Deformity Removal from Handwritten Text Documents using Variable Cycle GAN

Shivangi Nigam (rsi2018506@iiita.ac.in)
Indian Institute of Information Technology Allahabad
Adarsh Prasad Behera
Indian Institute of Information Technology Allahabad
Shekhar Verma
Indian Institute of Information Technology Allahabad
P Nagabhushan
Indian Institute of Information Technology Allahabad

Research Article

Keywords: Handwritten text, strike-off, semantics, generative adversarial network, image to image translation

Posted Date: June 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1488498/v1

Deformity Removal from Handwritten Text Documents using Variable Cycle GAN

Shivangi Nigam[1*], Adarsh Prasad Behera[1], Shekhar Verma[1] and P. Nagabhushan[1]

[1]Department of Information Technology, Indian Institute of Information Technology Allahabad, Jhalwa, Prayagraj, 211015, Uttar Pradesh, India.

*Corresponding author(s). E-mail(s): rsi2018506@iiita.ac.in;
Contributing authors: pwc2015004@iiita.ac.in; sverma@iiita.ac.in;
pnagabhushan@iiita.ac.in;

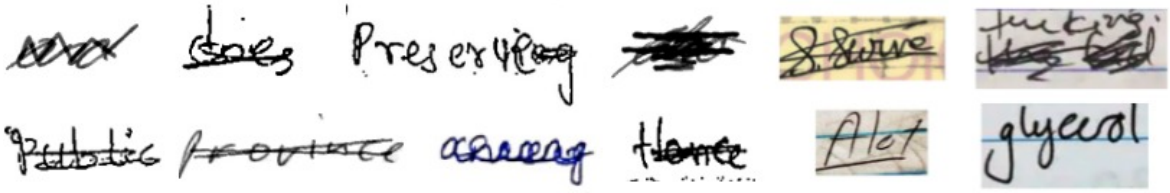
Abstract

Text document recognition systems perform well in the case of printed documents but fails to produce similar results for handwritten text documents. The significant challenges include different writing styles, various background complexities, added noise of image acquisition methods, and the presence of deformed text images such as strike-offs and underlines. Any deformity can be posed as a change in structural information, resulting in intensity variations of the original text. The restoration of deformed images aims to recover clean images while maintaining the structural information and preserving the semantic dependencies of the local pixels. The current adversarial networks are unable to preserve the structural and semantic dependencies as they consider each individual pixel-to-pixel variations and encourage the perceptually non-meaningful aspects of the images. We propose a Variable Cycle Generative adversarial network (VCGAN) to consider the perceptual quality of the images in the learning objective which is based on the variable content loss to preserve the dependencies. We propose a Top-$\boldsymbol{k}$ Variable Loss ($\boldsymbol{TV_k}$) to compute the similarity of images by accounting the intensity variations that do not interfere with image semantic structures. The results show that VCGAN is able to remove most of the deformities with an elevated $\boldsymbol{F1}$ score of **97.40%**. We also tested the images generated by VCGAN with a handwritten text recognition system. VCGAN outperforms the current state of the art algorithms with a character error rate of **7.64%** and word accuracy of **81.53%**.

Keywords: Handwritten text, strike-off, semantics, generative adversarial network, image to image translation

1 Introduction

Handwritten text recognition (HTR) is an active area of research because of its wide range of applications e.g., digitization for digital libraries [22], text restoration [5, 6, 31] etc. It aims to transform the text present in graphic forms such as images of handwritten notes, scene text, memos, whiteboards, medical records, and historical documents into its symbolic representation. The complexity of the problem is dictated by the constraints of handwritten text, i.e. diverse free flow writing styles of individuals, innumerable characters and their combinations, divergent backgrounds such as ruled pages, the image behind text (scene text) and various image acquisition practices which pose challenges to the recognition systems [7, 16, 26, 27].



a) Strikeoff words with clean background b) Strike-off words with complex background

Fig. 1: Strike-off words

The current state-of-the-art handwritten text recognition has been done on images acquired through ideal and supervised means, for example, on IAM data set [18]. The HTR systems have achieved high accuracy on such data sets; however, they lack the notion of the intricacies of handwritten text documents[20]. One such intricacy is strike-off components in a handwritten text. These are markdown indicators by the writer after a writing error in a handwritten document. An example of such occurrence is unrestricted handwritten text in students' examination notebooks which may have these strike-off errors. With such a sample, HTR may produce irrelevant outputs. Due to the lack of such unrestricted data, it was challenging to develop HTR systems that could perform well on these irregularities in data. Almost every handwritten document is expected to have such intricacies that a current OCR system cannot process.

Albeit all the diversities and complexities of handwritten text, almost all handwritten text recognition systems assume that the document texts are flawlessly written and captured. However, chances of errors in unconstrained handwriting are quite high. There may be various kinds of writing errors. Perhaps the most common is the strike-off error. The strike-off is a markdown indication to discard the concerning content of the text. It may be on a single character, single word, multiple characters, multiple words, or multiple lines. The style of strike off is a characteristic of individual writing fashion, which is profoundly indiscriminate. Some typical examples of the challenges are shown in Figure 1. Figure 1 a) show the strike-off images with a with various types of strokes. Various image acquisition and environment-related flaws(ruled pages, background image, blurriness, skewness) are major reasons for degrading document image quality. Most of the strike-off removal work done considers the data sets with images as in fig.1.a), while in reality, a handwritten document can have other distortions as well (Figure 1 b)). The strike-off elements may have various lengths and shapes. A larger portion of text is mostly struck off with straight lines like single-line strike-off, multi-line strike off, or cross strike offs. In contrast, the words may have many stroke types like a wave, zig-zag, cross, lined, scratch, etc. Also, different persons have different styles of strike off. [21].

