

Parameter-Efficient Transfer Learning for Image Fragment Reassembly: Robustness Analysis of CNN and Vision Transformer Approaches

Chuning LIU

*Department of Computer and Information Sciences
Auckland University of Technology
Auckland, New Zealand*

Wei Qi YAN

*Department of Computer and Information Sciences
Auckland University of Technology
Auckland, New Zealand*

Abstract—Pairwise fragment relation classification, which predicts the relative position of two image fragments, is the critical stage of most learned fragment reassembly systems, yet methods are evaluated almost exclusively on clean inputs. This paper compares a fully fine-tuned ViT-Small, ViT-Small adapted with LoRA and with Visual Prompt Tuning (VPT) at an identical 114K-parameter budget, and a ResNet-18 reference, on CelebA fragment pairs at 3×3 and 4×4 grids under Gaussian noise, edge erosion, and random occlusion applied at the fragment level. The first has been showed that, under a shared full fine-tuning recipe, parameter-efficient fine-tuning (PEFT) methods appeared to suffer from performance collapse on a finer grid (4×4)—with clean-data accuracy of only 45%—and exhibited vastly different performance characteristics depending on the type of degradation. After per-method learning-rate selection and training to convergence, LoRA reaches within 4.5 points of full fine-tuning at 0.5% trainable parameters, retains most of the backbone’s corruption robustness, and outperforms budget-matched VPT under nearly all conditions. Independent of protocol, robustness remains degradation-dependent: Transformer models degrade gracefully under noise and occlusion, whereas the CNN is far more robust to edge erosion, a reversal we trace largely to the boundary-strip input channel. Three experimental factors have been identified that can lead to biased conclusions: learning rate mismatch, insufficient training budget on finer grids, and noise injection across composite layers; any of these factors can potentially reverse conclusions regarding PEFT robustness.

Index Terms—image fragment relation classification, jigsaw puzzle, vision transformer, parameter-efficient fine-tuning, corruption robustness, LoRA, visual prompt tuning

I. INTRODUCTION

Image fragment reassembly reconstructs a complete image from its pieces and supports forensic document restoration and archaeological artefact reconstruction [14]. Modern learned systems solve it in two stages [9], [18]: a classifier first predicts the relative spatial relation of each fragment pair, and a search or optimisation stage then assembles the image from these pairwise predictions. Because every assembly decision depends on the classifier, the accuracy of pairwise relation classification bounds the performance of the whole pipeline. This paper studies that classification stage in isolation.

Although learned classifiers reach high accuracy on clean benchmark data, captured fragments contain sensor noise,

physical edge damage, and partial occlusion, and existing studies report results almost only on clean inputs. Two developments lead to a re-examination. First, Vision Transformers (ViT) [1] are reported to be more robust than CNNs to perturbation and occlusion in object classification [15]–[17], but fragment relation classification depends on fine boundary detail, so that advantage need not transfer. Second, parameter-efficient fine-tuning (PEFT) methods such as LoRA [3] and Visual Prompt Tuning (VPT) [4] adapt a frozen backbone through fewer than 1% of its parameters, which is attractive for the constrained hardware on which damaged inputs are most likely, yet whether an adapted model keeps the robustness of its backbone is largely untested.

There is a methodological trap when comparing PEFT methods to full fine-tuning. The optimal training recipe of a method that updates 0.5% of the parameters behind a frozen backbone differs from that of full fine-tuning, so a recipe shared across methods can systematically under-train PEFT and produce conclusions about capacity or robustness that are really conclusions about optimisation. Therefore, treating the evaluation protocol itself is seen as an object of study: every model is trained with a per-method learning rate and a convergence-based budget, and the same models trained under a shared recipe are kept as a control.

The contributions of this paper are:

- A benchmark for corruption robustness in pairwise fragment relation classification: source-disjoint train, validation, and test splits, per-method learning rates, training to a stated convergence criterion, and three corruption families applied at the fragment level, with test accuracy tracking validation within 0.2 points for every model.
- Evidence that robustness is degradation-dependent. Transformer models degrade gracefully under noise and occlusion, whereas the CNN is far more robust to edge erosion, and a composite-input ablation traces most of the Transformer’s erosion collapse to the boundary-strip channel rather than to the architecture alone.
- Three documented protocol artifacts in PEFT robustness evaluation. Under a shared recipe, PEFT appears to collapse on the finer grid whereas LoRA and VPT exhibited

markedly different patterns of robustness degradation. Following method-specific hyperparameter tuning, the performance gap between LoRA and full fine-tuning narrowed to within 4.5 points while retaining most of its robustness, outperforming VPT, which has a comparable computational budget under almost all conditions. The issue of insufficient budget on finer grids affected all models, not just PEFT.

