VIETNAM GENERAL CONFEDERATION OF LABOR

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**PHAM THAI AN BINH – 522K0031**

**PHAN DUC HIEU – 522K0032**

**MIDTERM PROJECT**

**COMPUTER VISION**

**HO CHI MINH CITY, 2025**

VIETNAM GENERAL CONFEDERATION OF LABOR

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**PHAM THAI AN BINH – 522K0031**

**PHAN DUC HIEU – 522K0032**

**MIDTERM PROJECT**

**COMPUTER VISION**

Advised by

**Dr. PHAM VAN HUY**

**HO CHI MINH CITY, 2025**

**ACKNOWLEDGEMENT**

We would like to express our sincere gratitude to Mr. Pham Van Huy, our instructor and mentor, for his valuable guidance and support throughout the midterm project report of our project.

Contents

[● Part 1: Research Models 3](#_Toc198485424)

[● Convolutional Neural Networks (CNNs): 3](#_Toc198485425)

[● R-CNN Family and Mask R-CNN 17](#_Toc198485426)

[● YOLO (You Only Look Once) 34](#_Toc198485427)

[● AutoEncoders (AE) and Variational AutoEncoders (VAE) 51](#_Toc198485428)

[● GAN 69](#_Toc198485429)

[● Vision Transformers (ViT) 69](#_Toc198485430)

[● Diffusion Models (e.g., Stable Diffusion) 69](#_Toc198485431)

[● Part 2: Demo Implementation & Explanations 70](#_Toc198485432)

[● Write here 70](#_Toc198485433)

[● Write here 70](#_Toc198485434)

# Part 1: Research Models

## Convolutional Neural Networks (CNNs):

CNNs are a specialized type of deep learning algorithm, particularly effective for processing and analyzing visual data like images and videos. Their architecture is inspired by the human visual cortex and allows them to automatically and adaptively learn spatial hierarchies of features.

**1. Input/Output (Problem Solved, Image Size, etc.)**

* Problem Solved: CNNs are primarily used for tasks involving visual data. The most common application is image classification, where the model assigns a label (e.g., "cat," "dog," "car") to an input image. They are also fundamental to other computer vision tasks like:
  + Object Detection: Identifying and locating multiple objects within an image (e.g., drawing bounding boxes around them and classifying them).
  + Image Segmentation: Classifying each pixel in an image to belong to a certain object or region.
  + Facial Recognition: Identifying or verifying a person from a digital image or a video frame.
  + Image Captioning: Generating a textual description of an image.
  + Video Analysis: Processing and understanding video content.
* Input:
  + Typically, the input is an image, represented as a 2D array of pixel values (for grayscale images) or a 3D array (for color images, e.g., with Red, Green, Blue channels).
  + The size of the input image can vary depending on the specific CNN architecture and the problem. Common input sizes for pre-trained models include 224x224 pixels, 227x227 pixels, or 299x299 pixels. However, CNNs can be designed to handle different image dimensions.
* Output:
  + Image Classification: A vector of probabilities, where each element corresponds to the likelihood that the input image belongs to a specific class. The class with the highest probability is typically chosen as the prediction.
  + Object Detection: A list of bounding boxes, class labels for each box, and confidence scores.
  + Image Segmentation: A segmentation map, which is an image where each pixel is assigned a class label.
  + Other tasks: The output format varies depending on the specific problem (e.g., a sequence of words for image captioning).

**2. Model Structures Used**

CNNs are a type of deep neural network characterized by their hierarchical structure, which typically includes several types of layers:

* Input Layer: This layer holds the raw pixel values of the input image.
* Convolutional Layer: This is the core building block of a CNN.
  + It applies a set of learnable filters (also called kernels) to the input image or feature maps from previous layers.
  + Each filter slides (convolves) across the width and height of the input, computing dot products between the filter entries and the input at any position.
  + This process produces a 2D activation map (or feature map) that highlights specific features (e.g., edges, corners, textures) detected by the filter.
  + Key characteristics:
    - Local Receptive Fields: Neurons in a convolutional layer are only connected to a small, local region of the input volume.
    - Parameter Sharing: The same filter is used across different spatial locations in the input, significantly reducing the number of parameters compared to fully connected networks and making the model more efficient and less prone to overfitting.
    - Multiple Filters: A convolutional layer typically learns multiple filters, each detecting a different feature.
* Activation Layer (e.g., ReLU):
  + After each convolutional operation, an activation function is applied element-wise to the feature map.
  + The Rectified Linear Unit (ReLU) (f(x) = max(0, x)) is the most commonly used activation function in CNNs. It introduces non-linearity into the model, allowing it to learn more complex patterns. Other activation functions like Sigmoid or Tanh can also be used but are less common in hidden layers of modern CNNs.
* Pooling Layer (e.g., Max Pooling, Average Pooling):
  + Pooling layers are used to reduce the spatial dimensions (width and height) of the feature maps, thereby reducing the number of parameters and computational complexity in the network.
  + They also help to make the representations learned by the network more robust to small translations in the input image (translation invariance).
  + Max Pooling: Selects the maximum value from a small rectangular region of the feature map.
  + Average Pooling: Calculates the average value from a small rectangular region.
* Fully Connected Layer (Dense Layer):
  + After several convolutional and pooling layers, the high-level reasoning in the neural network is done via fully connected layers.
  + Neurons in a fully connected layer have connections to all activations in the previous layer, as seen in regular Multi-Layer Perceptrons (MLPs).
  + The output of the convolutional and pooling layers (which are typically 3D volumes of feature maps) is flattened into a 1D vector before being fed into the fully connected layers.
  + These layers are typically used at the end of the network to perform the final classification or regression task based on the learned features.
* Output Layer:
  + The final layer of the network, which produces the desired output
  + For classification tasks, this layer often uses a Softmax activation function to output a probability distribution over the classes.
  + For regression tasks, it might use a linear activation function.

A typical CNN architecture stacks these layers sequentially: INPUT -> [[CONV -> ACT] \* N -> POOL?] \* M -> [FC -> ACT] \* K -> FC (OUTPUT) where *\*N, \*M, \*K* denote multiple repetitions, and *POOL?* indicates an optional pooling layer.

**3. Famous CNN Architectures:** LeNet-5, AlexNet, VGGNet, GoogLeNet (Inception), ResNet (Residual Networks), DenseNet.

**4. Datasets used**

CNNs are trained on large, labeled datasets. The nature of the dataset depends heavily on the specific task.

* **Image Classification Datasets:**
  + **MNIST:** A classic dataset of handwritten digits (0-9). It contains 60,000 training images and 10,000 testing images, each 28x28 pixels in grayscale. It's often used as a "hello world" for image classification.
  + **CIFAR-10/CIFAR-100:** Consists of 60,000 32x32 color images in 10 classes (CIFAR-10) or 100 classes (CIFAR-100), with 6,000 images per class. These are more complex than MNIST and are widely used for benchmarking.
  + **ImageNet:** A very large-scale dataset with millions of high-resolution color images belonging to thousands of classes (e.g., "cat," "dog," "car," "tree"). The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) played a crucial role in the advancement of deep learning, particularly CNNs. Models are often pre-trained on ImageNet and then fine-tuned for other tasks.
  + **Fashion-MNIST:** A dataset of Zalando's article images, intended as a direct drop-in replacement for the original MNIST dataset for benchmarking machine learning algorithms. It has the same image size, data format, and training/testing splits.
  + **SVHN (Street View House Numbers):** Images of house numbers collected by Google Street View. It's similar in style to MNIST but with more variability and a larger number of labeled example
* **Object Detection Datasets:**
  + **PASCAL VOC (Visual Object Classes):** Contains images with 20 object classes. Annotations include bounding boxes and class labels for objects in the images.
  + **COCO (Common Objects in Context):** A large-scale object detection, segmentation, and captioning dataset. It features 80 object categories with rich annotations, including bounding boxes and instance-level segmentation masks. It's known for its complexity due to multiple objects per image and contextual information.
  + **Open Images Dataset:** A very large dataset by Google with millions of images annotated with image-level labels, object bounding boxes, object segmentation masks, and visual relationships.
* **Image Segmentation Datasets:**
  + Many object detection datasets like PASCAL VOC and COCO also provide pixel-level segmentation masks.
  + **Cityscapes:** Focuses on semantic understanding of urban street scenes. It provides pixel-level annotations for 30 classes grouped into 8 categories (e.g., road, sky, vehicle, pedestrian).
* **Domain-Specific Datasets:** For specialized applications like medical imaging (e.g., datasets of X-rays, MRIs with annotations for tumors or diseases), satellite imagery, or autonomous driving, specific datasets are curated.

**5. Data Augmentation:** To increase the diversity and size of training data and prevent overfitting, various data augmentation techniques are commonly applied. These include: \* Flipping (horizontal/vertical) \* Rotation \* Scaling \* Cropping \* Translation \* Changing brightness, contrast, or saturation \* Adding noise

**6. How the models trained**

Training a CNN involves an iterative process of feeding training data to the model, calculating the error in its predictions, and adjusting the model's parameters (weights and biases of the filters and neurons) to minimize this error.

