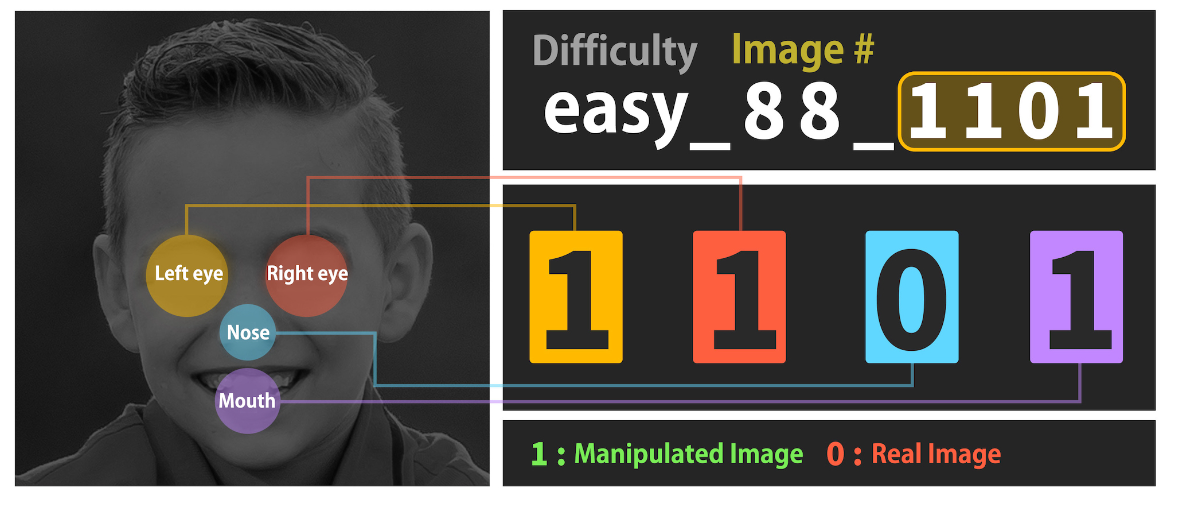
**Datasets:**

* **CelebFaces Attributes Dataset (CelebA):**
  + large-scale face attributes dataset
  + more than 200K celebrity images each with 40 attribute annotations.
  + covers large pose variations and background clutter.
  + has large diversities, large quantities, and rich annotations
  + 10,177 number of identities, 202,599 number of face images, 5 landmark locations, 40 binary attributes annotations per image.
  + Celeba [Link](https://www.dropbox.com/s/d1kjpkqklf0uw77/celeba.zip)
  + Celeba 128x128 [link](https://www.dropbox.com/s/7e966qq0nlxwte4/celeba-128x128-5attrs.zip)
  + Celeba 256x256 [link](https://www.dropbox.com/s/zdq6roqf63m0v5f/celeba-256x256-5attrs.zip)
  + CelebAMask-HQ [link](https://github.com/switchablenorms/CelebAMask-HQ)

* **Flickr-Faces-HQ (FFHQ)**
  + 70,000 high-quality PNG images
  + 1024×1024 resolution
  + contains considerable variation in terms of age, ethnicity and image background
  + good coverage of accessories such as eyeglasses, sunglasses, hats, etc.
  + images were crawled from Flickr , automatically aligned,cropped using dlib.
  + [Link](https://drive.google.com/drive/folders/1u2xu7bSrWxrbUxk-dT-UvEJq8IjdmNTP)
* **Real and Fake faces from Kaggle**
  + 70k REAL faces from the Flickr dataset collected by Nvidia
  + 70k fake faces sampled from the 1 Million FAKE faces (generated by StyleGAN) that was provided by Bojan.
  + resized all the images into 256px
  + [Link](https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces)
* **CIPLAB , YONSEI UNIVERSITY from Kaggle**

****

* + contains expert-generated high-quality photoshopped face images
  + composite of different faces, separated by eyes, nose, mouth, or whole face



##### 960 fake images

* + 1081 real images
  + [Link](https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection)

