



國立中山大學應用數學系

碩士論文

Department of Applied Mathematics

National Sun Yat-sen University

Master's Thesis

以深度學習張量特徵萃取實現超解析成像

Tensor feature extraction in deep learning for super-resolution
imaging

研究生： 黃鼎勳

Ding-Hsun Huang

指導教授： 李宗鍾 博士

Dr. Tsung-Lin Lee

中華民國 111 年 7 月

July 2022

國立中山大學研究生學位論文審定書

本校應用數學系碩士班

研究生黃鼎勳（學號：M092040011）所提論文

以深度學習張量特徵萃取實現超解析成像

Tensor feature extraction in deep learning for super-resolution imaging

於中華民國 111 年 6 月 23 日經本委員會審查並舉行口試，符合碩士學位論文標準。

學位考試委員簽章：

召集人 呂宗澤 呂宗澤 委員 李宗鍾 李宗鍾

委員 郭岳承 郭岳承 委員 卓建宏 卓建宏

委員 黃杰森 黃杰森 委員 _____

指導教授(李宗鍾) 李宗鍾 (簽名)

摘要

近年來在超解析成像（Super-resolution imaging）提高影像解析度之技術的領域，往往都是使用內插法將原始影像放大再做處理，而對於內插法的計算而言，其中隱藏許多低頻的雜訊。如何在超解析度成像的過程中，萃取出更完整的特徵，減少不必要的雜訊，使得成像能更準確、快速。因此，張量分解（tensor factorization）扮演著不可或缺的角色。

比較以往深度學習方法僅僅只將影像輸入模型訓練，本研究在前處理將影像使用高階奇異值分解（High-Order Singular Value Decomposition）萃取重要特徵，再通過模型學習 結果顯示影像先使用特徵分解之模型具有更好的成像，在相同的模型之下皆具有更好的表現。

關鍵詞：特徵萃取、高階奇異值分解、深度學習、影像處理、超解析成像。

Abstract

In recent years, in the field of super-resolution imaging technology to improve image resolution, interpolation is often used to enlarge the original image, and for the calculation of interpolation, many low-frequency noises are hidden. In the process of super-resolution imaging, more complete features are extracted, unnecessary noise is reduced, and imaging can be more accurate and fast. Therefore, tensor factorization plays an important role.

Compared with the deep learning CNN models, which only input original images into the model for training, we use HOSVD (High-Order Singular Value Decomposition) to extract important features from the images in the pre-processing, and then input them into the model. The experiments show that the image using the HOSVD has better original image, and it has better performance in the same CNN model.

Keywords: feature extraction, high-order singular value decomposition, deep learning, image processing, super-resolution imaging.

Contents

論文審定書	i
摘要	ii
Abstract	iii
1 Introduction	1
2 Preliminary	2
3 Singular Value Decomposition for tensor	2
3.1 Singular Value Decomposition	3
3.2 High-Order Singular Value Decomposition	4
4 Super-Resolution via Deep Learning	7
4.1 Super-Resolution CNN (SRCNN)	7
4.2 Very Deep Super-Resolution (VDSR)	8
4.3 ResNet (Deep residual network)	10
5 Experiments	15
6 Conclusion	18
References	19

List of Figures

1	Singular Value Decomposition	3
2	Feature extraction	4
3	Mapping the tensor to matrix	4
4	Mapping the matrix to tensor	5
5	SRCNN	7
6	VDSR	9
7	Receptive field	9
8	Training error of VDSR10 and VDSR18	10
9	淺層與深層網路架構	11
10	Residual learning unit	12
11	A simple neural network	13
12	Residual learning for a neural network	13
13	VDSR18 and ResNet18	14
14	Training error	15
15	流程圖	16

List of Tables

1	SRCNN in testing data	8
2	VDSR10 in testing data	10
3	VDSR10 and VDSR18 in testing data	11
4	ResNet18 in testing data	14
5	A simple image for low-resolution and high-resolution	16
6	A simple image in testing data	17
7	Average in testing data	18

1 Introduction

本研究的目的在於如何使用特徵萃取與深度學習的方法，來生成取得高解析度高畫質的影像。在深度學習（Deep learning）中，資料的特徵工程、重要變數分析和模型選擇等等，是訓練模型前重要的前處理過程，而特徵萃取能夠提取資料的重要特徵，並減少資料的記憶體儲存，使其能有效加快網路的學習。相較於輸入原始資料（原始影像），這對於訓練深度學習網路而言將是一個很好的初步步驟。

在特徵萃取的部分使用 SVD 分解（Singular Value Decomposition）提取重要特徵，擷取 99% 之重要資訊並對資料進行 1.5 至 2 倍不等的壓縮比例，同時進行資料壓縮和特徵萃取工程。而在 SVD 分解的過程中發現，去除那 1% 的特徵向量所對應的資料，有助於將影像去除噪音、低頻雜訊，使得影像在超解析成像（Super-resolution imaging）的過程中不會受到雜訊噪音的干擾，不管是在主觀的視覺上或是客觀的數值上都有更好的結果。

然而現今許多資料不再是向量（vector）而是以張量（tensor）的形式，例如一張灰階的影像將其數值化可以看做是一個矩陣，彩色影像以 RGB 數值化則可以看做是一個三維空間的張量。但以往所熟知的 SVD 分解方式，僅僅只能處理矩陣的資料，而不能處理更高維度的資料，因此將張量矩陣化（Matricization）[1]的方式便由此而生，HOSVD（High-Order Singular Value Decomposition）[2]成為能夠對高階張量做特徵工程的方式之一。HOSVD 的運算過程是將一個高階張量的資料對其矩陣化後，形成一個矩陣的資料再做 SVD 分解之特徵萃取，萃取完特徵後再以原本矩陣化的過程還原回高階張量的形式。

在生成高解析度高畫質的影像上，使用深度學習的卷積神經網路（Convolution Neural Network, CNN）[3]模型架構，並以人臉為資料訓練模型。本篇論文會使用三種 CNN 架構之模型，SRCNN（Super-Resolution CNN）[4]、VDSR（Very Deep Super-Resolution）[5]和 ResNet（Deep residual network）[6]，探討各個模型的架構之優劣和特性，並分別對資料有無使用 HOSVD 前處理進行實驗。

實驗以 64×64 彩色影像為輸入， 128×128 提升一倍之高解析度高畫質影像為輸出，並以峰值訊噪比（Peak signal-to-noise ratio, PSNR）[7]為主要指標，PSNR 越高代表輸出影像與原始高畫質影像越接近。結果顯示每個模型有使用 HOSVD

