I.J. Information Engineering and Electronic Business, 2021, 6, 48-61
Published Online December 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2021.06.05

A Novel Approach for Video Inpainting Using Autoencoders

Irfan Siddavatam
Department of Information Technology, K J Somaiya College of Engineering, Mumbai, Maharashtra, India
Email: irfansiddavatam@somaiya.edu

Ashwini Dalvi
Department of Information Technology, K J Somaiya College of Engineering, Mumbai, Maharashtra, India
Email: ashwinidalvi@somaiya.edu

Dipti Pawade
Department of Information Technology, K J Somaiya College of Engineering, Mumbai, Maharashtra, India
Email: diptipawade@somaiya.edu

Akshay Bhatt
Department of Information Technology, K J Somaiya College of Engineering, Mumbai, Maharashtra, India
Email: akshay.vb@somaiya.edu

Jyeshtha Vartak
Department of Information Technology, K J Somaiya College of Engineering, Mumbai, Maharashtra, India
Email: jyeshtha.v@somaiya.edu

Arnav Gupta
Department of Information Technology, K J Somaiya College of Engineering, Mumbai, Maharashtra, India
Email: arnav.gupta@somaiya.edu

Received: 04 September 2021; Accepted: 12 October 2021; Published: 08 December 2021

Abstract: Inpainting is a task undertaken to fill in damaged or missing parts of an image or video frame, with believable content. The aim of this operation is to realistically complete images or frames of videos for a variety of applications such as conservation and restoration of art, editing images and videos for aesthetic purposes, but might cause malpractices such as evidence tampering. From the image and video editing perspective, inpainting is used mainly in the context of generating content to fill the gaps left after removing a particular object from the image or the video. Video Inpainting, an extension of Image Inpainting, is a much more challenging task due to the constraint added by the time dimension. Several techniques do exist that achieve the task of removing an object from a given video, but they are still in a nascent stage. The major objective of this paper is to study the available approaches of inpainting and propose a solution to the limitations of existing inpainting techniques. After studying existing inpainting techniques, we realized that most of them make use of a ground truth frame to generate plausible results. A 'ground truth' frame is an image without the target object or in other words, an image that provides maximum information about the background, which is then used to fill spaces after object removal. In this paper, we propose an approach where there is no requirement of a 'ground truth' frame, provided that the video has enough contexts available about the background that is to be recreated. We would be using frames from the video in hand, to gather context for the background. As the position of the target object to be removed will vary from one frame to the next, each subsequent frame will reveal the region that was initially behind the object, and provide more information about the background as a whole. Later, we have also discussed the potential limitations of our approach and some workarounds for the same, while showing the direction for further research.

Index Terms: Object Removal, Image Inpainting, Video Inpainting, Background Regeneration, Autoencoders.

1. Introduction

Video editing is crucial in many walks of life today. From educational videos for toddlers to tampering criminal evidence, the applications of removing objects from videos are wide. Often, while preparing tutorial videos, or videos of presentations, there are disturbances in the frame which distract the viewer. Something as simple as a water bottle may cause the focus to shift from the presenter. In making a video summary or documenting work on a project in the form of a video, it may so happen that while shooting, the camera captures some unwanted or unrelated objects. Also, some universities ask their prospective students to send in their applications along with a video just as many companies schedule video interviews for employees. In such a make or break situation the video mustn't contain any objectionable material.

For the millennial generation, photo and video editing go hand-in-hand with having a good online presence. If someone participates in adventure sports and would like to show it off using a video, they wouldn't want anything else to capture the viewers' attention. Thus, there are instances of people removing unwanted objects from photos and videos of their vacations. Such an application also has the potential to tamper crucial criminal evidence or CCTV footage. Hence, this technology too has applications falling in the darker side and can be misused if it falls in bad hands.

This major objective of this paper is to develop an architecture that will successfully remove a target object from a given video and regenerate the background of the removed object while maintaining consistency with the rest of the video. To study the available systems, existing solutions and their limitations; through research is carried out in the literature survey section. In our unique contribution, we will be training an Autoencoder network to detect the object in the frames of the video, remove the object, and inpaint the missing region in the frame. Our approach to achieve this will require a set of reference images of the object to be removed, and a video which contains the object. The encoders are used for learning the features and representations of the inputs. In this case, we will have two sets of inputs; an input video and a set of reference images of the target object. The encoder will be tasked with learning the features of the background and the target object from the input video frames and reference images respectively. The decoder will then locate the object in the video frame, fade out its presence, and recreate the background by varying multiple loss functions. The output will be a video from which the target object has been successfully removed and no signs of editing or tampering are obvious to the viewer. Such an implementation will not only have a variety of applications but will also be a major contribution to the video inpainting domain which is still nascent. Additionally, our novel approach to the task of removing objects from a video will work even if a ground truth frame is not available. For a given video, a 'ground truth' frame is a frame where the object that is to be removed is absent, thereby containing the maximum possible information about the background.

The remaining sections of the paper are organized as follows: the existing techniques to achieve the same goal are discussed in the following section. Later, we elaborate on our proposed methodology and some of its potential limitations. Prospective countermeasures to overcome said limitations are described and a pathway to further exploration is shown.

2. Literature Review

Inpainting is an operation done to realistically complete images or video frames and restore corrupted parts if any. Consequently, it is frequently used for restoring ancient photographs. Keeping in mind the objective of this paper – removing objects from a video and regenerating the background, inpainting can be used to recreate the background after the object has been removed from the video frames. Video Inpainting is an extension of Image Inpainting and has an added time dimension.

DeepFakes are a relatively modern phenomenon in which a person present in a photo or video is replaced by other using Deep Learning techniques. In the process of creating DeepFakes, a set of images of faces of the people in the video, both source and target, are learned by the model in order to implement the swap. The model learns the facial features to detect the person and then modify the features of the source image. The proposed architecture extends the idea of learning features similar to the ones used for DeepFake creation. Thus, to have a profound understanding of existing inpainting techniques, both for images and videos, as well as the best models of DeepFake creation, we conducted a three-fold literature survey to determine a base model for our paper. The table 1 provides an overview of various research papers, the techniques used for Image inpainting. For the illustration purpose the benefits of each method over other techniques, and their limitations is discussed in detailed.