Section II reviews related work, Section III describes data, models, recipes, and the evaluation protocol, Section IV reports results, Section V discusses mechanisms and artifacts, and Section VI concludes.

II. RELATED WORK

A. Fragment and Jigsaw Puzzle Reassembly

Early studies scored piece boundaries with handcrafted compatibility measures [14]. Deep learning reformulated the task as pairwise relation prediction followed by global assembly; Deepzpple scores fragment pairs with a Siamese network and assembles them by shortest-path search [18]. Transformer-based and generative studies followed, including edge-aware permutation learning that targets eroded boundaries [9], direct coordinate prediction [10], diffusion transformers for missing pieces [6], and evolutionary [8] or vision language [7] studies for large puzzles with eroded gaps, each addressing one damage type through a specialised architecture. No prior study compares general backbones and fine-tuning strategies across several corruption types for this task under one shared protocol.

B. Robustness of Vision Transformers

ViT processes an image as a sequence of patches under global self-attention [1], and several studies report that it withstands occlusion, common corruptions, and spatial perturbations better than CNNs in object classification [15]–[17]. Those tasks are solved by global shape, whereas fragment relation classification depends on thin boundary evidence, so the transfer of this finding is an open question that has been tested directly.

C. Parameter-Efficient Fine-Tuning

PEFT methods update a small parameter set while the backbone stays frozen [5]. LoRA adds trainable low-rank matrices beside frozen projections [3]; VPT prepends learnable prompt tokens at each layer and leaves all weights unchanged [4]. Both approach full fine-tuning accuracy on standard benchmarks, but those evaluations measure clean accuracy, and the robustness that a frozen backbone passes on to its adaptation has received comparatively little attention. Moreover, PEFT papers commonly report a single learning rate shared with full fine-tuning; It is showed below that this choice alone can reverse conclusions about PEFT robustness.

III. METHODOLOGY

A. Task and Data

CelebA face images dataset has been used. [12] (178×218 pixels). Because neither dimension is divisible by 3 or 4, each image is cropped in center to the largest grid size (216×177 for 3×3 , 216×176 for 4×4) and cut into $N \times N$ equal fragments (59×72 and 44×54 pixels). A classifier receives an ordered pair of fragments and predicts one of $C=9$ classes: the eight adjacency directions or non-adjacent. The label space is fixed for every grid size, while difficulty rises with N . All ordered pairs of each image are enumerated, giving 72 pairs per 3×3 image and 240 per 4×4 image; on the 3×3 grid the non-adjacent class holds 44.4% of pairs, each orthogonal direction 8.33%, and each diagonal direction 5.56%.

Splits are made at the image level so that no source image contributes pairs to more than one split. The same 5,000 training and 1,000 validation images are used for both grids (yielding the 360K and 72K pairs per configuration); the 4×4 enumerations are subsampled to the same pair budget with a fixed seed. A further 1,000 images never used in any earlier experiment form the test set on which every result in this paper is reported, validation is used only for checkpoint and learning-rate selection.

B. Composite Input

Each pair is encoded as a single 224×224 composite so that one single-stream backbone serves every grid. The canvas is divided into four 112×112 quadrants: the two fragments, resized bilinearly, occupy the top row, and two boundary strips occupy the bottom row. The horizontal strip joins the 10-pixel right edge of the first resized fragment to the 10-pixel left edge of the second, the vertical strip joins the bottom and top edges likewise, and each strip is resized to fill its quadrant; the remaining canvas is uniform gray (value 128). Strip width is measured on the resized 112×112 fragments.

C. Models

Four configurations are compared (Table I). **ResNet-18** [2] (torchvision, ImageNet-1k weights) is fully fine-tuned with a two-layer head ($512 \rightarrow 256$, batch normalisation, ReLU, dropout 0.5, linear to C). **ViT-S-Full** is ViT-Small/16 [1] (timm, ImageNet-21k pre-trained and ImageNet-1k fine-tuned *augreg* weights) with all parameters trainable and a linear head. **ViT-S-LoRA** freezes the same backbone and attaches LoRA [3] with rank $r=4$ and $\alpha=8$ to the fused QKV and output projections of every block; the low-rank bypass and the linear head are the only trainable parameters. **ViT-S-VPT-24** freezes the backbone and prepends 24 learnable prompt tokens at every layer (VPT-Deep [4]), again with a trainable linear head. LoRA and VPT-24 train an identical 114,057 parameters (0.52%), so their comparison is budget-matched by construction. The CNN serves as a reference rather than a controlled contrast: it differs from ViT-S in parameter count, pre-training data, head, and receptive-field structure, so we do not attribute CNN–ViT differences to a single factor.

