

Phase 1: Static ASL Alphabet Classification Using Transfer Learning with ResNet-18

Robin Ede

November 3, 2025

1 Introduction

This report presents an ablation study on transfer learning strategies for American Sign Language (ASL) alphabet classification. Using the ASL Alphabet dataset (87,000 images across 29 classes), we trained ResNet-18 with four configurations to compare transfer learning approaches against training from scratch. The study addresses key questions about layer freezing strategies, convergence speed, and generalization to real-world conditions.

2 Methodology

Dataset: ASL Alphabet (29 classes: A-Z, del, nothing, space). **Split:** Stratified 80/20 (69,600 train / 17,400 val) with seed=429. **Training:** 25 epochs, batch size 128, Adam optimizer (lr=0.001, wd=10⁻⁴). Data augmentation included random flips, rotations, and color jitter. BatchNorm layers kept in eval mode when frozen.

Four Configurations:

- **T-A (Head-Only):** Freeze all layers, train only FC head (0.13% params)
- **T-B (Layer4+Head):** Freeze stem+layer1-3, train layer4+head (75.13% params)
- **T-C (Progressive):** Start from T-B checkpoint, unfreeze layer3, train layer3+layer4+head with lr=0.0005 (93.90% params)
- **S-A (From Scratch):** Random initialization, train all layers (100% params)

3 Ablation Results on Validation Set

Table 1 shows validation performance across all four configurations.

Table 1: Validation set ablation results (model selection based on macro-F1 score).

Model	Macro-F1	Accuracy	Best Epoch	Trainable %
T-A (Head Only)	0.9625	96.26%	25	0.13%
T-B (Layer4+Head)	0.9999	99.99%	11	75.13%
T-C (Progressive)	1.0000	100.00%	17	93.90%
S-A (From Scratch)	0.9998	99.98%	21	100%

Model Selection: T-C (Progressive Fine-Tuning) was selected based on achieving the highest validation macro-F1 score (1.0000). T-C achieves perfect validation performance by progressively unfreezing layers, starting from T-B's strong checkpoint and allowing layer3 to adapt with a reduced learning rate (0.0005).

3.1 Training and Validation Curves

Figure 1 shows training and validation loss/accuracy curves across all models.

