國立台北大學資訊工程學系專題報告

Enhancing 2D Pixels to 3D Models Generation Via Input Normalization

專題組員:Hsin-Fei Wu, Chieh-Ting Hsu, Pei-Ti Shi, Chang-Jui Lin

專題編號:PRJ-NTPUCSIE-114-009

執行期間:2025 年 7 月 至 2026 年 5 月

1.摘要

本研究旨在開發一套「單張 2D 影像轉 3D 模型並綁定骨架」之自動化流程系統。傳統 2D 轉 3D 生成技術在面對非標準化輸入源時，常因輸入影像品質差異與模型本體限制，導致幾何網格畸變及紋理貼圖染色等不穩定現象。

為解決此痛點，本系統於影像輸入端建構了一套自動化預處理流程，整合解析度提升、姿勢與視角標準化、以及智慧去背等模組，將任意輸入影像進行標準化處理。隨後導入 Step1X-3D 架構進行幾何與紋理之二階段生成，並結合 Unirig 進行自動化骨架綁定與權重配置。

實驗結果表明，本系統所提出之預處理機制能顯著提升 3D 模型生成的外形完整度、貼圖精確度以及骨架分布與權重穩定度。本專題亦開發了直觀之使用者介面，有效降低 3D 建模與動畫製作之技術門檻，在數位內容創作（UGC）與虛擬實境（VR）互動應用中具備高實用價值與發展潛力。

2.簡介

現實生活中，人們常透過相機拍攝照片記錄重要的人與動物，也會在 2D 平面中創作自己喜愛的角色。然而此類影像大多僅能以平面形式保存。

若能將 2D 影像自動化轉換為 3D 模型，不僅能提升視覺體驗的真實性與趣味性，亦可結合 3D 列印技術製成實體公仔。此外，若生成的 3D 模型具備骨架結構（Rigging），便能進一步調整姿勢、製作動畫，並廣泛應用於虛擬實境（VR）、遊戲開發等互動式場景為達成上述目標，本專題旨在建立一

套「單張 2D 影像轉 3D 模型並綁定骨架」之自動化流程系統。在早期實驗中，本團隊嘗試直接利用現有技術進行生成，發現其結果極不穩定，易產生幾何畸變或結構不合理之現象。

經深入分析，輸入影像的品質為決定 3D 模型生成質量的關鍵因素。由於無法限制使用者輸入之影像規格，本專題於「影像輸入」與「模型生成」之間建構了一套影像預處理橋樑。透過解析度提升、姿勢與視角標準化、以及背景去處等預處理模組，將輸入影像進行標準化，顯著提升後續 3D 生成模型的穩定度與精確度。

3.相關論文研究

本專題於規劃初期分為「3D 模型品質優化」與「自動化骨架技術」兩大研究主軸並行推進。在文獻回顧與前期實驗中，針對現有的 2D 轉 3D 生成模型與骨架綁定技術進行深入探討，並確立了本研究的切入點。

A. 現存技術瓶頸與挑戰

經文獻探討，現有的 2D 轉 3D 視覺模型在實際應用中普遍存在以下瓶頸

外觀生成不穩定：主流模型多專注

於「三維幾何網格」的推算，但在「紋理貼圖預測」與「色彩跨視角同步」上成效有限，導致外觀產出品質極不穩定。

泛用性受限：多數前沿技術屬於「特定物種專項模型」（例如僅限定於人類或特定四足動物），缺乏對泛用性物種的生成能力。

骨架綁定連帶缺陷： 由於前級生成的幾何網格在邊緣與關節處不夠精確，亦無法穩定串接紋理資訊，導致後續的自動化骨架綁定與權重配置品質大幅下降。

B. 關鍵技術導入與本研究之定位

為克服上述幾何與紋理脫節的困境，本專題導入了近年發表的開源前沿架構 Step1X-3D。該架構具備「幾何與紋理分離之二階段生成」特性，提供了一體化的網格與外觀生成系統，顯著提升了幾何形狀與紋理搭配的穩定度。

然而在實驗中我們發現，當面對結構較為複雜或非標準化的輸入影像時，Step1X-3D 仍會因輸入源的雜訊而產生幾何失真與紋理污染，並連帶影響後續骨架綁定之生成品質。

