èkphrasis

46° Convegno Internazionale
dei Docenti delle Discipline della Rappresentazione
Congresso della Unione Italiana per il Disegno

Spazio latente della rappresentazione e rappresentazione dello spazio nell'epoca dell'*èkphrasis* artificiale

Fabrizio Gay
Irene Cazzaro

Abstract

Il contributo indaga il concetto di Spazio Latente (LS) come struttura semantica e matematica alla base delle tecnologie di intelligenza artificiale generativa, in particolare nei *Large Multimodal Models* (LMM). In una prospettiva genealogica, il LS emerge non solo come astrazione computazionale, ma come forma di rappresentazione della conoscenza, capace di tradurre dati tra diverse sostanze espressive (testo, immagine, audio, video, 3D). La tesi sostenuta è che il Disegno, e più in generale la rappresentazione dello spazio, siano oggi forme di traduzione intersemiotica artificiale, analoghe ai processi di *èkphrasis* operati dagli LMM. L'articolazione del LS nei modelli generativi – *autoencoder*, VAE, GAN, modelli di diffusione, *Transformer* – mostra come le immagini artificiali non siano semplici artefatti visivi, ma proiezioni geometriche e semantiche di dati in spazi matematici multidimensionali. La ricerca mostra inoltre come la rappresentazione latente dello spazio consenta nuove modalità di trasformazione e sintesi visiva: dalla segmentazione e reinterpretazione di opere pittoriche, alla loro ricostruzione in modelli 3D e immersivi, utilizzando tecnologie come *Neural Radiance Fields* (NeRF), *Gaussian splatting* e algoritmi di stima della profondità. In questo scenario, il concetto tradizionale di immagine si espande fino a includere strutture latenti di significato, ridefinendo il ruolo del disegno e della rappresentazione nello studio della visione umana e artificiale.

Parole chiave
Spazio Latente, disegno artificiale, *èkphrasis* artificiale, traduzione intersemiotica, modelli generativi.

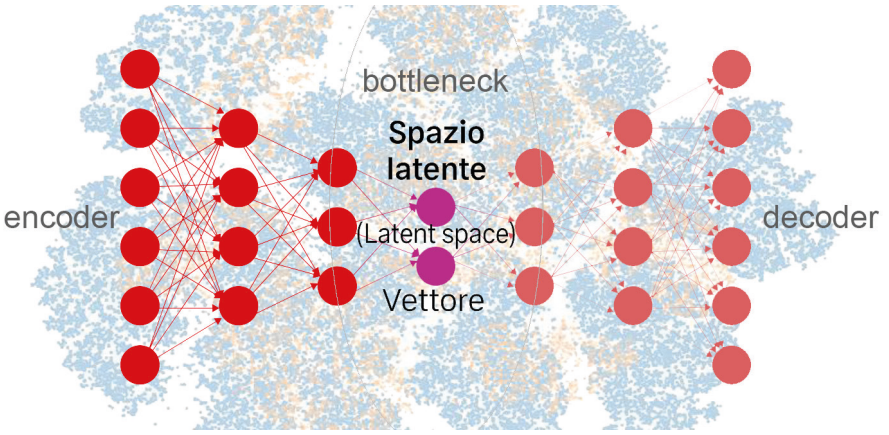


Diagramma sintetico dello spazio latente, che illustra il processo di compressione e ricostruzione: dal tensore ad alta dimensionalità al vettore latente e ritorno, secondo lo schema *input* > spazio latente > *output*.

doi.org/10.3280/oa-1430-c952.

Le concrete virtualità dello Spazio Latente

Col travolgente sviluppo dei *Large Multimodal Models* (LMM) e delle connesse applicazioni AI *generative Text-to-Image-to-Text*, *Text-to-Video-to-Text* (V2T), *Audio-to-Text-to-Audio*, *Image-to-Audio-to-image*, *Multimodal-to-Text*, *Text-to-3D*, *Image-to-3D*, *Video-to-3D*, ecc., la parola *image* (immagine) non denota solo le *picture*s (artefatti visuali tangibili), si riferisce anche agli invisibili pattern di dati tradotti e rielaborati nella comunicazione tra macchine intelligenti: invisibili entità digitali – *raster*, vettoriali, semantiche… – [Paglen 2016] trasdotte nelle più diverse sostanze espressive. Cos'è un'immagine visiva artificiale e che rapporto genetico ha con altre sostanze espressive (linguistiche e non linguistiche) dipende dal tipo di modello analitico o generativo che la tratta – VLMs, VAE, GANs, *Diffusion*, *Transformers*, … – e, conseguentemente, dipende dalla forma del cosiddetto 'Spazio Latente' (LS) del modello [Zylinska 2024].
Nello sviluppo delle reti 'neurali' la nozione di Spazio Latente (LS) è divenuta centrale a partire dai sistemi *autoencoder* e generativi basati sul *deep learning*, dove lo LS è sia il risultato dell'addestramento del modello (attraverso una rete neurale di *encoding*), sia l'ambiente in cui avviene la generazione di nuovi dati attraverso la rete di *decoding*. Lo LS è la 'mappa' del *dataset* di addestramento – di dimensioni (*feature*s) abnormi – in una struttura a dimensione molto inferiore che conserva solo l'essenziale dell'informazione acquisita per l'operabilità del modello.
Questo luogo d'informazione 'latente' è uno spazio immensamente più compatto dove si semplifica e chiarifica la comprensione dei dati, facilitando l'estrazione di pattern e la generazione di nuove istanze coerenti con il dataset originale. Lo LS è dunque lo spazio di tutte le possibili variazioni, interpolazioni e combinazioni di dati che il modello può effettivamente realizzare in *output*. Esso contiene tutte le virtualità che un modello generativo può potenzialmente realizzare.
Mentre nella linguistica e nella semiotica umana la 'competenza virtuale' immanente a ogni produzione e ricezione di fatti espressivi è una pura istanza della teoria – espressa nei concetti di '*Langue*' di Ferdinand de Saussure e di 'Enciclopedia' di Umberto Eco – nei sistemi di interpretazione artificiale basati sul *Deep Learning*, questa 'competenza' è concretamente implementata nello specifico LS e formalizzata nella sua specifica geometria: ex. quella di spazio vettoriale, o di distribuzione di probabilità, o di sequenza di trasformazioni…
Ogni modello ha una propria forma di LS – ex: quelle di *feature space, generative Latent Space, scene representation space, spectrogram Latent Space*, ecc. – corrispondente alla sua funzione e alle diverse architetture dell'interpretazione artificiale che opera (fig. 1).
Per esempio, la generazione di descrizioni testuali di date immagini visive, come l'operazione inversa, avviene grazie alla capacità dello LS di rappresentare semanticamente entità di contenuto verbale e visivo, ovvero di usare il verbale come metalinguaggio del visuale, e l'inverso. Le varie forme di *èkphrasis* artificiale operate dagli MML non sono che casi specifici di una forma molto più generale di 'traduzione intersemiotica artificiale' iscritta (implementata) nella geometria dello LS.
Anche l'odierna teoria e tecnica del Disegno – dal rilievo alle prefigurazioni progettuali – non sono che casi particolari di questa generalizzata 'traduzione intersemiotica artificiale'. Oltre ai modi attuali del disegno, anche la storia della rappresentazione visuale, in quanto studio delle genealogie della visione umana naturale e artificiale, dovrebbe indagare gli sviluppi della nozione di LS, che – seppur coniata recentemente – ha una genealogia molto antica, profonda e riconducibile a due tradizioni principali:
1. la genealogia della rappresentazione geometrica, che attraversa la cartografia, la matematica, la fisica e la statistica, riflettendo l'idea di mappare strutture complesse proiettandole in spazi di dimensione diversa per renderle più comprensibili;
2. la genealogia della modellazione geometrica della semantica, in cui grandezze semantiche vengono tradotte in entità di spazi (vettoriali, topologici, …). Qui, distanze, direzioni, cluster, traiettorie e regioni delineano non solo relazioni spaziali, ma anche relazioni di tipo logico o fisico (ad esempio relazioni causali) fra oggetti o concetti; rappresentano relazioni concettuali e similarità semantiche fornendo una struttura matematica alla rappresentazione del significato.

