information

Article

Detect with Style: A Contrastive Learning Framework for Detecting Computer-Generated Images

Georgios Karantaidis * **and Constantine Kotropoulos**

School of Informatics, Faculty of Sciences, Aristotle University of Thessaloniki, University Campus, 54124 Thessaloniki, Greece; costas@csd.auth.gr
* Correspondence: gkarantai@csd.auth.gr

Abstract: The detection of computer-generated (CG) multimedia content has become of utmost importance due to the advances in digital image processing and computer graphics. Realistic CG images could be used for fraudulent purposes due to the deceiving recognition capabilities of human eyes. So, there is a need to deploy algorithmic tools for distinguishing CG images from natural ones within multimedia forensics. Here, an end-to-end framework is proposed to tackle the problem of distinguishing CG images from natural ones by utilizing supervised contrastive learning and arbitrary style transfer by means of a two-stage deep neural network architecture. This architecture enables discrimination by leveraging per-class embeddings and generating multiple training samples to increase model capacity without the need for a vast amount of initial data. Stochastic weight averaging (SWA) is also employed to improve the generalization and stability of the proposed framework. Extensive experiments are conducted to investigate the impact of various noise conditions on the classification accuracy and the proposed framework's generalization ability. The conducted experiments demonstrate superior performance over the existing state-of-the-art methodologies on the public DSTok, Rahmouni, and LSCGB benchmark datasets. Hypothesis testing asserts that the improvements in detection accuracy are statistically significant.

Keywords: multimedia forensics; computer-generated images; supervised contrastive learning; style transfer

Citation: Karantaidis, G.; Kotropoulos, C. Detect with Style: A Contrastive Learning Framework for Detecting Computer-Generated Images. *Information* **2024**, *15*, 158. https://doi.org/10.3390/info15030158

Academic Editors: Ilias Theodorakopoulos and Nikolaos Mitianoudis

Received: 29 January 2024
Revised: 6 March 2024
Accepted: 8 March 2024
Published: 11 March 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Artificial intelligence is crucial for detecting disinformation, which threatens democratic values worldwide. Due to the rapid growth of social media content and the rapid development of image processing and machine learning, combating disinformation has become a priority.

Distinguishing computer-generated images (CGIs) from natural images (NIs) is especially pertinent for addressing the practical challenges posed by deepfakes. Deepfakes, being advanced forms of computer-generated imagery, underscore the critical need for robust identification methods. In practical terms, the ability to differentiate between authentic and manipulated content haws become indispensable for countering the proliferation of deepfakes across diverse domains. For instance, within media and entertainment, where deepfakes can deceive audiences by depicting fabricated scenarios or altering the appearance of individuals, the capability to discern these manipulations ensures the preservation of trust between content creators and their viewers. Moreover, in a forensic analysis and cybersecurity context, the application of techniques to identify alterations plays a pivotal role in verifying the authenticity of visual evidence and mitigating the potential for false information dissemination. This practical application of distinguishing CGIs from NIs extends its significance into various industries reliant on visual representation, fortifying the credibility and reliability of presented content, which is crucial in an era where deepfakes challenge the authenticity of visual information.

Due to the perpetual and exponential growth of multimedia technologies in conjunction with the advances in the deployment of tools for CGI creation, CGIs have become so realistic that individuals are not capable of distinguishing them from NIs with their naked eyes. A plethora of image processing techniques and 3D image rendering software packages have contributed to the creation of such sophisticated content. Various high-quality galleries of CGIs exist, such as the Autodesk A360 rendering gallery [1], the Artlantis gallery [2], the VRay gallery [3], and the Corona gallery [4]. Notwithstanding the multimedia forgery outbreak, realistic CGIs have come to be added to the arsenal of fraudsters. As a countermeasure, there is an urgent need to deploy algorithms that can accurately and reliably discriminate between CGIs and NIs. Thus, multimedia forensics draws the community's attention to methods to encounter all kinds of attacks within image forensics [5], including approaches for universal image forensics [6], copy–move forgery detection [7], splice detection [8], and face anti-spoofing detection [9]. Many approaches have also been introduced in the context of image forgery detection that leverage gradient-based illumination [10], decision fusion [11], pairwise relations [12], and transformed spaces based on image illuminant maps [13].

Digital forensics can be useful for determining the difference between NIs and CGIs. A scenario where CGIs can cause harm is through image manipulation for political propaganda, making authenticity validation a crucial aspect. Another challenging scenario is verifying the authenticity of images, particularly when offenders attempt to manipulate child pornography photos digitally so as to appear like CGIs. In all circumstances, attesting to the validity of the photographs is a key challenge in forensics.

Distinguishing CGIs from NIs can be treated as a classification task. Until recently, many approaches have proposed hand-crafted features [14–18] to cope with the aforementioned classification problem, while the majority of recent state-of-the-art methods utilize deep neural network (DNN) methods, e.g., [19–24]. The latter methods tend to be more efficient in discovering hidden patterns and structures in images. On top of that, the generalization ability of DNN methods allows for automation, which is crucial in real-life applications, even when large training datasets are unavailable.

In this paper, to take full advantage of NN methods in terms of learning complex data representations and automatically deriving highly accurate decisions, an end-to-end convolutional neural network (CNN)-based framework is proposed to discriminate between CGIs and NIs. To the best of the authors' knowledge, this is the first attempt to demonstrate the potential of supervised contrastive learning in the context of discrimination between CGIs and NIs. The proposed framework consists of two stages. First, a CNN is proposed, which is based on the ResNet-18 [25] architecture that employs the supervised contrastive (SupCon) loss presented in [26]. On top of this, and apart from the data augmentation, a complementary style transfer module is introduced to enhance training by enriching the network with additional negative samples to those of the original dataset. Handcrafted image augmentations (e.g., cropping, blurring, flipping) provide insufficient variation in visual features, limiting the performance of contrastive learning techniques that employ them. The style transfer module creates synthetic images (e.g., deepfakes) but they can also add artificial visual features to real images. The core idea behind integrating style transfer is to enable more accurate training by using only the original dataset, even when insufficient training samples exist. The paper demonstrates that style transfer improves the accuracy of contrastive learning. During the second stage, the trained model is fed to a linear classifier for further training using the cross-entropy loss.

Contrastive learning forces samples of the same class to stay close to each other, while samples that belong to different classes are pushed far away. Supervised contrastive learning leverages the label information, providing many positive samples to the network instead of self-supervised contrastive learning. Positive samples are fed into the classifier using data augmentation procedures. Moreover, stochastic weight averaging (SWA) is employed on the network outputs after each stage to improve robustness. The style transfer module operates in real time and takes advantage of the NIs that constitute the positive class. It introduces a progressive attentional manifold alignment. Thus, it can dynamically reposition the style features of some arbitrarily chosen CGIs by repeated attention operations to align the content manifold to the style manifold. With the contribution of the style transfer module, the training procedure is enriched with additional incoming samples, allowing models with datasets that consist of a limited number of training samples to be more robust and effective. Overall, the proposed framework aims at identifying and mitigating deceptive visuals, fortifying the trustworthiness and reliability of visual content across diverse applications and sectors.

The experimental results are disclosed on the public benchmark DSTok [27], Rahmouni [24], and LSCGB [28] datasets, demonstrating that the proposed framework accurately distinguishes CGIs and NIs, outperforming the state-of-the-art approaches and motivating further research. On top of that, the generalization ability of the proposed framework trained on the DSTok dataset is tested on the publicly available Rahmouni dataset. Moreover, CoStNet is trained on the most recent state-of-the-art LSCGB dataset and tested on the challenging DSTok dataset and is compared against state-of-the-art approaches. The impact of various parameters during the training is assessed. When the test samples are infected with salt-and-pepper or Gaussian noise, an extensive evaluation of the proposed approach is performed to attest to its ability to deliver accurate results under various conditions. When insufficient training samples are available, an ablation study is undertaken to examine the impact of the style transfer module. Furthermore, hypothesis testing is performed to assess whether the improvements in detection accuracy delivered by the proposed framework against state-of-the-art approaches are statistically significant.

The main contributions of the paper are as follows:

- A novel CNN-based framework is designed to discriminate between CGIs and NIs, abbreviated as CoStNet. To the best of the authors' knowledge, this is the first attempt to conduct such discrimination based on supervised contrastive learning and style transfer in the benchmark DSTok, Rahmouni, and LSCGB datasets.
- A complementary style transfer module, which operates in real-time, is employed to increase the training CGIs even when a limited number of training samples is available, thus enhancing the training procedure.
- CoStNet achieves state-of-the-art accuracies in the benchmark DSTok, Rahmouni, and LSCGB datasets, underscoring its remarkable advancement in the field.
- The generalization capability of CoStNet, initially trained on the LSCGB dataset, is evaluated through testing on the DSTok dataset. Additionally, CoStNet undergoes training on the DSTok dataset and is subsequently tested on the Rahmounis' dataset to assess its broader applicability.
- The proposed framework is robust against high salt-and-pepper and Gaussian noise at various corruption levels.
- Multiple tests are conducted to empirically demonstrate that CoStNet is less sensitive to modifications of the training parameters, such as the number of training epochs and the batch size.
- An ablation study is performed to assess the impact of the style transfer module when limited training samples are available.
- Hypothesis testing confirms that the improvements in detection accuracy between CoStNet and methods reported in the literature are statistically significant.

In summary, the proposed CoStNet framework is a CNN-based novel architecture that utilizes real-time style transfer and supervised contrastive learning to discriminate CGIs from NIs. CoStNet is demonstrated to accurately discriminate CGIs and NIs across benchmark datasets such as the DSTok, the Rahmouni, and the LSCGB datasets. The incorporation of the style transfer module allows for the augmentation of CGIs based on existing image content, thus offering additional training CGIs. By doing so, the challenge of training sample scarcity for CGIs prevalent in real-world forensic scenarios is addressed. CoStNet's robust performance in handling various noise levels and parameter settings, as well as its generalization ability in testing, further underscores its versatility and effectiveness under diverse conditions. CoStNet's resilience to variations is also evaluated in scenarios with limited training data through an ablation study, demonstrating its capabilities in CGI discrimination.

The rest of this paper is organized as follows. Section 2 briefly surveys the literature on the discrimination of CGIs from NIs. Section 3 details the proposed framework. Benchmark datasets are described in Section 4. Experimental evaluation is presented in Section 5. Conclusions are drawn, and limitations and future work are discussed in Section 6.

2. Related Work

The advances in multimedia forensics, on the one hand, and the sophisticated software that enables the ever-increasing creation of realistic CGIs, on the other hand, have challenged scientists to develop new methods to encounter fraudulent manipulations arising from such technological advances. A transfer learning and convolution block attention module, which considers both the shallow content features and the deep semantic features of the image, was introduced in [19] to tackle the problem of distinguishing NIs from CGIs. Parallel to the evolution of algorithms, ever-challenging datasets were released. In [28], the new large-scale CG benchmark dataset (LSCGB) was released, consisting of 71,168 CG and 71,168 natural annotated images. A baseline texture-aware network was proposed to address the discrimination problem on their benchmark dataset. A novel two-branch network was proposed to tackle the generalization problem in the blind detection of CGIs by introducing different initializations in the first layer so that more diverse features were extracted [20]. However, no prior knowledge of new distributions was used to develop a rigorous formulation. In [22], the color and texture characteristics of local patches were integrated within a dual-input CNN framework, and a directed acyclic graph recurrent neural network was employed to model the spatial dependence of local patterns.

