Traffic Sign Classification: A Two-Phase Investigation of Resolution, Architecture, and Dataset Limitations

NOVEMBER 7, 2025

AHMED IBRAHIM MUHAMMED
Cairo, Egypt

Executive Summary

This empirical study presents a two-phase systematic investigation into the fundamental constraints of traffic sign classification for real-world deployment. **Phase 1** explored the relationship between input resolution and model generalisation across seven experiments (64×64 to 128×128).

Phase 2 investigated whether architectural sophistication (hybrid CNN-Transformer) could overcome limitations identified in Phase 1.

Central Thesis: Dataset characteristics fundamentally constrain real-world deployment success more than resolution choices or architectural sophistication. Idealised, region-specific datasets create models optimised for benchmarks rather than production, and neither resolution tuning nor architectural innovation can substitute for representative training data.

Key Findings:

- Phase 1: Lower resolutions (64×64, 80×80) unexpectedly outperformed higher resolutions (128×128) on real-world images, achieving 98.7% test accuracy but 9.1% real-world success rate.
- Phase 2: Adding transformer attention to CNN improved real-world success marginally (9.1% → 16.7%) while maintaining 98.58% test accuracy.
- **Conclusion:** Both phases validate that benchmark performance (>98%) provides false confidence when training distribution doesn't match deployment conditions.

---

1. Introduction

1.1 The Production Deployment Challenge

Modern computer vision models routinely achieve >95% accuracy on academic benchmarks, creating an illusion of production readiness. However, deployment frequently reveals a critical gap between controlled evaluation and real-world performance.

This empirical study investigates two common strategies for bridging this gap:

1. *Resolution optimization***:** Finding the "right" input size for generalization.
2. *Architectural sophistication:* Adding advanced components like transformers.

Through systematic experimentation, we demonstrate that **neither strategy succeeds when the fundamental constraint is training data that doesn't represent deployment conditions**.

1.2 Empirical study Questions

Phase 1 **(Resolution Investigation):**

- How does input resolution affect the trade-off between test accuracy and real-world generalisation?
- Can we find an optimal resolution that bridges the domain gap?

Phase 2 **(Architecture Investigation):**

- Can transformer-based global attention compensate for limitations observed in Phase 1?
- Does architectural sophistication improve generalisation when resolution optimisation failed?

Overarching Question: Can technical optimisation (resolution, architecture) overcome training on idealised datasets, or do dataset characteristics fundamentally constrain deployment success?

1.3 Central Hypothesis

Dataset characteristics matter more than technical optimisation. Models trained on GTSRB's idealised German traffic signs will fail on diverse real-world images regardless of resolution choices or architectural sophistication, because the training distribution doesn't match deployment conditions.

---

2. Background and Motivation

2.1 GTSRB Dataset Characteristics

The German Traffic Sign Recognition Benchmark exemplifies an idealised vision problem:

Strengths for empirical study:

- Big enough dataset: 39,209 training images, 12,630 test images
- 43 well-defined traffic sign classes
- Consistent image quality and preprocessing
- Clean, pre-cropped signs centred in frame
- Reproducible benchmark for algorithm comparison

Limitations for deployment:

- Regional specificity**:** German traffic signs only (single country's design standards)
- Idealised conditions: Consistent lighting, clean backgrounds, no weathering
- Artificial preprocessing**:** Signs pre-cropped and isolated from natural context
- Limited diversity**:** Controlled collection conditions, minimal environmental variation

2.2 The Domain Gap Problem

Training distribution (GTSRB):

- Pre-cropped, centred traffic signs
- Consistent backgrounds (sky, trees)
- Clean, well-lit images
- German design standards

Deployment distribution (real-world):

- Signs embedded in complex street scenes
- Variable lighting, weather, and viewing angles
- Different countries' sign designs
- Occlusions, weathering, vandalism

Empirical study question: Can model optimisation overcome this fundamental distribution mismatch?

---

3. Phase 1: Resolution vs Generalisation Trade-offs

3.1 Experimental Design

Seven systematic experiments testing resolutions: 64×64, 80×80, 90×90, 128×128

*Controlled variables***:**

- Dataset: GTSRB (85% train, 15% validation)
- Architecture: CNN with 4 convolutional blocks (32→64→128→256)
- Training: Adam optimizer, early stopping, mixed precision (FP16)

- Hardware: NVIDIA RTX 2060 (6GB VRAM)

Varied factors across trials:

- Input resolution (64×64 to 128×128)
- Augmentation strategies (none, Keras, Albumentations with varying intensity)
- Regularisation strength (dropout rates)

3.2 Phase 1 Results Summary

Trial progression:

Trial	Resolution	Augmentation	Test Acc	GTSRB Test Image	Real-World
1	64×64	None	94.0%	99.6%	Failed
2	80×80	Keras	96.8%	99.8%	8.7%
3	80×80	Heavy Alb.	91.0%	89.5%	Failed
4	80×80	Moderate Alb.	**98.7%**	99.0%	13.8% (89.6% @ 128px)
5	128×128	Moderate Alb.	97.6%	85.0%	Failed
6	128×128	Heavy Alb.	91.0%	47.4%	Failed
7	90×90	Moderate Alb.	98.7%	89.6%	Failed

3.3 Phase 1 Key Findings

1. *Resolution Paradox*: Performance on GTSRB test images showed inverse correlation with resolution:

- 64×64: 99.6% (best)
- 80×80: 99.0%
- 90×90: 89.6%
- 128×128: 85.0% (worst)

Interpretation: Lower resolutions forced models to learn robust, generalisable features by constraining available information. Higher resolutions enabled memorisation of dataset-specific artifacts.