Dimensioni di rappresentazione

La genealogia del concetto di Spazio Latente (LS) nella storia della rappresentazione è legata all'evoluzione dei concetti di 'mappa' e 'dimensione'. Da nozione intuitiva legata alle tre dimensioni dell'estensione fenomenica, l'idea di 'dimensione' assume una concezione astratta per denotare solo il numero di parametri indipendenti (coordinate o gradi di libertà) di un modello, cioè, numero di categorie 'intensionali' di uno spazio artificiale di rappresentazione.

Modello	Natura dello spazio latente	Geometria dello LS	Rappresentazione Continua o Discreta	Modalità dominante
Autoencoder (AE)	**Deterministico** Ogni punto nello spazio latente corrisponde a una rappresentazione univoca dei dati	**Non strutturato** Casuale	**Continua** Punti vicini nello LS corrispondono a rappresentazioni simili	Visuale
VAE	**Probabilistico** I punti nello spazio latente non sono deterministici, ma vengono campionati da una distribuzione probabilistica	**Strutturato** Distribuzione probabilistica gaussiana	**Continua**	Visuale
GAN	**Semi-deterministico** Vettore di rumore casuale, campionato da una distribuzione uniforme o gaussiana	**Non strutturato** Tensori di pixel o altri tipi di dati numerici trattati come un vettore di rumore casuale in trasformazione	**Continua** Ogni vettore di rumore corrisponde a un punto nello LS, che viene trasformato in un output specifico dal generatore. L'input e l'output sono trattati come vettori continui o tensori continui	Visuale
Diffusion Models	**Rumore Gaussiano** Modellato come processo stocastico	**Altamente strutturato** Transizioni progressive continue tra stati	**Continua** Trasformazione di dati nella loro forma continua	Visuale
Transformers NLP (GPT, BERT)	**Embedding di patch/token** Spazio vettoriale	**Strutturato non lineare** Per catturare relazioni complesse tra i dati, come quelle tra testo e immagini	**Discreta** Insieme discreto di punti o categorie	Verbale
CLIP	**Embedding multi-modale** Spazio vettoriale	**Strutturato**	**Discreta** + Continua	Verbale + Visuale intersemiotica
NeRF	**Funzione continua** e **deterministica** di un oggetto o di una scena	**Regolare** Campo di radianza 5D	**Continua**	Visuale

Possibilità di controllo da testo e su attributi

– +

Fig. 1. Tabella di confronto tra i principali modelli di Spazio Latente, evidenziando le differenze nella natura dello spazio, geometria, rappresentazione (continua o discreta) e applicazioni dominanti, come visuale o verbale, grado di parametrizzazione categoriale.

Tale evoluzione diventa irrefutabile alla fine dell'Ottocento, specialmente con la geometria descrittiva generalizzata da Giuseppe Veronese, che estende la geometria tradizionale agli spazi di dimensione arbitraria attraverso un parallelo approccio algebrico. In questo contesto, la 'immersione di Veronese' è un'operazione opposta alla costruzione di un moderno LS: si tratta di una mappa polinomiale che rappresenta uno spazio proiettivo in una dimensione superiore, consentendo di studiare le proprietà geometriche delle sue strutture: le varietà algebriche.
All'opposto, l'analisi delle componenti principali (PCA) introdotta da Karl Pearson (1901) – usata anche in psicometria per l'analisi dei 'fattori latenti' – è un metodo per ridurre la dimensionalità di uno spazio dato, proiettando i dati iperdimensionali in uno spazio a dimensione inferiore, ma che preserva la massima varianza statistica tra i dati originari. Questo approccio è sviluppato ulteriormente da Harold Hotelling (1933), introducendo la fattorizzazione delle matrici: una tecnica che permette di rappresentare i dati in modo ancora più compatto e statisticamente efficiente, come negli odierni LS [Beniger, Robyn 1978; Pasquinelli 2023].
Col sorgere della Teoria dell'Informazione di Claude Shannon (1948), con l'introduzione del concetto di 'entropia' – come misura dell'incertezza di una sorgente di dati – e di 'capacità del canale' – come misura di quanta informazione può essere trasmessa senza perdita – si generalizza l'idea che l'informazione può essere codificata in modo più efficiente, con una trasformazione che separa gli aspetti essenziali dai dettagli ridondanti o irrilevanti. Tale rappresentazione compatta riguarda le fondamenta statistiche della teoria dell'informazione, nel cui ambito, negli anni '60, si formalizzano i Modelli di Markov Nascosti (HMM): modelli statistici capaci di rappresentare un sistema con stati nascosti (non osservabili) e con osservazioni visibili allo scopo di estrarre sequenze di dati in cui le dinamiche sottostanti non sono direttamente osservabili, cioè, inferendo informazioni latenti dalle osservazioni visibili.
La nozione matematica di spazio latente (LS) diventa esplicita con l'evoluzione dei modelli a reti neurali, ovvero con l'affermazione della teoria connessionista in psicologia cognitiva che culmina negli anni '80. È in questa fase che, passando dalla misurazione e compressione dell'informazione alla modellazione di stati nascosti (gli HMM), si giunge alle prime formulazioni esplicite delle odierne tecniche di apprendimento automatico, di rappresentazione compatta (*autoencoder*) e generazione di dati strutturati (modelli generativi).
L'idea fondamentale di *autoencoder* è formulata in particolare, tra molti altri, dallo psicologo e informatico Geoffrey Hinton, realizzando una rete neurale profonda *feed-forward* costituita da un *encoder* e un *decoder*, addestrata – con retro-propagazione dell'errore – a ricostruire l'informazione appresa dai dati di *input*.
Se, per esempio, l'*input* è una piccola immagine digitale di 256×256 pixel con 3 canali di colore (RGB) della *Pala di San Zaccaria* di Giovanni Bellini (figg. 2, 3), la dimensione del relativo tensore sarà 196.608. Gli strati della rete neurale dell'*encoder* analizzano l'immagine di *input* attraverso algoritmi di descrizione di 'basso livello' – cioè, secondo categorie plastiche: topologiche, cromatiche, eidetiche, testurali ecc. – e di 'alto livello', cioè secondo categorie iconiche: raffigurazioni di oggetti e di ambienti, di allegorie o di temi iconografici. Se il numero di queste caratteristiche è, per esempio, 128, l'*encoder* traduce l'immagine in un *array* unidimensionale (un vettore) di 128 numeri ordinati, che esprime il valore codificato di queste caratteristiche. L'*encoder* mappa così ogni input in un singolo punto (vettore) dello LS. Questo significa che per ogni *input*, è unico il vettore di caratteristiche latenti che lo rappresenta in forma compatta: in questo caso è 1536 volte più compatto.
Il *decoder* fa il percorso inverso, cerca di ricostruire il dato originale a partire dai punti (vettori) dello LS, nel modo più fedele possibile. Cioè, trasforma il vettore in LS nuovamente in un'immagine (tensore 256×256×3) che dovrebbe essere il più simile possibile all'immagine originale.
Nell'*autoencoder* di Hinton l'input veniva semplicemente compresso e ricostruito: elaborato come vettore continuo. La relazione tra ciascun *input*, lo LS e l'*output* è deterministica, ma

non necessariamente biunivoca a causa della compressione e della perdita di informazioni in questo processo di astrazione e concretizzazione successiva. L'obiettivo pratico di questo processo di 'derivazione' e 'integrazione' successiva era che l'*output* reintegrato fosse il più simile all'immagine originale nei termini posti dalle 128 caratteristiche essenziali nello LS. In questo modo il modello era già in grado di svolgere compiti di *denoising* e completamento integrativo di immagini o di testi consentendo le prime operazioni d'interpolazione semantica.

