Scarcity to Scale: Semantic Active Generative Augmentation (SAGA) for Amplifying Rare-Event Classification

Research Report for CSE 542 AMS 560 (Stony Brook University)

Group 7: Manav Ukani, Dhruv Shah, Param Patel, Neel Modi, Astha Soni

Abstract

The advancement of deep learning in computer vision is frequently obstructed by the scarcity of high-quality, balanced data, a challenge that is particularly acute in rare-event detection tasks. Real-world datasets often exhibit long-tail distributions where critical edge cases are severely underrepresented, leading to models that bias heavily toward majority classes. This report presents the research, development and comprehensive evaluation of SAGA (Semantic Active Generative Augmentation), a novel framework designed to systematically dismantle the barriers of class imbalance through controlled generative artificial intelligence. Crucially, the framework addresses the primary risk of generative augmentation: hallucination. SAGA integrates a rigorous Hallucination Guard, a multi-stage validation pipeline utilizing Structural Similarity (SSIM) for identity preservation, CLIP-based consistency checks for semantic alignment and a Classifier-in-the-Loop for decision boundary reinforcement. Applied to the AffectNet dataset, a benchmark notorious for its severe imbalance in complex emotional states, SAGA demonstrated a statistically significant improvement in model robustness. The baseline Vision Transformer (ViT) achieved a test accuracy of 60.17%, while the SAGA-augmented model reached 64.67%, an absolute improvement of 4.5%. More notably, the system achieved double-digit accuracy gains in minority classes.

1 Introduction

1.1 The Crisis of Data Scarcity and the Long-Tail

In the contemporary landscape of artificial intelligence, data is frequently cited as the new oil. However, unlike oil, the value of data is not uniform; it is heavily dependent on distribution. The "long-tail" problem remains a persistent and formidable bottleneck in the deployment of reliable computer vision systems. While collecting data for common occurrences (such as cars on a highway or "happy" faces in social media images) is trivial, capturing rare, anomalous or subtle events is exponentially more difficult. This creates a Zipfian distribution (variant of Power Law distribution) where a handful of classes dominate the dataset, while the "tail" contains a vast number of rare but often critical categories [1].

1.2 The Failure of Traditional Augmentation

Historically, researchers have attempted to mitigate class imbalance through resampling strategies (oversampling minority classes or undersampling majority ones) and traditional data augmentation. Traditional augmentation techniques, as shown in Figure 1 such as random rotation, flipping and color changes, serve to improve invariance to geometric and photometric shifts but

suffer from a fundamental limitation: they are semantically passive. For example, a rotated image of a “happy” face is, semantically, still just a “happy” face. These transformations do not introduce the new feature combinations required to teach a model the subtle muscle configurations associated with a “fearful” expression if those examples are missing from the training distribution. The augmentation process must be semantically active and capable of synthesizing entirely new instances of the rare class that possess valid semantic features.

