Decoupling-NeRF

Decompose the scene and renderer in NeRF

黃仁鴻

Outline

- Introduction
- Method
- Experiment
- Conclusion

Introduction

於 2020 年提出的神經輻射場(Neural Radiance Field, NeRF) 利用簡單的類神經網路結構來擬合 Volume Rendering 的 3D 模型。 但 NeRF 的設計會將 Renderer 與 Scene 嵌入於同一個類神經網路中。 導致兩者具有高度耦合性而無法拆分。 因此需要更換場景時，NeRF就需要重新進行訓練。

Introduction

然而，在一般 3D 場景的儲存與展示都是將 Scene 及 Renderer 拆分開來，並將 Scene 作為輸入以取得對應視角的照片。 這樣一來，Renderer 的部分就能重複利用於不同的 3D 場景上。

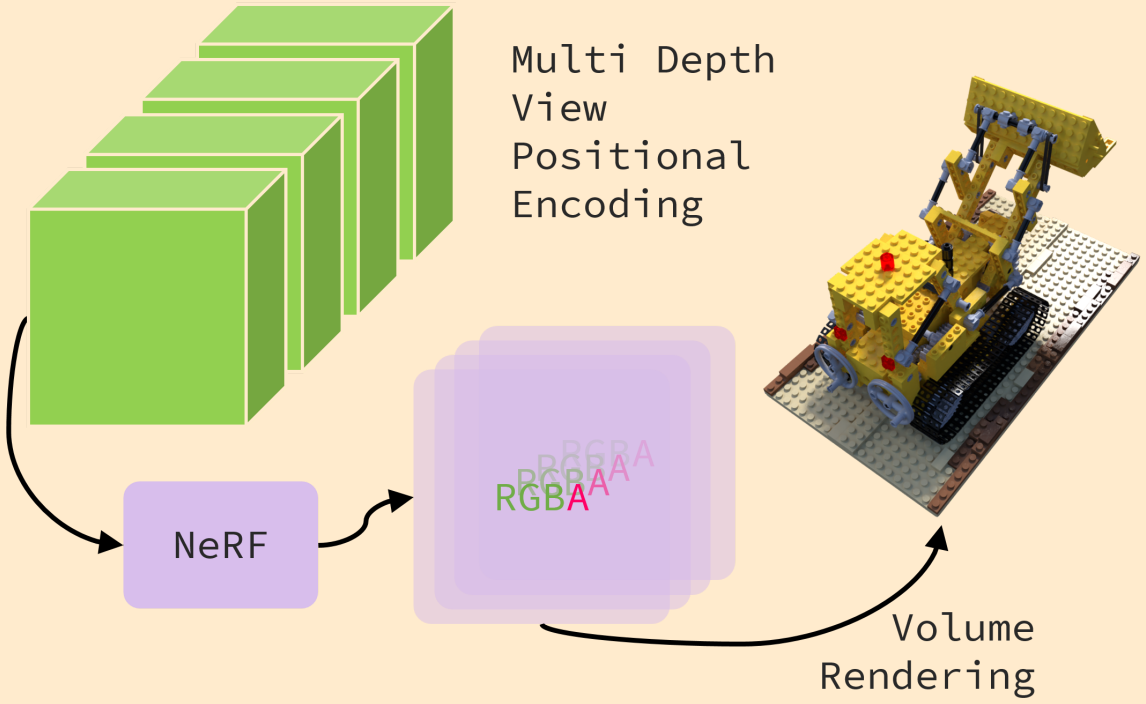
Introduction

本次專題研究提出的 Decoupling NeRF， 便是利用 Scene Encoding Block 將照片編碼為場景特徵後，與目標視角一同輸入至 Renderer Block 生成目標圖片， 藉此讓 NeRF 能快速應用在各種場景而不需要重新擬合。

