

Decoupling-NeRF: Decompose the scene and renderer in NeRF

黃仁鴻
P76094169
NCKU CSIE

1. Introduction

於 2020 年提出的神經輻射場(Neural Radiance Field, NeRF) [1]利用簡單的類神經網路結構來擬合 Volume Rendering 的 3D 模型。但 NeRF 的設計如 Fig. 1，相當於是將 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。因此每當需要更換場景時，NeRF 就需要重新進行訓練。與一般我們在深度學習中訓練完成後，即可套用於不同場景中的方法有所差別。

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。對應於原本 NeRF 中，訓練所使用的照片便相當於嵌入在 NeRF 內的場景，若可以將照片改用於模型的輸入，便可將 Scene 與 Renderer 解耦合。

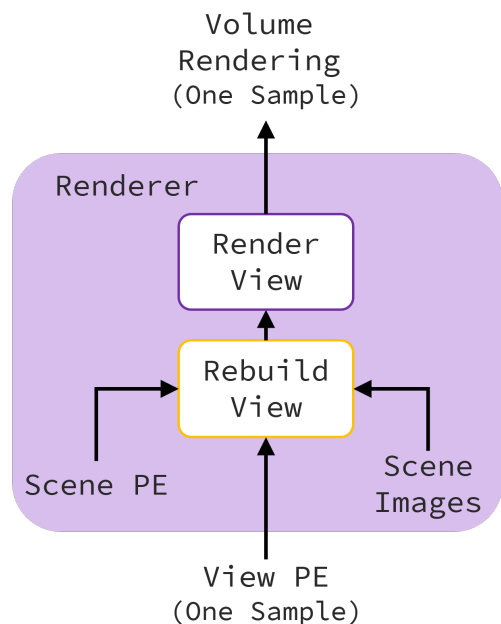


Figure 1. NeRF 的設計使 Scene 與 Renderer 具有高度耦合性，使 Scene 內嵌於模型當中無法分離。

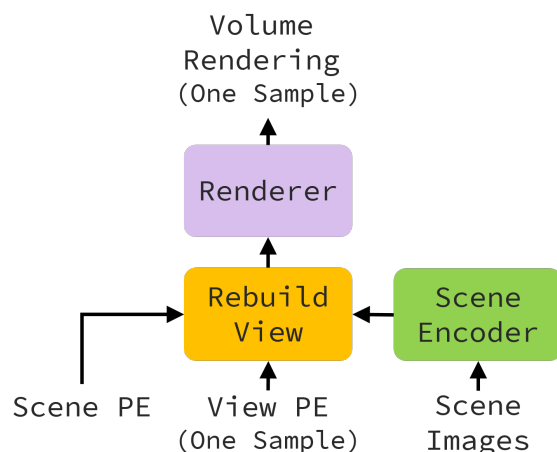


Figure 2. 本次專題所提出的 Decoupling-NeRF 就是希望將 Scene 資訊與 Rebuild View 獨立出來。當要替換繪製場景時就不需要再經過耗時的訓練，只需要變更輸入的 Scene Images 跟 Scene Position Encoding 即可。

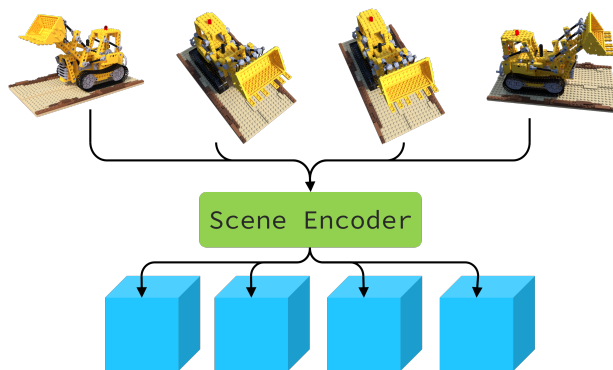


Figure 3. 將場景照片透過 Encoder 編碼成 Scene Embedding。

因此，本次專題研究目標便是提出 Decoupling-NeRF 這個架構，使其可以快速應用在各種場景而不需要重新擬合。Fig. 2 是本次專題的架構，將隱含於 NeRF 中的 Renderer 與 Rebuild 分開，並利用 Scene Encoder 對場景照片進行編碼，在 Rebuild View 使用 Multi