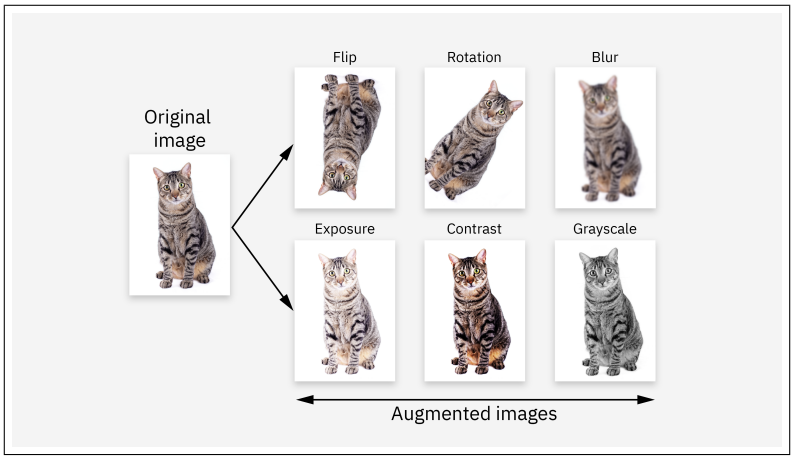


Figure 1: Traditional methods in data augmentation for images

1.3 The Generative AI Opportunity and the Hallucination Risk

The advent of generative artificial intelligence, particularly diffusion models and large multi-modal models, offers a theoretical solution to this impasse. These models have learned the statistical distribution of visual concepts from billions of image-text pairs and can, in principle, synthesize infinite variations of any given class. This capability presents an opportunity to artificially populate the ”tail” of the distribution with high-fidelity synthetic data [2]. However, the integration of generative AI into training pipelines is fraught with risk. Generative models are prone to ”hallucination”. The production of outputs that are statistically plausible but factually incorrect or semantically misaligned as seen in Figure 2.

Figure 2: Images Rejected by SAGA Framework (considered to be AI hallucinations)

1.4 The SAGA Framework

This report details the development of SAGA (Semantic Active Generative Augmentation), a framework designed to harness the power of generative AI while strictly mitigating its risks. SAGA moves beyond passive augmentation to active data creation, using a TargetBalancing-Manager to identify data deficits and a generative engine to fill them. The core innovation of SAGA is its Hallucination Guard, a rigorous, multi-modal filtering system that vets every

synthetic image for structural fidelity, semantic consistency and classifier confidence before it is permitted to enter the training set as seen in Figure 3. By solving the dual challenges of scarcity and quality, SAGA bridges the gap between theoretical generative capabilities and practical, industrial-grade model performance.

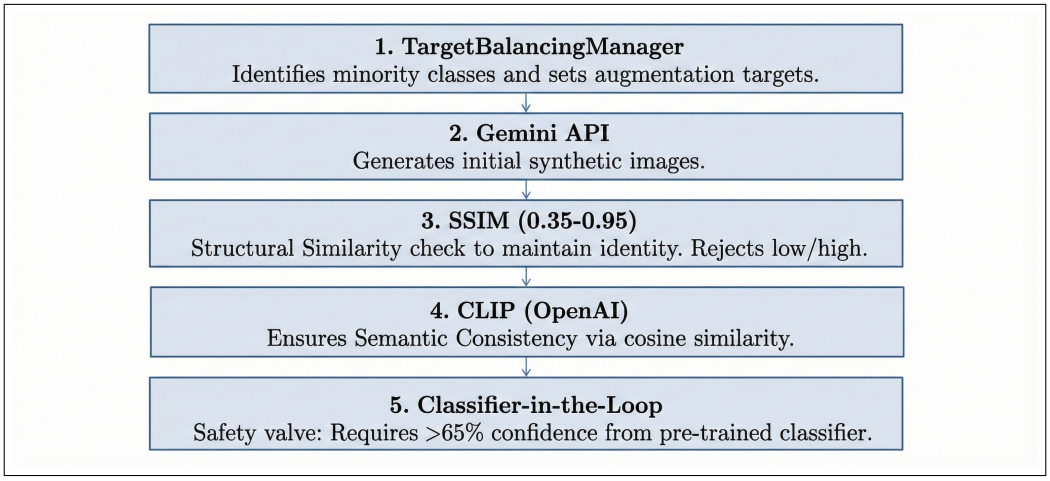


Figure 3: The SAGA Framework Architecture

2 Background & Motivation

2.1 Theoretical Underpinnings of Class Imbalance

The problem of class imbalance is rooted in the optimization dynamics of deep neural networks. Standard loss functions, such as Cross-Entropy Loss, treat all samples equally. When the majority class constitutes 90% of the data, the gradient descent process is overwhelmingly driven by the errors made on these majority samples. The network quickly learns that the path of least resistance to minimizing global loss is to predict the majority class for every input. This results in a model that may have 90% accuracy but a recall of 0% for the minority class-a completely unacceptable outcome for rare-event detection tasks [1] [3].

