sustainability

Article

Advancing Ancient Artifact Character Image Augmentation through Styleformer-ART for Sustainable Knowledge Preservation

Jamiu T. Suleiman and Im Y. Jung *

School of Electronic and Electrical Engineering, Kyungpook National University, Daegu 41566, Republic of Korea; jamiu.suleiman111@gmail.com

* Correspondence: iyjung@ee.knu.ac.kr

Abstract: The accurate detection of ancient artifacts is very crucial in recognizing and tracking the origin of these relics. The methodologies used in engraving characters onto these objects are different from the ones used in the modern era, prompting the need to develop tools that are accurately tailored to detect these characters. The challenge encountered in developing an object character recognition model for this purpose is the lack of sufficient data needed to train these models. In this work, we propose Styleformer-ART to augment the ancient artifact character images. To show the performance of Styleformer-ART, we compared Styleformer-ART with different state-of-the-art data augmentation techniques. To make a conclusion on the best augmentation method for this special dataset, we evaluated all the augmentation methods employed in this work using the Frétchet inception distance (FID) score between the reference images and the generated images. The methods were also evaluated on the recognition accuracy of a CNN model. The Styleformer-ART model achieved the best FID score of 210.72, and Styleformer-ART-generated images achieved a recognition accuracy with the CNN model of 84%, which is better than all the other reviewed image-generation models.

Keywords: imprinted ship characters; automatic recognition; recognition accuracy; dataset augmentation; machine learning classifiers

check for updates

Citation: Suleiman, J.T.; Jung, I.Y. Advancing Ancient Artifact Character Image Augmentation through Styleformer-ART for Sustainable Knowledge Preservation. *Sustainability* **2024**, *16*, 6455. https://doi.org/10.3390/su16156455

Academic Editor: Ripon Kumar Chakrabortty

Received: 2 June 2024
Revised: 11 July 2024
Accepted: 24 July 2024
Published: 28 July 2024

1. Introduction

The detection of characters engraved, embedded, or etched onto ancient artifacts is paramount in unraveling the mysteries of ancient civilizations. As shown in Figure 1, these characters serve as vital clues, enabling researchers to decipher the origins, meanings, and historical significance of artifacts. Epigraphy, the study of inscriptions directly engraved on durable materials like stone, pottery, and metal by ancient individuals, groups, and institutions, is fundamental to understanding historical texts. While thousands of inscriptions have survived to present times, many have suffered damage over the centuries, resulting in fragmented texts that are not easily discernible [1]. Additionally, inscriptions can be relocated or trafficked far from their original sites, making radiocarbon dating ineffective due to the inorganic nature of most inscribed materials and their automatic tracking requires a large amount of data [2,3]. Ancient artifacts and relics, often considered proprietary objects, are not readily available for extensive data collection and annotation, hindering the development of robust detection models. In this context, training machine learning models to accurately detect such characters poses a significant challenge due to the scarcity of labeled data [1,4]. In this research paper, we address the deficiency of labeled data for training object recognition models by leveraging data augmentation techniques to generate synthetic images [5,6]. By artificially expanding the available dataset, we aimed to enhance the performance and robustness of machine learning algorithms in detecting characters on ancient artifacts.

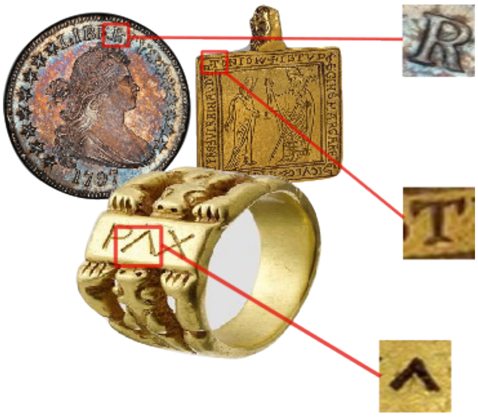

Figure 1. Extracted characters from ancient artifacts—ring, coin, and a tablet.

Our study proposes the use of synthetically generated images to alleviate the problem of the scarcity of data. In this study, we introduced a new model, Styleformer-ART, to augment the ancient artifact character images. We developed our model specifically for generating synthetic images tailored to augment the training data for artifact character detection. We explored the efficacy of our model in enhancing model performance and generalization capabilities in the context of archaeological research and artifact analysis. Moreover, we investigated a critical question: does the classical generative adversarial network (GAN) outperform newer generations of GANs that utilize transformers and diffusion architecture in augmenting artifact character images? This comparative analysis seeks to shed light on the effectiveness of different synthetic image generation approaches in enhancing the detection accuracy of machine learning models, particularly in scenarios with limited labeled data and inherent challenges such as image blurriness and artifact degradation over time [5–9]. Through this research endeavor, we aim to contribute to the advancement of computational methods in archaeology and cultural heritage preservation. By harnessing the power of synthetic data generation techniques and innovative machine learning architectures, we strive to overcome the data scarcity inherent in the study of ancient artifacts and empower researchers with enhanced tools for artifact analysis, origin tracing, and historical interpretation [1,2]. The following points highlight the main findings and contributions of our research:

- We introduce Styleformer-ART—a transformer-based model: we modified the encoder architecture of the Styleformer model with StyleGAN2 to suit the intricate requirements involved in generating high-quality synthetic artifact character images;
- We demonstrate the use of augmented images for training a recognition model for ancient artifact characters, demonstrating how the use of augmented images can enhance the training and performance of recognition model, which was, in our case, a CNN model for artifact character recognition tasks;
- Through an evaluation, we demonstrate that our Styleformer-ART model outperforms all the other cutting-edge GAN models that are suitable for artifact character recognition.

In the subsequent sections of this paper, we provide a comprehensive overview of the methodologies employed, the experimental setup, the results analysis, and the implications of our findings in the domain of ancient artifact character detection and interpretation. We believe that this research will not only enrich scholarly discourse in archaeology but also improve the development of more robust and accurate computational tools for studying and preserving our rich cultural heritage.