Dal continuo deterministico al discreto probabilistico

Negli anni '80, parallelamente agli *autoencoder* deterministici, si aprirono gli studi applicativi sulle 'reti bayesiane' e si svilupparono i Modelli di Markov Nascosti gerarchici con l'obiettivo di modellare relazioni probabilistiche tra variabili.
Le reti bayesiane o 'reti di credenza' – applicate alla diagnosi medica, al design o al processo giudiziario – sono 'grafi diretti aciclici' (DAG) i cui nodi rappresentano variabili e i cui archi rappresentano dipendenze in termini di probabilità condizionale esplicita (teorema di Bayes) tra le variabili casuali.
Invece le reti di credenza (*Belief Networks*) e i Modelli di Markov Nascosti gerarchici sviluppati da Michael I. Jordan estendono così il concetto di reti bayesiane introducendo variabili latenti che non sono direttamente osservabili ma influenzano le variabili osservabili. Le variabili latenti permettono di catturare strutture complesse nei dati, migliorando la capacità di modellare distribuzioni di probabilità intricate.
Si passava così dal determinismo dell'*autoencoder* classico – che codifica un input in un singolo punto latente – ai *Variational Autoencoders* (VAE) di Diederik P. Kingma e Max Welling [Kingma, Welling 2013] che mappa l'*input* come una distribuzione probabilistica (un vettore media e uno di varianza) introducendo una forma di discretizzazione probabilistica nello LS, poiché i punti nello LS non sono deterministici, ma vengono campionati da una distribuzione probabilistica.
Gli sviluppi successivi delle tecniche del *deep learning* portano alla possibilità di discretizzare anche in modo non supervisionato dei sempre più vasti *corpora* di *training data* provenienti da diverse sostanze espressive, specie di immagini 2D etichettate verbalmente, come *LabelMe*, *PASCAL VOC*, *COCO*, *ImageNet-2* (1,56 milioni di immagini etichettate in 30.000 categorie tratte da *WordNet*), *Open Images*, *JOGAN*, *Places365* (365 milioni di immagini di luoghi)… o di immagini 3D etichettate come *ARKitScenes*, *Structured3D*, *SceneFun3D*, *Habitat Synthetic Scene Dataset* (HSSD), *ImageNet3D*, *ShapeNet*…
Con i *training data* intermodali si apriva l'epoca dell'èkphrasis artificiale, quella della generazione artificiale di immagini e testi tramite altre immagini e/o *prompt* verbali. La generazione di immagini condizionata da immagini vede dal 2014 [Goodfellow *et al.* 2014] l'introduzione delle *Generative Adversarial Networks* (GAN) composte da due reti neurali in competizione:
1. un Generatore che prende in *input* un vettore di rumore casuale e lo mappa a un'immagine coerente;
2. un Discriminatore che confronta l'immagine generata con l'*input* ricevuto dal *dataset* di *training* aprendo un ciclo di *feedback* tra le due reti che porta il generatore a produrre immagini (o altre forme di dati sintetici) sempre più coerenti coi prototipi forniti.

Le limitazioni in termini di stabilità del *training* e qualità delle immagini generate dai GAN hanno portato al successo il ben diverso paradigma dei modelli di diffusione, introdotti nel 2015 [Ho, Ajay, Abbeel 2020], che utilizza un processo iterativo di aggiunta e rimozione di rumore per generare immagini di alta qualità, diventando il nuovo standard per la generazione di immagini.
Infine, i modelli *Transformer* introdotti nel 2017 da Google Brain [Vaswani *et al.* 2017] offrono migliori possibilità di controllo intersemiotico, inaugurando la genealogia dei *Large Multimodal Models* (LMM) come *CLIP* (OpenAI 2021), *DALL-E* (OpenAI 2021),

Fig. 2. Variazioni sul tema della *Pala di San Zaccaria* di Giovanni Bellini declinata artificialmente in altre culture visuali tramite modelli generativi *text-to-image* e *image-to-image*.

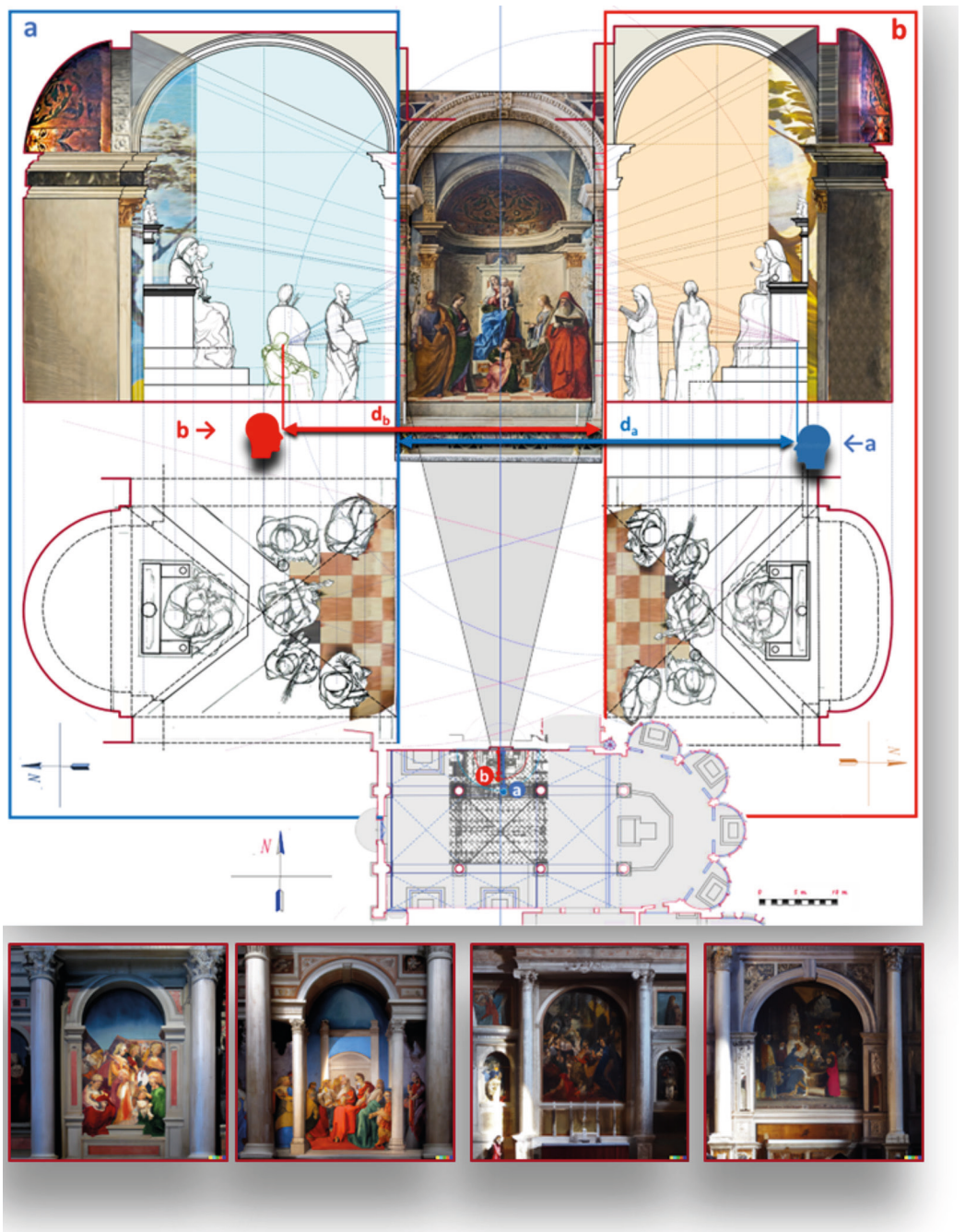


Fig. 3. Restituzione prospettica della *Pala di San Zaccaria* di Giovanni Bellini e sue variazioni contro-fattuali in modalità *blending* con esempi di pittura veneta del Cinquecento.

