國立中興大學電機工程學系

碩士學位論文

以區塊為基礎之轉換器於影像恢復之研究

The Study of Patch-Based Transformer for Image Restoration

指導教授：廖俊睿 Jan-Ray Liao

研 究 生：吳承勳 Chen-Xun Wu

中華民國 一百一十四年 七月

國立中興大學電機工程學系

碩士學位論文

題目(中文)：以區塊為基礎之轉換器於影像恢復之研究

題目(英文)：The Study of Patch-Based Transformer for Image Restoration

姓名：吳承勳　學號：7112064013

經　口　試　通　過　特　此　證　明

論文指導教授　[signature]

論文考試委員　[signature]

[signature]

[signature]

[signature]

[signature]

中　華　民　國　114　年　7　月　10　日

摘要

基於轉換器（Transformer）的架構在影像修復任務中已展現出強大潛力，然而其核心的自注意力（Self-Attention）機制因具有與影像解析度成二次方增長的計算複雜度，使其在高解析度應用中受到限制。現有的解決方案，如基於視窗（Window-based）的注意力犧牲了全局整合能力，而通道級（Channel-wise）注意力則間接建模空間關係，專注於特徵層面的統計一致性。這在高效能與全局空間建模能力之間留下了待解決的權衡問題。

為解決此問題，本研究提出一種新的區塊自注意力（Patch-Wise Self-Attention, PW）機制。此方法不以像素為單位，而是將影像特徵圖劃分為若干「區塊」(Patch)，並在這些區塊之間進行自注意力計算。此設計大幅減少了參與注意力運算的單元數量，從而在保持全局上下文建模能力的同時，有效降低了計算複雜度，使其適用於高解析度影像修復。我們將此模組整合於一個多尺度的 U-Net 架構中，以進行影像去雜訊任務的評估。

本論文在多個合成雜訊與真實雜訊的基準資料集上進行了詳盡的實驗。結果顯示，在合成雜訊任務中，所提出的區塊自注意力在所有雜訊水平下，其性能指標（PSNR/SSIM）均一致優於通道級與視窗級注意力模型，尤其在重建複雜的影像結構與紋理細節方面展現顯著優勢。對於更具挑戰性的真實雜訊任務，結合了通道級與區塊級注意力的混合模型取得了最佳表現，在 SIDD 資料集上達到 39.75 dB 的 PSNR。消融實驗進一步揭示了模型性能與計算複雜度之間的權衡，並發現稀疏的區塊採樣策略能在提升效率的同時維持優異性能。

本研究驗證了區塊自注意力是一種有效的設計範式。它透過引入強大的空間結構化統計歸納，為模型提供了直接建模區塊間關係的能力，這與通道注意力所側重的特徵間關係形成互補。本研究不僅為高效能影像修復轉換器提供了一個具競爭力的解決方案，也為未來視覺轉換器的架構設計提供了新的思路。

關鍵字： **影像修復、自注意力、轉換器、空間、特徵、複雜度**

Abstract

Transformer-based architectures have demonstrated strong performance in image restoration. However, their core component, self-attention, suffers from quadratic computational complexity with respect to image resolution, limiting scalability in high-resolution scenarios. Existing alternatives such as window-based attention reduce this complexity by restricting context to local windows, while channel-wise attention models statistical dependencies across feature channels but lacks direct spatial modeling.

To address these limitations, we propose a Patch-Wise Self-Attention (PW-SA) mechanism that performs attention across non-overlapping spatial patches instead of individual pixels. This reduces the number of tokens in the attention operation, achieving a favorable balance between computational efficiency and global context modeling. The proposed module is integrated into a multi-scale U-Net architecture and evaluated on both synthetic and real-world image denoising tasks.

Experiments on standard benchmarks demonstrate that PW-SA consistently outperforms window- and channel-based attention across various noise levels, particularly excelling in reconstructing fine textures and complex structures. On the SIDD dataset, a hybrid model combining channel-wise and patch-wise attention achieves state-of-the-art performance with a PSNR of 39.75 dB. Ablation studies reveal that sparse patch sampling can further reduce complexity with minimal performance degradation.

This study validates that patch-wise self-attention is an effective design paradigm. By introducing a strong, spatially structured inductive bias, it provides the model with the ability to directly model inter-patch relationships, which complements the inter-feature relationships emphasized by channel-wise attention. This research not only offers a competitive solution for efficient, high-performance Transformer-based image restoration but also provides new insights for the future design of vision transformers.

Keywords: Image Restoration, Self-Attention, Transformer, Spatial Modeling, Feature Modeling, Computational Complexity

目錄

圖目錄

表目錄

第一章 緒論

本章在 1.1 章介紹了影像修復任務的背景，還有技術的演變，在 1.2 章介紹本研究的動機與方法目的如何處理問題，在 1.3 章說明論文架構的安排。

1.1 研究問題

影像在取得過程中常會受到各種內在與外在因素干擾而產生劣化，最常見的就是雜訊（Noise）與模糊（Blur）。雜訊可能來自感測器熱雜訊、量化誤差、壓縮、傳輸過程中的干擾等[1]；而模糊則可能來自運動、對焦不準或光學系統缺陷。這些劣化大幅降低了影像的解析度與辨識度，不利於後續分析例如辨識、分割的準確度，因此影像修復（Image Restoration）是一項關鍵的基礎任務。

傳統方法多基於影像的先驗統計特性，例如非局部自相似性（Non-local Self-Similarity, NSS），代表性方法如 Non-local Means [2]、BM3D [3]。這些方法假設影像中存在重複結構，藉由比對與加權平均相似區塊進行降噪，但對於複雜紋理或真實世界雜訊表現有限。

深度學習方法如 DnCNN[4]、FFDNet[5]透過大量數據學習從噪聲影像到乾淨影像的映射，大幅提升了降噪品質。然而基於卷積神經網路（CNN）的方法受限於固定感受野與局部特徵建模能力，難以捕捉長距離依賴（Long-range Dependencies）[6]。

為了提升長距離空間建模能力，轉換器架構（Transformer）被引入圖像處理任務。其核心機制—自注意力（Self-Attention）能動態建模任意全局位置間的關聯，提供比捲積神經網路更靈活的全域上下文建模能力。然而，標準自注意力在影像中需考慮所有像素間關係，導致計算複雜度會與圖片的高寬成平方成長 $O(H^2W^2)$，其中 H,W 為空間尺寸，對高解析度影像不具實用性[7]。

1.2 研究動機與目的

轉換器的複雜度$O((HW)^2)$難以實現，現有方法多針對計算複雜度問題提出權衡，如移動視窗轉換器 SwinIR[8], [9]採用窗口內局部自注意力，降低複雜度為 $O((M)^2)$，M 為窗口大小內總像素數但犧牲了全局性。修復器 Restormer[10]採用通道注意力（Channel-wise Attention），將注意力計算從空間轉至通道維度，複雜度為通道的平方 $O(C^2)$，C為通道數，與解析度無關效率極高，但其建模空間關聯是間接的。這些策略分別代表「空間局部」與「通道全域」的兩個極端。我們關注的核心問題是：是否存在一種中間型態，能同時兼具空間建模能力與計算效率？

本研究探討過去轉換器的計算複雜度解決方案，提出一種更具彈性的注意力設計以「影像區塊」（Patch）作為建模單位元。我們認為，區塊是一個自然的影像結構，在雜訊像素獨立的前提下比單一像素更穩定，並且本身結構更明確，且數量遠少於整張圖像的像素。

本研究的目的如下，提出區塊級注意力模組，區塊為基本單位進行注意力計算，降低複雜度$O\left(\frac{(HW)^2}{P^2}\right)$，$P$為區塊大小的，這種設計能保留轉換器的全局建模能力，同時模型的計算複雜度適用於高解析度影像。

我們最後根據實驗結果，我們結合不同注意力來對雜訊建模。區塊級注意力在影像空間細節和紋理較好，通道注意力在影像平滑區域較好，設計的多種注意力，能兼具不同維度來建模，成為一種新穎且實用的架構。

1.3 論文架構

本論文共分為五章，各章內容安排如下：第一章為緒論，說明本研究欲解決的問題為何，以及我們打算採用什麼方法來處理。第二章為研究背景，介紹影像修復任務中所涉及的重要技術與相關文獻，包括目前常見的範式與方法。第三章說明本研究所提出的解決方法與模型細節，並解釋其設計原理、實作方式與可期望的效果。第四章展示各項實驗結果，包含與其他方法的比較分析，本研究的模型控制實驗，以及對模型效能與資源使用的探討。第五章總結本研究的成果與限制，方法上的分析總結，並提出未來可能的研究方向與應用潛力。

第二章 研究背景

本章在 2.1 介紹影像退化與復原模型如何建模，在 2.2 介紹基於機器學習如何建模在 2.3-2.6 介紹自注意力相關的知識，和來源模型，在 2.7-2.9 介紹層規範化、殘差學習、U-net 架構，這些重要且常見的方法。

2.1 影像退化與復原模型(Restoration model)

在影像處理的復原任務中，影像退化被建模為透過線性算子和加性雜訊的組合將原始（真實）影像 f(x,y)轉換為退化(degraded)影像 g(x,y)，如圖 1。

最常見的線性、平移不變模型公式如下：

$$g(x,y) = h(x,y) * f(x,y) + n(x,y),$$

$h(x,y)$是退化（例如運動模糊、散焦）的點擴展函數(Point spread function)，*表示二維卷積，$n(x,y)$是加性雜訊，通常假設為零均值且與 f(x,y)無相關：

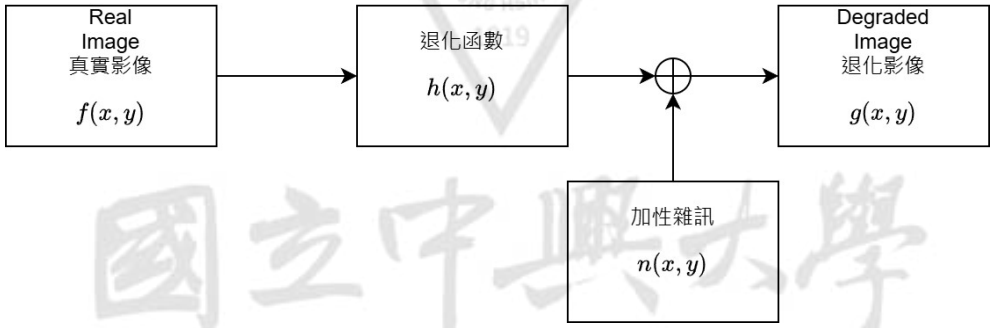


圖 1 影像的退化模型圖

恢復的目標是根據 $h(x,y)$ 和雜訊 $n(x,y)$統計資料來估計$\hat{f}(x,y)$。傳統方法包括濾波、逆濾波和基於正規化的方法。深度學習(Deep Learning)在近年來展現極佳的性能如：卷積神經網路（例如 DnCNN）[4]、生成對抗網路（例如 DeblurGAN）[11]以及基於 Transformer 的架構[9]等等模型，用於影像復原任務上。

2.2 基於機器學習的建模(Machine Learning-Based Modeling)

在影像復原領域，機器學習提供了一種資料驅動的方法，無需依賴退化模型的明確知識，即可學習從退化觀測值到清晰影像的逆映射。設 $x_{input} \in R^{m\times n\times c}$ 表示退化的輸入影像，$x_{output} \in R^{m\times n\times c}$ 表示對應的清晰(地面實況）影像。給定一個包含成對訓練樣本的資料集 $\mathcal{D} = \{(x_{input}^{(i)}, x_{output}^{(i)})\}_{i=1}^{N}$，目標是學習一個參數函數 $f_\theta : x_{input} \mapsto x_{output}$其函數能夠近似於輸入對輸出的真實映射。最小化一個合適的損失函數 $\mathcal{L}(f_\theta(x_{input}), x_{output})$，例如：均方誤差 (MSE)。這種公式使模型能夠直接從資料中學習複雜影像的先驗和復原函數，從而無需手動建立退化模型。

2.3 自注意力(Self-Attention)

深度神經網路中，尤其是卷積架構和基於注意力機制的模型，由於其強大的對空間進行建模的能力，已經成為影像恢復領域的主要建模方法。卷積架構受限於靜態權重和有限的感受視野，相比較下，注意力機制擁有全局建模和以資料動態調整權重的能力，並且根據推導注意力機制比卷積架構有更少的歸納偏差，所以注意力機制擁有更大的潛力，其中自注意力又是注意力機制在機器學習中的範式，自注意力使機器能夠自我特徵提取。

自注意力機制使模型能夠在處理高維度資料時動態地選擇和加權相關特徵。在影像處理中，注意力機制使網路能夠聚焦於顯著區域，從而提高影像修復[10]、物件偵測[12]和語義分割[13]等任務的效能

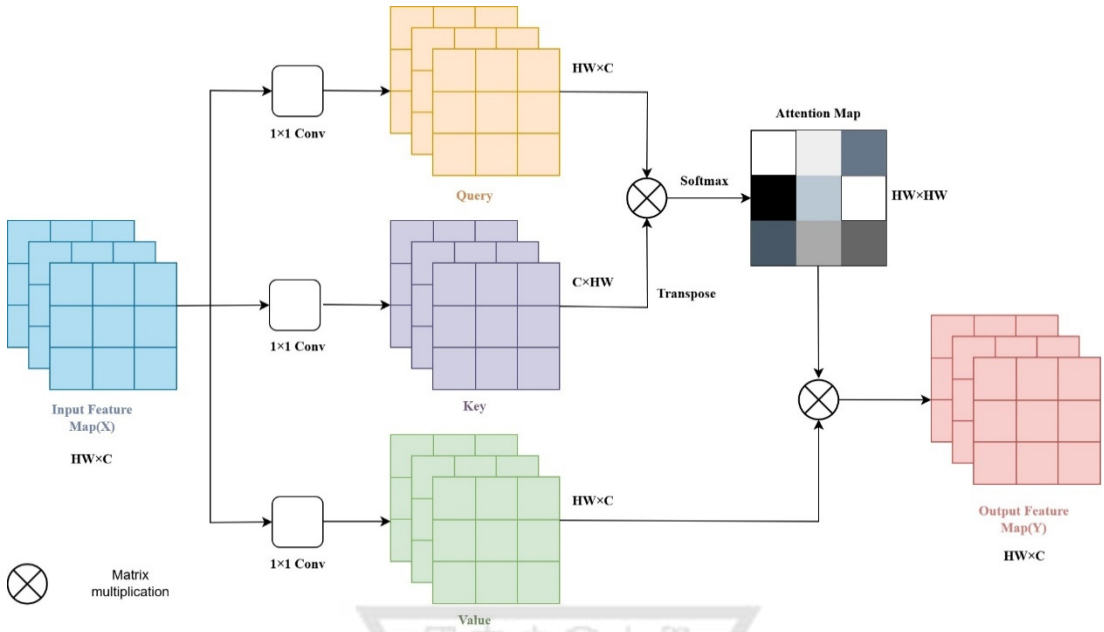


圖 2 自注意力機制在神經網路上架構圖

自注意力機制在神經網路上架構圖如圖 2，對 Q 和 K 矩陣縮放點積計算，能得到 Q 和 K 注意力圖(Q 和 K 相似程度的矩陣)，在以注意力圖為權重和 V 矩陣做點積計算，以概念來說，自注意力以像素間的相似程度(注意力圖) 作為權重，來計算每個像素的輸出值。具體而言，任意位置的輸出值，是空間中所有位置像素值(Value)的加權總和。

