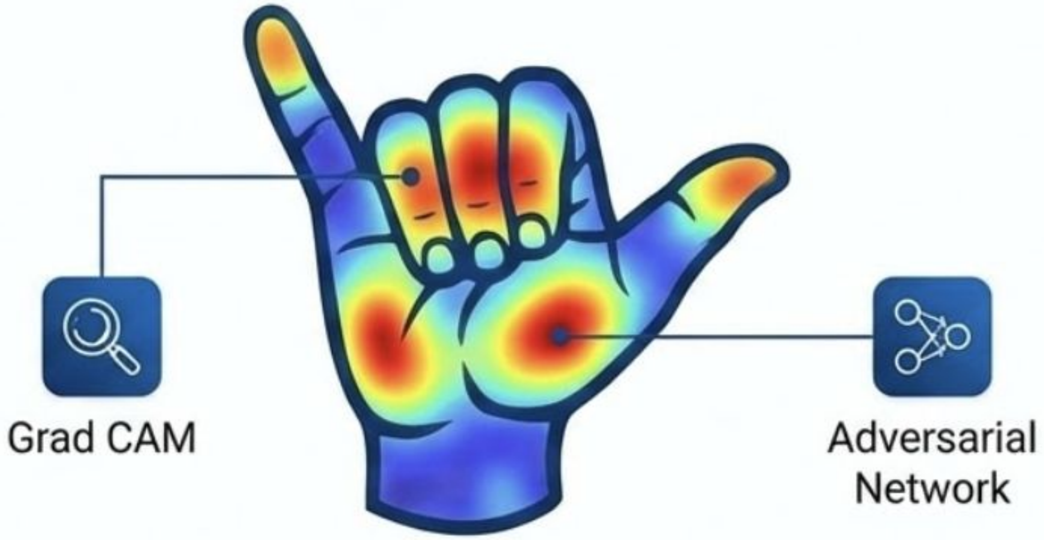
Reducing Bias in ASL Classification Models



The Problem: Skin-tone Bias

Lighter Skin Tone	Darker Skin Tone
	
✅Correctly Classified	❌Misclassified

We investigate how model performance varies across skin tone groups to identify bias

Goal & Expected Impact

- **Primary Goal:** Utilize Grad-CAM to demonstrate bias mitigation using adversarial networks on ASL letter classification

- **Expected Outcome:** Debias model without sacrificing predictive power and make debiasing more interpretable

Our Data

- **Dataset:** 42000 static images, “A-Z”, “Space”, “Del” taken by three individuals from different angles and lighting (per individual, not per image) manually labeled by us

- **Limitations:** Only contains three individuals and is not labeled, exposed to human bias. Lack of diversity means easy classification, not much bias present

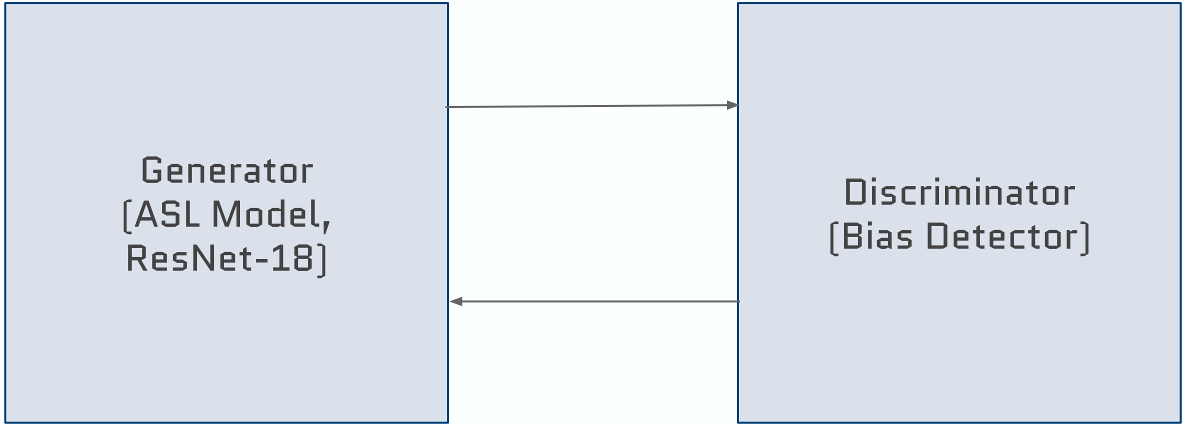
- **Future Work:** Diverse dataset with more imagery, diversity in background/lighting/shadows per image and letter per person, manually labeled demographics, similar to “ASL Citizen” Dataset

