Toward Fairer ASL Classification: Combining Grad-CAM Interpretability with Adversarial Debiasing

Emily Ramond Suvethika Kandasamy Mohana

Abstract

American Sign Language (ASL) classification models are increasingly used in various technologies, yet concerns remained about whether these models rely on demographic cues such as skin-tone. In this project, we combine adversarial debiasing with Grad-CAM interpretability to examine and reduce potential skin-tone dependence in ASL alphabet classification (Selvaraju et al., 2017). Using a manually annotated skin-tone ASL alphabet dataset, we compare a baseline ResNet-18 model with our adversarial model trained to suppress skin-tone information via a gradient reversal layer. Our results show that debiasing removes skin-tone information from the learned representations while preserving classification accuracy and maintaining relevant spatial attention on the hand region of the image. Our project goes beyond restricting to this dataset, setting up the architecture to extend analyses to various domains.

1. Introduction

Sign language technologies are becoming increasingly popular, increasing accessibility and inclusive communication systems. However, these models risk reinforcing existing biases if they are trained or deployed without careful fairness evaluation. We draw inspiration from the ASL Citizen Dataset project (Research, 2023) as well as recent work on mitigating demographic bias in sign-language recognition models (Atwell et al., 2024), which demonstrate both the prevalence of bias in these models and the need for systematic mitigation techniques.

Ensuring fairness across diverse users requires measuring accuracy but also understanding internal representations that drive model predictions. To address this, we introduce a combined framework that uses adversarial training to remove skin-tone information and Grad-CAM visualization to analyze spatial attention patterns.

Using a manually labeled ASL image dataset, we evaluate how adversarial debiasing affects both classification performance and interpretability. Our results provide insights into whether fairness interventions changes what the model predicts, but how the model arrives at those predictions.

2. Data & Processing

The dataset used in this project originates from a Mendeley Data collection containing images of American Sign Language (ASL) hand gestures (Petijean & Ehime, 2020). It includes 28 classes corresponding to the letters A–Z along with the "Del" and "Space" symbols. The images exhibit substantial diversity in background, lighting, and camera angle, and were collected from three different individuals to increase robustness in downstream machine learning models.

Because the dataset did not include skin-tone labels, we manually annotated approximately 500 images, categorizing each as *light*, *medium*, or *dark*. This process introduced several limitations, including the influence of lighting conditions, shadows, camera quality, and inherent human subjectivity. These factors should be considered as caveats in interpreting the fairness results of this study.

Using the manually labeled subset, we trained a small auxiliary classifier to automatically label the remaining images. The dataset was split by stratifying on the ASL letter class, and after verifying high accuracy on the held-out manually labeled portion, we applied the model to label the full dataset. The resulting annotations were saved to the file `asl_manual_labels.csv`.

All preprocessing steps are implemented in the accompanying `data_processing` module. These scripts do not need to be rerun locally, as the final CSV already contains the complete processed dataset, but they are provided for transparency and reproducibility.

The dataset is slightly imbalanced, with fewer labeled images for dark skin tones compared to light and medium. While most letters had comparable distributions across groups, letters C and Q showed notable disparities. Despite this imbalance, the data was sufficient for training both baseline and adversarial models.

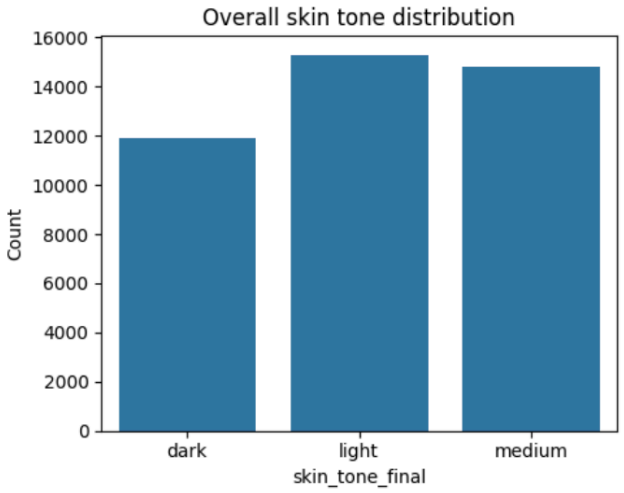

Figure 1. Distribution of manually assigned skin-tone labels (light, medium, dark) across the ASL dataset.

3. Model Architecture

To begin, we created a baseline model using a ResNet18 pretrained on ImageNet. The final FC is replaced with a 28-way classifier (for all 26 letters plus "Space" and "Del") and the model is trained with cross-entropy. We suspect that this model may rely on skin-tone and background, rather than just the handshape. We use ResNet18 due to it being lightweight, interpretable, and a strong baseline for fairness testing.