2.2 Vision Transformers: A Global Perspective

This project employs a Vision Transformer (ViT) architecture. The motivation for this choice lies in the inductive biases of CNN architectures, built on the convolution operation, which prioritizes local connectivity. Facial expressions, however, are inherently global. An emotion is defined not by the mouth alone or the eyes alone, but by the geometric relationship between them. A frown (mouth) combined with wide eyes might indicate "fear," while the same frown with furrowed brows indicates "anger." ViT processes an image as a sequence of patches and utilizes Self-Attention mechanisms to model the relationship between every patch and every other patch simultaneously.

2.3 The Necessity of Controlled Generation

The specific nature of the AffectNet task-emotion recognition-imposes strict constraints on generation. Emotion augmentation involves facial editing, we must take an existing face and change its emotion while preserving its identity. This Identity-Emotion Disentanglement is the central challenge. Therefore, any generative augmentation pipeline for this domain must include robust checks to ensure that the style (emotion) is changed while the content (identity) is preserved.

3 Literature Review

3.1 Traditional Data Augmentation

Standard data augmentation techniques have been the first line of defense against overfitting for decades. Geometric transformations (rotation, scaling, flipping) and photometric transformations (color jittering, brightness adjustment) are widely used to simulate variations in imaging conditions [4]. Advanced mixing techniques like MixUp (linear interpolation of pixel values) and CutMix (pasting patches of one image onto another) have shown promise in improving robustness. However, these methods produce unnatural, chimera-like images that lack semantic realism [5].

3.2 Oversampling Techniques in Pixel Space

Techniques like SMOTE (Synthetic Minority Over-sampling Technique) generate synthetic samples by interpolating between existing minority samples in the feature space. While highly effective for tabular data, SMOTE struggles in the high-dimensional pixel space of images. Linear interpolation between two images often results in "ghosting" or blurriness that destroys the fine-grained texture details necessary for emotion recognition [5].

3.3 Generative Augmentation: GANs vs. Diffusion

The rise of Generative Adversarial Networks (GANs) introduced the ability to synthesize realistic images. Conditional GANs (cGANs) and CycleGANs have been used to translate images between domains. However, GANs are notoriously difficult to train and suffer from mode collapse [6].

Diffusion Models (e.g., Stable Diffusion) operate by iteratively denoising a latent representation and have largely superseded GANs in terms of image fidelity and diversity. They are less prone to mode collapse and can generate highly diverse outputs from text prompts [7].

3.4 Hallucination Detection in Synthetic Data

The concept of "hallucination" in Generative AI is well-documented in Natural Language Processing (NLP) but is equally critical in Computer Vision [8]. Current literature suggests that automated metrics like Fréchet Inception Distance (FID) are insufficient for detecting individual hallucinations, as they measure distributional distance rather than per-sample correctness[9]. Emerging approaches advocate for reference-based evaluation and semantic consistency checks using Vision-Language models like CLIP, a strategy that heavily informs the SAGA methodology.

3.5 State-of-the-Art on AffectNet

AffectNet is widely regarded as one of the most challenging benchmarks in facial expression recognition. Current state-of-the-art methods typically achieve accuracies in the range of 58% to 63% on the 8-class classification task. Many of these methods still show significant confusion between "Contempt," "Neutral," and "Sadness," highlighting the persistent difficulty of the long-tail classes [10].

4 Methodology

4.1 Dataset: AffectNet

The project utilizes AffectNet, the largest database of facial expressions, valence and arousal in-the-wild. The original dataset has heavy-tailed distribution favouring "Happy" and "Surprise" classes, while "Contempt," "Disgust," and "Fear" are rare [3].

