Adobe

IMAGE CLASSIFICATION AND DEEPFAKE DETECTION

PRESENTED BY TEAM - 84

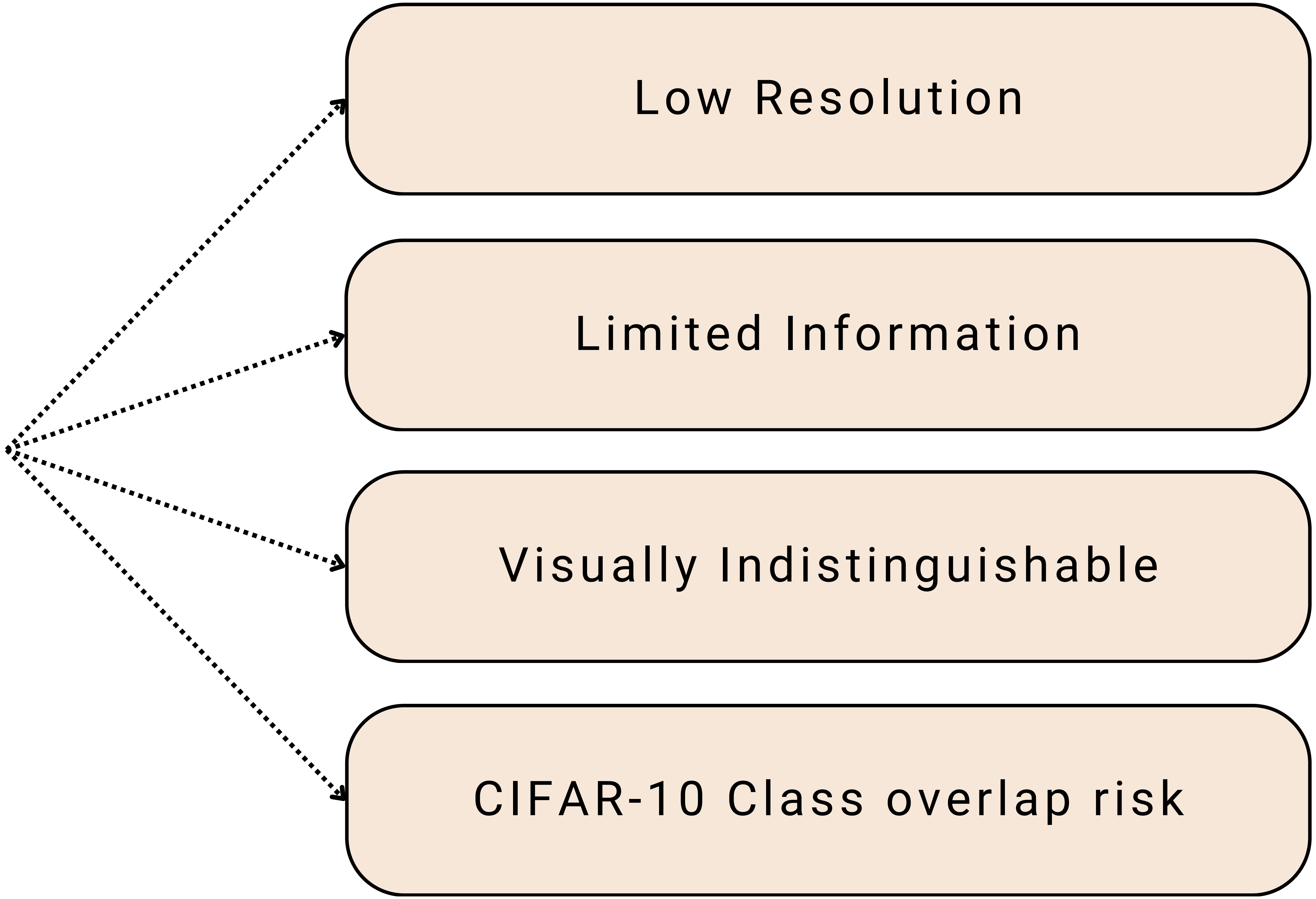
12 DECEMBER, 2024

INTER IIT Tech Meet 13.0

TASK 1

DEEPFAKE DETECTION

Challenges



- Low Resolution
- Limited Information
- Visually Indistinguishable
- CIFAR-10 Class overlap risk

Classification of 32 × 32 Images

INITIAL APPROACH - ENSEMBLE

APPROACH

Ensemble of shallow neural networks with weighted importance to different models - Tiny VGG, MobileNet v3

