國立清華大學 電機工程學系

實作專題研究成果報告

Indoor Scene Completion With Pretrained Diffusion Models

室內場景補全

專題領域：資工領域

組　　別：B381

指導教授：孫民

組員姓名：谷岳峰

研究期間：112 年 1 月 17 日至 112 年 11 月 27 日止，共 11 個月

1. 摘要

我們提出一種不需任何額外訓練 (training-free) 的方法，能以一個室內場景的少量 RGB-D 影像為輸入，對不完整的3D 室內場景進行補全。我們首先將輸入 RGB-D 影像投回三維空間取得有空缺的3D 網格模型(mesh)。接著我們從有空缺的3D 網格的中心取得有空缺的 RGB-D 全景圖並依序使用預訓練的文字轉圖片生成模型(Stable Diffusion Inpainting) 和預訓練的單目深度 (monocular depth) 預測模型對全景圖補全並投回3D 網格，這個階段將會補全大部分的空缺。最後，為了補全剩餘的小洞，我們逐步從預定或隨機的相機位置對3D 網格進行補全，最終生成完整的室內場景網格。我們在 ScanNet [6] 和 ArkitScenes [7] 資料集上做評估，並在視覺、幾何及語義方面的指標皆勝過前人的研究。

2. 研究背景

在虛擬實境、電腦圖學、遊戲開發等應用中，3D 室內場景都十分重要。目前多數 3D 場景重建的方法都需要大量關於場景的影像，且較無法還原或推敲出未被影像觀察到的部分。

近年來一類稱為 diffusion model 的圖片生成模型在訓練於大規模網路圖片後，展現出了生成真實、多樣且清晰圖片的能力，這類模型也解決圖片補全(inpainting)等等的相關應用 [1]。本專題希望借重預訓練在大量網路圖片的2D 圖片補全模型，對一個有空缺的3D 場景進行補全。

本專題的目標是使用預訓練圖片生成模型，給定數張室內場景的 RGB-D 影像 (D 代表深度資訊) 為輸入，重建出場景的3D 網格並在未被輸入影像觀察到的部分生成合理、風格一致的內容。

前人研究 RGBD2 [5] 訓練一個 Diffusion 模型來對場景進行補全，然而由於是訓練在與大規模網路資料相比較小的資料集上，輸出圖片的畫質和多樣性受到資料集所侷限，並且也需要用到為每個場景預先定義好的相機軌跡；Text2room [4] 提出使用預訓練文字轉圖片生成模型來生成3D 場景的方法，然而需要精心設計的相機軌跡以及文字描述(text prompt)。因此，我們提出一種新的方法，成功解決上述的問題。

3. 研究方法

輸入 N 張 RGB 影像 $\{I_i\}_{i=1}^{N}$、其深度圖(depth map) $\{\mathbf{D_i}\}_{i=1}^{N}$ 及其相機姿態(camera pose) $\{\mathbf{P_i}\}_{i=1}^{N}$，我們的方法可以生成完整的3D 場景三角網格模型(mesh) $M = (V, C, S)$，V、C、S 分別為網格頂點、顏色和面。

2-1. 流程概覽

我們的方法可大致分為四個步驟 (請參考 Fig. 1)：(1) 我們使用 textual inversion 方法取得能代表輸入影像風格的 text embedding。這些 text embedding 將會在後續作為圖片補全模型的輸入之一。(2) 我們將輸入 RGB 影像的像素根據其深度圖和相機姿態投至3D 空間，並將各個像素串聯起來成為三角網格的面，取得初始化有空缺的場景網格。(3) 接著我們從有空缺網格模型的中心取得有空缺的 RGB-D 全景圖，並依序使用預訓練的圖片補全模型(Stable Diffusion Inpainting [1]) 和預訓練的單目深度 (monocular depth) 預測模型對全景圖補全，最後將補全結果投回3D 網格，這個階段將會補全大部分的空缺。為此我們也提出一種使用預訓練圖片補全模型來補全全景圖的方法。(4) 最後，為