To ensure the integrity of the results, 1,200 images (150 per class) were isolated from others as the **Golden Test Set** prior to any training or augmentation. These images were strictly **unseen.**

4.2 Baseline Model Architecture

The backbone of the classification system is a Vision Transformer (ViT), specifically the `vit_base_patch16_224` variant. Input ($224 \times 224 \times 3$) is divided into 16×16 patches. Along with 12 stacked Transformer layers with Multi-Head Self-Attention (MHSA). The standard linear projection head was replaced with a custom Multi-Layer Perceptron (MLP) as shown in Figure 4

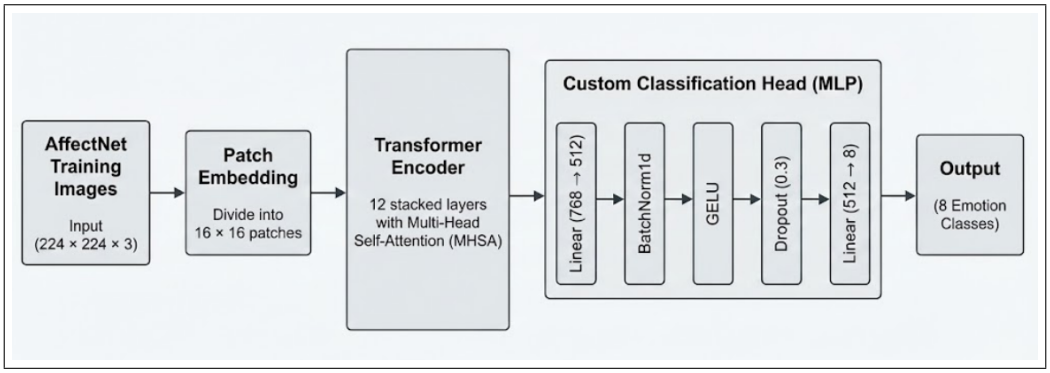


Figure 4: Baseline Architecture

4.3 SAGA Framework: The Augmentation Pipeline

The SAGA framework operates as an autonomous agent that actively monitors dataset health and intervenes to correct imbalances.

4.3.1 Phase 1: Target Balancing Manager

The TargetBalancingManager class autonomously calculates the deficit for each class. The manager cycles through target emotions via Round-Robin Scheduling to prioritize highest deficits and to ensure maximum source diversity across all target classes, preventing the injection of source-based feature bias into the synthetic dataset.

4.3.2 Phase 2: Generative Engine (Gemini 1.5 Flash)

We selected Google's Gemini 1.5 Flash model using Google Cloud API. The prompt used was engineered to enforce strict constraints:

> Edit this image. The person currently has a {current_emotion} expression. Change their facial expression to {target_emotion}. Maintain the exact same identity, lighting and background. Output only the modified image.

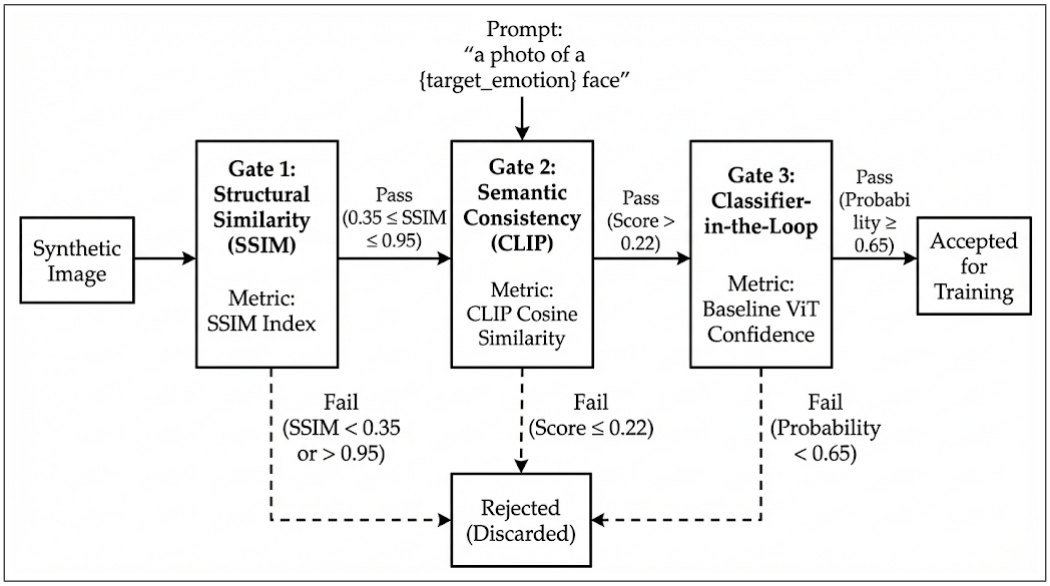


Figure 5: The Triple-Lock Hallucination Guard Mechanism.