# **iFakeFaceDB**

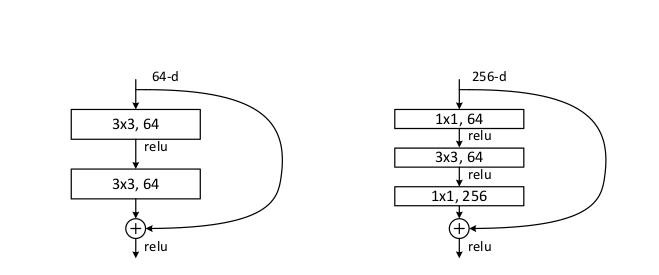
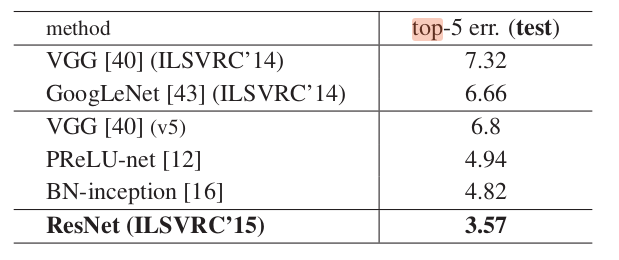
* + About 87,000 synthetic face images
  + generated by the Style-GAN model
  + transformed with the GANprintR approach.
  + All images aligned and resized to the size of 224 x 224.
  + [link](https://github.com/socialabubi/iFakeFaceDB)

**Feature Extraction Techniques:**

**(The first five methods are for general deepfake detection and also specific to image orientation like rotation, scaling ,translation.)**

1. **CNN :pretrained models**

* **ResNet**  
  The resnet is an ensemble model which is created to extract features at a very microscopic level.  
  But due to the ensemble way of training each layer in the resnet the overall complexity is lesser than VGG net but with that the error in test data set used the paper is also lesser than the VGG net (on test data set of imageNet we get 3.57% error) . This model won 1st place in the image classification competition held by ILSVRC in 2015.

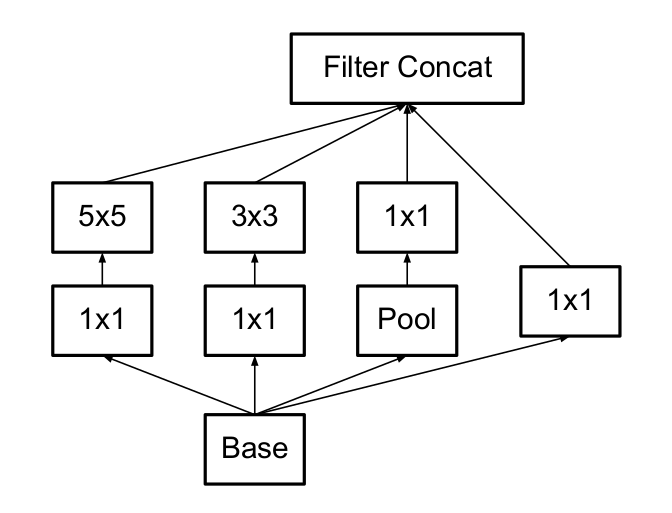
****

[**link to resnet paper**](https://ieeexplore.ieee.org/document/7780459)

* **Inception-Net**

As we can see in the previous model itself, the inception model has a higher error rate so why are we discussing it here? The answer is that independently it gives a higher error rate but when ensembled with 4 models and multicrop-evaluation(from paper) it gives a lower error rate at 3.5%.thus this proves that when used with other models it gives good efficiency.

