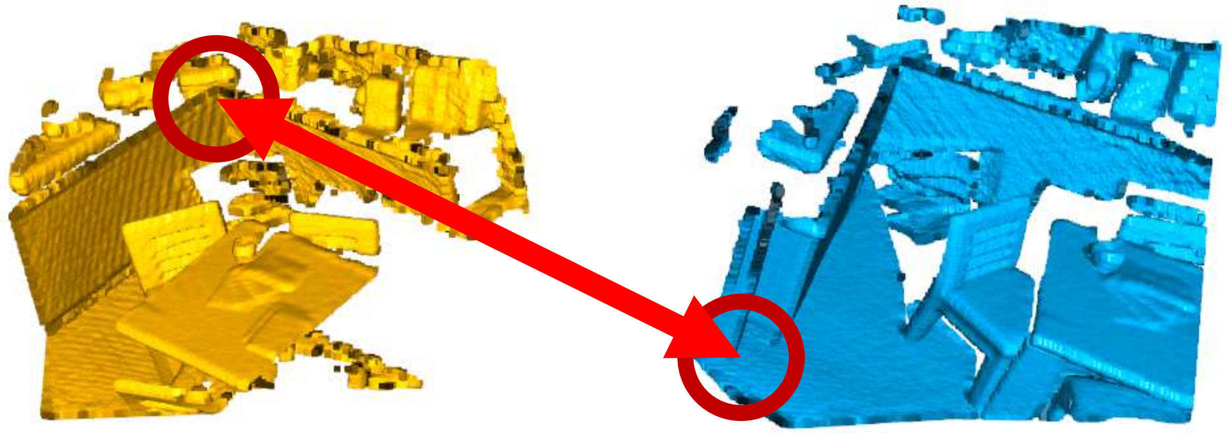
$$\begin{bmatrix} R & -I \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ x_2 \\ y_2 \\ z_2 \end{bmatrix} + \mathbf{t} \approx \mathbf{0}$$

$$\begin{bmatrix} r_{00} & r_{01} & r_{02} & -1 & 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ x_2 \\ y_2 \\ z_2 \end{bmatrix} + t_x = 0 \quad \begin{bmatrix} r_{10} & r_{11} & r_{12} & 0 & -1 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ x_2 \\ y_2 \\ z_2 \end{bmatrix} + t_y = 0 \quad \begin{bmatrix} r_{20} & r_{21} & r_{22} & 0 & -1 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ x_2 \\ y_2 \\ z_2 \end{bmatrix} + t_z = 0$$

$$\begin{bmatrix} R & -I \end{bmatrix}\begin{bmatrix} x_1 \\ y_1 \\ z_1 \\ x_2 \\ y_2 \\ z_2 \end{bmatrix} + \mathbf{t} \approx \mathbf{0}$$

Intersection of three hyperplanes





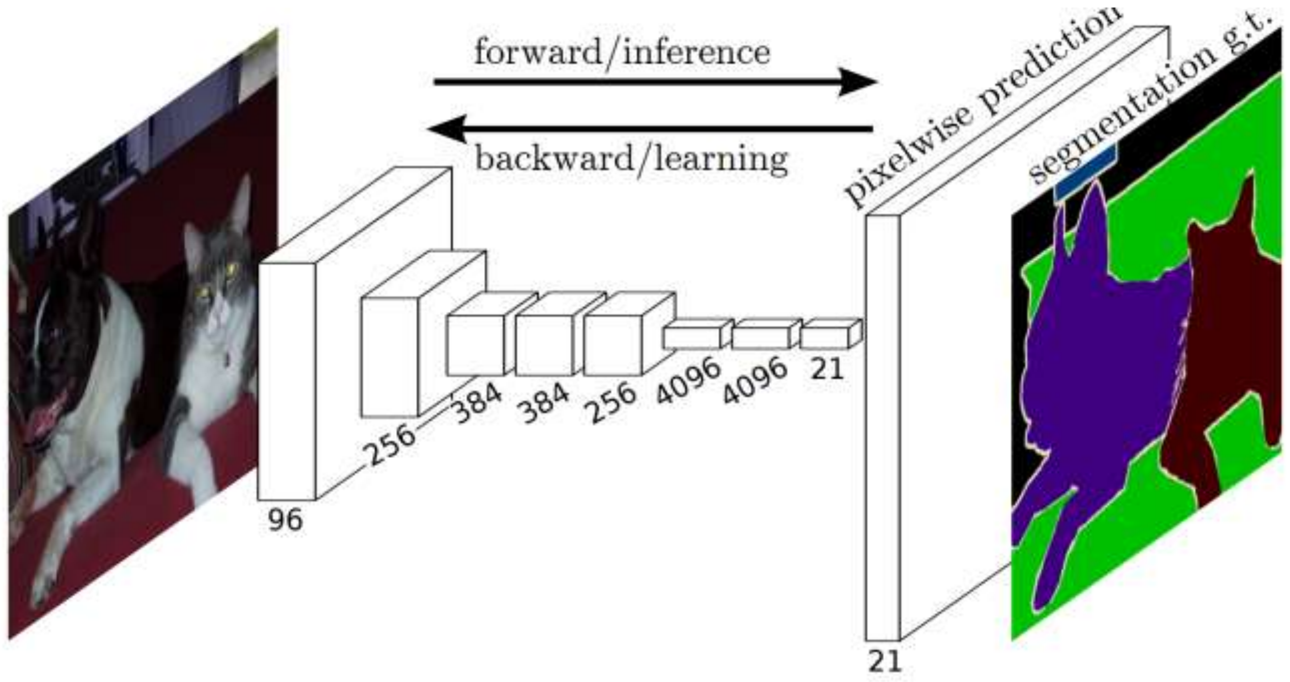
$$R \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} + \mathbf{t} \approx \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix}$$

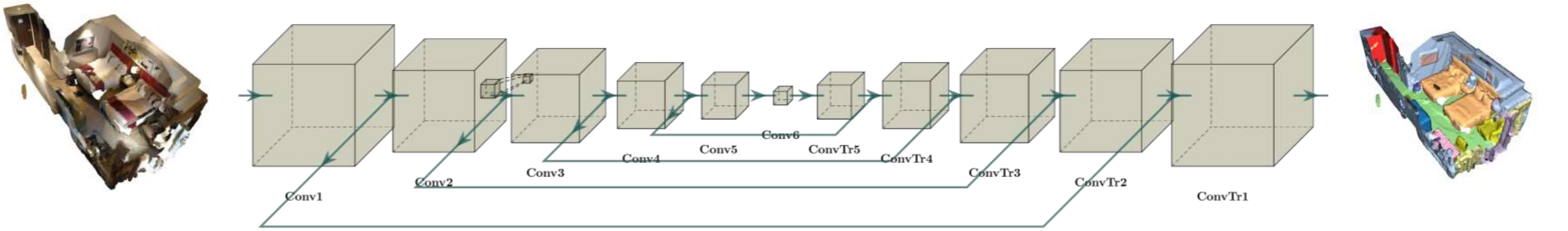
$$R \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} + \mathbf{t} \neq \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix}$$

Inliers: 3D subspace in 6D

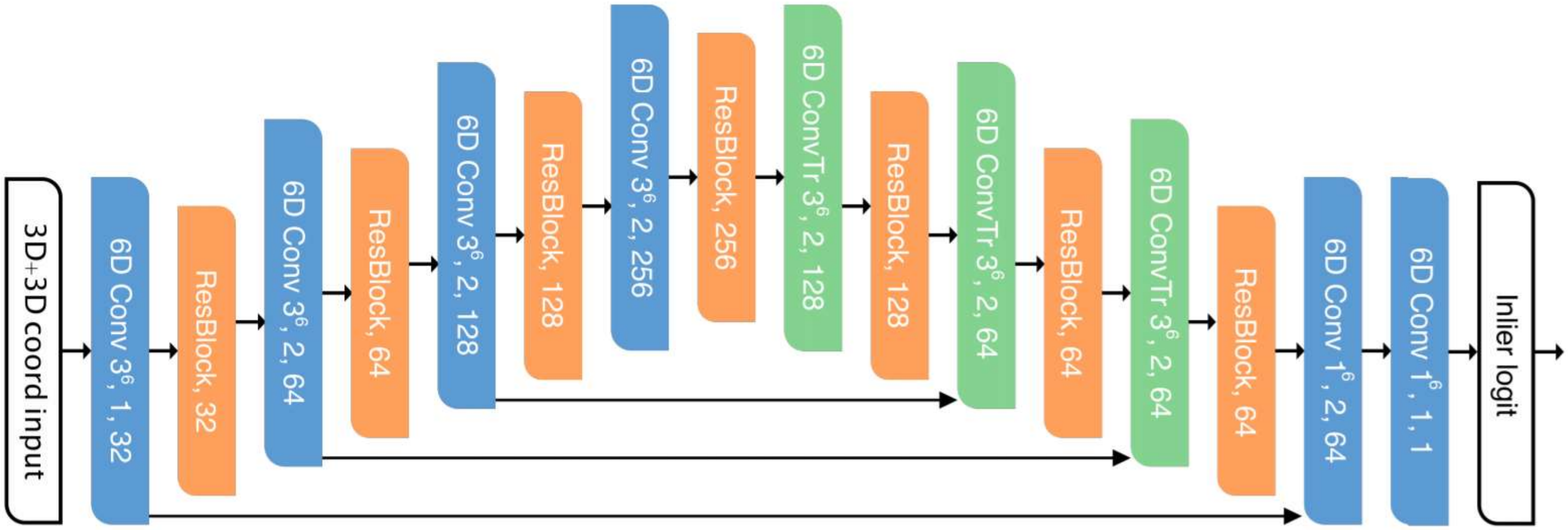
Outliers: Noise

Finding Inlier = Segmentation

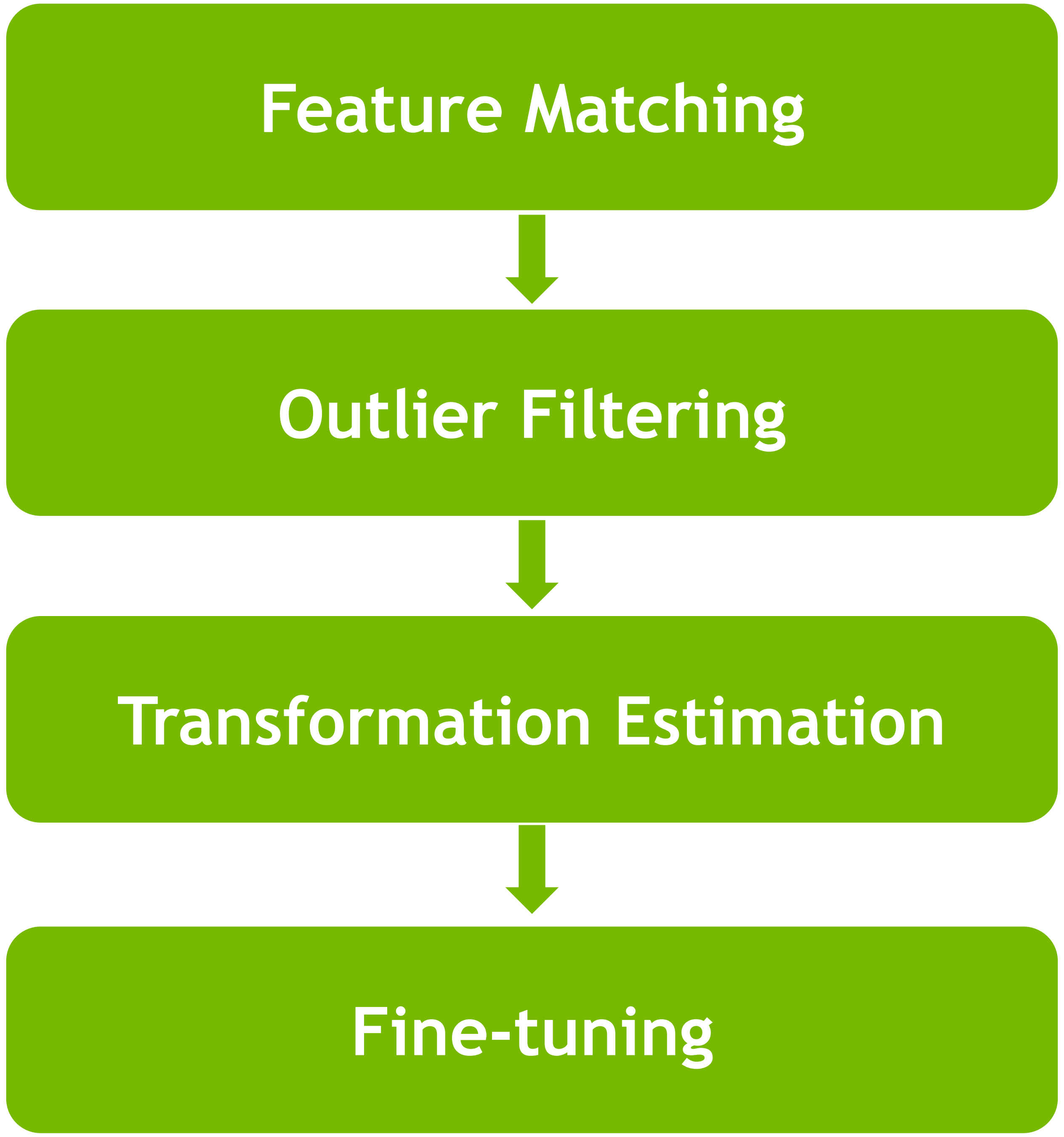

forward/inference
backward/learning
pixelwise prediction
segmentation g.t.
96
256
384
384
256
4096
4096
21
21


