# POLITECNICO DI TORINO

**Laurea Magistrale in Ingegneria del Cinema e dei Mezzi di Comunicazione**



**Tesi di Laurea Magistrale**

# AI e VFX: ottimizzazione dei processi di compositing e post-visualizzazione

**Relatore**

**Prof. Andrea BOTTINO**

**Candidato**

**Federico PALUMBO**

**Aprile 2025**

**Abstract**

Negli ultimi anni, l'intelligenza artificiale e il machine learning hanno rivoluzionato numerosi settori, compresa l'industria degli effetti visivi. Tra le applicazioni più promettenti si distingue la segmentazione video, un processo fondamentale per la creazione di maschere e canali alpha: elementi essenziali per il compositing e la post-visualizzazione. La seguente tesi si propone di individuare soluzioni basate sull'intelligenza artificiale e il machine learning per affrontare questa sfida. Partendo da un'analisi approfondita dello stato dell'arte per la segmentazione video, la ricerca si è evoluta nell'individuazione di strumenti in grado di generare maschere con un basso livello di interazione umana e nella progettazione e l'implementazione di un sistema di segmentazione automatizzata capace di generare maschere utilizzabili nei processi di post-visualizzazione e compositing, con lo scopo di testare l'efficacia del sistema in scenari realistici di produzione VFX. Infine il sistema è stato validato qualitativamente e quantitativamente dimostrando una significativa riduzione del tempo di lavoro manuale, pur presentando alcune limitazioni dovute alla giovinezza di questi nuovi strumenti, che al contempo evidenziano il potenziale di crescita e innovazione in questo settore.

# Indice

# List of Figures

# Acronimi

**VFX**
Visual effects

**AR**
Augmented reality

**SFX**
Special effects

**CG**
Computer grafica

**AI**
Artificial intelligence

**ROI**
Ritorno di investimento

**LSTM**
Long short-term memory

**CNN**
Convolutional neural network

**RnD**
Research and development

**ML**
Machine learning

**FX**

Effects

**CGI**

Computer generated imagery

**NN**

Neural network

**FNN**

Feed-forward neural network

**RNN**

Recurrent neural network

**ViT**

Vision transformer

**GAN**

Generative adversarial network

**VAE**

Variational autoencoder

**DL**

Deep learning

**CV**

Computer vision

**SST**

Sparse spatiotemporal transformer

**COCO**

Common objects in context

**GPU**

Graphics processing unit

**CPU**

Central processing unit

**LLM**

Large language model

**SAM**

Segment anything model

**NLP**

Natural language processing

**CLIP**

Contrastive language-image pre-training

**MLP**

Multi layer perceptron

**PVS**

Promptable visual segmentation

**MAE**

Mask autoencoder

**FIFO**

First in first out

# Chapter 1

# Introduzione

> "The term visual effect is used to describe any imagery created, altered, or enhanced for a film, or other moving media, that cannot be accomplished during live-action shooting"[1]

Il termine effetti visivi (VFX) è usato per descrivere ogni immagine creata, alterata o migliorata per un film o un altro media che sarebbe altrimenti impossibile da realizzare durante le riprese. I VFX sono ormai presenti in moltissime produzioni audiovisive, dal cinema alla pubblicità, dai videogiochi alle applicazioni immersive come la realtà virtuale (VR) e aumentata (AR).
Occorre distinguere tra gli effetti visivi e gli effetti speciali (SFX). Questi ultimi vengono realizzati fisicamente durante le riprese tramite l'uso di scenografie, esplosivi o meccanismi pratici. I VFX, al contrario, entrano in gioco prevalentemente durante la post-produzione. Essi sono impiegati in diversi ambiti, tra cui per esempio la creazione di ambientazioni digitali come le città distopiche di *Blade Runner (2017)*, l'integrazione di creature come i dinosauri di *Jurassic Park (1993)* o gli alieni di *Avatar (2009)*, la simulazione di eventi catastrofici come le tempeste in *The Day After Tomorrow (2004)* o i disastri spaziali in *Gravity (2013)*.
Queste applicazioni evidenziano come i VFX siano uno strumento potente per espandere i confini della narrazione visiva, consentendo ai creatori di immergere il pubblico in mondi straordinari.

## 1.1 L'evoluzione degli effetti visivi

L'evoluzione dei VFX è strettamente legata al processo tecnologico. Negli anni del cinema muto (1895-1927), gli effetti venivano creati attraverso tecniche artigianali, come le doppie esposizioni, i modelli in scala e i trucchi prospettici.
George Méliès, considerato il padre degli effetti visivi, utilizzò queste tecniche

nel suo celebre *Voyage dans la lune (1902)*, un'opera pioneristica che introdusse concetti innovativi per l'epoca.

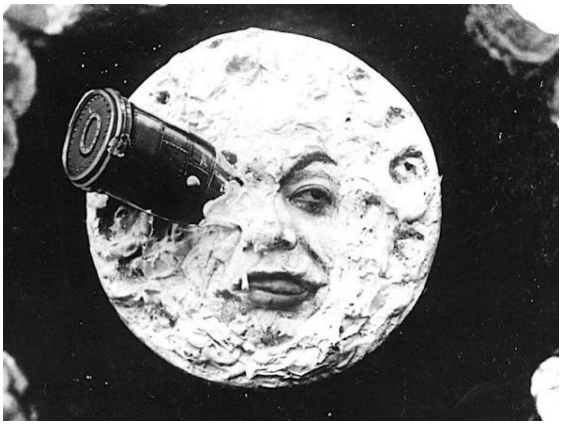

**Figure 1.1:** Voyage dans la lune (1902)
Fonte: `https://shorturl.at/URupm`

Con l'avvento del colore e del suono, l'industria cinematografica abbracciò nuove tecnologie come il matte painting e l'animazione stop-motion, rese celebri da opere come *King Kong (1933)*.
L'era digitale, iniziata negli anni '70, segnò una svolta radicale. Film come *Star Wars (1977)* introdussero l'uso di modellini e compositing ottico, mentre *Tron (1982)* fu uno dei primi a integrare immagini generate al computer (CGI).

**Figure 1.2:** Star Wars (1977)
Fonte: `https://shorturl.at/9jFGH`

Negli anni '90, la CGI divenne la protagonista indiscussa. Ne sono esempio film come *Jurassic Park (1993)* e *Terminator 2: Judgement Day (1991)*. Da allora, i VFX si sono evoluti ulteriormente, permettendo anche di creare interi mondi digitali, come quello visto in *Avatar (2009)*.

**Figure 1.3:** Avatar (2009)
Fonte: `https://shorturl.at/zWkFO`

## 1.2 Gli effetti visivi nell'era moderna

Negli ultimi anni i VFX hanno raggiunto un livello di realismo e complessità senza precedenti. I VFX sono diventati un elemento sempre più presente ma invisibile nella produzione cinematografica, televisiva e in numerosi altri settori come il gaming, la pubblicità e la formazione medica e ingegneristica.
La computer grafica (CG) è utilizzata non solo per generare creature o ambientazioni fantastiche ma anche per la creazione di effetti complessi, per esempio la simulazione di fenomeni naturali come fuoco, acqua e vento.
L'era moderna è sinonimo anche di motion capture e performance capture. Grazie a queste tecnologie è possibile registrare i movimenti di attori reali per trasferirli a personaggi digitali. Questa tecnologia innovativa è stata utilizzata in modo pioneristico per il personaggio di Gollum in *Lord of the Rings: The Fellowship of the Ring (2001)*. Grazie a questi strumenti è possibile catturare anche le micro-espressioni facciali, garantendo un maggiore realismo nelle performance digitali.
In aggiunta alla motion capture stanno prendendo sempre più campo la virtual production e il rendering in tempo reale. Queste tecnologie, inizialmente sviluppate nel settore dei videogiochi, stanno trovando sempre più applicazioni anche nel cinema e nella televisione. Strumenti come il motore grafico Unreal Engine permettono di creare scenari virtuali con un livello di dettaglio sorprendente. Un esempio emblematico è la serie *The Mandalorian (2019)* che utilizza la virtual production e la tecnologia StageCraft per generare ambienti virtuali foto-realistici in tempo reale migliorando il realismo, l'efficienza e la flessibilità sul set.

Queste nuove tecnologie hanno gettato le basi per l'incorporazione futura dell'AI nel settore del cinema e degli effetti visivi.

**Figure 1.4:** The Mandalorian (2019)
Fonte: `https://shorturl.at/I6xrq`

## 1.3 L'era dell'intelligenza artificiale

L'intelligenza artificiale si è oramai inserita in tutti i settori, compreso quello degli effetti visivi. Essa permette di automatizzare e velocizzare molti processi ed è particolarmente utile nel caso di task monotone, ripetitive o onerose in termini di tempo. Ma l'AI non solleva solamente dal carico di lavoro ripetitivo: è un ottimo strumento per affiancare l'artista e aiutarlo a generare idee creative o spunti di lavoro per migliorare il flusso creativo.
Film come *A.I. Artificial Intelligence (2001)* di Steven Spielberg hanno iniziato a trattare l'omonimo tema. Il film è ambientato in un mondo futuro afflitto dall'effetto serra e spaventato dalle innovazioni tecnologiche, dove gli esseri umani dividono ogni aspetto della loro vita con sofisticati robot da compagnia chiamati Mecca. Tratta temi legati all'AI, riflettendo un interesse crescente che va oltre la produzione cinematografica. Presenta uno spaccato su come l'AI potrebbe affrontare eventuali sentimenti ed emozioni, ricalcando le eterne domande esistenziali della nostra specie.
Nel 2004 esce *I, Robot*, film di Alex Proyas che tratta il tema dell'autonomia delle macchine e solleva domande sull'equilibrio della sicurezza umana. L'avvincente storia, con al centro il detective Del Spooner interpretato da Will Smith, mette in discussione il confine tra l'AI come strumento e l'AI come entità autonoma. Gli spunti di riflessione del film su come le decisioni etiche possano influenzare l'evoluzione della tecnologia risultano ancora oggi molto attuali. Le case di produzione stanno già sfruttando l'AI per analizzare le sceneggiature, scrutando variabili come trama,

personaggi e dialoghi. Questi strumenti non valutano solamente la qualità artistica, ma offrono la potente e al contempo pericolosa possibilità di valutare anche il potenziale successo commerciale di un film, fornendo un quadro completo che guida le decisioni di investimento. Ne è un esempio Cinelytic, uno strumento che permette di valutare la potenziale riuscita di un progetto audiovisivo, utilizzato dalla Warner Bros per il film *Logan (2017)*. Questo tool usa i dati esistenti dell'industria cinematografica per formulare previsioni sui rischi di un progetto futuro. Inoltre, permette di fornire delle previsioni riguardanti il casting, che lo strumento chiama analisi dei talenti. Questa funzione fornisce informazioni su popolarità e idoneità degli attori e delle attrici al progetto, fornendo informazioni sul loro punteggio box office, recensioni, età, interazioni sui social media e budget richiesto. In aggiunta, fornisce consigli riguardo la scelta del genere, i gusti del pubblico, la scelta del canale di distribuzione e previsioni sugli incassi, stime sui minimi e massimi al botteghino e il ritorno di investimento (ROI).
Un altro strumento di AI utilizzato è ScriptBook che, come suggerisce il nome, permette di analizzare le sceneggiature identificando il genere più adatto e l'analisi delle scene. Fornisce una valutazione delle scene in base al grado di emozionalità e studia i dialoghi per stabilire il mood complessivo dello script. Permette inoltre di analizzare la struttura dei personaggi, la loro presenza all'interno del film espressa in percentuale e le linee di battute che hanno. Fornisce anche l'interazione tra i due generi per valutare l'uguaglianza di genere.
Ma l'AI non si ferma alla semplice analisi. Nel 2016 la casa di produzione End Cue, con Allison Friedman and Andrew Swett, rilasciano *Sunspring*, un cortometraggio fantascientifico sperimentale interamente scritto da un'intelligenza artificiale usando la rete neurale chiamata long short-term memory (LSTM). Il corto narra la storia di tre persone, H, H2 e C, che vivono in un mondo futuristico tra omicidi e amore.

**Figure 1.5:** Sunspring (2016)
Fonte: `https://shorturl.at/4Hisc`

Sono molteplici i casi applicativi dell'intelligenza artificiale nel mondo del cinema, dalla pre-produzione, alla produzione e post-produzione, fino al doppiaggio. Un altro caso emblematico è l'uso del FaceDirector per controllare le espressioni facciali in *Avengers: Infinity War (2018)*. Questo strumento permette di avere la quasi completa libertà di agire sulle espressioni del volto in post-produzione, in modo tale da raggiungere con maggiore efficienza e flessibilità la visione creativa del regista.[2]


**Figure 1.6:** Face Director - Disney
Fonte: `https://shorturl.at/xqsoJ`

Rimanendo all'interno del tema dei volti modificabili in seguito alle riprese, è presente un ulteriore tool, il Deepfake, utilizzato per esempio nel film *The Irishman (2019)*. Con questa tecnologia è ora possibile sostituire i volti di controfigure o attuare la pratica nota come de-aging, ovvero il ringiovanimento degli attori o attrici. A tal proposito, nel film Here (2024) di Robert Zemeckis, è stata utilizzata questa tecnologia in real-time, affinché il regista potesse vedere in tempo reale sul set l'effetto sui volti degli attori. Lo sviluppo di queste tecnologie in real-time, come quella impiegata in questo film grazie a Metaphysic, apre le porte ad un maggior controllo e ad una maggiore consapevolezza durante le riprese del risultato finale.

**Figure 1.7:** Here (2024)
Fonte: `https://shorturl.at/iFF9u`

## 1.4 Contesto

Negli ultimi anni il panorama degli effetti visivi ha subito una trasformazione radicale, trainata dall'avvento dell'intelligenza artificiale e del machine learning. I VFX, che inizialmente servivano per superare i limiti della realtà nei film e nelle produzioni televisive, si sono evoluti fino a diventare una componente essenziale in una vasta gamma di applicazioni, dal cinema ai videogiochi, fino ai contenuti destinati alle piattaforme digitali e ai social media.
Questo settore ha vissuto un periodo di grande difficoltà durante la pandemia di COVID-19. La chiusura dei set cinematografici, il rallentamento delle produzioni e le restrizioni legate al lavoro in presenza hanno provocato un crollo significativo dell'industria dei VFX. Molte aziende, anche di grande rilievo, hanno subito gravi perdite economiche o sono state costrette a ridimensionare drasticamente le proprie attività. Tuttavia, le stesse difficoltà hanno spinto il settore a cercare nuove soluzioni tecnologiche per adattarsi a un contesto profondamente mutato.[3]
Durante il lockdown la produzione di film e serie TV è diminuita del 97,8%, circa 890.000 professionisti del settore hanno perso il lavoro. Regal Cinemas ha chiuso temporaneamente 663 sale (90% del suo fatturato) licenziando 40.000 dipendenti, Cinemark ha registrato un calo dei ricavi da 957,8 milioni di dollari a 9 milioni di dollari in un trimestre. Le entrate globali al botteghino nel 2020 sono state di 11,5 miliardi di dollari, contro i 42,5 miliardi di dollari del 2019 (decremento di oltre il 72%). Per quanto riguarda l'Italia nel 2020 si sono incassati 182.509.209€, per un numero di presenze in sala pari a 28.140.682. Rispetto al 2019 si è registrata una diminuzione degli incassi e delle presenze rispettivamente del 71,30% e del 71,18%. Tutto ciò ha inciso in maniera fortemente negativa sull'industria degli effetti visivi.[4] Solo a partire dal 2023 è iniziata una lenta e graduale ripresa del mondo del cinema, smorzata però dallo sciopero degli sceneggiatori e attori che ha posticipato o annullato molte produzioni impattando negativamente sugli studi di VFX. In Italia nel 2023 si sono incassati 495.692.418€ con un numero di presenze di 70.639.346, confermando ancora un calo del 16,3% degli incassi e del 23,2% delle presenze rispetto alla media del periodo 2017-2019. D'altro canto, i film prodotti in Italia nel 2023 sono stati 402 (+13% sul 2022), superando anche i livelli pre-pandemia (+23,7% sul 2019).[5]
Tra le innovazioni che hanno trovato terreno fertile in questo periodo, spiccano le tecnologie di produzione virtuale e l'uso crescente dell'intelligenza artificiale per automatizzare e migliorare i processi creativi. Tuttavia, è importante sottolineare che l'AI non sta sostituendo il lavoro degli artisti, ma piuttosto lo sta arricchendo. L'intelligenza artificiale si configura come uno strumento collaborativo, capace di amplificare le capacità umane e di semplificare attività ripetitive o particolarmente onerose. Ad esempio, mentre un algoritmo può velocizzare il processo di rotoscoping o di tracking, l'intervento umano rimane indispensabile per definire il contesto

creativo, prendere decisioni estetiche e garantire che il risultato finale rispetti una visione artistica coerente.
L'AI, quindi, non è un sostituto dell'intuizione, dell'esperienza o della sensibilità artistica, ma un supporto per liberare l'artista da compiti meccanici e consentirgli di concentrarsi sugli aspetti più creativi e innovativi del progetto. Grazie a queste tecnologie, gli artisti possono esplorare un numero maggiore di opzioni creative e raggiungere risultati visivamente sorprendenti in tempi più rapidi. Durante il periodo della pandemia l'adozione di questi strumenti ha subito un'accelerazione, spinta dalla necessità di adattarsi a condizioni di lavoro remoto e a produzioni che richiedevano una maggiore flessibilità. La capacità dell'AI di gestire grandi volumi di dati e di elaborare sequenze video complesse si è dimostrata fondamentale per mantenere alti standard qualitativi, anche in un contesto di risorse limitate e tempi ridotti. Un esempio emblematico è l'utilizzo di reti neurali convoluzionali (CNN) e dei modelli generativi per analizzare e processare sequenze video. Questi algoritmi possono non solo automatizzare fasi complesse del processo creativo, ma anche proporre soluzioni innovative che possono ispirare nuove idee agli artisti. Si tratta di un dialogo continuo tra umano e macchina, in cui l'artista guida il processo e l'intelligenza artificiale fornisce strumenti sempre più sofisticati per materializzare una visione.[6]
In definitiva, il mondo dei VFX si trova a un punto di svolta. Dopo le sfide senza precedenti affrontate durante la pandemia, il settore ha trovato nell'innovazione tecnologica una strada per reinventarsi. L'intelligenza artificiale e il machine learning non rappresentano soltanto un processo tecnico, ma una trasformazione culturale del modo in cui l'arte digitale viene concepita e realizzata. La collaborazione tra umano e macchina è oggi al centro di questa rivoluzione, aprendo nuove possibilità creative e immaginando un futuro in cui la sinergia tra sensibilità artistica e potenza computazionale ridefinisce i confini del possibile.
La seguente tesi si basa sul lavoro svolto durante quattro mesi di stage curriculare presso EDI Effetti Digitali Italiani, azienda leader in Italia nella realizzazione di effetti visivi per TV, cinema, serie e pubblicità, nonché casa di produzione cinematografica, attiva sia in progetti italiani che internazionali. Recentemente è stato istituito un nuovo reparto di Ricerca e Sviluppo (RnD), al quale ho avuto l'opportunità di appartenere per tutta la durata del mio stage. Tale iniziativa rappresenta un chiaro segno dell'impegno dell'azienda verso l'innovazione e la ricerca, evidenziando la sua volontà di investire in ambiti strategici per il continuo progresso e la competitività nel settore.