數學式表示上自注意力機制的查詢 Q、鍵 K 和值 V 來自於同一特徵，也就是說查詢 Q、鍵 K 和值 V 都來自同一特徵向量 X 的線性投影（linear projection），向量計算如下，其中W^Q, W^K, W^V是可訓練的權重參數：

$$X \in R^{T\times d},\quad (W^Q, W^K, W^V) \in R^{d\times d_k}$$

$$Q = XW^Q,\quad K = XW^K,\quad V = XW^V,$$

自注意力以縮放點積(Scaled Dot-Product)來計算輸出：

$$Y = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

其中 d_k 為 Q 和 K 的維度

2.4 像素級注意力之轉換器 Token-wise attention of Transformer

轉換器架構最初由 Vaswani 等人在開創性著作[14]中提出，如圖 3，它透過純粹基於自注意力機制的機制取代循環和卷積結構，標誌著序列建模的範式轉移。轉換器最初是為機器翻譯設計的，但其通用性使其廣泛應用於自然語言處理、語音以及視覺等領域。

轉換器優勢，轉換器自注意力(Self-Attention)架構允許每個符號(token)根據與所有其他符號的資訊，動態計算輸出值（Value）的加權方式。這種全域的 token-to-token 建模方式 使模型可以建立全域注意力無視位置距離，整合所有位置的資訊，尤其擅長捕捉長距離依賴關係。

轉換器的限制，在轉換器架構中自注意力的運算複雜度為$\mathcal{O}(\mathrm{H}^2\mathrm{W}^2)$。因此在高解析度影像或長序列輸入中，會導致計算與記憶體負擔迅速上升，限制應用於高解析影像。

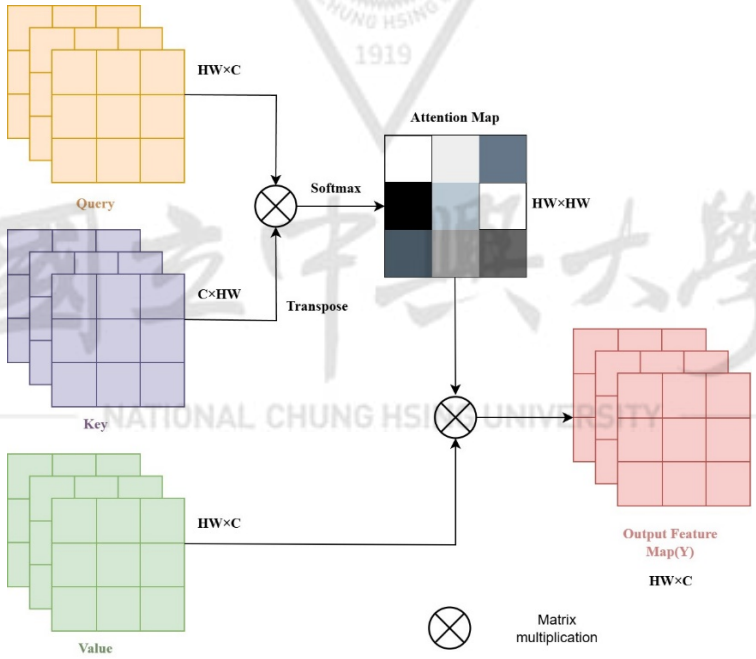


圖 3 像素間自注意力機制

2.4.1 多頭自注意力 Multi-Head Attention

多頭自注意力由[14]中提出，是一種在同一輸入內，並行使用多組注意力子空間來捕捉不同關係的機制。將輸入特徵向量分頭計算自注意力，然後將各頭的結果拼接回原維度，參考圖 4。

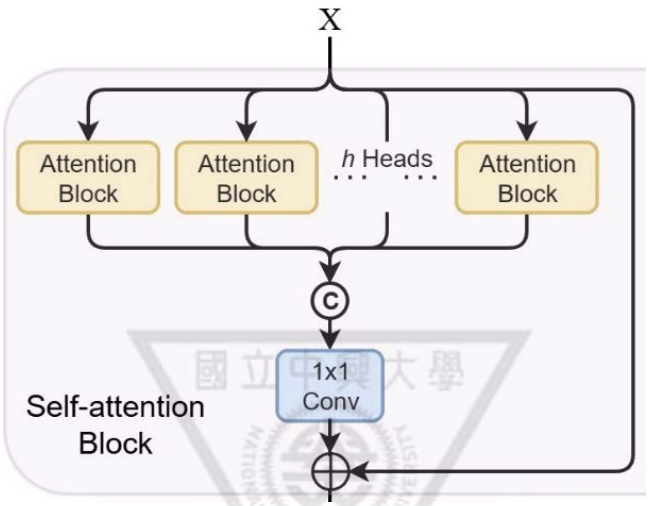


圖 4 多頭自注意力

令有 h 個注意力頭，每一頭 h 分別由其專屬的投影矩陣計算

$$W_h^Q, W_h^K, W_h^V \in R^{d_{model} \times \frac{d_{model}}{h}}$$

$$Q_h = X W_h^Q, \qquad K_h = X W_h^K, \qquad V_h = X W_h^V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O,$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V),$$

每一個注意力頭專注於不同的特徵投影子空間，使模型能同時捕捉多種關係，如局部細節、全局依賴、方向性模式等。相較於單頭注意力，多頭設計可分散學習壓力，減少單一注意力頭只關注特定特徵的風險，使整體訓練更穩定。

2.5 通道級注意力之修復器 Channel-wise attention of Restormer

修復器[10]提出一種針對影像復原任務的變體，如圖 5，將自注意力從像素間注意力轉為通道間（Channel-wise, CW）注意力，即在通道維度上計算特徵間的注意力關係。

修復器的優勢，這種自注意力架構設計使得注意力運算的複雜度從 $O((HW)^2)$ 降為 $O(C^2)$，避免了空間維度上的二次複雜度，大幅提升了高解析度場景的可行性。並且，通道注意力，對特徵向量之間進行注意力計算，有助於捕捉特徵意義代表例如邊緣、紋理的特徵關聯，維持特徵的統計一致性。

修復器的缺點，通道注意力是無法直接建模空間位置間的相關性，長距離空間關係需透過其他架構設計間接補償。

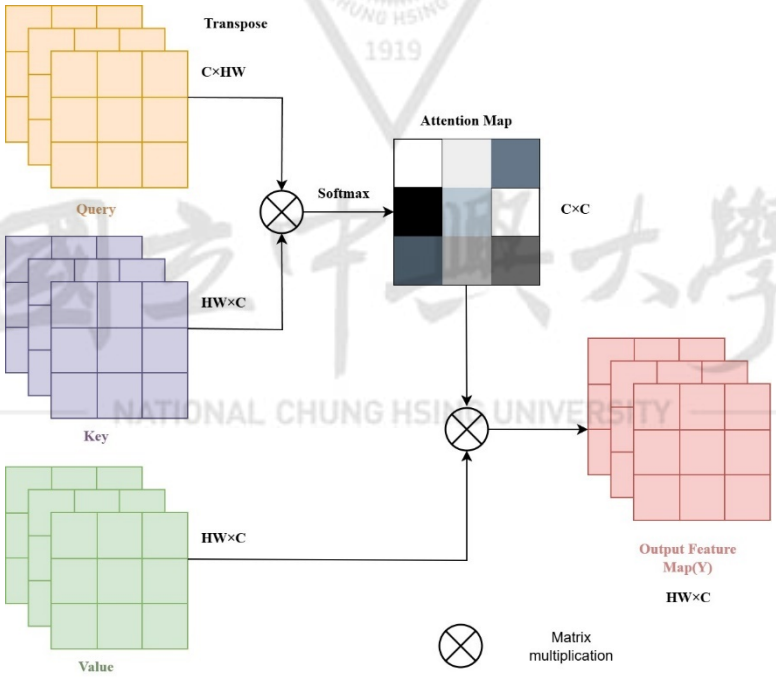


圖 5 通道間自注意力機制

2.6 基於移動視窗注意力之轉換器 Swin Transformer

移動視窗轉換器[8]則採用一種區域限制基於視窗（Window-Based, WB）的自注意力策略，如圖 6，將影像切成多個視窗大小，在固定尺寸的視窗內進行自注意力運算，計算複雜度從轉換器的$O((HW)^2)$降低至$O\left(\frac{(HW)^2}{(\mathrm{Wn})^2}\right)=O((M)^2)$，Wn 為總窗口數量，M 為窗口大小內總像素數，為克服視窗間語意割裂問題，引入視窗移動（ShiftedWindow）機制，彌補跨視窗資訊融合。

移動視窗轉換器的優勢，這種架構設計保留了轉換器的局部像素級建模能力，保留局部空間資訊的完整性與鄰近依賴關係，適合高解析應用同時兼顧運算效率，在影像修復[9]與辨識[8]任務中展現極高效能。

但移動視窗轉換器缺點為無法捕捉真正全域的注意力，模型架構設計較為複雜，需依賴位移與堆疊實現全域語意。

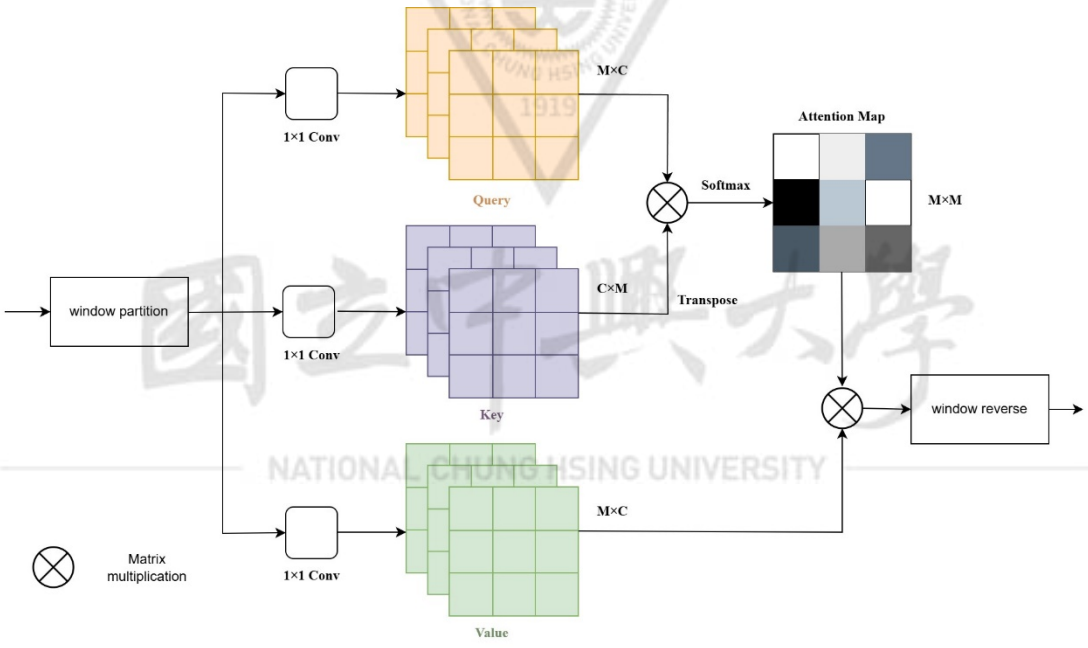


圖 6 基於視窗自注意力機制

2.7 Layer normalization

在深度學習模型的訓練過程中，內部協變偏移（Internal Covariate Shift）是造成訓練不穩定、收斂困難的主要原因之一。此現象指的是，隨著網路參數更新，每一層的輸入分布不斷變化，使後續層難以適應。為了穩定訓練流程，正規化（Normalization）技術被提出來以統一每層的輸入分布。

最初由 Ioffe 和 Szegedy 提出的 Batch Normalization（BN）[15]有效加速了卷積神經網路的訓練，但其效能高度依賴小批次的統計量，並在處理序列數據（如 RNN、Transformer）與小批次設定時表現不佳。為克服此限制，Ba 等人提出層規範化（Layer Normalization）[16]，直接在單一樣本內計算特徵間的均值與變異數，擺脫了對 mini-batch 統計的依賴，並成為 RNN、Transformer 等架構的標準配備。

考慮一個輸入特徵張量 $x_{b,c,h,w}$，Layer Normalization 對單一樣本的所有特徵維度進行標準化，其公式為：

$$\mu_b = \frac{1}{CHW}\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W} x_{b,c,h,w}$$

$$\sigma_b = \sqrt{\frac{1}{CHW}\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(x_{b,c,h,w} - \mu_b\right)^2 + \epsilon}$$

$$\mathrm{LayerNorm}\left(x_{b,c,h,w}\right) = \frac{x_{b,c,h,w} - \mu_b}{\sigma_b}$$

層規範化用於重新調整尺度與位移。這樣的操作在每個樣本獨立進行，因此特別適用於無法構成完整 batch 的訓練流程或是需要高模型穩定性的場景。

從訊號處理觀點來看，BN 更偏向於學習一個在群體中的上下文關係（contextual reference），而 LN 更像是針對個別輸入訊號的內部結構進行自我校準。

2.8 殘差學習(Residual Learning)