Flamingo (DeepMind 2022) e altri modelli che integrano testo, immagini, audio e video. In questi modelli, i *dataset* di addestramento grezzi sono direttamente discretizzati in *token* testuali o in *patch* (piccole porzioni d'immagine), poi trasformati da una rete neurale in *embeddings* nello Spazio Latente (LS). Lo LS è uno spazio vettoriale discreto e probabilistico, immagine di una semantica distribuzionale, dove la distanza semantica è geometrica, calcolata utilizzando metriche come la distanza coseno o euclidea tra i vettori di *embeddings*.

Conclusioni

La tesi che il Disegno sia una forma di 'traduzione intersemiotica' considera le immagini come se fossero 'lingue' con scritture non alfabetiche. Per esempio si potrebbe sperimenta-

re la trasposizione della spazialità rappresentata pittoricamente in forme espressive di altre culture visive.
Utilizzando un modello multimodale immagine-testo, come *CLIP*, integrato con il *Segment Anything Model* (SAM), si possono segmentare le figurazioni presenti nell'opera pittorica. Queste segmentazioni sono poi traducibili in nuove rappresentazioni visive attraverso diverse modalità sia con procedimenti *image-to-image* e tecniche di *blending*, sia con procedure *text-to-image*.
In una fase successiva le immagini 2D si possono tradurre in modelli 3D o 5D (fig. 4), utilizzando tecnologie avanzate come i *Neural Radiance Fields* (NeRF), i modelli di stima della profondità (MiDaS, DPT) e i modelli di generazione di *mesh* (Pix2Mesh).

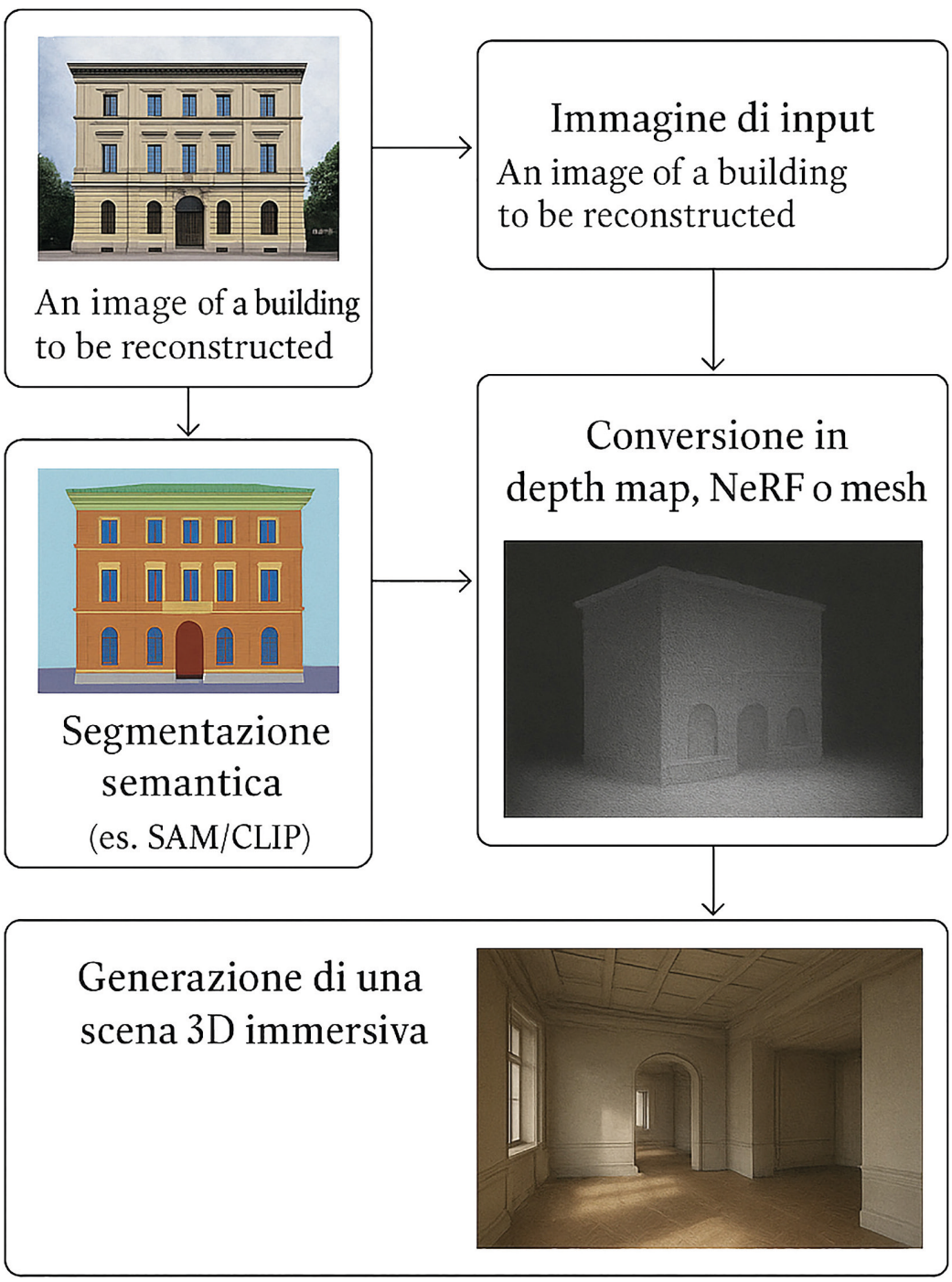


Fig. 4. *Pipeline* di un processo che trasforma una serie di immagini in un ambiente 3D immersivo, integrando algoritmi di intelligenza artificiale. Lo schema multi-livello comprende: immagine di *input* accompagnata da un prompt testuale; segmentazione semantica (es. SAM/CLIP); conversione in *depth map*, NeRF o *mesh*; generazione di una scena 3D immersiva come *output*.

In particolare, l'uso di NeRF [Mildenhall *et al.* 2021] consente di generare campi di radianza neurale, dove ogni punto dello spazio 3D è dotato di parametri aggiuntivi che regolano la visibilità e la trasparenza. Recentemente, l'emergere delle tecniche di *Gaussian splatting* [Kerbl *et al.* 2023] ha offerto un'alternativa efficiente e fotorealistica alla sintesi volumetrica neurale: la scena viene rappresentata tramite insiemi di punti gaussiani 3D dotati di opacità e colore, permettendo *rendering* ad alta qualità in tempo reale. Questo risulta in rappresentazioni spaziali che possono essere visualizzate in modalità immersive, come la realtà aumentata (AR) o la realtà virtuale (VR), offrendo un'esperienza visiva coinvolgente e realistica.
Tuttavia, queste tecniche acquisiscono il loro pieno significato solo all'interno di un progetto di rappresentazione profondamente sensato. Un esempio emblematico di questo concetto è la celebre *Pala di San Zaccaria* di Giovanni Bellini (figg. 2, 3). Questa opera non solo rappresenta uno spazio ideale, ma lo fa riferendosi all'architettura circostante della chiesa realizzata da Mauro Codussi. Gli elementi gotici e rinascimentali, come le arcate a sesto acuto e a tutto sesto, trovano un corrispettivo idealizzato nell'ambiguità prospettica del baldacchino raffigurato nella pala. Un dialogo visivo tra opera pittorica e ambiente reale che produce una coerenza semantica e percettiva per ogni osservatore.

Riferimenti bibliografici

Beniger, J. R., Robyn, D. (1978). Quantitative Graphics in Statistics: A Brief History. In *The American Statistician*, 32 n.1, pp. 1-11.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., e Bengio, Y. (2014). Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K.Q. Weinberger (Eds.). *Advances in Neural Information Processing Systems*. Proceedings of the conference NeurIPS 2014. Montréal, 8-13 December 2014, vol. 27, pp. 1-9. https://dl.acm.org/doi/10.1145/3422622.