## 1.5 Obiettivi

L'intelligenza artificiale è in continua e rapida evoluzione e sono molteplici i casi applicativi nell'industria degli effetti visivi già presenti. La seguente tesi si propone di esplorare soluzioni basate sull'AI e il ML per la segmentazione di sequenze video. In particolare l'obiettivo è quello di individuare, sviluppare e testare un sistema semi-automatizzato in grado di generare maschere utilizzabili nei processi di compositing e post-visualizzazione, a partire dall'analisi dello stato dell'arte. Questo lavoro si pone come obiettivo quello di ridurre il tempo di lavoro manuale, migliorare l'efficienza produttiva e valutare l'implementazione di questo sistema all'interno della pipeline di produzione di un'azienda di effetti visivi. Il lavoro si è inizialmente concentrato sullo studio dello stato dell'arte relativo alle tecnologie e agli strumenti attualmente utilizzati per la segmentazione video. Questo studio ha permesso di identificare fin da subito lo stato attuale, i limiti e le sfide che questi nuovi strumenti mettono a disposizione. Sulla base di queste osservazioni sono stati eseguiti dei test qualitativi e quantitativi, sui quali è stato svolto un confronto con i capi reparto di 2D ed effects (FX). In seguito è stato implementato un sistema in grado di generare maschere e canali alfa con una segmentazione semantica, utile a migliorare l'efficienza e la rapidità durante la post-visualizzazione.

# Chapter 2

# Stato dell'arte

In questo capitolo si offre una panoramica sullo stato dell'arte per la generazione di maschere (matte), sulle tecniche tradizionali che vengono sfruttate e sulla motivazione per cui queste maschere sono utili e, in taluni shots, essenziali. Si affrontano i canali colore, in particolare il canale alfa, essenziale e necessario per poter ottenere una matte, e si accenna la regola della pre-moltiplicazione, l'operazione matematica necessaria al fine di visualizzare il personaggio o l'oggetto scontornato. Inoltre viene presentata la pipeline tradizionale per la realizzazione degli effetti visivi, con particolare attenzione alla pre-visualizzazione, al rotoscoping e al compositing. Infine, vengono trattati la segmentazione video, il suo impiego all'interno del processo e i vari tipi di segmentazione.

## 2.1 Immagine digitale

Un'immagine digitale è la rappresentazione numerica di un'immagine bidimensionale. La rappresentazione può essere di tipo vettoriale oppure raster (detta anche bitmap). Nel primo caso sono descritti gli elementi primitivi, come linee e poligoni, che vanno a comporre l'immagine. Nel secondo caso, invece, l'immagine è composta da una matrice di punti, chiamati pixel, la cui colorazione è codificata tramite uno o più valori numerici, i bit.
All'interno dei dati dei pixel è solitamente presente un header, che contiene diverse informazioni sull'immagine a partire dal numero di righe e colonne di pixel. Queste informazioni sono necessarie per poter disporre la sequenza di pixel in linee, in modo tale da formare una griglia rettangolare di punti. La tecnica per rappresentare l'immagine in una matrice NxM (dove N è il numero delle righe e M delle colonne) è detta raster.
Le immagini bitmap possono essere memorizzate in diversi formati, spesso basati su un algoritmo di compressione, che può essere lossy (con perdita d'informazione),

come nelle immagini JPEG, oppure lossless (senza perdita d'informazione), come nel caso delle immagini GIF o PNG. Queste immagini possono essere generate da una grande varietà di dispositivi di acquisizione, come scanner e fotocamere digitali, le quali contengono sensori CCD o CMOS. Il campo dell'elaborazione digitale delle immagini studia gli algoritmi per modificare tali immagini.

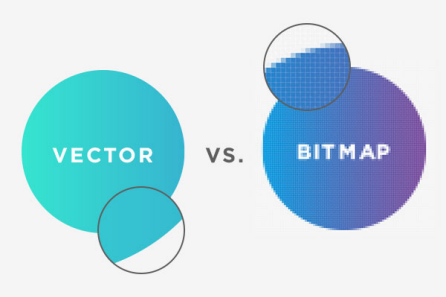


**Figure 2.1:** Vector vs Bitmap
Fonte: `https://shorturl.at/5pguf`

### 2.1.1 Canali colore

Nelle immagini a colori viene memorizzato il livello di intensità dei colori fondamentali. Nel modello di colore RGB per esempio vengono memorizzati il rosso, il verde e il blu. Nel modello CMYK, invece, vengono memorizzati ciano, magenta, giallo e nero (questo modello viene usato nella stampa). Nelle immagini monocromatiche, ovvero le immagini in scala di grigio, il valore indica l'intensità di grigio che varia dal nero al bianco.
Il numero di colori o di livelli di grigio possibili è chiamato profondità e dipende dal massimo numero di combinazioni permesse dalla quantità di bit utilizzata per questi dati. Ad esempio, un'immagine con 1 bit per pixel avrà al massimo due combinazioni possibili (0 e 1) e quindi potrà rappresentare solo due colori (bianco o nero). Un'immagine a 8 bit per pixel potrà rappresentare fino a 256 colori.

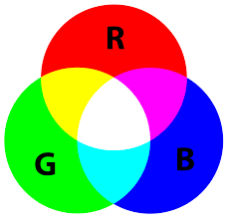


**Figure 2.2:** Canali colore
Fonte: `https://shorturl.at/7xeZV`

### 2.1.2 Canale alfa

Il canale alfa (alpha channel) è il canale che porta con sè l'informazione di trasparenza. È un'immagine in scala di grigi nella quale l'immagine in primo piano (foreground) è una silhouette. La silhouette può essere immaginata come nero su bianco o con i suoi valori invertiti, bianco su nero. Il canale alfa rappresenta la differenza in colore tra il colore che sta dietro al soggetto e i colori del soggetto in foreground. Il livello numerico della matte in ogni pixel è proporzionale alla visibilità del background. I valori dei pixel del canale alfa vanno da 0 a 1, dove lo zero indica che il canale alfa è nullo e l'uno che il canale alfa è massimo. I valori compresi tra questi due limiti sono chiamati pixel semi-trasparenti.

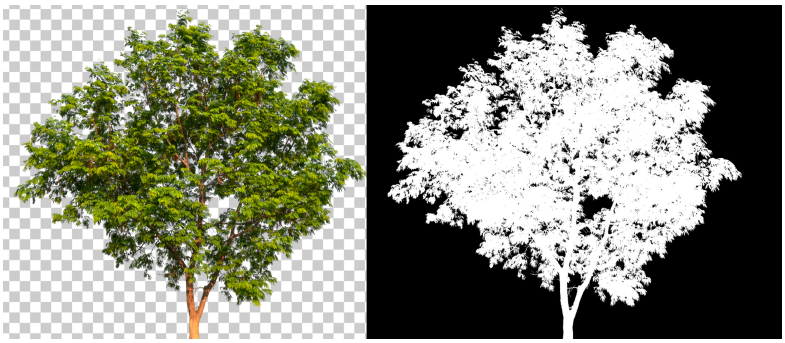

**Figure 2.3:** Canale alfa
Fonte: `https://shorturl.at/QYL16`

### 2.1.3 Premoltiplicazione

L'operazione matematica che permette di isolare un'area specifica dell'immagine grazie al suo canale alfa è detta premoltiplicazione (premult). Questa operazione consiste nel moltiplicare per ogni pixel dell'immagine il suo valore di alfa per tutti i valori dei canali colore.

$$R' = R \cdot \alpha, \quad G' = G \cdot \alpha, \quad B' = B \cdot \alpha$$

Ad esempio, un pixel che avrà valori RGBA pari a 255,0,0,1 rimarrà tale. Un pixel con valori 0,0,255,0 diventerà 0,0,0,0. Un pixel con valori 0,255,0,0.5 diventerà 0,127.5,0,0.5.
Vige una regola molto importante nella premoltiplicazione.[7] Qualora durante il compositing si dovesse applicare della color ad un elemento (ad esempio con un nodo di grade), questo non deve essere premoltiplicato. Non seguire correttamente questo ordine di operazioni potrebbe causare la comparsa di un bordino nero attorno all'elemento. Gli elementi CG quando vengono renderizzati, per loro natura, sono

già premoltiplicati. Ciò significa che, se in compositing è richiesta l'integrazione di un elemento realizzato in CG, prima di fare un matching dei colori è necessario svolgere l'operazione inversa alla premoltiplicazione, chiamata unpremult.

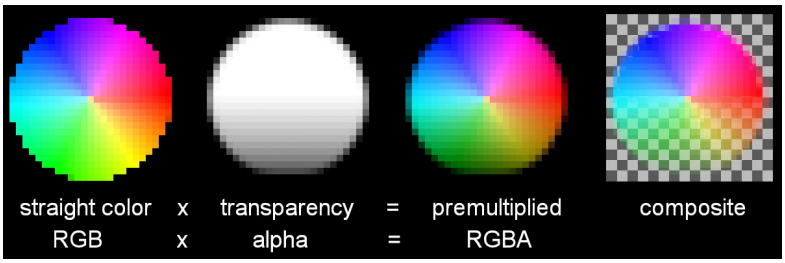


**Figure 2.4:** Premoltiplicazione
Fonte: `https://shorturl.at/tzLDB`

### 2.1.4 Maschera alfa

Le maschere consentono di isolare e manipolare aree specifiche di un'immagine o di un video. Permettono di attuare interventi mirati a una zona specifica di un'inquadratura. Una maschera è essenzialmente un filtro che separa una parte visiva dal resto della scena, permettendo modifiche selettive senza influire sul contenuto circostante. Oppure consentono di interporre un elemento tra l'area mascherata e lo sfondo. Ad esempio, è spesso utilizzata per rimuovere sfondi, integrare elementi in una scena o regolare i colori in maniera puntuale.

## 2.2 VFX pipeline

La pipeline degli effetti visivi si riferisce a tutte le fasi in cui i VFX sono richiesti in un film o in una serie TV. La pipeline serve ad organizzare ogni reparto in modo da definire i ruoli dei vari artisti che intervengono durante la lavorazione e stabilire il timing di lavorazione. Realizzare un effetto visivo richiede lavoro di squadra e richiede che tutti gli artisti siano allineati alla visione del regista e del VFX Supervisor del progetto, al fine di raggiungere insieme un unico obiettivo comune.
La pipeline vede una suddivisione condivisa con la suddivisione delle fasi di lavorazione di un film: pre-produzione, produzione e post-produzione. Nella pre-produzione viene sviluppata l'idea di un film, viene scritta la sceneggiatura, stipulati i budget e stabilite le timelines. Vengono poi definiti il cast e le location e vengono ottenuti i finanziamenti per produrre il film. Nella produzione vengono svolte le riprese, se richiesta viene impiegata la motion capture, la virtual production e gli

effetti speciali pratici. Durante la post-produzione vengono applicati gli effetti visivi al footage live-action, il sonoro e vengono fatti il montaggio e la color. Dopodiché il film viene distribuito.[8]

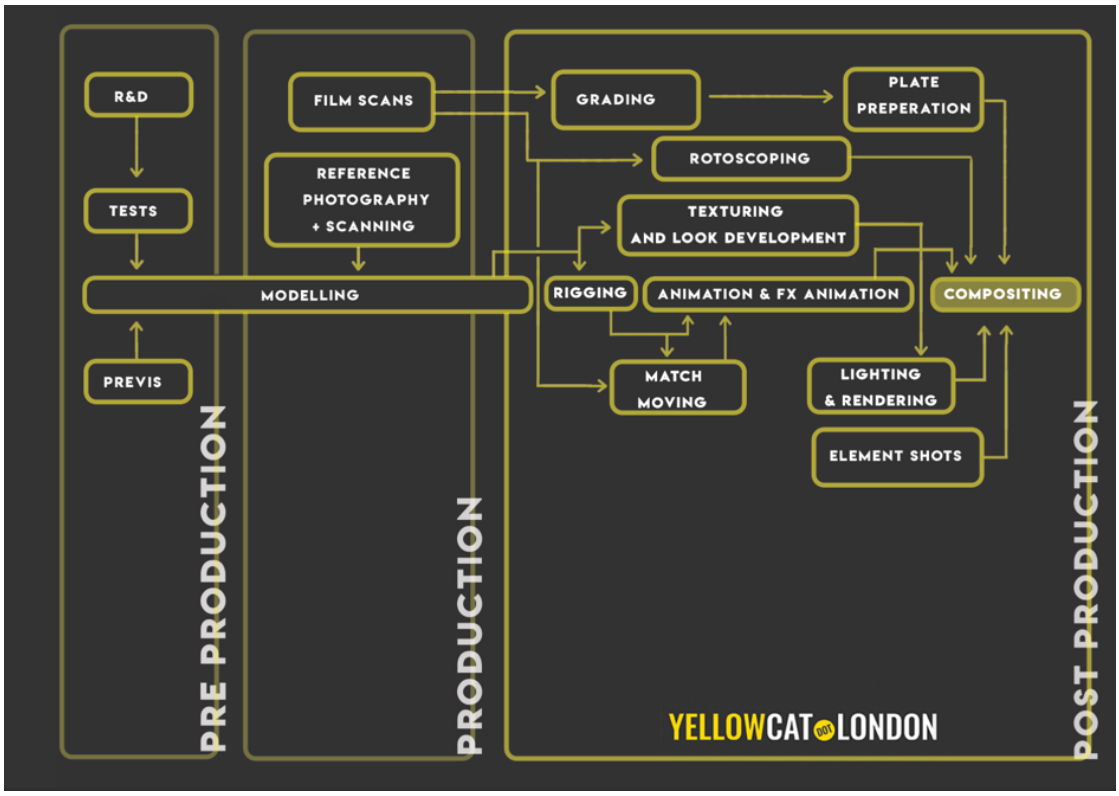


**Figure 2.5:** VFX pipeline
Fonte: `https://shorturl.at/hjkBf`

Si potrebbe racchiudere i reparti che intervengono durante l'intero processo di realizzazione degli effetti visivi nelle seguenti macro-categorie: Storyboard e animatic, previs, concept art e design, matchmove e camera tracking, layout e production design, modeling e creazione degli asset, ricerca e sviluppo, rigging, animazione, FX e simulazioni, lighting, look-dev e rendering, compositing.
Con lo storyboard e l'animatic gli artisti creano rappresentazioni visive delle azioni descritte nella sceneggiatura. Vengono poi realizzate la previsualizzazione (descritta nei paragrafi successivi) e illustrazioni che rappresentino al meglio l'idea, la visione e lo stile creativo che il regista ha in mente per il film o la serie. Nel momento in cui vengono svolte le riprese e si ha a disposizione il girato, viene svolta l'operazione di tracking e matchmove in cui viene analizzato lo shot e viene svolto il "solve" della camera, ovvero vengono tracciati i movimenti degli oggetti fermi all'interno dell'inquadratura per ottenere una nuvola di punti in uno spazio tridimensionale. Si

ottiene una prima rappresentazione 3D della scena che si è ripresa sul set. Questa è utile per posizionare in seguito elementi in CG, character, simulazioni, in modo tale che, una volta inseriti, rimangano ancorati alla scena creando l'illusione che siano realmente presenti nell'ambiente ripreso sul set. Con l'operazione di "layout" si posizionano i primi elementi nella scena 3D, compresa la camera virtuale. Inizialmente questi elementi saranno semplici forme tridimensionali, come cubi, sfere o parallelepipedi. Viene difatti svolta l'operazione chiamata "blocking", ovvero il posizionamento all'interno della scena 3D di forme geometriche semplici, che verranno poi sostituite dai modelli 3D e dagli asset finali. Questi modelli sono modellati su software di modellazione 3D come Maya o Blender e possono raggiungere anche milioni di poligoni. Durante queste fasi interviene anche il reparto di ricerca e sviluppo (R&D), un team composto da persone che hanno competenze trasversali, dalla programmazione alla scienza, fino alla matematica. Il loro lavoro è rendere fattibile qualcosa che prima non lo era o era molto complicato da realizzare. Il team di ricerca e sviluppo è a supporto degli altri reparti, fornisce tool, plugins, codice per trasformare in realtà le idee. Questo reparto lavora dalle primissime fasi di pre-produzione fino alla post-produzione. Il reparto di rigging, invece, si occupa di creare gli scheletri e le armature (rig) per i character o gli oggetti da animare in una sequenza. Senza questi rig gli animatori non avrebbero alcun tipo di controllo per i movimenti e le deformazioni. Viene creata una struttura con le articolazioni e giunti (joints) che poi viene animata. Qualora lo shot richieda l'inserimento di particelle, fluidi, fiamme o una dinamica di corpi rigidi interviene il team di FX. Vengono create le simulazioni che verranno poi inserite nelle sequenze video. Il reparto di lighting si occuperà di illuminare la scena, i look-dev artists di supervisionare il look complessivo del lavoro e il reparto di rendering, come dice la parola stessa, di renderizzare gli shots. Il rendering è il processo di generazione di un'immagine a partire da una descrizione matematica di una scena tridimensionale, interpretata da algoritmi che definiscono il colore di ogni pixel. Infine, con il compositing vengono assemblati i contributi realizzati da ogni reparto con il footage live-action. Questo passaggio verrà approfondito nei paragrafi seguenti.