Table 1. A comparison of existing Image Inpainting techniques.

[1] Context Encoders: Feature Learning by Inpainting	Objectives	– A context encoder is used, instead of a traditional encoder architecture, that combines semantics of the visual structures along with the appearance – Given a mask as an input that indicates the missing region, the context information of the features is encoded – The decoder uses the feature representation to predict missing content
	Benefits	– Fills in missing regions where it can get hints from the nearby pixels – Semantic inpainting and feature learning enable filling missing regions, even if the ground truth is not known to the model. This, however, results in blurry output
	Limitations	– Does not work for irregular holes because encoder considers the structure of the hole only during training, not inference – Results in blurry, distorted images as the CNN does not learn information from regions at a spatial distance from the hole in the image
[2] Semantic Image Inpainting with Deep Generative Models	Objectives	– Model is trained to learn from a large number of surrounding pixels and derive context – Missing content is predicted using context and prior losses
	Benefits	– Works well for large or arbitrary regions without retraining the model as is the case with context encoders – Can fill the missing region even in the absence of input masks for the missing region – The inference is drawn regardless of the structure of the missing region
	Limitations	– Fails if there is no information about the region to be reconstructed – The model works well for simple structures such as faces but isn't dense enough for complex scenes
[3] Generative Face Completion	Objectives	– Employs a generator and two discriminators to complete the facial features – The generator predicts an output which is then passed to the set of discriminators – One of them checks if it is consistent with the face, whereas the other checks for compatibility with the rest of the image
	Benefits	– Successfully regenerate semantically coherent and visually consistent facial parts from noise – Works well even with varying inputs of the masks
	Limitations	– The network doesn't recognize the orientation and locality of the faces in the image – Cannot deal with randomly cropped faces and unaligned faces – Does not fully comprehend the spatial correlations between adjacent pixels and details such as lip color
[4] General Deep Image Completion with Lightweight Conditional Generative Adversarial Networks	Objectives	– The training strategy generates four representative types of corruptions to enhance learning generalization that can complete various types of corrupted images
	Benefits	– Overcomes the unstable training problem of GAN where the training time of generator and discriminator is different
	Limitations	– Not generalized enough
[5] Patch-Based Image Inpainting with Generative Adversarial Networks	Objectives	– The generator of the GAN is trained to reconstruct the missing regions – The generator is modified to accommodate both local continuity and holistic features – Adversarial loss is optimized to obtain realistic outputs – Searches for best fitting patch
	Benefits	– Overcomes undesired artifacts and noise – Ensures local and global consistency
	Limitations	– Training of the generator and discriminator requires different hyperparameters – No visual semantics – Limited to available image statistics
[6] Generative Image Inpainting with Contextual Attention	Objectives	– First roughly estimates the missing content using a convolutional network with a reconstruction loss – Then uses contextual attention layers to process patches using known features – Contextual attention also boosts spatial coherence
	Benefits	– Works well for novel structures – Ensures both local and global consistency
	Limitations	– For very different test images, similar attention maps are returned which shows that the model may suffer from an optimization problem such as being stuck in a local minima – The reconstruction loss results in a blurry image, but cannot be ignored as it is essential in regularizing the GAN – Works for rectangular holes only

[7] Image Inpainting for Irregular Holes Using Partial Convolutions	Objectives	– Uses a Partial Convolutional Layer, comprising a masked and re-normalized convolution operation followed by a mask-update step – A binary mask is received as input and result depends only on the filled region – Masks are auto-updated after each step and after going through sufficient layers and updates, masked regions diminish
	Benefits	– Works for irregular masks while overcoming color discrepancies and blurriness – Fills missing regions that need not be in the center of the image
	Limitations	– Fails for sparsely structured images
[8] Structural inpainting	Objectives	– Combines PatchGAN and Context Encoder, which requires half the input than the traditional encoder-decoder. The decoder takes input from a bottleneck at the end of the encoder producing the half-sized output – Model is trained to reconstruct the half-sized square central part of a square natural-color image, that is the part being greyed out in the network's input – The loss is a combination of spatial pixel distance and feature space
	Benefits	– Overcomes blurriness – Gives better results by combining PatchGANs and Context Encoders
	Limitations	
[9] Probabilistic Semantic Inpainting with Pixel Constrained CNNs	Objectives	– Focuses not only on creating a diverse number of plausible inpaintings but also match pixel constraints and pick the most likely outcome – Model is based on sampling each pixel in an image conditioned on all the previously sampled pixels – Inputs are the whole image and a masked version of the image
	Benefits	– Able to generate multiple samples as a single task – Exhibits high sample diversity – Details such as facial symmetry and eye color are maintained
	Limitations	– Computationally intensive training and sampling – Slow to train thus limiting application to large scale images – Requires as many forward passes as there are pixels in the image which is much more than convention
[10] Deep Inception Generative Network for Cognitive Image Inpainting	Objectives	– Adopts a network-in-network approach – Utilize more complex structures to abstract the data within diverse receptive fields and explore enough cognitive understanding – The model has to decide based on broken artifacts what the target image is. Human vision can do it, computer vision not always
	Benefits	– Inpainting is no longer restricted to regular shape masks and rectangular regions – Robust for arbitrary completion including custom draw-in masks – Improves computer vision cognition
[11] Progressive Image Inpainting with Full-Resolution Residual Network	Objectives	– The strategy is that intermediate restorations should be of high quality to restrict error propagation – Consists of a full resolution residual (FRR) network, an N blocks 1 dilation strategy for mask updating and step losses – FRR Blocks maintains a full resolution branch for feature integration, and texture prediction – Assigning more FRR Blocks in a single mask updating step improves results – Step loss accounts for intermediate restoration
	Benefits	– Extends progressive inpainting strategy to irregular holes, and enables full utilization of prior information about data distribution – Proven to avoid vanishing gradient problem
[12] High-Resolution Image Inpainting Using Multi-Scale Neural Patch Synthesis	Objectives	– Joint optimization of global image content and local textures which predicts recurring details by matching and adjusting patches – Features extracted from middle convolution layers used to recreate missing content – Consists of a content network that takes an input image with the central region removed, a texture network that produces neural patches similar to neighbors, and a joint loss function
	Benefits	– Introduction of the texture network drives the generation of high-frequency details while the content network maintains global consistency – Can also be used for denoising, superresolution, retargeting, and so on
	Limitations	– Takes a long time to achieve desired results – Sometimes introduces discontinuities or artifacts