Ho, J., Ajay J., Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.). *Advances in Neural Information Processing Systems*. Proceedings of the conference NeurIPS 2020. Online conference, 6-12 December 2020, ,vol. 33, pp. 1-12. https://www.semanticscholar.org/paper/Denoising-Diffusion-Probabilistic-Models-Ho-Jain/5c126ae3421f05768d8edd97ecd44b1364e2c99a.

Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G. (2023). 3D Gaussian Splatting for Real-Time Radiance Field Rendering. In *ACM Transactions on Graphics*, vol. 42, issue 4, n. 139, pp. 1-14.

Kingma, D. P., Welling, M. (2013). Auto-Encoding Variational Bayes. In *CoRR*. https://www.semanticscholar.org/paper/Auto-Encoding-Variational-Bayes-Kingma-Welling/5f5dc5b9a2ba710937e2c413b37b053cd673df02.

Mildenhall, B., Srinivasan, P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R. (2021). NeRF: representing scenes as neural radiance fields for view synthesis. In *Communications of the ACM*, vol. 65, n. 1, pp. 99-106. https://dl.acm.org/doi/10.1145/3503250.

Paglen, T. (8 dicember 2016). Invisible Images (Your Pictures Are Looking at You). In *The New Inquiry*. https://thenewinquiry.com/invisible-images-your-pictures-are-looking-at-you/.

Pasquinelli, M. (2023). *The Eye of the Master: A Social History of Artificial Intelligence*. London: Verso Books.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.). *Advances in Neural Information Processing Systems*. Proceedings of the conference NeurIPS 2017. Long Beach, 4-9 December 2017, pp. 1-15. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Zylinska, J. (2024). Diffused Seeing: The Epistemological Challenge of Generative AI. In *Media Theory*, vol. 8, n. 1, pp. 229-258. https://doi.org/10.70064/mt.v8i1.1075.

Autori
Fabrizio Gay, Università IUAV di Venezia, fabrizio@iuav.it
Irene Cazzaro, Università IUAV di Venezia, icazzaro@iuav.it

Per citare questo capitolo: Fabrizio Gay, Irene Cazzaro (2025). Spazio latente della rappresentazione e rappresentazione dello spazio nell'epoca dell'*èkphrasis* artificiale. In L. Carlevaris *et al.* (a cura di). *èkphrasis. Descrizioni nello spazio della rappresentazione/èkphrasis. Descriptions in the space of representation*. Atti del 46° Convegno Internazionale dei Docenti delle Discipline della Rappresentazione. Milano: FrancoAngeli, pp. 3795-3814. DOI: 10.3280/oa-1430-c952.

 ISBN 9788835182412

ekphrasis

46th International Conference
of Representation Disciplines Teachers
Congress of Unione Italiana per il Disegno

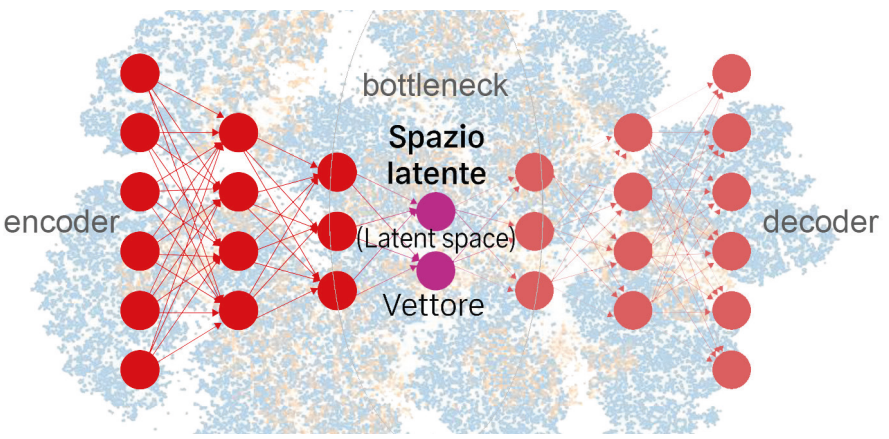
Latent Space of Representation and Representation of Space in the Era of Artificial *Èkphrasis*

Fabrizio Gay
Irene Cazzaro

Abstract

The contribution investigates the concept of Latent Space (LS) as a semantic and mathematical structure underlying current generative artificial intelligence technologies, particularly in Large Multimodal Models (LMMs). Through a genealogical perspective, the work demonstrates that LS is not merely a computational abstraction but a genuine form of knowledge representation capable of translating data across different expressive mediums (text, image, audio, video, 3D). The main thesis posited is that Design, and more broadly, spatial representation today, are specific forms of artificial intersemiotic translation, conceptually analogous to the ekphrastic processes performed by LMMs. The articulation of LS in current generative models –from autoencoder neural networks to probabilistic approaches (VAE, GAN, diffusion models, Transformers)– highlights how artificial images are no longer simple visual artefacts but geometric and semantic projections of datasets into multidimensional mathematical spaces. The research also shows how the latent representation of space enables new modes of visual transformation and synthesis: from the segmentation and reinterpretation of paintings to their reconstruction into 3D and immersive models, using technologies such as Neural Radiance Fields (NeRF), Gaussian splatting and depth estimation algorithms. In this scenario, the traditional concept of the image expands to include latent structures of meaning, redefining the role of drawing and representation in the study of human and artificial vision.

Keywords
Latent Space, artificial drawing, artificial *èkphrasis*, intersemiotic translation, generative models.



A synthetic diagram of the Latent Space, illustrating the process of compression and reconstruction: from the high-dimensional tensor to the latent vector and back, following the input > latent space > output scheme.

doi.org/10.3280/oa-1430-c952.

The concrete virtualities of Latent Space

With the overwhelming development of Large Multimodal Models (LMMs) and related generative AI applications such as Text-to-Image-to-Text, Text-to-Video-to-Text (V2T), Audio-to-Text-to-Audio, Image-to-Audio-to-Image, Multimodal-to-Text, Text-to-3D, Image-to-3D, Video-to-3D, and more, the term 'image' no longer refers solely to 'pictures' (tangible visual artefacts). It also denotes the invisible data patterns translated and reprocessed in the communication between intelligent machines: invisible digital entities –raster, vectorial, semantic, etc.– [Paglen 2016] transduced into various expressive forms. What constitutes an artificial visual image and its genetic relationship with other expressive forms (linguistic and non-linguistic) depends on the type of analytical or generative model handling it –VLMs, VAE, GANs, Diffusion, Transformers, etc.– and, consequently, on the form of the so-called 'Latent Space' (LS) of the model [Zylinska 2024].
In the development of 'neural' networks, the notion of 'Latent Space' (LS) has become central, starting from autoencoder systems and generative models based on deep learning, where the LS is both the result of the model's training (through a neural encoding network) and the environment where new data is generated through the decoding network. The LS is the 'map' of the training dataset –of enormous dimensions (features)– into a much lower-dimensional structure that retains only the essential information acquired for the model's operability.
This 'latent' information space is an immensely more compact space where data understanding is simplified and clarified, facilitating pattern extraction and the generation of new instances consistent with the original dataset. The LS is thus the space of all possible variations, interpolations, and combinations of data that the model can effectively produce as output. It contains all the virtualities that a generative model can potentially realise.
While in human linguistics and semiotics the 'virtual competence' immanent in every production and reception of expressive facts is a pure instance of theory –expressed in Ferdinand de Saussure's concept of '*Langue*' and Umberto Eco's '*Enciclopedya*'– in artificial interpretation systems based on Deep Learning, this 'competence' is concretely implemented in the specific LS and formalised in its specific geometry: for example, that of a vector space, probability distribution, sequence of transformations, etc.
Each model has its own form of LS –for example, feature space, generative Latent Space, scene representation space, spectrogram Latent Space, etc.– corresponding to its function and the different architectures of artificial interpretation it operates (fig. 1).
For example, the generation of textual descriptions of given visual images, as well as the inverse operation, occurs thanks to the LS's ability to semantically represent entities of verbal and visual content, in other words, to use the verbal as a metalanguage of the visual, and vice versa. The various forms of artificial *èkphrasis* operated by MMLs are but specific instances of a much more general form of 'artificial intersemiotic translation' inscribed (implemented) in the geometry of the LS.
Even today's theory and technique of Drawing –from surveying to project prefigurations– are but particular cases of this generalised 'artificial intersemiotic translation'. Beyond current drawing methods, the history of visual representation, as the study of genealogies of natural and artificial human vision, should investigate the developments of the LS notion, which –though recently coined– has a very ancient and deep genealogy, traceable to two main traditions:
1. the genealogy of geometric representation, which traverses cartography, mathematics, physics, and statistics, reflecting the idea of mapping complex structures by projecting them into spaces of different dimensions to make them more comprehensible;
2. the genealogy of geometric modelling of semantics, where semantic magnitudes are translated into entities of (vectorial, topological, etc.) spaces. Here, distances, directions, clusters, trajectories, and regions delineate not only spatial relationships but also logical or physical relationships (for example, causal relationships) between objects or concepts; they represent conceptual relationships and semantic similarities, providing a mathematical structure to the representation of meaning.