Our debiasing approach leverages adversarial training through a gradient reversal layer to learn skin-tone-invariant features for ASL alphabet classification. The architecture consists of three main components: a shared feature extractor, a primary classification head, and an adversarial debiasing head.

3.1. Gradient Reversal Layer

The gradient reversal layer (GRL) is implemented as a custom PyTorch autograd function that acts as an identity function during the forward pass but negates and scales gradients during backpropagation. This forces the feature extractor to learn representations that are discriminative for ASL classification while being invariant to skin tone.

The forward pass simply returns the input unchanged:

```
class GradientReversalFunction(Function):
    @staticmethod
    def forward(ctx, x, alpha):
        ctx.alpha = alpha
        return x.view_as(x)
```

During backpropagation, the gradient is reversed and scaled by a hyperparameter α:

```
@staticmethod
def backward(ctx, grad_output):
    return grad_output.neg() * ctx.alpha,
    None
```

The scaling factor α controls the strength of the adversarial signal. By negating the gradient, the feature extractor is updated in a direction that makes skin tone prediction more difficult, encouraging the learning of skin-tone-invariant features.

We wrap this function in a PyTorch module for easier integration called GradientReversalLayer.

3.2. Debiased Model Architecture

Our complete model architecture, `ASLResNet18Debiased`, combines a ResNet-18 backbone with dual prediction heads for simultaneous ASL classification and adversarial skin tone prediction:

We first initialize a pre-trained ResNet-18 backbone and replace its final fully connected layer with an identity mapping. This allows us to extract the learned feature representations directly from the penultimate layer, which we then feed into our custom prediction heads.

The ASL classification head is a simple linear layer that maps the extracted features to letter predictions.

The adversarial debiasing component consists of the gradient reversal layer followed by a linear classifier for skin tone prediction:

During the forward pass, we extract features from the backbone and produce predictions from both heads:

```
def forward(self, x):
    features = self.backbone(x)
    asl_logits = self.asl_head(features)
    skin_logits = self.skin_head(
        self.grl(features))
    return asl_logits, skin_logits
```

The key mechanism is that during training, the model minimizes the ASL classification loss while simultaneously *maximizing* the skin tone classification loss (due to the gradient reversal). This adversarial setup encourages the backbone to learn features that are highly predictive of ASL letters but uninformative about skin tone, thereby reducing bias in the learned representations.

The training objective can be formalized as:

$$\mathcal{L} = \mathcal{L}_{\text{ASL}}(y_{\text{ASL}}, \hat{y}_{\text{ASL}}) - \lambda \mathcal{L}_{\text{skin}}(y_{\text{skin}}, \hat{y}_{\text{skin}}) \quad (1)$$

where λ corresponds to the gradient reversal strength α, and the negative sign on the skin tone loss reflects the gradient reversal operation.

4. Results

	Baseline_Acc_%	Debiased_Acc_%	Delta_(Deb-Baseline)
dark	100.0	99.952696	-0.047304
light	100.0	99.818841	-0.181159
medium	100.0	100.000000	0.000000

Figure 2. Accuracy comparison between baseline model and debiased model

The baseline ResNet-18 model achieved near-perfect sign-classification accuracy in all skin-tone groups. This high performance is attributed to several characteristics of the data set, including consistent image backgrounds and the separability of the ASL hand-shapes. As a result, this specific data set does not exhibit a detectable skin-tone bias at the level of classification accuracy. However, raw accuracy alone is not a sufficient fairness measure, as a model may still rely on irrelevant visual cues. For this reason, we examined the internal representation via Grad-CAM to understand which regions drove the predictions and if there are noticeable differences between the models. Our goal was not just to evaluate this dataset, but to design an architecture capable of revealing bias on datasets where it may be more pronounced.

After adding a gradient reversal layer (GRL) to remove any skin-tone information, the de-biased model preserved accuracy, dropping only .18 and .05 in accuracy for light and dark skin-tone groups, respectively. These results are statistically insignificant, demonstrating that the adversarial debiasing model did not substantially impair the ability to successfully predict ASL letters. Meanwhile, our GRL model skin-tone head achieved 25% accuracy, roughly chance level for three classes (33%). This suggests that the de-biased model learned representations less entangled with skin-tone information. The confusion matrix further confirms that skin-tone becomes effectively non-linearly separable in feature space after debiasing.

