# 2022 年臺灣國際科學展覽會
# 優勝作品專輯

作品編號　**190019**

參展科別　電腦科學與資訊工程

作品名稱　**Art Recovery through PConv (Partial Convolutions) and GLCIC (Globally and Locally Consistent Image Completion)**

得獎獎項　三等獎


就讀學校　臺北美國學校

指導教師　**Jude Clapper**

作者姓名　郭馥愷



關鍵詞　**Image Inpainting、Machine Learning**

# 作者簡介

我是郭馥愷，目前就讀台北美國學校十一年級。十年級在學校選修 AP Computer Science 才開始接觸電腦科學領域，從此對這個領域產生極高的興趣，所以也透過線上課程進修相關程式語言與類神經網路。

音樂、大提琴、藝術、電腦科學都是我的興趣。我很幸福，無論我想做什麼，我的父母都無條件地支持；也很幸運，求學過程中遇到很多鼓勵我的老師。感謝教授的指導，讓我能將機器學習與藝術做結合，完成這次的科展作品。

研究時數據分析結果有時不如預期，但當成果出現時就會非常開心，我喜歡這種挑戰！

# 研究摘要

在生成性模型(Generative Models)中的一個主要應用就是“影像修復” (Image Inpainting) 也稱為“影像完成”(Image completion)。 影像修復經常被應用於許多影像處理，包含在生活照片中移除背景不必要的物件再回填移除後缺損的影像。 但是，或許之前的研究較多著墨於技術而非美學，至目前為止，很少有影像修復的研究著重於藝術作品的重建應用。

所以，本研究計畫提出三個新的模型來針對藝術作品做更優化的影像修復，以達到較一般處理日常照片所使用的如 Places2 和 ImageNet 等影像修復技術在視覺上更為自然逼真的處理： 第一種模型是 PConv (Partial Convolutions)，它利用部分旋積(partial convolution) 來避免一般由於遮蔽區域中畫素起始值設定而常見的影像模糊問題。 第二種模型是 GLCIC (Globally and Locally Consistent Image Completion)，是一種以 GAN (Generative Adversarial Network) 為基礎，進一步在全域鑑別器 (global discriminator) 之上，再建構一個區域性鑑別器(local discriminator)，以確保在整體畫面與細部畫面的合理與一致性的方法。 最後一個模型是一個在本研究中所提出的全新、整合性的模型－PConv-GAN。 在這個創新的模型中，我們將 GLCIC 模型中常用於旋積過程中”補零”(zero padding) 的手法，以 PConv 模型中部分旋積的方式來取代。最後我們會利用一系列的印象派畫作為例，以 $L^1$ loss 和 PSNR (peak signal-to-noise ratio) 兩種方法來評估這三個模型。

# Abstract

One of the main applications of generative models has been image inpainting, or image completion. Image inpainting has been utilized for various purposes, of which include removing background objects and restoring damage from photos of daily life. However, very few image inpainting methods up to date have applied inpainting to the reconstruction of artworks, possibly due to a mainstream focus on the technical rather than applicative aspects of image completion.

Thus, we propose three inpainting models for the recovery of artworks as the majority of inpainting models are currently evaluated on real-life images such as Places2 and ImageNet and hence may not produce visually plausible results for art pieces. The first two models are as follows: the partial convolutional inpainting model (PConv), which avoids the problem of blurriness inherent in fixing initial pixel values in the masked area of an image, and Globally and Locally Consistent Image Completion (GLCIC), a GAN (generative adversarial network) with an added local discriminator on top of a global discriminator for both local and global visual consistency. The final model evaluated is a novel, integrated model - PConv-GAN - where the standard zero padding of the convolutional layers of GLCIC is replaced with partial convolutional-based padding. The three models are then evaluated on a collection of Impressionist artworks by $L^1$ loss and PSNR (peak signal-to-noise ratio).

## 1. Introduction

Generative models are deep-learning models that are capable of producing images through modeling patterns in the image dataset they are given. They have many applications, of which include turning low resolution images to high resolution ones or black and white images to colored ones, among image-to-image translation, text-to-image translation, video frame prediction, and much more. One of the applications of generative models is inpainting, and although the origins of inpainting without machine learning can be traced back to as far back as a few centuries since photos were developed [1], inpainting with artificial intelligence only started developing considerably several years ago starting with Context Encoders [2]. Numerous approaches to inpainting subsequently have been made over the course of the past few years. Notably, nearly all those approaches were evaluated on real life images of datasets such as but not limited to ImageNet [3] and Places2 [4]. This trend can be attributed to a general focus on creating state-of-the-art models for benchmark datasets that all virtually contain real life images. Thus, approaches with respect to artworks are relatively scarce, even though artwork recovery itself is an arguably important topic – it preserves works that may communicate important messages or reflect historical trends. Hence, this paper emphasizes the potential application of deep-learning inpainting to art recovery while comparing two models - PConv (Partial Convolutions) [5] and GLCIC (Globally and Locally Consistent Image Completion) [6] - evaluated on a custom art dataset.

