Semantic Segmentation & Detection Using Multi-Sensory Data

How to obtain such SAR annotations?

- **Option 1:** Manual (or somewhat interactive) annotation is one potential solution
  - Often requires expert's knowledge and
  - Easily becomes impractical when large scenes need to be labelled

- **Option 2:** Employ SAR simulation-based models as proposed e.g., in (Auer et al., 2010) (Tao et al., 2014)
  - However, such methods have their own limitations in a sense, they typically require accurate models (3-D building models and/or accurate DSMs) to precisely generate such ground truth data which, in most cases, is not available

- **Option 3:** Use of existing geographic information systems (GIS) data to obtain direct annotations in SAR images

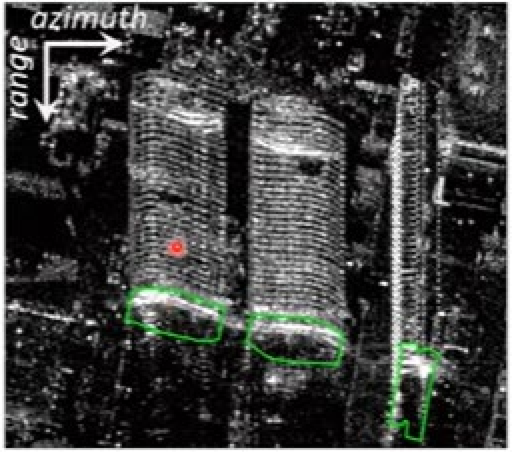
Data Science in Earth Observation
TUM School of Engineering and Design
Technical University of Munich

Layover in SAR Images

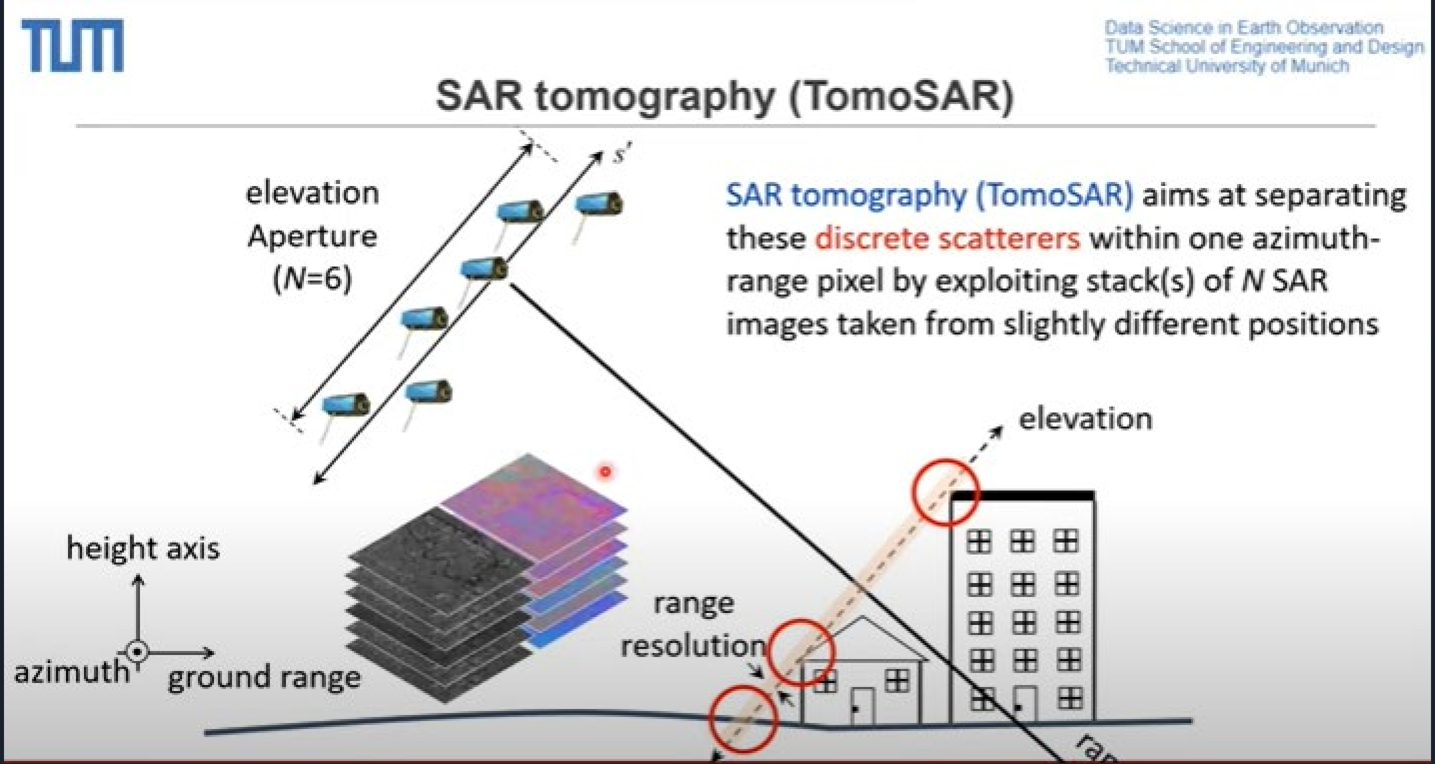
SAR is a ***side looking*** imaging radar that projects a 3-D scene onto 2-D image with two native coordinates: *range* and *azimuth*

