Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Informatica

Analisi Multimodale dell'Arte Generata da AI: Riconoscimento dello Stile e Valutazione dell'Aderenza Semantica con CLIP

Relatore:
Prof. Andrea Asperti

Presentata da:
Nico Wu

Sessione IV
Anno Accademico 2025

Sommario

L'avvento dei modelli Vision-Language, come CLIP (Contrastive Language-Image Pre-training), ha trasformato la Computer Vision, aprendo nuove frontiere nella generazione e nell'analisi di contenuti visivi. Tuttavia, l'efficacia di tali modelli nel comprendere concetti astratti propri del dominio artistico quali lo stile pittorico e la qualità estetica rimane una questione aperta.

Questa tesi si inserisce in un progetto di ricerca collaborativo volto a valutare le rappresentazioni latenti di CLIP nel dominio dell'arte, analizzando sia opere reali che sintetiche su dataset eterogenei (NGA, WikiArt, AI-ArtBench e AI-Pastiche). Il lavoro complessivo indaga tre dimensioni fondamentali: l'allineamento testo-immagine, il riconoscimento dello stile e la valutazione della qualità generativa.

Nello specifico, il contributo originale di questo elaborato si focalizza sull'analisi delle immagini generate artificialmente. In una prima fase, è stato validato l'allineamento semantico tra prompt e immagini sintetiche. Successivamente, si è esaminata la capacità di CLIP di riconoscere stili artistici complessi, dimostrando che tecniche di supervisione leggera come il *Linear Probing* ottengono prestazioni superiori rispetto ad approcci *Zero-Shot* o *Few-Shot*. Infine, è stato condotto uno studio inedito sull'aderenza estetica, confrontando la similarità di CLIP con il giudizio umano. I risultati evidenziano una discrepanza significativa: il modello tende a ignorare artefatti visivi e difetti strutturali determinanti per la percezione umana. L'integrazione sperimentale di un *defect score* ha permesso di migliorare l'allineamento, suggerendo che le attuali rappresentazioni multimodali necessitino di segnali ausiliari per fungere da giudici di qualità affidabili.

I risultati concludono che, sebbene CLIP eccella nel catturare la semantica denotativa, la comprensione delle sfumature stilistiche e qualitative nell'arte generata richiede ancora interventi mirati e supervisione specifica.

Elenco delle figure

Indice

Capitolo 1

Introduzione

1.1 Contesto e Rilevanza dei Modelli Multimodali

L'avanzamento delle architetture di apprendimento profondo ha consolidato i **modelli multimodali** come pilastri fondamentali nelle moderne applicazioni di intelligenza artificiale. Questi sistemi, capaci di integrare e correlare informazioni provenienti da domini diversi, in particolare testo e immagini, hanno rivoluzionato il campo della *computer vision*. Tra questi, **CLIP** (Contrastive Language–Image Pretraining), sviluppato da OpenAI, si distingue come un'innovazione architetturale significativa [1].

Addestrato attraverso un compito di pre-addestramento contrastivo su una vasta collezione di coppie (*immagine*, *testo*) estratte dal web, CLIP ha dimostrato una notevole capacità di apprendere **rappresentazioni visive trasferibili** attraverso la supervisione del linguaggio naturale. Questa caratteristica lo ha reso uno strumento versatile e ampiamente adottato in molteplici ambiti della *computer vision*, tra cui la classificazione di immagini *zero-shot*, l'identificazione di oggetti (es. *image re-identification*) e l'*image retrieval*. Inoltre, la sua capacità di misurare la coerenza semantica tra un *prompt* testuale e un'immagine è stata cruciale, portando alla sua adozione come **funzione di guida (*guidance loss*)** in modelli generativi all'avanguardia, come *DALL-E* e *Stable Diffusion* [2] (sebbene CLIP funga anche da *prior guidance* in altri contesti).

1.2 Motivazione e Obiettivi della Ricerca

Nonostante l'eccellenza prestazionale di CLIP in compiti di visione generici, permane la necessità di investigare a fondo le sue abilità e i suoi limiti quando applicato a domini visivi complessi e altamente specializzati. Il **contesto artistico**, caratterizzato da astrazione, sottigliezze stilistiche e una ricchezza di dettagli formali, rappresenta un banco di prova ideale per valutare la vera profondità della comprensione visiva del modello.

La presente ricerca si pone l'obiettivo di **indagare criticamente le capacità di CLIP** nell'analisi delle opere d'arte, al fine di determinare la sua affidabilità come meccanismo di guida per la generazione di opere artistiche. Specificamente, intendiamo valutare se l'attuale rappresentazione multimodale appresa da CLIP sia sufficiente per catturare le sfumature che definiscono la qualità e l'identità di un'opera d'arte.

L'ipotesi di partenza, supportata dai risultati preliminari, suggerisce che, sebbene CLIP dimostri una robusta capacità di riconoscere il **contenuto semantico** (i soggetti) delle opere, essa fatica a cogliere **dettagli fini** e **concetti astratti** cruciali, quali la tessitura delle pennellate, l'aderenza stilistica a movimenti specifici e la corretta collocazione temporale (periodo storico) dell'opera. Questa limitazione suggerisce l'esigenza di sviluppare metodologie più espressive o di sfruttare in modo più mirato le *feature* nascoste nello spazio latente codificato da CLIP.

1.3 Struttura Sperimentale

Per ottenere risultati oggettivi e replicabili, i nostri esperimenti sono stati strutturati per dissezionare i vari livelli di comprensione visiva di CLIP nel dominio artistico. A tal fine, è stato mantenuto **CLIP come componente fissa e congelata (*zero-shot evaluation*)**, senza alcuna fase di *fine-tuning*, per isolare e valutare le sue capacità intrinseche apprese dal pre-addestramento originale.

La metodologia di valutazione si articola in tre livelli progressivi di complessità:

1. **Livello I: Content Retrieval.** L'obiettivo primario è verificare l'abilità di CLIP nell'identificazione accurata dei **soggetti principali** all'interno delle opere artistiche, testando la comprensione del contenuto denotativo dell'immagine.

2. **Livello II: Style Recognition.** Dopo aver validato la comprensione del contenuto, si procede a un livello di astrazione superiore, verificando la capacità di CLIP di discriminare e categorizzare opere basandosi su **concetti stilistici astratti** (es. Cubismo, Impressionismo) (per un contesto sulla rappresentazione neurale dello stile, si veda anche [3]).

3. **Livello III: Artifacts Detection and Human Adherence.** Questo esperimento mira a valutare se CLIP possieda la capacità di fungere da **giudice di qualità** autonomo, similmente alla percezione umana. L'analisi si concentra sulla sua abilità di identificare artefatti visivi, difetti o incoerenze formali all'interno delle immagini, un requisito fondamentale per una guida efficace dei modelli generativi. (Si veda anche la ricerca sulle metriche percettive per il rilevamento di *artefatti* [4]).

Per garantire la massima obiettività e la generalizzabilità dei risultati, gli esperimenti sono condotti utilizzando quattro *dataset* distinti, rappresentativi di diverse sfaccettature dell'arte digitale e tradizionale: **NGAD** (National Gallery of Art), **WikiArt**, **AI-pastiche** e **AI-artBench**.

I risultati di questa ricerca mirano a fornire un contributo significativo alla comunità scientifica, offrendo una **valutazione quantitativa e qualitativa** delle prestazioni di CLIP in ambito artistico, informando futuri sviluppi metodologici per una più efficace integrazione dei modelli multimodali nella creazione e nell'analisi dell'arte digitale.

1.4 Datasets Utilizzati

Per la conduzione degli esperimenti diagnostici volti a investigare le *feature* latenti di CLIP, abbiamo deliberatamente incorporato sia *dataset* di opere artistiche reali sia collezioni di opere generate artificialmente. Questo approccio metodologico non solo permette un'analisi più accurata della capacità di codifica di CLIP, ma consente anche di valutare se il modello riesca a percepire e a codificare i difetti visivi (artefatti) intrinseci alla generazione artificiale di opere d'arte.

1.4.1 National Gallery of Art (NGA)

Il *dataset* della **National Gallery of Art** (NGA) è stato costruito utilizzando 2.693 opere disponibili pubblicamente sul sito web della National Gallery of Art di Washington [5]. Ogni opera è corredata da una ricca dotazione di metadati essenziali, tra cui l'OID (*Object Identifier*), il titolo, lo stile, l'autore, una descrizione e un collegamento all'immagine ad alta risoluzione. Il vantaggio distintivo nell'impiego di un *dataset* di arte reale e curata come NGA risiede nella disponibilità di **descrizioni accurate e professionali** fornite da critici e storici dell'arte. Tale livello di dettaglio e autorità critica è fondamentale per valutare la percezione di CLIP anche da una prospettiva estetica e storico-artistica.

Nonostante la ricchezza dei metadati, un limite metodologico è rappresentato dalla sua **non-bilanciatura per classe** (stile o periodo), un fattore che può introdurre *bias* non intenzionali nell'addestramento di modelli di classificazione di stile.

1.4.2 AI-Pastiche

Il *dataset* **AI-Pastiche** è una collezione di 953 opere generate artificialmente, concepita per offrire un ambiente ricco e variegato per la valutazione dei modelli di *computer vision* nel contesto dell'arte sintetica. Le opere spaziano su oltre 19 stili e sono state generate in modo eterogeneo utilizzando 12 differenti modelli generativi, partendo da 73 *prompts* meticolosamente progettati.