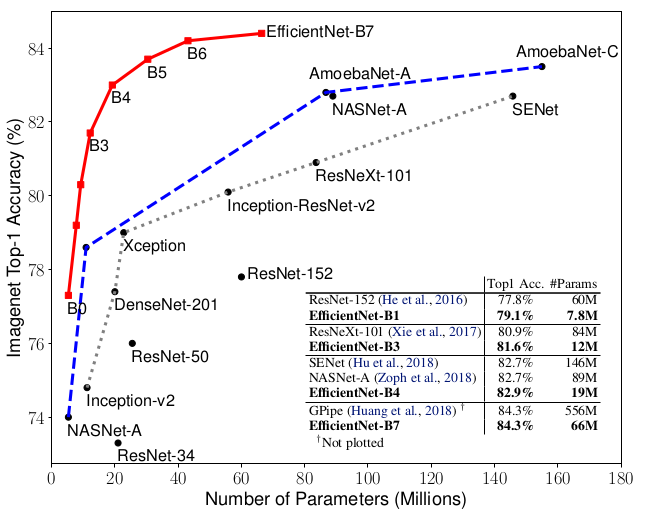
Inception net uses separate inception modules as a fundamental unit rather than conv2D layers here the module concatenates various outputs from different stride values done in convolution.

****

[**Link to inceptionNet paper**](https://arxiv.org/abs/1512.00567v3)

* **Efficient-Net**

These are versions of AlexNets,XceptionNets, Etc but have the balancing in their depth of the neural layers and the number of neurons in each layer. This normalization provides the highly efficient version of its form. We can see in the graph below the increase in accuracy.



[Link to EfficientNet paper](https://arxiv.org/abs/1905.11946)

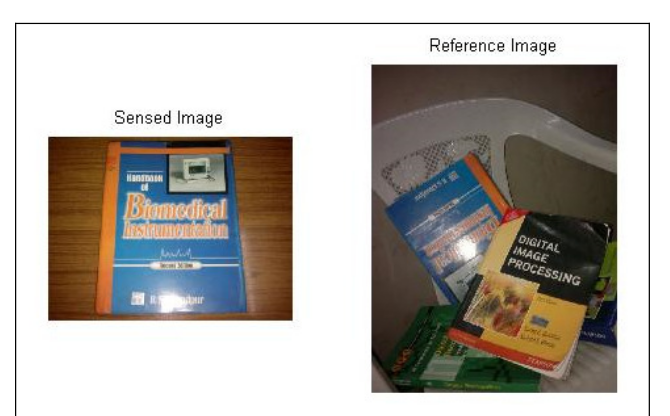
**2. Rotation-Invariant Descriptors**

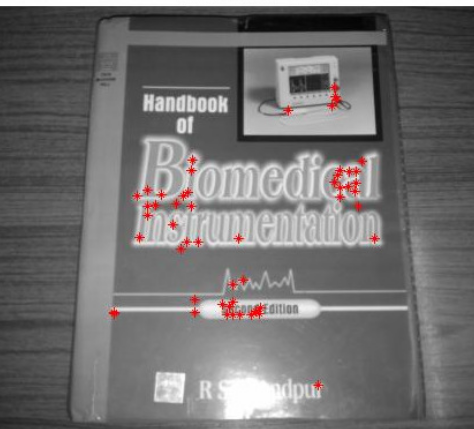
* **Scale-Invariant Feature Transform(SIFT)**

This is a method where we extract features that have the same meaning even though they are translated ,rotated ,or scaled to a different component .

Firstly it identifies the region of interest , then applies key point localization which would find out the coordinates of the keypoints (pixels which stand out in the image component and can be matched with the trained image).

Then the process is to get a mapping between the keypoints and consistent orientation Then the gradient magnitude is calculated for the keypoints and then a specific algorithm is used to find out if the image is the same or different if the orientation is corrected . But if needed to be used in deep face detection it could be used to correct the orientation of the image or map to a reference size of image and then use other models to classify as deep fake or real.