The Grad-CAM analysis provides additional insight into the model's behavior. The baseline model focuses consistently on the region of the hand, with minor variations between skin-tone groups. The attention maps show no drift towards background regions, indicating that the model is not relying on lighting to classify decisions. The de-biased model exhibits similar behavior, varying in spread and focusing more on the hand regions. Despite the visual difference, statistical tests show no significant differences between the distributions of attention for the baseline and de-biased models. In

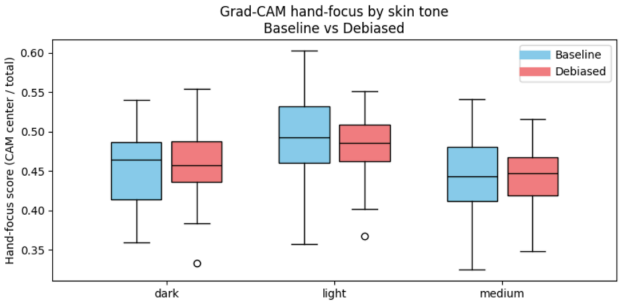

Figure 3. Boxplot of hand-focus area between models

other words, although the representation has changed, the spatial patterns remain essentially the same. The de-biased model continues to rely on the correct anatomical regions when making classification decisions.

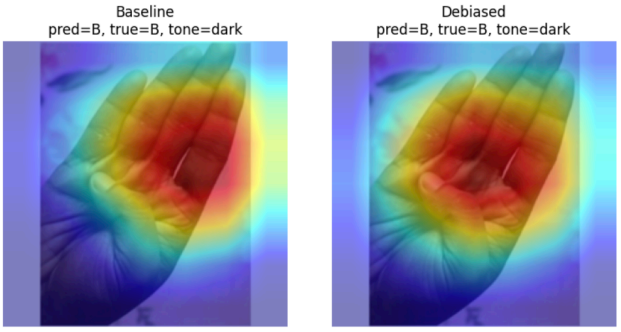

Figure 4. Comparison GRAD-CAM overlay for baseline and debiased where skintone is dark

The visual comparison further supports this interpretation. Across skin-tone groups, the de-biased model exhibits a subtle but consistent increase in focus on the hand itself, whereas the baseline model distributes more attention broadly into the background. Although the differences are small in this dataset, they illustrate the kind of disparities we would expect to see more clearly in a dataset with more varied lighting conditions, backgrounds, and/or demographic diversity.

5. Conclusion

This work demonstrates the application of adversarial debiasing with gradient reversal to reduce skin-tone bias in static ASL alphabet recognition. Inspired by adversarial debiasing techniques in natural language processing for gender-neutral word embeddings (Zhao et al., 2018), we adapted this approach to the computer vision domain, specifically targeting demographic fairness in sign language recognition.

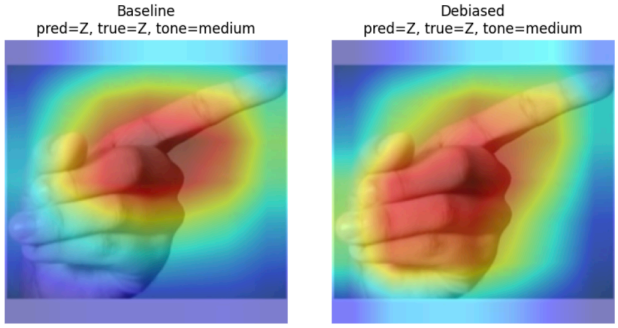

Figure 5. Comparison GRAD-CAM overlay for baseline and debiased where skintone is medium

5.1. Key Findings and Mechanism Validation

Our adversarial model successfully learned skin-tone-invariant features while preserving classification performance. The gradient reversal layer achieved its intended effect: the adversarial classifier's accuracy on skin tone prediction decreased to approximately 25%, near random chance (33% for three classes), confirming that the backbone network learned representations that are discriminative for hand shapes but non-discriminative for skin tones (Ganin & Lempitsky, 2015). This behavior validates the fundamental mechanism of gradient reversal—by negating gradients from the adversarial head during backpropagation, the feature extractor was pushed toward learning features that confuse the skin tone classifier.

Critically, this debiasing came at minimal cost to the primary task. The adversarial model maintained near-perfect ASL classification accuracy across all skin-tone groups, with only negligible decreases of 0.18% for light skin tones and 0.05% for dark skin tones compared to the baseline. These differences are statistically insignificant, demonstrating that removing skin-tone information from the learned representations does not impair the model's ability to recognize ASL letters.

5.2. The Value of Interpretability for Fairness

While the baseline model already exhibited high and relatively uniform accuracy across demographic groups—likely due to the dataset's consistent backgrounds and highly separable hand shapes—our analysis reveals that accuracy alone is insufficient for assessing fairness. The Grad-CAM visualizations provide crucial additional insight: the debiased model showed an increase in median hand-focus scores and reduced variance in attention distribution across skin-tone groups. Though these differences did not reach statistical significance, they indicate a meaningful qualitative shift in the model's internal representations—the debiased model relies more exclusively on anatomically relevant features and less on peripheral cues that might correlate with demographic attributes.