NeRF

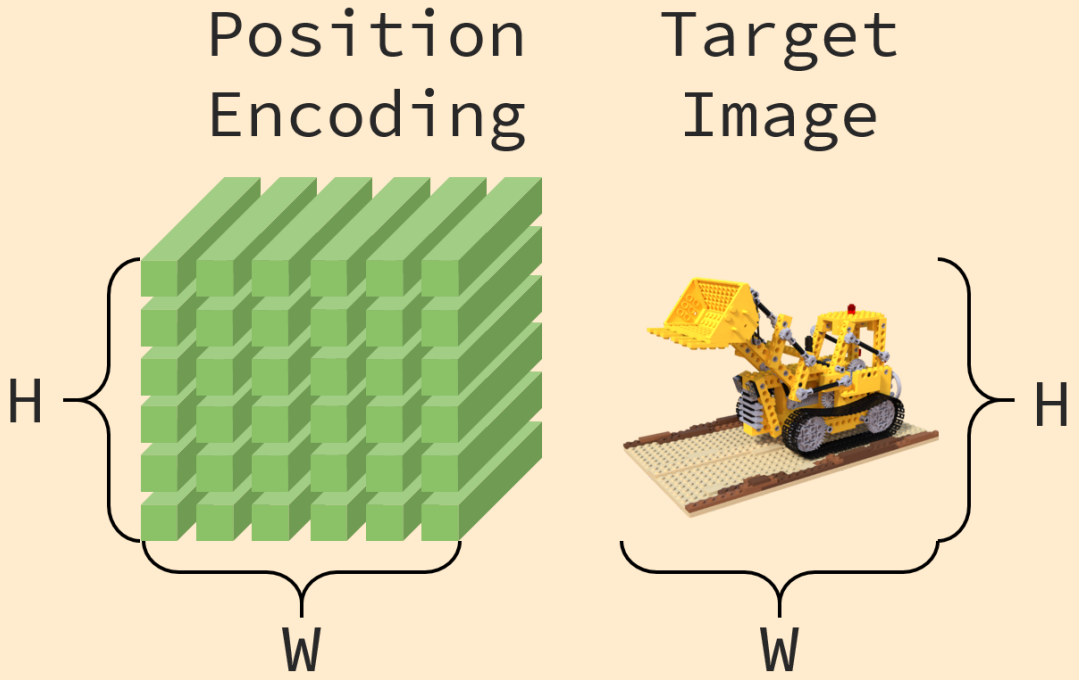
NeRF 在算繪時，會將每個體素的 Positional Encoding 輸入進類神經網路中，並獲取對應的 RGBA。

在利用體素與鏡頭的距離跟得到的 RGBA 值，合成出最終的視圖。



Positional Encoding

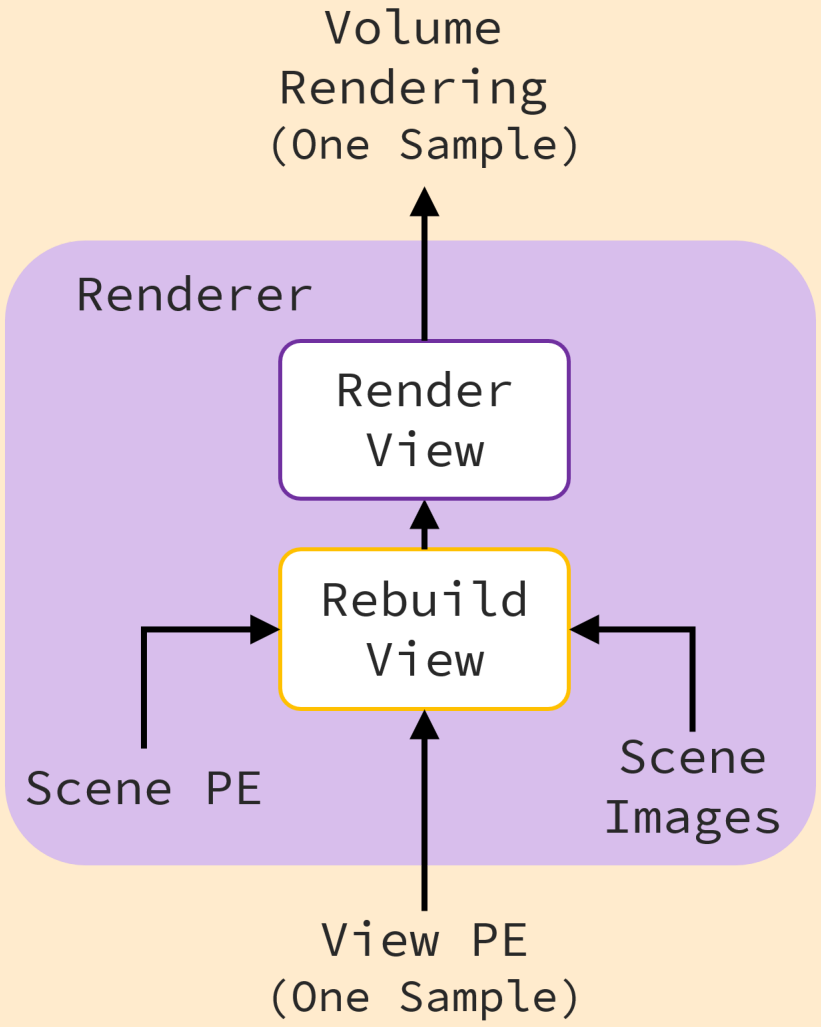
先利用相機姿態矩陣計算出視圖中每個像素的座標，再將座標通過下式轉換為 Positional Encoding。



$$PE(p) = Concat\begin{pmatrix} sin(2^0\pi p), cos(2^0\pi p), \dots, \\ sin(2^{L-1}\pi p), cos(2^{L-1}\pi p) \end{pmatrix}$$