Methodology: Grad-CAM

Original Image	Grad-CAM Overlay
	

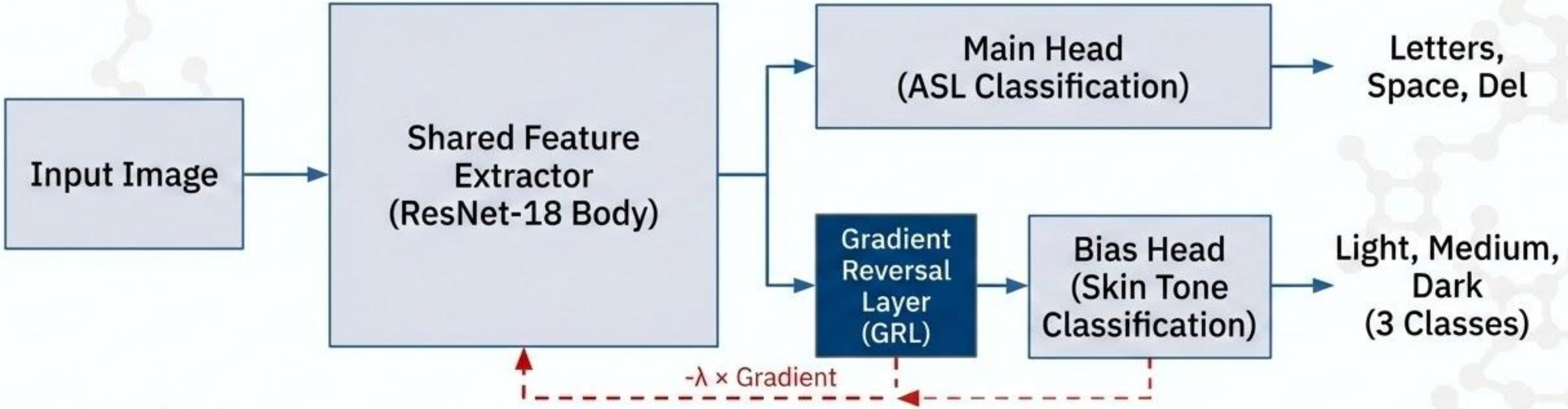
Grad-CAM visualizes which regions of the image are most important for the model's prediction

Methodology: Adversarial Network



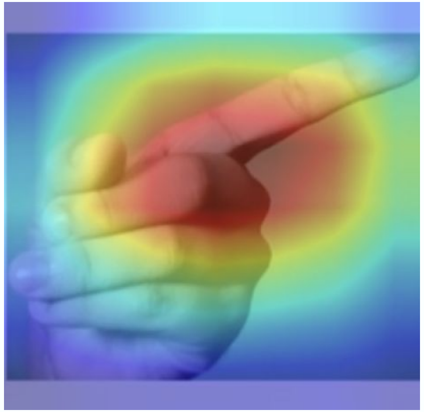
The adversarial network consists of a generator (our baseline model) and a discriminator. The discriminator tries to detect skin-tone bias in the model's output, and this feedback is used to train the generator to be less biased.

Methodology: Gradient Reversal Layer (GRL)



During backpropagation, the GRL reverses the gradient from the Bias Head (multiplies by -λ). This forces the Shared Feature Extractor to learn representations that are ***invariant*** to skin tone, effectively "unlearning" the bias information.