### 2.2.1 Rotoscoping

Le prime tecniche per la creazione di maschere risalgono all'epoca analogica. Venivano tracciati manualmente i contorni di un soggetto, fotogramma per fotogramma. Sebbene questo approccio fosse accurato, era anche altrettanto dispendioso in termini di tempo. Questa tecnica era chiamata rotoscopio (dall'inglese rotoscope), ideata da Max Fleischer. Il disegnatore ricalcava le scene a partire da una pellicola filmata in precedenza.
Questa tecnica fu impiegata da Fleischer per la prima volta nella serie *Out of the Inkwell (1918-1929)* assieme a suo fratello Dave che, nelle vesti di un clown,

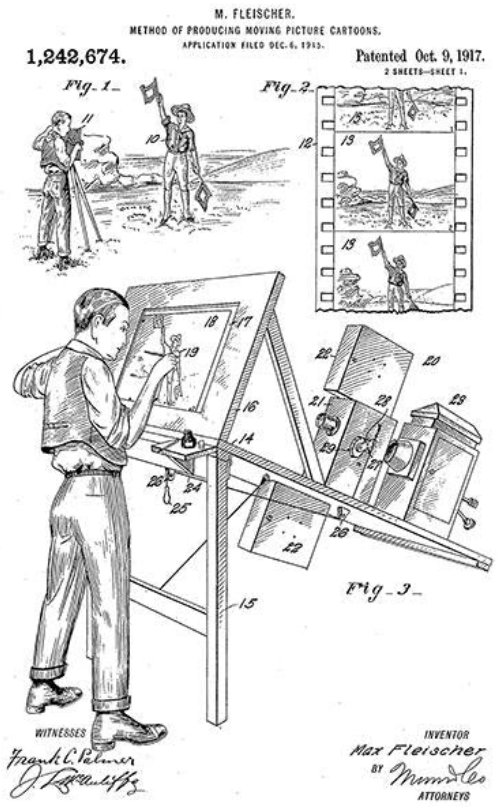


**Figure 2.6:** Rotoscopio
Fonte: `https://shorturl.at/e67dH`

interpretava il noto personaggio Koko il Clown, per poi essere utilizzata in numerosi cartoni Disney tra cui *Biancaneve e i sette nani (1937)*. Il rotoscopio può causare piccole imperfezioni rispetto alle sagome originali che cambiano nel tempo: i singoli tratti, quando disegnati e animati attraverso di esso, sembrano vibrare innaturalmente. Questo effetto è stato volontariamente ricercato durante la realizzazione del videoclip degli a-ha *Take on Me (1985)*.

Con il compositing digitale questa tecnica è stata rivoluzionata da software come Nuke, Fusion, Mocha e Silhouette.
Prima di parlare di rotoscoping è doveroso accennare le definizioni di alcuni elementi che verranno nominati durante questo processo: inquadratura, fotogramma, interpolazione, shapes, tracking, motion blur, occlusione. Un film è composto da scene, che sono composte da inquadrature, a loro volta composte da fotogrammi. Un'inquadratura (shot) è la porzione di spazio fisico inquadrata dalla macchina da presa o fotocamera. La durata di un'inquadratura è la durata che intercorre tra l'inizio di quella inquadratura e lo stacco. Un fotogramma (frame) è, invece,

**Figure 2.7:** Take On Me - a-ha (1985)
Fonte: `https://shorturl.at/7YxRa`

la parte più piccola di un film. In matematica l'interpolazione è un metodo per determinare i valori di una funzione all'interno di un intervallo nel quale ne sono noti solo alcuni. Una shape è una forma geometrica che viene disegnata dall'artista, con punti e linee, che definisce l'area che si vuole scontornare. Nel momento in cui si deve scontornare una persona, per esempio, non si disegnerà una shape unica per tutto il corpo bensì tante piccole shapes in corrispondenza dei giunti/snodi del corpo umano (gomito, avambraccio, polso, ecc...).

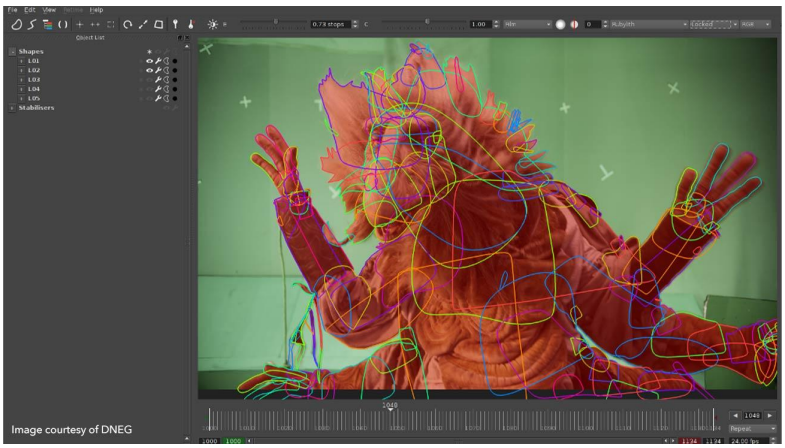


**Figure 2.8:** Roto shapes
Fonte: `https://shorturl.at/PKybJ`

Il tracking è l'operazione che permette di tracciare il movimento di un singolo punto, di più punti o di una superficie. Il motion blur è l'effetto di scia, mosso, sfocato che si ottiene quando un elemento, la camera o entrambi sono in movimento durante l'esposizione.
Un altro elemento che introduce ulteriore difficoltà, aumentando il tempo di lavoro, è l'occlusione. Questa si verifica quando, durante la scena, un elemento passa davanti a quello che si sta scontornando. In questo caso è necessario modificare la maschera in modo tale che non vada a sovrapporsi all'elemento occludente. Può rendersi necessaria la creazione di una maschera di esclusione o di ritaglio dell'oggetto occludente (holdout mask) che verrà sottratta alla maschera dell'elemento occluso.
In Nuke il rotoscoping basa le sue fondamenta sull'uso di due nodi: Roto e RotoPaint.

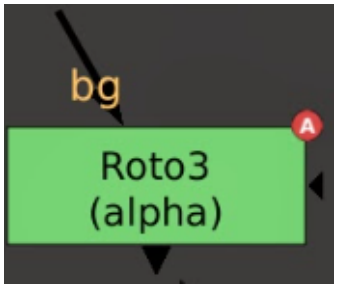


**Figure 2.9:** Nodo Roto

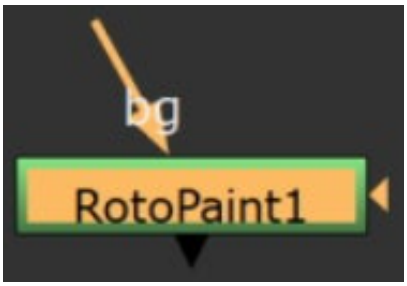


**Figure 2.10:** Nodo RotoPaint

Il processo inizia disegnando le forme (shapes) con il nodo Roto sui fotogrammi chiave e prosegue con un'interpolazione tra di essi. Queste maschere possono essere formate da forme circolari, quadrate, bezier o spline. Se il soggetto da scontornare è in movimento è necessario fare un tracking della maschera, che può essere fatto manualmente o con l'impiego di un tracker spostando la maschera in modo che in ogni fotogramma essa si trovi nella posizione corretta. Più il soggetto è in movimento più sarà complicato realizzare una buona maschera; questo è dovuto al fatto che maggiore sarà il movimento, maggiore sarà il motion blur presente.
Il nodo di RotoPaint, invece, è un nodo basato su vettori (vector-based) che viene utilizzato per il rotoscoping, la rimozione dei rig (rig removal) e altre tasks legate al compositing. Si possono disegnare Bezier e B-Spline con attributi individuali e gruppi di livello, includendo sfumature, motion blur, modalità di blending e trasformazioni 2D individuali o gerarchiche. Molti controlli sono condivisi con il nodo di Roto ma il RotoPaint presenta molti più strumenti. Questo nodo ha anche la possibilità di utilizzare il pennello (brush) per clonare aree dell'immagine.
Il rotoscoping può essere fatto anche sui software Fusion, Mocha e Silhouette. Su Fusion bisogna importare il video nella MediaIn, aggiungere il nodo Polygon o BSpline e tracciare il contorno dell'oggetto che si desidera scontornare, utilizzando

i punti di controllo. Assicurandosi che ci sia l'impostazione Auto-Key abilitata, è necessario spostarsi nel tempo attraverso la timeline per modificare ed adattare la forma della maschera all'oggetto in movimento. Se quest'ultimo è particolarmente complesso è possibile aiutarsi con l'impiego del tracker per tracciarne il movimento. Con il parametro Soft Edge e altre tecniche è poi possibile sfumare i bordi per renderli meno netti. Esiste un'ulteriore funzionalità: il tracking planare. Essa permette di tracciare i movimenti di un piano. Tra le caratteristiche del software Silhouette spiccano alcuni strumenti di intelligenza artificiale e machine learning realizzati ad hoc per migliorare il rotoscoping.

### 2.2.2 Pre-visualizzazione e post-visualizzazione

La pre-visualizzazione (pre-visualization, abbreviata previs) è il passo successivo alla realizzazione di uno storyboard. L'immagine in movimento che fornisce la previs, rispetto alla staticità di uno storyboard, permette al regista di avere un'idea più chiara di quello che potrà essere il risultato finale. Permette di progettare non solo la composizione visiva dell'immagine ma anche il timing e il movimento di camera. Lo storyboard e la previs sono utilizzati anche per mantenere fissi e coerenti il design, in modo da avere consistenza nelle fasi di progettazione future. La previs può contenere storyboard animati, shot, elementi in CG, ambienti virtuali. Al fine di pianificare al meglio l'impiego di risorse e di budget per la realizzazione di uno shot, la previs permette di sperimentare ed esplorare possibili soluzioni tracciando la strada per il compimento dello stesso.

> "La previs è un processo collaborativo che genera versioni preliminari di inquadrature o sequenze, prevalentemente utilizzando strumenti di animazione 3D e ambienti virtuali. Permette ai filmmakers di esplorare visivamente le idee creative, pianificare soluzioni tecniche e comunicare una visione condivisa per una produzione efficiente"[1]

Esistono vari tipi di previs, differenziati in base al focus della loro utilità: pitchvis, technical previs, on-set previs, postvis e D-vis. La pitchvis illustra il potenziale di un progetto prima che sia stato interamente finanziato. La technical previs incorpora anche la camera, le luci e il layout della scena per aiutare a definire i requisiti produttivi. L'on-set previs, come suggerisce il termine, racchiude tutte le visualizzazioni fatte sulla location per aiutare il regista, il VFX Supervisor e la crew a valutare la resa di alcuni effetti direttamente sul set e spesso in real-time. La postvis combina elementi digitali e il girato live-action per definire con maggiore precisione gli effetti visivi da sviluppare e i loro costi. La D-vis (design visualization), infine, utilizza un framework virtuale in cui è possibile testare i requisiti di produzione e fare location scouting.
Negli anni i film sono stati pre-visualizzati usando una grande varietà di tecniche

tra cui gli storyboards, test live-action, fotografie e concept. Non è strettamente necessario utilizzare storyboard o previs nella pre-produzione. Sebbene alcuni registi come Alfred Hitchcock o Orson Welles fossero rinomati per la meticolosa pianificazione dei loro film, altri registi si sono affidati a una maggiore improvvisazione. Questi due strumenti però sono divenuti presto una modalità standard per evidenziare fin da subito eventuali difficoltà del progetto. I montatori e gli artisti dei VFX usano spesso materiale pre-esistente per rifinire gli effetti visivi quando sono ancora nella fase di progettazione. Un esempio di tecnica utilizzata in passato è il "rip-o-matic", ovvero un montaggio di materiale proveniente da film, riprese o video animati pre-esistenti. L'esempio più famoso è il rip-o-matic di *Star Wars (1977)* in cui George Lucas usò footage della battaglia aerea della Seconda Guerra Mondiale. Negli anni '80 la CGI iniziò ad essere utilizzata anche per la previs. Per *Tron (1982)* Bill Kroyer e Jerry Reese usarono una combinazione di storyboards disegnati e linee vettoriali fatte in computer grafica per alcune scene come l'inseguimento tra moto. Un altro impiego fu nel film *The Boy Who Could Fly (1986)*. Il production designer Jim Bissell, per rappresentare la sequenza in cui due bambini volano sopra una festa scolastica, fece creare una rappresentazione virtuale della scena; essa gli permise di vedere dove potesse avere una migliore gestione del set. La previs si evolse e cominciò ad essere uno strumento molto utile per fornire qualche informazione tecnica alla crew come per esempio i dati sulla camera, i rig e il movimento di camera. Ne è un esempio *Panic Room (2002)* in cui David Fincher ebbe la possibilità di sperimentare cranes virtuali.
La postvis, a differenza della previs in cui si ha un approccio molto più orientato ad un maggiore utilizzo della CG, è molto utile per visualizzare le scene successivamente alle riprese usando, ad esempio, una combinazione tra il girato (plate), miniature e CG. Contrariamente alla previs in cui si ha molta più libertà di sperimentare ed esplorare nuove idee creative, nuovi effetti e movimenti di camera, nella postvis si è fortemente limitati al plate e, di conseguenza, al movimento di camera già presente.
L'uso della postvis è diventato parte essenziale della realizzazione di un film. Quando si ha a che fare con un personaggio in CG o un effetto cruciale per lo storytelling, può risultare complicato girare lo shot senza essere al corrente dell'ingombro o dell'interazione con l'ambiente circostante. L'impiego della postvis permette di vedere come gli elementi girati in live-action interagiscono e lavorano con gli elementi inseriti in post-produzione e fornisce uno specchietto sul risultato finale.

Il lavoro di rotoscoping, il più delle volte, è necessario durante la postvis. Per integrare in background elementi in CG, ambienti virtuali o matte paintings è necessario che gli attori in foreground siano su di un layer separato rispetto al resto della scena. Dunque è necessario avere il canale alpha dei soggetti in modo tale da poterli posizionare davanti agli elementi aggiunti in digitale. Per avere il canale alpha, come detto in precedenza, bisogna ricorrere al rotoscoping o alla

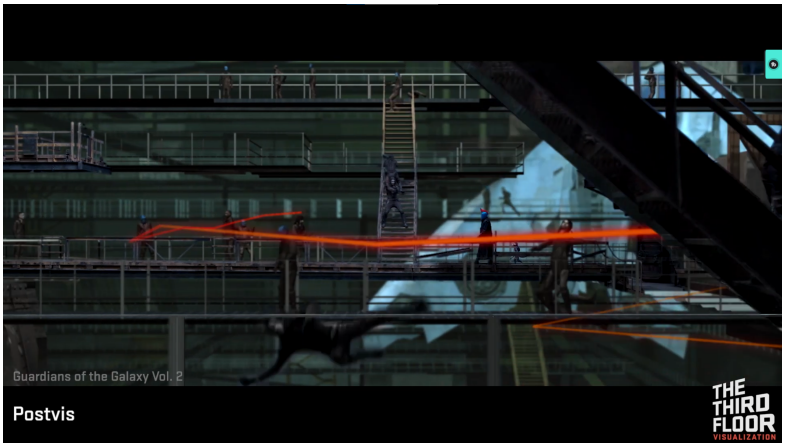


**Figure 2.11:** Postvis
Fonte: `https://youtu.be/KCORIBoj8NY`

segmentazione video, che verrà trattata in seguito. Le maschere dei soggetti, in questa fase, possono anche essere approssimative e poco precise. Durante la postvis è importante riuscire a raggiungere un buon risultato in poco tempo, in modo da poter mostrare al cliente o al regista i progressi di lavorazione per ottenere dei feedback sulla potenziale resa degli effetti visivi finali.
Nell'era dell'intelligenza artificiale sono molti gli strumenti che permettono di velocizzare questo processo. Nei paragrafi e capitoli successivi verranno mostrati metodi e tecniche con cui segmentare sequenze video attraverso un sistema quasi interamente automatizzato, al fine di ottimizzare il processo di post-visualizzazione dal punto di vista della generazione di maschere per questa importantissima fase di lavorazione.