Ciascuna opera è annotata con il modello generativo, il *prompt* originale, il soggetto, lo stile, il periodo storico, ed è corredata da punteggi quantitativi che misurano la presenza di **difetti, autenticità e aderenza al tema**. L'inclusione di questi **punteggi sugli artefatti** rappresenta un vantaggio cruciale: si rivelerà infatti essenziale per gli esperimenti sull'aderenza, che mirano a determinare se le rappresentazioni latenti di CLIP siano in grado di codificare e catturare i difetti visivi intrinseci ai processi generativi.

A causa del vincolo di **77 *token***, anche i *prompts* associati a questo *dataset* sono stati abbreviati utilizzando la medesima tecnica di riassunto basata su `GPT4o-mini` applicata al *dataset* NGA, garantendo l'uniformità metodologica.

Il valore di AI-Pastiche risiede nella **dettagliata progettazione manuale dei *prompts***, che garantisce un elevato controllo semantico sulle immagini generate. Ulteriore vantaggio è l'ampia varietà dei dati, con diversificazione di modelli, stili e soggetti. Sebbene i generatori di immagini utilizzati a volte presentino difetti visivi evidenti, tale caratteristica non costituisce un ostacolo, ma è al contrario un elemento diagnostico centrale della presente ricerca. Nella Figure 1.1 riportiamo un esempio di immagini generato dallo stesso prompt.

Figura 1.1: Esempi di immagini provenienti dal dataset AI-Pastiche. Tutte le immagini sono state generate a partire dal seguente prompt:"Generate a view of venice in the vedutism style of the first half of the XVIII century, focusing on a scene along the Grand Canal. The composition features detailed classical architecture with grand domes and facades, and gondolas moving along the canal. Add soft clouds to the sky and ensure there is little fading in the horizon, providing clear visibility of distant buildings. The color palette should include very soft blues and warm earth tones, avoiding saturated colors. The atmosphere remains calm and luminous, with minimal light-and-shadow effects, capturing the beauty and grandeur of Venice from a broad perspective."

1.4.3 AI-ArtBench

AI-ArtBench è stato originariamente sviluppato nell'ambito del progetto ArtBrain [6] da ricercatori della Carnegie Mellon University e dell'UC Berkeley, con l'obiettivo di fungere da **benchmark** per la classificazione degli stili e la distinzione tra opere umane e artificiali. Si tratta di una collezione massiva contenente oltre 185.000 immagini, distribuite su 10 stili. Di queste, 60.000 sono opere create da umani (6.000 immagini per stile) e 125.015 sono riproduzioni sintetiche generate utilizzando *Latent Diffusion* e *Stable Diffusion* con *prompt* condizionati. Le immagini umane hanno dimensioni 256×256, mentre quelle sintetiche sono disponibili nelle risoluzioni 256×256 e 768×768.

I punti di forza di questo *dataset* risiedono in:

1. **Robustezza e Bilanciamento:** ArtBench [7] è uno dei primi *dataset* bilanciati per classe nel dominio artistico. Un *dataset* bilanciato riduce il rischio che i modelli allenati sviluppino *bias* evidenti nella classificazione, conferendo maggiore robustezza al modello addestrato su di esso.

2. **Qualità e Annotazione:** Le immagini umane possiedono un'alta qualità e sono corredate da *captions* chiare e concise che ne descrivono il contenuto. L'intero *dataset* è stato sottoposto a un rigoroso processo di selezione e standardizzazione.

3. **Espansione del Dominio:** L'inclusione di opere artistiche sintetiche arricchisce il dominio del modello, consentendogli di adattare la sua comprensione anche alle rappresentazioni artificiali, raggiungendo un livello superiore di generalizzazione.

Tuttavia, AI-ArtBench presenta alcune criticità legate alla generazione delle opere sintetiche. Gli autori hanno adottato le prime versioni dei modelli *Stable Diffusion* e *Latent Diffusion*, che non erano ancora pienamente perfezionati e che, di conseguenza, producono spesso difetti evidenti. Inoltre, per la generazione, gli autori hanno utilizzato lo stesso *prompt* ("an artwork in {style} art style"), variando unicamente i **semi** di generazione. Il risultato è una collezione di opere artificiali che tendono a essere **rappresentazioni stereotipiche e ripetitive** di un certo stile artistico, con soggetti e composizioni visivamente troppo simili tra loro. Questa omogeneità non favorisce l'arricchimento del *dataset* in termini di diversità, rendendo il modello più incline all'*overfitting*. Sarebbe stato auspicabile l'impiego di *prompt* più ricchi e diversificati per introdurre una maggiore variabilità interpretativa e visiva. Alcuni esempi delle immagini contenute in questo dataset sono illustrati in (Figure 1.2)

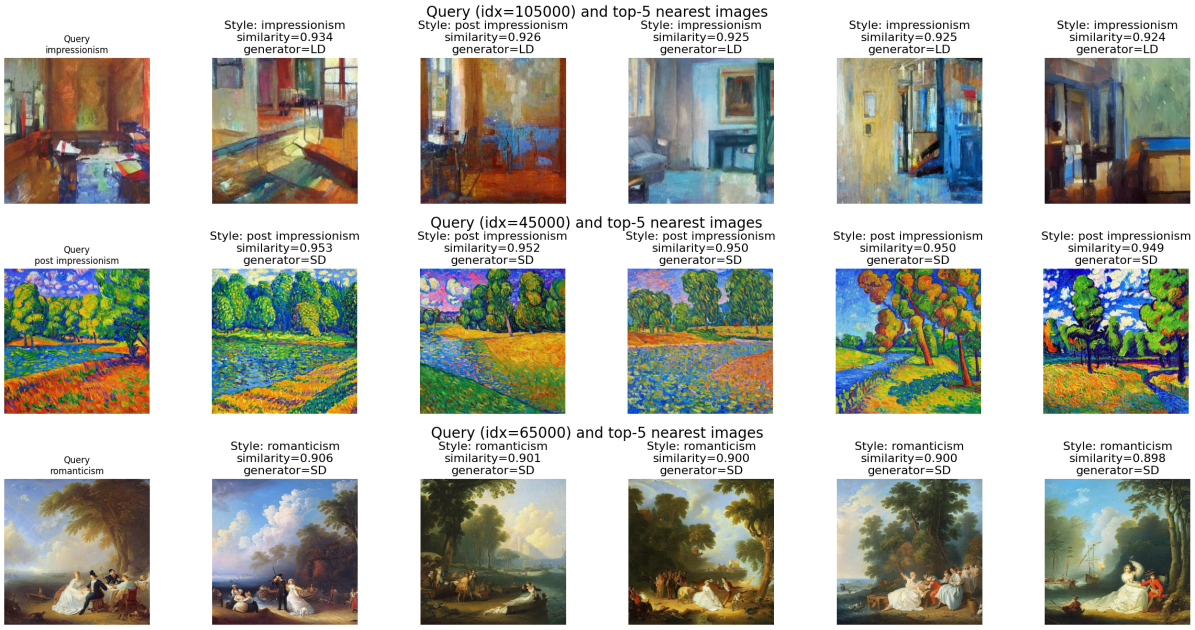

Figura 1.2: Esempi di immagini generate da AI-ArtBench e dei loro *nearest neighbor*, che mostrano la somiglianza visiva responsabile dell'incremento artificiale delle metriche di performance.

1.4.4 AI-WikiArt

AI-WikiArt (o *WikiArt-VLM*) è un *dataset* concepito per valutare le *performance* dei *Vision-Language Models* (VLM) nell'attribuzione degli autori di opere artistiche. Si tratta di una collezione che comprende 39.530 opere reali tratte dal vasto *dataset* pubblico **WikiArt**, con l'aggiunta di tre imitazioni sintetiche per ogni opera reale (39.530 $\times$ 3), generate utilizzando modelli come *Flux*, *Stable Diffusion* e *F-Lite*. Il *dataset* reale è riccamente diversificato, includendo 128 artisti famosi, 10 generi e 27 stili.

La componente sintetica è stata generata tramite una *pipeline* in due fasi:

1. L'opera originale da imitare è stata processata da *GPT4.1-mini* per generare una descrizione dettagliata.

2. Questa descrizione è stata utilizzata come *prompt* di generazione per i modelli, istruendoli a "produrre un'immagine che assomigli a un dipinto di [pittore corretto], ma non una copia esatta tra le sue opere: [descrizione]".

Questo processo mirava a produrre opere sintetiche allineate a quelle reali per stile e contenuto, pronte per la valutazione del modello. Il vantaggio principale del *dataset* risiede nelle **opere reali ricche di metadati**, inclusa l'attribuzione dell'autore. I difetti, tuttavia, si concentrano sulla parte sintetica:

1. I generatori impiegati erano relativamente modesti rispetto agli standard attuali, producendo riproduzioni meno realistiche e contenenti **tracce evidenti di artefatti** visivi e stilistici.

2. Le *captions* generate da *GPT4.1-mini* sono risultate carenti di dettagli stilistici, concentrandosi prevalentemente sul contenuto. Il risultato è che le immagini riprodotte non possiedono una fedeltà stilistica sufficiente all'originale.