了補全剩餘的小洞，我們逐步從預定或隨機的相機位置對網格模型進行補全，最終生成完整的室內場景網格。

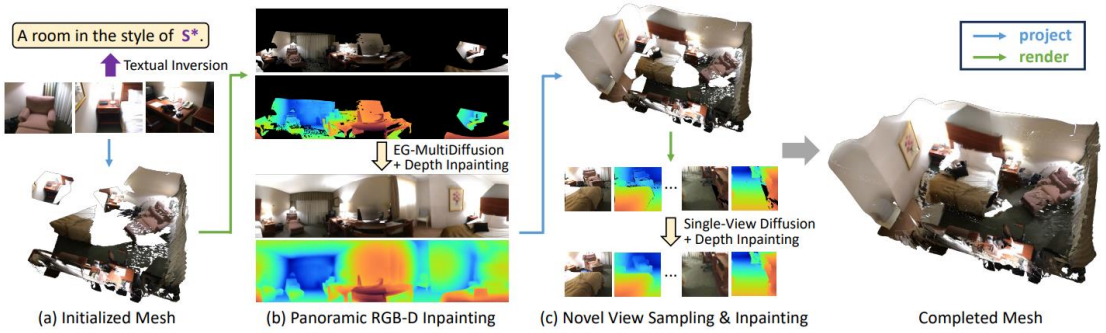

Figure 1. Pipeline of our method

2-2. 使用2D 圖片補全模型補全 3D 網格

輸入一張有空缺的圖片 **I**、補全遮罩(inpainting mask) **m** 和一段文字描述(text prompt) **T**，圖片補全模型可以在遮罩為1的部分生成合理的內容，將原先的輸入圖片 **I** 補全為完整的圖片 $\mathbf{I}'$，即 $\mathbf{I}' = G(\mathbf{I}, \mathbf{m}, \mathbf{T})$。RGBD2 [5] 和 Text2room [4] 提出可以從一段相機軌跡的 L 個相機姿態 $\{\hat{P}_i\}_{i=1}^{L}$逐步補全3D 網格：在每個相機姿態$\hat{P}$ render 出一張有空缺的圖片$\hat{I}_i$及表示哪裡沒 render 到網格的遮罩$\boldsymbol{m_i}$。接著使用上述的圖片補全模型生成補全完整的圖片$\hat{I'}_i$並使用單目深度預測模型預測這張圖片的深度圖 $\hat{D}_t'$。最後，在依照深度圖$\hat{D}_t'$和相機姿態$\hat{P}$將圖片$\hat{I'}_i$投回3D 空間，並連結像素點成為三角網格面。

然而，這類方法需要精心設計的相機軌跡$\{\hat{P}_i\}_{i=1}^{L}$使得每個相機姿態都必須觀察到一部份存在的網格，且存在 loop-closure 的問題。因此，我們提出先對場景中心的全景圖進行補全以解決這些問題。

2-3. 使用 Multi-view Diffusion 生成全景圖

MultiDiffusion [3] 可以使用預訓練的文字轉圖片模型生成全景圖。MultiDiffusion [3] 將原局的全景圖分為多個互相重疊的局部視窗，並使用文字轉圖片模型同時對這些局部視窗做 denoising。然而生出來的結果並不符合 equirectangular panorama 全景圖的幾何特性，如 Fig. 3。

我們首先使用從全景圖取得 M 個互相重疊的 perspective 圖片$\{x^i\}_{i=1}^{M}$及其 inpainting mask。對於第 i 個 view，定義$\{x_T^i, \dots, x_0^i\}_{i=1}^{M}$為由總共 T 個 step 的 reverse diffusion process 產生的 noisy latent images，其中$\{x_T^i\}_{i=1}^{M}$ 初始化為 Gaussian noise 且 $\{x_0^i\}_{i=1}^{M}$ 是最終補全的圖片。對於 reverse diffusion process 的第 t 個 step，我們預測 noise ϵ_θ 並 de-noise 出 $\widehat{x_0^i}$，對於每個 view i：

$$\widehat{x_0^i} = \frac{x_t^i - \sqrt{1-\alpha_t}\epsilon_\theta\left(x_t^i, t, \boldsymbol{I}, \boldsymbol{m}, \boldsymbol{T}\right)}{\sqrt{\alpha_t}} \tag{1}$$