### 2.2.3 Compositing

Il compositing è un elemento del processo di post-produzione che è ormai presente nella maggior parte dei film, serie TV e pubblicità. Le attività di compositing consistono nel combinare in un'unica immagine diversi elementi visivi provenienti da diverse fonti, come ad esempio rimuovere il blue screen dietro a un attore e rimpiazzarlo con uno sfondo generato digitalmente. La parte difficile è fare in modo che questi elementi, diversi tra loro, risultino parte dello stesso ambiente narrativo e siano coerenti tra loro in termini di prospettiva, colori, fuoco e distorsione introdotta dalla lente. Deve essere creata l'illusione che questi elementi siano stati filmati tutti nello stesso momento e luogo.
Il compositing tradizionale, anche chiamato compositing non digitale, si riferisce al compositing chimico, ottico e fisico. Nella prima metà del Novecento furono

realizzate delle stampanti ottiche per usare la tecnica della doppia esposizione che consisteva nell'esporre la pellicola due volte per combinare diversi elementi. In seguito furono utilizzate le travelling mattes. Quando veniva filmata la porzione frontale della scena lo sfondo non veniva esposto. Dopodiché veniva svolta la stessa operazione coprendo il foreground ed esponendo il background. Questa tecnica è anche chiamata in-camera matte. Un esempio dell'impiego di questa tecnica si vede in *The Great Train Robbery (1903)* di Edwin S. Porter, dove è stata usata per inserire un treno fuori dalla finestra di un ufficio postale.

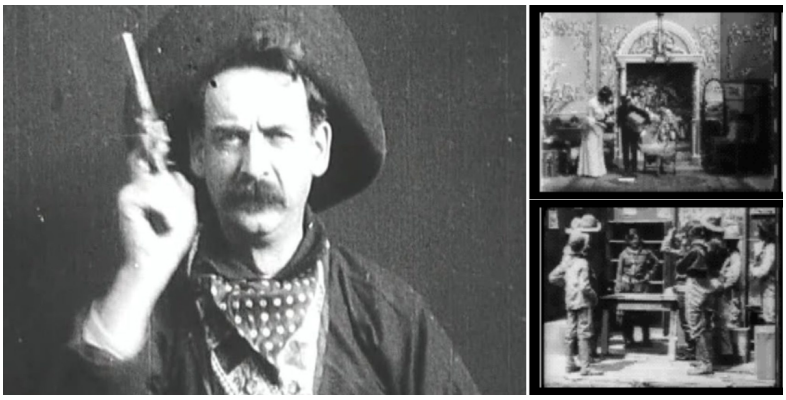

**Figure 2.12:** The Great Train Robbery (1903)
Fonte: `https://shorturl.at/HBWxw`

Furono anche utilizzati dei dipinti su vetro piazzati davanti alla camera per risultare in foreground. Aggiungendo degli elementi dietro al vetro veniva creata l'illusione di profondità. Il limite di questa tecnica era la staticità degli elementi in primo piano. Ne è un esempio *Gone with the Wind (1939)* di Victor Fleming, in cui la piantagione e i campi sono dipinti sul vetro mentre la strada e i soggetti in movimento sono posizionati dietro al pannello.

Un'altra tecnica comunemente utilizzata è quella delle miniature. Inserire all'interno di una scena un edificio come una torre alta può essere molto costoso, poco pratico o talvolta impossibile. L'utilizzo delle miniature può risolvere questo problema, inserendo un modellino in scala nell'inquadratura con la prospettiva adeguata. Nel 2011 è stato pubblicato il documentario *Sense of Scale (2011)* di Berton Pierce interamente dedicato al mondo degli effetti speciali e agli artisti che costruiscono miniature. In questo documentario viene citato anche l'impiego delle miniature in *Star Wars: A New Hope (1977)* di George Lucas. La miniatura del Millennium Falcon è stata ripresa su un blue screen da una camera i cui movimenti sono controllati da un computer. Sono stati poi aggiunti il background ed altri elementi in movimento tramite la doppia o tripla esposizione. Queste miniature sono state anche fatte esplodere attraverso delle mini cariche per simulare la loro distruzione

**Figure 2.13:** Gone With The Wind (1939)
Fonte: `https://shorturl.at/q2G02`

all'interno dei combattimenti del film.
Altre tecniche di compositing non digitale sono la proiezione frontale e la retro-proiezione (front and rear projection). Consistono nel proiettare frontalmente o posteriormente il foreground o il background. Queste tecniche sono state ampiamente utilizzate durante le scene in auto per simulare lo sfondo in movimento, come per esempio in *Dr. No (1962)*.


**Figure 2.14:** Dr. No (1962)
Fonte: `https://shorturl.at/RjDHB`

Infine una tecnica ormai superata è il chroma keying analogico, un processo molto lento e complesso paragonato alla controparte digitale, ma che permetteva effetti spettacolari. Per realizzare questo effetto l'oggetto in foreground era ripreso su un blue screen e nuovamente ripreso su pellicole con dei filtri colore che rendevano il

blue screen nero. Ogni oggetto indesiderato (come i supporti delle camere) erano mascherati con delle card nere. Il filmato risultante era una foreground matte. Per ottenere la background matte era necessario ripetere l'operazione con colori invertiti. Infine questi due rullini di pellicola potevano essere sovrapposti creando il compositing finale.
Il compositing digitale è l'evoluzione della controparte analogica. Funziona basandosi all'incirca sugli stessi principi implementando quelli che Lev Manovich nel suo libro "Il Linguaggio dei Nuovi Media" (2001) descrive come principi ispiratori dei nuovi media: la rappresentazione numerica, la modularità, l'automazione, la variabilità e la transcodifica. Un'immagine si può descrivere attraverso una funzione matematica ed è soggetta a manipolazione. I pixel che la compongono sono campioni discontinui che assemblati in una struttura continuano a mantenere le loro identità separate, secondo il principio di modularità. I primi due principi consentono l'automazione di molte operazioni necessarie per la creazione, la manipolazione e l'accesso.

> "Quindi l'intenzionalità umana può essere rimossa, almeno in parte, dal processo"[9]

Negli anni Novanta l'area dei nuovi media in cui era presente l'intelligenza artificiale fu per lo più il mondo dei videogiochi. Quasi tutti i videogiochi includevano una componente chiamata "AI engine", ovvero il codice del gioco che controlla i personaggi. Proprio perché i videogiochi sono altamente codificati e strettamente basati sulle regole, sono facilmente programmabili. Un nuovo oggetto mediale, inoltre, non è qualcosa che rimane identico a sé stesso all'infinito ma è qualcosa che può essere declinato in versioni molte diverse tra loro, secondo il principio di variabilità. Il compositing non digitale implicava l'immodificabilità del risultato. Il compositing digitale è variabile, mutabile e liquido.
Un film digitale può essere copiato e distribuito infinite volte senza alcuna perdita di qualità. Dal momento che manipolare un'immagine non significa intervenire permanentemente sull'immagine stessa, l'artista è più libero di sperimentare. Il compositing digitale è composto da un insieme di operazioni che in qualche modo alterano l'immagine. Queste operazioni sono funzioni matematiche che manipolano le informazioni contenute nei pixel dell'immagine.
I software di montaggio e compositing sono numerosi. Si dividono principalmente in due categorie: i software basati su nodi (node-based) e quelli basati su livelli (layer-based). Un node-based software gestisce le tasks di compositing con una struttura di nodi collegati tra loro. Ogni nodo definisce un'operazione matematica che altera l'immagine. L'insieme di questi nodi costituisce l'albero di nodi (node tree o node graph) che appare similmente a un diagramma di flusso (flowchart).
Un software layer-based, invece, presenta una struttura a livelli. Ogni elemento è organizzato uno al di sopra dell'altro. Ogni effetto applicato è assegnato a un layer

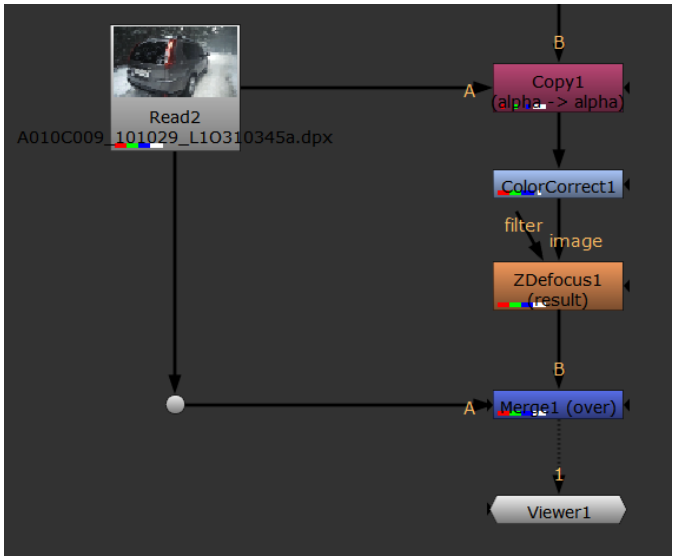


**Figure 2.15:** Node Graph Nuke.
Fonte: `https://shorturl.at/iyq1Z`

specifico. Viene così a formarsi una lista gerarchica di livelli in cui il livello sopra agisce sul livello sottostante e così via.[10]

## 2.3 Machine learning

Il machine learning (ML) è una branca dell'intelligenza artificiale che si concentra su sistemi capaci di imparare dai dati e migliorare le loro performance nel tempo, senza dover essere esplicitamente programmati. Il machine learning include al suo interno algoritmi che sono in grado di riconoscere dei pattern, ossia degli schemi ricorrenti, prendere decisioni e fare predizioni sui risultati sulla base dei dati in input.
L'obiettivo principale dell'apprendimento automatico è che una macchina sia in grado di generalizzare dalla propria esperienza e sia in grado di svolgere ragionamenti induttivi. Per generalizzazione si intende l'abilità della macchina di portare a termine un compito su materiale nuovo, attraverso l'esperienza acquisita durante l'apprendimento.
Esistono vari tipi di machine learning suddivisi per il grado di intervento dell'umano all'interno del processo: supervised learning, unsupervised learning, semi-supervised learning e reinforcement learning. Con il supervised learning l'algoritmo impara dai dati etichettati (labeled data) e dalle coppie di input-output forniti. Alcune applicazioni del supervised learning trovano spazio nel rilevamento di spam, classificazione di immagini e predizioni sul prezzo di magazzino.

L'algoritmo che sta alla base dell'unsupervised learning, invece, lavora con dati non etichettati (unlabeled data) e identifica pattern o strutture. Questo tipo di apprendimento viene utilizzato per operazioni di clustering come la segmentazione dei clienti e il rilevamento di anomalie.
Il semi-supervised learning, come suggerisce la parola, combina sia l'impiego di dati etichettati sia non etichettati per incrementare l'accuratezza durante l'apprendimento. Alcuni impieghi possono essere la classificazione del testo e il riconoscimento del parlato.
L'ultimo tipo di apprendimento, il reinforcement learning, prevede che il modello impari a prendere decisioni interagendo con un ambiente e ricevendo ricompense o penalità. Questo tipo di apprendimento è tipico della robotica.
Esiste un'ulteriore classificazione in base al grado di apprendimento richiesto per arrivare al risultato desiderato. Tra queste tipologie troviamo il "one-shot learning" e lo "zero-shot learning". Il primo è un apprendimento che richiede un singolo esempio affinché possa giungere a un risultato che più gli assomigli, affinché qualora venga sottoposto a materiale mai visto prima sia in grado di replicare ciò che ha appreso. Lo "zero-shot learning", d'altro canto, usa dati supervisionati durante l'apprendimento per riuscire a generalizzare senza alcun esempio richiesto.
Il workflow del machine learning segue diversi steps. Inizialmente occorre identificare e definire il problema, collezionare e pre-processare i dati, selezionare le features ed effettuare il training. In base alle necessità occorre scegliere il modello appropriato che dovrà essere allenato sulla base dei dati di training forniti e al termine dell'apprendimento andranno analizzati i dati per valutarne l'accuratezza, la precisione e il risultato finale. Il processo di machine learning prevede anche una fase importantissima di tuning, in cui si regolano i parametri sulla base dei risultati ottenuti, al fine di migliorare le performance del modello.
Il machine learning, in definitiva, può aiutare a trasformare la mole di dati già disponibili in un grande valore.[11]

### 2.3.1 Rete neurale

Le reti neurali (neural networks, NN) sono le fondamenta degli algoritmi di machine learning. Le differenze tra i neuroni biologici umani e quelli artificiali sono sottili. Entrambi i sistemi consistono in unità interconnesse che processano e trasmettono informazioni tramite strati di nodi interconnessi.
Nella maggior parte dei casi una rete neurale artificiale è un sistema adattivo che cambia la propria struttura in base a informazioni esterne o interne che scorrono attraverso la rete in fase di apprendimento. In termini pratici le reti neurali sono strutture non-lineari in cui ogni nodo elabora i segnali ricevuti e trasmette il risultato ai nodi successivi.
Le reti neurali sono un sistema di tipo statistico dotato di una buona immunità al

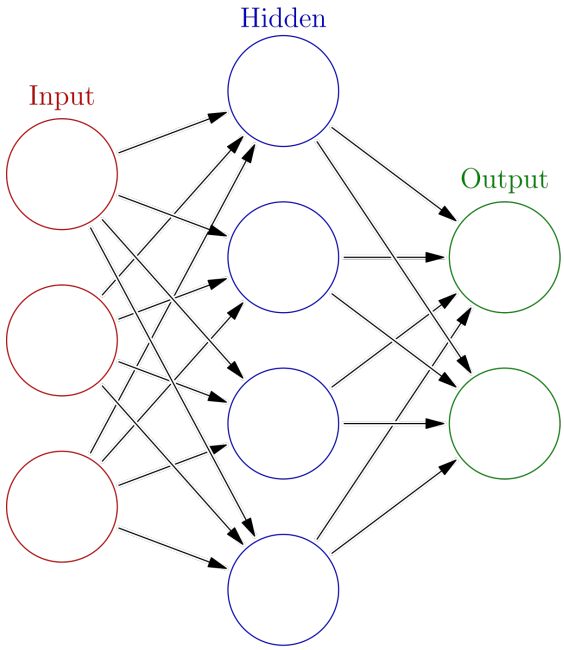


**Figure 2.16:** Rete neurale.
Fonte: `https://shorturl.at/rSMiu`

rumore. Ciò significa che se alcune unità del sistema dovessero funzionare male, la rete nel suo complesso avrebbe delle riduzioni di prestazioni ma difficilmente andrebbe incontro ad un blocco del sistema.
Di contro, le reti neurali vengono spesso definite "black box", per indicare la loro scarsa interpretabilità del processo interno che porta alle loro decisioni. Questo è dovuto al fatto che possono avere anche milioni o miliardi di parametri ed è complicato per un umano comprendere come ogni peso contribuisca al risultato finale.
Esistono diversi tipi di reti neurali. Alcuni di questi sono: reti feed-forward, reti convoluzionali, reti ricorrenti, reti basate su attenzione e transformer, reti generative e autoencoders.
Le reti feed-forward (FNN) presentano un grafico aciclico diretto e sono le prime reti messe a punto. Queste reti hanno un layer di input che riceve i dati grezzi e li trasmette ai layers successivi, passando per layers nascosti e arrivando al layer di output. Questa rete è migliorata con la discesa del gradiente (gradient descent), una tecnica che permette di automatizzare i parametri dei nodi aggiustando pesi per minimizzare la funzione di perdita. Questa discesa è guidata dalla retro-propagazione dell'errore (backpropagation) che calcola il gradiente in ordine inverso, partendo dall'output layer e muovendosi indietro tramite i layers nascosti.

