AI-aided Design? Processi *text-to-image* per il disegno di architettura

Matteo Flavio Mancini, Sofia Menconero

Abstract

L'Intelligenza Artificiale (AI) sta segnando una svolta in molti campi della vita umana ed è opportuno interrogarsi sulla sua possibilità di utilizzo nei processi di rappresentazione del progetto di architettura.
Il contributo presenta una breve digressione sul passato recente delle tecnologie AI al fine di spiegarne il funzionamento, una fotografia sull'attuale stato dell'arte dai processi text-to-image *a quelli* image-to-3D, *concentrandosi in particolare sulla piattaforma* StableDiffusion, *oltre a proporre una panoramica sui più recenti studi nel campo del progetto di architettura. La successiva sperimentazione diventa occasione per mostrare le potenzialità dell'AI quanto al processo di co-creazione e alla possibilità di simulare diverse tecniche grafiche, fino alla visualizzazione fotorealistica. D'altro canto, vengono presentati i limiti che, allo stato attuale dello sviluppo, invalidano talvolta i risultati dei processi* text-to-image *per quanto riguarda gli aspetti scientifici della rappresentazione.*
Le conclusioni propongono una riflessione sulle differenze tra intelligenza umana e artificiale, sul tema dell'autorialità condivisa uomo-macchina e sulle loro conseguenze per il progetto d'architettura.

Parole chiave: intelligenza artificiale, text-to-image, *disegno di progetto, autorialità,* stablediffusion.

Introduzione

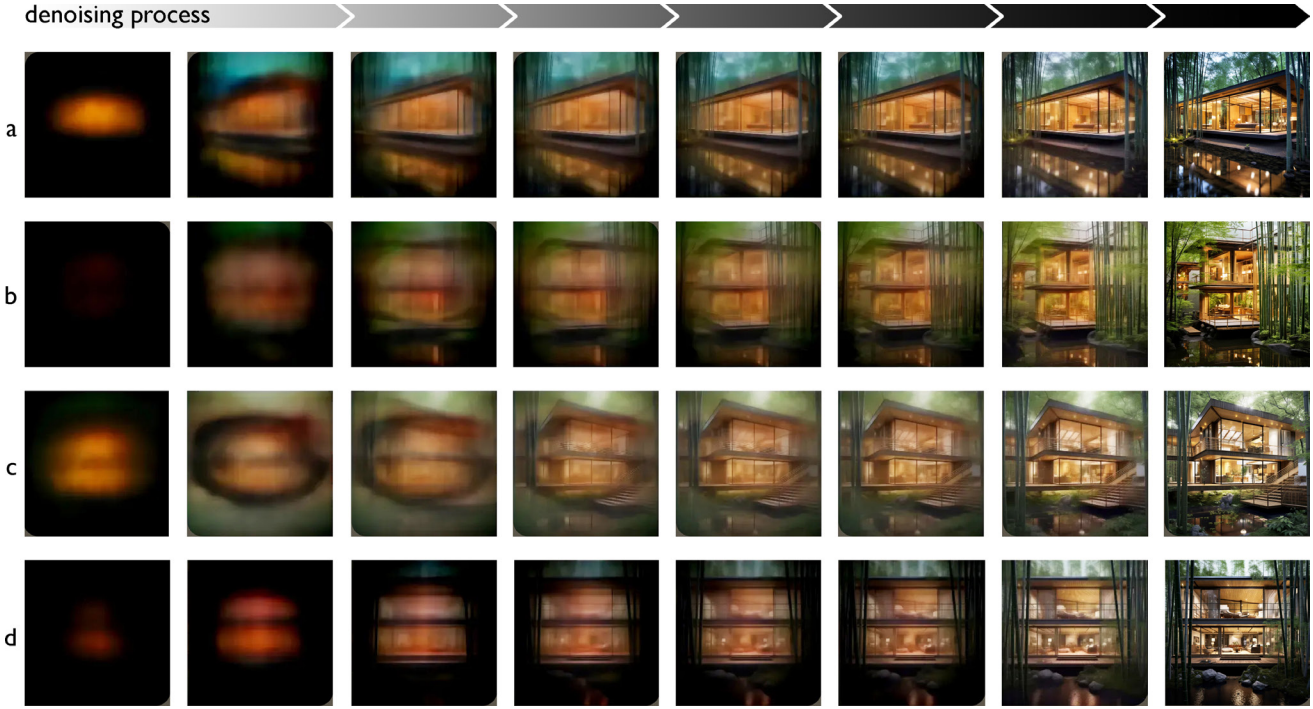
L'architettura e il disegno di architettura hanno attraversato negli ultimi trent'anni svolte importantissime. Il *first digital turn* [Carpo 2013] ha visto l'introduzione della rappresentazione digitale negli anni Novanta del XX secolo mentre il *second digital turn* [Carpo 2017] si è avviato con la diffusione di algoritmi e *big data* a partire dagli anni '10 del XXI secolo. Dieci anni dopo, stiamo assistendo a un'altra potenziale svolta dovuta a una repentina accelerazione dello sviluppo e della diffusione di strumenti di Intelligenza Artificiale (AI), già in uso nei maggiori studi di architettura come Coop Himmelb(l)au [Prix et al. 2022], Zaha Hadid Architects [Wallish 2022] e Foster + Partners [Tsigkari et al. 2021].
Una branca dell'AI, basata su processi di tipo *text-to-image*, propone soluzioni facilmente accessibili e dedicate alla creazione di immagini. Si tratta di un modello di *machine learning* che usa un linguaggio naturale descrittivo come input e produce un'immagine basata sull'elaborazione della descrizione fornita. I risultati ottenuti da queste piattaforme sono sorprendenti in termini di corrispondenza ai *prompt* testuali inseriti e di flessibilità delle tecniche grafiche che sono in grado di (ri)produrre.
Partendo dal presupposto che, allo stato attuale, queste AI non hanno alcuna coscienza creativa né un'effettiva capacità di comprendere le regole compositive e proiettive o la spazialità rappresentata nelle immagini, è comunque opportuno interrogarsi sulle loro possibilità di utilizzo nei processi di rappresentazione del progetto di architettura. Con questo obiettivo e tenendo conto delle caratteristiche

intrinseche di questa tecnologia, che verranno esposte nei successivi paragrafi, si propone la sperimentazione dell'AI *text-to-image* attraverso la piattaforma open-source *StableDiffusion* per la realizzazione di immagini prospettiche capaci di contribuire alle fasi preliminari di ideazione del progetto.