Dimensions of Representation

The genealogy of the Latent Space (LS) concept in the history of representation is linked to the evolution of the concepts of 'map' and 'dimension'. From an intuitive notion tied to the three dimensions of phenomenal extension, the idea of 'dimension' assumes an abstract

Model	Nature of latent space	Geometry of LS	Continuous or Discrete Representation	Dominant modality
Autoencoder (AE)	**Deterministic** Each point in the latent space corresponds to an unambiguous representation of the data	**Unstructured** Random	**Continuous** Nearby points in the latent space correspond to similar representations	Visual
VAE	**Probabilistic** Points in the latent space are not deterministic, but sampled from a probabilistic distribution	**Structured** Gaussian probabilistic distribution	**Continuous**	Visual
GAN	**Semi-deterministic** Random noise vector, sampled from a uniform or Gaussian distribution	**Unstructured** Pixel tensors or other types of numerical data treated as a random noise vector in transformation	**Continuous** Each noise vector corresponds to a point in the latent space, which is transformed into a specific generator output. Both input and output are treated as continuous vectors or tensors	Visual
Diffusion Models	**Gaussian Noise** Modelled as a stochastic project	**Highly structured** Progressive continuous transitions between states	**Continuous** Transformation of data into their continuous form	Visual
Transformers NLP (GPT, BERT)	**Embedding of patch/token** Vector space	**Structured, non-linear** To capture complex relationships between data, such as those between text and images	**Discrete** A discrete set of points or categories	Verbal
CLIP	**Multi-modal embedding** Vector space	**Structured**	**Discrete** + Continuous	Verbal + Visual intersemiotic
NeRF	**Continuous and deterministic function** Of an object or a scene	**Regular** 5D radiance field	**Continuous**	Visual

Possibility of control from text and on attributes

– +

Fig. 1. Comparative table of the main latent space models, highlighting differences in the nature of the space, geometry, representation (continuous or discrete), and dominant applications, such as visual or verbal, and the degree of categorical parameterisation.

conception to denote only the number of independent parameters (coordinates or degrees of freedom) of a model, in other words, the number of 'intensional' categories of an artificial representation space.
This evolution becomes undeniable by the end of the 19th century, especially with the descriptive geometry generalised by Giuseppe Veronese, which extends traditional geometry to spaces of arbitrary dimension through a parallel algebraic approach. In this context, the 'Veronese embedding' is an operation opposite to the construction of a modern LS: it is a polynomial map that represents a projective space in a higher dimension, allowing the study of the geometric properties of its structures: algebraic varieties.
Conversely, Principal Component Analysis (PCA) introduced by Karl Pearson (1901) –also used in psychometrics for the analysis of 'latent factors'– is a method for reducing the dimensionality of a given space, projecting hyperdimensional data into a lower-dimensional space while preserving the maximum statistical variance among the original data. This approach is further developed by Harold Hotelling (1933), introducing matrix factorisation: a technique that allows representing data in an even more compact and statistically efficient manner, as in today's LS [Beniger, Robyn 1978; Pasquinelli 2023].
With the rise of Claude Shannon's Information Theory (1948), with the introduction of the concept of 'entropy' –as a measure of the uncertainty of a data source– and 'channel capacity' –as a measure of how much information can be transmitted without loss– the idea that information can be encoded more efficiently, with a transformation that separates essential aspects from redundant or irrelevant details, becomes generalised. This compact representation concerns the statistical foundations of information theory, within which, in the 1960s, Hidden Markov Models (HMMs) are formalised: statistical models capable of representing a system with hidden (unobservable) states and visible observations, aiming to extract data sequences where the underlying dynamics are not directly observable, in other words, inferring latent information from visible observations.
The mathematical notion of Latent Space (LS) becomes explicit with the evolution of neural network models that is with the affirmation of connectionist theory in cognitive psychology culminating in the 1980s. It is in this phase that, moving from the measurement and compression of information to the modelling of hidden states (HMMs), we arrive at the first explicit formulations of today's techniques of automatic learning, compact representation (autoencoder), and generation of structured data (generative models).
The fundamental idea of 'autoencoder' is formulated, among many others, by the psychologist and computer scientist Geoffrey Hinton, realising a deep feed-forward neural network consisting of an 'encoder' and a 'decoder', trained –with error backpropagation– to reconstruct the information learned from input data.
If, for example, the input is a small digital image of 256×256 pixels with 3 colour channels (RGB) of Giovanni Bellini's *Pala di San Zaccaria* (figs. 2, 3), the dimension of the related tensor will be 196,608. The layers of the neural network encoder analyse the input image through 'low-level' description algorithms –that is according to plastic categories: topological, chromatic, eidetic, textural, etc.– and 'high-level' ones, in other words, according to iconic categories: depictions of objects and environments, allegories, or iconographic themes. If the number of these features is, for example, 128, the encoder translates the image into a one-dimensional array (a vector) of 128 ordered numbers, expressing the encoded value of these features. The encoder thus maps each input to a single point (vector) in the LS. This means that for each input, there is a unique vector of latent features representing it in a compact form: in this case, it is 1,536 times more compact. The decoder does the reverse path, attempting to reconstruct the original data from the points (vectors) in the LS, as faithfully as possible. That is, it transforms the vector in the LS back into an image (256×256×3 tensor) that should be as similar as possible to the original image.
In Hinton's autoencoder, the input was simply compressed and reconstructed: processed as a continuous vector. The relationship between each input, the LS, and the output is deterministic, but not necessarily bijective due to the compression and loss of information in this process of abstraction and subsequent concretisation. The practical

goal of this process of 'derivation' and 'subsequent integration' was that the reintegrated output would be as similar as possible to the original image in terms of the 128 essential features in the LS. In this way, the model was already capable of performing tasks of denoising and integrative completion of images or texts, allowing the first operations of semantic interpolation.