殘差學習（Residual Learning）最早由 He et al.在 ResNet 架構[17]中提出，用於解決極深神經網路中的退化問題，其核心思想是透過殘差映射（residual mapping）簡化學習目標。

經過實驗證實殘差訊號學習已成為現代深度學習架構的基本範式，尤其適用於去雜訊[4]、超解析度和去模糊等影像修復任務。殘差學習並非直接預測乾淨的目標訊號，而是專注於對退化輸入與真實值之間的差異（殘差）進行建模。如圖 7，給定退化輸入 x_{input} 和真實值輸出x_{output}，殘差定義為 $r = x_{output} - x_{input}$。然後，模型學習函數 f_θ，使得：

$$\hat{x}_{output} = x_{input} + f_\theta(x_{input}),\ f_\theta(x_{input}) \approx r$$

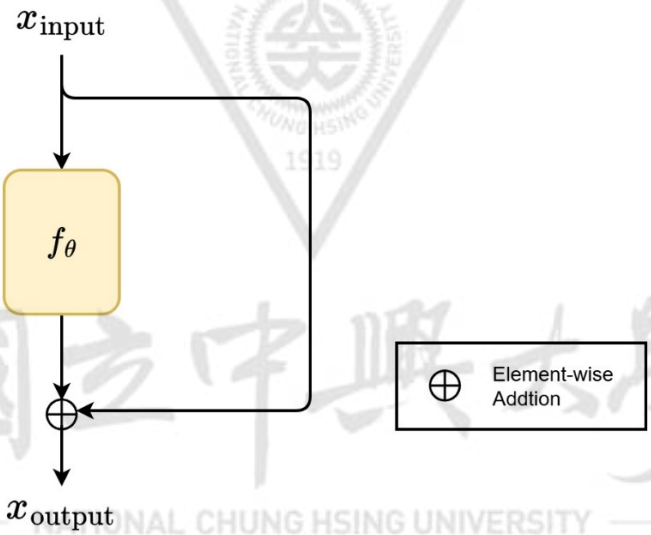


圖 7 殘差學習

這種策略有幾個優點：

1. 簡化了學習目標，殘差r通常是稀疏的（例如在復原任務中高斯雜訊幅度小），因此比直接預測完整影像更容易收斂。
2. 強化對高頻訊號的建模，直接學習殘差鼓勵模型聚焦於雜訊與邊緣細節等高頻成分，而非重複學習輸入中已存在的低頻背景。
3. 維持資訊流通性，殘差結構中的捷徑（skip connection）允許深層網路層仍可存取輸入資訊，有助於避免梯度消失，提升訓練穩定性。

2.9 U-net 架構

U-Net 由 Ronneberger 等人於影像分割首次提出[18]，採用對稱的「U型」設計，包含一個收縮編碼器路徑和一個擴張解碼器路徑，並透過跳躍連接相互連接。將 U-Net 的網路架構和注意力機制或殘差學習結合，在其他影像領域也表現出優異的結果包括影像分割、影像修復[19]和語意場景理解[20]，以 U-Net 框架為代表的編碼器-解碼器架構已成為影像領域像素級任務的範式。

在編碼器路徑中，模型逐步降低空間解析度，同時增加特徵維度。對圖片進行下取樣，這種分層編碼方式能夠捕捉輸入的低頻部分。相反地，在解碼器路徑中，網路透過一系列上取樣逐步恢復空間解析。同時減少特徵維度。

U-Net 採用跳躍連接，將編碼器的每個輸出特徵向量與相應的解碼器特徵向量整合起來。這種編碼器特徵向量的直接傳輸保留了原本可能遺失的空間解析度，從而實現精確回復。

在 U-Net 模型的任務解釋上，不同編碼層將修復問題視為對不同空間解析度進行處理。U-Net 中的跳躍連接則有助於整合學習低級紋理和高級上下文訊息，從而產生精確的像素級的重建。

2.10 結語

本章介紹後續研究的理論背景。先介紹影像退化與復原數學模型，隨後介紹現代基於機器學習的數據驅動範式。核心部分聚焦於自注意力機制，解析其運作原理，並比較了數種自注意力架構，包括原始 Transformer 的像素級注意力、Restormer 的通道級注意力以及 Swin Transformer 的視窗級注意力。最後介紹了層規範化、殘差學習和 U-Net 架構等在當代深度學習模型中不可或缺的設計。總體而言，本章回顧的內容為設計高效能的影像復原模型提供了必要的知識。

第三章 實驗方法

本章在 3.1 模型架構說明架構設計原理和原因，在 3.2 轉換器模組簡述模組化元件，並在 3.2.1、3.2.2 詳細介紹設計的區塊自注意力是如何運作的，在 3.2.3 門控深度卷積前饋模組說明設計原理和原因

3.1 模型架構（Model Architecture）

本研究架構參考修復器[10]所採用之影像恢復模型建立於一個對稱的 U-Net 式架構之上，整體設計結合了編碼-解碼 Encoder-Decoder 架構多層次化的特徵提取、上下採樣操作、模組化的自注意力模組與門控深度卷積前饋模組。此結構在保持空間資訊與跨層特徵整合的同時，也具備良好的模組替換性與擴展潛力，適合作為影像復原任務的基礎架構。

如圖 8，編碼-解碼架構與上下採樣策略的部分，整體架構由編碼器、解碼器組成。編碼模組自輸入影像開始，透過多層 Transformer block 提取逐層抽象的表示，並使用像素解混洗（pixel-unshuffle）操作進行空間解析度降低$\frac{1}{2}\times\frac{1}{2}$倍與通道擴展 2 倍，逐步將解析度由原始尺寸降至$\frac{1}{4}\times\frac{1}{4}$倍空間尺度；Decoder 則採用像素混洗 pixel-shuffle 操作進行對應的上採樣，並搭配 skip connection 將對應編碼層的特徵融合至 decoder 層，以彌補復原過程中的低頻與高頻資訊缺失。此對稱的編碼-解碼架構可視為一種「多尺度資訊融合機制」，不僅能有效擷取局部細節與全局語意特徵，更透過 skip connection 保留低層次的空間紋理，進而強化高品質影像的恢復能力。每一層的特徵融合後皆通過 1×1 卷積進行通道壓縮，以維持通道維度數量並保持編碼-解碼的特徵比重，提升記憶體效率與訓練穩定性。

解碼結束後的高層次特徵會進一步輸入至細化模組與高解析度優化的部分，此模組亦由多個自注意力模組組成，並在高解析度下運行以處理細節。透過維持空間完整性，此階段得以針對紋理細節、邊緣連續性與低頻偽影等空間細節進行最後優化。refinement 後的特徵會透過一個 1×1 卷積層轉換為殘差影像R並加回輸入影像I得到預測輸出：$\hat{I} = I + R$.

此架構以模組化設計，架構具有高度耦合能力，自注意力模組可替換為其他注意力機制，或調整 block 數量以適應不同運算資源與任務需求，具有良好的架構泛化性與應用彈性。

總結來說，這些設計共同促成了影像復原任務中所需的細節保留、殘差影像R一致性、整體視覺品質的提升和容易設計實驗的架構。

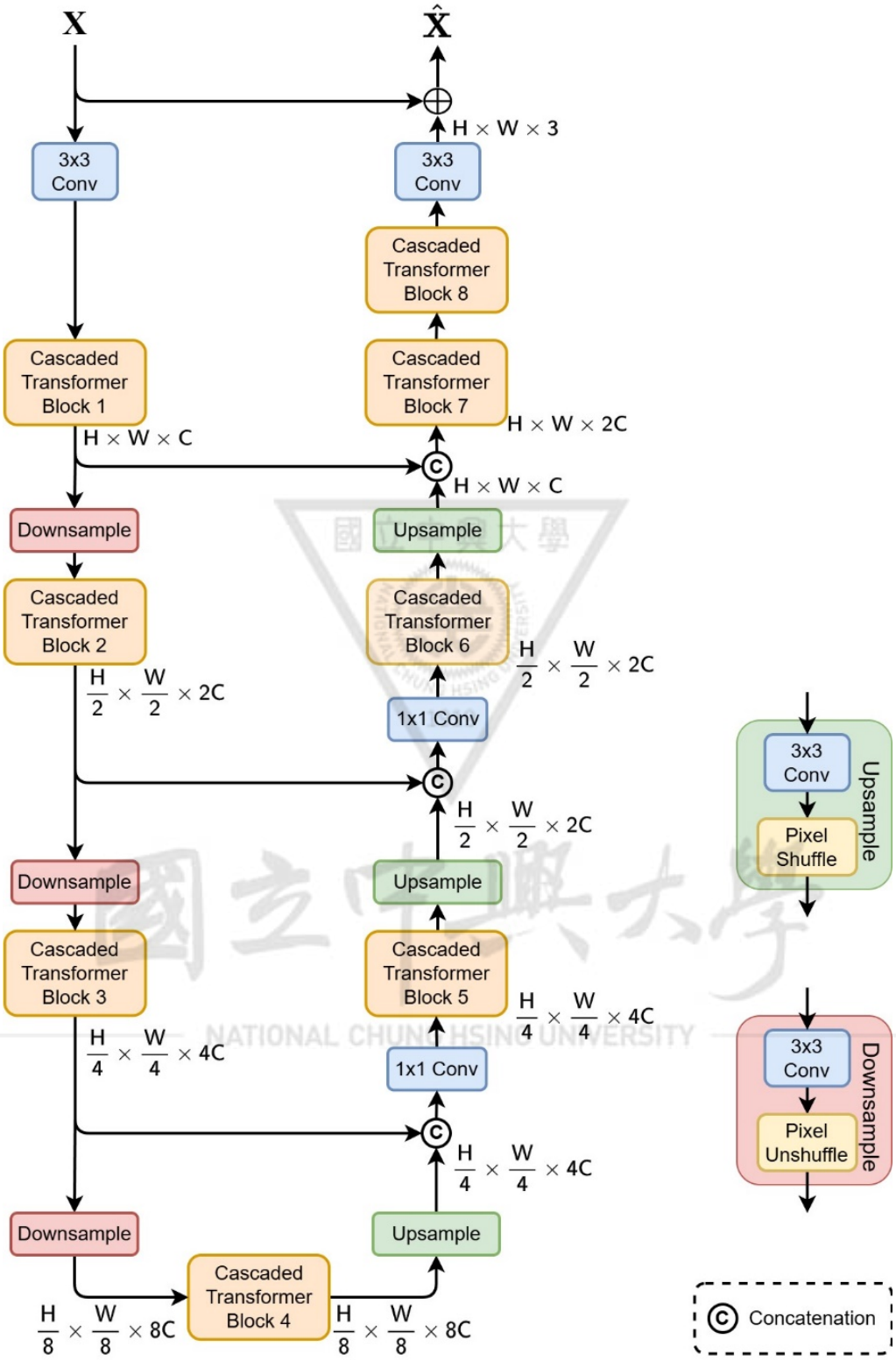


圖 8 區塊自注意力之轉換器模型 U-net 架構

3.2 轉換器模組 Transformer block

轉換器模組以自注意力模組和前饋神經網路構成，如圖 9，在本章會介紹我們設計的區塊自注意力模組（Patch-wise Self-attention Block），以及沿用修復器的門控深度卷積前饋模組（Gated-Dconv Feed-Forward Network）。架構定位與角色互補性下，自注意力模組和門控深度卷積前饋模組並非獨立設計。自注意力模組著重於建構跨空間位置的全域關聯與語意對齊，因為計算了各個位置的相關性；門控深度卷積前饋模組則專注於局部的非線性轉換與資訊控制，做特徵提取和篩選。

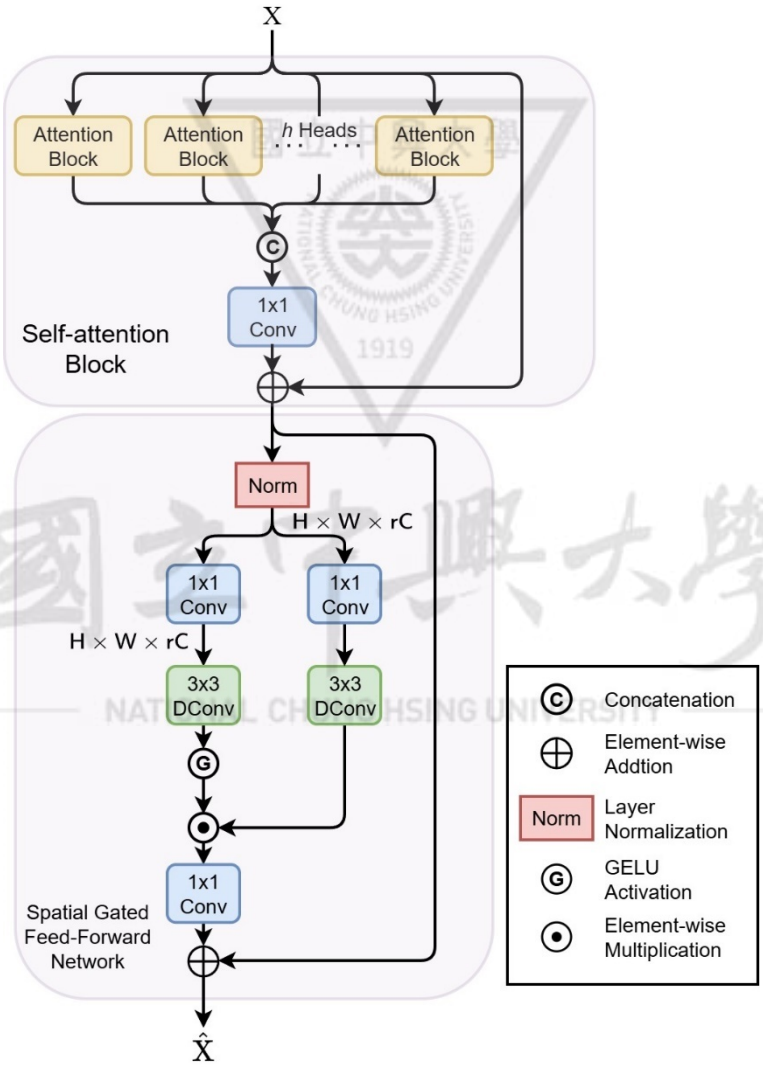


圖 9 轉換器模組 Transformer block

3.2.1 展開和折疊 Unfold and Fold

展開（Unfold），如圖 10，將輸入張量按照給定參數(kernel size、stride、padding) 進行視窗核心滑動分區，每個局部區塊攤平成特徵向量。參數介紹：