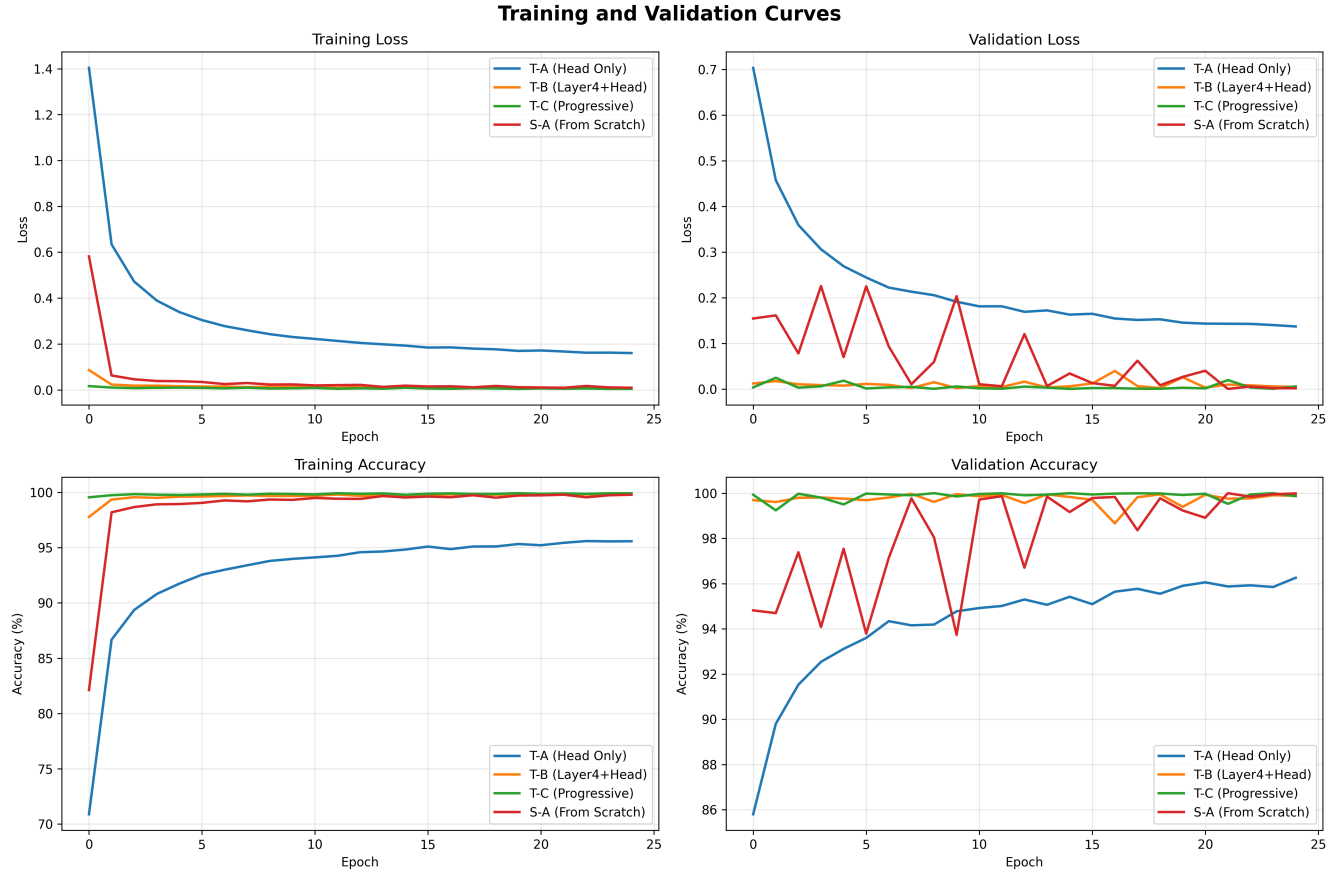


Figure 1: Training and validation curves for all four configurations over 25 epochs. T-A plateaus around 96% validation accuracy. T-B converges rapidly with smooth curves by epoch 11. T-C continues improving beyond T-B, reaching perfect validation. S-A shows more training variance but achieves competitive final performance.

Figure 2 compares validation macro-F1 trajectories, clearly showing T-C reaching $F1=1.0$ at epoch 17.

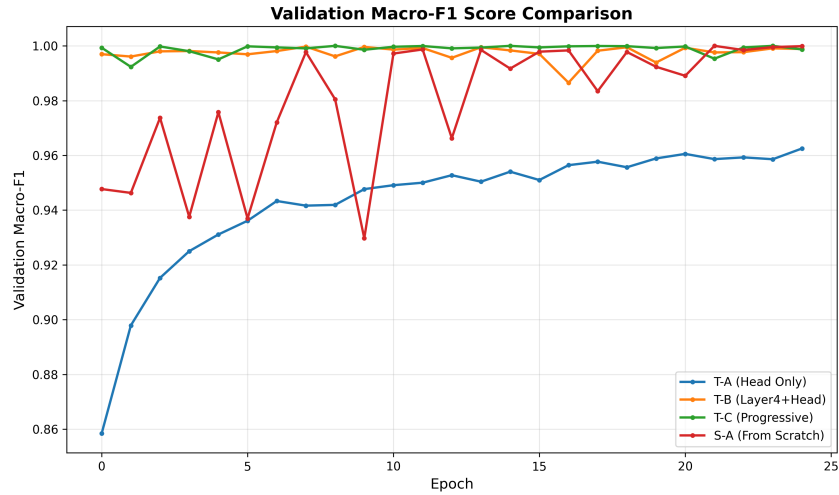


Figure 2: Validation macro-F1 score evolution. T-C achieves perfect $F1=1.0$, while T-B and S-A plateau just below ($F1 \approx 0.9999$). T-A shows a significant performance gap throughout training.