[13] Free-Form Image Inpainting With Gated Convolution	Objectives	– Introduces a new patch-based spectral normalization GAN loss that is formulated by applying the discriminator on image patches, trains faster and more stable – Gated convolutions used instead of vanilla convolutions to avoid ambiguity during training due to valid and invalid pixels in masked regions
	Benefits	– Provides a learnable dynamic feature selection framework for each channel at the spatial location instead of treating all input pixels as valid like vanilla convolutions – Generates higher quality, more flexible outputs
	Limitations	– Gated convolutions require additional parameters and model needs to be slimmed
[14] Foreground-Aware Image Inpainting	Objectives	– Learns foreground first, then inpaints. Foreground contour is detected, completed and then the whole image is completed – Completed contour, along with the input image then guides to fill other holes – Multiple loss functions such as focal loss to rank the pixels by their importance, SN-GAN loss to obtain sharper results, content loss, and adversarial loss. However, content and adversarial losses are not applied simultaneously and curriculum training is used
	Benefits	– Uses contour prediction to guide inpainting as other learning-based methods may not predict the areas where holes overlap or touch foreground objects as they are trained to fill random masks
[15] EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning	Objectives	– Generates edges first, then image completion. Separate networks for each task, both of an adversarial framework, first to predict edges and then adjust RGB pixel intensities accordingly – Edge generator follows a GAN architecture where the generator uses dilated convolutions and discriminators are PatchGAN – The image completion network utilizes the edge map and the incomplete input image to deduce the missing regions by combining the original background and predicted edges, and uses a perceptual loss with style loss
	Benefits	– Maximizing patch similarity or propagating background data sometimes fails to reconstruct complex details or leave blurry object outlines. In such cases, the edge generation approach works best
	Limitations	– When edge maps are provided by multiple images, the output tends to share characteristics of both images – Edge generation fails in highly textured areas or in cases of large arbitrary holes

We discussed the various techniques of Image Inpainting in table 1 along with their advantages and disadvantages. It was observed that most of these methods used generative models or generative adversarial networks, to recreate the missing parts of an image. The difference between these methods is in the implementation of the generative networks – patch-based GANs, different loss functions to optimize, novel architectures, and so on. Moreover, they all have a common advantage of successfully recreating the missing parts of an image even when the damaged areas are of a shape that has not been encountered before. The other techniques discussed are semantic in nature and include encoders in the form of CNNs, inception networks, and the like, which work well for irregular patches but are computationally intensive.

Another vertical for our literature survey is to explore the video inpainting techniques. Video inpainting is an extension of image inpainting with larger search space and temporally consistent constraints. These additional constraints make video inpainting a much more difficult task compared to conventional image inpainting. As a result, the video inpainting algorithms available at present are few in number and relatively nascent. Owing to these added aspects specific to video inpainting, table 1 discuss the various techniques used for video inpainting, the benefits of those methods over others, and their limitations.

Table 2. A comparison of existing Video Inpainting techniques.

[16] Video inpainting under constrained camera motion	Objectives	– Holes in frames are filled by prior information from other important frames and then the leftover portion is filled – The moving objects are restored first independent of the background – The background is later filled in by directly copying information from other existing frames or by extending texture synthesis techniques to the spatiotemporal domain
	Benefits	– The technique combines motion-based and spatial inpainting to produce three image mosaics that deal with motion as well as speed up the overall process
	Limitations	– Since the background is dynamic the simple segmentation video did not replicate the current boundaries for moving objects in the video – A simple segmentation video was not able to replicate the background when the background was dynamic