## 2. Related Work

Approaches to inpainting are generally separated into two categories: non-machine learning and learning. Non-machine learning approaches solely rely on propagating appearance information from neighboring pixels, which results in several acknowledged problems. The first is that such models can only produce visually plausible results when given small holes in which the color and texture variance is small. When bigger areas are missing from the image, the models generally produce outputs that may be over-smoothed or structurally unsound. This problem is particularly amplified by the models' lack of semantic awareness. Another issue is the models' inability to produce novel objects, which may also be especially a problem when much information from the image is missing. Examples of non-learning approaches to inpainting include patch-based methods, where the model iteratively searches for relevant patches of information in the image's non-hole regions. Such patch-based approaches are often computationally expensive, and although PatchMatch [7] speeds it up with a faster algorithm, it is still not real time and the problems inherent in non-learning methods are still prevalent.

Deep learning approaches have proved to produce superior results compared to non-learning approaches in inpainting. They are able to generate novel objects in the missing regions and display a knowledge of semantics. However, many are prone to blurry outputs, and at times, illogical structures. Liu et al. [5] recognize the issue of blurriness with inpainting methods and attributed it to the application of convolutional filters over both valid pixels and pixels with substituted values in mask. The solution was to apply convolutions only to valid pixels, creating the approach of partial convolutions. Their model based on partial convolutions were able to reach state-of-the-art inpainting results – the first model proposed in this paper will be the partial convolutional architecture evaluated on artworks.

Another deep learning approach in inpainting, GLCIC (Globally and Locally Consistent Image Completion) was proposed by Iizuka et al. [6], which built upon the encoder-decoder architecture proposed by Context Encoders by including two discriminators, global and local. The global discriminator views the entire image while the local one evaluates the completed region, ensuring both global and local consistency. Essentially, Iizuka et al.'s model is a GAN but with specialized global and local discriminators. The model was able to handle arbitrary inpainting masks and produce high resolution images – thus, the second model proposed in this paper is based on GLCIC.

## 3. Methodology

### 3.1 Dataset and Environment

The dataset in which the two models were trained on was a custom dataset created by filtering WikiArt [8] for Impressionist landscape and cityscape works in the public domain. The dataset also includes files found on a Claude Monet dataset provider by Varnez on Kaggle [9]. There are 5,124 total images in the dataset, 3756 from WikiArt and 1368 from Kaggle, with a 90% to 10% train to test split.

As seen by comparing Fig 1 and Fig 2, the contents of the custom dataset are noticeably different from the natural images ImageNet contains.

Fig 1. Sample images from custom Impressionist art dataset.

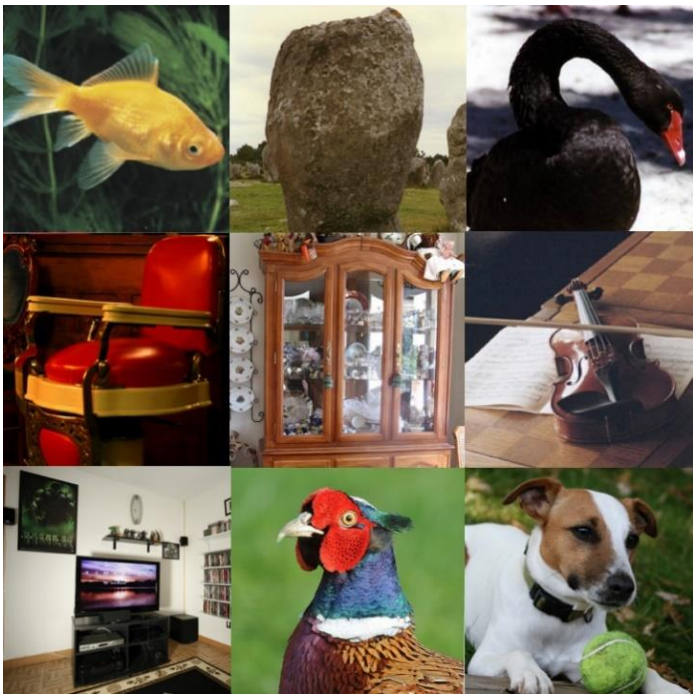

Fig 2. Sample images from ImageNet.

The two models were each trained on a NVIDIA Tesla P100-PCIE GPU with 32 GB of RAM. The integrated development environment was Google Colab Pro.