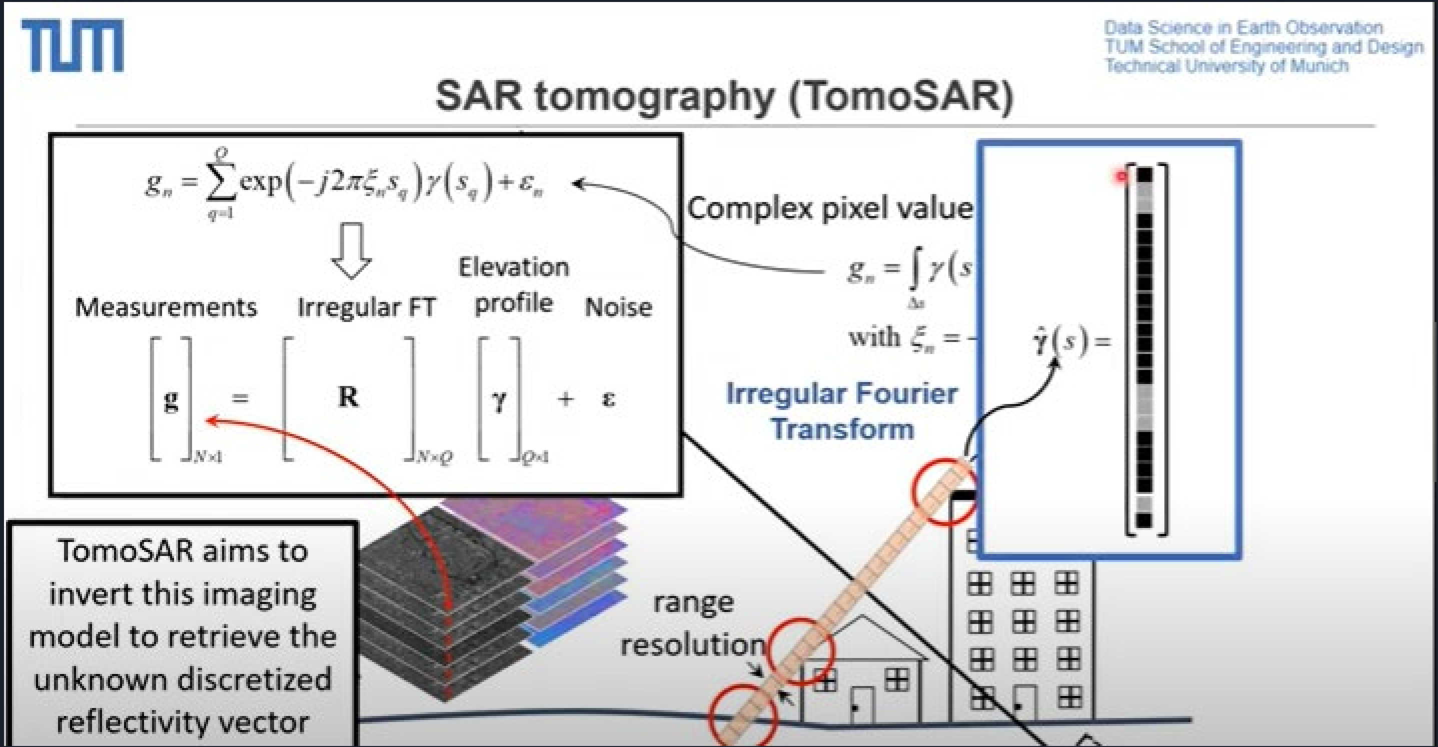
flight direction → Look direction

