electronics

MDPI

Article

Enhancing Image Copy Detection through Dynamic Augmentation and Efficient Sampling with Minimal Data

Mohamed Fawzy *, Noha S. Tawfik and Sherine Nagy Saleh

Computer Engineering Department, College of Engineering and Technology, Arab Academy for Science, Technology and Maritime Transport, Alexandria 1029, Egypt; noha.abdelsalam@aast.edu (N.S.T.); sherine_nagi@aast.edu (S.N.S.)

* Correspondence: mfawzyy2302@aast.edu; Tel.: +20-1116-341-931

Abstract: Social networks have become deeply integrated into our daily lives, leading to an increase in image sharing across different platforms. Simultaneously, the existence of robust and user-friendly media editors not only facilitates artistic innovation, but also raises concerns regarding the ease of creating misleading media. This highlights the need for developing new advanced techniques for the image copy detection task, which involves evaluating whether photos or videos originate from the same source. This research introduces a novel application of the Vision Transformer (ViT) model to the image copy detection task on the DISC21 dataset. Our approach involves innovative strategic sampling of the extensive DISC21 training set using K-means clustering to achieve a representative subset. Additionally, we employ complex augmentation pipelines applied while training with varying intensities. Our methodology follows the instance discrimination concept, where the Vision Transformer model is used as a classifier to map different augmentations of the same image to the same class. Next, the trained ViT model extracts descriptors of original and manipulated images that subsequently underwent post-processing to reduce dimensionality. Our best-achieving model, tested on a refined query set of 10K augmented images from the DISC21 dataset, attained a state-of-the-art micro-average precision of 0.79, demonstrating the effectiveness and innovation of our approach.

Keywords: artificial intelligence; copy detection; deep learning; supervised learning; vision transformers

Citation: Fawzy, M.; Tawfik, N.S.; Saleh, S.N. Enhancing Image Copy Detection through Dynamic Augmentation and Efficient Sampling with Minimal Data. *Electronics* **2024**, *13*, 3125. https://doi.org/10.3390/electronics13163125

Academic Editor: Francesco Beritelli

Received: 13 June 2024
Revised: 20 July 2024
Accepted: 31 July 2024
Published: 7 August 2024

1. Introduction

Social media has become an integral part of daily life, remarkably influencing how we communicate, share information, and perceive the world, particularly through the use of and engagement with images. This phenomenon extends to the modification of images, which plays a crucial role in shaping personal and social narratives. The blending of text and images in spreading information has been shown to enhance the perceived credibility of content, making the detection and correction of disinformation increasingly challenging [1]. Additionally, the widespread influence of digitally altered images has been found to affect individual body satisfaction and beauty goals, maintaining unrealistic standards despite public awareness of these modifications [2].

Copy detection has emerged as a fundamental task in the digital era. As defined by Pizzi et al. [3], this process involves determining whether pieces of media, such as photographs or videos, are derived from the same original source. The relevance of copy detection expands in the context of image usage and manipulation, a common occurrence in our digital world where image modification and distribution are common.

Image manipulation, while prevalent in the digital age, is not a new phenomenon. Historically, it has been used for various purposes. Joseph Stalin manipulated photographs to eliminate political rivals from Soviet records, as highlighted by source [4]. In today's digital landscape, however, the simplicity of altering images brings additional challenges. The easy access to sophisticated, yet user-friendly, image editing tools poses risks related to

historical accuracy and the propagation of false information. Modern examples, including the alteration of images of Martin Luther King Jr. to depict them inappropriately and the manipulation of crowd sizes in photographs from President Trump's inauguration, serve as examples of how easily historical facts can be misrepresented [5]. Furthermore, an analysis of over 13 million Facebook posts, found that about 23% of political images were corrupted with misinformation [6]. This included various forms, such as altered images, misleading memes, unmodified images with incorrect captions, and screenshots of misleading social media posts.

The increasing sophistication in the manipulation of digital content, as evidenced by the study of Khalil et al. [7], highlights the urgent need for advanced models capable of distinguishing between genuine and altered media, underscoring the critical role of technological advancements in maintaining the authenticity of digital media. Using deepfake technology in the diffusion of personal intimate images, especially through sharing personal images without consent, is one example of unauthorized image manipulation approaches. The use of such technology violates personal privacy and may lead to one being involved in legal and social problems [8].

With the rising use of social media and the vast amount of photos and videos shared every day, it has recently become a challenge to detect false visuals. The threat of such false media going viral while being unnoticed may lead to the spread of false, biased truths and people losing trust in the media. This has brought about the need for developing reliable image copy detection models to be used by journalists and social media platforms in order to spot and combat these false visuals. Therefore, having effective image copy detection tools is essential, as it ensures the ethical sharing of original images, which in turn helps maintain the trustworthiness of online content. These tools also protect the rights of those who create visual content against unauthorized copying.

This article attempts to find a solution to the image copy detection problem through the use of Vision Transformers (ViT), particularly focusing on the descriptor track of the Image Similarity Challenge 2021 (ISC21) [9], which will be explained in the upcoming section. Our approach features the following innovative aspects:

1. Dynamic augmentation pipelines: Introducing a dynamic augmentation strategy with varying levels of augmentation difficulty, applied on the fly during training. This dynamic augmentation not only enhances the model's robustness against various image transformations, but also provides a fresh set of training examples for each epoch.
2. Data-efficient performance: training with significantly fewer images (only 3% of the DISC21 dataset [9]) and less time, our approach still matches the performance of the winning solutions in the ISC21 challenge's descriptor track, achieving comparable micro-average precision.
3. Optimized subset selection: Employing a clustering algorithm to segment the training data, we systematically select representative samples from the total DISC21 training data available. The clustering approach ensures comprehensive exposure across all variations within the selected training subset, simulating full data coverage and aiding in better generalization even with less data.