[Link to SIFT paper](https://ieeexplore.ieee.org/document/8376448)

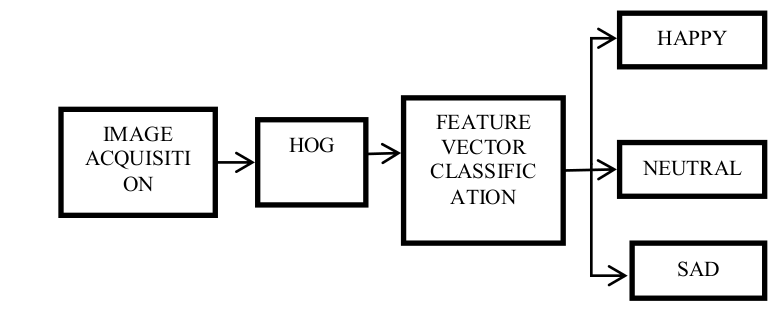
* **Speeded-Up Robust Features(SURF)**

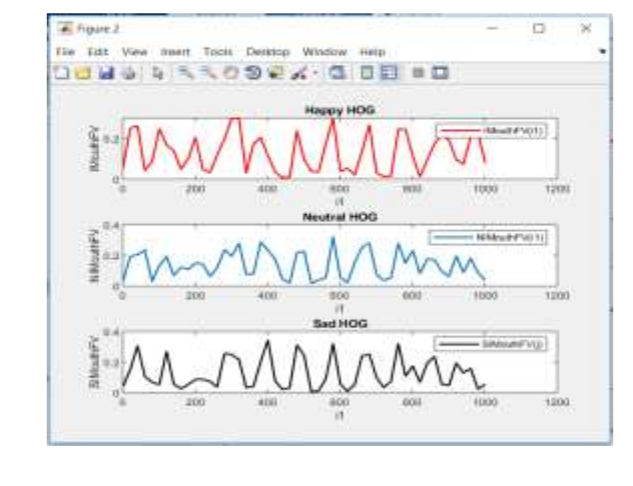
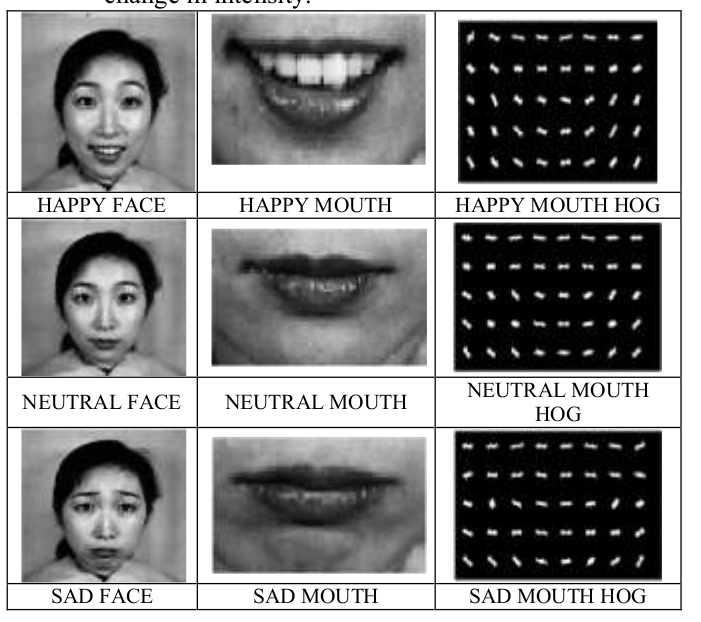
This is a new and improved version of the SIFT model that works the same but along with scaling,translation,rotation it is also robust to different occlusion of image components.

[Link to SURF paper](https://ieeexplore.ieee.org/document/8666894)

**3. Histogram of Oriented Gradients(HOG)**

This method is robust to orientations like in SIFT the difference is that there in SIFT we use Key points to compare but her we consider all the points that is all the pixels or superPixels to create a grid of gradients in its corresponding pixels or superpixels.This gradients also contain the local orientation information in them stored which can be used in further levels.Then use them to create a histogram of those gradients. In Paper they have used these histograms to classify if the face is happy or sad etc.