Recent developments have seen the impressive use of generative modelling [12] for Document Image enhancement [29] and handwritten text generation tasks [11] [9]. Generative models are unsupervised probabilistic models which attempt to learn the patterns in a dataset and use these observations to generate new data. Generative adversarial networks (GAN)[12] use these generative models in a supervised fashion by classifying the generated samples into real and fake. GANs have been useful for generating novel and meaningful text with preserving the semantic and syntactic properties used for natural language processing and other document related applications. Handwritten text Image generation using Generative adversarial networks (GANs) have been useful in reducing the gap of data sets for training deep learning models [9, 20, 21]. Lately, image to image translation (I2I) has been useful for translating from a source domain representation to target domain representation. The research works in [14], [24, 25] have used I2I for translating from strike off image to a clean version of the corresponding image. The images produced by these [14, 24, 25] algorithms achieve acceptable perceived visual quality while not precisely matching the ground truth. The restoration task addressed by these works is supervised by simple element loss functions based on Mean Squared Error (MSE)

or Structural Similarity Indices (SSIM)[24, 25]. These loss functions encourage the perceptually meaningless aspects of the input by accounting for the overall structure of the input. Hence the methodologies do not contemplate the natural denotation of the problem and limit the performance of restoration.

In this work, we pose the strike-off strokes as changes in structural information and the semantic content that is reflected in intensity variations in a clean handwritten text. A well-known way to evaluate intensity variations is the $L1$ norm. However, it cannot consider the importance of perceptual quality as it includes perceptually non meaning-full aspects. The perceived visual quality of a cleaned image is directly related to the removal of strike-off strokes and the preservation of inter-dependency of spatially close pixels. Consequently, we need a norm to evaluate the intensity variations while maintaining the structural information of the local pixels. Here we propose Top-k Variance TV_k norm as a new norm to measure similarity between images. TV_k focuses on the topmost intensity variations to account for the significant structural differences between strike-off image and its clean counterpart. The major contributions of this work are:

1. Strike off removal using Unpaired Image to image translation with a weakly supervised Adversarial model: VCGAN.
2. Content Loss: Top-k Variance TV_k for measuring the similarity of strike-off image and its clean counterpart.
3. A CNN-LSTM-CTC model to perform recognition tasks on the cleaned images generated by VCGAN.

The rest of the paper is structured as follows. Section 2 presents some of the recent developments and related works. In section 3, some preliminaries or essential background knowledge is presented. The objectives and problem definition are stated in section 4. Section 5 contains the proposed methodology. The data sets and implementation details are explained in section 6. In section 7, the experimental and comparison results are discussed, and finally, section 8 concludes the paper.

2 Related Works

The research in strike-off identification have primarily used manual handcrafted methods such as SVM, and HMM [3] and then image in-painting methods have been utilized to restore the text. It had been a comparatively less explored area due to the insufficiency of handwritten strike-off data sets. Recently, to resolve the data scarcity, data augmentation techniques [23, 32] have been adequate to augment input data and produce new data in the input data space. Salient data augmentation techniques such as cropping, adding noise, resizing, flipping, rotating, and changing the colour of an image were formerly used. Although there is a drawback that there is no introduction of new data and the data so created is not enough for improving model's generalizability [23]. Lately, many data augmentation techniques have been explored on the pretext of generating strike-off datasets. Recently a Resnet-BiLSTM-CTC based method was proposed for strike-off text generation in [20, 28].

The studies conducted on strike off text processing has been specific to scripts or styles of strike off. A very early work [2] used K-Nearest Neighbor(K-NN) to identify and reject the noise elements as in scribbles, crossed-outs and isolated strokes. In [19], authors present a Markov random field(MRF) based MAP framework to determine joint energy distribution between labels and observation fields. In [4], a probabilistic contextual relationship model using a patch-based MRF was proposed for restoring the printed documents having degradations such as cuts, merges, blobs and erosion. Another work in [17] proposed HMM-based wave and line stroke recognition, although the detection is not considered in this work. Brink et al.[8] used a decision tree-based binary classifier for the removal of crossed-out handwritten text components. A US patent [30] claimed to recognize crossed out English characters by a feature-based classifier. Chaudhary et al., [1] have utilized morphological and graph-based features computed from deformed text images to identify and remove strike-offs. This work considers most of the strokes possible in a handwritten text, but the solution requires apriori knowledge of the strike-off being handled. In contrast to these manually crafted measures, deep learning approaches have been explored to address the

problem of strike-off identification and recognition. The research works in [14], [24, 25] have used I2I for translating from strike off image to a clean version of the corresponding image. These works have created synthetic data set for this purpose under supervised conditions. These data sets do not acknowledge that free form handwritten text has much more obstructions than posed in their work. Hence the methodologies proposed does not contemplate the natural denotation of the problem.

3 Preliminaries/Background

3.1 GANs for Image to image translation

In image to image translation, the models seek to learn the mapping from input domain $\mathcal{X}$ to a target domain $\mathcal{Y}$, using paired or unpaired training samples $\mathcal{S} = \{(x_i, y_j)_{i,j}^{N,M}\}$ where $x \epsilon \mathcal{X}$ and $y \epsilon \mathcal{Y}$. In an adversarial training model there is a *Generator* $Gen(\mathcal{X})$, which seeks to learns the mapping $h : \mathcal{X} \rightarrow \mathcal{Y}$ of the natural image manifold during a min-max game with *Discriminator* $Disc(\mathcal{Y})$. The generator $Gen(\mathcal{X})$ takes input from the domain of input images $\mathcal{X}$ and translates it to the domain of target images $\mathcal{Y}$ such that the image produced is indistinguishable from the input image. The adversarial loss is imposed to penalise the fake samples, which are identified by the discriminator $Disc(\mathcal{Y})$. The generator $Gen(\mathcal{X})$ competes with the discriminator $Disc(\mathcal{Y})$ with an intent to fool the discriminator by producing an indistinguishable fake sample. The generator and discriminator compete for a common min-max objective

$$\mathcal{L}_{adv_x} = E_{x\sim P_d}[log(D(x))]+E_{z\sim P_z}[log(1-(D(z)))] \quad (1)$$

Estimating probabilities of real and generated images gives the cross-entropy between real and generated distributions. The discriminator aims to maximize its probability estimate whether the x is real or fake. The translation task aims to preserve the source content features and translate them into the target domain's style. In the unpaired image to image translation, the loss computed between generated fake samples and the original samples is not enough to ensure that the input x_i will be mapped to the specific target y_i. The model regularizes by penalizing the inconsistencies of different domain translations with cycle consistent loss. Without cycle consistent loss, the generator was producing images in the target domain but could not be translated between the two domains. The mappings $h : \mathcal{X} \rightarrow \mathcal{Y}$ and $f : Y \rightarrow X$ aid the translation task to cycle between the two domains X, Y to and fro. The cycle consistency is imposed by ensuring that the image to image translation is transitive i.e. $x \rightarrow h(x) \rightarrow f(h(x)) \sim x$. The loss $\mathcal{L}_{cycle_x}$ is computed by using the L1 norm between the generated sample and the original sample.