4.3.3 Phase 3: Hallucination Guard

An image is accepted only if it passes all three gates as shown in Figure 5.

1. **Gate 1: Structural Similarity (SSIM)**
   Metric: SSIM Index.
   Acceptance Band: $0.35 - 0.95$. Ensures the person is the same (identity preserved) but the image has changed enough to reflect a new expression.

2. **Gate 2: Semantic Consistency (CLIP)**
   Metric: CLIP Cosine Similarity.
   Threshold: Score > 0.22. Ensures the image semantically aligns with the text prompt "a photo of a {target_emotion} face."

3. **Gate 3: Classifier-in-the-Loop**
   Metric: Baseline ViT Confidence.
   Threshold: Probability ≥ 0.65 from the baseline model. Ensures the image reinforces the decision boundary and the expression in augmented image is distinct enough to learn from with respect to the original image and target emotion.

4.4 Team Contributions

- **Astha:** Dataset curation, preprocessing and initial baseline methods.
- **Neel:** Trying different models, finalizing Vision Transformer architecture, custom MLP head and baseline training pipeline.
- **Manav:** AI hallucination mitigation research, working on SAGA augmentation pipeline to resolve challenges and Gemini integration.
- **Dhruv:** Retraining pipelines, confusion matrix analysis and results synthesis.
- **Param:** Investigation into alternative generative models (OpenAI, MidJourney) and industrial scalability analysis.

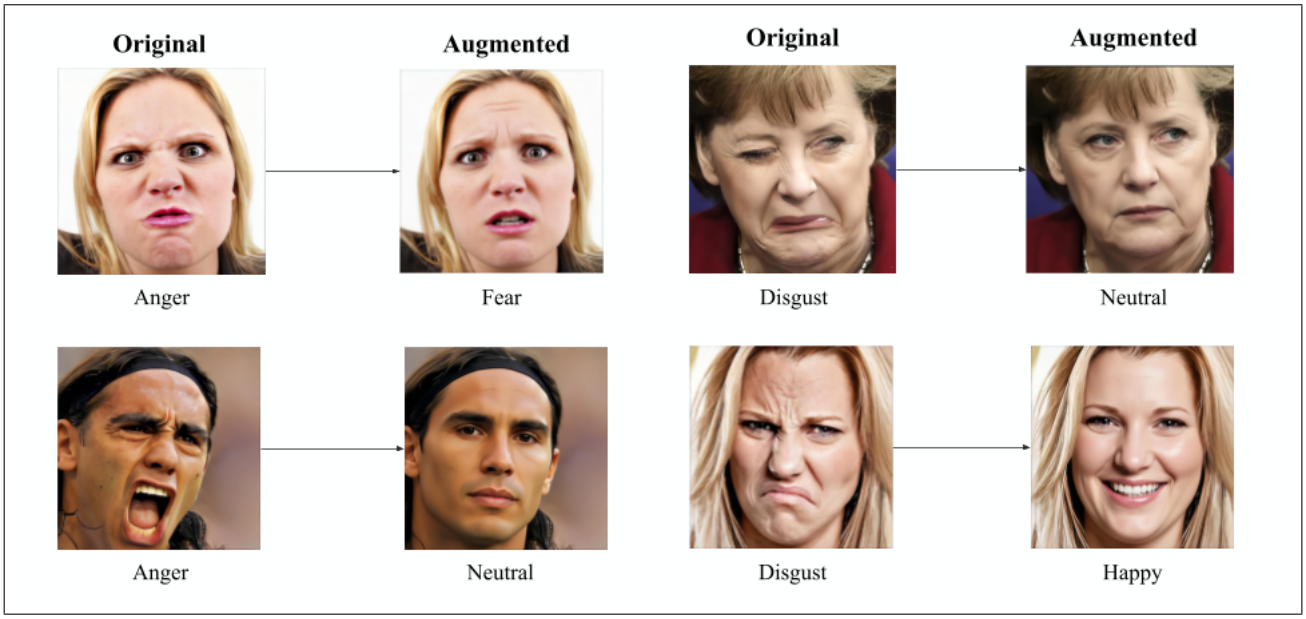


Figure 6: Source images and their high-fidelity, accepted augmented counterparts, verified by the SAGA framework

