HCMUS_TheFangs at NakbaArchiveClassifier Shared Task: Foundation Models and Advanced Training Strategies for Conflict Damage Classification

Dao Sy Duy Minh* **Huynh Trung Kiet***
Tran Chi Nguyen **Nguyen Lam Phu Quy** **Pham Phu Hoa**
*Joint first authors
Faculty of Information Technology, University of Science,
Vietnam National University Ho Chi Minh City
Ho Chi Minh City, Vietnam
{23122041, 23122039, 23122044, 23122048, 23122030}@student.hcmus.edu.vn

Abstract

We present our system for the NakbaArchiveClassifier shared task at Nakba-NLP 2026, which requires classifying Instagram images from Gaza as showing destroyed or damaged infrastructure versus intact surroundings. Working with a small, imbalanced dataset (1,400 training images; 1.83:1 class ratio), we conduct a systematic empirical study of six model-training combinations spanning five architecture families: standard CNNs (EfficientNet-B4), self-supervised ViTs (DINOv2-ViT-L), hybrid multi-axis Transformers (MaxViT-Base), masked-image-modelling ViTs (EVA-02-Base), and large-kernel CNNs (UniRepLKNet). For our best performing configuration–MaxViT-Base with focal loss, MixUp, and a rich geometric augmentation pipeline–we provide a detailed component analysis. Our system achieves a macro F1 of **0.899** on the public test set, ranking **1st** on the competition leaderboard. We additionally report findings from novel experiments including a Kolmogorov-Arnold Network (KAN) classification head and VLM-regularized training with BLIP-2-generated captions, offering insights into what does and does not transfer to conflict-domain imagery under severe data scarcity.

Keywords: image classification, conflict damage, foundation models, MaxViT, focal loss, KAN, data augmentation, class imbalance, Nakba

1. Introduction

The SaltPillar archive (Tech for Palestine) is collecting millions of Gaza conflict images from Instagram to form a searchable historical record. The NakbaArchiveClassifier shared task (Abrahams et al., 2026) formalises this as binary classification-destruction vs. not_destruction-evaluated by macro F1. Three properties make it non-trivial: only 1,400 labeled training images (1.83:1 class imbalance), visually distinctive domain content (rubble microstructure, JPEG compression artefacts absent from ImageNet), and a **photographer-bias** risk: 118 unique content creators appear in both train and validation splits, so a powerful model can maximise validation F1 by memorising author-specific style rather than damage semantics. Our system, HCMUS_TheFangs, addresses these jointly through backbone selection, focal-loss training, and a bias-targeted augmentation cascade, achieving macro F1 = **0.899** (1st place). We additionally report the first application of a KAN (Liu et al., 2024) classification head and VLM-regularised training (Li et al., 2023) to conflict imagery, with frank failure analyses of both.

2. Task and Dataset

The task is binary classification of Instagram images from Gaza (Abrahams et al., 2026): predict destruction vs. not_destruction, evaluated by macro F1. Table 1 shows the split statistics. The 1.83:1 imbalance is perfectly consistent across all three splits. A key EDA finding: **118 photographers** appear in both train and validation, creating a shortcut risk that motivates the photographer-bias defences in Section 3.

Split	Total	destruction	not_destr.
Train	1,400	494 (35.3%)	906 (64.7%)
Val	199	70 (35.2%)	129 (64.8%)
Test	402	141 (35.1%)	261 (64.9%)

Table 1: Class distribution across splits (Abrahams et al., 2026). The 1.83:1 imbalance is perfectly consistent, ruling out split-specific artifacts.

3. Methods

Our system architecture is shown in Figure 1. A shared training protocol is applied across all backbone variants to ensure fair comparison; differences in the results table can therefore be attributed to architecture and loss function choices alone.