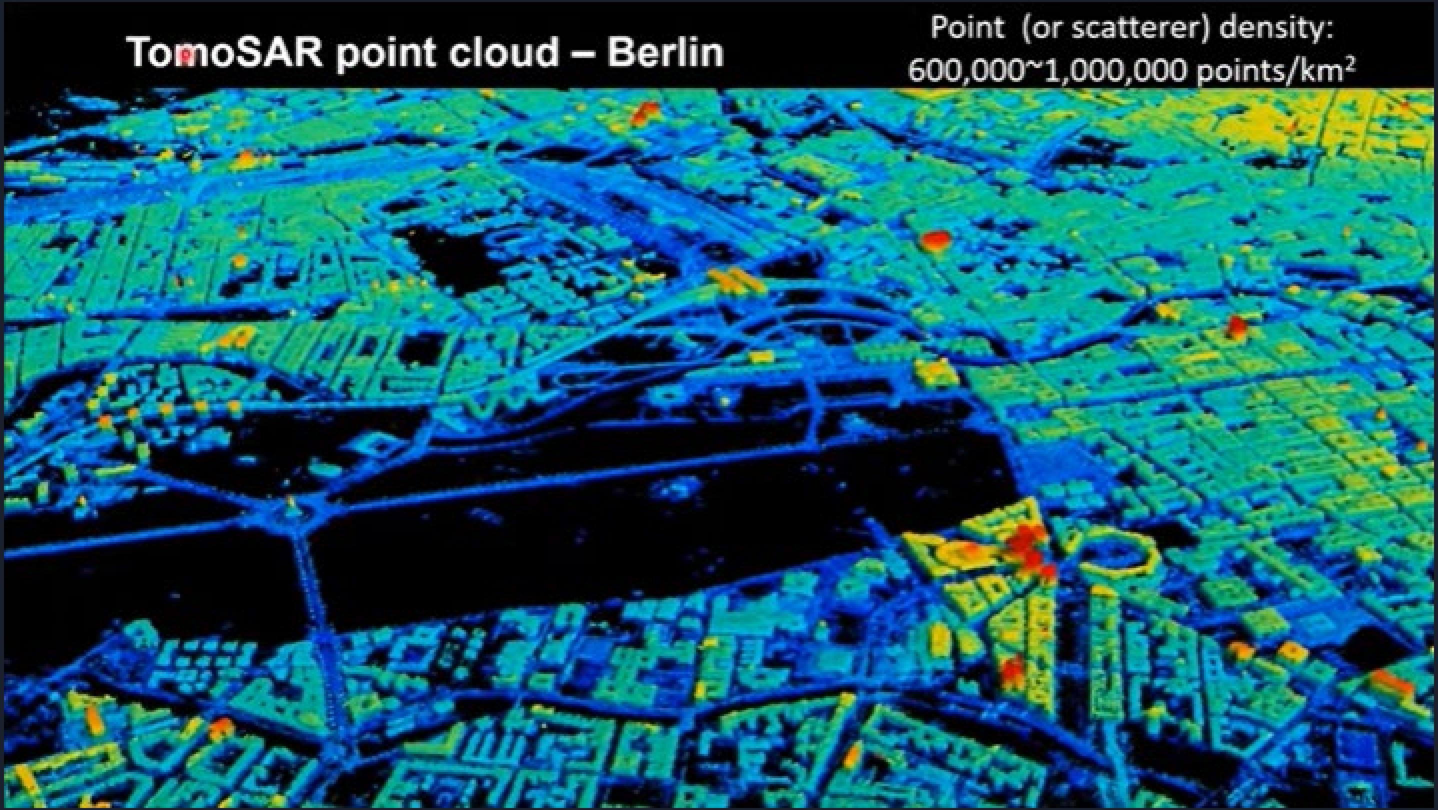
3-D scene (Top view)



