Image-Guided Image Style Transfer with Diffusion model

張銀軒、黃永毅、賴柏勛 指導老師:丁培毅

➢ 研究背景與目的

擴散模型在圖像風格轉換中的表現一直都很出色，但最大的問題是模型本身速度不快，且擴散模型的隨機性也影響了產出的內容。大部分現有的方法需要對擴散模型進行微調，或者用額外的神經網絡。而我們使用了一種不需要額外訓練，直接使用額外的 loss function 來試圖將預訓練擴散模型的輸出導向到想要的方向。通過這種方法來提高使用者建構的速度，而不用花太多時間來微調擴散模型，並與其他現有不同的圖像風格轉換方法來做比較。

➢ 研究方法

我們使用 OpenAI 開發的 guided diffusion，其優點為在 sampling 時對 Diffusion Model 的輸出進行條件約束，而無需為每個具體情境重新訓練網絡，而 loss 可分為保留圖片內容的 content loss 以及確保風格轉換正確性的 style loss

總損失 (Total Loss) :

$$L_{total} = L_{content} + L_{style}$$

content loss 可分為三部分

1. 計算原始圖片與生成圖片的 MSE

$$L_{content1} = \frac{1}{N*M}\sum_{i=1}^{N}\sum_{j=1}^{M}(I_{original}[i,j]-I_{generated}[i,j])^2$$

2. 計算原始圖片與生成圖片在 VGG feature map 中的 MSE

$$L_{content2} = \frac{1}{K*L}\sum_{k=1}^{K}\sum_{l=1}^{L}(F(I_{original})[k,l]-F(I_{generated})[k,l])^2$$

3.根據 ZeCon 在文字導向風格轉換裡的主要貢獻

style loss 結合了文字導向風格轉換使用的 CLIP score，改為將兩張圖片放進 CLIP Space 進行比較，以及風格轉換常用的計算兩張圖片的 Gram matrix 再計算相似度的方法，α和β代表兩種做法的權重