2. Related Work

The task of character detection on ancient artifacts using machine learning techniques has garnered significant interest in recent years, driven by the need to automate and enhance the analysis of archaeological materials. In this section, we review relevant studies and methodologies that contribute to the understanding of character detection, data augmentation, and synthetic image generation within the context of artifact analysis.

2.1. Character Detection on Artifacts

Prior research in the field of archaeology and computer vision has explored various approaches to character detection on artifacts. For instance, Fontanella et al. [10] employed deep learning models for character recognition on historical steles, demonstrating the feasibility of applying convolutional neural networks (CNNs) to decipher ancient inscriptions. Similarly, Huang et al. [3] utilized object detection techniques to identify and classify characters on archaeological relics, highlighting the importance of the accurate localization and recognition of textual elements on fragmented surfaces. Yalin et al. [11] introduced the method of enhancing the denoising of ancient Chinese character images by adding four local branches to the global branch. The global branch captures overall noise, while local branches focus on specific regions near stroke structures, improving detail preservation. Simulated ancient document noise and various loss functions are used for adversarial training and optimizing the model to produce high-quality, authentic denoised images.

2.2. Data Augmentation Techniques

The scarcity of labeled data in archaeological studies necessitates the use of data augmentation to expand training datasets. Traditional augmentation methods, such as rotation, translation, and scaling, have been widely adopted [12]. Recent advancements in generative models, particularly GANs, have enabled the generation of synthetic images that closely resemble real artifacts, offering a promising avenue for enhancing model performance in character detection tasks. Studies have explored the effectiveness of synthetic image generation in various domains, including medical imaging and natural scene understanding. Ding et al. introduced CCGAN, a conditional GAN architecture for image-to-image translation, which has been adapted for generating characters [13,14]. Moreover, advancements in transformer-based models [15] and diffusion architectures [16] have shown promise in generating high-quality synthetic images, albeit with varying degrees of success in preserving artifact-specific details and textual characteristics. To address the question of whether classical GANs outperform newer generations of GANs in augmenting blurry artifact character images, recent studies have compared the performance of different GAN architectures. For example, Karras et al. [8] introduced StyleGAN2, an improved version of GANs that exhibits superior image fidelity and diversity compared to traditional GAN models. Park et al. introduced Styleformer, a transformer-based model that incorporated StyleGAN2 with the transformer architecture as the generator. This model achieved state-of-the-art performance [17,18]. However, the model performed poorly on the artifact character dataset.

2.3. Challenges and Opportunities

While the existing literature provides valuable insights into character detection and synthetic image generation, several challenges persist [1–3,19]. These include the preservation of artifact-specific textures and details, the mitigation of data imbalance in archaeological datasets, and the adaptation of machine learning techniques to accommodate varying levels of artifact degradation and deterioration over time. In this work, we modified the encoder architecture in the Styleformer [17] model to accurately capture the global structure of the limited artifact character dataset and proposed Styleformer-Art, which is tailored to generating augmented images for artifact character images.

3. Dataset

Curating a dataset of images of ancient artifacts with engraved or etched characters, as described in the given context, poses several significant challenges:

- The Scarcity of Data: The primary challenge lies in the limited availability of labeled images of ancient artifacts. Such data are not readily available in large quantities,

which is crucial for training machine learning models effectively [2,11]. Ancient artifacts with detailed and verified annotations are rare, making the collection process difficult and time-consuming;
- Historical and Geographical Diversity: The dataset encompasses artifacts from various historical periods and geographical locations, introducing significant variability in artifact types and character styles [1,11]. This diversity, while valuable for training robust models, complicates the collection process as it requires sourcing from a wide array of periods and locations;
- Condition of Artifacts: Many artifacts may be in poor condition due to age, leading to incomplete or degraded engravings. This necessitates careful selection and possibly even digital restoration efforts to ensure the data are usable for model training [20,21];
- Character Frequency: The restriction to the 10 most frequently encountered English characters highlights the difficulty in obtaining a balanced dataset with all English alphabet characters. The distribution of characters, as shown in Table 1, is uneven, with some characters being much more prevalent than others. This imbalance can affect the model's ability to generalize across all characters.

In this work, we curated a dataset consisting of images of ancient artifacts containing engraved or etched characters. Due to the limited availability of labeled data, we supplemented our dataset with publicly accessible archaeological image repositories from the Internet. The dataset encompasses artifacts from diverse historical periods and geographical locations, ensuring variability in artifact types and character styles. In this work, we focused on artifacts with engravings in the English language due to their availability in public repositories on the Internet. Due to the restricted availability of all English alphabet characters, our focus was on the 10 most frequently encountered English characters that were accessible. Table 1 illustrates the distribution of each of these 10 characters.

Table 1. Number of samples for each character.

Characters	Number of Samples
A	85
D	24
E	75
I	46
L	37
N	45
T	58
S	31
R	57
O	43

4. Methodology

Figure 2 highlights the various steps involved in our methodology. We collected the artifact images and processed them. This was followed by model design. We then generated images with the augmentation techniques using various cutting-edge GANs. For the final step, the augmented images were evaluated with the FID score and the accuracy of the CNN model when trained on the augmented images. A detailed discussion of each step is presented in the subsections below.

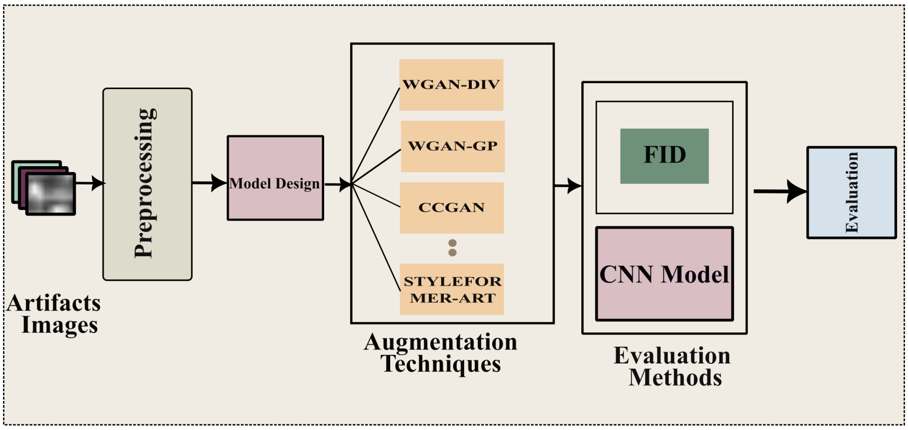

Figure 2. Overview of the steps involved in our approach.