* **Loss Function:** A loss function quantifies the difference between the model's prediction and the actual ground truth label.
  + For classification: **Cross-Entropy Loss** (or Log Loss) is commonly used.
  + For regression tasks (e.g., predicting bounding box coordinates): Mean Squared Error (MSE) or Smooth L1 Loss might be used.
* **Optimizer:** An optimization algorithm is used to update the model's parameters in the direction that minimizes the loss function.
  + **Stochastic Gradient Descent (SGD):** A foundational optimization algorithm.
  + **Adam (Adaptive Moment Estimation):** A popular and often effective optimizer that adapts the learning rate for each parameter.
  + **RMSprop, Adagrad, Adadelta:** Other adaptive learning rate optimizers.
* **Backpropagation:** This is the algorithm used to calculate the gradients of the loss function with respect to each parameter in the network. These gradients indicate how much each parameter contributed to the error. The optimizer then uses these gradients to update the parameters.
* **Batch Training:** Instead of processing one image at a time (which can be noisy) or the entire dataset at once (which can be computationally expensive), training is typically done in **mini-batches**. A small batch of images is processed, the average loss is computed, and parameters are updated.
* **Epochs:** An epoch is one complete pass through the entire training dataset. Training usually involves running for multiple epochs until the model's performance on a validation set stops improving (or starts to degrade, indicating overfitting).
* **Learning Rate:** A crucial hyperparameter that controls the step size during parameter updates. Setting an appropriate learning rate is important for effective training. Learning rate schedules (e.g., decreasing the learning rate over time) are often used.
* **Initialization:** The initial values of the weights are important. Poor initialization can lead to slow convergence or getting stuck in poor local minima. Common initialization strategies include Xavier/Glorot initialization or He initialization.
* **Regularization:** Techniques to prevent overfitting, where the model performs well on training data but poorly on unseen data.
  + **L1/L2 Regularization (Weight Decay):** Adds a penalty to the loss function based on the magnitude of the weights.
  + **Dropout:** Randomly "drops out" (sets to zero) a fraction of neurons during training, forcing the network to learn more robust features.
  + **Batch Normalization:** Normalizes the activations of a layer, which can help stabilize training, speed up convergence, and act as a regularizer.
  + **Early Stopping:** Monitor the model's performance on a validation set during training and stop training when the validation performance starts to degrade.

**7. How to implement (the algorithm, the configuration/settings…)**

Implementing CNNs typically involves using deep learning frameworks, which provide pre-built layers, optimizers, and training loops.

* **Frameworks:**
  + **TensorFlow (with Keras API):** A popular open-source library developed by Google. Keras is a high-level API for building and training models that is now integrated into TensorFlow.
  + **PyTorch:** An open-source library developed by Facebook's AI Research lab (FAIR). Known for its flexibility and Pythonic feel, especially popular in research.
  + **JAX:** A Google library for high-performance numerical computing and machine learning research, often used for its autograd and XLA (Accelerated Linear Algebra) compiler.
* **General Steps for Implementation:**
  + **Load and Preprocess Data:**
    - Load images from disk or a dataset repository.
    - Resize images to a consistent input size for the network.
    - Normalize pixel values (e.g., scale to [0, 1] or [-1, 1], or standardize by subtracting the mean and dividing by the standard deviation of the training set).
    - Convert labels to the appropriate format (e.g., one-hot encoding for classification).
    - Split data into training, validation, and test sets.
    - Apply data augmentation to the training set.
  + **Define the Model Architecture:**
    - Use the framework's API to stack layers: convolutional, activation, pooling, fully connected.
    - Specify parameters for each layer (e.g., number of filters, filter size, stride, padding for convolutional layers; number of units for fully connected layers).
    - Choose activation functions.
  + **Compile the Model (TensorFlow/Keras specific step):**
    - Specify the optimizer (e.g., Adam, SGD).
    - Specify the loss function (e.g., categorical cross-entropy).
    - Specify metrics to monitor during training (e.g., accuracy).
  + **Train the Model (Fit the Model):**
    - Provide the training data and validation data.
    - Specify the batch size and number of epochs.
    - The framework handles the forward pass, loss calculation, backpropagation, and parameter updates.
    - Callbacks can be used for tasks like saving the best model, adjusting the learning rate, or early stopping.
  + **Evaluate the Model:**
    - Use the test set (which the model has not seen during training) to assess its generalization performance.
    - Calculate metrics like accuracy, precision, recall, F1-score, confusion matrix.
  + **Make Predictions (Inference):**
    - Use the trained model to predict new, unseen images.
* **Configuration/Settings (Hyperparameters):**
  + **Network Architecture:** Number of layers, types of layers, number of filters/units in each layer, filter sizes, stride, padding.
  + **Optimizer:** Choice of optimizer (Adam, SGD, etc.), learning rate, momentum (for SGD), beta1/beta2 (for Adam).
  + **Batch Size:** Number of samples processed before the model is updated.
  + **Number of Epochs:** Number of times the entire training dataset is passed through the model.
  + **Regularization Parameters:** Dropout rate, L1/L2 penalty strength.
  + **Initialization Scheme:** How initial weights are set.

Finding the optimal set of hyperparameters often requires experimentation and techniques like grid search, random search or more advanced hyperparameter optimization methods.

**8. How is the model performance? (accuracy, …)**

CNN performance is highly dependent on the specific architecture, dataset, task, and amount of training.

* **Metrics:**
  + **Accuracy:** The most common metric for classification tasks. It's the ratio of correctly classified images to the total number of images.
    - Accuracy = (True Positives + True Negatives) / (Total Samples)
  + **Top-k Accuracy:** For multi-class classification, an image is considered correctly classified if the true label is among the top k predictions made by the model (e.g., Top-5 accuracy for ImageNet).
  + **Precision, Recall, F1-Score:** Useful especially for imbalanced datasets.
    - Precision = True Positives / (True Positives + False Positives)
    - Recall (Sensitivity) = True Positives / (True Positives + False Negatives)
    - F1-Score = 2 \* (Precision \* Recall) / (Precision + Recall)
  + **Confusion Matrix:** A table that visualizes the performance of a classification model, showing the counts of true positive, true negative, false positive, and false negative predictions for each class.
  + **Area Under the ROC Curve (AUC-ROC):** A measure of a classifier's ability to distinguish between classes.
  + **For Object Detection:** Mean Average Precision (mAP) is a standard metric, which involves calculating the average precisionacross all classes and/or different Intersection over Union (IoU) thresholds.
  + **For Segmentation:** Pixel Accuracy, Mean IoU (Intersection over Union).
* **General Performance Trends:**
  + **Deeper and more complex models** (like ResNet, Inception) generally achieve higher accuracy on challenging datasets like ImageNet but require more computational resources and data.
  + **Transfer Learning:** Using a CNN pre-trained on a large dataset (like ImageNet) and then fine-tuning it on a smaller, specific dataset often leads to significantly better performance than training from scratch, especially when the target dataset is small.
  + **State-of-the-art (SOTA) models** on benchmarks like ImageNet have achieved very high Top-1 and Top-5 accuracies, often surpassing human performance in specific image classification tasks. For example, accuracies well above 80% (Top-1) and 95% (Top-5) are common for SOTA models on ImageNet.
  + Performance can be affected by factors like dataset quality and size, class imbalance, choice of hyperparameters, and regularization techniques.

**9. Some outputs**

* **Image Classification:**
  + Input: Image of a cat.
  + Output: {"class": "cat", "probability": 0.95} or a vector like [0.01 (dog), 0.95 (cat), 0.02 (bird), 0.02 (car)]
* **Object Detection:**
  + Input: Image with a dog and a car.
  + Output: A list of detections, e.g.:
    - {"bounding\_box": [x1, y1, x2, y2], "class": "dog", "score": 0.92}
    - {"bounding\_box": [x3, y3, x4, y4], "class": "car", "score": 0.88}
* **Image Segmentation:**
  + Input: Image of a street scene.
  + Output: A segmentation map (an image of the same size as the input) where each pixel is colored or labeled according to its class (e.g., pixels belonging to "road" are colored blue, "car" pixels are red, "sky" pixels are green).

## R-CNN Family and Mask R-CNN

These models represent an evolution in accurately detecting and classifying objects within an image, with Mask R-CNN extending this to instance segmentation (pixel-level masks for each object).

**1. R-CNN (Regions with CNN features) Family**

The R-CNN family of models revolutionized object detection by combining region proposals with Convolutional Neural Networks.