Le reti convoluzionali (CNN) sono progettate con diversi strati ognuno specializzato in un diverso livello di rilevamento delle features. Questi strati convoluzionali applicano dei filtri alle immagini in input, rilevando caratteristiche semplici come bordi o angoli. Gli strati successivi sono poi in grado di effettuare un rilevamento più complesso, a partire dalle features semplici rilevate. L'innovazione introdotta dalle CCNs sta nel loro campo recettivo locale. Come la visione umana esse sono in grado di concentrarsi su una zona locale dell'immagine, riconoscendo oggetti che variano di posizione, dimensione e orientamento. Come detto in precedenza, alla base del funzionamento di queste reti ci sono i filtri convoluzionali, anche conosciuti come kernels. Questi filtri scorrono attraverso l'immagine in input, svolgendo operazioni matematiche per rilevare features specifiche, divenendo in grado di riconoscere patterns. Dopo aver estratto le features una rete CNN è anche in grado di classificare indicando la presenza o meno di patterns all'interno dell'immagine. Alcuni ambiti applicativi di questa tipologia di rete neurale nella computer vision possono essere il riconoscimento di un oggetto all'interno di un'immagine (modelli come ResNet), il riconoscimento facciale, la segmentazione semantica (che verrà trattata in seguito) e il miglioramento della qualità delle immagini (SRGAN).
Le reti neurali ricorrenti (RNN) lavorano molto bene con i dati sequenziali in quanto sono molto abili a riconoscere le dipendenze temporali. Sono reti che mantengono memoria degli steps precedenti grazie alla struttura formata da loops. Questo meccanismo di memoria è fondamentale per analizzare il contesto ed individuare le dipendenze tra sequenze.
Le reti neurali attention-based sono un'evoluzione ulteriore. Tra queste troviamo il transformer, un modello di rete neurale che assegna un peso alle diverse parti dell'input in base alla loro rilevanza. Questo modello usa una tecnica nota come "self-attention". Invece di utilizzare una rete sequenziale per codificare le informazioni, elabora tutte le parole dell'input in parallelo tramite una matrice di attenzione. I Vision Transformer (ViT) classificano immagini sfruttando il meccanismo di attention e sono utilizzati per classificare oggetti o scene. Inoltre queste reti neurali sono utilizzate per la generazione di immagini come ad esempio i modelli DALL-E o Stable Diffusion basati su transformer.
Le reti generative (GAN) prevedono un continuo dualismo tra generatore e discriminatore per il raggiungimento dell'output. Le GAN sono le reti che hanno abilitato la possibilità dello style transfer, un processo in cui è possibile combinare lo stile di un'immagine o di un video e il contenuto di un altro.
Infine gli autoencoders sono degli strumenti utili a risolvere tasks generative. Aprono le porte alla compressione delle informazioni e al processo di denoising. I Variational Autoencoders (VAE) estendono il concetto introducendo una meccanica probabilistica. Invece di apprendere la rappresentazione latente deterministica, i VAE imparano la distribuzione probabilistica degli spazi latenti.[12]

### 2.3.2 Deep learning

> "Il deep learning (DL) è un sottoinsieme di machine learning che utilizza reti neurali multilivello, chiamate reti neurali profonde, per simulare il complesso potere decisionale del cervello umano."

Il deep learning si differenzia dal machine learning per la struttura dell'architettura della rete neurale su cui si basa. Il machine learning prevede uno o due livelli computazionali mentre il deep learning da tre in su, di solito centinaia o migliaia. I modelli di deep learning possono utilizzare l'apprendimento non supervisionato e quindi estrarre features da dati non elaborati. Nel tempo, un algoritmo di deep learning diventa sempre più accurato grazie alla progressione di calcoli attraverso i livelli. Il deep learning richiede grande capacità computazionale, fornita da schede video prestanti. Molte delle reti neurali viste in precedenza sono impiegate negli algoritmi di deep learning, come le CNN, le RNN, le GAN, i transformer e gli autoencoders. Anche i modelli di diffusione, ossia i modelli generativi, rientrano nel deep learning. Sono addestrati utilizzando il processo di diffusione in avanti e in senso inverso di addizione progressiva del rumore e negazione. Aggiungono gradualmente rumore gaussiano ai dati di addestramento fino a renderli riconoscibili, poi svolgono un processo di "denoising" inverso in grado di sintetizzare l'output (principalmente immagini) da un input di rumore casuale o derivante da un'altra immagine. Rispetto alle GAN i modelli di diffusione sono più stabili e producono risultati migliori, al costo di un maggiore calcolo computazionale.

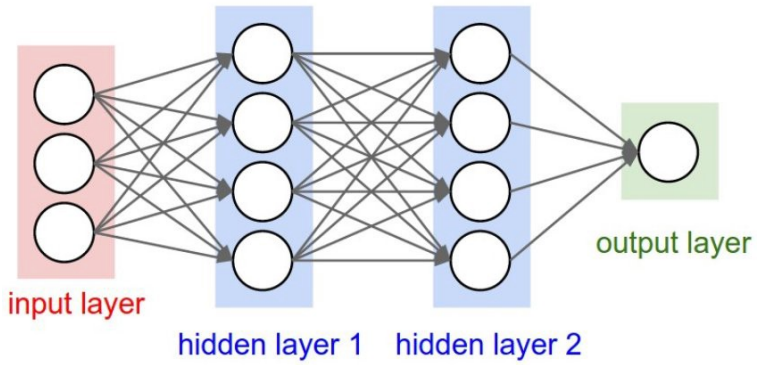


**Figure 2.17:** Deep learning.
Fonte: `https://shorturl.at/9iVpB`

## 2.4 Computer vision

> "La computer vision (CV) è un campo dell'intelligenza artificiale che utilizza il machine learning e le reti neurali per insegnare ai computer e ai sistemi a ricavare informazioni significative da immagini digitali, video e altri input visivi e a formulare raccomandazioni o intraprendere azioni quando vengono identificati dei patterns"

I primi esperimenti in ambito computer vision furono fatti nel 1959, quando alcuni neurofisiologi mostrarono ad un gatto una serie di immagini, nel tentativo di correlare una risposta nel suo cervello. Scoprirono che esso rispondeva prima ai bordi o alle linee rette, provando che la sua elaborazione delle immagini iniziava con forme semplici. Nello stesso periodo fu sviluppata la prima tecnologia che permetteva di scansionare delle immagini tramite computer. Nel 1963 i computer furono finalmente in grado di trasformare immagini bidimensionali in forme tridimensionali, mentre nel 1974 fu introdotta la tecnologia di riconoscimento ottico dei caratteri (OCR), in grado di riconoscere il testo stampato. Con lo sviluppo del riconoscimento intelligente dei caratteri (ICR) questi due tipi di riconoscimenti sono stati impiegati nell'elaborazione di documenti e fatture, nel riconoscimento delle targhe dei veicoli e nei pagamenti con i dispositivi mobili. Dal 1982 in avanti furono sviluppati algoritmi per rilevare bordi, angoli, curve e forme e furono integrate al loro interno le reti neurali convoluzionali. Dal 2000 fu possibile anche effettuare il riconoscimento di oggetti (object recognition) e l'anno successivo fu il momento del riconoscimento facciale in tempo reale.
Anche l'industria degli effetti visivi ha beneficiato di queste nuove tecnologie e strumenti. Una delle applicazioni maggiormente utili della computer vision è la segmentazione di sequenze video.

### 2.4.1 Segmentazione video

La segmentazione video (video segmentation) è il processo di partizionamento di un video in regioni multiple basate su caratteristiche specifiche, come per esempio i bordi di un oggetto, il movimento, il colore, la texture o altre features visive. L'obiettivo è identificare e separare oggetti diversi dal background in sequenze video. La segmentazione video può agire su vari livelli di granularità, partendo dalla segmentazione di oggetti in uno shot fino alla segmentazione di intere inquadrature o scene.
Sebbene i modelli di segmentazione delle immagini condividano determinati utilizzi con i modelli di rilevamento degli oggetti, essi differiscono per un aspetto fondamentale: identificano le diverse entità contenute in un'immagine a livello di pixel, anziché approssimarle con un riquadro di contorno (bounding box). In sostanza,


**Figure 2.18:** Computer vision.
Fonte: `https://shorturl.at/2eimJ`

un modello di classificazione può definire gli oggetti contenuti in un'immagine, un modello di rilevamento di oggetti ne identifica la loro posizione e un modello di segmentazione ne identifica le forme e i confini.
Il funzionamento che sta alla base di questa tipologia di modelli è un processo di analisi dell'immagine end-to-end che divide l'immagine digitale in più segmenti e classifica le informazioni contenute in ciascuna regione.
Esistono vari tipi di video segmentation: la segmentazione di oggetti, la segmentazione semantica, la segmentazione per istanze e la segmentazione panottica.

### Segmentazione oggetti

La segmentazione per oggetti di una sequenza video ha come obiettivo segmentare un oggetto e farne il tracking per l'intera durata della sequenza.
Esistono tre tipologie di object segmentation che variano in base al grado di supervisione: unsupervised, semi-supervised, interactive.
Quella non supervisionata non prevede alcuna informazione etichettata e utilizza un algoritmo di optical flow per stabilire i vettori di movimento dei pixel e quindi il movimento degli oggetti.
Il processo semi-supervised, invece, utilizza pochi dati etichettati per una maggiore precisione. Un esempio di questo processo è il modello Sparse Spatiotemporal Transformers (SST) che include un passaggio feedforward in un meccanismo

attention-based.
Infine, l'interactive video object segmentation è molto utile quando si ha la necessità di agire in real-time.

### Segmentazione semantica

La segmentazione semantica etichetta ogni singolo pixel contenuto in un'immagine in base alla sua classe semantica. Questi modelli creano una mappa di segmentazione, cioè una ricostruzione dell'immagine fornita in input in cui ogni pixel viene codificato con i colori della classe semantica, creando così delle maschere di segmentazione. Una maschera di segmentazione è una porzione dell'immagine che è stata differenziata dalle altre regioni.[13]

**Figure 2.19:** Segmentazione semantica.
Fonte: `https://shorturl.at/CyzlC`

Ogni classe semantica è rappresentata da una matrice di pixel le cui celle assumono il valore uno laddove è presente quella classe e il valore zero dove essa non è presente.
Un approccio comune nella segmentazione semantica è quello di creare una SegNet, una rete basata su architettura CNN. Questa CNN, attraverso un'implementazione inversa della rete, l'aggiunta di sovracampionamento e sottocampionamento, è in grado di classificare l'immagine a livello di pixel. Un layer di output, infine, classifica i pixel associandoli a una classe specifica.
Questi modelli richiedono l'utilizzo di grandi dataset per l'apprendimento, pre-etichettati da umani. Molti dataset sono disponibili anche open source, tra cui il Pascal Visual Object Classes (Pascal VOC), l'MS COCO e i dataset sui paesaggi

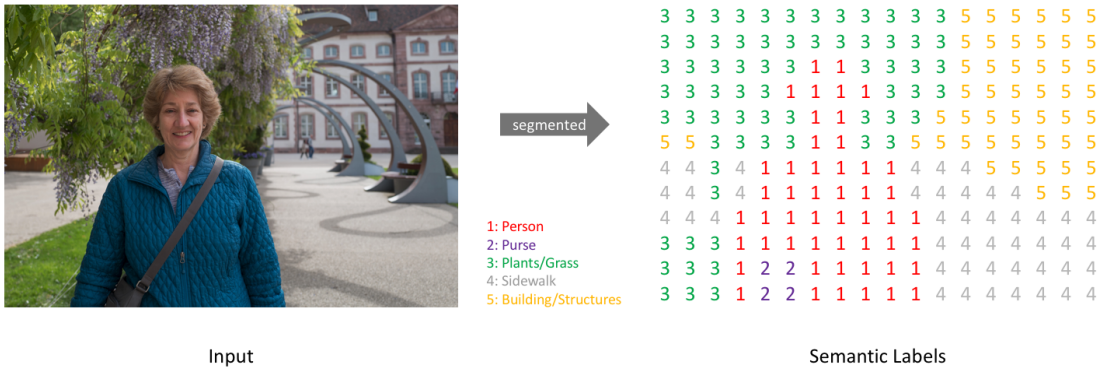


**Figure 2.20:** Semantic labels.
Fonte: `https://shorturl.at/17cdj`

urbani.
Con l'impiego di reti neurali complesse è possibile utilizzare questi modelli su shots mai visti prima per ricavarne maschere di segmentazione.

### Segmentazione istanze

La segmentazione per istanze segmenta ogni oggetto e tutte le sue istanze in maschere separate, pur appartenendo alla stessa classe.

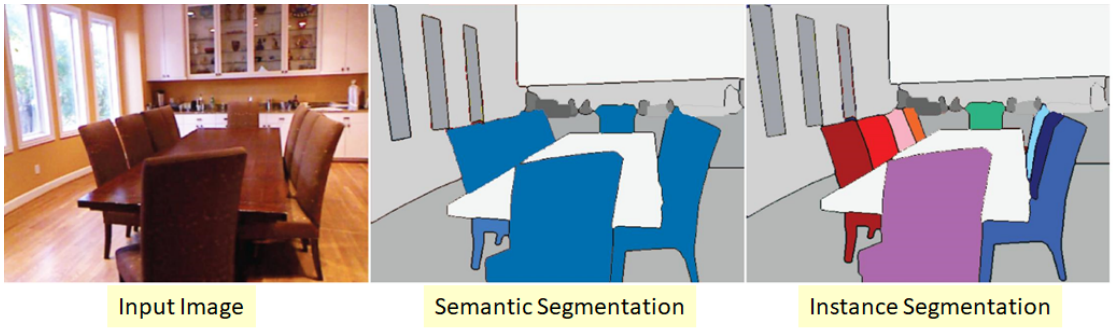


**Figure 2.21:** Segmentazione istanze.
Fonte: `https://shorturl.at/WcWyD`

### Segmentazione panottica

La segmentazione panottica è praticamente l'unione della semantic e dell'instance segmentation.

(**a**) Image (**b**) Semantic segmentation

(**c**) Instance segmentation (**d**) Panoptic segmentation

**Figure 2.22:** Tipi di segmentazione.
Fonte: `https://shorturl.at/p7ELq`

# Chapter 3

# Fase di test

A seguito di una fase iniziale in cui è stato studiato ed analizzato lo stato dell'arte dell'intelligenza artificiale applicata al campo degli effetti visivi, si è passati all'individuazione degli strumenti più utili al fine di segmentare sequenze video e ottenere maschere utilizzabili nei processi di compositing e post-visualizzazione. Questi tools sono stati testati su shots di produzione per verificarne l'efficienza e l'efficacia in contesti reali di produzione. Questa fase è stata progettata in modo tale da rispecchiare gli scenari tipici di lavorazione utilizzando software compatibili con la pipeline già sviluppata e instaurata in azienda, così da poterne valutare le performance e il possibile inserimento di queste nuove tecniche di mascheramento. In questo capitolo vengono prima mostrati i componenti del computer su cui sono stati svolti i test per stabilire un rapporto tra la velocità di calcolo e la potenza computazionale del PC, vengono descritti i software impiegati in questa fase e gli strumenti testati e infine vengono presentati gli shots su cui sono stati effettuati i test con i relativi risultati ottenuti.

## 3.1 Hardware utilizzato

Tutto questo lavoro in azienda è stato svolto su un computer fisso con il processore AMD Ryzen 7950X3D, 128 GB di RAM e la scheda video NVIDIA Geforce RTX 4080 SUPER. I componenti utilizzati sono importanti al fine di dare un contesto alle tempistiche di elaborazione degli strumenti testati e dunque al tempo impiegato per ottenere un risultato. Inoltre, è da tenere in considerazione che questi tool richiedono una buona capacità di calcolo per poter funzionare.
Possedere una buona GPU è di fondamentale importanza nell'impiego di strumenti AI o di Machine Learning. Le schede video sono progettate per effettuare il calcolo parallelo che, rispetto al calcolo sequenziale tipico delle CPU, velocizza di molto i tempi di calcolo. Molti modelli di Deep Learning sono caratterizzati da un elevato

numero di parametri e livelli. Le GPU sono in grado di gestire questi tipi di modelli in maniera efficiente, consentendo anche l'addestramento di reti neurali profonde. Oltre all'addestramento di questi modelli, le schede video sono importanti per la fase di inferenza, ossia l'utilizzo di modelli addestrati per fare predictions o prendere decisioni. Le schede video permettono di accelerare questo processo.

## 3.2 Software utilizzati

I software utilizzati sono gli standard per la realizzazione degli effetti visivi e sono i programmi utilizzati in EDI, così da poter offrire un confronto con la pipeline classica e quella con l'impiego dell'AI e poter valutare l'implementazione di queste nuove tecniche all'interno della lavorazione già instaurata in azienda.
Sono stati utilizzati i seguenti software: Nuke, Silhouette e ComfyUI. L'unico tra questi a non essere impiegato all'interno della pipeline è ComfyUI.