特徵萃取相比於沒有使用之 PSNR 皆有所提升，且在主觀視覺上有明顯變好的趨勢。

2 Preliminary

一階張量可以看成一個向量，二階張量可以看成一個矩陣。一個張量稱為高階張量，如果張量的階數（也稱之為維度）超過兩個，那麼將此張量稱之為高階張量，因此考慮一個高階張量的一個多維數組，具有 N 維數組的張量，稱之為 N 階張量。其中使用的符號包括純量（scalars）為 x ，向量（vectors）為 \mathbf{x} ，矩陣為 X ，張量為 \mathcal{X} 。

纖維（Fiber）是從張量中抽取向量的操作。在矩陣中固定其中一個維度，以得到行或者列，稱之為 1-mode fiber 或者 2-mode fiber，表示為 $\mathbf{x}_{:j}$ 和 $\mathbf{x}_{i:}$ 。在三階張量（third-order tensor）中，1-mode fiber 表示為 $\mathbf{x}_{:jk}$ ，2-mode fiber 表示為 $\mathbf{x}_{i:k}$ ，3-mode fiber 表示為 $\mathbf{x}_{ij:}$ 。

矩陣化（Matricization），也稱為展開（unfolding）或展平（flattening），是將 tensor (i_1, i_2, \dots, i_N) 映射到 matrix (i_n, j) 的轉換

$$j = 1 + \sum_{\substack{t=1 \\ t \neq n}}^N (i_t - 1) J_t \text{ with } J_t = \prod_{\substack{k=1 \\ k \neq n}}^{t-1} I_k.$$

所以一個 tensor \mathcal{X} 的 n -mode unfolding 其表示為 $X_{(n)} \in \mathbb{R}^{I_n \times (I_1 \times \dots \times I_{n-1} \times I_{n+1} \times \dots \times I_N)}$ 。

一個 third-order tensor $\mathcal{X} \in \mathbb{R}^{2 \times 2 \times 2}$ 的三個 n -mode unfolding 表示為

$$X_{(1)} = [\mathbf{x}_{:11} \quad \mathbf{x}_{:21} \quad \mathbf{x}_{:12} \quad \mathbf{x}_{:22}],$$

$$X_{(2)} = [\mathbf{x}_{1:1} \quad \mathbf{x}_{2:1} \quad \mathbf{x}_{1:2} \quad \mathbf{x}_{2:2}],$$

$$X_{(3)} = [\mathbf{x}_{11:} \quad \mathbf{x}_{21:} \quad \mathbf{x}_{12:} \quad \mathbf{x}_{22:}].$$

不難發現每次 n -mode unfolding 只是將 n -mode fiber 依序排成一列。

3 Singular Value Decomposition for tensor

在影像資訊中 Singular Value Decomposition (SVD) 是用於特徵萃取或降維的主要方法之一，其目的是找出影像的部分最大特徵值所對應之特徵向量，保留影像重要成分，並去除影響較小的特徵，以利於減少影像存取的空間和去除雜訊。

然而現今影現今彩色影像多以 RGB 三原色之形式儲存，意指將彩色影像數值化後會得到一個三維張量的資料，而傳統 SVD 分解法只能處理二維矩陣資料。因此更多將張量展開成矩陣再進行特徵分解的方式被開發出來，可以有效保留影像的資訊和完整性。接下來會先回顧 SVD，再接著介紹 HOSVD。

3.1 Singular Value Decomposition

假設 A 是一個 $m \times n$ 階矩陣，且矩陣的元素都是實數。如此則存在一個 Singular Value Decomposition (SVD 分解) 使得

$$A = U\Sigma V^T$$

其中 U 是 $m \times m$ 階正交矩陣， Σ 是 $m \times n$ 非負實數對角矩陣，而 V^T 即 V 的共軛轉置，是 $n \times n$ 階正交矩陣。

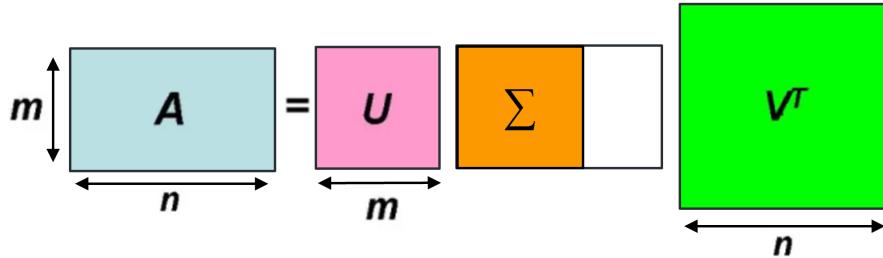


Figure 1: Singular Value Decomposition

SVD 分解主要有兩點好處。第一，利用矩陣的 SVD 大多元素為 0 的性質，可以使儲存量大幅減少，來做數據壓縮。第二，因奇異值由大到小且快速遞減的性質，部分的奇異值就佔了奇異值之和 90% 甚至 99% 的比例，進而達到去噪和降維的功用。

對於分解後的 A 矩陣而言，取前 r 個 Σ 的奇異值，並將其餘奇異值設為 0，使得一些不重要、影響較小的資訊去除，可以得到 A 的近似矩陣 $\tilde{A} = U_r \Sigma_r V_r^T$ 。在儲存資料上，將一個較大的矩陣以三個較小的矩陣方式儲存，甚至 Σ_r 可以用向量來儲存，而不是矩陣。