$$\mathcal{L}_{cycle_x} = E_{x\sim p_{(X)}} \|fake(x) - x\|_1 \quad (2)$$

The $\mathcal{L}_{cycle}$ gives the total of all the absolute errors between each pixel. This means that each pixel value is measured with other values to produce a total representation of the translated pixel loss. The total cycle loss is the sum of both translations $\mathcal{X} \rightarrow \mathcal{Y}$ and $\mathcal{Y} \rightarrow \mathcal{X}$. Here the pixel-wise loss is determined by average loss across all pixels. The total loss of cycle GAN is :

$$Total\ loss = \mathcal{L}_{adv_x} + \mathcal{L}_{Cycle_{x,y}} \quad (3)$$

4 Problem Definition

A distorted image can be considered a sum of an un-distorted reference and an error component. An accepted assumption is that loss of perceptual quality is related to the visibility of the error component. MSE objectively quantifies the strength of the error and has precise physical meanings. However, different distorted images with different visible or invisible errors may have the same MSE. Images are highly structured, and spatially proximate pixels carry information about the structure of the content and, therefore, exhibit strong dependencies. MSE is independent of the underlying signal structure and does not match perceived visual quality. Different perceptual image quality assessment measures weigh different aspects of errors to give different quantitative measures. The content loss of current adversarial networks is unable to preserve the structure and semantic dependencies. Metrics like L1, MSE, SSIM consider each individual pixel-to-pixel variations.

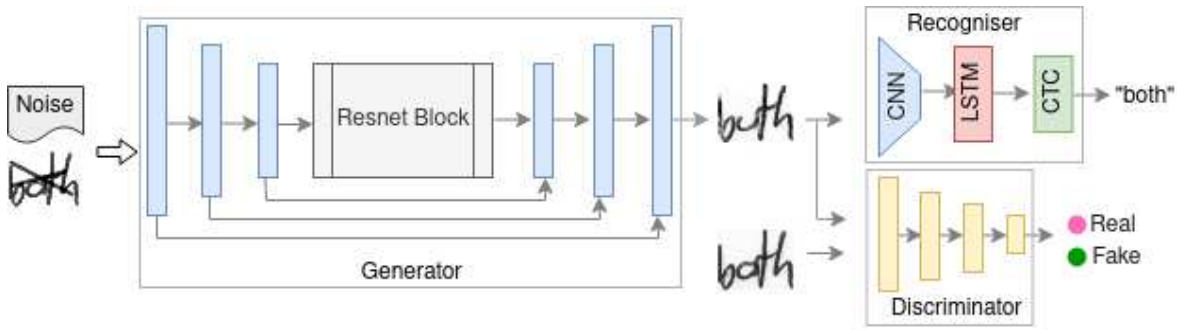


Fig. 2: A network architecture of VCGAN with Generator, Discriminator and Recogniser

These matrices encourage the perceptually non-meaningful aspects of the images and thus limit the performance of the restoration.

5 VCGAN and Handwritten text recovery

The objective of the restoration of strike-off images is to recover clean images while maintaining the structural information of the local pixels. This problem can be optimized to achieve perceptually pleasing target images, although the perceived visual quality of these target images should match the ground truth distribution. There is one common denominator in all the metrics such as MSE, SSIM, etc. These assessments measure the amount of structural information preserved in the distorted version of the reference image. Lost structural information leads to a lower quality score. These indices are applied locally as distortions may vary spatially, and statistical features of the image may be spatially non-stationary. For instance, the SSIM Index defines the structural information in an image as those attributes representing the structure of objects in the scene, independent of the local average luminance and contrast. The objective structural similarity indices can capture the characteristics of subjective measures and yield better assessment as compared to MSE. However, these indices try to discount distortions that do not affect the local structures. Thus, we need indices based on intensity variations and structural information that can be used to measure the similarity between images that overcome the limitations of the intensity-based MSE index. The focus of this work is to preserve the inter-dependence of spatially close pixels while removing the strike-off strokes. To adhere to this goal, we design an objective function such that the target images are on the raw image manifold while maintaining the similarity to the ground truth distribution.

In this work, we propose Variable Cycle GAN (VCGAN) which seeks to learn the mapping from the input domain of strike-off images $\mathcal{S}$ to a target domain of clean images $\mathcal{C}$. The generators $Gen(\mathcal{S})$ and $Gen(\mathcal{C})$ aim to translate from one domain to the other such that the images produced are indistinguishable from the images in respective domains. The idea is to utilize the unpaired image to image translation ability of adversarial models [33] for strike-off text image restoration. The model is weakly supervised as it uses unpaired training data samples $\{(s_i, c_j)_{i,j}^{N,M}\}$ where $s \epsilon \mathcal{S}$ and $c \epsilon \mathcal{C}$.. Consequently, we do not have the exact ground truth pair of the strike text images we are training with. Persuading with this information, rather than having the target image exactly match the ground truth, we encourage the similarity of underlying semantic structural distributions. This is achieved by a new similarity norm Top-k Variance $\mathcal{TV}_k$.

5.1 VCGAN Adversarial Loss

The Adversarial Loss is computed between the generated image and the real image where the generator aims to generate samples as good as real so that they are indistinguishable to the discriminator, while the discriminator seeks to differentiate the real and generated samples. The discriminator takes real $x \epsilon X$ and fake $z \epsilon Z \sim p_{data}(\mathcal{X})$ inputs and aims to maximize its estimate of probability of real data $logD(x)$ and fake data generated by

generator $log(1 - D(G(x)))$.The generator generates fake data $G(z)$ with the given noise z to fool the discriminator. The generator cannot impact the estimations of the discriminator on real data, so it tries to minimize its estimate of fake data by minimizing $log(1 - D(G(x)))$. The objectives of discriminator and generator can be defined by the equation :

$$\mathcal{L}_{adv}(\mathcal{G}, \mathcal{D}, \mathcal{X}, \mathcal{Y}) = min_{\mathcal{G}} max_{\mathcal{D}}[\mathcal{L}_{adv_x} + \mathcal{L}_{adv_y}] \tag{4}$$

5.2 VCGAN Content Loss

The image similarity measure involves computing some distance metric to evaluate the differences between corresponding pixels of two images. The Sum of absolute error (SAE) or $L1$ norm computes the absolute difference between corresponding pixels and then takes mean over all pixels. The average loss is the most widely used metric to reach a fair approximation of abrupt anomalies. The pixel-wise loss between two images (a, b) of dimension $(m * n * 3)$ with $L1$ is given by:

$$\|L\|_1 = \frac{1}{p} \sum_{i=0}^{p} l(a_i, b_i) \tag{5}$$

where $p = (m * n * 3)$ and $l(a_i, b_i)$ is the L1 norm between every pixel of a and b.