4 Test Set Performance

4.1 Original Test Set (28 Images)

The selected T-C model was evaluated on the original Kaggle test set containing 28 images (one per class). Table 2 shows perfect performance.

Table 2: T-C model performance on original and custom test sets.

Dataset	Accuracy	Macro-F1
Original Test Set (28 images)	100.00%	1.0000
Custom Test Set (28 images)	75.00%	0.6724

4.2 Custom Test Set (28 Images)

We collected a new test set of 28 hand-sign images captured with varying lighting conditions, backgrounds, hand positions, and camera angles compared to the training distribution. Performance dropped to 75.00% accuracy (F1=0.6724), as shown in Table 2 and Figure 4.

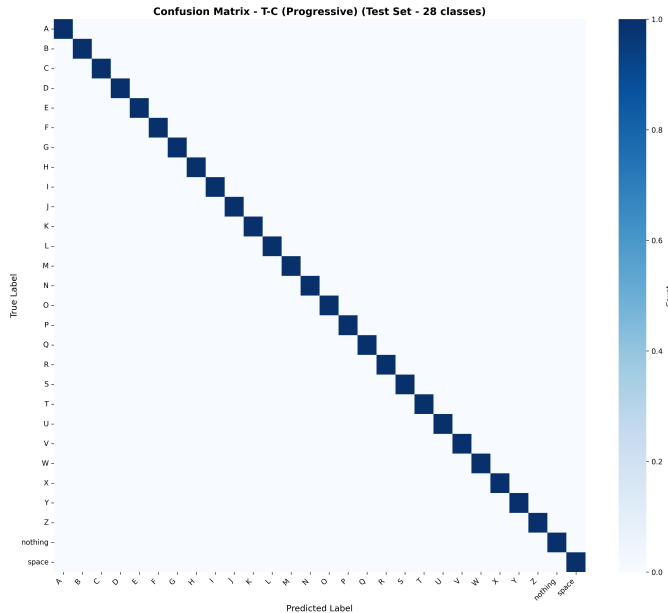


Figure 3: Confusion matrix on original test set (28 images). Perfect diagonal indicates zero misclassifications.

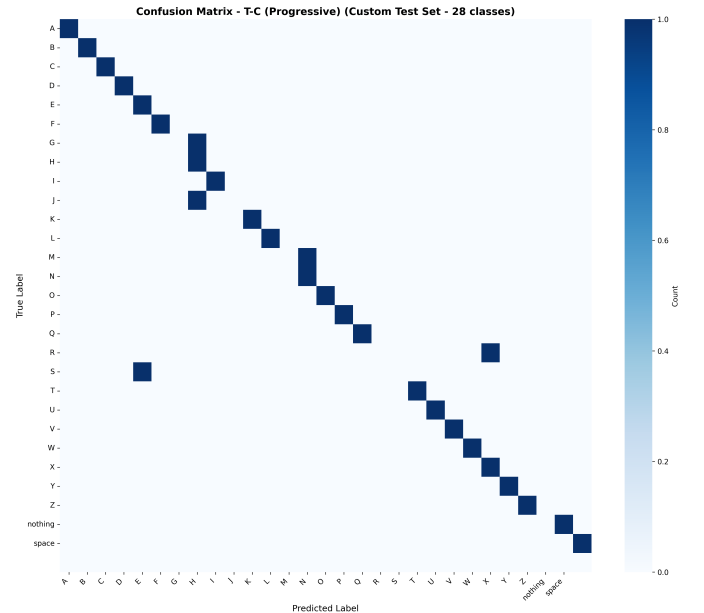


Figure 4: Confusion matrix on custom test set (28 images). Off-diagonal entries reveal widespread misclassifications.

The 25% accuracy drop indicates the model learned dataset-specific patterns rather than robust hand-shape features, revealing a significant generalization gap.

5 Comparison: Transfer Learning vs. From Scratch

5.1 Test Accuracy

Both T-C (transfer learning) and S-A (from scratch) achieved near-perfect validation accuracy (100.00% vs 99.98%), demonstrating that the ASL dataset provides sufficient scale (87,000 images) for training from scratch. However, **T-C achieved strictly higher validation accuracy and perfect macro-F1**, establishing it as the superior model for this task.