$$p = (x, y, z)$$

NeRF is content coupling



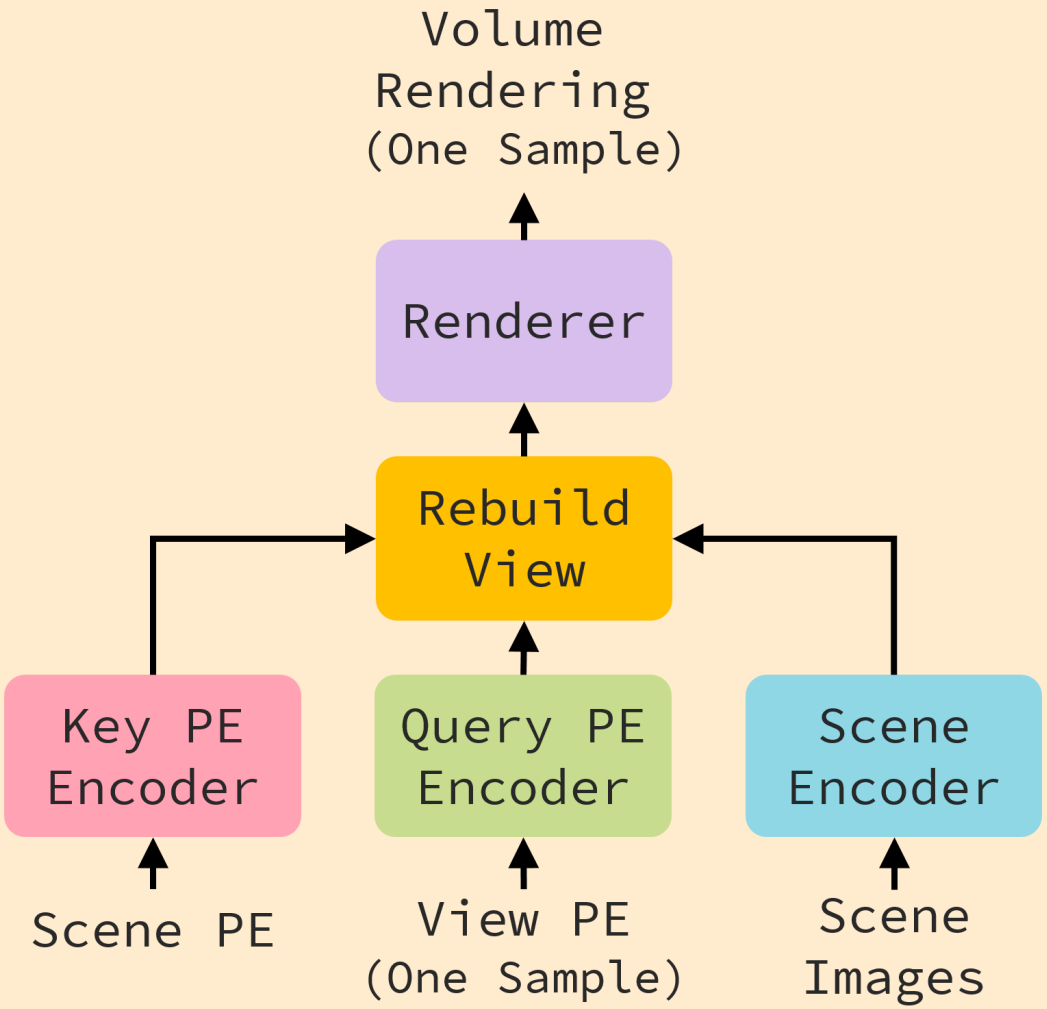
NeRF 的做法會隱式的將場景資訊記錄到類神經網路中。再藉由輸入的位置編碼 (Positional Encoding, PE) 提取出該點的資訊。

