

**DEEP LEARNING PROJECT**

**CSE 4006**

**LITERATURE REVIEW**

**Project Title: INVISIBLE MAN USING MASK- RCNN**

**Team:  
1.Kashish(22BCE8559)  
2.Vanama Dhhanush Krishna Kumor(22BCE7763)  
3.Parepalli Kesava Naga Sai Durga Vardhan(22BCE20530)**

**Existing Model 1: Mask R-CNN**

**Description**

Mask R-CNN (Regional Convolutional Neural Network) is a deep learning model designed for object detection, classification, and instance segmentation. It extends Faster R-CNN by adding a mask output, allowing for pixel-wise segmentation of objects in addition to bounding box predictions.

**Purpose**

Mask R-CNN is an advanced deep learning model designed for instance segmentation, object detection and keypoint detection. Unlike traditional object detection models, which only predict bounding boxes and object labels, Mask R-CNN also predicts a pixel-wise mask for each detected object. This allows the model to not only locate objects in an image but also to provide a precise segmentation of the object’s shape.

**Key Components**

1. **Backbone Network:**

Typically, a deep convolutional neural network (e.g., ResNet) is used to extract high-level features from the input image..

1. **Region Proposal Network (RPN):**

Mask R-CNN builds on Faster R-CNN by using the same Region Proposal Network (RPN) to generate region proposals, which are areas of the image likely to contain objects.

The RPN predicts whether each proposed region contains an object and provides a rough bounding box for the object.

**RoIAlign (Region of Interest Alignment):**

Unlike Faster R-CNN, which uses RoIPool to extract features for each region, Mask R-CNN introduces RoIAlign.

RoIAlign precisely maps the region proposals onto the feature map, avoiding the quantization issues of RoIPool. This improves the accuracy of pixel-level tasks like instance segmentation.

1. **Object Detection Head:**

The features extracted from each region proposal are passed through fully connected layers that predict the object class and refine the bounding box.

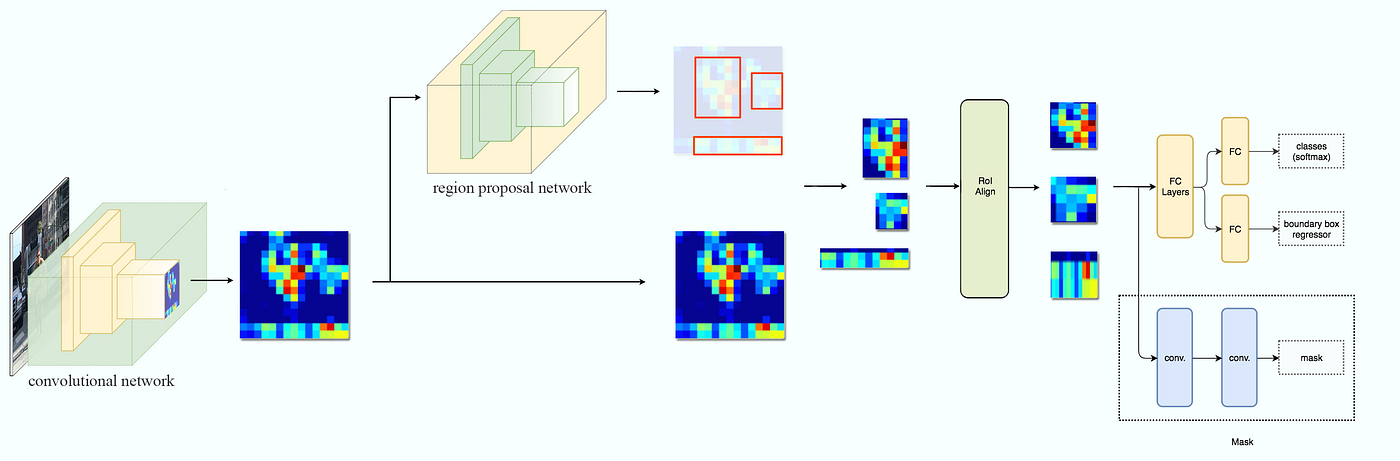
The output includes a class label and an improved bounding box for each detected object.

1. **Mask Prediction Head:**

In addition to the bounding box and classification, a binary mask is predicted for each region, indicating the exact pixels belonging to the object within the bounding box.

This mask head is a small convolutional network that operates parallel to the object detection head.