Per i motivi sopracitati, l'affidabilità di questo *dataset* per l'addestramento di modelli per l'attribuzione dell'autore è limitata. Per la generazione di immagini sintetiche più fedeli, sarebbe stato necessario integrare la *pipeline* con l'intervento di esperti umani o l'uso di descrizioni artistiche più dettagliate per arricchire i *prompt* e produrre imitazioni più accurate delle opere originali. Gli esempi di questo dataset sono illustrati nella (Figure 1.3)

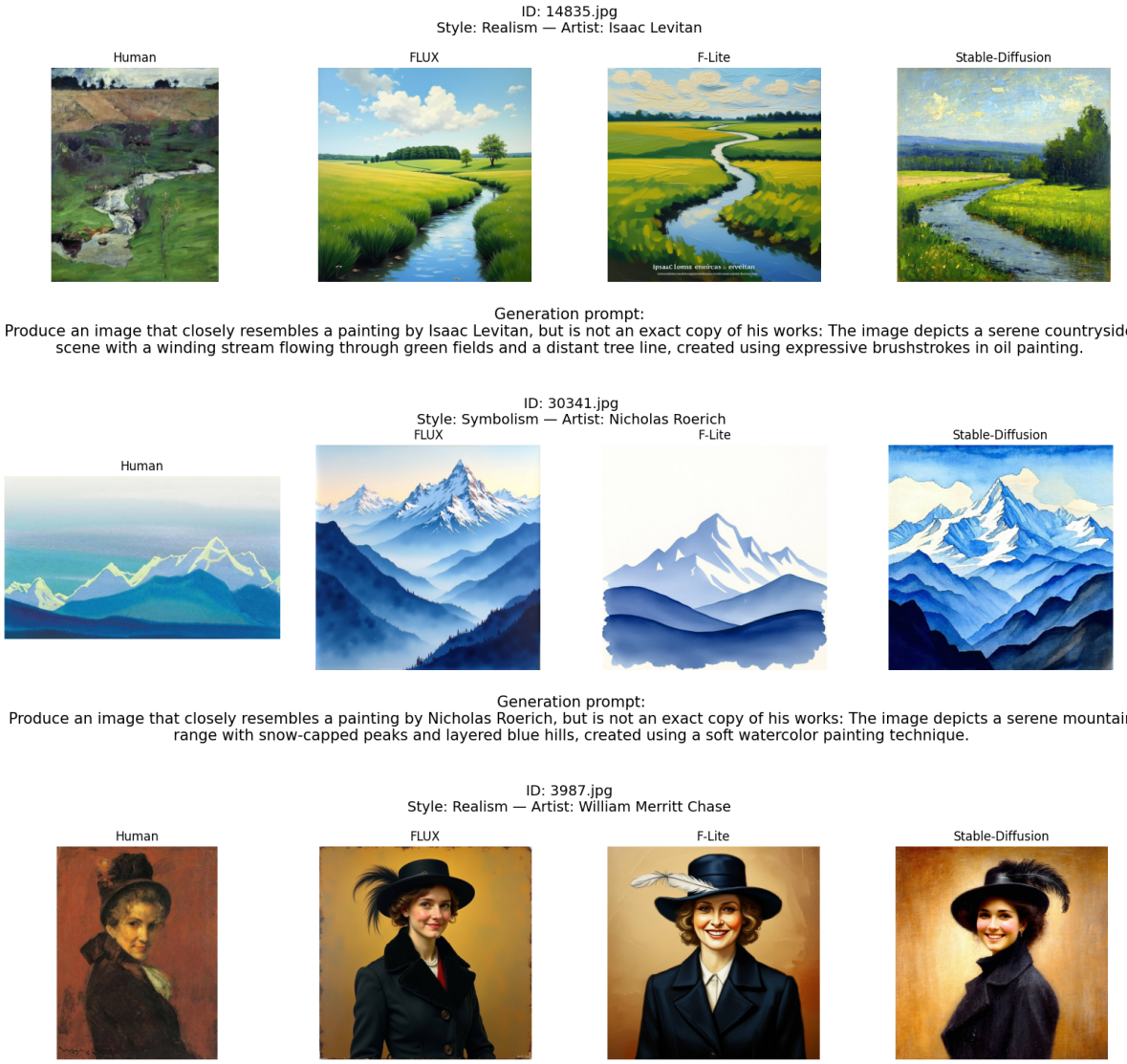

Figura 1.3: Esempi di immagini provenienti dal dataset AI-WikiArt

Capitolo 2

Il Modello CLIP: Architettura e Funzionamento

2.1 Nascita di CLIP

I modelli tradizionali di *Computer Vision* sono stati a lungo basati su un insieme fisso e limitato di categorie di oggetti, come nel caso di ImageNet [8]. Sebbene tali modelli abbiano raggiunto prestazioni elevate nei compiti per cui sono stati addestrati, la loro capacità di generalizzare a domini diversi rimane ridotta. Quando si presentano nuovi oggetti o classi non incluse nel *dataset* di addestramento, è infatti necessario introdurre manualmente nuove etichette e riaddestrare il modello. Questa dipendenza da una supervisione fortemente vincolata ha rappresentato per anni uno dei principali limiti della visione artificiale.

Per superare tali restrizioni, il team di OpenAI ha introdotto **CLIP** (*Contrastive Language-Image Pre-training*) [1], un modello multimodale in grado di apprendere rappresentazioni visive e testuali condivise a partire da una vasta collezione di dati naturali. L'idea centrale di CLIP è di sfruttare il linguaggio naturale come forma di supervisione generalizzabile, anziché basarsi su etichette fisse e predefinite.

L'architettura di CLIP è composta da due moduli principali, addestrati congiuntamente secondo un approccio contrastivo:

1. **Encoder di immagine (*Image Encoder*)**: utilizza varianti di architetture note come ResNet (RN50, RN101) o, più frequentemente, il **Vision Transformer (ViT)** [9], per mappare un'immagine in un vettore di *embedding*.

2. **Encoder di testo (*Text Encoder*)**: si basa su un modello *Transformer* (simile a GPT) adattato per rappresentare frasi o didascalie come vettori nello stesso spazio latente dell'encoder visivo.

Durante il pre-addestramento, CLIP viene esposto a circa 400 milioni di coppie (immagine, testo). Dato un *batch* di N immagini e N descrizioni testuali, il modello deve individuare le corrispondenze corrette tra tutte le N^2 possibili combinazioni. A tale scopo viene utilizzata una **funzione di perdita contrastiva** che massimizza la similarità del coseno tra i vettori di immagini e testi corrispondenti, e la minimizza per le coppie non corrispondenti. In questo modo, i due encoder imparano a rappresentare immagini e testi in uno **spazio latente condiviso**, dove elementi semantici simili risultano vicini.

Uno dei risultati più notevoli di questo addestramento è la capacità di **generalizzazione zero-shot**. Una volta addestrato, CLIP può affrontare nuovi compiti di classificazione senza ulteriore *fine-tuning*, semplicemente confrontando l'immagine con descrizioni testuali formulate in linguaggio naturale. Il processo di classificazione zero-shot avviene come segue:

1. Si definisce un insieme di frasi che descrivono le classi d'interesse (ad esempio, "una foto di un gatto", "una foto di un cane").
2. L'immagine da classificare viene convertita nel suo *embedding* visivo.
3. Si calcola la similarità del coseno tra tale vettore e gli *embedding* testuali delle descrizioni.
4. L'immagine viene assegnata alla classe con la similarità maggiore.

Questo meccanismo consente a CLIP di raggiungere o superare modelli supervisionati su decine di *dataset* di *Computer Vision*, dimostrando un'elevata **robustezza e capacità di trasferimento** tra domini.

2.2 Impatto e Applicazioni nel Panorama della Computer Vision

L'introduzione di CLIP ha avuto un impatto profondo sul campo della visione artificiale, estendendo il suo impiego a una grande varietà di compiti multimodali.

- **Guida per modelli generativi:** CLIP viene utilizzato come metrica di allineamento semantico tra testo e immagine nei modelli *text-to-image*, come *Stable Diffusion* [2, 10], *DALL-E* [11] e GLIDE [12], dove funge da funzione di perdita di guida [13, 14] (*guidance loss*).
- **Classificazione zero-shot fine-grained:** grazie alla sua capacità di cogliere differenze sottili, CLIP si è dimostrato efficace nella classificazione di categorie visivamente simili [15, 16, 17].

- **Image retrieval e ricerca semantica:** lo spazio multimodale di CLIP è ampiamente impiegato nella *semantic search* [18, 19, 20] e nella generazione di immagini basati su prompts [21, 22], consentendo di cercare immagini tramite descrizioni testuali complesse.
- **Altri ambiti applicativi:** CLIP è stato applicato con successo anche a compiti di *Optical Character Recognition* (OCR), riconoscimento di azioni nei video e geolocalizzazione visiva.

2.3 Limiti e criticità del modello

Nonostante gli eccellenti risultati ottenuti, il modello CLIP presenta numerose limitazioni, soprattutto quando viene applicato a domini complessi come quello artistico. Diversi studi hanno infatti evidenziato che le sue prestazioni derivano spesso da **correlazioni superficiali** tra elementi visivi e testuali, più che da una reale comprensione semantica o composizionale. In particolare, CLIP mostra difficoltà nel trattare concetti come la **negazione**, l'**ordine degli oggetti**, il **conteggio** e l'associazione corretta tra **attributi e oggetti** [23, 24, 25].

Esperimenti condotti su scene contenenti più oggetti hanno inoltre messo in luce la presenza di **bias strutturali** nei due encoder che compongono il modello [26]:

- l'*encoder* visivo tende a privilegiare gli oggetti di dimensioni maggiori;
- l'*encoder* testuale assegna un peso maggiore agli elementi menzionati per primi nel *prompt*.

Tali asimmetrie compromettono la capacità del modello di cogliere le relazioni spaziali e narrative tra gli elementi di una scena, un aspetto particolarmente rilevante nell'analisi delle opere artistiche, dove la disposizione e l'interazione visiva tra i soggetti rivestono un ruolo fondamentale.