Conv1
Conv2
Conv3
Conv4
Conv5
Conv6
ConvTr5
ConvTr4
ConvTr3
ConvTr2
ConvTr1

6D CONVOLUTIONAL NETWORK



REGISTRATION PIPELINE



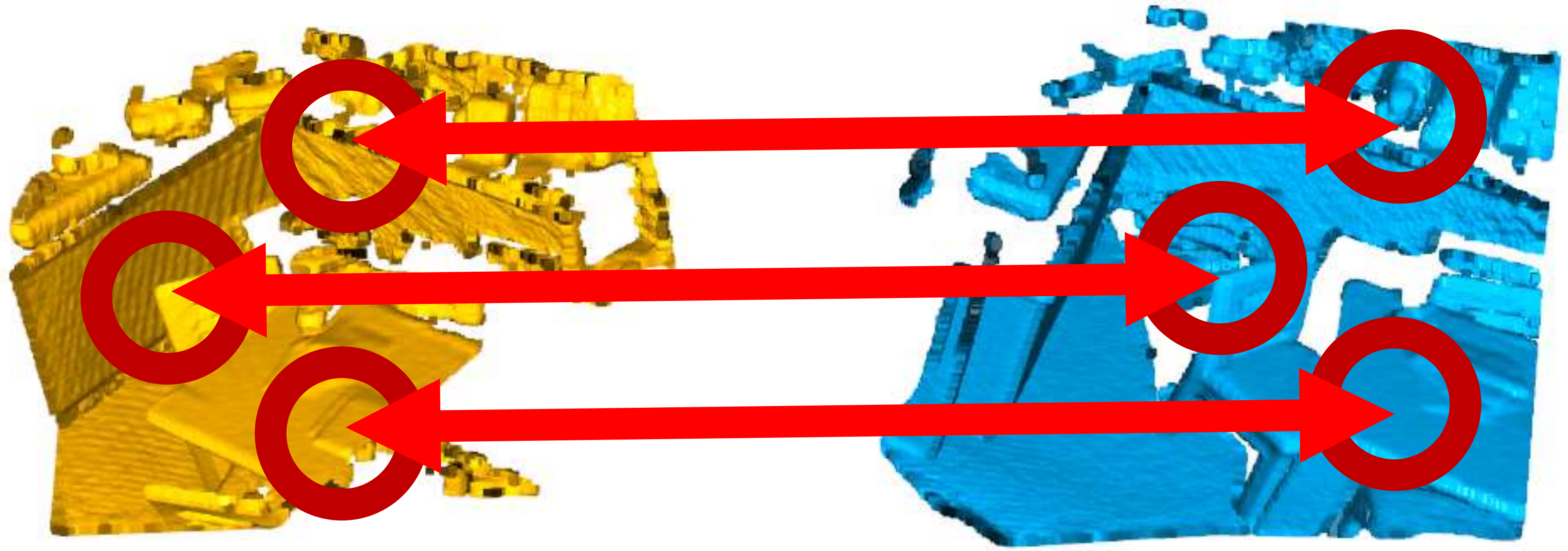
Feature Matching

Outlier Filtering

Transformation Estimation

Fine-tuning

Transformation Estimation

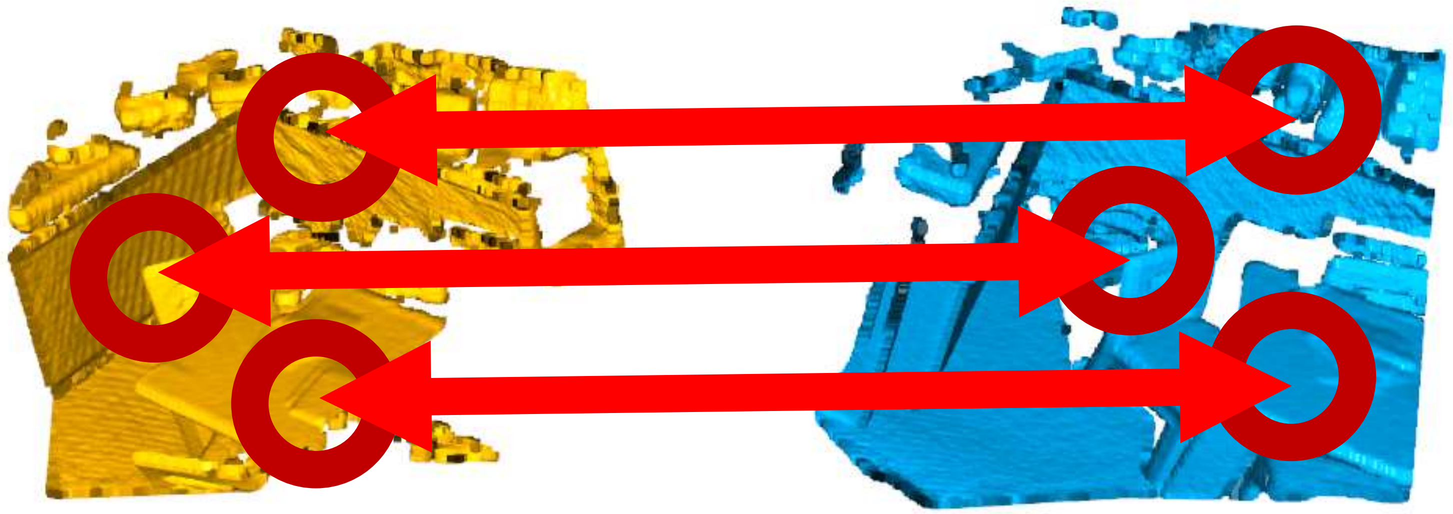


(x_1, y_1, z_1) (x_2, y_2, z_2)

$$\underset{R,\mathbf{t}}{\operatorname{argmin}} \sum_i \left(R \begin{bmatrix} x_1^i \\ y_1^i \\ z_1^i \end{bmatrix} + \mathbf{t} - \begin{bmatrix} x_2^i \\ y_2^i \\ z_2^i \end{bmatrix} \right)^2$$

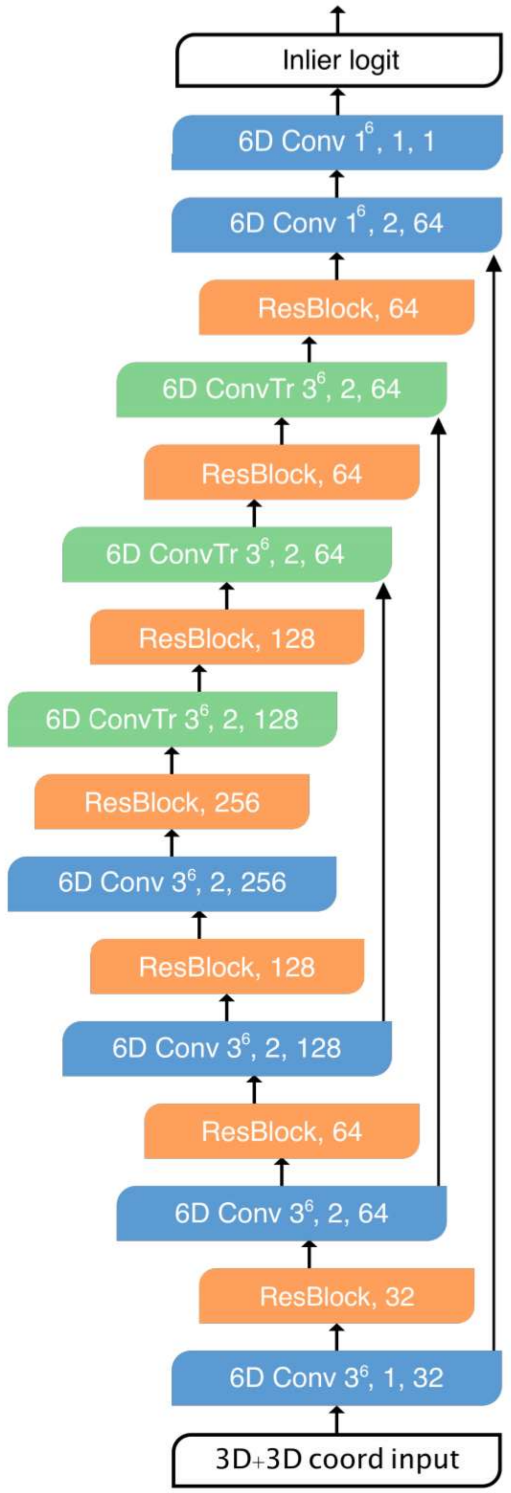
Procrustes Analysis

Transformation Estimation

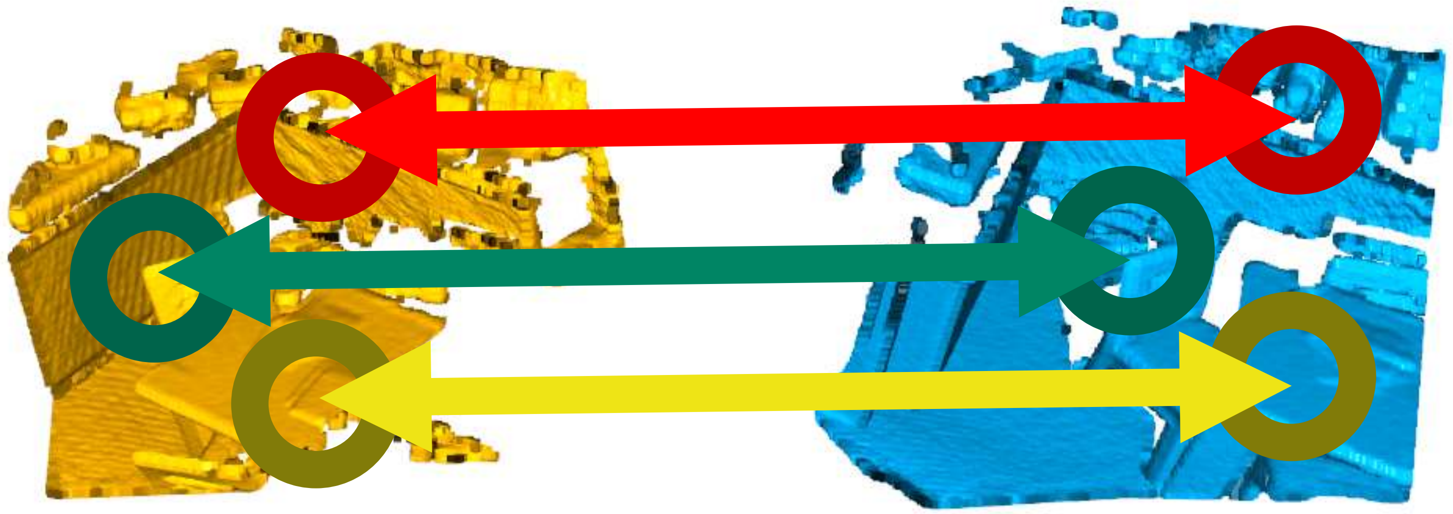


(x_1, y_1, z_1) (x_2, y_2, z_2)

Transformation Estimation



$$w_i = p\left(\begin{bmatrix} x_1^i \\ y_1^i \\ z_1^i \end{bmatrix}, \begin{bmatrix} x_2^i \\ y_2^i \\ z_2^i \end{bmatrix}\right)$$