**Architecture Diagram**



**Existing Model 2: DeepLab v3**

**Description**

DeepLab v3 is a deep learning model designed for semantic image segmentation. Semantic segmentation involves classifying each pixel in an image into a specific category (e.g., background, car, person). The model is widely used in computer vision tasks like autonomous driving, medical imaging, and scene understanding, where precise pixel-wise classification is essential.

**Purpose**

Semantic Segmentation: Assigns a label (class) to every pixel in the image, allowing for detailed object and scene understanding.

Multi-Scale Context Aggregation: Uses different dilation rates to capture information from various spatial scales, improving segmentation accuracy for objects of different sizes.

Edge Refinement: Focuses on improving the accuracy along object boundaries where segmentation is often tricky.

**Key Components**

1. **Convolutional Neural Network (Backbone)**:

Typically, ResNet or MobileNet is used as the backbone for feature extraction from the input image.

The backbone captures basic features like edges, textures, and patterns from the image and reduces its spatial dimensions to create feature maps.

1. **Atrous (Dilated) Convolution**:

This convolution technique allows the model to maintain high-resolution feature maps while expanding the receptive field.

It helps in capturing global context by adjusting the dilation rates in convolutional layers without increasing the computational cost.

1. **Atrous Spatial Pyramid Pooling (ASPP)**:

One of the core innovations in DeepLab v3, ASPP captures contextual information at multiple scales by applying parallel convolutional layers with different dilation rates.

This module improves the model's ability to understand objects at various scales, ensuring accurate segmentation for both small and large objects.

1. **Batch Normalization & Dropout**:

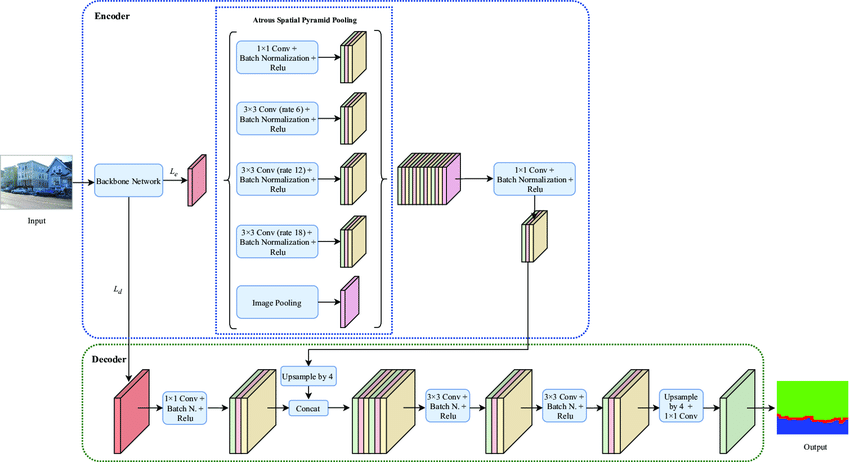
These layers are added between the convolution layers to help stabilize training, prevent overfitting, and improve generalization.

1. **Upsampling (Decoder)**:

After processing the image through convolution and ASPP layers, the spatial resolution is restored using upsampling techniques like bilinear interpolation.

This allows the model to provide pixel-wise predictions at the same resolution as the input image.

**Architecture Diagram**



**Proposed Model : INVISIBLE MAN USING MASK- RCNN**

**Description**

The newly proposed model is an integration of Mask R-CNN and custom background replacement techniques. It combines human segmentation and inpainting methods to remove humans from a live video stream and seamlessly replace them with the pre-recorded background.

**Purpose**

The new model’s primary goal is to remove humans from a video feed in real-time and replace the removed area with a pre-recorded static background. This enhances existing models by:

- Focusing on real-time processing.

- Providing seamless background inpainting without relying on complex deep learning methods for the background filling part.

- Optimizing for practical use cases such as virtual reality or surveillance where human presence needs to be masked.

**Key Features**

1. Real-Time Human Detection and Segmentation: Leverages Mask R-CNN for detecting and masking humans in video frames.

2. Background Inpainting: The model uses custom OpenCV-based logic to fill the removed human areas with corresponding background pixels.

3. Performance Optimization: Optional GPU acceleration using OpenCV’s CUDA backend is integrated to enhance the processing speed.