4.1. Dataset Preparation

Before model training, as shown in Figure 3, we performed preprocessing steps to enhance the quality and interpretability of the artifact images. This involved noise reduction and contrast enhancement tailored to preserve textual details and enhance character visibility.

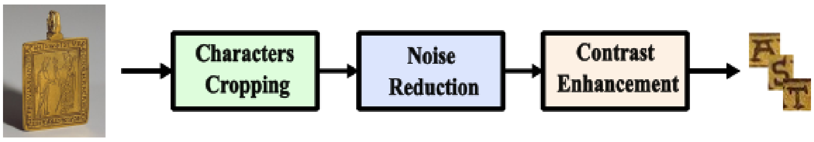

Figure 3. Preprocessing steps: cropping, noise reduction, and contrast enhancement to generate the character images.

Additionally, we conducted data normalization and standardization to ensure consistency in image features and facilitate model convergence during training.

4.2. Model Design

In this section, we discuss the architecture of our proposed Styleformer-ART model and the design of the CNN model used in the evaluation.

4.2.1. Styleformer-ART for Data Augmentation

Our image augmentation model consists of a generator and a discriminator established by Goodfellow et al. [5] and utilizes a transformer instead of a convolution neural network [22–24]. The Styleformer encoder network [17] serves as the fundamental block of Styleformer. As shown in Figure 4a, our generator is conditioned on a learnable constant input similar to the synthesis networks in existing StyleGAN models. However, unlike the conventional approach, the continuous input (originally 8 × 8) is flattened to 64 before entering the transformer-based encoder. Subsequently, this input is augmented with learnable positional encoding and then traverses through the Styleformer-Art encoder. Based on the Styleformer encoder [17], Styleformer-ART was implemented by modifying styleformer [17] to enhance its efficiency in artifact character image generation, as detailed in Figure 4.

The Styleformer [25] encoder implements a modified residual connection and scales the input feature map with a style vector (Mod Input in Figure 4b). In Attention Style Injection, unlike the vanilla GAN, StyleGAN2 utilizes layer-wise style vectors to enable controllable generation through style vectors. In the modulation process for self-attention, similar to the StyleGAN2 [8] style block, the input feature map in the Styleformer encoder is scaled by a style vector. However, unlike the convolution operation in StyleGAN2 [8], the self-attention operation involves two steps: the dot product of query and the key to create an attention map (i.e., the kernel), and the weighted sum of the value with the calculated attention map. The style vector applied to each operation step should be different; thus, style modulation is performed twice (Mod Input, Mod Value). Similarly, in demodulation for the query, key, and value, demodulation operations are necessary to remove the scaled effect of the input. Additionally, demodulation operations are applied to the encoder output to maintain unit standard deviation.

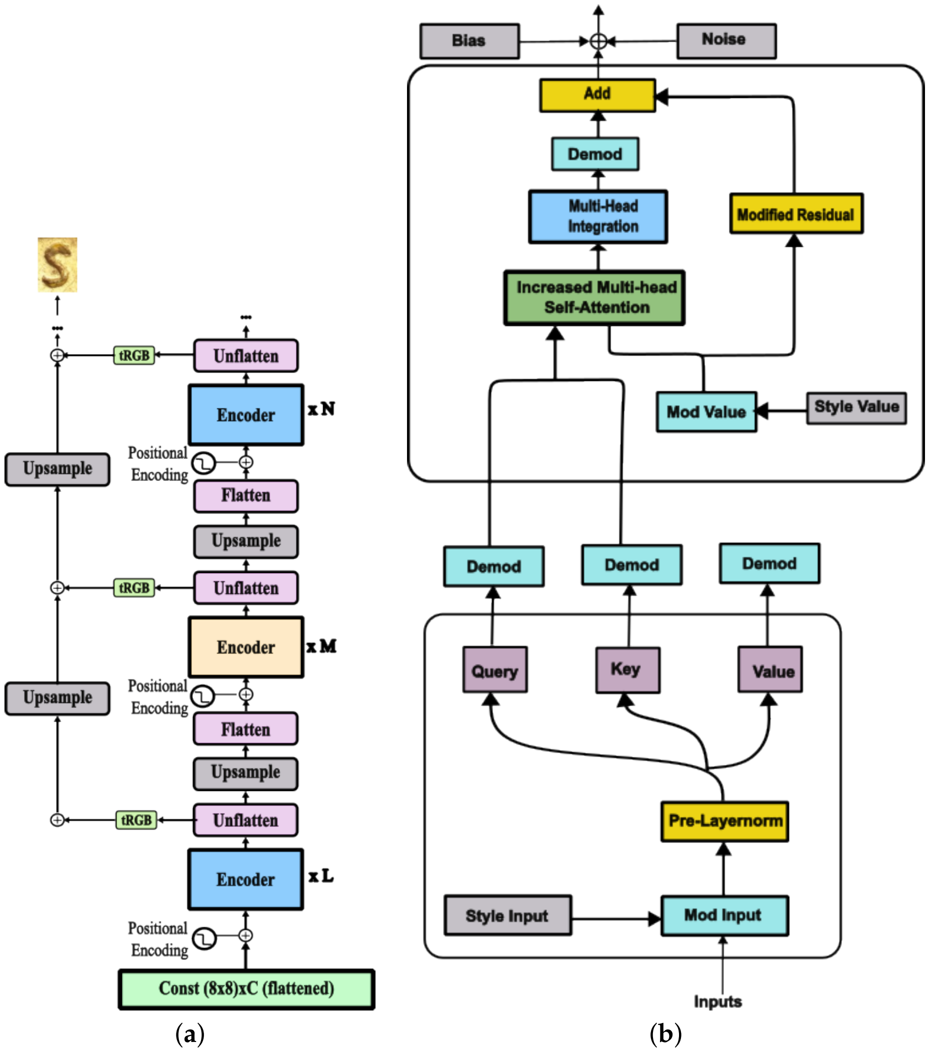

Figure 4. The architecture of the components of the Styleformer-ART generator: (**a**) the architecture of Styleformer-ART; (**b**) the Styleformer-ART encoder.

