

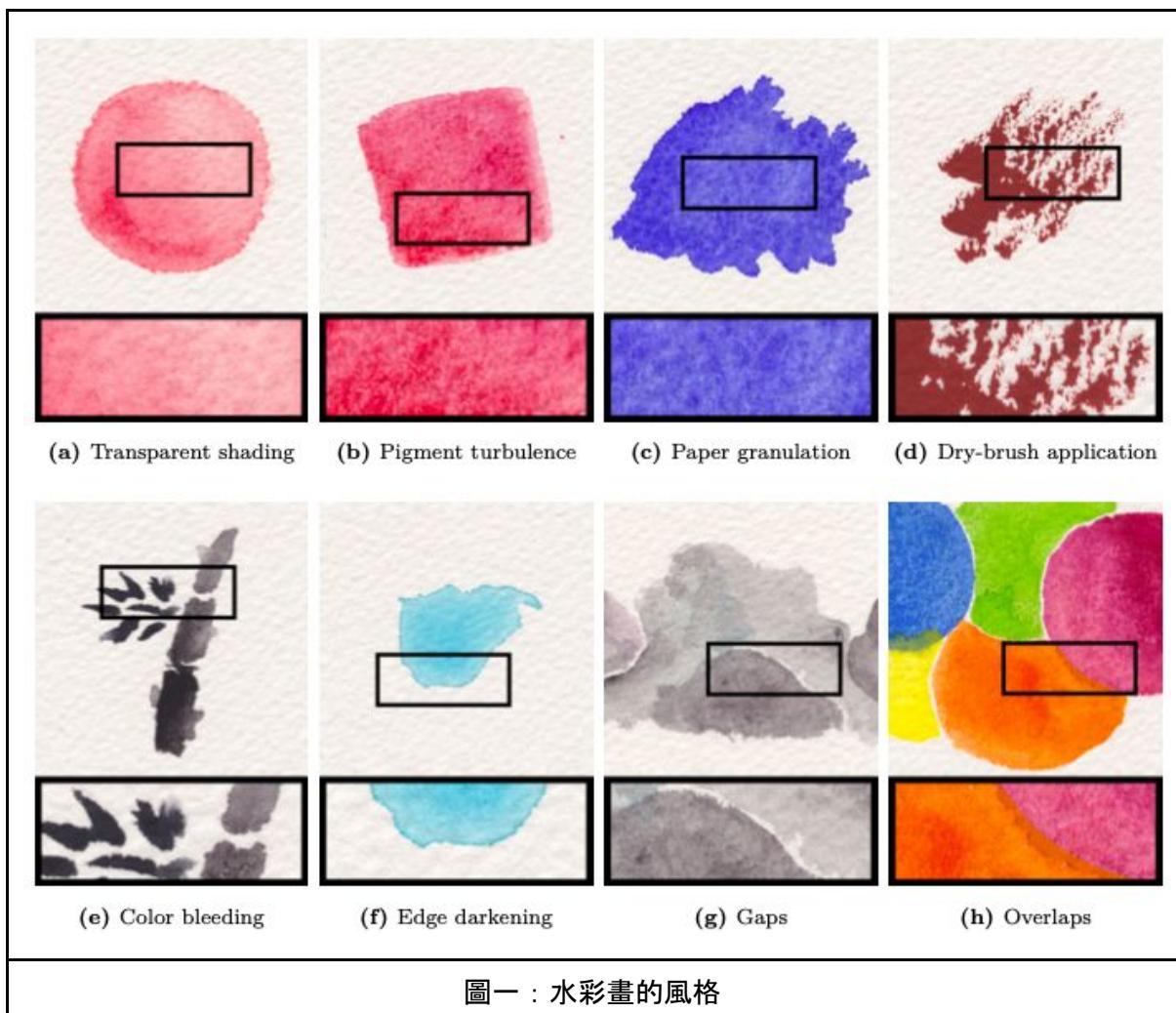
DIP 2020 Final Report

Style Transfer (Group 12)