TABLE I

TRAINING PARAMETERS (LEARNING RATE SELECTED ON VALIDATION FROM $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$ FOR PEFT; COSINE BUDGET EXTENDED UNTIL THE FINAL EPOCH VALIDATION GAIN FALLS BELOW 0.3 POINTS), AND CLEAN TEST SET METRICS.

Model	Params	lr	Grid	Ep.	Acc.	Bal.	M-F1
ResNet-18	11.3M (100%)	10^{-4}	3×3	8	99.31	99.36	99.15
			4×4	16	99.23	99.75	98.90
ViT-S-Full	21.7M (100%)	10^{-4}	3×3	8	99.11	99.27	98.92
			4×4	24	96.99	98.89	95.58
ViT-S-LoRA	114K (0.5%)	10^{-3}	3×3	16	98.73	99.10	98.50
			4×4	24	92.46	97.95	90.03
ViT-S-VPT-24	114K (0.5%)	10^{-3}	3×3	16	97.32	98.40	96.95
			4×4	32	79.38	94.90	78.36
Majority class			3×3		44.44	11.11	–
			4×4		64.81	11.11	–

D. Training

All models are trained on clean data with AdamW [13] (weight decay 10^{-4}), batch size 256, cosine annealing to 10^{-6} , focal loss [11] ($\gamma=2$, class weights proportional to inverse frequency) to counter the non-adjacent majority, and colour jitter (brightness and contrast 0.1) as the only augmentation, geometric augmentations are excluded because flips would corrupt directional labels. The best checkpoint is kept and the random seed is fixed. Every model uses learning rate 10^{-4} and 8 epochs, the setting chosen for full fine-tuning.

For LoRA and VPT the learning rate is selected on validation from $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$, 10^{-3} is best for both. Full fine-tuning retains 10^{-4} . The cosine budget is then extended (8, 16, 24, or 32 epochs, each a fresh run) until the final epoch validation gain falls below 0.3 points. Table I lists the finalized epoch counts: the 3×3 full models converge within 8 epochs, whereas on 4×4 ResNet-18 needs 16, ViT-S-Full 24, LoRA 24, and VPT 32. All results in Tables I, II, and IV use the training configurations specific to each method, Table III presents a comparison between the two. Training used a single NVIDIA A100.

E. Corruption Protocol and Metrics

Robustness is measured on the test set with corruption applied at test time only, to the raw fragments before resizing and composite construction, so that the boundary strips are built from corrupted pixels exactly as they would be in a damaged capture. Gaussian noise with $\sigma \in \{10, 20, 30, 50\}$ on the $[0, 255]$ scale is clipped to valid pixel values. Edge erosion replaces a band of 1, 2, 4, or 8 pixels along every fragment border with gray (128), measured on the original fragment (1–8 px correspond to 1.7–13.6% of a 3×3 fragment edge and 2.3–18.2% of a 4×4 edge). Random occlusion masks a rectangle covering 10, 20, or 30% of each fragment with gray, at an independent position per fragment drawn from a fixed per-pair seed. Accuracy has been reported for every condition, on clean data reported balanced accuracy and macro-F1, because the non-adjacent class holds 44.4% (3×3) and 64.8% (4×4) of test pairs.

IV. EXPERIMENTS AND RESULTS

A. Clean Accuracy

Table I reports clean test metrics. On the 3×3 grid all four models exceed 97% accuracy. On the 4×4 grid the rank is ResNet-18 (99.23%), ViT-S-Full (96.99%), ViT-S-LoRA (92.46%), and ViT-S-VPT-24 (79.38%). The balanced accuracies are markedly higher for the two PEFT models (97.95% and 94.90%), which shows that their errors are concentrated in one region of the confusion matrix rather than spread across classes, Section V-B identifies that region. Test accuracy tracks the validation accuracy used for checkpoint selection within 0.2 points for all eight configurations, so the selection procedure did not inflate the reported numbers.

B. Robustness Under Corruption

Table II and Fig. 1 report the full matrix.