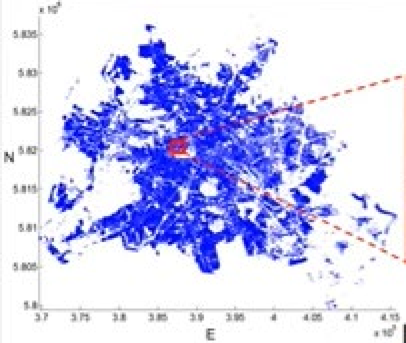
Corresponding SAR image (TerraSAR-X)

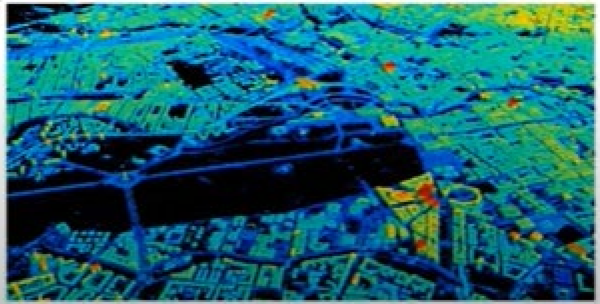

TUM
Data Science in Earth Observation
TUM School of Engineering and Design
Technical University of Munich
SAR tomography (TomoSAR)
SAR tomography (TomoSAR) aims at separating these discrete scatterers within one azimuth-range pixel by exploiting stack(s) of N SAR images taken from slightly different positions
elevation
Aperture
(N=6)
s'
elevation
height axis
azimuth
ground range
range
resolution


TUM
Data Science in Earth Observation
TUM School of Engineering and Design
Technical University of Munich
SAR tomography (TomoSAR)
$g_n = \sum_{q=1}^{Q} \exp(-j2\pi\xi_n s_q)\gamma(s_q) + \varepsilon_n$
Complex pixel value
Measurements
Irregular FT
Elevation profile
Noise
$[\mathbf{g}]_{N\times 1} = [\mathbf{R}]_{N\times Q} [\boldsymbol{\gamma}]_{Q\times 1} + \boldsymbol{\varepsilon}$
$g_n = \int_{\Delta s} \gamma(s$
with $\xi_n = -$
$\hat{\boldsymbol{\gamma}}(s) =$
Irregular Fourier Transform
TomoSAR aims to invert this imaging model to retrieve the unknown discretized reflectivity vector
range resolution


TomoSAR point cloud – Berlin
Point (or scatterer) density:
600,000~1,000,000 points/km²