Passato (recente) e presente della generazione di immagini basata sull'AI

I sistemi generativi basati sull'AI che è possibile fruire nel campo dell'architettura e del design sono in rapida evoluzione. Un punto di svolta nella ricerca sulla generazione di immagini è segnato dall'invenzione delle *Generative Adversarial Networks* (GAN) nel 2014 [Goodfellow et al. 2014]. Si tratta di un'architettura di *deep learning* in cui due reti neurali antagoniste (un generatore e un discriminatore) interagiscono reiterativamente durante l'addestramento al fine di giungere al punto in cui il discriminatore non sia più in grado di distinguere le immagini sintetiche prodotte dal generatore rispetto alle immagini reali immesse come dati di addestramento. Nel 2016 è stata sviluppata un'architettura GAN in grado di generare immagini plausibili da descrizioni testuali dettagliate [Reed et al. 2016] avviando di fatto il sistema AI *text-to-image*. Un ulteriore avanzamento è stato segnato da un più efficiente metodo di apprendimento basato sull'elaborazione del linguaggio naturale, chiamato CLIP (*Contrastive Language-Image Pre-training*) [Radford et al. 2021]. Questo modello di classificazione delle immagini identifica gli oggetti imparando dal testo associato a un'immagine (piuttosto che da etichette assegnate manualmente) ed è stato allenato su 400 milioni di coppie immagini-testo estratte dal web. I modelli CLIP

Fig. 1. Processo di denoising durante la generazione di quattro varianti (a, b, c, d) in Midjourney *attraverso il prompt: a modern Japanese house in a bamboo forest in spring (elaborazione degli autori).*



sono in grado di stimare la conformità di un'immagine generata per un *prompt* testuale [Colton et al. 2021]. Un esempio di questo abbinamento è il sistema di generazione di immagini chiamato *VQGAN-CLIP*, che utilizza una rete neurale GAN ancora più potente. I contributi significativi dell'architettura *VQGAN-CLIP* riguardano la qualità visiva sia nella generazione che nella manipolazione delle immagini, la fedeltà semantica tra testo e immagine generata, l'efficienza dovuta al fatto che il metodo non richiede ulteriore addestramento oltre ai modelli pre-allenati e il valore dello sviluppo e della scienza aperta [Crowson et al. 2022, p. 2]. Successivamente, i sistemi basati su GAN sono stati sostituiti con i *diffusion model*, ovvero modelli probabilistici di *machine learning* addestrati a eliminare il rumore delle immagini precedentemente introdotto, imparando a invertire il processo di diffusione [Dhariwal et al. 2021]. L'addestramento di questi modelli li rende in grado di utilizzare i metodi di *denoising* per sintetizzare nuove immagini, prive di rumore, da input casuali (fig. 1).
Alcune applicazioni dell'AI fruibili nel campo dell'architettura e del design sono le seguenti:

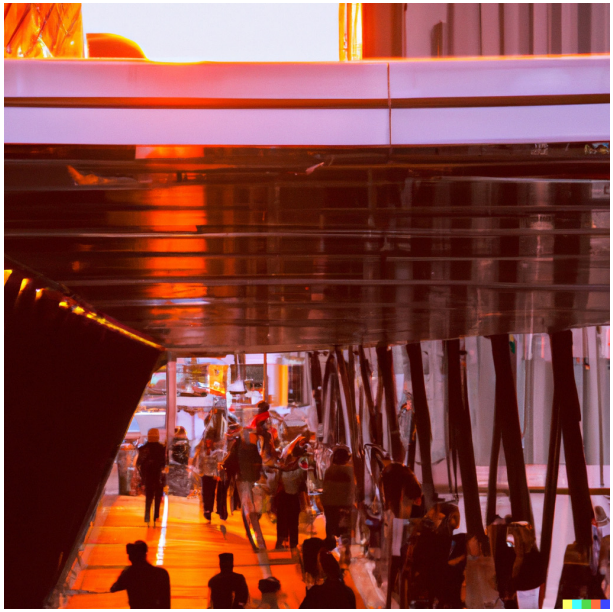
- *text-to-image*, la più diffusa operazione di generazione di immagini attraverso una descrizione testuale, spesso associata ad altre funzionalità;
- *image-to-image*, per la trasformazione di un'immagine input in modo che corrisponda alle caratteristiche di un'immagine di destinazione, può essere usata per trasferire uno stile, per modificare o rimuovere oggetti dalle immagini (*inpainting*), per trasformare a colori un'immagine in bianco e nero, per aumentare la risoluzione di un'immagine (*upscaling*);
- *text* o *image-to-video*, per creare video da un *prompt* testuale (ad esempio *Make-a-Video* [Singer et al. 2022], o *CogVideo* [Hong et al. 2022]) oppure per creare un'animazione grazie al montaggio di immagini generate attraverso l'*image-to-image* (ad esempio *Deforum*) con effetti simili a un video in *stop-motion*.
- *text* o *image-to-3D*, per generare modelli 3D da un *prompt* testuale (ad esempio *Point-E* per generare nuvole di punti [Nichol et al. 2022], *Shape-E* per mesh texturizzate [Jun et al. 2023]), oppure la generazione dei modelli 3D può avvenire a partire da un'immagine (ad esempio *Kaedim*).

L'incredibile recente diffusione dell'AI *text-to-image* deriva dall'attivazione di alcune piattaforme con interfacce semplici da utilizzare, anche da parte di fruitori non esperti, come *DALL-E 2*, *Midjourney* e *StableDiffusion*.

Le principali piattaforme per l'AI *text-to-image*

DALL-E è la prima tra le tre piattaforme a essere stata presentata nel gennaio 2021 (l'attuale versione 2 è di aprile 2022) da *OpenAI* [Ramesh et al. 2021], gli stessi sviluppatori di *ChatGPT*. La piattaforma, fruibile online in abbonamento, propone quattro funzioni: la generazione di immagini realistiche e artistiche da una descrizione testuale che può combinare concetti, attributi e stili (fig. 2); l'*outpainting*, ovvero l'espansione dell'immagine oltre i margini originali attraverso la creazione di una nuova composizione; l'*inpainting*, tramite il quale è possibile modificare porzioni di immagine aggiungendo o eliminando degli oggetti attraverso una descrizione testuale e mantenendo coerente il resto della scena; la generazione di variazioni ispirate a un'immagine di input.
Midjourney, rilasciata il 12 luglio 2022, è attualmente giunta alla versione 5.2 con notevoli miglioramenti rispetto agli esordi in termini di aderenza al *prompt* e fotorealismo (fig. 3) e conta, dopo un anno, oltre 15 milioni di utenti [1].

Fig. 2. Immagine generata con DALL-E 2 *attraverso il prompt: a modern building on a crowded street at sunset (elaborazione degli autori).*



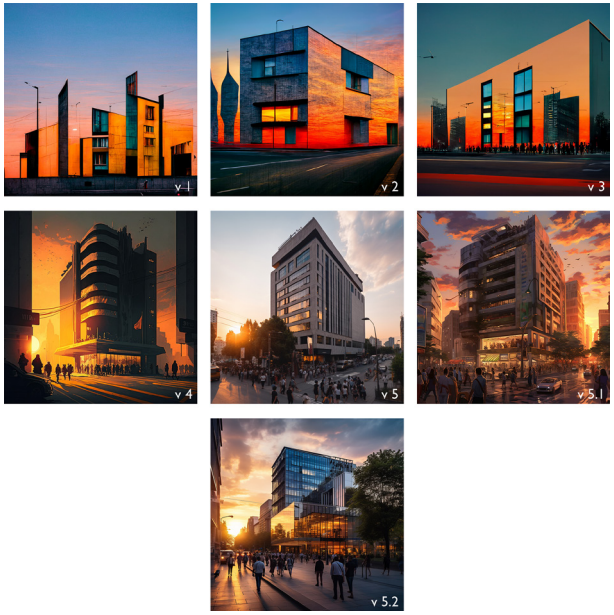


Fig. 3. Confronto tra diverse immagini generate a partire dallo stesso testo (prompt: a modern building on a crowded street at sunset) relative a diverse versioni di Midjourney *(elaborazione degli autori).*

Fig. 4. Confronto tra diverse immagini generate a partire dallo stesso testo (prompt: a modern building on a crowded street at sunset) relative a diverse versioni di StableDiffusion *(elaborazione degli autori).*