Although an important issue in designing an objective function is to deal with the high level inter-dependent semantic content of an image, the perceived visual quality of an image is related to the visibility of any deformity (strike-off) in the image and the semantic structural content of the image. Such content is more salient to maintain the perceptual quality of an image. Thus the judgement of perceptual similarity is influenced by removing the deformities and preserving the salient semantic structures. To address this goal, we draw inspiration from the work [10] to design an objective function that is sensitive to the structural information of an image. We propose Top-k Variance $\mathcal{TV}_k$ to compute the similarity of images by the intensity variation between strike-off and its clean counterpart. The variation detects the structural information change, accounting only those variations which should not interfere with the inter-dependency of spatially close pixels. To address the shortcomings of the $L1$ norm, we allow $\mathcal{TV}_k$ to better measure the similarity by allowing only top k intensity variations. By doing this, we ignore the variations that interfere with the images' semantic structures.

$$\|\mathcal{TV}_k\| = \frac{1}{k}[\underset{1 \le j \le k}{max} \sum_{i=0}^{p} l(a_i, b_i)] \tag{6}$$

$\mathcal{TV}_k$ can be scaled to match the standard $L1$ norm when $(k = p)$. The coefficient k is a meta-parameter that provides flexibility to adapt to different types of data distributions

5.3 VCGAN Objective function

The objective function of VCGAN is :

$$\begin{aligned} \mathcal{L}_{adv}(\mathcal{G}, \mathcal{D}, \mathcal{S}, \mathcal{C}) = \mathcal{L}_{adv_s} + \mathcal{L}_{adv_c} \\ +\lambda(\mathcal{TV}_{k_s} + \mathcal{TV}_{k_c}) \end{aligned} \tag{7}$$

where λ is a coefficient to control the relative importance of the two losses in the objective function. The aim is to solve for:

$$\mathcal{G}^*, \mathcal{D}^* = arg\ min_G, max_D[\mathcal{L}_{adv}(\mathcal{G}, \mathcal{D}, \mathcal{S}, \mathcal{C})] \tag{8}$$

The generators $Gen(\mathcal{S})$ and $Gen(\mathcal{C})$ learns the mapping $h : \mathcal{S} \rightarrow \mathcal{C}$ and $f : \mathcal{C} \rightarrow \mathcal{S}$ during a min-max game with the discriminators $Disc(\mathcal{C})$ $Disc(\mathcal{S})$ respectively. The model tries restore strike-off image $\hat{x}$ to cleaned image $\hat{y}$ with minimum restoration error while encouraging the target $\hat{y}$ to be preceptually similar to the ground truth distribution of x.

6 Implementation details

6.1 Adversarial Network

The generative networks of our model are adapted from Johnson et al. [15]. The generator network contains stride-2 convolutions for down-sampling and several residual blocks, which are used to capture relevant information and flow that information from the initial layers to the last ones and at last two $\frac{1}{2}$ convolutions for up-sampling. We use the instance normalization technique as used in [15]. The discriminator network uses 70×70 Patch GAN to identify the real and fake images.

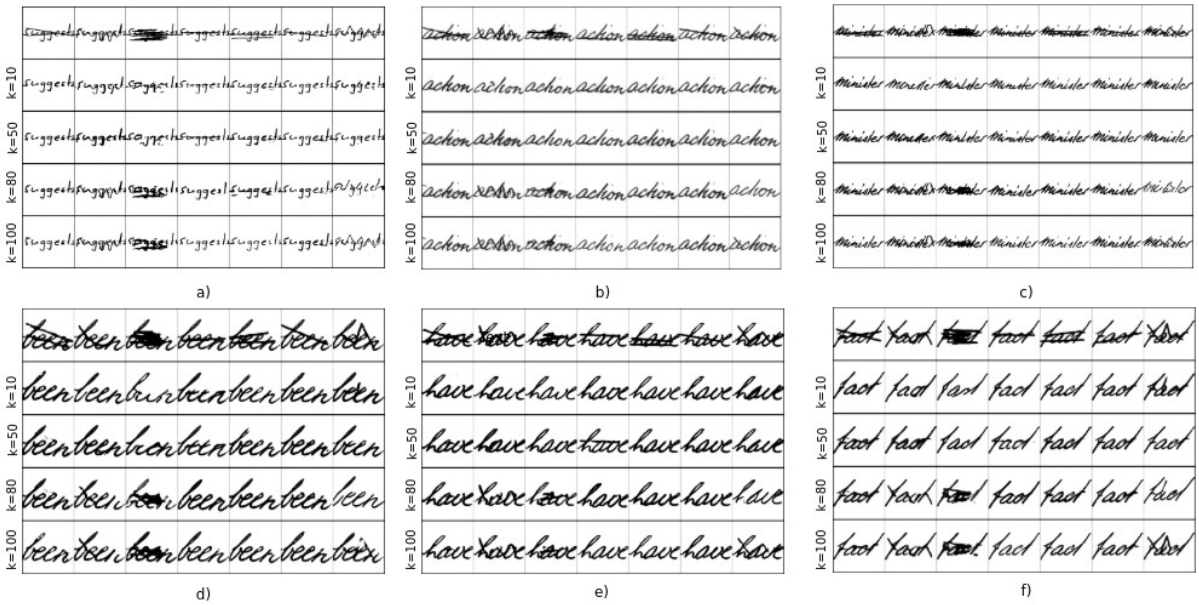


Fig. 3: Strike-off removal with VCGAN for various values of coefficient k

Unlike a normal GAN discriminator, a Patch GAN outputs a $N \times N$ output array O in which O_{ij} signifies whether the respective patch belongs to real or fake.