[17] VORNet: Spatio-temporally Consistent Video Inpainting for Object Removal	Objectives	– Combines the information from previous frames and generates results in the current frame. Prior information is collected by using the optical flow to capture background motion and recover the removed foreground part by warping the previous background accordingly. For the constantly occluded region, existing image-based inpainting models could generate plausible results. – A refinement network is designed to select and refine these candidates to derive a spatially and temporally consistent result. – Training is done using reconstruction loss, perceptual loss and two designed GAN losses to evaluate the quality of videos with mean square error, a learned perceptual metric and visual results
	Benefits	– Generate visually plausible and temporally coherent results online, without post-processing – It generates clear videos even for diverse datasets.
	Limitations	– Uses optical flow to get information from the previous frames, which results in extra execution time and parameters – Still unable to capture the object motions in detail and there is an unavoidable occlusion problem, which makes the warped frames blurry.
[18] Deep Blind Video Decaptioning by Temporal Aggregation and Recurrence	Objectives	– A neural network model is used where the encoder takes input in the form of source frames to extract visible pixels. These are then fed to the decoder as input – Model is improved with recurrent feedback that achieves temporal coherence and provides knowledge about missing pixels
	Benefits	– Video frames are successfully recreated even when there is a lot of activity in the video – Minute intricacies and textures are regenerated even if there is frequent change in lighting – Most effective to remove text from video
	Limitations	– In case of images with solid shadows, results are blurry
[19] Deep Flow-Guided Video Inpainting	Objectives	– A spatial and temporal coherent optical flow field is generated using a deep flow completion network. – Missing pixels can then be passed on and warped from available data and then small pixels can be filled using predictions
	Benefits	– In comparison to other methods, the runtime speed is fast and no assumptions are required for missing regions and movement of the video content.
	Limitations	– The flow of video inpainting is inaccurate on the edge of a car and similarly for other irregular geometrical objects
[20] Deep Video Inpainting	Objectives	– Built on a model using a feed-forward deep network, a network is designed to gather and fine-tune information from adjacent frames and recreate missing areas. – The output is made to be temporally coherent by a recurrent feedback and memory module
	Benefits	– If sufficient memory is available, the video produces excellent results as the temporal consistency is improved.
	Limitations	– Where there is a large occlusion in a video, the color saturates irregularly. The discrepancy error leads to inaccurate warping. Also, there will be cases where memory would be insufficient and, in such cases, results would be elusive
[21] Frame-Recurrent Video Inpainting by Robust Optical Flow Inference	Objectives	– The proposed model combines ConvLSTM and motion for creating the spatio-temporal coherency in the input video – Computation is optimized by dealing with larger frame videos by streaming in real-time
	Benefits	– There is no problem with video length, produces results in real-time streaming, and can deal with large motions. Also, the framework can produce inpainted video frames with spatial details and temporal coherence where the advantage lies in the strong capability of ConvLSTM and it can also fix up holes if the given set of input is proper which has to be fed with two sources of flow.
[22] Free-form Video Inpainting with 3D Gated Convolution and Temporal PatchGAN	Objectives	– The generator uses a 3D convolutional layer to service the masked video by learning the difference between unmasked, filled, and masked areas. – The discriminator is a temporal Patch GAN discriminator that penalizes high-frequency spatial-temporal features and improves coherency by combining various losses
	Benefits	– Improved on preexisting two-stage adversarial models (EdgeConnect) and patch-based methods of free form inpainting that had very high computation times and were limited to repetitive patterns. This method overcomes these limitations by modeling the distribution of real videos and generating realistic results only by forward inference, without searching.
	Limitations	– Does not work when the video is far from training data or masked area is too thick

[23] Recurrent Temporal Aggregation Framework for Deep Video Inpainting	Objectives	– A deep feed-forward network with temporal aggregation and a timely propagation of useful information from previous reference frames in sequence – Video decaptioning network automatically detects and inpaints over text by residual learning and loss functions. Residual learning deals only with the corrupted pixels in the frames. Multiple frames are taken as input by the encoder, but only the middle frame is reconstructed to form the memory layer. Missing pixels are deduced by temporal pooling – The video inpainting network learns to align references onto target frames instead of following 3D convolutions. Visible patches are then aggregated to predict missing content
	Benefits	– To overcome intensive computation and dependence on previously computed motion, a 3D encoder-2D decoder is used so that features are traced from the video itself. – Video inpainting network works better than existing methods for large, arbitrary masks by handling dynamic content with temporal aggregation and recurrence.
	Limitations	– Results are blurry in case of solid shadows – Where there is a large occlusion in a video, the color saturates irregularly – Regions not in the time frame also end up being blurry
[24] Copy-and-Paste Networks for Deep Video Inpainting	Objectives	– The network copies pixels from reference frames and pastes them in the target frames – The network focuses on alignment and context matching to combine frames based on similarity – A self-supervised alignment network that can accommodate images with large missing areas – Frame by frame processing of the video wherein target frames are aligned with the references, pixels are copied from the reference to the copy-network, outputs of copy-network are used to fill missing regions, and the completed frame may be used as a reference
	Benefits	– Uses affine matrix instead of optical flow so that it works for large occlusions or near-static objects – It also works for extracting pixels from temporally distant frames.
[25] An Internal Learning Approach to Video Inpainting	Objectives	– The network follows a consistency-aware training strategy that captures motion consistency and traverses it across various time frames thereby handling overall consistency – Input is the video and mask representing known regions. A combination of the two is used to generate the target from a noise frame by the generator – Once the generator, which is an encoder-decoder network is trained, it can be used to generate all frames
	Benefits	– Internal learning for using the recurrence of visual patterns in images. Knowledge of images is encoded – Does not require dataset hence overcomes the difference between video data and image data
	Limitations	– Long processing time – May fail for large holes or object is mostly static with respect to the background

After reviewing various video inpainting approaches in table 2, we observed that most of these techniques require a particular frame, called a 'ground truth' frame, in their dataset that already has the content which is to be recreated in the frames where there is damage or missing areas. The way this particular frame is used differs from one approach to another. Some methods use this frame for training or to learn features from it, whereas some others directly use data from this frame to complete the damaged frames. This frame is also used by some networks to propagate data over a series of subsequent frames. Thus, methods that use a ground truth frame are known to achieve good spatial as well as temporal consistency in their output. That being said, the presence of such a frame in every input video or dataset cannot be guaranteed, which creates the need for a method that doesn't depend on such ground truth frames.

DeepFake videos are a result of applying Deep Learning techniques to generate faces swapped videos. They do not require a ground truth frame for successfully producing face-swapped outputs. Thus, in our pursuit to develop a technique of video inpainting that bypasses the requirement of a ground truth frame, table 3 discuss the techniques used for DeepFake creation and detection, the benefits of those methods over other techniques, and their limitations.

Table 3. An overview of popular DeepFake creation and detection techniques.