A statistical model for NIs was proposed in [29], built upon a wavelet-like decomposition. Higher-order wavelet statistics showed substantial differences that made it possible to distinguish between CGIs and NIs. In [16], a geometry-based model was proposed that utilized the physical characteristics of CGIs and NIs in the classification process. Local patches of the image intensity function were employed to form a patch distribution, which enabled, in combination with the geometry model, to uncover the distinctive physical characteristics of NIs and CGIs. The statistical characteristics of local edge patches were examined in [18], and a visual language was created to handle the discrimination between CGIs and NIs. In [30], a technique based on sensor pattern noise was developed that used three high-pass filters to filter out low-frequency signals. A five-layer CNN was utilized to classify the input image patches, and a majority vote scheme was employed to extend the classification results to the full-sized images. A CNN was presented in [31], promoting the so-called local-to-global strategy. Forensic decisions were derived from local patches, and a global decision based on majority voting of the full-sized images was implemented. In [24], a CNN with a custom pooling layer was proposed. Local estimates of the class probabilities were employed to predict the full-size image label. A deep convolutional recurrent attention model was proposed to classify CGIs and NIs, employing a local-to-global strategy [32]. Image patches were trained, and the full-sized images were classified using the simple majority vote rule. An attention-based dual-branch CNN with fused color components was proposed in [33]. There, raw RGB components and their noisy versions were given as input to the network, while the attention-based model optimized the output features from the two branches in combination to perform detection. A method for distinguishing between CGIs and NIs based on DNN and transfer learning was presented in [23]. A qualitative examination of ResNet-50 bottleneck characteristics for CGI detection was performed. Comprehensive reviews of various methods for discriminating between CGIs and NIs can be found in [34,35].

3. Proposed Framework

3.1. Framework Overview

The proposed framework for detecting and discriminating CGIs from NIs comprises two modules, as shown in Figure 1. Input CGIs are passed through the first module, namely the style transfer module, which generates additional CGIs added to the training set, thus enriching the training procedure. When there are not enough training samples given a CGI as a basis for style semantics, style transfer can create as many CGIs as NIs, whose content semantics remain unaltered. The style transfer module leverages a pre-trained VGG network [36] to encode the content image and imbue it with stylistic patterns derived from a separate style image, yielding distinct features for each one of them. Adhering to the framework introduced in [37], these features undergo a transformative process facilitated by the attentional manifold alignment (AMA) block to achieve stylization. This block encompasses a channel alignment module, an attention module, and a spatial interpolation module. Once processed through three iterations of the AMA blocks, the aligned content feature is fed into the decoder, resulting in the generation of the stylized image. In terms of practical implementation, the role of the style transfer module within CoStNet is pivotal. This module operates by transferring the visual characteristics of one image (e.g., texture, color, and style) onto another while preserving its content. Particularly in scenarios where the availability of CGIs is limited, the style transfer module enhances the diversity of CGIs by synthesizing new images based on the content of NIs. In such cases, the augmentation enriches the training data, thereby improving the robustness of the framework against variations in CGI appearance. More details on style transfer learning are given in Section 3.2. An example of a generated CGI based on an NI through the style transfer module is depicted in Figure 2. Details for the style transfer module can be found in Figure 3.

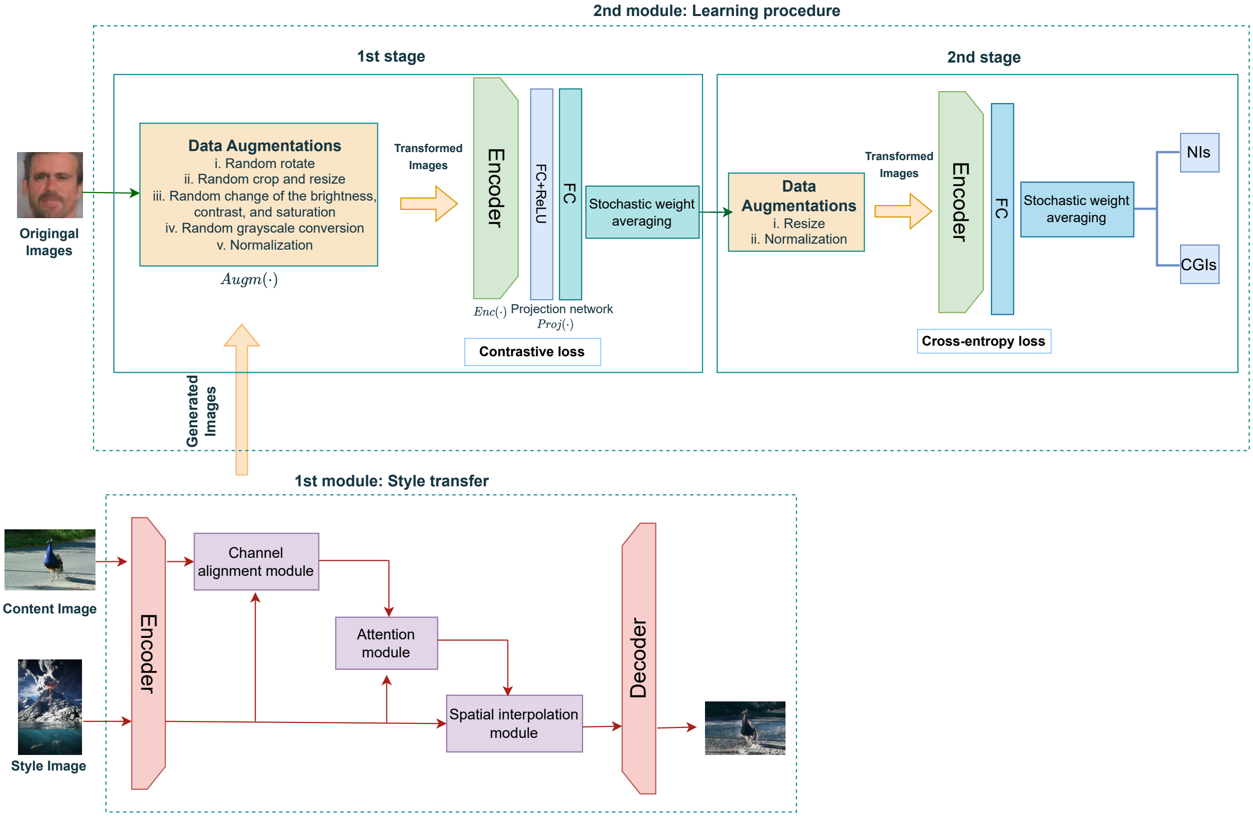


Figure 1. Architecture of the proposed CoStNet.

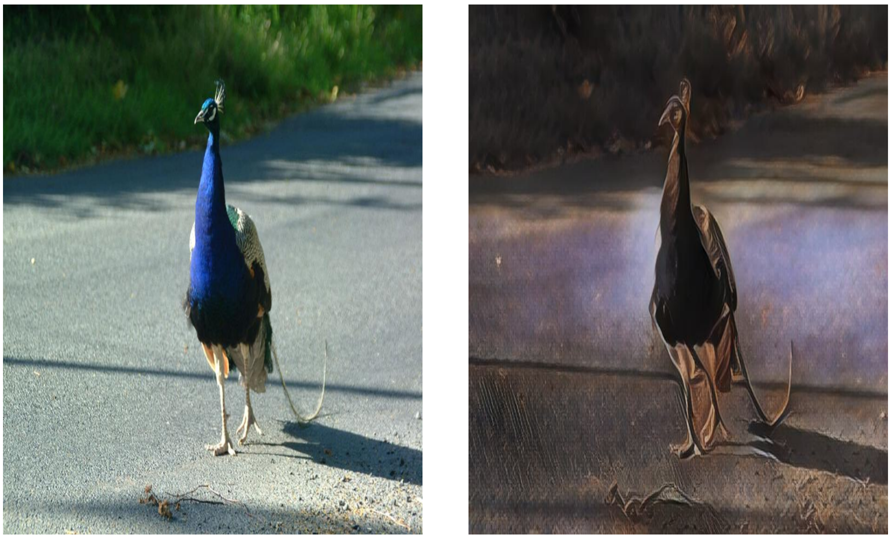


Figure 2. On the (**left**), a natural image is depicted. On the (**right**), a computer-generated image is shown that was generated by the style transfer module.

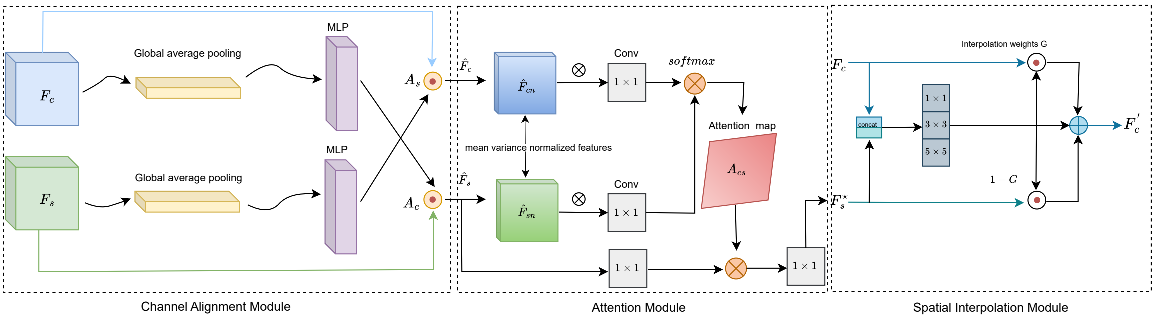


Figure 3. Components of style transfer module.

Having increased the number of CGIs by employing the style transfer module, the total amount of input images is passed to the second module, which implements the learning procedure. It consists of two distinct stages. The first stage comprises an encoder with a CNN architecture, namely the ResNet-18 architecture, employing the supervised contrastive loss. Details can be found in Section 3.3. Upon receiving an input batch of the enriched data, random data augmentation is applied twice to yield two batch duplicates representing a different data view. Both duplicates are then forwarded through the encoder network, generating a normalized embedding. During training, this representation is further passed through a projection network, which is disregarded during inference. The outputs of the projection network are used to compute the supervised contrastive loss, as proposed in [26]. For classification purposes, the output of the first stage is fed into an encoder network identical to that of the first stage and then to a linear classifier, which is trained on top of the fixed representations using cross-entropy loss, allowing the trained model to be employed for classification tasks. After each training stage, SWA is applied to improve the model's generalization and stability [38].