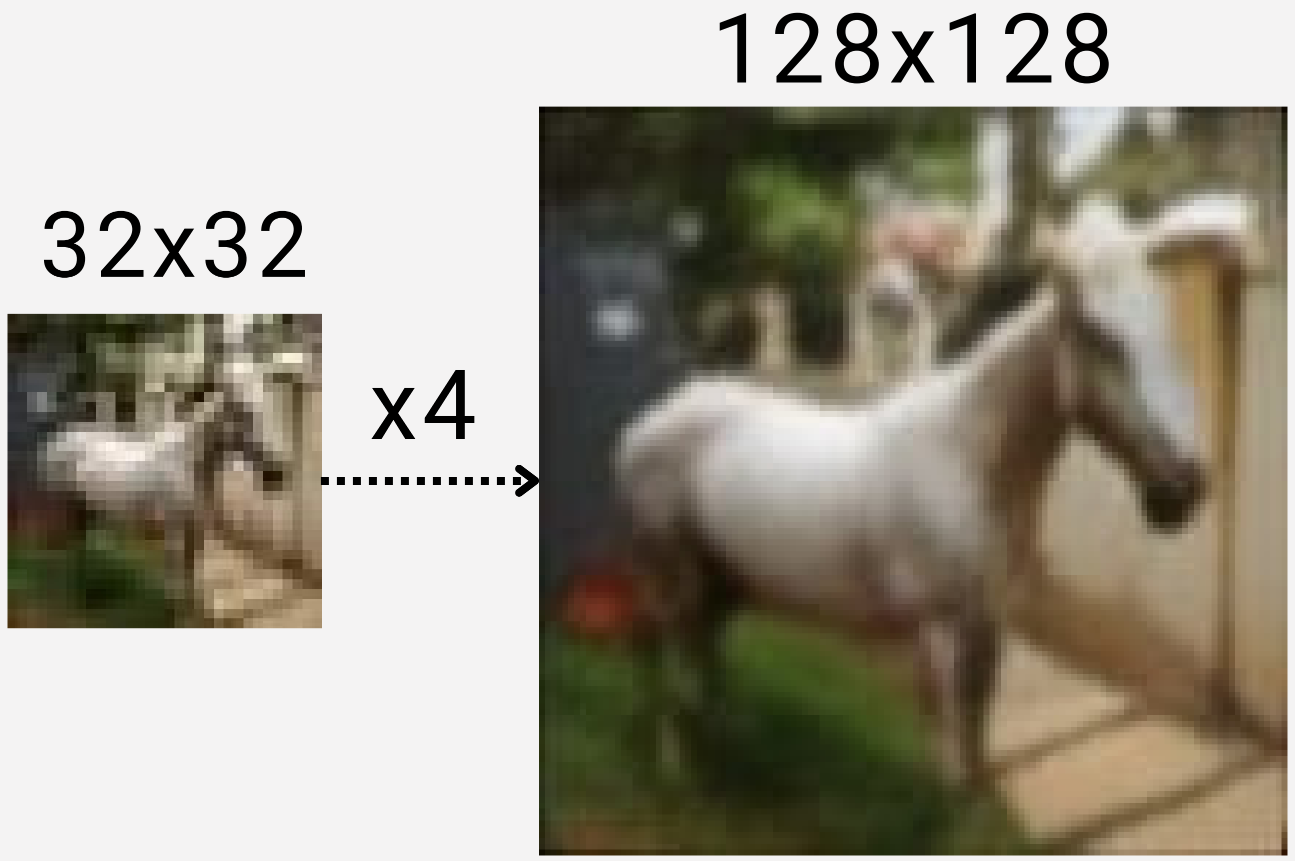
ISSUES

Low Resolution limited the potential for achieving superior results
Lack of depth made networks vulnerable to adversarial attacks

RESULT

Acheived Accuracy - 93.5 % on CIFAR Test Dataset
Adversarial attacks significantly reduced accuracy to further 85%

Super Resolution – Real-ESRGAN



MOTIVATION

Considering the challenge of adversarial attacks, Real-ESRGAN emerged as the best fit due to its superior performance in handling complex degradation processes compared to other models.

KEY CHALLENGE

Super-resolving images without introducing artifacts is crucial for deepfake and artifact detection.

OUR SOLUTION

- Since real images having minimal artifacts, we fine-tuned Real-ESRGAN using only high-low resolution real image pairs.
- This prevented the model from learning to super-resolve artifacts by excluding artifact-heavy data.
- Ensured artifact-free super-resolution to maintain detection accuracy.

Classification of 128×128 Images

SUPER RESOLUTION & CLASSIFICATION

PRE-TRAINING

DenseNet121 model pre-trained on ImageNet dataset.

FINE-TUNING

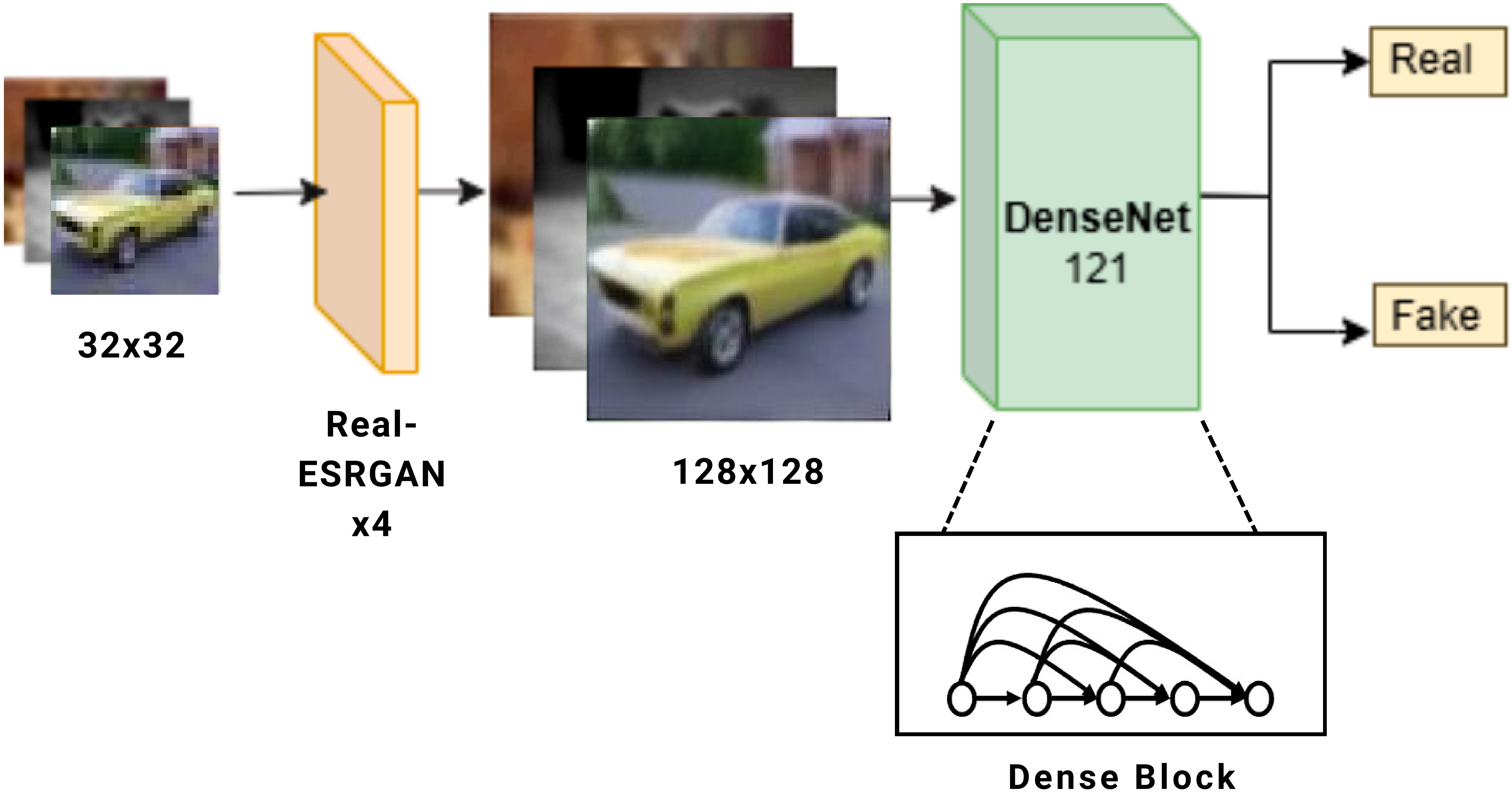
Using Real-Fake pairs of Super-resoluted CIFAKE dataset