The novelty of our work lies in the combination of these techniques to achieve state-of-the-art performance in image copy detection with minimal data and computational resources.

The rest of the article is structured into four main sections. First, the related work reviews current literature on image copy detection and the ISC21 challenge, followed by a proposed model section illustrating the details of the architecture and techniques of our model. The experimental results section presents the achieved results on the DISC21 dataset along with their comparison to others. Finally, there is a section to conclude our findings and outline future research directions for the image copy detection task.

2. Related Work

The development of the Copydays dataset by Douze et al. [10] is one of the early attempts in image copy detection research. Not only did they introduce a dataset specifically

for the copy detection task, but they also provided essential insights into the robustness of Global Image Feature (GIST) descriptors [11] against various image alterations.

Image copy detection has undergone significant advancements, leveraging evolving computational methods and machine learning technologies. The method of GIST-PCA hashing has emerged as a significant development. Kim et al. [12] introduced a scalable approach for near-duplicate image detection, utilizing a combination of GIST descriptor and Principal Component Analysis (PCA) [13] hashing. This method deals with image variations such as cropping and accurate border framing.

With the advancements in deep learning, the MultiGrain network introduced by Berman et al. [14] innovatively combines image classification and particular object retrieval into a unified framework. Based on a standard classification trunk with a pooling layer adaptable to high-resolution images, this network is trained with both cross-entropy loss for classification and a ranking loss for retrieval.

Highlighting the practical application of these advancements, the ISC21 challenge, a key event in the field of image copy detection, showcased several innovative approaches that significantly advanced the domain. Among the winners, distinct methodologies stood out, each contributing uniquely to the field.

Yokoo et al. [15] focused on enhancing the discriminative power of image embeddings, essential for accurate image copy detection. Employing EfficientNetV2, they innovated a robust training pipeline inspired by progressive learning, which escalated the input image resolution and regularization as training advanced. They harnessed contrastive loss with cross-batch memory for training, outperforming other metric learning losses. A significant part of their methodology was the negative embedding subtraction post-process, which enhanced copy detection performance by isolating target samples from similar negative samples.

On the other hand, Papadakis and Addicam [16] took a different approach. They initially employed a triplet-based training method, and later transitioned to the Additive Angular Margin Loss (ArcFace) approach for its practicality and enhanced performance. This adaptation involved experimenting with different image sizes and resolutions to optimize the models' effectiveness. The core of their methodology utilized established Convolutional Neural Network (CNN) architectures, including EfficientNetV2 l, EfficientNetV2 s, EfficientNet b5, and NfNet l1. These models processed the modified images through a self-adaptive pooling layer, followed by a linear layer, to generate uniform image signatures for consistent recognition. To improve facial image detection, they incorporated extensive datasets, including ImageNet and a synthetic facial image collection. Their innovative 'Drip Training' method progressively introduced complexity, enabling gradual adaptation and preventing model overload.

Wang et al. [17] presented a novel approach utilizing unsupervised pre-training with Barlow-Twins and deep metric learning. Their method departs from supervised learning models, addressing the unstructured and diverse nature of online images. They introduce 'descriptor stretching' to adjust model outputs for consistency and employ a dual-loss function, combining triplet loss with hard sample mining and cross-entropy loss, for balanced learning. The baseline model includes a Generalized Mean (GeM) pooling layer, WaveBlock enhancements when using the ResNet50 architecture, and a high-dimension projector transforming 2048-dim features into 8192-dim. A learnable matrix then reduces these features to 256-dim for efficient processing.

While the previously mentioned methodologies summarize the winning solutions and highlight key findings recognized in the ISC21 challenge, it is also important to acknowledge significant contributions outside of the challenge. Pizzi et al. [18], introduced the SSCD model, which adapts self-supervised contrastive training for copy detection, incorporating a ResNet-50 trunk for feature extraction and GeM pooling to enhance descriptor discrimination. A notable feature of SSCD is the use of entropy regularization, which is crucial for maintaining distinct separation between descriptor vectors.

The invention of Vision Transformers (ViTs) has had a revolutionary impact on image analysis. Originally discovered by Dosovitskiy et al. [19], ViTs process images by dividing them into patches similar to the approach of Natural Language Processing. Recent studies like the work of Horváth et al. [20] and Jang et al. [21] have explored the potential of ViTs in several applications. Jang et al. introduce the concept of Self-Distilled Self-Supervised Learning (SDSSL), which has improved the performance of ViTs in several tasks, including image copy detection. Furthermore, Khan et al.'s work [22] demonstrates the applicability of ViTs in a wide spectrum of image processing tasks. The work of Coccomini et al. [23] compares ViTs and CNNs in the context of deepfake image detection; their observations show that while CNNs, EfficientNetV2 to be specific, provide better results during training, ViTs have better generalization ability. Examining all the previously mentioned studies suggests that ViTs can achieve good results in the detection of manipulated image copies.

Despite the diverse methodologies in image copy detection, recent studies have focused only on contrastive learning approaches and specific applications of Vision Transformers in contexts like satellite image manipulation and deepfake detection. However, these approaches have not explored the potential of ViTs in a classification framework tailored to the challenges of image copy detection. Our research fills this gap by employing ViTs not just as a feature extractor, but as a central component of a classification-based model. This model capitalizes on the strengths of ViTs in handling the complexity of modern digital environments.