5 Novelty of Our Approach

The SAGA framework diverges from existing literature in three fundamental ways:

1. **Active vs. Passive Augmentation:** SAGA uses a TargetBalancingManager to surgically fill data deficits rather than applying uniform transformations.
2. **Emotion as Style Transfer:** By framing the task as Identity-Preserving Semantic Editing, SAGA avoids mode collapse and preserves the rich "in-the-wild" variance of AffectNet.
3. **The Triple-Lock Safety Mechanism:** The combination of SSIM, CLIP and Classifier Confidence transforms Generative AI from a "black box" into a reliable component of the training pipeline.

6 Results

6.1 Quantitative Analysis

Table 1: Global Performance Metrics

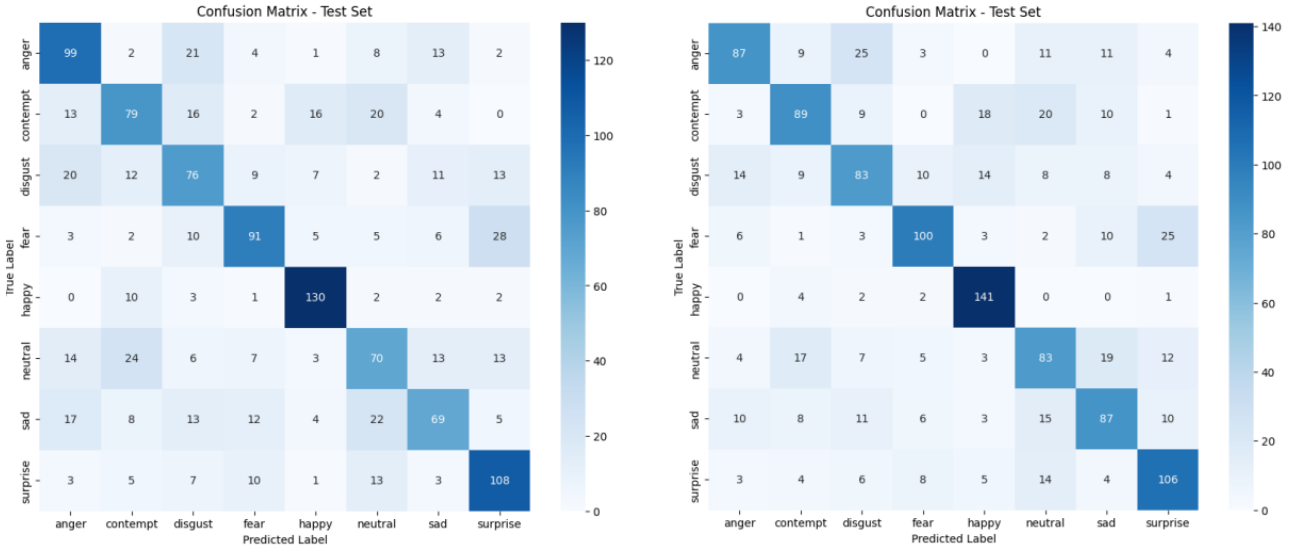
Metric	Baseline Model	SAGA Augmented	Absolute Improvement
Test Accuracy	60.17%	64.67%	+4.50%
Validation Accuracy	69.39%	64.62%	-4.77% (baseline overfitting)

The **4.5% absolute improvement** on the strictly **unseen test set** is statistically significant. The baseline's higher validation accuracy suggests it was overfitting to the easy majority classes.