Generation of Automatic SAR Annotation Mask





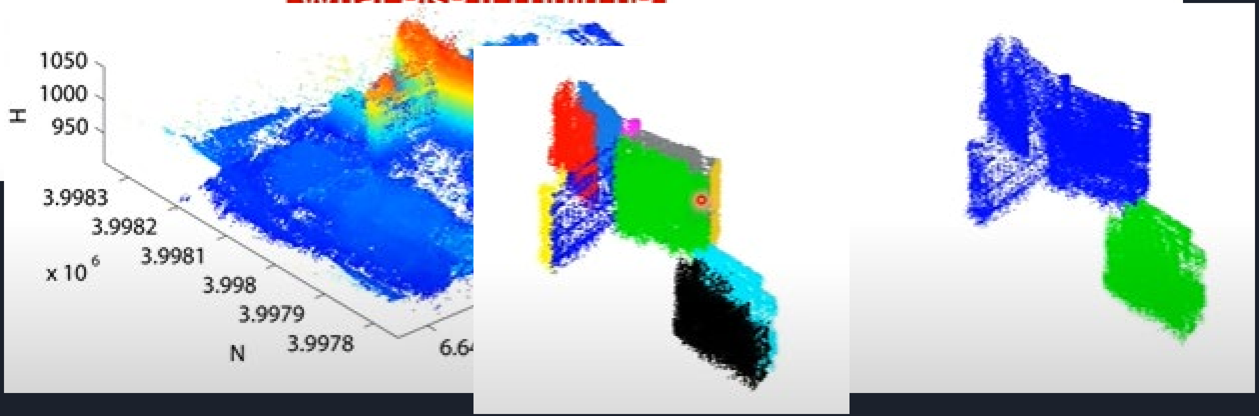
- Extract all those TomoSAR points that lie within the OSM building polygons
- Project these building points back to SAR image coordinates (i.e., range and azimuth) to yield the building mask

Data Science in Earth Observation
TUM School of Engineering and Design
Technical University of Munich

Why to do 3D Point Cloud Segmentation?

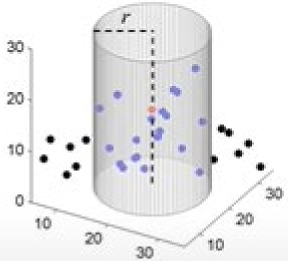
- Demonstrated the use-case where 3D point cloud segmentation could assist in creating large-scale SAR annotation masks

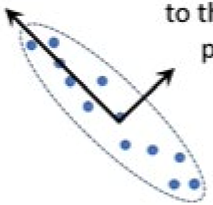
What else it could be



Conventional – 3D Features (Handcrafted)

- **Conventional approach**
  - Extract hand-crafted features + unsupervised/superv



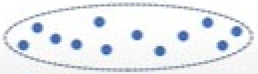


Eigenvector corresponding to the **smallest** eigenvalue points in the **normal** direction

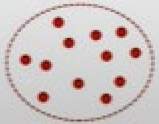
Aspect ratio is a measure of ***"flatness"***

Eigenvalues of 3D structure denoted by $\lambda_1 \geq \lambda_2 \geq \lambda_3 \in \mathbb{R}$

$\frac{\lambda_2}{\lambda_1}$ is small



$\frac{\lambda_2}{\lambda_1}$ is larger



Features	Definitions
Sum of eigenvalues	$\sum \lambda_i$
Omnivariance	$(\prod \lambda_i)^{\frac{1}{3}}$
Eigenentropy	$-\sum \lambda_i \ln(\lambda_i)$
Linearity	$(\lambda_1 - \lambda_2)/\lambda_1$
Planarity	$(\lambda_2 - \lambda_3)/\lambda_1$
Sphericity	λ_3/λ_1
Change of curvature	$\lambda_3/(\lambda_1 + \lambda_2 + \lambda_3)$
Verticality (x2)	$\left\lvert \frac{\pi}{2} - \text{angle}(\mathbf{e}_i, \mathbf{e}_z) \right\rvert_{i\in(0,2)}$
Absolute moment (x6)	$\frac{1}{\lvert\mathcal{N}\rvert}\left\lvert \sum \langle \mathbf{p} - \mathbf{p}_0, \mathbf{e}_i \rangle^k \right\rvert_{i\in(0,1,2)}$
Vertical moment (x2)	$\frac{1}{\lvert\mathcal{N}\rvert}\sum \langle \mathbf{p} - \mathbf{p}_0, \mathbf{e}_z \rangle^k$
Number of points	$\lvert\mathcal{N}\rvert$
Average color (x3)	$\frac{1}{\lvert\mathcal{N}\rvert}\sum c$
Color variance (x3)	$\frac{1}{\lvert\mathcal{N}\rvert - 1}\sum (c - \bar{c})^2$

(Thomas et al., 3DV, 2018)

- Region growing algorithms
- Gaussian sphere + meanshift clustering
- Energy minimization frameworks
- ...