Come *DALL-E*, è fruibile online in abbonamento. Le tre principali attività generative su *Midjourney* sono: un'immagine a partire da un *prompt* testuale; una descrizione a partire da un'immagine; un'immagine di sintesi a partire da due fino a cinque immagini in input. *Midjourney* (come *StableDiffusion*) permette di utilizzare anche un *prompt* negativo nel caso non si vogliano specifici elementi nell'immagine generata.
StableDiffusion, rilasciata ad agosto 2022, è l'unica delle tre piattaforme a essere open-source e si basa su un *diffusion model* chiamato *latent diffusion model* [Rombach et al. 2022]. L'attuale versione beta XL è disponibile solo online in abbonamento, mentre le precedenti versioni possono anche essere installate in locale gratuitamente. *StableDiffusion* supporta la generazione di immagini attraverso l'uso di un *prompt* di testo che descrive gli elementi da includere o escludere dall'output (fig. 4), l'*inpainting* e l'*outpainting*, la generazione *image-to-image* e l'*upscaling*. È inoltre possibile associare a *StableDiffusion* delle estensioni come *ControlNet*, che genera variazioni di un'immagine di input attraverso descrizioni testuali, e *Deforum*, che attraverso la funzione *image-to-image* genera una serie di immagini, applicando piccole trasformazioni, e le cuce insieme per creare un video.
Il vantaggio più grande di *StableDiffusion* rispetto alle altre piattaforme è la possibilità che gli utenti finali possano implementare un addestramento aggiuntivo (*fine-tuning*) per ottimizzare gli output di generazione in modo che corrispondano a casi d'uso più specifici. Ad esempio, negli studi di architettura dove l'AI è entrata a far parte del processo creativo, la rete neurale viene allenata con immagini mirate del repertorio progettuale dello studio al fine di ottenere risultati più in linea con il linguaggio architettonico e grafico.
Dunque, a differenza delle precedenti piattaforme, *StableDiffusion* permette una maggiore libertà di utilizzo in termini di personalizzazione del processo generativo, per questo motivo è stata scelta per la successiva sperimentazione, associata all'estensione *ControlNet* [Zhang et al. 2023] la quale migliora il controllo degli output.

Quest'ultima è una struttura di rete neurale progettata per gestire modelli di diffusione incorporando condizioni aggiuntive: manipolando le condizioni di input dei blocchi riesce a controllare ulteriormente il comportamento generale di un'intera rete neurale. *ControlNet* agisce a partire da un'immagine in input e una descrizione testuale, e permette di ottenere delle immagini che sono variazioni conformi all'input dal punto di vista compositivo ma che seguono anche la descrizione impostata. Il processo a cui i dati sono sottoposti prevede, innanzitutto, la generazione di una mappa basata sull'immagine input (chiamata fase di annotazione o pre-processamento) la quale viene usata dalla rete per generare le varianti con le caratteristiche descritte testualmente (fig. 5).

Avanti e indietro nello spazio latente: tra allenamento e generazione nel modello di diffusione di *StableDiffusion*

Per approcciare correttamente le sperimentazioni che seguiranno nei prossimi paragrafi è importante cercare di comprendere non tanto gli aspetti prettamente tecnico-informatici quanto i processi attuati da questo tipo di AI, nello specifico *StableDiffusion*, nei due momenti distinti dell'allenamento e della generazione, poiché da essi dipendono sia l'uso appropriato che l'interpretazione critica di questa tecnologia.
I modelli di diffusione mutuano dalla termodinamica il concetto di diffusione, cioè il fenomeno per cui le particelle di un fluido si muovono randomicamente all'interno di un altro fluido con diversa concentrazione, fino a raggiungere una nuova condizione di equilibrio. Allo stesso modo, le immagini delle AI durante la generazione sembrano progressivamente emergere dal caos del rumore digitale. Il principio della diffusione viene usato sia in fase di allenamento (*forward diffusion*), che in fase di generazione (*reverse diffusion*). In *StableDiffusion*, entrambi questi processi avvengono nel *latent space*, uno spazio numerico/informativo in cui le immagini vengono tradotte in tensori (matrici a più dimensioni) per lavorare su una loro versione compressa, più leggera del *pixel space* iniziale delle immagini. Anche i testi che descrivono le immagini subiscono una simile traduzione e compressione. L'analogia tra rappresentazione latente dei testi e delle immagini è importante

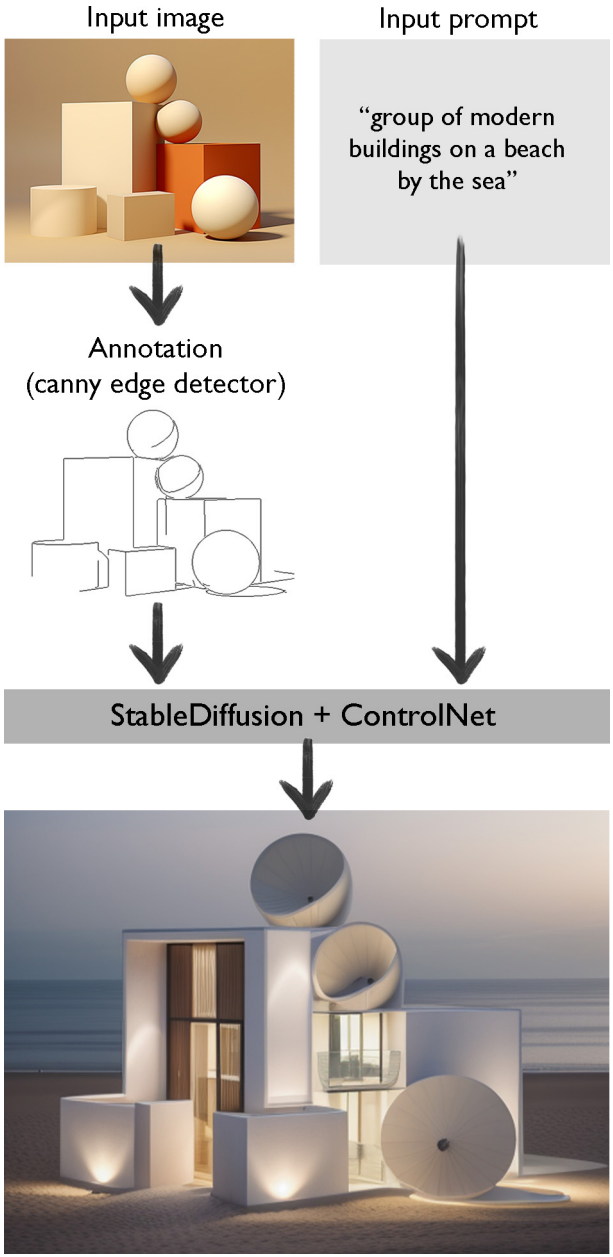
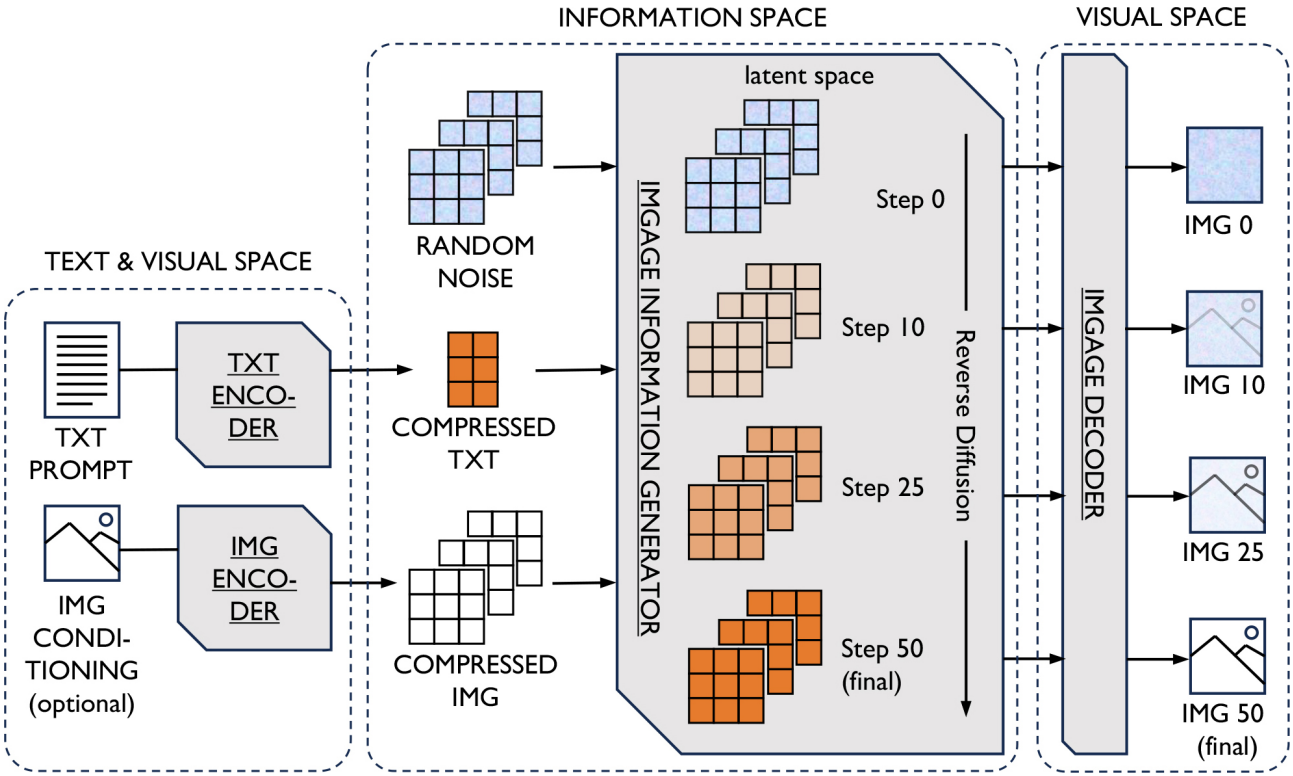