Results

Baseline	De-biased
	
Less Focus	**More Focus**

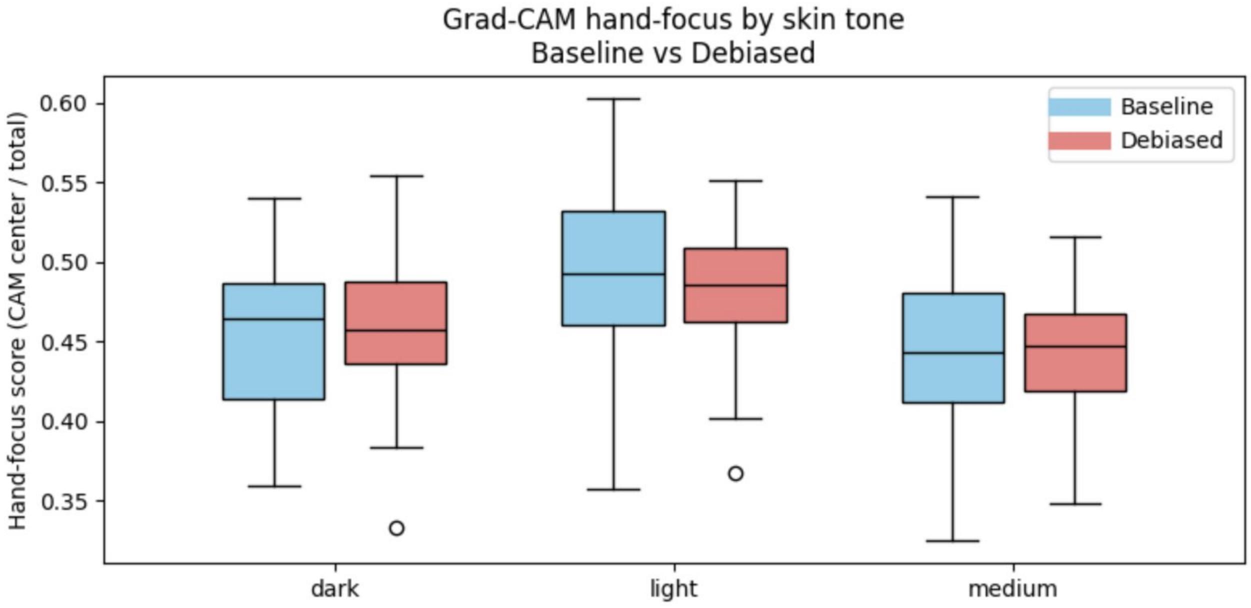
The de-biased model expands hand-focus region for both medium and dark skin-tone groups.

Results

Baseline	De-biased
99 % Accuracy	-.05% Accuracy

Accuracy maintained between both models with no significant differences

Results



Hand-focus Consistency Improved for Debiased Model

Challenge: High Baseline Accuracy

- Baseline model already achieves very high accuracy (>99%) on the test set.
- Difficult to find a metric that significantly differentiates the baseline and debiased models.
- Standard accuracy metrics may not capture subtle improvements in fairness or interpretability.
- Requiring alternative evaluation methods beyond simple accuracy scores.

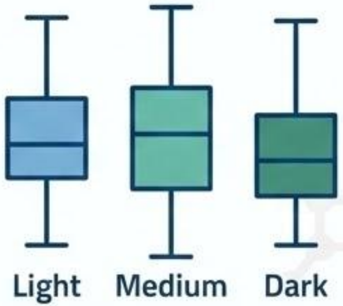
Limitations & Challenges

Initial Interpretation Challenge

- Grad-CAM intended for interpretability, but adversarial network behavior was initially unclear.
- Comparing baseline accuracy didn't clearly show effective debiasing.
- Difficulty assessing if the model was genuinely "unlearning" bias.

Solution: Boxplot Visualization

- Visualizing accuracy across skin tones using boxplots provided clarity.
- Boxplots revealed performance distribution for Light, Medium, and Dark groups.
- Enabled concrete assessment of debiasing effectiveness.



Conclusion

- Adversarial networks are a useful tool to mitigate bias while maintaining predictive accuracy and power

- There is a need for more robust, diverse, and labeled image datasets to train and evaluate fairness

Future Directions & Broader Impact

Diverse Datasets & Evaluation

- Need for larger, more diverse ASL datasets with various demographics.
- Utilize resources like Meta's FACET dataset for comprehensive fairness evaluation.

Methodological Exploration

- Adapt adversarial debiasing for continuous sign language recognition from video (e.g., video transformers, 3D CNNs).
- Investigate theoretical properties of gradient reversal (convergence, hyperparameter selection).

Broader Applications

- Apply methodology to other critical computer vision tasks.
- Examples: Medical image analysis, facial recognition, emotion recognition, gesture-based interfaces.