接著我們將各個 view denoise 的結果平均起來：

$$\widehat{x_0^i}' = \frac{\sum_j W_{j\to i}\left(x_0^j\right)}{\sum_j m_{j\to i}} \tag{2}$$

其中$W_{j\to i}$表示從第 j 個 view warp 到第 i 個 view 的轉換。

最終，我們加入 random noise ϵ 取得 x_{t-1}^i：

$$x_{t-1}^i = \sqrt{\alpha_{t-1}}\,\widehat{x_0^i}' + \sqrt{1-\alpha_{t-1}}\epsilon \tag{3}$$

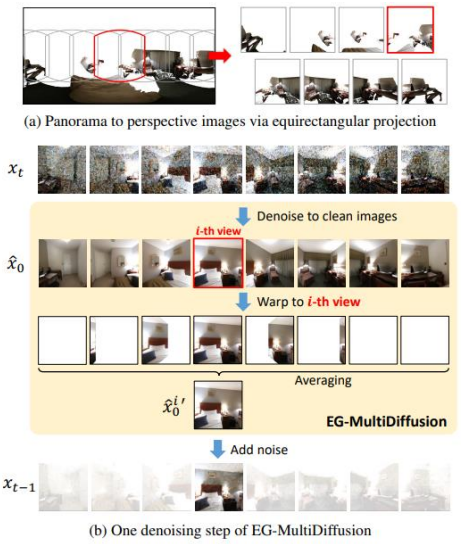

Figure 2. Multi-view diffusion with equirectangular geometry

Figure 3. Comparison of methods for panorama generation

Texture Refinement. 由於每次 warping 都需要對圖片做 interpolation，過程會有資訊流失，導致最終生成結果失去高頻率的細節。因此在最終的 F 個 step 我們使用 MultiDiffusion [3] 做 denoising。如此既可確保全景圖幾何符合 equirectangular geometry，又能保留較高頻率的細節，如 Fig. 3。

本專題設定 M=8, T=50 和 F=20，並使用 Stable Diffusion Inpainting 模型 [1] 作為圖片補全模型。

2-4. 全景圖深度圖補全

本專題使用單目深度預測模型 IronDepth [2] 來預測全景圖的深度。此模型的流程包含兩個階段：initial depth prediction 和20個循環的 depth refinement 。我們使用類似2-3的策略，先對全景圖中數個相互重疊的 perspective images 預測深度，經過與從3D 網格 render 出有空缺的深度圖對齊後，使用該模型的 depth refinement 模型對預測深度做20次 refinement。在每次 refine 過後對各個 perspective images 的預測深度在重疊處做平均，確保深度的一致性。

2-5. 網格補全

經過全景圖補全後，3D 網格大部分的空缺已被補全。取決於使用情境，我們的方法可以根據預先定義的相機軌跡對剩餘縫隙進行補全(請參考2-2)，或是隨機選取相機位置逐步補全3D 網格：每次先在場景內部隨機採樣多個相機姿態，並選擇一個最佳的相機姿態，再以2-2描述的方法進行補全。選擇最佳相機姿態的策略是挑選空缺比例和深度圖最小值的乘積最大者，並且會過濾掉空缺比例<1%或>50%、backface 比例>1%以及深度圖最小值<1公尺的相機姿態。這樣通常能優先捕捉到場景中剩餘最大的洞。

4. 研究結果

我們在 ScanNet [6] 資料集上對我們的方法於前人方法 RGBD2 [5] 和 Text2room* [4] 進行比較。另外，由於 RGBD2 [5] 是訓練在 ScanNet 資料集，我們也使用 ArkitScenes [7] 資料集驗證我們方法在不同 domain 泛化的能力。Text2room [4] 原先的情境是根據文字敘述生成3D 場景，因此我們將其修改為 Text2room* [4] 以符合我們的情境設定。