因此，本專題並非盲目追求修改底層擴散模型之參數，而是將研究方向轉向「逆向問題探討與輸入源優化」。本研究聚焦於架構一套前級影像預處理管線，作為複雜輸入影像與 3D 生成模型之間的標準化橋樑，以此全面提升既有開源架構在泛用場景下的生成品質。

4. 問題分析

A. 解析度問題

我們在實驗中發現，輸入影像的解析度對於最終模型的貼圖質感具有決定性的影響。

低解析度影像導致前級演算法無法精準擷取銳利邊緣，使得生成的 3D 網格表面流暢度不足。當後續紋理擴散模型嘗試將材質投影至該網格時，幾何不對齊會導致 UV 貼圖產生嚴重的拉伸與邊緣模糊。即便後端的多視角同步演算法再強，它也只能在多個視角之間同步「模糊的顏色塊」，而無法在 3D 的 UV 貼圖上還原出清晰、立體的材質細節。

圖一、解析度模糊的影像(左) 生成模型對照(右)

B. 姿勢及視角問題

a. 姿勢問題:

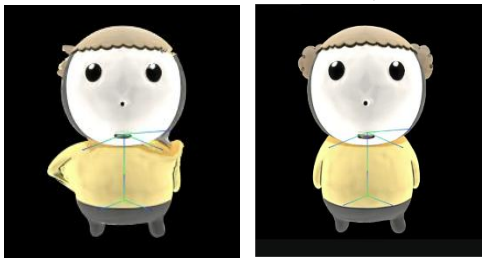
在 3D 幾何與骨架預測階段，輸入影像中主體的姿態對最終模型的可用性具決定性影響。若主體呈現非標準姿態（如坐姿、趴臥）或產生肢體自我遮蔽與黏連時，系統將面臨以下技術瓶頸：

解剖結構辨識失效：模型難以精準區分四肢、尾巴等特徵物件的邊界與比例，導致生成的 3D 網格產生非預期的拓撲結構黏連。

關節點拓撲預測缺失：由於特徵點被遮蔽，後續骨架生成流程無法精確定位核心關節點，導致預測的關節軸數量不足，無法進行高自由度的動畫應用。

蒙皮權重分配不均：當幾何網格因姿態重疊而畸變時，骨架與網格間的權重綁定將發生偏誤。在後續調整姿勢時，會引發嚴重的網格拉伸、扭曲或不自然穿面等現象。

(圖一)未經處理的輸入影像



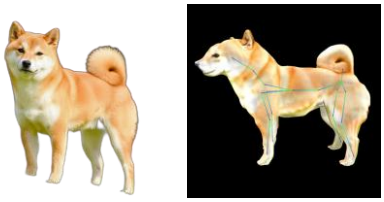
(圖二)關節點拓樸預測缺失(左)連帶影響後續骨骼處理(右)，包含關節點缺失、蒙皮權重不均與結構粘連。

b.視角問題

單張 2D 影像本質上缺乏三維空間的深度資訊。當輸入影像的主體未處於標準視角(如完全正側面或正背面)，而是呈現任意面向鏡頭的非標準角度時，將產生以下限制:

深度估算失準：生成模型在推算軸向深度與形體厚度時，易產生視覺縮短誤差，導致 3D 身體比例計算失準。

非對稱幾何拉伸：為預測影像中不可見的面，擴散模型必須依據輸入角度進行跨視角外推。然而，若輸入視角具備傾角，演算法在進行網格重建時，會導致視覺盲區的網格產生「部分拉伸過度、部分壓縮不足」的非對稱形變，最終破壞模型的整體對稱性與美觀度。



(圖三)輸入為未經處理的影像(左)，產生如(右)之頭部比例扭曲之模型

C.背景雜訊問題

StepIX-3D 採用多視角擴散模型(Multi-view Diffusion)在潛在空間中進行跨視角同步，以此確保從前、後、左、右看過去的 3D 貼圖是連貫的。

在進行多視角旋轉預測時，AI 的注意力機制會混淆「主體邊緣」與「背景邊緣」。這會導致在嘗試「Bake」3D 貼圖時，誤把背景的色彩與紋理當作主體的一部分一起投影上去。這就是為什麼生成的 3D 模型腹部經常會黏著背景顏色的原因。(如圖)