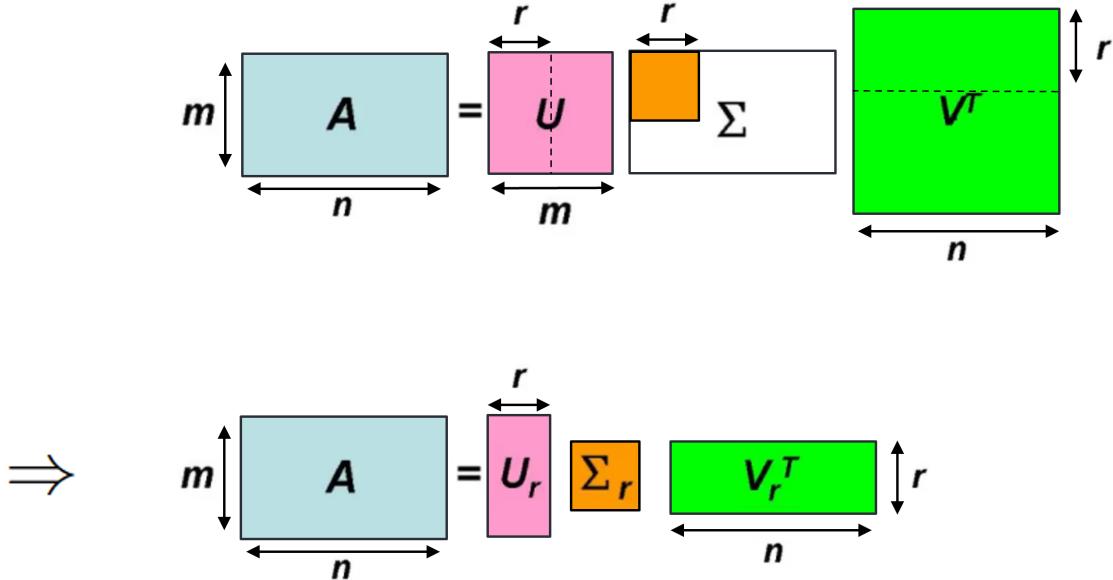


Figure 2: Feature extraction

其中，資料壓縮比 = $\frac{m \times n}{(m+n+1) \times r}$ [8]，節省空間率 = $1 - \frac{(m+n+1) \times r}{m \times n}$ 。

3.2 High-Order Singular Value Decomposition

對三維張量的資料，HOSVD 的目標在於對其進行 SVD 特徵分解。在一個彩色影像中，我們可以將其視為一個三維張量，且第三維度分別為 RGB 三個通道的三張影像組合而成。

HOSVD 的過程是結合了張量矩陣化（Tensor matricization）與 SVD 分解，將張量展開成矩陣的方式，使其能夠進行特徵萃取工程。而如何將彩色影像展開且不影響到影像原有的結構，是接下來要探討的目標。

假設一個 $m \times m$ 大小之彩色影像，其三維張量表示為 $\mathcal{X} \in \mathbb{R}^{m \times m \times 3}$ ，為了保持影像完整性，以利於影像特徵萃取，我們將其進行 1-mode unfolding，使得 \mathcal{X} 映射到 $X_{(1)} \in \mathbb{R}^{m \times 3m}$

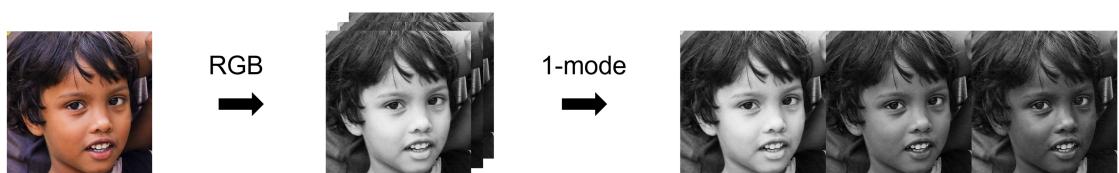


Figure 3: Mapping the tensor to matrix

可以發現在影像矩陣化的結果，也就是將 RGB 三個通道的影像排列在一起，獲得一個較寬且較扁的矩陣。就影像而言，這樣的矩陣化良好的保存了影像的結構，使得之後特徵萃取的過程有效的保留影像重要資訊。

接著對得到的 $X_{(1)} \in \mathbb{R}^{m \times 3m}$ 做 SVD 分解萃取特徵，使得 $X_{(1)} = U\Sigma V^T$ ，得到其中的特徵值 $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m$ 。然後將 Σ 取前 r 個特徵值使得

$$\frac{\sum_{i=1}^r \sigma_i}{\sum_{i=1}^m \sigma_i} = 99\%$$

並令 $\sigma_{r+1} = \sigma_{r+2} = \dots = \sigma_m = 0$ ，得到 Σ_r ，則 $\tilde{X}_{(1)} = U\Sigma_r V^T$ ，且 $\tilde{X}_{(1)} \approx X_{(1)}$ 。最後再將得到的 $\tilde{X}_{(1)}$ 映射回三維張量 $\tilde{\mathcal{X}}$ ，使得 $\tilde{\mathcal{X}} \approx \mathcal{X}$ 。

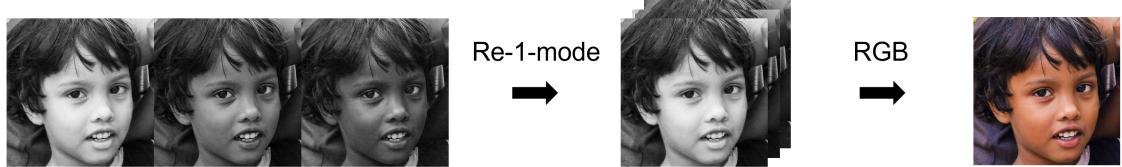


Figure 4: Mapping the matrix to tensor

對 Figure 3 和 Figure 4 的 128×128 彩色影像為例使用 HOSVD，進行 1-mode unfolding 後得到 128×384 的灰階影像，並對其使用 SVD 分解。在 128 個奇異值中，僅僅只取前 50 奇異值就包含了 99% 奇異值之和的比例。此時資料壓縮比為 1.9162 及節省空間率為 0.4781，並且將其重建回彩色影像後，仍然保留大部分影像資訊，說明 HOSVD 可以在維持彩色影像品質之下，對其進行了良好的特徵萃取。

Algorithm 1: HOSVD

Input: A tensor of image $\mathcal{X} \in \mathbb{R}^{m \times m \times 3}$.

Output: A tensor of image $\tilde{\mathcal{X}} \approx \mathcal{X}$.

Step 1 (Matricization):

Calculate 1-mode unfolding on the tensor \mathcal{X} to get the matrix $X_{(1)} \in \mathbb{R}^{m \times 3m}$.

Step 2 (Singular Value Decomposition):

Calculate $X_{(1)} = U\Sigma V^T$ of the singular value decomposition.

Extract important singular values σ_i for $i = 1, \dots, m$ from $\Sigma \in \mathbb{R}^{m \times 3m}$.

for $r = 1 : m$

 Calculate total sum of the singular values $y_m = \sum_{i=1}^m \sigma_i$.

 Calculate first r sums of the singular values $y_r = \sum_{i=1}^r \sigma_i$.

if $y_r/y_m \geq 99\%$

break

end

end

Set $\sigma_{r+1} = \dots = \sigma_m = 0$ to get Σ_r .

Calculate $X_{(1)} = U\Sigma_r V^T$.