3. Dataset and Challenge Background

The field of image copy detection has gained notable attention in parallel with the introduction of focused datasets, such as Copydays developed by Douze et al. [10]. It contained only 157 original images and 229 transformed versions of such images. Some examples of augmentations include resizing, cropping, and more complex edits. Despite its utility, Copydays' limited dataset size and range of manipulations have drawn attention to the need for more advanced data that would allow for overcoming contemporary challenges in image copy detection.

The ISC21 challenge [9], conducted at NeurIPS'21, served as a major event for the image copy detection task. The goal was to identify modified copies of a corpus of one million images. This challenge focused on carrying out sophisticated image-matching tasks, ranging from identifying near-exact copies to detecting subtle manipulations. Thus, the challenge aimed at offering a comprehensive assessment of the image recognition technologies used then.

The challenge introduced the DISC21 dataset [9], a new standard for large-scale image copy detection that includes reference, query, and training sets. The dataset encompasses varying augmentations that mirror the complexity found in current social media content. These augmentations range from basic re-encoding and resizing to carrying out more complex edits like overlaying images and random objects, geometric changes, and deepfake creations. These varied modifications challenge algorithms in figuring out the origin of augmented images.

The ISC21 challenge has significantly advanced image copy detection procedures. It did not only aim at identifying modified copies within this vast image collection, but also emphasized the crucial role of the descriptor track. This track specializes in developing powerful descriptors that handle diverse image manipulations.

In this research, we investigate the validity of our proposed model on the DISC21 dataset instead of earlier limited ones like Copydays, as it presents the current challenges in the image copy detection field with its large-scale inclusions of images and augmentation varieties.

4. Proposed Model

The core of our proposed model is employing Vision Transformers in a classification framework, which we adapt and fine-tune for the specific task of image copy detection. Our

proposed model follows the instance discrimination approach [24,25], where each image and its augmented versions are considered a single class. After training, the fine-tuned ViT model serves as a feature extractor to derive embeddings for both original and augmented images, addressing the problem from a descriptor perspective. Subsequently, we compare the post-processed embeddings of each original and augmented image pair for similarity to determine the origin of each query image accurately. An illustration of the proposed model is depicted in Figure 1.

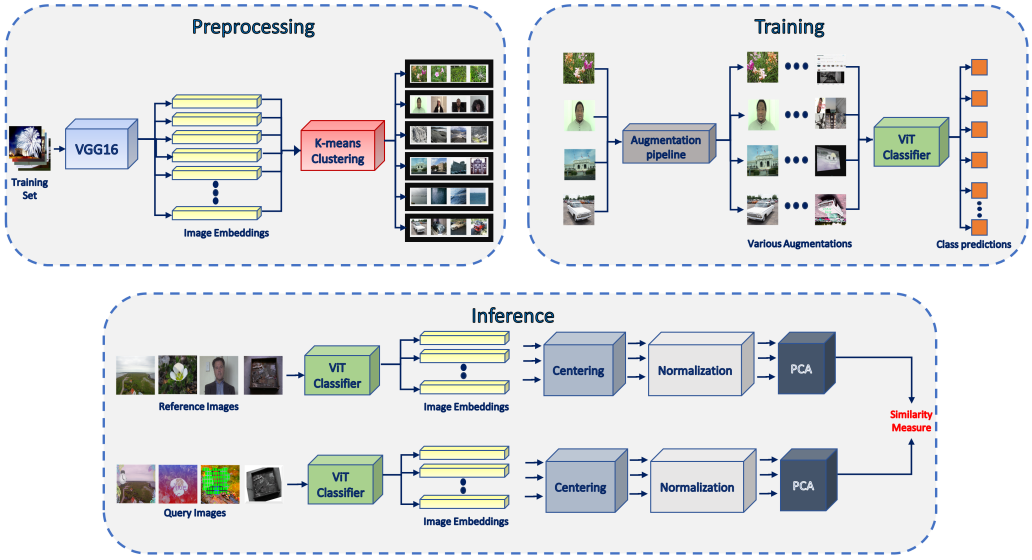


Figure 1. Structure of the proposed model divided into three phases: preprocessing, training, and inference.

Our approach also offers a significant computational advantage in selecting a representative subset of the DISC21 dataset. Unlike conventional methodologies that rely on the exhaustive use of large datasets, our model demonstrates that a more targeted approach can yield comparable results. Additionally, our strategy incorporates a diverse range of image augmentations, creating a robust model capable of handling various forms of image manipulation, which is a critical aspect of copy detection.

4.1. Dataset Selection and Preprocessing

In addressing the challenge posed by the ISC21 competition, we opted to construct a representative subset of images from the extensive DISC21 dataset. The selection process involved the use of the K-means algorithm [26] to cluster the DISC21 training set, with pre-trained VGG16 [27] neural network embeddings as the basis for clustering. This approach allowed for grouping images based on visual similarity, thus ensuring a representative and diverse sample of the entire dataset. The clustering was guided by the elbow method [28], which determined the optimal number of clusters to segment the one million training images, eventually settling on six distinct clusters. Our sampling technique is motivated by the fact that during the learning process of a machine learning model, the model learns its parameters from the training instances provided. Redundant instances do not improve the predictive performance of the model, and can sometimes even degrade it. We further select samples from the distinct clusters to construct the training subsets for our experiments. Those subsets are significantly smaller in size compared to the original DISC21; however, they capture the essential patterns and information with the least redundancy among the chosen representatives.