* R-CNN (Original):
  + Idea: First, generate a set of candidate object regions (region proposals) in an image using an algorithm like Selective Search. Then, for each region, warp it to a fixed size and feed it into a CNN to extract features. Finally, classify these features with SVMs and refine the bounding boxes with a linear regressor.
  + Drawbacks: Very slow due to processing thousands of region proposals independently through a CNN, and a multi-stage training pipeline.
* Fast R-CNN:
  + Improvement: Instead of feeding each region proposal to the CNN independently, Fast R-CNN feeds the entire image to the CNN once to generate a feature map. Then, for each region proposal, it extracts a fixed-length feature vector from the feature map using a RoI (Region of Interest) Pooling layer. These features are then fed into fully connected layers for classification and bounding box regression.
  + Advantage: Much faster than R-CNN and allows for end-to-end training (except for region proposal generation).
* Faster R-CNN:
  + Improvement: Introduced a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. The RPN is trained to predict object bounds and "objectness" scores at each location.
  + Advantage: Achieved significantly faster object detection by integrating region proposal generation into the neural network. This made object detection systems truly end-to-end trainable and much faster.
* **Mask R-CNN**

1. Input/Output (Problem Solved, Image Size, etc.)

* **Problem Solved:** Mask R-CNN is a framework for instance segmentation. This means it not only detects and classifies individual objects in an image (like Faster R-CNN) but also generates a high-quality segmentation mask for each detected instance (i.e., it tells you which pixels belong to each object).
  + It can perform:
    - Object Detection: Locating objects with bounding boxes and classifying them.
    - Instance Segmentation: Generating pixel-level masks for each distinct object instance.
    - Keypoint Detection (Optionally): Can be extended to predict keypoints for objects like human poses.
* Input:
  + An image, typically a 3-channel color image (RGB).
  + The size of the input image can vary. Mask R-CNN architectures (often built on backbones like ResNet) can handle various input sizes, though images are usually resized or padded to fit certain dimensions for batch processing, while maintaining aspect ratio. Common input sizes might be around 800x800 or 1024x1024 pixels, but this is flexible.
* Output: For each detected object instance in the image:
  + Class Label: The category of the object (e.g., "person," "car," "tree").
  + Confidence Score: A score indicating the model's confidence in the detection.
  + Bounding Box: Coordinates (x, y, width, height) defining the rectangular box around the object.
  + Segmentation Mask: A binary mask (an image of the same size as the object's bounding box or the original image, depending on implementation) where pixels belonging to the object instance are marked (e.g., as 1) and background pixels are marked as 0. This mask precisely outlines the shape of the object

**2. Model Structures Used (briefly introduce about the model)**

Mask R-CNN extends Faster R-CNN by adding a branch for predicting segmentation masks in parallel with the existing branches for classification and bounding box regression.

The overall architecture consists of several key components:

* Backbone Network:
  + This is a standard CNN (e.g., ResNet-50, ResNet-101, ResNeXt) that extracts low-level and mid-level features from the input image
  + Often, a Feature Pyramid Network (FPN) is used on top of the backbone. FPN generates multiple levels of feature maps with different spatial resolutions, allowing the model to detect objects at various scales more effectively. The FPN creates a top-down pathway with lateral connections to combine semantically strong features from deeper layers with spatially precise features from shallower layers.
* Region Proposal Network (RPN):
  + This component is identical to the RPN in Faster R-CNN.
  + It slides a small network over the feature maps generated by the backbone (or FPN) to propose candidate regions (Regions of Interest - RoIs) that might contain objects.
  + For each candidate region, it outputs an "objectness" score (probability of containing any object) and bounding box refinements.
* RoIAlign (Region of Interest Align):
  + This is a crucial improvement over RoIPool (used in Fast R-CNN).
  + RoIPooling involves quantization of RoI coordinates, which can lead to misalignment between the extracted features and the original RoI, negatively impacting mask prediction accuracy.
  + RoIAlign avoids this quantization. It uses bilinear interpolation to compute the exact values of the input features at four regularly sampled locations within each RoI bin and then aggregates them (e.g., using max or average pooling). This preserves precise spatial locations.
* Heads (Branches for each task):
  + After RoIAlign extracts features for each proposed RoI, these features are fed into several parallel "heads":
    1. Classification Head: Predicts the class label of the object within the RoI (e.g., using a fully connected layer followed by a softmax).
    2. Bounding Box Regression Head: Refines the coordinates of the bounding box proposed by the RPN to more accurately fit the object.
    3. Mask Prediction Head: This is the key addition in Mask R-CNN.
       - It's typically a small Fully Convolutional Network (FCN) applied to each RoI.
       - It predicts a binary segmentation mask (e.g., 28x28 pixels) for each RoI and for each class. This means if there are K classes, it outputs K masks, and the mask corresponding to the predicted class is chosen.
       - The small FCN allows the mask head to learn pixel-to-pixel correspondence, generating masks with fine spatial detail. The predicted small masks are then scaled up to the RoI size and binarized to get the final instance mask.

Decoupling of Mask and Class Prediction: An important design choice in Mask R-CNN is that the mask branch predicts a binary mask for each class independently, without competition among classes (using a per-pixel sigmoid activation followed by a binary loss for each class). The class prediction from the classification head is then used to select which mask to use. This decoupling improves instance segmentation performance.

**3. Describe something about the datasets used**

Mask R-CNN, being an instance segmentation model, requires datasets with more detailed annotations than just bounding boxes. Specifically, pixel-level masks for each object instance are needed.

* **COCO (Common Objects in Context):**
  + This is the primary dataset used for training and benchmarking Mask R-CNN and many other instance segmentation models.
  + COCO contains a large number of images (over 330K images, with over 200K labeled) featuring 80 common object categories (e.g., person, car, cat, banana, chair).
  + Crucially, it provides **pixel-level segmentation masks** for every object instance in the images, along with bounding boxes and class labels.
  + The dataset is challenging due to the presence of multiple objects per image, objects at various scales, and complex scenes with occlusions.
  + Standard evaluation metrics for instance segmentation on COCO are based on Average Precision (AP) over different Intersection over Union (IoU) thresholds for the masks.
* **PASCAL VOC (Visual Object Classes):**
  + While PASCAL VOC (especially VOC 2012) has segmentation annotations, it's primarily for semantic segmentation (labeling pixels by class, not by instance) and object detection.
  + Some efforts have been made to extend it or use it for instance segmentation, but COCO is more comprehensive for this task.
* **Cityscapes:**
  + This dataset focuses on semantic understanding of urban street scenes, primarily for autonomous driving applications.
  + It provides high-quality, dense pixel-level annotations for 30 classes grouped into 8 categories (road, car, pedestrian, etc.).
  + It also includes instance-level annotations for "vehicle" and "person" classes, making it suitable for training and evaluating instance segmentation models in urban environments.
* **LVIS (Large Vocabulary Instance Segmentation):**
  + A more recent dataset that aims to address the long-tail distribution of object categories found in the real world.
  + It contains over 160,000 images and annotations for over 1,200 object categories, including many rare objects.
  + This dataset is designed to push instance segmentation models to handle a much larger and more diverse set of objects.
* **Domain-Specific Datasets:**
  + For specialized applications, custom datasets are often created. Examples include:
    - **Medical Imaging:** Datasets of CT scans, MRIs, or microscopy images with instance-level masks for cells, tumors, organs, or other biological structures.
    - **Satellite Imagery:** Datasets with instance masks for buildings, roads, or other geographical features.
    - **Retail:** Datasets of products on shelves with instance masks for inventory management or automated checkout.
  + Creating these datasets is labor-intensive as it requires manual pixel-level annotation for each object instance.

**Data Augmentation:** Similar to CNNs, data augmentation is crucial for training robust Mask R-CNN models. Common techniques include: \* Horizontal flipping (masks must also be flipped). \* Scaling and resizing (masks and bounding boxes must be adjusted accordingly).

\* Rotation.

\* Color jittering (adjusting brightness, contrast, saturation). \* Sometimes, more advanced techniques like "copy-paste" augmentation (pasting object instances from one image onto another) are used, especially for rare classes.

**4. How the models trained**

Training Mask R-CNN involves optimizing a multi-task loss function that combines the losses from its different heads (classification, bounding box regression, and mask prediction).

* **Multi-Task Loss:** The total loss L is a weighted sum of the losses for each task:
  + L = L\_cls + L\_box + L\_mask
  + **L\_cls (Classification Loss):** This is the log loss (cross-entropy) for the predicted class of each RoI. It's calculated on the classification head.
  + **L\_box (Bounding Box Regression Loss):** This measures the error between the predicted bounding box coordinates and the ground-truth bounding box for positive RoIs. Smooth L1 loss is commonly used, similar to Faster R-CNN. It's calculated on the bounding box regression head.
  + **L\_mask (Mask Prediction Loss):** This is the core new loss component.
    - For each RoI associated with a ground-truth object, the mask head outputs a K x m x m mask, where K is the number of classes and m x m is the mask resolution (e.g., 28x28).
    - The loss L\_mask is defined as the average binary cross-entropy loss. It's applied only to the k-th mask, where k is the ground-truth class of the RoI.
    - This encourages the network to generate a precise binary mask for the correct class, independent of other classes.
* **Training Process:**
  + **Backbone Pre-training:** The backbone network (e.g., ResNet-FPN) is usually initialized with weights pre-trained on a large image classification dataset like ImageNet. This significantly speeds up convergence and improves performance.
  + **End-to-End Training:** The entire Mask R-CNN model (backbone, RPN, and heads) can be trained end-to-end.
  + **Positive/Negative RoIs:**
    - During training, RoIs proposed by the RPN are labeled as positive if their IoU (Intersection over Union) with a ground-truth bounding box is above a certain threshold (e.g., 0.5), and negative otherwise.
    - The classification and bounding box regression losses are calculated only for positive RoIs.
    - The mask loss is calculated only for positive RoIs that have a corresponding ground-truth mask.
  + **Optimizer:** Adam or SGD with momentum are commonly used.
  + **Learning Rate Schedule:** The learning rate is often decreased at certain epochs or based on validation performance.
  + **Batch Size:** Images are processed in mini-batches.
  + **Epochs:** The model is trained for a certain number of epochs, monitoring performance on a validation set to prevent overfitting and for early stopping.
* **RPN Training:** The RPN itself is trained with its own loss function, which includes a binary classification loss (object vs. non-object) for anchors and a regression loss for anchor box refinement. The RPN and the main Mask R-CNN network can be trained jointly or in an alternating fashion. In modern implementations, joint training is common.

