

3DCV Term Project Report

R09922074 潘奕廷 R09922031 黃子源 R09944030 高晟璋

I. Motivation

近年來，深度學習在電腦視覺領域的快速發展，使得許多問題透過深度學習的方法能得到比傳統方法更好的結果。隨著 2020 年於 ECCV 上發表的一篇 oral paper, NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis[8]，有許多基於此論文的衍伸研究相繼發表，使得 view synthesis 領域成為了近期熱門的一個主題之一。因此，我們組想將 view synthesis 的概念與近年另外一個熱門的主題 - Augmented Reality(AR) 結合，實作出一款有趣的特效應用，模擬知名系列電影-哈利波特魔法世界中，於畫框或是報紙中的人物能與真實世界的人互動的效果。



互動特效於《哈利波特》電影

II. Related Work

3D reconstruction

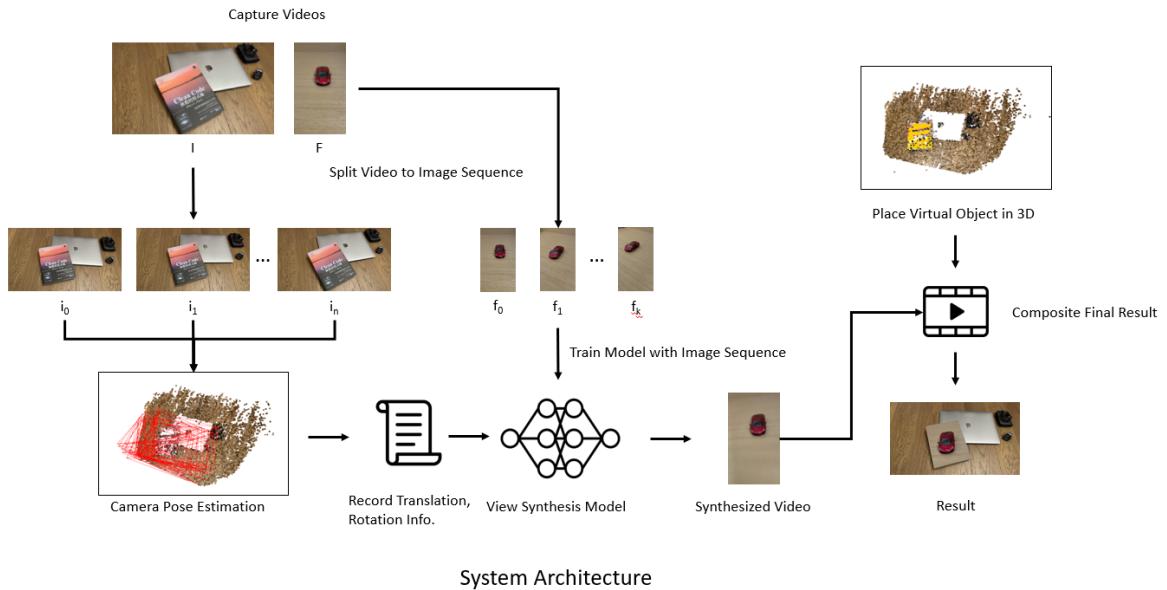
透過拍攝一個物體多個視角下的影像，在已知拍攝相機的 camera pose 前提下，我們能夠分析影像間點與點的對應關係，搭配 triangulation 的技巧去估算並還原物體在 3D 空間中的輪廓及位置；而在 camera pose 屬於未知的情況下，則大大增加了還原物體 3D 資訊的挑戰性。為解決上述問題，因此發展出了 structure from motion (SfM) 的技術，透過 SIFT 等特徵匹配演算法匹配影像間點與點的關係、分解 essential matrix 與 fundamental matrix 求得拍攝相機間的 R, t 關係，接著使用 triangulation 的方式求得 point cloud，再使用 bundle adjustment 的技術修正投影誤差，透過迭代的方式求得最佳解，一步步還原場景的 3D 結構。而已有許多套件實現了此套算法，能夠幫助我們快速得到不錯的 3D 結構資訊，例如 COLMAP[9]、OpenMVG[10] 等，使用者僅需提供場景於不同視角下的一系列拍攝影像，以及相機的內參資訊，這類套件便可快速地還原出場景的 point cloud、mesh 等資訊，大大加速了相關應用的開發。