Image augmentation is an essential concept for training models for copy detection. Modifications like brightness adjustment, rotation, and subject shifting are all important for making a model capable of differentiating between modified and original images. Having

such a diversity of augmentations increases the variety of instances seen by the model, which improves its performance. In this study, we employed a variety of augmentation techniques, each applied with a specific probability and across several levels of hardness. Some of these augmentations were implemented using functions from the AugLy [29] library. Examples of such augmentations are:

- Random collage: assembles a composite from different pictures in a grid, varying from 1×1 to 3×3, with each segment hosting a distinct picture (Figure 2b).
- Legofy: transforms images to look as if they were built out of Lego bricks (Figure 2c).
- Noise: introduces random noise or graininess (Figure 2d).
- Overlay stripes: adds stripe patterns with varying characteristics (Figure 2e).
- Overlay text: overlays randomly generated text onto images (Figure 2f).
- Overlay image: involves overlaying one image over another (Figure 2g).
- Overlay onto screenshot: places images into screenshot templates from social media platforms (Figure 2h).
- Emoji overlay: superimposes random emojis onto training images (Figure 2i).
- Random filter: applies a random filter from a set, including ones like MaxFilter, UnsharpMask, Contour, SMOOTH, etc., to images (Figure 2j).
- Pad square: adds padding around images to form a square shape (Figure 2k).
- Edge enhance: enhances image edges by applying a filter (Figure 2l).

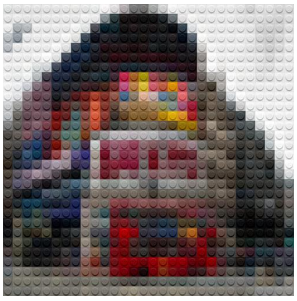
In addition to the specialized augmentations detailed previously, other fundamental image transformations were used. These included standard augmentations like color jitter, grayscale conversion, and horizontal flips, among others. These basic transformations, though simpler, are still crucial, and their inclusion enhanced our model performance. This balanced mixture of both complex and fundamental augmentations was key to achieving a comprehensive training regime.

Our approach to training employs a multi-tiered augmentation strategy, systematically categorizing transformations based on their complexity into four intensity levels: moderate, hard, harder, and hardest, as demonstrated in Figure 3. Moderate augmentations slightly alter images by using minimal changes like minor crops and flips, preserving much of their original structure. As we advance to harder augmentation levels, the model is gradually introduced to more complex transformations. The hard level introduces techniques like overlaying images and legofy, whereas the harder and hardest levels further intensify these changes, presenting the model with more challenging scenarios.

To ensure a balanced and comprehensive training process, we maintain an equal number of augmentations at each difficulty level for every image. This approach is vital for teaching the model the concept of class consistency across different degrees of augmentation. It aids the model in understanding that both minimally and heavily augmented images correspond to the same original image, a key aspect for the downstream task of accurate image mapping as part of the instance discrimination training. Moreover, augmentations are not applied in a static or predetermined sequence. Instead, each augmentation has a certain probability of being applied, and they are randomly shuffled and varied for each training epoch. This approach ensures that the model encounters a wide array of visual challenges, preventing overfitting to specific patterns or transformations. It is also important to note that each tier of augmentation contained the preprocessing steps required by the model, which include image resizing and normalization.

(**a**) Original Image

(**b**) Random Collage



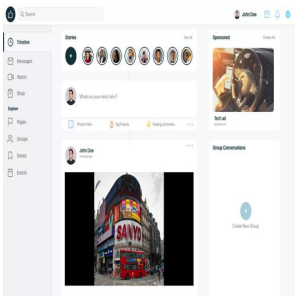
(**c**) Legofy

(**d**) Noise

(**e**) Overlaying Stripes

(**f**) Overlaying Text

(**g**) Overlaying Image



(**h**) Overlaying Onto Screenshot

(**i**) Overlaying Emoji

(**j**) Random Filter

(**k**) Square Padding

(**l**) Edge Enhancing

Figure 2. original image and its augmented versions, including various augmentation techniques.

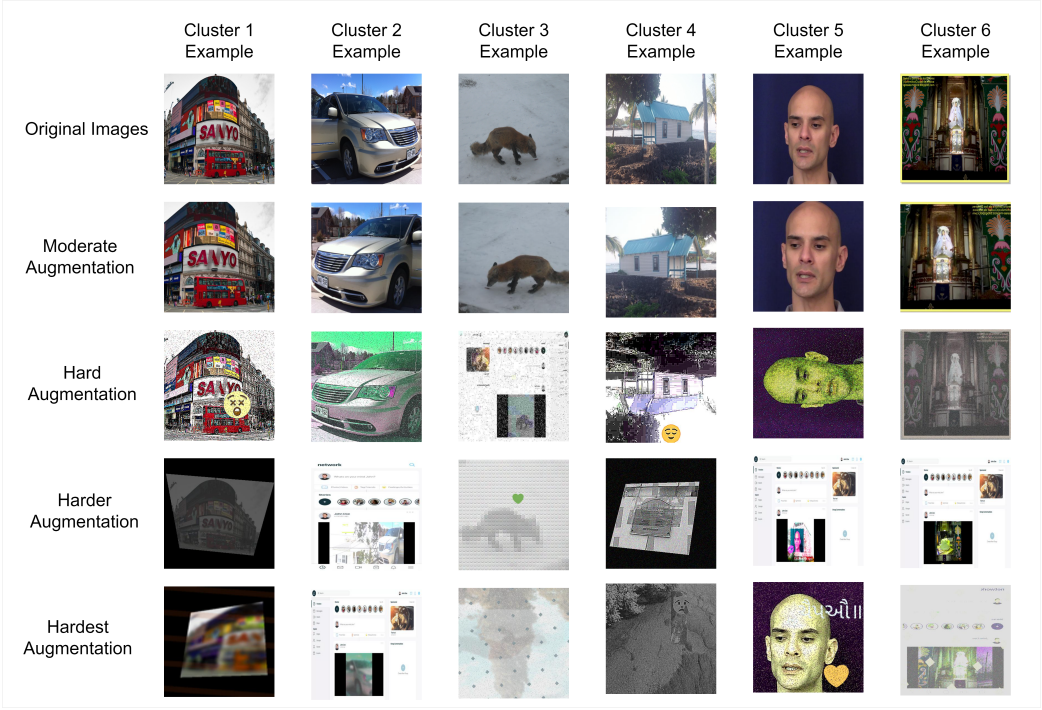


Figure 3. Examples of original images from six different clusters and their augmented versions using different augmentation pipelines.