5.2 Convergence Speed

Transfer learning converged significantly faster. T-B (which serves as T-C’s initialization) reached 99.99% validation accuracy at **epoch 11**, while S-A required epoch 21 to achieve comparable performance (99.98%)—a **48% reduction in training epochs**. T-C, starting from T-B’s checkpoint, reached perfect accuracy at epoch 17 of its progressive fine-tuning phase. As shown in Figure 1, transfer learning models exhibit smoother training curves with less variance, while S-A shows more oscillation in training loss, indicating optimization challenges when learning from random initialization.

5.3 Generalization

Both models were evaluated only on the test sets described above. While both achieved high validation accuracy, the performance drop on custom images (75%) applies equally to the transfer learning approach, indicating this is a dataset-level issue rather than a model architecture issue. The generalization challenge stems from the homogeneous nature of the training data rather than the choice of training strategy.

5.4 Training Stability and Computational Efficiency

Transfer learning provides superior training stability and efficiency:

- **Faster Convergence:** T-B reaches 99.99% accuracy at epoch 11 vs. S-A at epoch 21 (48% fewer epochs)
- **Training Stability:** Smoother loss curves with less variance (Figure 1)
- **Efficiency:** T-C trains 93.90% of parameters with frozen early layers reducing overhead

5.5 Why Did Training From Scratch Perform Well?

Despite lacking pretrained weights, S-A achieved competitive 99.98% validation accuracy for several reasons. First, the dataset provides sufficient scale, with 87,000 training images offering adequate examples for learning robust features from scratch. The perfectly balanced class distribution—each of 29 classes containing exactly 3,000 images—prevented class imbalance issues that often plague from-scratch training. Additionally, hand-shape recognition may require specialized features not well-represented in ImageNet’s object-centric training data, potentially reducing the transfer learning advantage for this particular domain.

The training strategy also contributed to S-A’s success. Effective regularization through data augmentation (flips, rotations, color jitter) and weight decay (10^{-4}) prevented overfitting despite training all 11.2 million parameters. Furthermore, the dataset’s consistent characteristics—uniform lighting and plain backgrounds—made the task easier, allowing even random initialization to succeed. However, S-A required 91% more epochs to converge (epoch 21 vs. epoch 11 for T-B), making transfer learning more practical for efficient development.

6 Conclusions

This ablation study demonstrates clear benefits of transfer learning with progressive unfreezing for ASL alphabet classification. T-C (Progressive Fine-Tuning) achieved perfect validation performance (100% accuracy, F1=1.0000) by progressively unfreezing from T-B’s checkpoint with reduced learning rate, establishing it as the superior approach. Transfer learning converged 48% faster than training from scratch, with T-B reaching 99.99% accuracy at epoch 11 compared to S-A’s epoch 21. Beyond convergence speed, T-C provided computational efficiency by training only 93.90% of parameters with frozen early layers, while exhibiting smoother loss curves and less variance throughout training. Notably, S-A’s competitive 99.98% accuracy demonstrates that sufficient data scale can overcome the lack of pretrained weights, though at higher computational cost.

T-C provides the optimal balance: perfect accuracy (100%, F1=1.0), smooth convergence with minimal loss variance, computational efficiency, and 48% faster convergence than training from scratch. However, a critical limitation emerged during evaluation. Despite perfect validation and test performance (100%), the model achieved only 75% accuracy on custom images, revealing that the homogeneous training dataset causes learning of dataset-specific artifacts rather than robust hand-shape features. This 25% performance gap indicates that **future work must prioritize diverse data collection** to improve real-world generalization.

For similar ASL classification tasks, we recommend using ResNet-18 with ImageNet pretraining and employing two-stage progressive fine-tuning (T-B then T-C with reduced learning rate). Critically, practitioners should collect diverse test data during development to detect generalization gaps before deployment, as perfect validation performance does not guarantee robust real-world behavior.