Alternate epochs involved training on FGSM attacked images

RESULT

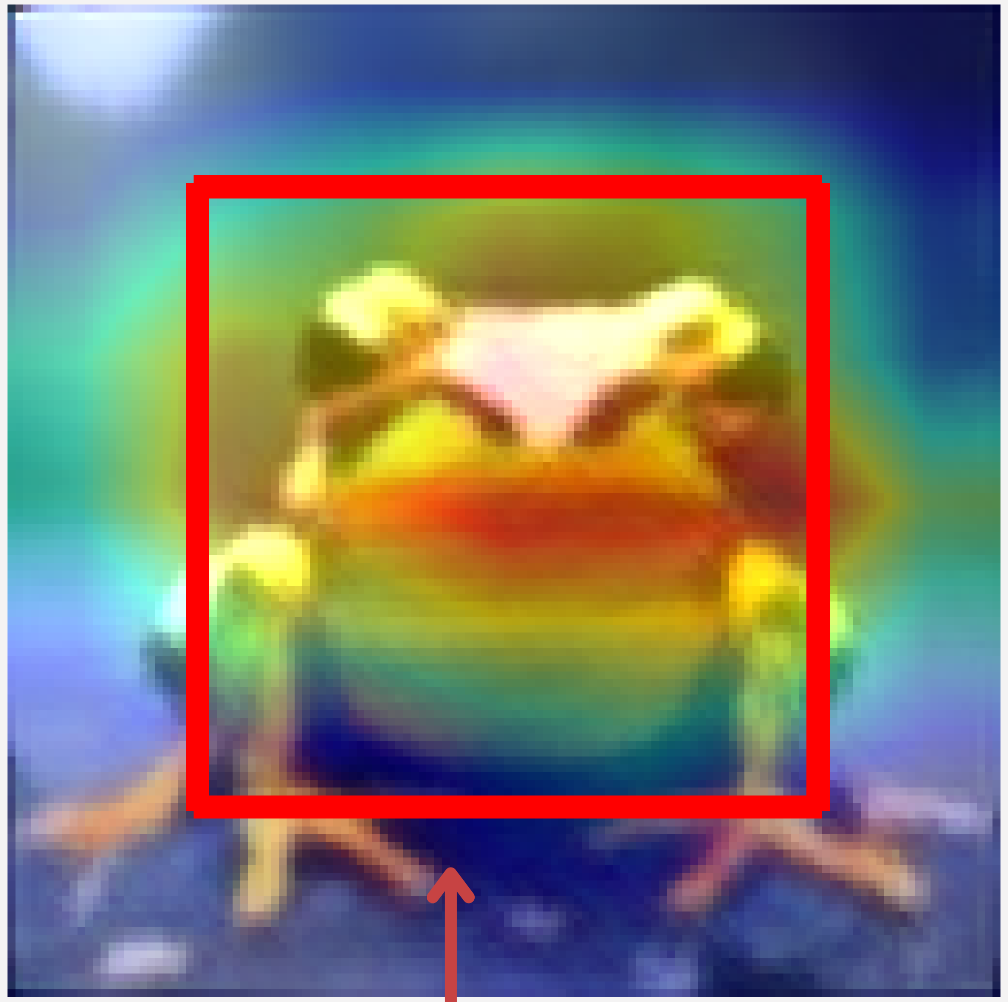
97.9% accuracy on CIFAKE test dataset.

PIPELINE FOR TASK 1



GRAD-CAM

HOW TO LOCALISE REGIONS OF INTEREST (ROI)?



MARKING A ROI

Marked a red box of 80x80 pixels around the region of highest activation (denoted by dark red) and then passed it through the VLM

GRAD-CAM

Used on the last convolution layer of DenseNet121 to generate a heatmap, superimposed on the image.