[26] Deep Learning for Deepfakes Creation and Detection	Objectives	– A two encoder-decoder network is used to swap faces. Each pair trains on an image set and the encoder's features are shared between the pair. Thus, the encoder learns the similarity between the two sets and is now able to reconstruct faces or even swap them – Later versions also accommodate adversarial and perpetual losses
	Benefits	– Weakness cannot be exploited easily as training is much more generalized than detection techniques
	Limitations	– Can be detected by a variety of methods if the method of creation is known to the detector

[27] Media Forensics and DeepFakes: an overview	Objectives	– The object can be spliced from a different or same image and pasted – Fake content can also be generated with segmentation maps, text descriptions or even sketches by using style transfer GANs and autoencoders
	Benefits	– Some methods hide traces in the Fourier domain – Sometimes noise is injected in the target image. Such noise designed to fool a particular architecture of detectors is not transferable and lives in the crux of images
	Limitations	– Can easily be detected by blind methods, one-class classifiers, and supervised methods – Deep learning detectors used facial asymmetry, spatio-temporal modifications in videos, variation in color of faces dues to blood flow for detection
[28] Everybody Dance Now	Objectives	– The first method to enable motion transfer between two videos or targets by image-to-image translation – Pose detection is done and stick figures are learned, global pose normalization is done to learn the differences between source and target bodies, locations, and motions – In the generator, the mappings from localized stick figures are learned and applied to target with adversarial training – A discriminator is then used to classify the output as real or fake
	Benefits	– Create novel frames instead of altering existing ones – Uses motion transfer among 2D subjects to mitigate the retargeting problem – Able to distinguish movement from appearance and transfer facial expressions
	Limitations	– Fails in case of missing or noisy locations from the pose detection phase which transfer the errors in later stages – If motion speed is different in input video and videos used for training. Results in jittery outputs – Scaling does not account for the difference in length of limbs or body
[29] A Style-Based Generator Architecture for Generative Adversarial Networks	Objectives	– Instead of giving the features to the generator via an input layer, a learned constant is passed – The input is then normalized and mapped to obtain an intermediate output – Instances of this intermediate output guide the generator as learned affine transforms which then undergo a number of convolutions and style transfer operations with noise introduced to it after each step in the generator – This noise is generated stochastically and the noise image is finally broadcasted to all feature maps after appropriate scaling and added to the output of each convolution
[30] On Hallucinating Context and Background Pixels from a Face Mask using Multi-scale GANs	Objectives	– The proposed cascade GAN model is used to generate context and background pixels from a face mask only – The outputs are produced at multiple resolutions by taking the weighted sum of the various loss functions used in training – The model takes face mask inputs during training and sets its weights such that they are parameterized by the ground truth.
	Benefits	– Requires only a few training images instead of images with each attribute such as age, gender, pose, hairstyle, and so on – The model generates own training data by introducing variations in gender, ethnicity, backgrounds. Also capable of generating stock images
[31] Deepfake Video Detection Using Recurrent Neural Networks	Objectives	– Two training image sets are used: Set of original faces, and a set of extracted faces from videos – Two autoencoders are trained separately and a latent representation learned from the face extracted from the input video to the decoder network to train on the target face to be inserted in the video
	Limitations	– Scene inconsistency due to insufficient information – The output is not temporally consistent since the video is processed frame-by-frame and lacks previous frame data

To summarize, in table 1 gives brief overview about the image inpainting techniques, table 2 discuss about the video inpainting techniques while table 3 discuss the techniques available for deepfake creation. After reviewing the existing literature on Image and Video Inpainting (discussed in table 1, 2 and 3), we can infer that there are various existing models designed in recent years to recreate missing parts of images and subsequently videos. Our observation is that the majority of these techniques rely on the presence of a frame in the input video which does not contain the object at all, that is a ground truth image. This frame is then used as a point of reference when the target object is removed and the background needs to be regenerated. Therefore, in this paper, we come up with a model that would work without any such constraint on the video which is discussed in the section 3.

3. Proposed Methodology and Discussion

As we observed from the literature survey, most of the existing methods for image and video inpainting rely on a ground truth frame to successfully inpaint the missing regions. However, such systems may be fallible in cases of novel scenes, or the absence of ground truth frames. Thus, there is a need for a method that is not dependent on such ground truth frames. To achieve this first logical step would be learning the background of the video. We can understand that as the object moves through the frames of the video, more portion of the background is revealed, and more contexts can be derived. As a result, enough contexts can be gathered from the input video frames, which are then used to recreate the background, once the target object is removed. Broadly, the entire process can be divided into three modules-

- Learning the context or background from the input video,
- Learning the features of the object so that they can be used to detect the object,
- Object removal and background reconstruction.

As proof of concept, we developed a model that only takes a video containing the target object as its input. There is no restriction on the video in terms of availability of the ground truth frame, that is, a video with no ground truth frame, works just as well with this implementation. In predicting the background, the information is gathered about the background while the object moves and doesn't stay in one place for too long. The limitation with approach is obvious that if the object doesn't move to provide enough contexts for the background, the prediction may not be accurate. For the demonstration purpose, we have considered the video of a horse running across a field. From that video, a frame shown in fig 1a is extracted and feed as an input to system. The system is now supposed to learn the background of this image so the other frames in the same video sequence are extracted and processed so that all the details should be considered while generating the background. Finally, the median of all extracted frames is taken and an estimate of the background was produced as seen in Fig 1b. Alternatively, taking the mode of all frames also produces plausible results. Now this background can be used to mask any image in the frame.

Fig. 1a. Input video Frame to learn background

Fig. 1b. Predicted background for upcoming frames

After background prediction, each frame is individually processed using the pre-trained YOLO object detection model to locate the target object in the frames. The YOLO model returns the coordinates of a bounding box that contains the object. Once the object is located, all pixels of the box are set to zero, that is, the object is removed and then taking the predicted background from the previous step, the pixels are filled and the frame is completed. In the sample shown in fig 1a, the object is the horse. Using YOLLO model boundaries for horse is detected and later it is inpainted using the background detected in fig 1b and as a result frame in fig 2 is generated.

Fig 2. Background recreated and frame completed

The output frames are then put in the correct sequence and stitched to make the output video. Alternatively, one can also use a pre-trained Mask-RCNN instead of the YOLO model. Even though this program gives plausible results and validates our approach, we observed that it isn't infallible either and has limitations such as the constraint on the target object to be in motion throughout its presence in the video, slight change in color observed on the edges of the target object bounding box (as seen in Fig 3), and masking of the objects behind the target object in some cases (Fig 4). The following result was produced using the Mask-RCNN model.

