1. Train YOLO and RCNN on the data and then deploy the model into an interface that allows some stakeholder to interact with it – similar to <https://www.erdosinstitute.org/_files/ugd/987354_c14304f235e84ec288d694b9a11e7917.pdf>
2. Building a modular hybrid model that runs a single-image detector (YOLO or Faster R-CNN) and adds a temporal head (LSTM or transformer) for sequence data. – The question this model would answer is detection but it is able to understand videos and single frames
   1. Understand the performance of the Baseline model over time. (Do we see a bad performance especially in the beginning of the time?)
3. The performance seems to be related to the sizes of the polyps. Figuring out bounding box threshold based on the sizes (or all the sizes of that image).
4. Generalizability improvement:
   1. Something important that I stumbled upon to consider is the difference between White Light and Narrow Band Imaging. They are two different ways of exploring the same anatomy and they highlight different aspects of the image. Thus, for generalizability, it is important that the model knows both WL and NBI. A trick we could use is to do augmentation on WL images to mimic NBI images. Maybe we need to reconsider how we stratify our data?
   2. Potentially using Reinhard or gray-world filter to reduce training noise on data centers so the model is actually learning polyps.
   3. Convert masks to tight image boxes to get more data
   4. Model:
      1. We can make a hybrid of the detection backbones, that is, combine RT-DETR/DINO-DETR (transformer based, excellent localization) and YOLOv8 (fast, good recall, overdetects). Train both of these on the same dataset and then we can fuse their predictions using weighted boxes fusion.
      2. Varifocal/focal loss (good for small polyps) for classification, GIoU/DIoU (penalizes misalignment more than L1 or IoU) for boxes
      3. Even though we are not doing segmentation, we could add a tiny segmentation head (U-Net branch) that predicts mask to regularize prediction regions
      4. Adding a temporal booster for video images (EMA smoothing, ConvLSTM)
   5. Training:
      1. Warm up on easy datasets like Kvasir SEG with heavy color jitter and blur. No NBI only WL.
      2. Domain mix up – slow add in of new data and turn down the augmentation when we have more variety in the epochs
      3. Use CycleGAN on each center
   6. Inference – check EndoCV metric