3.2. *Style Transfer Learning*

CoStNet integrates a style transfer module to render a content image with style patterns from a reference image. By doing so, training samples are augmented, and the algorithm performance improves even when few training samples are available. A state-of-the-art arbitrary style transfer framework, called Progressive Attentional Manifold Alignment [37], is employed, which gradually aligns the content and style manifolds using an attention mechanism for consistent stylization across semantic regions. The loss function in the progressive manifold alignment approach is comprised of several stages. Let L_{ss} denote the content loss, while L_r, L_m, and L_h denote the style losses. At each stage, the loss is calculated as a weighted sum [37]:

$$L = \sum_{i=1}^{3}(\lambda^i_{ss}L^i_{ss} + \lambda^i_r L^i_r + \lambda^i_m L^i_m + \lambda^i_h L^i_h) + L_{ae} \tag{1}$$

where λ^i_ξ refers to a weight parameter for L_ξ, $\xi \in \{ss, r, m, h\}$ in the ith stage of the procedure as described in [37], and L_{ae} stands for the autoencoder loss.

The content loss L_{ss} employs the ℓ_1 norm between the self-similarity matrices of the content feature $\boldsymbol{F}_c$ and the VGG feature of the stylized image $\boldsymbol{F}_{cs}$ [36]. Let also H_x and W_x denote the height and the width of the feature $\boldsymbol{F}_x$ with $x \in \{c, s\}$, respectively. The L_{ss} is given by

$$L_{ss} = \frac{1}{H_c W_c}\sum_{ij}\left|\frac{D^c_{ij}}{\sum_i D^c_{ij}} - \frac{D^{cs}_{ij}}{\sum_j D^{cs}_{ij}}\right| \tag{2}$$

where $\boldsymbol{D}^c = [D^c_{ij}]$ and $\boldsymbol{D}^{cs} = [D^{cs}_{ij}]$ are the pairwise cosine distance matrices of content features $\boldsymbol{F}_c$ and VGG features of the stylized image $\boldsymbol{F}_{cs}$. The cosine distance matrix is defined as 1 minus the cosine similarity between $\boldsymbol{F}_c$ and $\boldsymbol{F}_{cs}$ [39].

Let L_r denote the relaxed earth mover distance [40,41] to align the content manifold to the style manifold:

$$L_r = \max\left(\frac{1}{H_s W_s}\sum_i \min_j C_{ij}, \frac{1}{H_c W_c}\sum_j \min_i C_{ij}\right) \tag{3}$$

where C_{ij} denotes the pairwise cosine distance matrix between $\boldsymbol{F}_{cs}$ and $\boldsymbol{F}_s$. The statistic of the style feature is represented by the subscript s, while the statistic of the VGG feature of the stylization result is represented by the subscript cs.

In order to regularize the magnitude of features, the moment matching loss was employed [37]:

$$L_m = ||\boldsymbol{\mu}_{cs} - \boldsymbol{\mu}_s||_1 + ||\boldsymbol{\Sigma}_{cs} - \boldsymbol{\Sigma}_s||_1 \tag{4}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the mean vector and the covariance matrix of the feature vectors.

Let L_h be the differentiable color histogram loss introduced in [42]:

$$L_h = \frac{1}{\sqrt{2}}||\boldsymbol{H}_s^{\frac{1}{2}} - \boldsymbol{H}_{cs}^{\frac{1}{2}}||_2 \tag{5}$$

where $\boldsymbol{H}$ refers to the color histogram feature and $\boldsymbol{H}^{\frac{1}{2}}$ denotes the element-wise square root. The color histogram feature proposed in [42] is employed to control the distribution of colors in the generated images. The color histogram loss function encourages the generated images to match a specified color histogram, which is a representation of the distribution of colors in an image.

Moreover, an autoencoder loss L_{ae} is proposed to preserve the shared space during manifold alignment. Let $\boldsymbol{I}_{rc}$ and $\boldsymbol{I}_{rs}$ denote the reconstructed content and style images from the encoded features. The loss is given by [37]:

$$L_{ae} = \lambda_{ae}(||\boldsymbol{I}_{rc} - \boldsymbol{I}_c||_2 + ||(\boldsymbol{I}_{rs} - \boldsymbol{I}_s)||_2) + \sum_i(||\phi_i(\boldsymbol{I}_{rc}) - \phi_i(\boldsymbol{I}_c)||_2 + ||\phi_i(\boldsymbol{I}_{rs}) - \phi_i(\boldsymbol{I}_s)||_2) \tag{6}$$

where λ_{ae} denotes a weight parameter which is kept fixed. $\phi_i(\boldsymbol{I})$ refers to a Rectified Linear Unit *ReLU_i_1* layer VGG feature of image $\boldsymbol{I}$, where *ReLU_i_j* denotes the result of *conv_i_j* with ReLU activation. The loss Equation (6) forces the decoder to reconstruct features in the VGG space, which in turn restricts all features between the encoder and decoder to lie within this space. The loss Equation (6) retains a common space for aligning the content and style manifolds from the standpoint of manifold alignment.

The proposed module is a completely autonomous part of the CoStNet. It comprises a channel alignment module designed to accentuate related content and style semantics, an attention module facilitating the establishment of feature correspondences, and a spatial interpolation module aimed at dynamically aligning the manifold structures.

The channel alignment module utilizes a combination of global average pooling and a multilayer perceptron (MLP) to embed $F \in R^{H \times W \times C}$ into R^C and derive the corresponding channel weights. H, W, and C denote the height, width, and the channels of F. These weights, denoted as $A_c \in R^C$ and $A_s \in R^C$, are computed based on both the content feature F_c and the style feature F_s. Subsequently, the features F_c and F_s undergo cross-weighting with A_s and A_c, respectively, resulting in aligned features $\hat{F}_c$ and $\hat{F}_s$. The attention module utilizes 1×1 convolutional blocks for feature embedding, along with mean variance normalization, to compute the attention map A_{cs}. The attention map captures pairwise similarities between features. Subsequently, the style feature vectors are redistributed based on the content feature $F_s^{\star}$ according to the computed attention map A_{cs}. The spatial interpolation module synthesizes spatial information for adaptive interpolation between the content feature F_c and the redistributed style feature $\hat{F}_s^{\star}$. Specifically, the dense operation employs multiscale convolution kernels on the concatenated feature to compute interpolation weights G. By concatenating the features, local discrepancies between corresponding content and style features are identified, enabling the determination of appropriate interpolation strengths. Consequently, the spatial interpolation module effectively merges the most similar content and style feature vectors, facilitating manifold alignment through linear redistribution of the style feature and interpolation of its linear components with the content feature. A visual workflow of style transfer module components is depicted in Figure 3. A detailed analysis of each component can be found in [37]. The style transfer module is executed before feeding the training samples into the first stage employing the CNN. It can be implemented in real-time, depending on whether additional samples are needed due to a lack of training CG samples.

3.3. Supervised Contrastive Learning

Self-supervised contrastive learning tries to maximize the similarity of two normalized vector representations (i.e., embeddings), pulling together the normalized embeddings that belong to the same class while pushing away the normalized embeddings that belong to different classes. In [26], label information was leveraged, and self-supervised contrastive learning was extended to fully supervised contrastive learning, enabling the consideration of many positives and negatives per anchor. On the contrary, in self-supervised learning, a single positive is only considered. Here, the extension in [26] is exploited by including the SupCon loss Equation (9) to the challenging application of distinguishing between CGIs and NIs.

From a practical point of view, contrastive learning embeds data points into a latent space, where similar instances are brought closer together while dissimilar instances are pushed apart. Specifically, CoStNet takes advantage of the supervised contrastive learning, where the framework is trained to minimize the contrastive loss between positive pairs (i.e., samples from the same class, e.g., NIs) and maximize the margin between negative pairs (i.e., samples from different classes, such as CGIs and NIs). This process facilitates the learning of discriminative features necessary to accurately distinguish between NIs and CGIs. This capability is particularly important in real-world practical forensic applications, where the accurate discrimination is essential for reliable analysis and interpretation.

Following the notation in [26], let us consider a set of N image/label pairs, $\{\boldsymbol{x}_k, \boldsymbol{y}_k\}_{k=1,\dots,N}$ with their corresponding training batch $\{\tilde{\boldsymbol{x}}_l, \tilde{\boldsymbol{y}}_l\}_{l=1,\dots,2N}$, where $\tilde{\boldsymbol{x}}_{2k}$ and $\tilde{\boldsymbol{x}}_{2k-1}$ are 2 augmentations of $\boldsymbol{x}_k$ and $\tilde{\boldsymbol{y}}_{2k-1} = \tilde{\boldsymbol{y}}_{2k} = \boldsymbol{y}_k$.

Let $I \equiv \{1, 2, \ldots, 2N\}$. For $i \in I$, let $A(i) \equiv I \setminus \{i\}$. If $\tau \in \mathbb{R}^+$ denotes a scalar temperature parameter, define

$$P_{i\,x} \equiv \frac{\exp\left(\frac{\boldsymbol{z}_i^\top \boldsymbol{z}_x}{\tau}\right)}{\sum_{\alpha \in A(i)} \exp\left(\frac{\boldsymbol{z}_i^\top \boldsymbol{z}_\alpha}{\tau}\right)}. \tag{7}$$

In Equation (7), $\boldsymbol{z}_l = \mathrm{Proj}(\mathrm{Enc}(\tilde{\boldsymbol{x}}_l)) \in \mathbb{R}^{D_p}$, where D_p is the size of a single linear layer, $\mathrm{Enc}(\tilde{\boldsymbol{x}}_l)$ maps $\tilde{\boldsymbol{x}}$ to a representation vector $\boldsymbol{r}_l$, and $\mathrm{Proj}(\boldsymbol{r}_l)$ maps $\boldsymbol{r}_l$ to vector $\boldsymbol{z}_l$.

Let i be the anchor and $j(i)$ denote the index of another augmented sample in the same set known as the *positive*. In the self-supervised approach, the loss function is formulated as [26]:

$$\mathcal{L}^{self} = \sum_{i \in I} \mathcal{L}_i^{self} = -\sum_{i \in I} \log P_{i\,j(i)} \tag{8}$$

The remaining $2(N-1)$ indices in $A(i) \setminus \{j(i)\}$ are called the *negatives*.

The SupCon loss is a generalization of Equation (8), which leverages the label information [26]. The SupCon loss is formulated as follows:

$$\mathcal{L}_{out}^{sup} = \sum_{i \in I} \mathcal{L}_{out,i}^{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log P_{i\,p} \tag{9}$$

where $P(i) \equiv \{p \in A(i) : \tilde{\boldsymbol{y}}_p = \tilde{\boldsymbol{y}}_i\}$ denotes the set of indices of all positives in the set of augmented samples that are distinct from i and $|P(i)|$ stands for the cardinality of set $P(i)$.