Un'ulteriore criticità riguarda la **robustezza** del modello. Sebbene CLIP si dimostri più resistente rispetto ai modelli basati su ImageNet di fronte a variazioni visive comuni [27], esso rimane vulnerabile a:

- **attacchi avversari** (*adversarial attacks*), in grado di manipolare le predizioni del modello mediante perturbazioni visive impercettibili [28];
- **sensibilità al prompt**, ovvero variazioni significative nelle prestazioni dovute a differenze anche minime nella formulazione testuale delle query [29].

Infine, poiché CLIP è stato addestrato su un ampio corpus di dati non filtrati provenienti dal web, esso riflette e talvolta amplifica i **bias sociali e culturali** presenti nel dataset di addestramento [30]. Tale fenomeno risulta particolarmente problematico in

ambiti estetici: se, ad esempio, le descrizioni di "arte classica" nei dati di training fanno riferimento prevalentemente a determinati periodi storici o stili, il modello può mostrare difficoltà nel riconoscere correttamente varianti stilistiche contemporanee o ibride.

2.4 Tecniche di Ottimizzazione e Adattamento

Per superare i limiti evidenziati, la ricerca recente ha sviluppato diverse strategie di adattamento di CLIP, mirate a migliorarne le prestazioni in compiti specifici preservando al contempo le conoscenze acquisite durante il pre-addestramento.

Un approccio diffuso consiste nell'inserire **moduli leggeri** all'interno dell'architettura del modello per eseguire un *fine-tuning* efficiente su nuovi compiti, senza modificare i pesi originali. Esempi di questa famiglia sono *CLIP-Adapter* [31], *TIP-Adapter* [32] e *LIxP* [33].

Altri metodi, come *CoOp* [34] e *CoCoOp* [35], mirano a ridurre la sensibilità al *prompt* apprendendo vettori di *token* aggiuntivi che fungono da contesto adattivo. Tuttavia, nel dominio estetico tali approcci risultano meno efficaci, poiché il divario semantico tra descrizioni testuali e caratteristiche visive stilistiche è spesso troppo ampio per essere colmato da un semplice *prompt tuning*.

Un'ulteriore linea di ricerca propone di proiettare le *feature* latenti di CLIP in **sottospazi a dimensione ridotta** o specifici per il compito, come nei modelli *APE* e *CLIP-Subspace* [36, 37]. Questo tipo di approccio è particolarmente interessante in quanto:

- la similarità del coseno non è invariante rispetto a trasformazioni lineari;
- uno strato di proiezione può enfatizzare le *feature* più rilevanti (ad esempio, quelle legate allo stile o al periodo storico) senza modificare il *core* del modello.

Tali strategie di proiezione sono anche alla base dei più recenti modelli *Vision-Language* come *LLaVA* [38], *Qwen-VL* [39] e *InstructBLIP* [40], che impiegano proiezioni allenabili — ad esempio il *Q-Former* introdotto in *BLIP-2* [41] — per collegare efficacemente le rappresentazioni visive e linguistiche. Tuttavia, questi modelli sono maggiormente orientati all'*instruction-following* (cioè la capacità di seguire istruzioni testuali) piuttosto che all'analisi interpretativa delle immagini, e pertanto non rientrano nel focus della presente ricerca.

Infine, studi basati su tecniche di *gradient ascent* [42, 43, 44] hanno esplorato la possibilità di **ricostruire immagini a partire dagli embedding di CLIP** [45]. Questi metodi consentono di indagare ciò che il modello "vede" e come struttura le proprie rappresentazioni interne, offrendo strumenti preziosi per valutarne la trasparenza interpretativa. Tale prospettiva è particolarmente utile nello studio dell'arte, dove comprendere i meccanismi percettivi del modello è tanto importante quanto misurarne l'accuratezza.

Nel complesso, tali approcci riflettono un crescente sforzo della comunità scientifica per ampliare l'utilità a valle di CLIP e, al contempo, per accedere e interpretare meglio la struttura del suo spazio di embedding. Nel contesto della presente ricerca, tuttavia, si è scelto di evitare metodologie che implichino processi di *fine-tuning*, concentrandosi invece su analisi basate su proiezioni lineari, utili per comprendere più a fondo la natura e l'organizzazione intrinseca dello spazio latente di CLIP.

Capitolo 3

Text–Image Alignment

3.1 Obiettivi e Metodologia

Il primo passo della nostra analisi consiste nel valutare la capacità di CLIP di associare immagini e descrizioni testuali. Questo esperimento fornisce una baseline utile per comprendere in che misura il modello riesca a cogliere la coerenza semantica fra i due domini prima di affrontare compiti più complessi come il riconoscimento dello stile artistico.

La procedura è stata sviluppata in collaborazione con il gruppo di lavoro e si basa sul confronto fra gli embeddings testuali e visivi. Dati $A \in \mathbb{R}^{n \times d}$ (embeddings testuali) e $B \in \mathbb{R}^{m \times d}$ (embeddings visivi), la matrice $C = AB^{\top}$ contiene per ogni coppia immagine–testo la loro similitudine coseno. Per ogni immagine si seleziona la descrizione con punteggio più alto e si valuta la correttezza della predizione tramite *accuracy* o *recall@K* a seconda della dimensione del dataset.

Il mio contributo principale riguarda l'intera pipeline di text–image alignment sul dataset **AI-Pastiche**: preprocessing, valutazione con diversi modelli CLIP e analisi dettagliata dei risultati.

3.2 Datasets Utilizzati

L'esperimento è stato condotto su tre dataset:

- **AI-Pastiche**: 73 descrizioni generate tramite prompt dettagliati; questo è il dataset sul quale ho svolto completamente l'analisi.
- **NGAD**: circa 1.500 descrizioni; analisi svolta in collaborazione con il gruppo.
- **AI-WikiArt**: circa 40.000 testi; analisi svolta in collaborazione con il gruppo.

Il dataset **AI-ArtBench** è stato escluso poiché le sue descrizioni estremamente sintetiche non consentono una valutazione semantica significativa dell'allineamento.

3.3 Limitazioni dei Prompt e Preprocessing

I dataset NGAD e AI-WikiArt includono descrizioni spesso lunghe. Per rispettare il limite dei 77 token dell'encoder testuale di CLIP, le descrizioni sono state riassunte mediante un sistema di summarization automatica basato su `GPT4o-mini`. Questa parte è stata svolta congiuntamente dal gruppo.

Per tutte le immagini è stato inoltre applicato il preprocessing standard di CLIP (ridimensionamento, center crop, normalizzazione). Tale pipeline può rimuovere porzioni rilevanti dell'immagine, soprattutto in caso di rapporti di forma estremi, influenzando direttamente le performance dei modelli.

3.4 Risultati

3.4.1 Risultati su NGAD

Il dataset NGAD permette una valutazione basata su *recall@K*. I risultati, ottenuti nell'ambito del lavoro collettivo, sono riportati in Tab. 3.1.

Model	**recall@1**	**recall@5**	**recall@10**
RN50	0.663	0.915	0.966
RN101	0.693	0.926	0.966
RN50x4	0.741	0.946	0.978
RN50x16	0.791	0.964	0.988
RN50x64	**0.828**	0.970	0.990
ViT-B/32	0.678	0.925	0.970
ViT-B/16	0.709	0.928	0.969
ViT-L/14	0.794	0.972	0.989
ViT-L/14@336px	0.814	**0.974**	**0.991**

Tabella 3.1: Summary–image alignment per NGAD.

I modelli più grandi (RN50x64, ViT-L/14@336px) forniscono le prestazioni migliori, confermando la dipendenza del compito dalla capacità rappresentativa del modello.

3.4.2 Risultati su AI-Pastiche (mio contributo)

Questa sezione rappresenta il contributo principale del mio lavoro. L'analisi completa del dataset AI-Pastiche è stata svolta da me, includendo preprocessing, implementazione della pipeline di allineamento e valutazione dei modelli.

Model	Accuracy
RN50	0.866
RN101	0.887
RN50x4	0.891
RN50x16	0.893
RN50x64	**0.896**
ViT-B/32	0.881
ViT-B/16	0.880
ViT-L/14	**0.896**
ViT-L/14@336px	**0.896**

Tabella 3.2: Accuracy su AI-Pastiche (mio contributo).

Le performance sono elevate per tutti i modelli, con un'accuracy massima pari a 0.896. In particolare, RN50x64 e ViT-L/14@336px si confermano i modelli più efficaci. L'elevata coerenza delle immagini generate con i prompt rende il compito relativamente meno ambiguo rispetto ad altri dataset.

3.4.3 Risultati su AI-WikiArt

I risultati su AI-WikiArt (ottenuti con ViT-L/14@336px) sono riportati in Tab. 3.3. Questa parte è stata condotta dal gruppo.

	Recall@1	**Recall@5**	**Recall@10**
WikiArt (AI-generated)	0.4258	0.6565	0.7416

Tabella 3.3: Prompt–image alignment su AI-WikiArt.

Le prestazioni risultano inferiori a causa della scarsa qualità semantica delle descrizioni sintetiche e della bassa coerenza delle immagini generate.

3.5 Conclusioni

L'esperimento mostra che CLIP è efficace nell'associare immagini e descrizioni quando i testi sono ricchi e le immagini rappresentano fedelmente il contenuto. Il mio contributo principale è stato l'intero processo di valutazione del dataset AI-Pastiche, i cui risultati confermano le capacità del modello nel contesto di immagini generate di alta qualità. I risultati su NGAD e AI-WikiArt, ottenuti in collaborazione, forniscono un quadro comparativo utile e mostrano i limiti introdotti da dataset più rumorosi o semantici meno strutturati.

