Intelligenza Artificiale Generativa per Digital Twin Urbanistico partecipato

Marco Pesic[1], Giovanni Lughi[1], Matteo Roffilli[1]
[1] Bioretics srl, [m.pesic, g.lughi, m.roffilli]@bioretics.com

Abstract. La ricostruzione digitale fotorealistica di monumenti, edifici e arredi urbani generici è fondamentale per rendere il patrimonio culturale e urbanistico accessibile a un pubblico ampio, oltre i confini dei soli professionisti del settore. Le tecniche tradizionali di fotogrammetria sono efficaci ma non efficienti perché comportano costi elevati dovuti alla necessità di attrezzature specifiche per l'acquisizione dei point cloud oltre a software dedicato per il processing e molte ore di calcolo. È necessario inoltre avere a disposizione l'oggetto da scansionare e non sono quindi applicabili *ex-post* a realtà non più disponibili. Allo scopo di superare queste limitazioni e rendere il processo di cloning volumetrico digitale alla portata di tutti i cittadini, in questo lavoro presentiamo l'utilizzo di tecniche innovative di Intelligenza Artificiale Generativa (GenAI) [1] quali NeRF [2] e 3DGS [3], che consentono la creazione di digital twin a partire unicamente da un insieme di fotografie o video, anche quando acquisiti con uno smartphone economico.

Introduzione

L'attuale fotogrammetria ad alta precisione è una pratica costosa: economicamente perché necessita di particolare hardware di acquisizione; temporalmente perché richiede lunghe e complesse preparazioni (Ground Control Point, …) e configurazioni hw/sw, oltre che una padronanza non banale della materia.

Difatti, oltre a dispositivi tutto sommato comuni, come fotocamere reflex e droni, si utilizzano strumenti avanzati, quali scanner laser 3D e sistemi GNSS (Global Navigation Satellite System). Queste attrezzature, tra l'altro molto costose, generano grandi quantità di dati, che richiedono, quasi sempre nella pratica, software proprietari per gestirne l'elaborazione, aumentando così anche i costi legati a storage e licenze software.

Come soluzioni ad oggi meno costose, troviamo alcuni particolari dispositivi mobili – smartphone - dotati di sensori di profondità e LiDAR. Tuttavia, tali dispositivi sono limitati alla ricostruzione di oggetti di piccole dimensioni. Il principale vantaggio rispetto, ad esempio, a una fotocamera reflex è dato dalla possibilità di utilizzare l'odometria visiva. Tale tecnica combina dati provenienti da vari sensori (immagini, LiDAR, accelerometro, giroscopio, …) per determinare la posizione della fotocamera durante la scansione e ciò offre un vantaggio rispetto ai metodi che utilizzano solo immagini, come il Structure from Motion (SfM) di COLMAP, che non riescono a mantenere la scala reale dell'oggetto.

L'obiettivo principale di questi metodi è ottenere misurazioni precise degli oggetti catturati, ma presentano limitazioni nella resa visiva fotorealistica. In particolare, hanno difficoltà a riprodurre superfici riflettenti o specchi d'acqua. Per ovviare a queste problematiche, si stanno diffondendo nuovi modelli di intelligenza artificiale generativa, come "Neural Radiance Field" (NeRF) e "3D Gaussian Splatting" (3DGS), che permettono di ricostruire oggetti in modo fotorealistico utilizzando semplici immagini. Poiché l'accuratezza delle misure non è il focus di questi modelli, l'hardware richiesto può essere più economico e può includere dispositivi mobili con capacità di catturare immagini in alta definizione o fotocamere reflex.

Un altro vantaggio di questi metodi, basati su modelli statistici, è la maggiore velocità nella creazione dei modelli 3D rispetto ai metodi *tradizionali* di fotogrammetria.