### 3.2 U-Net Architecture as Baseline for PConv

The partial convolutional network for inpainting has U-Net [10] as its baseline. U-Net was originally intended for medical image segmentation, but its applications can extend to inpainting as well. The architecture of U-Net consists of convolutional layers with ReLU (Rectified Linear Unit) as their activation that downsample the input and then upsample it at the end, effectively creating a U-shape. This choice in design allows U-Net to map contextual information end-to-end. The skip connections that are employed in U-Net further strengthen the model's capability in learning the high-level and low-level semantics of an image through concatenating feature maps of previous layers to later ones with the same dimension. Thus, U-Net is remarkable in its awareness of both high-level and low-level features, making it ideal for inpainting models that aim to capture important contextual information while also retaining fine-grained details in the output.

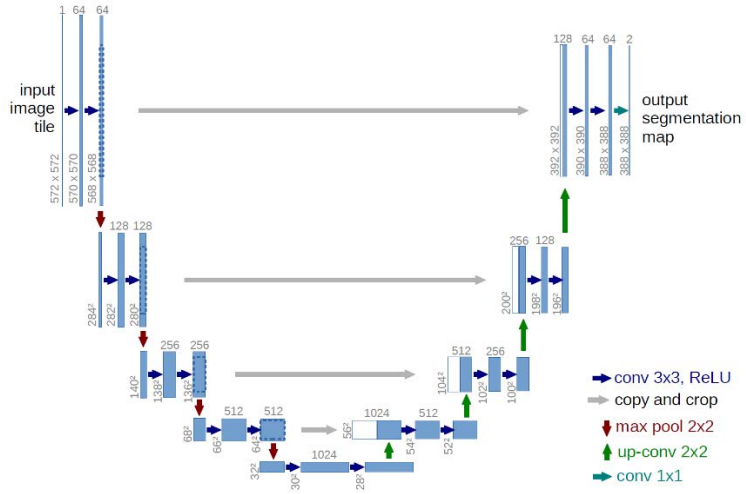


Fig 3. Architecture of U-Net.

The partial convolutional network for inpainting replaces the convolutional layers in U-Net with partial convolutions to generate more accurate pixel values in the masked area.

### 3.3 Intuition behind PConv

Many inpainting approaches fix initial values inside the masked, or covered, region of the image, often with the mean pixel value of the dataset. Then, convolutional layers are applied over the whole image, essentially treating both the real pixel values and the fixed pixel values as valid. As a result, the upsampling process results in blurred outputs in the masked areas. Partial convolutions address this problem by only performing convolutions on the "valid" regions of the image, defined as regions with at least one pixel that is not masked. After a partial convolution is performed, there will be more pixels produced around the masked area, and these pixels would be considered as valid.

Consequently, after enough iterations, the masked area would finally be replaced by valid pixels, producing the final result.

### 3.4 Loss functions for PConv

Let $\mathbf{I}_{in}$ be the input image with the hole, $\mathbf{I}_{out}$ the output of the generator/completion network, and $\mathbf{M}$ the initial binary mask (0 for holes). Let $\mathbf{I}_{comp}$ be equivalent to the image of $\mathbf{I}_{out}$ but with non-hole pixels set to ground truth. Let $K_n$ be the normalization factor for the nth selected layer.

Below are the $L^1$ losses, where $L_{hole}$ and $L_{valid}$ denote the loss for the hole and non-hole pixels respectively:

$$\mathcal{L}_{hole} = \| (1 - M) \odot (\mathbf{I}_{out} - \mathbf{I}_{gt}) \|_1$$

$$\mathcal{L}_{valid} = \| M \odot (\mathbf{I}_{out} - \mathbf{I}_{gt}) \|_1$$

$L^1$ losses target per-pixel reconstruction accuracy. Note that $L_{valid}$ is needed because the model produces not only the initially masked area when it is fed an input, but also reproduces the areas outside of the mask due to its architecture.

Below is the perceptual loss:

$$\mathcal{L}_{perceptual} = \sum_{n=0}^{N-1} \|\Psi_n(\mathbf{I}_{out}) - \Psi_n(\mathbf{I}_{gt})\|_1 + \sum_{n=0}^{N-1} \|\Psi_n(\mathbf{I}_{comp}) - \Psi_n(\mathbf{I}_{gt})\|_1$$

Perceptual loss allows for comparison of high-level features, and thus may at times be more reliable than $L^1$ per-pixel loss. $\Psi_n$ represents the activation map of the nth selected layer of ImageNet-pretrained ResNet50 [11]. ResNet50 was not pretrained on the custom dataset due to the dataset's limited size.