因此，每個 NeRF 模型只會紀錄一種場景而且無法輕易切換。

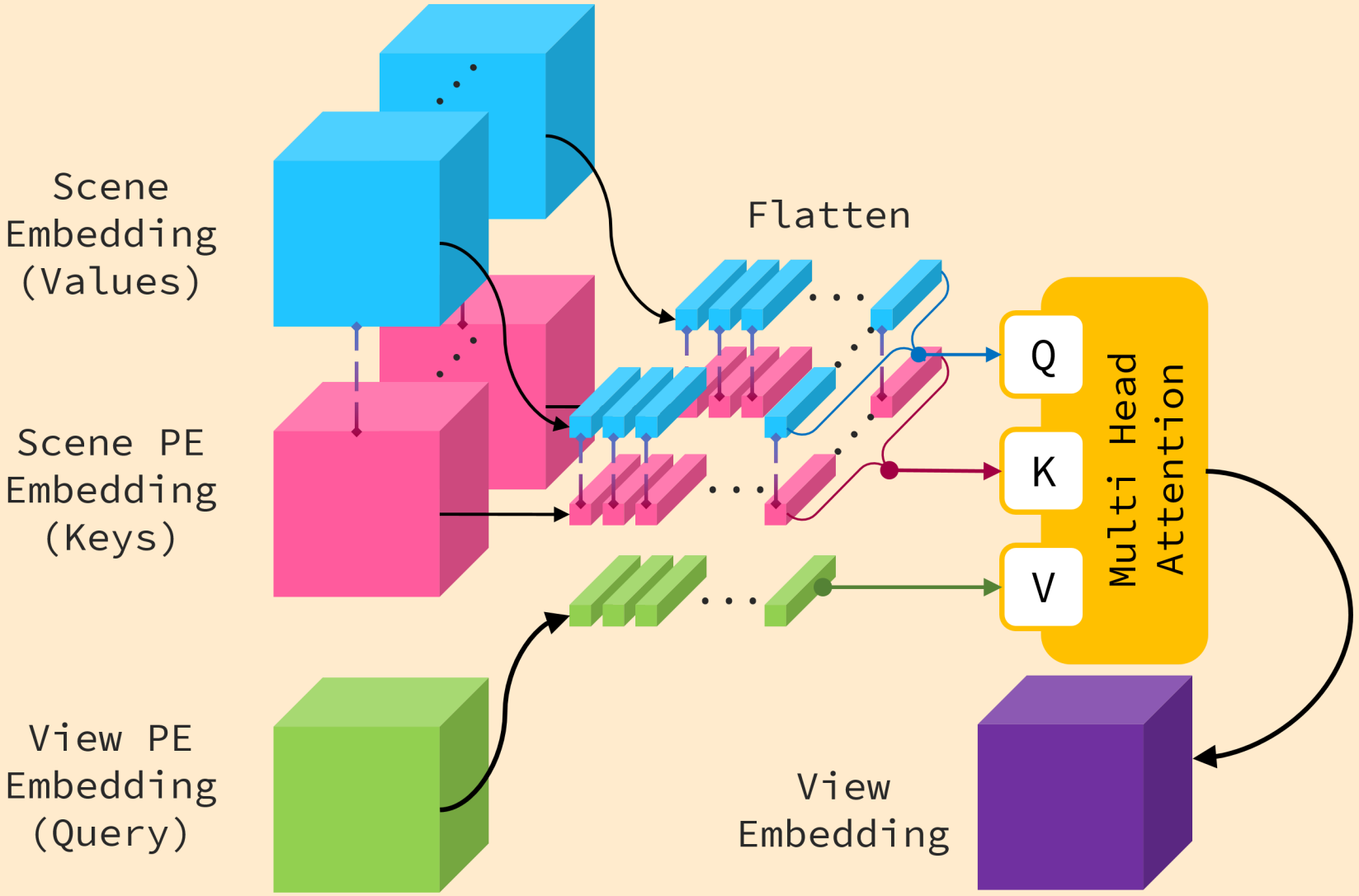
Decoupling NeRF

本專題提出的 Decoupling NeRF 便是將場景資訊、場景彩現拆分開。

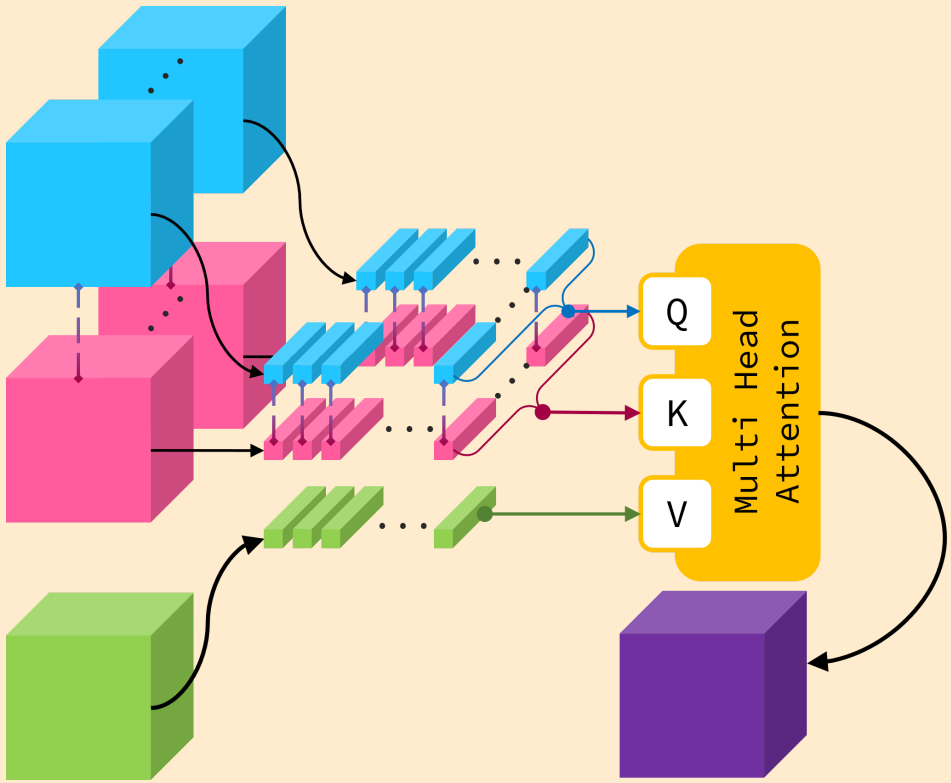
在 Rebuild View 中會對場景編碼重新組織，再交由 Renderer Block 輸出畫面。