視窗大小 kernelsize 定義每個區塊的高與寬，即每次抽取的 patch 視窗大小。

補丁 padding 在輸入張量邊緣補零，使得邊界處也能形成完整的區塊。

步長 stride 定義視窗每次滑動的步長，控制相鄰區塊之間的重疊程度。

對於輸入張量$X \in R^{B\times C\times H\times W}$操作會在空間維度 H,W 上，以滑動視窗的方式，依參數進行視窗核心滑動分區，每個局部區塊攤平成特徵向量。視窗會依照參數設定，從左上角開始滑動步長距離直到圖像補丁邊緣，每次取出視窗長為P_h和寬為$P_w$patch，並對該 patch 所涵蓋的 C 通道進行攤平得到：

$$\text{Unfold (X)} = \text{Y} \in R^{B\times(C\cdot P_h\cdot \text{P}_w)\times N}$$

其中 N 為區塊總數。

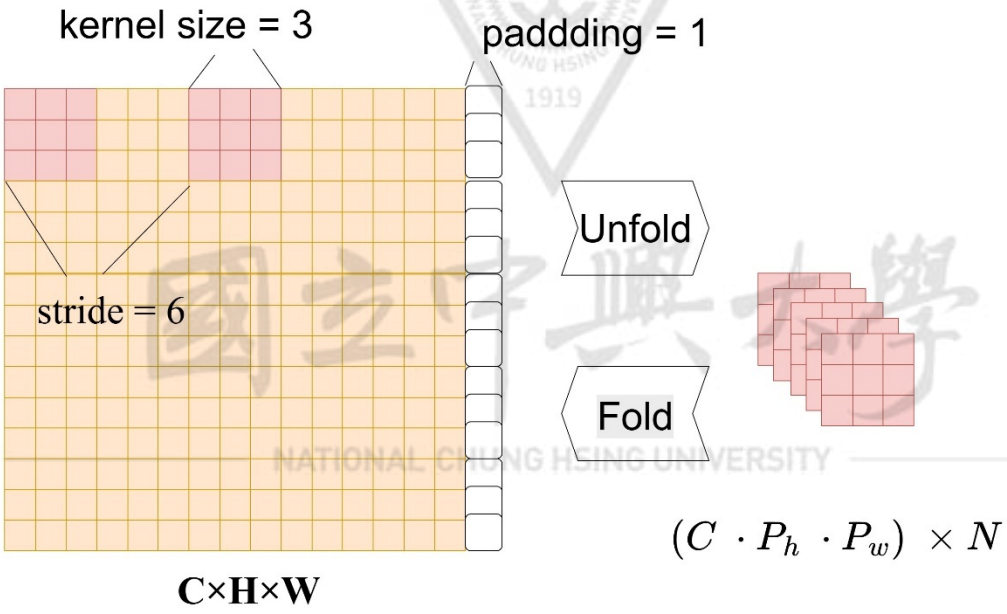


圖 10 折疊和展開

折疊（Fold），如圖 10，將上述展開後的張量，依照相同的參數設定還原至空間維度

$$\text{Y} \in R^{B\times(C\cdot P_h\cdot \text{P}_w)\times N},$$

$$\text{Fold(Y)} = \text{X} \in R^{B\times C\times H\times W}$$

折疊操作透過區塊還原原來的對應位置，若存在重疊需要進行額外處理，將相鄰重疊區塊加總進行重建。

3.2.2 區塊級自注意力模組（Patch-wise self-attention block, PW）

在高解析度影像復原任務中，轉換器[14]所採用的標準自注意力機制會有計算複雜度瓶頸問題，其計算複雜度與空間呈二次成長$\mathcal{O}(H^2W^2C)$。為解決此問題，本研究參考 ViT[7] 及 COLA-Net [21] 的設計理念，提出一種區塊級自注意力模組，其核心設計為將特徵圖劃分為非重疊的空間區塊（non-overlapping spatial blocks），並對區塊之間執行自注意力操作，從而降低注意力計算的空間複雜度。

輸入特徵圖表示為 $X \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$，其中$H$、$W$為空間維度，$C$為通道數。區塊自注意力機制先進行計算 QKV，注意力機制的運作方式如下，先做特徵嵌入(feature embedding)：

$$\mathbf{Q} = W_d^Q\left(W_p^Q\mathbf{X}\right),$$
$$\mathbf{K} = W_d^K\left(W_p^K\mathbf{X}\right),$$
$$\mathbf{V} = W_d^V\left(W_p^V\mathbf{X}\right),$$

其中，$W_p(\cdot)$ 表示 1×1 逐點卷積（point-wise convolution），其在每個空間位置聚合跨通道上下文；$W_d(\cdot)$ 表示 3×3 深度卷積（depth-wise convolution），其在每個通道內獨立編碼空間依賴關係。得到特徵向量$\mathbf{Q},\mathbf{K},\mathbf{V} \in R^{H\times W\times C}$。

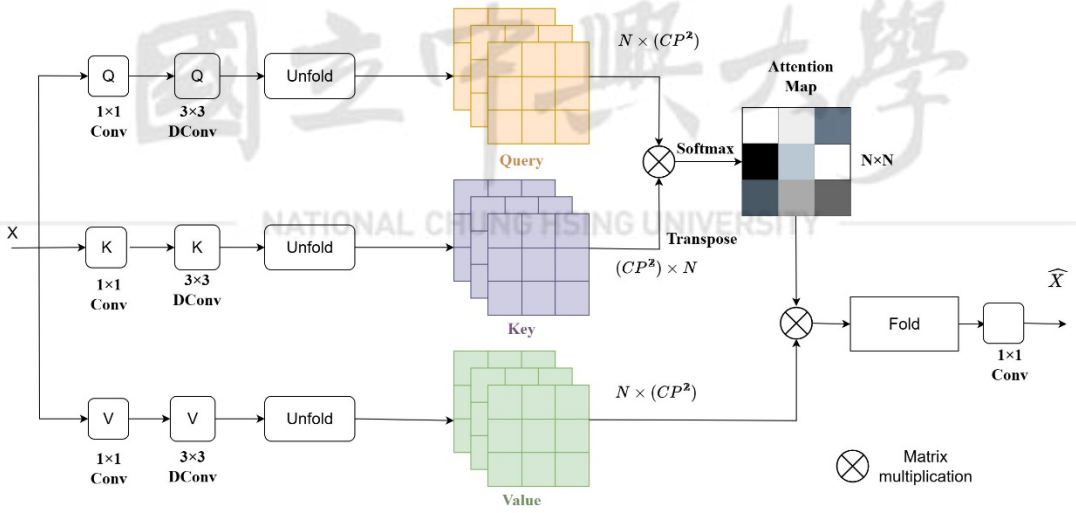


圖 11 區塊自注意力模組

再把每個$\mathbf{Q},\mathbf{K},\mathbf{V}$單獨展開$Unfold\,(\mathbf{Q},\mathbf{K},\mathbf{V}) \in R^{\mathrm{N}\times(CP^2)}$，展開操作如圖 10，把輸入特徵圖$\mathbf{Q},\mathbf{K},\mathbf{V}$轉為對應的 N 個扁平化的區塊。展開後以扁平化的區塊特徵進行自注意力機制圖 11，再將輸出折疊回去$fold\,(\dot{X}) \in R^{HW\times C}$，

折疊操作如圖 10，再以 1×1 逐點卷積做特徵嵌入得到輸出$\hat{X} \in R^{H\times W\times 3}$。

$$\text{Attention Map}(\mathbf{Q},\mathbf{K}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{t}\right) \in R^{N\times N}$$

$$\dot{X} = \text{Attention Map}(\mathbf{Q},\mathbf{K})\mathbf{V} \in R^{\mathrm{N}\times C*P}$$

$$\hat{X} = W^{X}\, fold\,(\dot{X}) + \mathbf{X}$$

其中，$W^{X}(\cdot)$ 表示 1×1 逐點卷積（point-wise convolution），其在每個空間位置聚合跨通道上下文，t 是對通道計算 Q、K 的長度，所以 QK 計算會從縮放點積變成弦積相似(cosine similarity)。

區塊注意力在視窗大小等於步長時，把原本 transformer 像素自注意力計算複雜度$\mathcal{O}((HW)^2)$降低到$\mathcal{O}(N^2) = \mathcal{O}\left(\frac{HW^2}{P^2}\right)$，$N$為區塊數，

$P = P_h \times P_w$，P_h和P_w是區塊長和寬，P 為視窗大小總像素數，也就是說計算複雜度隨區塊窗口呈平方倍反比變化。

區塊注意力設計確保每個區塊注意力模組關注區塊間的空間關係，並保持注意力全局上下文建模，這種區塊注意力促進了空間結構化的歸納偏差，這對於局部紋理和結構一致性至關重要的任務非常有用。

3.2.3 門控深度卷積前饋模組 Gated-Dconv Feed-Forward Block

在典型的轉換器架構中，特徵學習主要由全連接的前饋網路（Feed-Forward Network, FFN）完成，但這種設計在序列資料中有效，但在影像復原任務中卻存在結構性限制：其僅在通道維度操作，無法感知像素的空間鄰近關係。

修復器架構[10]中提出了門控深度卷積前饋網路（GDFN），如圖 12，以取代傳統只有線性映射前饋網路，以兩項機制取代傳統方式：

1. 1 深度可分離卷積（Depthwise Separable Convolution），使模型只對單一通道維度來運算卷積，利用空間鄰域的資訊進行特徵提取，並減少計算複雜度。這一改變使 GDFN 不再是空間不變（spatially invariant）的轉換模組，而能直接建模鄰近像素間的結構，如紋理與邊緣，對於影像修復任務中尤為重要。
2. 2 閘控機制（Gating Mechanism）導入雙路徑的閘控設計。輸入進入兩個平行分支，分別通過不同的線性變換與卷積處理，最終以元素乘法（Element-wise Multiplication）方式合併兩路輸出：此閘控機制的架構設計可讓網路根據輸入資料自動選擇性通過訊息，對於特徵提取做一個特徵篩選操作。

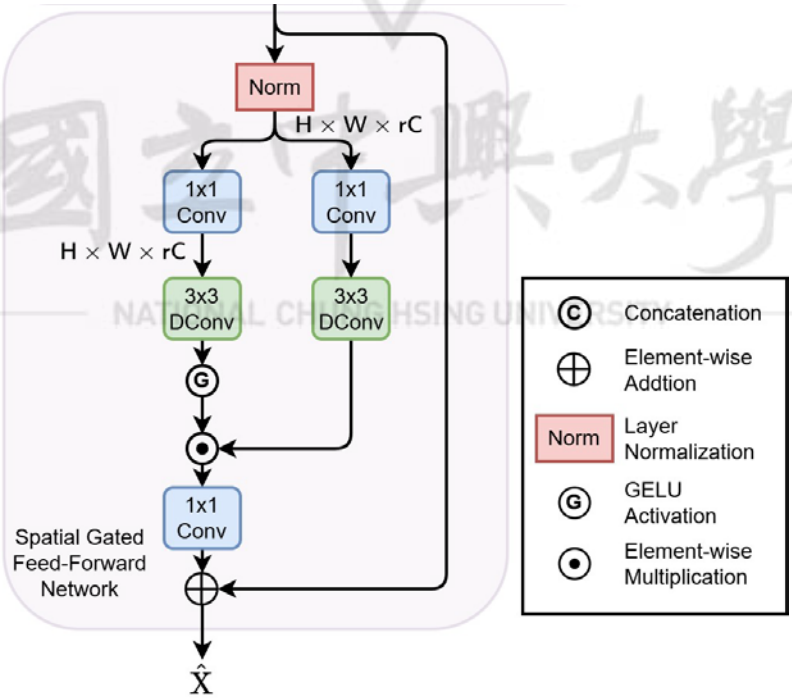


圖 12 門控深度卷積前饋模組

3.3 多種注意力機制之結合策略

在探討各種注意力機制的特性後，本研究進一步探索通道和區塊注意力機制之間是否存在互補性。單一類型的自注意力帶有其固有的歸納偏差（Inductive Bias），使其在特定類型的資訊處理上具有優勢。例如：通道自注意力（CW）其歸納偏差傾向於建模特徵通道間的關聯，擅長捕捉全局的特徵統計一致性。不同的，區塊自注意力（PW）其歸納偏差傾向以區塊於建模空間上的結構關係。它擅長捕捉全局的空間依賴性，形成一種空間結構化的先驗。我們也實驗出(1.21.4 區塊大小的自注意力不變性)區塊注意力所建構的空間依賴結構，具有自注意力上的區塊大小不變性（patch-size invariance），可望具備良好的模型泛化與結構轉移能力。

因此，我們假設透過結合不同類型的自注意力機制，可以讓模型學習到更全面、更泛化的影像先驗知識。為了驗證此假設，本研究設計了兩種不同的結合策略，以探討它們在影像修復任務中的表現。

3.3.1 串接 (Serial Connection)

第一種策略是將通道注意力（CW）與區塊注意力（PW）模組進行串接。具體而言，我們將模型中原有的單一自注意力模組，替換為一個由 CW 模組與 PW 模組前後相連構成的複合模組。