Here, the gradient of SupCon loss Equation (9) is given by:

$$\begin{aligned}
\frac{\partial \mathcal{L}_{out}^{sup}}{\partial \boldsymbol{z}_i} &= \frac{-1}{|P(i)|} \sum_{p \in P(i)} \frac{\partial}{\partial \boldsymbol{z}_i} \left\{ \frac{\boldsymbol{z}_i^\top \boldsymbol{z}_p}{\tau} - \log \sum_{\alpha \in A(i)} \exp(\boldsymbol{z}_i^\top \boldsymbol{z}_\alpha / \tau) \right\} \\
&= \frac{-1}{\tau |P(i)|} \sum_{p \in P(i)} \left\{ \boldsymbol{z}_p - \frac{\sum_{\alpha \in A(i)} \boldsymbol{z}_\alpha \, \exp(\boldsymbol{z}_i^\top \boldsymbol{z}_\alpha / \tau)}{\sum_{\alpha \in A(i)} \exp(\boldsymbol{z}_i^\top \boldsymbol{z}_\alpha / \tau)} \right\} \\
&= \frac{-1}{\tau |P(i)|} \left\{ \sum_{p \in P(i)} \boldsymbol{z}_p - \sum_{p \in P(i)} \sum_{p' \in P(i)} \boldsymbol{z}_{p'} \, P_{ip'} - \sum_{p \in P(i)} \sum_{n \in N(i)} \boldsymbol{z}_n \, P_{in} \right\} \\
&= \frac{-1}{\tau |P(i)|} \left\{ \sum_{p \in P(i)} \boldsymbol{z}_p - \sum_{p' \in P(i)} |P(i)| \boldsymbol{z}_{p'} \, P_{ip'} - \sum_{n \in N(i)} |P(i)| \boldsymbol{z}_n \, P_{in} \right\} \\
&= \frac{1}{\tau} \left\{ \sum_{p \in P(i)} \boldsymbol{z}_p \left(P_{ip} - \frac{1}{|P(i)|} \right) + \sum_{n \in N(i)} \boldsymbol{z}_n \, P_{in} \right\}
\end{aligned} \tag{10}$$

where $N(i) \equiv \{n \in A(i) : \tilde{\boldsymbol{y}}_n \neq \tilde{\boldsymbol{y}}_i\}$ is the set of indices of all negatives in the set of the augmented samples. A detailed visual representation of the second module of the proposed framework is illustrated in Figure 4.

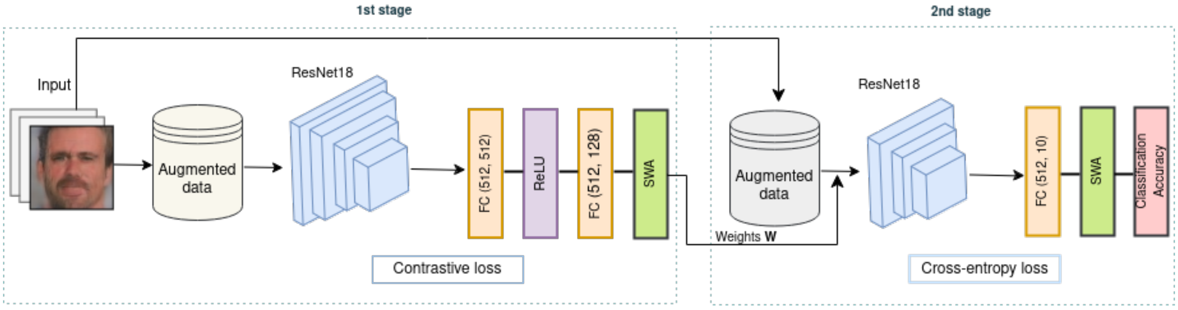


Figure 4. Learning procedure of the proposed framework.

4. Datasets

In applications such as the discrimination of CGIs from NIs, which degenerate into a binary classification problem, dataset selection acts crucially in the overall system accuracy. This is due to the fact that the network should be trained on incoming data that resemble real-life scenarios to achieve generalization. The need for proper and accurate dataset selection is becoming more apparent as more efficient and sophisticated methods and algorithms are released. The ability to handle more complex and challenging datasets, including CGIs and NIs, which are difficult to distinguish with the naked eye, is required in the most recent deep learning network-based methods. Here, we employ the DSTok [27], the Rahmouni [24], and the LSCGB [28] datasets, three datasets that are commonly used in the literature, to assess the performance of the proposed CoStNet. A set of challenging images of the DSTok dataset is depicted in Figure 5. Starting with the aforementioned datasets, we introduce additional ones appearing in the literature. The benchmark datasets are summarized in Table 1.

- **DSTok dataset** [27]: The DSTok dataset comprises a total of 4850 CGIs and 4850 NIs sourced from the Internet. NIs encompass diverse indoor and outdoor landscapes captured by various devices, while CGs exhibit photorealistic qualities. This collection boasts high-resolution images, ranging from 609 × 603 to 3507 × 2737, showcasing significant inter-class diversity. Such characteristics position the DSTok dataset as a pivotal resource for research in CG image detection, emphasizing its prominence in the literature.
- **Rahmouni's dataset** [24]: Rahmouni's dataset consists of 1800 high-resolution CGIs of size 1920 × 1080 pixels downloaded from the Level-Design Reference Database [43]. These CGIs were taken from photorealistic video games (i.e., Uncharted 4, Battlefield Bad Company 2, The Witcher 3, Battlefield 4, and Grand Theft Auto 5). Only these five distinct video games were deemed to exhibit a sufficient level of photorealism and thus they were employed. On the other hand, 1800 high-resolution NIs with a size of 4928 × 3264 pixels were obtained from the RAISE dataset [44] comprising a diverse array of settings, including outdoor and indoor scenes such as monuments, houses, landscapes, people bodies and faces, and forests.
- **LSCGB dataset** [28]: It is one of the most recent datasets. Its size is orders of magnitude larger than that of the preceding datasets. It consists of 71,168 CGIs and 71,168 NIs. It is characterized by high diversity and small bias regarding the distribution of color, tone, brightness, and saturation.
- He's dataset [22]: He's dataset consists of 6800 CGIs downloaded from the Internet. The images were created using a variety of rendering software packages, such as Maya, AutoCAD, etc. Another 6800 NIs were included in the dataset, which were captured under various indoor and outdoor circumstances. All images were stored in jpeg format, and their size ranges from 266 × 199 to 2048 × 3200.
- Columbia dataset [45]: The Columbia dataset consists of four sets of 800 images, resulting in a total of 3200 images. It consists of 800 NIs captured using the professional single-lens reflex Canon 10D and Nikon *D*70. These images demonstrate content diversity regarding indoor and outdoor scenes, various lighting conditions, etc. Another 800 NIs were retrieved from the Internet using Google Image Search based on keywords matching the CGI set's categories. A total of 800 CGIs were downloaded from the Internet. The images were classified based on their content, such as nature, objects, architecture, etc. Various rendering software packages were employed to create them. Another 800 CGIs were recaptured from the monitor while displaying the set of 800 previous CGIs.

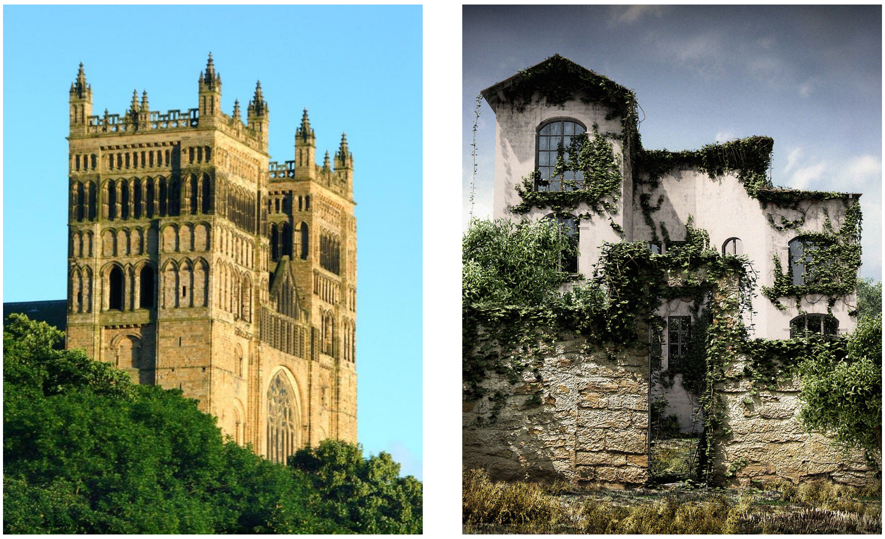


Figure 5. Sample images of DSTok dataset. On the (**left**), a natural image is depicted. On the (**right**), a computer-generated image is shown. It is difficult to determine that the image on the right is computer-generated with the naked eye.

Table 1. Benchmark datasets.

Dataset	# of CGIs	# of NIs	CGI Sources	NI Sources	Year
DSTok [27]	4850	4850	3D models	Photo-sharing websites	2013
Rahmouni [24]	1800	1800	3D models games	Existing benchmarks	2017
LSCGB [28]	71,168	71,168	Models, games, movies, GANs	Existing benchmarks, movies, photo-sharing websites	2020
He [22]	6800	6800	3D models	Personal collection	2018
Columbia [45]	1600	1600	3D models	Personal collection, Google Image Search	2005

5. Experimental Evaluation

5.1. Experimental Setup and Augmentations

CoStNet works effectively in real-life applications (code is available at https://github.com/geokarant/CoStNet, accessed on 28 January 2024).During the first stage, the network was trained for 100 epochs, employing a batch size = 200, while the second stage of the linear classifier was trained for 100 epochs employing a batch size = 20. The Stochastic Gradient Descent [46] was employed with a learning rate of 0.1 and 0.01 for the first and second stages, respectively. The cosine annealing scheduler was employed to adjust the learning rate during training. The maximum numbers of iterations were set to 100 and 20 and the minimum learning rates were set to 0.01 and 0.001 for the first and second stage, respectively. In evaluating the performance of the proposed framework, classification accuracy was utilized as a primary metric in accordance with the literature and measured using the formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{11}$$

where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives.

The proposed approach was implemented using the PyTorch 1.7.1 framework (https://pytorch.org/, accessed on 28 January 2024), and the hardware settings are indicated in Table 2.

Table 2. Hardware settings.

Details	Configuration
CPU	i9-7900X @ 3.3 GHz
GPU	RTX 2080 Ti
RAM	126 GB

Data augmentation positively affects the training procedure and contributes to the accurate classification of CGIs and NIs. Three CGIs were employed to operate as reference images for content and style semantics during the style transfer module preceding the CNN module. All NIs in the training set were passed through the style transfer module. Consequently, NIs' content manifold was aligned to the style manifold of CGIs, and a new set of CGIs was created to enhance the training procedure. Afterwards, a standard series of data augmentation procedures was applied to the dataset images. The input images were (i) randomly cropped and resized to 224 × 224 pixels; (ii) randomly rotated; (iii) randomly changed in brightness, contrast, and saturation; (iv) converted to grayscale with a probability of 0.2; and (v) normalized so that pixel values $\in [0,1]$. CoStNet was tested on various benchmark datasets, as described in Section 5.2, and several series of experiments were conducted, including parameters assessment and generalization ability (Section 5.3), robustness capability (Section 5.4), style transfer module impact contribution (Section 5.5), and statistical significance evaluation (Section 5.6).