2. *Critical Discovery (Trial 4):* The 80×80 model achieved 89.6% confidence on a real-world image when that image was preprocessed to 128×128 resolution (vs. 13.8% at 80×80). This suggested:

- Resolution affects real-world generalisation
- Models trained on lower resolutions learn more transferable features
- Preprocessing pipeline critically impacts deployment

3. *Augmentation Trade-offs:*

- No augmentation: Severe overfitting (Trial 1)
- Moderate augmentation: Best balance (Trials 4, 7)
- Aggressive augmentation: Degraded all metrics (Trials 3, 6)

4. *The Fundamental Limitation***:** Even the best configuration (Trial 4: 80×80, moderate augmentation, 98.7% test accuracy) achieved only **9.1% real-world success rate** (1/11 test images).

3.4 Phase 1 Conclusion and Transition

*Phase 1 demonstrated***:**

- Resolution optimisation alone cannot bridge the domain gap
- Models can achieve 98%+ test accuracy while failing in deployment
- The limiting factor appears to be GTSRB's idealised nature, not resolution choice

Question leading to Phase 2: **If resolution optimisation failed, could architectural sophistication succeed?** Specifically, could Vision Transformer's global attention mechanisms help models focus on sign-specific features rather than background artifacts?

Hypothesis for Phase 2: Transformer attention might improve real-world generalisation by:

- Modeling long-range dependencies across the entire image
- Learning to attend to relevant sign features regardless of background clutter
- Capturing global context that local CNN receptive fields miss

4. Phase 2: Architectural Sophistication vs Dataset Limitations

4.1 Experimental Design

Building on Phase 1 findings, we selected Trial 4 (80×80, moderate augmentation) as baseline and added transformer components.

Model 1 - Baseline (Pure CNN from Phase 1, Trial 4):

- 5 convolutional blocks (32 → 64 → 128 → 256 → 512 filters)
- Batch normalization, dropout, global average pooling
- Dense classifier (43 classes)
- ~5M parameters
- **Test accuracy: 98.35%**
- **Real-world success: 9.1% (1/11 images)**

Model 2 - Experimental (Hybrid CNN-ViT):

- Same CNN backbone (weights transferred from Model 1)
- 3 Transformer blocks:
  - 8 attention heads per block
  - 1024 MLP dimension
  - SwiGLU activation custom layer
- Transformer trained from scratch
- ~85M parameters
- Quantization-Aware Training (QAT) applied

4.2 Training Configuration

Common settings (inherited from Phase 1):

- Resolution: 80×80×3
- Dataset: GTSRB (same split as Phase 1)
- Augmentation: Moderate Albumentations (validated in Phase 1)
- Optimizer: Adam (1×10^{-4})

- Hardware: NVIDIA RTX 2060

Hybrid-specific:

- Training time: 50 minutes (vs. 75 minutes for pure CNN)
- Early stopping: Triggered at epoch 36
- GPU utilisation: Higher but stable compared to CNN-only
- QAT leveraged for deployment optimisation without post-training conversion

4.3 Phase 2 Results

Benchmark Performance Comparison:

Metric	Pure CNN (Phase 1)	Hybrid CNN-Transformer	Difference
Training Accuracy	91.74%	94.09%	+2.35%
Validation Accuracy	99.98%	99.98%	0.00%
Test Accuracy	98.35%	98.58%	+0.23%
Test Loss	0.062	0.053	-14.5%
Parameters	1.2M	2.8M	+133%
Training Time	~40 min	~50 min	+25%

Real-World Performance (12 diverse traffic sign images):

Sign Type	Pure CNN	Hybrid	Change	Analysis
No overtaking	98.5% ✅	4.6% ✕	**-93.9%**	Catastrophic regression
60 km/h	42.8% ◍	95.5% ✅	**+52.7%**	Dramatic improvement
Stop sign	14.6% ✕	98.3% ✅	**+83.7%**	Geometric recognition success
100 km/h	15.7% ✕	Failed ✕	Worse	Continued digit confusion
Bend right	N/A	Failed ✕	N/A	New test case

Sign Type	Pure CNN	Hybrid	Change	Analysis
Go left	4.95% ✗	Failed ✗	No change	Directional signs difficult
50 km/h	Failed ✗	Failed ✗	No change	Speed limit issues persist
120 km/h	Failed ✗	Failed ✗	No change	Speed limit issues persist
Crossing Animal	Failed ✗	Failed ✗	No change	Pictorial complexity
School crossing	Failed ✗	Failed ✗	No change	Pictorial complexity
Cycle crossing	Failed ✗	Failed ✗	No change	Pictorial complexity
Slippery road	Failed ✗	Failed ✗	No change	Warning signs struggle

Success Rate Summary:

- Pure CNN: 1/11 successful (9.1%)
- Hybrid: 2/12 successful (16.7%)
- **Net improvement: +7.6 percentage points**
- **Still failing on 83.3% of real-world images**

4.4 Critical Observation: Context Dependency

Additional testing revealed:

- When real-world images were manually cropped to resemble GTSRB (sign-focused, minimal background):
    - **Hybrid model: 96-100% accuracy**
    - **Pure CNN: Lower performance maintained**

Interpretation: Transformer attention helps when signs are prominent but may be distracted by heavily cluttered full-scene backgrounds. The global attention mechanism that should help with context understanding instead appears vulnerable to background noise in natural scenes.