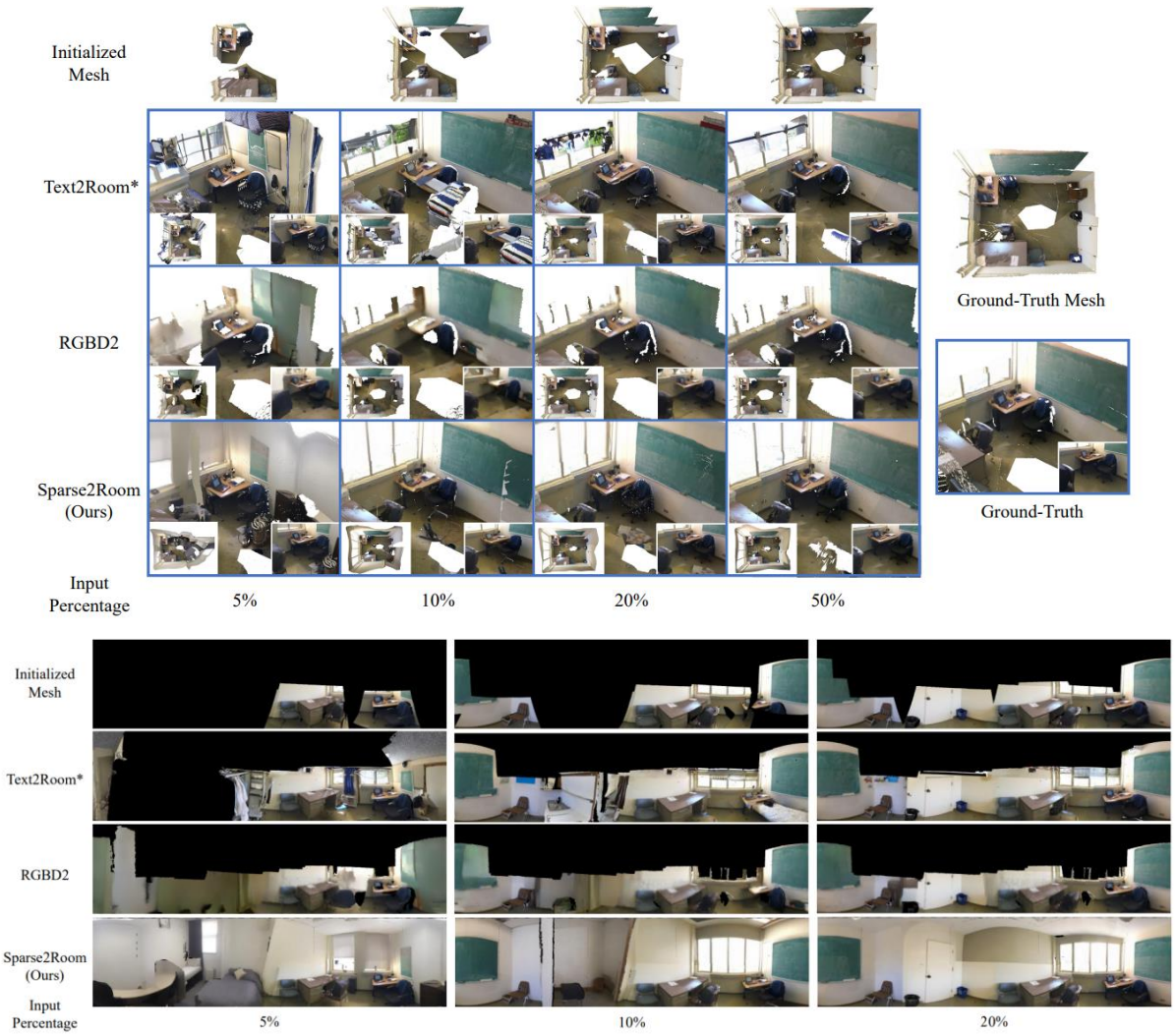

Figure 4. Qualitative comparisons on ScanNet

Methods	Visual												Geometric				Semantic			
	$PSNR_{color}(\uparrow)$				$SSIM_{color}(\uparrow)$				$LPIPS_{color}(\downarrow)$				$MSE_{depth}(\downarrow)$				$CS_{input}(\uparrow)$			
	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%
Text2Room* [15]	13.2	14.9	16.7	17.2	0.460	0.543	0.598	0.615	0.566	0.488	0.439	0.417	0.55	0.21	0.14	0.12	0.75	0.79	**0.81**	**0.80**
RGBD2 [18]	13.7	16.0	17.5	**18.6**	0.501	0.562	0.599	0.609	0.564	0.488	0.446	0.417	0.38	0.15	**0.08**	**0.06**	0.71	0.72	0.72	0.73
Ours	**14.4**	**16.7**	**17.6**	18.2	**0.524**	**0.599**	**0.628**	**0.633**	**0.531**	**0.441**	**0.410**	**0.400**	**0.27**	**0.13**	0.09	0.09	**0.79**	**0.81**	**0.81**	**0.80**