5.2. Evaluation Results on the Benchmark Datasets

To evaluate CoStNet for differentiating between CGIs and NIs, we initially employed the public benchmark DSTok [27] dataset. CoStNet was compared against state-of-the-art methods with respect to classification accuracy. The classification of 14 state-of-the-art methods is that reported in [19,28]. CoStNet achieved a classification accuracy of 97.11% exceeding by 1.05% the CGNet proposed in [19], which was based on transfer learning and attained an accuracy of 96.1%. The method proposed in [20] was lagging behind, reaching an accuracy of 95.30%. The four best-performing methods were concluded by the inclusion of the method in [23], which resulted in an accuracy of 95.02%. The accuracy reported herein was achieved after 100 epochs of training with a batch size equal to 200. The accuracies of all methods employed are listed in Table 3. It is worth mentioning that the first training stage provided a well-trained CNN yielding a high validation performance before representations were fed into the linear classifier in the second training stage. During the first training stage, the training loss decayed rapidly after approximately 10 epochs, resulting in high accuracy, as demonstrated by the experiments conducted in Section 5.3. The training loss of the CNN, as well as its validation accuracy, are plotted in Figures 6 and 7, respectively.

In the context of the Rahmouni dataset, CoStNet demonstrated remarkable efficacy, achieving a remarkable accuracy of 100.00%. The derived accuracy positions CoStNet as the leading method among the compared approaches, showcasing its prowess in distinguishing CGIs from NIs. Notably, CoStNet surpassed all other methods, including the closest contender Bai [28], which attained an accuracy of 99.94%. This substantial margin underscores the robustness of CoStNet in CGI detection, outperforming well-established methodologies, such as Meena [47], Zhang [21], Nguyen [48], and Huang [49], among others.

The proposed CoStNet achieves an accuracy of 89.91% on the benchmark LSCGB dataset, showcasing its robust performance in detecting CGIs. While CoStNet slightly lags behind the method proposed in [28], which holds the highest accuracy at 91.45%, the

1.54% difference is relatively modest in the broader context of CGIs detection. CoStNet's competitive standing underscores its effectiveness and reliability in addressing the challenges posed by the large-scale LSCGB dataset. Notably, CoStNet outperforms several other state-of-the-art methods listed in Table 3, positioning it as a strong contender for practical applications in image forensics. The subtle variations in accuracy underscore the competitiveness of both methods in tackling the intricacies of the LSCGB dataset. Moreover, the proposed CoStNet contributes to the diversity of high-performing algorithms, offering a viable alternative for practitioners in need of reliable image forensics tools.

The accuracies achieved by CoStNet on the benchmark DSTok and Rahmouni datasets positions it as a cutting-edge solution in image manipulation detection. While facing a slightly more competitive landscape on the LSCGB dataset, CoStNet remains at the forefront of advancements in this domain, contributing significantly to the state-of-the-art methodologies in the field of CGIs detection.

Table 3. Detection accuracy (%) of state-of-the-art methods on benchmark datasets. Accuracies with † were obtained from [19], while the rest were obtained from [28].

Algorithms	**DSTok [27]**	**Rahmouni [24]**	**LSCGB [28]**
Rahmouni [24]	75.49 †	85.39	77.45
Quan [31]	93.74 †	90.49	82.80
Yao [30]	93.35 †	92.93	82.91
Gando [50]	85.50 †	-	-
De Rezende [23]	95.02 †	-	-
He [22]	91.58 †	-	-
Quan [20]	95.30 †	-	-
Zhang [21]	91.97 †	99.72	90.42
Chawla [51]	85.11	94.46	77.12
Nguyen [48]	94.42	99.71	90.02
Huang [49]	94.24	99.56	90.18
Meena [47]	93.65	99.70	90.09
Yao [19]	96.10 †	-	-
Bai [28]	96.35	99.94	**91.45**
CoStNet (Proposed)	**97.11**	**100.00**	89.91

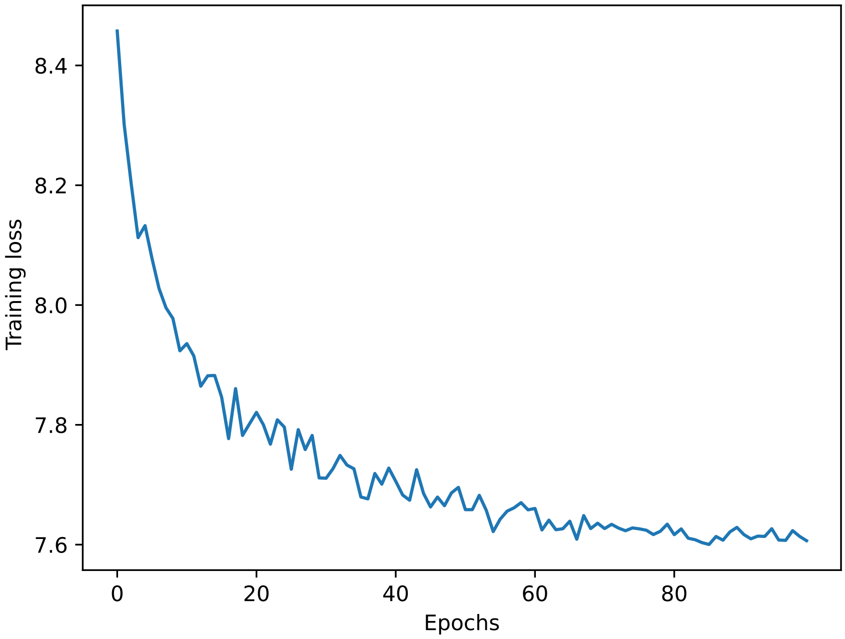


Figure 6. Training loss versus epochs during the first training stage of CoStNet on the DSTok dataset.

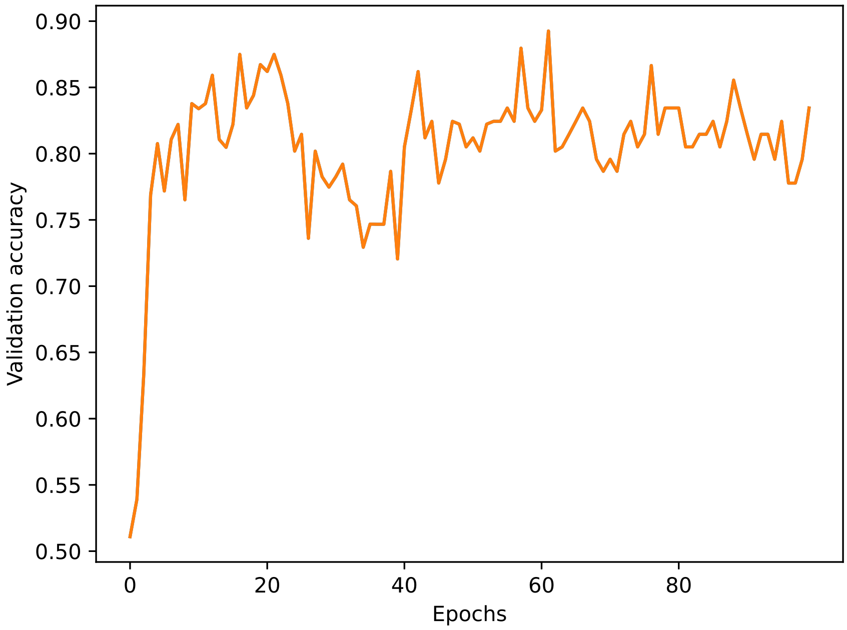


Figure 7. Validation accuracy versus epochs during the first training stage of CoStNet on the DSTok dataset.

5.3. Parameters' Assessment and Generalization Ability

The performance of CoStNet depends on various parameters. An extensive study was performed to apprehend which parameters affect it. A series of experiments was carried out to investigate how the number of epochs affects the detection results in the two stages. During the experiments, the batch size was fixed to 200. A top accuracy of 97.11% was measured when 100 epochs were employed in both stages. It is worth noting that an accuracy of 96.63% was achieved after 20 training epochs in both stages, outperforming the method proposed in [19]. It is also interesting to note that after only 10 training epochs, the proposed approach derived an accuracy of 95.81%, which, although slightly inferior to that reported in [19], is still rated as the second-best method. The accuracy results for various epochs are listed in Table 4.

Table 4. Detection accuracy (%) for various numbers of training epochs on the DSTok dataset.

Epochs	10	20	30	50	60	70	100
Accuracy	95.81	96.63	96.91	96.83	96.89	96.92	**97.11**

A second set of experiments was conducted to assess the contribution of the batch size in the classification accuracy. Various values for batch size were tested, while the number of epochs was kept fixed at 100. The result in accuracy for a batch size of 250 still remained above the 97% bound. For batch sizes smaller than 250, the accuracy was above 96%, demonstrating that CoStNet is not severely affected by the batch size. It is noteworthy that when a small batch size of 20 samples was employed, an accuracy of 96.75% was measured, still outperforming the state-of-the-art method [19] and demonstrating the classification ability of CoStNet. The accuracy results for various batch sizes are listed in Table 5.

Table 5. Detection accuracy (%) for various batch sizes on the DSTok dataset.

Batch size	20	30	50	60	70	100	250
Accuracy	96.75	96.59	96.83	96.56	96.47	96.81	**97.03**

A very important aspect of the model is related to its generalization ability, i.e., the proficiency to accurately classify unfamiliar data derived from a diverse array of setups. In pursuit of this, we harnessed the prowess of our trained model on the DSTok dataset and subjected it to rigorous evaluation on the well-known Rahmouni dataset, transcending boundaries with cross-dataset testing. In Table 6, the accuracies of CoStNet trained on the DSTok dataset and tested on Rahmouni's test set are summarized. We present the garnered accuracies of the CoStNet model trained on the DSTok dataset, rigorously tested on Rahmouni's distinguished test set. While the accuracy of CoStNet registers at 73.67%, it takes its place as the third top-performing contender, maintaining its stature even amid more formidable challenges.

The discrepancy between the top ranking within the DSTok dataset and the subsequent third-place position in the cross-dataset testing illuminates a significant challenge in deep learning methodologies: their susceptibility to dataset variations. The substantial disparity between the performance metrics achieved within the DSTok dataset, where the proposed method and the majority of the models surpassed the 90% accuracy threshold, and the diminished performance observed on Rahmouni's dataset underscores the considerable impact of dataset dissimilarities. Rahmouni's dataset presents limitations due to its divergent stylistic attributes, diverse content structures, and potentially distinct contextual elements compared to the DSTok dataset. These disparities extend beyond quantitative differences and significantly affect model adaptability when confronting unfamiliar data distributions.

The principal issue lies in the models' challenge in generalizing effectively across dissimilar datasets. Despite exhibiting commendable performance within the familiar confines of the DSTok training data, the noticeable deterioration in detection accuracy across all models in the cross-dataset assessment highlights the substantial divergence between the datasets, emphasizing the critical need to fortify models against such variations.

Table 6. Detection accuracy (%) in cross-dataset testing. State-of-the-art methods and the proposed CoStNet are trained on the DSTok dataset and tested on Rahmouni's dataset.