- Deep learning based 3D point segmentation

Problem – 3D point clouds are "*Unstructured*"

- Lack of grid-structure
- Permutation-invariance
- Sparsity
- Highly variable density
- Acquisition artifacts
- Occlusions
- Data volume considerable

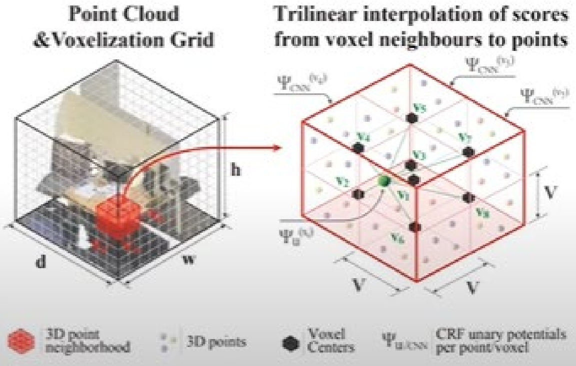
Data Science in Earth Observation
TUM School of Engineering and Design
Technical University of Munich

Can Voxelization help?

- Lack of grid-structure
- Permutation-invariance
- Sparsity

Idea: Generalize 2D convolutions to regular 3D grids

▸ Voxelization + 3D convNets



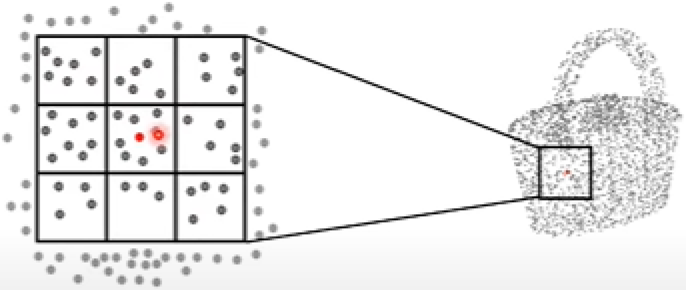
Drawbacks:
- Additional **computation effort**
- Makes the data unnecessarily **voluminous**
- Introduce **quantization errors** that may not only hinder in extracting implicit 3D shape information but also in capturing the essential data invariances for the required segmentation task

Can we perform convolution without Voxelization?

3D Point-wise Convolution to Extract Deep Features

- Inspired by (Hua et al., 2018), we employed a point-wise convolution into a residual framework to perform the semantic segmentation and object classification

$$p_k^{l+1} = \sum_{i=1}^{u\times v\times q}\left[\frac{1}{|\xi_k(i)|}\sum_{j=1}^{|\xi_k(i)|} p_j^l\right]\cdot w_i$$



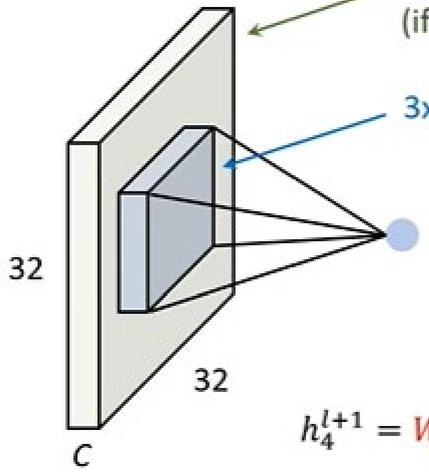
Point-wise convolution

$\xi_k(i)$ denotes the *i*th grid cell
w_i represents the kernel weight at the *i*th grid cell $\xi_k(i)$
$p_j^l \in \xi_k(i)$ the value of any *j*th point lying within *i*th grid cell $\xi_k(i)$ in the previous layer l