黃振修 王煥智 蕭皖文 莫易喆

1. Artistic Effect Observations

在專題進行的前期，我們嘗試了許多的實驗，發現若使用任一風格影像對輸入影像進行風格轉換，結果常差別人意。我們閱讀許多文獻，發現 Montesdeoca [4] 對水彩的特性做了非常仔細的觀察。以水彩風格為例（如圖一），作者認為一張畫要有水彩的感覺至少需要考慮透明度、顏料湍流、紙的顆粒度、乾筆刷感、顏色渲染、邊緣深化、間隙留白、色料疊加等八項元素。因此，對影像進行風格轉換，有非常多的細節需要考量。



對此，我們詳細觀察不同 style image 做出來的成果。以油畫為例，我們挑選了多位畫家不同風格的畫作，例如梵谷（Van Gogh）的鳶尾花（Irises）、俄羅斯畫家阿爾森米斯庫爾班諾夫（Arsen Kurbanov）的奇蹟發生（Miracles Happen）……等作品。最後我們發現梵谷（Van Gogh）的鳶尾花（Irises）有著明顯的筆觸，還有相近於湖面的藍色調，因此做為 style image

應用在湖畔風景照做出來的風格轉換就有不錯的油畫風格表現（如圖二）。但若使用 Arsen Kurbanov 的奇蹟發生（Miracles Happen）進行風格轉換（如圖三），由於筆觸不明顯，整體色調偏白，因此效果不如鳶尾花來得好。



圖二：以鳶尾花對 lake.jpg 進行風格轉換

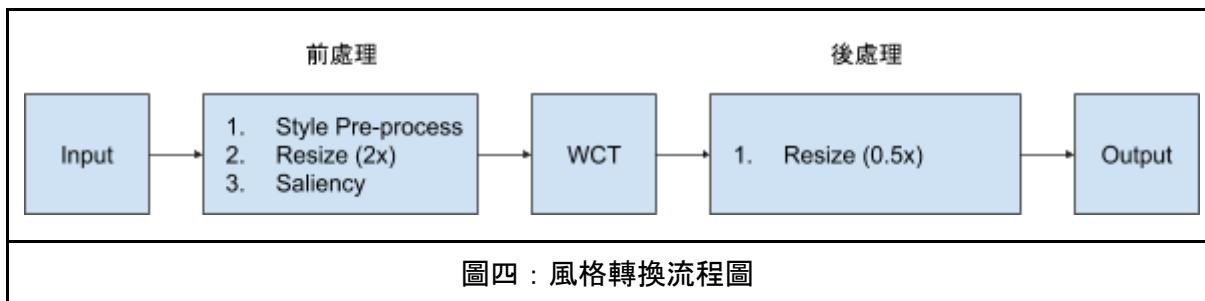


圖三：以奇蹟發生對 lake.jpg 進行風格轉換

由上述實驗，為了獲得良好的結果，我們獲得了兩個重要的結論：我們需挑選適合的 style image 做為輸入，並且對影像進行預處理（Pre-processing）與後處理（Post-processing）。在嚴博士的分享裡也提到類似的觀點。我們產生最終成果的方法將基於這兩個結論，並在下面做說明。

2. Method

我們的方法流程圖如下方圖四。在前處理的環節，我們會針對不同的風格進行不同的預處理、對影像放大以調整筆觸大小、使用 VGG 找尋影像裡的重要的部份（Saliency）。之後，使用 WCT 方法進行風格轉換，並在後處理的環節，將原先在前處理放大的影像縮小回來，成為最後輸出的影像。以下將針對各個處理方法做說明。