Fig. 5. Schema della generazione di un'immagine attraverso StableDiffusion *con il condizionamento (canny edge) di ControlNet (elaborazione degli autori).*

perché fa capire come le AI non immagazzinino e non elaborino raccolte di sillabe, parole o porzioni di immagini ma operino su rappresentazioni numeriche astratte delle caratteristiche delle immagini, degli oggetti rappresentati, delle possibili situazioni ambientali e delle varie tecniche e stili. Il *latent space* può essere immaginato come il luogo dove l'AI memorizza, in fase di allenamento, e da cui richiama, in fase di generazione, le proprie 'conoscenze'.
L'allenamento di queste AI non è progressivo nel tempo ma avviene prima della pubblicazione, pertanto le loro "conoscenze" sono statiche e aggiornate periodicamente con il progredire delle versioni. Ad esempio, *StableDiffusion* è allenato sul dataset open-access LAION 5B, composto da 5 miliardi di coppie immagine-didascalia il cui contenuto è esplorabile, a partire dall'immissione di un *prompt* testuale, attraverso un apposito portale [2]. La consultazione del dataset di allenamento permette di farsi un'idea sulle corrispondenze tra termini e immagini e, quindi, su cosa sia possibile aspettarsi dai risultati delle elaborazioni: una ricerca che non restituisca risultati coerenti indica che non potranno essere generate delle immagini che corrispondano alle aspettative per quell'input testuale. In fase di addestramento, le immagini del database vengono elaborate introducendo pattern casuali di rumore di diversa intensità. Le immagini così elaborate, insieme alle didascalie corrispondenti, vengono sottoposte all'AI per allenarla a individuare il tipo di pattern adottato, la quantità di rumore introdotta e a rimuovere entrambi per migliorare la qualità delle immagini. In questo modo, attraverso il processo del *forward diffusion*, l'AI apprende contemporaneamente sia

Fig. 6. Schema delle fasi del processo di generazione di un'immagine attraverso il modello di diffusione adottato da StableDiffusion *(elaborazione degli autori).*



come ottenere immagini prive di rumore sia le corrispondenze tra immagini e testi.
Quanto appreso viene utilizzato da *StableDiffusion* per sviluppare un processo generativo che parte e termina in uno spazio in cui i dati (testi e immagini) sono adatti alla percezione umana, attraversa uno spazio puramente informativo (spazio latente) in cui i dati sono rappresentati da *token* (testi) e *tensori* (immagini).
Il processo generativo può essere suddiviso in tre blocchi fondamentali (fig. 6). Il primo prevede la compressione e trasformazione in informazioni numeriche attraverso *encoder* (reti neurali appositamente allenate) dei dati input inseriti per condizionare la generazione dell'immagine. In *StableDiffusion*, grazie all'estensione *ControlNet*, gli input testuali possono essere integrati da condizionamenti grafici opzionali. Nel secondo blocco, attraverso il processo di *reverse diffusion*, avviene l'elaborazione degli input in relazione alle conoscenze note. Tale processo è reiterativo e passa attraverso diversi step di *denoising* per affinare la corrispondenza tra gli input immessi e l'immagine generata. In questa fase l'elaborazione avviene a livello di informazioni numeriche e non c'è alcuna elaborazione grafica di immagini. Quest'ultima avviene nel terzo blocco, dove le rappresentazioni numeriche vengono tradotte da una rete neurale con funzione di *decoder* in immagini visualmente percepibili [Rombach et al. 2022].

Studi relativi all'AI applicata al progetto di architettura

Alcuni studi che riguardano l'AI nell'ambito della progettazione architettonica sono incentrati nell'evidenziare le potenzialità e i limiti della tecnologia. Nella maggior parte dei casi il potenziale è riscontrato come supporto nel processo creativo [Jaruga-Rozdolska 2022; Paananen et al. 2023]. Tra le altre potenzialità sono citate l'abilità di poter immaginare forme astratte, re-immaginare l'architettura biomimetica, rivisitare l'architettura tradizionale e visualizzare avanzamenti fotorealistici a partire da schizzi architettonici. I limiti individuati sono relativi alla possibilità di controllo e personalizzazione dei processi, alla scarsa considerazione per quanto riguarda gli aspetti di fattibilità strutturale, all'eventuale incoerenza stilistico-architettonica dei risultati generati [Hegazy et al. 2023]. I casi studio relativi al progetto di architettura su cui l'AI è stata applicata riguardano la fase ideativa, la generazione di schizzi con specifici stili grafici, l'aggiunta di persone e oggetti in immagini esistenti, la combinazione di varie parti di immagini in una composizione coerente, la variazione di un'immagine iniziale, la variazione dello stile grafico di un'immagine esistente, il disegno planimetrico, il design di esterni e interni, la creazione di texture, il progetto urbano [Ploenning et al. 2022; Yildirim 2022].
Uno studio didattico riguarda l'integrazione delle tecniche di AI alle tecniche tradizionali in un corso di rappresentazione del design al primo anno universitario, dove gli autori hanno notato un miglioramento delle capacità interpretative e compositive degli studenti [Tong et al. 2023]. Agli studenti era stato chiesto di creare una composizione di solidi e di disegnare a mano proiezioni ortogonali e assonometria isometrica; poi di generare una serie di immagini con *Midjourney* attraverso alcune parole chiave; infine di combinare le due produzioni precedenti mediante varie tecniche.