Step 3 (Reconstruct image):

Calculate Re-1-mode unfolding on the matrix $X_{(1)}$.

Then get the tensor of image $\tilde{\mathcal{X}} \approx \mathcal{X}$.

4 Super-Resolution via Deep Learning

近年來深度學習的神經網路架構越來越強大、越來越廣泛，包含物件分類、圖像視覺、語言處理等。其中卷積神經網路（Convolutional Neural Network, CNN）是廣泛用於影像上的神經網路，它的每一層 CNN 都可以覆蓋一部分小範圍內的神經元，使其可以學習區塊範圍的特徵，對於大型圖像處理有出色表現。

以往使用內插法（Interpolation）直接求得更高解析度的成像，往往得到的效果差強人意。因此在內插法後進行 CNN 神經網路，學習更深層更複雜的影像特徵所得到的成像，在視覺上和數值結果上表現得更好。

本文將依序介紹 Super-Resolution CNN (SRCNN) 、Very Deep Super-Resolution (VDSR) 及 ResNet，三種神經網路架構在超解析成像（Super-Resolution）的架構及運作原理。實驗以 Department of Computer Science, Yonsei University 提供的開源人臉影像庫，訓練模型提高人臉影像的解析度，並以 PSNR 值（數值越大代表與目標影像越相近）為指標。

4.1 Super-Resolution CNN (SRCNN)

SRCNN (Super-resolution convolution neural network) 是一個神經網路，輸入為低解析度（視覺上）的圖像，輸出為高解析度的圖像。在將圖像餵進神經網路前，需要先經過一個預處理 bicubic interpolation（雙三次插值算法經常用於圖像或者影片的縮放），將原始圖片變成跟想要的高解析度圖像一樣大小後，再餵進神經網路中。而神經網路做的事情，主要分成三個步驟，區塊特徵抽取與表達（Patch extraction and representation）、非線性對應（non-linear mapping）和重建（reconstruction）。

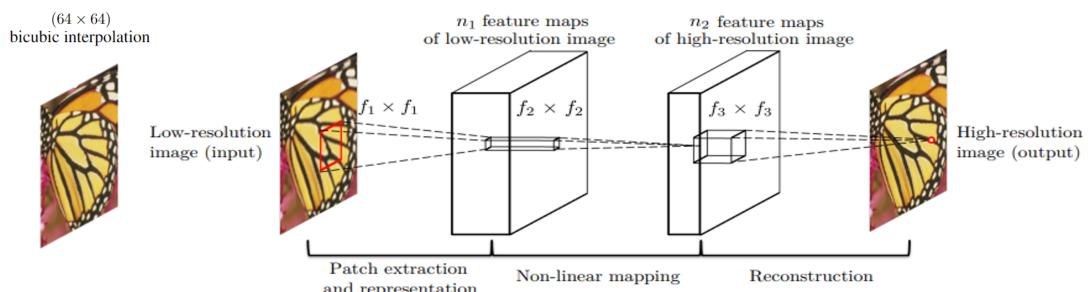


Figure 5: SRCNN

區塊特徵抽取與表達（Patch extraction and representation）是通過 CNN 的 filter 提取影像的區塊特徵，儲存成多維向量（維數等於 filter 數量），且所有的特徵向量組合成特徵矩陣（feature maps），使用 64 個 9×9 大小的 filter 形成 64 張 feature maps。非線性對應（non-linear mapping）是將 n_1 維的特徵矩陣進一步進行 CNN 提取更複雜的特徵，形成另一個 n_2 維的特徵矩陣，使用 32 個 5×5 大小的 filter 形成 32 張 feature maps。重建（reconstruction）等於是一個反卷積（Deconvolution）[9]，將 n_2 維的特徵矩陣還原回 RGB 的彩色影像，使用 3 個 5×5 大小的 filter 還原回 RGB 三張影像。

在 SRCNN 中所採用的損失函數（Loss Function）[10]是基於平均方根差（Mean Square Error），定義為重建後的相片每一個像素與真正的圖片的每一個像素的差異，

$$L(W) = \frac{1}{n} \sum_{i=1}^n \|F(X_i, W) - Y_i\|^2$$

並使用隨機梯度下降（Stochastic gradient descent, SGD）[11]和標準反向傳播（backward propagation）[12]來更新參數和最小化損失函數。

$$\Delta_{i+1} = \Delta_i - \eta \cdot \frac{\partial L}{\partial W_i}, \quad W_{i+1} = W_i + \Delta_{i+1}$$

在人臉超解析成像的實驗中，SRCNN 對比只有 Bicubic interpolation 後的影像效果來的好。PSNR 越高表示與原始影像越接近，影像復原的越好。

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{255}{\text{MSE}} \right)$$

Average	Bicubic	SRCNN
PSNR	32.8726 db	33.3617 db

Table 1: SRCNN in testing data

4.2 Very Deep Super-Resolution (VDSR)

VDSR (Very Deep Super-resolution) 是一個相對於 SRCNN 還要深層的神經網路，其中發現到網路的加深可以使得輸出影像的精準度顯著的提高，得到更高解析度、品質更好的影像。VDSR 主要解決 SRCNN 的兩個問題，第一，網路層數過

淺導致感受視野過小，實驗中使用 10 和 18 層 CNN 做為神經網路架構；第二，在加深網路的同時解決收斂速率變慢的情況發生。

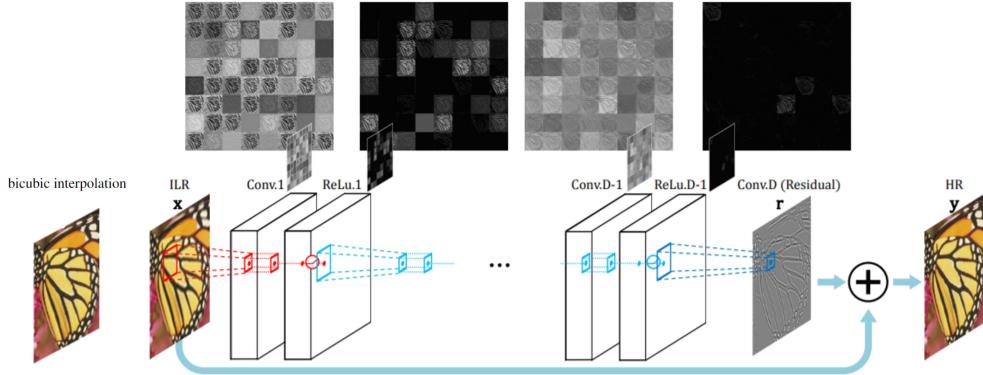
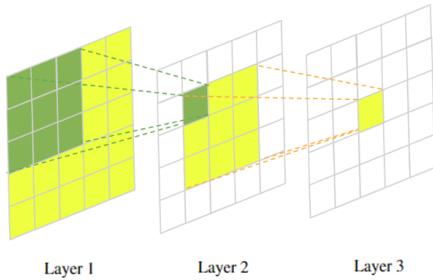


