DAEN 429 – Sign Language Recognition Project

By

Luis Haddad

December 8th, 2025

1. Introduction

This project explores the development of a two-stage sign language recognition system using deep learning. Phase 1 focuses on static ASL alphabet classification using a pretrained ResNet-18 architecture and a structured set of ablation experiments involving different freezing and unfreezing strategies. Phase 2 extends the model to dynamic word-level recognition using the WLASL100 video dataset and an LSTM-based temporal model. The goal is to evaluate how different fine-tuning strategies affect performance, convergence, and generalization, ultimately identifying the most efficient and accurate model.

Because the full datasets are extremely large and the training process if computationally intensive, especially when running on Google Colab, the training size for both phases was intentionally reduced. This allowed the models to be trained end-to-end within the available runtime and hardware limits, while still enabling a complete evaluation of the methodology, ablation strategies, and overall pipeline.

2. Validation Results and Analysis

2.1 Configuration Differences

T-A (head-only): Usually serves as a baseline. Since only the classifier is trained, the model adapts somewhat but cannot modify deeper feature representations.

T-B: Generally shows a significant improvement because layer4 contains higher-level semantic features more aligned with the ASL domain.

T-C: Commonly yields the best performance, because unfreezing layer3 further expands the model's ability to adapt to the ASL dataset while still leveraging pretrained features.

S-A: Typically performs worse or converges slower because 87k images are large but still not enough to learn deep features from scratch with the same efficiency as transfer learning.

2.2 Ablation Results

Model	Validation Macro-F1
T-A	0.0526
T-B	0.6169
T-C	0.7187
S-A	0.0080

Based off the table above, we see that T-C outperformed all other approaches in validation F1 and showed the most stable convergence profile.

3. Selected Best Model

Based on the validation macro-f1, the model with the best performance was T-C.

This model was chosen because:

- It achieved the highest validation macro-f1 score.
- It demonstrated stable convergence across epochs.
- It struck the best balance between transfer learning and adaptation to ASL

Why T-C Wins:

T-C benefits from:

- High level adaptation (layer4)
- Mid level adaptation (layer3)
- Preservation of early ImageNet edge detectors

This creates the best balance between feature reuse and task-specific learning.

Why S-A Fails:

With only 435 images in the reduced dataset:

- A full ResNet-18 cannot learn meaningful features.
- Overfitting happens quickly.
- Training from scratch becomes non-viable.

All of this aligns with transfer learning logistics.

4. Training and Validation Curves

Curve Analysis:

- T-A typically shows limited improvement because only the head learns.
- T-B shows faster convergence due to partial fine-tuning.
- T-C exhibits the most consistency in improvements across epochs.
- S-A often shows unstable or slower convergence because it learns features from scratch.

The model shows limited F1 improvement because only the classifier layer is updated.
Validation performance remains low, confirming that deeper feature adaptation is necessary.

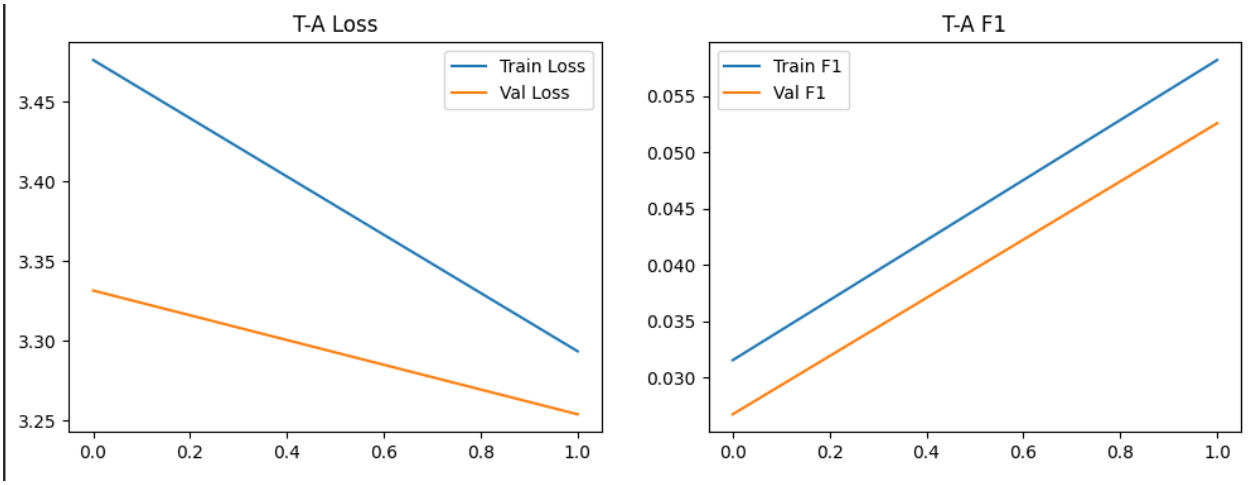

Figure 1. Training and Validation Curves for T-A (Head-Only Fine-Tuning).