Potenzialità dell'AI *text-to-image* per il disegno di architettura

Per sperimentare il possibile contributo dell'AI nella fase preliminare del progetto, il momento in cui la rappresentazione contribuisce all'ideazione e alla prefigurazione, si è deciso di lavorare sia riguardo alla definizione dell'idea che alla sua visualizzazione.
Sono stati ipotizzati tre diversi input grafici: due viste prospettiche esterne di un modello tridimensionale volumetrico e uno schizzo al tratto di un interno, tutti volutamente privi di caratterizzazioni se non quelle minime indispensabili per la definizione spaziale e l'impostazione dell'inquadratura. Questi input grafici, grazie all'estensione *ControlNet*, hanno il compito di inserire nel processo generativo l'impostazione morfologica generale del progetto mentre gli input testuali vengo utilizzati per descrivere le tecniche grafiche desiderate ed eventuali caratteristiche delle architetture in fatto di materiali, contesto e ulteriori caratteristiche stilistiche che si desidera inserire. I risultati di queste prime sperimentazioni dimostrano la notevole flessibilità dell'AI nel (ri)creare tecniche grafiche diverse, che variano dal disegno a lapis, alle matite colorate, all'acquerello, con una notevole capacità di integrazione di elementi di contesto sia naturali che artificiali. Contemporaneamente, l'aggiunta da parte dell'AI di elementi di dettaglio quali trame, bucature e materiali, contribuisce all'avanzamento dell'ideazione in un processo in cui si può ipotizzare che alcuni di questi elementi possano essere effettivamente inseriti nel prosieguo del progetto, in uno scambio uomo-macchina reiterato (fig. 7).

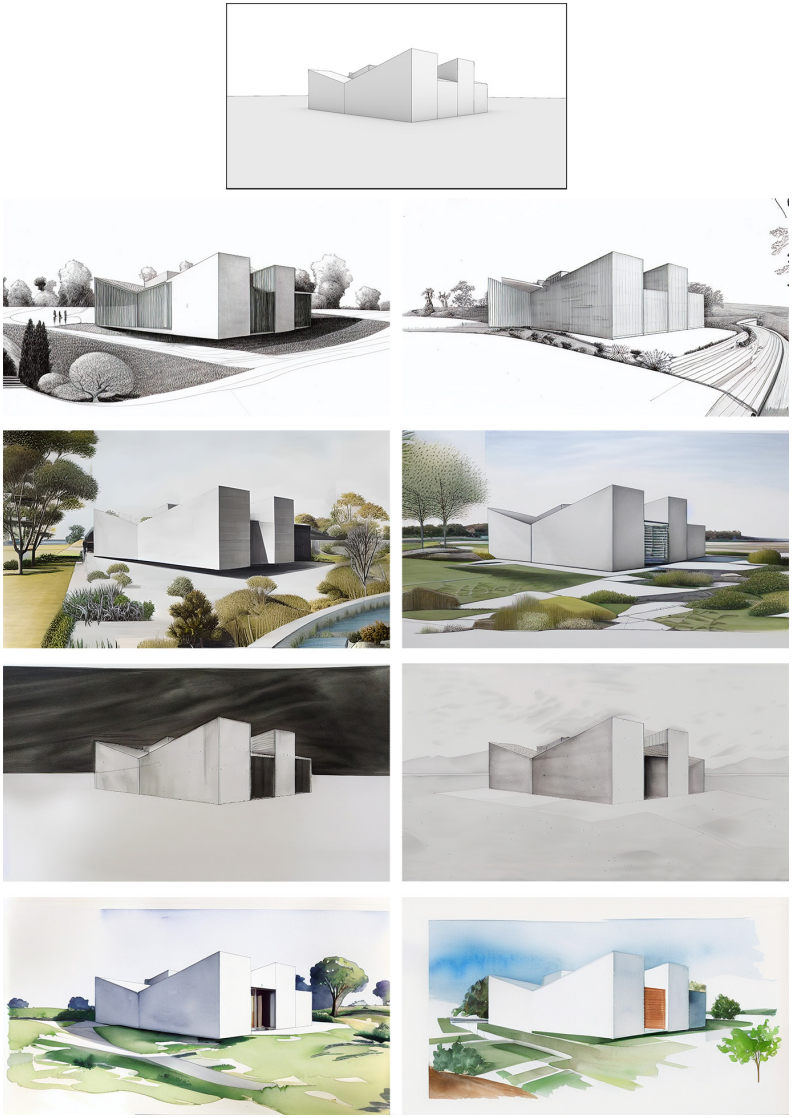


Fig. 7. Immagini generate con StableDiffusion *per simulare diverse tecniche grafiche. Dall'alto verso il basso: lapis, matite colorate, acquerello monocromatico e a colori. Prompt: linear, exterior view, contemporary architecture, highly detailed architecture, large windows, concrete, architectural drawings, technical drawings, [tecnica grafica desiderata], line drawings, working drawings, architectural sketches, conceptual style, abstract (elaborazione degli autori).*