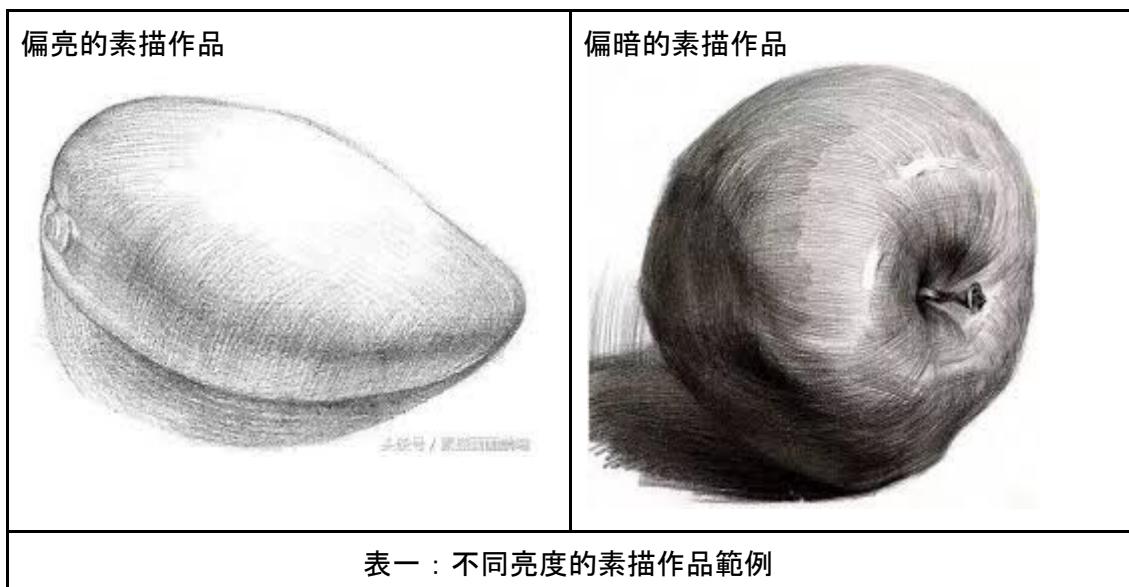


a. Style Pre-process

根據以上我們對於各種藝術畫風的分析，我們對於這四種風格分別做了以下處理：

素描：基於我們設定採用黑白鉛筆素描，而非色鉛筆素描，因此我們對於輸入的 content image 會先把他轉換成灰階，避免原始照片上過多的色彩影響到素描 style image 的最終呈現（有時候會有很不搭配的狀況）。

如表一，由於素描作品整體色調可能過淡白或過濃黑，這會讓原本的影像轉移後偏亮或是偏暗而失去原始的亮度對比：

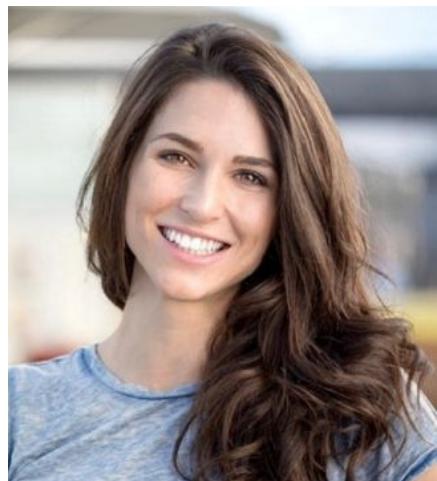


因此我們會把素描的 style image 針對轉成灰階後的 content image 做 histogram matching，讓轉換後的色調忠於原始內容，如表二所示。



表二：經過 histogram matching 的影像

水彩：水彩由於是使用水來調整顏料的一種繪畫方式，經過加水後的顏色通長期飽和度都會降低，而亮度因為加了水而有部份提高，因此針對水彩的風格轉換，我們會把原始影像轉到 HSV 顏色域，然後把 saturation 調低，把 value 調高，先使其色彩上就有水彩的感覺，然後在讓 style image 上面水彩的筆觸與渲染，水流融等感覺呈現出來，如表三所示。



表三：模擬水彩色調調整前後的影像比較

油畫：油畫則是與水彩相反，由於其本身是使用油料來融化有機顏料，其顏色本身就比較飽和不透明，因此我們會把原始影像轉到 HSV 顏色域，然後把 saturation 調高，把 value 調低，使其看起來有油畫的感覺。同時因為顏料本身是用畫刀括上去的，我們經過實驗，也發現可以套用 bilateral filter，可以在保留 edge 的情況下，讓顏色更平滑而去除一般拍照會成現出的顆粒感，如表四所示。