Below is the style loss:

$$\mathcal{L}_{style_{out}} = \sum_{n=0}^{N-1} \left|\left| K_n \left( (\Psi_n(\mathbf{I}_{out}))^\intercal (\Psi_n(\mathbf{I}_{out})) - (\Psi_n(\mathbf{I}_{gt}))^\intercal (\Psi_n(\mathbf{I}_{gt})) \right) \right|\right|_1$$

$$\mathcal{L}_{style_{comp}} = \sum_{n=0}^{N-1} \left|\left| K_n \left( (\Psi_n(\mathbf{I}_{comp}))^\intercal (\Psi_n(\mathbf{I}_{comp})) - (\Psi_n(\mathbf{I}_{gt}))^\intercal (\Psi_n(\mathbf{I}_{gt})) \right) \right|\right|_1$$

Style loss serves a similar purpose as perceptual loss. Gram matrices for autocorrelation are calculated for each feature map.

Below is the total variation loss:

$$\mathcal{L}_{tv} = \sum_{(i,j)\in P,(i,j+1)\in P} \|\mathbf{I}_{comp}^{i,j+1} - \mathbf{I}_{comp}^{i,j}\|_1 + \sum_{(i,j)\in P,(i+1,j)\in P} \|\mathbf{I}_{comp}^{i+1,j} - \mathbf{I}_{comp}^{i,j}\|_1$$

$L_{tv}$ serves as a smoothing penalty for P, which is the 1-pixel dilation of the hole region. This encourages the model to produce images with less noise.

The final loss function used for optimization with tuned hyperparameters is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{valid} + 6\mathcal{L}_{hole} + 0.05\mathcal{L}_{perceptual} + 120(\mathcal{L}_{style_{out}} + \mathcal{L}_{style_{comp}}) + 0.1\mathcal{L}_{tv}$$

## 3.5 GAN Architecture as Baseline for GLCIC

GANs contain two components – a generator which forges images and a discriminator which discerns whether the input images are real (i.e. from the training dataset) or fake (i.e. from the generator). The generator receives the prediction of the discriminator as feedback for improvement. As the discriminator improves its classification of real or fake, the generator is forced to model images that are closer to the distribution of real images. The design of GANs thus represents a zero-sum game between the generator and the discriminator, where the ultimate goal is to reach Nash equilibrium where both models maximize their optimality.

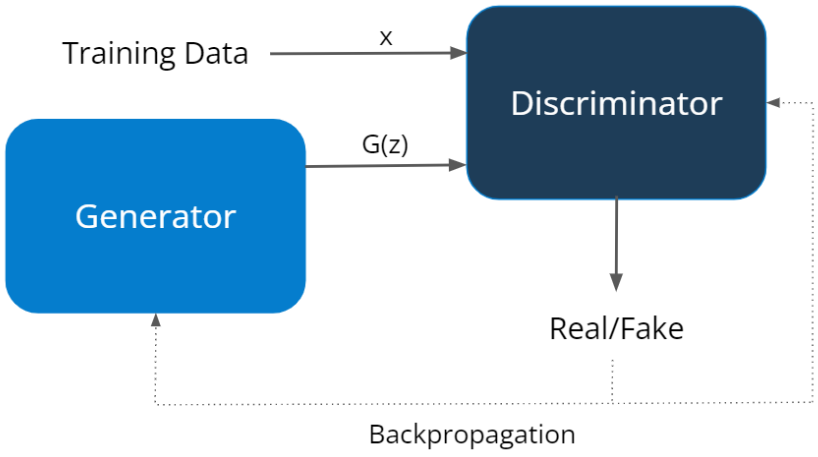


Fig 4. Architecture of a vanilla GAN

GLCIC builds upon the vanilla GAN by having two specialized discriminators – global and local. This choice in design enforces structural consistency of the whole image as well as the masked region. GLCIC also has an encoder and decoder architecture for its generator to capture semantic information from the input to model possible pixel values in the hole. It should also be noted that GLCIC fixes mean pixel values of the dataset in the masked area for the input – this point will be brought up again later in the results section.

Table 1. Architecture of the generator in GLCIC. "Output" represents the number of output channels for that layer. Each convolutional layer is followed by a batch norm and has ReLU as its activation except the last layer. The last convolutional layer is followed by a sigmoid activation to predict normalized RGB pixel values.

