DLRG@ NakbaArchiveClassifier Shared Task: Deep Transfer Learning for Destruction Detection in Nakba Archive Images Using EfficientNet-B3

Ramesh Kannan R, Ratnavel Rajalakshmi
School of Computer Science and Engineering,
Vellore Institute of Technology, Chennai,
India - 600127.
kanannrameshr@gmail.com, rajalakshmi.r@vit.ac.in

Abstract

Automatic identification of destruction in conflict-affected regions is an important task for humanitarian monitoring and historical documentation. Visual analysis of destruction scenes can assist researchers and policy makers in understanding the extent of damage in affected areas. This paper presents a deep learning-based image classification approach for identifying destruction and non-destruction scenes in Nakba-related images. The problem is formulated as binary image classification on Nakba images. A transfer learning approach using EfficientNet-B3 is adopted to learn discriminative visual features from Nakba images. Experimental evaluation shows that the proposed model achieved an Weighted F1-score of 83.87 % and an overall classification accuracy of 85.57 % and secured 10^{th} rank in the competition. The results demonstrate that our proposed pre-trained method can effectively capture structural damage patterns and visual cues associated with destruction scenes. Code: https://github.com/kannanrrk/NakbaImageClassifier

Keywords: EfficientNet-B3, Nakba Image Classification, Destruction Detection, Deep Learning, Transfer Learning, Computer Vision

1. Introduction

The rapid proliferation of visual content on social media has transformed the way conflicts are documented, disseminated and interpreted. In contemporary crises, images often shows formal reporting, shaping public opinion and informing journalistic, humanitarian and policy responses in near real time. However, the overwhelming volume of user-generated and news imagery articles has created a critical bottleneck. The lack of reliable, scalable systems for automatically identifying and categorizing scenes of destruction and violence.

Automatic destruction detection lies at the intersection of computer vision, multimedia analysis and humanitarian technology. The breakthrough of deep convolutional neural networks (CNNs) (Krizhevsky et al., 2012) for large-scale image recognition established the foundation for modern visual classification systems. Subsequent architectural innovations such as residual learning (He et al., 2016) further improved training stability and performance in deep networks. These advances enabled the development of specialized systems for crisis and disaster imagery analysis. Social media based damage assessment and crisis related image classification have demonstrated the feasibility of automatically identifying destruction severity (Nguyen et al., 2017; Alam et al., 2018) and humanitarian needs from visual streams. More recently, transformer-based vision architectures have shown strong capability in modeling global visual context (Dosovitskiy et al., 2021), offering alternative approaches to convolution-based systems to attention based systems (Vaswani et al., 2023).

To advance the research in these area, NakbaArchiveClassifier (Abrahams et al., 2026) organized a shared task on destruction detection classifiers as part of NakbaNLP at LREC 2026. The shared task focuses on binary and fine grained classification of destruction related images. The organizers encouraged participants to develop robust model which is capable of handling visual ambiguity, varying image quality and domain-specific challenges.

As a competitive and computationally efficient baseline, the use of EfficientNet-B3 was proposed. EfficientNet introduced a principled compound scaling method that uniformly balances network depth, width and resolution, achieving state of the art accuracy with significantly fewer parameters than conventional CNN architectures (Tan and Le, 2019). The B3 variant offers a favorable trade off between representational power and computational efficiency, making it well suited for large scale image classification tasks under realistic resource constraints. It's improved resolution scaling enhances sensitivity to fine grained visual cues such as debris patterns, smoke, structural damage and injury indicators.

In this proposed work, EfficientNet-B3 acts as the backbone architecture for destruction detection,

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 229–233

fine tuned on the Nakba-related dataset. Through this shared task the work aims to establish standardized benchmark, encourages methodological diversity and stimulate further research on automated destruction detection in historically and politically significant visual archives.

2. Related Work

Recent advances in vision and multimodal modeling have substantially influenced how visual content is represented and classified. Early work in vision language joint representations laid the foundation for models capable of integrating textual and visual cues, with models such as ViLBERT demonstrating co-attentional modeling of images and text for vision-and-language tasks (Lu et al., 2019). Similarly, universal encoders like Unicoder VL introduced cross modal pretraining strategies that learned joint visual linguistic representations useful for retrieval and reasoning (Li et al., 2019).

In the realm of large scale multimodal pretraining, Contrastive Language–Image Pre-training (CLIP) emerged as a pivotal model, enabling zero-shot classification across diverse domains and learning transferable image representations through natural language supervision (Radford et al., 2021). Building on this, models such as BEiT-3 unified vision and language tasks through masked visual representation (Imglish) prediction and joint modeling, achieving strong transfer across classification and detection tasks (Wang et al., 2022).

More recent work has focused on further tightening vision language alignment and performance. Bootstrapping Language-Image Pretraining (BLIP) and its successor BLIP-2 improved multimodal pretraining. It combined strong image encoders with language generation objectives to support both understanding and generation tasks (Li et al., 2023). Hybrid multimodal architectures such as the Bilateral Vision Transformer (BVA-Transformer) leverage BLIP's learned features to fuse image and text representations for improved multimodal classification performance (Zhang et al., 2024). Recent research also explores post training enhancements for foundational multimodal models like CLIP, using techniques such as diffusion based feedback to improve fine grained visual perception and classification accuracy (Wang et al., 2024b).