**5. How to implement (the algorithm, the configuration/settings…)**

Implementing Mask R-CNN typically involves using established deep learning frameworks and often leveraging well-maintained open-source implementations.

* **Frameworks:**
  + **PyTorch:** Very popular for Mask R-CNN implementations. torchvision.models.detection.mask\_rcnn provides a pre-built Mask R-CNN model. Many research projects and custom implementations use PyTorch.
  + **TensorFlow:** TensorFlow Object Detection API provides implementations of Mask R-CNN. The original Matterport Mask R-CNN implementation was built on Keras and TensorFlow 1.x. TensorFlow 2.x also supports such models.
  + **Detectron2 (PyTorch-based):** Developed by Facebook AI Research (FAIR), Detectron2 is a ground-up rewrite of Detectron. It's a popular platform for object detection and segmentation research and provides highly optimized and modular implementations of Mask R-CNN and other models.
  + **MMDetection (PyTorch-based):** An open-source object detection toolbox based on PyTorch, which includes Mask R-CNN and a wide array of other detection and segmentation models.
* **Key Implementation Steps & Configuration:**
  + **Dataset Preparation:**
    - Format annotations correctly (bounding boxes, class labels, pixel-level masks). COCO format is a common standard.
    - Implement data loaders that handle image loading, preprocessing (resizing, normalization), and data augmentation.
  + **Model Definition:**
    - **Choose a Backbone:** ResNet-50/101 with FPN is a common choice.
    - **Configure RPN:** Set anchor scales and ratios, NMS (Non-Maximum Suppression) thresholds.
    - **Configure RoIAlign:** Output resolution of RoIAlign.
    - **Configure Heads:** Define the architecture of the classification, bounding box, and mask heads (number of layers, channels). The mask head typically outputs a K x m x m tensor.
  + **Loss Functions:** Implement or use framework-provided versions of L\_cls, L\_box, and L\_mask.
  + **Training Loop:**
    - Forward pass through the network.
    - Compute the multi-task loss.
    - Backward pass (backpropagation).
    - Optimizer step.
  + **Inference (Prediction):**
    - Feed an image through the backbone and RPN to get RoIs.
    - For each RoI, pass it through RoIAlign and the heads to get class scores, refined bounding boxes, and raw mask predictions (e.g., 28x28 per class).
    - Apply NMS to filter overlapping bounding boxes based on class scores.
    - For the final detected objects, select the mask corresponding to the predicted class.
    - Resize the predicted mask (e.g., 28x28) to the size of the detected bounding box in the original image.
    - Apply a threshold (e.g., 0.5) to the mask values to get a binary mask.
  + **Hyperparameters (examples):**
    - Learning rate, weight decay, momentum.
    - Batch size.
    - Number of training epochs.
    - RPN: anchor sizes/ratios, NMS threshold, number of proposals.
    - RoI selection: IoU thresholds for positive/negative samples, batch size per image for RoIs.
    - Mask head: output resolution (e.g., 28x28).
    - Pre-trained weights for the backbone.
* **Popular Open-Source Implementations:**
  + **Matterport Mask\_RCNN (TensorFlow 1.x, Keras):** One of the earliest and widely used open-source implementations. While a bit dated, it's well-documented and good for understanding the basics. (Link: https://github.com/matterport/Mask\_RCNN)
  + **Detectron2 (PyTorch):** Provides state-of-the-art, highly optimized implementations. (Link: https://github.com/facebookresearch/detectron2)
  + **MMDetection (PyTorch):** Another comprehensive toolbox. (Link: https://github.com/open-mmlab/mmdetection)
  + **Torchvision (PyTorch):** torchvision.models.detection.maskrcnn\_resnet50\_fpn offers a convenient pre-trained model.

**6. How is the model performance? (accuracy, …)**

Mask R-CNN set a new state-of-the-art for instance segmentation upon its release and remains a strong baseline.

* **Metrics:**
  + The primary metric for instance segmentation is **Average Precision (AP)**, typically evaluated on the COCO dataset.
  + AP is calculated for different IoU (Intersection over Union) thresholds between the predicted mask and the ground-truth mask. For example:
    - AP: Averaged over multiple IoU thresholds (e.g., from 0.5 to 0.95 with a step of 0.05). This is the main COCO metric.
    - AP\_50: AP at IoU threshold of 0.5 (looser localization).
    - AP\_75: AP at IoU threshold of 0.75 (stricter localization).
  + AP is also reported for different object sizes: AP\_small, AP\_medium, AP\_large.
  + For the object detection part, standard detection AP metrics (based on bounding box IoU) are also reported (AP\_bb).
* **Performance Levels:**
  + On the COCO dataset, Mask R-CNN with a ResNet-101-FPN backbone can achieve around **37-40 AP** for instance segmentation (mask AP) and a similar or slightly higher AP for bounding box detection.
  + Stronger backbones (e.g., ResNeXt-101-FPN) and various training improvements (longer schedules, more data augmentation, multi-scale training/testing) can push the performance higher, into the low to mid 40s for mask AP.
  + The original Mask R-CNN paper reported around 37.1 mask AP with ResNet-101-FPN on COCO test-dev.
  + Performance depends heavily on the backbone, training data, hyperparameters, and implementation details.
  + It generally provides high-quality masks that accurately delineate object boundaries.
* **Speed:**
  + Mask R-CNN is computationally more intensive than object detection-only models like Faster R-CNN or YOLO due to the added mask prediction branch.
  + Inference speed can be around 5 frames per second (FPS) on a high-end GPU with a ResNet-101 backbone, making it suitable for many applications but potentially too slow for real-time requirements on constrained hardware. Lighter backbones can improve speed at the cost of some accuracy.

**7. Some outputs**

* **Input:** An image containing multiple people, cars, and a dog.
* **Output:** For each detected instance:
  + **Instance 1 (Person):**
    - class: "person"
    - score: 0.98
    - bounding\_box: [x1, y1, w1, h1]
    - mask: a binary 2D array representing the pixels belonging to this specific person.
  + **Instance 2 (Car):**
    - class: "car"
    - score: 0.95
    - bounding\_box: [x2, y2, w2, h2]
    - mask: a binary 2D array representing the pixels belonging to this specific car.
  + **Instance 3 (Person):**
    - class: "person"
    - score: 0.92
    - bounding\_box: [x3, y3, w3, h3]
    - mask: a binary 2D array representing the pixels belonging to this other person (distinct from Instance 1).
  + **Instance 4 (Dog):**
    - class: "dog"
    - score: 0.89
    - bounding\_box: [x4, y4, w4, h4]
    - mask: a binary 2D array representing the pixels belonging to this dog.

Visually, the output would be the original image with bounding boxes drawn around each detected object, the class label and score displayed next to the box, and the object itself highlighted with its specific segmentation mask (often shown as a colored overlay).

## YOLO (You Only Look Once)

YOLO is a popular family of real-time object detection algorithms. The core idea is to reframe object detection as a single regression problem, directly from image pixels to bounding box coordinates and class probabilities. This contrasts with region-proposal-based methods like R-CNN and its variants.

**1. Input/Output (Problem Solved, Image Size, etc.)**

* Problem Solved: YOLO is designed for real-time object detection. It aims to identify all objects within an image and draw bounding boxes around them, along with classifying each object. Unlike two-stage detectors (like R-CNN family) that first propose regions and then classify/refine them, YOLO performs these tasks in a single pass through the network.
* Input:
  + Typically a 3-channel color image (RGB).
  + YOLO models require the input image to be resized to a fixed square dimension. Common input sizes for different YOLO versions include:
    - YOLOv1: 448x448 pixels
    - YOLOv2 (YOLO9000): Often 416x416 or other multiples of 32.
    - YOLOv3: Commonly 320x320, 416x416, or 608x608 pixels. The choice often balances speed and accuracy.
    - YOLOv4, YOLOv5, and subsequent versions (YOLOv6, v7, v8, YOLO-NAS, etc.): Continue to use similar square input sizes, often configurable (e.g., 640x640 is very common for YOLOv5 and later).
  + The aspect ratio of the original image is usually maintained by padding the shorter dimension after resizing.
* Output:
  + YOLO outputs a set of bounding boxes for detected objects, along with a class label and a confidence score for each box.
  + Specifically, the output is typically a tensor (or a set of tensors in later versions that detect at multiple scales) that encodes:
    - Bounding Box Coordinates: For each detected object, usually represented as (x\_center, y\_center, width, height), normalized with respect to the image dimensions or grid cell size.
    - Objectness Score (or Confidence Score): A score indicating the probability that the bounding box actually contains an object and how well the predicted box fits the object. It's often a product of P(Object) \* IoU(predicted, truth).
    - Class Probabilities: For each bounding box, a vector of probabilities for each class the model can detect (e.g., if the model detects 80 classes, this is an 80-element vector). The class with the highest probability is chosen as the object's class.

**2. Model Structures Used (briefly introduce about the model)**

YOLO's core idea is to reframe object detection as a single regression problem, directly from image pixels to bounding box coordinates and class probabilities.