Algorithms	Rahmouni's Dataset
Rahmouni [24]	60.85
Quan [31]	56.43
Yao [30]	78.37
Gando [50]	67.48
De Rezende [23]	73.00
He [22]	56.78
Zhang [21]	61.78
Quan [20]	59.36
Yao [19]	**82.41**
CoStNet (Proposed)	73.67

In the pursuit of enhancing CoStNet's generalization capabilities, the model was systematically trained on the complex LSCGB dataset and subsequently evaluated through cross-dataset testing on the DSTok dataset. The results demonstrate a significant performance milestone, with CoStNet surpassing established state-of-the-art methods by achieving a detection accuracy of 93.03%, as shown in Table 7. This denotes a substantial advancement compared to prior assessments on Rahmouni's dataset, affirming CoStNet's adaptability and resilience to diverse and challenging data distributions.

Notably, CoStNet achieves a detection accuracy of 93.03%, outperforming state-of-the-art algorithms. Compared to the highest-performing state-of-the-art algorithm, Bai [28], which attains an accuracy of 83.95%, CoStNet demonstrates a substantial improvement of 11.21%. Furthermore, when contrasted with the mean accuracy of the baseline methods (approximately 78.63%), CoStNet exhibits an impressive percentage increase of approximately 18.31%. These results underscore the notable efficacy of CoStNet in surpassing established algorithms, showcasing its proficiency in handling the cross-dataset challenges posed by the DSTok dataset.

CoStNet's robust generalization is evident when trained on challenging datasets, such as the LSCGB dataset. Exposure to increased complexity and diverse data modalities enhances the model's adaptability. This observed phenomenon highlights CoStNet's capacity to discern intricate patterns, facilitating adept generalization across diverse contexts. Systematic training on challenging datasets plays a crucial role in fortifying the model against overfitting and enabling it to capture underlying structures that transcend dataset-specific nuances. This scientific observation emphasizes the pragmatic utility of subjecting deep learning models to progressively complex training scenarios for enhanced real-world applicability.

Table 7. Detection accuracy (%) in cross-dataset testing. State-of-the-art methods and the proposed CoStNet are trained on the LSCGB dataset and tested on the DSTok one.

Algorithms	DSTok Dataset
Nguyen [48]	78.71
Huang [49]	80.78
Zhang [21]	72.57
VGG-19 [36]	77.16
Bai [28]	83.95
CoStNet (Proposed)	**93.03**

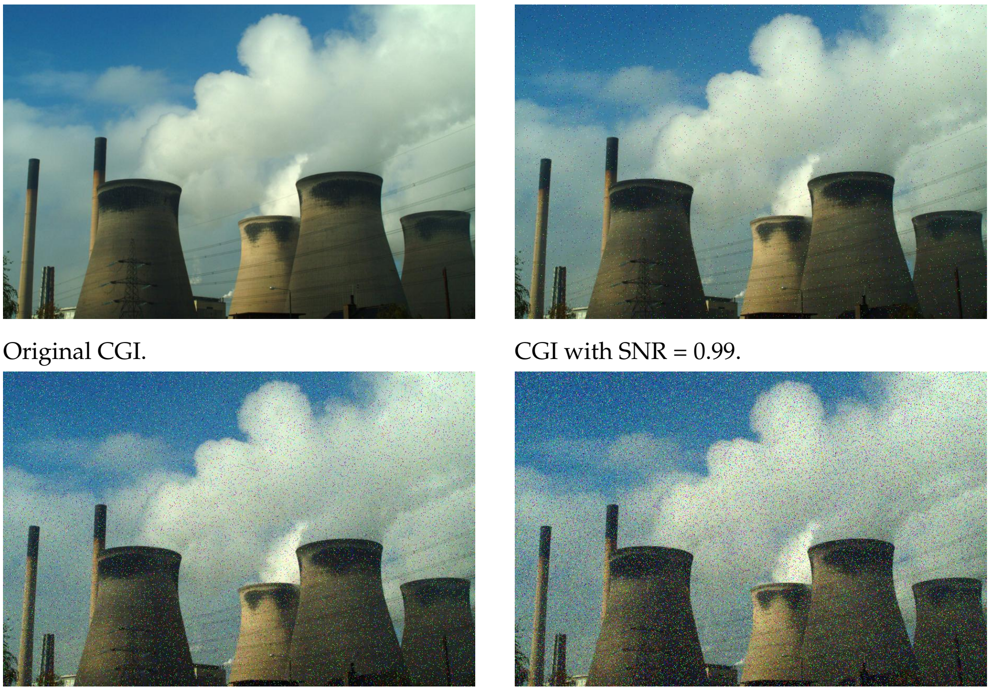
5.4. Robustness Capability

In the context of real-world digital forensics, an essential criterion for a comprehensive system is its ability to exhibit robustness against a spectrum of noise types and levels. To ascertain the effectiveness of our CoStNet model, we conducted a set of meticulously designed experiments, aligning with established literature, to facilitate a direct comparison with existing methods. This approach allowed us to evaluate the model's performance under diverse conditions.

Within the initial experimental set, we introduced salt-and-pepper noise to the test samples from the DSTok dataset, mirroring the methodology outlined in [19], while maintaining consistent signal-to-noise ratios (SNRs) at 0.99, 0.95, and 0.9. The summary in Table 8 provides a quantitative overview of the outcomes. An example of injected noise is depicted in Figure 8. When the test samples were infected with salt-and-pepper noise with SNR = 0.99, the proposed approach yielded an accuracy of 95.20% demonstrating its potential, while the method proposed in [19] was lagging behind with an accuracy of 93.08%. When the SNR of the injected noise was decreased to 0.95, CoStNet achieved an accuracy of 92.97%, outperforming its competitors. Even in the most challenging condition of SNR 0.9, CoStNet maintained an accuracy of 90.23%, notably exceeding the second-best method's accuracy of 82.38% [19]. It is worth mentioning that 6 out of 10 approaches achieved an accuracy of about 50%, while their accuracy exceeded 90% in the original experiments. For example, the detection accuracy of the Quan [20] algorithm exhibited a decrement from 95.30% to 55.58% subsequent to exposure to salt-and-pepper noise, indicating a susceptibility to perturbations mirroring real-world scenarios, thus reflecting a limited robustness under such conditions. Similar behavior was noticed also in Rahmouni [24], Yao [30], He [22], Zhang [21], and Quan [31]. Amid the complexities of noise interference, CoStNet emerges as an exemplar of adaptability, underscoring its potential to thrive under demanding real-world conditions.

Table 8. Detection accuracy (%) after salt-and-pepper noise attack on the test images.

Algorithms	SNR = 0.99	SNR = 0.95	SNR = 0.9
Rahmouni [24]	52.59	51.36	50.73
Quan [31]	50.27	50.02	49.99
Yao [30]	47.96	45.44	50.00
Gando [50]	79.01	70.52	64.53
De Rezende [23]	92.19	86.63	80.55
He [22]	50.18	50.01	50.05
Zhang [21]	57.50	52.67	51.93
Quan [20]	55.58	48.79	49.35
Yao [19]	93.08	88.91	82.38
CoStNet (Proposed)	**95.20**	**92.97**	**90.23**



Original CGI.

CGI with SNR = 0.99.

CGI with SNR = 0.95.

CGI with SNR = 0.9.

Figure 8. Original CGI and the CGIs altered by injecting salt-and-pepper noise at various SNRs.

A subsequent series of experiments was undertaken, involving the introduction of Gaussian noise. Drawing inspiration from the methodology outlined in [19], we set the Gaussian noise's mean value to 0 while maintaining a signal-to-noise ratio (SNR) of 0.7. This experimental protocol also entailed the exploration of three distinct standard deviations (SD) for the noise, specifically 10, 30, and 50. The results of detection accuracy are presented in Table 9 and elucidate the model's performance across varying degrees of Gaussian perturbations. When SD = 10, the proposed approach yielded an accuracy of 66.13%, being placed in the fifth place with respect to accuracy. The top-performing approach was the method proposed in [23]. When the SD was increased to 30, CoStNet was rated as the fourth top-performing out of the ten methods with an accuracy of 62.03%. Notably, when the SD was increased to 50, CoStNet resulted in an accuracy of 60.97% ranked as the third best performing method. This fact demonstrates that the greater SD of noise, the better the ranking of the proposed method. When the SD increases, De Rezende's [23] approach maintains a high detection accuracy of 96.63%, demonstrating its robustness in such kind of attack. We argue that this occurrence stems from the preprocessing methodology utilized by this method. Such preprocessing involves the deduction of the mean RGB value of the ImageNet dataset from each pixel during the preprocessing phase. The detection accuracy when SD increases notably deteriorates, demonstrating that this form of attack profoundly impacts the overall performance of the models.

Relative deteriorations in accuracies across varying levels of Gaussian noise attacks reveal notable trends among the evaluated methods. De Rezende's approach showcased considerable vulnerability, experiencing a 19.27% deterioration from SD = 10 to SD = 30 and a total 30.21% decrease from SD = 10 to SD = 50. Similarly, the Yao [19] method exhibited significant susceptibility, with deteriorations of 9.61% from SD = 10 to SD = 30 and 18.50% from SD = 10 to SD = 50. In contrast, the proposed CoStNet method demonstrated relatively better robustness, showcasing deteriorations of 6.20% from SD = 10 to SD = 30 and 7.80% from SD = 10 to SD = 50. These observations underline the varying degrees of resilience among the evaluated algorithms against escalating levels of Gaussian noise, with De Rezende displaying the most pronounced sensitivity and CoStNet illustrating relatively improved stability in the face of increasing noise levels.

Table 9. Detection accuracy (%) after Gaussian noise attack on test images.

Algorithms	SD = 10	SD = 30	SD = 50
Rahmouni [24]	52.19	50.00	50.25
Quan [31]	52.95	49.66	48.91
Yao [30]	44.31	41.23	50.00
Gando [50]	75.00	65.08	57.50
De Rezende [23]	**96.63**	78.00	67.44
He [22]	72.38	57.47	54.41
Zhang [21]	54.64	50.30	49.12
Quan [20]	51.39	50.24	49.12
Yao [19]	88.54	**80.03**	**72.16**
CoStNet (Proposed)	66.13	62.03	60.97

5.5. Impact of Style Transfer (Ablation Study)

The proposed CoStNet benefits from the style transfer module, which acts in a complementary manner, enriching the training procedure with additional training CG samples. The research question that arises refers to the contribution of the style transfer module in cases of reduced training samples. Four different experiments were conducted in which the style transfer module had different quantitative contributions to training samples, as depicted in Figure 9. Specifically, 75%, 50%, 25%, and 10% of the initial CG training samples were randomly removed and replaced with the same percentages using the style transfer module in the DSTok dataset such that the original number of training samples remains unchanged. The best accuracy was observed when the style transfer module replaced 10% of the training samples with CGIs, reaching an accuracy of 97.09%, outperforming the CGNet [19], which derived an accuracy of 96.10%. When 25% of the original CG training samples were removed and replaced by the style module, CoStNet achieved an accuracy of 96.56%. When half of the training samples were randomly removed and replaced by the style transfer module, CoStNet reached an accuracy of 95.75%, performing accurately and being placed second. Finally, when only 25% of the original training samples were retained and the style transfer module completed the rest of the samples, CoStNet reached an accuracy of 94.72%. The results of the ablation study demonstrate the significant contribution of the style transfer module in achieving improved accuracy with reduced training samples. This finding supports the claim that incorporating the style transfer module into the proposed architecture can lead to more effective and accurate predictions. Such insights provide valuable guidance for future research in this area and suggest that the proposed approach has the potential to enhance the performance of a wide range of machine learning applications in the context of discriminating CGIs from NIs.