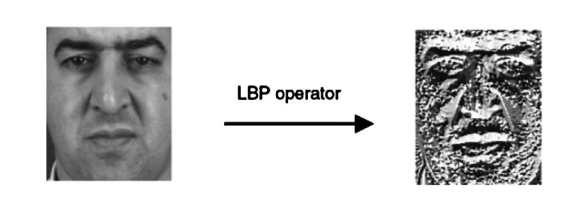
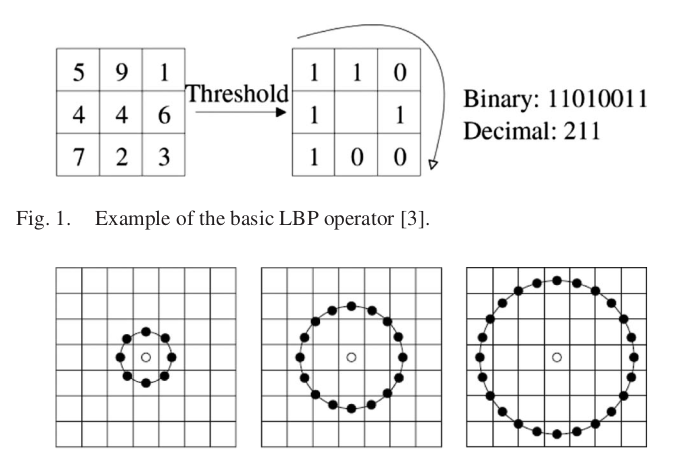




[Link to HOG paper](https://ieeexplore.ieee.org/document/8821814)

**4. Local Binary Patterns**

For each pixel in the image a predefined neighborhood is defined and is considered first and then using the gray scale values of the central pixel and neighbor pixels a binary code is generated and then passed onto next layers.



[Link to Local Binary Patterns paper](https://ieeexplore.ieee.org/abstract/document/5739539?casa_token=SlOM10m6dfMAAAAA:d1dUhkXn3ZECHDDF_8B1XKPcaK9g04nL18U_F0Q76aO2Edid1p8qthIYZeTYTzHnBU9XhWn5ZcU)

**5. Rotation-Invariant CNN Architectures**

We create a rotation-invariant and Fisher discriminative CNN (RIFD-CNN) model (Exception : this is a feature extractor + classification model here) .Specifically,the rotation-invariant layer is trained by imposing an explicit regularization constraint on the objective function that enforces invariance on the CNN features before and after rotating. The Fisher discriminative layer is trained by imposing the Fisher discrimination criterion on the CNN features so that they have small within-class scatter but large between-class separation.

[Link to Rot-Inv CNN paper](https://ieeexplore.ieee.org/document/7780684)

**(The following two are specific for the face-pose orientation in images)**

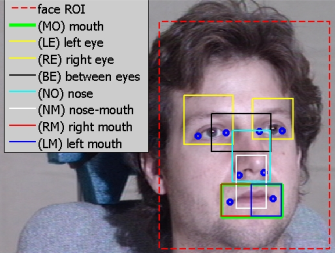
**6. Facial Landmark Detection**

The result from this feature extractor is a bounding box for high level features like eyes,nose,etc.This model is robust to the illumination, face orientation as well.



The first step is the assign the likelihood value of a pixel belonging to the different feature landmarks we have(left eye,right part of the mouth,nose,etc) based on these values we assign the threshold to get the region enclosing the landmark.