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 121–127

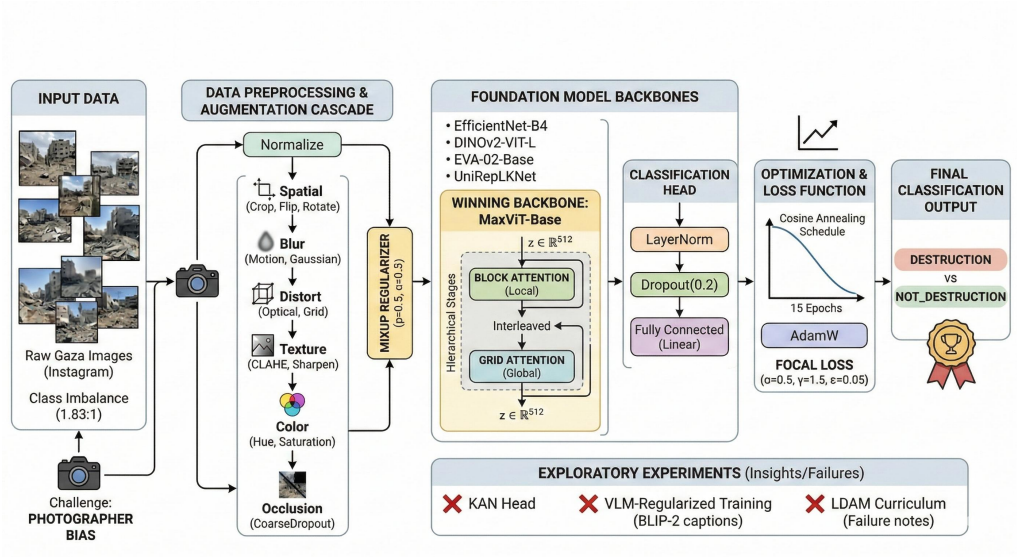

Figure 1: Full training pipeline of our best system (EXP009). The top row shows the end-to-end flow during training: each input image passes through a six-step augmentation cascade, is stochastically mixed with a second image via MixUp ($p = 0.5$), forwarded through the MaxViT-Base backbone, projected by a lightweight classification head, and trained with Focal Loss. At inference, the MixUp branch is skipped and the best-checkpoint model (selected on validation macro F1) is used. The bottom-left panel details each augmentation step; the bottom-right panel shows the four-stage MaxViT architecture, where each stage interleaves an MBConv block with local Block Attention and global Grid Attention before progressive spatial downsampling.

3.1. Shared Training Protocol

To ensure fair backbone comparison, we fix a shared training protocol across all six experiments: AdamW optimisation with a cosine-annealing learning-rate schedule, 15 training epochs with best-checkpoint selection on validation macro F1, and ImageNet-standard normalisation. Full hyper-parameter values-learning rates, batch sizes, weight decay, normalisation constants, and hardware details-are reported in Appendix A.

3.2. Backbone Architectures

We benchmark five architecture families spanning the modern representation landscape. At one end sits the supervised CNN baseline, **EfficientNet-B4** (Tan and Le, 2019), which provides a solid starting point but is limited by ImageNet-only pretraining. At the other extreme, **DINOv2-ViT-L/14** (Oquab et al., 2024) brings the largest capacity (307M parameters) and self-supervised features, but its sheer size proves a liability on our small dataset. Between these poles, we test three architectures that trade off inductive biases in different ways: **EVA-02-Base** (Fang et al., 2023), which reconstructs CLIP (Radford et al., 2021) features via masked image modelling; **UniRepLKNet** (Ding et al., 2024), which replaces attention entirely with large-kernel convolutions up to 31×31; and our winning backbone, **MaxViT-Base** (Tu et al., 2022), detailed in Section 3.3. Table 2 summarises the results; the discussion that follows focuses on the insights these comparisons reveal.