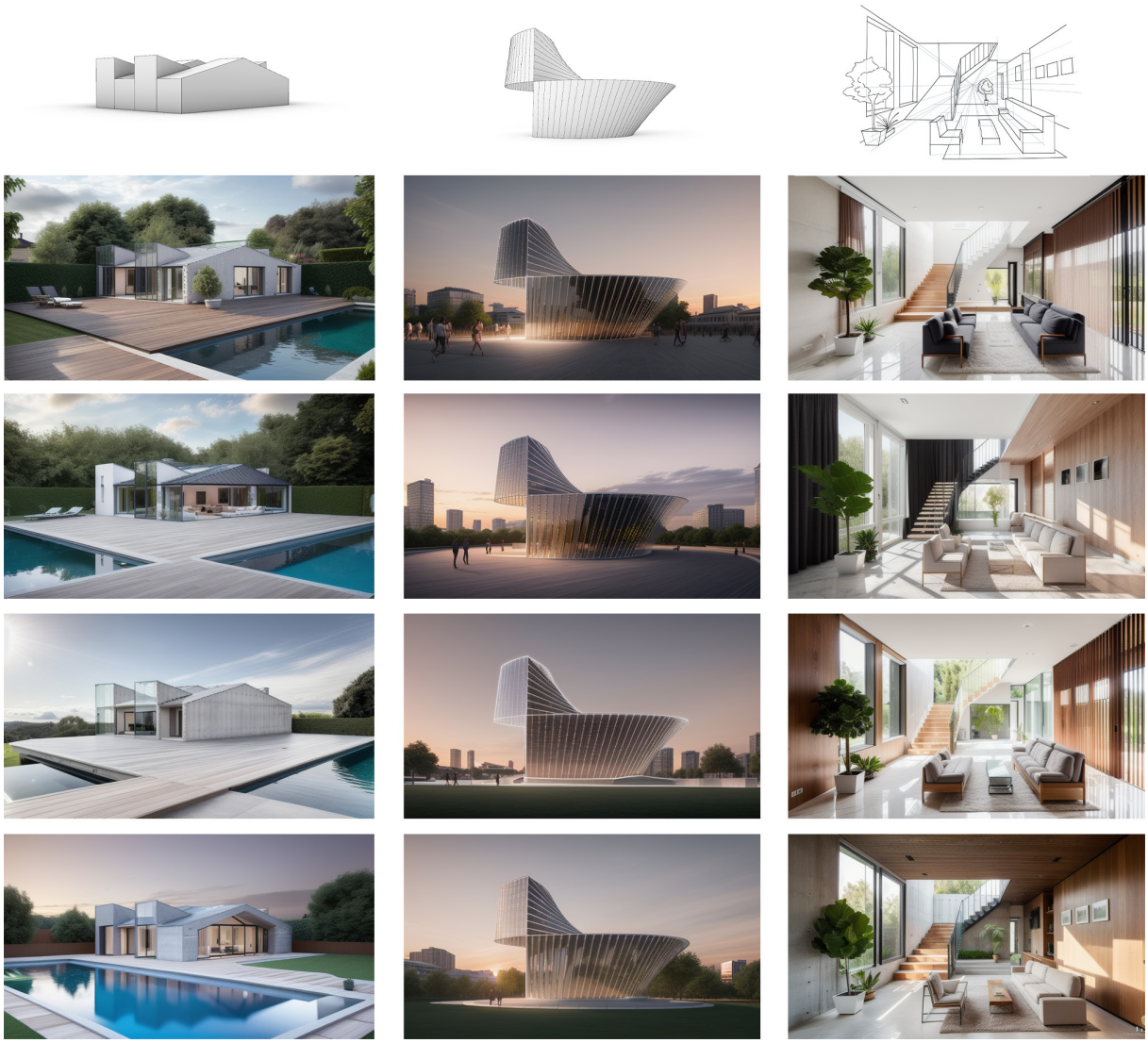


Fig. 8. Immagini generate con StableDiffusion *per visualizzazioni fotorealistiche. A sinistra, viste esterne, prompt: exterior home view, concrete walls and roof, large glass windows, small rectangular swimming pool, garden, garden furniture, clouds. Al centro, viste esterne con superfici curve, prompt: a pavillion in a contemporary architecture style, covered with reflective panels, surrounded by a round pool with people and trees. A destra, viste interne, prompt: home interior view, modern architecture, large glass windows with curtains, timber framing, wood flooring, concrete ceiling, steel staircase, large sofa with pillows, armchair, coffee table with flowerpot, carpet, plants, lamp, minimalist style furniture, sunlight from windows, daylight (elaborazione degli autori).*

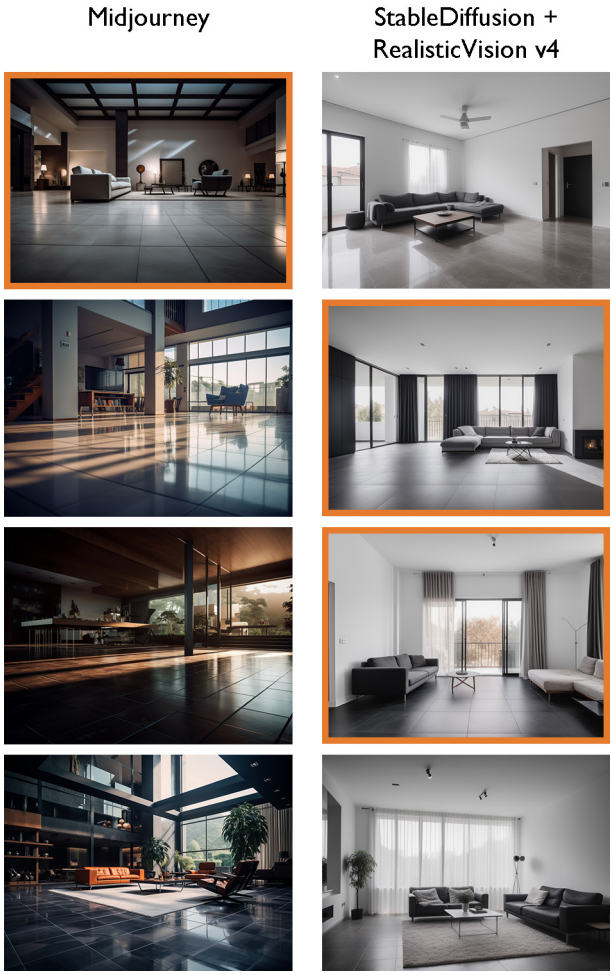


Fig. 9. Immagini generate attraverso il prompt: central perspective, home interior view, floor with regular dark square tiles, modern architecture, minimalist style furniture, daylight (elaborazione degli autori).

Il possibile contributo in termini di definizione dell'idea attraverso la rapida generazione di varianti è più evidente se si richiede all'AI di produrre immagini fotorealistiche. In questo caso si apprezza maggiormente la capacità di proporre variazioni a partire da quanto richiesto tramite il *prompt* testuale. Le sperimentazioni condotte sulle viste esterne mostrano la varietà di materiali e interpretazioni dei semplici schemi volumetrici proposti come input, nonché l'abilità nella creazione di contesti di ambientazione (fig. 8). Analogamente, le sperimentazioni basate su uno schizzo digitale al tratto di un ambiente interno mettono in luce la capacità di accostamento cromatico e dei materiali ma anche la propensione ad aggiungere elementi quali tende e sopralzi del pavimento. Compaiono anche elementi di piccole dimensioni, quali punti luce e complementi di arredo. La distribuzione di queste integrazioni appare in linea di massima coerente con l'impostazione generale.

Limiti dell'AI *text-to-image* nella rappresentazione

I limiti indagati nel presente paragrafo [3] riguardano in particolare l'aspetto della rappresentazione architettonica (prospettiva, riflessioni, illuminazione/ombre).
Allo stato attuale, l'AI non ha alcuna coscienza delle regole proiettive sottese a una corretta costruzione prospettica. Se da un lato questa asserzione era deducibile sulla base dei principi teorici dietro la tecnologia, essa trova anche conferma su base sperimentale. Andando ad aggiungere, nel *prompt* testuale, una parte descrittiva riguardante il metodo di rappresentazione (*central perspective*) [4] si giunge a dei risultati in cui la prospettiva centrale è presente solo in una parte delle immagini generate (fig. 9). Andando successivamente ad analizzare l'impianto prospettico di due delle precedenti immagini generate, si osserva che le linee di fuga delle piastrelle quadrate del pavimento (rette orizzontali perpendicolari al piano di quadro) non individuano un punto univoco di convergenza (fig. 10). Inoltre, tracciando le diagonali delle piastrelle quadrate dai due estremi visibili nelle immagini, si nota che le intersezioni intermedie non sono perfettamente allineate alle diagonali. Dunque, le prospettive sono percettivamente efficaci ma non sono proiettivamente corrette. I risultati della sperimentazione prospettica fanno supporre che l'AI non sia stata addestrata a riconoscere correttamente i diversi metodi della rappresentazione.