To address the efficiency problem with image resolution, Styleformer [17] introduces two techniques: the Linformer application and the combination of Styleformer [17] and StyleGAN2 [8]. Linformer reduces the time and space complexity from $O(n^2)$ to $O(nk)$, where k is fixed to 256, making it feasible for high-resolution image generation. Combining Styleformer and StyleGAN2 [8] serves as the discriminator and allows for generating extremely high-resolution images, leveraging the strengths of both models.

In our investigation, focusing on the domain of artifact character augmentation, characterized by low-resolution images (below 32 × 32 pixels), we posited that the utilization of Linformer may not yield a discernible improvement in image quality. To empirically validate this hypothesis, we conducted a comparative analysis of images generated using Styleformer with Linformer and Styleformer without Linformer. We named the resulting model without Linformer, "Styleformer-Art". We then conducted an ablation study to select the right hyperparameter for training the Styleformer-Art, as shown in Table 2. This study was conducted over three (3) characters: A, D, and E, and the average of the FID score was computed.

Table 2. Results of the ablation study on the Styleformer-Art generator architecture, which shows the performance of Linformer models with different depths and minimum heads.

Linformer	Depth	Minimum Heads	FID
✓	32	1	191
✗	32	1	141
✗	64	1	140
✓	64	1	198
✓	32	2	194
✗	32	2	165
✓	64	2	201
✗	64	2	175

As shown in Table 2, the configuration without Linformer outperformed all the other ones that incorporate Linformer. This observation from the ablation study corroborates our initial hypothesis, i.e., removing Linformer will result in better qualities of generated images.

Across all configurations, the removal of Linformer consistently resulted in lower (better) FID scores. For instance, at a depth of 32 and with one attention head, the FID score improved from 191 to 141 upon removing Linformer. Similarly, at a depth of 64 with one attention head, the FID score improved from 198 to 140 without Linformer.

Depth Variation: Both depths (32 and 64) showed improvements when Linformer was removed. This indicates that the positive impact of removing Linformer was consistent regardless of the model's depth. Notably, the lowest FID score (140) was achieved with a depth of 64 and one attention head, indicating that while increasing depth generally benefited the model, the exclusion of Linformer still held the key to better performance.

Number of Attention Heads: Increasing the minimum number of attention heads from one to two generally resulted in higher (worse) FID scores for both configurations with and without Linformer. However, even in these cases, the models without Linformer performed better. For example, with two heads and a depth of 32, the FID score improved from 194 to 165 without Linformer. This suggests that for low-resolution images, a simpler attention mechanism with fewer heads may be more effective.

The ablation study provides clear evidence that removing Linformer from the Styleformer architecture results in better performance for low-resolution image generation. This improvement is consistent across different depths and numbers of attention heads, suggesting that the benefits of excluding Linformer are robust and not dependent on specific hyperparameter settings. The outstanding performance of Styleformer-Art (without Linformer) can be attributed to the simplified architecture, which might be more suited to handle the low complexity of low-resolution images. The overhead introduced by Linformer, designed to handle high-dimensional data efficiently, may not be necessary for this domain and could introduce unnecessary computational complexity.

4.2.2. CNN Model for Data Recognition

As shown in Figure 5, the architecture of the CNN model, which is the most commonly used model in character recognition and artifact recognition [25–27], was developed to evaluate the performance of each image augmentation method. We chose it due to its simplistic nature and the conclusion made will be generalizable across other recognition models. The first layer has 64 filters/kernels of size 3 × 3 with ReLU activation, using a uniform initializer and takes input images of size 32 × 32 with one channel (grayscale). The subsequent layer has 128 filters of size 3 × 3 with ReLU activation and a uniform initializer. These layers perform max pooling with a pool size of 2 × 2 to downsample the spatial dimensions of the feature maps. The flatten layer flattens the 2D feature maps into a 1D vector to prepare for the fully connected (Dense) layers. The model was incorporated with a ReLU activation and He uniform initialization. The second Dense layer includes L2 regularization (with a penalty of 0.001) to help prevent overfitting. The final Dense layer has 10 units with SoftMax activation for multi-class classification (outputting probabilities of each class) since we were calculating on 10 classes. Dropout layers with a rate of 0.5 were added after the first and second Dense layers. The Dropout helps prevent overfitting by randomly setting a fraction of input units to zero during training [12,28,29]. Finally, the model was compiled with an Adam optimizer (with learning rate = 0.001) and the categorical cross-entropy loss function in Equation (1) below:

$$L = \sum_{i=1}^{N} \sum_{j=1}^{k} y_{ij} \log(p(ij)) \tag{1}$$

which penalizes the difference between the predicted and the true class, and the accuracy metric for evaluation during training.

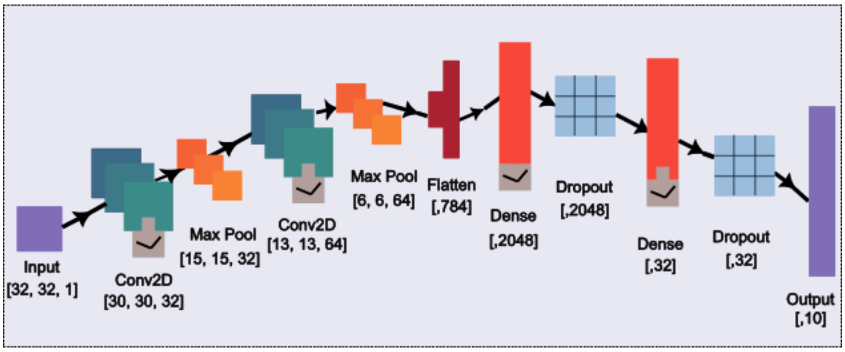

Figure 5. The architecture of the CNN model.