3.3. Winning Model in Detail: MaxViT-Base with Focal Loss

MaxViT (Tu et al., 2022) interleaves **Block Attention** (local 8×8 windows) and **Grid Attention** (dilated global sampling) at $\mathcal{O}(HW)$ cost per stage, giving simultaneous fine-grained texture sensitivity and global scene context-exactly what destruction classification requires. Four hierarchical stages (64/128/256/512 ch) are detailed in Figure 1. The backbone output $\mathbf{z} \in \mathbb{R}^{512}$ feeds a slim head:

$$h(\mathbf{z}) = \mathrm{Linear}_{512\to2}\big(\mathrm{Dropout}(0.2)\big(\mathrm{LayerNorm}(\mathbf{z})\big)\big). \quad (1)$$

The augmentation cascade (Figure 1, left panel) is designed in three tiers. Standard geometric/blur transforms (groups 1–2) cover camera diversity. Optical/grid distortions and texture randomisation (groups 3–4, each $p = 0.3$) are our key photographer-bias defences: non-linear spatial warping destroys low-frequency stylistic cues (vignetting, aspect-ratio preferences) while texture jitter breaks post-processing fingerprints. Colour jitter and CoarseDropout (groups 5–6) handle illumination and occlusion. Batch-level `MixUp` (Zhang et al., 2018) ($p = 0.5$, $\lambda \sim \mathrm{Beta}(0.2, 0.2)$) is the primary capacity regulariser.

Focal Loss (Lin et al., 2017) with $\alpha = 0.5$, $\gamma = 1.5$, $\epsilon_{\text{smooth}} = 0.05$ handles the 1.83:1 imbalance:

$$\mathcal{L}_{\text{focal}}(\hat{p}, y) = -\alpha\,(1 - \hat{p}_y)^{\gamma}\,\log \hat{p}_y. \qquad (2)$$

When MixUp is active: $\mathcal{L} = \lambda\,\mathcal{L}_{\text{focal}}(\cdot, y_a) + (1-\lambda)\,\mathcal{L}_{\text{focal}}(\cdot, y_b)$. Optimiser: AdamW (Loshchilov and Hutter, 2019), LR 2×10^{-5}, CosineAnnealing ($T_{\max} = 15$).

3.4. Additional Experiments

Beyond backbone selection, we probe three research questions. Can *language supervision* regularise visual features by aligning them with BLIP-2 (Li et al., 2023) captions? Can a *curriculum-based margin loss* (Cao et al., 2019) handle the 1.83:1 imbalance better than focal loss? And can a *KAN* (Liu et al., 2024) classification head-which replaces fixed activations with learnable B-spline functions-outperform a standard MLP? We additionally explore domain adversarial training (Ganin et al., 2016) to penalise photographer-identity prediction. Full configurations are provided in Appendix A.

4. Experiments and Results

4.1. Implementation Details

All experiments use PyTorch (Paszke et al., 2019) with `timm` (Wightman, 2019) and `albumentations` (Buslaev et al., 2020), following the shared protocol of Section 3. Full hyperparameters and per-experiment configurations are in Appendix A. Code: `https://github.com/hcmus-thefangs/nakba-classifier`.

4.2. Main Results

Table 2 presents results across all configurations. Our MaxViT-Base system achieves macro F1 = 0.899 on the public test set, securing 1st place on the leaderboard. Two immediate patterns stand out. First, the moderately-sized backbones-EVA-02 and UniRepLKNet-perform competitively (0.884 and 0.886 respectively), demonstrating that neither masked-image-modelling pretraining nor attention mechanisms hold a monopoly on strong features for this domain. Second, and more revealingly, the DINOv2-based experiments exhibit the largest gaps between validation and test performance, despite being the highest-capacity models tested. This divergence is the clearest fingerprint of the photographer-bias problem described in Section 2: these large models memorise stylistic signatures of individual content creators rather than learning generalisable damage semantics. The most dramatic case is the KAN classification head, which achieves a strong validation F1 of 0.865 but collapses to 0.552 on the test set-a failure we analyse in detail in Section 4.5.

EXP	Model / Strategy	Val F1	Test F1
001	EfficientNet-B4 + Focal Loss	0.829	0.870
006	DINOv2-ViT-L + VLM Regularisation	0.859	0.804
007	DINOv2-ViT-L + LDAM Curriculum	0.747	0.709
008	DINOv2-ViT-B + KAN Head	0.865	0.552
010	EVA-02-Base (MIM ViT)	0.885	0.884
015	UniRepLKNet + Focal Loss	0.881	0.886
009	**MaxViT-Base + Focal Loss**	**0.881**	**0.899**

Table 2: Macro F1 on validation and public test sets. EXP009 achieved 1st place on the leaderboard.