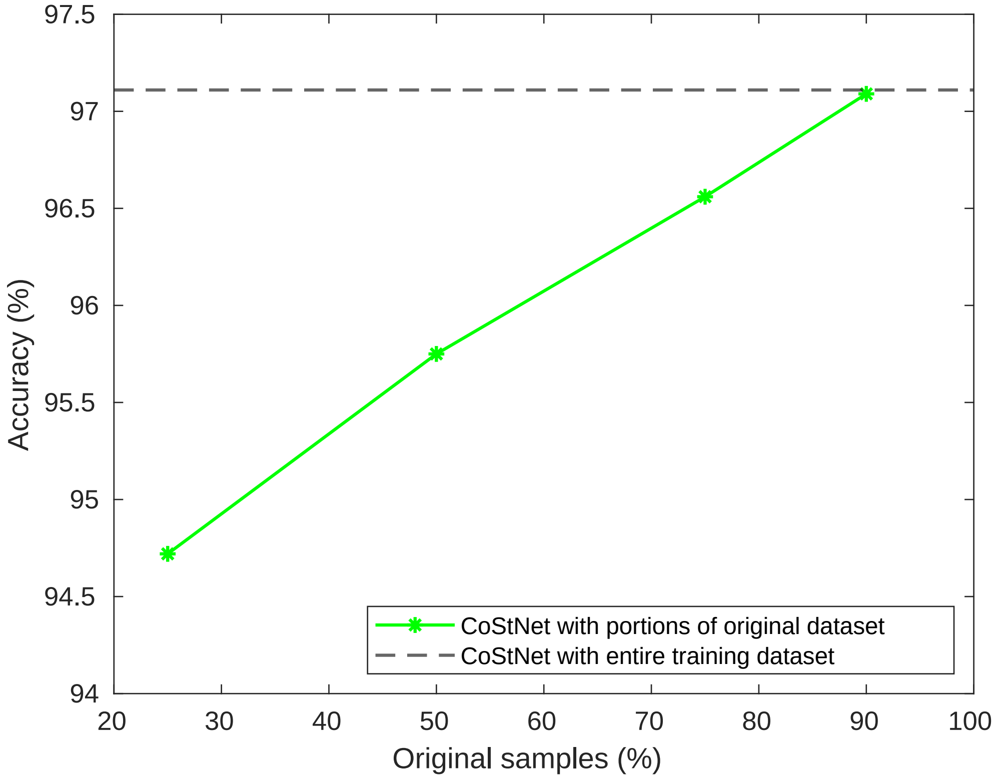


Figure 9. Accuracy of the proposed CoStNet on the DSTok dataset when several portions of the original dataset are retained.

5.6. Statistical Significance

It is of paramount importance to underscore a nuanced but impactful observation in the realm of accuracy disparities between our proposed method and the state-of-the-art methodology expounded in [19] on the DSTok dataset. Specifically, a marginal deviation of 1.01% is discernible when circumstances entail the integration of the style transfer module to augment the reservoir of training samples, as meticulously delineated in Section 5.2. Moreover, a commensurate distinction of 0.99% surfaces when the style transfer module assumes a pivotal role in replenishing 10% of the training set encompassing CGIs. This strategic recalibration is aimed at aligning with the numerical representation stipulated in [19].

This nuanced differential serves to accentuate the meticulous precision and unwavering stability inherent in our methodology. Furthermore, it serves as a testament to the method's remarkable consistency across diverse settings and methodologies for assimilating supplementary data. These insightful differentials aptly underscore the method's inherent robustness and reliability.

To check whether the accuracy difference of 0.99% is statistically significant, the approximate analysis in [52] is applied. The accuracies ϖ_1 and ϖ_2 are binomially distributed random variables. If $\hat{\varpi}_1, \hat{\varpi}_2$ denote the empirical accuracies, and $\overline{\varpi} = \frac{\hat{\varpi}_1+\hat{\varpi}_2}{2}$, the hypothesis $H_0 : \varpi_1 = \varpi_2 = \overline{\varpi}$ is tested at 95% level of significance. The accuracy difference has a variance of $\beta = 2\frac{\overline{\varpi}(1-\overline{\varpi})}{N}$, where N is the number of images. For $\zeta = 1.65\sqrt{\beta}$, if $|\hat{\varpi}_1 - \hat{\varpi}_2| \geq \zeta$, H_0 is rejected with a risk 5% of being wrong. Similarly, there is sufficient evidence to warrant the rejection of the claim that both the CGNet [19] and CoStNet methods attain the same accuracy. Accordingly, in 95% of repetitions of the experiment, CoStNet is expected to outperform CGNet [19]. The aforementioned analysis certifies that, in our case, the obtained $\zeta = 0.24\%$ indicates that the observed 0.99% accuracy difference between the proposed framework and the state-of-the-art CGNet reported in [19] is statistically significant.

The same procedure was employed to check whether the accuracy difference of 0.76% between the proposed CoStNet and the method presented in [28] when both were trained and tested on the DSTok dataset is statistically significant. In that case, the obtained $\zeta = 0.23\%$ indicates that the observed 0.76% accuracy difference between the proposed framework and the method reported in [28] is statistically significant. This analysis provides strong evidence that the performance enhancements achieved by our method over the state-of-the-art CGNet [19] and the method presented in [28] is not merely incidental, but rather statistically validated. This statistical rigor not only complements the empirical observations but also reinforces the credibility of claims regarding the method's effectiveness, lending support to the notion of the robustness and reliability of the proposed methodology.

5.7. Real-World Forensic Applications

The proposed CoStNet offers substantial practical utility in the age of deepfakes, with extensive applicability in real-world scenarios, notably within the domain of multimedia forensics. By accurately distinguishing between CGIs and NIs, the proposed framework offers enhanced capabilities to detect deepfakes. The ability to differentiate between CGIs and NIs is crucial for identifying manipulated or forged images, thus preserving the integrity of digital evidence and ensuring the reliability of forensic conclusions, indeed. Moreover, CoStNet's potential extends beyond forensic analysis to combating misinformation in digital media platforms. With the proliferation of manipulated images in social networks, the method provides a valuable tool for researchers and media professionals to detect and flag potentially deceptive content, thus safeguarding the integrity of information dissemination. Additionally, it holds promise for multimedia authentication applications. By accurately verifying the authenticity of digital images, the method enhances trust in photographic evidence, particularly in fields such as journalism where the credibility of visual content is paramount for reporting factual information. Furthermore, CoStNet's applications transcend traditional forensics to include fraud detection in various domains. In banking and document authentication, for instance, the method can be employed to verify the authenticity of scanned signed documents, thereby mitigating identity theft and financial fraud. By leveraging supervised contrastive learning and incorporating the style transfer module, the method pushes the boundaries of deep learning techniques, fostering innovation in multimedia analysis and paving the way for future research endeavors.

6. Conclusions, Limitations, and Future Directions

An end-to-end deep learning framework, denoted as CoStNet, has been introduced as a novel solution for the application of distinguishing NIs from CGIs. The innovation combines the principles of supervised contrastive learning, arbitrary style transfer, and the ResNet-18 architecture within a unique two-module framework. Through the integration of contrastive learning, CoStNet circumvents the necessity for hand-engineered features and adeptly captures intricate feature representations inherent in the training data, thereby enabling precise classification. Notably, the incorporation of the style transfer module extends the efficacy of training by enriching the dataset with an amplified array of negative samples beyond the confines of the original dataset. The robustness and efficacy of CoStNet are substantiated through a comprehensive series of experiments, leveraging the benchmark DSTok, Rahmouni, and LSCGB datasets. Furthermore, its prowess is evaluated in terms of both its generalization capacity and resilience through cross-dataset testing. By benchmarking CoStNet's detection accuracy against state-of-the-art methods, its competence is reaffirmed across various parameter configurations, encompassing batch sizes and epochs. Notably, even with modest training epochs and compact batch sizes, CoStNet emerges as an adept classifier, surpassing state-of-the-art methodologies. An in-depth ablation study elucidates the pivotal role played by the style transfer module, particularly in scenarios with constrained training data availability. Empirical results corroborate CoStNet's performance equivalence to state-of-the-art methods, while its superiority in distinguishing NIs from CGIs is underscored by a remarkable accuracy, outperforming the state-of-the-art approaches. Significantly, statistical tests substantiate the statistical significance of these performance enhancements.

While CoStNet demonstrates promising performance in distinguishing natural images NIs from CGIs in many different setups, there are some noteworthy limitations that warrant consideration. The disparities identified in Rahmouni's dataset, which is characterized by divergent stylistic attributes, diverse content structures, and potentially distinct contextual elements compared to the DSTok dataset, extend beyond quantitative differences and significantly impact the adaptability of the CoStNet model when faced with unfamiliar data distributions. Consequently, cross-dataset testing, especially when CoStNet was trained on the DSTok dataset and tested on Rahmouni's dataset, poses significant challenges for model generalization. Additionally, the framework is not robust in the presence of Gaussian noise during testing, leading to performance deterioration compared to scenarios involving salt-and-pepper noise. These limitations underscore the need for future research aimed at enhancing the framework's efficacy and resilience in practical applications.

The effectiveness and efficiency demonstrated by the proposed framework chart a compelling trajectory for future research endeavors. A pursuit to enhance the robustness of the framework is evident, aiming to address inherent limitations. This could involve the development of more robust CNN architectures tailored to handle even more diverse datasets and noise conditions. Advanced noise reduction techniques or regularization methods could be explored to improve the model's resilience to Gaussian noise. Additionally, investigating transfer learning strategies may enhance model generalization across different datasets, ultimately advancing the framework's applicability in real-world scenarios. Moreover, there is an imperative drive towards the development and integration of lightweight models, a strategic approach poised to tackle real-world temporal constraints and to cater to applications necessitating near real-time operation.

Author Contributions: Conceptualization, G.K. and C.K.; methodology, G.K.; software, G.K.; validation, G.K.; formal analysis, G.K. and C.K.; investigation, G.K.; resources, G.K. and C.K.; data curation, G.K.; writing—original draft preparation, G.K.; writing—review and editing, G.K. and C.K.; visualization, G.K.; supervision, C.K.; project administration, C.K.; funding acquisition, G.K. and C.K. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI PhD Fellowship grant (Fellowship Number: 820).