4.2. Training Methodology

Our training methodology is centered around the Vision Transformer, more specifically the ViT L16 architecture, which is particularly suited for handling complex image data. Since its introduction in 2020, the Vision Transformer model [19] marks a notable shift from the conventional convolutional neural networks that are typically used in image recognition tasks. The ViT architecture leverages the capabilities of transformer models, originally popularized in natural language processing, and adapts them for computer vision tasks, including object detection, segmentation, image and scene generation, and many more [22]. The overall process of our training methodology is visualized in Figure 1.

Key components and processes of ViT:

- Image processing: ViT processes images by dividing them into fixed-size patches, flattening these patches, and projecting them into a high-dimensional space embedded with positional information, transforming the image into a sequence of vectorized patches.
- Classification token: ViT introduces a learnable classification token to the sequence of embedded patches which, after processing through the transformer blocks, serves as the global image representation for classification.
- Self-attention mechanism: the self-attention mechanism in ViT computes a weighted sum of the input data, focusing on more relevant features to capture informative representations.
- Transformer encoder structure: ViT includes multiple transformer blocks, each with a multi-head self-attention layer and a feed-forward layer, processing image patches effectively.
- Additional layers: the architecture also incorporates Multilayer Perceptron (MLP) layers and Layer Normalization (LN).
- Class prediction: in our adapted approach, each non-augmented image is treated as a unique class. Post-training, the classification token's embedding is used to determine the similarity between the original and augmented images.

A comprehensive depiction of the Vision Transformer architecture, as employed in our study, is presented in Figure 4.

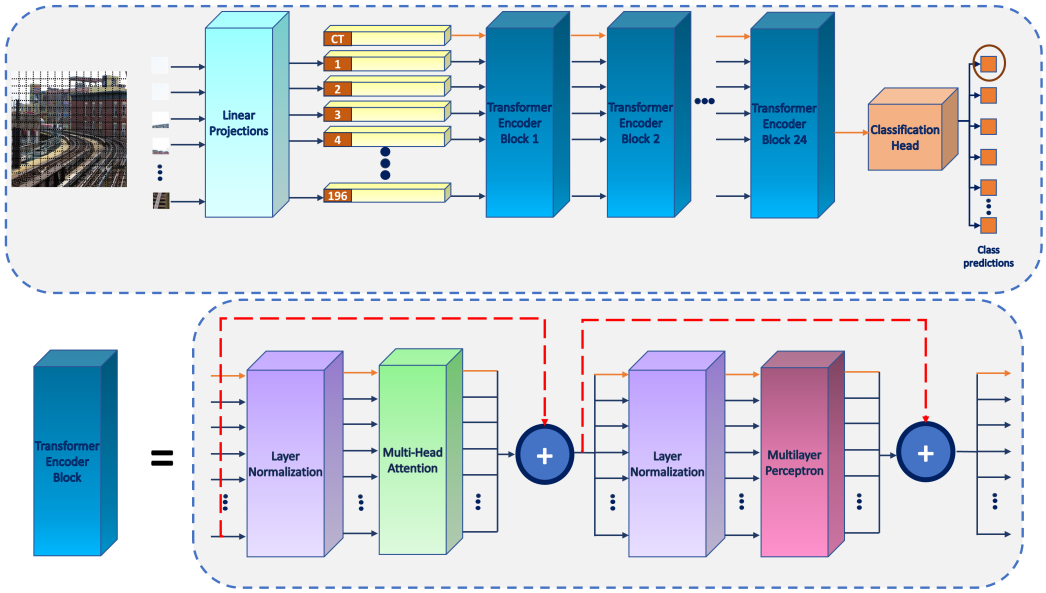


Figure 4. The upper part of the figure demonstrates the architecture of the ViT L16 model. ViT L16 features 24 transformer encoder blocks; the lower part details the architecture of each one of them.

We initialized our model with ViT L16's pre-trained weights, utilizing its extensive learning from large image datasets like ImageNet [30]. The model's classification head was modified to align with our classification approach, where the number of classes is equal to the number of images in the selected subset of the training set.

In the training phase, we fed the model with the images selected from the DISC21 dataset's clusters, which were subjected to various augmentations. For each epoch, the model learned to classify these augmented images into their respective original classes. We used Cross-Entropy as our loss function, which is effective for multi-class classification problems [31]. The optimization of the model was carried out using Stochastic Gradient Descent (SGD) [32] with a momentum of 0.9. We set a learning rate of 4×10^{-4}, and included weight decay to prevent overfitting.

4.3. Inference and Postprocessing

After training the model, we shifted to the evaluation phase to assess the model's performance in detecting image copies by removing the classification head and utilizing it as a feature extractor. In this stage, the model extracted 1024-dimensional embeddings from the DISC21 dataset's reference and query sets, corresponding to original and augmented images, respectively. A series of post-processing steps were conducted afterward, which include centering, normalization, and PCA [13] with whitening to reduce the dimensionality to 512. Resultant embeddings were used to conduct a similarity search, using the FAISS [33] library, between reference and query sets, enabling an evaluation of the model's effectiveness in accurately mapping between original images and their modified counterparts.

