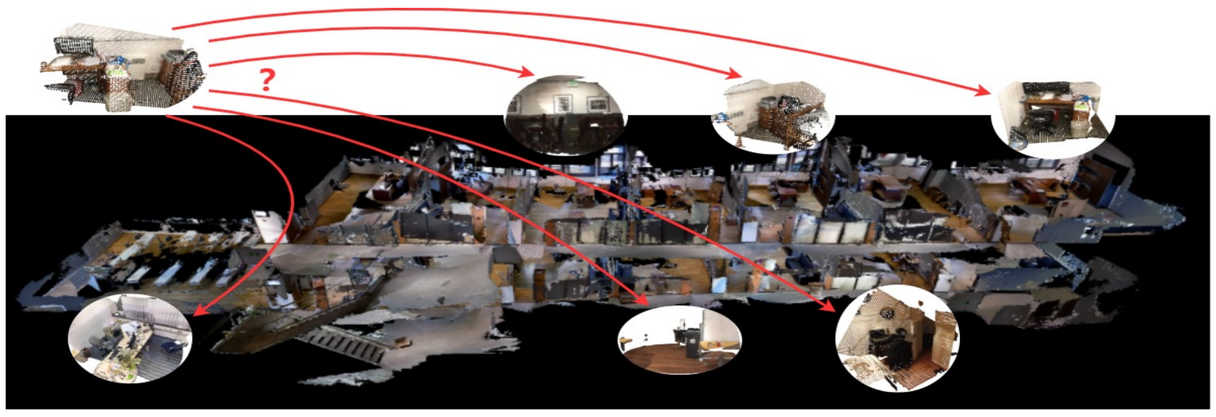
AEGIS-Net: Attention-Guided Multi-Level Feature Aggregation for Indoor Place Recognition

Yuhang Ming[1,2*], Jian Ma[3*], Xingrui Yang[4], Weichen Dai[1,2], Yong Peng[1,2], Wanzeng Kong[1,2]

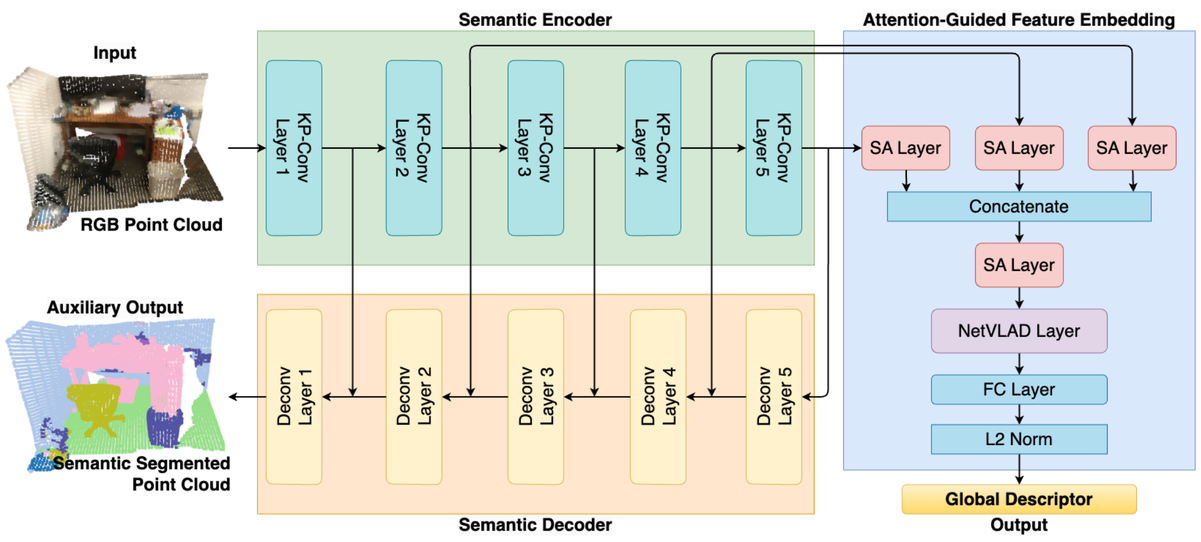
1 School of Computer Science, Hangzhou Dianzi University;
2 Zhejiang Key Laboratory of Brain-Machine Collaborative Intelligence;
3 Unaffiliated; 4 High-speed Aerodynamics Institute, CARDC.