Data Availability Statement: Raw data for DSTok dataset are not publicly available. They can be provided upon request from the authors of the original paper [27]. The LSCGB dataset is publicly available at https://github.com/wmbai/LSCGB, accessed on 12 February 2023. Rahmouni's dataset can be found in https://github.com/NicoRahm/CGvsPhoto, accessed on 7 January 2023.

Acknowledgments: The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI PhD Fellowship grant (Fellowship Number: 820).

H.F.R.I.
Hellenic Foundation for
Research & Innovation

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

References

1. Autodesk A360 Rendering Gallery. Available online: https://gallery.autodesk.com/a360rendering/ (accessed on 24 January 2020).
2. Artlantis Gallery. Available online: https://artlantis.com/en/gallery/ (accessed on 24 January 2020).
3. Learn VRay. Available online: https://www.learnvray.com/fotogallery/ (accessed on 24 January 2020).
4. Corona Renderer Gallery. Available online: https://corona-renderer.com/gallery (accessed on 24 January 2020).
5. Yang, P.; Baracchi, D.; Ni, R.; Zhao, Y.; Argenti, F.; Piva, A. A survey of deep learning-based source image forensics. *J. Imaging* **2020**, *6*, 9. [CrossRef]
6. Mazumdar, A.; Bora, P.K. Siamese convolutional neural network-based approach towards universal image forensics. *IET Image Process.* **2020**, *14*, 3105–3116. [CrossRef]
7. Goel, N.; Kaur, S.; Bala, R. Dual branch convolutional neural network for copy move forgery detection. *IET Image Process.* **2021**, *15*, 656–665. [CrossRef]

8. Rhee, K.H. Detection of spliced image forensics using texture analysis of median filter residual. *IEEE Access* **2020**, *8*, 103374–103384. [CrossRef]
9. Chang, H.; Yeh, C. Face anti-spoofing detection based on multi-scale image quality assessment. *Image Vis. Comput.* **2022**, *121*, 104428. [CrossRef]
10. Matern, F.; Riess, C.; Stamminger, M. Gradient-based illumination description for image forgery detection. *IEEE Trans. Inf. Forensics Secur.* **2019**, *15*, 1303–1317. [CrossRef]
11. Chen, J.; Liao, X.; Qin, Z. Identifying tampering operations in image operator chains based on decision fusion. *Signal Process. Image Commun.* **2021**, *95*, 116287. [CrossRef]
12. Zhang, X.; Sun, Z.; Karaman, S.; Chang, S. Discovering image manipulation history by pairwise relation and forensics tools. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 1012–1023. [CrossRef]
13. Carvalho, T.; Faria, F.; Pedrini, H.; Torres, R.; Rocha, A. Illuminant-based transformed spaces for image forensics. *IEEE Trans. Inf. Forensics Secur.* **2015**, *11*, 720–733. [CrossRef]
14. Wang, J.; Li, T.; Shi, Y.; Lian, S.; Ye, J. Forensics feature analysis in quaternion wavelet domain for distinguishing photographic images and computer graphics. *Multim. Tools Appl.* **2017**, *76*, 23721–23737. [CrossRef]
15. Peng, F.; Zhou, D.; Long, M.; Sun, X. Discrimination of natural images and computer generated graphics based on multi-fractal and regression analysis. *AEU Int. J. Electron. Commun.* **2017**, *71*, 72–81. [CrossRef]
16. Ng, T.; Chang, S.; Hsu, J.; Xie, L.; Tsui, M. Physics-motivated features for distinguishing photographic images and computer graphics. In Proceedings of the 13th Annual CM International Conference on Multimedia, Singapore, 28 November–30 December 2005; pp. 239–248.
17. Chen, W.; Shi, Y.Q.; Xuan, G. Identifying computer graphics using HSV color model and statistical moments of characteristic functions. In Proceedings of the IEEE International Conference on Multimedia and Expo, Beijing, China, 2–5 July 2007; pp. 1123–1126.
18. Zhang, R.; Wang, R.; Ng, T. Distinguishing photographic images and photorealistic computer graphics using visual vocabulary on local image edges. In *Digital Watermarking Techniques in Curvelet and Ridgelet Domain*; Springer: Cham, Switzerland, 2011; pp. 292–305.
19. Yao, Y.; Zhang, Z.; Ni, X.; Shen, Z.; Chen, L.; Xu, D. CGNet: Detecting computer-generated images based on transfer learning with attention module. *Signal Process. Image Commun.* **2022**, *105*, 116692.
20. Quan, W.; Wang, K.; Yan, D.M.; Zhang, X.; Pellerin, D. Learn with diversity and from harder samples: Improving the generalization of CNN-Based detection of computer-generated images. *Forensic Sci. Int. Digit. Investig.* **2020**, *35*, 301023. [CrossRef]
21. Zhang, R.; Quan, W.; Fan, L.; Hu, L.; Yan, D. Distinguishing computer-generated images from natural images using channel and pixel correlation. *J. Comput. Sci. Technol.* **2020**, *35*, 592–602. [CrossRef]
22. He, P.; Jiang, X.; Sun, T.; Li, H. Computer graphics identification combining convolutional and recurrent neural networks. *IEEE Signal Process. Lett.* **2018**, *25*, 1369–1373. [CrossRef]
23. De Rezende, E.R.; Ruppert, G.C.; Theophilo, A.; Tokuda, E.K.; Carvalho, T. Exposing computer generated images by using deep convolutional neural networks. *Signal Process. Image Commun.* **2018**, *66*, 113–126. [CrossRef]
24. Rahmouni, N.; Nozick, V.; Yamagishi, J.; Echizen, I. Distinguishing computer graphics from natural images using convolution neural networks. In Proceedings of the IEEE Workshop on Information Forensics and Security (WIFS), Rennes, France, 4–7 December 2017; pp. 1–6. [CrossRef]
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
26. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: New York, NY, USA, 2020; Volume 33, pp. 18661–18673.
27. Tokuda, E.; Pedrini, H.; Rocha, A. Computer generated images vs. digital photographs: A synergetic feature and classifier combination approach. *J. Visual Commun. Image Repres.* **2013**, *24*, 1276–1292.
28. Bai, W.; Zhang, Z.; Li, B.; Wang, P.; Li, Y.; Zhang, C.; Hu, W. Robust texture-aware computer-generated image forensic: Benchmark and algorithm. *IEEE Trans. Image Process.* **2021**, *30*, 8439–8453. [CrossRef]
29. Lyu, S.; Farid, H. How realistic is photorealistic? *IEEE Trans. Signal Process.* **2005**, *53*, 845–850. [CrossRef]
30. Yao, Y.; Hu, W.; Zhang, W.; Wu, T.; Shi, Y. Distinguishing computer-generated graphics from natural images based on sensor pattern noise and deep learning. *Sensors* **2018**, *18*, 1296. [CrossRef]
31. Quan, W.; Wang, K.; Yan, D.; Zhang, X. Distinguishing between natural and computer-generated images using convolutional neural networks. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2772–2787. [CrossRef]
32. Tariang, D.B.; Senguptab, P.; Roy, A.; Chakraborty, R.S.; Naskar, R. Classification of Computer Generated and Natural Images based on Efficient Deep Convolutional Recurrent Attention Model. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 146–152. [CrossRef]
33. He, P.; Li, H.; Wang, H.; Zhang, R. Detection of computer graphics using attention-based dual-branch convolutional neural network from fused color components. *Sensors* **2020**, *20*, 4743.
34. Meena, K.B.; Tyagi, V. Methods to distinguish photorealistic computer generated images from photographic images: A review. In Proceedings of the Advances and Applications in Computer Science, Electronics and Industrial Engineering, Ghazibad, India, 12–13 April 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 64–82. [CrossRef]

35. Ni, X.; Chen, L.; Yuan, L.; Wu, G.; Yao, Y. An evaluation of deep learning-based computer generated image detection approaches. *IEEE Access* **2019**, *7*, 130830–130840.
36. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. [CrossRef]
37. Luo, X.; Han, Z.; Yang, L. Progressive Attentional Manifold Alignment for Arbitrary Style Transfer. In Proceedings of the Asian Conference on Computer Vision, Macao, China, 4–8 December 2022; pp. 3206–3222.
38. Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D.; Wilson, A. Averaging weights leads to wider optima and better generalization. *arXiv* **2018**, arXiv:1803.05407.
39. Manning, C.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008.
40. Kolkin, N.; Salavon, J.; Shakhnarovich, G. Style transfer by relaxed optimal transport and self-similarity. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10051–10060.
41. Qiu, T.; Ni, B.; Liu, Z.; Chen, X. Fast optimal transport artistic style transfer. In Proceedings of the 27th International Conference on Multimedia Modeling, Prague, Czech Republic, 22–24 June 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 37–49.
42. Afifi, M.; Brubaker, M.; Brown, M. Histogan: Controlling colors of gan-generated and real images via color histograms. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 7941–7950.
43. Piaskiewicz, M. Level-Design Reference Database. 2017. Available online: http://level-design.org/referencedb (accessed on 24 January 2020).
44. Dang-Nguyen, D.; Pasquini, C.; Conotter, V.; Boato, G. RAISE: A raw images dataset for digital image forensics. In Proceedings of the 6th ACM Multimedia Systems Conference, Portland, OR, USA, 18–20 March 2015; pp. 219–224.
45. Ng, T.; Chang, S.; Hsu, J.; Pepeljugoski, M. Columbia photographic images and photorealistic computer graphics dataset. *ADVENT Technical Report*; Columbia University: New York, NY, USA, 2005; pp. 205–2004.
46. Amari, S. A theory of adaptive pattern classifiers. *IEEE Trans. Electron. Comput.* **1967**, *3* , 299–307.
47. Meena, K.B.; Tyagi, V. Distinguishing computer-generated images from photographic images using two-stream convolutional neural network. *Appl. Soft Comput.* **2021**, *100*, 107025. [CrossRef]
48. Nguyen, H.; Tieu, T.; Nguyen-Son, H.; Nozick, V.; Yamagishi, J.; Echizen, I. Modular convolutional neural network for discriminating between computer-generated images and photographic images. In Proceedings of the 13th International Conference on Availability, Reliability and Security, Hamburg, Germany, 27–30 August 2018; pp. 1–10. [CrossRef]
49. Huang, R.; Fang, F.; Nguyen, H.; Yamagishi, J.; Echizen, I. A method for identifying origin of digital images using a convolutional neural network. In Proceedings of the IEEE 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Auckland, New Zealand, 7–10 December 2020; pp. 1293–1299.
50. Gando, G.; Yamada, T.; Sato, H.; Oyama, S.; Kurihara, M. Fine-tuning deep convolutional neural networks for distinguishing illustrations from photographs. *Expert Syst. Appl.* **2016**, *66*, 295–301.
51. Chawla, C.; Panwar, D.; Anand, G.S.; Bhatia, M. Classification of computer generated images from photographic images using convolutional neural networks. In Proceedings of the IEEE 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Bangalore, India, 19–22 September 2018; pp. 1053–1057. [CrossRef]
52. Guyon, I.; Makhoul, J.; Schwartz, R.; Vapnik, V. What size test set gives good error rate estimates? *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 52–64.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.