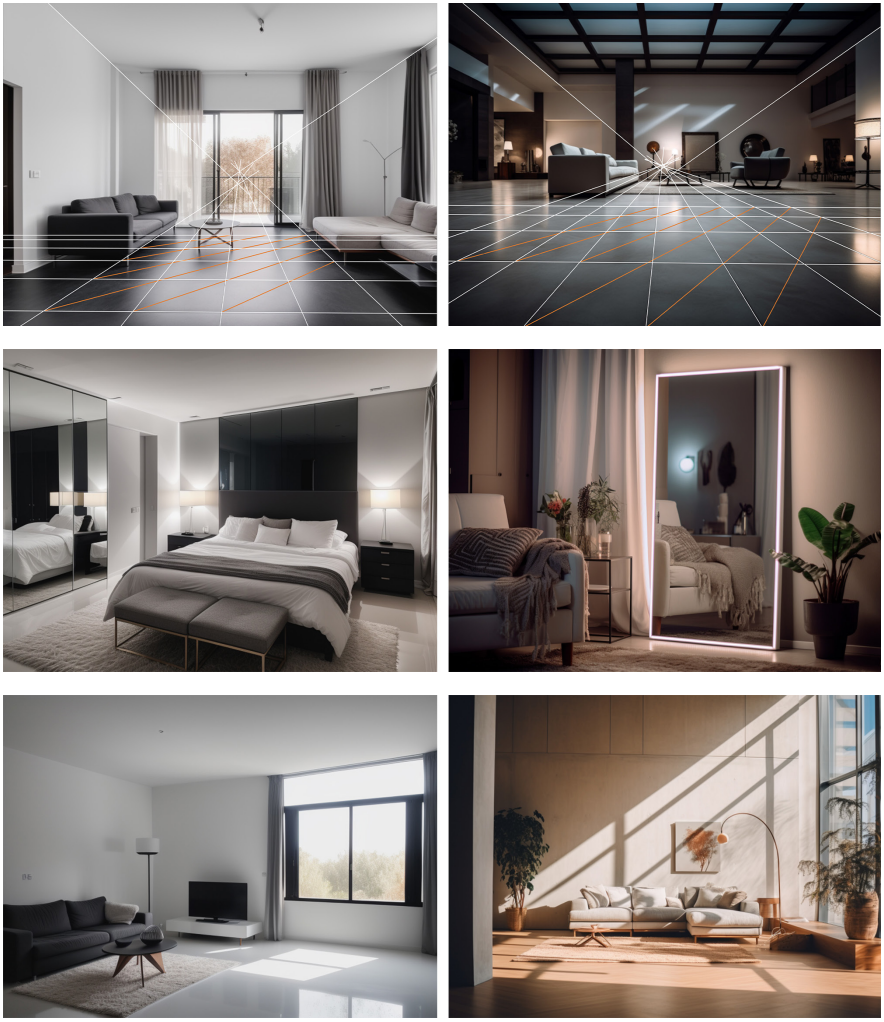


Fig. 10. Analisi prospettica di due delle precedenti immagini generate con StableDiffusion *a sinistra e* Midjourney *a destra (elaborazione degli autori).*

Fig. 11. Immagini generate con StableDiffusion *a sinistra e* Midjourney *a destra attraverso il prompt: home interior view, bedroom, modern architecture, minimalist style furniture, mirror with reflections (elaborazione degli autori).*

Fig. 12. Immagini generate con StableDiffusion *a sinistra e* Midjourney *a destra attraverso il prompt: home interior view, modern architecture, minimalist style furniture, dramatic lighting and shadows (elaborazione degli autori).*

L'analisi di elementi riflessi negli specchi ha presentato l'incoerenza di alcune soluzioni, sottolineando l'inconsapevolezza spaziale dell'AI. In particolare, il riflesso dello specchio manca di alcuni elementi presenti nella scena (come la coperta stesa sul letto in fig. 11 a sinistra) oppure mostra tali elementi in posizione incoerente (sempre la coperta in fig. 11 a destra, che nella scena è appoggiata al bracciolo del divano mentre nel riflesso è stesa dall'altro lato).
Lo studio delle ombre è un'ulteriore conferma del fatto che la costruzione delle immagini generate dall'AI non abbia consapevolezza dello spazio tridimensionale che rappresenta. Molto spesso i raggi di luce che penetrano dalle finestre producono delle ombre che non sono coerenti con gli infissi (fig. 12).

Conclusioni

L'analisi del funzionamento delle AI *text-to-image*, insieme alla ricognizione degli studi sul tema e alle sperimentazioni condotte, permette di tracciare delle prime riflessioni sul loro possibile ruolo per la rappresentazione del progetto di architettura.
Le sperimentazioni mettono in risalto sia potenzialità che punti di debolezza. Tra le prime possono certamente essere annoverate la grande velocità di generazione che permette, in poche decine di secondi, di disporre di immagini dall'elevata qualità visuale, la flessibilità delle tecniche grafiche (ri)prodotte e la coerenza con i *prompt* proposti. Queste potenzialità rendono possibili dei rapidi salti avanti-indietro nel processo progettuale, dalla fase preliminare ideativa a quella di visualizzazione avanzata dell'idea. Tra le seconde, oltre ai già analizzati limiti in termini di correttezza della rappresentazione, bisogna annoverare le criticità segnalate da più parti riguardo alla legittimità in termini di diritti d'autore dei metodi adottati per creare i database per l'allenamento [5] e la presenza di potenziali *bias* culturali indotti nelle AI. A tal proposito, è sufficiente notare come le piazze proposte in fig. 8 al centro rispecchino chiaramente modelli nord americani, lontani dalla concezione europea di spazio pubblico. Inoltre, si evidenzia come questo tipo di AI non siano attualmente idonee per contribuire alla realizzazione di elaborati tecnici di progetto.
Esiste una sostanziale differenza concettuale tra l'intelligenza artificiale e quella umana, che consegue dal processo di allenamento e generazione delle AI. Queste ultime sono intelligenze di tipo interpolativo, cioè efficientissime nell'interpolare valori esistenti nell'ambito del database di allenamento e generare un valore non presente ma che non gli sarà mai del tutto estraneo. Non sono cioè in grado di estrapolare nuovi valori, non solo non presenti ma del tutto alieni al database. Questa forma di intelligenza è, invece, tipicamente umana [Del Campo 2022a]. La potenzialità della co-creazione uomo-AI in fase ideativa sembra quindi risiedere proprio nella collaborazione tra due tipi diversi di intelligenze, in cui quella interpolativa, avviata e guidata dagli input umani, propone immagini «familiari ma strane» [Del Campo 2022b, p. 28] dalle quali l'intelligenza umana può cogliere suggestioni da sviluppare in idee innovative. Questa ipotesi rinnova la problematica riguardante l'autorialità che già la progettazione/rappresentazione algoritmica ha aperto, portando all'idea che un'autorialità condivisa da più agenti (umani o artificiali) sia connaturata al progredire della rivoluzione digitale in architettura. In questo contesto, l'autorialità umana va comunque intesa come "primaria" poiché ha il ruolo di creare le regole generali, gli "oggetili" deleuziani, dalle quali le autorialità artificiali 'secondarie' deriveranno le singole forme, gli oggetti [Carpo 2011, p. 40, 123-128]. La transizione dalla figura di architetto come progettista di singole forme a quella di progettista di regole generali è già in corso e ha portato all'ampliamento dello spettro dei linguaggi adottati. A partire dalla seconda svolta digitale, gli architetti hanno imparato a comporre *script* e algoritmi, affiancandoli alla rappresentazione grafica, e ora, con l'avvento delle AI basate su *prompt*, sono chiamati a integrare il linguaggio naturale in forma scritta tra i loro metodi di progettazione.
Quest'ultima sfida posta dalla rivoluzione digitale deve far riflettere sui linguaggi di rappresentazione in senso più ampio e sul loro insegnamento, uno dei possibili campi di ricerca interdisciplinare del presente e del prossimo futuro del disegno d'architettura.