表四：模擬油畫色彩調整前後的影像比較

對於水彩及油畫來說，由於所選擇的 style image 其本身的色調可能相對於遠本的影像過於強烈或是平淡，因此仿照素描，我們也會針對 style image 做對 content image 的 histogram matching，讓轉換後的結果更忠於原始影像，如表五所示。



表五：先進行 histogram matching 再進行油畫風格轉換的影像結果

水墨：由於水墨作品也有色彩的呈現，被非一定是黑白，所以我們就沒有做任何灰階的處理。加上水墨的作品比較寫意，會對畫面做比較大幅度整體的創作，因此我們也不針對 content image 或是 style image 做任何的 histogram matching，而是讓所選擇的 style image 本身的創作意涵套用到原本的影像。

b. Resize & Resolution Matching

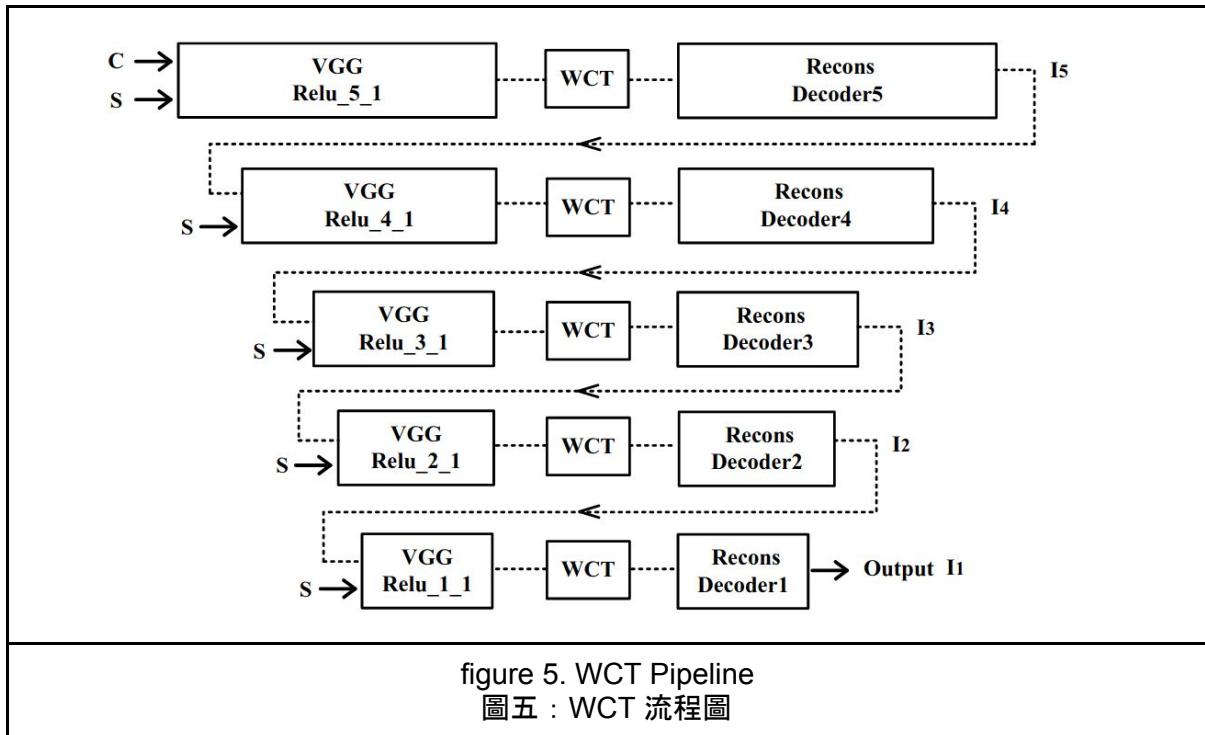
由於我們採用的 WCT 方法，在使用 autoencoder 轉換到 latent code 套用 style 再重建回來的過程中需要 content 與 style image 檔案大小一樣，因此我們通常會把 style image 用 bicubic resize 方式調整大小到跟 content 一樣。然後我們觀察到，由於我們所選擇的 style image 通常具有一定的解析度（因為希望轉換效果能呈現的好），反而會讓兩張測試影像 woman & lake 其本身因為 size 就不大（450 & 280），這會讓 style image 上的筆觸與線條被過度 downsize，導致效果不顯著，因此我們會探斷輸入影像其最長邊是否小於 500 像素，若是的