Beyond classification and retrieval, large multimodal foundation models continue to evolve rapidly with models integrating vision capabilities directly into language model backbones. Qwen-VL and Gemini variants, shown broad reasoning and understanding across visual and textual inputs (Wang et al., 2024a). Surveys of multimodal vision models further highlights the paradigm shift towards unified multimodal reasoning frameworks that leverage joint visual language training objectives to address complex semantic tasks, including image classification, organization and multimodal information detection (Singh et al., 2026; Lupaşcu et al., 2026). Vision-based approaches have been applied in meme-related studies, including MMOD-Meme (Ramesh Kannan et al., 2022), which focuses on face emotion recognition, and TrollMeme (Ramesh Kannan and Rajalakshmi, 2022) classification. These highlights provides various implementation details on vision based approach models. Research must be expanded both in depth and breadth to build a robust vision based model for detecting destruction images.

3. Methodology

3.1. Problem Formulation

The task is formulated as a binary image classification problem. Let $\mathcal{X}$ denote the space of input images and $\mathcal{Y} = \{0, 1\}$ the label space, where

$$0 \rightarrow \text{Non-destruction}, \quad 1 \rightarrow \text{Destruction}.$$

Given an input image $I \in \mathcal{X}$, the goal is to learn a function

$$f_\theta : \mathcal{X} \rightarrow \mathcal{Y},$$

parameterized by θ, such that the predicted label is

$$\hat{y} = f_\theta(I), \quad \hat{y} \in \{0, 1\}.$$

Equivalently, in probabilistic form, the model estimates

$$p_\theta(y \mid I),$$

and the final prediction is obtained as

$$\hat{y} = \arg \max_{y \in \{0,1\}} p_\theta(y \mid I).$$

3.2. EfficientNet-B3 Model

EfficientNet-B3 is used as the backbone network for feature extraction. EfficientNet models scale network depth, width and resolution in a balanced manner, enabling efficient learning with fewer parameters. Transfer learning is applied by initializing the EfficientNet-B3 network with pretrained weights and fine tuning the final layers for binary classification. The architecture consists of EfficientNet-B3 Network with convolutional layer as a backbone, followed by Global Average Pooling layer, Fully connected dense layer, Softmax layer. The final layer produces probability scores for the two classes. Initially, all training images and their corresponding labels were organized, ensuring proper alignment between image names and class labels. Label encoding was performed using a LabelEncoder to convert categorical labels into numerical form suitable for model training.

To handle the data efficiently, a custom Dataset class was implemented, which loads each image, applies transformations and returns it alongside its label. The images were preprocessed using the torchvision library, which involved resizing them to a fixed dimension of 300×300 pixels and converting them to tensors. Data were loaded with DataLoader with batching, shuffling and multi-threaded loading to enhance training performance. The model architecture was built upon a pretrained EfficientNet-B3, with the original classification head removed and replaced by a fully connected layer matching the number of target classes. The model was trained using the AdamW optimizer with a cross-entropy loss function. During training, gradients were computed and back propagated to update the model parameters iteratively. After training, the model weights were saved for inference. For testing, a separate TestDataset was released by the organizers (Abrahams et al., 2026). The unseen images predictions were generated using the trained model. This pipeline ensures a systematic and reproducible workflow for NakbaImage Destruction Image. Table 1 shows the parameter details of the proposed approach.

Parameter	Value
Input size	300 × 300
Batch size	16
Optimizer	Adam
Learning rate	1e-4
Loss Function	CrossEntropyLoss

Table 1: Training Parameters

4. Dataset

The Nakba dataset (Abrahams et al., 2026) was manually collected from publicly available online sources and historical archives by SaltPillar. Nakba Dataset contains large collection of images sourced from social media and news platforms, with a focus on scenes related to conflict and destruction. The images vary in content and context, ranging from direct depictions of violence such as airstrikes, shootings, injuries, death, blood, terror, grief. Some of them are destruction based images with indirect or contextual scenes, including children playing, adults engaging in daily activities, or journalists reporting abroad. The primary goal of this shared task is to support the development of automated systems capable of detecting and categorizing scenes of destruction. Such systems are essential for journalists, engineers, humanitarian organizations and decision-makers who rely on timely, accurate visual information for analysis, reporting and crisis response. Images are labeled to distinguish between directly relevant content like violent image or destructive image and indirectly relevant content to contextual scene or non-violent scenes. The organizers released dataset is shown in Table 2, where 70 % of the data is Training set and 10 % data is Validation set and remaining 20 % data is Test set.

