TALL: Deepfake Video Detection via Thumbnail Layout Learning

Dataset: Subset of Celeb-DF v2

📌 PRESENTED BY

ARYA PANDEY
ADHIKANSH GOEL

INTRODUCTION

- The rapid evolution of deepfake technology has made it easier to produce highly realistic fake videos.
- These pose significant threats in misinformation, privacy, and security.
- Traditional detection models analyze frames independently and often fail to capture temporal and contextual information.
- Our solution leverages TALL (Thumbnail Layout Learning) with the Swin Transformer to capture both spatial and temporal patterns.
- ResNet, a CNN-based model, is used as a baseline for comparison.

Data Pipeline

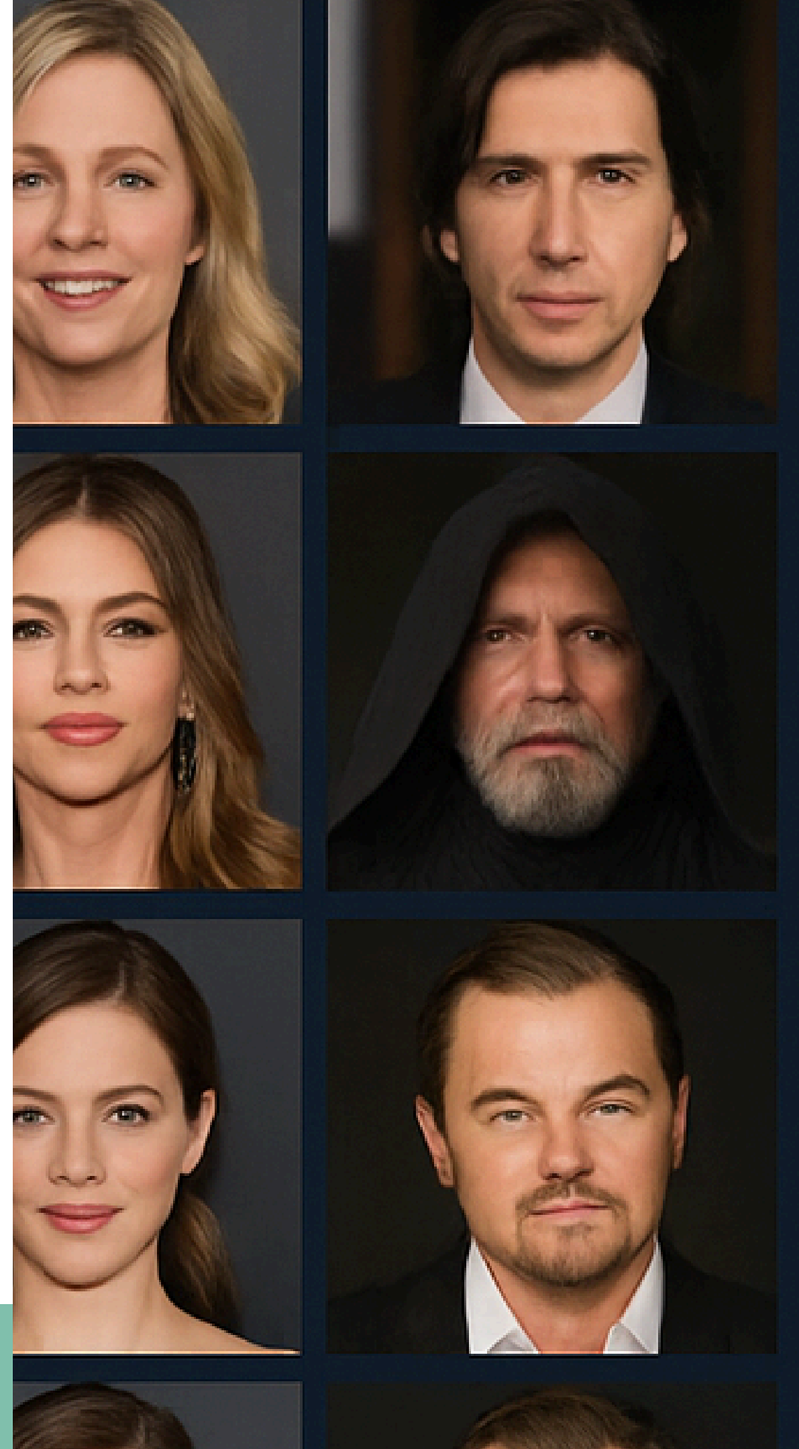
Step-by-step:

1. Load Celeb-DF subset (real/fake)
2. Sample N frames uniformly from each video
3. Resize each frame to 224 x 224
4. Stitch frames into a grid layout (e.g., 2 x 2)
5. Save as thumbnail layout image
6. Assign label: Real or Fake

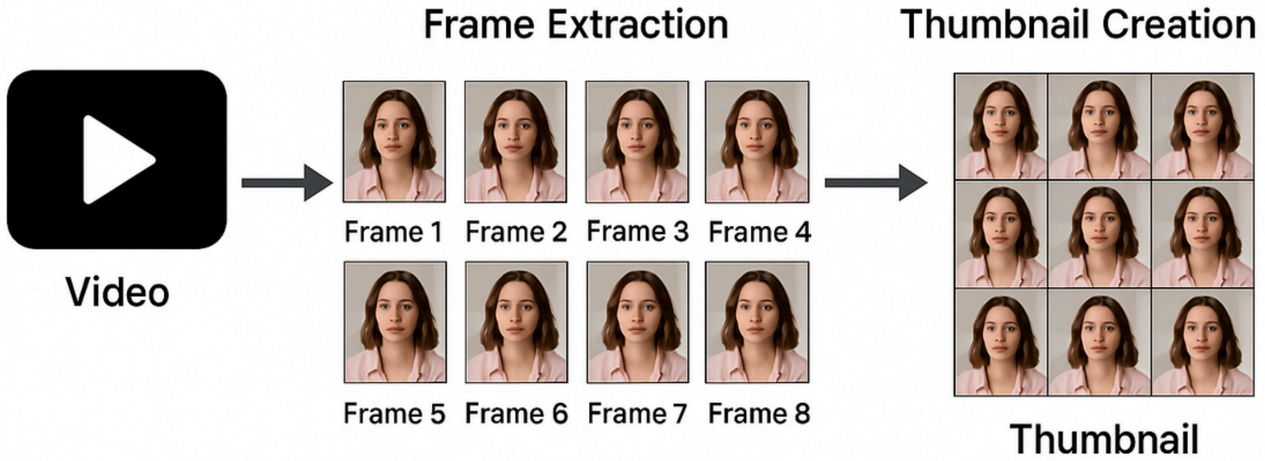
Why Celeb-DF v2 Subset?

Celeb-DF v2 is a challenging benchmark dataset for deepfake detection.
High-resolution, realistic manipulations unlike earlier datasets (FaceForensics++).

- We used a small but diverse subset:
  - 126 real + 900 fake videos
  - Chosen for variety in lighting, compression, and identities.



Thumbnail Creation Process



- Frame Extraction: 4 equally spaced frames per video are captured using OpenCV.
- Grid Construction:
  - Frames arranged into 2 rows and 2 columns.
  - Each resized to 112 x 112.
  - Final thumbnail dimension: 224 x 224.

Tools: OpenCV for frame extraction, PIL for image composition.