WORKING

Stores activations from a specific layer and calculates gradients with respect to the prediction

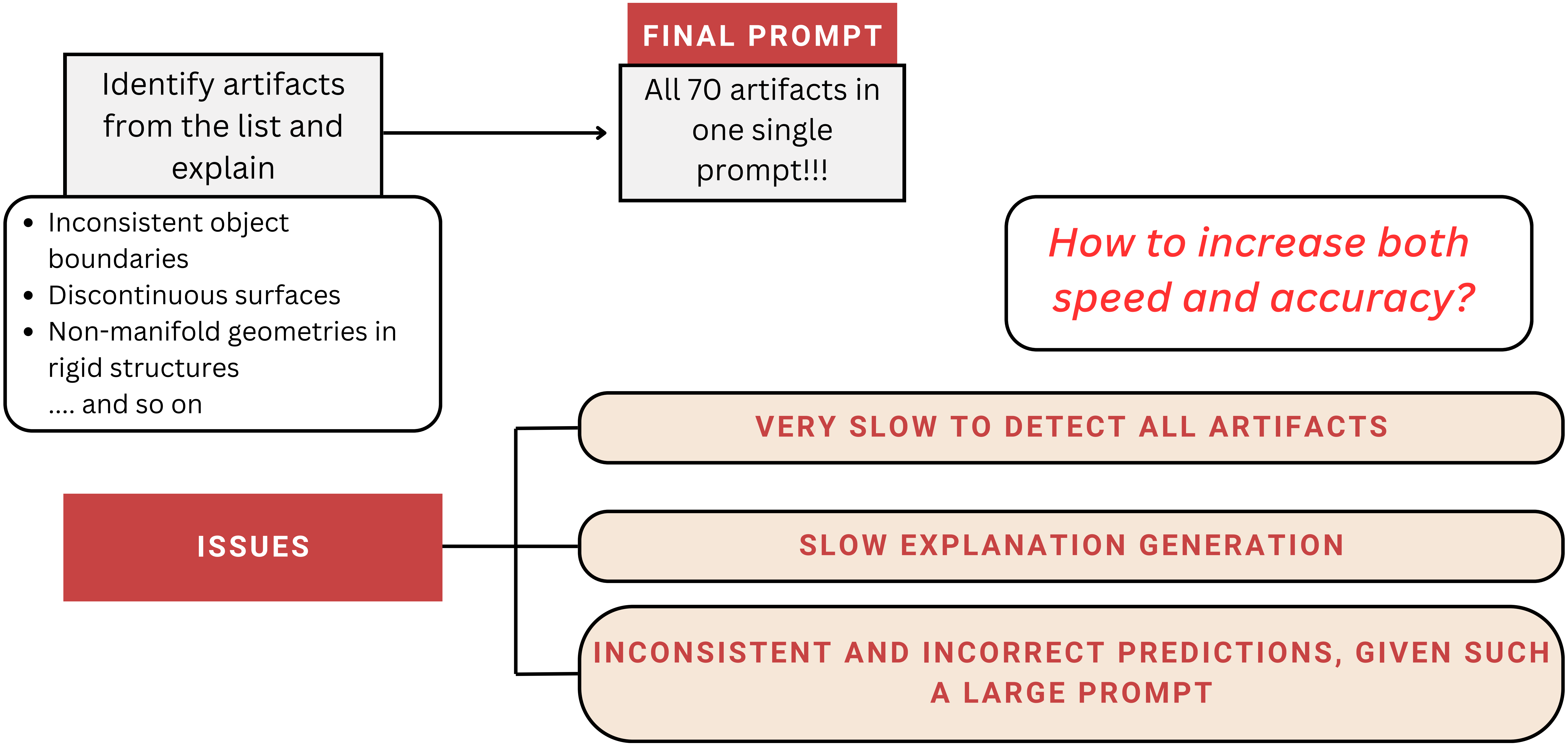
IMPORTANCE MAPPING

ed region should ideally contain the artifact present inimage.Stores activations from a specific layer and calculates gradients with respect to the prediction

TASK 2

ARTIFACT DETECTION AND EXPLANATION GENERATION

Issues with Simple & Direct Prompting



Identify artifacts from the list and explain

- Inconsistent object boundaries
- Discontinuous surfaces
- Non-manifold geometries in rigid structures