4.3. Data Augmentation

Given the scarcity of labeled data, we employed data augmentation techniques to expand our training dataset. We conducted an extensive review of both traditional augmentation methods, including rotation, translation, scaling, and flipping, and recent augmentation techniques, including the generative adversarial network, CCGAN [13], WGANGP [7], WGANDIV [30], and ACGAN [31], to introduce variability and increase the diversity of artifact images. These models were selected because they represent the image generation benchmark and are most suitable for character augmentation. Additionally, we explored the use of transformer-based models, and introduced Styleformer-ART for synthesizing realistic artifact-like character images from a limited dataset. We compared different GAN architectures, including the classical ACGAN [13,31], WGANDIV [30], WGANGP [8], and transformer-based GANs, with Styleformer to generate synthetic artifact images with varying levels of fidelity and textual clarity [17].

4.4. Experimental Setup

Our experiments were conducted using Pytorch Version 2.4 (a Python-based deep learning framework) on two Nvidia TITAN RTX GPU-accelerated computing platforms manufactured by Nvidia Inc. for model training and inference. Hyperparameter tuning and optimization were performed to maximize model performance and convergence. We employed a systematic approach to experimental design, ensuring the reproducibility and reliability of our results.

4.5. Evaluation Metrics

To assess the performance of the synthetic image generation model, we evaluated the performance of the generated images by FID and visual inspection. While evaluating our CNN models, we utilized standard evaluation metrics, including precision, recall, and F1-score, for character detection tasks. We conducted cross-validation experiments and evaluated model robustness across different artifact categories and degradation levels. Furthermore, we compared the effectiveness of synthetic image augmentation techniques in improving model generalization and reducing overfitting on limited labeled data.

5. Result and Evaluation

Figure 6 presents the original images, and the top-performing augmentation techniques: Styleformer-Art, WGANDIV, and WGANGP. Upon visual inspection, it becomes evident that the images generated by Styleformer-Art closely resemble the original images in terms of visual characteristics.

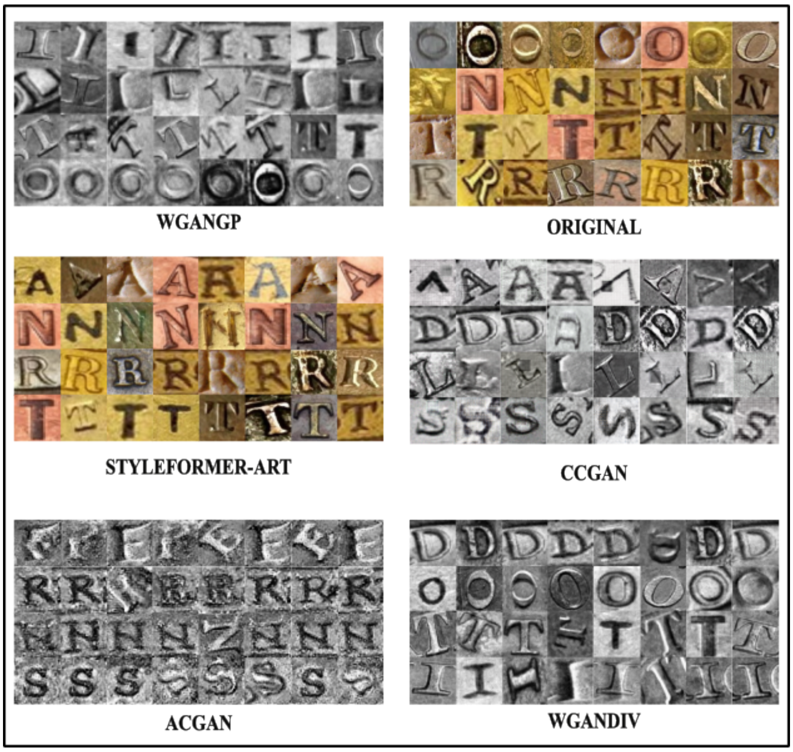

Figure 6. Visualization of the top-performing augmentation techniques.

We evaluated the generated images on the Frétchet inception distance (FID) score. The FID score for each augmentation method is provided in Table 3. The results of different synthetic image generation techniques for the task of detecting characters are associated with specific letters on artifacts [32–34]. The evaluation metric used is the FID, which measures the similarity between real and generated images based on their feature representations extracted from a pre-trained deep neural network (Inception Net) [33,35]. The FID between two multivariate Gaussians has a closed formula. The features were extracted from both the real and fake images by the inception network at the pool3 layer and were fitted to a Gaussian distribution and computed as follows:

$$\mathrm{FID_{avg}} = \frac{1}{N}\sum_{i=1}^{N}\left(\|\mu_{\mathrm{real},i} - \mu_{\mathrm{gen},i}\|^2 + \mathrm{Tr}\left(\Sigma_{\mathrm{real},i} + \Sigma_{\mathrm{gen},i} - 2\cdot\sqrt{\Sigma_{\mathrm{real},i}\Sigma_{\mathrm{gen},i}}\right)\right) \quad (2)$$

where N is the total number of the character categories, Tr is the trace of a matrix, and $N(\mu_{\mathrm{r}}, \Sigma_{\mathrm{r}})$ and $N(\mu_{\mathrm{g}}, \Sigma_{\mathrm{g}})$ are the Gaussian fitted to the real and generated images, respectively.