### 3.2.1 Nuke

Nuke è un software di compositing basato su una struttura a nodi, sviluppato per la prima volta nel 1993 da Bill Spitzak, nell'azienda Digital Domain. Fu da subito impiegato per film come *True Lies (1994)*, *Apollo 13 (1995)* e *Titanic (1997)*. Nel mentre, in un garage dell'Inghilterra, Simon Robinson e Bruno Nicoletti fondavano The Foundry, ribrandizzata come Foundry nel 2017.
Nel 2007 Nuke divenne parte di Foundry e negli anni a seguire espanse i propri limiti aggiungendo centinaia di nuove features, come un camera tracker interno, il denoise, il deep compositing e strumenti per la gestione della stereoscopia. Estese il suo core con Python, Qt e il supporto multi-platform. Presto divenne uno dei software industry standard per il compositing.
In seguito a numerosi premi e riconoscimenti ottenuti, nel 2021 viene rilasciato Nuke 13.0 che introduce per la prima volta un toolset di Machine Learning nella versione NukeX. Con questa nuova release vengono inseriti due nuovi nodi, CopyCat e Inference, aprendo le porte al machine learning legato al compositing.[14]


**Figure 3.1:** Nuke.
Fonte: `https://shorturl.at/2Ftmq`

### 3.2.2 Silhouette

Silhouette è un software sviluppato da Boris FX ed è un programma industry standard per le operazioni di Roto e Paint.
Nel 2024, con la versione 2024.5, sono stati introdotti nuovi nodi che impiegano il machine learning e l'intelligenza artificiale all'interno del software: Matte Assist ML, Optical Flow ML, Retime ML, EZ Mask e Mask ML.[15]

### 3.2.3 ComfyUI

ComfyUI è un programma open source e node-based che permette agli utenti di generare immagini, video e audio con la GenAI. Con GenAI si intende un tipo di AI che è in grado di generare testo, immagini, video e audio in risposta a richieste dette "prompt".
Questi sistemi utilizzano modelli generativi, un esempio tra questi sono i modelli linguistici di grandi dimensioni (Large Language Models, LLMs) che producono dati a partire da un dataset di addestramento. Tra i sistemi di GenAI è compreso anche ChatGPT, basato sui modelli GPT-3 e GPT-4.
ComfyUI è stato rilasciato su GitHub a gennaio 2023 dall'utente comfyanonymous.[16] A luglio 2024, NVIDIA ha annunciato il supporto per ComfyUI all'interno del suo software RTX Remix, mentre ad agosto 2024 è stato implementato il supporto per il modello Flux, un modello di diffusione sviluppato da Black Forest Labs.[17] Attualmente ComfyUI è la UI più popolare per Stable Diffusion (un modello di deep learning per generare immagini), assieme ad Automatic1111.


**Figure 3.2:** Silhouette.
Fonte: `https://shorturl.at/tYcGZ`

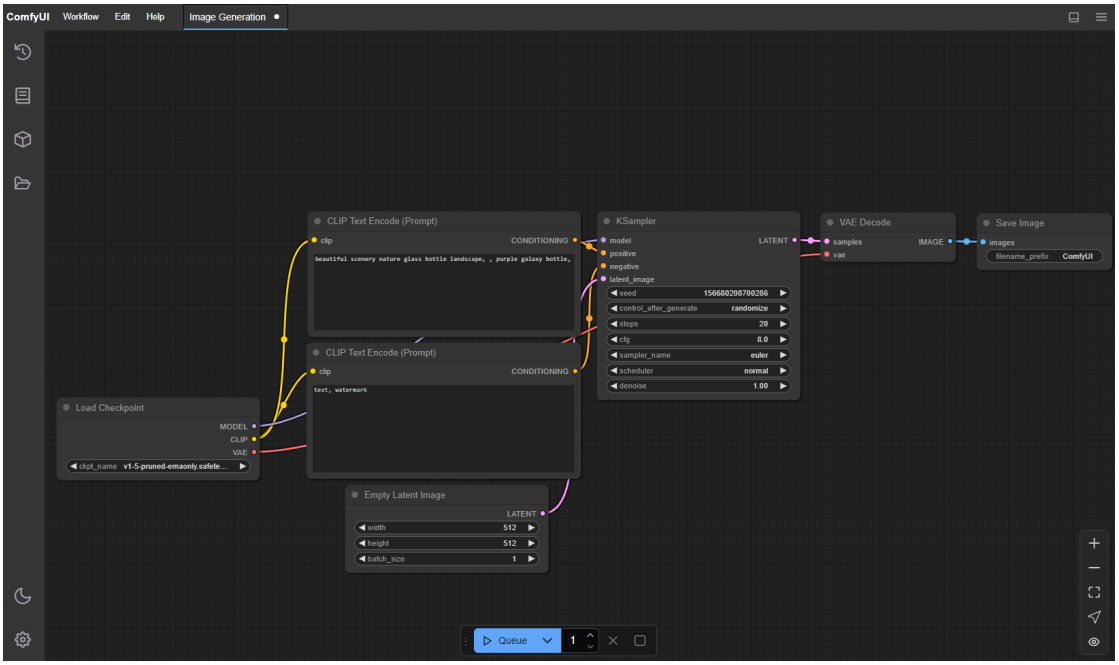


**Figure 3.3:** ComfyUI.
Fonte: `https://shorturl.at/7JQKB`

ComfyUI è da immaginare come una tela digitale su cui si possono tessere i propri flussi di lavoro connettendo diversi nodi, ognuno dei quali rappresenta una funzione o un'operazione specifica.

Durante lo stage curriculare è stato integrato ComfyUI all'interno del software Nuke, con l'impiego del metodo sviluppato dall'utente vinavfx.[18] Con questa integrazione ogni artista o compositor di EDI che utilizza Nuke può usufruire della quasi totalità dei nodi di ComfyUI senza mai uscire da Nuke. L'integrazione è stata possibile grazie alle API (Application Programming Interface) di ComfyUI.

## 3.3 Strumenti utilizzati

In seguito a un attento studio degli strumenti di segmentazione video e generazione di maschere con l'impiego di intelligenza artificiale o machine learning, sono stati individuati i seguenti: Segment Anything (SA), Segment Anything 2 (SA2), Matte Assist ML, EZ Mask, Mask ML e CopyCat.

### 3.3.1 Segment Anything

Il progetto di Segment Anything è stato sviluppato nell'aprile 2023 da Meta, con la costruzione del dataset di segmentazione più grande a quel tempo. Esso conta oltre 1 bilione di maschere su 11 milioni di immagini, rispettando le licenze e la privacy. Il modello Segment Anything Model (SAM) è progettato e addestrato per essere gestibile tramite prompt, in modo tale che possa compiere task zero-shot.
I Large Language Models pre-addestrati sui dataset web-scale stanno rivoluzionando il "Natural Language Processing" (NLP), con una generalizzazione zero-shot o few-shot. Il NLP è una sottobranca di linguistica, informatica e AI che tratta l'interazione tra i computer e il linguaggio umano, in particolare su come programmare i computer per elaborare e analizzare grandi quantità di dati di linguaggio naturale. Questa capacità è spesso implementata con il "prompt engineering", cioè il processo di strutturazione e creazione di un'istruzione al fine di produrre il miglior risultato con l'AI. Un prompt è un testo in linguaggio naturale che descrive il compito che un'intelligenza artificiale dovrebbe eseguire. Un prompt di alta qualità, completo e ben congegnato, incide a sua volta sulla qualità dei contenuti generati dall'AI, che si tratti di immagini, codice o testi.[19]

#### Funzionamento

Il funzionamento del Segment Anything si basa su un modello che supporta prompting flessibile e in grado di generare maschere di segmentazione in real-time in maniera interattiva. Per raggiungere questo risultato è necessario definire in maniera chiara la "promptable segmentation task", cioè l'attività di segmentazione definita tramite un prompt. Il prompt in questo caso è definito dall'individuazione e dalla scelta dell'area che si desidera segmentare. Questo modello ha come obiettivo quello di fornire una maschera valida in output. Ciò significa che, nel caso in cui il prompt

sia ambiguo e si riferisca a più oggetti (ad esempio un punto su una maglietta indica sia la maglietta sia la persona intera), il modello fornisce comunque in output almeno una maschera tra le possibili.
Il modello comprende un image encoder e un prompt encoder. Le due informazioni in uscita vengono poi combinate da un decoder che predice le maschere di segmentazione.
Per raggiungere una forte generalizzazione il SAM è stato addestrato su un vasto set di maschere durante tre fasi: manuale assistita, semi-automatica e completamente automatica. Nel primo stage il modello assiste l'utente nell'annotazione delle maschere, nel secondo riesce a generare automaticamente maschere sulla base di prompt che indicano la posizione degli oggetti da segmentare e nel terzo stage SAM è in grado di generare circa 100 maschere di alta qualità per immagine, sulla base di una griglia di punti in foreground.

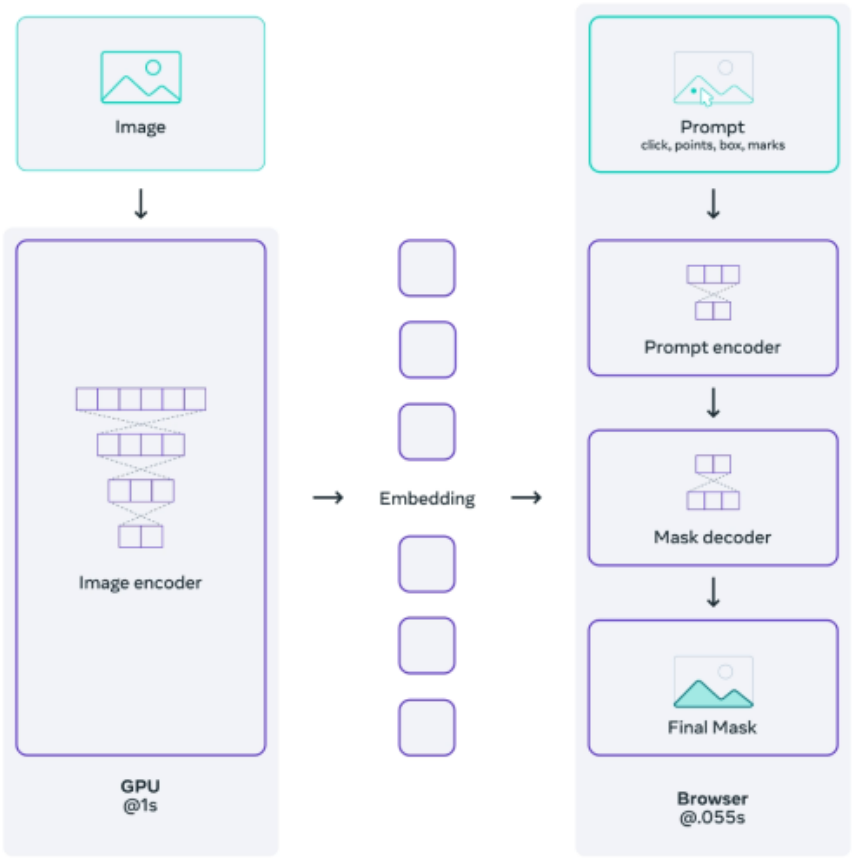


**Figure 3.4:** Funzionamento SAM.
Fonte: `https://shorturl.at/wiyJn`

SAM è composto da tre componenti: un image encoder, un prompt encoder flessibile e un decoder veloce per la maschera. Il modello è stato costruito sui modelli Transformer Vision (ViT) con dei trade-off per un risultato real-time. L'image encoder è un ViTMAE pre-addestrato, ossia un Vision Transformer Mask Autoeconder. Dopo che l'immagine è stata codificata agisce il prompt encoder che considera due tipologie di prompt: sparse (punti, riquadri, testo) e dense (maschere). I punti e i riquadri sono rappresentati con codifiche di posizione mentre il testo con un text encoder CLIP (Contrastive Language-Image Pre-Training). I dense prompts, invece, sono incorporati usando convoluzioni. Il mask decoder, infine, mappa l'image embedding (la rappresentazione numerica dell'immagine), i prompt embeddings e un token di output alla maschera. In questo processo è inserita anche la self-attention e la cross-attention bi-direzionale (prompt to image, image to prompt). Dopo aver svolto questa operazione l'immagine viene sovracampionata e un MLP (Multi Layer Perceptron), una rete neurale feedforward, mappa l'output token a un classificatore dinamico lineare che, in seguito, calcola la probabilità della maschera in foreground in ogni posizione dell'immagine.
Per risolvere l'ambiguità è stato appurato che tre maschere in output sono sufficienti per coprire il più dei casi (le maschere annidate spesso hanno tre livelli di profondità: intera, una parte e una sottoparte). Durante l'apprendimento si tiene in considerazione solo la perdita minima. Per stabilire la classifica delle maschere il modello predice un punteggio di attendibilità (confidence score) per ogni maschera.

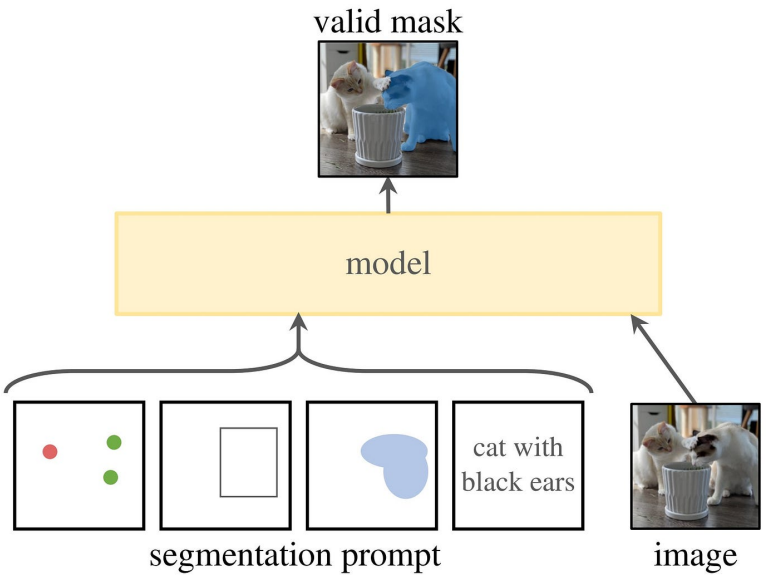


**Figure 3.5:** Maschera valida SAM.
Fonte: `https://shorturl.at/uh4Ib`

Foundry mette a disposizione una libreria open source di modelli di machine learning convertiti in file .cat che sono in grado di lavorare nativamente dentro Nuke. Questa libreria si chiama Cattery. Tutti i file all'interno della libreria sono sotto la licenza BSD 3-Clause, che consente la ridistribuzione e l'uso in forma di codice sorgente o binaria con o senza modifiche, a condizione che conservino l'avviso di copyright, l'elenco di condizioni e il disclaimer.
Cattery comprende numerosi file .cat, per vari casi applicativi.

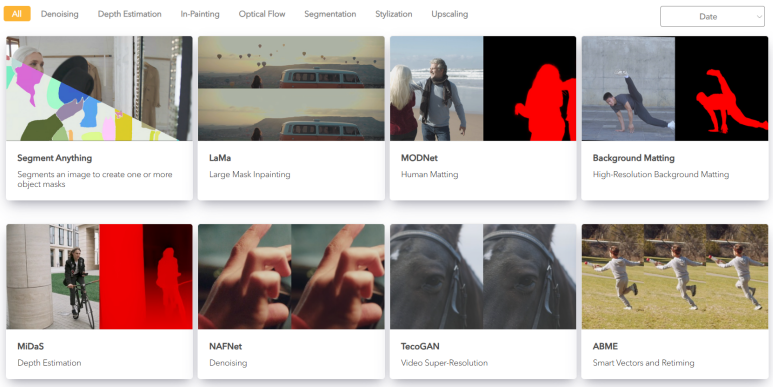


**Figure 3.6:** Cattery Nuke

Tra questi spicca il Segment Anything.
Il suo funzionamento prevede la connessione del relativo nodo al plate e l'impostazione dei parametri all'interno delle sue proprietà.
Attualmente il SegmentAnything di Cattery non è temporalmente consistente in modalità Multi Mask (in questa modalità la maschera di segmentazione varia ad ogni frame della sequenza), dunque il setup della CryptoMatte (un'immagine in cui ogni oggetto è contraddistinto da un'ID univoco, utile in fase di compositing) non è facilmente realizzabile in quanto il Mask ID varia in ogni frame. È stato dunque utilizzato in modalità Single Mask che genera un canale alpha unico per il soggetto che si desidera mascherare.
Esiste una versione migliorata di questo strumento, integrabile in Nuke grazie all'utente rafaelperez.[20] Il risultato ottenuto è molto simile in entrambi i casi. Quello di Rafael Perez, però, risulta essere più leggero, veloce e intuitivo da utilizzare, poiché si ha la possibilità di visualizzare la maschera in modalità di trasparenza "over" direttamente sul plate.