Gaussian noise. Noise separates the architectures most sharply. On 3×3, ResNet-18 falls from 99.31% to 37.80% at $\sigma=30$ while ViT-S-Full keeps 87.33%, a gap of 49.5 points; on 4×4 the CNN is near the majority class floor from $\sigma=20$ onward whereas ViT-S-Full still holds 64.08% at $\sigma=30$. LoRA retains most of this tolerance (75.16% at $\sigma=30$ on 3×3; 53.37% on 4×4), and VPT retains less (71.18% and 30.11%).

Edge erosion. Erosion reverses the ranking. On 3×3 ResNet-18 keeps 86.06% at 1 px and 47.85% at 8 px, whereas ViT-S-Full falls to 71.39% and then to 10.21%, close to the 8.33% constant-prediction floor. On 4×4 the CNN leads clearly at 1 and 2 px (72.42% against 43.92%) and both models approach the floor from 4 px. One inversion is stable across protocols: at 8 px on 3×3 the frozen-backbone LoRA model (22.36%) and VPT model (16.23%) both exceed the fully fine-tuned ViT (10.21%).

Random occlusion. Occlusion is the mildest condition for every model, and the Transformer advantage grows with severity and grid: at 30% occlusion ViT-S-Full keeps 96.95% against 93.54% for the CNN on 3×3 and 93.60% against 85.98% on 4×4. LoRA loses 3.6 and 5.7 points from clean on the two grids, and VPT, although lowest in absolute accuracy, shows the smallest relative drop on 4×4 (3.3 points).

C. Budget-Matched Comparison of LoRA and VPT

The LoRA and VPT-24 rows of Table II share a budget of 114,057 parameters. LoRA leads on clean data by 1.4 points (3×3) and 13.1 points (4×4). VPT is ahead only under light noise on 3×3 ($\sigma \leq 20$; 90.58% against 86.77% at $\sigma=10$) and at 4 px erosion; LoRA leads every other cell on both grids, including all of 4×4. Under occlusion LoRA leads in absolute accuracy while VPT shows the smaller relative drop. Once both are tuned, LoRA is the stronger adaptation under nearly every condition.

D. Sensitivity to the Training Protocol

Table III evaluates the same four architectures on 4×4 under the identical final test protocol, trained either with the shared configuration or with the per-method configuration; Fig. 2 shows the validation curves behind two of the rows. Under

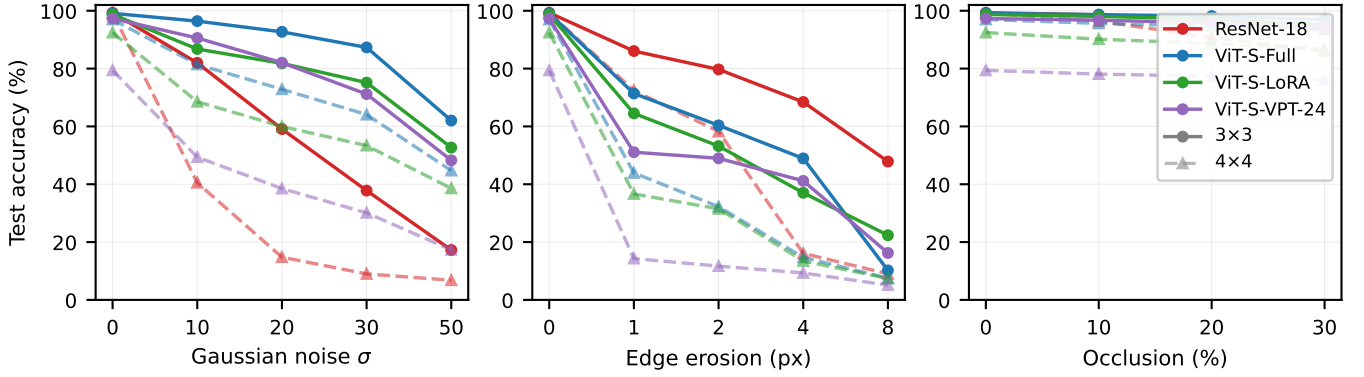


Fig. 1. Test accuracy under increasing severity of Gaussian noise (left), edge erosion (centre), and random occlusion (right). Solid lines with circles: 3×3 grid; dashed lines with triangles: 4×4 grid. Severity 0 is the clean condition.