Novel View Synthesis

輸入一個場景中不同視角下所拍攝的影像集，如何合成出一個不屬於此影像集中的新視角下所看到的影像是一个困擾研究者們的困難問題之一。傳統的作法透過影像間的匹配、光流的分析，以及影像內插(interpolation) 等方式，能處理在兩已知視角間小幅度移動的新視角下的影像，但對於影像集外的視角合成結果

則不甚理想。近年來，隨著深度學習方法的興起，StereoMag[2] 提出使用 multiplane images(MPIS) 的方式來表示場景，SynSin[7] 則是輸入單張影像，透過 differentiable point cloud renderer 與 generative module 來生成新視角下的影像。3D photography[1] 將影像拆解成顏色、深度、邊界三個子問題探討，並提出 layered depth image(LDI) 的方法將上述三個資訊整合保存，將輸入影像透過深度與邊界資訊合成出新視角下的影像，並透過 inpainting 的模型對新視角中有缺失的顏色、深度等資訊進行填補，進而得到優於傳統做法的結果。但上述做法仍然無法處理大幅度變換視角的情形，因此 NeRF[8] 提出使用 MLP 的方式保存場景資訊，輸入相機位置與視角即可合成出新視角下的影像，一定程度解決了不能大幅度移動的缺點。而 Nex[6] 則於 NeRF 的相法上進一步將場景的顏色拆解，改為使用 basis 組合的方式表示場景顏色資訊，改善 NeRF 僅能表示 diffuse 不能表示 reflection 等光影的缺點，也大幅減少訓練時所需的模型參數量，加速一個場景於 training 與 testing 所需的時間。

III. Methodology



我們將輸入的影片轉為 image sequence 來處理。為了要合成影片特效，需要拍攝兩段影片，分別模擬兩個世界的影像，我們將模擬真實世界的影像定義 I，模擬第二世界的影像定義為 F，轉為 image sequence 後的影像定義為 $i_0 \sim i_n$ 以及 $f_0 \sim f_k$ 。

我們的系統分成三個階段：1. 3D reconstruction 2. novel view synthesis 3. video composition。

於第一階段，我們使用 structure from motion(SfM) 的技術，透過輸入 $i_0 \sim i_n$ 以及相機內參資訊，我們建立拍攝場景的 point cloud，並還原拍攝影片時的 camera pose，接著我們將還原的 camera pose 中的 translation 以及 rotation 資訊額外輸出成 json 檔保存。第二階段我們將 $f_0 \sim f_k$ 作為輸入訓練我們的模型，待訓練完畢後，我們將第一階段生成的 json 檔作為輸入，來使我們的模型生成對應視角的 novel view result。第三階段我們首先利用第一階段生成的 point cloud 來建

立 3D 空間，接著於空間中加入一個可調整的長方體，用於決定合成第二世界影像的 3D 位置，最後利用已知的相機內外參建立每一個影像的投影矩陣，將長方體的位置投影回 2D 影像平面上，並將影像中屬於長方體範圍內的區域用第二階段產生的 novel view 所取代，合成出最終的特效影像。