Motivation

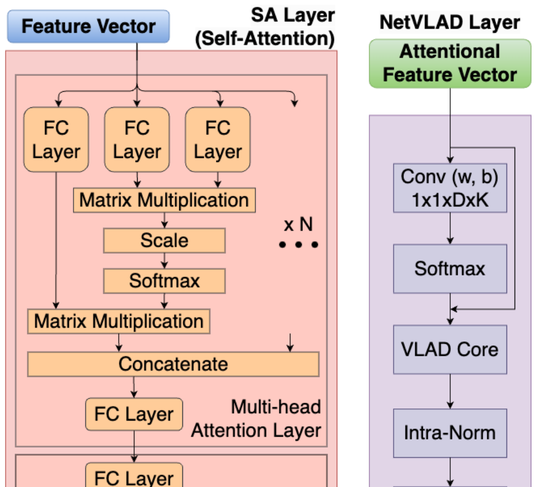


- **Goal**
  Indoor place recognition with RGB point clouds
- **Challenges**
  - Limited amount of information caused by the close proximity of the sensor
  - Similar appearance and structure among difference places
- **Contributions**
  - Propose a 2-stage multi-task learning approach to utilize geometry, colour, implicit semantic features
  - Introduced an adaptive feature aggregation with self-attention layers
  - Comparison with a traditional feature-based method and four state-of-the-art deep learning-based methods and significantly outperform all six methods.

Network Architecture & Training Procedure



- **Main Components**
  - **Semantic Encoder:** 5 deformable kernel point convolution (KP-Conv) layers;
  - **Semantic Decoder:** 5 layers of nearest upsampling followed by unary convolution;
  - **Attention-Guided Feature Embedding:** 4 self-attention layers with 3 applied on features extracted from layer 2, 4, 5 and the other applied on their concatenation; and 1 NetVLAD layer for feature aggregation.



- **Training Procedure**
  - **Stage 1:** train the semantic encoder and semantic decoder on the semantic segmentation task
  - **Stage 2:** discard the semantic decoder and train the attention-guided feature embedding with the weights in the semantic encoder fixed
- **Loss**
  - Lazy quadruplet loss

$$\mathcal{L}_{LazyQuad}(\mathcal{T}) = \max_{i,j}([\alpha + \delta_i^{pos} - \delta_j^{neg}]_+) + \max_{i,k}([\beta + \delta_i^{pos} - \delta_k^{*}]_+)$$