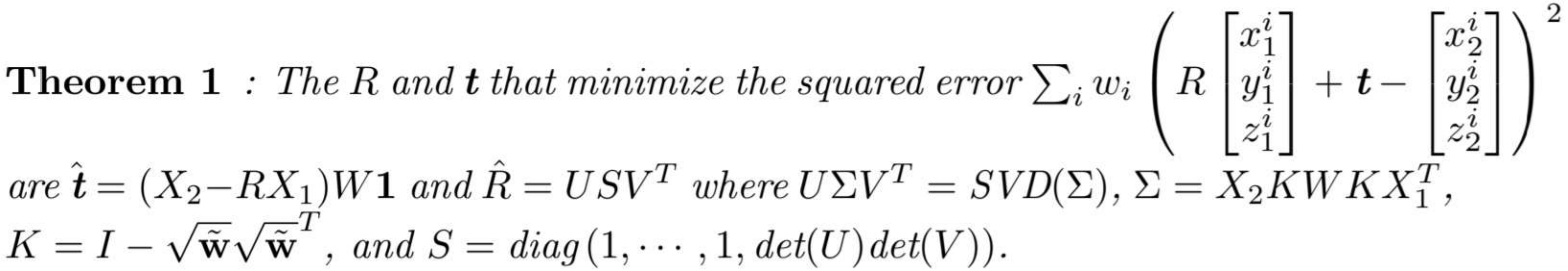


(x_1, y_1, z_1) (x_2, y_2, z_2)

$$\operatorname*{arg\,min}_{R,\mathbf{t}} \sum_i w_i \left(R \begin{bmatrix} x_1^i \\ y_1^i \\ z_1^i \end{bmatrix} + \mathbf{t} - \begin{bmatrix} x_2^i \\ y_2^i \\ z_2^i \end{bmatrix} \right)^2$$

Transformation Estimation

$$\underset{R,\mathbf{t}}{\operatorname{argmin}} \sum_i w_i \left(R \begin{bmatrix} x_1^i \\ y_1^i \\ z_1^i \end{bmatrix} + \mathbf{t} - \begin{bmatrix} x_2^i \\ y_2^i \\ z_2^i \end{bmatrix} \right)^2$$



Theorem 1 : *The R and $\mathbf{t}$ that minimize the squared error* $\sum_i w_i \left(R \begin{bmatrix} x_1^i \\ y_1^i \\ z_1^i \end{bmatrix} + \mathbf{t} - \begin{bmatrix} x_2^i \\ y_2^i \\ z_2^i \end{bmatrix} \right)^2$ *are* $\hat{\mathbf{t}} = (X_2 - RX_1)W\mathbf{1}$ *and* $\hat{R} = USV^T$ *where* $U\Sigma V^T = SVD(\Sigma)$, $\Sigma = X_2KWKX_1^T$, $K = I - \sqrt{\tilde{\mathbf{w}}}\sqrt{\tilde{\mathbf{w}}}^T$, *and* $S = diag(1, \cdots, 1, det(U)det(V))$.

$$\hat{R} = USV^T$$

$$\hat{\mathbf{t}} = (X_2 - \hat{R}X_1)W\mathbf{1}$$

Transformation Estimation

$$\hat{R} = USV^T \qquad \hat{\mathbf{t}} = (X_2 - \hat{R}X_1)W\mathbf{1}$$

Weighted Procrustes

Differentiable w.r.t W, inlier probability

Transformation Estimation

$$\hat{R} = USV^T \qquad \hat{\mathbf{t}} = (X_2 - \hat{R}X_1)W\mathbf{1}$$

Weighted Procrustes

Differentiable w.r.t W, inlier probability

1. **Complexity linear to num. correspondences**
   - High-resolution correspondences
2. **Scans with partial overlap**
   - No 1-1 mapping since weight can be 0

Transformation Estimation

$$\hat{R} = USV^T \qquad \hat{\mathbf{t}} = (X_2 - \hat{R}X_1)W\mathbf{1}$$

Weighted Procrustes

Differentiable w.r.t W, inlier probability

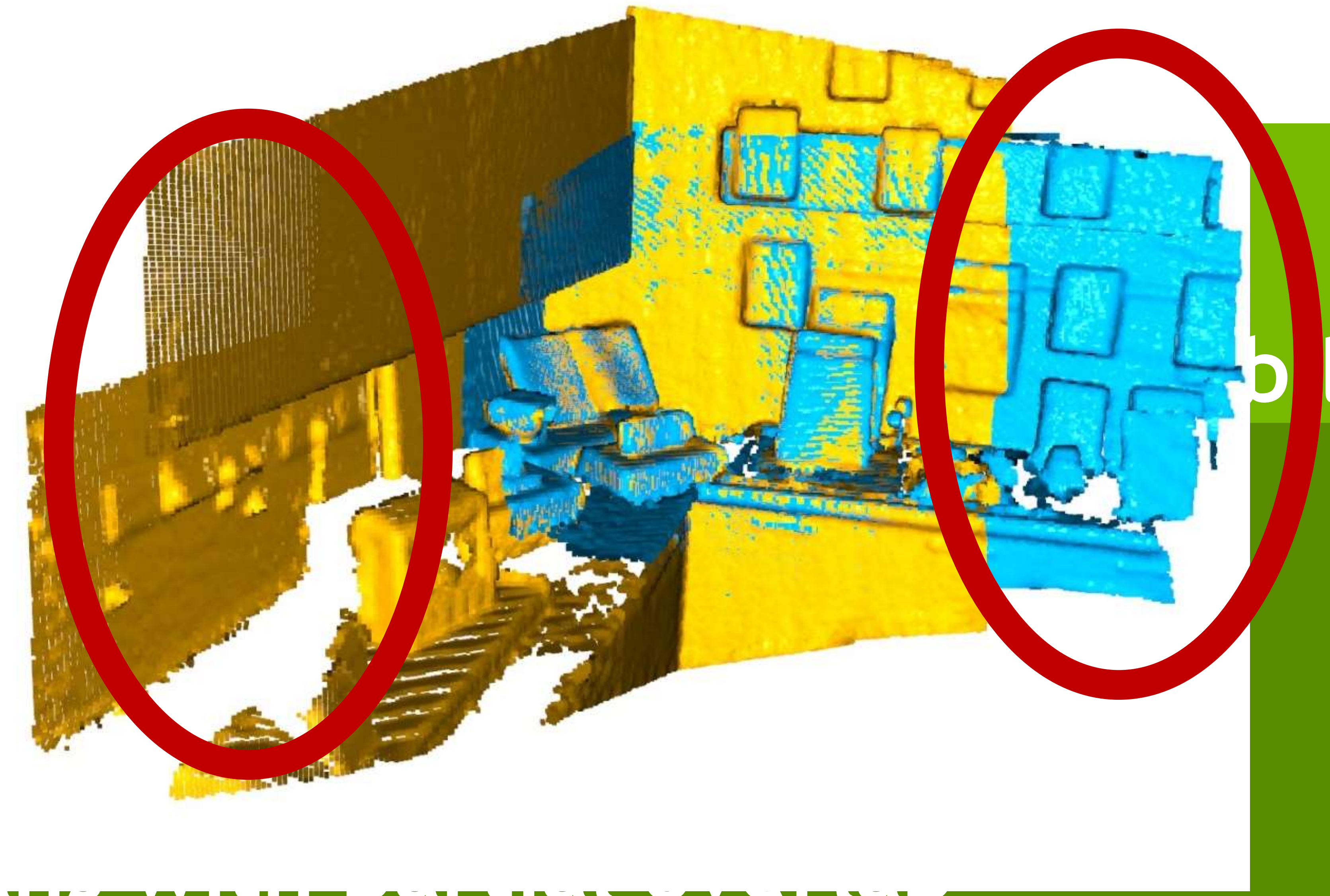
1. **Complexity linear to num. correspondences**
   - High-resolution correspondences
2. **Scans with partial overlap**
   - No 1-1 mapping since weight can be 0

Transformation Estimation

$$\hat{R} = USV^T \qquad \hat{\mathbf{t}} = (X_2 - \hat{R}X_1)W\mathbf{1}$$

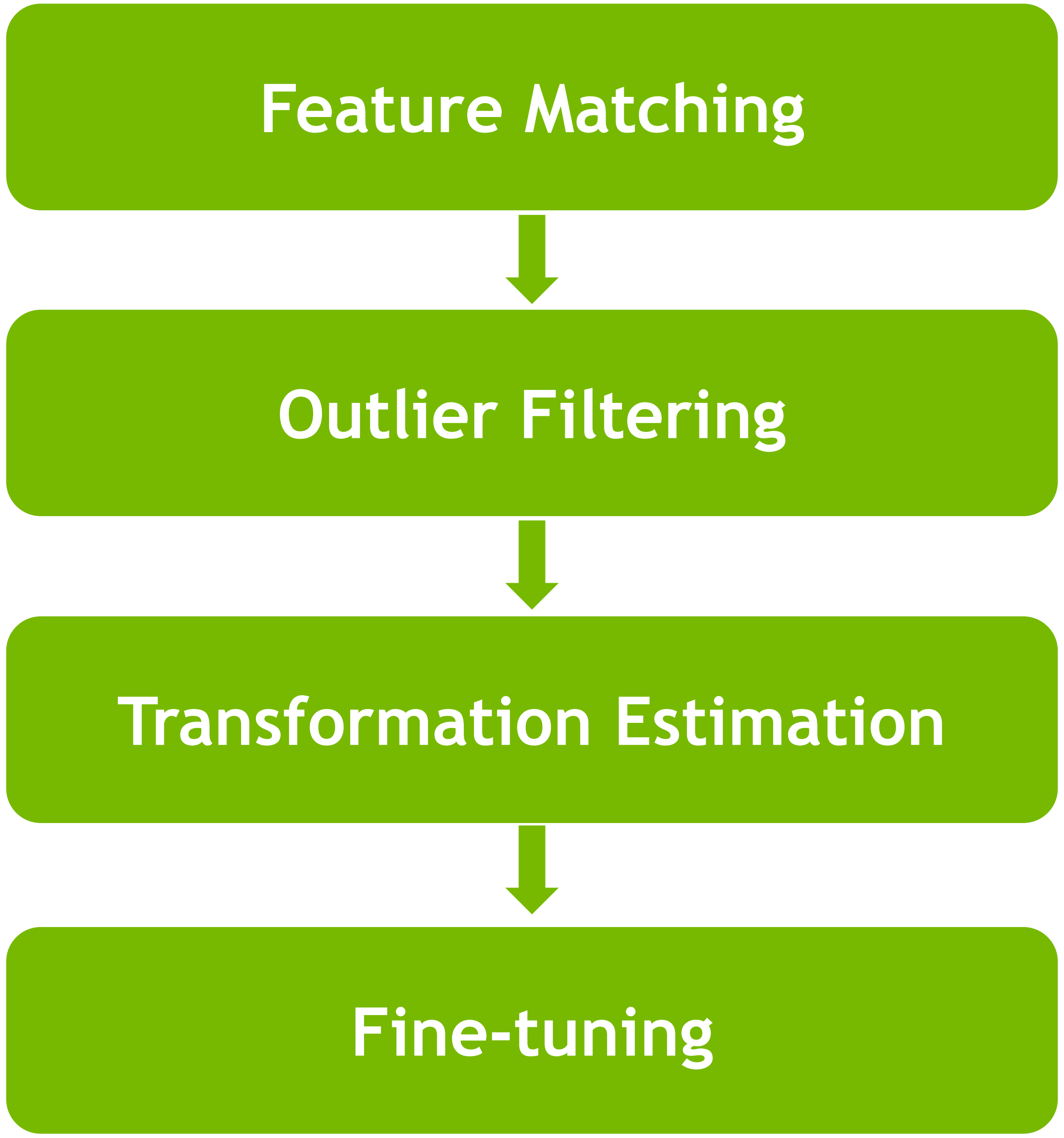
Diffe bility

1. Co co
   -



2. Scans with partial overlap
   - No 1-1 mapping since weight can be 0

REGISTRATION PIPELINE



Feature Matching

Outlier Filtering

Transformation Estimation

Fine-tuning

Fine-tuning

$$\underset{R,\mathbf{t}}{\operatorname{argmin}} \sum_i w_i L\left(R \begin{bmatrix} x_1^i \\ y_1^i \\ z_1^i \end{bmatrix} + \mathbf{t}, \begin{bmatrix} x_2^i \\ y_2^i \\ z_2^i \end{bmatrix} \right)$$