Figure 6: VDSR

第一，CNN 感受視野（receptive field）[13]表達了兩個 feature maps 間神經元的對應關係，以了解神經元在前一個 feature maps 上的運算中包含了多少區域。如果 CNN 神經網路沒有足夠大的感受視野，則影像上一些較大的特徵可能未被檢測到。



$$\text{Let } s_i = 1 \text{ and } k_i = 3 \text{ for all } i$$

$$r_3 = 1, \text{ receptive field} = 1$$

$$r_2 = 1 \times (1 - 1) + 3, \text{ receptive field} = 3$$

$$r_1 = 1 \times (3 - 1) + 3, \text{ receptive field} = 5$$

Figure 7: Receptive field

感受視野的定義為 $r_i = s_i \cdot (r_{i+1} - 1) + k_i$ 。其中 r_i 為第 i 層的感受區域， s_i 為第 i 層的步長（stride）， k_i 為第 i 層的 kernel size。

第二，為了收斂緩慢之改善而採用了殘差學習（Residual-Learning），主要學習輸出與輸入的殘差（即高頻的特徵）取代直接學習全部。因為高解析度的影像中其實就包含著低解析度的資訊，因此直接學習兩者間的殘差再將其與低解析度影像相加，便能得到我們的輸出，不僅運算變快，也能讓反向傳播更新參數的收

效能加速。

$$\text{Loss function : } \frac{1}{2} \|\mathbf{y} - f(\mathbf{x})\| \implies \frac{1}{2} \|\mathbf{r} - f(\mathbf{x})\|, \quad \mathbf{r} = \mathbf{y} - f(\mathbf{x})$$

實驗中 VDSR 是每一層 CNN layer 都使用 64 個 3×3 大小的 filter 並加入 Relu 激勵函式所組成。VDSR10 包含了 10 層 CNN layer 且可以發現層數增加對於成像效果明顯變好。

Average	Bicubic	SRCCN	VDSR10
PSNR	32.8726 db	33.3617 db	35.2216 db

Table 2: VDSR10 in testing data

4.3 ResNet (Deep residual network)

ResNet 是一個非常深度且使用非常多 Residual-Learning 的神經網路模型，其主要目的是要處理神經網路的退化 (degradation) 問題，退化是當模型的層數越多越深時，誤差值卻反而提高，造成了複雜的模型效果卻不好。而退化問題的產生歸咎於參數更新，當模型越深，越前面層數的參數通過反向傳播和梯度下降的梯度值呈指數遞減，導致了參數更新緩慢，甚至無法更新參數，使得模型效果不佳。

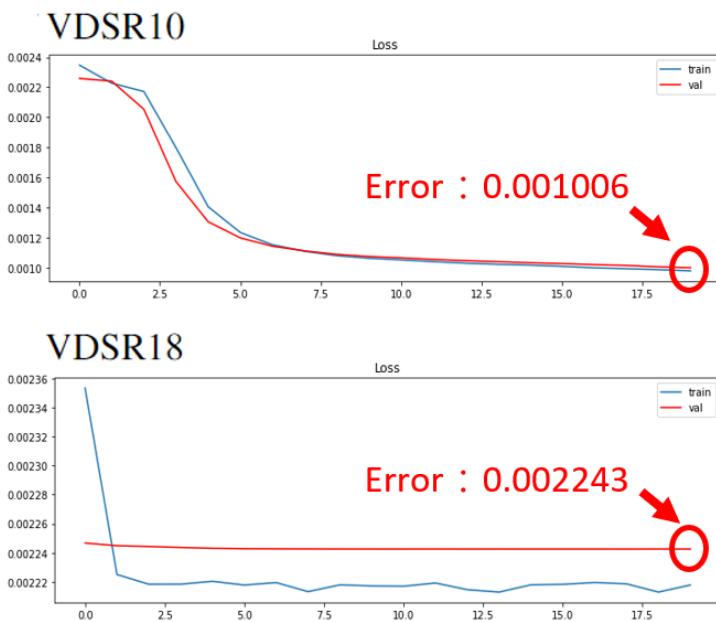


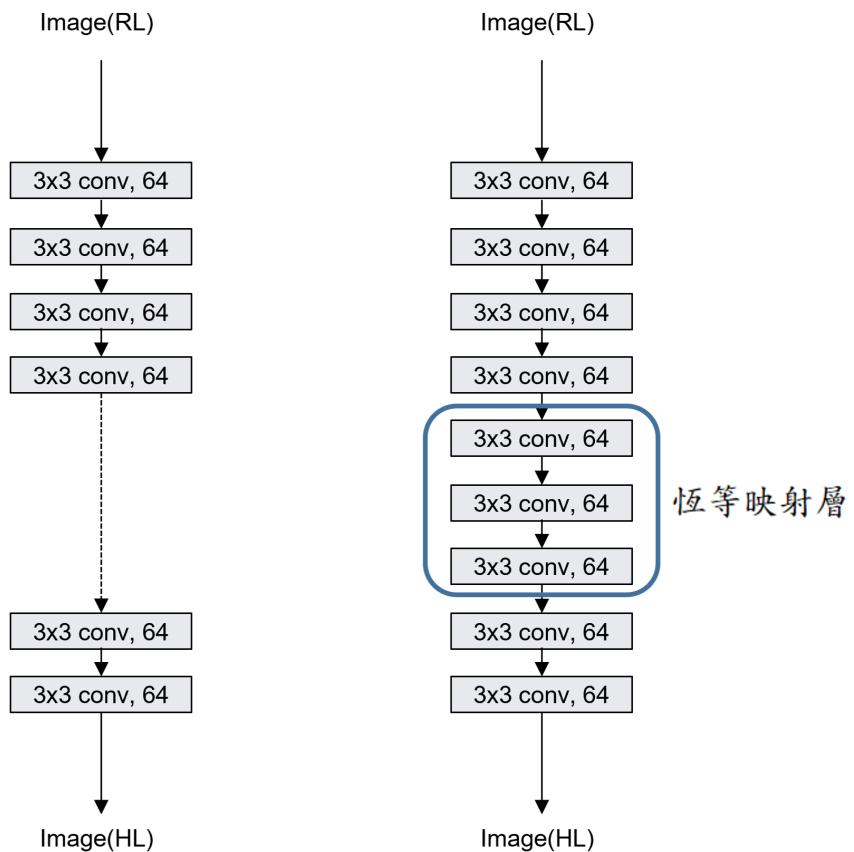
Figure 8: Training error of VDSR10 and VDSR18