6.2 Recognition Network

To evaluate the performance of VCGAN, we use a subsequent recognition network. A CNN-LSTM-CTC network is implemented for recognition tasks. The CNN layers are used to perform feature extraction, which is used by LSTM layers that identify the temporal patterns in the feature set. The network is composed of 5 CNN layers, 2 LSTM layers and at last connectionist temporal classification (CTC) is used to predict the final output.

6.3 Data set

In this work, we have used a synthetic data set that incorporates mostly all types of strokes as in single line, double line, cross, wave, zig-zag and scratch[13]. We have created strike-off words by superimposing actual strokes over the clean word images by using the algorithm proposed in [14]. The handwritten text images are obtained from student notebooks which are not biased towards any supervised conditions. These documents are further segmented into words for training the proposed model. We have also included IAM [18] data set to increase the diversity of training samples.

6.4 Training details

We have applied some techniques for optimizing the training of the GAN model. First, The objective of the generator is to minimize the probability of images being predicted as fake. In other words, the generator seeks to maximize the probability of images being predicted as real. Thus non-saturating loss function of generator is to maximize $log(D(G(x)))$

$$\mathcal{L}_G(X, Y) = E_{z \sim P_z}[log((D(G(z))))] \quad (9)$$

Secondly, to reduce the model oscillation while training, we have used two approaches a)Instead of learning with a fixed learning rate, we use a learning rate scheduler to produce a warm and to reduce the learning curve. This is a way to reduce the primacy effect of the early training examples. Without it, you may need to run a few extra epochs to get the convergence desired, as the model un-trains those early superstitions. b) Provide discriminator with a set of recently produced images instead of just providing the latest ones

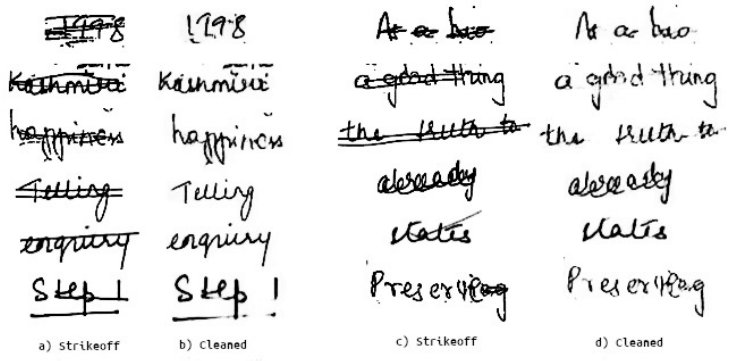


Fig. 4: Strike-off samples taken form student notebooks and corresponding generated cleaned text

7 Results

In this work, we have considered seven types of deformities (strike-offs) in the handwritten text: cross, wave, scratch, single line, double line, diagonal, and zig-zag. The model is tested on the proportional combination of all types of deformities. The proposed model is also tested on underlined text images. It can very well differentiate between a strike-off and an underline. We have observed high performance of some deformities like a cross, single line, double line and diagonal, while scratch, wave and zig-zag have seen average performances. Figure 3 shows some samples of the generated strike-off removal. We have also tested our approach on a handwritten data set collected from student notebooks. It can be observed in figure 4 that our approach achieves state-of-the-art results on those as well. The implemented approach is tested with respect to two objectives:

- Performance of strike off detection and removal
- Performance of Handwritten text recognition

7.1 Performance of strike-off removal

An authentic image quality assessment compares the target images with the ground truth images. Due to the absence of ground truth data of all the test data, we access the image quality by using various reference metrics such as Image similarity metrics (MSE (Mean Square Error), PSNR (Peak Signal to Noise Ratio), SSIM (Structured Similarity Index Method)), Pixel-based metrics (Precision, Recall and $F1$ score) and Image restoration metrics (Deformity Detection rate, Restoration accuracy, F measure) to analyse the performance of the strike-off image restoration. These Image quality assessment techniques measure the deviation of quality of generated clean images with respect to the ideal/ground truth clean images.

In a generated clean image, we define:
True Positives (TP): the foreground area of the image which is correctly identified as strike off and is removed in generated clean image
False Positives (FP): the foreground area of the image which is incorrectly labelled as strike off and hence is removed in the generated clean image.
False Negatives (FN): the missing area of the image which could not be labelled as strike off and hence is not removed from the generated clean image.
Intersection measure (IM) : Intersecting pixels of generated image and true image
True pixels (Tr) : Foreground pixels in true image
Pred pixels (Pr) : Foreground pixels in generated image

We present the experimentation results in the form of following measures which are shown in Tables 1, 2, 3, 4 representing various values of coefficient $k = (10, 50, 80, 100)$.

- Pixel based measures
  - Precision (Pr)= $TP/TP + FP$
  - Recall (Re): $TP/TP + FP$
  - $F1$ score ($F1$s): $2 * Precision * Recall/(Precision + Recall)$

The results in the pixel-based methods show that we have achieved an average $F1score = 93.75(\pm 2.86)$, while Precision and Recall are $92.78(\pm 3.018), 94.759(\pm 2.698)$. We have achieved the best Precision of 94.371% and best Recall of

Stroke type	MSE	PSNR	SSIM
cross	0.067	11.957	0.706
wave	0.070	11.712	0.780
scratch	0.064	12.149	0.799
single line	0.063	12.178	0.704
double line	0.063	12.178	0.707
diagonal	0.063	12.178	0.710
zig-zag	0.065	12.064	0.703

(a) Image Similarity Metrics

DDR	RA	FM
97.856	96.963	97.407
91.589	90.591	91.087
89.576	88.580	89.075
95.401	96.817	96.104
96.747	95.849	96.296
95.333	94.121	94.723
90.541	89.136	89.833

(b) Deformity detection Metrics

Pr	Re	F1s
96.971	97.824	97.396
90.187	92.723	91.437
88.581	90.591	89.575
95.810	97.389	96.593
94.827	97.445	96.118
94.074	95.821	94.939
89.010	91.519	90.247