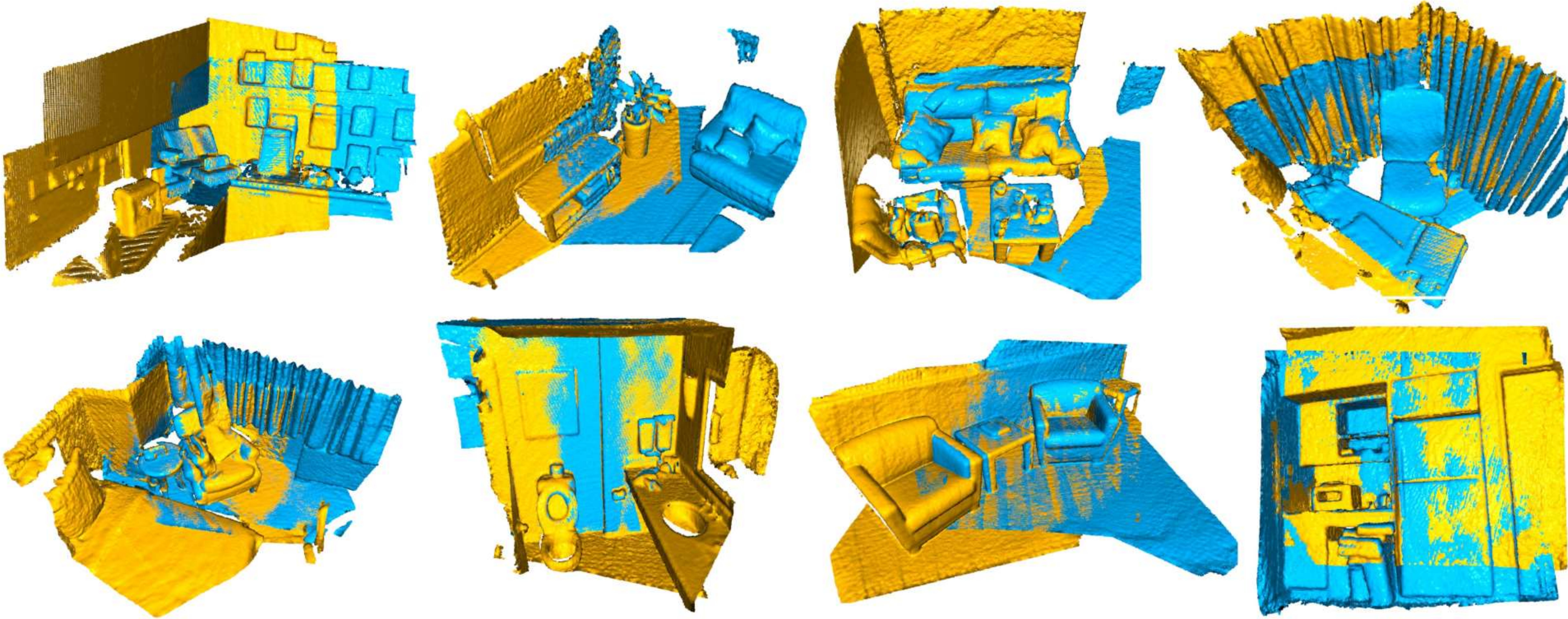
- Gradient-based optimization
- Continuous 6D representation [Zhou et al.]

$$f : \mathbb{R}^6 \rightarrow \mathrm{SO}(3)$$

$$\underset{\mathbf{a},\mathbf{t}}{\operatorname{argmin}} \sum_i w_i L\left(f(\mathbf{a}) \begin{bmatrix} x_1^i \\ y_1^i \\ z_1^i \end{bmatrix} + \mathbf{t}, \begin{bmatrix} x_2^i \\ y_2^i \\ z_2^i \end{bmatrix} \right)$$

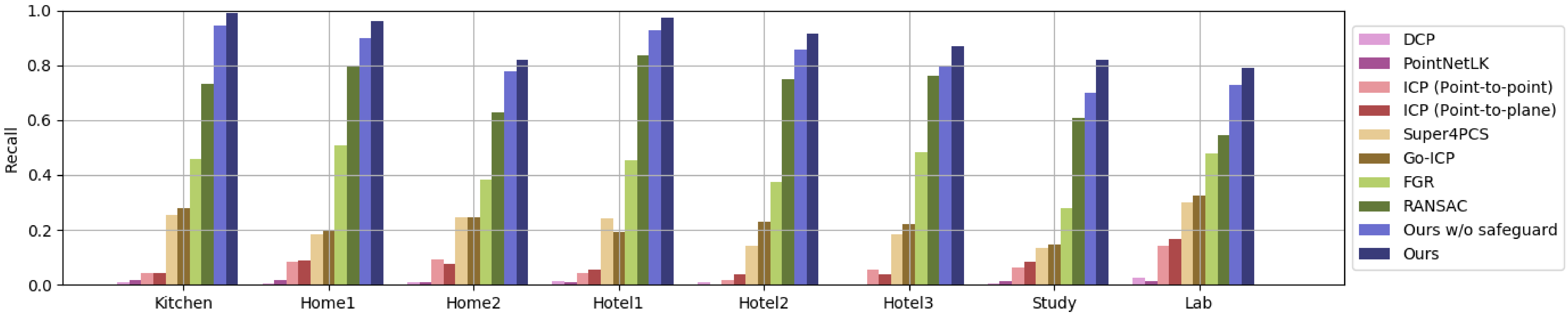
Zhou et al., **On the Continuity of Rotation Representations in Neural Networks**, CVPR’19

TRAINED AND TESTED ON 3DMATCH



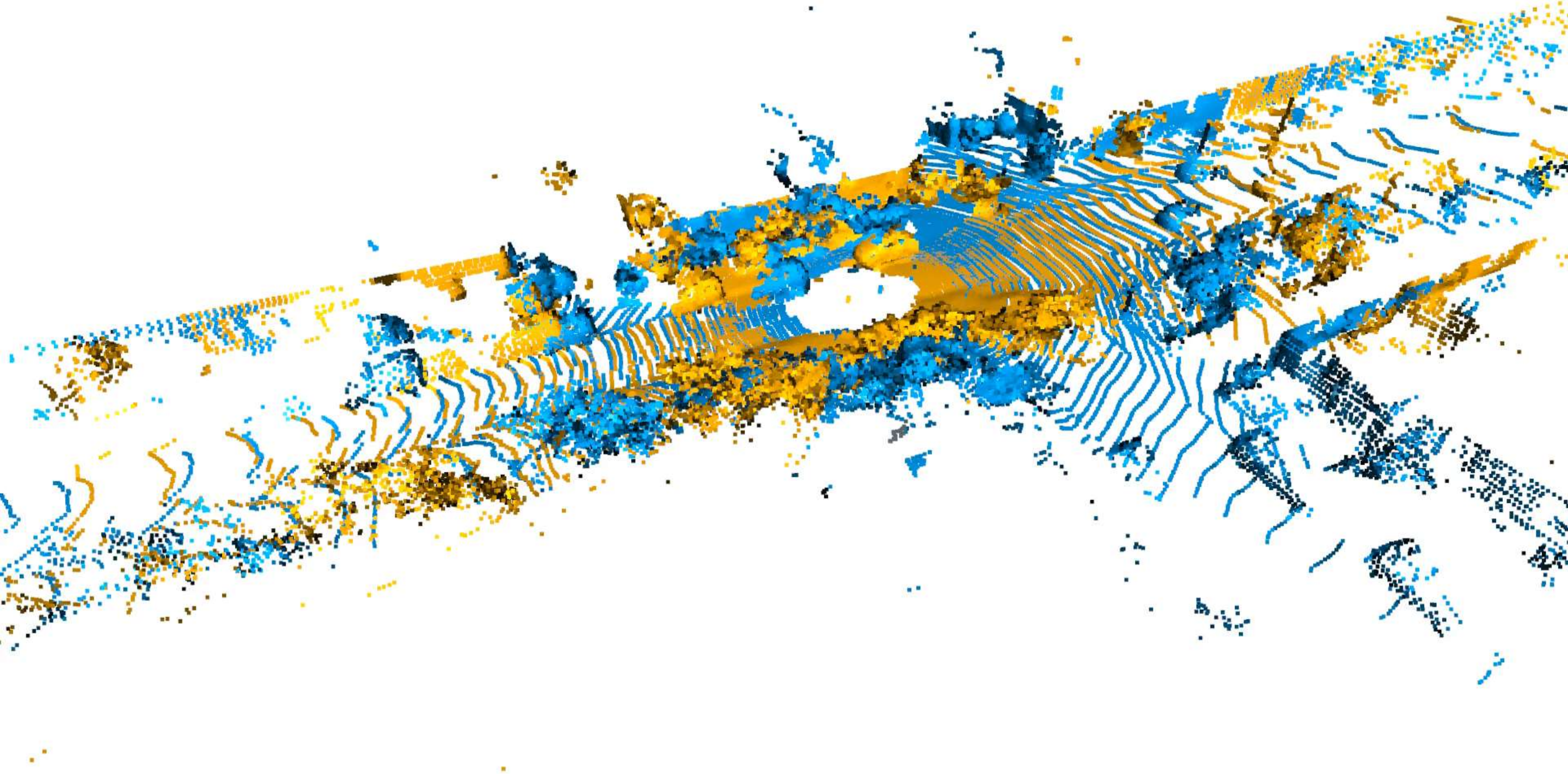
3D REGISTRATION: DEEP GLOBAL REGISTRATION

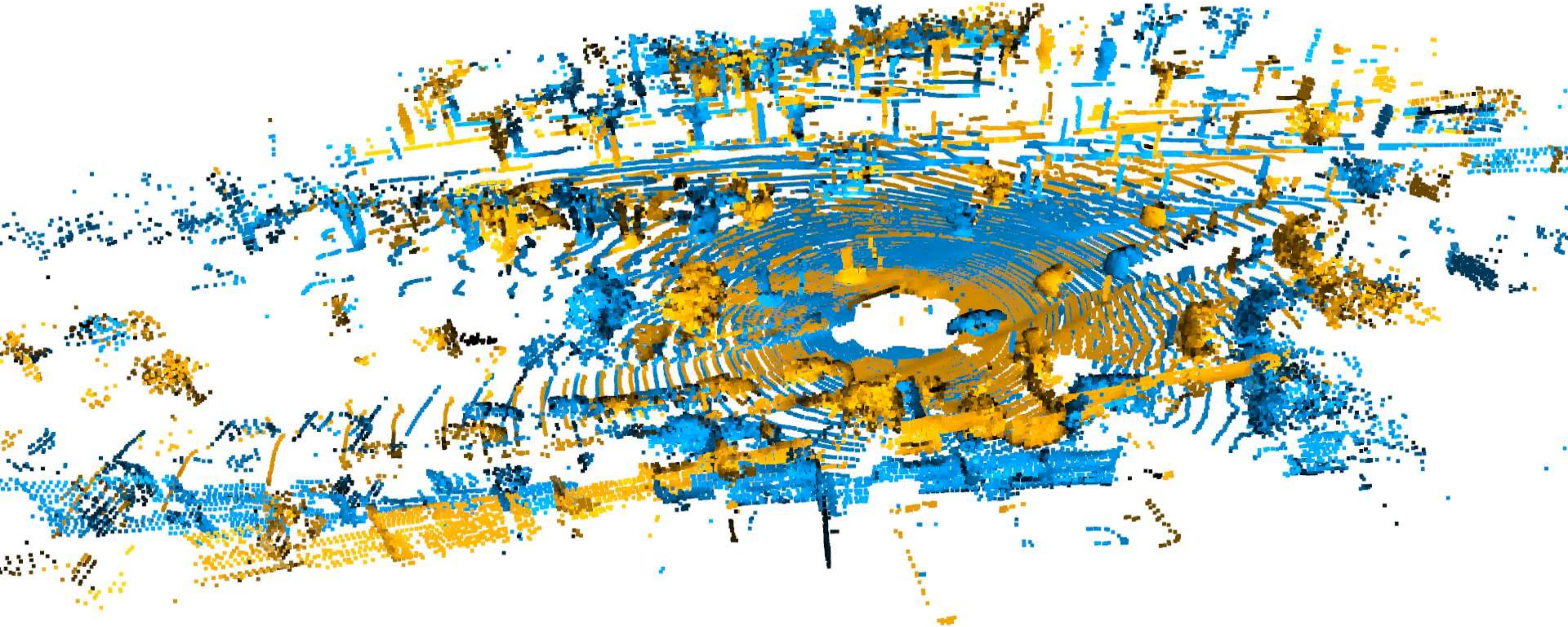
Learning the Structure of the correspondences