Data Science in Earth Observation
TUM School of Engineering and Design
Technical University of Munich

Graph Convolution

Images are a **structured graph of pixels!**

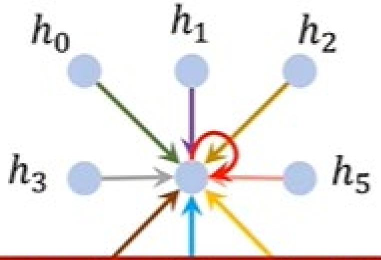


32x32xC activations
(if image then C = 3)

3x3xC filter, K filters

W_0	W_1	W_2
W_3	W_4	W_5
W_6	W_7	W_8

$W_i \in \mathbb{R}^{K \times C}$



$h_i \in \mathbb{R}^C$

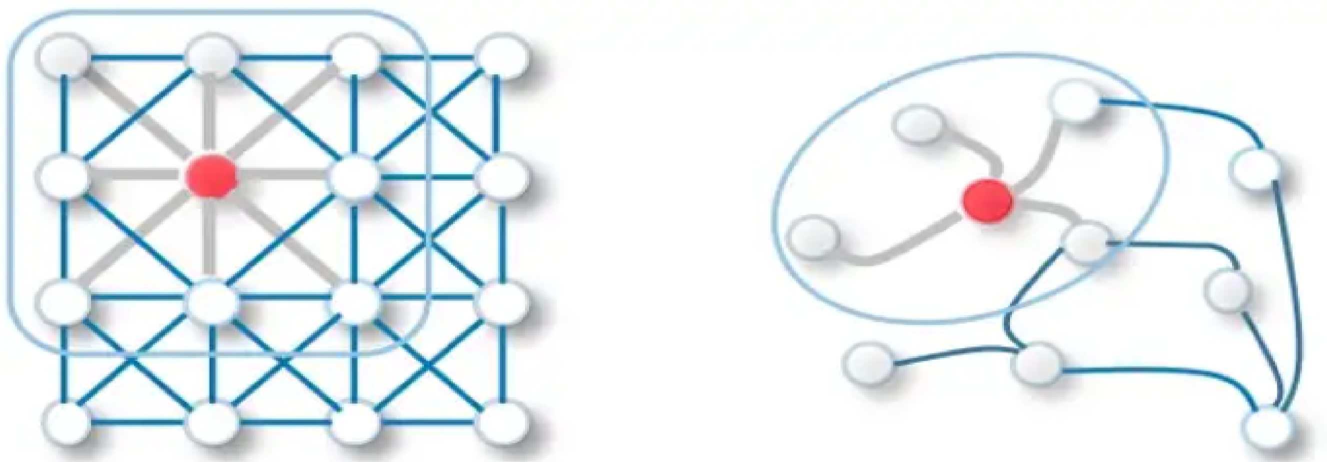
- **Ordering of neighbors** should **NOT** matter
- **Number of neighbors** should **NOT** matter