Materiali e Metodi

NeRF

Neural Radiance Field rappresenta una scena come un campo di radianza, imparando una funzione $f(x,\theta) \rightarrow (c,\sigma)$. La funzione prende come input una posizione 3D **x** e la direzione di vista θ e come output restituisce il colore del punto **c** con la sua rispettiva densità σ (uno scalare che rappresenta la presenza o meno di un oggetto in quel determinato punto nello spazio **x**). Per ogni punto 3D nello spazio il suo colore può essere renderizzato attraverso un raggio che parte dall'origine della camera $\mathbf{o} r(t) = o + td$, questo permette il campionamento di punti lungo il raggio, tra il near bound e il far bound. Il colore finale che sarà visualizzato in un determinato pixel dato un raggio che passa per esso sarà: $I(r) = \sum_{i=1}^{N} T_i \left(1 - e^{-\sigma_i \delta_i}\right) c_i$ dove $T_i = exp\left(\sum_{j=1}^{i-1} \sigma_j \, \delta_j\right)$
Dove **T** rappresenta la trasmittanza accumulata lungo il raggio e $\delta_j = t_{j+1} - t_j$si riferisce alla distanza tra punti campionati adiacentemente.

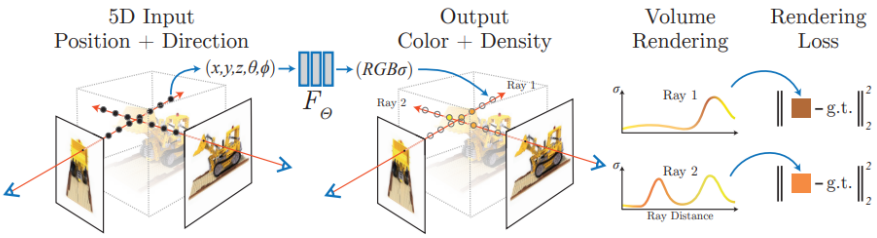

Figura 1: Immagine presa da [2] che mostra la pipeline di allenamento di un modello NeRF

3D Gaussian Splatting

3DGS rappresenta una scena basata su primitive volumetriche (distribuzioni Gaussiane tridimensionali) che hanno un insieme di parametri aggiustabili: posizione μ, matrice di covarianza Σ (che in pratica è decomposta in scala e rotazione), opacità (o), ed armoniche sferiche (SH) che sono dei coefficienti utilizzati per rappresentare il colore. Queste primitive 3D vengono proiettate su uno spazio 2D, (μ′) e (Σ′) sono la media e la matrice di covarianza 2D e rasterizzate usando α-blending. Dove i pesi α-blending vengono calcolati in questo modo

$$\alpha = oG$$

G è la gaussiana proiettata nello spazio 2D dove (x,y) sono gli indici dei pixel dello schermo

$$G(x,y) = e^{\left(\left(\frac{-1}{2}\right)([x,y]^T-\mu')^T \Sigma([x,y]^T-\mu')\right)}$$

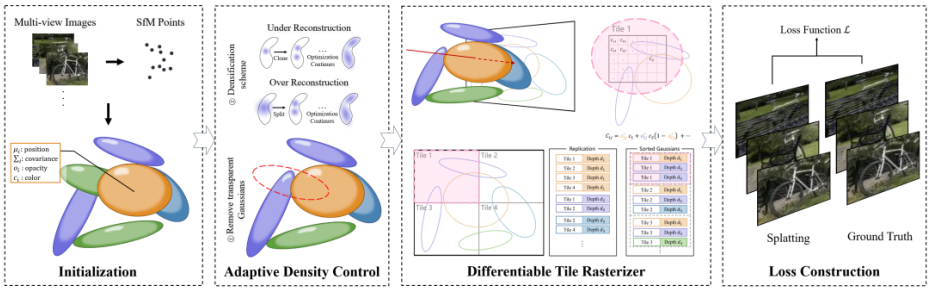

Figura 2: Immagine presa da [10] che mostra la pipeline di allenamento di un modello 3DGS