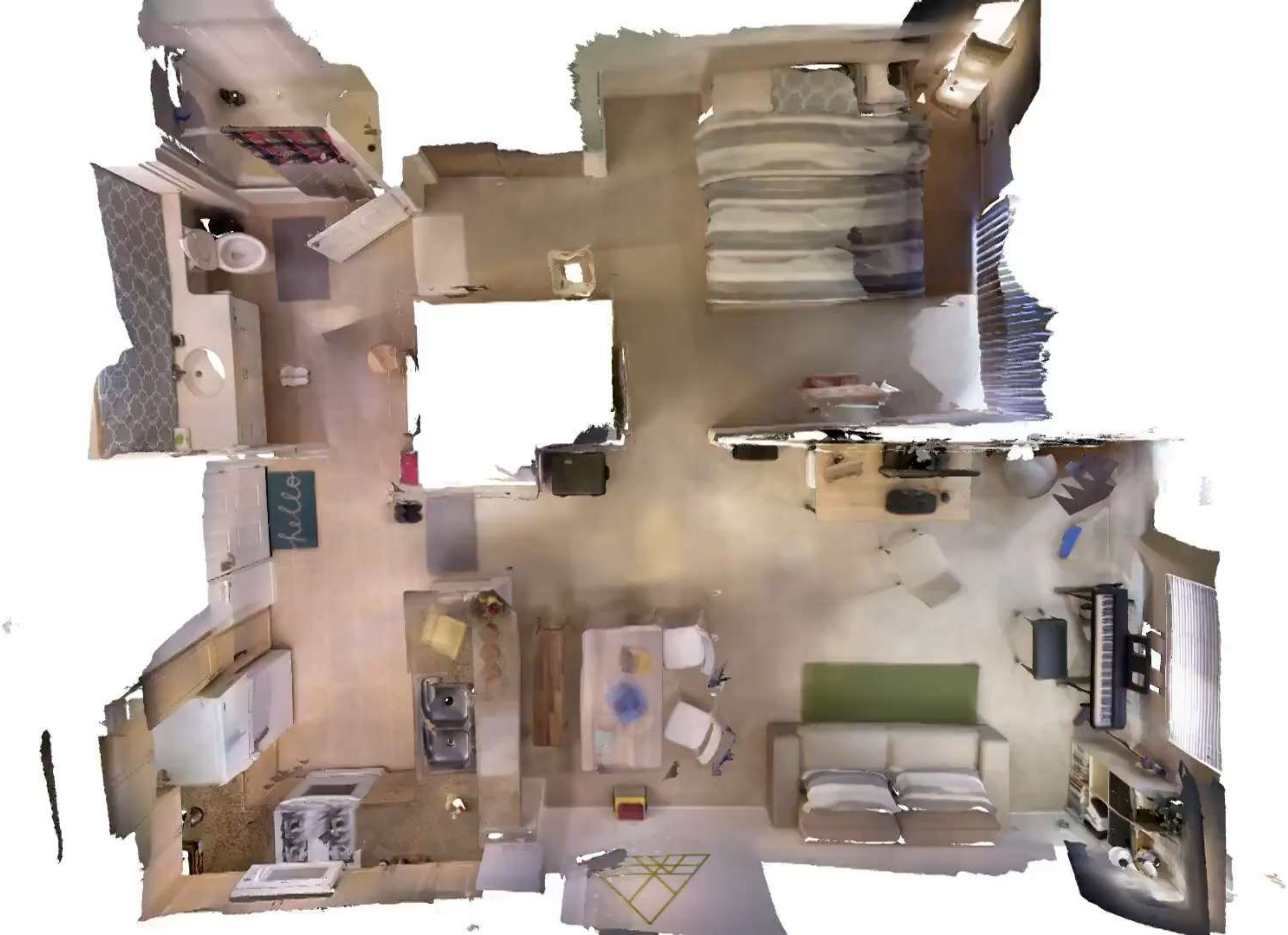


Chris Choy, JunYoung Gwak, Silvio Savarese, **4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks**, CVPR'19

Chris Choy, Wei Dong, Vladlen Koltun, **Deep Global Registration**, CVPR'20 Oral



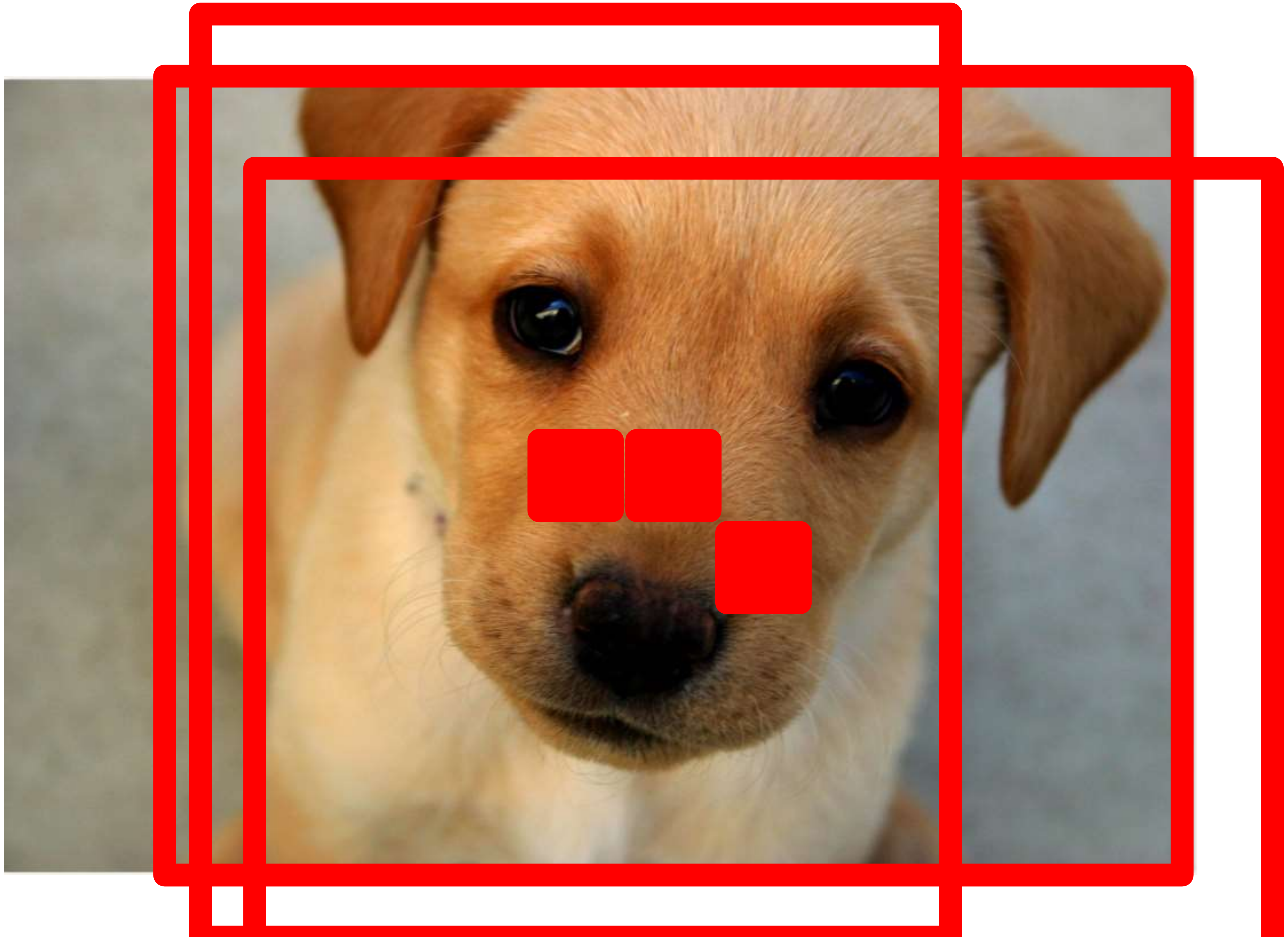



hello

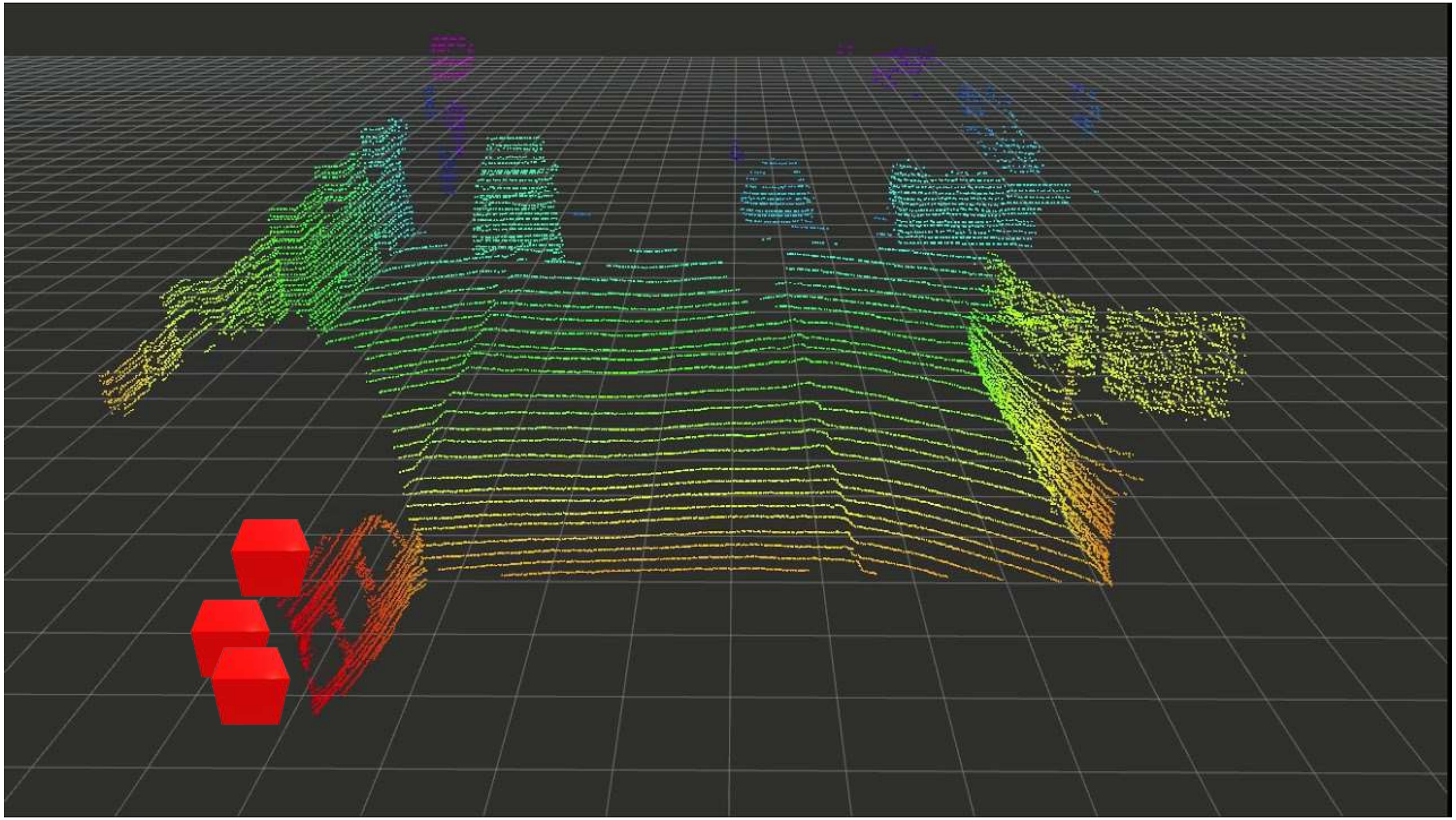
GENERATIVE SPARSE DETECTION NETWORKS

Gwak et al., **Generative Sparse Detection Networks for 3D Single-shot Object Detection**, preprint 2020