TABLE II

TEST ACCURACY (%) ON THE SOURCE-DISJOINT TEST SET (1,000 UNSEEN IMAGES; 72,000 PAIRS PER GRID) UNDER THE UNIFIED FRAGMENT-LEVEL CORRUPTION PROTOCOL. ALL MODELS ARE TRAINED TO CONVERGENCE WITH PER-METHOD LEARNING RATES (TABLE I). ViT-S-LoRA AND ViT-S-VPT-24 TRAIN AN IDENTICAL 114,057 PARAMETERS, SO THEIR ROWS ALSO CONSTITUTE THE MATCHED-BUDGET COMPARISON. BEST VALUE PER COLUMN AND GRID IN BOLD.

Model	Grid	Clean	Gaussian noise (σ)				Edge erosion (px)				Occlusion		
			10	20	30	50	1	2	4	8	10%	20%	30%
ResNet-18	3×3	99.31	81.98	59.09	37.80	17.26	86.06	79.72	68.45	47.85	98.61	96.87	93.54
ViT-S-Full	3×3	99.11	96.42	92.71	87.33	62.02	71.39	60.37	48.93	10.21	98.66	98.15	96.95
ViT-S-LoRA	3×3	98.73	86.77	81.81	75.16	52.69	64.50	53.20	37.04	22.36	98.06	97.08	95.18
ViT-S-VPT-24	3×3	97.32	90.58	82.03	71.18	48.27	51.08	48.97	41.17	16.23	96.74	95.82	94.07
ResNet-18	4×4	99.23	40.51	14.79	8.98	6.81	72.42	58.25	16.09	9.02	96.41	90.81	85.98
ViT-S-Full	4×4	96.99	81.43	72.85	64.08	44.73	43.92	32.25	14.70	7.52	95.63	94.64	93.60
ViT-S-LoRA	4×4	92.46	68.54	59.98	53.37	38.58	36.64	31.48	13.44	7.40	90.16	88.68	86.74
ViT-S-VPT-24	4×4	79.38	49.38	38.50	30.11	17.31	14.29	11.68	9.31	5.10	78.09	77.35	76.09

the shared configuration LoRA and VPT reach only 45.19% and 45.56% clean accuracy and appear to be capacity-limited. Under the per-method configuration the same architectures reach 92.46% and 79.38%; the change is a learning rate one order of magnitude larger and a budget long enough to converge. The full models are affected too: at 8 epochs ResNet-18 and ViT-S-Full were still improving by 0.9 and 1.9 points per epoch, and extending their budgets raises clean accuracy from 97.13% to 99.23% and from 91.47% to 96.99%. The corruption columns move accordingly, and the LoRA-VPT comparison changes qualitatively: within two points of each other under every corruption before tuning, LoRA ahead by 23 points under noise and 10 under occlusion after.

E. Composite-Input Ablation

To locate the adjacency signal, ResNet-18 and ViT-S-Full were retrained on 3×3 with a fragments-only composite (strip quadrants gray) and a strips-only composite (fragment quadrants gray), and evaluated with the same masking (Table IV). On clean data fragments alone reach 98.89% and 98.97%, within 0.4 points of the full composite, and strips alone still reach about 96%, so the two channels are largely redundant when the input is intact. Under erosion they separate. The strips-only models fall to about 49% at 1 px and to exactly 8.33% at 8 px, the share of a single orthogonal class, because

TABLE III

TRAINING-PROTOCOL SENSITIVITY ON THE 4×4 GRID. THE SAME ARCHITECTURES EVALUATED UNDER THE IDENTICAL FINAL PROTOCOL (TEST SET, FRAGMENT-LEVEL CORRUPTIONS), TRAINED EITHER WITH THE SHARED FULL FINE-TUNING CONFIGURATION (LR 10^{-4} , 8 EPOCHS) OR WITH PER-METHOD LEARNING RATES AND CONVERGENCE-BASED BUDGETS (LR, EPOCHS). THE APPARENT PEFT COLLAPSE AND MUCH OF THE GRID-DIFFICULTY GAP ARE ARTIFACTS OF THE SHARED CONFIGURATION.