(c) Pixel based Metrics

Table 1: Performance Metrics for strike off removal for $k = 10$

Stroke type	MSE	PSNR	SSIM
cross	0.069	12.835	0.684
wave	0.076	12.991	0.676
scratch	0.074	11.991	0.670
single line	0.069	12.172	0.702
double line	0.068	11.902	0.711
diagonal	0.072	12.961	0.600
zig-zag	0.079	12.799	0.687

(a) Image Similarity Metrics

DDR	RA	FM
97.711	95.952	96.824
91.565	90.562	91.061
89.565	88.562	89.061
95.372	96.805	96.083
96.438	95.827	96.132
95.328	94.069	94.694
90.526	89.008	89.761

(b) Deformity Detection Metrics

Pr	Re	F1s
94.341	97.711	95.996
90.164	92.710	91.419
88.562	90.565	89.552
95.805	97.372	96.582
94.827	97.438	96.115
94.069	95.828	94.940
89.008	91.526	90.249

(c) Pixel based Metrics

Table 2: Performance Metrics for strike off removal for $k = 50$

97.724%. We can see that Tables 1, 2 are more close to the best values of precision and recall. The best $F1$ score of our method is 97.39. The best of these values are obtained for cross-type of strike-offs, whereas the scratch stroke has observed the score of 89.57.

- Image similarity measures
  - Mean squared error (MSE) between generated $g(x, y)$ and true $t(x, y)$ image as defined in equation.
    $MSE = \frac{1}{MN} \sum_{n=0}^{N} \sum_{m=0}^{M} [t(n, m) - g(n, m)]^2$
  - Peak Signal to Noise Ratio (PSNR)
    $PSNR = 10 \cdot \log_{10} \left(peak_{value}^2\right) \left(MSE\right)$
  - Structural similarity index measure (SSIM) computes luminance $\mathcal{L}$, contrast $\mathcal{C}$ and structure $\mathcal{S}$ of reference images x, y.
    $SSIM(x, y) = [\mathcal{L}(x, y)^{\alpha} \cdot \mathcal{C}(x, y)^{\beta} \cdot \mathcal{S}(x, y)^{\gamma}]$

The values of Image similarity metrics 1a, 2a, 3a, 4a have almost consistent results. MSE and PSNR are absolute errors, and these can have the same values for different deformations in an image. It cannot discriminate the structural content of images. However, SSIM better captures perception and saliency-based variations; thus, we can see that for $k = 10$ and $k = 50$ we have achieved slightly better values of SSIM.

- Image Restoration measures [1]
  - Deformity detection Rate: $DDR = IM/Tr$
  - Reconstruction Accuracy: $RA = IM/Pr$

Stroke type	MSE	PSNR	SSIM
cross	0.081	14.913	0.582
wave	0.082	14.865	0.575
scratch	0.080	14.664	0.584
single line	0.072	13.580	0.584
double line	0.078	13.834	0.594
diagonal	0.075	13.080	0.599
zig-zag	0.080	14.889	0.634

(a) Image Similarity Metrics

DDR	RA	FM
94.555	88.791	91.582
89.766	87.834	88.789
87.653	85.157	86.387
90.621	88.492	89.544
90.399	88.637	89.509
90.741	89.069	89.897
86.999	86.959	86.979

(b) Deformity Detection Metrics

Pr	Re	F1s
89.852	93.781	91.774
87.185	93.718	90.334
86.568	92.175	89.284
88.441	93.512	90.906
88.525	93.367	90.882
89.121	94.741	91.845
86.872	88.787	87.819

(c) Pixel based Metrics

Table 3: Performance Metrics for strike off removal for $k = 80$

Stroke type	MSE	PSNR	SSIM
cross	0.086	12.023	0.574
wave	0.086	11.963	0.565
scratch	0.084	11.682	0.532
single line	0.073	12.215	0.517
double line	0.079	12.216	0.521
diagonal	0.087	11.907	0.556
zig zag	0.085	12.037	0.537

(a) Image Similarity Metrics

DDR	RA	FM
94.456	88.758	91.518
89.732	87.821	88.766
87.642	85.145	86.375
90.61	88.449	89.516
90.313	88.619	89.458
90.702	89.043	89.865
86.579	86.847	86.713

(b) Deformity detection Metrics

Pr	Re	F!s
89.832	93.564	91.660
87.185	93.563	90.261
86.568	92.144	89.269
88.441	93.566	90.931
88.525	93.290	90.845
89.121	94.712	91.831
86.872	88.787	87.819

(c) Pixel based Metrics

Table 4: Performance Metrics for strike off removal for $k = 100$

– F- Measure: $FM = (2*DR*DR)/(DR+RA)$

The deformity detection rate depicts the model's performance on the detection of strike-off regions, while the restoration accuracy measures how well the text is recovered from the strike-off image. The F measure conveys the overall accuracy of detection as well as restoration. Tables 1, 2 show good recognition accuracy and F measure, while the tables 3, 4 have mixed responses to these metrics. The best score of DDR is 97.85%, the reconstruction accuracy is 96.96%, and the corresponding F Measure is 97.407%. The overall $F1$ score is $93.50(\pm 3.01)$, with detection rate and restoration accuracy as $93.863(\pm 2.824), 93.151(\pm 3.185)$.

The results show that the proposed model is consistent over various deformities. We have applied above stated metrics on the generated cleaned images vs clean handwritten images. In figure 4, we show some examples of students notebooks. These examples include multiple word strike-off, multiple line strike-off, partial strike-off and underlined samples. Another figure 5 shows comparison of various reference works [1, 14, 25] with our proposed model. The recent work [25] proposed TexRGAN based on cyclegan, although they have not considered three strike-off categories, scratch, wave and zig-zag, which are present in most handwritten documents. While the work in [1] have not considered scratch strike-off although the work tends to produce comparable results on other strike-offs considered in this work. The work in [14] has also proposed a similar model [25] but their model does not account for the semantics of the content into their objective function. Due to this, the translation proposed in

Strike off type	Measure	$k = 10$	$k = 50$	$k = 80$	$k = 100$
Clean	CER%	1.76			
	WA%	89.52			
Cross	CER%	**7.64**	8.45	30.87	34.01
	WA%	**72.45**	71.38	20.56	19.09
Wave	CER%	**20.31**	23.54	22.74	26.95
	WA%	**54.35**	51.31	17.16	15.67
Diagonal	CER%	**4.76**	6.12	18.91	19.34
	WA%	**81.53**	80.15	18.61	17.38
Double Line	CER%	7.96	**7.81**	15.47	17.29
	WA%	70.12	**75.78**	28.95	25.23
Single line	CER%	10.78	**10.57**	16.85	18.74
	WA%	70.96	**71.19**	27.39	19.82
Scratch	CER%	**32.65**	39.10	55.67	59.48
	WA%	**31.47**	27.91	14.14	15.00
Zig-zag	CER%	**15.41**	19.89	58.70	60.88
	WA%	52.36	**53.23**	18.03	15.11

Table 5: Handwritten text recognition (Character error rate (CER) and word accuracy(WA)) of generated images for strike off removal on various k values

the paper is not seem-less. A comparison of various $F1$ scores is shown in 6.