SINGLE SHOT OBJECT DETECTION: ANCHORS



3D SCANS

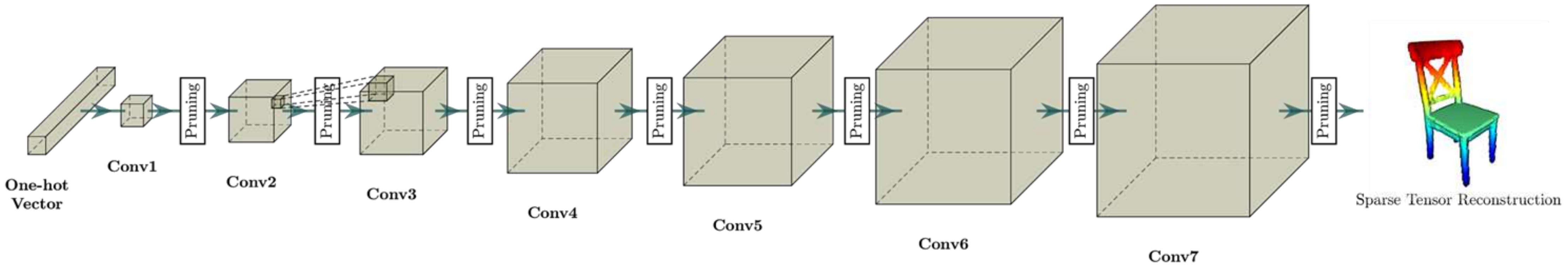


GENERATION NETWORKS

Generating Geometry / Sparsity Pattern

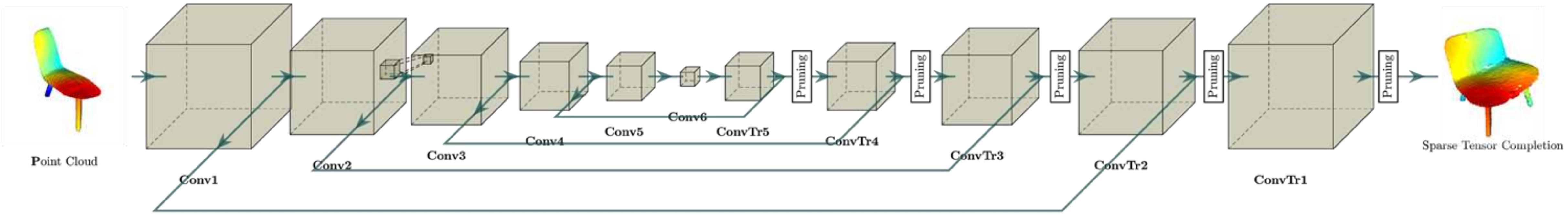
Full Reconstruction

Feature Vec. to 3D Object



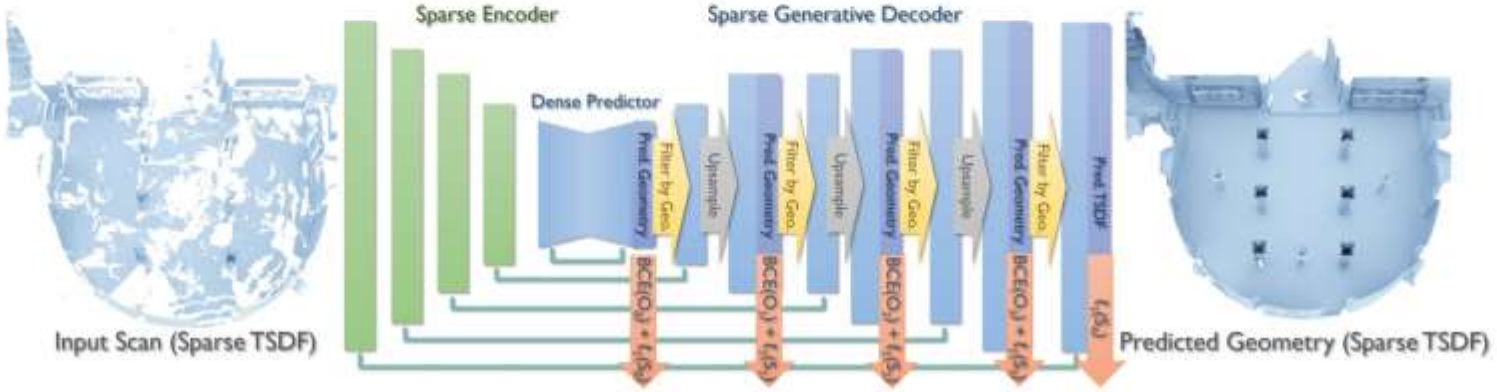
Completion

Partial 3D Object to Complete 3D Object



Scene Completion

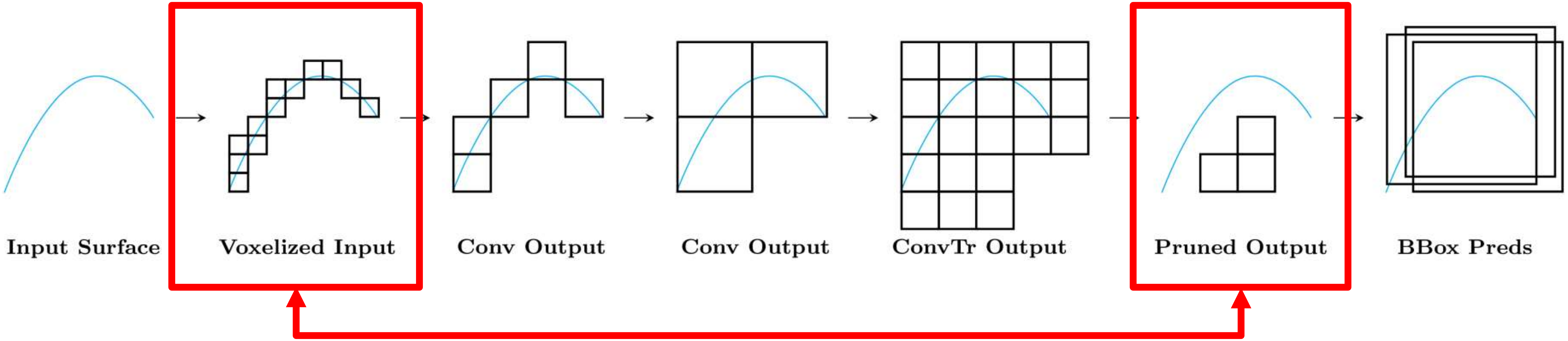
Dai et al. Sparse Generative NN



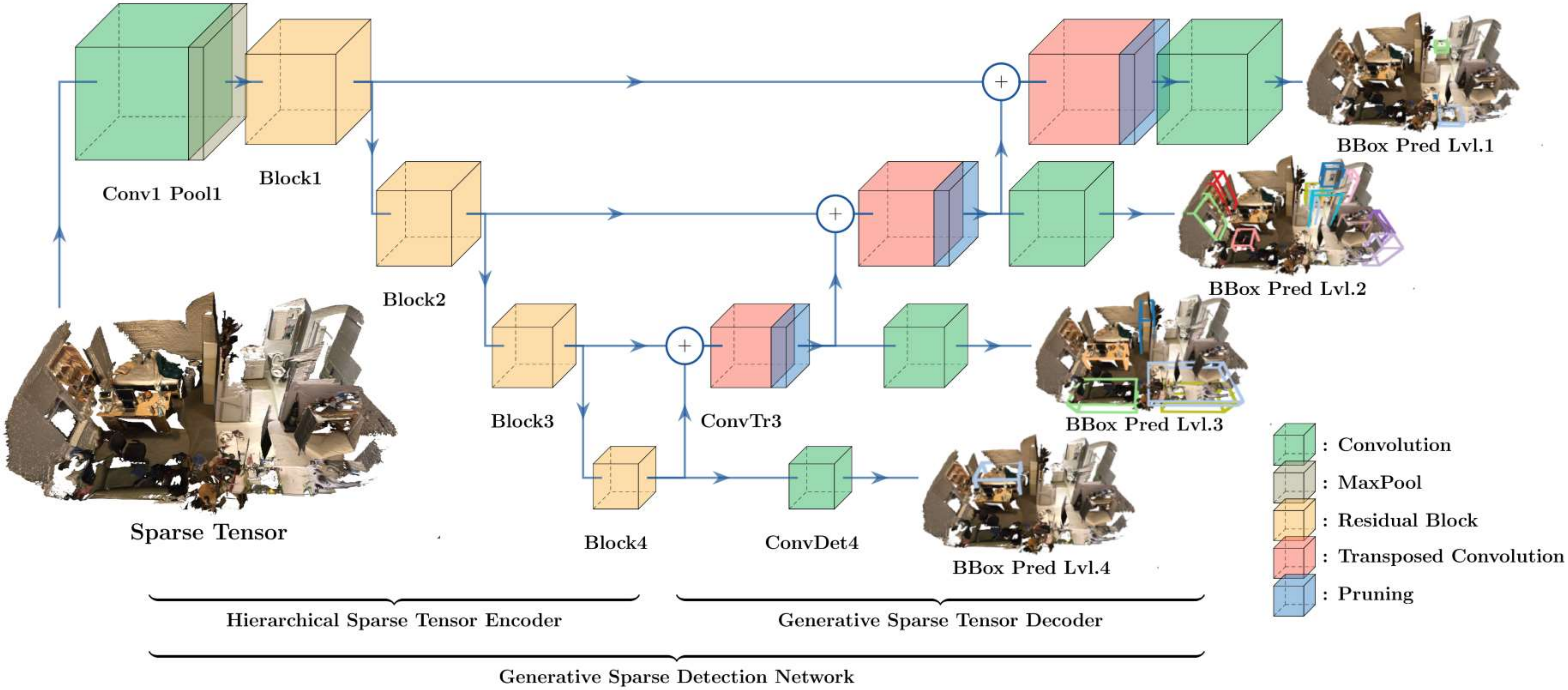
Choy et al., **4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks**, CVPR'19

Dai et al., **SG-NN: Sparse Generative Neural Networks for Self-Supervised Scene Completion of RGB-D Scans**, arXiv'20

GENERATING BOUNDING BOX ANCHORS



JunYoung Gwak, Chris Choy, Silvio Savarese, **Generative Sparse Detection Networks for 3D Single-shot Object Detection**, preprint 2020



