**Purpose:** One barrier to the clinical deployment of deep learning-based image segmentation models is the existence of “anomalous” cases where the segmentation model may fail, despite external validation. We developed methods that enable the detection of images on which the segmentation models are likely to fail, thereby aiding in clinical deployment of these models.

**Methods:** We designed an anomaly detector that approximates the performance of a liver segmentation model on CT images by determining whether the images have a high likelihood of falling within the model’s training distribution. We consider an image to be within the training distribution if our anomaly detection model can reconstruct the image. We use backpropagation to map input images to a latent space and a GAN to reconstruct the images. We used a StyleGAN2 network, due to its cutting-edge performance on high-resolution images. We will score reconstructed images using a residual loss. If this score is over a specified threshold, the image will be classified as anomalous. We trained the detection model on 96 CT scans and have 50 unseen CT scans available for anomaly detection. We use standard image similarity metrics for evaluation.

**Results:** We successfully trained the StyleGAN2 network on slices extracted from non-contrast enhanced abdominal CT images. The network achieved a Fréchet-Inception score of 31.806. It also attained a Wasserstein distance of 1.715, a KL-divergence of 0.010, and a mutual information score of 0.227 between the pixel distributions of 120 real and 120 generated images. After successful reconstruction of images, a radiologist classified 10 images as real or generated with specificity of 0.6.

**Conclusion:** We defined an anomaly detection paradigm that determines when a segmentation model is likely to fail on an image. This paradigm can aid the clinical deployment of segmentation models, diversify medical imaging datasets, and prioritize images for labeling.