Table 1. Qualitative results on ScanNet [6]

Methods	Visual												Geometric				Semantic			
	$PSNR_{color}(\uparrow)$				$SSIM_{color}(\uparrow)$				$LPIPS_{color}(\downarrow)$				$MSE_{depth}(\downarrow)$				$CS_{input}(\uparrow)$			
	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%
Text2Room* [15]	11.4	12.9	14.3	15.2	0.383	0.433	0.498	0.534	0.672	0.601	0.523	0.480	0.92	0.70	0.32	0.28	0.83	0.85	0.86	**0.87**
RGBD2 [18]	12.2	13.9	15.2	16.6	0.463	0.502	0.541	0.564	0.665	0.596	0.532	0.474	0.51	0.29	0.20	0.11	0.80	0.81	0.81	0.81
Ours	**13.2**	**14.7**	**15.8**	**16.8**	**0.504**	**0.545**	**0.572**	**0.592**	**0.630**	**0.555**	**0.499**	**0.466**	**0.41**	**0.27**	**0.14**	**0.09**	**0.87**	**0.87**	**0.87**	**0.87**

Table 2. Qualitative results on ArkitScenes [7]

5. 總結

我們提出一種不需任何額外訓練 (training-free) 的方法，能以一個室內場景的少量 RGB-D 影像為輸入，對不完整的3D 室內場景進行補全。我們首先將輸入 RGB-D 影像投回三維空間取得不完整的三維網格模型(mesh)。接著我們從不完整網格模型的中心取得不完整的 RGB-D 全景圖並依序使用預訓練的文字轉圖片生成模型(Stable Diffusion Inpainting) 和預訓練的單目深度 (monocular depth) 預測模型對全景圖補全並投回三維網格模型，這個階段將會補全大部分的空缺。最後，為了補全剩餘的小洞，我們逐步從預定或隨機的相機位置對網格模型進行補全，最終生成完整的室內場景網格

6. 心得感想

在加入實驗室的第一個學期主要是在閱讀各個電腦視覺領域的文獻，並與其他學長和專題生同學分享與輪流報告看到的 paper。坦白說，最初在看 paper 階段時我天真的以為有些電腦視覺方法的流程都很直觀，應該不難想到，直到後來加入了這個跟兩位學長合作的 project 實際動手時，才發現很多問題乍看之下很好解決，但實際上很多實作的想法都會因為各種意想不到原因而達不到預期的效果。幸好有老師的指導和學長們的討論，我們才能逐步的找出導致方法失敗的原因並逐一解決。經過這一年的閱讀、討論、研究和嘗試，從一開始看 NeRF、visual grounding、到這個 project 原本是 NeRF 跟 Diffusion 結合但 NeRF 再度默默消失(XD)，我學到很多電腦視覺領域的知識、也了解到進行研究的流程與技巧。感謝老師、實驗室學長們和專題生同學一路以來的指導與討論。

7. 參考資料

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image syn-¨ thesis with latent diffusion models. In *CVPR*, pages 10684– 10695, 2022.

[2] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Irondepth: Iterative refinement of single-view depth using surface normal and its uncertainty. *arXiv preprint arXiv:2210.03676*, 2022.

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *ICML*, 2023.

[4] Lukas Hollein, Ang Cao, Andrew Owens, Justin Johnson,¨ and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. *arXiv preprint arXiv:2303.11989*, 2023.

[5] Jiabao Lei, Jiapeng Tang, and Kui Jia. Rgbd2: Generative scene synthesis via incremental view inpainting using rgbd diffusion models. In *CVPR*, pages 8422–8434, 2023.

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017.

[7] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.