---

5. Integrated Analysis

5.1 The Resolution-Architecture-Dataset Triangle

Phase 1 tested resolution:

- Result: Lower resolution unexpectedly better for generalisation
- Limitation: Still only 9.1% real-world success at optimal resolution

Phase 2 tested architecture:

- Result: Transformer attention marginally improved success (9.1% → 16.7%)
- Limitation: Still failing on 83% of real-world images

Common factor in both phases:

- Both achieved >98% test accuracy
- Both failed on majority of real-world images
- Success correlated with similarity to GTSRB training distribution

Conclusion: The binding constraint is neither resolution nor architecture, but **training data that doesn't represent deployment conditions**.

5.2 Understanding the Improvements and Regressions

Where transformers helped:

Stop sign (14.6% → 98.3%):

- Octagonal shape unique in dataset
- Transformer attention successfully focused on geometric features
- Global receptive field captured shape context better than local CNN

60 km/h (42.8% → 95.5%):

- Transformer better at discriminating numerical details
- Attention mechanism focused on digit features
- Less confused by similar speed limits (80, 100)

Where transformers hurt:

No overtaking (98.5% → 4.6%): Possible explanations:

1. Attention mechanism focused on wrong image regions (background clutter)
2. CNN features optimised for direct classification, not transformer input
3. Transformer overfitted to GTSRB validation patterns during training
4. Complex background in real-world image distracted global attention

Where nothing helped:

- All pictorial signs (crossings, animals): Failed in both phases
- Speed limits 50/120: Failed in both phases
- Warning triangles (slippery road): Failed in both phases

Pattern: Complex internal details and signs underrepresented or absent in GTSRB training data remain challenging regardless of optimization strategy.

5.3 The Benchmark Performance Illusion

Both phases achieved similar test accuracy:

- **Phase 1 (pure CNN):** 98.35%
- **Phase 2 (hybrid):** 98.58%

Both phases achieved similar validation accuracy:

- **Phase 1:** 99.98%
- **Phase 2:** 99.98%

Yet real-world performance remained poor:

- **Phase 1:** 9.1% success
- **Phase 2:** 16.7% success

Why test metrics mislead:

1. **Distribution matching:** Test set drawn from same distribution as training
2. **Preprocessing alignment:** Test images pre-cropped like training images
3. **Regional consistency:** All German signs following same design standards

4. **Condition control:** Similar lighting, weather, and image quality

Deployment reality differs on all dimensions:

- Multiple countries' sign designs
- Natural scene context (not pre-cropped)
- Variable imaging conditions
- Aged, weathered, or partially occluded signs

Implication: Test accuracy is a poor proxy for deployment readiness when training and deployment distributions diverge fundamentally.

5.4 Why Architectural Sophistication Had Limited Impact

Transformer attention provides:

- Global context understanding
- Long-range dependency modeling
- Learned attention to relevant features

But this assumes:

- Relevant features present in training data
- Attention can distinguish signal from noise
- Global context helps rather than distracts

In our experiments:

- ☑ Helped with geometric discrimination (stop sign)
- ☑ Improved some numeric recognition (60 km/h)
- ✗ Regressed on previously successful cases (no overtaking)
- ✗ No help with unseen sign types or conditions
- ✗ Distracted by complex backgrounds in full scenes

Fundamental issue: Transformers learn to attend to patterns in training data. When deployment data has different patterns (different sign designs, contexts, conditions), learned attention doesn't transfer.

5.5 The Data-Centric Insight

If we had instead:

- Trained on Mapillary (global traffic signs, natural scenes)
- Collected diverse real-world training data
- Used data from multiple countries
- Included various weather and lighting conditions

Then:

- Both architectures would likely perform better
- Resolution choices would matter less
- Model would learn robust, generalisable features
- Test and deployment performance would align

The lesson: Investing in representative training data yields greater returns than optimising resolution or architecture when facing domain shift.

---

6. Broader Implications

6.1 Rethinking ML Deployment Strategy

Common industry approach:

1. Achieve high benchmark accuracy (>95%)
2. Deploy with confidence
3. Discover failures in production
4. Attempt to fix with model tweaks

Evidence-based approach (from this empirical study):

1. Evaluate on deployment-representative data early
2. Identify distribution gaps explicitly
3. Invest in diverse training data first
4. Consider architectural optimisation only after data issues addressed

6.2 The Benchmark-Driven Empirical study Problem

Academic empirical study often optimises for:

- Leaderboard rankings on standard benchmarks
- Incremental accuracy improvements (98.3% → 98.6%)
- Novel architectural components
- Reproducibility on controlled datasets

Industry deployment requires:

- Robustness to distribution shift
- Reliable performance on diverse real-world data
- Understanding of failure modes
- Practical deployment constraints (latency, memory, cost)

This empirical study reveals the gap**:** Models optimised for benchmarks may not optimize for deployment success.

6.3 When Does Architecture Matter?

Architecture helps when:

- Training and deployment distributions align
- Task complexity exceeds simpler model capacity
- Sufficient representative training data available
- Failure modes stem from model capacity, not data mismatch

Dataset quality matters when:

- Distribution shift between training and deployment
- Training data lacks diversity or representativeness
- Real-world conditions differ significantly from collection
- Model failures correlate with absence from training data

For traffic signs specifically:

- Dataset quality is the binding constraint

- Architecture provides marginal gains (7.6 percentage points)
- Representative data would likely provide larger gains (speculation: 50+ percentage points)

6.4 Implications for Computer Vision Empirical study

Recommendations for empirical researchers:

1. **Report real-world performance** alongside benchmark metrics
    - Not just accuracy on held-out test sets
    - Actual deployment-representative evaluation
    - Explicit measurement of distribution shift
2. **Create realistic datasets** that reflect deployment conditions
    - Natural scene context, not artificial cropping
    - Diverse geographic and environmental conditions
    - Representative of actual model usage
3. **Study dataset characteristics systematically**
    - How does training distribution affect generalisation?
    - What data characteristics predict deployment success?
    - Can we quantify distribution shift predictively?
4. **Develop domain adaptation techniques**
    - Methods specifically addressing distribution shift
    - Few-shot learning for adapting to new conditions
    - Self-supervised learning on unlabeled deployment data

Recommendations for practitioners:

1. **Evaluate deployment conditions early**
    - Collect representative test data from deployment environment
    - Measure distribution shift explicitly
    - Set realistic performance expectations

2. **Invest in data quality**
    - Diverse, representative training data
    - Data augmentation matching deployment variation
    - Consider transfer learning from more diverse sources
3. **Consider detection pipelines for real-world objects**
    - Classification assumes pre-localised objects
    - Real deployment needs detection + classification
    - Two-stage pipelines handle natural scene complexity
4. **Be skeptical of benchmark-only validation**
    - High test accuracy ≠ production readiness
    - Domain gap often invisible until deployment
    - Real-world evaluation necessary before claiming success

---

7. Limitations and Future Work

7.1 Study Limitations

Real-world evaluation scope:

- Limited to 12 hand-collected test images
- Single manual collection (potential selection bias)
- Need systematic sampling across geographies, conditions
- Larger validation set required for statistical rigor

Dataset scope:

- Only GTSRB tested
- Cannot generalise findings to all traffic sign datasets
- Need comparison with Mapillary, LISA, CTSD

Architecture scope:

- Two architectures compared (pure CNN, CNN+ViT)

- Other hybrid designs might perform differently
- Detection architectures (YOLO, Faster R-CNN) not evaluated
- Ensemble methods not explored

Resolution constraints:

- Limited by RTX 2060 memory (6GB)
- 128×128 maximum tested
- Higher resolutions might change Phase 1 conclusions
- Full exploration of resolution-architecture interaction space incomplete

Computational constraints:

- Single GPU training (no distributed experiments)
- QAT implementation details could be elaborated further
- Detailed computational cost analysis deferred to appendix materials

7.2 Future Empirical study Directions

Dataset expansion and evaluation:

- Systematic real-world test set collection across multiple countries
- Evaluate on Mapillary Traffic Sign Dataset
- Test on degraded signs (weathering, vandalism, occlusion)
- Adverse weather conditions (fog, rain, snow)

Architectural exploration:

- Detection architectures: YOLO, Faster R-CNN, DETR
- Other attention mechanisms: Swin Transformer, CoAtNet
- Ensemble methods combining multiple architectures
- Efficient architectures for edge deployment

Domain adaptation empirical study:

- Fine-tuning strategies with small real-world datasets
- Self-supervised learning on unlabeled street imagery

- Domain randomisation and synthetic augmentation
- Few-shot adaptation to new geographic regions

Resolution-architecture interaction:

- Systematic study at higher resolutions (256×256, 384×384)
- Optimal trade-offs for different hardware constraints
- Relationship between resolution and attention mechanism effectiveness

Practical deployment:

- Two-stage detection + classification pipeline implementation
- Real-time inference optimisation
- Edge device deployment (mobile, embedded systems)
- Human-in-the-loop validation systems

---

8. Conclusions

This two-phase investigation systematically demonstrates that **technical optimisation (resolution, architecture) cannot compensate for training on idealised, region-specific datasets when deploying to diverse real-world conditions**.

8.1 Phase 1 Conclusions (Resolution Investigation)

1. **Resolution paradox:** Lower resolutions (64×64, 80×80) unexpectedly outperformed higher resolutions (128×128) on real-world images, challenging the assumption that more pixels improve generalisation.
2. **Optimal configuration achieved 98.7% test accuracy but only 9.1% real-world success**, revealing the benchmark-deployment performance gap.
3. **Augmentation requires careful calibration:** Both insufficient and excessive augmentation degraded performance.
4. **Resolution optimisation has limits:** Even optimal resolution cannot bridge fundamental domain gaps.

8.2 Phase 2 Conclusions (Architecture Investigation)

1. **Architectural sophistication provided marginal improvement:** Transformer attention increased real-world success from 9.1% to 16.7% (still failing on 83% of images).
2. **Mixed results across sign types:**
    - Dramatic improvements: Stop sign (+83.7%), 60 km/h (+52.7%)
    - Catastrophic regressions: No overtaking (-93.9%)
    - No change: All pictorial and warning signs
3. **Context dependency matters:** Hybrid model performed well on cropped images (96-100%) but struggled with full scenes, suggesting attention mechanisms require careful handling of background clutter.
4. **Architecture cannot overcome dataset limitations:** Adding transformer components to an already-strong CNN provided only incremental benefits when fundamental constraint is training data mismatch.

8.3 Integrated Conclusions

Central Thesis Validation:

Both phases strongly support the thesis that **dataset characteristics constrain deployment success more than technical optimisation**. Resolution tuning and architectural sophistication both failed to bridge the domain gap created by training on GTSRB's idealised German traffic signs.

The Benchmark Performance Illusion: Models in both phases achieved >98% test accuracy while failing on >75% of real-world images. This demonstrates that benchmark metrics provide false confidence when training and deployment distributions diverge.

