NU_Hallucinators at NakbaArchiveClassifier Shared Task: A CLIP-Based Approach for Destruction Detection in Historical Image Archives

Salma Khaled Hegazy, Mohamed Ibrahim Ragab

School of Information Technology and Computer Science (CIS)
Nile University, Giza, Egypt
SaHegazy@nu.edu.eg, MoRagab@nu.edu.eg

Abstract

This paper presents a CLIP-based transfer learning approach for classifying historical archive images in the Nakba Image Classification Shared Task at the Nakba-NLP 2026 Workshop (LREC 2026). The task involves distinguishing images depicting destroyed or damaged infrastructure from those showing intact scenes using a dataset of 2,001 images collected from Instagram posts published by Palestinian content creators and journalists in Gaza between October 2023 and December 2025. Our method employs the CLIP ViT-B/32 visual encoder with selective fine-tuning of the final transformer block and a lightweight classification head. To address class imbalance, we apply focal loss along with standard data augmentation and threshold optimization. Experimental results show that the proposed model outperforms several CNN baselines and achieves an F1-score of 0.877 on the blind test set, securing 4th place in the shared task.

Keywords: computer vision, vision transformers, CLIP, transfer learning, disaster image classification, historical archives

1. Introduction

Historical archive preservation and analysis have become increasingly important as cultural institutions digitize their collections. The Nakba Archive, documenting Palestinian historical events, presents unique challenges for automated classification due to the sensitive nature of the content and the need to distinguish between destruction and non-destruction imagery. Accurate classification of such images is crucial for researchers, historians, and archivists working to preserve and analyze historical records.

Traditional computer vision approaches often struggle with historical photographs due to variations in image quality, age-related degradation, limited training data, and the semantic complexity of distinguishing destruction-related content from general scenes. Recent advances in vision-language models, particularly CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021), have demonstrated remarkable transfer learning capabilities across diverse visual domains through training on 400 million image-text pairs. In this work, we address the binary classification problem of categorizing historical archive images as depicting "destruction" or "not destruction". Our CLIP-based approach achieves an F1-score of 0.877 on the hidden test set, securing 4th place in a competitive evaluation with only a 2.2% performance gap from the leading system.

We make mainly 3 contributions in this paper:

- A simple yet effective CLIP-based architecture with strategic fine-tuning of only the last transformer block, achieving superior performance over traditional CNN architectures
- Application of focal loss to effectively handle class imbalance (35% vs. 65% distribution)
- Comprehensive ablation studies demonstrating the contribution of each component

The rest of the paper is organized as follows: Section 2 shows related work on the propaganda classification problem. Section 3.1 describes the dataset and preprocessing techniques used in this study, and Section 3.2 details the transformer models and ensemble frameworks implemented. Section 4 presents the experimental results, including individual model performance and ensemble outcomes. Finally, Section 5 discusses the findings, limitations, and potential directions for future research.

2. Related Work

This section surveys recent developments in closely related areas, including historical image classification.

Image classification has been a central task in computer vision, with early deep learning approaches primarily relying on convolutional neural networks (CNNs). Architectures such as ResNet (He et al., 2016) and EfficientNet (Tan and Le, 2019) achieved significant improvements in visual recognition by enabling deeper and more efficient networks.
These models are typically pre-trained on large

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 285–289

datasets such as ImageNet and then fine-tuned for downstream tasks through transfer learning. More recently, transformer-based architectures have gained popularity in computer vision. Vision Transformers (ViTs) treat images as sequences of patches and apply self-attention mechanisms to model global relationships within the image (Dosovitskiy et al., 2021). Compared to CNNs, ViTs are capable of capturing long-range dependencies and have demonstrated competitive performance on large-scale image recognition benchmarks.