* Grid System:
  + YOLO divides the input image into an S x S grid of cells (e.g., 7x7 in YOLOv1, 13x13, 26x26, 52x52 in later versions for multi-scale detection).
  + Each grid cell is responsible for detecting objects whose centers fall within that cell.
* Unified Architecture (Single CNN):
  + The entire image is processed by a single Convolutional Neural Network (CNN) in one forward pass.
  + This CNN backbone extracts features from the image. Early versions used custom architectures (like a modified GoogLeNet in YOLOv1, Darknet-19 in YOLOv2, Darknet-53 in YOLOv3). Later versions (YOLOv4, v5, and beyond) often use more advanced backbones (e.g., CSPDarknet, EfficientNet-based, or custom lightweight backbones) and incorporate architectural improvements like Feature Pyramid Networks (FPN), Path Aggregation Networks (PANet), and other neck structures to better fuse features from different scales.
* Bounding Box Prediction:
  + For each grid cell, the model predicts:
    - B bounding boxes (e.g., B=2 in YOLOv1, B=3 or more in later versions per scale).
    - For each of these B boxes, it predicts:
      * Coordinates (x, y, w, h) relative to the grid cell or image dimensions.
      * An objectness/confidence score.
  + Anchor Boxes (introduced in YOLOv2 and used thereafter): Instead of directly predicting the width and height of bounding boxes, YOLOv2 and later versions predict offsets relative to pre-defined default box shapes called "anchor boxes" (or "prior boxes"). These anchors are determined by running k-means clustering on the bounding box dimensions from the training dataset. Using anchors helps the network learn to predict common object shapes more easily and improves localization accuracy. Each grid cell is associated with a set of anchor boxes of different aspect ratios and scales.
* Class Prediction:
  + Independently of the bounding box predictions, each grid cell (in YOLOv1) or each anchor box (in later versions) also predicts a set of class probabilities P(Class\_i|Object). This is conditioned on an object being present in the grid cell/anchor.
* Multi-Scale Detection (YOLOv3 and later):
  + To better detect objects of various sizes, YOLOv3 and subsequent versions make predictions at multiple feature map resolutions (scales). For example, predictions might be made on feature maps of size 13x13 (for large objects), 26x26 (for medium objects), and 52x52 (for small objects). This is often achieved using an FPN-like structure. Each scale uses a different set of anchor boxes appropriate for the object sizes it's trying to detect.
* Output Tensor:
  + The final output of the network is typically a tensor of shape

S x S x (B \* (5 + C)) for YOLOv1, or a set of such tensors for multi-scale versions.

* + - S x S: Grid size.
    - B: Number of bounding boxes predicted per grid cell.
    - 5: Corresponds to 4 bounding box coordinates (x, y, w, h) and 1 objectness/confidence score.
    - C: Number of object classes.
  + For versions with anchor boxes, the structure is more like S x S x Num\_Anchors x (5 + C).
* Non-Maximum Suppression (NMS):
  + Since multiple grid cells or anchors might predict bounding boxes for the same object, a post-processing step called Non-Maximum Suppression (NMS) is applied.
  + NMS filters out redundant, overlapping bounding boxes. It keeps the box with the highest confidence score for an object and suppresses other boxes that have a high IoU (Intersection over Union) with it and detect the same class.

Evolution of YOLO Architectures:

* YOLOv1: The original, simple concept. Used a custom network based on GoogLeNet. Suffered from lower recall and localization errors compared to two-stage detectors.
* YOLOv2 (YOLO9000): Introduced anchor boxes, batch normalization, higher resolution input, and a new backbone (Darknet-19). Significantly improved accuracy and speed. YOLO9000 variant could detect over 9000 categories by jointly training on detection and classification datasets.
* YOLOv3: Used a deeper backbone (Darknet-53), introduced multi-scale predictions (using an FPN-like structure), and made changes to the loss function. Further improved accuracy, especially for small objects, while maintaining good speed.
* YOLOv4: Focused on optimizing training and architecture by incorporating a "bag of freebies" (BoF - techniques that improve accuracy without increasing inference cost, like data augmentation) and "bag of specials" (BoS - techniques that slightly increase inference cost but significantly improve accuracy, like new activation functions, attention mechanisms). Used CSPDarknet53 backbone, PANet neck, and SPP (Spatial Pyramid Pooling).
* YOLOv5 (Ultralytics): Not a direct research paper successor but a very popular open-source implementation that continued to improve upon YOLO concepts with a focus on usability, speed, and accuracy. Implemented in PyTorch. Introduced auto-anchor generation and further architectural refinements. Many subsequent "YOLO" versions have been released by various groups (e.g., YOLOv6, v7, v8, YOLOX, PP-YOLO, DAMO-YOLO, YOLO-NAS). These often introduce new backbones, neck designs, label assignment strategies, and loss functions to push performance boundaries.

**3. Describe something about the datasets used**

YOLO models, being focused on object detection, are trained on datasets that provide images along with bounding box annotations and class labels for objects within those images.

* **PASCAL VOC (Visual Object Classes):**
  + The VOC 2007 and VOC 2012 datasets were commonly used for training and benchmarking early versions of YOLO.
  + These datasets contain images with 20 object classes (e.g., person, car, bird, cat, chair).
  + Annotations include bounding boxes and class labels.
* **COCO (Common Objects in Context):**
  + This is the most widely used dataset for training and evaluating modern YOLO models (YOLOv3 and later).
  + COCO features 80 object categories and provides a large number of images with complex scenes, multiple objects per image, and objects at various scales.
  + It has become the standard benchmark for object detection. Training on COCO allows YOLO models to learn a diverse set of object appearances and contexts.
* **ImageNet (for pre-training):**
  + While ImageNet is primarily an image classification dataset, the backbones of YOLO models (e.g., Darknet, CSPDarknet) are often pre-trained on ImageNet. This pre-training helps the network learn general visual features, which can then be fine-tuned for the object detection task on datasets like COCO or PASCAL VOC. This significantly improves convergence and final performance.
* **Open Images Dataset (OID):**
  + A very large dataset by Google, containing millions of images with annotations for a wide variety of object classes (around 600 categories with bounding boxes).
  + Its scale and diversity make it suitable for training more general and robust object detection models. Some YOLO variants or custom-trained models might use OID.
* **Custom Datasets:**
  + A major strength of the YOLO ecosystem (especially with frameworks like Ultralytics YOLOv5, YOLOv8) is the ease of training on custom datasets.
  + Users can collect their own images and annotate them (or use annotation tools/services like Roboflow, LabelImg, CVAT) to create datasets for specific applications, such as:
    - Detecting specific types of defects in manufacturing.
    - Identifying particular animal species in wildlife monitoring.
    - Detecting traffic signs or pedestrians for autonomous driving research.
    - Recognizing products on retail shelves.
  + The YOLO data format is relatively simple, typically requiring a text file for each image, where each line specifies class\_id x\_center y\_center width height (normalized coordinates) for an object.
* **Data Augmentation:**
  + Extensive data augmentation is crucial for training high-performing YOLO models, especially to prevent overfitting and improve generalization. Common techniques include:
    - **Geometric augmentations:** Scaling, translation, rotation, shearing, flipping (horizontal).
    - **Photometric augmentations:** Adjusting brightness, contrast, saturation, hue.
    - **Mosaic augmentation (popular in YOLOv4/v5 and later):** Combines four training images into one, teaching the model to identify objects at smaller scales and in unusual contexts.
    - **MixUp:** Creates new training samples by linearly interpolating between pairs of images and their labels.
    - **Copy-Paste:** Pasting objects from one image onto another.

**4. How the models trained**

Training YOLO involves optimizing a complex loss function that accounts for errors in bounding box localization, objectness confidence, and class prediction.

* **Loss Function:** The YOLO loss function is a multi-part loss that sums errors from different aspects of the prediction.
  + **Localization Loss (Bounding Box Regression Loss):** Penalizes errors in the predicted bounding box coordinates (x, y, w, h). Typically uses Sum of Squared Errors (SSE) in early versions or Complete IoU (CIoU) / Distance IoU (DIoU) / Generalized IoU (GIoU) loss in later versions. This loss is only calculated if an object is present in a grid cell/anchor.
  + **Confidence Loss (Objectness Loss):**
    - Penalizes the model if it's not confident when an object is present in a grid cell/anchor (P(Object) should be high).
    - Penalizes the model if it's confident when no object is present (P(Object) should be low).
    - This is often a binary cross-entropy or similar loss. Different weights are usually applied to cells containing objects versus those not containing objects (the λ\_noobj parameter, typically small, addresses the imbalance as most cells don't contain objects).
  + **Classification Loss:** If an object is present, this penalizes errors in the predicted class probabilities. Typically uses cross-entropy loss.

The overall loss is a weighted sum of these components. For example, in YOLOv1: L = λ\_coord \* L\_localization + L\_confidence\_obj + λ\_noobj \* L\_confidence\_noobj + λ\_class \* L\_classification (Later versions refine these components and their weighting).

* **Training Process:**
  + **Backbone Pre-training:** The CNN backbone is often pre-trained on ImageNet.
  + **End-to-End Training on Detection Dataset:** The entire network is then trained on an object detection dataset like COCO or PASCAL VOC.
  + **Anchor Box Assignment:** For versions using anchor boxes (YOLOv2+), each ground-truth object is assigned to a specific grid cell (the one containing the object's center) and a specific anchor box within that cell (usually the one with the highest IoU with the ground-truth box). Only this assigned anchor box is responsible for predicting that object.
  + **Optimization:** SGD with momentum or Adam optimizer.
  + **Learning Rate Schedule:** Often starts with a "warm-up" phase (gradually increasing the learning rate) followed by a schedule that decreases the learning rate over time (e.g., cosine annealing, step decay).
  + **Batch Size:** Training is done in mini-batches.
  + **Epochs:** Trained for a specific number of epochs.
  + **Multi-scale Training (YOLOv3 and later):** To improve robustness to object scale variations, some YOLO versions change the input image resolution randomly during training.