Rebuild View



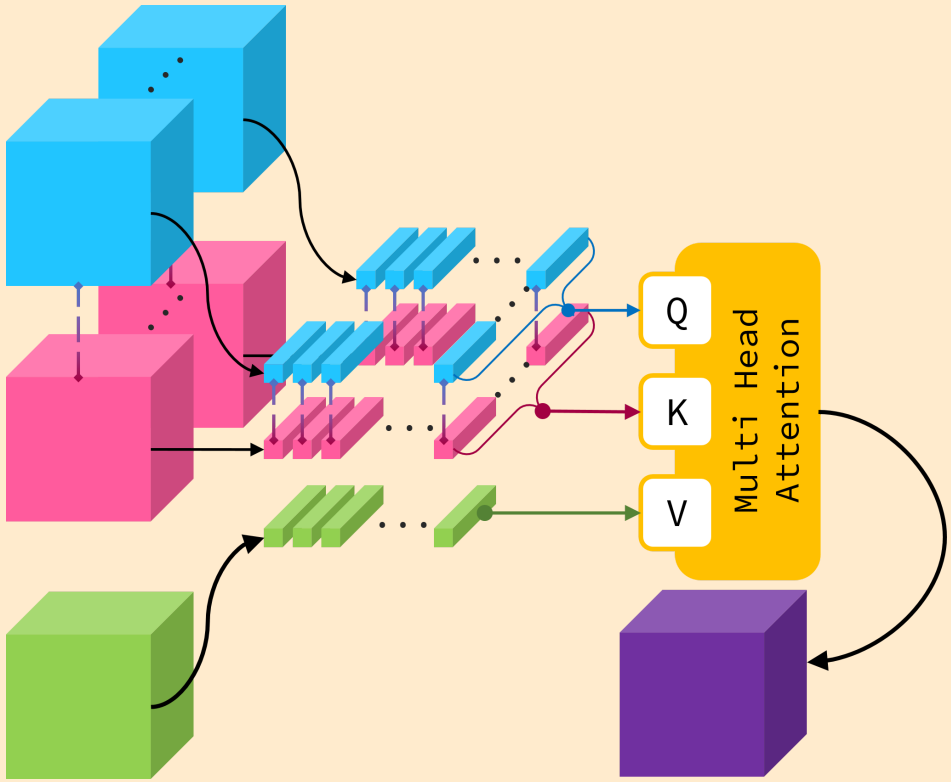
Rebuild View



將 View PE、Scene PE 與 Scene Image 分別編碼成 Query、Key 及 Value。

並利用 Multihead Attention 的方法把 Value(Scene Embedding) 合成為目標視角的場景結構編碼。

Rebuild View



而重構的編碼不直接包含視角資訊，因此可以避免類神經網路直接學習將 PE 映射成 3D 模型，使其強制從場景編碼生成目標景象。

Multihead Attention

$$MHA(Q,K,V) = Concat(head_1, \dots, head_h)W^O$$

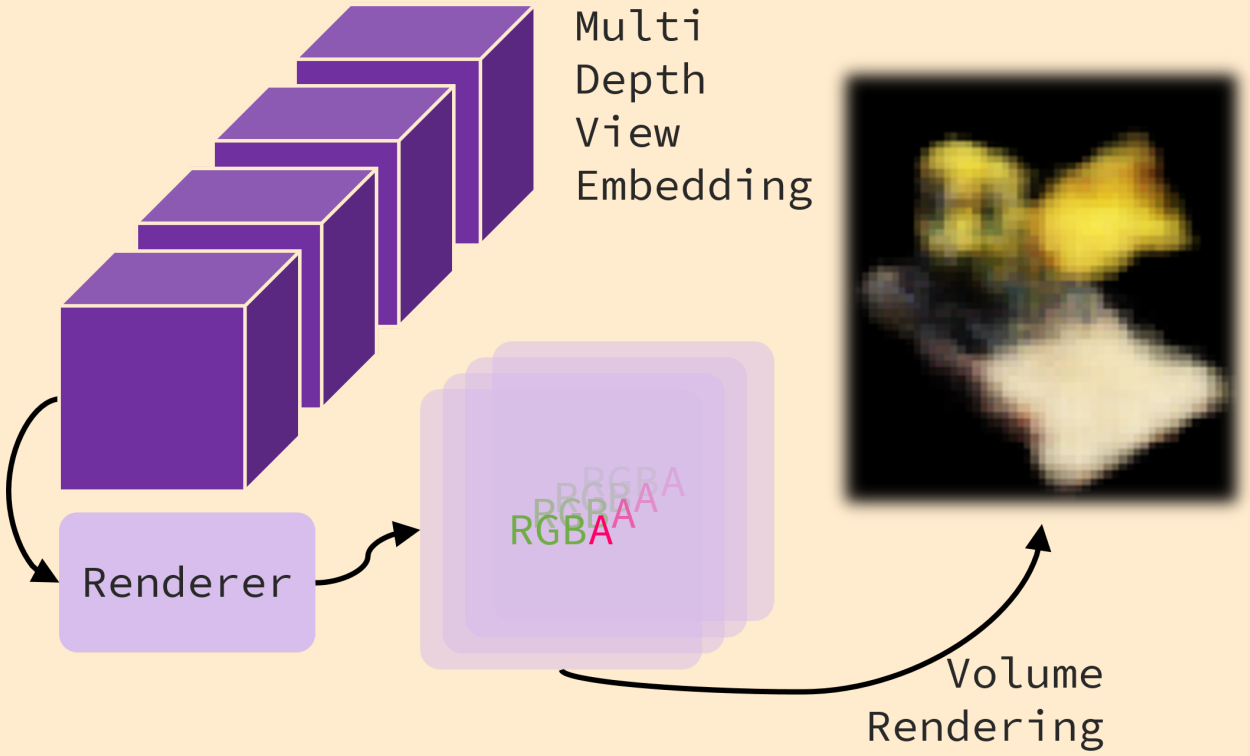
$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