Capitolo 4

Riconoscimento dello Stile Artistico

4.1 Evoluzione delle tecniche di classificazione

L'attribuzione automatica dello stile artistico ha subito una profonda evoluzione parallela ai progressi della Computer Vision. Inizialmente affidata all'analisi di esperti, la classificazione ha visto i primi tentativi di automazione tramite l'ingegnerizzazione manuale di *feature* (colori, texture, composizione) processate da algoritmi classici come SVM o k-NN [46, 47]. Tuttavia, la natura astratta dei concetti artistici rendeva questi approcci poco generalizzabili.

La svolta è giunta con le Reti Neurali Convoluzionali (CNN). L'uso del *Transfer Learning* e del *Fine-tuning* su reti pre-addestrate (es. su ImageNet) ha permesso di ottenere rappresentazioni molto più ricche, superando i metodi precedenti [48]. Recentemente, il paradigma si è spostato verso i modelli multimodali *Vision-Language* (VLM) come CLIP, che apprendono uno spazio latente condiviso tra immagini e testo. Sebbene promettenti, resta da chiarire se tali modelli, allenati su dati generici del web, possiedano la granularità necessaria per distinguere stili pittorici complessi senza soffrire di bias semantici o allucinazioni visive, come evidenziato in recenti studi su LMM [49].

Il presente capitolo indaga l'efficacia di CLIP come *feature extractor* per lo stile. Il lavoro qui presentato è frutto di una ricerca collaborativa; nello specifico, il contributo dell'autore si concentra sull'analisi delle performance relative ai dataset sintetici (in particolare **AI-Pastiche**) e sulla valutazione della capacità di generalizzazione del modello tra dominio reale e artificiale (esperimenti *Inter-Dataset*).

4.2 Metodologia Sperimentale

Per valutare le capacità di CLIP, abbiamo adottato un approccio incrementale, partendo dalle capacità native del modello fino all'addestramento di classificatori dedicati.

4.2.1 Approccio Zero-Shot (Baseline)

Come punto di partenza, abbiamo testato la capacità "nativa" di CLIP di classificare lo stile confrontando la similarità coseno tra l'embedding dell'immagine e i prompt testuali del tipo *"an artwork in {style} style"*. Questo serve a quantificare il *modality gap*, ovvero la distanza tra la rappresentazione visiva dello stile e la sua descrizione testuale.

4.2.2 Linear Probing (Focus della Tesi)

Per superare i limiti dello Zero-Shot, abbiamo utilizzato la tecnica del *Linear Probing*: addestrare un classificatore lineare (Regressione Logistica) direttamente sugli embedding visivi estratti da CLIP. Questo approccio ci permette di capire se l'informazione stilistica è presente nello spazio latente, indipendentemente dalla capacità del modello di collegarla al testo. Gli esperimenti sono stati condotti in due configurazioni:

- **Intra-Dataset:** Addestramento e test sullo stesso dataset (split 80/20). Questa fase, svolta in collaborazione, ha coinvolto i dataset AI-WikiArt, AI-Artbench e AI-Pastiche.
- **Inter-Dataset (Cross-Domain):** Addestramento su opere reali (AI-WikiArt Human) e test su tutti gli altri domini, inclusi quelli sintetici. Questa fase rappresenta il nucleo del contributo dell'autore, mirato a verificare se la nozione di "stile" appresa su quadri reali sia trasferibile alle imitazioni generate dalle IA.

4.2.3 Datasets

Le analisi hanno coinvolto quattro dataset principali: **NGAD** e **AI-WikiArt** (opere reali), **AI-ArtBench** (misto) e **AI-Pastiche** (interamente sintetico). L'uso di AI-Pastiche è centrale per valutare il comportamento del modello su generazioni artificiali controllate.

4.3 Analisi dei Risultati

Sulla base delle analisi preliminari di *Text Alignment*, il modello di riferimento utilizzato è `ViT-L/14@336px`.

4.3.1 Limiti della classificazione Zero-Shot

I risultati Zero-Shot (Tabella 4.1) fungono da *baseline*. Le performance sono modeste (circa 30% su NGAD), evidenziando che CLIP, senza adattamento, fatica a collegare l'immagine al nome dello stile.

L'analisi visiva tramite UMAP (Figura 4.1) conferma un netto *modality gap*: gli embedding delle immagini e quelli dei testi risiedono in regioni distinte dello spazio

Dataset	Accuracy	Recall@1	Recall@5
NGAD	0.3006	0.3006	0.7189
AI-Pastiche	0.4974	0.4974	0.8429
AI-WikiArt (Human)	0.3664	0.3664	0.7189
AI-ArtBench (Human)	0.5516	0.5516	0.9370

Tabella 4.1: Sintesi dei risultati Zero-Shot (Baseline).

vettoriale. Inoltre, si nota che CLIP tende a raggruppare le opere più per contenuto semantico (soggetto) che per stile artistico.

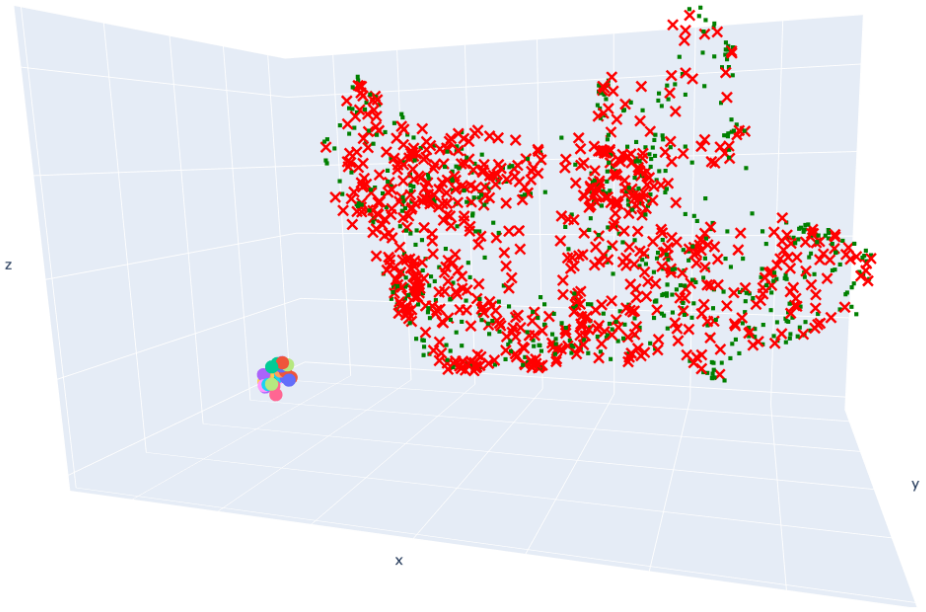

Figura 4.1: Proiezione UMAP: separazione tra spazio visivo e testuale.

4.3.2 Efficacia del Linear Probing

L'addestramento di un classificatore lineare ha rivelato che le informazioni stilistiche sono effettivamente codificate negli embedding visivi, ma necessitano di supervisione per emergere.

Analisi Intra-Dataset

Negli esperimenti condotti in collaborazione su AI-WikiArt e AI-ArtBench, e specificamente su AI-Pastiche, l'uso del Linear Probing con embedding non normalizzati ha portato a un notevole incremento delle performance, con accuratezze che oscillano tra il

74% e il 77% per i dataset reali (Tabella 4.2). Questo dimostra che CLIP possiede una ricca rappresentazione latente dello stile.

Dataset	Accuracy	F1
NGAD	0.7771	0.7775
AI-WikiArt (Human)	0.7528	0.7543
AI-ArtBench (Human)	0.7495	0.7491

Tabella 4.2: Risultati Intra-Dataset (Linear Probing non normalizzato).

Generalizzazione Cross-Domain (Inter-Dataset)

Questa sezione descrive il nucleo dell'indagine condotta dall'autore: valutare come un modello addestrato su arte umana (**AI-WikiArt Human**) si comporti su domini diversi, specialmente quelli sintetici (**AI-Pastiche**).

Dataset di Test	Accuracy	F1
Dominio Reale (Controllo)		
AI-WikiArt Human (Test split)	0.7654	0.7670
NGAD	0.4729	0.4831
Dominio Sintetico (Target)		
AI-Pastiche	**0.3452**	0.3682
AI-ArtBench (AI)	0.4905	0.5093
AI-WikiArt (AI)	0.3088	0.3087

Tabella 4.3: Generalizzazione Inter-Dataset (Train su AI-WikiArt Human).

Come evidenziato in Tabella 4.3, si osserva un calo di prestazioni generalizzato. Tuttavia, il dato più rilevante riguarda **AI-Pastiche**, che crolla al 34.52%. Questo risultato suggerisce una conclusione fondamentale: lo "stile" generato dalle IA possiede caratteristiche statistiche distributive diverse dallo stile umano originale. Sebbene visivamente simile, l'imitazione sintetica non inganna lo spazio latente di un classificatore addestrato su opere reali. Il risultato leggermente migliore su AI-ArtBench (AI) è probabilmente dovuto al fenomeno del *shared inductive bias*, dato che i generatori usati per quel dataset sono strettamente legati all'architettura di CLIP.

4.3.3 Confronto con metodi Few-Shot (Contesto Collaborativo)

Per validare la scelta del Linear Probing, il gruppo di ricerca ha effettuato un confronto con *Adaptive Prior Refinement* (APE) [36], una tecnica avanzata di *few-shot learning*. Gli esperimenti su NGAD hanno mostrato che APE, anche con 16 esempi per classe, raggiunge un'accuratezza massima del 53.42%, ben lontana dal 77.71% ottenuto con il

nostro approccio a supervisione completa. Questo confronto conferma che per compiti a grana fine come lo stile artistico, le tecniche few-shot non riescono ancora a estrarre adeguatamente le feature latenti, rendendo necessario l'addestramento di classificatori dedicati (Linear Probing) come svolto in questa tesi.