Crediti

Gli autori hanno condiviso tutte le fasi della ricerca in modo equo. Ai fini della stesura dell'articolo: M.F.M. ha scritto *Introduzione*, *Avanti e indietro nello spazio latente*, *Potenzialità* e *Conclusioni*; S.M. ha scritto *Passato e presente*, *Le principali piattaforme*, *Studi relativi* e *Limiti*.

Note

[1] Dato relativo a luglio 2023, fonte <https://discord.com/servers> (consultato il 24 luglio 2023).

[2] <https://rom1504.github.io/clip-retrieval/?back=https%3A%2F%2Fknn.laion.ai&index=laion5B-H-14&useMclip=false> (consultato il 24 luglio 2023).

[3] La sperimentazione è stata condotta su due piattaforme di AI: *Midjourney* e *StableDiffusion* associato a un ulteriore modello di addestramento chiamato *RealisticVision* v. 4.

[4] Nel *prompt* è stata anche inserita una specifica relativa alla presenza di un pavimento composto da piastrelle quadrate in modo da permettere la successiva analisi prospettica.

[5] <https://www.egair.eu/> (consultato il 24 luglio 2023).

Autori

Matteo Flavio Mancini, Dipartimento di Architettura, Università degli Studi Roma Tre, matteoflavio.mancini@uniroma3.it
Sofia Menconero, Dipartimento di Storia, Disegno e Restauro dell'Architettura, Sapienza Università di Roma, sofia.menconero@uniroma1.it

Riferimenti bibliografici

Carpo, M. (2011). *The alphabet and the algorithm*. Cambridge - London: The MIT Press.

Carpo, M. (ed.). (2013). *The digital turn in architecture 1992-2012*. Chichester: John Wiley & Sons.

Carpo, M. (2017). *The second digital turn: design beyond intelligence*. Cambridge - London: The MIT Press.

Colton, S. et al. (2021). Generative Search Engines: Initial Experiments. In A. Gómez de Silva Garza et al. (a cura di). *Proceedings of the 12th International Conference on Computational Creativity*, Mexico City, 14-18 settembre 2021, pp. 237-246. Mexico City: ACC.

Crowson, K. et al. (2022). *VQGAN-CLIP*: Open Domain Image Generation and Editing with Natural Language Guidance. In *arXiv*. <https://arxiv.org/abs/2204.08583> (consultato il 18 luglio 2023).

Del Campo, M. (2022a). When Robots Dreams. In Conversation with Alexandra Carlson. In *Architectural Design*, n. 03, v. 92, pp. 47-53.

Del Campo, M. (2022b). *Neural Architecture. Design and Artificial Intelligence*. Novato: Oro Editions.

Dhariwal, P., Nichol, A. (2021). Diffusion Models beat GANs on Image Synthesis. In M. Ranzato et al. (eds). *Advances in Neural Information Processing Systems*, v. 34 , pp. 1-15. Cambridge: MIT Press.

Goodfellow, I. et al. (2014). Generative Adversarial Nets. In Z. Ghahramani et al. (eds). *Advances in Neural Information Processing Systems*, v. 29 , pp. 1-9. Cambridge: MIT Press.

Hegazy, M., Saleh, A.M. (2023). Evolution of AI role in architectural design: between parametric exploration and machine hallucination. In *MSA Engineering Journal*, v. 2, n. 2, pp. 262-288. <www.doi.org/10.21608/MSAENG.2023.291873> (consultato il 18 luglio 2023).

Hong, W. et al. (2022). CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. In *arXiv*. <https://arxiv.org/abs/2205.15868> (consultato il 18 luglio 2023).

Jaruga-Rozdolska, A. (2022). Artificial intelligence as part of future practices in the architect's work: MidJourney generative tool as part of a process of creating an architectural form. In *Architectus*, v. 3, n. 71, pp. 95-104.

Jun, H., Nichol, A. (2023). *Shape-E*: Generating Conditional 3D Implicit Functions. In *arXiv*. <https://arxiv.org/abs/2305.02463> (consultato il 18 luglio 2023).

Nichol. A. et al. (2022). *Point-E*: A System for Generating 3D Point Clouds from Complex Prompts. In *arXiv*. <https://arxiv.org/abs/2212.08751> (consultato il 18 luglio 2023).

Paananen, V. *et al.* (2023). Using Text-to-Image Generation for Architectural Design Ideation. In *arXiv*. <https://arxiv.org/abs/2304.10182> (consultato il 18 luglio 2023).

Ploennings, J., Berger, M. (2022). AI Art in Architecture. In *arXiv*. <https://arxiv.org/abs/2212.09399> (consultato il 18 luglio 2023).

Prix, W. et al. (2022). The Legacy Sketch Machine. From Artificial to Architectural Intelligence. In *AD, Machine Hallucinations: Architecture and Artificial Intelligence*, v. 92, n. 3, pp. 14-21.

Ramesh, A. et al. (5 gennaio 2021). DALL-E: Creating images from text. <https://openai.com/research/dall-e> (consultato il 18 luglio 2023).

Radford, A. et al. (2021). Learning Transferable Visual Models from Natural Language Supervision. In M. Meila, T. Zhang (a cura di). *Proceedings of the 38th International Conference on Machine Learning*. Virtuale, 18-24 luglio, v. 139, pp. 8748-8763. Maastricht: ML Research Press.

Reed, S. et al. (2016). Generative Adversarial Text to Image Synthesis. In M. F. Balcan, K. O. Weinberger (a cura di). *Proceedings of the 33rd International Conference on Machine Learning*, v. 48, pp. 1060-1069. Maastricht: ML Research Press.

Rombach, R. et al. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 giugno, pp. 10674-10685. New York: IEEE.

Singer, U. et al. (2022). *Make-A-Video*: Text-to-Video Generation without Text-Video data. In *arXiv*. <https://arxiv.org/abs/2209.14792> (consultato il 18 luglio 2023).

Tong, H. et al. (2023), An attempt to integrate AI-based techniques into first year design representation course. In K. Vaes, J. Verlinden (a cura di). *Connectivity and Creativity in times of Conflicts. Cumulus Conference Proceedings*. Anversa, 12-15 aprile, pp. 1-5. Anversa: University of Antwerp.

Tsigkari, M. et al. (29 marzo 2021). Towards Artificial Intelligence in Architecture: How machine learning can change the way we approach design. In *Plus Journal*, <https://www.fosterandpartners.com/insights/plus-journal/towards-artificial-intelligence-in-architecture-how-machine-learning-can-change-the-way-we-approach-design> (consultato il 18 luglio 2023).

Wallish, S. (2022). GAN Hadid. In S. Carta (a cura di). *Machine Learning and the City: Applications in Architecture and Urban Design*, pp. 477-481. Hoboken-Chichester: John Wiley & Sons.

Yildirim, E. (2022). Text-to-image generation A.I. in architecture. In H. Hale Kozlu (a cura di). *Art and Architecture: Theory, Practice and Experience*, pp. 97-119. Lyon: Livre de Lyon.

Zhang, L., Agrawala, M. (2023). Adding Conditional Control to Text-to-Image Diffusion Models. <https://arxiv.org/abs/2302.05543> (consultato il 18 luglio 2023).