| **Layer** | **Kernel** | **Dilation (η)** | **Stride** | **Padding** | **Output** |
|---|---|---|---|---|---|
| Conv. | 5 x 5 | 1 | 1 x 1 | 2 | 64 |
| Conv. | 3 x 3 | 1 | 2 x 2 | 1 | 128 |
| Conv. | 3 x 3 | 1 | 1 x 1 | 1 | 128 |
| Conv. | 3 x 3 | 1 | 2 x 2 | 1 | 256 |
| Conv. | 3 x 3 | 1 | 1 x 1 | 1 | 256 |
| Conv. | 3 x 3 | 1 | 1 x 1 | 1 | 256 |
| Dilated Conv. | 3 x 3 | 2 | 1 x 1 | 2 | 256 |
| Dilated Conv. | 3 x 3 | 4 | 1 x 1 | 4 | 256 |
| Dilated Conv. | 3 x 3 | 8 | 1 x 1 | 8 | 256 |
| Dilated Conv. | 3 x 3 | 16 | 1 x 1 | 16 | 256 |
| Conv. | 3 x 3 | 1 | 1 x 1 | 1 | 256 |
| Conv. | 3 x 3 | 1 | 1 x 1 | 1 | 256 |
| Deconv. | 4 x 4 | 1 | ½ x ½ | 1 | 128 |
| Conv. | 3 x 3 | 1 | 1 x 1 | 1 | 128 |
| Deconv. | 4 x 4 | 1 | ½ x ½ | 1 | 64 |
| Conv. | 3 x 3 | 1 | 1 x 1 | 1 | 32 |
| Output | 3 x 3 | 1 | 1 x 1 | 1 | 3 |

Table 2. Architecture of the discriminators in GLCIC. FC stands for a fully-connected layer. Each convolutional layer is followed by a batch norm and has ReLU as its activation except the last layer. The last convolutional layer is followed by a sigmoid activation to predict whether an image is real or fake. *Only the global discriminator has a sixth convolutional layer.

| Layer | Kernel | Stride | Padding | Output |
|---|---|---|---|---|
| Conv. | 5 x 5 | 2 x 2 | 2 | 64 |
| Conv. | 5 x 5 | 2 x 2 | 2 | 128 |
| Conv. | 5 x 5 | 2 x 2 | 2 | 256 |
| Conv. | 5 x 5 | 2 x 2 | 2 | 512 |
| Conv. | 5 x 5 | 2 x 2 | 2 | 512 |
| *Conv. | 5 x 5 | 2 x 2 | 2 | 512 |
| FC | - | - | - | 1024 |

3.6 Loss Functions for GLCIC

Let $\mathbf{C(x, M_c)}$ be the generator in functional form, with **x** as the input image and $\mathbf{M_c}$ as the binary mask (1 for holes). Let **D(x, M)** be the discriminator in functional form.

The loss function of GLCIC consists of a weighted MSE (mean squared error) loss and a BCE (binary cross entropy) or GAN loss. The combination of these two loss functions were shown to provide training stability for Context Encoders, the backbone of GLCIC. Thus, GLCIC uses the same loss function.

Below is the MSE Loss:

$$\mathcal{L}(x, M) = \| M \odot (\mathbf{C}(x, M) - x) \|^2$$

Below is the BCE Loss:

$$\min_c \max_d \mathrm{E}\ [\ \log(\mathbf{D}(x, M)) + \log(1 - \mathbf{D}(\mathbf{C}(x, M), M)\ ]$$

E refers to the expected value over all instances. Minimax refers to the minimization of loss for the generator and the maximization of the loss for the discriminator. A higher loss of the discriminator in most cases means that the generator is producing images that are similar to the real image distribution, hence lowering its loss.

3.7 Partial Convolution-Based Padding GAN

We propose a new inpainting model based on Globally & Locally Consistent Image Completion. Instead of the standard zero padding that was originally used in the GLCIC model, we utilize partial convolution-based padding in this model.

Padding is an important feature of convolutional layers as it allows us to design deeper networks by decelerating volume size while preserving information at the boundary of an image. However, standard zero padding has the problem of adding extrapolated data to an input image. Such added features may deteriorate the performance of the network - thus, we replace zero based padding in GLCIC with partial convolution-based padding which solely conditions convolutional output on valid input pixels (non-extrapolated).