4.4 Conclusioni

L'analisi ha dimostrato che CLIP codifica informazioni stilistiche significative, ma che queste rimangono latenti e richiedono supervisione per essere attivate (Linear Probing). Il contributo specifico sugli esperimenti Inter-Dataset e su AI-Pastiche ha evidenziato una discrepanza tra arte reale e sintetica: il modello, pur riconoscendo lo stile su dati omogenei, fatica a generalizzare dalle opere umane alle loro imitazioni artificiali. Ciò indica che i generatori attuali, seppur potenti, non replicano fedelmente la "firma" statistica dello stile presente nello spazio latente di CLIP.

Capitolo 5

Valutazione Computazionale dell'Estetica

5.1 Evoluzione dello stato dell'arte

L'applicazione di modelli di visione artificiale per la stima della qualità estetica definisce il campo della *Computational Aesthetics*. Storicamente, questa disciplina ha cercato di emulare il giudizio umano attraverso approcci progressivi. Le prime ricerche tentavano di codificare la bellezza tramite formule matematiche rigide basate su simmetria e complessità (Birkhoff, 1933) o mediante l'estrazione manuale di feature fotografiche come la regola dei terzi [50, 51]. Tuttavia, l'astrazione soggettiva del bello si è rivelata refrattaria a tali modellizzazioni deterministiche.

La svolta è arrivata con il Deep Learning: l'introduzione delle CNN e l'uso di dataset su larga scala come AVA hanno permesso modelli come RAPID [52] di superare i limiti delle feature artigianali. Più recentemente, l'avvento di CLIP ha segnato un ulteriore passo avanti, mostrando correlazioni con il giudizio umano superiori ai metodi precedenti [53]. Nonostante ciò, resta aperta una questione cruciale: l'efficacia di CLIP nel valutare immagini sintetiche. A differenza delle foto naturali, le immagini generate da IA soffrono di artefatti specifici (es. incongruenze geometriche) che l'occhio umano penalizza immediatamente. Verificare se CLIP possieda una simile sensibilità ai difetti è l'obiettivo centrale di questo capitolo.

5.2 Protocollo Metodologico

Per determinare se CLIP possa fungere da giudice di qualità per l'arte sintetica, abbiamo strutturato un'indagine basata sul confronto diretto con annotazioni umane [54]. Il protocollo si divide in tre fasi analitiche:

1. **Analisi Zero-shot:** Valutazione della coerenza tra il punteggio di aderenza calcolato da CLIP e il giudizio umano.

2. **Probing degli Artefatti:** Verifica della presenza di informazioni relative ai difetti visivi all'interno dello spazio latente del modello.

3. **Integrazione Supervisionata:** Test sull'efficacia di combinare la similarità semantica con un punteggio esplicito dei difetti.

5.2.1 Fase 1: Stima dell'Aderenza Semantica

Nel primo esperimento, abbiamo raccolto valutazioni umane sull'aderenza tra immagine e prompt nel dataset AI-Pastiche, utilizzando una scala ternaria (Bad/Neutral/Good). Definito P come il prompt e $\{I_1, \ldots, I_n\}$ le immagini generate, calcoliamo gli embedding CLIP v_i (visivo) e t (testuale). La similarità coseno grezza è data da:

$$s_i = \cos(v_i, t) = \frac{v_i \cdot t}{\|v_i\| \, \|t\|}.$$

Per garantire la comparabilità con i giudizi umani h_i (media delle valutazioni normalizzate), riscaliamo la similarità in $[-1, 1]$ ottenendo $\tilde{s}_i$. L'efficacia del modello è quantificata dall'allineamento coseno tra i vettori dei punteggi:

$$\text{Align} = \cos(\tilde{s}, h).$$

Un'osservazione preliminare fondamentale è che gli umani tendono a penalizzare fortemente le immagini che, pur semanticamente corrette, presentano gravi difetti visivi.

5.2.2 Fase 2: Rilevamento Latente dei Difetti

Per accertare se CLIP "veda" i difetti, abbiamo classificato gli artefatti in tre livelli: *None*, *Minor* (imperfezioni lievi) e *Major* (errori anatomici gravi). Abbiamo quindi tentato di predire il punteggio di difettosità umano d_i addestrando un regressore lineare direttamente sugli embedding visivi v_i:

$$\hat{d}_i = w^\top v_i.$$

I pesi w sono ottimizzati minimizzando l'errore quadratico medio. Un valore basso del coefficiente R^2 in questo esperimento indicherebbe che l'informazione sulla qualità tecnica non è linearmente accessibile nello spazio latente di CLIP.

5.2.3 Fase 3: Modello Ibrido con Defect Score

Data l'ipotesi che CLIP ignori i difetti, abbiamo proposto un modello correttivo che integra esplicitamente il giudizio sui difetti d_i con la similarità semantica $\hat{s}_i$:

$$\tilde{y}_i = a\hat{s}_i + bd_i + c.$$

I parametri vengono appresi per massimizzare la somiglianza con il giudizio umano complessivo y_i. Un miglioramento delle performance in questa fase confermerebbe che la mancanza di sensibilità agli artefatti è un limite primario di CLIP.

5.3 Dataset di Riferimento

L'analisi è stata condotta esclusivamente su **AI-Pastiche** (vedi Sez. 1.4.2). La scelta è obbligata: è l'unico dataset in nostro possesso che fornisce annotazioni umane granulari sia sull'aderenza al prompt che sulla presenza di artefatti visivi, permettendo di disaccoppiare la valutazione semantica da quella tecnica.

5.4 Analisi dei Risultati

Riportiamo di seguito gli esiti quantitativi del confronto tra percezione macchina e umana.

Modello	**Allineamento con Umani**	**Post-Integrazione Defect Score**
RN50	0.406	0.478
RN50x64	0.428	0.484
ViT-B/32	0.411	0.481
ViT-L/14	0.425	0.482
ViT-L/14@336px	**0.437**	**0.497**

Tabella 5.1: Confronto dell'allineamento coseno: performance base (colonna 1) vs integrazione manuale dei difetti (colonna 2).

Analisi Zero-shot: Il modello **ViT-L/14@336px** ottiene il miglior allineamento (0.437), ma il valore assoluto indica una correlazione solo moderata. È emersa una sistematica divergenza di scala: gli umani tendono a essere più indulgenti (media voti ≈ 0.7) rispetto alle stime di similarità più conservative di CLIP.

Cecità ai Difetti: Il secondo esperimento ha prodotto un risultato negativo netto: la regressione lineare sugli embedding ha restituito valori di $R^2 \approx 0$. Questo conferma che CLIP non codifica, almeno linearmente, la distinzione tra un'immagine tecnicamente perfetta e una affetta da gravi artefatti generativi.

Impatto della Correzione: Come evidenziato nella seconda colonna della Tabella 5.1, l'iniezione manuale del *defect score* migliora le prestazioni di tutte le architetture, portando il ViT-L/14@336px a un allineamento di 0.497. Questo guadagno dimostra che la componente "qualità tecnica" è una variabile latente fondamentale che CLIP trascura. Tuttavia, il fatto che l'allineamento non raggiunga valori prossimi a 1 suggerisce che esistano ulteriori fattori (probabilmente stilistici o emotivi) che influenzano il giudizio umano ma sfuggono ancora alla comprensione del modello.

5.5 Conclusioni

L'indagine ha evidenziato che l'allineamento estetico tra CLIP e l'uomo è presente ma incompleto. Il modello eccelle nel valutare la coerenza semantica (il contenuto), ma fallisce nel penalizzare gli artefatti strutturali (la forma), che sono invece determinanti per l'osservatore umano. L'incapacità di estrarre informazioni sui difetti dagli embedding visivi suggerisce che CLIP, nel suo pre-addestramento, impara a ignorare il "rumore" visivo per concentrarsi sul concetto, diventando così cieco agli errori generativi. I risultati del modello ibrido indicano che per costruire metriche di valutazione automatiche affidabili per l'arte sintetica, è necessario affiancare ai VLM attuali dei moduli specializzati nella rilevazione della coerenza strutturale e degli artefatti.

5.6 Conclusioni

La nostra indagine sulla capacità di CLIP di interpretare le opere d'arte — abbracciando sia la produzione umana che le immagini generate dall'IA — delinea un modello di ampia portata, sebbene ancora limitato nel cogliere la profondità e le sfumature proprie della comprensione estetica umana. Se da un lato CLIP dimostra competenza nell'associare immagini a macro-categorie semantiche e nel produrre descrizioni testuali plausibili, dall'altro incontra sistematiche difficoltà nel gestire le dimensioni più soggettive dell'analisi artistica, quali lo stile, l'intento autoriale, il tono emotivo, il contesto culturale e la precisione tecnica.

Un aspetto critico emerso riguarda l'assenza di una capacità affidabile nel rilevare gli artefatti genuini del processo generativo, come le distorsioni anatomiche di mani o volti e altre incoerenze strutturali tipiche dell'immaginario sintetico. Tali difetti, sebbene salienti per un osservatore umano, non trovano un riscontro coerente all'interno degli embedding visivi di CLIP.