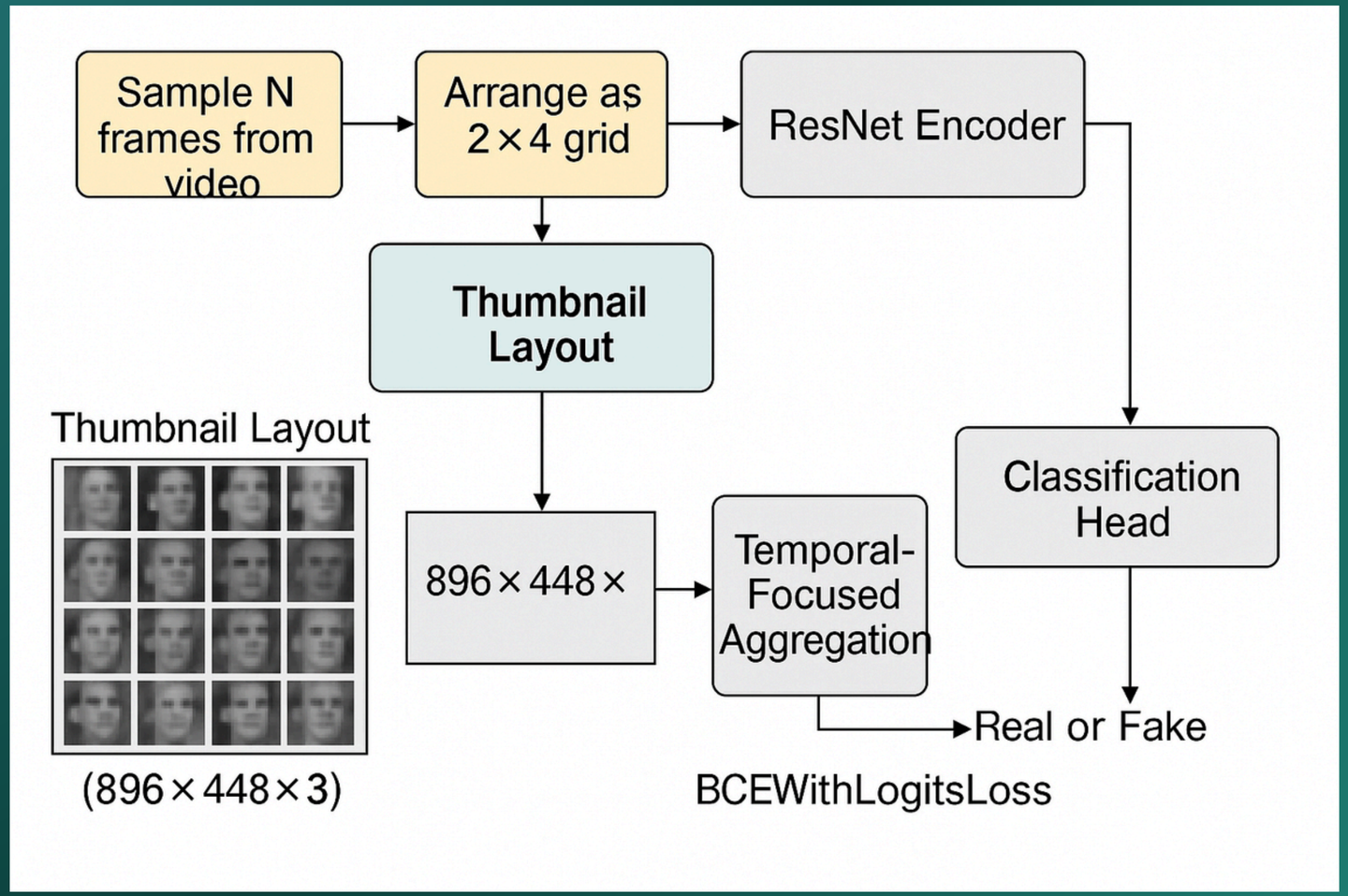
TALL + Swin Transformer Model

- TALL Concept: Use grid of frames to model both space and time in a single input.
- Swin Transformer:
  - Operates on non-overlapping patches.
  - Uses a hierarchical structure with shifted windows.
  - Suitable for high-resolution images and layout-based inputs.
- Input thumbnails resized to 224x224 for training with the Swin base model.

ResNet Baseline Model

- Used ResNet-50 architecture for comparison.
- Pretrained on ImageNet.
- Final layer modified for binary classification.
- Input thumbnails were same as those used in the TALL-Swin setup.
- Limitation: ResNet lacks the ability to explicitly capture inter-frame relationships.

Architecture Overview

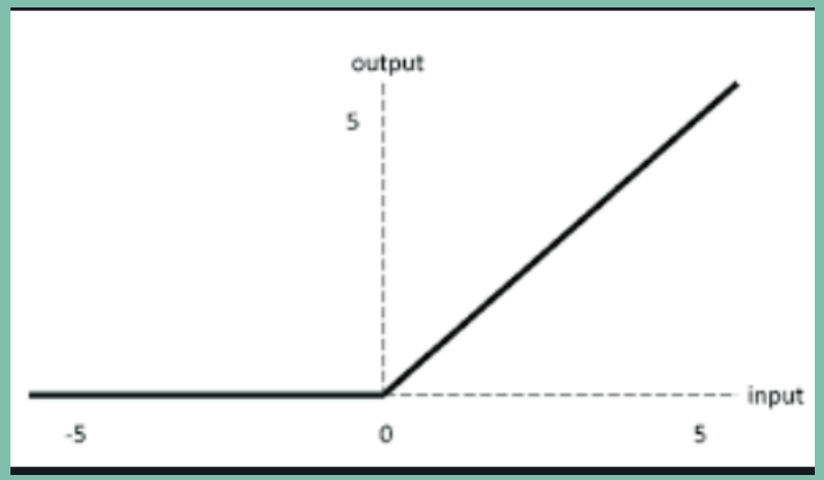


Loss Function and Activation Function

The cross-entropy loss is employed to optimize the TALL-Swin, which is defined as:

$$\mathcal{L}_{CE} = -\frac{1}{n}\sum_{i=1}^{n} y_i \log \mathcal{F}(x_i) + (1 - y_i)(\log (1 - \mathcal{F}(x_i))$$