Hierarchical Sparse Tensor Encoder

Generative Sparse Tensor Decoder

Generative Sparse Detection Network

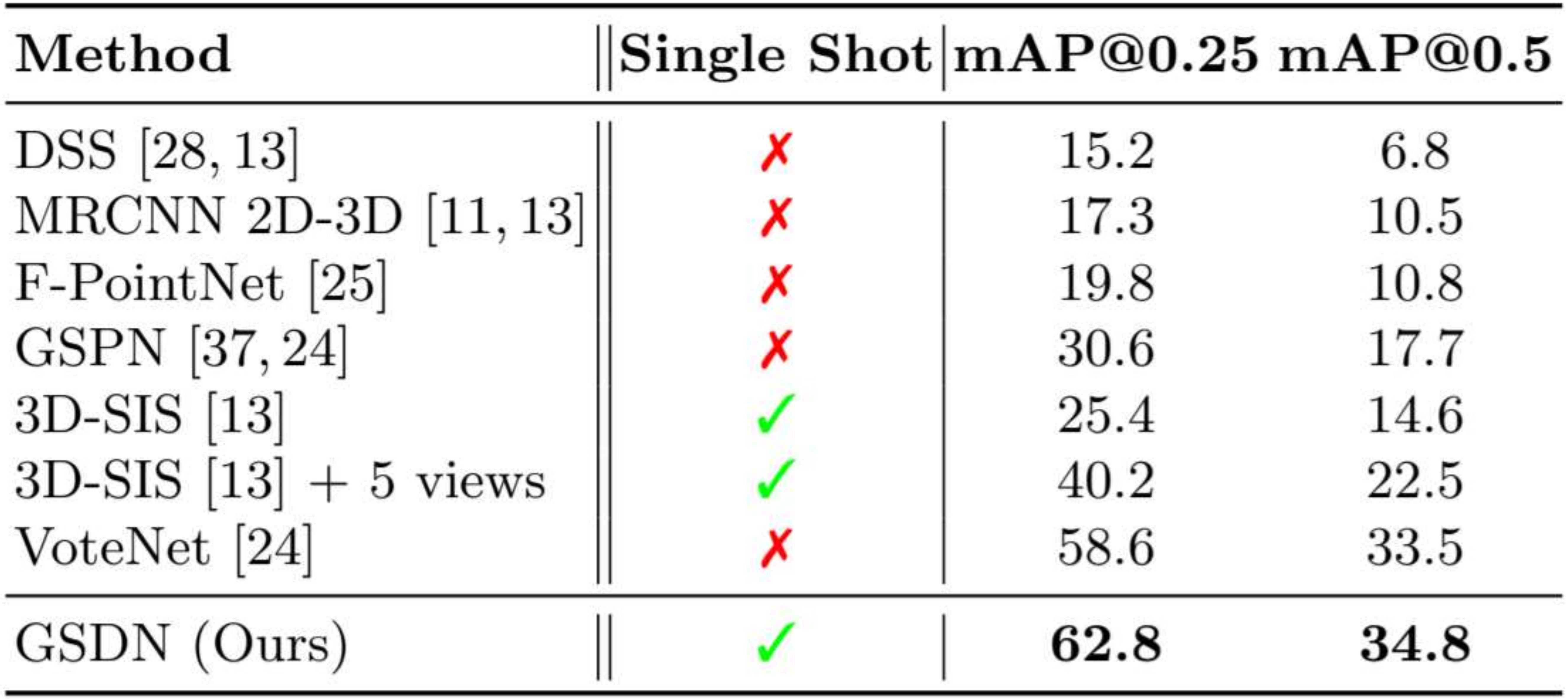
🟩 : Convolution

▨ : MaxPool

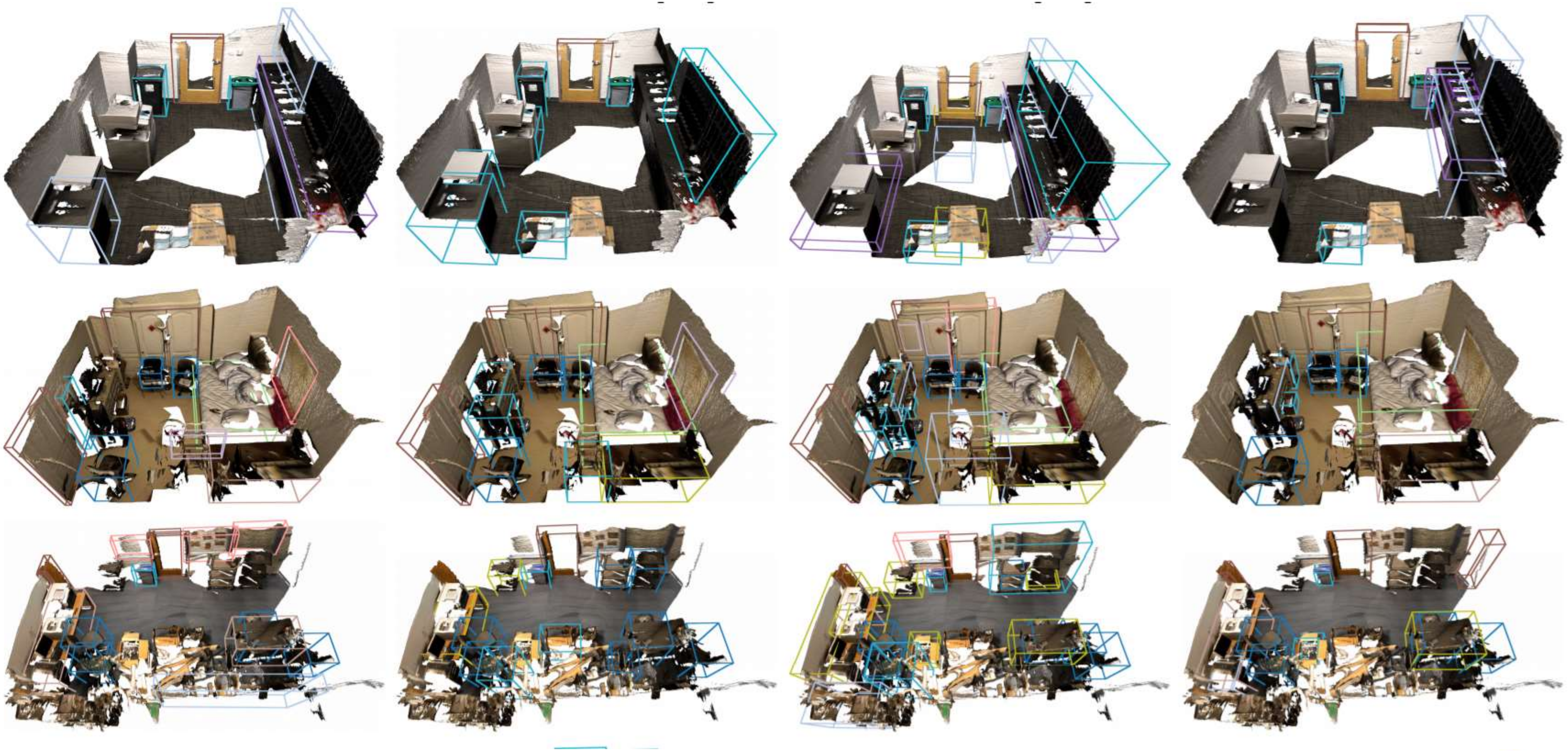
🟧 : Residual Block

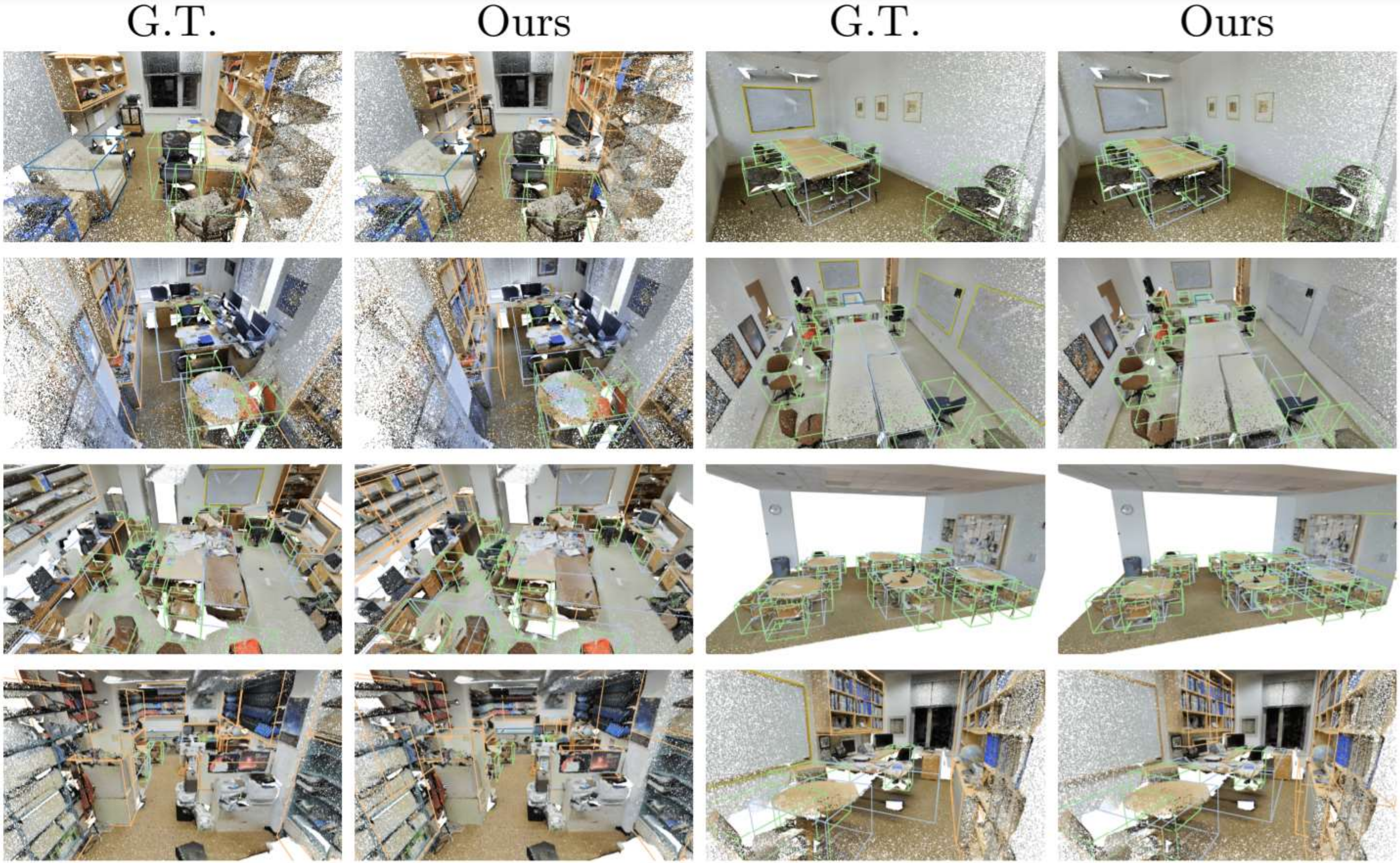
🟥 : Transposed Convolution

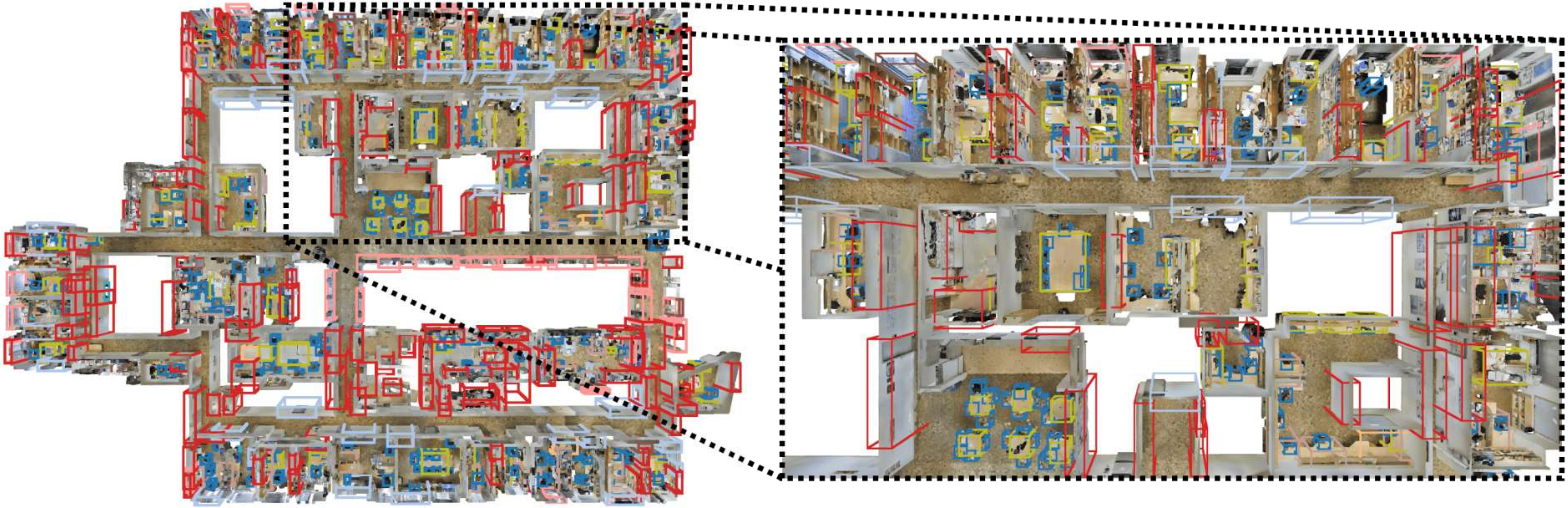
🟦 : Pruning

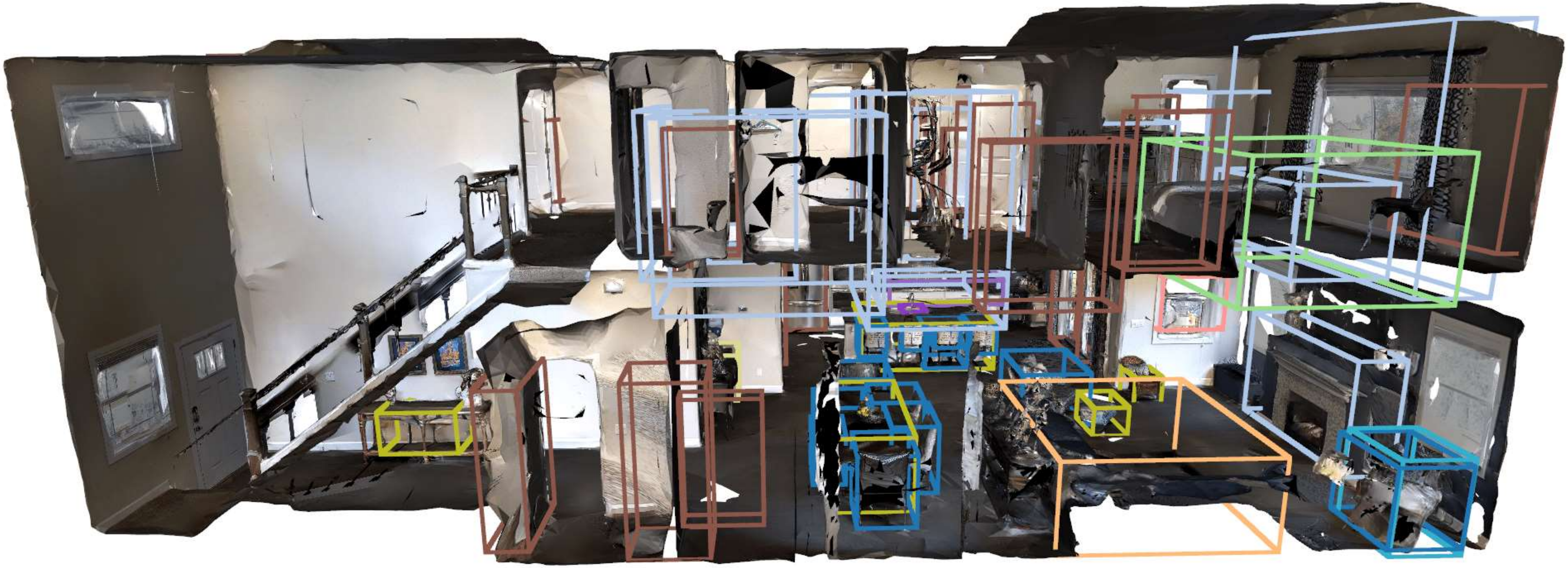


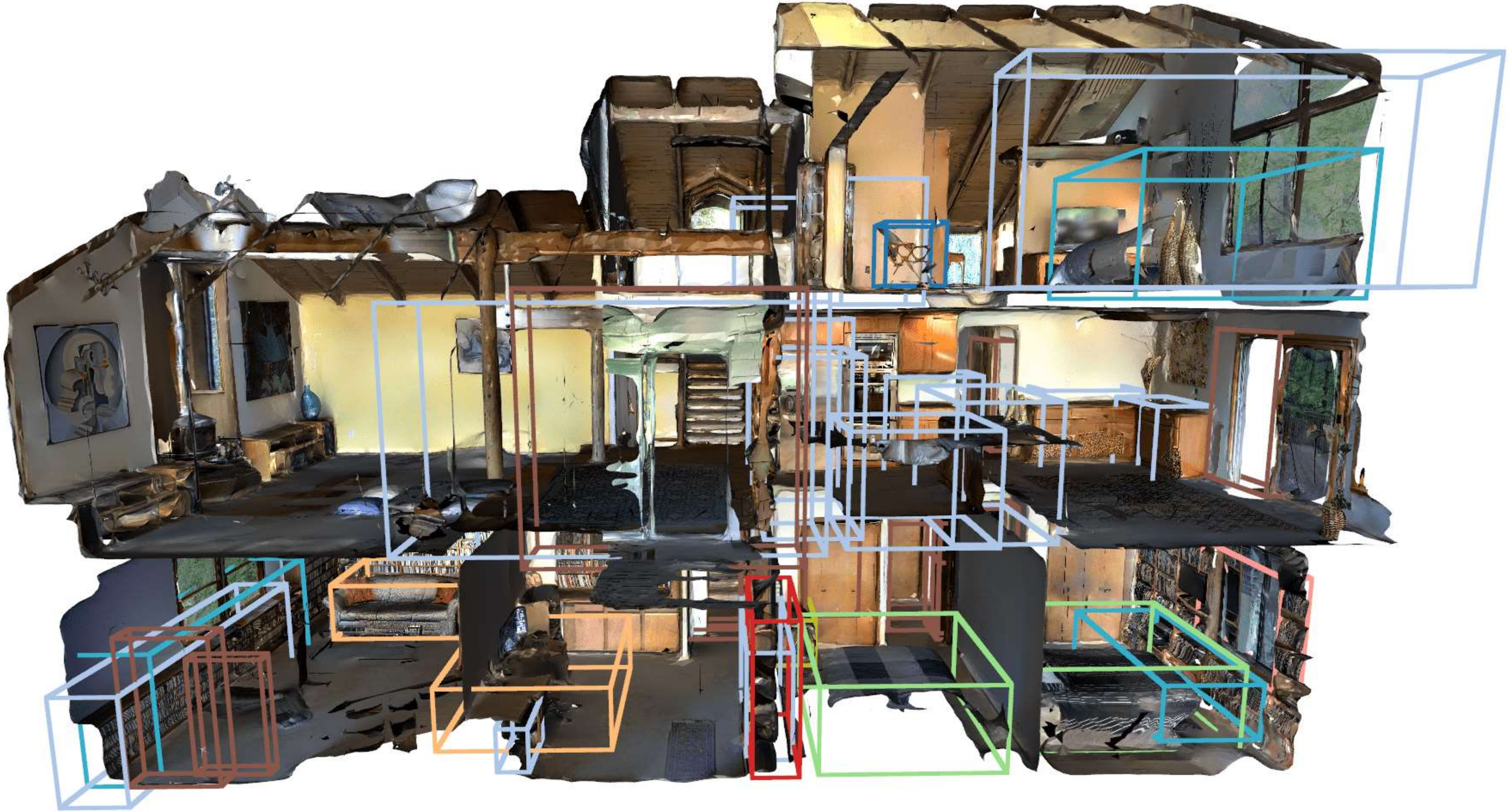
Method	Single Shot	mAP@0.25	mAP@0.5
DSS [28, 13]	✗	15.2	6.8
MRCNN 2D-3D [11, 13]	✗	17.3	10.5
F-PointNet [25]	✗	19.8	10.8
GSPN [37, 24]	✗	30.6	17.7
3D-SIS [13]	✓	25.4	14.6
3D-SIS [13] + 5 views	✓	40.2	22.5
VoteNet [24]	✗	58.6	33.5
GSDN (Ours)	✓	**62.8**	**34.8**




G.T.
Ours
G.T.
Ours







CONCLUSION

CONCLUSION