4.3. Ablation Study

Table 3 isolates each component of EXP009 by changing one element at a time. Loss function choice is the most consequential variable: swapping focal loss for LDAM/DRW costs 13.4 F1 points, confirming that the moderate 1.83:1 imbalance sits squarely in focal loss's operating regime, while LDAM's large margin terms destabilise training at this data scale. Within augmentation, MixUp delivers the largest single gain (+1.5 points), functioning as the primary batch-level regulariser. The architecture swap from MaxViT-Base to EfficientNet-B4 costs 5.2 points-the single largest contributor-confirming that backbone choice dominates all other decisions.

Configuration	Val F1	Δ
Full EXP009 (reference)	0.881	-
Loss function:		
Focal $\rightarrow$ CrossEntropy	0.862	-0.019
Focal $\rightarrow$ LDAM+DRW (3-stage)	0.747	-0.134
Augmentation:		
No geometric distortions	0.871	-0.010
No texture transforms	0.868	-0.013
No CoarseDropout	0.874	-0.007
No MixUp	0.866	-0.015
Baseline only (Flip + Rotate)	0.852	-0.029
Backbone (same training config):		
EfficientNet-B4	0.829	-0.052
EVA-02-Base	0.885	$+0.004$
UniRepLKNet	0.881	±0.000

Table 3: Component ablation of EXP009. Each row modifies one element; all other settings are held fixed. Δ = change in Val macro F1 vs. the full configuration.

4.4. Leaderboard

Table 4 shows the top-5 of the final public leaderboard. Our system leads by 0.4 F1 points over the

second-ranked team; the top-5 span just 3.3 points, indicating a genuinely competitive field.

#	Team	F1	Acc.
1	**HCMUS_TheFangs**	**0.899**	**0.908**
2	rahaf_jaber	0.895	0.905
3	zahira_blr	0.889	0.901
4	salmakh	0.877	0.886
5	mohamedfathy	0.866	0.878

Table 4: Top-5 public leaderboard (macro F1). Full standings at the shared task website.

4.5. Analysis and Discussion

Destruction recognition is fundamentally a multi-scale problem-a model must detect local texture cues (cracked concrete, rubble geometry) while understanding global scene coherence-and *architectural fit* proves more important than raw capacity in addressing it. MaxViT's interleaved Block and Grid attention captures both neighbourhood detail and long-range layout at linear cost, which is why it outperforms DINOv2-ViT-L despite the latter having $2.5\times$ more parameters. The validation-to-test gap makes this concrete: DINOv2 drops 5.5 F1 points (memorising photographer shortcuts instead of damage semantics), while MaxViT actually *improves* by 1.8 points-resilience we attribute to our bias-targeted augmentation cascade that strips away low-frequency stylistic cues before they can be memorised.

Our exploratory experiments reinforce this theme. The KAN classification head (Liu et al., 2024) shows genuine promise when the backbone is frozen, demonstrating that learnable B-spline activations can capture non-monotonic decision boundaries useful for damage features. However, once we unfreeze the backbone for joint fine-tuning, the system collapses due to a gradient-scale mismatch: backbone weights need learning rates around $\sim 10^{-5}$ while B-spline grid parameters need $\sim 10^{-3}$, and a single shared rate causes catastrophic spline deformation. This is not a flaw of KAN itself but a practical engineering gap that per-parameter-group optimisation would likely resolve.

VLM regularisation via BLIP-2 (Li et al., 2023) captions exposes a more fundamental limitation. The captioner, trained on web-scale data, produces generic descriptions ("a photo of a building") that lack conflict-domain vocabulary, so InfoNCE (van den Oord et al., 2018) alignment anchors features to imprecise semantics and actively hurts generalisation. The LDAM (Cao et al., 2019) curriculum similarly fails through gradient oscillations at its stage transitions, producing the lowest score among all configurations. Both failures, together with the KAN collapse, share a root cause: techniques designed for large-scale, domain-general benchmarks can silently break when applied to data that is small, culturally specific, and visually unlike their pretraining distribution.