Inoltre, si osserva un marcato divario di dominio (*domain gap*) nel tentativo di generalizzare la classificazione stilistica attraverso dataset distinti. Sebbene CLIP offra prestazioni adeguate all'interno di singole collezioni, sia l'accuratezza che la coerenza degradano quando le etichette stilistiche vengono trasferite tra dataset che differiscono per distribuzione delle immagini, pratiche curatoriali o schemi di annotazione. Questa limitazione suggerisce una scarsa robustezza nell'astrazione delle feature stilistiche al di fuori dei contesti specifici di apprendimento, una sfida che deve essere affrontata affinché tali modelli possano supportare applicazioni come il *retrieval* inter-collezione, l'analisi culturale o il supporto alle decisioni curatoriali.

A un livello più fondazionale, questa analisi solleva considerazioni rilevanti sulla struttura delle rappresentazioni multimodali. Sebbene lo spazio di embedding congiunto di CLIP aiuti a colmare il divario tra visione e linguaggio, questa connessione rimane asimmetrica: gli embedding testuali sono intrinsecamente strutturati e interpretabili, mentre quelli delle immagini risultano più opachi e spazialmente "aggrovigliati" (*entangled*). Di conseguenza, valutare la comprensione visiva attraverso la similarità testuale può condurre a conclusioni fuorvianti, in particolare nei domini artistici, dove il significato è spesso veicolato tramite segnali visivi non verbali, ambigui o dipendenti dal contesto. Elementi come la testura, la composizione, la gestualità e l'affettività spesso resistono a una codifica linguistica diretta, esponendo i limiti chiave delle attuali metodologie di valutazione per i modelli multimodali.

In prospettiva, l'avanzamento dei sistemi visione-linguaggio richiederà un passaggio da un allineamento superficiale verso un modello di percezione visiva più profondo e olistico. Tali sistemi dovranno essere capaci di ragionare sulle immagini non solo in termini di oggetti e stili, ma anche rispetto ai riferimenti storici, all'intento artistico e alla narrativa visiva. Il raggiungimento di questo obiettivo potrebbe richiedere nuove forme di supervisione che vadano oltre l'addestramento basato su didascalie (*caption-based*),

per includere commenti di esperti, metadati storico-artistici, annotazioni culturalmente informate e interazioni multimodali dialogiche.

Poiché l'intelligenza artificiale diventa sempre più integrata nella produzione, interpretazione e disseminazione della cultura visiva, è essenziale considerare non solo *cosa* i modelli percepiscano, ma *come* lo facciano e quali prospettive vengano implicitamente codificate. CLIP offre una base solida, ma non approssima ancora la percezione umana nelle arti. Piuttosto, dovrebbe essere considerato come una lente interpretativa parziale e distorta, capace tuttavia di offrire intuizioni preziose sia sulla visione artificiale che sui modi umani di vedere.

Bibliografia

[1] A. Radford et al., «Learning Transferable Visual Models From Natural Language Supervision,» in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, M. Meila e T. Zhang, cur., ser. Proceedings of Machine Learning Research, Accessed: 2025-02-13, vol. 139, PMLR, 2021, pp. 8748–8763. indirizzo: http://proceedings.mlr.press/v139/radford21a.html.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser e B. Ommer, «High-resolution image synthesis with latent diffusion models,» in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, 2022, pp. 10 684–10 695.

[3] L. A. Gatys, A. S. Ecker e M. Bethge, «A Neural Algorithm of Artistic Style,» *CoRR*, vol. abs/1508.06576, 2015. indirizzo: http://arxiv.org/abs/1508.06576.

[4] R. Zhang, P. Isola, A. A. Efros, E. Shechtman e O. Wang, «The Unreasonable Effectiveness of Deep Features as a Perceptual Metric,» in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 586–595. DOI: 10.1109/CVPR.2018.00068. indirizzo: http://openaccess.thecvf.com/content%5C_cvpr%5C_2018/html/Zhang%5C_The%5C_Unreasonable%5C_Effectiveness%5C_CVPR%5C_2018%5C_paper.html.

[5] N. G. of Art, *National Gallery of Art Open Data Program*, Accessed: 2024-01-29, gen. 2024. indirizzo: https://www.nga.gov/open-access-images/open-data.html.

[6] R. S. R. Silva, A. Lotfi, I. K. Ihianle, G. Shahtahmassebi e J. J. Bird, «ArtBrain: An Explainable end-to-end Toolkit for Classification and Attribution of AI-Generated Art and Style,» *CoRR*, vol. abs/2412.01512, pp. 1–20, 2024. DOI: 10.48550/ARXIV.2412.01512. arXiv: 2412.01512. indirizzo: https://doi.org/10.48550/arXiv.2412.01512.

[7] P. Liao, X. Li, X. Liu e K. Keutzer, «The ArtBench Dataset: Benchmarking Generative Models with Artworks,» *CoRR*, vol. abs/2206.11404, 2022. DOI: 10.48550/ARXIV.2206.11404. arXiv: 2206.11404. indirizzo: https://doi.org/10.48550/arXiv.2206.11404.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li e L. Fei-Fei, «ImageNet: A large-scale hierarchical image database,» in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[9] A. Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2021. arXiv: 2010.11929 [cs.CV]. indirizzo: https://arxiv.org/abs/2010.11929.

[10] P. Esser et al., «Scaling Rectified Flow Transformers for High-Resolution Image Synthesis,» *CoRR*, vol. abs/2403.03206, pp. 1–28, 2024. DOI: 10.48550/ARXIV.2403.03206.

[11] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu e M. Chen, «Hierarchical Text-Conditional Image Generation with CLIP Latents,» *CoRR*, vol. abs/2204.06125, pp. 1–27, 2022. DOI: 10.48550/ARXIV.2204.06125. indirizzo: https://doi.org/10.48550/arXiv.2204.06125.

[12] A. Q. Nichol et al., «GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models,» in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, PMLR, 2022, pp. 16 784–16 804. indirizzo: https://proceedings.mlr.press/v162/nichol22a.html.

[13] X. Liu et al., «More Control for Free! Image Synthesis with Semantic Diffusion Guidance,» in *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, IEEE, 2023, pp. 289–299. DOI: 10.1109/WACV56688.2023.00037. indirizzo: https://doi.org/10.1109/WACV56688.2023.00037.

[14] A. Taghipour et al., «Faster Image2Video Generation: A Closer Look at CLIP Image Embedding's Impact on Spatio-Temporal Cross-Attentions,» *CoRR*, vol. abs/2407.19205, pp. 1–11, 2024. DOI: 10.48550/ARXIV.2407.19205. arXiv: 2407.19205. indirizzo: https://doi.org/10.48550/arXiv.2407.19205.

[15] P. Wang, D. Li, X. Hu, Y. Wang e Y. Zhang, «CLIPMulti: Explore the performance of multimodal enhanced CLIP for zero-shot text classification,» *Comput. Speech Lang.*, vol. 90, p. 101 748, 2025. indirizzo: https://doi.org/10.1016/j.csl.2024.101748.

[16] S. Li, J. Cao, P. Ye, Y. Ding, C. Tu e T. Chen, «ClipSAM: CLIP and SAM collaboration for zero-shot anomaly segmentation,» *Neurocomputing*, vol. 618, p. 129 122, 2025. indirizzo: https://doi.org/10.1016/j.neucom.2024.129122.

[17] H. Yang, N. Wang, H. Li, L. Wang e Z. Wang, «Application of CLIP for efficient zero-shot learning,» *Multim. Syst.*, vol. 30, n. 4, p. 219, 2024. indirizzo: https://doi.org/10.1007/s00530-024-01414-9.

[18] V. Lytvyn, R. Peleshchak, I. Rishnyak, B. Kopach e Y. Gal, «Detection of Similarity Between Images Based on Contrastive Language-Image Pre-Training Neural Network,» in *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems. Volume I: Machine Learning Workshop, Lviv, Ukraine, April 12-13, 2024*, V. Lytvyn, A. Kowalska-Styczen e V. Vysotska, cur., ser. CEUR Workshop Proceedings, vol. 3664, CEUR-WS.org, 2024, pp. 94–104. indirizzo: https://ceur-ws.org/Vol-3664/paper8.pdf.

[19] F. Peng, X. Yang, L. Xiao, Y. Wang e C. Xu, «SgVA-CLIP: Semantic-Guided Visual Adapting of Vision-Language Models for Few-Shot Image Classification,» *IEEE Trans. Multim.*, vol. 26, pp. 3469–3480, 2024. DOI: 10.1109/TMM.2023.3311646. indirizzo: https://doi.org/10.1109/TMM.2023.3311646.

[20] Q. Zhou, C. Du, S. Wang e H. He, «CLIP-MUSED: CLIP-Guided Multi-Subject Visual Neural Information Semantic Decoding,» in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, ICLR, 2024, pp. 1–17. indirizzo: https://openreview.net/forum?id=lKxL5zkssv.

[21] Y. Surapaneni e C. Bhagvati, «Scene Text Image Super-Resolution with CLIP Prior Guidance,» in *Pattern Recognition - 27th International Conference, ICPR 2024, Kolkata, India, December 1-5, 2024, Proceedings, Part XXXII*, Elsevier, 2024, pp. 17–32. indirizzo: https://doi.org/10.1007/978-3-031-78125-4%5C_2.

[22] Z. Huang, A. Zhou, Z. Lin, M. Cai, H. Wang e Y. J. Lee, «A Sentence Speaks a Thousand Images: Domain Generalization through Distilling CLIP with Language Guidance,» in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, IEEE, 2023, pp. 11 651–11 661. indirizzo: https://doi.org/10.1109/ICCV51070.2023.01073.

[23] K. Alhamoud et al., *Vision-Language Models Do Not Understand Negation*, 2025. arXiv: 2501.09425 [cs.CV]. indirizzo: https://arxiv.org/abs/2501.09425.