- A sparse tensor is a powerful representation : discretization has more pros than cons
- Combining discrete representations with continuous representations
  - LIDAR pointclouds, RGB-D scans, voxel-downsampled
  - Hierarchical representation by downsampling points
    - Lose the resolution anyway
  - Discrete for intermediate layers, continuous for the first and last

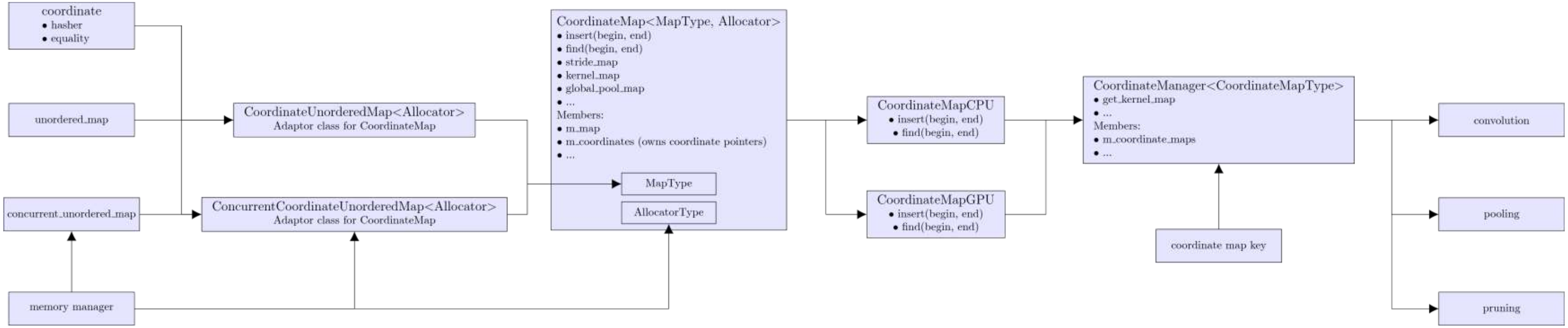
Benjamin Graham, **Sparse 3D convolutional neural networks**, BMVC’15

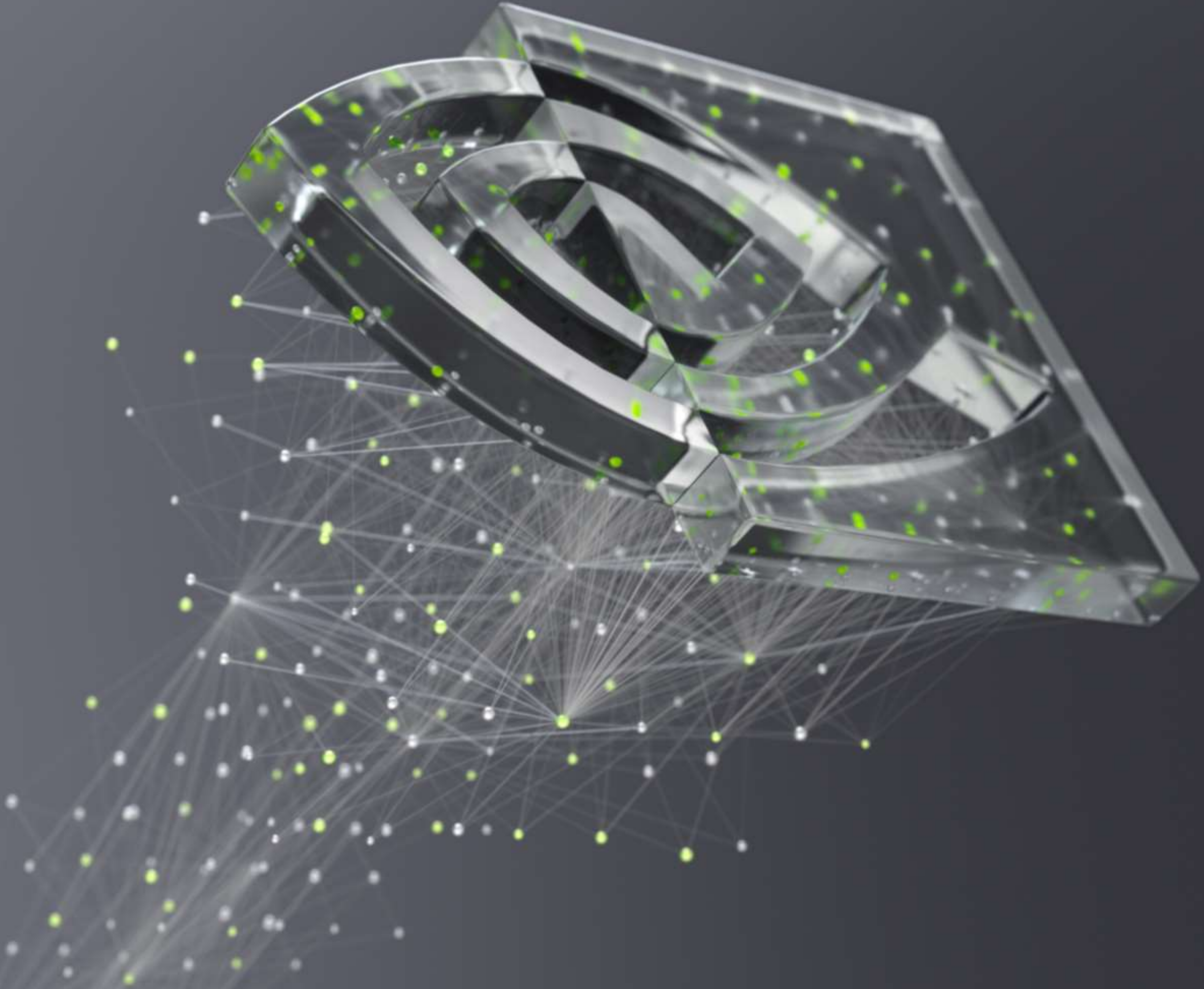
Dai et al., **SG-NN: Sparse Generative Neural Networks for Self-Supervised Scene Completion of RGB-D Scans**, arXiv’20

Peng et al., **Convolutional Occupancy Networks**, arXiv’20

MINKOWSKI ENGINE

- Support for various backends
  - GPU/CPU hashtable





NVIDIA