IV. Implementation

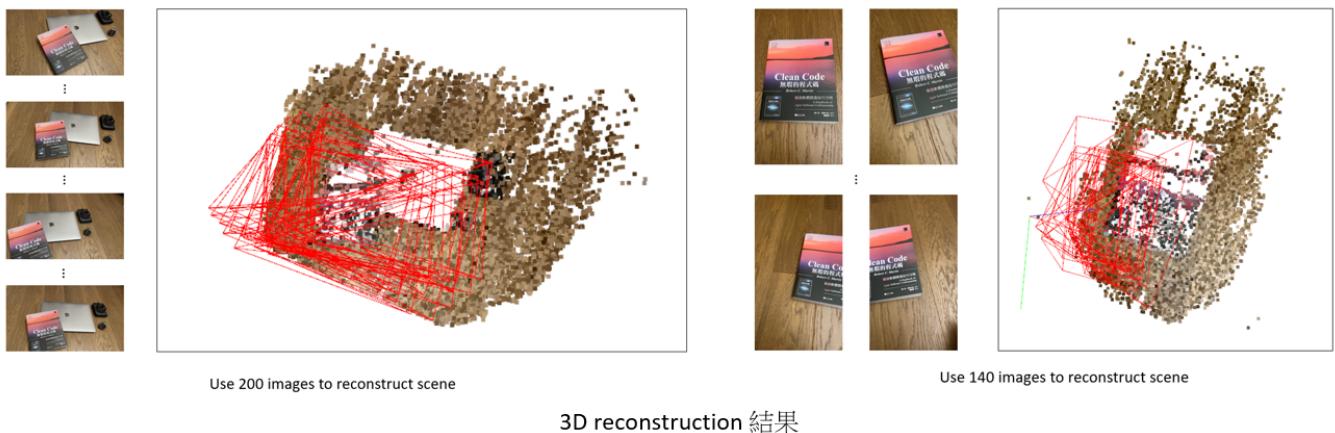
3D Reconstruction

我們使用 OpenCV Python 套件將所拍攝的影片轉為 image sequence 來處理，接著使用 OpenMVG[10] structure from motion 的功能還原拍攝場景的 point cloud 資訊，同時得到每一個影像的 R, t 資訊，利用此資訊，我們能生成每個影像的投影矩陣。值得注意的是，OpenMVG 計算得到 camera pose 並不是直接給出了相機在世界座標下的 rotation 與 translation，根據官方文件說明，其關係如下：

$$[R|t] = [R] - RC$$

因此在使用時還需經過轉換後才會得到正確的 pose。

在得到場景的 point cloud 與 camera pose 後，我們使用 Open3D 的套件建立 3D 場景，並於場景內擺入立方體用於調整合適的 AR 應用位置，在調整完後，最終於 video composition 階段計算立方體經過投影矩陣轉換後的 2D 位置，並將其範圍內的 pixel 改由第二世界的影像取代。完成 3D reconstruction 後的結果如下圖所示。



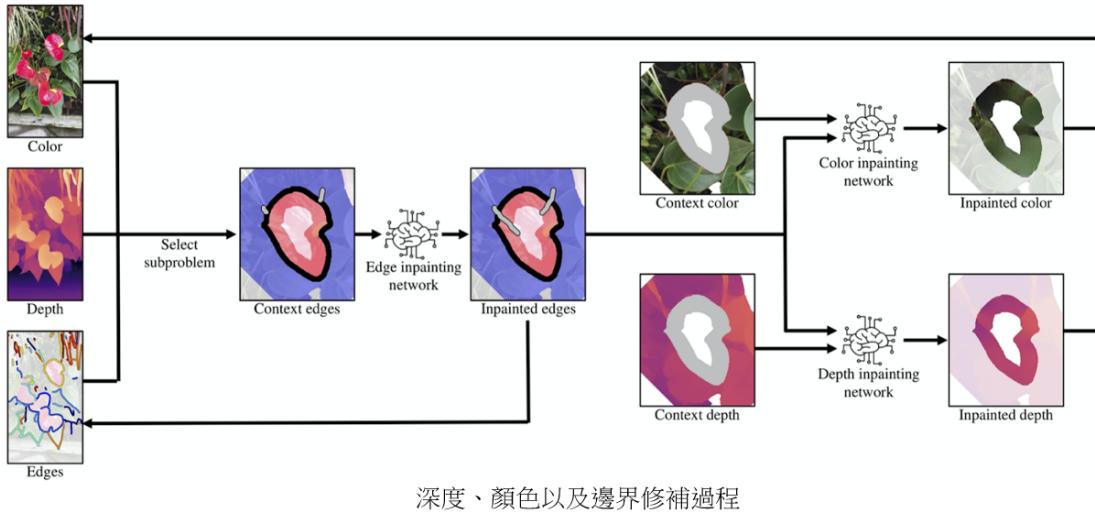
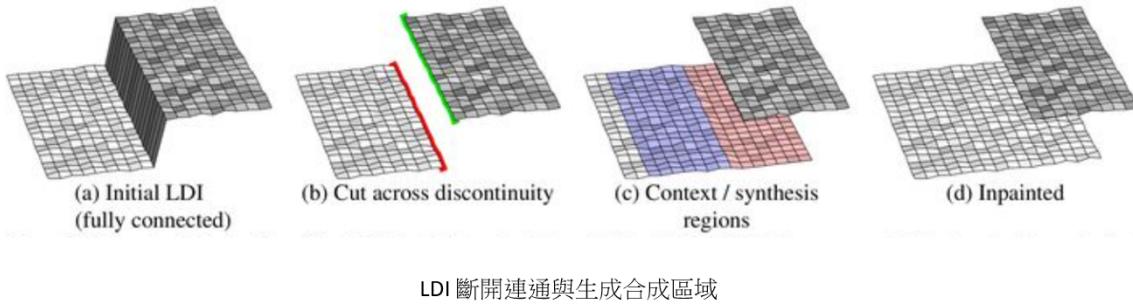
Novel View Synthesis