Metodologia

Il metodo utilizzato per ricreare una scena a partire da un video o una serie di immagini segue i seguenti passaggi: 1) acquisizione video della scena da ricostruire; 2) identificazione delle posizioni delle camere; 3) creazione di un modello volumetrico (ad esempio con NeRF o 3DGS); 4) [opzionale] esportazione del modello 3D nel formato desiderato (Point cloud, mesh, ...). Il quarto passaggio è spesso necessario per integrare i modelli generati in motori di rendering preesistenti. Tuttavia, per visualizzare i risultati non è sempre necessario convertire il modello in un mesh, poiché è possibile utilizzare visualizzatori volumetrici che offrono una resa fotorealistica e immersiva più accurata rispetto al rendering tradizionale con mesh e texture. Questo approccio è stato applicato per la creazione di modelli 3D di statue e monumenti. Allo stato attuale, questi modelli dalla resa di altissima qualità visiva, preservano, come premesso, una minor qualità della precisione delle misure. Tale problema potrebbe essere mitigato adottando, oltre che eventuale hardware specifico, modelli di depth estimation [11], che permettono di estrarre mappe di profondità utilizzando le informazioni presenti nelle sole immagini. Tali mappe di profondità aiuterebbero notevolmente il processo di identificazione delle

posizioni della camera, garantendo una maggiore precisione nel determinare la distanza e le posizioni relative dell'oggetto osservato.

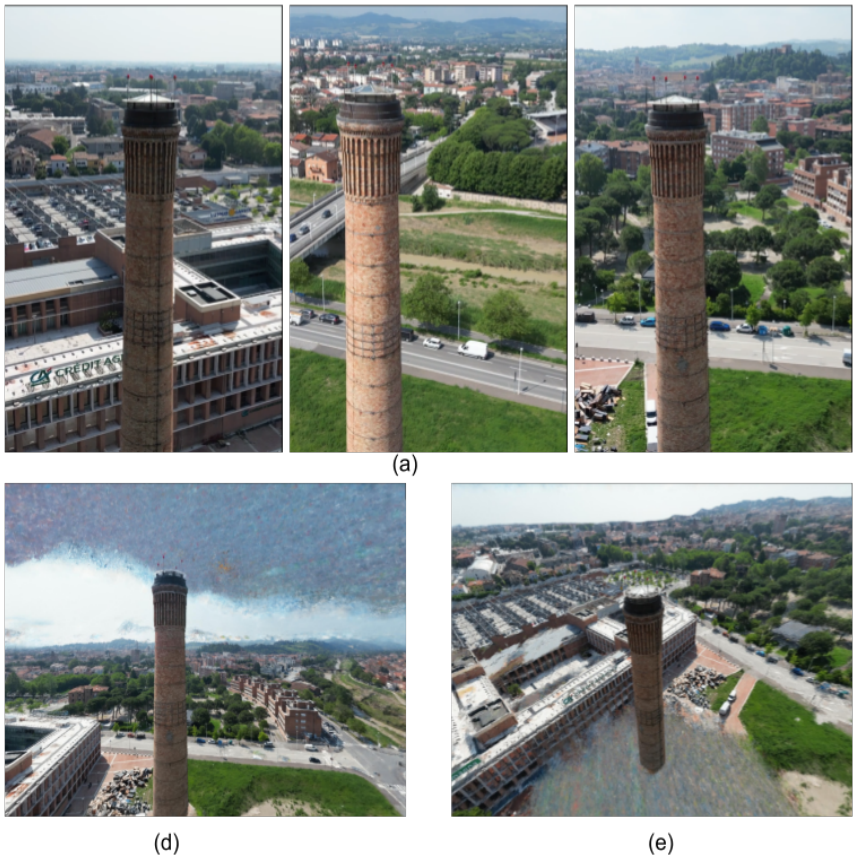

Figura 3: (a) immagine della ex fornace di Cesena catturate da un drone. (d)(e) Immagini generate dalla GenAI in tempo reale utilizzando le immagini catturate dal drone.

Risultati

Attualmente, i due principali metodi volumetrici utilizzati sono NeRF e 3DGS. Come accennato in precedenza, ciascuno di essi presenta vantaggi e svantaggi. NeRF offre una resa fotorealistica leggermente superiore rispetto a 3DGS, ma quest'ultimo si distingue per l'efficienza computazionale, risultando più performante. Grazie a questa efficienza, i modelli generati con 3DGS consentono una visualizzazione fluida, real-time, rendendolo ideale per applicazioni che richiedono reattività immediata.