4.4. Evaluation Metric: Micro Average Precision

We assessed the model's performance using the micro-average precision (µAP) metric, which was the standard method of determining the winners of the ISC21 challenge. The µAP offers a comprehensive measure of the model's accuracy across various confidence thresholds by calculating the average precision, essentially representing the area under the precision-recall curve.

The µAP is defined as:

$$\text{Micro Average Precision (µAP)} = \sum_{i=1}^{N} p(i)\Delta r(i) \quad (1)$$

- $p(i)$ is the precision at position i in the sorted list of detected pairs.
- $\Delta r(i)$ is the change in recall from the previous position.
- N is the total number of detected pairs for all queries.

This calculation, involving all returned pairs for all queries, is a robust metric for evaluating the overall effectiveness of the model in detecting various instances of image manipulation [9].

5. Experimental Results

In this part of our research, we focus on validating the performance of our proposed model for image copy detection on the DISC21 dataset. A series of experiments were designed to assess the model's efficiency under different configurations. Additionally, our analysis includes a comparison with the state-of-the-art methods that have been previously applied to the DISC21 dataset.

Our experimental analysis was conducted on the DISC21 dataset, specifically focusing on the phase two data, which was pivotal in determining the challenge's winners. It is crucial to note that, unlike the standard test set used in the challenge, we adopted a modified approach for our experiments. The original challenge format involved a reference set of 1 million non-augmented images and a phase 2 query set of 50,000 augmented images. In this setup, not all query images had corresponding origins in the reference set; specifically, only 10,000 of the 50,000 query images were linked to an origin in the reference set. Our experiments validate the performance of our model by refining the test set to 10,000 query images and a corresponding set of 10,000 reference images, denoted as $DISC21_{ref}$. In this set, we ensure that each query image is directly mapped to a singular, corresponding image in the reference set. Furthermore, to maintain the integrity and comparability of our results, we reproduce the results from winners and other published literature by using their publicly released models on the newly refined query and reference set.

Several experiments were initially conducted to ensure optimal model parameters. We focused on smaller-scale evaluations to identify the best configuration employing a targeted parameter tuning process due to computational limitations for larger experiments. This involved exploring various epoch sizes (ranging between 1 and 10), tuning fixed learning rates (ranging between 1×10^{-4} and 6×10^{-4}), weight decay (ranging between 1×10^{-4} and 1×10^{-8}) and experimenting with learning rate scheduling techniques. The batch size was restricted by the hardware resources; thus, values 16, 32 and 64 were tested. These experiments guided us to select the most effective parameter combination, which is detailed in Table 1.

Table 1. Training parameters.

Parameter	Value
Framework	PyTorch
GPU	NVIDIA RTX 4060 Ti
Model Architecture	ViT L16
Epochs	7
Batch Size	64
Learning Rate	4×10^{-4}
Weight Decay	1×10^{-6}
Image Resolution	224×224 pixels

Table 2 presents a detailed analysis of our model's performance across varied training configurations. It outlines the outcomes of all experiments conducted, each quantified by

the µAP metric, under distinctive setups. Rows labeled 1K through 30K correspond to the count of original images used in each experiment, which also matches the number of classes. The columns, ranging from 20 to 200, indicate the number of augmentations. Augmentations are applied to each original image in the training data with equal distribution among the four intensity levels (moderate, hard, harder, and hardest). The table shows that our best model achieved a µAP of 0.79 on the $DISC21_{ref}$.

Table 2. Micro average precision of various training configurations with clustering.

Images \ Augs	20	40	60	80	120	160	200
1K	0.52	0.56	0.58	0.60	0.61	0.61	0.52
5K	0.60	0.62	0.63	0.64	0.66	0.67	0.58
10K	0.61	0.63	0.65	0.69	0.74	0.74	0.70
20K	0.66	0.69	0.75	0.75	0.75	0.77	0.78
30K	0.68	0.71	0.73	0.76	0.78	0.79	0.78

While it is logical to assume that increasing the volume of data can enhance model performance, as observed by higher µAP in bigger training sets; however, it is also important to recognize the importance of the augmentation pipeline. Increasing the number of instances per class through different intensities of augmentation often offers considerable benefits compared to merely increasing the number of classes by expanding the dataset size. Our findings show that the augmentation effect increased the model's performance across all training scenarios, with a maximum increase of 0.13 in the 10K training set. A detailed examination of the data presented in Table 2 highlights an interesting pattern between the values in the last columns for the initial rows and those in the first columns for the latter rows. For example, the precision values at lower augmentations (20, 40) for higher image counts (20K, 30K) are comparable to those at higher augmentations (160, 200) for lower image counts (1K, 5K). In other words, the model performs almost the same when it is trained on the same total number of data, whether these data are augmented instances or instances of new classes. This leads us to the conclusion that, as the number of instances per class increases, it compensates for the overall quantity of images. This observation is particularly advantageous, considering the exhaustive and time-consuming task of collecting training data for machine learning models. It suggests that systematic augmentation and class instance balancing can effectively enhance model performance without the need for extensive data collection. On the other hand, it is also important to note that the benefits of augmentations plateau beyond a threshold, indicating the presence of an optimal augmentation level beyond which model performance may not significantly improve, as observed from the last columns in the table.

Similarly, Table 3 presents a comparison between the µAP achieved using our proposed clustering approach against a random non-clustering framework. Specifically, the clustering approach involves selecting a balanced number of images from each cluster, whereas the random approach involves drawing images arbitrarily from the entire pool of 1 million training images. To investigate the effectiveness of choosing a representative subset, we selected several experiments for comparison, as detailed in the table covering most training configurations. This led us to compare early configurations (1K $\times$ 20 to 30K $\times$ 20) across both frameworks. The consistently superior performance of the clustering approach in these initial tests provided a strong motivation to develop the clustering models fully, as seen in Table 3. Subsequently, we revisited the non-clustering approach for our best-performing configuration, 30K $\times$ 160, to test our model. The clustering approach also demonstrated superior results, confirming our approach's effectiveness.