我們主要嘗試了兩種不同的 novel view synthesis 方法來實作此次專題，並於最後探討了兩種方法各自的優缺點以及其限制。

- **3D Photography**

3D photography 的方法我們主要是參考 [CVPR 2020] 3D Photography using Context-aware Layered Depth Inpainting [1] 這篇論文，此方法首先對影像進行深度估計，得到 RGB-D 的影像後，將問題拆分成三個子網路來處理，並使用 Layered Depth Image (LDI) 來儲存結果，其中除了紀

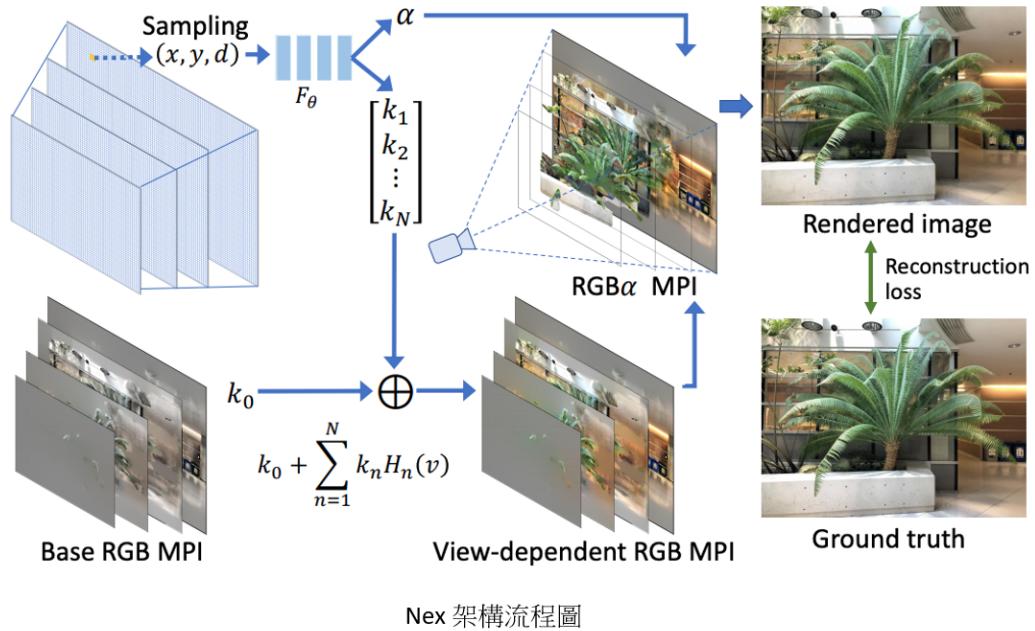
錄每個 pixel 的顏色資訊外，額外紀錄了與相鄰 pixel 的深度連通關係。而三個子網路分別處理的是深度、顏色，以及邊界，在計算時，首先會根據深度資訊初始化 LDI，並根據 LDI 中所記錄的深度連通關係將深度不連續的部分連通性斷開，並將斷開的區域切分為前、背景，生成一塊合成區域用以表示被前景遮蔽的部分，接著便依序使用三個子網路處理合成區域的資訊。首先使用邊界子網路預測此區域邊界關係，能夠推斷合成區域的結構關係並用於引導後續子網路預測結果；而依靠邊界子網路的預測結果，深度子網路與顏色子網路則依序修復合成區域的資訊，並將修復後的結果更新回 LDI 中。因此，場景資訊最終以 LDI 的形式呈現，在給定新的視角作為輸入時，便能依靠 LDI 中紀錄的深度連通關係來合成新視角的影像及顏色，達到 novel view 的效果。