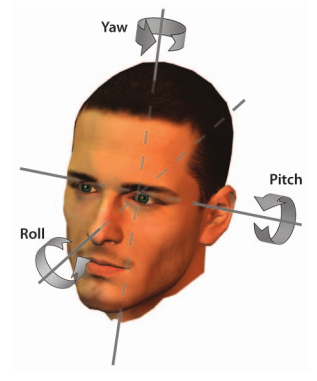
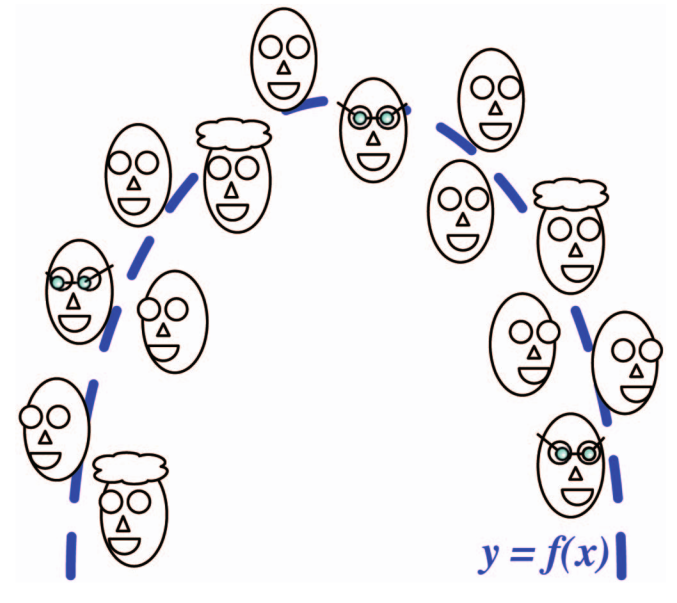
Based on the orientation information from the landmarks we can get the orientation of the entire face and thus can be used as a feature in deep face detection.

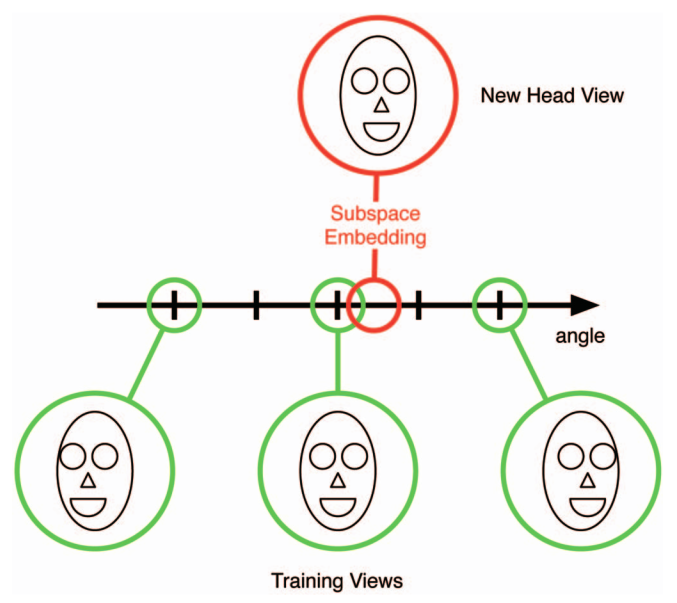
 [Link to FLD paper](https://ieeexplore.ieee.org/document/5771411)

**7. Pose Estimation**

Ideal pose Estimator would be invariant of various actors including biological appearance, facial expression and presence of accessories(specs,hat,etc).

It estimates the input face in any one of many discrete angular views where there may be more than one degree of freedom.(like vertical axis ,horizontal,etc)

 [Link to Pose Estimation paper](https://ieeexplore.ieee.org/document/4497208)

**ML/DL models:**

* **SVM:**

SVM is a supervised learning based machine learning model that is particularly effective in dealing with binary classification problems. To utilize SVM for deepfake classification, a set of features extracted from the input data is required. These features can include both visual and temporal characteristics, such as texture, color distribution, motion patterns, or frequency components. The SVM algorithm then maps these feature vectors onto a high-dimensional space and finds the optimal hyperplane that best separates the real and fake samples.

* **Decision Trees:**

Same as SVM, same input data can be passed here and from that a tree can be constructed and classification can be done.

* **InceptionNet:**

It introduced the concept of "inception modules" that allow for efficient and parallelized feature extraction at different scales within the network. InceptionNet aims to address the trade-off between depth and computational efficiency in neural networks. It achieves this by employing a combination of 1x1, 3x3, and 5x5 convolutions in parallel within each inception module, followed by max pooling and concatenation of their outputs. This design allows the network to capture both local and global features effectively. Furthermore, InceptionNet significantly reduces the computational cost by incorporating dimensionality reduction through 1x1 convolutions. By utilizing multiple inception modules stacked together, the network can learn a hierarchy of complex features. InceptionNet has achieved state-of-the-art performance on various image recognition tasks, while maintaining reasonable computational requirements, making it a widely adopted architecture in computer vision applications.