As shown in Table 3, ACGAN and CCGAN, consistently demonstrated lower FID scores, achieving an average of 459 and 363, respectively, indicating greater divergence from the real images across most characters. WGANDIV, WGANGP, and Styleformer-ART achieved average FID scores of 277.36, 274.29, and 210.72, respectively, indicating a closer resemblance of the generated images to the real image distributions. Notably, characters such as 'S' and 'D' exhibited substantially improved performance with Styleformer-ART, emphasizing the effectiveness of specific approaches for certain character types. Characters like 'A' and 'I' consistently displayed higher FID scores across all techniques, indicative of specific challenges associated with generating these characters faithfully. The range of FID scores, spanning from as low as 130 to as high as 460, reflects the diversity inherent in our dataset, a characteristic not typically encountered in larger datasets such as ImageNet [33,35,36]. This diversity prompted the exploration of alternative metrics for evaluating the fidelity of generated images.

Table 3. The FID score of the image generated by each synthetic image generation technique.

Characters	WGANGP [7]	WGANDIV [30]	ACGAN [31]	CCGAN [13]	Styleformer-ART
A	315.88	314.92	456.87	399.71	**146.88**
D	224.14	225.74	421.86	283.15	**106.95**
E	291.64	293.71	415.85	352.91	**155.80**
I	287.84	291.25	488.31	379.81	**248.59**
L	276.03	282.71	491.18	409.40	**215.04**
N	301.49	307.07	488.31	408.13	**291.00**
O	249.46	249.24	396.67	341.99	**216.54**
R	286.43	288.35	476.93	347.79	**135.83**
S	224.90	229.34	469.72	341.50	**131.50**
T	285.08	291.28	484.30	372.63	**267.30**
Average	274.29	277.36	459.0	363.70	**210.72**

To further assess the quality of the generated images, we employed a convolutional neural network (CNN) model trained on images generated using each augmentation technique. Figure 7 illustrates the loss trajectory of the CNN model trained on datasets augmented by various techniques. The graph indicates that the Styleformer-Art method attained the lowest validation loss, averaging 0.0060, with the WGANDIV technique yielding a close second, at an average loss of 0.0065. This further outlines the quality of the images generated by our augmentation model.

The loss graph in Figure 7 shows that Styleformer-Art significantly outperformed both the original dataset and the augmented dataset created with WGANDIV, WGANDIV, CCGAN, and ACGAN in terms of loss reduction over 30 epochs. Styleformer demonstrated a sharp and consistent decrease in loss from the beginning to the end of the training period, highlighting its superior learning capabilities and stability. In contrast, the original dataset, although it improved substantially over time, started with a much higher loss and did not achieve as low a final loss as the augmented dataset. This indicates that augmented artifact character images are more effective and efficient in learning compared to the original dataset.

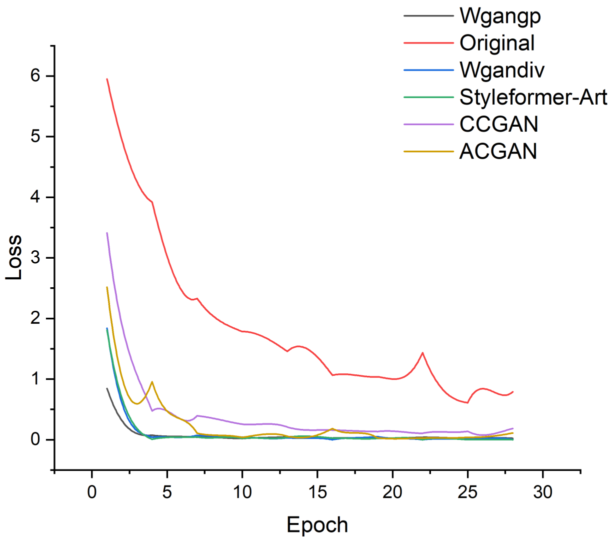

Figure 7. The validation loss graph of the CNN model for each augmentation technique.

6. Discussion

When comparing Styleformer-Art to the other augmentation methods, it is clear that Styleformer-Art maintained a distinct advantage. Techniques like WGANDIV and Wgangp also showed significant improvements in reducing loss over the epochs, but they experienced more variability and fluctuations. Similarly, CCGAN and ACGAN exhibited greater variability while demonstrating effective learning and did not achieve final loss values that were as low as those of Styleformer-Art. This suggests that, while these augmentation methods are beneficial, they are not as stable or as efficient as Styleformer-Art in minimizing loss.

Styleformer-Art's consistent performance and ability to maintain low loss values throughout the training period underscore its robustness and effectiveness. In contrast, WGANDIV, WGANDIV and CCGAN, despite their improvements, showed more fluctuations and higher final loss values. This comparative analysis highlights the superiority of Styleformer-Art in achieving minimal loss and optimal learning performance, making it the most preferable model among those evaluated.

Subsequently, Table 4 presents the performance of various data augmentation methods applied to the CNN model, evaluating the key metrics of accuracy, precision, recall, and F1-score. Initially, the CNN model's performance on the original dataset was notably poor, with all metrics reflecting suboptimal performance. The baseline accuracy, precision, recall, and F1-score were recorded at 0.23, 0.26, 0.23, and 0.10, respectively, indicating that the model struggled to learn from the unaugmented dataset. It is obvious from the table that, after transitioning to the augmented datasets, WGANDIV and WGANGP yielded substantial improvements. Both methods achieved an accuracy of 0.75, a precision of 0.77, and a recall of approximately 0.76–0.77. The F1 scores for these methods were also significantly enhanced, at 0.75 for WGANDIV and 0.76 for WGANGP. These results suggest that the GAN-based augmentation techniques effectively enhanced the model's capability to generalize from the data, contributing to improved performance across all the evaluated metrics.

Table 4. The performance of each augmentation method on the CNN model.

Metrics	Original Dataset	WGANDIV [30]	WGANGP [7]	ACGAN [31]	CCGAN [13]	Styleformer-ART
Accuracy	0.23	0.75	0.75	0.40	0.73	**0.84**
Precision	0.26	0.77	0.77	0.53	0.74	**0.83**
Recall	0.23	0.76	0.77	0.41	0.77	**0.84**
F1-Score	0.10	0.75	0.76	0.39	0.72	**0.82**

In comparison, ACGAN showed moderate success, with an accuracy of 0.40 and a precision of 0.53. Its recall stood at 0.41, its F1-score was slightly lower at 0.39. Although ACGAN performed better than the original dataset, it did not reach the performance levels seen with WGANDIV or WGANGP, indicating that while ACGAN could enhance the model's performance, it was less effective than other GAN variants.