Average	Bicubic	SRCNN	VDSR10	VDSR18
PSNR	32.8726 db	33.3617 db	35.2216 db	34.6484 db

Table 3: VDSR10 and VDSR18 in testing data

在訓練中 VDSR 神經網路 18 層相較於 10 層的誤差要來的大，甚至出現誤差沒有下降的趨勢。而成像的結果上 VDSR18 來的比 VDSR10 的 PSNR 值低，這是因為退化問題所導致。

退化問題說明了並非一味的加深網路就會有好的效果。假使目前有一個較淺的網路結構，將這個淺層網路增加恆等映射層（identity mapping layers）形成一個較深的網路，使得經過恆等映射層的輸入與輸出是一樣的，直覺上來說兩個網路的結果應該一樣。但事實上，深層網路經過訓練後，無法將參數訓練成輸入輸出恆等。



假設將一個網路結構視爲一個函數 $\mathcal{H}(x)$ ，想要訓練後 $\mathcal{H}(x) \rightarrow x$ 這樣的逼近似乎很困難，因此將網路結構從 $\mathcal{H}(x)$ 轉變成 $\mathcal{F}(x) = \mathcal{H}(x) - x$ ，結果這樣的逼近變得相對容易，在深層的網路中也不會有退化的狀況發生，這樣的型態 $\mathcal{F}(x) = \mathcal{H}(x) - x$ 稱之爲 residual mapping。經過訓練後發現這樣的 residual mapping 變得更容易優化，若是想要衡等映射的話，可以將 $\mathcal{F}(x)$ 趨近到 0，也就達到衡等輸出了。

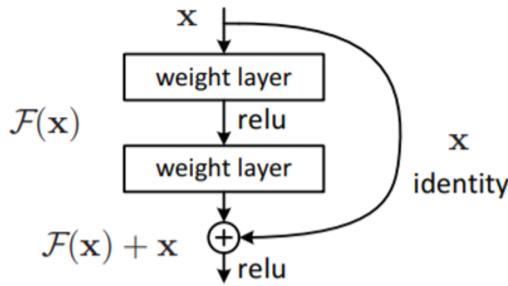


Figure 10: Residual learning unit

Residual learning 有主要以下優點。第一，如果圖中兩層網路對結果無益，那麼直接跳過即可（即參數設爲 0）；第二，參數更加容易學習， $\mathcal{F}(x)$ 相比於原本直接訓練來的更小，使得參數大多接近 0，甚至通常爲稀疏矩陣，而參數初始化一般都在 0 附近，更容易訓練；第三，不會出現梯度消失問題。

梯度消失問題（Vanishing gradient problem）[14]是一種深度學習中的難題，出現在想要更新參數並計算梯度值的時候。每當經過一輪的訓練，神經網路參數的梯度更新值與損失函數的偏導數成比例，因爲參數是隨著層數不斷累乘，使得越前面的參數經過連鎖法則（chain rule）[15]，其梯度值會消失，使得參數無法更新，甚至整個網路都可能沒辦法繼續訓練。在深度學習中，如果其梯度值在 $(0, 1)$ 範圍內，反向傳播以鏈式法則來計算梯度，則越前面的參數不斷累乘起來就會趨近於 0。

$$\text{Relu} : R(x) = \max(x, 0), \text{Input} : x_1, \text{Output} : y$$

$$y = R(w_3x_3), x_3 = R(w_2x_2), x_2 = R(w_1x_1)$$

$$\frac{\partial y}{\partial w_1} = w_3 \cdot R'(w_3x_3) \times w_2 \cdot R'(w_2x_2) \times R'(w_1x_1)$$

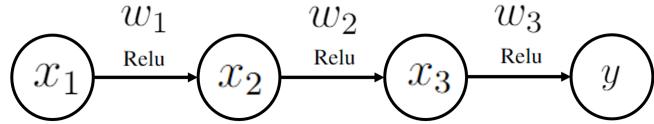


Figure 11: A simple neural network

Figure 11 中，當我們反向傳播計算 w_1 的梯度值時，會將後面 w_2 和 w_3 的參數直累乘，若參數介於 $(0, 1)$ 的範圍內且神經網路非常深，將會導致 w_1 的梯度值接近 0，從而發生梯度消失問題。反之，參數若大於 1， w_1 的梯度值接近無窮大，則會發生梯度爆炸。

$$y = w_1x_1 + R(w_3x_3)$$

$$\frac{\partial y}{\partial w_1} = x_1 + \frac{\partial R(w_3x_3)}{\partial w_1}$$

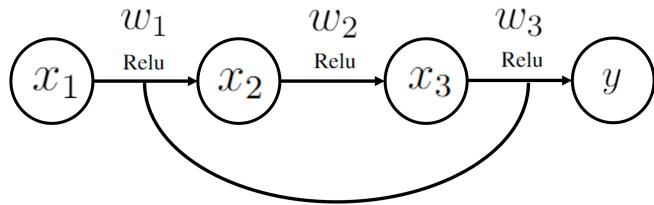


Figure 12: Residual learning for a neural network

當加入 Residual learning 後， w_1 的梯度值多了 x_1 也就是輸入值，使得就算後一項的偏微分趨近於 0 也還是有 x_1 ，除非一開始的輸入就為 0 了。此神經網路架構很好的避免了梯度消失問題，就算網路架構多深都能夠訓練得起來。

在設計網路架構上，以每兩層 CNN 就使用 Residual learning unit 來建構網路。以 VDSR18 來做對比，設計 ResNet18 網路架構。