話，我們會直接把 content image 放大兩倍，然後轉換風格，轉換之後再 downsize 1/2 回原本的影像尺寸，此種作法的效果會比較好。

同樣的，針對特別大的原始影像，例如 street 這張影像，也應該挑解析度特別高的 style image，或是先 downsize 原始影像然後再放大，但因為該照片上有些人的臉本來就比較小，downsize 之後在轉換風格時容易被扭曲，所以最終我們沒有如此做，而是針對此一觀察提出將來針對高解析度照片的處理方式做出指引。

c. WCT

In this project, we adopt and improve Whitening & Coloring Transform (WCT) {Li et al.} [5] -- a universal style transfer method. WCT is a simple and effective method that can generalize to unseen styles. This method's key idea is to treat style transfer as a reconstruction problem while using feature transforms (whitening and coloring) at the intermediate level to transform the content features w.r.t to the style feature statistics. WCT style transfer pipeline is shown in the figure 5.



Multi-level style transfer. The figure 5 is directly adopted from {Li et al.}.

P2S1 In WCT {Li et al.} [5] VGG-19 is employed as the encoder, and for each of the VGG ReLU layers used (Relu_X_1) a corresponding feature decoder is trained (Recons DecoderX). After the initial training, all encoding and decoding layers are fixed. Then, to perform style transfer, a learning-free scheme consisting of whitening and coloring steps is used.

Whitening and coloring transform. Given content image I_c and style image I_s , VGG decoder extracts their feature vectors $f_c \in \mathbb{R}^{C \times H_c \times W_c}$ and $f_s \in \mathbb{R}^{C \times H_s \times W_s}$.

Whitening step. First, f_c is centered by subtracting its mean vector m_c . Then f_c is transformed by the whitening transform to \hat{f}_c so that the feature maps are uncorrelated ($f_c f_c^T = I$):

$$\hat{f}_c = E_c D_c^{-\frac{1}{2}} E_c^T f_c$$

where D_c is a diagonal matrix with the eigenvalues of the matrix $f_c f_c^T$, and E_c is the corresponding matrix of eigenvectors (i.e., $f_c f_c^T = E_c D_c E_c^T$)

Coloring step. Similarly, in the coloring step f_s is centered by subtracting its mean vector m_s and transformed by the coloring transform, which is the inverse of the whitening step:

$$\hat{f}_{cs} = E_s D_s^{\frac{1}{2}} E_s^T \hat{f}_c$$

where D_s is a diagonal matrix with the eigenvalues of the matrix $f_s f_s^T$, and E_s is the corresponding matrix of eigenvectors.

In this step, \hat{f}_c is colored w.r.t. the statistics of the style vector, that is the resulting representation \hat{f}_{cs} has the same correlation between its feature maps as f_s , $\hat{f}_{cs} \hat{f}_{cs}^T = f_s f_s^T$. In other words, its Gram matrix or covariance matrix that encodes the information about the style matches the Gram matrix of the style features.

Finally, the feature vector \hat{f}_{cs} is recentered by m_s .

The WCT method we use is further improved by recent works {Lu et al.} [6] and {Wyenn et al.} [7]. {Lu et al.} [6] proposed a new improved feature transform that leads to better contour preservation and {Wyenn et al.} [7] introduced two new parameters to control the trade-off between the amount of the style transferred and the content preservation at each level of the multi-level style transfer.