CCGAN demonstrated notable results, with an accuracy of 0.73, a precision of 0.74, a recall of 0.77, and an F1-score of 0.72. These metrics indicate that CCGAN provided a well-rounded improvement in the model's performance, closely paralleling the effectiveness of WGANDIV and WGANGP. An F-test was carried out to determine if there was a significant difference in performance across the different augmentation methods. The test result, F (5, 18) = 145.01, $p < 0.01$, indicates that the performance varied across the different augmentation techniques. This is also evident in the heat map in Figure 8, which highlights the significant difference among the augmentation methods.

Similarly, the Nemenyi test was also carried out to investigate the differences further. The critical difference was 3.61, backing up the initial result from the previous test results. Table 5 shows the rank of each augmentation technique across the evaluation metrics. The rank sum of 4 for the Styleformer-ART indicates the model's superiority over the other techniques.

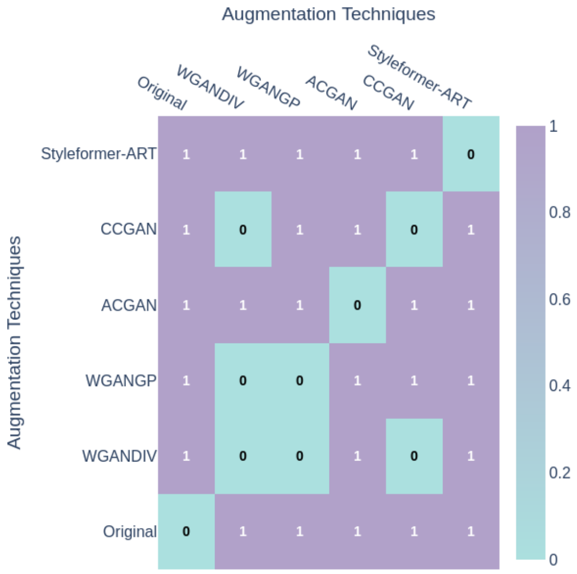

Figure 8. The pairwise comparison matrix of the augmentation techniques.

Table 5. The ranks of each augmentation technique across the evaluation metrics.

Metrics	Original Dataset	WGANDIV [30]	WGANDIV [7]	ACGAN [31]	CCGAN [13]	Styleformer-ART
Accuracy	6.0	2.5	2.5	5.0	4.0	**1.0**
Precision	6.0	2.5	2.5	5.0	4.0	**1.0**
Recall	6.0	4.0	2.5	5.0	2.5	**1.0**
F1-Score	6.0	3.0	2.0	5.0	4.0	**1.0**

Our model, the Styleformer-ART method, emerged as the most effective augmentation technique. It achieved the highest scores across all metrics: an accuracy of 0.84, a precision of 0.83, a recall of 0.84, and an F1-score of 0.82. These results signify that Styleformer-ART not only improved the CNN model's performance but also did so more efficiently than the other methods. The superior performance of Styleformer-ART can be attributed to its advanced augmentation techniques, which likely provide more diverse and representative training samples.

As observed in Figure 7 and Tables 3 and 4, the experimental findings reveal that augmenting the artifact character images using WGANDIV, WGANGP, CCGAN, and Styleformer-ART led to much better performance. The significant accuracy increase observed underscores the efficacy of these synthetic image generation techniques in enhancing CNN models and other types of recognition models in recognizing artifact characters. Moreover, augmenting the training dataset with synthetic images generated by Styleformer-ART led to substantial performance improvements, with a notable increase of up to 60%.

7. Conclusions and Future Work

Styleformer-ART achieved state-of-the-art performance compared to the other image generation models. To arrive at our conclusion regarding the optimal augmentation method, we assessed all methods employed in this study by comparing the Frétchet inception distance (FID) scores between reference and generated images. Additionally, we evaluated the methods based on the CNN model's accuracy when trained on the augmented data. The Styleformer-ART model demonstrated the highest FID score of 210.72 and achieved a CNN accuracy of 84%. These results have important implications for artifact analysis and cultural heritage preservation, as they demonstrate the value of leveraging synthetic image generation techniques to overcome challenges related to data scarcity and artifact degradation. However, our work was confined to the English language. This raises questions about the potential application of these techniques in other languages, including syllable-based languages like Korean and non-alphabetic ones like Chinese. Further research is needed to explore how these methods can be adapted and optimized for different linguistic and cultural contexts, considering the unique characteristics and challenges associated with each language. In the future, we will explore other character sets from other languages and evaluate how Styleformer-Art performs in generating good-quality images.

Author Contributions: J.T.S. and I.Y.J. conceived and designed the experiments; J.T.S. performed the experiments; J.T.S. and I.Y.J. analyzed the data; J.T.S. wrote the paper, and I.Y.J. reorganized and corrected the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (No. 2021R1F1A1064345) and by the BK21 FOUR project funded by the Ministry of Education, Korea (No. 4199990113966).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in https://github.com/Tundeh/Artifacts_Character_Dataset (accessed on 23 July 2024).