Table 3. Comparison of micro average precision for clustering and random non clustering approaches.

Configuration	Random	Clustering
1K × 20	0.35	0.52
5K × 20	0.41	0.60
10K × 20	0.55	0.61
20K × 20	0.57	0.66
30K × 20	0.51	0.68
30K × 160	0.76	0.79

We also performed experiments to test the importance of each block in the post-processing stage. Table 4 summarizes the impact of various post-processing blocks on our proposed model. Each row represents a different combination of the components, marked with ✓ for included blocks. By comparing the results achieved, it can be noted that normalization is crucial, significantly boosting performance (shown in experiments 3, 5, 6, and 7). While centering offered a slight further improvement when added to normalization (experiment 5), PCA (experiment 6) did not provide any additional benefit. Combining all three components (experiment 8) yielded a significant 2% improvement in overall performance compared to normalization alone.

In Table 5, we compare the performance of our best model against that of the top three winners from the descriptor track of the challenge [15–17] and the work of Pizzi et al. [18], which is the most recent study on this dataset on the $DISC21_{ref}$. It is important to note that, as previously mentioned, $DISC21_{ref}$ is a refined query and reference set, and hence, we re-implemented the evaluation scripts of winner models within our validation scheme. The table illustrates the standing of our model in the context of other state-of-the-art models. Not only did our proposed model outperform all other models with only 3% of the training data other models employed, but it also surpassed models that produce an even bigger embedding vector size.

Table 4. Analysis of the effect of each component in the post-processing stage.

Exp. ID	Centering	Normalization	PCA	µAP
1	×	×	×	0.69
2	✓	×	×	0.69
3	×	✓	×	0.76
4	×	×	✓	0.66
5	✓	✓	×	0.77
6	✓	×	✓	0.67
7	×	✓	✓	0.76
8	✓	✓	✓	0.79

The results from Tables 2 and 3 show key aspects of our model's performance. The integration of the ViT architecture, along with our dynamic augmentation strategy and a well-structured classification framework, played a pivotal role in producing cut-edge results with a fraction of the training data used previously. Further, the comparative analysis between the clustering and non-clustering approaches, as shown in Table 3, validates the efficacy of our proposed clustering method. This enhanced performance is likely attributed to the prevalence of highly similar images in the training set. The findings are in agreement with our hypothesis that less data can still yield more improvements in performance when carefully selected, augmented, and trained.

Table 5. Comparison of model performances in terms of micro average precision (µAP) on the $DISC21_{ref}$.

Model	Embedding Size	µAP
Yokoo et al. [15]	256	0.76
Papadakis and Addicam [16]	256	0.74
Wang et al. [17]	256	0.73
Pizzi et al. [18] SSCD_DISC_Large	1024	0.78
Pizzi et al. [18] SSCD_DISC_Mixup	512	0.74
Our Best Model	512	0.79

Finally, to assess the generalization ability of our model pre-trained on the DISC21 data, we conducted an additional experiment, testing it on the Copydays dataset. The results were comparable to those presented by Berman et al. [14] and Touvron et al. [34]. Berman et al. introduced two ResNet-based models with differing resolution sizes achieving µAP of 0.75 and 0.825, respectively, with a vector size of 2048, while Touvron et al. presented two different ViT B/16 models, achieving 0.764 and 0.818 with a vector size of 1536. Notably, our proposed model outperformed both by achieving 0.84, despite having a smaller embedding size of only 1024.

6. Conclusions

In this research, we aimed to tackle the image copy detection problem, which is critical in our digital world. Our innovative approach merged K-means clustering with VGG16 embeddings to strategically group the images of the DISC21 training set, allowing us to choose a smaller yet effective training set. Multi-tiered augmentation pipelines with varying intensity levels are introduced to form the training set, which is used to train the ViT L16 model in a classification framework. In evaluation, we adopted a descriptor strategy such that the trained model extracts feature vectors from original and manipulated images, which were post-processed and compared for similarity. Our proposed pipeline, featuring dynamic augmentation and optimized subset selection, achieved state-of-the-art results with our best model trained with only 30K images of the DISC21 training set, achieving a µAP of 0.79, outperforming state-of-the-art and recent literature. These results demonstrate the effectiveness and novelty of our integrated clustering and augmentation approach, which significantly improved our results over random samples. Our approach is further validated by achieving comparable performance on the Copydays dataset, demonstrating its generalizability across different copy detection benchmarks.

For future research, we aim to further improve our findings by applying other clustering algorithms and enlarging our augmentation pipeline to adapt to any new trends in image manipulation, including the use of advanced generative models for transformations such as image-to-image translation, occlusion and inpainting, and sample generation. Another set of experiments will include testing other ViT architectures to try to enhance our performance further. Moreover, we intend to expand our scope and investigate the validity of the proposed model with other digital media formats, such as video.