Table 2: Class-Wise Breakdown (Minority Classes)

Class	Baseline Accuracy	SAGA Accuracy	Improvement
Sadness	46.00%	58.00%	+12.00%
Contempt	52.67%	59.33%	+6.66%
Fear	60.67%	66.67%	+6.00%
Happy	86.67%	94.00%	+7.33%

Sadness (+12%) represents the most dramatic result. The synthetic data likely provided the "connective tissue" needed to distinguish the subtle droop of eyelids or mouth corners separating "Sadness" from "Neutral."



(a) Baseline Model

(b) Augmented (SAGA) Model

Figure 7: Side-by-side comparison of Baseline vs SAGA Confusion Matrices

6.2 Qualitative Analysis

We conducted a Google Forms survey, through the Piazza where human participants (students) guessed on ambiguous edge-case images (resulting in almost 50/50 splits), however SAGA model confidently classified them correctly. Figure 8 suggests SAGA enables the detection of super-human micro-features.

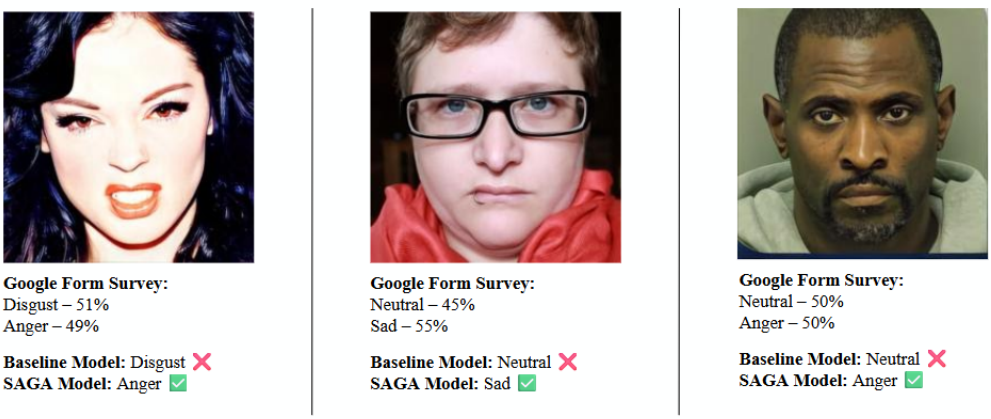


Figure 8: Human Survey Benchmarking

Apart from that, as seen in Figure 9, we also developed a Streamlit UI where the model's improved performance can be experienced with live low latency video feeds.