Fig 3. Mask-RCNN input frame

Fig 4. Mask-RCNN output

To check the feasibility of our method, we implement a python program that would take an input video, extract frames from it, detect the target object in it using a pre-trained object detection model - YOLO or Mask-RCNN, and remove the object. The background is predicted using the input frames, and finally used to fill the missing spaces. This implementation is successful in giving plausible results with an accuracy of 93% with the YOLO model and 65% with the Mask-RCNN model, even though Mask-RCNN provides a more visually appealing result.

Thus, we infer that with this approach the limitation on the technique, that is the requirement of ground truth frame, is overcome. However, since there are some visible shortcomings as discussed in the previous paragraph, this approach needs to be refined to meet our objective.

3.1 Proposed Approach

On studying the process of creating DeepFakes and learning about the various state of the art neural network architectures, we deduce that an Autoencoder architecture will be an appropriate option to reach our goal. An Autoencoder is characterized by one or more encoders and decoders. Its objective is to learn the representation of the data provided as input. It does so by reducing the input to an encoding of a smaller dimension and then tries to generate an output as close to the original input, from the encoding obtained. In this case, we will have two sets of inputs; an input video and reference images of the target object, and the two encoders will be tasked with learning the features of the background and the target object from the respective inputs. The decoder then fades out the target object from the frame and recreates the background by optimizing multiple loss functions. Thus, with this approach, we intend on treating video inpainting as analogous to a DeepFake creation task.

The proposed model takes an input video and a set of reference images of the target object, to learn the features of the background and the target object. Frames are then extracted from the input video and the encoder-decoder pair trains to learn the background from the video. Another encoder-decoder pair is simultaneously trained to learn the features of the target object from the set of reference images. Since object removal and background reconstruction requires representations learned by both of these, we suggest using a single decoder among both these pairs.

The autoencoder to be used here will have two encoders, corresponding to the two inputs, and a single decoder producing three outputs. With this approach, the task of removing the object and recreating the background can be done by optimizing three interdependent loss functions. The network successfully learns the representation of the background (from the extracted frames of the video) and the target object by optimizing two losses – frame reconstruction loss and object reconstruction loss respectively (L_a, L_b). The third custom loss function, object removal loss (L_c) can be computed such that it minimizes the frame reconstruction loss, but maximizes the object reconstruction loss, thereby gradually removing the object and recreating the background in the frame. Such a function can be defined as:

$$L_c = -\big(\big(L_a \times \log(L_a)\big) - \big((1 - L_b) \times \log(1 - L_b)\big)\big) \quad (1)$$

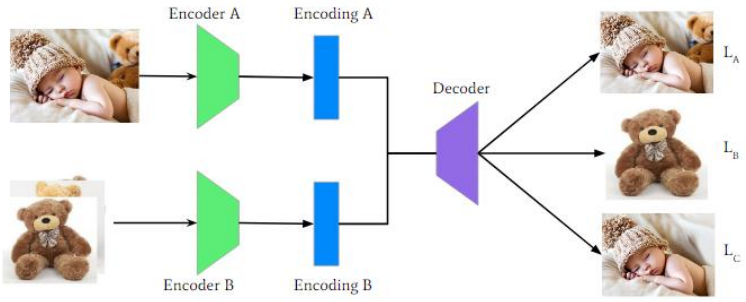

Fig 5. Initially proposed architecture

3.2 Limitations Of Proposed Architecture

A potential problem with the architecture shown in Fig 5 is the impact of the third loss function L_c, on the learning capability of the encoders. In optimizing all loss functions simultaneously, the model could fail to learn the individual representations of the input frames and the target object. Thus, the three loss functions may need to be optimized independently.

To achieve that, instead of having a single model do all the work, one could try splitting the tasks as shown in Fig 6. By using two encoder-decoder pairs and a third decoder, the effect of L_c on the other encoders can be eliminated. The encoders each learn the representation of the input frames of the video and the target object. The two encoder-decoder pairs should be trained simultaneously using a custom training loop. The encodings of the input frame obtained from the encoder are used by the third decoder to construct the output image according to the optimized loss function. In the same training loop, losses of each of the two networks (L_a, L_b) will be used to compute the removal loss (L_c) at each iteration. Thus, all three losses can now be optimized simultaneously, but now independent of each other.

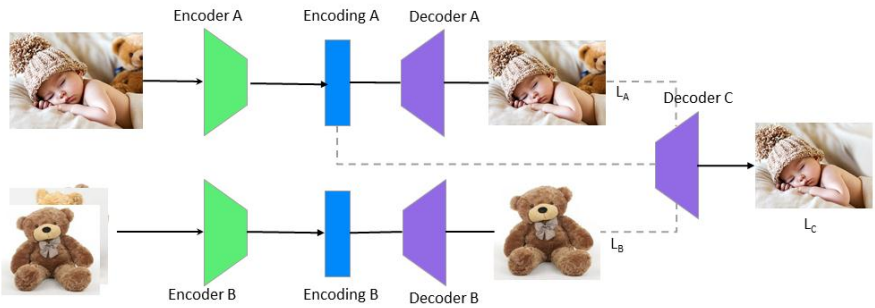

Fig 6. Modified architecture

Another potential limitation could be the loss tending to zero due to logarithmic computation in the custom loss function. Such a value results in a black output. One way to overcome this limitation could be to define the function such that as the reconstruction loss decreases, the removal loss should increase. Hence, a decreasing function in the range of the reconstruction loss could be a suitable alternative. One such function is the negative log function. We also suggest adding a loss which is a combination of two competing losses, one of which is used in removal and the other in reconstructing the background, to the function. This new loss can be computed as the mean absolute error between the output produced (*C*) and the input frame (*A*). Thus, the new object removal loss is computed as:

$$L_c = 0.5 \times \left(meanabsoluterror(C, A)\right) - 0.5 \times \log(L_b) \quad (2)$$

Since this is a firsthand approach for video inpainting, there is scope for further assessment of the same. We have come to observe that the removal loss function plays a very crucial role in determining the output of the model as a whole, and are confident that with the correct function, the desired output can be obtained. Thus, a significant improvement in the concerned loss functions and if required, the existing autoencoder model could give promising results.