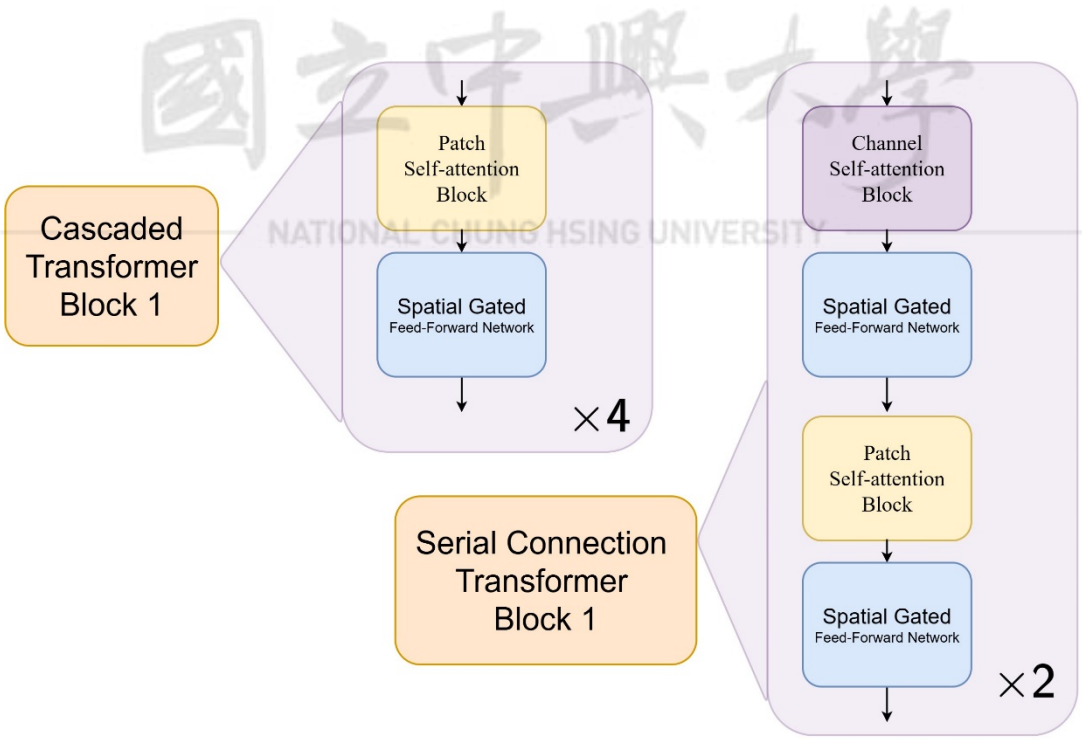


圖 13 門控深度卷積前饋模組

設計原理：此設計的理論基礎是建立「先特徵提取，後建模空間」的流程。我們假設通道注意力（CW）作為一個高效的特徵提取工具，能先對各個像素的特徵向量進行提煉與加權，增強其語意表達能力。再來，對這些經過語意增強的特徵輸入到區塊注意力（PW）中，進行空間結構上的關係建模。我們將在實驗中統一採用 CW 在前，PW 在後的順序。

3.3.2 編碼通道-解碼區塊 (Enoder Restormer-Decoder PW)

第二種策略是利用 U-Net 架構的特性，在模型的不同部分使用不同的注意力機制。我們將其稱為編碼-解碼融合，具體實現是透過特徵圖的拼接（Concat）完成。

設計原理：此理論基礎一樣以「先特徵提取，後建模空間」為流程。方法利用了 U-Net 架構的編碼器-解碼器分離特性，為不同階段分配最適合的注意力機制。

編碼器階段：負責將影像壓縮為多層次的抽象特徵。在此階段，我們全部採用通道注意力（CW），利用其計算效率高、擅長捕捉全局特徵關係的優點。

解碼器階段：核心任務是從抽象特徵中精確地重建影像的空間細節。在此階段，我們全部採用區塊注意力（PW），利用其空間結構建模能力來恢復高頻紋理與邊緣。

透過這兩種結合策略的實驗，本研究在探討不同歸納偏差的組合是否能產生「1+1>2」的效果，並找出在不同任務（如合成雜訊 vs. 真實雜訊）下最為有效的模型架構。

3.4 結語

本章建構了一套專為影像修復任務設計的實驗方法。核心貢獻在於提出了一種區塊級自注意力（PW）機制，有效降低了自注意力在高解析度場景下的計算負擔，同時保留了對空間結構的建模能力。此外，本章基於實驗分析，設計了兩種注意力融合策略（串接與編碼器-解碼器分層），旨在探討結合不同歸納偏差對模型性能的影響。

第四章 實驗結果分析

本章在 4.1 介紹訓練和測試資料集，在 4.2 介紹用到的影像品質指標，在 4.3 介紹實驗的模型和訓練設定，在 4.4 展示模型計算複雜度和參數量，在 4.5, 4.6 則展示模型在不同資料集的結果。

本章亦嘗試探討此區塊注意力的建模特性，並與其他注意力形式（如全域注意力、視窗注意力等）進行比較，以理解其在捕捉局部與全域語意依賴上的優劣與適用性。

4.1 訓練和測試資料集

本研究針對三項影像復原任務：合成雜訊去雜訊（Additive Gaussian Noise Denoising）、真實影像去雜訊（Real Image Denoising）與影像去模糊（Deblurring），使用多個標準基準資料集(benchmark dataset)進行模型訓練與評估。下列分別說明各任務所採用的資料來源與數量。

4.1.1 合成雜訊去雜訊（Additive Gaussian Noise Denoising）

針對合成雜訊的復原任務，為了評估模型在可控雜訊環境下的表現，我們使用多個資料集訓練模型於合成高斯雜訊任務中以增加模型泛化：

訓練集：我們在這些影像中加入標準高斯雜訊（$\sigma = 0 \sim 50$ 之間）

資料集	訓練影像數量	備註
DIV2K[21]	800	高品質超解析度影像集
Flickr2K[21]	2,650	網路照片數據集，常與 DIV2K 合併
WaterlooED [22]	4,744	品質失真評估數據集
BSD400 [23]	400	經典去雜訊訓練資料集

測試集：CBSD68、Kodak24、McMaster 和 Urban100 為主要測試資料集，我們在這些影像中加入標準高斯雜訊（$\sigma = 15, 25, 50$）進行性能評估。

資料集	測試影像數量	備註
CBSD68 [23]	68	BSD 系列之測試集，常用於合成高斯雜訊基準評估
Kodak24	24	高品質彩色圖像，常用於測試
McMaster	18	高彩色飽和度圖像集
Urban100 [24]	100	高結構性與細節豐富圖像

4.1.2 真實雜訊去雜訊（Real Image Denoising）

為處理實際攝影過程中產生的非理想雜訊（非加性、高斯非穩態），我們採用 Smartphone Image Denoising Dataset (SIDD)。SIDD 包含來自 5 種智慧型手機、10 種場景、不同照明條件下的共 30000 張帶有實際雜訊的圖像。每張影像對應一張對齊的乾淨 Ground Truth。

在本研究中，我們從 SIDD 訓練集選擇了 320 張圖像進行模型訓練，測試集則採用官方提供的標準補丁：

訓練集：320 張完整圖像裁切成，共 30608 個 512×512 的圖像區塊

測試集： 40 張完整圖像裁切成，共 1,280 個 256×256 的圖像區塊

4.2 影像品質指標（Image quality metric）

影像品質評估是影像處理與生成任務中的重要環節，能夠量化模型輸出的影像與參考影像之間的差異。常見的全參考指標包括 PSNR、SSIM 與 LPIPS，各具優缺點與適用情境。

4.2.1 峰值訊噪比（Peak Signal-to-Noise Ratio, PSNR）

峰值訊噪比是影像品質評估中最常用的傳統指標之一，用於衡量輸出影像與參考影像之間的像素誤差強度，單位為分貝（dB），數值越高表示兩者越相似、失真越小。

峰值訊噪比的數學定義，基於均方誤差（MSE）定義的對數比率，其中 $\mathrm{MAX_I} = 255$ 對於 8-bit 影像：

$$\mathrm{MSE_{RGB}} = \frac{1}{3HW} \sum_{c\in\{R,G,B\}} \sum_{i=1}^{H} \sum_{j=1}^{W} \left[I_c(i,j) - \hat{I}_c(i,j) \right]^2$$

$$\mathrm{PSNR_{RGB}} = 10 \cdot \log_{10} \left(\frac{\mathrm{MAX_I^2}}{\mathrm{MSE_{RGB}}} \right)$$

峰值訊噪比的優點是計算快速，物理意義明確，但缺點是無法有效反映人類對紋理、結構等細節的感知。所以需要下個指標。

4.2.2 結構相似性指標（Structural Similarity Index Measure, SSIM）

Wang 等人提出結構相似性指標[]作為更符合人眼感知機制的影像品質評估方法。與僅關注像素誤差的 PSNR 不同，SSIM 從亮度（Luminance）、對比度（Contrast）與結構（Structure）三個層面進行比較，模擬人類視覺系統（Human Visual System）對影像的感知方式。

結構相似性指標的數學定義是一種區塊指標，其計算通常基於影像的滑動視窗。對於參考影像區域 x 與待評估影像區域 y，其 SSIM 可表示為：

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

將 R、G、B 的值相加平均，即為彩色影像的 SSIM，其中：

μ_x, μ_y：分別為區塊與的平均

σ_x, σ_y：為各自的標準差，用以估計對比度

σ_{xy}：為x, y間的協方差，用以評估結構相似性

C_1, C_2：是穩定項，用於避免分母為零

能捕捉結構與對比特徵，比 PSNR 更接近人類視覺評估，適用於多種影像處理任務，如壓縮、去雜訊與超解析度。但缺點是對於非結構性扭曲（如色彩偏移、旋轉）不夠敏感。

4.3 模型和訓練設定

所有實驗均使用相同模型架構設定、轉換器模組設定和訓練參數設定，除非另有註明。

架構設定：

參考修復器模型[10]架構的設定，本模型採用 4 層編碼解碼架構，每一層皆由若干 Transformer blocks 組成，並搭配多頭注意力與門控深度卷積前饋模組進行特徵轉換與聚合。模型架構各層數值如下：

Transformer blocks 轉換器模組數量（levels 1–4）： [4 , 6 , 6 , 8]

多頭注意力頭數： [1 , 2 , 4 , 8]

特徵通道數量： [48 , 96 , 192 , 384]

Refinement stage：包含 4 個 Transformer blocks

GDFN 通道擴展倍率： $\gamma = 2.66$

此設計兼顧運算效率與表現能力，並透過多層編碼解碼層以強化影像中不同尺度資訊的建模能力。

轉換器模組設定：

對於區塊自注意力模組，視窗大小、補丁、步長設定都為：4；

基於視窗自注意力模組，視窗大小：7，轉移大小：0；

訓練設定：訓練在 Python 11.9.3 的 PyTorch，使用 Python 套件 skimage 實作 PSNR 和 SSIM 的計算。

損失函數： L_1 損失函數，適用於像素重建任務；

優化器：AdamW[25]，參數設為 $\beta_1 = 0.9 , \beta_2 = 0.999$, weight decay $= 1 \times 10^{-4}$ ；

學習率：採用 cosine annealing[26] 機制初始為 3×10^{-4} 遞減至 1×10^{-6} ；

資料擴增方式：隨機水平與垂直翻轉；

訓練輪數：訓練共進行 300K iterations。並使用漸進式訓練（Progressive Learning），自小至大的 patch size 搭配遞減 batch size，以提升訓練穩定性與空間泛化能力；

表格 1 漸進式訓練參數設定

Iteration	Patch Size	Batch Size
0–92K	128×128	8
92K–156K	160×160	5
156K–204K	192×192	4
204K–240K	240×240	2
240K–276K	240×240	1
276K–300K	256×256	1

4.4 SIDD 真實影像去雜訊

對於真實影像去雜訊任務中，主要是減少影像訊噪比，所以在性能的衡量上本研究將針對 PSNR 指標為第一參考指標。

4.4.1 不同自注意力的影響

控制變數全部參數按照 1.20 模型和訓練設定。自變數為轉換器模組中的不同自注意力模組，有修復器 Restormer[10]的通道(Channel-Wise, CW)自注意力(架構同修復器)、本研究的區塊(Patch-Wise, PW)自注意力、SwinIR[8]的基於視窗(Window-Based, WB)自注意力和不使用(None)注意力直接將 QKV 特徵相加。其中應變數最優 PSNR 和 SSIM 以螢光顯示。

從表格 2 中可以看到，本研究的區塊自注意力有最好的表現，這是改良自注意力的優勢展現。。

而根據其他自注意力可以推論，基於視窗自注意力和不使用(None)注意力在學習表現與其他注意力差 0.2dB，兩者彼此卻差異不大 0.08dB，注意力的全局性在這裡影響大。

表格 2　SIDD 資料集上模型對於不同自注意力的影響

	PSNR	SSIM
Restormer	39.62	0.912
PW	**39.74**	**0.913**
SwinIR	39.48	0.909
None attention	39.41	0.909

4.4.2 多模型結合的影響

控制變數全部參數按照 1.20 模型和訓練設定。自變數為修復器[10]的通道(Channel-Wise, CW)自注意力、本研究的區塊(Patch-Wise, PW)自注意力和兩者的結合方式，參考 1.16 多種注意力機制之結合策略，第一種，以 Restormer 的 CW 作為自注意力模組在編碼器，PW 作為自注意力模組在後續的融合並列(Fusion Concat)結合。第二種，以通道注意力在前作為自注意力，而區塊注意力在後串接（Connection），倆倆交替串接以此算作 2 個自注意力模組全部替換 2 個單一自注意力模組。其中 CW 都在 PW 前，因為理論上 CW 是比較好的特徵提取工具。其中應變數最優 PSNR 和 SSIM 以螢光顯示。

表格 3 排序 PSNR 從小到大，最好的模型將 CW 在前作為自注意力模組，PW 在後串接，串接模組會有最好的表現。

表格 3 SIDD 資料集上模型對於多模型結合的影響

	PSNR	SSIM
Restormer	39.6229	0.9119
Enoder Restormer-Decoder PW	39.6653	0.9119
PW	39.7373	0.9129
Restormer &PW Connection	**39.7475**	**0.9131**

4.4.3 基準性能比較

根據[27]中 SIDD 資料集上基準，本研究列出其他模型基準指標，來與區塊自注意力做比較和對照，於表格 4，其中為了和其他模型基準對齊本研究用 MATLAB 來計算指標，而非 Python。

神經網路模型概念原理上，修復器(Restormer), DAGL[28], Ours, 都是基於轉換器，DANet+[29]是基於對抗網路，SADNet[30]是基於卷積網路。