7.2 Performance of handwritten text recognition (HTR)

To evaluate the recognition capability of generated images by the proposed model, we have used a CNN-LSTM-CTC based handwritten text recognition module. We pass the generated images to our HTR module, which produces transcribed text. The HTR module is pre-trained on the IAM words dataset. The text output so generated by the HTR module is evaluated with Character Error rate (CER) and Word Accuracy(WA). These error rates provide granularity to the measure of the difference between transcribed text and actual text labels. We have manually labelled the generated images into their respective transcribed text labels to perform recognition tasks. The computation of these errors involves computing Levenshtein distance to compute the distance between two string sequences using character/-word level operations(insertion, deletion, substitution) to transform the output string sequence into a target string sequence.

- Character Error Rate (CER):

$$CER = (S + D + I)/N \tag{10}$$

  where, S is number of substitutions, D is number of Deletions and I is number of Insertions. N represents total operations need to be performed. We require lower CERs for better recognition.

- Word Accuracy (WA):

$$WA = Words_{OK}/Total\ words \tag{11}$$

  WA represents the ratio of correctly transcribed words. The words correctly transcribed are labelled as OK. The higher the accuracy, the better is the recognition performance.

Table 5 shows the CER's and WA's for various strokes. The diagonal stroke type achieves best accuracy of 81.53% for $k = 10$ followed by double line with 75.78% for $k = 50$ and the least is for scratch stroke with 31.47%. The lowest CER is 4.76 for diagonal stroke with $k = 10$. The higher values of k have observed higher CER values and lower accuracies.

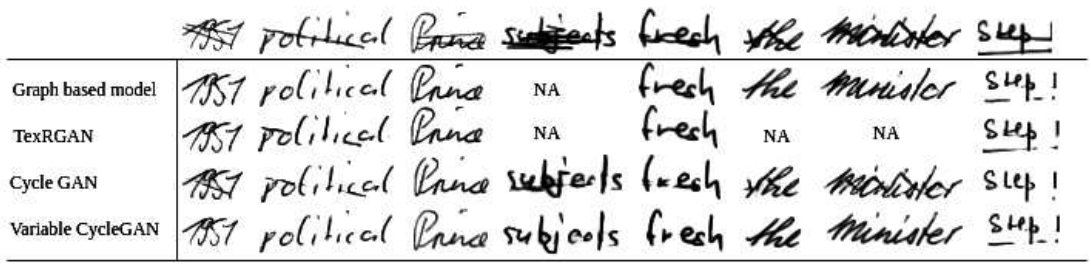


Fig. 5: Comparison of state-of-the-art methods on various stroke types

Approaches	$F1$ score	CER%	WA %
Graph based model[1]	89.44	-	-
TexRGAN[25]	96.76	12.74	65.28
CycleGAN[14]	97.01	-	-
VC GAN(ours)	**97.40**	**7.64**	**81.53**

Table 6: Comparison of performance measures of various state-of-the-art methods

Fig. 6: Some examples of challenging samples of strike-off text with corresponding generated and ground truth text.

7.3 Discussion

The Top-k Variance TV_k norm focuses on topmost intensity variations to capture the deformities while maintaining the perceived visual quality of the images. The coefficient k is a meta-parameter which provides flexibility to adapt to different types of data distributions. In this work we have tested for $k = 10, 50, 80, 100$. We have observed that k ranging from $10 - 50$ gives the best results for removal of strike-off while greater values of k tend to include the meaningless intensity variations when measuring the similarity of images. The flexibility of k allows it to extended to match the Cycle Loss of CycleGAN when $k = 100$.

The trained model on a Synthetic strike-off data set with various values of coefficient k produces clean images of the strike-off words in the English language. It can be observed in figure 3 that variation in k significantly changes the perceptual quality of the image. The results on unconstrained collection of data can be seen in figure 4. The approach is language agnostic and script/grammar independent. It can be extended to other scripts (such as Bengali, Devanagari, etc.). We have tested the model in seven types of strike-offs as cross, wave, scratch, diagonal, single line, double line, zig-zag. The results showed that VCGAN removed most of the strike-offs, although stroke types wave, scratch and zig-zag are challenging to remove. We obtained best results on cross, diagonal, single line and double line stroke types with $F1$ score of 97.40%. We used a CNN-LSTM-CTC module to perform handwritten text recognition tasks on the cleaned image generated by VCGAN, that produced a character error rate of 7.64% and word accuracy of 81.53%. Our approach is not able to clean the strike-off thoroughly 6. The most complex strike-offs are scratch, wave and zigzag. The scratch type of stroke is the most difficult one to remove. However, our approach is able to recover most of the instances of scratched ones.

8 Conclusion

In this work, we posed the strokes as intensity variations in structural information of the clean text. TV_k norm measures the similarity between two images while ignoring the intensity variations which interfere with image semantic structures. It was observed that when k value was kept small i.e. in the range $10 - 50$, the model was better at accounting the intensity variations due to various deformities. For smaller k, only the pixels with much higher intensity variations were selected so the most affected pixels due to strike-offs would be picked. In this range, VCGAN was able to preserve

the variations, which were a part of image semantic structures. With k above this range, VCGAN could not differentiate between the semantic structures and deformities, and thus the deformities that were spatially very close to the original text structures were also preserved while computing image similarity. VCGAN could recover strike-off text significantly even with the complex cases of strike-off strokes like wave, zig-zag and scratch. It outperformed the state of the art methodologies and achieved an $F1$ score of 97.40% on the generated images, and corresponding HTR achieves a character error rate of 7.64% and word accuracy of 81.53%. The proposed VCGAN can be extended to other applications of image to image translations as well. The objective loss function TV_k empowers the VCGAN with flexibility of expanse of semantic dependencies to be preserved in order to improve perceived image quality in various applications. In future works, we wish to optimize the selection procedure of the auxiliary coefficient k to foreshorten the domain of k values which contribute to enhance perceived visual quality of images.