At each level l starting from $l = L$ (coarse to fine stylization), given a pair of encoders and decoders at level l , (e_l, d_l) , stylized image \hat{I}_l is obtained by the equation below:

$$\hat{I}_l = d_l(\gamma(\delta C_l^s(e_l(\hat{I}_{l+1}))) + (1 - \delta)C_l^s(e_l(I^c))) + (1 - \gamma)e_l(I^c)$$

Where $\gamma \in (0, 1)$ controls the degree of stylization and $\delta \in (0, 1)$ controls the degree of detail preservation.

In our project, we further improve this parametrization method by introducing saliency map to obtain new adapted parameters γ_s and δ_s .

d. Saliency

Not all regions are equally important for human perception -- the viewer's eye is drawn to more interesting areas, such as eyes, mouth, etc. That's where the method should be able to preserve more details. On the other hand, Motivated by this, we propose to use saliency map -- a map localizing "the most interesting" pixels in the image -- to adapt γ_s , δ_s according to this principle. We do this by modifying the original formulation by {Wyenn

et al.} [7]. Given the precomputed saliency map $I_{saliency}$ and parameters δ , γ selected by the user, we define the output at each level l as:

$$\hat{I}_l = d_l(\gamma_s(\delta_s C_l^s(e_l(\hat{I}_{l+1}))) + (1 - \delta_s)C_l^s(e_l(I^c))) + (1 - \gamma_s)e_l(I^c)$$

where

$$\delta_s = \delta - w I_{saliency}, \delta \in (0, 1) \text{ and}$$

$$\gamma_s = \gamma - \frac{w}{2} I_{saliency}, \gamma \in (0, 1).$$

We set w to $w = 0.2$.

In our project, we experiment with both a traditional and deep-learning approach for detecting salient pixels. As for the traditional approach, we employ the Spectral Residual method by {Hou and Zhang} [8], and for the deep-learning approach, we re-use VGG19 used as the encoder in WCT. The Spectral Residual method detects salient pixels by analyzing log-spectra representations of the image. This approach does not involve any prior knowledge of objects in the image or their classes. On the other hand, a deep learning model will typically involve such information if it's trained in a supervised way in tasks such as image classification. Intuitively, Convolutional Neural Network (CNN) learns to extract features that are important for recognition and thus can be used to approximate human cognition to some extent. We use VGG19 (up to the 5th layer) to extract feature maps and transform feature maps to saliency maps by preserving the maximal feature activation channel-wise. More sophisticated methods such as Grad-CAM {Selvaraju et al.} [9] can be used in the future.



Figure 6. Comparison between the traditional saliency detection approach (left) and VGG19 extracted “saliency” map (right). Note that the face is less affected by undesired style transfer artifacts using the deep-learning approach.

最終我們整理針對輸入的各種藝術風格，我們所採用的各種前處理與後處理於下表六：

表六：風格轉換方法列表				
	Oil	Watercolor	Pencil	Ink
Grayscale			V	
Increase Saturation	V			

Reduce Saturation		V		
Bilateral filtering	V			
Style Image histogram matching	V	V	V	
Upscale and Downscale	O	O	O	O
Weighted by saliency map	V	V	V	V