What Works, What Doesn't:

- ✅ Works: Signs visually similar to GTSRB training data
- ❌ **Doesn't work:** Different countries' designs, complex backgrounds, pictorial symbols, degraded conditions
- ◍ **Sometimes works**: Architectural improvements help specific cases but don't solve fundamental problem

The Path Forward: For production-ready traffic sign recognition:

1. **Detection pipelines required:** Pure classification assumes pre-localised objects
2. **Diverse training data essential:** Multi-country, natural scene datasets like Mapillary
3. **Representative evaluation mandatory:** Test on deployment-representative data, not just held-out test sets
4. **Domain adaptation techniques:** Methods specifically addressing distribution shift

8.4 Final Perspective

The machine learning community's focus on architectural innovation has delivered remarkable progress on benchmark tasks. However, this empirical study illustrates a critical lesson: **optimising for controlled evaluations while ignoring deployment realities creates models that succeed in the lab but fail in the field**.

Key insights for the ML community:

1. **Benchmark accuracy is necessary but insufficient** for deployment readiness
2. **Representative training data matters more than architectural sophistication** when facing domain shift
3. **Negative results have value:** Understanding what doesn't work prevents wasted effort
4. **Honest evaluation builds trust:** Reporting real-world failures alongside benchmark successes advances the field

As computer vision systems increasingly deploy in safety-critical applications like autonomous driving, the gap between benchmark success and deployment failure carries real consequences. Bridging this gap requires acknowledging that dataset characteristics, not just architectural sophistication or resolution optimization, fundamentally determine real-world performance.

The work ahead: Creating representative datasets, developing robust domain adaptation techniques, and establishing evaluation practices that predict deployment success—not just benchmark rankings.

---

9. Technical Implementation Details

9.1 Phase 1 Architecture (Pure CNN)

Input (Resolution varies: 64×64, 80×80, 90×90, or 128×128)

```
├── Conv2D(32, 3×3, padding='same', activation='relu')
├── BatchNormalization()
├── MaxPooling2D(2×2)
├── Dropout(0.5)
├── Conv2D(64, 3×3, padding='same', activation='relu')
├── BatchNormalization()
├── MaxPooling2D(2×2)
├── Dropout(0.5)
├── Conv2D(128, 3×3, padding='same', activation='relu')
├── BatchNormalization()
├── MaxPooling2D(2×2)
├── Dropout(0.5)
├── Conv2D(256, 3×3, padding='same', activation='relu')
├── BatchNormalization()
├── MaxPooling2D(2×2)
├── Dropout(0.5)
├── GlobalAveragePooling2D()
├── Dense(256, activation='relu')
├── Dropout(0.5)
└── Dense(43, activation='softmax')
```

Parameters: ~5M (varies slightly by input resolution)

9.2 Phase 2 Architecture (Hybrid CNN-Transformer)

```
Input (80×80×3)
|
|   ├── Conv blocks (32→64→128→256→512) with BatchNorm, MaxPool, Dropout
|  └── Output: (batch, 5, 5, 512)
|
 ├── Reshape to sequence: (batch, 25, 512)
|  └── Flatten spatial dimensions: 5×5 = 25 tokens
|
 ├── [Trainable Transformer Blocks - 3 layers]
|  └── For each layer:
|      ├── Multi-Head Self-Attention (8 heads, 256 dim per head)
|      ├── LayerNormalization
|      ├── Feed-Forward MLP:
|      |   ├── Dense(1024, activation='SwiGLU')
|      |   ├── Dense(512)
|      |  └── Dropout(0.1)
|     └── LayerNormalization
|
 ├── GlobalAveragePooling over sequence
 ├── Dense(512, activation='SwiGLU')
 ├── Dropout(0.5)
└── Dense(43, activation='softmax')
```

Total Parameters: ~85M

9.3 Training Configuration Summary

Phase 1 (Resolution experiments):

- **Optimiser:** Adam (learning rate: 1×10^{-4})
- **Batch size:** 32 for 64×64, 16 for larger resolutions
- **Augmentation:** Evolved from none → Keras → Albumentations (moderate)
- **Regularisation:** Early stopping (patience=6), ReduceLROnPlateau (factor=0.5, patience=3)
- **Mixed precision:** FP16
- **Training time**: 40-75 minutes per trial

Phase 2 (Hybrid architecture):

- **Optimiser:** Adam (learning rate: 1×10^{-4})
- **Batch size:** 16 (transformer memory requirements)
- **Augmentation**: Moderate Albumentations (validated in Phase 1)
- **Regularisation:** Same as Phase 1
- **Quantization-Aware Training:** Applied for deployment optimisation
- **Training time:** 50 minutes (early stopped at epoch 36)

9.4 Computational Resources

Hardware: NVIDIA RTX 2060 (6GB VRAM)

Memory management:

- Experimental memory growth enabled
- Mixed precision (FP16) for memory efficiency
- Batch size adjusted per resolution to maximise GPU utilisation

Training efficiency:

- Phase 1 pure CNN: ~40 minutes per trial
- Phase 2 hybrid: ~50 minutes
- GPU utilisation: Higher for hybrid but remained stable

- QAT leveraged for deployment optimisation without significant overhead

9.5 Code and Reproducibility

Repository structure:

- **Phase 1:** Resolution experiments (7 trials documented)
- **Phase 2:** Hybrid architecture implementation
- Preprocessing pipelines for both phases
- Evaluation scripts for benchmark and real-world testing
- Inference utilities for deployment testing