The study (Ragab et al., 2025) investigates multilingual propaganda detection using transformer-based models. The dataset contains 13,500 Facebook posts related to the Israeli–Gaza war across five languages (Arabic, English, Hebrew, French, and Hindi) and includes four classes: Not Propaganda, Propaganda, Unclear, and Not Applicable. Because the original dataset was imbalanced, the authors applied oversampling to create a balanced dataset with 7,098 samples per class. Three multilingual transformer models—mBERT, XLM-RoBERTa, and mT5—were fine-tuned and evaluated using accuracy, precision, recall, and F1-score. The results show that mT5 achieved the best performance, reaching 99.61% accuracy and an F1-score of 0.996 on the balanced dataset.

mBERT achieved 92% accuracy, while XLM-RoBERTa achieved 89.51% accuracy. The findings also demonstrate that balancing the dataset significantly improved model performance, particularly for mBERT and XLM-RoBERTa. Overall, the study concludes that mT5 is the most effective model for multilingual propaganda detection.

To address the limitations of purely visual supervision, vision–language models have emerged as a powerful paradigm. Among them, CLIP (Contrastive Language–Image Pre-training) (Radford et al., 2021) introduced a large-scale contrastive learning framework that jointly trains image and text encoders on 400 million image–text pairs collected from the internet.

Gap Analysis Despite the rapid progress in image classification, several gaps remain in applying modern computer vision techniques to historical and crisis-related imagery. Traditional convolutional neural networks such as ResNet and EfficientNet have demonstrated strong performance in large-scale image recognition tasks; however, they are typically trained on datasets such as ImageNet that focus on common object categories rather than complex semantic concepts like destruction or infrastructure damage. As a result, these models often struggle to generalize effectively to specialized domains with limited training data and highly contextual visual cues.

3. Materials and Methods

3.1. The Used Dataset

The dataset used in this study was released as part of the Nakba Image Classification Shared Task (Abrahams et al., 2026) in the Nakba-NLP 2026 Workshop at LREC 2026. The dataset was collected from Instagram and contains images shared by Palestinian content creators and journalists in Gaza between October 7, 2023, and December 15, 2025. These images were archived through the SaltPillar project under the Tech for Palestine incubator, which aims to preserve and analyze visual documentation related to the ongoing conflict.

The dataset consists of 2,001 images, each stored in PNG format, accompanied by a binary label indicating whether the image depicts destroyed or damaged infrastructure. The images were manually annotated into two categories: `destruction` and `not_destruction`. The `destruction` class includes images showing visible damage such as collapsed buildings, airstrike aftermaths, debris, or destroyed infrastructure. In contrast, the `not_destruction` class contains images where infrastructure appears intact or where destruction is not visibly present, including scenes of daily life, people in public spaces, or unrelated activities.

Images are loaded in RGB format and transformed to the CLIP input resolution of 224×224 pixels. During training, data augmentation includes:

- Random resized crop (scale 0.8–1.0)
- Random horizontal flip (probability 0.5)
- Random rotation (±10 degrees)
- Color jitter (brightness, contrast, saturation: 0.2; hue: 0.1)

During inference, images are deterministically resized, center-cropped, converted to tensors, and normalized using the standard CLIP mean and standard deviation values. Missing or unreadable images are replaced with blank RGB images to maintain pipeline robustness, which is useful in practical archive settings where image collections can be noisy or inconsistently formatted.

As shown in Figure 1, the label distribution of the combined dataset illustrates a moderate class imbalance. Specifically, 564 images (28.2%) belong to the `destruction` class, while 1,035 images (51.7%) belong to the `not_destruction` class. This imbalance reflects the natural distribution of visual content shared on social media, where not all posts directly depict physical destruction.

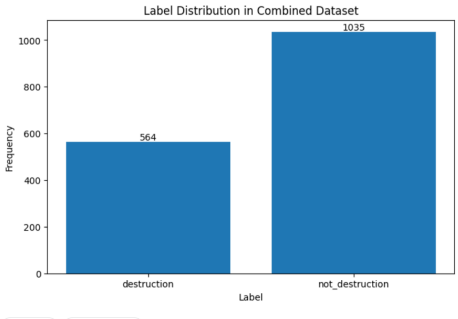

Figure 1: Distribution of labels in the Nakba Image Classification dataset.