V: 採用的前後處理

O: 當輸入影像最長邊小於 500 像素時採用

3.Photo Results

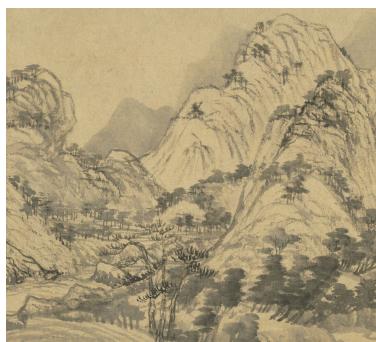
這個章節將一一列出所有風格轉換的影像。



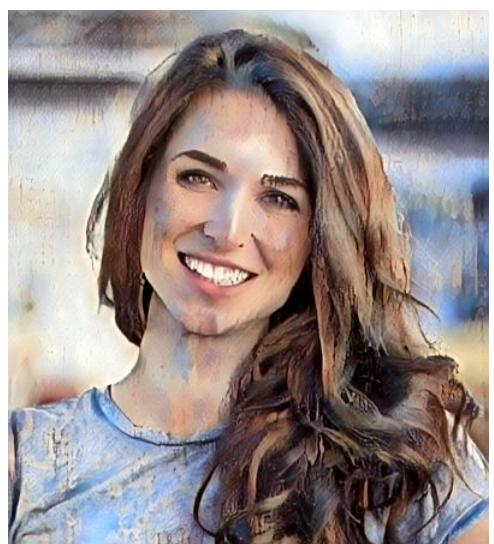
素描



水墨



水彩



Lake.jpg

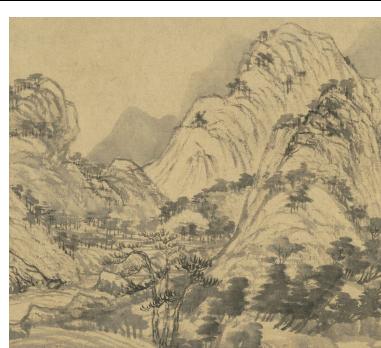
油畫



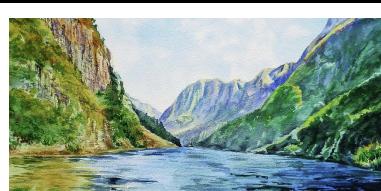
素描



水墨



水彩



Street.jpg

油畫		
水墨		

4. Video Results

在處理影片的時候，我們注意到影片畢竟和圖片有很大的不同，首先由於圖片是靜態的，人們在看的時候會四處轉移焦點觀察影像，因此 style 的強度很重要，又必須作到不扭曲重要實際內容。但是影片因為是動態的，人們會被畫面的 motion 所吸引，而風格的感受則取決於連續畫面累積所感受到的視覺殘留，因此處理 video 畫面時，我們不能選用跟靜態影像一樣的參數 $\text{gamma} = 0.9$, $\text{delta} = 0.95$ 。而使選用了 $\text{gamma} = 0.05$, **$\text{delta} = 0.6$** ，特別是 delta 減少了 CNN 累積每一層的 style 效果，讓因為 style 而導致的畫面扭曲減少，避免影片閃爍。此外，為了更加平順影片觀賞效果，我們套用了一個類高斯 low pass filter 的 $[1, 2, 1]$ 三張 frame 的加權平均，可以在穩定畫面與避免過度模糊之間取得平衡。以下是我們做出來的影片效果：

以下將列出風格轉換後的影片的其中一個影格，以展示套用風格轉換後的樣貌。整體風格轉換影片請見連結。

油畫

https://drive.google.com/drive/folders/1mn3T5I_EKU8NVwSkUcZOMUQov3rgf9VI?fbclid=IwAR2uyijLOAEUwXvwwVlZmM2AtikjYApXWgNULqTs8YGr6_o2hakn

wUk9WM



素描

https://drive.google.com/drive/folders/1mn3T5l_EKU8NVwSkUcZOMUQov3rgf9VI?fbclid=IwAR2uyijLOAEUwXwwwVIZmM2AtikjYApWgNULqTs8YGr6_o2hknwUk9WM



水彩

https://drive.google.com/drive/folders/1mn3T5l_EKU8NVwSkUcZOMUQov3rgf9VI?fbclid=IwAR2uyijLOAEUwXwwwVIZmM2AtikjYApWgNULqTs8YGr6_o2hknwUk9WM



水墨

https://drive.google.com/drive/folders/1mn3T5I_EKU8NVwSkUcZOMUQov3rgf9VI?fbclid=IwAR2uyijLOAEUwXwwwVIZmM2AtikjIYApwGnULqTs8YGr6_o2haknwUk9WM