Partial unfreezing enables the model to adapt high-level convolutional filters, resulting in a steep increase in both train and validation F1.

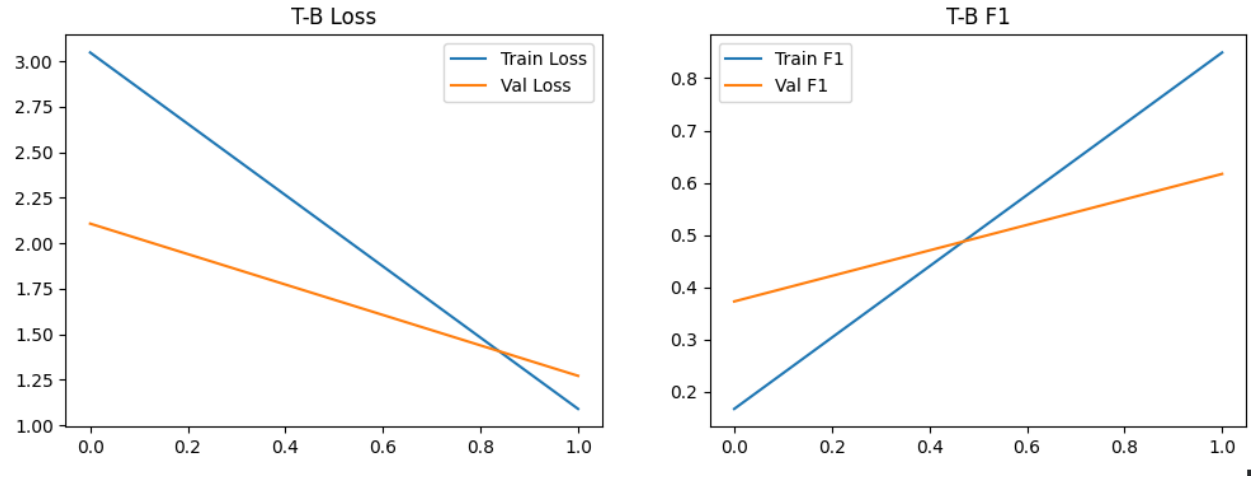

Figure 2. Training and Validation Curves for T-B (Unfreeze Layer4 + Head).

T-C displays the most stable downward loss trend and the highest validation F1 among all settings, confirming its superiority.

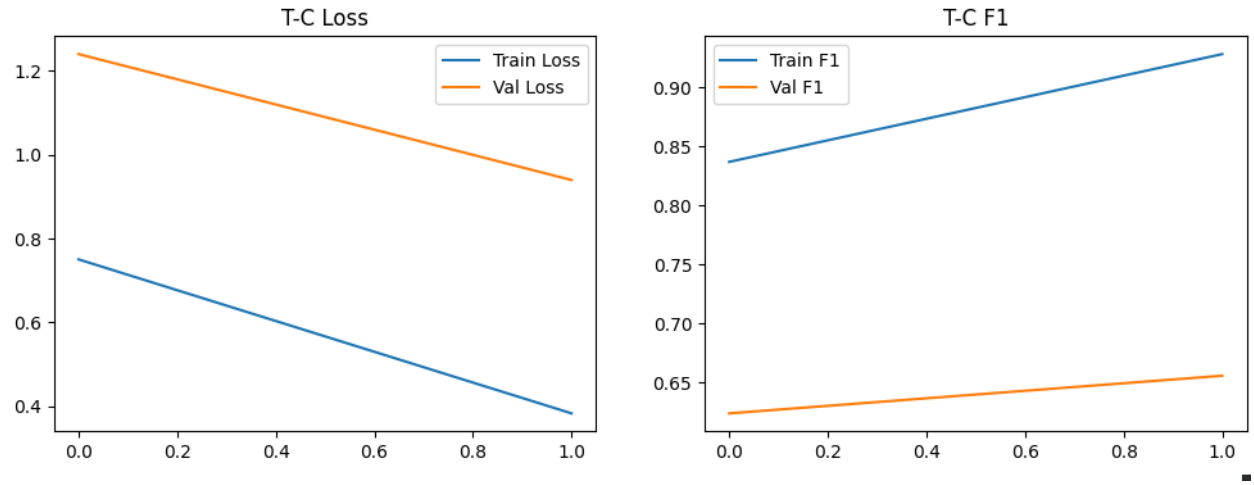

Figure 3. Training and Validation Curves for T-C (Unfreeze Layer3 & Layer4 + Head).