4. Conclusion

Video inpainting is an upcoming research area with deep learning techniques being the latest technology to be used for the task. On reviewing recently published works in the domain, we have observed that most of the techniques make use of a ground truth frame for a variety of purposes in the entire inpainting process.

Our proposed method aims at overcoming this dependency on ground truth frames and employs neural networks for learning and inpainting the video frames. The neural network architecture we have proposed is that of an autoencoder. The inspiration for such an architecture stems from neural network models used for DeepFake creation. An autoencoder is generally characterized by a combination of encoders and decoders. It learns the representations of its inputs by downsampling them and has the capability of fabricating desired outputs by varying the parameters with which the learned encodings are upsampled.

We intend to utilize this property of autoencoders to learn representations of input video frames and reference images of the target object. The target object can then be detected in the video frames, and by optimizing a combination of loss functions, we propose to remove the object and simultaneously complete the background. However, in modeling video inpainting as a DeepFake creation task, there are a couple of potential limitations with the architecture and the loss functions that we have pointed out which could pave the way for further study.

References

[1] Feature Learning by Inpainting. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2536-2544 http://openaccess.thecvf.com/content_cvpr_2016/html/Pathak_Context_Encoders_Feature_CVPR_2016_paper.html

[2] Raymond A. Yeh, Chen Chen, Teck Yian Lim, Alexander G. Schwing, Mark Hasegawa-Johnson, Minh N. Do (2017). Semantic Image Inpainting with Deep Generative Models. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5485-5493. http://openaccess.thecvf.com/content_cvpr_2017/html/Yeh_Semantic_Image_Inpainting_CVPR_2017_paper.html

[3] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang (2017). Generative Face Completion. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3911-3919. http://openaccess.thecvf.com/content_cvpr_2017/html/Li_Generative_Face_Completion_CVPR_2017_paper.html

[4] Ching-Wei Tseng, Hung Jin Lin, Shang-Hong Lai (2017). General Deep Image Completion with Lightweight Conditional Generative Adversarial Networks. British Machine Vision Conference 2017, London, UK, September 4-7, 2017. http://www.bmva.org/bmvc/2017/papers/paper080/paper080.pdf

[5] Ugur Demir, Gozde Unal (2018). Patch-Based Image Inpainting with Generative Adversarial Networks. arXiv preprint arXiv:1803.07422. https://arxiv.org/abs/1803.07422

[6] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas S. Huang (2018). Generative Image Inpainting with Contextual Attention. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5505-5514. http://openaccess.thecvf.com/content_cvpr_2018/html/Yu_Generative_Image_Inpainting_CVPR_2018_paper.html

[7] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, Bryan Catanzaro (2018). Image Inpainting for Irregular Holes Using Partial Convolutions. The European Conference on Computer Vision (ECCV), 2018, pp. 85-100. http://openaccess.thecvf.com/content_ECCV_2018/html/Guilin_Liu_Image_Inpainting_for_ECCV_2018_paper.html

[8] Huy V. Vo, Ngoc Q. K. Duong, Patrick Perez (2018). Structural inpainting. Proceedings of the 26th ACM international conference on Multimedia, Pages 1948–1956. https://doi.org/10.1145/3240508.3240678

[9] Emilien Dupont, Suhas Suresha (2019). Probabilistic Semantic Inpainting with Pixel Constrained CNNs. arXiv preprint arXiv:1810.03728. https://arxiv.org/abs/1810.03728

[10] Qingguo Xiao, Guangyao Li, Qiaochuan Chen (2019). Deep Inception Generative Network for Cognitive Image Inpainting. arXiv preprint arXiv:1812.01458. https://arxiv.org/abs/1812.01458

[11] Zongyu Guo, Zhibo Chen, Tao Yu, Jiale Chen, Sen Liu (2019). Progressive Image Inpainting with Full-Resolution Residual Network. Proceedings of the 27th ACM International Conference on Multimedia, Pages 2496–2504. https://doi.org/10.1145/3343031.3351022

[12] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, Hao Li (2017). High-Resolution Image Inpainting Using Multi-Scale Neural Patch Synthesis. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6721-6729. http://openaccess.thecvf.com/content_cvpr_2017/html/Yang_High-Resolution_Image_Inpainting_CVPR_2017_paper.html

[13] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas S. Huang (2019). Free-Form Image Inpainting With Gated Convolution. The IEEE International Conference on Computer Vision (ICCV), 2019, pp. 4471-4480.

http://openaccess.thecvf.com/content_ICCV_2019/html/Yu_Free-Form_Image_Inpainting_With_Gated_Convolution_ICCV_2019_paper.html

[14] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, Jiebo Luo (2019). Foreground-Aware Image Inpainting. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5840-5848. http://openaccess.thecvf.com/content_CVPR_2019/html/Xiong_Foreground-Aware_Image_Inpainting_CVPR_2019_paper.html

[15] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, Mehran Ebrahimi (2019). EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning. arXiv preprint arXiv:1901.00212. https://arxiv.org/abs/1901.00212

[16] Kedar A. Patwardhan, Guillermo Sapiro, Marcelo Bertalmío (2007). Video inpainting under constrained camera motion. IEEE Transactions on Image Processing, vol. 16, no. 2, pp. 545-553. https://ieeexplore.ieee.org/abstract/document/4060949/

[17] Y. Chang, Z. Y. Liu, W. Hsu (2019). VORNet: Spatio-temporally Consistent Video Inpainting for Object Removal. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 2019, pp. 1785-1794. https://ieeexplore.ieee.org/document/9025383

[18] Dahun Kim, Sanghyun Woo, Joon-Young Lee (2019). Deep Blind Video Decaptioning by Temporal Aggregation and Recurrence. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4263-4272. http://openaccess.thecvf.com/content_CVPR_2019/html/Kim_Deep_Blind_Video_Decaptioning_by_Temporal_Aggregation_and_Recurrence_CVPR_2019_paper.html