### 3.3.2 Segment Anything 2

Segment Anything Model 2 (SAM 2) è un modello per la segmentazione di immagini e video. Come per il modello precedente, è stato costruito dagli sviluppatori un

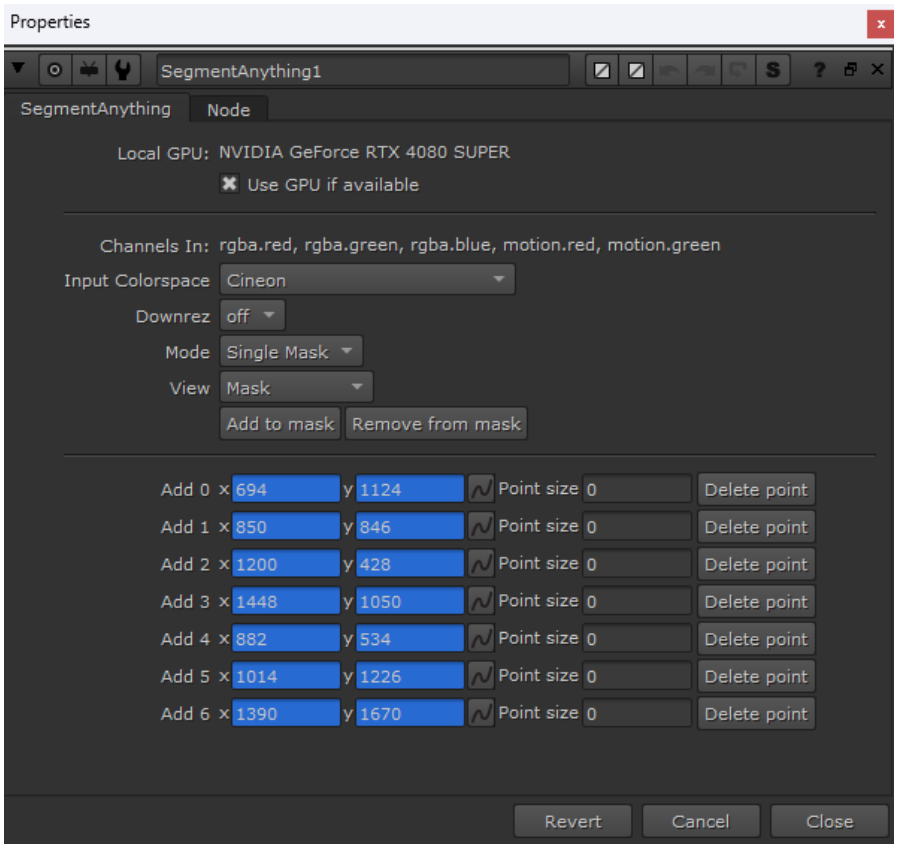


**Figure 3.7:** Proprietà Segment Anything

data engine, il quale migliora il modello e le informazioni attraverso l'interazione con l'umano, per collezionare il più grande dataset di segmentazione video. SAM 2 è basato su un'architettura transformer con una memoria per il processamento di video in real-time. Questo modello ha dimostrato grandi performance su una grande varietà di task. Nella segmentazione video è stata osservata un'accuratezza maggiore, usando un terzo delle interazioni rispetto all'approccio precedente. Nella segmentazione di immagini SAM 2 è sei volte più veloce rispetto a SAM.
Con la rapida crescita dei contenuti multimediali, si è reso necessario estendere il modello SAM ai video oltre che alle immagini. Viene considerata, perciò, un'ulteriore dimensione per le informazioni video: il tempo. La segmentazione nei video mira a determinare il cambiamento spazio-temporale delle entità, una sfida unica rispetto alle semplici immagini. Le entità possono subire grandi modifiche tra un frame e l'altro a causa del movimento, della deformazione, dell'occlusione, dei cambiamenti di luce e di altri fattori. SAM 2 lavora con una Promptable Visual Segmentation (PVS) task che generalizza la segmentazione di immagini al campo dei video. La task prende in input punti, riquadri o maschere in qualsiasi

**Figure 3.8:** Segment Anything Rafael Perez

frame del video per definire un segmento di interesse per il quale una maschera spazio-temporale viene predetta. Una volta che la maschera è stata predetta può essere iterativamente rifinita fornendo ulteriori prompt negli altri frame.[21]

### Funzionamento

SAM 2 può essere visto come una generalizzazione di SAM per i video che genera una maschera per un frame e la estende spazio-temporalmente per l'intera sequenza. Spazialmente il modello agisce come SAM. L'architettura di questo modello prevede i seguenti blocchi: image encoder, memory attention, prompt encoder e mask decoder, memory encoder e memory bank.
Per un processing real-time di video di durata arbitraria è stato impiegato un approccio simile al campo dello streaming, trattando i frames una volta che sono disponibili.
Il ruolo della memory attention è di condizionare le features del frame corrente sulla base delle features dei frames precedenti e delle predizioni, così come dei nuovi prompt nel frame corrente. Vengono utilizzati meccanismi di self e cross attention e vengono immagazzinate le informazioni in un banco di memoria, seguito da un MLP (Multi Layer Perceptron).
Il prompt encoder è identico a SAM e può essere sia positivo sia negativo. La gestione di prompt ambigui è gestita come in SAM, con la particolarità che in un video vengono predette maschere multiple per ogni frame. Se nessun prompt aggiuntivo risolve l'ambiguità, viene fornita in output la maschera con lo score di

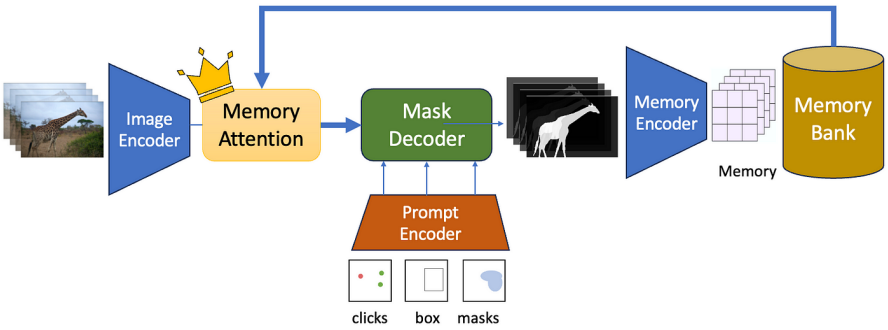


**Figure 3.9:** Architettura SAM2.
Fonte: `https://shorturl.at/5kmtH`

affidabilità più elevato.
Rispetto a SAM c'è una sostanziale differenza: non è sempre presente un oggetto valido da segmentare sulla base di un prompt fornito. Questo è dovuto all'occlusione che può essere presente in alcuni frame.
Il banco di memoria mantiene le informazioni dai frame precedenti, tenendo una coda FIFO (First In First Out), cioè un metodo di transito di oggetti in cui il primo ad entrare è il primo ad uscire. L'ordine di uscita dunque è uguale a quello di entrata. Viene mantenuta una memoria spaziale come una mappa di features dei frame e una lista di puntatori di oggetti come vettori di spostamento degli oggetti tra fotogrammi. In questo modo il modello è in grado di rappresentare il movimento a breve termine degli oggetti ma non in quelli in cui è presente del prompt, perché in quei frame sono presenti dense prompts molto complessi da generalizzare per l'intera sequenza.
Il dataset collezionato con il data engine di SAM 2 è il SA-V e comprende 50.9K video con 642.6K maschere ottenute. Le riprese sono per il 54% indoor e 46% outdoor, di una durata media di 14 secondi. Per le maschere sono state fatte 190.9K maschere a mano e sono state generate 451.7K maschere automatiche.
Grazie all'utente kijai su GitHub è disponibile una versione del Segment Anything 2 integrabile in ComfyUI.[22]
Per installare dei Custom Node su ComfyUI è necessario entrare nel Manager.

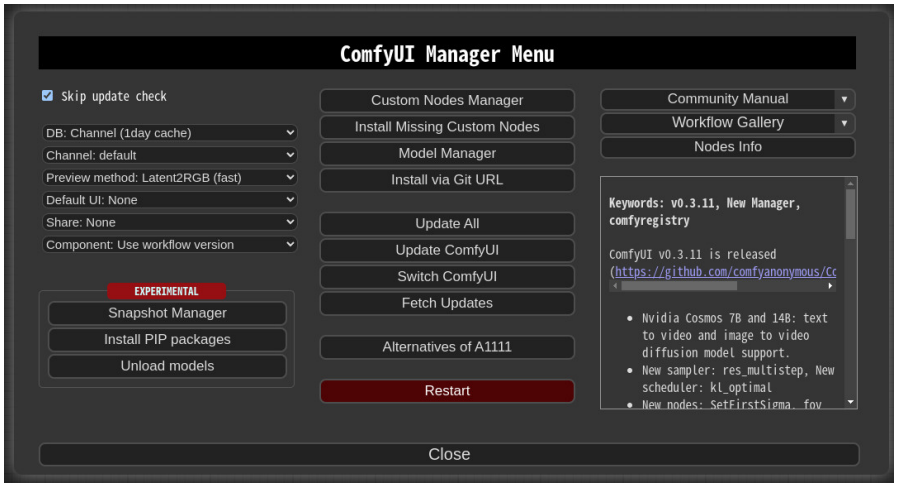


**Figure 3.10:** ComfyUI Manager.
Fonte: `https://shorturl.at/DYfVs`

Bisogna entrare nella sezione "Custom Nodes Manager", cercare il nome del nodo e cliccare su "Install".
Infine occorre fare "Restart" e il nodo apparirà tra quelli utilizzabili.
Il funzionamento di SAM2 su ComfyUI è ancora macchinoso. ComfyUI al momento dei test lavora solamente con immagini, video o sequenze di immagini .exr, pertanto, se si ha una sequenza .dpx bisogna prima convertirla. Un'altra difficoltà nell'utilizzare questo strumento su ComfyUI è la poca reattività in real-time. Ogni qualvolta si effettua una modifica, per vederla applicata è necessario far partire l'elaborazione premendo su "Queue prompt".

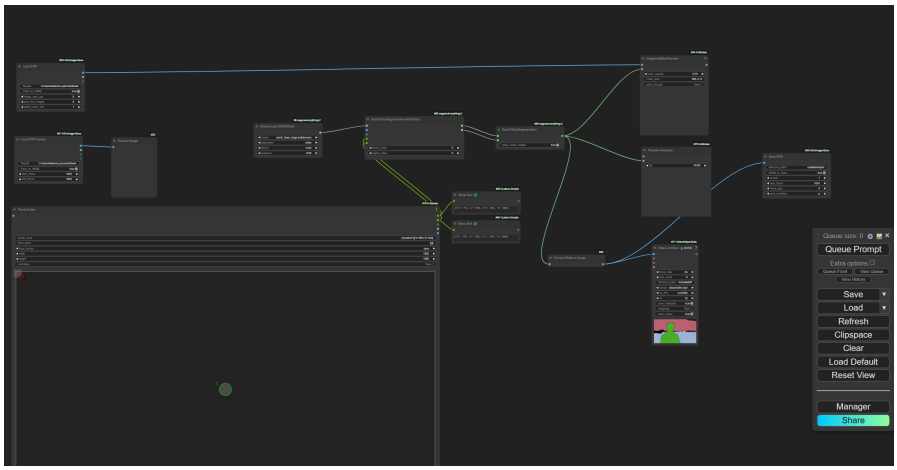


**Figure 3.11:** SAM2 su ComfyUI

### 3.3.3 ViTMatte

Il ViTMatte è uno strumento in grado di ricalcolare la maschera lungo i bordi del soggetto per aumentare l'informazione e rendere la matte più precisa e accurata. ViTMatte è una rete neurale naturale in grado di estrarre un'alfa di migliore qualità a partire da una garbage matte (una maschera rough).[23] Sebbene ViTMatte funzioni meglio su immagini fisse e non abbia stabilità temporale, può comunque essere utile per ottenere maschere difficili, in particolare quelle con dettagli fini come capelli o peli.

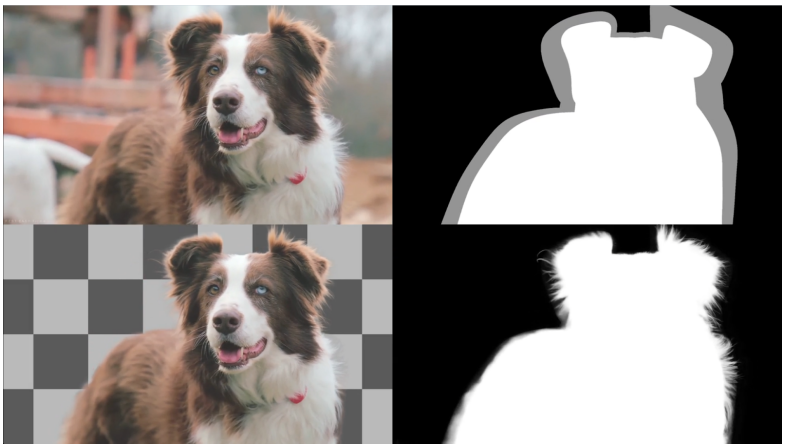

**Figure 3.12:** ViTMatte

#### Funzionamento

ViTMatte crea una matte di alta qualità da matte semplici in pochi clic, utilizzando una trimap per identificare i bordi e le aree semitrasparenti. Una trimap è un'immagine in scala di grigi che aiuta ViTMatte a sapere su cosa concentrarsi. Il colore nero significa trasparente, il bianco significa opaco e il grigio significa che quell'area "necessita di lavoro". Per crearla basta fare una Roto con due Bezier, una completamente opaca e l'altra con opacità ridotta, ad esempio 0.3. Esistono ormai dei nodi che calcolano in maniera automatica la trimap.
Grazie agli utenti rafaelperez e vinavfx è stata sviluppata una versione del ViTMatte integrabile e utilizzabile direttamente in Nuke, a partire dalla versione 13.2.[24]

### 3.3.4 Matte Assist ML

Il Matte Assist ML è un nuovo nodo introdotto nella versione 2024.5 di Silhouette. Con questo strumento è possibile estendere automaticamente una maschera per l'intera durata della sequenza, a partire da una o più maschere generate in uno o

più frame del video. Il Matte Assist ML usa il machine learning, la segmentazione a oggetti e la propagazione. Per creare una maschera con i bordi più naturali è possibile usare il Matte Assist in combinazione con Power Matte e Trimap.

### Funzionamento

Il funzionamento di questo nodo coinvolge diversi parametri: mask source, reference frames, propagation mode e start frame.
La mask source indica la sorgente della maschera, ossia la maschera fornita in input da estendere per la durata della sequenza. Si può selezionare Input Alpha se possiede una maschera creata con i nodi Mask ML, EZ Mask o Paint. Questa modalità è da utilizzare solo se si vuole segmentare un singolo oggetto. La modalità Roto invece permette di estendere una roto creata in precedenza. Ogni singola roto viene considerata un oggetto separato da segmentare.[25]

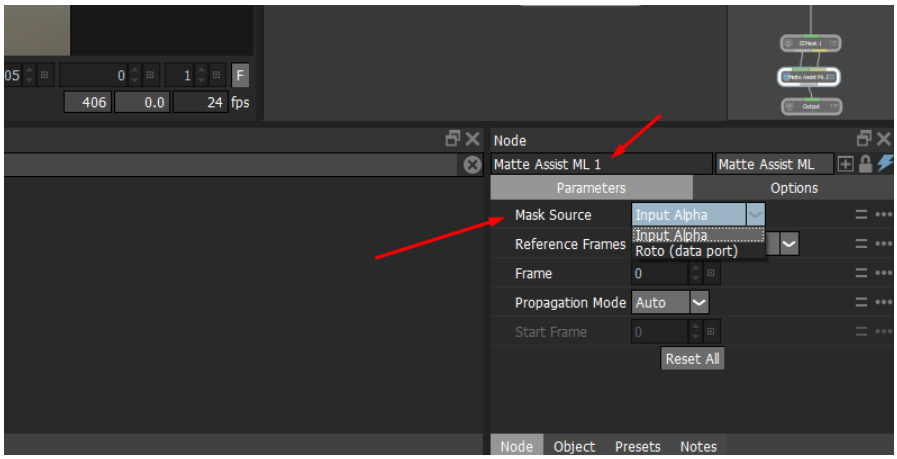


**Figure 3.13:** Matte Assist ML.
Fonte: `https://shorturl.at/T2Akq`

Se si seleziona la modalità Input Alpha nel Mask Source è possibile scegliere nel parametro "reference frames" tra Frame, Markers o Keyframes, come riferimento per la generazione della matte. Frame indica il frame di riferimento da cui generare la maschera, Markers permette di generarla a partire dai markers e Keyframes permette di generare la maschera sulla base del Paint o EZ Mask.
La Propagation Mode indica la modalità con cui propagare la maschera per la durata della sequenza. Può essere Auto se si desidera che parta dal primo frame oppure Custom se si vuole personalizzare il frame di partenza.
Il frame di partenza viene indicato nel parametro Start Frame.