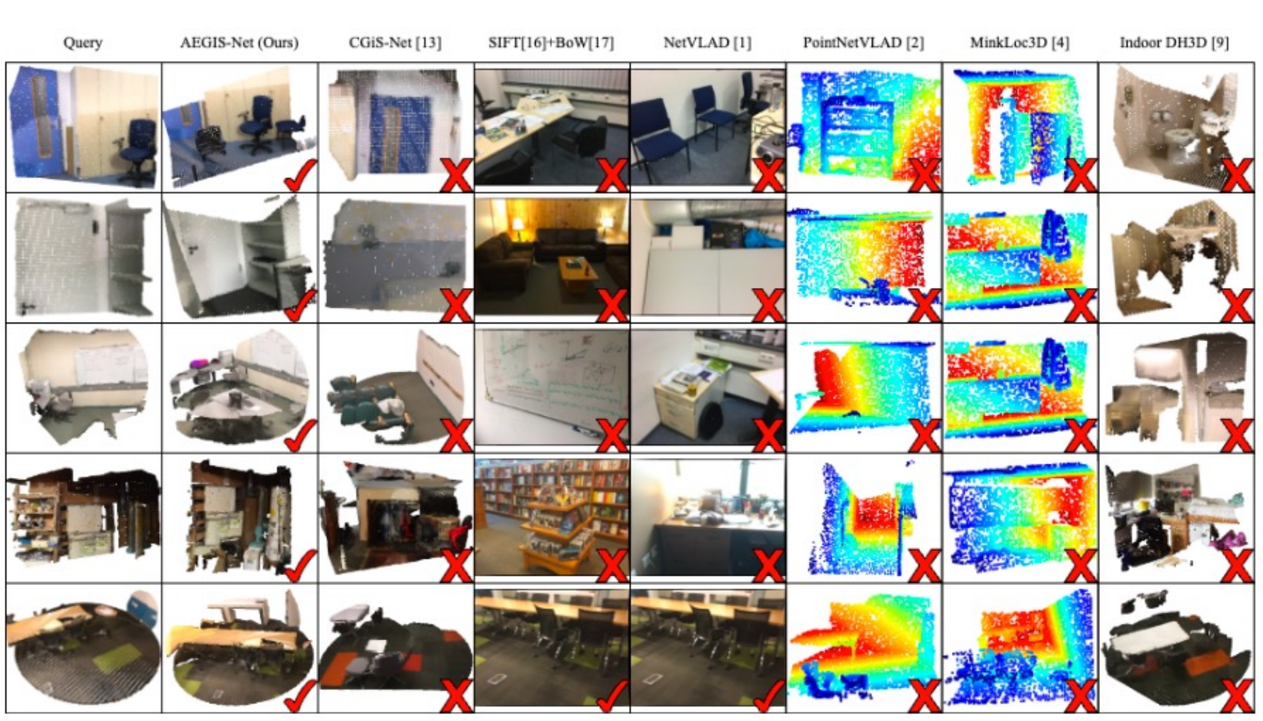
Dataset

- **ScanNetPR:** derived from ScanNet with keyframe selection based on camera movement, then the RGB point clouds are generated from these selected keyframes
  - Total: 807 different indoor scenes and 1,613 RGB-D scans
  - Training: 565 scenes, 1,201 scans and **35,102** keyframes
  - Validation: 142 scenes, 312 scans and **9,693** keyframes
  - Test: 100 scenes, 100 scans and **3,608** keyframes

Evaluation

- **Generating place recognition database**
  - Database point clouds are selected at least 3 meters apart
  - Result in 236 database point clouds, 3,372 test point clouds
- **Evaluation Criterion**
  - Threshold for successful retrievals: 3 meters
  - Criterion: Top-K average recall rate (%)
- **Comparison Counterparts**
  - Baseline: SIFT + BoW
  - Advanced learning-based methods: CGiS-Net, Indoor DH3D, MinkLoc3D, PointNetVLAD, NetVLAD.
- **Ablation Study**
  - **Efficiency:** AEGIS-Net converges after 20 epochs, CGiS-Net, on the other hand, needs 60 epochs. When restricting the training epochs to 20, CGiS-Net exhibits a significant performance degradation
  - **Attention Mechanism:** Removing the self-attention layer before and after feature concatenation results a drastic drop in the place recognition performance.

Methods	R@1	R@2	R@3
AEGIS-Net (Ours)	**65.09**	**74.26**	**79.06**
CGiS-Net [13]	61.12	70.23	75.06
SIFT [16] + BoW [17]	16.16	21.17	24.38
NetVLAD [1]	21.77	33.81	41.49
PointNetVLAD [2]	5.31	7.50	9.99
MinkLoc3D [4]	3.32	5.81	8.27
Indoor DH3D [9]	16.10	21.92	25.30
CGiS-Net-20 [13]	56.82	66.46	71.74
AEGIS-Net (w/o attention)	55.13	66.19	71.95



Contact: yuhang.ming@hdu.edu.cn