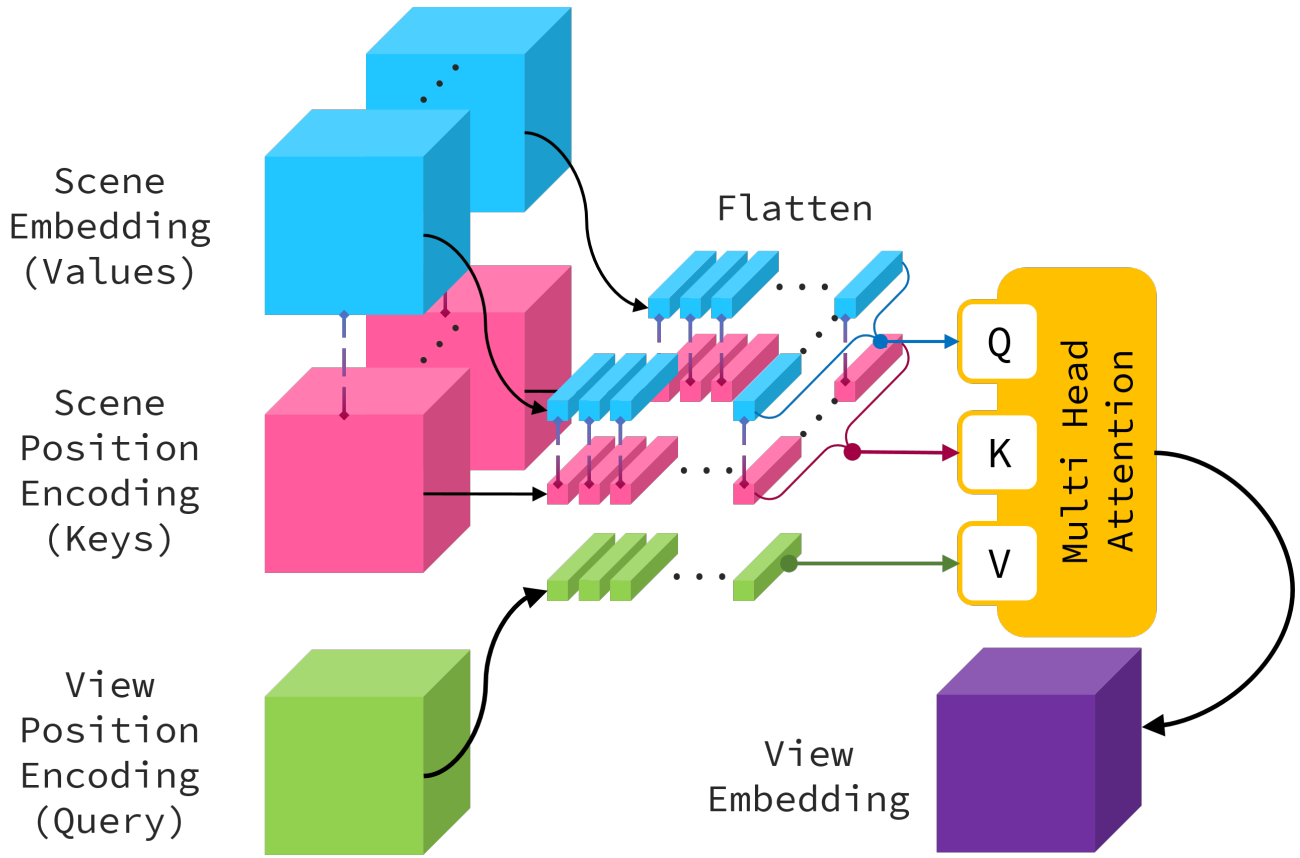


Figure 4. 將目標視角的位置編碼作為 Querys，與 Scene Embedding 及其對應的位置編碼計算 Multi Head Attention 來取得 View Embedding。