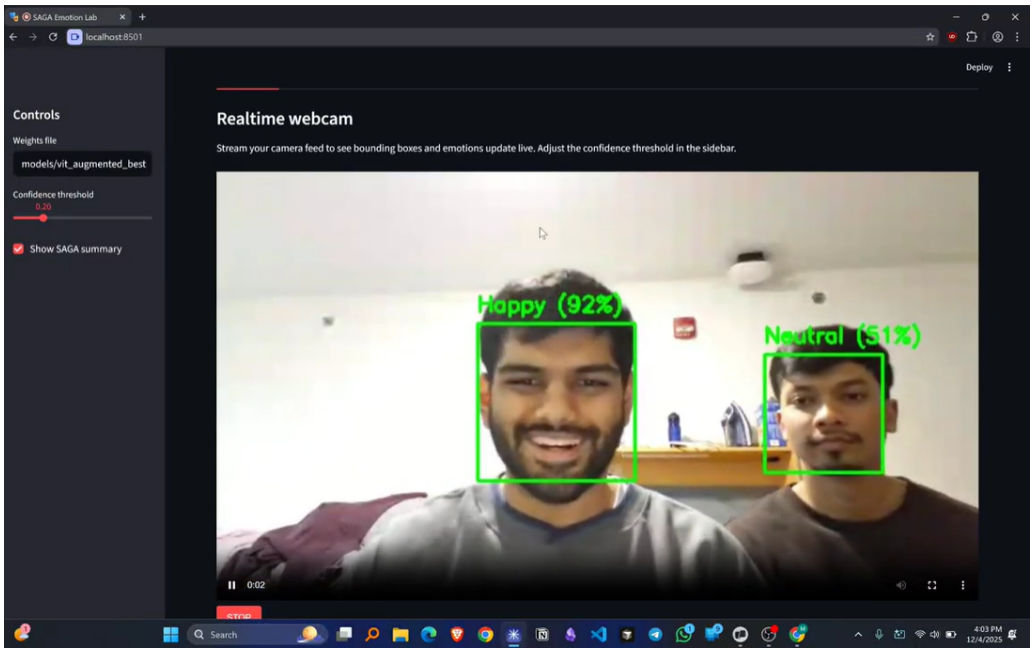


Figure 9: Real-Time Demonstration with deployed model on Streamlit UI

7 Conclusion & Future Work

This research establishes SAGA as a rigorous solution to class imbalance. By wrapping a generative engine in a **Triple-Lock Hallucination Guard**, the framework successfully synthesizes high-fidelity training data. Future work includes domain expansion to medical imaging, automated thresholding via meta-learning and diffusion fine-tuning (e.g., LoRA).

A Appendix

A.1 List of Project Deliverables

The following items constitute the complete delivery of the project:

- **Training Notebooks**: Pipeline for training the reference model and fine-tuned model on the combined dataset (available in GitHub).
- **Augmentation Pipeline Code:** Python script containing the `TargetBalancingManager` and `HallucinationGuard` classes for image generation (available in GitHub).
- **Real-Time Demo UI:** Streamlit application code for live inference through webcam/upload. Available in GitHub.
- **Fine-tuned Model Weights:** Baseline and SAGA-Augmented ViT, access on Kaggle.
- **Combined (Augmented+Original) Dataset:** Final dataset, with `combined_labels.csv`. Available on Kaggle.
- **Slide Deck:** Presentation summarizing architecture and results, also on GitHub.

A.2 Code & Data Links

- **Code Repository:** https://github.com/manavukani/saga-rare-event-detection
- **Kaggle Dataset:** https://www.kaggle.com/datasets/manavukani/affectnet
- **Trained Model Weights:** https://www.kaggle.com/datasets/manavukani/my-models

A.3 Instructions for Running the Code

The project is modular and distributed in 3 folders (training, generation, ui). The setup instructions are detailed in the GitHub README files for each of them. Refer to these links:

- Image Generation (Augmentation): Link
- Model Training: Link
- Streamlit UI: Link

References

[1] Optimizing Class Imbalance in Facial Expression Recognition Using Dynamic Intra-Class Clustering. *PMC - NIH*. Accessed Dec 5, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC12109554/

[2] Hallucination (artificial intelligence). *Wikipedia*. Accessed Dec 5, 2025.

[3] AffectNet+: A Database for Enhancing Facial Expression Recognition with Soft-Labels. *arXiv*, 2024. https://arxiv.org/html/2410.22506v1

[4] A Complete Guide to Data Augmentation. *DataCamp*. Accessed Dec 5, 2025.

[5] A Comprehensive Survey on Imbalanced Data Learning. *arXiv*, 2025. https://arxiv.org/html/2502.08960v3

[6] GANs vs Diffusion Generative AI Comparison. *SabrePC Blog*. Accessed Dec 5, 2025.

[7] DREAM: On hallucinations in AI-generated content for nuclear medicine imaging. *arXiv*, 2025. https://arxiv.org/html/2506.13995v2

[8] Mitigating Object Hallucination via Data Augmented Contrastive Tuning. *arXiv*, 2024. https://arxiv.org/html/2405.18654v1

[9] Rethinking the Learning Paradigm for Facial Expression Recognition. *arXiv*, 2022. https://arxiv.org/html/2209.15402v3

[10] Facial Expression Recognition Based on Weighted-Cluster Loss and Deep Transfer Learning Using a Highly Imbalanced Dataset. *MDPI*, 2020. https://www.mdpi.com/1424-8220/20/9/2639