* **XceptionNet:**

Has shown reliable results in identifying forged videos.Instead of using a single CNN, it makes use of inception model that decouples different tasks that can operate in non overlapping features and concatenate the results in the end.

* **EfficientNet (EffNet):**

Focuses on efficiently scaling convolution networks to improve performance and speed while keeping the size minimum. EfficientNet-B7 performed 6x times faster while being 8.4x times smaller than a traditional ConvNet

* **MesoNet:**

There are two architectures in this: Meso-4 and MesoInception4. They are based on well-performing networks for image classification that alternate layers of convolutions and pooling for feature extraction and a dense network for classification.

* Meso-4: This network begins with a sequence of four layers of successive convolutions and pooling, and is followed by a dense network with one hidden layer. To improve generalization, the convolutional layers use ReLU activation function that introduce non-linearities and Batch Normalisation to regularize their output and prevent the vanishing gradient effect, and the fully-connected layers use Dropout to regularize and improve their robustness. There are 27,977 trainable parameters for this network.
* MesoInception-4: This has been created by replacing the first two convolutional layers of Meso4 by a variant of the inception module. The idea of the module is to stack the output of several convolutional layers with different kernel shapes and thus increase the function space in which the model is optimized. This network has 28,615 trainable parameters overall.
* **ResNet:**

ResNet, short for Residual Neural Network, is a groundbreaking deep learning architecture that revolutionized the field of computer vision. ResNet addresses the challenge of training extremely deep neural networks by utilizing skip connections or residual connections. These connections allow the network to learn residual mappings, focusing on the incremental changes to the input. By doing so, ResNet mitigates the problem of vanishing gradients that hindered the training of deep networks. ResNet architectures consist of multiple residual blocks, each containing convolutional layers, batch normalization, and ReLU activation functions. The skip connections ensure the smooth propagation of information, enabling the network to learn increasingly complex representations. ResNet has achieved remarkable performance in various image recognition tasks and has become a fundamental model for many subsequent advancements in deep learning.

* **Vision Transformers:**

Vision Transformers, also known as ViTs were initially introduced for sequential data, but vision transformers adapt the architecture for image-based data. ViTs revolutionize computer vision by replacing traditional convolutional neural networks (CNNs) with self-attention mechanisms. In this, the input images are divided into fixed-size patches, which are then linearly embedded into a sequence of tokens. These tokens represent local visual information and are fed into a transformer encoder. The transformer encoder comprises multiple self-attention layers and feed-forward neural networks. The self-attention mechanism allows the model to capture global dependencies and establish relationships between different patches. This global context enables ViTs to effectively recognize objects, spatial relationships, and context in images.

**Metrics:**

* Since deepfake detection is a binary classification task one of the primary methods to evaluate it can be use of **Confusion matrix,** to summarize the success and failures in the detection tasks
  + Metrics that can be derived from confusion matrix are:

1. Precision and recall
2. Sensitivity and specificity
3. Accuracy and F-Score

* **Receiver Operating Characteristics(ROC) and Area Under curve** are used to measure the performance of binary classifiers that output a score of prediction.
* **Log Loss** is a widely used performance metric for binary classifiers that can return the probability score for the predicted label.

**Top 3 image papers:**

1. **Learning Self-Consistency for Deepfake Detection** <https://openaccess.thecvf.com/content/ICCV2021/papers/Zhao_Learning_Self-Consistency_for_Deepfake_Detection_ICCV_2021_paper.pdf>
2. **Fighting Deepfakes by exposing convolutional traces on images:** <https://ieeexplore.ieee.org/document/9189772>
3. **DeepFake Face Image Detection based on Improved VGG Convolutional Neural Network** <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9189596>