.... and so on

FINAL PROMPT

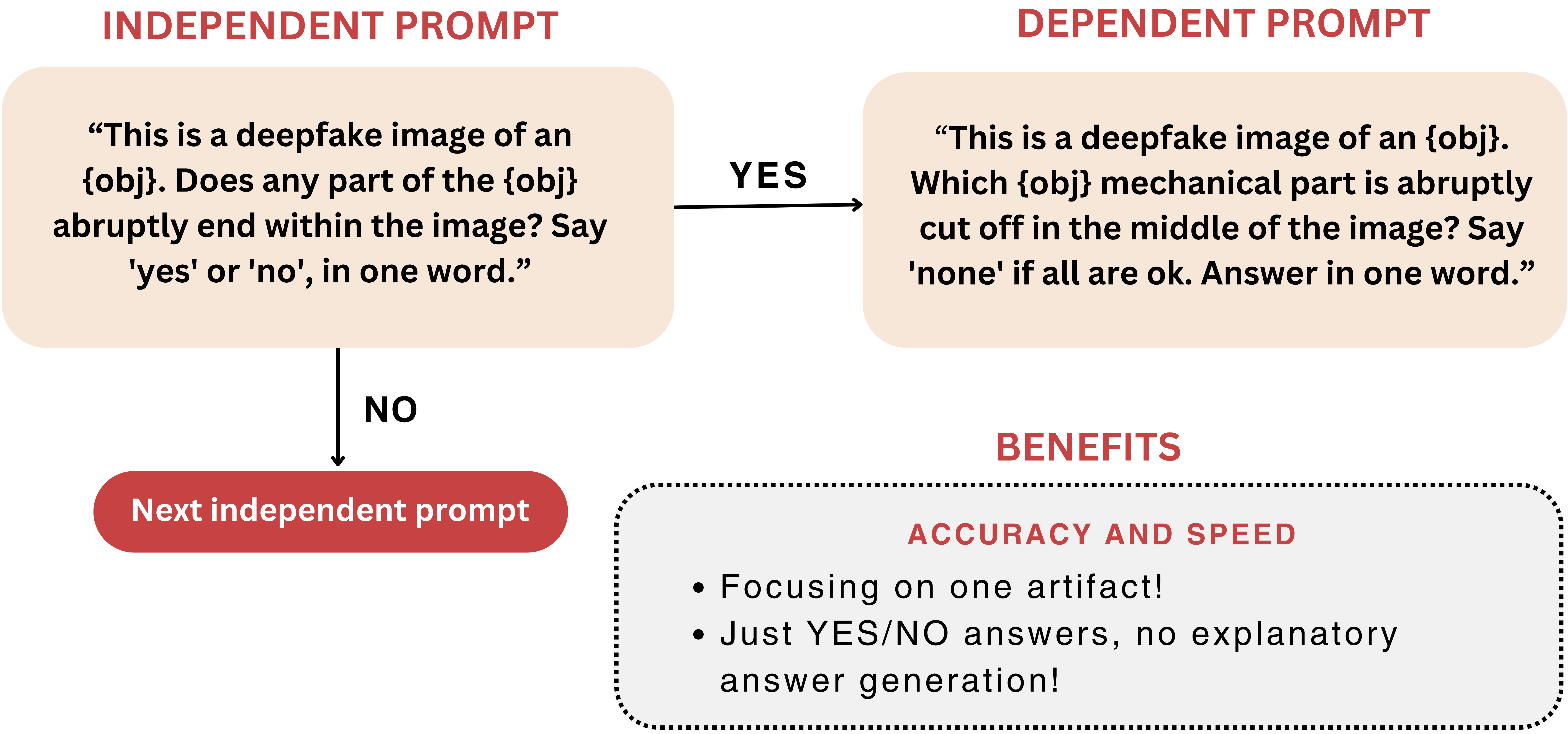
All 70 artifacts in one single prompt!!!

How to increase both speed and accuracy?

ISSUES

- VERY SLOW TO DETECT ALL ARTIFACTS
- SLOW EXPLANATION GENERATION
- INCONSISTENT AND INCORRECT PREDICTIONS, GIVEN SUCH A LARGE PROMPT

Yes/No Question Hierarchy



INDEPENDENT PROMPT

"This is a deepfake image of an {obj}. Does any part of the {obj} abruptly end within the image? Say 'yes' or 'no', in one word."

YES

DEPENDENT PROMPT

"This is a deepfake image of an {obj}. Which {obj} mechanical part is abruptly cut off in the middle of the image? Say 'none' if all are ok. Answer in one word."

NO

Next independent prompt

BENEFITS

ACCURACY AND SPEED

- Focusing on one artifact!
- Just YES/NO answers, no explanatory answer generation!

Artifact Categorization

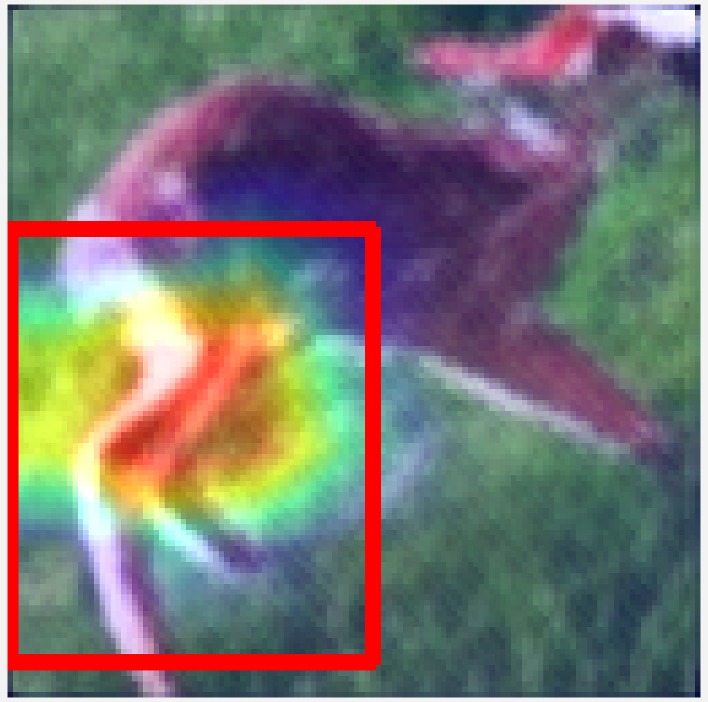
Still, 70 Questions per image? NO!

Well an average of 7 questions per image! How?

Not all artifacts apply to every image. They can be categorized based on:

- Object Type: Living or non-living.
- Artifact Location: Present on the animal's face, body, or background.
- Object Class: Specific object category.

This **hierarchical approach** along with **categorization of artifacts w.r.t class and region** reduces the sample set of potential artifacts for each image, significantly **improved processing speed.**



[✓ Living] or Non-Living?