From deterministic continuity to probabilistic discreteness

In the 1980s, parallel to deterministic autoencoders, applied studies on Bayesian networks opened, and hierarchical Hidden Markov Models developed with the aim of modelling probabilistic relationships between variables.
Bayesian networks or 'belief networks' –applied to medical diagnosis, design, or judicial processes– are 'directed acyclic graphs' (DAGs) whose nodes represent variables and whose arcs represent dependencies in terms of explicit conditional probability (Bayes' theorem) between random variables.
Instead, belief networks and hierarchical Hidden Markov Models developed by Michael I. Jordan extend the concept of Bayesian networks by introducing latent variables that are not directly observable but influence observable variables. Latent variables allow capturing complex structures in data, improving the ability to model intricate probability distributions.
Thus, we moved from the determinism of the classical autoencoder –which encodes an input into a single latent point– to the Variational Autoencoders (VAE) of Diederik P. Kingma and Max Welling [Kingma, Welling 2013], which map the input as a probability distribution (a mean vector and a variance vector), introducing a form of probabilistic discretisation in the LS, since the points in the LS are not deterministic but are sampled from a probability distribution.
Subsequent developments in deep learning techniques led to the possibility of discretising, even in an unsupervised manner, increasingly vast corpora of training data from different expressive forms, especially verbally labelled 2D images, such as *LabelMe*, *PASCAL VOC*, *COCO*, *ImageNet-2* (1.56 million images labelled in 30,000 categories drawn from WordNet), *Open Images*, *JOGAN*, *Places365* (365 million images of places), etc., or 3D labelled images such as *ARKitScenes*, *Structured3D*, *SceneFun3D*, *Habitat Synthetic Scene Dataset* (HSSD), *ImageNet3D*, *ShapeNet*, etc.
With intermodal training data, the era of artificial *èkphrasis* emerged –the one of the artificial generation of images and texts through other images and/or verbal prompts. Image generation conditioned by images saw the introduction of *Generative Adversarial Networks* (GANs) in 2014 [Goodfellow *et al.* 2014], composed of two competing neural networks:
1. a Generator that takes a random noise vector as input and maps it to a coherent image;
2. a Discriminator that compares the generated image with the input received from the training dataset, opening a feedback cycle between the two networks that leads the generator to produce images (or other forms of synthetic data) increasingly coherent with the provided prototypes.

The limitations in terms of training stability and quality of images generated by GANs led to the success of the very different paradigm of diffusion models, introduced in 2015 [Ho, Ajay, Abbeel 2020], which uses an iterative process of adding and removing noise to generate high-quality images, becoming the new standard for image generation.
Finally, Transformer models introduced in 2017 by Google Brain [Vaswani *et al.* 2017] offer better possibilities of intersemiotic control, inaugurating the genealogy of Large Multimodal Models (LMMs) such as *CLIP* (OpenAI, 2021), *DALL-E* (OpenAI, 2021), *Flamingo* (DeepMind, 2022), and other models integrating text, images, audio, and video. In these models, raw training datasets are directly discretised into textual 'tokens' or 'patches' (small image portions), then transformed by a neural network into embeddings in the Latent Space (LS).

Fig. 2. Artificial variations on the theme of Giovanni Bellini's *San Zaccaria* Altarpiece, reimagined in other visual cultures through text-to-image and image-to-image generative models.

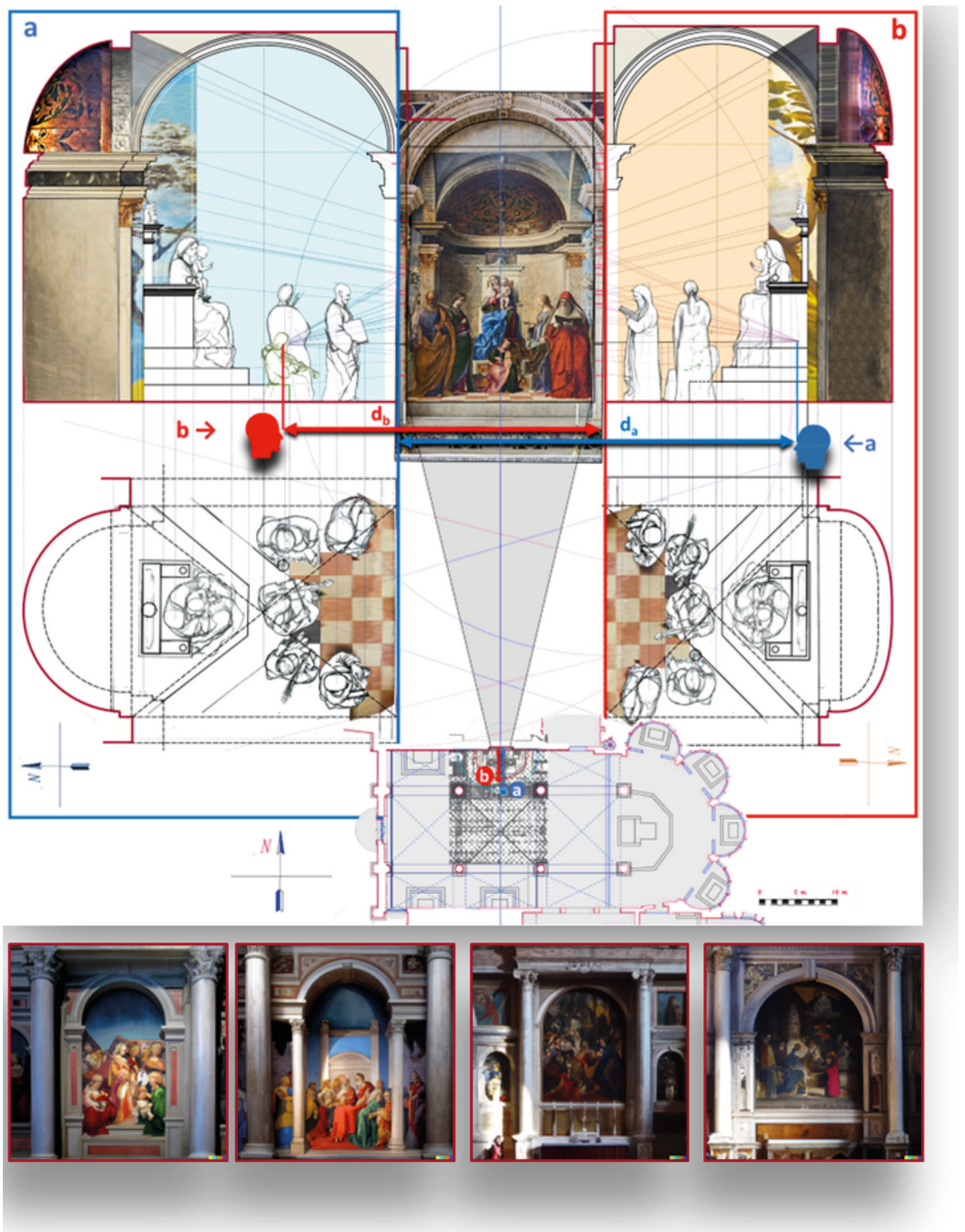


Fig. 3. Elementary photogrammetric perspective rendering of Giovanni Bellini's *San Zaccaria* Altarpiece and its counterfactual variations in blending mode with examples of Venetian painting from the 16th century.

The LS is a discrete and probabilistic vector space, an image of a distributional semantics, where semantic distance is geometric, calculated using metrics such as cosine or Euclidean distance between embedding vectors.