**5. How to implement (the algorithm, the configuration/settings…)**

Implementing YOLO from scratch is a complex undertaking. Most users leverage existing well-maintained frameworks and open-source implementations.

* **Frameworks & Implementations:**
  + **Darknet:** The original framework for YOLO (up to YOLOv4), written in C and CUDA by Joseph Redmon. It's known for being efficient but can be harder to modify for research compared to Python-based frameworks.
  + **Ultralytics YOLO (YOLOv5, YOLOv8, etc.):** Extremely popular PyTorch-based implementations. They are well-documented, easy to use for training on custom datasets, offer pre-trained models, and have a large community. (e.g., https://github.com/ultralytics/yolov5, https://github.com/ultralytics/ultralytics)
  + **MMDetection (PyTorch):** A comprehensive open-source object detection toolbox that includes implementations of various YOLO versions and other detectors.
  + **TensorFlow Hub / TensorFlow Object Detection API:** Provide pre-trained YOLO models and tools for training.
  + Many other independent PyTorch and TensorFlow implementations of different YOLO versions exist on GitHub.
* **General Steps for Using/Implementing YOLO (e.g., with Ultralytics YOLO):**
  + **Setup Environment:** Install Python, PyTorch (with CUDA support for GPU training), and clone the YOLO repository (e.g., Ultralytics).
  + **Prepare Dataset:**
    - Organize images and create annotation files in the YOLO format (e.g., .txt file per image with class\_id x\_center y\_center width height).
    - Create a dataset configuration file (e.g., a .yaml file) that specifies paths to training/validation images and lists the class names.
  + **Choose a Model Configuration:**
    - YOLO frameworks provide different model sizes (e.g., YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, or YOLOv8n, YOLOv8s, etc.). Smaller models are faster but less accurate; larger models are more accurate but slower.
    - Select a pre-trained model (e.g., weights pre-trained on COCO) as a starting point for fine-tuning on a custom dataset, or train from scratch.
  + **Train the Model:**
    - Use the provided training script/command, specifying:
      * Path to the dataset configuration file.
      * Model configuration/weights to use.
      * Batch size.
      * Number of epochs.
      * Image size.
      * Other hyperparameters (learning rate, optimizer, data augmentation settings).
    - Training progress (loss, metrics like mAP) is usually logged, and model checkpoints are saved.
  + **Evaluate the Model:**
    - Use the validation set to evaluate the trained model's performance (mAP, precision, recall).
  + **Make Predictions (Inference):**
    - Use the trained model weights to detect objects in new images, videos, or camera streams.
    - Inference scripts usually allow setting a confidence threshold (to filter out low-confidence detections) and an IoU threshold for NMS.
* **Key Configuration/Settings (Hyperparameters):**
  + **Model Architecture:** Choice of YOLO version and size (e.g., YOLOv5s vs. YOLOv5x).
  + **Input Image Size:** (e.g., 320, 416, 640).
  + **Batch Size.**
  + **Number of Epochs.**
  + **Optimizer:** (e.g., Adam, SGD).
  + **Learning Rate:** Initial learning rate, learning rate schedule.
  + **Data Augmentation Settings:** Enabling/disabling specific augmentations and their parameters (e.g., mosaic, mixup, flip probability, scale factor).
  + **Anchor Boxes:** While modern implementations like YOLOv5/v8 can automatically learn optimal anchor boxes for a custom dataset, these could also be manually configured.
  + **Loss Function Hyperparameters:** Weights for different loss components (e.g., λ\_coord, λ\_noobj).

**6. How is the model performance? (accuracy, …)**

YOLO models are known for their excellent balance between speed and accuracy, particularly for real-time applications.

* **Metrics:**
  + **mAP (mean Average Precision):** The primary metric for object detection accuracy. Commonly reported at an IoU threshold of 0.5 (mAP@0.5 or mAP50) and averaged over multiple IoU thresholds from 0.5 to 0.95 (mAP@[.5:.95] or mAP50-95), following COCO evaluation standards.
  + **Precision and Recall:** Also important, especially for understanding trade-offs.
  + **F1-Score:** The harmonic mean of precision and recall.
  + **FPS (Frames Per Second):** A critical metric for real-time performance, indicating how many images the model can process per second. This heavily depends on the hardware (GPU/CPU), model size, and input resolution.
* **General Performance Trends:**
  + **Speed:** YOLO models are significantly faster than two-stage detectors like Faster R-CNN. Many YOLO variants can achieve real-time performance (>30 FPS) even on moderate GPUs, with lightweight versions running even faster.
  + **Accuracy:**
    - Early YOLO versions (v1, v2) were faster but generally less accurate than contemporary two-stage detectors, especially for small objects.
    - YOLOv3 significantly improved accuracy while maintaining speed.
    - YOLOv4, YOLOv5, and subsequent versions have pushed accuracy to levels comparable to or even exceeding many two-stage detectors, while often being much faster. For example, on the COCO dataset:
      * YOLOv3 achieved around 33.0 mAP@[.5:.95].
      * YOLOv4 achieved around 43.5 mAP@[.5:.95].
      * YOLOv5 models (e.g., YOLOv5l) can achieve mAP@[.5:.95] in the high 40s to low 50s.
      * YOLOv8 models show further improvements, with larger variants reaching well into the 50s for mAP@[.5:.95] on COCO.
  + **Trade-off:** There's a clear trade-off between model size/complexity, speed, and accuracy. Smaller YOLO models (e.g., YOLOv8n - nano) are very fast but less accurate, while larger models (e.g., YOLOv8x - extra large) are more accurate but slower and require more computational resources. This allows users to choose a model that best fits their specific application requirements.

**7. Some outputs**

* **Input:** An image from a street scene.
* **Output (typically after NMS, visualized on the image):**
  + A bounding box around a car with label "car" and confidence score "0.92".
  + A bounding box around a pedestrian with label "person" and confidence score "0.88".
  + A bounding box around a traffic light with label "traffic light” and confidence score "0.75".
* **Raw Output (conceptual, before NMS and rendering):** A list of detected objects, each with:
  + bounding\_box: [x\_center, y\_center, width, height] (normalized)
  + objectness\_score: 0.95
  + class\_probabilities: [0.01 (person), 0.92 (car), ..., 0.02 (traffic light)] (leading to "car" as the predicted class)

## AutoEncoders (AE) and Variational AutoEncoders (VAE)

AutoEncoders are unsupervised neural networks that learn efficient data codings (representations) in a compressed, lower-dimensional form. Variational AutoEncoders are a generative extension of AEs that learn a probabilistic distribution of the latent space.

**1. Input/Output (Problem Solved, Image Size, etc.)**

* **Problem Solved:**
  + **Dimensionality Reduction/Feature Learning:** Autoencoders are trained to learn a compressed representation (encoding) of the input data. The goal is to reconstruct the original input from this compressed representation as accurately as possible. The learned encoding can serve as a lower-dimensional feature set.
  + **Data Denoising:** If trained with noisy input and clean target output, autoencoders can learn to remove noise from data.
  + **Anomaly Detection:** Autoencoders trained on "normal" data will have higher reconstruction errors for anomalous or out-of-distribution inputs, which can be used to detect anomalies.
  + **Data Compression (conceptual):** While they learn compressed representations, they are not always as efficient as dedicated compression algorithms like JPEG or PNG for images, but the learned features can be more semantically meaningful.
* **Input:**
  + Can be various types of data: images, vectors (e.g., embeddings, sensor readings), sequences.
  + For images, this would be the raw pixel data, typically flattened into a vector for basic autoencoders, or processed by convolutional layers in Convolutional Autoencoders.
  + Image size can vary. If using fully connected layers, images are often resized to a fixed dimension and flattened. Convolutional autoencoders can handle image dimensions more naturally.
* **Output:**
  + The primary output is a **reconstruction of the input data**. The goal is for this output to be as close to the original input as possible.
  + The **bottleneck layer's activation (the encoding)** is also a crucial output, representing the compressed, lower-dimensional representation of the input.

**2. Model Structures Used (briefly introduce about the model)**

An autoencoder consists of two main parts: an encoder and a decoder.