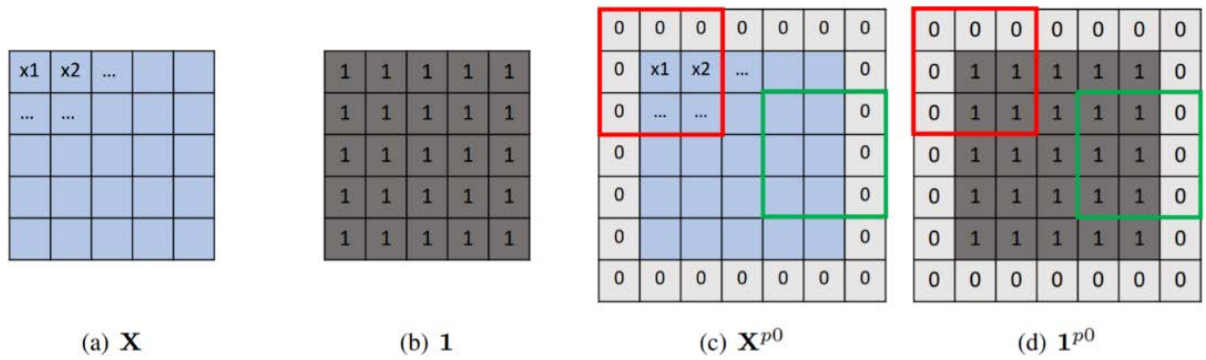


Fig 5. **X** represents the input image while **1** represents the matrix of ones that is of the same dimension as **X**. Partial convolution-based padding utilizes $\mathbf{X^{p0}}$, the zero padded input image, and $\mathbf{1^{p0}}$, a matrix that indicates which pixels to perform convolutions on (the ones being valid pixels and the zeroes invalid), while standard zero padding only utilizes $\mathbf{X^{p0}}$. The red and green boxes represent sliding convolution windows.

## 4. Results and Discussion

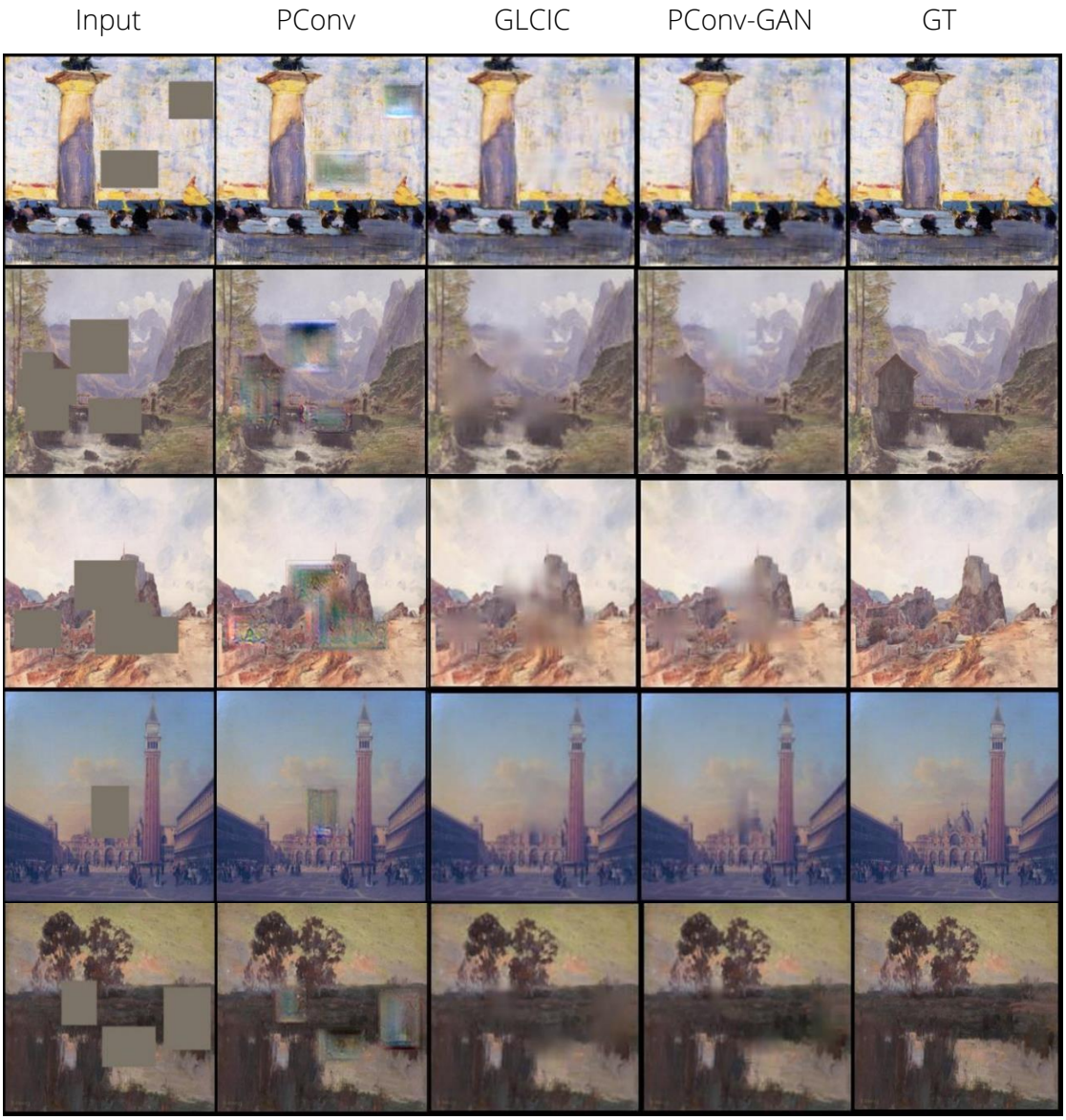


Fig 6. Comparison of inference results using regular-shaped masks on aforementioned impressionism artwork dataset. GT refers to the ground truth.