Conclusions

The thesis that Drawing is a form of 'intersemiotic translation' considers images as if they were 'languages' with non-alphabetic writings. For example, one could experiment with the transposition of spatiality represented pictorially into expressive forms of other visual cultures.
Using a multimodal image-text model, such as *CLIP*, integrated with the *Segment Anything Model* (SAM), one can segment the figurations present in a painting. These segmentations

are then translatable into new visual representations through different modalities, both with image-to-image procedures and blending techniques, and with text-to-image procedures.
In a subsequent phase, 2D images can be translated into 3D or 5D models (fig. 4), using advanced technologies such as *Neural Radiance Fields* (NeRF), depth estimation models (MiDaS, DPT), and mesh generation models (Pix2Mesh). In particular, the use of NeRF [Mildenhall *et al.* 2021] allows generating neural radiance fields, where each point in 3D space is endowed with additional parameters regulating visibility and transparency.
Recently, the emergence of Gaussian splatting techniques [Kerbl *et al.* 2023] has provided an efficient and photo-realistic alternative to neural volumetric synthesis: the scene is represented through sets of 3D Gaussian points with opacity and colour, enabling high-quality

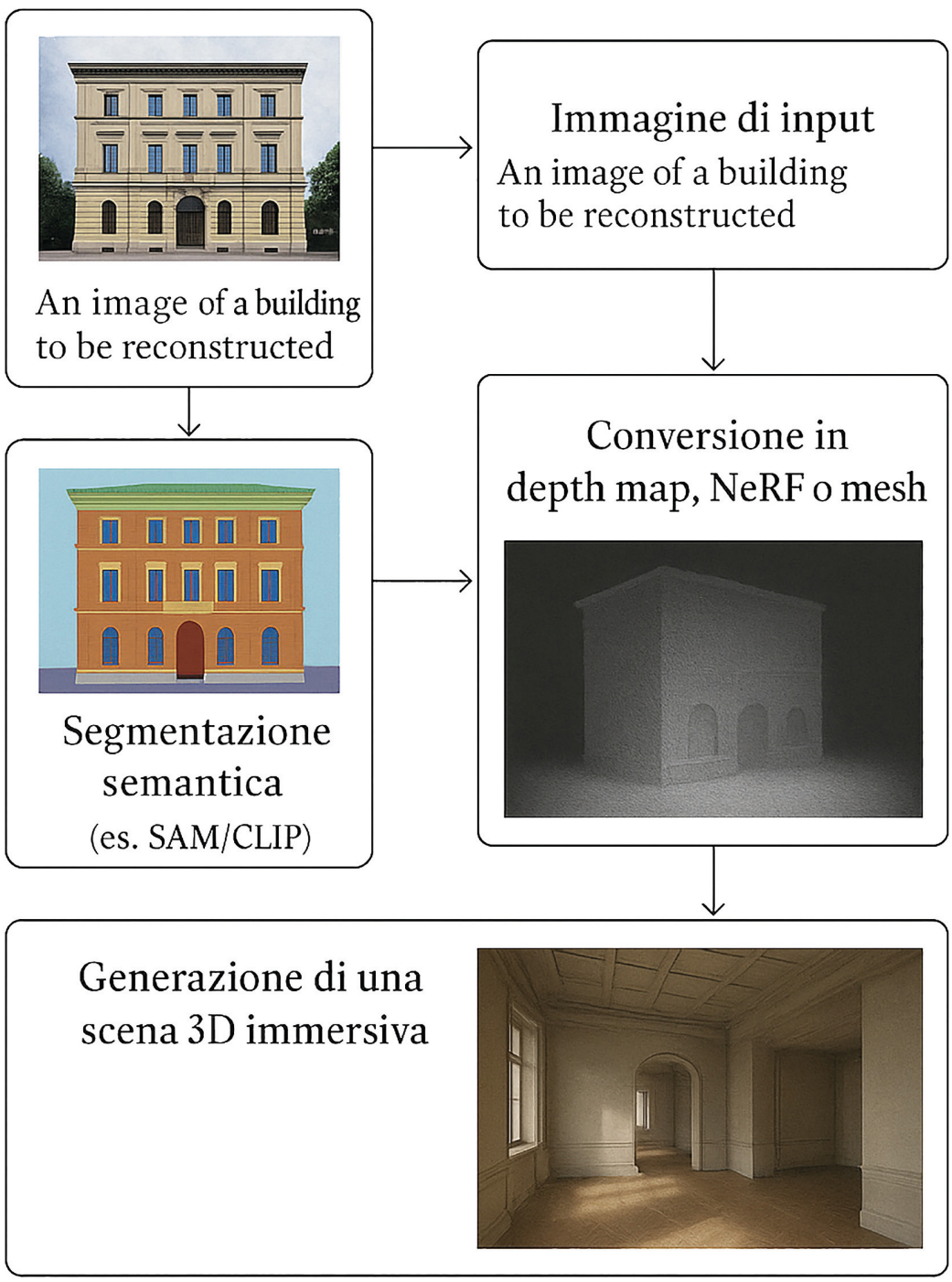


Fig. 4. Pipeline of a process that transforms a series of images into an immersive 3D environment, integrating artificial intelligence algorithms. The multilayered diagram includes: input image accompanied by a textual prompt; semantic segmentation (e.g., SAM/CLIP); conversion into depth map, NeRF, or mesh; generation of an immersive 3D scene as output.

real-time rendering. This results in spatial representations that can be visualised in immersive modes, such as Augmented Reality (AR) or Virtual Reality (VR), offering an engaging and realistic visual experience.
However, these techniques acquire their full meaning only within a deeply meaningful representation project. An emblematic example of this concept is the famous *Pala di San Zaccaria* by Giovanni Bellini (figs. 2, 3). This work not only represents an ideal space but does so by referring to the surrounding architecture of the church realised by Mauro Codussi. The Gothic and Renaissance elements, such as the pointed and round arches, find an idealised counterpart in the perspectival ambiguity of the canopy depicted in the altarpiece. A visual dialogue between the painting and the real environment that produces semantic and perceptual coherence for every observer.

Reference List

Beniger, J. R., Robyn, D. (1978). Quantitative Graphics in Statistics: A Brief History. In *The American Statistician*, 32 n.1, pp. 1-11.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., e Bengio, Y. (2014). Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K.Q. Weinberger (Eds.). *Advances in Neural Information Processing Systems*. Proceedings of the conference NeurIPS 2014. Montréal, 8-13 December 2014, vol. 27, pp. 1-9. https://dl.acm.org/doi/10.1145/3422622.

Ho, J., Ajay J., Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.). *Advances in Neural Information Processing Systems*. Proceedings of the conference NeurIPS 2020. Online conference, 6-12 December 2020, ,vol. 33, pp. 1-12. https://www.semanticscholar.org/paper/Denoising-Diffusion-Probabilistic-Models-Ho-Jain/5c126ae3421f05768d8edd97ecd44b1364e2c99a.

Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G. (2023). 3D Gaussian Splatting for Real-Time Radiance Field Rendering. In *ACM Transactions on Graphics*, vol. 42, issue 4, n. 139, pp. 1-14.

Kingma, D. P., Welling, M. (2013). Auto-Encoding Variational Bayes. In *CoRR*. https://www.semanticscholar.org/paper/Auto-Encoding-Variational-Bayes-Kingma-Welling/5f5dc5b9a2ba710937e2c413b37b053cd673df02.

Mildenhall, B., Srinivasan, P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R. (2021). NeRF: representing scenes as neural radiance fields for view synthesis. In *Communications of the ACM*, vol. 65, n. 1, pp. 99-106. https://dl.acm.org/doi/10.1145/3503250.

Paglen, T. (8 dicember 2016). Invisible Images (Your Pictures Are Looking at You). In *The New Inquiry*. https://thenewinquiry.com/invisible-images-your-pictures-are-looking-at-you/.

Pasquinelli, M. (2023). *The Eye of the Master: A Social History of Artificial Intelligence*. London: Verso Books.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.). *Advances in Neural Information Processing Systems*. Proceedings of the conference NeurIPS 2017. Long Beach, 4-9 December 2017, pp. 1-15. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Zylinska, J. (2024). Diffused Seeing: The Epistemological Challenge of Generative AI. In *Media Theory*, vol. 8, n. 1, pp. 229-258. https://doi.org/10.70064/mt.v8i1.1075.

Authors
Fabrizio Gay, Università IUAV di Venezia, fabrizio@iuav.it
Irene Cazzaro, Università IUAV di Venezia, icazzaro@iuav.it

To cite this chapter: Fabrizio Gay, Irene Cazzaro (2025). Latent Space of Representation and Representation of Space in the Era of Artificial Èkphrasis. In L. Carlevaris *et al.* (Eds.). *èkphrasis. Descrizioni nello spazio della rappresentazione/èkphrasis. Descriptions in the space of representation.* Proceedings of the 46th International Conference of Representation Disciplines Teachers. Milano: FrancoAngeli, pp. 3795-3814. DOI: 10.3280/oa-1430-c952.

ISBN 9788835182412