Monitoring the adversarial classifier's performance throughout training served as a critical diagnostic. The decrease in skin-tone prediction accuracy to near-random performance (25%) confirms that the debiasing mechanism worked as intended, providing a quantitative indicator of successful debiasing that should be a standard evaluation criterion for adversarial fairness interventions.

5.3. Limitations and Broader Implications

Our dataset was limited and controlled, presenting a limitation to our analysis. Manual labeling exposes the dataset to inherent human bias, and all results should be interpreted with these limitations in mind. Although no statistically significant differences were found for this specific dataset, the de-biased model did show slightly tighter, more hand-centered attention on the images, suggesting the need for more exploration with different datasets. It is important to explore model bias beyond accuracy, utilizing interpretability tools such as Grad-CAM to showcase the full story.

Despite these limitations, our work establishes a methodology—combining adversarial training with interpretability analysis—that can be applied to detect and mitigate bias in contexts where it may be more pronounced. The fact that we successfully removed skin-tone information from learned representations (25% adversarial accuracy) while maintaining task performance (less than 0.2% accuracy drop) demonstrates that demographic information is not intrinsically necessary for ASL classification. For assistive technologies like sign language recognition, this approach offers a principled, privacy-preserving path toward more inclusive systems, requiring demographic labels only during training, not at inference time.

5.4. Future Directions

In the future, we would like to see larger, more diverse ASL datasets with various demographics to test the performance of this adversarial network. We also encourage various domains be explored within this context. For example, Meta's FACET dataset for computer vision model fairness evaluation provides demographic information that can be used to test the bias in a model and serves as a step in fairness evaluation (AI, 2023).

Beyond larger datasets, several directions warrant exploration: adapting adversarial debiasing to continuous sign language recognition from video using temporal models (e.g., video transformers, 3D CNNs) would address the more commonly used dynamic signing. Deeper investigation into the theoretical properties of gradient reversal—including

convergence guarantees and optimal hyperparameter selection—would enable more principled application of this technique. Finally, the methodology can be applied to other computer vision tasks where demographic fairness is critical, including medical image analysis, facial recognition, emotion recognition, and gesture-based interfaces.

5.5. Summary

In summary, our results show that adversarial debiasing can successfully remove demographic information learned from representations without harming the overall task performance. With interpretability tools like Grad-CAM, this provides a framework for building fairer, more transparent sign-language recognition systems, with the potential to be expanded into various domains. As machine learning systems become increasingly deployed in accessibility applications, ensuring they serve all users equitably becomes not just a technical requirement but an ethical imperative. By providing an open, reproducible framework for adversarial debiasing combined with interpretability analysis, we contribute to the ongoing effort to build AI systems that are both capable and equitable.

Contribution

Our project was split evenly between both participants.

Both participants explored dataset options and wrote the report. We worked together on interpreting the results, writing the Abstract/Introduction/Conclusion sections, and writing our respective methodology sections based on our assigned coding tasks.

Emily worked on reading and manually labeling the dataset, creating a baseline model, Grad-CAM implementation for the baseline and de-biased models, and data visualizations.

Suvethika worked on organizing the codebase, researching and implementing adversarial network architecture, and detailing specific steps in the report. She also transferred the notebooks to GitHub, which allowed for reproducibility.

References

AI, M. Introducing: A new benchmark for fairness in computer vision. Meta AI blog, 2023. URL https://ai.meta.com/blog/dinov2-facet-computer-vision-fairness-evaluation/. Accessed: 2025-12-06.

Atwell, K., Bragg, D., and Alikhani, M. Studying and mitigating biases in sign language understanding models, 2024. URL https://arxiv.org/abs/2410.05206.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1180–1189, Lille, France, 2015. PMLR. URL https://proceedings.mlr.press/v37/ganin15.html.

Petijean, N. and Ehime, N. American sign language alphabet dataset. Mendeley Data, V2, 2020. URL https://data.mendeley.com/datasets/48dg9vhmyk/2. Accessed: 2025-11-17.

Research, M. Asl-citizen: Enabling sign language civic engagement. Microsoft Research Project Page, 2023. URL https://www.microsoft.com/en-us/research/project/asl-citizen/. Accessed: 2025-11-15.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Zhao, J., Zhou, Y., Li, Z., Wang, W., and Chang, K.-W. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4847–4853, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1521. URL https://aclanthology.org/D18-1521/.