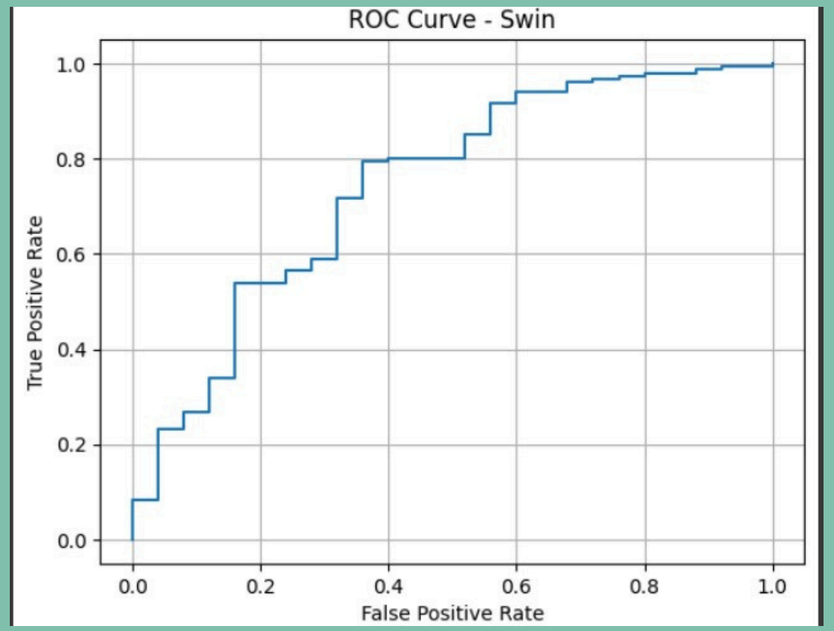
ReLU (Rectified Linear Unit) is used in Swin Transformers primarily because it's a computationally efficient and effective activation function that addresses the vanishing gradient problem, a common challenge in training deep neural networks.

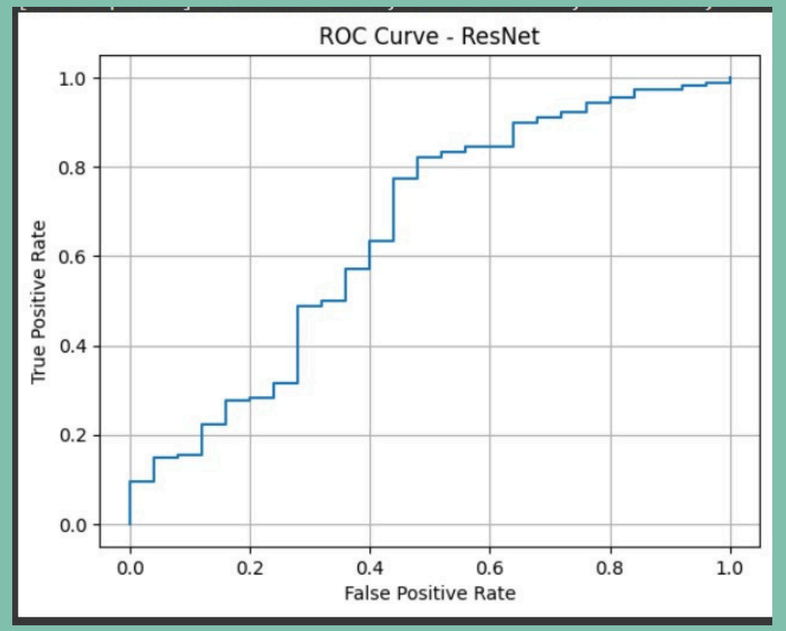


$$f(x) = \max(0,x)$$

Results

	Accuracy	AUC
Swin	**0.8780**	**0.7418**
ResNet50	**0.8683**	**0.6942**





Conclusion

Swin Transformer outperforms ResNet50 in terms of accuracy and AUC, showing stronger ability to capture spatial inconsistencies common in synthetic videos.

✅ TALL (Thumbnail Layout) enables efficient spatial-temporal compression of video information into a single image — making real-time deepfake detection more practical.

🚀 This modular pipeline offers a scalable foundation for deepfake detection and can be extended with larger models, motion cues, or real-world deployment strategies.

Thank You