Nel mese di Maggio 2024 Bioretics ha vinto una competizione [7], per la categoria “educazione e turismo”, grazie all’utilizzo di modelli volumetrici per la ricostruzione di monumenti e opere storiche. Questi modelli possono essere inseriti nel metaverso

(tema centrale della challenge), visualizzati tramite visori AR, utilizzati in qualsiasi visualizzatore che supporti mesh 3D. In particolare, trattandosi di modelli georeferenziati e che aumentano di significato se correttamente posizionati nello spazio, è possibile utilizzarli tramite l'ecosistema Cesium [8](figura 6) per la loro visualizzazione ed interazione: CesiumJS, Cesium for Omniverse, Cesium for Unreal, Cesium for Unity.

Un diverso caso d'uso, purtroppo tristemente attuale, è l'impiego del rendering volumetrico per una visualizzazione più rapida ed immersiva di eventi naturali, come le frane(figura 4), interruzioni di strade, ecc. Questo approccio permetterebbe agli operatori di avere più informazione e più rapidamente anche da remoto, ad esempio con una ripresa aerea fatta con un drone nel primo sopralluogo dei tecnici o addirittura con il video di uno smartphone registrato da un cittadino, usati per generare automaticamente un modello volumetrico che consenta di esplorare lo spazio da e fare le prime constatazioni dalla propria sede o dalla control-room, eventualmente dotata di apparecchi per la realtà virtuale.

Un possibile sviluppo futuro su cui stiamo ragionando insieme a partner istituzionali, prevede l'uso del rendering volumetrico per la generazione di una una mappa delle città alla stregua di "Google Street View", ma più dettagliata e con maggiori possibilità di interazione (come mostrato in [9]).

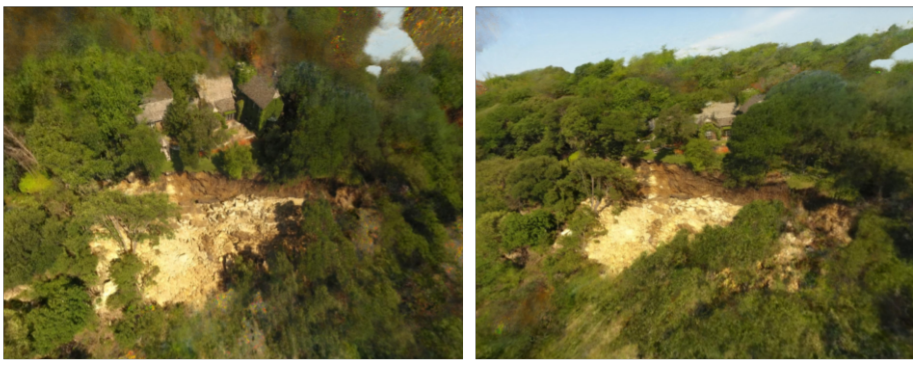

Figura 4: Immagini ricostruite attraverso metodo NeRF, prese dal video di una frana [12]

Altre possibilità riguardano l'adozione di metodi GenAI per rendere nuovamente interagibili realtà passate o non più esistenti ed eventualmente permettere agli utenti un facile raffronto di una certa realtà a distanza di tempo: generazione di modelli a partire da video di qualcosa non più esistente, confronto della stessa realtà a distanza di tempo (ad esempio la crescita di un bosco, di un albero monumentale, di un edificio, ecc.).

Conclusioni

La tecnica proposta si pone quindi sia in alternativa ai metodi attuali sia in estensione agli stessi permettendo la ricostruzione digitale di realtà disponibili solo come foto o video storici. I metodi statistici di GenAI offrono vantaggi significativi rispetto alla fotogrammetria geometrica tradizionale, tra cui tempi di produzione decisamente più rapidi, qualità visiva molto elevata e costi irrisori, a discapito attualmente della precisione numerica e geometrica dei modelli prodotti, che potrebbe in alcuni casi non essere sufficiente per utilizzi di precisione (*e.g.* il calcolo strutturale [4]). Questa tecnologia inoltre favorisce la cittadinanza partecipata in quanto abilita il singolo cittadino ad estendere le attuali piattaforme GIS con oggetti volumetrici di alta qualità visiva (vedi Figura 5) e la creazione di timeline storiche per un confronto temporale tra istanti significativi. Un ulteriore utilizzo previsto è la collocazione in realtà aumentata o metaverso di reperti archeologi, ora nelle teche dei musei, nella loro originaria posizione geografica. Utilizzando il nostro servizio web [6] è possibile riprodurre le immagini qui presentate a partire dai video originali.