Model	Configuration	Clean	$\sigma=30$	4 px	30%
ResNet-18	shared, 8 ep	97.13	8.78	12.55	72.16
	tuned (10^{-4} , 16 ep)	99.23	8.98	16.09	85.98
ViT-S-Full	shared, 8 ep	91.47	56.62	12.74	86.81
	tuned (10^{-4} , 24 ep)	96.99	64.08	14.70	93.60
ViT-S-LoRA	shared, 8 ep	45.19	25.69	9.36	40.62
	tuned (10^{-3} , 24 ep)	92.46	53.37	13.44	86.74
ViT-S-VPT-24	shared, 8 ep	45.56	23.91	5.88	39.58
	tuned (10^{-3} , 32 ep)	79.38	30.11	9.31	76.09

their only information source has been erased and they collapse to a constant prediction. The fragments-only models are more erosion-robust than the full composite at almost every level; the clearest case is ViT-S-Full, which keeps 55.42% at 8 px without strips against 10.21% with them, and ResNet-18 keeps 67.41% against 47.85%. Under occlusion all variants stay within about 1.5 points of each other. The strip channel is therefore nearly free on clean data but actively amplifies erosion damage.

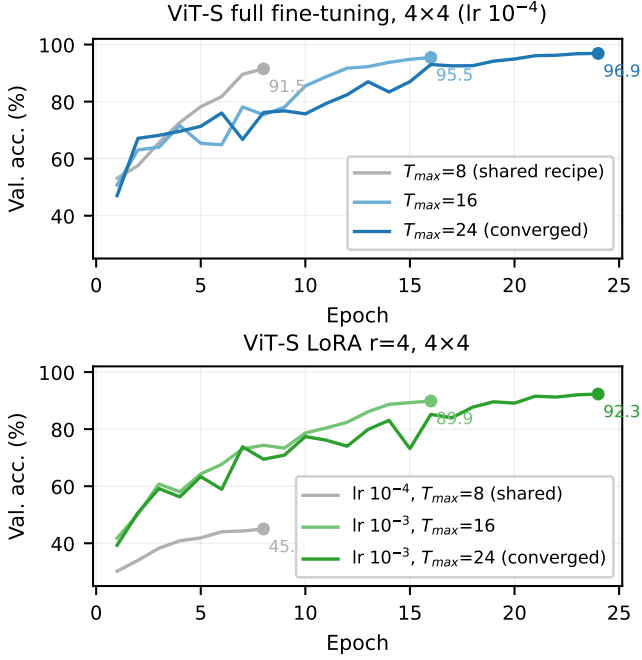


Fig. 2. Validation accuracy per epoch on the 4x4 grid for ViT-S full fine-tuning (top) and ViT-S-LoRA (bottom) under the shared configuration and under progressively longer cosine budgets; markers give each run’s final accuracy. Both were far from converged at 8 epochs, and LoRA additionally required a larger learning rate.

TABLE IV
COMPOSITE-INPUT ABLATION ON THE 3x3 GRID (TEST SET).
DEGRADATION IS APPLIED TO FRAGMENTS BEFORE COMPOSITE
CONSTRUCTION AND MASKING, MATCHING EACH VARIANT’S TRAINING
PIPELINE. STRIPS-ONLY MODELS COLLAPSE TO THE
CONSTANT-PREDICTION FLOOR OF 8.33% (THE SHARE OF EACH
ORTHOGONAL-DIRECTION CLASS IN THE FULL PAIR ENUMERATION).

Model	Input	Clean	Erosion (px)				Occlusion		
			1	2	4	8	10%	20%	30%
ResNet-18	Full	99.31	86.06	79.72	68.45	47.85	98.61	96.87	93.54
	Fragments	98.89	93.21	76.35	69.58	67.41	97.93	96.22	92.13
	Strips	95.82	48.42	36.71	10.04	8.33	95.15	94.40	93.16
ViT-S-Full	Full	99.11	71.39	60.37	48.93	10.21	98.66	98.15	96.95
	Fragments	98.97	76.32	67.97	65.41	55.42	98.58	97.95	96.73
	Strips	96.23	50.33	46.65	24.01	8.33	95.54	94.79	93.55

V. DISCUSSION

A. Degradation Dependent Robustness

The noise and occlusion results are consistent with the account offered for object classification [15], [17], in which global attention pools evidence from every patch while a convolutional decision built from local windows degrades in proportion to the damaged area.

Erosion reverses the ranking through two distinguishable factors. The first is the input design. An 8-pixel erosion on a 59-pixel 3x3 fragment edge covers 13.6% of the fragment, which after resizing to 112 pixels spans about 15 pixels, wider than the 10 pixel strip. Consequently, depending on the construction, the quadrant containing the strip turns uniformly

gray, misleading models that had learned to assign weight to these regions, rather than merely causing a loss of partial information. The ablation quantifies this: the same ViT backbone keeps 55.42% at 8 px without strips and 10.21% with them. The second factor relates to architecture, though confounding variables exist: when using only image patches as input, ResNet-18 consistently outperforms ViT across various erosion levels (e.g., 93.21% vs. 76.32% at 1-pixel erosion). Given the differences between the two backbones regarding pre-training data, classification heads, and parameter counts—and the 1–8 pixel erosion widths are obvious smaller than the 16-pixel patch size, this residual performance gap should be due to the specific CNN-ViT combinations and their tokenization methods, rather than to a universal characteristic of the Transformer architecture itself.