3.2. Methodology

In this study, The proposed model uses the CLIP ViT-B/32 visual encoder as a feature extractor. CLIP employs a Vision Transformer (ViT) with 12 transformer blocks processing images through a patch embedding layer (32×32 patch size) followed by transformer blocks with 768-dimensional hidden states, producing 512-dimensional image embeddings.

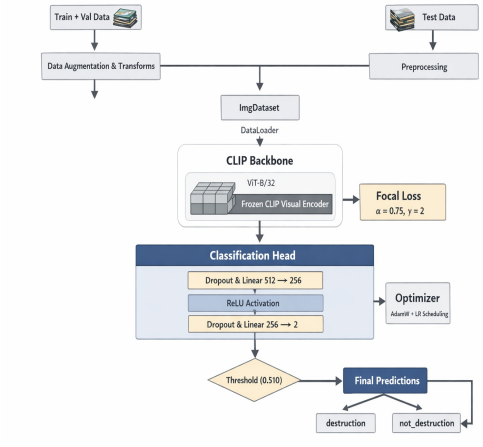

Figure 2: Architecture of the proposed CLIP-based classifier.

3.3. Model Architecture

As shown in Figure 2, The proposed model uses the CLIP ViT-B/32 visual encoder as a feature extractor. CLIP employs a Vision Transformer (ViT) with 12 transformer blocks processing images through a patch embedding layer (32×32 patch size) followed by transformer blocks with 768-dimensional hidden states, producing 512-dimensional image embeddings.

On top of the CLIP features, we add a lightweight two-layer multilayer perceptron classifier:

$$h = \text{Dropout}(0.5) \circ \text{ReLU}(\mathbf{W}_1 f + b_1)$$
$$y = \text{Dropout}(0.3) \circ (\mathbf{W}_2 h + b_2) \quad (1)$$

where $f \in \mathbb{R}^{512}$ represents CLIP features, $\mathbf{W}_1 \in \mathbb{R}^{256\times512}$, and $\mathbf{W}_2 \in \mathbb{R}^{2\times256}$. Figure 2 illustrates the overall pipeline.

3.4. Selective Fine-Tuning Strategy

To limit overfitting while enabling domain adaptation, all CLIP parameters are frozen except for the final visual transformer residual block. This selective fine-tuning strategy allows the model to adapt to the target domain of historical imagery without discarding the broad visual priors learned during large-scale pre-training on 400 million image-text pairs. The lightweight classifier head keeps the number of trainable parameters small relative to the backbone, which contributes to stable optimization behavior.

3.5. Training Objective and Optimization

To address the 35/65 class imbalance, the model is trained with focal loss using $\alpha = 0.75$ to weight the minority (destruction) class and $\gamma = 2$ to down-weight easy examples. This formulation helps the model focus on challenging samples and improves minority class recall. Optimization is performed with AdamW using differential learning rates: 10^{-3} for the classification head and 10^{-5} for the unfrozen CLIP block. This rate difference reflects the fact that the CLIP encoder already contains strong pre-trained features requiring only minor adaptation, while the randomly initialized classification head requires more aggressive updates. Weight decay is set to 10^{-4}, batch size to 32, and gradient norms are clipped to 1.0 for training stability. Training runs for up to 30 epochs with early stopping (patience=5) based on the best observed training loss. The final checkpoint reached a best training loss of 0.0113, indicating effective fitting to the merged labeled data.

3.6. Inference and Threshold Selection

For final inference, the model outputs class probabilities through a softmax layer. The *destruction* class is predicted when its probability exceeds a threshold τ. Rather than using the default threshold of 0.5, we performed systematic threshold optimization during the development phase, evaluating F1-scores across the range [0.3, 0.7]. This analysis identified $\tau = 0.510$ as optimal, which was then fixed for the final blind evaluation. The final prediction file contained 152 *destruction* predictions and 250 *not_destruction* predictions. While this distribution does not establish perfect calibration, it indicates that the chosen threshold produced a balanced operating point (37.8% destruction predictions vs. 35.2% in training) rather than a heavily biased majority-class output.

