Baseline Model: ResNet-50 for Real vs AI-Generated Image Classification

Wenyang Qiu

Concordia University

November 28, 2025

Problem Setup

Task: Binary classification of images into Real vs AI-generated.

- Input: face images from a mixed corpus of real photos and generative models.
- Output: probability of "fake" ($P(\text{fake})$) and predicted label (real / fake).
- Goal: build a strong **baseline** and understand its strengths/weaknesses before proposing a new CV model.

Dataset: DeepDetect-2025 (Kaggle)

- Source: `ayushmandatta1/deepdetect-2025` (Kaggle).
- Split:
  - Train: 90,409 images (real + fake).
  - Test: 21,776 images.
  - Balanced binary labels: `0 = real`, `1 = fake`.
- Preprocessing:
  - Resize to 224×224.
  - Normalize with ImageNet mean / std.
  - Standard PyTorch `DataLoader`, batch size $= 32$.

Baseline Model: ResNet-50

- Backbone: **ResNet-50**, initialized with ImageNet-1K weights.
- Replace final FC layer:

$$\texttt{fc:} \quad 2048 \to 2 \quad (\text{real}, \text{fake})$$

- Loss: cross-entropy on 2-way softmax outputs.
- Optimizer: Adam, learning rate 1×10^{-4}.
- Training:
  - Device: GPU (Colab).
  - Runs: 1–5 runs, 60 epochs each (best checkpoint saved on validation accuracy).

Quantitative Results (Test Set)

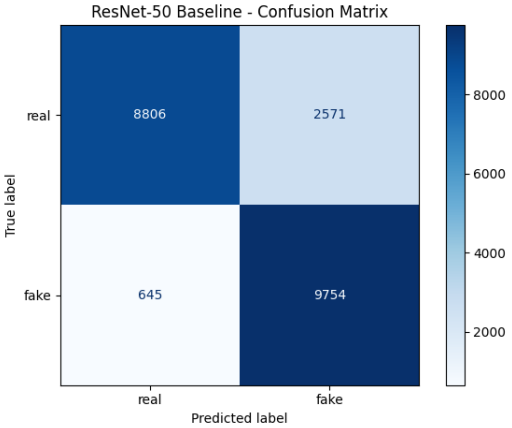
- Overall performance of the ResNet-50 baseline on the held-out test split:

Model	Accuracy	ROC-AUC	PR-AUC	F1 (fake)
ResNet-50 baseline	0.85	0.96	0.96	0.86

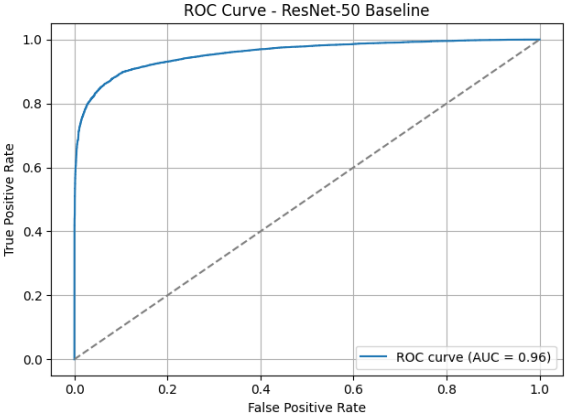
- The baseline already achieves strong discrimination between real and AI-generated images.
- Next: understand where it *succeeds* and where it *fails*, to motivate a stronger model.