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

References

1. Assael, Y.; Sommerschield, T.; Shillingford, B.; Bordbar, M.; Pavlopoulos, J.; Chatzipanagiotou, M.; Androutsopoulos, I.; Prag, J.; de Freitas, N. Restoring and attributing ancient texts using deep neural networks. *Nature* **2022**, *603*, 280–283. [CrossRef] [PubMed]
2. Narang, S.R.; Kumar, M.; Jindal, M.K. DeepNetDevanagari: A deep learning model for Devanagari ancient character recognition. *Multimed. Tools Appl.* **2021**, *80*, 20671–20686. [CrossRef]
3. Huang, H.; Yang, D.; Dai, G.; Han, Z.; Wang, Y.; Lam, K.M.; Yang, F.; Huang, S.; Liu, Y.; He, M. AGTGAN: Unpaired Image Translation for Photographic Ancient Character Generation. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022.
4. Casini, L.; Marchetti, N.; Montanucci, A.; Orrù, V.; Roccetti, M. A human–AI collaboration workflow for archaeological sites detection. *Sci. Rep.* **2023**, *13*, 8699. [CrossRef] [PubMed]
5. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Neural Information Processing Systems, Cambridge, MA, USA, 8–13 December 2014.
6. Alqahtani, H.; Kavakli-Thorne, M.; Kumar, G. Applications of Generative Adversarial Networks (GANs): An Updated Review. *Arch. Comput. Methods Eng.* **2019**, *28*, 525–552. [CrossRef]
7. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved Training of Wasserstein GANs. In Proceedings of the Neural Information Processing Systems, Red Hook, NY, USA, 4–9 December 2017.
8. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and Improving the Image Quality of StyleGAN. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8107–8116.
9. Warde-Farley, D.; Bengio, Y. Improving Generative Adversarial Networks with Denoising Feature Matching. In Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.
10. Fontanella, F.; Colace, F.; Molinara, M.; di Freca, A.S.; Stanco, F. Pattern recognition and artificial intelligence techniques for cultural heritage. *Pattern Recognit. Lett.* **2020**, *138*, 23–29. [CrossRef]
11. Yalin, M.; Li, L.; Ji, Y.; Li, G. Research on denoising method of chinese ancient character image based on chinese character writing standard model. *Sci. Rep.* **2022**, *12*, 19795. [CrossRef] [PubMed]
12. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
13. Ding, X.; Wang, Y.; Xu, Z.; Welch, W.J.; Wang, Z.J. CcGAN: Continuous Conditional Generative Adversarial Networks for Image Generation. *arXiv* **2020**, arXiv:2011.07466.
14. Midoh, Y.; Nakamae, K. Image quality enhancement of a CD-SEM image using conditional generative adversarial networks. In Proceedings of the Advanced Lithography, San Jose, CA, USA, 24–28 February 2019.
15. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
16. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *arXiv* **2020**, arXiv:2006.11239.
17. Park, J.; Kim, Y. Styleformer: Transformer-based Generative Adversarial Networks with Style Vector. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 8973–8982.
18. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 87–110. [CrossRef]

19. Abdulraheem, A.; Suleiman, J.T.; Jung, I.Y. Generative Adversarial Network Models for Augmenting Digit and Character Datasets Embedded in Standard Markings on Ship Bodies. *Electronics* **2023**, *12*, 3668. [CrossRef]
20. Hidayat, A.A.; Purwandari, K.; Cenggoro, T.W.; Pardamean, B. A Convolutional Neural Network-based Ancient Sundanese Character Classifier with Data Augmentation. *Procedia Comput. Sci.* **2021**, *179*, 195–201. [CrossRef]
21. Jindal, A.; Ghosh, R. An optimized CNN system to recognize handwritten characters in ancient documents in Grantha script. *Int. J. Inf. Technol.* **2023**, *15*, 1975–1983. [CrossRef]
22. 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 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
23. Cazenavette, G.; de Guevara, M.L. MixerGAN: An MLP-Based Architecture for Unpaired Image-to-Image Translation. *arXiv* **2021**, arXiv:2105.14110.
24. Emami, H.; Aliabadi, M.M.; Dong, M.; Chinnam, R.B. SPA-GAN: Spatial Attention GAN for Image-to-Image Translation. *IEEE Trans. Multimed.* **2019**, *23*, 391–401. [CrossRef]
25. Guha, R.; Das, N.; Kundu, M.; Nasipuri, M.; Santosh, K.C. DevNet: An Efficient CNN Architecture for Handwritten Devanagari Character Recognition. *Int. J. Pattern Recognit. Artif. Intell.* **2020**, *34*, 2052009. [CrossRef]
26. Driss, S.B.; Soua, M.; Kachouri, R.; Akil, M. A comparison study between MLP and convolutional neural network models for character recognition. In Proceedings of the Commercial + Scientific Sensing and Imaging, Anaheim, CA, USA, 9–13 April 2017.
27. Bhardwaj, A. An Accurate Deep-Learning Model for Handwritten Devanagari Character Recognition. *Int. J. Mech. Eng.* **2022**, *7*, 1317–1328.
28. Abdulraheem, A.; Jung, I.Y. Effective Digital Technology Enabling Automatic Recognition of Special-Type Marking of Expiry Dates. *Sustainability* **2023**, *15*, 12915. [CrossRef]
29. Corazza, M.; Tamburini, F.; Valério, M.; Ferrara, S. Unsupervised deep learning supports reclassification of Bronze age cypriot writing system. *PLoS ONE* **2022**, *17*, e0269544. [CrossRef]
30. Wu, J.; Huang, Z.; Thoma, J.; Acharya, D.; Gool, L.V. Wasserstein Divergence for GANs. In Proceedings of the European Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
31. Odena, A.; Olah, C.; Shlens, J. Conditional Image Synthesis with Auxiliary Classifier GANs. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016.
32. Dimitrakopoulos, P.; Sfikas, G.; Nikou, C. Wind: Wasserstein Inception Distance For Evaluating Generative Adversarial Network Performance. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3182–3186.
33. Yu, Y.; Zhang, W.; Deng, Y. *Frechet Inception Distance (fid) for Evaluating Gans*; China University of Mining Technology Beijing Graduate School: Xuzhou, China, 2021.
34. Benny, Y.; Galanti, T.; Benaim, S.; Wolf, L. Evaluation Metrics for Conditional Image Generation. *Int. J. Comput. Vis.* **2020**, *129*, 1712–1731. [CrossRef]
35. Betzalel, E.; Penso, C.; Navon, A.; Fetaya, E. A Study on the Evaluation of Generative Models. *arXiv* **2022**, arXiv:2206.10935.
36. Kynkaanniemi, T.; Karras, T.; Aittala, M.; Aila, T.; Lehtinen, J. The Role of ImageNet Classes in Fréchet Inception Distance. *arXiv* **2022**, arXiv:2203.06026.

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.