5. Some other experiments

在專題進行當中，我們也嘗試了一些方法。這些方法雖不盡我們滿意，但仍在這裡做整理。

a. 抹除部份非重要區域而得到主體或『意象』的感覺

由於水墨或是素描被非像水彩或是油畫一樣是滿版的作品，水墨著重意象，只會畫出作者想要呈現最中心的主體，其他部份留白。因此我們可以根據偵測出來的 saliency map，把周圍不重要的區域白化抹除，如表七所示：

			
沒有強調主體	根據 spectral saliency 強調房子與近處的草堆	根據 VGG16 saliency 淡化部份非主體	
		把周圍不重要的區域白化抹除的原始影像	
偵測到的 spectral saliency map	表七：saliency 比較		

但由於計算 saliency 本身是個非常困難的問題，而且不容易廣泛的針對所有影像找對真正人們認為的主體，因此我們只回報這個實驗的結果，在最終的專案實做中沒有真的去白化抹除原始的內容，而是只用來調整 WCT 這個方法的全局 gamma & delta 值，使其根據 saliency map 微調成每個 pixel 都不同的局部值，使其跟能針對內容本身的重要性調整 style 的強度。

b. Avatar-net

我們除了使用 WCT 方法，我們另外有嘗試使用 Avatar-net [1]，來與 WCT 方法來比較。我們發現到 Avatar-net 所產生的圖片上有明顯的小點筆觸，覆蓋在整張影像上。如表八的兩張影像所示，可以看出影像上有點具有類似於訊號受到 aliasing 影響的結果。與 WCT 相比，這樣的結果顯得不那麼乾淨，因此我們最後不採用 Avatar-net 方法。

此外，使用 Avatar-net 產生影片時，由於是每張影格每張影格去轉換，因此會有影片閃爍的問題。雖然使用藉由前後影格平均可以減輕閃爍的問題，但依然不如 WCT 方法平順，因此，最後我們決定不採用 Avatar-net 方法來做為我們最後結果的呈現。



表八：Avatar-net 產生的結果，左圖為油畫效果，右圖為水彩效果。

6. Conclusions

在這次的專題裡，我們發現到對於不同的風格類型，在轉換前需要先進行相對應不同的前處理。另外，雖然眾多萬用的風格轉換模型，我們建議要尋找與輸入影像較相似的風格影像來做轉移，效果才會比較好。再者，若輸入影像與風格影像的解析度差距過大，建議可以先將輸入影像放大，使得風格影像的筆觸在輸出影像上不會顯得過大。並且，著重輸入影像裡的部份區域進行風格轉換可以讓結果更好，但如何劃分重要的區域出來是困難的。在進行影片的風格影像轉移時，以每個影格各自進行風格轉移常會有影像閃爍的問題。對此，我們認為可以建立影格與影格之間的關聯性，讓相同的物體有相同的風格轉換結果，以減少影片閃爍的問題。

7. References

1. Avatar-net (<https://arxiv.org/pdf/1805.03857.pdf>)
2. Joseph Zbukvic's painting: <https://www.josephzbukvic.com/paintings/>
3. Eugene von Guerard's painting: <https://www.wikiart.org/en/eugene-von-guerard>
4. Art-directed Watercolor Rendered Animation (
https://www.researchgate.net/publication/303486747_Art-directed_Watercolor_Rendered_Animation)
5. {Li et al.} Universal Style Transfer via Feature Transforms
(<https://arxiv.org/pdf/1705.08086.pdf>)
6. {Lu et al.} A Closed-form Solution to Universal Style Transfer
(<https://arxiv.org/pdf/1906.00668.pdf>)
7. {Wyenn et al.} Unsupervised Learning of Artistic Styles with Archetypal Style Analysis (<https://arxiv.org/pdf/1805.11155.pdf>)
8. {Hou and Zhang} Saliency Detection: A Spectral Residual Approach
(<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4270292>)

9.{Selvaraju et al.} Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (<https://arxiv.org/pdf/1610.02391.pdf>)