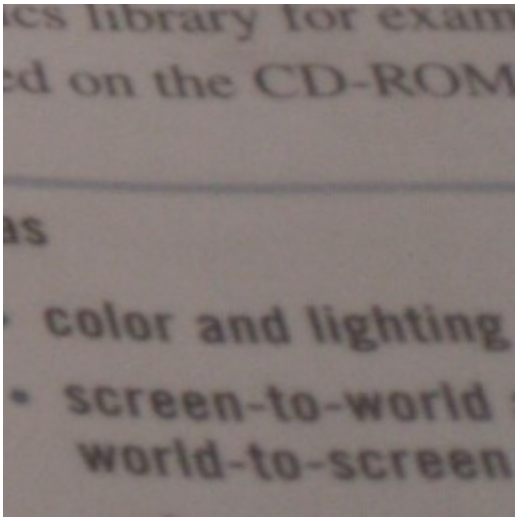
我們的區塊自注意力全部參數按照 1.20 模型和訓練設定，除了在 1.21.1 不同自注意力的影響，發現 CW&PW Connection 具有最好的效能，所以模型的自注意力模組採用其模組，以通道注意力在前作為自注意力，而區塊注意力在後串接設計。

受限於設備記憶體限制本研究的 1.20 模型和訓練設定無法與修復器一樣，所以 CW(通道注意力)也就是修復器本研究的訓練設定下，表現結果比我們的模型差顯示在表格 4，但修復器[10]中的結果比我們的模型高 0.27dB。這只是因為訓練設定的影響。

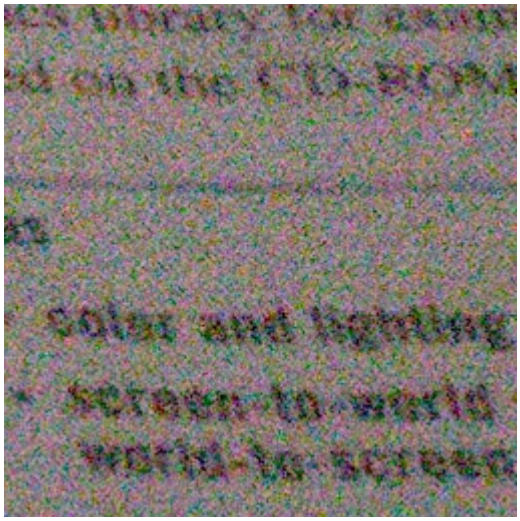
圖 14 SIDD 資料集上紋理細節圖的比較和圖 15 SIDD 資料集上平滑區域圖的比較，直接說明了空間上的資訊和特徵上的資訊的差別，紋理細節圖需要空間資訊，所以修復器[10]不擅長，而我們的模型有區塊自注意力來處理空間資訊，所以有較佳表現。相反平滑區域圖是修復器[10]通道自注意力和 DANet+[29]對抗網路擅長，尤其通道自注意力大大的領先 1dB。

表格 4　SIDD 資料集上基準性能比較

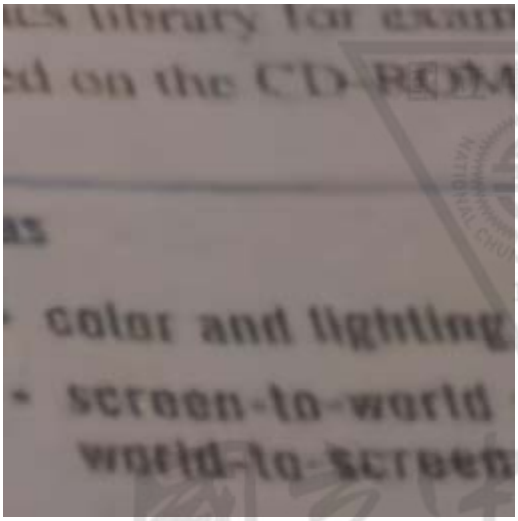
	Restormer	Ours	SADNet	DANet+	DAGL
PSNR	**40.02**	39.75	39.46	39.47	39.20
SSIM	**0.960**	0.959	0.957	0.957	0.957



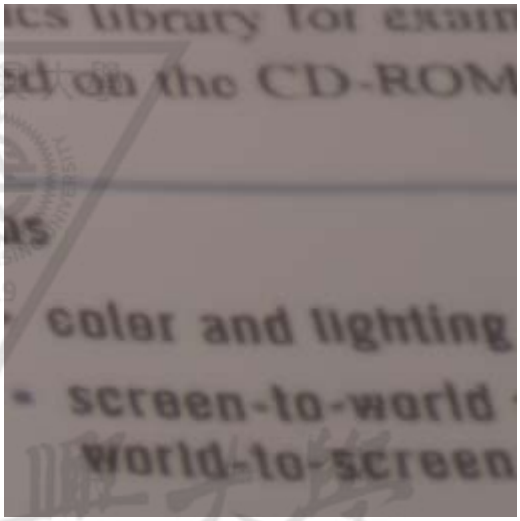
原始影像 Groun Truth



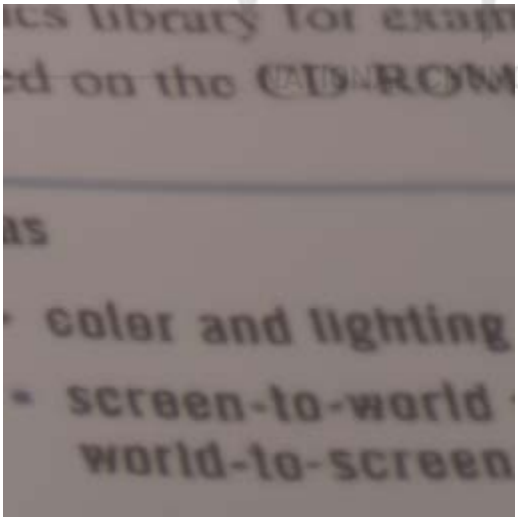
雜訊影像　Noisy Image



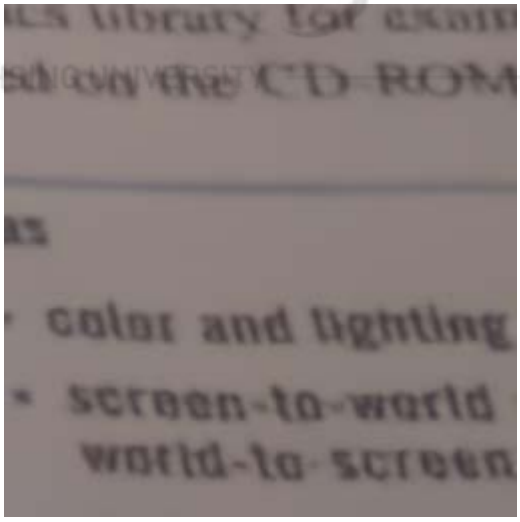
DAGL　34.96　0.936



DANET+　35.03　0.940



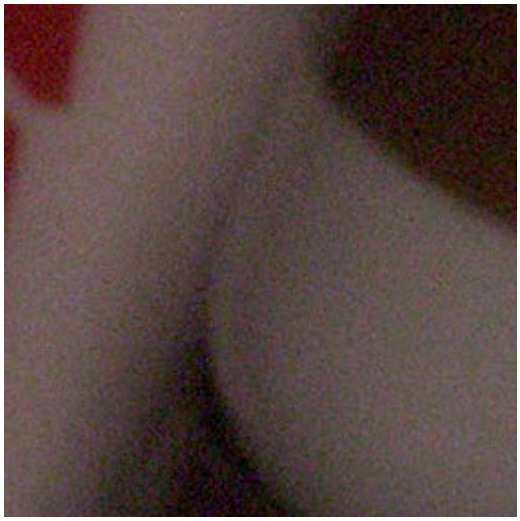
Restormer　35.03　0.940



Ours　35.40　0.942

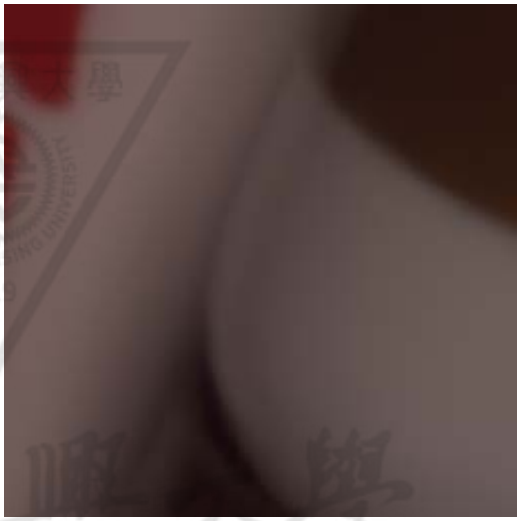
圖 14 SIDD 資料集上紋理細節圖的比較

原始影像 Groun Truth

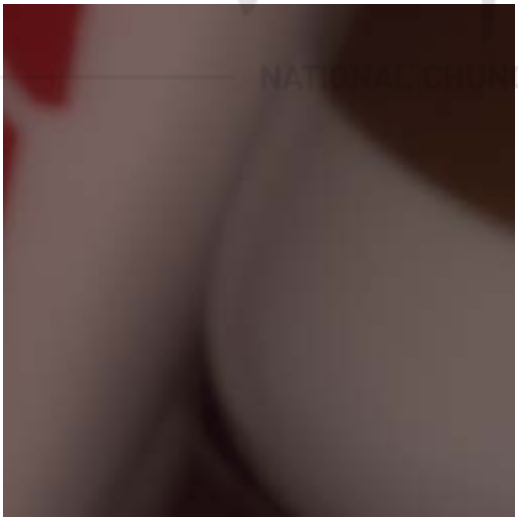


雜訊影像　Noisy Image

DAGL　45.23　0.994



DANET+　45.92　0.994



Restormer　46.76　0.995

Ours　45.82　0.994

圖 15 SIDD 資料集上平滑區域圖的比較

4.4.4 區塊大小的自注意力不變性

我將在參數按照 1.20 模型和訓練設定，區塊(Patch-Wise, PW)自注意力的區塊大小為 4 訓練後，模型權重結果拿來用在區塊大小 3 和 5 推論，此非對稱的訓練推論區塊大小上的結果在表格 5，發現 PSNR 只下降 1dB，SSIM 只下降 0.008，這表示了注意力的空間上的泛化能力，注意力的根據空間動態調整特徵加權，使特徵適應不同自注意力。

表格 5　SIDD 資料集上自注意力於區塊大小的非對稱訓練推論

區塊大小	PSNR	SSIM
4	**39.74**	**0.913**
4 to 3	38.83	0.905
4 to 5	38.71	0.905

近一步探討注意力的空間和通道間的泛化能力，參數按照 1.20 訓練後的模型權重結果拿來用在不同自注意力間的非對稱訓練推論上，結果在表格 5，有修復器[10]的通道(Channel-Wise, CW)自注意力(架構同修復器)、本研究的區塊(Patch-Wise, PW)自注意力、SwinIR[8]的基於視窗(Window-Based, WB)自注意力和不使用(None)注意力直接將 QKV 中的 Value 特徵導出。

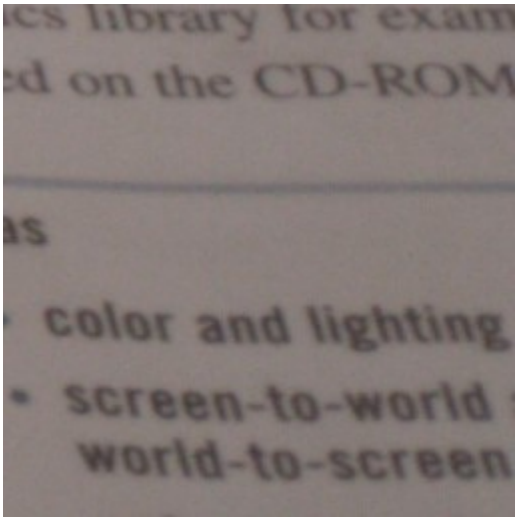
對照 SIDD 影像測試集的基線雜訊(noisy baseline)的值為：

PSNR	23.66
SSIM	0.328

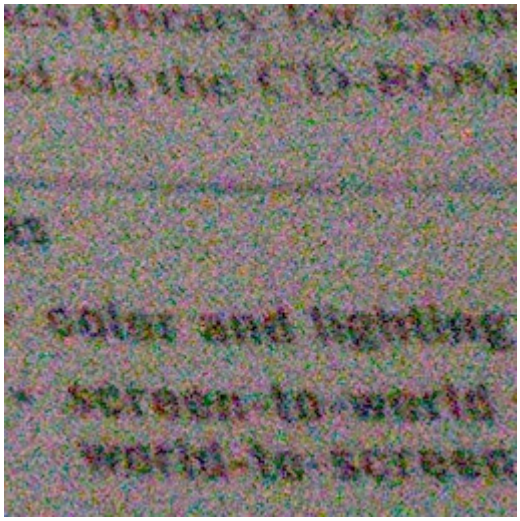
可以看到自注意力空間和通道間特徵不具有遷移(transfer)能力，但 PW 到 WB 因為都是對空間做自注意力，所以還有一定的相似性。而不使用(None)注意力直接導出區塊注意力中的值(value)，則沒有任何影像修復。

表格 6　SIDD 資料集上自注意力於不同自注意力的非對稱訓練推論

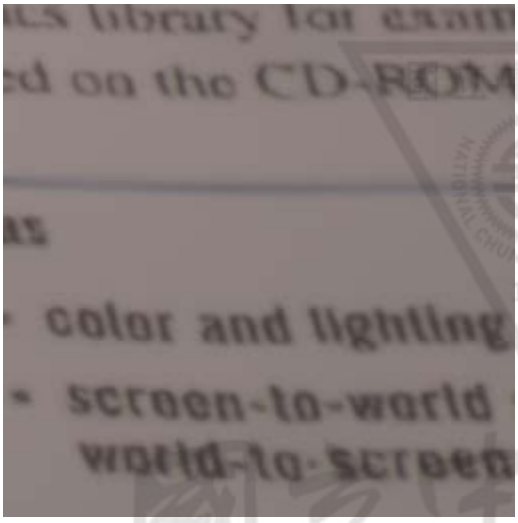
區塊大小	PSNR	SSIM
PW	39.74	0.913
PW to Restormer	18.67	0.128
PW to SwinIR	29.81	0.636
PW to Value	22.23	0.236