B. What PEFT Retains and How It Fails

After per-method tuning, LoRA retains most of the backbone’s robustness: on 3x3 it keeps 86% of the full model’s accuracy at $\sigma=30$ and 98% under 30% occlusion, and it exceeds the full model under deep erosion. The residual 4x4 gap has a specific shape. The confusion matrices show that adjacency classes are recognised with recall above 97% by every model, and that almost all PEFT errors are non-adjacent pairs assigned to one of the four diagonal classes: of 46,664 non-adjacent test pairs, VPT-24 assigns about 3,300 to each diagonal class, LoRA about 1,200, ViT-S-Full about 400, and ResNet-18 fewer than 140. Diagonal neighbours share only a corner, so the boundary strips carry almost no continuity evidence and the decision rests on content priors, which is exactly where a frozen backbone is weakest; the inverse-frequency class weights further bias predictions away from the non-adjacent majority. Balanced accuracy (Table I) separates this concentrated confusion from a uniform loss of competence. VPT is also the slowest to converge, needing 32 epochs against 24 for LoRA, which is consistent with adaptation through prompts acting on attention only indirectly.

C. Protocol Artifacts

First, a learning rate suitable for full-parameter fine-tuning causes models with fewer than 1% of the total parameters to suffer from under-training by an order of magnitude, the apparent PEFT collapse on 4x4 (45%) is entirely an optimisation effect. Second, an 8-epoch budget under-trains every model on the finer grid, including the full models (Table III), and for ResNet-18 the 4x4 difficulty gap vanishes once converged. The most consequential artifact is qualitative: under the shared recipe, budget-matched LoRA and VPT displayed different performances: VPT excels in the presence of noise, whereas LoRA performs better against image erosion. However, after thorough fine-tuning, VPT maintains an advantage only under conditions of slight noise, while LoRA holds the upper hand in almost all scenarios.

D. Implications

The design of the backbone network and input should be based on the expected type of damage: a CNN taking

only fragments as input is appropriate when edge damage dominates, whereas a Transformer is preferable when noise or occlusion is the primary issue; in scenarios where erosion may occur, the use of a "strip channel" should be avoided as this eliminates a potential failure mode at negligible cost on clean data. Fine-tuned LoRA represents a reliable and parameter-efficient solution; with a finer grid setting, its performance lags behind full fine-tuning by only about 4.5 points and outperforms VPT under an equivalent budget. Methodologically, when comparing PEFT techniques, distinct learning rates should be set for each method, model convergence must be verified prior to assessing capacity or robustness, and class-balanced metrics should be reported in cases of class imbalance.

E. Limitations

Every configuration is trained once, so differences below about two points, including several 3×3 LoRA–VPT cells, are not interpreted. All images are aligned faces; the fragments-only result (98.9% without any boundary information) shows that content position priors carry much of the signal, so absolute accuracies will be lower on unaligned domains and the redundancy finding needs validation there. Splits are image-disjoint but not identity-disjoint, so a person may appear in both training and test images. The learning-rate sweep covered the PEFT methods only. One CNN and one ViT were compared, with the confounds noted in Section III-C.

VI. CONCLUSION

This paper compared full fine-tuning, LoRA, and VPT on a ViT-Small backbone, with a ResNet-18 reference, for pairwise image-fragment relation classification under fragment-level noise, erosion, and occlusion, on a source-disjoint test set and under a protocol with per-method learning rates and convergence-based budgets. Robustness is degradation dependent: Transformer models degrade gracefully under noise and occlusion while the CNN is far more robust to erosion, and most of the Transformer's erosion collapse is caused by the boundary-strip input channel. Tuned LoRA is within 4.5 points of full fine-tuning at 0.5% trainable parameters, retains most of its robustness, and outperforms budget-matched VPT under nearly all conditions. Future work should add multiple seeds, unaligned domains, non-uniform damage, and evaluation through a full reassembly pipeline.