Figura 5: (a) immagine della statua del Bufalini in Cesena generata dalla GenAI in tempo reale. (b) le uniche viste di Google Maps e Google Street View. (c) altri punti di vista generati da GenAI.

I modelli ottenuti possono poi essere esportati nei comuni formati già in uso per utilizzi tecnici e di videoproduzione.

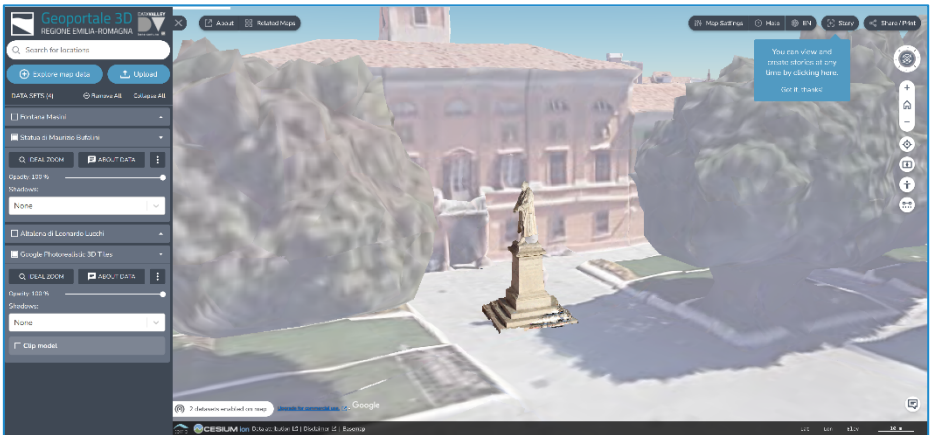

Figura 6: statua del Bufalini in mesh generata dal metodo proposto, inserita nella copia del Geoportale della Regione Emilia-Romagna [5] con attivo il layer 3D di Google Maps.

Riferimenti bibliografici

1. Christopher M. Bishop, Hugh Bishop, Deep Learning Foundations and Concepts, Springer Cham, https://doi.org/10.1007/978-3-031-45468-4 (2024).
2. Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. arXiv:2003.08934 (2020).
3. Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis: 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ArXiv 2308.04079 (2023).
4. Mazzacca, G., Karami, A., Rigon, S., Farella, E. M., Trybala, P., and Remondino, F.: NERF for heritage 3D reconstruction, Int. Arch. Photogramm. Remote Sens. SpatialInf.Sci. (2023).
5. Bioretics GIS https://twin.bioretics.com, consultato il 24/05/2024.
6. Bioretics Cloneme https://cloneme.bioretics.com, consultato il 24/05/2024.
7. Clust-er https://innovate.clust-er.it/challenge-metaverso-e-realta-immersive-per-start-up-e-pmi/
8. Cesium https://cesium.com/
9. Kerbl, Bernhard and Meuleman, Andreas and Kopanas, Georgios and Wimmer, Michael and Lanvin, Alexandre and Drettakis, George: A Hierarchical 3D Gaussian Representation for Real-Time Rendering of Very Large Datasets http://www-sop.inria.fr/reves/Basilic/2024/KMKWLD24
10. Siting Zhu and Guangming Wang and Dezhi Kong and Hesheng Wang: 3D Gaussian Splatting in Robotics: A Survey https://arxiv.org/abs/2410.12262
11. Lihe Yang and Bingyi Kang and Zilong Huang and Zhen Zhao and Xiaogang Xu and Jiashi Feng and Hengshuang Zhao: Depth Anything V2 https://arxiv.org/abs/2406.09414
12. Drone video from Austin Watershed Protection Dept. shows 2018 Shoal Creek landslide damage | KVUE https://youtu.be/ySlF8dQzyZM?si=47xGG-7FwfkQ6tQe