# Methodology

The proposed methodology employs a two-stage approach consisting of a keypoint prediction module and a spatio-temporal keypoint refinement module [xu2023graph].

In the keypoint prediction module, real-time video serves as input, and a ResNet18 network, pre-trained on the ImageNet dataset, is used as the backbone to extract image features. A segmentation branch is then applied to distinguish surgical tools from the background. Once segmented, a vector pixel voting process utilizes a vector field to predict the keypoint locations of the surgical tool[xu2023graph].

Following keypoint prediction, graph information is constructed for the identified keypoints. Temporal information is first captured using a Temporal Convolutional Network (TCN) [zeng2020srnet], which models the relationships between consecutive frames. Then, a Graph Convolutional Network (GCN) [ yan2018spatial] extracts spatial relationships among the keypoints, refining their positions based on the graph structure [xu2023graph]. The final 2D keypoint outputs from the model are subsequently converted into 3D coordinates using the PnP algorithm [yun2017object].

## Reference：

@inproceedings{zeng2020srnet,

title={Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach},

author={Zeng, Ailing and Sun, Xiao and Huang, Fuyang and Liu, Minhao and Xu, Qiang and Lin, Stephen},

booktitle={Computer Vision--ECCV 2020: 16th European Conference, Glasgow, UK, August 23--28, 2020, Proceedings, Part XIV 16},

pages={507--523},

year={2020},

organization={Springer}

}

@inproceedings{yan2018spatial,

title={Spatial temporal graph convolutional networks for skeleton-based action recognition},

author={Yan, Sijie and Xiong, Yuanjun and Lin, Dahua},

booktitle={Proceedings of the AAAI conference on artificial intelligence},

volume={32},

number={1},

year={2018}

}

TCN : Temporal Convolutional Network

GCN : Graph Convolutional Network