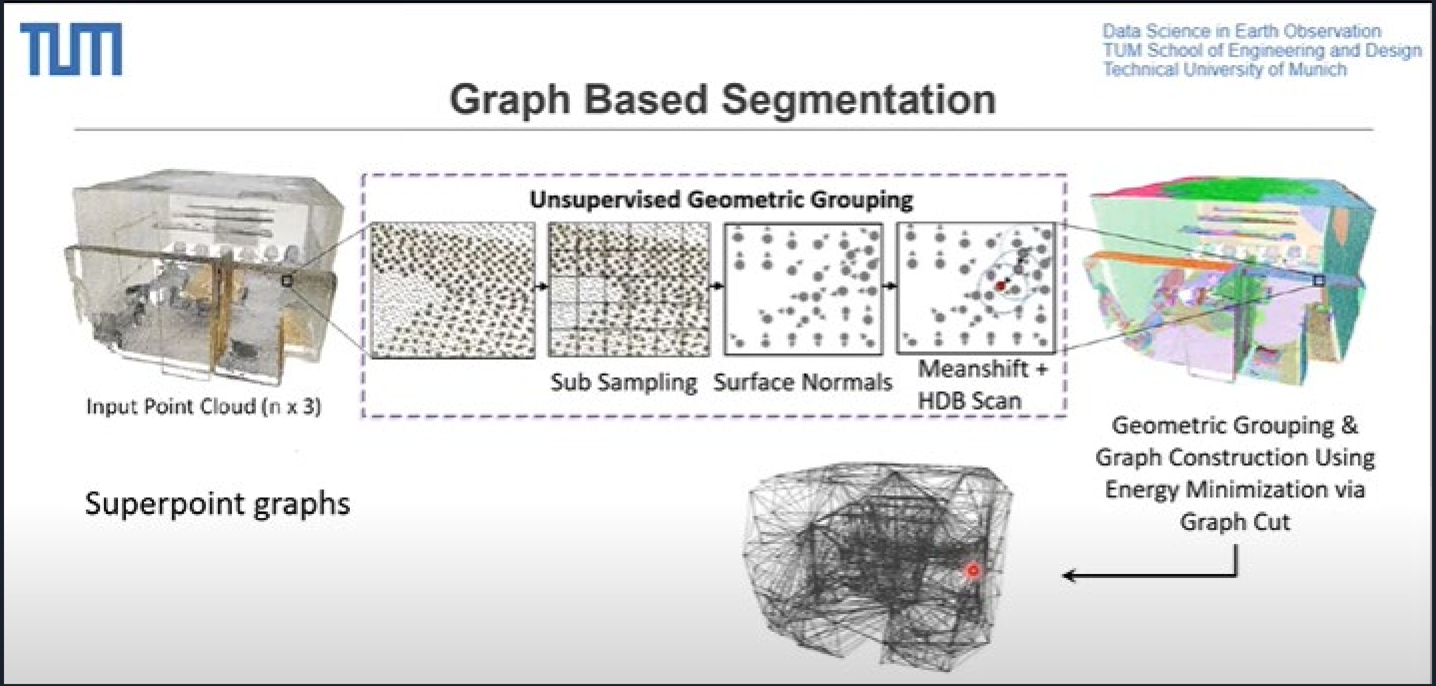
Can this formulation work over **unstructured** graph?

$$h_4^{l+1} = W_4 h_4^{l+1} + W_0 h_0^{l+1} + W_1 h_1^{l+1} + \dots + W_5 h_5^{l+1} + W_7 h_7^{l+1} + W_8 h_8^{l+1}$$

Credit: (Fei-Fei Li, CS231n)

GCNs perform similar operations where the model learns the features by inspecting neighboring nodes. The major difference between CNNs and GNNs is that CNNs are specially built to operate on regular (Euclidean) structured data, while GNNs are the generalized version of CNNs where the numbers of nodes connections vary and the nodes are unordered (irregular on non-Euclidean structured data).




Data Science in Earth Observation
TUM School of Engineering and Design
Technical University of Munich
Graph Based Segmentation
Unsupervised Geometric Grouping
Input Point Cloud (n x 3)
Sub Sampling
Surface Normals
Meanshift +
HDB Scan
Geometric Grouping &
Graph Construction Using
Energy Minimization via
Graph Cut
Superpoint graphs

Object Detection: *Natural* vs. *Remote Sensing* Images

- Many state-of-the-art object detection methods exist
  - Region CNNs and its variants
  - Yolo Only Look Once (YOLO) family
  - Single Shot MultiBox (SSD) Detector
  - RetinaNet
  - ...

- Object detection is highly challenging in remote sensing because
  - Images are acquired from high altitudes causing atmospheric distortions,
  - Illumination and viewpoint variations,
  - Partial occlusions, and
  - Clutter (especially in urban environments)


Object Detection: Natural vs. Remote Sensing Image
Airplane
Ship
Vehicle
PASCAL VOC Dataset
DIOR Dataset