5. Conclusion

We presented HCMUS_TheFangs, the top-ranked system (macro F1 = 0.899) for the NakbaArchive-Classifier task. Evaluating five architectures on just 1,400 images reveals a critical insight: the binding constraint in conflict-domain vision is *domain fit*, not model capacity. MaxViT's (Tu et al., 2022) multi-scale attention elegantly captures both local rubble textures and global structural collapse, while avoiding the photographer-bias trap that ensnares larger models. Furthermore, bias-targeted augmentations prove as influential as backbone choice itself, underscoring the need for domain-specific regularisation.

Our exploratory failures are equally telling. KAN (Liu et al., 2024) heads show promise but require nuanced optimisation, while VLM regularisation falters because generic captioners lack conflict-domain vocabulary. These limitations highlight the urgent need for specialised humanitarian vision-language models. Looking ahead, the SaltPillar archive's scale and temporal depth offer rich opportunities for semi-supervised learning and change-detection. Ultimately, bridging the gap between web-scale pretraining priors and conflict-domain reality demands evaluation protocols that explicitly penalise style overfitting.

6. Ethics Statement

This work uses images depicting conflict and destruction in Gaza. All images were collected and shared by the task organizers under appropriate agreements. We recognize that automated classification of conflict imagery carries inherent ethical weight: misclassification (both false positives and false negatives) can have real consequences for archiving, historical record, and humanitarian decision-making. We caution against deployment without human oversight. Our code and models are released for research purposes only.

7. Acknowledgements

We thank the Nakba-NLP 2026 shared task organizers for their efforts in curating this important dataset and hosting the competition. We also acknowledge the contributions of the broader Palestinian digital archive community in preserving this visual record.

8. Bibliographical References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The nakbaarchive-classifier shared task on nakba image classification. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexey A. Kalinin. 2020. Albumentations: Fast and flexible image augmentations. *Information*, 11(2):125.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, Xiangyu Yue, and Ying Shan. 2024. Unireplknet: A universal perception large-kernel convnet for audio, video, point cloud, time-series and image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5513–5524.

Yuxin Fang, Jiaqi Sun, Jianian Wang, Xiaosong Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva-02: A visual representation-level-mimicri transformer for image recognition. *arXiv preprint arXiv:2303.11331*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. volume 17, pages 2096–2030.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. 2024. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Vasiljevic, Pauline Sun, Peter Schwartz, Akhil Batra, Vincent Brandao, et al. 2024. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Jian Sun, Laurent Schampers, Ang Li, Christian Szegedy, Caroline Pantofaru, Golnaz Ghiasi, et al. 2022. Maxvit: Multi-axis vision transformer. *arXiv preprint arXiv:2204.01697*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. In *arXiv preprint arXiv:1807.03748*.

Ross Wightman. 2019. Pytorch image models. https://github.com/rwightman/pytorch-image-models.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization strategy to train strong classifiers with localizable features.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8024–8033.

Hongyi Zhang, Moustapha Cai, Yann Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

9. Language Resource References

A. Implementation Details

All experiments were conducted on a single NVIDIA P100 GPU (16 GB) via Kaggle. We used PyTorch 2.0 (Paszke et al., 2019) with `timm` (Wightman, 2019) for backbone loading and `albumentations` (Buslaev et al., 2020) for augmentation. All random seeds are fixed to 42 across `numpy`, `torch`, and `cudnn` (deterministic mode). Code and full configurations are publicly available at `https://github.com/technoob05/hcmus-thefangs-nakba-classifier`.

A.1. Shared Training Protocol

The shared protocol uses AdamW (Loshchilov and Hutter, 2019) ($\beta_1 = 0.9$, $\beta_2 = 0.999$, wd$= 10^{-2}$) with cosine-annealing LR schedule ($\eta_{\min} = 10^{-6}$, linear warm-up over epoch 1), 15 training epochs, and best-checkpoint selection on validation macro F1. Inputs are normalised with ImageNet (Deng et al., 2009) statistics (mean $[0.485, 0.456, 0.406]$, std $[0.229, 0.224, 0.225]$).