References

[1] Adak C, Chaudhuri BB (2014) An approach of strike-through text identification from handwritten documents. In: 2014 14th International Conference on Frontiers in Handwriting Recognition, IEEE, pp 643–648

[2] Arlandis J, Pérez-Cortes JC, Cano J (2002) Rejection strategies and confidence measures for a k-nn classifier in an ocr task. In: Object recognition supported by user interaction for service robots, IEEE, pp 576–579

[3] Banerjee J, Namboodiri AM, Jawahar C (2009) Contextual restoration of severely degraded document images. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 517–524

[4] Banerjee J, Namboodiri AM, Jawahar C (2009) Contextual restoration of severely degraded document images. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 517–524

[5] Bannigidad P, Gudada C (2016) Restoration of degraded historical kannada handwritten document images using image enhancement techniques. In: International Conference on Soft Computing and Pattern Recognition, Springer, pp 498–508

[6] Bannigidad P, Gudada C (2017) Restoration of degraded kannada handwritten paper inscriptions (hastaprati) using image enhancement techniques. In: 2017 International Conference on Computer Communication and Informatics (ICCCI), IEEE, pp 1–6

[7] Bathla AK, Gupta SK, Jindal MK (2016) Challenges in recognition of devanagari scripts due to segmentation of handwritten text. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, pp 2711–2715

[8] Brink A, van der Klauw H, Schomaker L (2008) Automatic removal of crossed-out handwritten text and the effect on writer verification and identification. In: Document Recognition and Retrieval XV, International Society for Optics and Photonics, p 68150A

[9] Eltay M, Zidouri A, Ahmad I, et al (2022) Generative adversarial network based adaptive data augmentation for handwritten arabic text recognition. PeerJ Computer Science 8:e861

[10] Fan Y, Lyu S, Ying Y, et al (2017) Learning with average top-k loss. Advances in neural information processing systems 30

[11] Fogel S, Averbuch-Elor H, Cohen S, et al (2020) Scrabblegan: Semi-supervised varying length handwritten text generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4324–4333

[12] Goodfellow I, Pouget-Abadie J, Mirza M, et al (2014) Generative adversarial nets. Advances in neural information processing systems 27

[13] Heil R, Vats E, Hast A (2021) Strikethrough removal from handwritten words using cyclegans. In: Lladós J, Lopresti D, Uchida S (eds) Document Analysis and Recognition – ICDAR 2021. Springer International Publishing, Cham, pp 572–586

[14] Heil R, Vats E, Hast A (2022) Paired image to image translation for strikethrough removal from handwritten words. arXiv preprint arXiv:220109633

[15] Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision, Springer, pp 694–711

[16] Khobragade RN, Koli NA, Lanjewar VT (2020) Challenges in recognition of online and off-line compound handwritten characters: a review. Smart Trends in Computing and Communications pp 375–383

[17] Liao M, Shi B, Bai X, et al (2017) Textboxes: A fast text detector with a single deep neural network. In: Thirty-First AAAI Conference on Artificial Intelligence

[18] Marti UV, Bunke H (2002) The iam-database: an english sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition 5(1):39–46

[19] Nicolas S, Paquet T, Heutte L (2006) Markov random field models to extract the layout of complex handwritten documents. In: Tenth International Workshop on Frontiers in Handwriting Recognition, Suvisoft

[20] Nisa H, Thom JA, Ciesielski V, et al (2019) A deep learning approach to handwritten text recognition in the presence of struck-out text. In: 2019 International Conference on Image and Vision Computing New Zealand (IVCNZ), IEEE, pp 1–6

[21] Nisa H, Ciesielski V, Thom J, et al (2021) Annotation of struck-out text in handwritten documents. In: Proceedings of the 25th Australasian Document Computing Symposium, pp 1–7

[22] Pande SD, Jadhav PP, Joshi R, et al (2022) Digitization of handwritten devanagari text using cnn transfer learning–a better customer service support. Neuroscience Informatics 2(3):100,016

[23] Perez L, Wang J (2017) The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:171204621

[24] Poddar A, Chakraborty A, Mukhopadhyay J, et al (2021) Detection and localisation of struck-out-strokes in handwritten manuscripts. In: International Conference on Document Analysis and Recognition, Springer, pp 98–112

[25] Poddar A, Chakraborty A, Mukhopadhyay J, et al (2021) Texrgan: a deep adversarial framework for text restoration from deformed handwritten documents. In: Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing, pp 1–9

[26] Rajiv KS, Amardeep SD (2010) Challenges in segmentation of text in handwritten gurmukhi script. In: International Conference on Business Administration and Information Processing, Springer, pp 388–392

[27] Rusu AI, Govindaraju V (2005) On the challenges that handwritten text images pose to computers and new practical applications. In: Document Recognition and Retrieval XII, International Society for Optics and Photonics, pp 84–91

[28] Shonenkov A, Karachev D, Novopoltsev M, et al (2021) Handwritten text generation and strikethrough characters augmentation. arXiv preprint arXiv:211207395

[29] Souibgui MA, Kessentini Y (2020) De-gan: A conditional generative adversarial network for document enhancement. IEEE Transactions on Pattern Analysis and Machine Intelligence

[30] Tuganbaev D, Deriaguine D (2013) Method of stricken-out character recognition in handwritten text. US Patent 8,472,719

[31] Wadhwani M, Kundu D, Chakraborty D, et al (2021) Text extraction and restoration of old handwritten documents. In: Digital Techniques for Heritage Presentation and Preservation. Springer, p 109–132

[32] Wigington C, Stewart S, Davis B, et al (2017) Data augmentation for recognition of handwritten words and lines using a cnn-lstm network. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), IEEE, pp 639–645

[33] Zhu JY, Park T, Isola P, et al (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232