↓

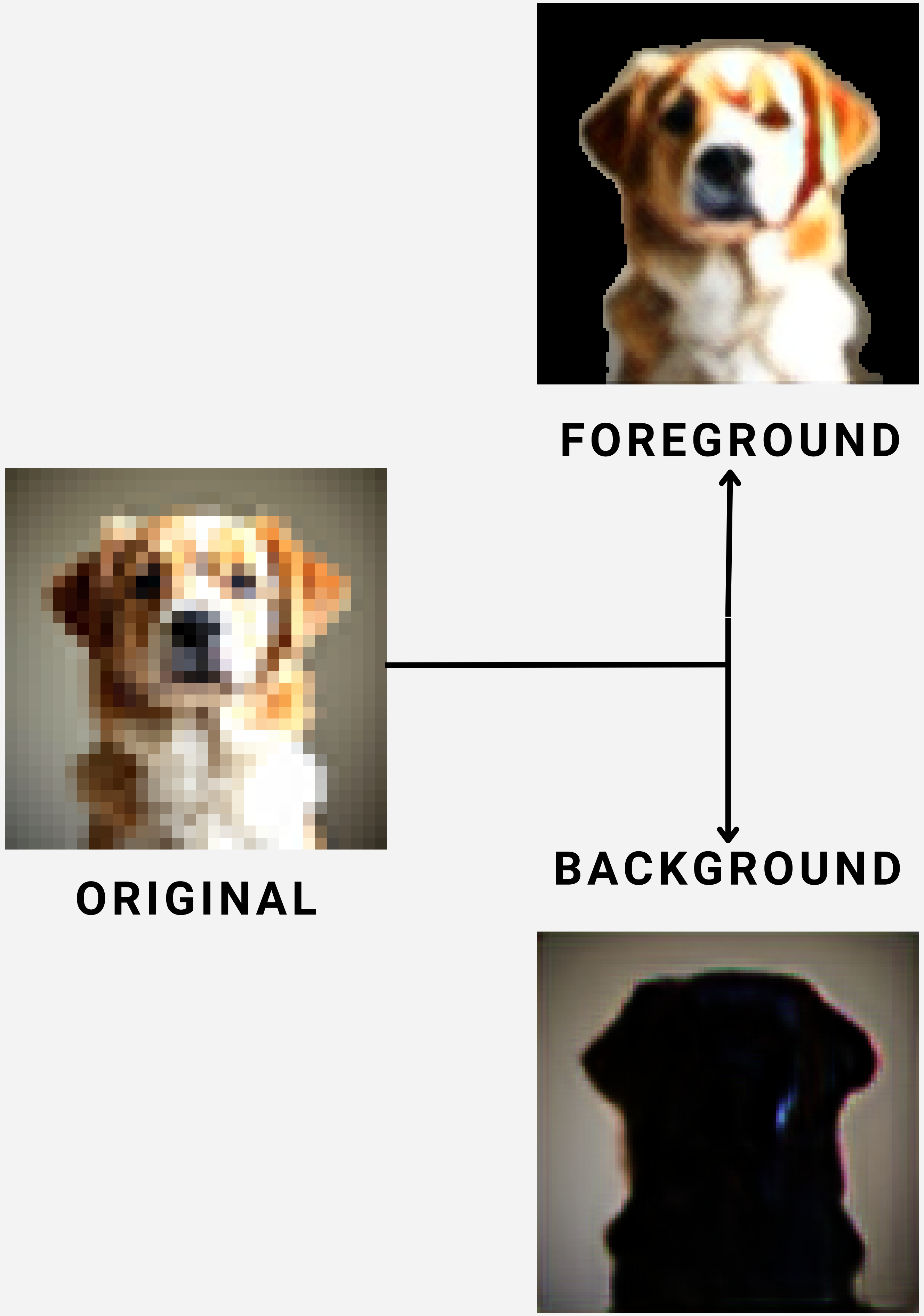
Face, [✓ Body] or Background?