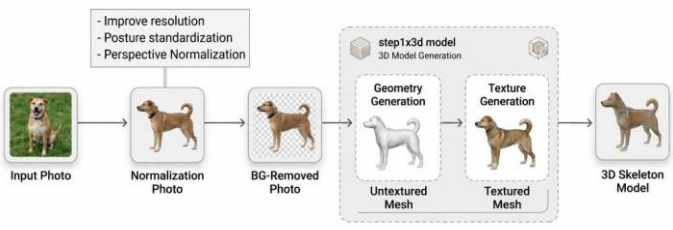
(圖四)還未去背的影像(左) 由左圖生成的模型可以看到渲染到背景顏色(右)

5.實作方法

A. 流程總覽

首先先將輸入的影像進行預處理，分別是解析度調整、姿勢標準化、角度標準化，之後再進行去背。

再將處理好的影像傳至模型分別依序生成 3D 模型及骨架生成與權重綁定。



(圖五)pipeline 示意圖

B. **影像預處理**

解析度提升:

(圖六)解析度提升前(左)解析度提升後(右)

姿勢與視角標準化:

(圖七)視角標準化前(左)視角標準化後(右)

去背:

(圖八)去背前(左)去背後(右)

C. **2D 轉 3D 視覺技術**

在 3D 生成階段，本專題導入了前沿的 Step1X-3D 架構。Step1X-3D 論文的核心原理在於提出了「幾何與紋理分離之二階段生成流程」與「潛在空間同步化技術」：模型第一階段專注於「三維幾何形狀」的推算。它採用了 3D 變分自編碼器（3D VAE） 結合 擴散變壓器（DiT,Diffusion Transformer 架構。幾何網格確定後，第二階段負責將 2D 高清影像的紋理「Bake」至該網格上。

D. 3D 圖學與骨架綁定(Unirig)

在自動化骨架綁定與蒙皮權重配置階段，本專題導入前沿之 UniRig 架構 [2]。該架構之核心優勢在於具備「類別無關（Category-Agnostic）」之泛化能力，無需預先指定物體類別(如人類或特定四足動物)，即可針對任意 3D 網格自動推導出層級化骨架拓撲（Rigging Skeleton）與預測蒙皮權重（Skinning Weights）。

在本系統之自動化管線中，UniRig 作為後級模組，其預測精確度高度依賴前級生成網格的質量。

技術痛點：若直接輸入未經預處理的畸變網格，UniRig 的幾何編碼器易受黏連背景或非對稱肢體干擾，導致生成的關節點偏移或權重嚴重扭曲（引發動畫拉伸與穿面問題）。

整合效益：透過本專題提出之前級影像預處理，Step1X-3D 產出之網格具備良好的對稱性與清晰邊界。這使得 UniRig 能精確定位核心關節，生成結構完整、權重分配均勻，且可直接導入 Blender 或遊戲引擎（如 Unity、UE5）中使用的骨架模型。

E. **軟體工程與流程自動化**

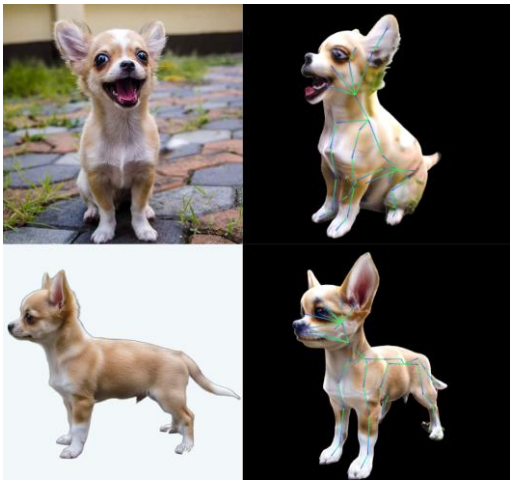
本專題之核心貢獻除前級影像優化外，亦包含將多個異質模型與工具鏈進行深度系統整合。由於從影像預處理、Step1X-3D 模型生成到後端 UniRig 骨架綁定，各模組所依賴的運行環境存在嚴重版本衝突，本章節