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

Multihead Attention 是於 2017 年由 Vaswani et al 所提出的，最早是用在 NLP 任務上，近年也開始在 CV 領域上流行。

Volume Rendering

Renderer Block 將不同深度的 View Embedding 轉換成多個 RGBA 圖層後，會再利用 Volume Rendering 將其合成為單一影像。



Loss

Target Pred

SSIM( , )

本次專案使用結構相似性指標 (Structural SIMilarity, SSIM) 作為損失函數。

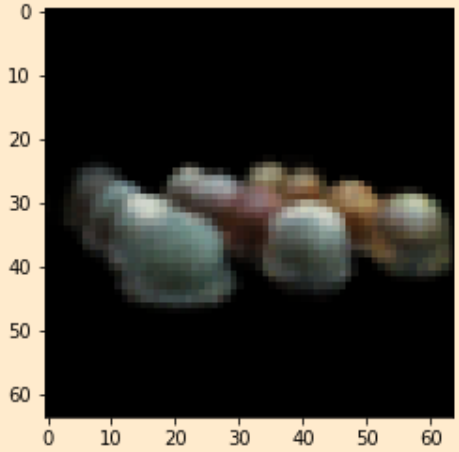
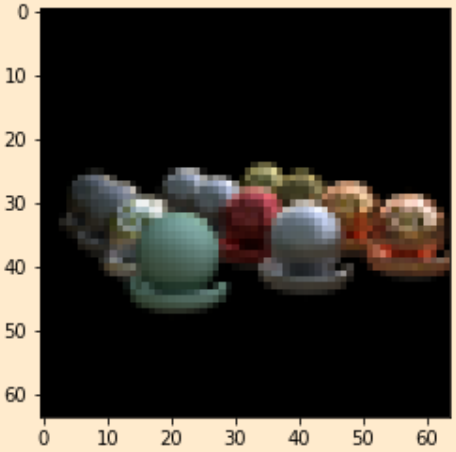
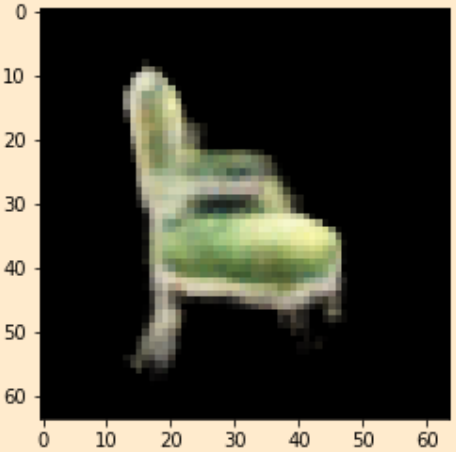
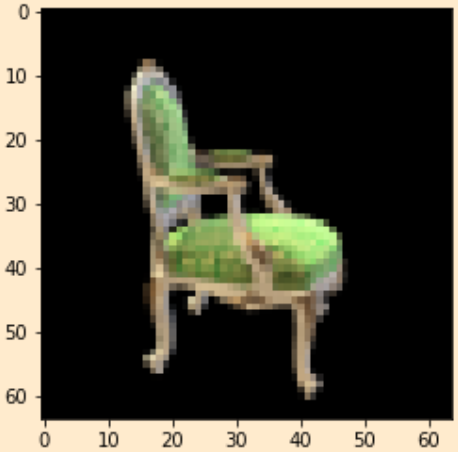
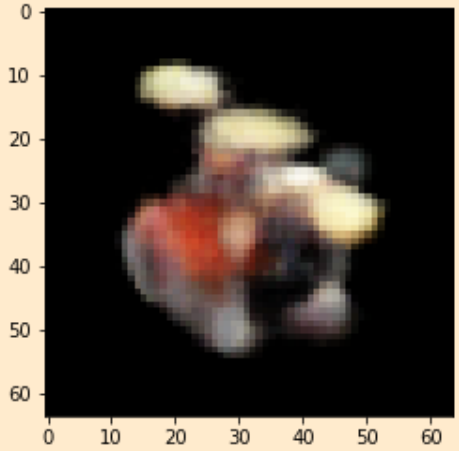
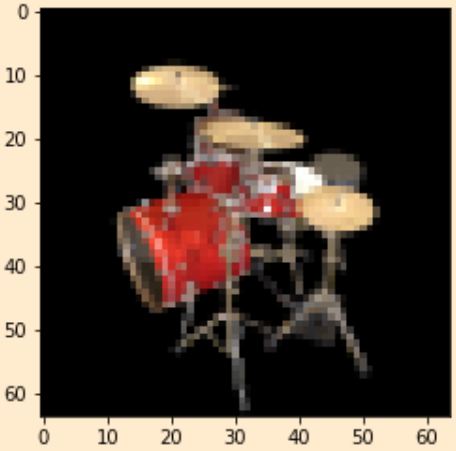
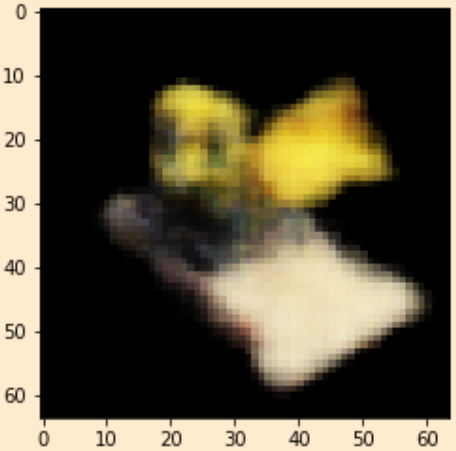
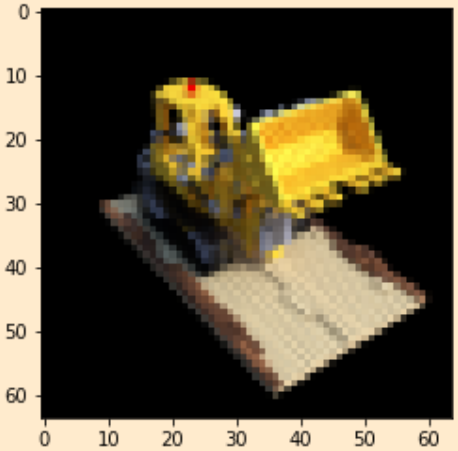
Data