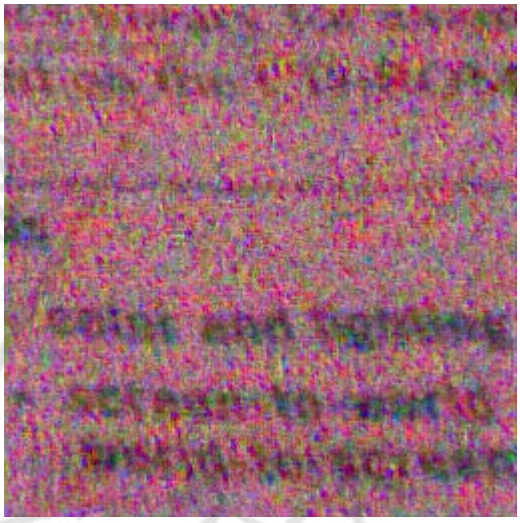
原始影像 Groun Truth



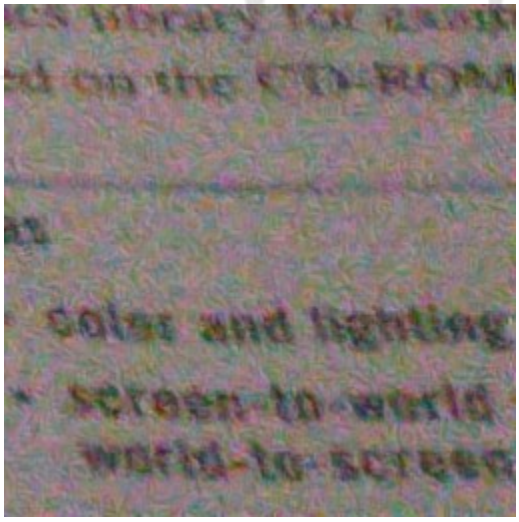
雜訊影像



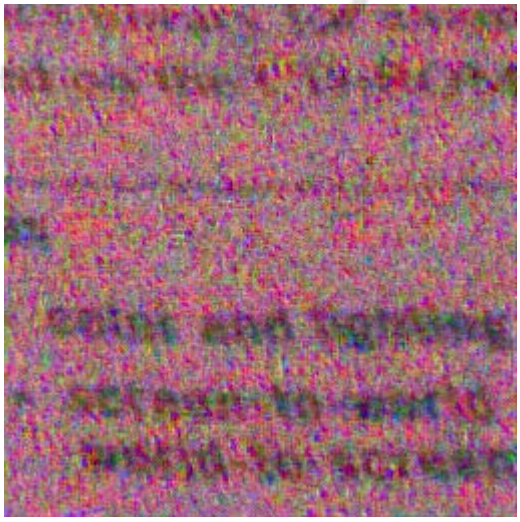
PW 35.37 0.918



PW to Restormer 19.29 0.118



PW to SwinIR 25.66 0.537



PW to None

圖 16 SIDD 資料集上不同自注意力的非對稱訓練推論的比較

4.5 合成高斯影像去雜訊 Gaussian Noise Color Denoising

對於合成影像去高斯雜訊任務中，雜訊以可加性高斯白噪音被加進原始影像，去高斯雜訊任務主要是減少影像訊噪比 PSNR。

4.5.1 區塊自注意力中區塊大小的影響

控制變數全部參數按照 1.20 模型和訓練設定。自變數為區塊自注意力中區塊的大小，也就是視窗大小、補丁、步長設定共同的數值大小，從記憶體可負擔的區塊長寬 4 一直以倍遞增至 32。其中應變數最優 PSNR 和 SSIM 以螢光顯示。

從表格 7 到表格 10 展現明顯的一致性，那就是隨著自變數區塊的變大，應變數 PSNR 和 SSIM 會跟著下降，所以可以推論越小的區塊甚至到區塊 1(等於 Token-wise)輸出的性能指標越好，但計算複雜度會跟著上升，因此如何利用性能和複雜度是一個問題。

根據圖 17 CBSD68 資料集 σ=50 上紋理細節圖的比較，可以看到將區塊設大就會失去修復辨識空間上的細節線條和紋理。

CBSD68 資料集：

表格 7　CBSD68 資料集上模型對於區塊大小的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
4	**33.92**	**31.0**	**26.80**	**0.9335**	**0.8872**	**0.784**
8	33.86	30.94	26.74	0.9321	0.8849	0.7808
16	33.86	30.92	26.69	0.9329	0.8856	0.7794
32	33.85	30.9	26.66	0.9328	0.8852	0.7782

Kodak 資料集：

表格 8　Kodak 資料集上模型對於區塊大小的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
4	**34.999**	**32.313**	**28.194**	**0.929**	**0.889**	**0.801**
8	34.901	32.224	28.130	0.927	0.886	0.798
16	34.902	32.206	28.065	0.928	0.887	0.797
32	34.896	32.183	28.034	0.928	0.886	0.796

McMaster 資料集：

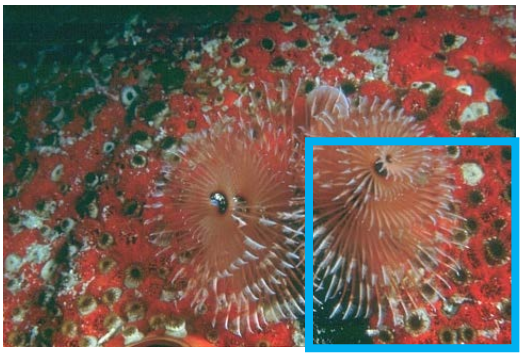
表格 9　McMaster 資料集上模型對於區塊大小的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
4	**34.75**	**31.70**	**26.60**	**0.9121**	**0.8579**	0.7441
8	34.64	31.64	26.59	0.9103	0.8568	**0.7479**
16	34.62	31.56	26.47	0.9105	0.8557	0.7426
32	34.65	31.56	26.43	0.9112	0.8561	0.7402

Urban100 資料集：

表格 10　Urban100 資料集上模型對於區塊大小的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
4	**34.38**	**31.62**	**26.90**	**0.9350**	**0.9050**	**0.8374**
8	34.17	31.40	26.72	0.9328	0.9019	0.8330
16	34.10	31.28	26.56	0.9327	0.9009	0.8275
32	34.10	31.26	26.51	0.9330	0.9007	0.8256



原始影像 Groun Truth

原始放大影像 Groun Truth

區塊大小=4　27.68　0.766

區塊大小=8　27.63　0.768

區塊大小=16　27.55　0.760

區塊大小=32　27.46　0.744

圖 17 CBSD68 資料集 σ=50 上紋理細節圖的比較

4.5.2 區塊自注意力中區塊和步長比例大小的影響

控制變數全部參數按照 1.20 模型和訓練設定中的數值。自變數為區塊自注意力中區塊比例的大小,區塊視窗大小設為 16 而區塊步長從 2 到 16,從區塊 16:16(區塊視窗大小:步長比例),一直以 2 倍遞減至 2:16。其中應變數最優 PSNR 和 SSIM 以螢光顯示。

從表格 11 到表格 14,即自變數區塊和步長比例在 2:16 和 4:16 應變數 PSNR 和 SSIM 比較好而 16:16 幾乎最差,區塊小於步長等於我們只把影像一部份比例拿去學習,其他影像部份忽略,這樣有差異不大的性能指標,原因是什麼原因有待調查。但這樣做大幅降低計算複雜度。

CBSD68 資料集:

表格 11　CBSD68 資料集上模型對於區塊和步長比例大小的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
2:16	**33.879**	30.942	26.742	**0.9334**	**0.8866**	0.7837
4:16	33.873	30.941	**26.750**	0.9333	0.8865	**0.7839**
8:16	33.875	**30.943**	26.743	0.9333	0.8864	0.7836
16:16	33.861	30.921	26.690	0.9329	0.8856	0.7794

Kodak 資料集:

表格 12　Kodak 資料集上模型對於區塊和步長比例大小的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
2:16	**34.918**	32.217	28.107	**0.9289**	**0.8879**	0.7998
4:16	34.916	32.217	**28.116**	0.9289	0.8878	**0.8001**
8:16	34.915	**32.218**	28.105	0.9288	0.8877	0.7998
16:16	34.902	32.206	28.065	0.9284	0.8870	0.7968

McMaster 資料集：

表格 13　McMaster 資料集上模型對於區塊和步長比例大小的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
2:16	**34.654**	31.591	26.513	0.9117	0.8574	0.7468
4:16	34.653	**31.603**	**26.539**	**0.9118**	**0.8580**	**0.7483**
8:16	34.645	31.578	26.506	0.9116	0.8572	0.7468
16:16	34.619	31.562	26.471	0.9105	0.8557	0.7426

Urban100 資料集：

表格 14　Urban100 資料集上模型對於區塊和步長比例大小的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
2:16	34.082	31.257	26.572	0.9332	0.9013	0.8299
4:16	34.082	31.256	**26.580**	**0.9333**	**0.9014**	**0.8304**
8:16	34.088	31.262	26.575	0.9332	0.9014	0.8302
16:16	**34.104**	**31.284**	26.558	0.9327	0.9009	0.8275

4.5.3 不同自注意力的影響

控制變數全部參數按照 1.20 模型和訓練設定。自變數為轉換器模組中的自注意力模組不同，有修復器[10]的通道(Channel-Wise, CW)自注意力(等同於 Restormer)、本研究的區塊(Patch-Wise, PW)自注意力、SwinIR [8]的基於視窗(Window-Based, WB)自注意力和不使用(None)注意力直接將 QKV 特徵相加。其中應變數最優 PSNR 和 SSIM 以螢光顯示。

從表格 15 表格 16 表格 17 表格 18 中可以看到，本研究的區塊自注意力有最好的表現，這是改良自注意力的優勢展現。

根據圖 18。再次驗證區塊(Patch-Wise, PW)自注意力比通道(Channel-Wise, CW)自注意力適合空間細節處理。

CBSD68 資料集：

表格 15　CBSD68 資料集上模型對於不同自注意力的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
Restormer	33.900	30.942	26.674	0.9335	0.8860	0.7773
PW	**33.923**	**31.000**	**26.795**	**0.9335**	**0.8872**	**0.7840**
SwinIR	33.872	30.930	26.734	0.9332	0.8863	0.7832
None	31.276	29.511	26.074	0.8542	0.8163	0.7169

Kodak 資料集：

表格 16　Kodak 資料集上模型對於不同自注意力的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
Restormer	34.966	32.238	28.037	0.9292	0.8877	0.7948
PW	**34.999**	**32.313**	**28.194**	**0.9295**	**0.8891**	**0.8010**
SwinIR	34.927	32.221	28.115	0.9287	0.8875	0.7995
None	31.766	30.347	27.155	0.8319	0.7976	0.7103

McMaster 資料集：

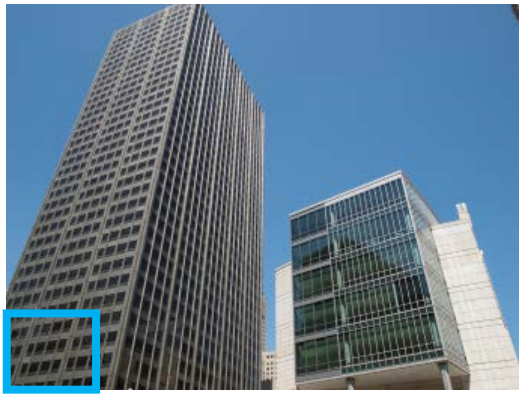
表格 17　McMaster 資料集上模型對於不同自注意力的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
Restormer	34.677	31.564	26.427	0.9107	0.8542	0.7353
PW	**34.751**	**31.699**	**26.603**	**0.9121**	**0.8579**	0.7441
SwinIR	34.662	31.593	26.538	0.9116	0.8568	**0.7461**
None	31.837	30.069	25.923	0.8286	0.7831	0.6764

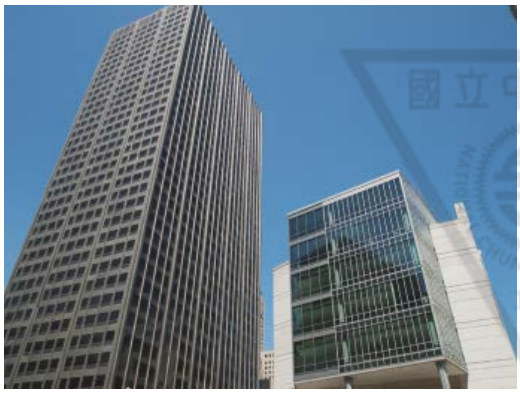
Urban100 資料集：

表格 18　Urban100 資料集上模型對於不同自注意力的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
Restormer	34.233	31.391	26.614	0.9339	0.9020	0.8265
PW	**34.381**	**31.617**	**26.899**	**0.9350**	**0.9050**	**0.8374**
SwinIR	34.113	31.274	26.571	0.9333	0.9010	0.8287
None	30.696	29.355	25.622	0.8524	0.8280	0.7529



原始影像 Groun Truth　　原始放大影像 Groun Truth



CW　29.93　0.941　　CW

PW　30.74　0.951　　PW

圖 18 Urban100 資料集 σ=50 上紋理細節圖的比較

4.5.4 多模型結合的影響

控制變數全部參數按照 1.20 模型和訓練設定。自變數為修復器 Restormer[10]的通道(Channel-Wise, CW)自注意力、本研究的區塊(Patch-Wise, PW)自注意力和兩者的結合方式，參考 1.16 多種注意力機制之結合策略，第一種，以 CW 作為自注意力模組在編碼器，PW 作為自注意力模組在後續解碼器的融合(Fusion Concat)結合。第二種，以 CW 在前作為自注意力模組 PW 在後串接（Connection），兩兩串接以此算作 2 個自注意力模組全部替換 2 個單一自注意力模組。其中 CW 都在 PW 前，因為理論上 CW 是比較好的特徵提取工具。其中應變數最優 PSNR 和 SSIM 以螢光顯示。

從表格 19、表格 20、表格 21、表格 22 中可以看到，本研究的區塊自注意力有最好的表現，這是改良自注意力的優勢展現，不需要結合其他模型來引入歸納偏差，和在 SIDD 資料集上表現不同。而其他自注意力結合方式可以推論，串接比編碼通道自注意力並解碼區塊自注意力好。

CBSD68 資料集：

表格 19　CBSD68 資料集上模型對於多模型結合的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
Restormer	33.900	30.942	26.674	0.9335	0.8860	0.7773
PW	**33.923**	**31.000**	**26.795**	0.9335	**0.8872**	**0.7840**
CW&PW Connection	33.923	30.972	26.728	**0.9338**	0.8866	0.7792
Enoder CW-Decoder PW	33.905	30.952	26.704	0.9333	0.8856	0.7774