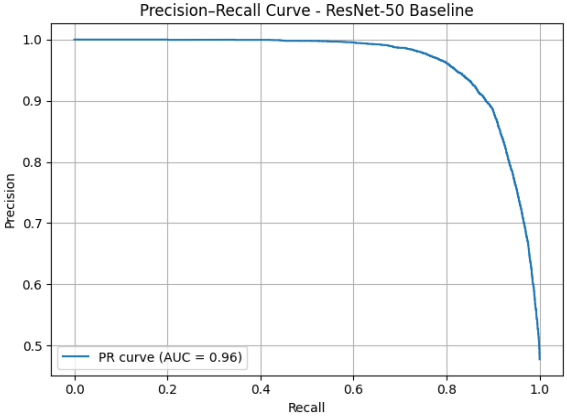
Confusion Matrix

- Confusion matrix on the test set:
  - True Real vs Predicted Real/Fake.
  - True Fake vs Predicted Real/Fake.
- Shows the trade-off between:
  - **False Positives** (real $\rightarrow$ fake)
  - **False Negatives** (fake $\rightarrow$ real)



ROC and Precision–Recall Curves



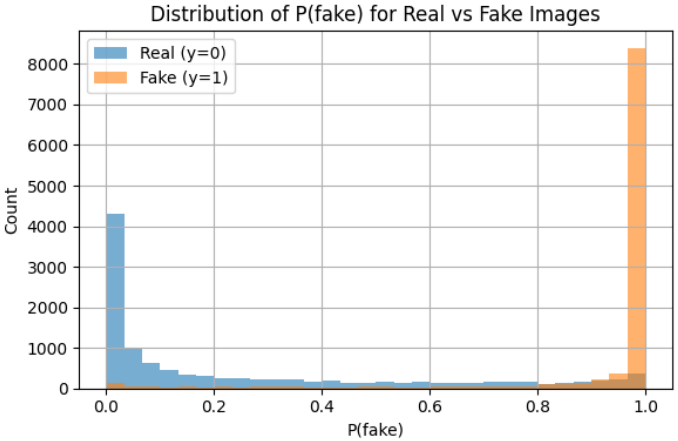


ROC curve (AUC = 0.96)

Precision–Recall curve (AUC = 0.96)

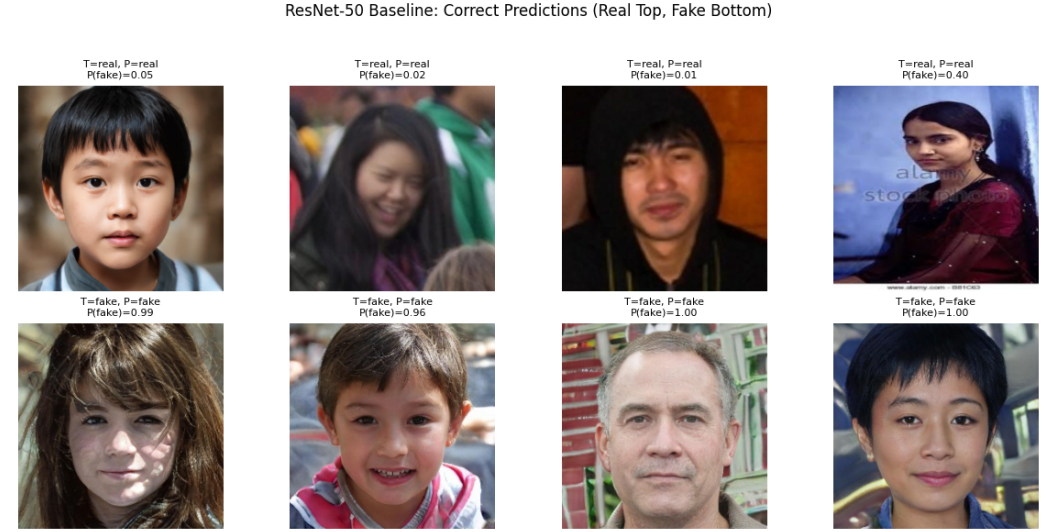
These curves summarize the baseline's behavior over all decision thresholds for the "fake" class.