• NeX

NeX 的方法主要是參考 [CVPR2021 Oral] NeX: Real-time View Synthesis with Neural Basis Expansion [6] 這篇論文。在該篇論文中他們主要是利用 multiplane image (MPI) 的方式將一個場景合成多張 plane 組成的深度場景來還原 novel view。不同的是他們把每個 MPI 的 pixel 都表示成一個 multi layer perceptron 的 output。因此該 MPI 可以適

應特殊材質(如湯匙、玻璃...)等的反光效果。整個方法 implement 的流程如下圖所示。MPI 可以表示成多張 RGBa 的 image, 其中 alpha為該 pixel 的透明度。每個 pixel 在經過 MLP 後會輸出該 pixel 位置的 α 值與 $k_1 \dots k_n$, n 個 k 用以表示定義好的 n 個 view dependent 的 basis function 的權重。將所有 basis function outputs 加權在一起即是最後的RGB 值。最後根據產生的 image 與 ground truth 之間的 loss 即可更新此 model 架構。值得注意的是對於每個不同的場景, Nex 都必須重新訓練一遍 (Instance based method), 並且每個場景都需要至少輸入 12 張不同角度拍出的影像作為 trianing image set, 這使得此方法無法即時套用在不同的場景之中, 只能於 offline 的效果生成所使用, 這是其較大的限制之一。



Nex 架構流程圖

V. Experiment

Novel View Video Generation

為了重現哈利波特電影中報紙裡人與真實世界角色互動的效果, 我們希望能將多個連續的場景 (例如一個人在揮手的影片) 都套用 novel view synthesis, 如此一來, 就能重現一段影片隨著真實世界旋轉平移的效果。我們嘗試使用 Nex 來重現此效果, 由於 Nex 需要以一個場景的 image set 做 training input, 其中每個場景各自算出來 world coordinate system 的 R, t 都會有 scale 上的不同, 因此需要完全對齊每個 scene。我們嘗試了兩種方法:(1) 透過腳架來固定拍攝的 image set 的角度, 並用第一個場景算出來的 R, t 來當作所有 image set 的旋轉平移。(2) 算出每個 image set 的 world coordinate 後對齊他們並找出各自的 scale, 並在轉換成真實世界座標時使用該 scale 做 normalize。

以結論來說，最後兩種方法的結果都不太好。我們推測原因是因為對於每個場景訓練出來的 multiplane image, 就算 R, t 的 scale 相同，得到的相同 R, t 的 output 也會有些微的不同，例如比較不 sharp 的細節不同，或是一些物品擺放位置也有些微差異。這些差異在 image sequence 連續放在一起的時候就會看出明顯的不同。因此我們認為 Nex 並不適合用來實作 novel view video generation。相對的，3D photography 只需要一張 image 當 input, 且不需要重新訓練模型，因此比 Nex 更適合拿來實作該效果。

Demonstration

- **3D photography**

3D Photography method 影片連結: <https://youtu.be/OY2GQGhqWXo>

在這個影片中，我們嘗試用 3D Photography 來實作 Nex 訓練失敗的 novel view video generation, 首先我們將影片中的各個 frame 抽取出來各自 inference 一段 novel view video, 因為拍攝時視角固定且使用同樣的 R, t 去 render, 產生出的影片會有同樣的 novel view, 差別只有每部影片紀錄的是不同 frame 下的 motion, 因此我們再將對應的 frame 取出(第 i 個 frame 產生出的影片就抽取該影片的第 i 個 frame)並重新合成一部新影片，就能夠在產生 novel view 的同時將影片中的 motion 記錄下來。

因為 input 為 single image, 透過此方法只能 render 出周圍一小部份的 novel view, 再加上我們所套用的 R, t 也是來自於拍攝的影片，會有些許的晃動，導致在 demo 影片中的 novel view 效果看起來不是特別明顯，但我們可以從物體的陰影處看出 model 確實有在物體的移動過程中，將不同角度的 background color render 出來。