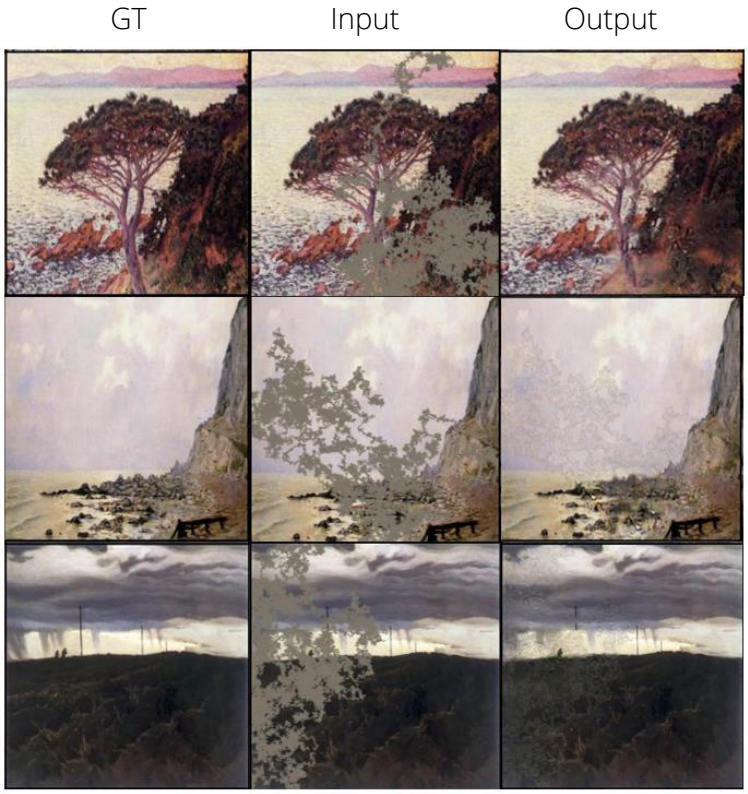


Fig 7. Inference results of the partial convolutional inpainting model using irregular-shaped masks on custom impressionism artwork dataset

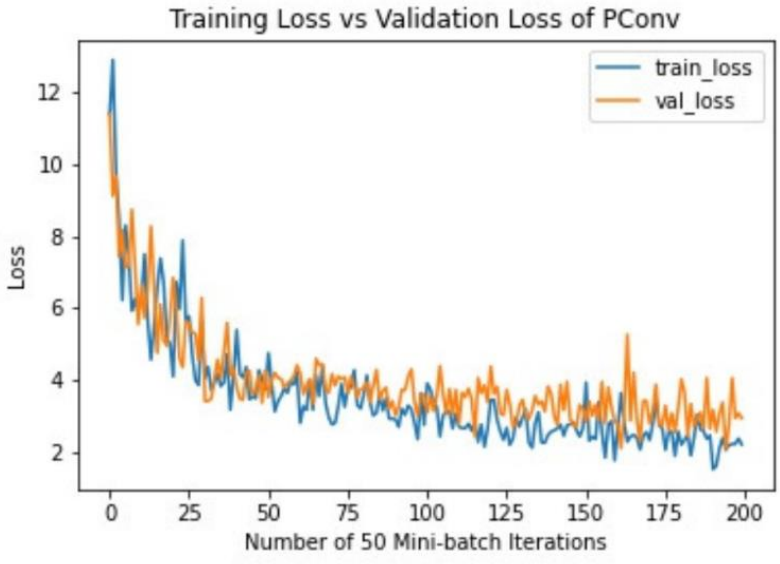


Fig 8. Loss curve of the partial convolutional model

Table 3. Comparison of models with PSNR (peak signal-to-noise ration) and relative $L^1$ evaluation metrics. A lower score for $L^1$ is favorable, while a higher score for PSNR is favorable.

| | L1 | PSNR |
|---|---|---|
| PConv | 0.0462 | 22.9 |
| GLCIC | 0.0127 | 31.1 |
| PConv-GAN | **0.00879** | **33.1** |

We can see the loss curve for the partial convolutional model in Fig 8 - training was halted as soon as validation losses plateaued and training losses fell below the validation losses to prevent overfitting. We do not display the loss curve for GLCIC and our proposed model because a GAN's loss curve, which oscillates due to the minimax loss function, does not accurately represent the generator's optimization and is hence not meaningful.