Head Attention [2] 將其重新組建為 View Embedding，最後透過單一 Neural Renderer 生成場景照片。

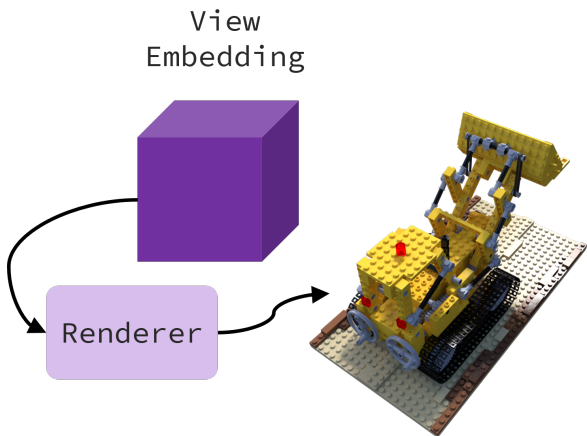


Figure 5. 將 Fig. 4 得到的 View Embedding 交由 Renderer 進行 Volume Rendering。

2. System framework

在原版的 NeRF 中，會使用目標視角的位置編碼(Position Encoding, PE)作為輸入，以生成該視角會拍攝到場景照片，而 Position Encoding 是將相機的位置資訊 Eq. 1 轉換所得:

$$PE(p) = [\sin(2^0 \pi p), \cos(2^0 \pi p), \dots, \sin(2^{L-1} \pi p), \cos(2^{L-1} \pi p)] \quad (1)$$

本專題架構如 Fig. 2 所示，可以分成 Scene Encoder、Rebuild View 與 Renderer 三個區塊。在進行照片生成時，先將多張場景照片編碼成 Scene Embedding(Fig. 3)，再配合對應的 Scene PE 及生成目標的 View PE，將這三者以 Fig. 4 的形式建立出 View Embedding，最後再交由 Renderer 產生視圖(Fig. 5)。

在 Fig. 4 中，利用了 Vaswani et al. [2] 所提出的 Multi Head Attention 來將 Scene Embedding 重新組織成需要

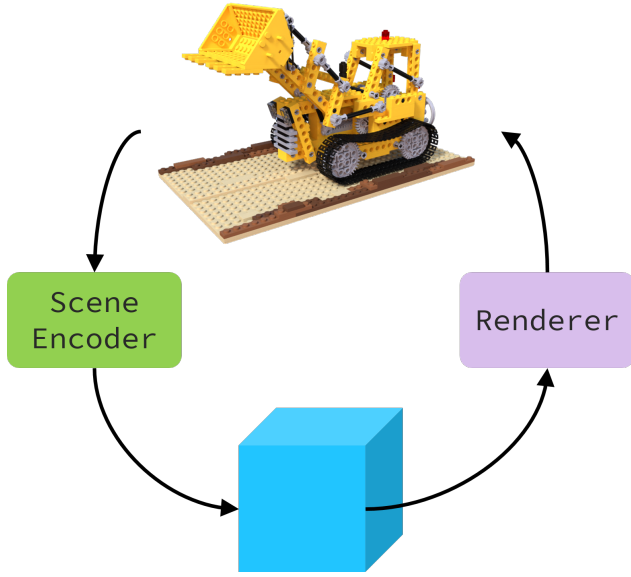


Figure 6. 預先訓練 Auto Encoder。

的 View Embedding。其計算方法如 Eq. 2:

$$\begin{aligned} MHA(Q, K, V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \\ \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \\ \text{Attention}(Q, K, V) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \end{aligned} \quad (2)$$

在訓練階段則可分為兩個步驟，第一步會把架構中的 Scene Encoder 及 Renderer 以 Auto Encoder 的形式進行預訓練(如 Fig. 6)。而步驟二不僅要訓練 Rebuild View，還會讓 Renderer 再次學習。

3. Expected results

本專題將會使用 NeRF 提供的資料集¹進行訓練與測試，其中提供了 8 種場景與 8 個物件的多張照片及拍攝視角，並已區分成訓練、驗證跟測試集。

在將模型訓練完成後，會測試不同數量的場景照片對模型生成能力所造成的影響。期望能將其製作成網頁應用，在計算能力有限的行動裝置下也能使用。

References

- [1] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.
- [2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefin dukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of*

¹Link https://drive.google.com/drive/folders/128yBriW1IG_3NJ5Rp7APSTZsJqdJdfc1