Author Contributions: Conceptualization, M.F., N.S.T. and S.N.S.; methodology, M.F.; software, M.F.; validation, M.F., N.S.T. and S.N.S.; formal analysis, M.F.; investigation, M.F.; resources, M.F.; data curation, M.F.; writing—original draft preparation, M.F.; writing—review and editing, M.F., N.S.T. and S.N.S.; visualization, M.F.; supervision, N.S.T. and S.N.S.; project administration, N.S.T. and S.N.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are openly available in the DISC21 dataset repository at https://ai.meta.com/datasets/disc21-dataset/ (accessed on 30 July 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hameleers, M.; Powell, T.E.; Van Der Meer, T.G.; Bos, L. A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Commun.* **2020**, *37*, 281–301. [CrossRef]
2. MacCallum, F.; Widdows, H. Altered images: Understanding the influence of unrealistic images and beauty aspirations. *Health Care Anal.* **2018**, *26*, 235–245. [CrossRef] [PubMed]
3. Pizzi, E.; Kordopatis-Zilos, G.; Patel, H.; Postelnicu, G.; Ravindra, S.N.; Gupta, A.; Papadopoulos, S.; Tolias, G.; Douze, M. The 2023 Video Similarity Dataset and Challenge. *arXiv* **2023**, arXiv:2306.09489.
4. Blakemore, E. *How Photos Became a Weapon in Stalin's Great Purge*; Canal História; A&E Television Networks: New York, NY, USA, 2019.
5. Thomson, T.; Angus, D.; Dootson, P.; Hurcombe, E.; Smith, A. Visual mis/disinformation in journalism and public communications: Current verification practices, challenges, and future opportunities. *J. Pract.* **2022**, *16*, 938–962. [CrossRef]
6. Yang, Y.; Davis, T.; Hindman, M. Visual misinformation on Facebook. *J. Commun.* **2023**, *73*, 316–328. [CrossRef]
7. Khalil, S.S.; Youssef, S.M.; Saleh, S.N. A Multi-Layer Capsule-Based Forensics Model for Fake Detection of Digital Visual Media. In Proceedings of the 2020 International Conference on Communications, Signal Processing, and Their Applications (ICCSPA), Sharjah, United Arab Emirates, 16–18 March 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
8. Viola, M.; Voto, C. Designed to abuse? Deepfakes and the non-consensual diffusion of intimate images. *Synthese* **2023**, *201*, 30. [CrossRef]
9. Douze, M.; Tolias, G.; Pizzi, E.; Papakipos, Z.; Chanussot, L.; Radenovic, F.; Jenicek, T.; Maximov, M.; Leal-Taixé, L.; Elezi, I.; et al. The 2021 image similarity dataset and challenge. *arXiv* **2021**, arXiv:2106.09672.
10. Douze, M.; Jégou, H.; Sandhawalia, H.; Amsaleg, L.; Schmid, C. Evaluation of gist descriptors for web-scale image search. In Proceedings of the ACM International Conference on Image and Video Retrieval, Fira, Greece, 8–10 July 2009; pp. 1–8.
11. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [CrossRef]
12. Kim, H.; Sohn, S.; Kim, J. Revisiting Gist-PCA hashing for near duplicate image detection. *J. Signal Process. Syst.* **2019**, *91*, 575–586. [CrossRef]
13. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
14. Berman, M.; Jégou, H.; Vedaldi, A.; Kokkinos, I.; Douze, M. Multigrain: A unified image embedding for classes and instances. *arXiv* **2019**, arXiv:1902.05509.
15. Yokoo, S. Contrastive learning with large memory bank and negative embedding subtraction for accurate copy detection. *arXiv* **2021**, arXiv:2112.04323.
16. Papadakis, S.M.; Addicam, S. Producing augmentation-invariant embeddings from real-life imagery. *arXiv* **2021**, arXiv:2112.03415.
17. Wang, W.; Zhang, W.; Sun, Y.; Yang, Y. Bag of tricks and a strong baseline for image copy detection. *arXiv* **2021**, arXiv:2111.08004.
18. Pizzi, E.; Roy, S.D.; Ravindra, S.N.; Goyal, P.; Douze, M. A self-supervised descriptor for image copy detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14532–14542.
19. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
20. Horváth, J.; Baireddy, S.; Hao, H.; Montserrat, D.M.; Delp, E.J. Manipulation detection in satellite images using vision transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Sharjah, United Arab Emirates, 16–18 March 2021; pp. 1032–1041.
21. Jang, J.; Kim, S.; Yoo, K.; Kong, C.; Kim, J.; Kwak, N. Self-distilled self-supervised representation learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 2829–2839.
22. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *ACM Comput. Surv. (CSUR)* **2022**, *54*, 1–41. [CrossRef]
23. Coccomini, D.A.; Caldelli, R.; Falchi, F.; Gennaro, C.; Amato, G. Cross-forgery analysis of vision transformers and CNNs for Deepfake Image detection. In Proceedings of the 1st International Workshop on Multimedia AI against Disinformation, Newark, NJ, USA, 27–30 June 2022; pp. 52–58.
24. Wu, Z.; Xiong, Y.; Yu, S.X.; Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3733–3742.
25. Zhao, N.; Wu, Z.; Lau, R.W.; Lin, S. What makes instance discrimination good for transfer learning? *arXiv* **2020**, arXiv:2006.06606.
26. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]
27. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
28. Thorndike, R.L. Who belongs in the family? *Psychometrika* **1953**, *18*, 267–276. [CrossRef]
29. Papakipos, Z.; Bitton, J. Augly: Data augmentations for robustness. *arXiv* **2022**, arXiv:2201.06494.
30. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.

31. Demirkaya, A.; Chen, J.; Oymak, S. Exploring the role of loss functions in multiclass classification. In Proceedings of the 2020 54th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, 18–20 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–5.
32. Robbins, H.; Monro, S. A stochastic approximation method. *Ann. Math. Stat.* **1951**, *22*, 400–407. [CrossRef]
33. Johnson, J.; Douze, M.; Jégou, H. Billion-scale similarity search with gpus. *IEEE Trans. Big Data* **2019**, *7*, 535–547. [CrossRef]
34. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 9650–9660.