* **Encoder:**
  + This part of the network takes the input data and maps it to a lower-dimensional representation called the **latent space** or **bottleneck** (also known as the encoding or code).
  + It typically consists of a series of layers (e.g., fully connected layers or convolutional layers) that progressively reduce the dimensionality of the input.
  + Activation functions (e.g., ReLU, sigmoid) are used in these layers.
  + h = f(x) where x is the input and h is the encoding.
* **Bottleneck (Latent Space):**
  + This is the layer with the smallest number of neurons in the network (for a standard undercomplete autoencoder). It holds the compressed representation h.
  + The dimensionality of this layer determines the degree of compression.
* **Decoder:**
  + This part of the network takes the latent representation h from the bottleneck and attempts to reconstruct the original input data from it.
  + It's typically a mirror image of the encoder, consisting of layers that progressively increase the dimensionality back to the original input's dimension.
  + If the encoder uses convolutional layers and pooling, the decoder will use deconvolutional (or transposed convolutional) layers and upsampling.
  + r = g(h) where r is the reconstructed input.
* **Types of Autoencoders:**
  + **Undercomplete Autoencoder:** The bottleneck layer has a smaller dimension than the input. This forces the AE to learn the most salient features.
  + **Overcomplete Autoencoder:** The bottleneck layer has a larger dimension than the input. To prevent them from simply learning the identity function, they are typically used with regularization (e.g., sparse autoencoders).
  + **Sparse Autoencoder:** A type of autoencoder (can be undercomplete or overcomplete) where a sparsity penalty is added to the loss function. This penalty encourages the network to activate only a small number of neurons in the hidden layers (often the bottleneck), leading to more specialized feature learning.
  + **Denoising Autoencoder (DAE):** The encoder receives a corrupted (e.g., noisy) version of the input, but the autoencoder is trained to reconstruct the original, clean input. This forces the model to learn robust features that are not affected by noise.
  + **Contractive Autoencoder (CAE):** Adds a penalty to the loss function that encourages the learned mapping (encoder) to be contractive, meaning it's less sensitive to small changes in the input. This helps learn features that capture the local structure of the data
  + **Convolutional Autoencoder (CAE):** Uses convolutional layers in the encoder (for feature extraction from grid-like data like images) and deconvolutional (transposed convolutional) layers in the decoder. This is better suited for image data as it preserves spatial structure.
* **Variational AutoEncoders (VAE)**

VAEs are a more advanced, generative type of autoencoder.

**1. Input/Output (Problem Solved, Image Size, etc.)**

* **Problem Solved:**
  + **Generative Modeling:** VAEs are primarily used to learn the underlying probability distribution of the input data. Once trained, they can generate new data samples that resemble the training data. For example, generate new images of faces after being trained on a dataset of faces.
  + **Learning Smooth Latent Representations:** VAEs enforce a specific structure (e.g., a Gaussian distribution) on the latent space. This makes the latent space more continuous and meaningful, allowing for smooth interpolation between data points in the latent space, which translates to smooth transitions in the generated output.
  + Also used for dimensionality reduction and feature learning, similar to AEs, but with a focus on generative capabilities.
* **Input:**
  + Similar to AEs: images, vectors, etc.
  + For images, typically raw pixel data. Convolutional VAEs are common for images.
* **Output:**
  + **Reconstruction of the input:** Like AEs, VAEs also reconstruct the input.
  + **Generated Samples:** The decoder part of a VAE can be used as a generative model. By sampling points from the learned latent distribution (e.g., from a standard Gaussian N(0, I)) and passing them through the decoder, new data samples can be generated.
  + The **parameters of the latent distribution (mean and variance)** for each input are also key outputs of the encoder.

**2. Model Structures Used (briefly introduce about the model)**

VAEs also have an encoder-decoder structure but with a probabilistic twist to the latent space.

* **Encoder (Inference Network or Recognition Network):**
  + Instead of mapping the input x to a single point h in the latent space, the VAE encoder maps the input x to the parameters of a probability distribution in the latent space.
  + Typically, this distribution is assumed to be a Gaussian. So, for each input x, the encoder outputs two vectors:
    - μ\_z (mean vector)
    - σ\_z (standard deviation vector, or often log(σ\_z^2) for numerical stability)
  + These parameters define a Gaussian distribution q(z|x) = N(z | μ\_z(x), σ\_z^2(x)I) for the latent variable z given input x.
* **Latent Space (Probabilistic):**
  + To get a latent vector z for reconstruction, a sample is drawn from this learned distribution q(z|x).
  + This sampling process introduces stochasticity.
  + **Reparameterization Trick:** Directly backpropagating through a sampling node is not possible. The reparameterization trick is used: *z = μ\_z + σ\_z \* ε,* where ε is a random sample from a standard normal distribution N(0, I). This way, μ\_z and σ\_z are deterministic nodes that gradients can flow through, and the stochasticity comes from ε.
* **Decoder (Generative Network):**
  + The decoder takes a point z sampled from the latent space and maps it back to the parameters of a distribution for the original data p(x|z).
  + For example, if the input data is binary, the decoder might output the parameters of a Bernoulli distribution for each pixel. If the input is real-valued (like pixel intensities normalized to [0,1]), it might output the mean of a Gaussian distribution (often assuming a fixed variance) or parameters of a Beta distribution.
  + It attempts to reconstruct the original input x from the sampled latent vector z.
  + x\_reconstructed = g(z)
* **Key Idea of VAE:** VAEs are trained to maximize the evidence lower bound (ELBO) on the log-likelihood of the data. The loss function has two terms:
  + **Reconstruction Loss:** Encourages the decoder to accurately reconstruct the input data given the latent representation (e.g., binary cross-entropy for binary data, or mean squared error for real-valued data).

This is E\_{q(z|x)}[log p(x|z)].

* + **KL Divergence Regularization Term:** This term acts as a regularizer on the latent space. It encourages the learned distribution q(z|x) (output by the encoder for a given x) to be close to a prior distribution p(z), which is typically chosen as a standard normal distribution N(0, I). This is -D\_KL(q(z|x) || p(z)). This term ensures that the latent space is somewhat smooth and well-behaved, allowing for meaningful generation.

**3. Describe something about the datasets used**

Since AEs and VAEs are primarily unsupervised (or self-supervised), they don't strictly require labeled data in the same way supervised models like classifiers do. However, the nature of the data is still crucial. For generative tasks with VAEs, the dataset defines what kind of new samples will be generated.

* **Image Datasets:**
  + **MNIST:** A very common dataset for introductory examples of AEs and VAEs. It consists of grayscale images of handwritten digits (28x28 pixels). Good for visualizing learned latent spaces and generating new digit images.
  + **Fashion-MNIST:** Another drop-in replacement for MNIST, containing grayscale images of fashion items (28x28 pixels).
  + **CIFAR-10:** Contains 32x32 color images across 10 classes. More complex than MNIST for AEs/VAEs due to color and more varied object shapes.
  + **CelebA (CelebFaces Attributes Dataset):** A large-scale dataset of celebrity face images (around 200,000 images), each annotated with attributes. Excellent for training VAEs to generate realistic human faces and for exploring latent space manipulations (e.g., changing facial attributes like "smiling" or "wearing glasses" by moving in the latent space).
  + **LSUN (Large-scale Scene Understanding):** Contains millions of images across various scene categories (e.g., bedrooms, churches, towers). Used for training large-scale generative models.
  + Higher resolution image datasets can also be used, but they require more complex convolutional AE/VAE architectures and significant computational resources.
* **Other Data Types:**
  + **Tabular Data:** AEs can be used for dimensionality reduction or anomaly detection on tabular datasets (e.g., sensor readings, financial data, user behavior logs). Each row would be an input vector.
  + **Sequential Data:** Recurrent Autoencoders (using LSTMs or GRUs in the encoder and decoder) can be applied to sequences like time series data, text (for learning sentence embeddings or generating text, though other models are often preferred for text generation), or motion capture data.
  + **Embeddings:** AEs can be trained on pre-trained word embeddings (like Word2Vec or GloVe) or other types of embeddings to learn more compressed or specialized representations.
* **Data Preprocessing:**
  + **Normalization:** Pixel values in images are typically normalized (e.g., to [0, 1] or [-1, 1]). For other data types, standardization (zero mean, unit variance) might be applied.
  + **Flattening:** For basic fully connected autoencoders, images are often flattened into 1D vectors. For convolutional AEs/VAEs, the 2D/3D structure is preserved.
  + No explicit labels are needed for the core AE/VAE training, but if using them for a downstream supervised task (e.g., classification on the learned latent features), then labels would berequired for that subsequent stage.

**4. How the models trained**

* **AutoEncoders (AE):**
  + **Loss Function:** The primary goal is to minimize the **reconstruction error**, which is the difference between the original input x and the reconstructed output r (or x\_reconstructed).
    - For real-valued inputs (like image pixel intensities normalized to [0,1]): **Mean Squared Error (MSE)** is commonly used: L(x, r) = ||x - r||^2.
    - For binary inputs (e.g., black and white images): **Binary Cross-Entropy (BCE)** is often used: L(x, r) = - Σ [x\_i \* log(r\_i) + (1 - x\_i) \* log(1 - r\_i)].
  + **Training Process:**
    - The input data is fed into the encoder to get the latent representation h.
    - The latent representation h is fed into the decoder to get the reconstructed input r.
    - The loss function (e.g., MSE or BCE) is computed between x and r.
    - Backpropagation is used to update the weights and biases of both the encoder and decoder to minimize this reconstruction loss.
    - Optimizers like Adam or SGD are used.
  + **For Denoising Autoencoders:** The input x\_corrupted is a noisy version of x. The loss is calculated between the reconstruction r (obtained from x\_corrupted) and the original clean x.
  + **For Sparse Autoencoders:** A sparsity penalty (e.g., L1 regularization on the hidden layer activations or KL divergence between average activation and a small target sparsity) is added to the reconstruction loss.
* **Variational AutoEncoders (VAE):**
  + **Loss Function (Evidence Lower Bound - ELBO):** The VAE loss function is more complex and consists of two terms:
    - **Reconstruction Loss:** This term encourages the decoder to accurately reconstruct the input. It's the expectation of log p(x|z) where z is sampled from q(z|x). Similar to AEs, this can be MSE or BCE depending on the nature of the data x.
    - **KL Divergence (Regularization Term):** This term D\_KL(q(z|x) || p(z)) measures how much the learned latent distribution q(z|x) (output by the encoder for input x) diverges from a prior distribution p(z) (usually a standard normal distribution N(0,I)). Minimizing this term pushes the encoder to learn latent distributions that are close to the prior, making the latent space more structured and continuous, which is good for generation. The overall loss to minimize is typically:

*L\_VAE = ReconstructionLoss - D\_KL(q(z|x) || p(z)).* (Note: ELBO is maximized, so loss is negative ELBO). Often, the KL divergence term is weighted by a hyperparameter β (Beta-VAE) to control the trade-off between reconstruction quality and the "niceness" of the latent space.