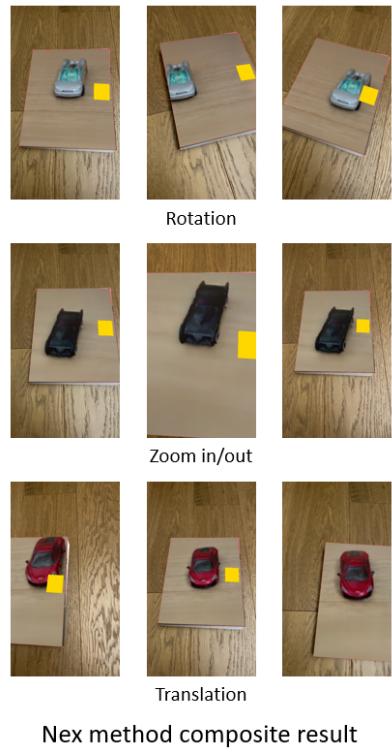
3D photography method composite result

- **Nex**

NeX method 影片連結: https://youtu.be/lygbEjOcS_E

在這個 demo 影片中，我們模擬一個 3D 線上車展的雜誌。能根據使用者的視角來多角度觀察一台新車。如果喜歡便可加入購物車。我們以第一人稱視角拍攝影片，可以看到車子的視角會隨著攝影機移動一起移動。也可以看出 Nex 在重現車子上能夠相當的 sharp。但對於一些特定視角或是顏色較深的材質還是可能會有細節模糊的情況產生。

從影片中也可看出，對於拍攝時的一些 vibration，在虛擬環境中那些手震都會被放大，因為兩個世界座標在模擬時，為了讓兩邊的 motion 與視角看起來清楚且合理，我們要經過一些 scale 的轉換。而這個 scale 的轉換也讓手震在虛擬世界中看起來更加明顯。如果想要調整 scale 來消除手震，則視角會變得不明顯且失焦。因此，最好的方法是能夠試著用其他人做的 video stabilization 的 work。這個領域如今也已經發展得相當純熟，相信能夠有效解決著個問題。



Nex method composite result

VI. Future Work

在我們現今使用的方法中，還是存在著一些 limitation:(1) 3D photography: 對於深度較為複雜的圖片可能會估算錯誤，在產生出的影片中可以明顯看出物體有錯位的情形，另外此方法只能處理小部分的 novel view，若旋轉平移太大則會 render 不出來，使影片有空洞的地方(2) Nex: 影片細節的 pixel 有時候會模糊，而且一個 scene 需要多張image sequence 去重新訓練一個模型。未來我們希望能夠有效解決這些 limitation，來讓整體的實作更真實且方便。

在 demo 影片方面，除了能夠去除手震外，我們還希望能加大整體 scene 的複雜度，讓整個虛擬世界看起來深度更有層次，如此一來 demo 的效果也會更好。

VII. Teamwork Distribution

潘奕廷: 3D reconstruction & AR 合成效果實作、report 撰寫

黃子源: Nex 方法實作、report 撰寫

高晟璋: 3D photography 方法實作、report 撰寫

VIII. Reference

1. [Shih, et al. 3d photography using context-aware layered depth inpainting. CVPR, 2020](#)
2. [Zhou, et al. Stereo Magnification: Learning view synthesis using multiplane images. SIGGRAPH, 2018](#)
3. [Yang, Nan, et al. D3VO: Deep Depth, Deep Pose and Deep](#)

[Uncertainty for Monocular Visual Odometry. CVPR 2020](#)

4. [Zhou, Tinghui, et al. Unsupervised learning of depth and ego-motion from video. CVPR 2017](#)
5. [Srinivasan, Pratul P., et al. Pushing the boundaries of view extrapolation with multiplane images. CVPR 2019](#)
6. [Suttisak Wizadwongsa, et al. NeX: Real-time View Synthesis with Neural Basis Expansion. CVPR 2021](#)
7. [Olivia Wiles, et al. Synsin: End-to-end view synthesis from a single image. CoRR 2019](#)
8. [Ben Mildenhall, et al. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. ECCV 2020 Oral](#)
9. [COLMAP - Structure-from-Motion and Multi-View Stereo](#)
10. [Open Multiple View Geometry library. Basis for 3D computer vision and Structure from Motion.](#)