We chose two metrics for evaluating the models – $L^1$ and PSNR (peak signal-to-noise ratio). $L^1$ targets per-pixel reconstruction accuracy between the output and the ground truth, while PSNR targets the corrupting noise present in an output that affects its fidelity given a reference/ground truth image. Based on both our metrics for $L^1$ and PSNR, our proposed inpainting model, Partial Convolution-Based Padding GAN (PConv-GAN), performs better than both the partial convolutional inpainting model and Globally & Locally Consistent Image Completion for regular masks on our custom artwork dataset, as seen in Table 3. GLCIC for both metrics performs better than the partial convolutional model, but the completed area is slightly blurry, possibly due to fixing the mean pixel value, the problem PConv intended to fix. This issue is also apparent, though to a lesser degree, for our model.

The partial convolutional model, which was proposed in 2018, later than GLCIC, performed surprisingly poor for regular masks in our experiment. However, we can see that the model performed reasonably for irregular masks, which is unexpected given how it would initially seem that irregular masks seem more difficult for an inpainting model to deal with than regular ones. Thus, we hypothesized that because the partial convolutional model was trained with irregular masks that allowed for more sporadic information, the model primarily learned to interpolate information between intermittent pixels, an act that is not as easily done with regular masks where there is less information exposed. The original PConv model proposed by Liu et al. was intended to be trained on irregular masks, which may imply the authors found that the model was more suitable for such a particular type of mask as well. However, their official implementation of this model has not been released yet, so this hypothesis remains inconclusive.

## 5. Conclusion and Future Works

Inpainting is an important problem in computer vision and has been researched by numerous technology companies such as NVIDIA and Adobe. Nevertheless, there is a lack of attention concerning artwork-based inpainting compared to that of natural images. We find that our proposed model is suitable for art recovery and performs better than the previously proposed inpainting models of partial convolutions and Globally & Locally Consistent Image Completion at least on impressionist artworks such as the ones in our dataset. Though the partial convolutional model does not produce visually plausible outputs for regular masks, it is able to create structurally coherent images when given irregular masks. GLCIC produces visually coherent outputs for regular masks as well. Future works may include the design of novel models for art recovery or inpainting in general, given deep learning inpainting is a relatively new field with numerous possibilities for experimentation.

## Bibliography


[1] Darrow, E. J. (2009, October 6). *Pietro Edwards and the restoration of the public pictures of Venice, 1778–1819: necessity introduced these arts*. University of Washington Libraries Digital Collections. https://digital.lib.washington.edu/researchworks/handle/1773/6223

[2] Pathak, D. (2016, April 25). *Context Encoders: Feature Learning by Inpainting*. ArXiv.Org. https://arxiv.org/abs/1604.07379

[3] *ImageNet*. (2007). ImageNet. https://www.image-net.org/index.php

[4] Khosla, A. (n.d.). *Places2: A Large-Scale Database for Scene Understanding*. Places2. http://places2.csail.mit.edu/download.html

[5] Liu, G. (2018, April 20). *Image Inpainting for Irregular Holes Using Partial Convolutions*. ArXiv.Org. https://arxiv.org/abs/1804.07723

[6] Iizuka. (2017). *Globally and Locally Consistent Image Completion*. http://iizuka.cs.tsukuba.ac.jp/projects/completion/data/completion_sig2017.pdf

[7] *PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing*. (2009). Princeton Computer Science PIXL Group Publications. https://gfx.cs.princeton.edu/pubs/Barnes_2009_PAR/index.php

[8] *WikiArt.org - Visual Art Encyclopedia*. (n.d.). https://www.wikiart.org/

[9] Varnez. (2020, July 6). Claude Monet Dataset Provider. Kaggle. https://www.kaggle.com/varnez/claude-monet-dataset-provider

[10] Ronneberger. (2015). *U-Net*. U-Net. https://arxiv.org/pdf/1505.04597.pdf

[11] He, K. (2015, December 10). *Deep Residual Learning for Image Recognition*. ArXiv.Org. https://arxiv.org/abs/1512.03385

[12] *U-Net: Convolutional Networks for Biomedical Image Segmentation*. (n.d.). U-Net. https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/

[13] Liu, G. (2018, November 28). Partial Convolution based Padding. ArXiv.Org. https://arxiv.org/abs/1811.11718

# 【評語】190019

1. 利用深度學習 GLCIC 模型的局部與全域特性的兩個鑑別器做影像修復效果看起來還不錯，可以用在古畫毀損自動修復，有應用價值。

2. 但是較缺乏對於模型為何有此優勢之解釋，做進一探討分析或與其他 baseline 做比較。

3. 英文表達能力佳。