Split	not_destruction	destruction
Training	906	494
Validation	129	70
Testing	260	142
Total	**1295**	**706**

Table 2: Dataset distribution

5. Experiments and Results

Multiple convolutional based architectures were investigated to attain the effectiveness of deep learning models for automated destruction detection in Nakba archival images from Table 3. Several Convolutional Neural Network (CNN) architectures were evaluated, including a simple CNN, ResNet(He et al., 2016), RegNet (Radosavovic et al., 2020) and EfficientNet-B3 (Tan and Le, 2019). The experimental results demonstrate that deep transfer learning models significantly outperforms the basic CNN baseline models.

Models	F1 Score
Basic CNN	0.64
ResNet	0.82
RegNet	0.82
EfficientNet-B3	**0.834**

Table 3: Results

The simple CNN achieved an F1-score of 0.64, indicating limited capability in capturing complex visual patterns. In contrast, deeper architectures such as ResNet and RegNet improved the performance substantially, both achieving an F1-score of 0.82. These models benefit from deeper feature extraction and improved representation learning, which enhances their ability to capture structural and contextual patterns in the images.

Among the evaluated models, EfficientNet-B3 achieved the highest F1-score of 0.834, demonstrating superior performance in classification accuracy of 0.8557. This improvement can be attributed to EfficientNet compound scaling strategy, which jointly optimizes network depth, width and resolution to achieve better feature representation with improved computational efficiency. The results suggest that transfer learning using EfficientNet-B3 provides a robust framework for detecting destruction patterns in a historical Palestine Nakba images.

Overall, the findings indicate that advanced pre-trained architectures significantly enhance classification performance compared to conventional CNN models. Particularly EfficientNet-B3 offers an effective and scalable solution for automated destruction detection in Nakba image archives.

6. Conclusion

Thus, the study presents a deep learning-based framework for the automated identification of destruction in conflict affected regions using Nakba-related images from Palestine. By formulating the task as a binary image classification problem and leveraging transfer learning with EfficientNet-B3, the proposed approach effectively learns discriminative visual features associated with structural damage and non-destruction scenes. The experimental results demonstrate the effectiveness of the model, achieving an F1-score of 83.87 % and an overall accuracy of 85.57 %, securing the 10^{th} rank in the competition. These findings indicate that transfer learning with EfficientNet-B3 is capable of capturing meaningful structural patterns and contextual cues relevant to destruction analysis, even with a manually collected and annotated dataset. Moreover, the study highlights the potential of deep learning techniques to support humanitarian monitoring, digital archiving, and historical documentation efforts.

Ethics Statement

This study does not involve human subjects, personal data, or sensitive information.

Limitations: The dataset is relatively small and manually curated, which may limit the model's generalization to other conflict-affected regions or diverse image sources. Visual context may not always be fully sufficient to understand destruction patterns. Other metadata information like geospatial information could enhance performance.

Future Work: Future research can focus on expanding the dataset, data augmentation strategies, incorporating multimodal information, incorporating multi-class damage severity levels, exploring ensemble or attention-based architectures in historical visual images.

7. Acknowledgements

We thank the Management for their guidance and support, which enabled us to continue the research.

8. References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. Nakbaarchiveclassifier: 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.

Firoj Alam, Fethi Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, 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)*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. 2019. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *arXiv preprint*.

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 *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

Marian Lupaşcu, Ana-Cristina Rogoz, Mihai Sorin Stupariu, and Radu Tudor Ionescu. 2026. Large multimodal models for low-resource languages: A survey. *Information Fusion*, 131:104189.

Dat Tien Nguyen, Fethi Ofli, Muhammad Imran, and Prasenjit Mitra. 2017. Damage assessment from social media imagery data during disasters. In

Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, and Gabriel Goh. 2021. Learning transferable visual models from natural language supervision.

Ilija Radosavovic, Ramprasaath R. Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10428–10436.

R Ramesh Kannan and Ratnavel Rajalakshmi. 2022. Multimodal code-mixed tamil troll meme classification using feature fusion. In *Proceedings of the First Workshop on Multimodal Machine Learning in Low-resource Languages*, pages 1–8.

R Ramesh Kannan, Manikandan Ravikiran, and Ratnavel Rajalakshmi. 2022. Mmod-meme: A dataset for multimodal face emotion recognition on code-mixed tamil memes. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 335–345. Springer International Publishing.

Gurpreet Singh, Lamia Qamar Athanickal Palassery, Nicholas Volta, Amruta Velamuri, and Aya Khanyile. 2026. Vision–language foundation models and multimodal large language models: A comprehensive survey of architectures, benchmarks, and open challenges. *Preprints*.

Mingxing Tan and Quoc V Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pages 6105–6114.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Ke-Yang Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *ArXiv*, abs/2409.12191.

Wenhui Wang, Hangbo Bao, Li Dong, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks.

Wenxuan Wang, Quan Sun, et al. 2024b. Diffusion feedback helps clip see better.

Kaiyu Zhang, Fei Wu, Guowei Zhang, Jiawei Liu, and Min Li. 2024. Bva-transformer: Image-text multimodal classification and dialogue model architecture based on blip and visual attention mechanism. *Displays*, 83:102710.