

Lack of systematic and quantitative evaluations of Interpretability in Semantic Segmentation Models

Perturbation-Based

- LIME [Ribeiro et al., 2016]
- SHAP [Lundberg and Lee, 2017]
- Occlusion [Samek et al., 2021]
- EBAnO (quatification) [Ventura et al., 2023]

Perturbation itself may introduce artifacts since disturbed images may be out-of-distribution

Backpropagation-based

- Activation Maximization [Stergiou, 2021]
- Layer-Wise Relevance Propagation [Jung et al., 2021]

Difficult to interpret, assume increased activation

Class activation

- Grad-CAM [Selvaraju et al., 2016]
- Grad-CAM++ [Chattopadhyay et al., 2017]
- Layer-CAM Jiang et al., 2021]
- Shap-CAM [Zheng et al., 2022]

Local-layer explainability

For semantic segmentation

- SHAP values for SS [Dardouillet et al., 2022]
- Prototypes [Sacha et al., 2023]
- SAU-Net proposed in [Sun et al., 2020]

No quantitative explanations

Classification models