Kodak 資料集：

表格 20　Kodak 資料集上模型對於多模型結合的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
Restormer	34.966	32.238	28.037	0.9292	0.8877	0.7948
PW	34.999	**32.313**	**28.194**	0.9295	**0.8891**	**0.8010**
CW&PW Connection	**35.002**	32.282	28.100	**0.9297**	0.8885	0.7953
Enoder CW-Decoder PW	34.978	32.262	28.068	0.9290	0.8873	0.7929

McMaster 資料集：

表格 21　McMaster 資料集上模型對於多模型結合的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
Restormer	34.677	31.564	26.427	0.9107	0.8542	0.7353
PW	**34.751**	**31.699**	**26.603**	**0.9121**	**0.8579**	**0.7441**
CW&PW Connection	34.716	31.619	26.510	0.9113	0.8550	0.7371
Enoder CW-Decoder PW	34.708	31.599	26.452	0.9112	0.8543	0.7335

Urban100 資料集：

表格 22　Urban100 資料集上模型對於多模型結合的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
Restormer	34.233	31.391	26.614	0.9339	0.9020	0.8265
PW	**34.381**	**31.617**	**26.899**	**0.9350**	**0.9050**	**0.8374**
CW&PW Connection	34.351	31.545	26.791	0.9348	0.9038	0.8317
Enoder CW-Decoder PW	34.339	31.542	26.781	0.9345	0.9033	0.8301

4.5.5 基準性能比較

將我們的模型區塊自注意力，設定如 1.20 模型和訓練設定。還有其他模型進行比較，Ours、Restormer、SwinIR 都是轉換器架構，DRUNet 是卷積神經網路 U-Net 式架構。

CBSD68 資料集：

表格 23　CBSD68 資料集上模型對於多模型結合的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
Ours	34.25	31.63	28.45	0.936	0.895	0.811
Restormer	34.39	31.78	28.59	0.935	0.894	0.813
SwinIR	**34.42**	**31.78**	**28.56**	**0.936**	**0.894**	**0.812**
DRUNet	34.30	31.69	28.51	0.934	0.893	0.810

Kodak 資料集：

表格 24　Kodak 資料集上模型對於多模型結合的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
Ours	35.06	32.67	29.64	0.927	0.888	0.812
Restormer	35.44	**33.02**	**30**	0.93	**0.894**	**0.823**
SwinIR	**35.46**	33	29.93	**0.93**	0.893	0.822
DRUNet	35.31	32.89	29.87	0.929	0.892	0.821

McMaster 資料集：

表格 25　McMaster 資料集上模型對於多模型結合的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
Ours	35.17	32.93	29.88	0.932	0.902	0.843
Restormer	35.55	33.31	**30.29**	0.934	0.906	**0.852**
SwinIR	**35.61**	**33.31**	30.2	**0.935**	**0.906**	0.849
DRUNet	35.41	33.15	30.08	0.932	0.903	0.846

Urban100 資料集：

表格 26　McMaster 資料集上模型對於多模型結合的影響

	PSNR			SSIM		
	σ =15	σ =25	σ =50	σ =15	σ =25	σ =50
Ours	34.73	32.54	29.55	0.950	0.928	0.881
Restormer	35.06	32.91	30.02	0.934	0.906	0.852
SwinIR	**35.22**	**33.06**	**30.06**	**0.935**	**0.906**	0.849
DRUNet	34.83	32.61	29.61	0.932	0.903	0.846

4.6 模型計算複雜度和參數量

模型使用的全部參數按照 1.20 模型和訓練設定。實驗跑在 256*256 大小的圖上，實驗記錄計算複雜度(FLOPs)和參數量(Params)。改動注意力不影響參數量，所以其值都一樣。

自變數為區塊自注意力中區塊的大小，也就是視窗大小、補丁、步長設定共同的數值大小，從 4 以 2 倍遞增至 16。

表格 27 計算複雜度對於區塊的大小

	FLOPs(G)	Params(M)
4	784.50	26.1117
8	325.31	26.1117
16	201.91	26.1117
32	167.04	26.1117

自變數為區塊自注意力中區塊比例的大小，從記憶體可負擔的區塊和步長比例 16:16 一直以 2 倍遞減至 2:16。

表格 28 計算複雜度對於區塊比例的大小

	FLOPs(G)	Params(M)
2:16	142.54	26.1117
4:16	144.70	26.1117
8:16	153.32	26.1117
16:16	201.91	26.1117

自變數為轉換器模組中的自注意力模組不同，有修復器[10]的通道(Channel-Wise, CW)自注意力(等同於 Restormer)、本研究的區塊(Patch-Wise, PW)自注意力、SwinIR[8]的基於視窗(Window-Based, WB)自注意力和不使用(None)注意力直接將 QKV 特徵相加。

表格 29 計算複雜度對於不同注意力

	FLOPs(G)	Params(M)
Restormer	220.3	26.1117
PW	784.88	26.1117
SwinIR	156.2	26.1117
None	142.38	26.1117

4.7 實驗總結

本研究透過在合成高斯雜訊與真實影像去雜訊資料集上的一系列實驗，對所提出的區塊自注意力（Patch-Wise Self-Attention, PW）機制進行了全面的評估。實驗結果可歸納為以下幾點核心發現：

4. 區塊自注意力（PW）的性能優越性：在合成雜訊去雜訊任務中，與通道自注意力（CW）、基於視窗的自注意力（WB）及不使用注意力的模型相比，本研究提出的區塊自注意力在所有測試資料集（CBSD68, Kodak, McMaster, Urban100）及不同雜訊水平下，均取得了最高的 PSNR 與 SSIM 指標（1.22.3 不同自注意力的影響）。這表明 PW 在捕捉影像復原關鍵特徵方面具有顯著優勢。
5. 模型組態對性能與效率的影響：
   區塊大小：實驗顯示，較小的區塊尺寸（如 4x4）能帶來更佳的去雜訊性能，但同時也導致計算複雜度（FLOPs）與圖片成平方增加（4.5.1 區塊自注意力中區塊大小的影響）。這顯示了模型性能與運算效率之間的權衡關係。
   區塊與步長比例：一個值得注意的現象是，採用稀疏採樣的區塊/步長比例（如 2:16）時，模型性能普遍優於密集採樣（16:16），且計算量更低（4.5.2 區塊自注意力中區塊和步長比例大小的影響）。此結果顯示對特徵圖進行選擇性的注意力計算可能是一種更高效的策略。
6. 不同任務下的最佳模型架構：
   在合成雜訊任務中，單獨使用區塊自注意力（PW）的模型表現最佳，優於任何模型結合方式（4.5.4 多模型結合的影響）。
   在更複雜的真實雜訊任務（SIDD）中，採用「通道注意力（CW）在前，區塊注意力（PW）在後」的串接（Connection）混合模型，取得了最佳的性能表現（PSNR 39.7475 dB），超越了單獨使用任一注意力的模型（4.4.2 多模型結合的影響）。
7. 基準性能比較：本研究的最佳模型在 SIDD 資料集上，使用 MATLAB 評估取得了 39.75 dB 的 PSNR，展現了具備競爭力的性能。質化分析進一步表明，本模型在紋理細節(圖 14)重建上優於修復器，而修復器則在平滑區域(圖 15)的處理上更具優勢。
8. 自注意力的動態選擇：本研究的在 1.21.4 區塊大小的自注意力不變性發現，自注意力具有空間泛化能力，自注意力圖能動態的對特徵向量值 Value 做加權計算，此結果為融合自注意力來獲取不同的歸納偏差帶來想法。

第五章 結論

本研究成功開發並驗證了一種新穎的區塊自注意力（Patch-Wise Self-Attention, PW）機制，並闡明其在影像去雜訊任務中的有效性與潛力。本研究的核心貢獻在於證明了區塊自注意力能夠引入一種強大的空間結構化歸納偏差。區塊注意力設計確保每個區塊注意力模組關注區塊間的空間關係，並保持注意力全局上下文建模，這種區塊注意力促進了空間結構化的歸納偏差，這對於局部紋理細節和結構一致性至關重要的任務非常有用。

不同於僅關注通道間關係的通道自注意力（CW），只關注特徵通道間的特徵關係，依賴特徵特性與統計的一致性。區塊自注意力直接對空間區塊之間的依賴性進行建模，使其在重建影像的局部紋理與結構方面表現卓越。這一點從本模型（PW）在紋理細節上優於 Restormer 的質化比較中，還有在高雜訊情況（σ=50）下，得到了充分驗證。這也揭示了不同注意力機制在處理不同影像內容（紋理 vs. 平滑區域）時的內在互補性。

實驗結果進一步表明，雖然單純的 PW 機制在合成雜訊任務中已足夠強大，但面對更具挑戰性（可以認為需要假設雜訊本身不具有空間結構，spatial correlation(near pixel)=0）的真實世界雜訊時，結合不同注意力的混合架構是更優的解決方案。將 CW 作為高效的全局特徵提取器，再由 PW 進行精細的空間結構建模，這種串接設計充分發揮了兩者的優勢，是應對複雜影像退化問題的有效途徑。

模型限制上，本研究也存在一些局限性。首先，區塊自注意力相較於其他機制（如 CW、WB）具有更高的計算複雜度，這在資源受限的應用中可能成為挑戰。

未來研究上，方向可包含：(1) 開發更高效的區塊注意力實現方式，以降低計算成本；(2) 實驗中觀察到的「稀疏採樣反而帶來更佳性能」的現象，深入研究稀疏注意力採樣的理論基礎，並將其發展為一種通用的模型優化策略；(3) 設計更先進的動態混合各種注意力架構，使其能根據輸入內容自適應地調配不同注意力的比重。

總體而言，本研究為基於轉換器 Transformer 的影像復原領域提供了一個有效的設計思路與堅實的實驗基礎。

參考文獻

[1] R. C. Gonzalez and R. E. Woods, *Digital image processing*, Fourth edition, Global edition. New York: Pearson, 2017.

[2] A. Buades, B. Coll, and J.-M. Morel, "A Non-Local Algorithm for Image Denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA: IEEE, 2005, pp. 60–65. doi: 10.1109/CVPR.2005.38.

[3] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007, doi: 10.1109/TIP.2007.901238.

[4] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017, doi: 10.1109/TIP.2017.2662206.

[5] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a Fast and Flexible Solution for CNN based Image Denoising," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4608–4622, Sep. 2018, doi: 10.1109/TIP.2018.2839891.

[6] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," Jul. 10, 2017, *arXiv*: arXiv:1707.02921. doi: 10.48550/arXiv.1707.02921.

[7] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 03, 2021, *arXiv*: arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929.

[8] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," Aug. 17, 2021, *arXiv*: arXiv:2103.14030. doi: 10.48550/arXiv.2103.14030.

[9] J. Liang, J. Cao, G. Sun, K. Zhang, L. V. Gool, and R. Timofte, "SwinIR: Image Restoration Using Swin Transformer," Aug. 23, 2021, *arXiv*: arXiv:2108.10257. doi: 10.48550/arXiv.2108.10257.

[10] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient Transformer for High-Resolution Image Restoration," Mar. 11, 2022, *arXiv*: arXiv:2111.09881. doi: 10.48550/arXiv.2111.09881.

[11] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks," Apr. 03, 2018, *arXiv*: arXiv:1711.07064. doi: 10.48550/arXiv.1711.07064.

[12] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," May 28, 2020, *arXiv*: arXiv:2005.12872. doi: 10.48550/arXiv.2005.12872.

[13] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," Oct. 28, 2021, *arXiv*: arXiv:2105.15203. doi: 10.48550/arXiv.2105.15203.

[14] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 02, 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

[15] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," Mar. 02, 2015, *arXiv*: arXiv:1502.03167. doi: 10.48550/arXiv.1502.03167.

[16] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," Jul. 21, 2016, *arXiv*: arXiv:1607.06450. doi: 10.48550/arXiv.1607.06450.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 10, 2015, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 18, 2015, *arXiv*: arXiv:1505.04597. doi: 10.48550/arXiv.1505.04597.

[19] J. Gurrola-Ramos, O. Dalmau, and T. E. Alarcon, "A Residual Dense U-Net Neural Network for Image Denoising," *IEEE Access*, vol. 9, pp. 31742–31754, 2021, doi: 10.1109/ACCESS.2021.3061062.

[20]O. Oktay *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas," May 20, 2018, *arXiv*: arXiv:1804.03999. doi: 10.48550/arXiv.1804.03999.

[21] R. Timofte *et al.*, "NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA: IEEE, Jul. 2017, pp. 1110–1121. doi: 10.1109/CVPRW.2017.149.

[22] R. W. Tucker and T. J. Walton, "On Gravitational Chirality as the Genesis of Astrophysical Jets," *Class. Quantum Gravity*, vol. 34, no. 3, p. 035005, Feb. 2017, doi: 10.1088/1361-6382/aa5325.

[23] S. Arfaoui and A. B. Mabrouk, "Some Generalized Clifford-Jacobi Polynomials and Associated Spheroidal Wavelets," Apr. 06, 2017, *arXiv*: arXiv:1704.03513. doi: 10.48550/arXiv.1704.03513.

[24] C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," Jul. 31, 2015, *arXiv*: arXiv:1501.00092. doi: 10.48550/arXiv.1501.00092.

[25] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," Jan. 04, 2019, *arXiv*: arXiv:1711.05101. doi: 10.48550/arXiv.1711.05101.

[26] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," May 03, 2017, *arXiv*: arXiv:1608.03983. doi: 10.48550/arXiv.1608.03983.

[27] *https://paperswithcode.com/*.

[28] C. Mou, J. Zhang, and Z. Wu, "Dynamic Attentive Graph Learning for Image Restoration," Sep. 14, 2021, *arXiv*: arXiv:2109.06620. doi: 10.48550/arXiv.2109.06620.

[29] Z. Yue, Q. Zhao, L. Zhang, and D. Meng, "Dual Adversarial Network: Toward Real-world Noise Removal and Noise Generation," Jul. 12, 2020, *arXiv*: arXiv:2007.05946. doi: 10.48550/arXiv.2007.05946.

[30] M. Chang, Q. Li, H. Feng, and Z. Xu, "Spatial-Adaptive Network for Single Image Denoising," Jul. 14, 2020, *arXiv*: arXiv:2001.10291. doi: 10.48550/arXiv.2001.10291.