$$L_{style} = \alpha \cdot C_{score} + \beta \cdot G_{score}$$

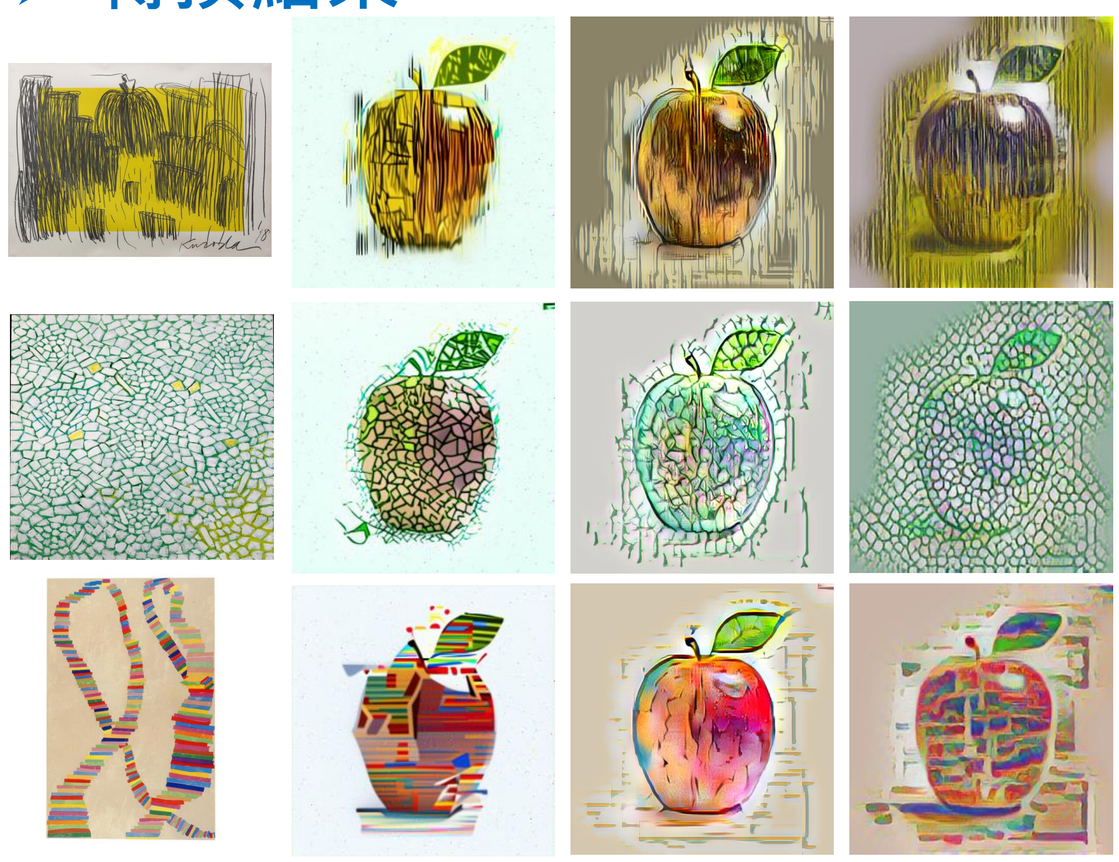
1. 使用 CLIP Score 比較兩張圖片在 CLIP 空間中的相似度

$$C_{score} = C(I_{original}, I_{generated})$$

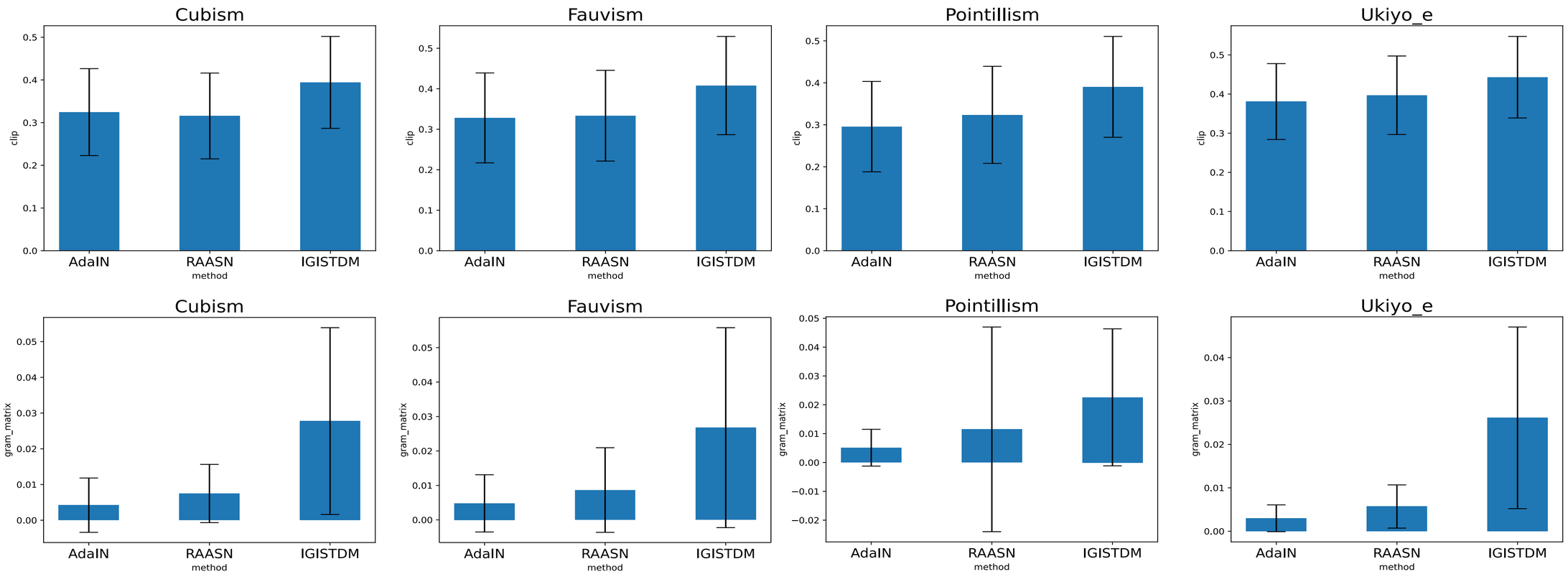
2. 使用 Gram Matrix 計算兩張圖片的風格相似度

$$G_{score} = G(I_{original}) * G(I_{generated})$$

➢ 轉換結果



➢ 分析



➢ 結論

在研究進度中，遇使用 CLIP 作為 Style loss 時的挑戰，由於 CLIP 將整個風格導向圖片 encode 進 CLIP space，導致生成結果可能包含風格圖片的內容或輪廓。此外，CLIP 對顏色訊息的較少理解也是一個問題，使生成結果的顏色相對保守。同時也觀察到 Gram matrix 在這方面有一定的優勢，能夠補足 CLIP 的不足，提供更豐富的顏色訊息。

雖然目前的結果顯示此方法在數值上較其他做法差，但這也為未來的優化提供了方向。我們認為對於超參數設計還有改進的空間，在 Content loss 和 Style loss 之間取得更好的平衡，將會對生成的結果有碩大的影響。此外，也能嘗試微調生成模型，以尋找此作法潛在於特定領域的應用。

總體而言，雖然目前的結果不如預期，但這次研究為相關領域提供了一個有價值的參考範例，並且未來也能透過優化相關參數以及 loss 架構的探討來更近此作法。