Distribution of $P(\text{fake})$ for Real vs Fake Images

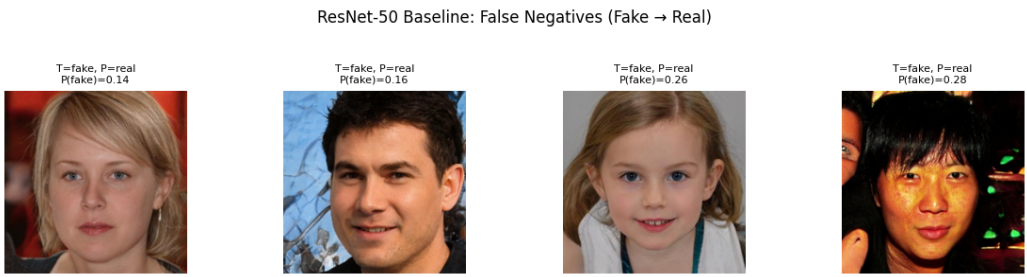


- Most real images have $P(\text{fake})$ close to 0.
- Most AI-generated images have $P(\text{fake})$ close to 1.
- Overlap region around 0.3–0.7 corresponds to **ambiguous cases**:
  - high-quality fakes, low-quality reals.
  - motivates more robust features (local, frequency, EXIF).

Qualitative Examples: Correct Predictions

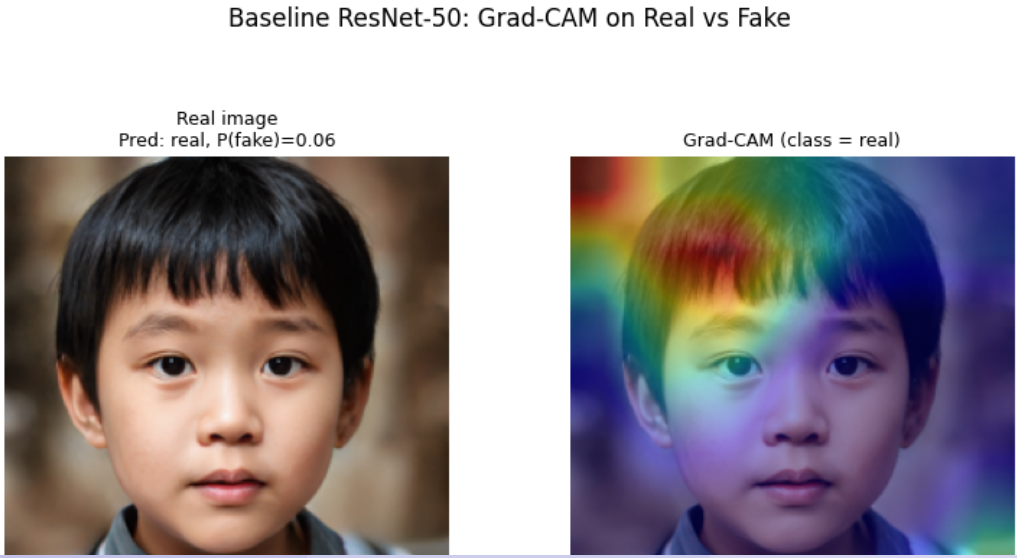


Error Cases: Fake → Real (False Negatives)



- All examples are AI-generated faces (**T=fake**) that the baseline classifies as real (**P=real**).
- The predicted $P(\text{fake})$ is relatively low (e.g., 0.13–0.28), indicating strong confusion.
- These high-quality fakes are visually very close to real faces, showing the limitation of relying only on global CNN appearance features.
- This motivates adding richer cues (local patches, frequency artifacts, EXIF) in the next model.

Grad-CAM Visualization of the Baseline ResNet-50



Summary of Baseline and Next Steps

What we have now

- A strong ResNet-50 baseline for real vs AI-generated image classification.
- Solid quantitative evaluation (accuracy, ROC-AUC, PR-AUC, F1).
- Visualization toolkit:
  - Confusion matrix, ROC/PR curves, $P(\text{fake})$ distribution.
  - Qualitative examples (TP/TN/FP/FN) and Grad-CAM.

Next steps (for the full paper)

- Design a multi-branch model combining:
  - Global features (backbone CNN/ViT),
  - Local patch-level cues,
  - Frequency-domain artifacts (FFT),
  - EXIF / metadata signals.
- Evaluate on stronger forensics benchmarks and compare with SOTA.