↓

Class - {obj} : Deer

↓

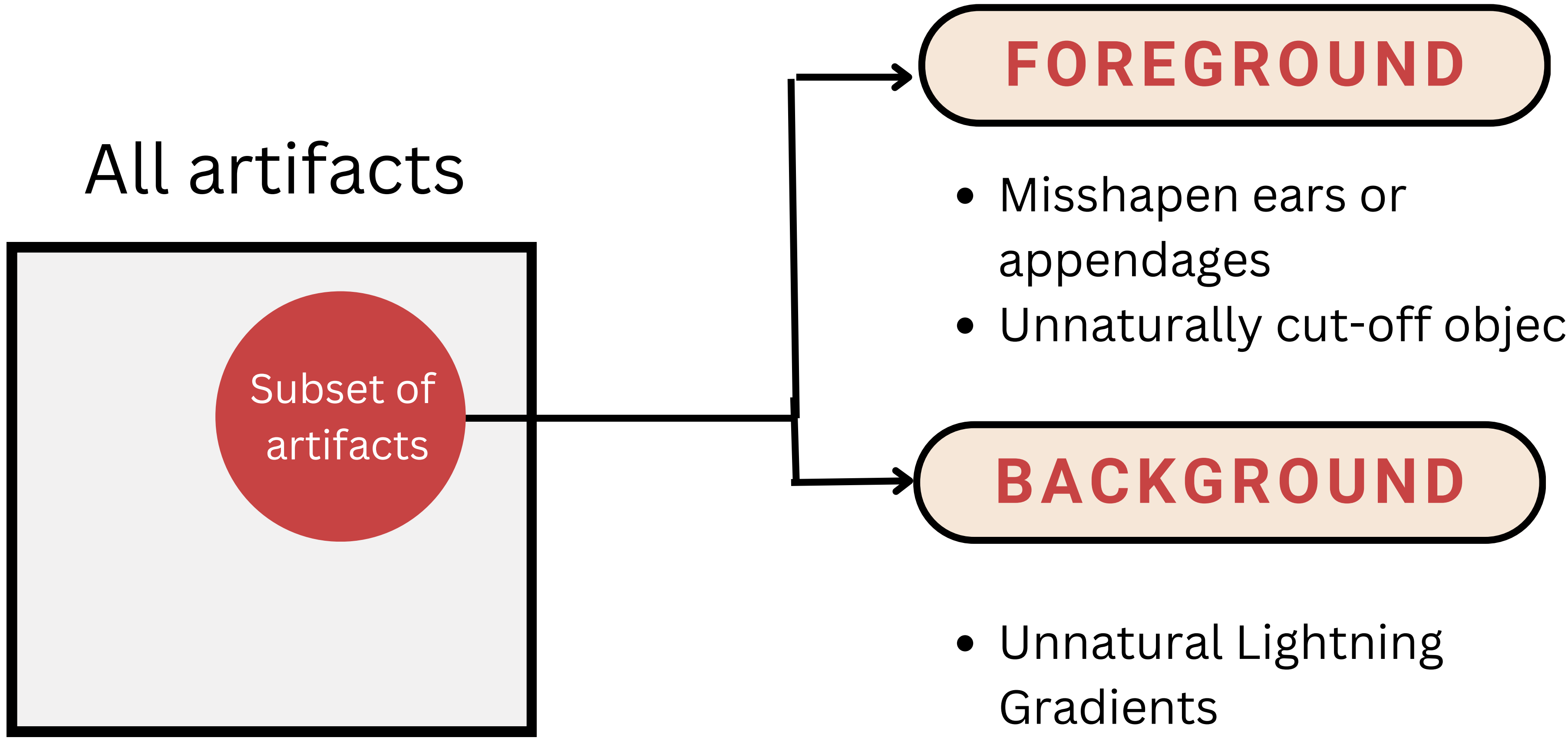
Deer Body Specific Queries!



FOREGROUND

ORIGINAL

BACKGROUND



All artifacts

Subset of artifacts

FOREGROUND

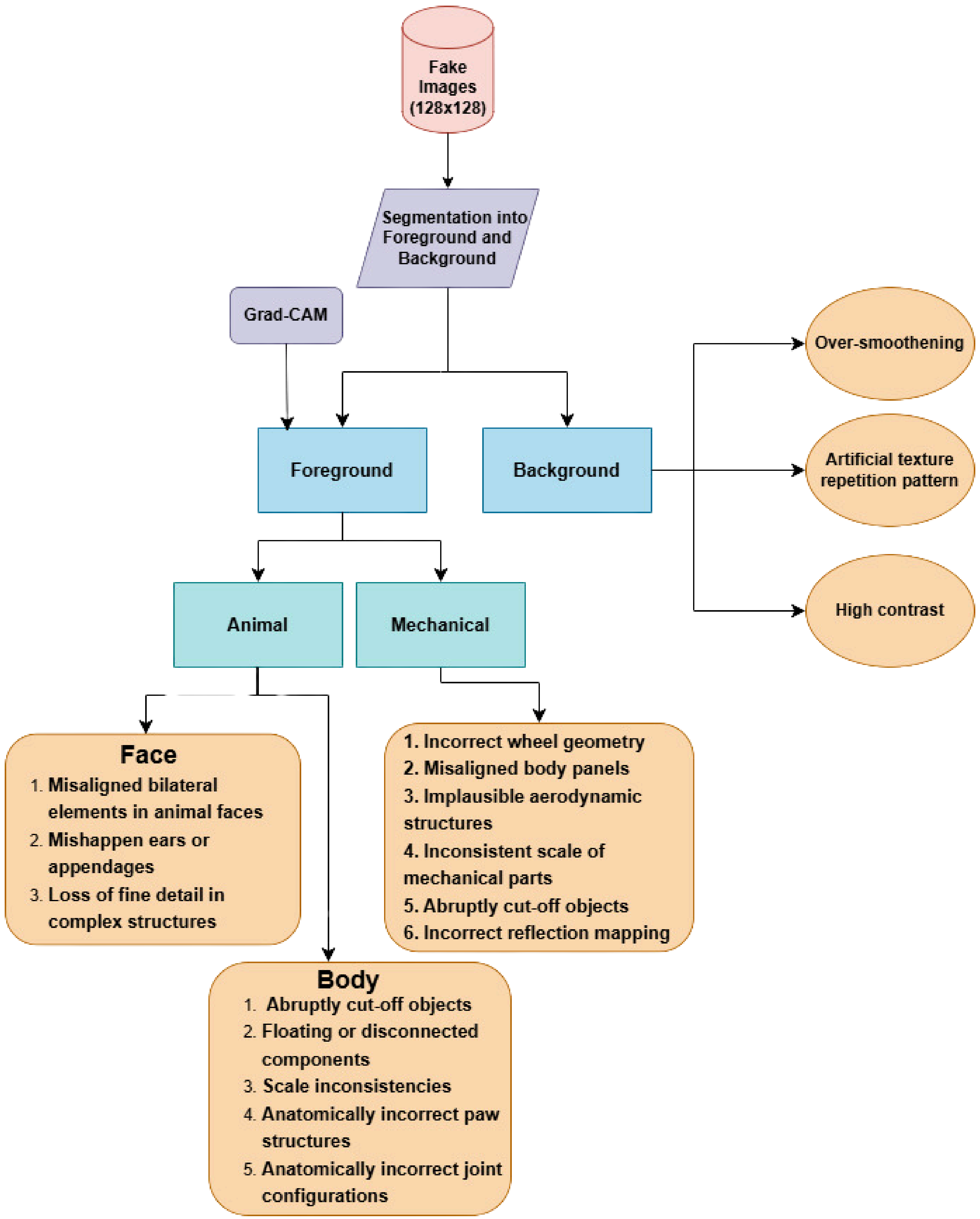
- Misshapen ears or appendages
- Unnaturally cut-off objects

BACKGROUND

- Unnatural Lightning Gradients
- Cinematization effects
- Artificial Smoothness ....