A.2. Per-Experiment Hyper-Parameters

Table 5 summarises backbone-specific settings. All other hyper-parameters follow the shared protocol above.

EXP	Backbone	BS	LR	Img	Notes
001	EfficientNet-B4	16	10^{-3}	384	Focal (α=0.5, γ=1.0)
006	DINOv2-ViT-L	8	10^{-4}	518	Frozen 10 ep; +InfoNCE
007	DINOv2-ViT-L	8	10^{-4}	518	Frozen 10 ep; LDAM+DRW
008	DINOv2-ViT-B	8	10^{-4}	518	KAN head (grid=5)
009	MaxViT-Base	8	2×10^{-5}	384	Focal+MixUp (p=0.5, α=0.2)
010	EVA-02-Base	4	10^{-5}	448	MixUp+CutMix (Yun et al., 2019)
015	UniRepLKNet	8	5×10^{-5}	384	Label smoothing 0.1

Table 5: Per-experiment hyper-parameters. BS=batch size, LR=peak learning rate, Img=input resolution.

A.3. Augmentation Cascade (EXP009)

The six-step `albumentations` (Buslaev et al., 2020) cascade: (1) RandomResizedCrop (scale 0.8–1.0) + HFlip + VFlip + Rotate $\pm15°$; (2) OneOf{MotionBlur, GaussianBlur} $p = 0.3$; (3) OneOf{OpticalDistortion, GridDistortion} $p = 0.3$; (4) OneOf{CLAHE, Sharpen, Emboss} $p = 0.3$; (5) HueSaturationValue + RandomBrightnessContrast $p = 0.3$; (6) CoarseDropout (8 holes, 32×32) $p = 0.3$. All other experiments use only flips, rotation, and colour jitter.

A.4. VLM-Regularised Training (EXP006)

EXP006 adds an auxiliary InfoNCE loss (van den Oord et al., 2018) ($\lambda = 0.3$, $\tau = 0.07$) that aligns DINOv2 visual features with text encodings produced by BLIP-2 (Li et al., 2023) and CLIP (Radford et al., 2021). Synthetic captions are generated for each training image; the contrastive objective pulls image representations toward their corresponding captions and pushes away mismatched pairs. The total loss is $\mathcal{L} = \mathcal{L}_{\text{CE}} + 0.3 \cdot \mathcal{L}_{\text{InfoNCE}}$. The intent is to suppress photographer-style memorisation by grounding features in semantic descriptions rather than visual style.

A.5. LDAM Curriculum (EXP007)

EXP007 implements a three-stage loss curriculum for imbalanced learning (Cao et al., 2019). Stage 1 (epochs 1–5): standard cross-entropy to learn initial representations. Stage 2 (epochs 6–10): LDAM loss with class-dependent margins $m_i = C_{\max}/n_i^{1/4}$. Stage 3 (epochs 11–15): LDAM with deferred re-weighting (DRW), where per-class weights are inversely proportional to class frequency. The curriculum transitions aim to first build stable representations before enforcing margin constraints.

A.6. KAN Classification Head (EXP008)

EXP008 replaces the standard MLP head with a two-layer Kolmogorov-Arnold Network (Liu et al., 2024): $768 \to 512 \to 256 \to 2$, using cubic B-spline activations with grid size 5. Unlike standard MLPs which use fixed activation functions (ReLU), KAN layers learn the activation shape itself via spline parameters, enabling non-monotonic decision boundaries. Training proceeds in two phases: backbone frozen (10 epochs), then full fine-tuning (5 epochs).

A.7. Domain Adversarial Training (EXP005)

EXP005 adds a Gradient Reversal Layer (GRL) (Ganin et al., 2016) to the DINOv2 backbone. A photographer-identity discriminator is trained adversarially via the GRL, which negates discriminator gradients during backpropagation, forcing backbone features to become uninformative for author identity whilst remaining informative for damage prediction. The adversarial coefficient is annealed as $\lambda_p = 2/(1 + e^{-10p}) - 1$ where $p \in [0, 1]$

is fractional training progress. This experiment is exploratory and its result is not included in the main comparison table.