[24] Z. Zhang, Z. Liu, M. Feng e C. Xu, *Can CLIP Count Stars? An Empirical Study on Quantity Bias in CLIP*, 2024. arXiv: 2409.15035 [cs.CV]. indirizzo: https://arxiv.org/abs/2409.15035.

[25] Y. Yamada, Y. Tang, Y. Zhang e I. Yildirim, *When are Lemons Purple? The Concept Association Bias of Vision-Language Models*, 2024. arXiv: 2212.12043 [cs.CV]. indirizzo: https://arxiv.org/abs/2212.12043.

[26] R. Abbasi, A. Nazari, A. Sefid, M. Banayeeanzade, M. H. Rohban e M. S. Baghshah, «CLIP Under the Microscope: A Fine-Grained Analysis of Multi-Object Representation,» in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, IEEE, 2025, pp. 9308–9317. indirizzo: https://openaccess.thecvf.com/content/CVPR2025/html/Abbasi%5C_CLIP%5C_Under%5C_the%5C_Microscope%5C_A%5C_Fine-Grained%5C_Analysis%5C_of%5C_Multi-Object%5C_Representation%5C_CVPR%5C_2025%5C_paper.html.

[27] W. Tu, W. Deng e T. Gedeon, «Toward a Holistic Evaluation of Robustness in CLIP Models,» *CoRR*, vol. abs/2410.01534, pp. 1–18, 2024. DOI: 10.48550/ARXIV.2410.01534. arXiv: 2410.01534. indirizzo: https://doi.org/10.48550/arXiv.2410.01534.

[28] V. D. Rosa, F. Guillaro, G. Poggi, D. Cozzolino e L. Verdoliva, «Exploring the Adversarial Robustness of CLIP for AI-generated Image Detection,» in *IEEE International Workshop on Information Forensics and Security, WIFS 2024, Rome, Italy, December 2-5, 2024*, IEEE, 2024, pp. 1–6. DOI: 10.1109/WIFS61860.2024.10810719. indirizzo: https://doi.org/10.1109/WIFS61860.2024.10810719.

[29] M. I. Ismithdeen, M. U. Khattak e S. Khan, *Promptception: How Sensitive Are Large Multimodal Models to Prompts?* 2025. arXiv: 2509.03986 [cs.CV]. indirizzo: https://arxiv.org/abs/2509.03986.

[30] V. Baherwani e J. J. Vincent, «Racial and Gender Stereotypes Encoded Into CLIP Representations,» in *The Second Tiny Papers Track at ICLR 2024, Tiny Papers @ ICLR 2024, Vienna, Austria, May 11, 2024*, OpenReview.net, 2024. indirizzo: https://openreview.net/forum?id=hQb6ts30wv.

[31] P. Gao et al., «CLIP-Adapter: Better Vision-Language Models with Feature Adapters,» *Int. J. Comput. Vis.*, vol. 132, n. 2, pp. 581–595, 2024. indirizzo: https://doi.org/10.1007/s11263-023-01891-x.

[32] T. Yu, Z. Lu, X. Jin, Z. Chen e X. Wang, «Task Residual for Tuning Vision-Language Models,» in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, Los Alamitos, CA, USA: IEEE, 2023, pp. 10 899–10 909. indirizzo: https://doi.org/10.1109/CVPR52729.2023.01049.

[33] K. Roth, Z. Akata, D. Damen, I. Balazevic e O. J. Hénaff, «Context-Aware Multimodal Pretraining,» *CoRR*, vol. abs/2411.15099, pp. 1–15, 2024. DOI: 10.48550/ARXIV.2411.15099. indirizzo: https://doi.org/10.48550/arXiv.2411.15099.

[34] K. Zhou, J. Yang, C. C. Loy e Z. L. Liu, «Learning to Prompt for Vision-Language Models,» *International Journal of Computer Vision*, vol. 130, pp. 2337–2348, 2022. indirizzo: https://link.springer.com/article/10.1007/s11263-022-01653-1.

[35] K. Zhou, J. Yang, C. C. Loy e Z. Liu, « Conditional Prompt Learning for Vision-Language Models,» in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, giu. 2022, pp. 16 795–16 804. indirizzo: https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01631.

[36] X. Zhu et al., «Not All Features Matter: Enhancing Few-shot CLIP with Adaptive Prior Refinement,» in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, Los Alamitos, CA, USA: IEEE Computer Society, 2023, pp. 2605–2615. indirizzo: https://doi.org/10.1109/ICCV51070.2023.00246.

[37] X. Zhu, B. Zhu, Y. Tan, S. Wang, Y. Hao e H. Zhang, «Selective Vision-Language Subspace Projection for Few-shot CLIP,» in *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, J. Cai et al., cur., ACM, 2024, pp. 3848–3857. DOI: 10.1145/3664647.3680885. indirizzo: https://doi.org/10.1145/3664647.3680885.

[38] H. Liu, C. Li, Q. Wu e Y. J. Lee, «Visual Instruction Tuning,» in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, NIPS Foundation, 2023, pp. 1–25. indirizzo: http://papers.nips.cc/paper%5C_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

[39] S. Bai et al., «Qwen2.5-VL Technical Report,» *CoRR*, vol. abs/2502.13923, pp. 1–23, 2025. indirizzo: https://doi.org/10.48550/arXiv.2502.13923.

[40] W. Dai et al., «InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning,» in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, NIPS Foundation, 2023, indirizzo: http://papers.nips.cc/paper%5C_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html.

[41] J. Li, D. Li, S. Savarese e S. C. H. Hoi, «BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models,» in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, PMLR, 2023, pp. 19 730–19 742. indirizzo: https://proceedings.mlr.press/v202/li23q.html.

[42] A. Mahendran e A. Vedaldi, «Understanding deep image representations by inverting them,» in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, IEEE Computer Society, 2015, pp. 5188–5196. DOI: 10.1109/CVPR.2015.7299155. indirizzo: https://doi.org/10.1109/CVPR.2015.7299155.

[43] A. Dosovitskiy e T. Brox, «Inverting Visual Representations with Convolutional Networks,» in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, IEEE Computer Society, 2016, pp. 4829–4837. indirizzo: https://doi.org/10.1109/CVPR.2016.522.

[44] H. Yin et al., «Dreaming to Distill: Data-Free Knowledge Transfer via DeepInversion,» in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, Computer Vision Foundation / IEEE, 2020, pp. 8712–8721. DOI: 10.1109/CVPR42600.2020.00874.

[45] H. Kazemi, A. M. Chegini, J. Geiping, S. Feizi e T. Goldstein, «What do we learn from inverting CLIP models?» *CoRR*, vol. abs/2403.02580, pp. 1–14, 2024. arXiv: 2403.02580. indirizzo: https://doi.org/10.48550/arXiv.2403.02580.

[46] R. S. Arora e A. Elgammal, «Towards automated classification of fine-art painting style: A comparative study,» in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012, pp. 3541–3544.

[47] J. Zujovic, L. Gandy, S. Friedman, B. Pardo e T. N. Pappas, «Classifying paintings by artistic genre: An analysis of features & classifiers,» in *2009 IEEE International Workshop on Multimedia Signal Processing*, 2009, pp. 1–5. DOI: 10.1109/MMSP.2009.5293271.

[48] C. Hentschel, T. P. Wiradarma e H. Sack, «Fine tuning CNNS with scarce training data — Adapting imagenet to art epoch classification,» in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 3693–3697. DOI: 10.1109/ICIP.2016.7533049.

[49] Y. Bin et al., «GalleryGPT: Analyzing Paintings with Large Multimodal Models,» in *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, ACM, 2024, pp. 7734–7743. indirizzo: https://doi.org/10.1145/3664647.3681656.

[50] P. Meng, Y. Liu, L. Zhang e X. Li, «Computational Aesthetics of Visual Artworks: Review and Outlook,» in *Cognitive Computing and Internet of Things. AHFE (2022) International Conference*, L. Paletta e H. Ayaz, cur., ser. AHFE Open Access, vol. 43, USA: AHFE International, 2022. DOI: 10.54941/ahfe1001833. indirizzo: https://doi.org/10.54941/ahfe1001833.

[51] J. Zhang, Y. Miao e J. Yu, «A Comprehensive Survey on Computational Aesthetic Evaluation of Visual Art Images: Metrics and Challenges,» *IEEE Access*, vol. 9, pp. 77 164–77 187, 2021. DOI: 10.1109/ACCESS.2021.3083075.

[52] X. Lu, Z. Lin, H. Jin, J. Yang e J. Z. Wang, «Rating Image Aesthetics Using Deep Learning,» *IEEE Transactions on Multimedia*, vol. 17, n. 11, pp. 2021–2034, 2015. DOI: 10.1109/TMM.2015.2477040.

[53] S. Hentschel, K. Kobs e A. Hotho, «CLIP knows image aesthetics,» *Frontiers in Artificial Intelligence*, vol. Volume 5 - 2022, 2022, TLDR: Comparing the usefulness of features extracted by CLIP compared to features obtained from the last layer of a comparable ImageNet classification model suggests that CLIP is better suited as a base model for IAA methods than ImageNet pretrained networks., ISSN: 2624-8212. DOI: 10.3389/frai.2022.976235. indirizzo: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2022.976235.

[54] A. Asperti, L. Dessı, M. C. Tonetti e N. Wu, «Does CLIP percevie art as we do?» In *Proceedings of the 22nd International Conference on Content-Based Multimedia Indexing (CBMI 2025)*, IEEE, 2025, to appear.