MASKING THE IMAGE

- **Model Implementation**: Implemented a **C-GAN** for the task.
- **Data Preparation**: Manually annotated a small number of background and foreground on super-resolved images.
- **Training Details**: Trained the model with this dataset for ~300 epochs.
- **Inference Limitation**: Inference results were constrained by the initial dataset quality.



Fake Images (128x128)

Segmentation into Foreground and Background

Grad-CAM

Foreground

Background

Over-smoothening

Artificial texture repetition pattern

High contrast

Animal

Mechanical

Face

1. Misaligned bilateral elements in animal faces
2. Mishappen ears or appendages
3. Loss of fine detail in complex structures

Body

1. Abruptly cut-off objects
2. Floating or disconnected components
3. Scale inconsistencies
4. Anatomically incorrect paw structures
5. Anatomically incorrect joint configurations

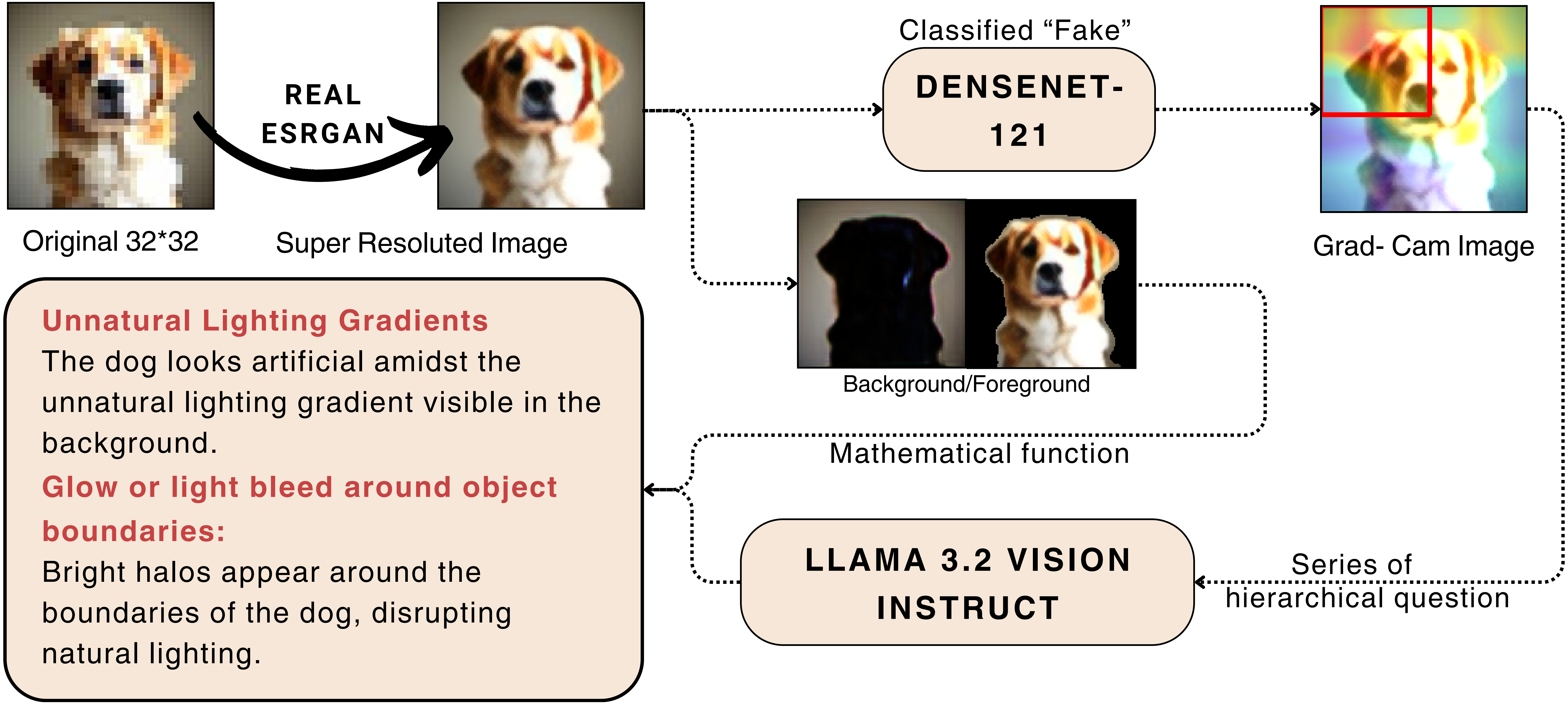
1. Incorrect wheel geometry
2. Misaligned body panels
3. Implausible aerodynamic structures
4. Inconsistent scale of mechanical parts
5. Abruptly cut-off objects
6. Incorrect reflection mapping

Performance:

- Achieves extremely fast inference speeds, approximately **5 seconds per image**.
- Direct Prompting with all 70 artifacts took 3 mins per image on average.

on NVIDIA A100 GPU

RESULTS



Unnatural Lighting Gradients

The dog looks artificial amidst the unnatural lighting gradient visible in the background.

Glow or light bleed around object boundaries:

Bright halos appear around the boundaries of the dog, disrupting natural lighting.

THANK YOU