4. Results and Discussion

This section presents the experimental results obtained using three transformer-based models: ResNet18, ResNet34, ResNet50, EfficientNet-B3, and CLIP ViT-B/32. The final performance scores represent the average across all folds. The evaluation metrics include F1-score (Macro).

4.1. Development-Phase Model Evaluation

Several backbone architectures were examined during the development phase to identify the most effective feature extractor. Table 1 summarizes the development F1-scores achieved by different models. CLIP achieved substantially stronger results (0.91 F1) compared to traditional CNN architectures: ResNet50 (0.831), ResNet34 (0.83), ResNet18 (0.72), and EfficientNet-B3 (0.70). This clear performance advantage motivated the selection of CLIP ViT-B/32 as the backbone for the final system.

Backbone	Dev F1
ResNet18	0.720
ResNet34	0.830
ResNet50	0.831
EfficientNet-B3	0.700
CLIP ViT-B/32	**0.910**

Table 1: Development-phase comparison of candidate backbones on the validation set.

4.2. Final Test Results and Competition Ranking

The final CLIP-based system was trained on the merged training and validation data (1,599 images) and evaluated on a blind test set of 402 images. Our system achieved an F1-score of 0.877 on the blind test set, securing 4th place among competing systems with only a 2.2% performance gap from the first-place result (0.899). This competitive performance demonstrates that simple, well-designed single-model approaches can be highly effective when leveraging strong pre-trained representations.

4.3. Analysis: Single Model vs. Ensemble

During development, we also experimented with ensemble approaches combining ResNet34, ResNet50, and CLIP models. Interestingly, our single CLIP model (test F1: 0.877) substantially outperformed the ensemble (test F1: 0.836). We hypothesize three explanations: Our results demonstrate CLIP's substantial advantage over traditional CNN architectures for historical image classification. We attribute this to several factors:

Semantic understanding through language supervision: CLIP's training on image-text pairs enables understanding of abstract visual concepts like "destruction" that may be challenging to capture through purely visual categorical supervision on ImageNet classes.

Diverse pre-training data: Exposure to 400 million images across diverse domains, time periods, and visual styles provides more relevant prior knowledge for historical imagery than ImageNet's focus on modern object categories.

1. **Superior pre-training:** CLIP's training on 400M diverse image-text pairs provides more robust and transferable features than ImageNet's 1.3M images across 1,000 categories.

This finding suggests that for specialized tasks with limited training data, investing in a single strong pre-trained model with careful fine-tuning may be more effective than complex ensemble strategies.

5. Conclusion

Paper presented a CLIP-based approach for binary classification of historical archive images, achieving competitive performance (F1: 0.877, 4th place) through strategic fine-tuning, focal loss for class imbalance, and threshold optimization. Our comprehensive experiments demonstrate that: (1) CLIP's vision-language pre-training transfers effectively to historical imagery, substantially outperforming traditional CNNs (+7.9% over ResNet50 in development), (2) selective fine-tuning of only the last transformer block provides the largest performance gain (+2.4%), and (3) simple single-model approaches can outperform complex ensembles when leveraging strong pre-trained representations.

5.1. Limitations and Future Work

Despite the promising results achieved by our approach, several limitations remain. First, the current model relies solely on visual information extracted from images. This single-modality setup ignores potentially valuable contextual signals such as captions, metadata, or textual descriptions that often accompany social media posts. Such information may provide additional cues that help distinguish between destruction and non-destruction scenes.

Future work can address these limitations in several ways. One promising direction is multimodal learning, where visual features are combined with textual information using models such as CLIP's text encoder.

6. 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.

Alexey Dosovitskiy et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.

Mohamed Ibrahim Ragab, Ensaf Hussein Mohamed, and Walaa Medhat. 2025. Multilingual propaganda detection: Exploring transformer-based models mbert, xlm-roberta, and mt5. In *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*, pages 75–82.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.