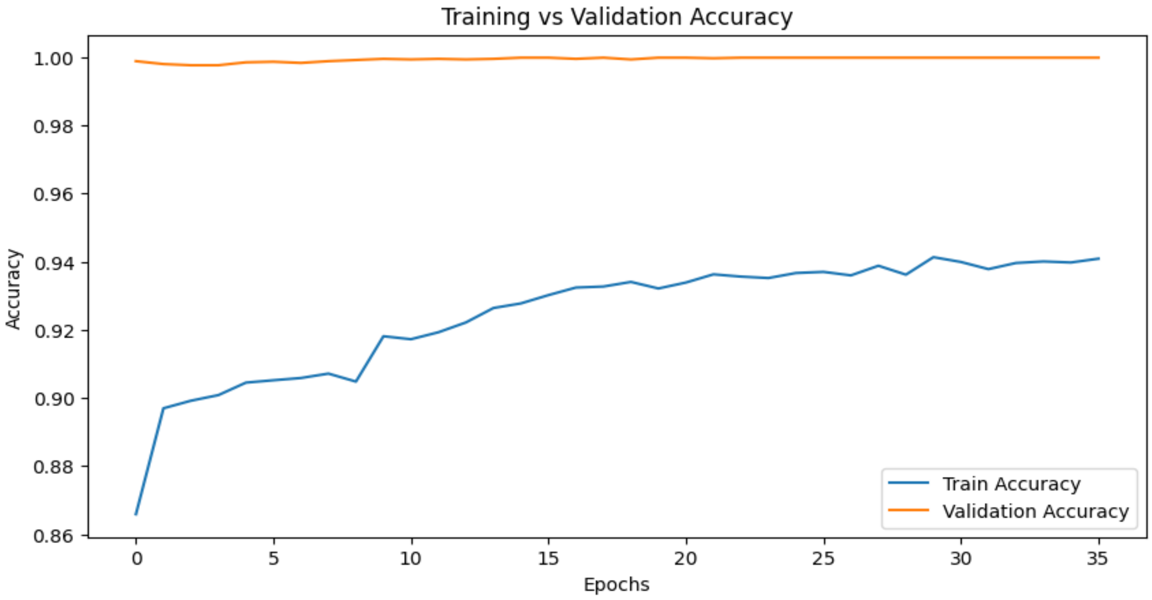
Reproducibility:

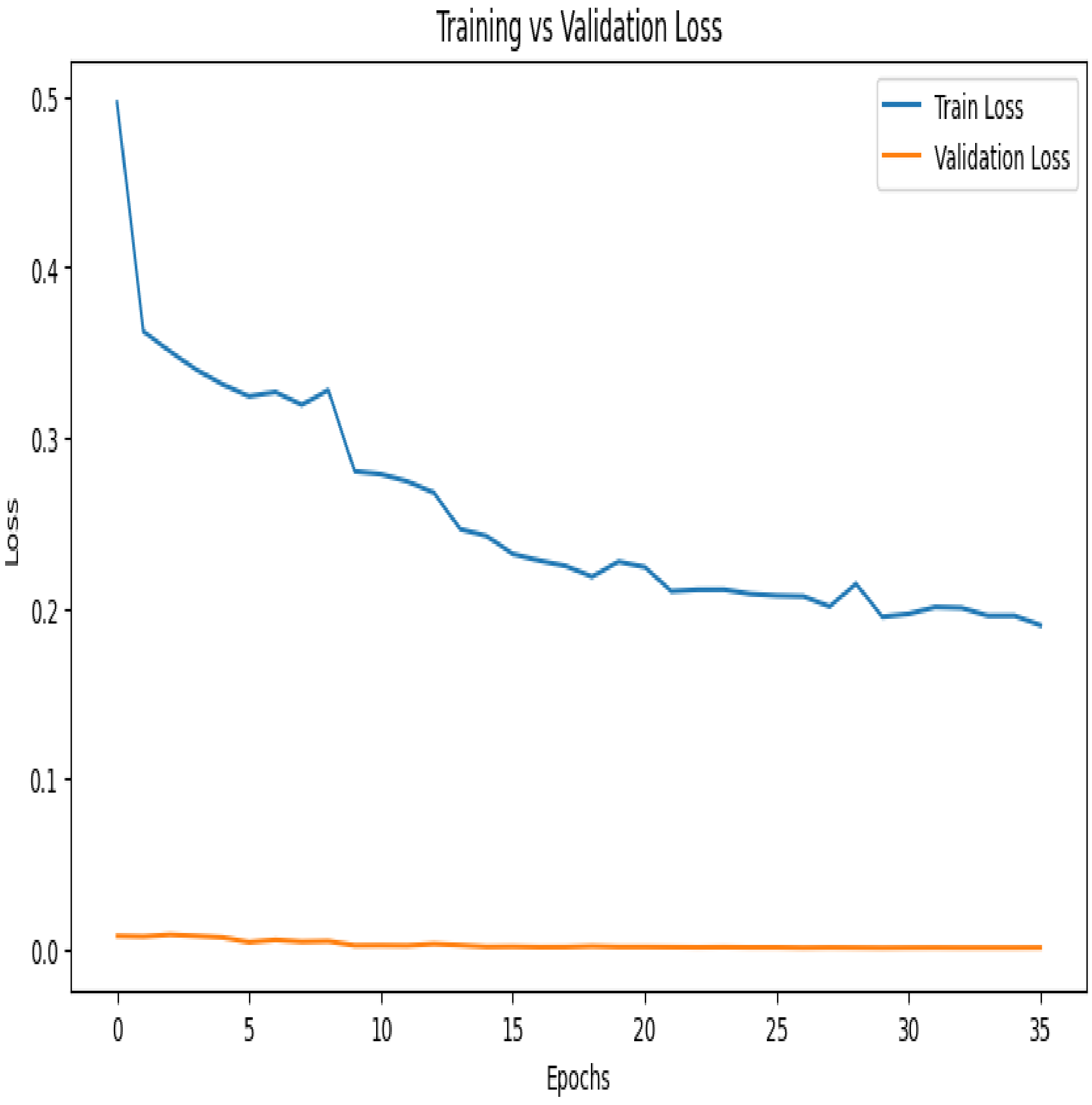
- Random seed: 42 – 7331 (hybrid model)
- Complete hyperparameter documentation
- Dataset split preservation between phases
- Preprocessing pipeline versioning