將說明本系統如何透過軟體工程架構克服異質環境串接、實作自動化管線，並設計直觀之使用者介面。

異質環境隔離與微服務架構:
為了合理解決各開源模型間的環境衝突，本系統採用多重環境隔離之微服務架構：
環境隔離策略：針對「影像預處理」、「Step1X-3D」與「UniRig」分別建構獨立的 Conda 虛擬環境，確保各模組的 CUDA 核心與深度學習相依套件互不干擾。
模組化資料交換：各獨立環境之間透過輕量化格式進行通訊，由主程式傳遞控制參數與 JSON 格式的中介資料，並透過系統檔案路徑交換產出的 .png 影像、.obj 幾何網格與 .fbx 骨架模型。
自動化管線整合與例外處理系統核心開發了一套主控端腳本，負責管控端到端（End-to-End）的完整自動化流程。
循序工作流控制：使用 Python 的 subprocess 模組依序動態觸發各虛擬環境中的推論腳本。當使用者上傳影像後，系統將自動執行：解析度提升--姿勢與視角標準化--智慧去背--Step1X-3D 生成 — UniRig 綁定之完整流程。
使用者介面設計與互動優化(以 Gradio 實作)為達成「無痛單鍵操作」之目標，本專題基於 Gradio 框架建構了直觀的 WebUI，將後端複雜的 AI 管線封裝為平易近人的軟體體驗。利用 Gradio 的事件監聽機制（如 btn.click()），使用者只需透過 gr.Image 上傳單張 2D 影像並點擊生成，後端即自動觸發非同步執行緒，將複雜的技術細節對使用者完全隱藏。最終生成的 3D 模型則整合 gr.Model3D 組件，讓使用者無需安裝外部軟體，即可在網頁瀏覽器上直接進行 360 度旋轉與縮放檢視。

6.主要成果與評估

預處理優化之結果:



(圖九)、優化前(上)優化後(下)生成模型對照。

(圖十)、優化前(上)優化後(下)生成模型對照。

(圖十一)、優化前(上)優化後(下)生成模型對照。

7.結語與展望

這篇專題主要實作了一套「將單張 2D 影像轉變成 3D 模型」的自動化流程系統並透過一連串預處理影像解決了因為 input 的異常因素導致的貼圖渲染與生成模型骨架的偏差問題。

解決了複雜的串接：將多個本來環境互不相容的技術互相串接，並設計了直觀易用的使用者介面（UI），大幅降低了 3D 建模與骨架綁定的技術門檻，使不具備美術背景或 Blender 操作經驗的一般使用者，亦能透過單鍵操作生成高品質的 3D 互動模型。

未來，我們可能只要隨手拍一張家裡的寵物，就能直接把牠變成動畫裡的主角，或是放進 VR 遊戲進行互動；甚至還可以結合 3D 列印技術，生成一個專屬的公仔模型。本專題之成果成功將傳統需由專業建模師耗費大量工時之工作流程進行自動化，為大眾化 3D 內容創作(UGC)提供了可行的解決方案。

8.銘謝

本專題之順利完成，由衷感謝指導教授在研究期間的悉心指導與方向指引，每逢技術串接與實驗遭遇瓶頸時，老師皆能給予關鍵的建議，使本研究得以順利收斂。

同時，感謝實驗室學長姐們在實作過程中提供的環境建置經驗與技術分享，讓我們能克服異質環境串接與實作方面的挑戰。特此表達誠摯謝忱。

9.參考文獻

[1] Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, Xiao Chen, Feipeng Tian, Jianxiong Pan, Zeming Li, Gang Yu, Xiangyu Zhang, Daxin Jiang, Ping Tan. “Step1X-3D: Towards High-Fidelity and Controllable Generation of Textured 3D Assets” 12 May 2025. https://arxiv.org/abs/2505.07747

[2] Jia-Peng Zhang, Cheng-Feng Pu, Meng-Hao Guo, Yan-Pei Cao, Shi-Min Hu. “One Model to Rig Them All: Diverse Skeleton Rigging with UniRig” 16 Apr 2025. https://arxiv.org/abs/2504.1

[3] Yi-Chuan Huang, Jiewen Chan, Hao-Jen Chien, Yu-Lun Liu. “Voxify3D: Pixel Art Meets Volumetric Rendering” 26 Apr 2026. https://arxiv.org/pdf/2512.07834

[4] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, Nicu Sebe. “Bilateral Reference for High-Resolution Dichotomous Image Segmentation” 25 Jul 2025. https://arxiv.org/pdf/2401.03407