nerf synthetic	
Train Object	chair, drums, ficus, hotdog, lego, materials
Test Object	mic, ship
Size	resize to 64x64 pixels

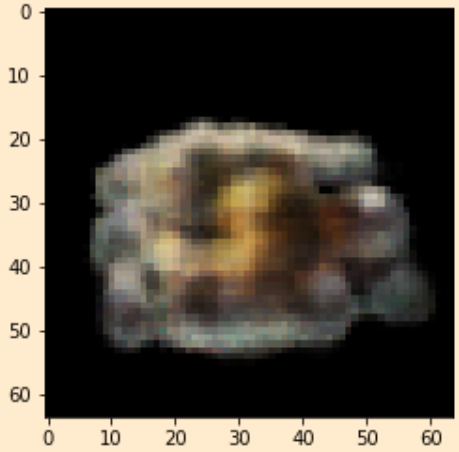
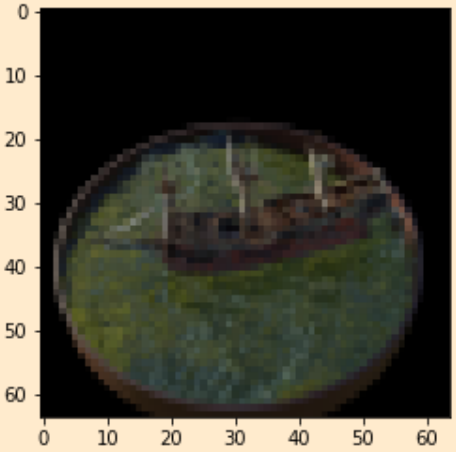
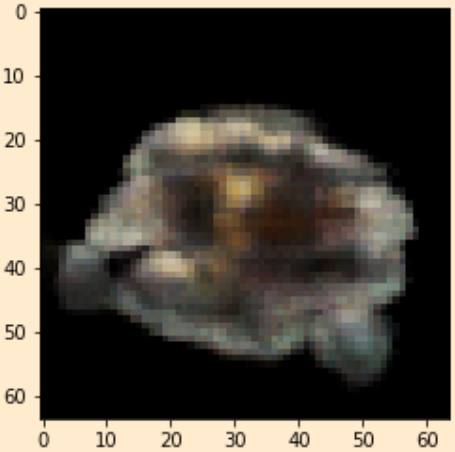
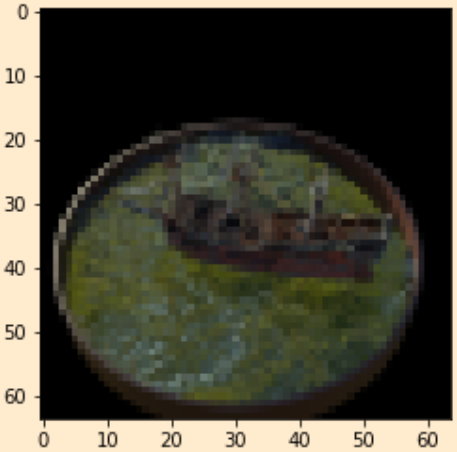
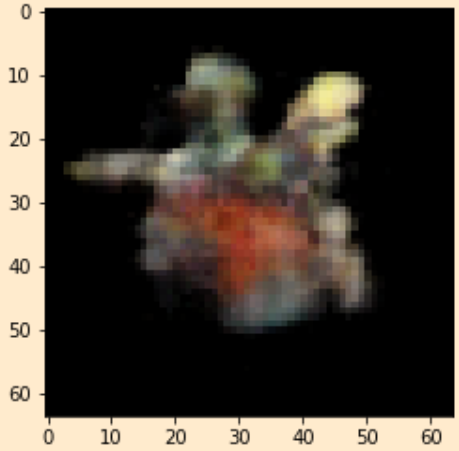
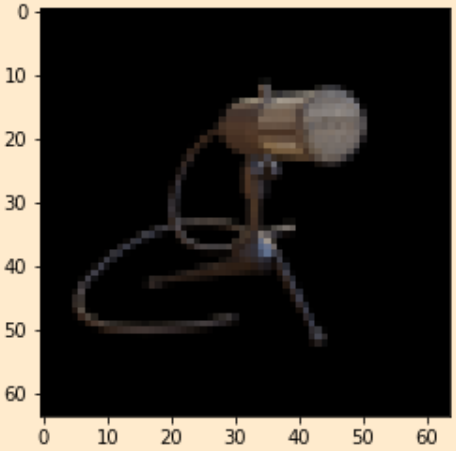
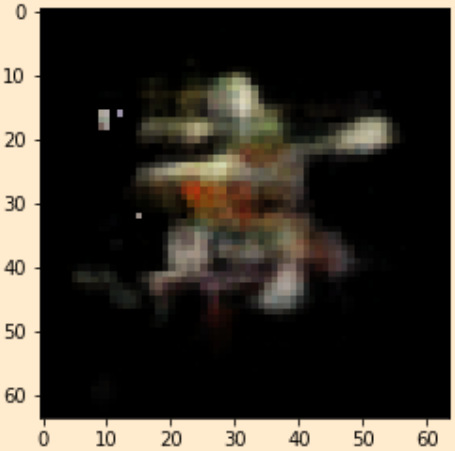
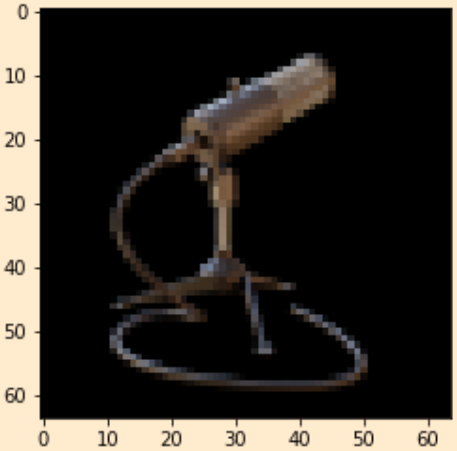
Hyperparameter

Optimizer:Adamax	lr	betas
	0.002	(0.9, 0.999)
Batch Size	4x8 (基本 batch size 為 4，但會累計 8 個 batch 才更新一次權重)	
PE：L_{embed}	12	
dim	256	
Multihead Attention	head number	head dim
	32	32
Down/Upsample	8	

Train Object

Pred	Target	Pred	Target
			
			

Test Object

Pred	Target	Pred	Target
			
			

Conclusion

- 目前只有在訓練時看過的物件才能成功重建，而沒有看過的物件在重建後會被扭曲成看過的物件。
- 推測原因有可能是在 Multihead Attention 時，類神經網路將位置資訊混進輸出中。導致模型退化成 NeRF + Object Condition，使其缺乏泛化能力。

Todo

- 研究如何能確實泛化到沒看過的物件。
- 使用 VQVAE 作為 Encoder 與 Renderer。
- 嘗試在特徵層級就使用 Contrastive Learning 進行約束。

Reference

- Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020.

- Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.