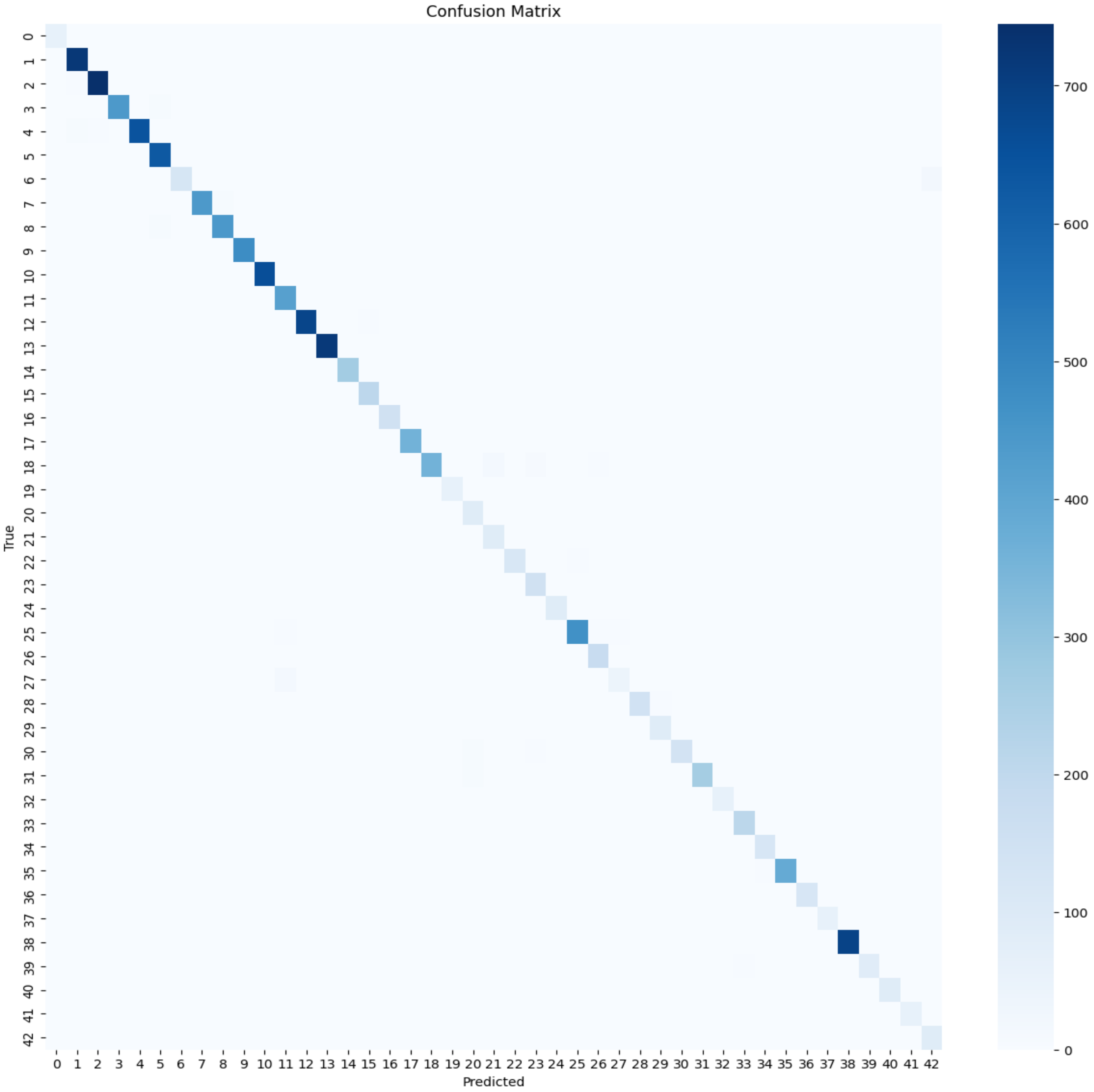
Availability: Complete implementation available at https://github.com/AhmadElsisy