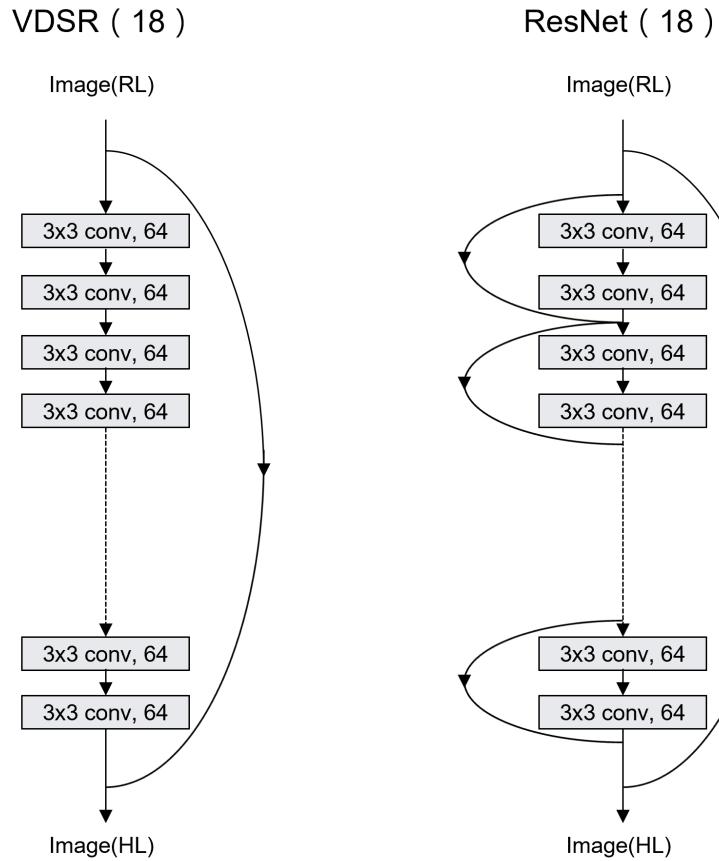


Figure 13: VDSR18 and ResNet18

在人臉影像之超解析成像的實驗中，訓練模型得到的訓練誤差上，發現誤差得以持續下降，表示加了 Residual learning unit 的網路在訓練上解決了網路越深越難以訓練的退化問題。另一方面在 PSNR 值的表現上 ResNet18 的結果是全部模型中最好且最高的。

Average	Bicubic	SRCNN	VDSR10	VDSR18	ResNet18
PSNR	32.8726 db	33.3617 db	35.2216 db	34.6484 db	35.5566 db

Table 4: ResNet18 in testing data

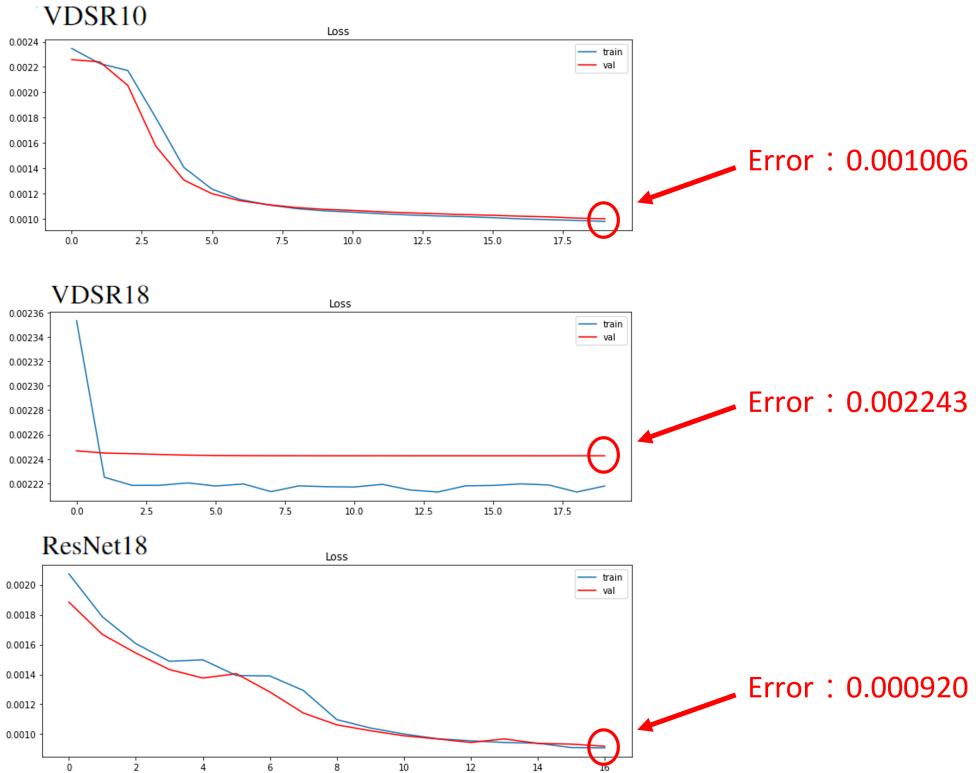


Figure 14: Training error

5 Experiments

在實驗上，結合 HOSVD 與三個深度學習 CNN 模型來實現超解析成像，將影像解析度提高兩倍為目標，並觀察各個模型中前處理有無使用 HOSVD 的差異。其中包含資料壓縮比、節省空間率和超解析成像結果的好壞，並以主觀影像（視覺上）和客觀 PSNR 值（數值越大代表與目標影像越相近），兩個指標來評估。

公式如下，

$$\text{資料壓縮比} = \frac{m \times n}{(m + n + 1) \times r}$$

$$\text{節省空間率} = 1 - \frac{(m + n + 1) \times r}{m \times n}$$

$$\text{均方誤差 (mean-square error, MSE)} = \frac{1}{n} \sum_{i=1}^n \|F(X_i, W) - Y_i\|^2$$

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{255}{\text{MSE}} \right)$$

實驗流程如下，將 Data 先經過 bicubic interpolation 初始放大兩倍使其與目標影像相同大小，直接放進 SRCNN、VDSR10、VDSR18 和 ResNet18 四種模型中分別作訓練，訓練沒有經過 HOSVD 的模型。另一方面，經過 bicubic interpolation 後做 HOSVD 萃取特徵後，再放進模型訓練經過 HOSVD 的 Data。最終分別做比較兩者間的差異，以及模型間性能的好壞。