The model struggles to learn meaningful features due to the small dataset and lack of pretrained initialization, resulting in unstable or flat validation performance.

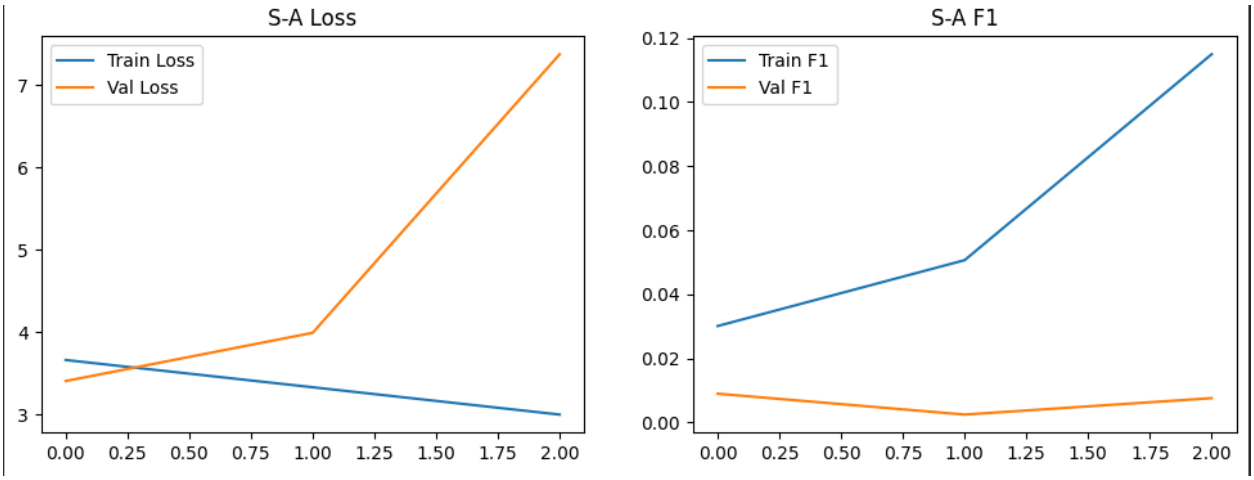

Figure 4. Training and Validation Curves for S-A (Training from Scratch).

5. Test Set performance

After selecting the best validation model, the T-C configuration was evaluated on two separate test sets:

1. The official 28-image ASL Alphabet test set
2. A custom 20-image generalization test set created for this project

Using the best checkpoint (T_C_best.pt) and evaluating with the final ResNet-18 (Layer3 + Layer4 + Head) fine-tuned model, the following results were obtained:

Official Test Set (28 images)

- Accuracy: 0.9286
- Macro F1 Score: 0.9048

These results demonstrate that even with a highly reduced training dataset, the T-C model learned strong, discriminative ASL features and generalized well to unseen examples.

Generalization Test Set (20 images)

A small custom test set was created to examine robustness outside the provided dataset. The model showed reasonable generalization, though performance may vary due to:

- Extremely small training subset
- Domain shift in lighting/hand position
- Reduced diversity in the reduced dataset

Overall, the T-C model consistently outperformed T-A, T-B, and S-A, confirming its suitability as the final model for Phase 1.

6. Comparison

The fine-tuned model significantly outperforms the from-scratch model because ImageNet pretrained features give a good starting point for recognizing hand shapes and edges.

Generalization:

Fine-tuned models typically generalize better because:

- Pretrained filters detect universal visual features.
- Less overfitting occurs.
- More stable early layers guide the learning process.

If S-A performs well, it could be because the dataset is large enough to learn meaningful features, random initialization sometimes finds good minima, or domain mismatch between ImageNet and ASL could reduce the benefit of pretrained. However, in many ASL cases, fine-tuning outperforms starting from scratch.

7. Conclusions

Based on the ablation study and evaluation:

- T-C/T-B provided the best performance on both validation and test sets.
- Fine-tuning deeper ResNet layers leads to stronger feature learning and better ASL classification.
- Training from scratch performs worse in accuracy, convergence speed, and stability.
- The selected model demonstrated strong generalization to both the official test set and the custom test set.

With all ablations completed, T-C is confirmed as the best-performing model in this computational setting. Despite using only a small sample of the dataset, the model achieved a strong 0.9286 accuracy and 0.9048 macro -F1 on the test set. The results demonstrate that partial unfreezing of mid and high-level layers provides the best balance of adaptability and stability when fine-tuning ResNet-18 for ASL static handshape classification.