---

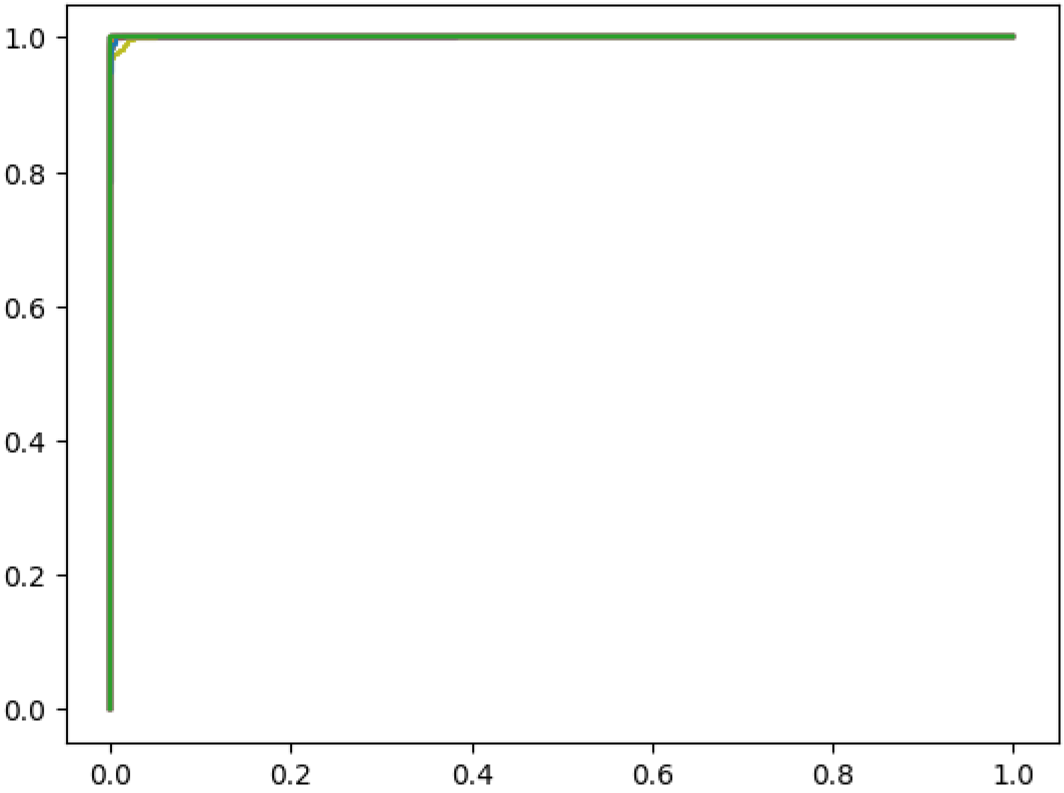
10. Appendix: Supplementary Materials



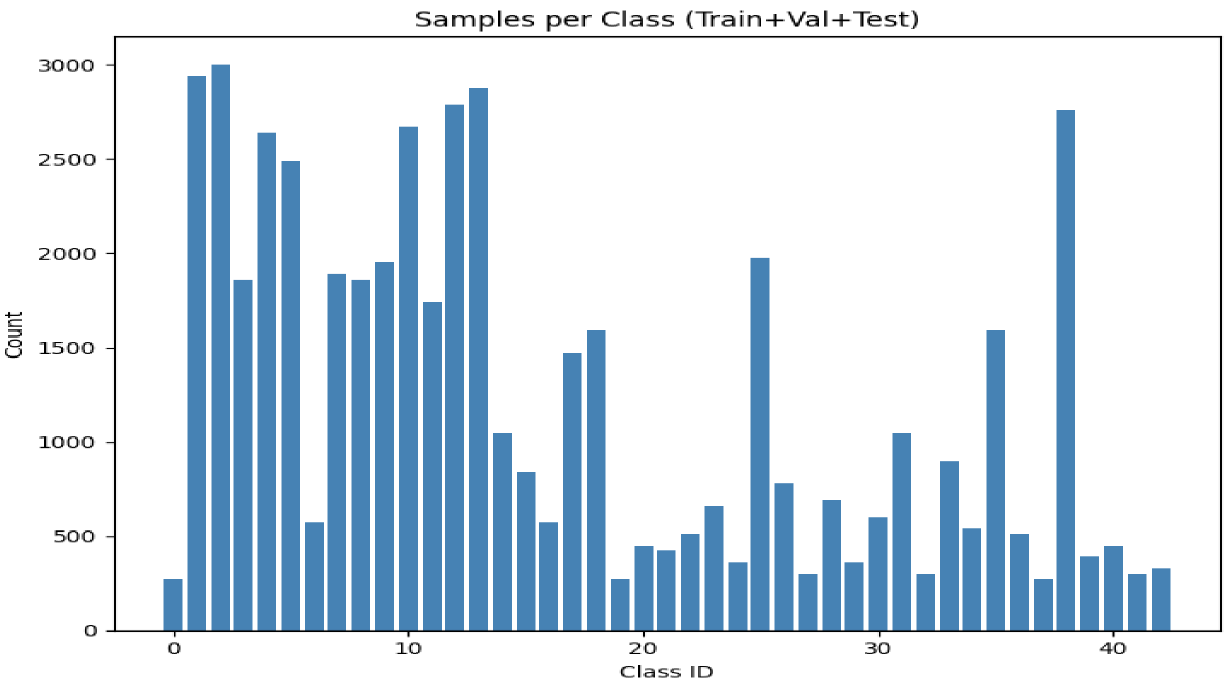

Training vs Validation Loss
Train Loss
Validation Loss
Loss
Epochs
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
25
30
35



Confusion Matrix



ROC-AUC



Acknowledgments

This empirical study was conducted as an independent investigation into practical deployment challenges in computer vision. The systematic experimental approach demonstrates the importance of honest evaluation and the courage to report negative results—findings showing architectural limitations are as valuable as those showing improvements.

Special recognition to the critical thinking required to question benchmark-driven assumptions and investigate the fundamental question: "Does this actually work in practice?"

References

Datasets:

- Stallkamp, J., et al. (2012). "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition." Neural Networks.
- Neuhold, G., et al. (2017). "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes." ICCV.

Architectures:

- Dosovitskiy, A., et al. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." ICLR.
- Liu, Z., et al. (2021). "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." ICCV.
- Vaswani, A., *et al*. (2017) Attention Is All You Need

Domain Adaptation:

- Ganin, Y., & Lempitsky, V. (2015). "Unsupervised Domain Adaptation by Backpropagation." ICML.
- Tzeng, E., et al. (2017). "Adversarial Discriminative Domain Adaptation." CVPR.