**Components and Workflow:**

**1.** **Input Stream:** Captures video from a webcam.

**2. Pre-recorded Background: A** clean background frame is captured without any moving objects.

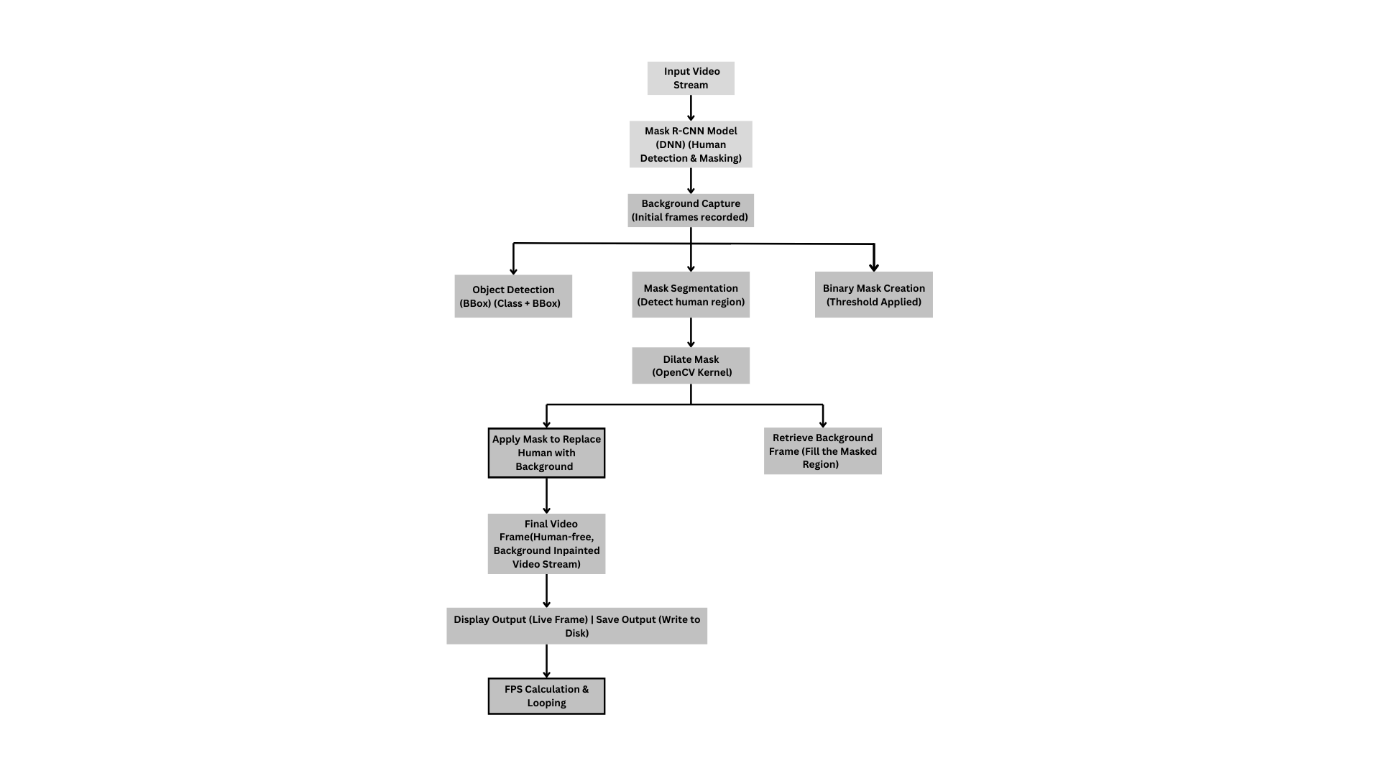
**3. Mask R-CNN Segmentation:** Humans are detected and masked using the Mask R-CNN model.

**4. Mask Refinement:** Morphological operations such as dilation are applied to clean up the mask.

**5. Background Filling:** Pixels corresponding to humans are replaced by background pixels, creating a seamless effect.

**6. Output Stream:** The final video, with humans removed, is displayed and optionally saved.

**Architecture Diagram**

****

**Conclusion**

The combination of existing model Mask R-CNN with background inpainting techniques creates a powerful real-time human removal system. This hybrid method works well for real-time applications like surveillance and video conferencing because it can accurately segment humans and replace backgrounds seamlessly.

While the models show promise, future challenges could include improving the inpainting for dynamic backgrounds or handling edge cases where human detection or segmentation fails. The model may need to be tested in a variety of settings and have its performance adjusted for less powerful hardware in next steps.