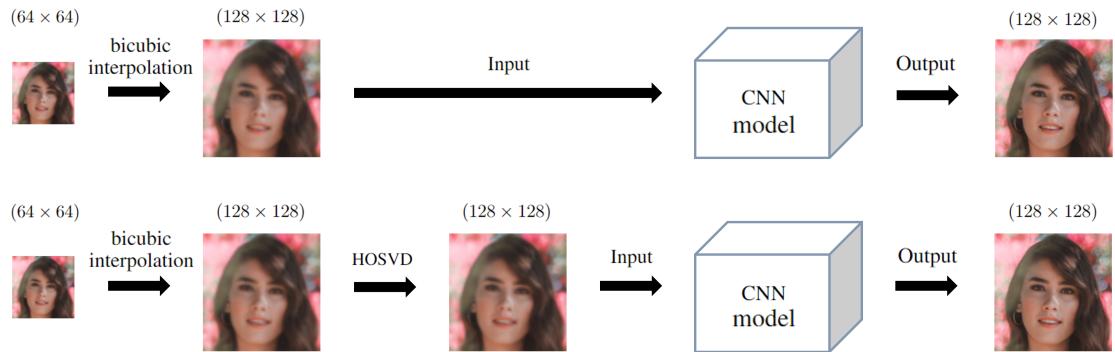


Figure 15: 流程圖

實驗數據由 Department of Computer Science, Yonsei University 提供的 Real and Fake Face Detection 影像庫[16]，包含 1081 張人臉影像。將 Data 分為 900 張 training data 、 100 張 validation data 及 81 張 testing data ，並將影像初始化為 128×128 大小做為目標影像，相同影像縮小為 64×64 做為訓練影像。

訓練影像經過 bicubic interpolation 成 128×128 低解析度影像後使用 HOSVD 特徵分解，取前 50 個奇異值，包含奇異值之和的 99% 比例。經過 HOSVD 特徵萃取後的影像壓縮比為 1.9162 及節省空間率為 0.4781，接著放入 CNN 模型做訓練。



Table 5: A simple image for low-resolution and high-resolution

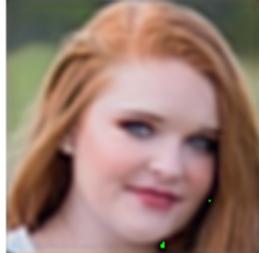
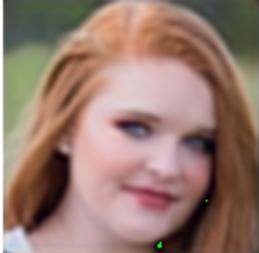
	PSNR (without HOSVD)	PSNR (HOSVD)
Bicubic		
	32.2250 db	32.1290 db
SRCNN		
	33.5956 db	33.6039 db
VDSR10		
	35.8561 db	36.0619 db
VDSR18		
	34.1401 db	34.4861 db
ResNet18		
	36.1401 db	36.4861 db

Table 6: A simple image in testing data

Average	Bicubic	SRCNN	VDSR10	VDSR18	ResNet18
PSNR (without HOSVD)	32.8726 db	33.3617 db	35.2216 db	34.6484 db	35.5566 db
PSNR (HOSVD)	32.8197 db	33.8426 db	36.3986 db	35.2902 db	36.5249 db

Table 7: Average in testing data

從 Table 6 可以看到每一個在有經過 HOSVD 特徵萃取的影像，通過訓練好的 CNN 模型所輸出的影像，在直觀（視覺）上相對於沒有 HOSVD 來的清楚、解析度來的更高，尤其 VDSR18 可以發現成像明顯更加的清晰不會模糊。另一方面，從 Table 7 的 testing data 平均值來看，雖然還未放入 CNN 模型前 Bicubic 插值法的影像相比於經過 HOSVD 的 PSNR 值要好上一點，因為 HOSVD 的特徵萃取本身會捨去小部分的資訊。但在 CNN 模型的訓練上，有使用 HOSVD 相比於沒有使用都是有顯著的提升的，甚至 VDSR10 和 ResNet18 上更是提升將近 1 db 值。

6 Conclusion

本文提出了結合 HOSVD 和 CNN 模型來提取影像數據的特徵，並實現超解析成像生成高畫質影像。在 CNN 模型中，隨著模型深度的加深，其成像結果越好。而在加入 HOSVD 處理輸入影像後，整體的成像效果有著明顯的提升，其關鍵在於 HOSVD 特徵萃取時，除去了不必要的雜訊，避免了高頻雜訊對於成像的干擾，而且 HOSVD 分解後的矩陣減少了多達 47% 儲存空間。實驗結果顯示，使用 HOSVD 生成的影像在主觀視覺上更加清楚，且在測試集上 PSNR 的平均值皆更好。

References

- [1] D. Letexier, S. Bourennane, and J. Blanc-Talon. “Nonorthogonal tensor matricization for hyperspectral image filtering.” *IEEE Geoscience and Remote Sensing Letters* 5.1 (2008): 3-7.
- [2] A. Rajwade, A. Rangarajan, and A. Banerjee. “Image denoising using the higher order singular value decomposition.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.4 (2012): 849-862.
- [3] S. Albawi, T.A. Mohammed, and S. Al-Zawi. “Understanding of a convolutional neural network.” *2017 international conference on engineering and technology (ICET)*. Ieee, 2017.
- [4] C. Dong, C. C. Loy, K. He, and X. Tang. “Image super-resolution using deep convolutional networks.” *IEEE transactions on pattern analysis and machine intelligence* 38.2 (2015): 295-307.
- [5] J. Kim, J.K. Lee, and K.M. Lee. “Accurate image super-resolution using very deep convolutional networks.” *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- [6] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition.” *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- [7] A. Hore, and D. Ziou. “Image quality metrics: PSNR vs. SSIM.” *2010 20th international conference on pattern recognition*. IEEE, 2010.
- [8] Z.N. Li, M.S. Drew, and J. Liu. *Fundamentals of multimedia*. Upper Saddle River (NJ):: Pearson Prentice Hall, 2004.
- [9] P.A. Jansson, ed. *Deconvolution of images and spectra*. Courier Corporation, 2014.

- [10] G. Seif, and D. Androutsos. “Edge-based loss function for single image super-resolution.” *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.
- [11] L. Bottou. “Stochastic gradient descent tricks.” *Neural networks: Tricks of the trade*. Springer, Berlin, Heidelberg, 2012. 421-436.
- [12] M.A. Nielsen. *Neural networks and deep learning*. Vol. 25. San Francisco, CA, USA: Determination press, 2015.
- [13] M.D. Zeiler, and R. Fergus. “Visualizing and understanding convolutional networks.” *European conference on computer vision*. Springer, Cham, 2014.
- [14] S. Hochreiter. “The vanishing gradient problem during learning recurrent neural nets and problem solutions.” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6.02 (1998): 107-116.
- [15] L. Ambrosio, and G.D. Maso. “A general chain rule for distributional derivatives.” *Proceedings of the American Mathematical Society* 108.3 (1990): 691-702.
- [16] Department of Computer Science, Yonsei University, Real and Fake Face Detection, available from <https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection>.