[19] Rui Xu, Xiaoxiao Li, Bolei Zhou, Chen Change Loy (2019). Deep Flow-Guided Video Inpainting. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3723-3732. http://openaccess.thecvf.com/content_CVPR_2019/html/Xu_Deep_Flow-Guided_Video_Inpainting_CVPR_2019_paper.html

[20] Dahun Kim, Sanghyun Woo, Joon-Young Lee, In So Kweon (2019). Deep Video Inpainting. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5792-5801. http://openaccess.thecvf.com/content_CVPR_2019/html/Kim_Deep_Video_Inpainting_CVPR_2019_paper.html

[21] Yifan Ding, Chuan Wang, Haibin Huang, Jiaming Liu, Jue Wang, Liqiang Wang (2019). Frame-Recurrent Video Inpainting by Robust Optical Flow Inference. arXiv preprint arXiv:1905.02882. https://arxiv.org/abs/1905.02882

[22] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, Winston Hsu (2019). Free-form Video Inpainting with 3D Gated Convolution and Temporal PatchGAN. The IEEE International Conference on Computer Vision (ICCV), 2019, pp. 9066-9075. http://openaccess.thecvf.com/content_ICCV_2019/html/Chang_Free-Form_Video_Inpainting_With_3D_Gated_Convolution_and_Temporal_PatchGAN_ICCV_2019_paper.html

[23] D. Kim, S. Woo, J. Lee, and I. S. Kweon (2019). Recurrent Temporal Aggregation Framework for Deep Video Inpainting. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 5, pp. 1038-1052. https://ieeexplore.ieee.org/abstract/document/8931251/

[24] Haotian Zhang, Long Mai, Ning Xu, Zhaowen Wang, John Collomosse, Hailin Jin (2019). An Internal Learning Approach to Video Inpainting. The IEEE International Conference on Computer Vision (ICCV), 2019, pp. 2720-2729. http://openaccess.thecvf.com/content_ICCV_2019/html/Zhang_An_Internal_Learning_Approach_to_Video_Inpainting_ICCV_2019_paper.html

[25] Sungho Lee, Seoung Wug Oh, DaeYeun Won, Seon Joo Kim (2019). Copy-and-Paste Networks for Deep Video Inpainting. The IEEE International Conference on Computer Vision (ICCV), 2019, pp. 4413-4421. http://openaccess.thecvf.com/content_ICCV_2019/html/Lee_Copy-and-Paste_Networks_for_Deep_Video_Inpainting_ICCV_2019_paper.html

[26] Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi (2019). Deep Learning for Deepfakes Creation and Detection. arXiv preprint arXiv:1909.11573. https://arxiv.org/abs/1909.11573

[27] Luisa Verdoliva (2020). Media Forensics and DeepFakes: an overview. arXiv preprint arXiv:2001.06564. https://arxiv.org/abs/2001.06564

[28] Caroline Chan, Shiry Ginosar, Tinghui Zhou, Alexei A. Efros (2019). Everybody Dance Now. The IEEE International Conference on Computer Vision (ICCV), 2019, pp. 5933-5942 http://openaccess.thecvf.com/content_ICCV_2019/html/Chan_Everybody_Dance_Now_ICCV_2019_paper.html

[29] Tero Karras, Samuli Laine, Timo Aila (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4401-4410. http://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html

[30] Sandipan Banerjee, Walter Scheirer, Kevin Bowyer, Patrick Flynn (2020). On Hallucinating Context and Background Pixels from a Face Mask using Multi-scale GANs. The IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 300-309. http://openaccess.thecvf.com/content_WACV_2020/html/Banerjee_On_Hallucinating_Context_and_Background_Pixels_from_a_Face_Mask_WACV_2020_paper.html

[31] D. Güera and E. J. Delp (2018). Deepfake Video Detection Using Recurrent Neural Networks. 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6, doi: 10.1109/AVSS.2018.8639163. https://ieeexplore.ieee.org/abstract/document/8639163/

Authors' Profiles

Dr Irfan Siddavatam received his PhD in Electronics Engineering from Veermata Jijabai Technological Institute, Mumbai in 2018. He is Associate Professor in Department of Information Technology at K.J.Somaiya College of Engineering, where he has been since 2001. His research interest includes Information and Cybersecurity, Artificial Intelligence, Machine Learning, Blockchain. His work includes developing intelligent solutions to ensure secure cyberspace not only for critical infrastructure but also for children and women safeguarding Irfan has published research papers in domain of Artificial Intelligence, Security. Irfan has worked with reputable government and software industries, including, India Smart Grid Forum (ISGF), Tata Power Delhi Distribution Limited (TPDDL), ABB GGISPL GLOBAL INDUSTRIES AND SERVICES Private Limited.

Ashwini Dalvi is pursuing her PH.D. Degree from VJTI, affiliated to Mumbai University. She joined the Department of Information Technology, K. J. Somaiya College of Engineering, Mumbai in 2006 as an Assistant Professor. She has published over 25 journal and conference papers in the areas of Security, Intelligent applications

Dipti Yogesh Pawade received B.E. degree in Computer Science and Engineering from Sant Gadge Baba University in 2009 and M.E. degree in Embedded System and Computing from G. H. Raisoni College of Engineering, Nagpur in 2012. Since 2012 she is an Assistant Professor in the Department of Information Technology at K J Somaiya College of Engineering, Vidyavihar, Mumbai. Her interest includes Machine learning, Web Security, and Web Application Development.

How to cite this paper: Irfan Siddavatam, Ashwini Dalvi, Dipti Pawade, Akshay Bhatt, Jyeshtha Vartak, Arnav Gupta, "A Novel Approach for Video Inpainting Using Autoencoders", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.13, No.6, pp. 48-61, 2021. DOI: 10.5815/ijieeb.2021.06.05