REFERENCES

- [1] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
- [3] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.
- [4] M. Jia *et al.*, "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 709–727.
- [5] Y. Xin *et al.*, "Parameter-efficient fine-tuning for pre-trained vision models: A survey and benchmark," *arXiv preprint arXiv:2402.02242*, 2024.
- [6] J. Liu, W. Teshome, S. Ghimire, M. Sznaiier, and O. Camps, "Solving masked jigsaw puzzles with diffusion vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 23009–23018.
- [7] Z. Xu and X. Liu, "VLHSA: Vision-language hierarchical semantic alignment for jigsaw puzzle solving with eroded gaps," *arXiv preprint arXiv:2509.25202*, 2025.
- [8] X. Song *et al.*, "ERL-MPP: Evolutionary reinforcement learning with multi-head puzzle perception for solving large-scale jigsaw puzzles of eroded gaps," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 7, 2025, pp. 6968–6977.
- [9] G. Heck, N. Lermé, and S. Le Hégarat-Masclé, "Solving jigsaw puzzles with vision transformers," *Pattern Anal. Appl.*, vol. 28, no. 2, p. 110, 2025.
- [10] G. Kim, H. Cho, and H. Nam, "Solving jigsaw puzzles by predicting fragment's coordinate based on vision transformer," *Expert Syst. Appl.*, vol. 272, p. 126776, 2025.
- [11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2980–2988.
- [12] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 3730–3738.
- [13] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.
- [14] T. S. Cho, S. Avidan, and W. T. Freeman, "A probabilistic image jigsaw puzzle solver," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2010, pp. 183–190.
- [15] M. Naseer *et al.*, "Intriguing properties of vision transformers," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 23296–23308.
- [16] S. Bhojanapalli *et al.*, "Understanding robustness of transformers for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10231–10241.
- [17] S. Paul and P.-Y. Chen, "Vision transformers are robust learners," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 2, 2022, pp. 2071–2081.
- [18] M.-M. Paumard, D. Picard, and H. Tabia, "Deepzle: Solving visual jigsaw puzzles with deep learning and shortest path optimization," *IEEE Trans. Image Process.*, vol. 29, pp. 3569–3581, 2020.
- [19] W. Yan, *Robotic Vision: Robotic Vision: From Deep Learning to Autonomous Systems*. Springer, 2026.
- [20] W. Yan, *Computational Methods for Deep Learning: Theory, Algorithms, and Implementations*. Springer, 2023.
- [21] W. Yan, *Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics*. Springer, 2019.
- [22] N. Gowdra, R. Sinha, S. MacDonell, W. Yan, "Maximum Categorical Cross Entropy (MCCE): A noise-robust alternative loss function to mitigate racial bias in Convolutional Neural Networks (CNNs) by reducing overfitting," *Pattern Recognition*, 2021.
- [23] B. Xiao, M. Nguyen, W. Yan, "Fruit ripeness identification using transformers," *Applied Intelligence*, 2023.
- [24] B. Xiao, M. Nguyen, W. Yan, "Apple ripeness identification from digital images using transformer," *Multimedia Tools and Applications*, 2023.
- [25] Y. Liu, P. Nand, M. Hossain, M. Nguyen, W. Yan, "Sign language recognition from digital videos using feature pyramid network with Detection Transformer," *Multimedia Tools and Applications*, 2023.
- [26] J. Qi, M. Nguyen, W. Yan, "Small visual object detection in smart waste classification using Transformers with deep learning," *IVCNZ*, 2022.
- [27] Z. Luo, M. Nguyen, W. Yan, "Kayak and sailboat detection based on the improved YOLO with Transformer," *ACM ICCV* 2022.
- [28] M. Al-Sarayreh, M. Reis, W. Yan, R. Klette, "A sequential CNN approach for foreign object detection in hyperspectral images," *Pacific Rim Conference on Multimedia*, 2019.
- [29] X. Wang, W. Yan, "Gait recognition using multichannel convolutional neural networks," *Neural Computing and Applications*, 2019.
- [30] R. Le, M. Nguyen, W. Yan, "Training a convolutional neural network for transportation sign detection using synthetic dataset," *IVCNZ*, 2021.
- [31] M. Nguyen, W. Yan, "Temporal colour-coded facial-expression recognition using convolutional neural network. International Summit Smart City 360°: Science and Technologies for Smart Cities," 2022.
- [32] D. Shen, C. Xin, M. Nguyen, W. Yan, "Flame detection using deep learning," *IEEE International Conference on Control, Automation and Robotics (ICCAR)*, 2018.