* + **Training Process:**
    - An input x is fed to the encoder, which outputs parameters μ\_z(x) and σ\_z(x) for the latent distribution q(z|x).
    - A latent vector z is sampled from q(z|x) using the reparameterization trick: z = μ\_z + σ\_z \* ε, where ε ~ N(0, I).
    - The sampled z is fed to the decoder, which outputs the parameters for the distribution of the reconstructed data p(x|z) (e.g., the mean of a Gaussian, or probabilities for a Bernoulli). Let's call the reconstructed sample x\_reconstructed.
    - The reconstruction loss (e.g., MSE or BCE between x and x\_reconstructed) is calculated.
    - The KL divergence between q(z|x) and p(z) is calculated analytically (for Gaussian distributions, there's a closed-form solution).
    - The total VAE loss is computed.
    - Backpropagation is used to update the weights of the encoder and decoder.

**5. How to implement (the algorithm, the configuration/settings…)**

* **Frameworks:** PyTorch and TensorFlow (with Keras) are commonly used.
* **Implementation Steps for a basic AE:**
  1. **Define Encoder:** Create a sequence of layers (e.g., Dense in Keras, Linear in PyTorch, or Conv2D) that reduce dimensionality.
  2. **Define Decoder:** Create a sequence of layers that mirror the encoder to increase dimensionality back to the input shape (e.g., using Conv2DTranspose for convolutional decoders).
  3. **Combine into Autoencoder Model:** The encoder output feeds into the decoder.
  4. **Compile/Prepare for Training:** Specify the optimizer (e.g., Adam) and the loss function (e.g., 'mse' or 'binary\_crossentropy')
  5. **Train:** Fit the model using the input data as both the input and the target output.
* **Implementation Steps for a VAE:**
  1. **Define Encoder:**
     + Layers to process input x.
     + Two separate output layers (or a single layer split) from the encoder's features: one for μ\_z and one for log(σ\_z^2) (or σ\_z).
  2. **Sampling Layer (Reparameterization Trick):**
     + A custom layer or function that takes μ\_z and log\_sigma\_sq\_z as input, samples ε ~ N(0, I), and computes z = μ\_z + exp(0.5 \* log\_sigma\_sq\_z) \* ε.
  3. **Define Decoder:** Takes sampled z as input and reconstructs x.
  4. **Define VAE Model:** Connect encoder, sampling, and decoder.
  5. **Define Custom Loss Function:**
     + Calculate reconstruction loss (e.g., binary\_crossentropy(x\_true, x\_reconstructed\_params)). Sum or average over features/pixels.
     + Calculate KL divergence term: 0.5 \* sum(1 + log\_sigma\_sq\_z - μ\_z^2 - exp(log\_sigma\_sq\_z)). Sum or average over latent dimensions.
     + Combine them: vae\_loss = reconstruction\_loss - kl\_loss. (Or + kl\_loss if kl\_loss is defined as negative KL divergence, or if the framework expects a loss to be minimized and ELBO is to be maximized).
  6. **Compile/Train:** Use the custom loss function and an optimizer.
* **Key Configuration/Settings:**
  1. **Latent Dimension Size:** A crucial hyperparameter. Too small, and it might not capture enough information for good reconstruction. Too large (for undercomplete AEs), and it might learn the identity function too easily.
  2. **Number of Layers and Units/Filters per Layer:** Defines the capacity of the encoder and decoder.
  3. **Activation Functions:** (e.g., ReLU, sigmoid, tanh). Sigmoid is often used in the final decoder layer if inputs are normalized to [0,1].
  4. **Optimizer and Learning Rate.**
  5. **Batch Size and Number of Epochs.**
  6. **For VAEs:** The weight β for the KL divergence term (in Beta-VAEs).

**6. How is the model performance? (accuracy, …)**

* **AutoEncoders (AE):**
  + **Reconstruction Error:** The primary metric is the value of the loss function on a test set (e.g., MSE or BCE). Lower is better.
  + **Visual Quality of Reconstructions:** For image data, visually inspecting the reconstructed images compared to the originals.
  + **Performance on Downstream Tasks:** If the learned latent features are used for tasks like classification or clustering, the performance on those tasks (e.g., accuracy, NMI) indicates the quality of the learned representations.
  + **For Denoising AEs:** Reconstruction error between the output and the *original clean* data.
* **Variational AutoEncoders (VAE):**
  + **Reconstruction Error:** Similar to AEs.
  + **KL Divergence Value:** Indicates how close the learned latent distributions are to the prior. This is part of the loss.
  + **Log-Likelihood (or ELBO value):** Higher ELBO on test data is better. Estimating the true log-likelihood p(x) can be computationally intensive, so ELBO is often used as a proxy.
  + **Quality of Generated Samples:** For generative VAEs, this is often subjective but crucial. Visual inspection of generated samples for realism, diversity, and coherence.
  + **Quantitative Metrics for Generative Models:**
    - **Fréchet Inception Distance (FID):** Measures the similarity between the distribution of generated samples and the distribution of real samples in terms of features extracted by a pre-trained Inception network. Lower FID is better.
    - **Inception Score (IS):** Measures both the quality (clarity) and diversity of generated images. Higher IS is better. (Less favored now compared to FID).
  + **Smoothness/Continuity of Latent Space:** Can be qualitatively assessed by interpolating between points in the latent space and observing if the transitions in the generated output are smooth and meaningful.

**7. Some outputs**

* **AutoEncoder (AE):**
  + **Input:** An image of a handwritten digit "7".
  + **Output (Reconstruction):** An image that looks very similar to the input "7", possibly slightly blurrier or with minor artifacts depending on the AE's capacity and training.
  + **Output (Latent Vector):** A low-dimensional vector (e.g., 32 numbers if latent\_dim=32) that represents the compressed form of the digit "7".
* **Denoising AutoEncoder (DAE):**
  + **Input:** An image of a handwritten digit "7" with added random noise (e.g., salt-and-pepper noise).
  + **Output (Reconstruction):** An image that ideally looks like a clean "7", with the noise removed or significantly reduced.
* **Variational AutoEncoder (VAE):**
  + **Input:** An image of a face.
  + **Output (Reconstruction):** A reconstruction of the input face, similar to an AE.
  + **Output (Latent Distribution Parameters):** μ\_z and σ\_z vectors that define a Gaussian distribution in the latent space for that input face.
  + **Output (Generated Sample):**
    - If you sample a random vector z\_sample from N(0, I).
    - Feed z\_sample to the VAE's decoder.
    - **Output:** A brand new image of a face that was not in the training set but shares characteristics with the faces the VAE was trained on.
  + **Output (Latent Space Interpolation):**
    - Take two input images (e.g., face A and face B).
    - Encode them to get z\_A and z\_B (or their distribution means).
    - Linearly interpolate between z\_A and z\_B in the latent space: z\_interp = α\*z\_A + (1-α)\*z\_B for α from 0 to 1.
    - Decode each z\_interp.
    - **Output:** A sequence of images showing a smooth transition from face A to face B.

# Part 2: Demo Implementation & Explanations

## Comparison Table

| **Model** | **Best For** | **Typical Image Size** | **Example Use Cases** |
| --- | --- | --- | --- |
| CNN | Image classification | 32x32 to 224x224 | Identifying objects, simple scene classification |
| Mask R-CNN | Instance segmentation | 800+ pixels | Precise object boundaries, multi-object scenes |
| YOLO | Fast object detection | 416x416 to 608x608 | Real-time detection, multiple object tracking |
| AE/ VAE | Image generation/ reconstruction | 28x28 to 128x128 | Face generation, anomaly detection, image denoising |

## CNN

**Demo:** Classifies uploaded images using a CNN pre-trained on ImageNet.

**Result:**

**A screenshot of a computer

AI-generated content may be incorrect.**

## Mask R-CNN

**Demo:** Detects and segments objects in uploaded images using a pre-trained Mask R-CNN

**Result:**

A person and dog walking in a park

AI-generated content may be incorrect.

## YOLO

**Demo:** Performs fast object detection on uploaded images using YOLOv8.

**Result:**

A collage of people crossing a street

AI-generated content may be incorrect.

## Autoencoder

**Demo:** Trains an autoencoder on uploaded images for reconstruction or denoising.

**Result:**

A license plate with numbers and letters

AI-generated content may be incorrect.

A pixelated image of a person's face

AI-generated content may be incorrect.

A close-up of a license plate

AI-generated content may be incorrect.