### 3.3.5 EZ Mask

L'EZ Mask è un nodo introdotto nella versione 2024.5 di Silhouette.
Permette di creare una maschera in maniera interattiva ed è in grado di gestire i bordi con semi-trasparenza per mantenere dettaglio anche nei capelli, nel fumo o nei riflessi. Il nodo EZ Mask, iterativamente, stima i valori di trasparenza per ogni pixel dell'immagine basandosi su una piccola porzione del foreground e del background e sulle pennellate che disegna l'utente.[26]

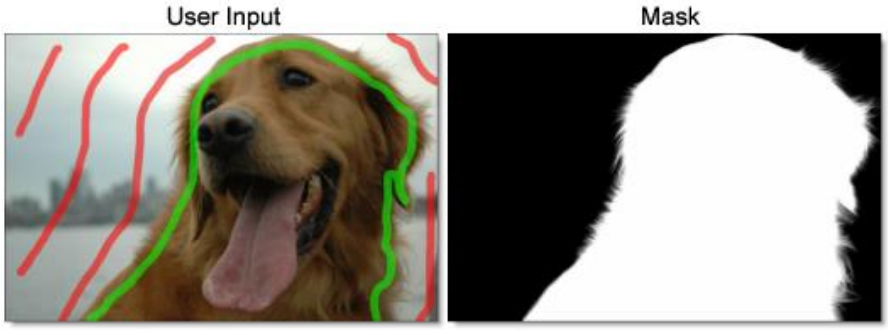


**Figure 3.14:** EZ Mask.
Fonte: `https://shorturl.at/7VNHv`

#### Funzionamento

Per la creazione della maschera ci sono diversi parametri e strumenti che si possono utilizzare.
L'EZ Mask crea una maschera utilizzando una trimap. Possono essere utilizzati due metodi di quest'ultima: Stroke e Filled.
La modalità Stroke permette di disegnare delle pennellate e richiede l'interazione dell'utente. Gli stroke possono essere di foreground e di background, indicando accuratamente cosa si intende segmentare e cosa no. Lo stroke di foreground deve essere vicino ai bordi del soggetto ma non troppo vicino all'edge. Se il foreground o il background hanno diversi colori, occorre disegnare degli stroke che coprano tutte queste variazioni di colore. Se si sono persi dei dettagli sui bordi è possibile disegnare delle linee con il pennello Unknown, per indicare che quell'area va trattata con semi trasparenza.
La modalità Filled, invece, permette di selezionare il foreground, il background e l'Unknown area. Prima di tutto è consigliabile disegnare con il Paint Unknown brush l'area lungo i bordi del soggetto da segmentare. Dopodiché, con il Paint Foreground brush e il Fill tool si deve selezionare l'area di foreground cliccandoci sopra. Infine, usando il Paint Background brush e il Fill tool bisogna fare lo stesso

con il background. In questo modo il foreground sarà dipinto di verde, l'Unknown area di blu e il background di rosso.
Oltre ai pennelli enunciati in precedenza e al Fill tool sono presenti anche il Paint Missing, che serve ad indicare le aree di dettaglio mancante (come per esempio i capelli) e l'Eraser per cancellare gli strokes esistenti.
I brush hanno i loro parametri: size e paint overlay opacity. Il primo indica la dimensione del pennello e il secondo l'opacità con cui si desidera dipingere.
Sono presenti anche delle opzioni di processamento che includono il Deartifact, utile per limitare artefatti visivi se presenti nell'immagine e la sensibilità del Missing Brush, utile per gestirne il valore. Maggiore è il suo valore maggiore sarà il dettaglio recuperato.

### 3.3.6 Mask ML

Il Mask ML è un nodo introdotto nella versione 2024.5 di Silhouette.
Attraverso lo strumento punta e clicca, permette di selezionare l'area di foreground che si desidera segmentare. In automatico verrà creata una maschera per il foreground con un solo click. Questo strumento è molto utile in combinazione con il Matte Assist ML.[27]

**Figure 3.15:** Mask ML.
Fonte: `https://shorturl.at/PqeKo`

#### Funzionamento

Ci sono alcuni strumenti utilizzabili con questo nodo come per esempio la possibilità di scegliere se aggiungere alla Matte di foreground o sottrarre un'area dell'immagine

alla maschera. Inoltre, è possibile disegnare un rettangolo per selezionare un oggetto intero oppure gestire l'opacità del punto con cui si clicca.

### 3.3.7 CopyCat

CopyCat è un tool di machine learning presente su Nuke dalla versione Nuke 15.1. Esso copia gli effetti applicati ad alcuni frame di un video (come per esempio la creazione di maschere, il lavoro di beauty o la rimozione della sfocatura) e li replica per i restanti frame della sequenza. L'output di CopyCat è una rete addestrata che viene salvata in un file .cat, pronto per applicare gli effetti tramite un nodo di Inference.
L'Inference è la fase successiva al training di un modello di machine learning. Durante questa fase, il modello lavora su nuovi dati per produrre un output sulla base di quanto appreso durante il training.

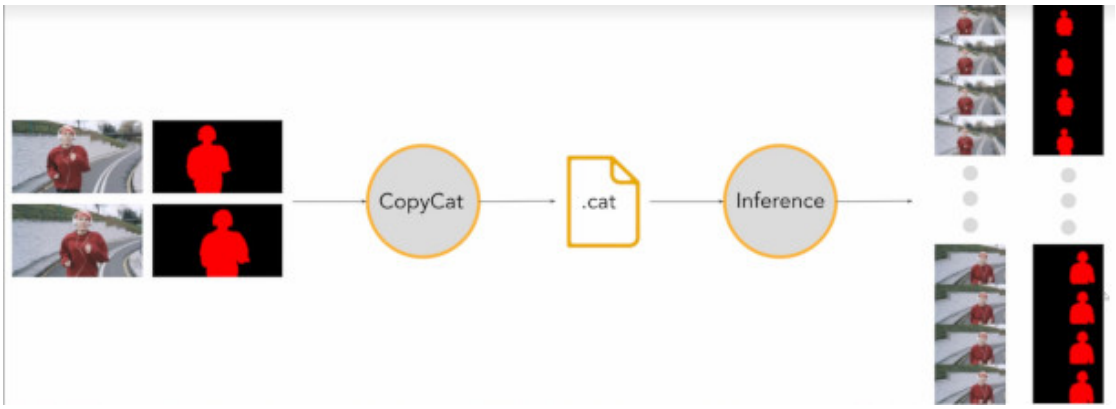


**Figure 3.16:** CopyCat.
Fonte: `https://shorturl.at/DLXIp`

Prima di descrivere il funzionamento di CopyCat, sono necessarie alcune definizioni: Input, Ground Truth, Model Weights e Loss.
L'input è l'immagine originale, la ground truth è l'immagine dopo l'applicazione dell'effetto, i model weights sono i pesi/parametri che costituiscono il modello e si aggiornano ad ogni step per raggiungere il risultato, la loss è la distanza tra output e ground truth.[28]
Il training loop di CopyCat prevede che durante l'addestramento vengano forniti dei frames (ground truth frames) sui quali sia stato applicato l'effetto che si desidera replicare per il resto della sequenza. Durante gli steps di training, CopyCat cercherà di replicare l'effetto sui frames restanti (input frames). Inizialmente l'output sarà di scarsa qualità e con il passare del tempo il modello affinerà sempre più il suo risultato, cercando di avvicinare l'output il più possibile alla ground truth.

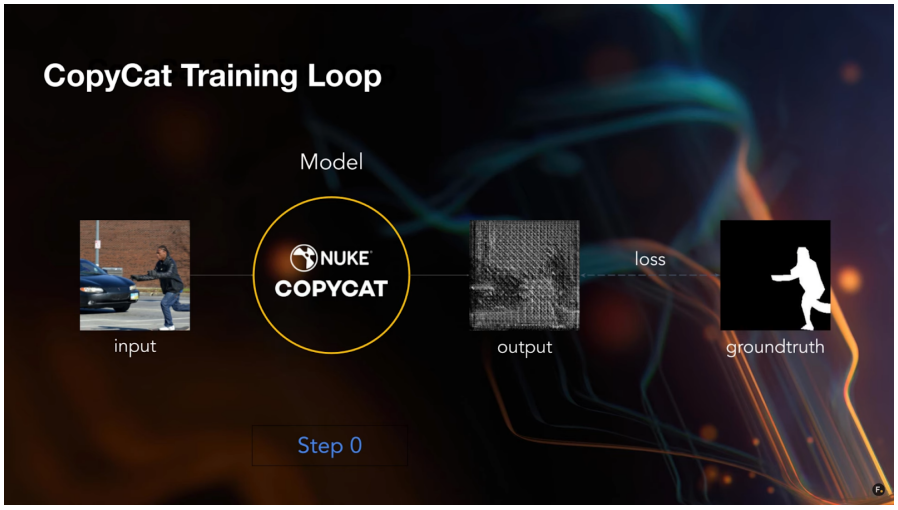


**Figure 3.17:** CopyCat training loop.
Fonte: `https://shorturl.at/BnV9S`

Il processo iniziale che prevede la scelta dei frames di ground truth da utilizzare è molto importante. I frames su cui effettuare il training di CopyCat sono chiamati frames rappresentativi. Un frame è rappresentativo della sequenza quando rappresenta al meglio l'intera sequenza. Tra i frames di una sequenza ci possono essere variazioni che vanno tenute in considerazione. Le principali variazioni sono dovute al movimento di camera o del soggetto, le variazioni di illuminazione e i cambi di fuoco. Durante il training bisogna dunque includere alcuni frames che contengono queste tipologie di cambiamenti, in modo tale da allenare l'algoritmo a gestirle.

CopyCat durante il training utilizza il Random Cropping, selezionando randomicamente delle porzioni dei frame da analizzare.
Il Crop è un ritaglio dell'immagine, una porzione. Fornire come Ground Truth anche dei frames ritagliati attorno all'area d'interesse incrementa il numero di volte che CopyCat analizza quell'area, aiutando a ridurre i tempi di calcolo e a migliorare i risultati in output. Bisogna ricordarsi anche di fornire almeno due frames non ritagliati (dimensione del fotogramma originale) affinché CopyCat abbia anche una "visione d'insieme" sui fotogrammi da lavorare. Fornire dei frames ritagliati può tornare utile anche quando si ha la necessità di "avvisare" CopyCat che la porzione del frame inclusa nel crop non deve essere modificata (negative prompt). Per esempio, se all'interno di una sequenza ci sono due persone con del sangue in volto e si vuole rimuovere il sangue solo da una delle due, per fare in modo che CopyCat non lo rimuova anche dall'altra si possono fornire dei frames ritagliati del volto che deve rimanere invariato.


**Figure 3.18:** CopyCat bruise removal.
Fonte: `https://shorturl.at/BnV9S`

Nel momento in cui si lavora sui frames da utilizzare come ground truth è possibile sfruttare a proprio vantaggio una tecnica chiamata Data Augmentation. L'augmentation è la pratica con la quale si aumentano le informazioni e i dati da fornire all'algoritmo, a partire dai frames già utilizzati come ground truth. Si possono per esempio apportare alcune semplici modifiche di transform augmentation (rotation, flip) e color augmentation (grade, contrast, saturation). L'augmentation può servire per simulare le variazioni di movimento, illuminazione e fuoco senza dover lavorare su frames aggiuntivi. Per esempio, aggiungendo un nodo di Grade a un frame e aumentando il valore della gamma, si andrà a schiarire l'immagine per poter simulare una condizione di maggiore luce all'interno della scena.
Altre due considerazioni da fare riguardano i Superwhite Pixels e il Colorspace. Il machine learning si aspetta sempre valori tra 0 e 1. Avere dei valori molto vicini a 1 non causa troppi problemi ma avere dei valori super wide (5, 10+) può risultare problematico durante il training. Si hanno due soluzioni per gestire i pixel superwhite: applicare un nodo di clamp o spostarsi in uno spazio colore logaritmico. Il clamp permette di convertire in 1 tutti i valori che lo eccedono. Lo spazio logaritmico, invece, permette di comprimere i valori in un range ristretto. Ciò può significare, di contro, un tempo necessario maggiore per CopyCat nel distinguere questi valori e, dunque, un maggior tempo di calcolo.
Per quanto riguarda lo spazio colore bisogna ricordarsi di allenare CopyCat nello stesso spazio colore in cui si vuole fare Inference.

### Funzionamento

CopyCat è un tool molto potente quanto delicato e per poterne sfruttare al meglio le potenzialità bisogna conoscere ogni parametro.
Le epochs sono la metrica standard nel machine learning. Non è possibile impostare direttamente il numero di steps di training desiderati ma è possibile arrivarci variando il numero di epochs. Il numero totale degli steps è dato dal numero di epoches per la la grandezza del dataset sulla grandezza delle batch (3.1), cioè dal numero dei frames di ground truth sul numero di input crops processati ad ogni step.

$$\text{Total Steps} = \frac{\text{Epochs} \times \text{Data Set}}{\text{Batch Size}} \tag{3.1}$$

Aumentando il numero di epochs aumenterà la qualità del risultato finale, al costo di un maggiore tempo di calcolo.
La batch size può essere impostata su small, medium o large ed indica il numero di random crops che viene calcolato ad ogni step del training. Per un training generalizzato si consiglia di utilizzare una batch size elevata.

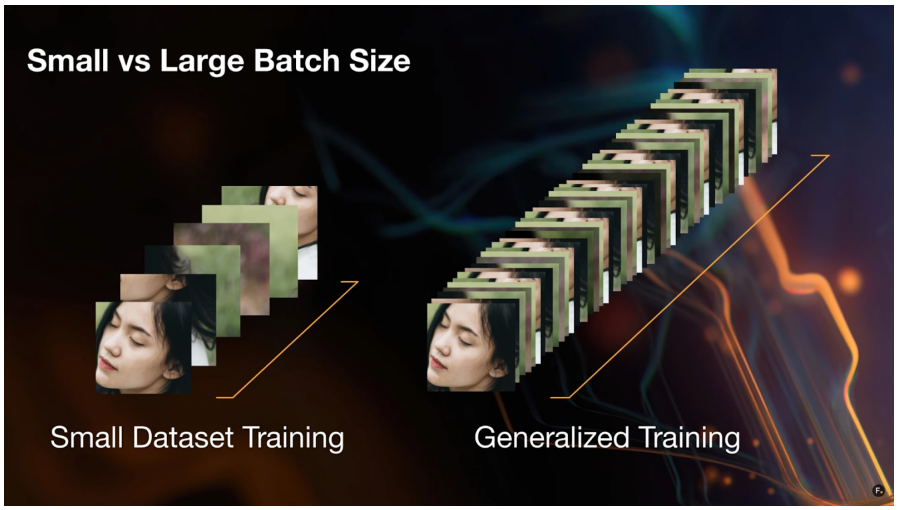


**Figure 3.19:** CopyCat batch size.
Fonte: `https://shorturl.at/BnV9S`

Gli initial weights sono un ottimo punto di partenza per il training perché sono dei pesi che vengono applicati al modello sulla base di pre-addestramenti. Attualmente CopyCat dispone di tre pre-trained weights: deblur, upscale e human matting. Il primo è utile nel caso in cui l'obiettivo sia rimuovere la sfocatura, il secondo nel caso in cui si desideri fare un upscale dell'immagine e l'ultimo si adatta perfettamente alla generazione di maschere di persone. È possibile utilizzare, inoltre, un checkpoint

personale basato su un training custom svolto in precedenza.
Un altro parametro che può essere impostato è la model size, cioè la grandezza del modello che può essere small, medium o large. Una size impostata su small è adatta ad operazioni pixel-based come per esempio il deblur mentre una size impostata su large si abbina perfettamente a task di tipo semantico come ad esempio la segmentazione.
È possibile, infine, selezionare la crop size, ossia la dimensione dei crop randomici che vengono analizzati durante ogni step del training.
In base alla documentazione di CopyCat, la crop size va impostata in maniera simile alla model size. È consigliabile dunque impostarla su small per le tasks pixel-based e su large per le operazioni che richiedono maggiore contesto globale. Ne conviene che maggiore è la dimensione impostata, maggiori saranno il tempo di calcolo e la durata del training.
Una volta avviato il training è molto importante controllare il suo andamento, descritto con una funzione di loss. La loss (perdita) è la metrica per la performance del modello.

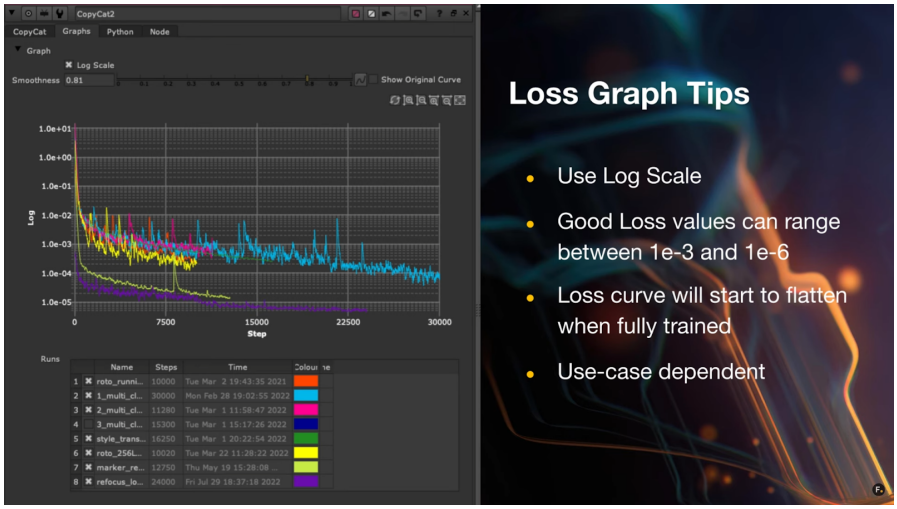


**Figure 3.20:** CopyCat loss.
Fonte: `https://shorturl.at/BnV9S`

Sull'asse delle ascisse è presente il numero di steps svolti durante il training, mentre sulle ordinate il valore della loss. Valori buoni di perdita sono nel range tra 1e-3 e 1e-6. La curva, man mano che il training si completa, si appiattisce. Se la curva inizia subito ad appiattirsi su valori alti, significa che il modello non è in grado di lavorare correttamente; probabilmente c'è un problema con i settings in input come per esempio un mismatch tra frames in input e ground truth.
Un ulteriore metodo per visualizzare l'andamento del training, da un punto di vista

qualitativo, è il contact sheet. Questo è una griglia, ove le colonne rappresentano da sinistra a destra l'input, la ground truth e l'output.[29]
Tra i molteplici casi applicativi di CopyCat troviamo anche la segmentazione.
Il grafo a nodi generico per un training basato su roto prevede l'impiego di diversi nodi: read, frame range, frame hold, roto, append clip, shuffle, remove e copycat.

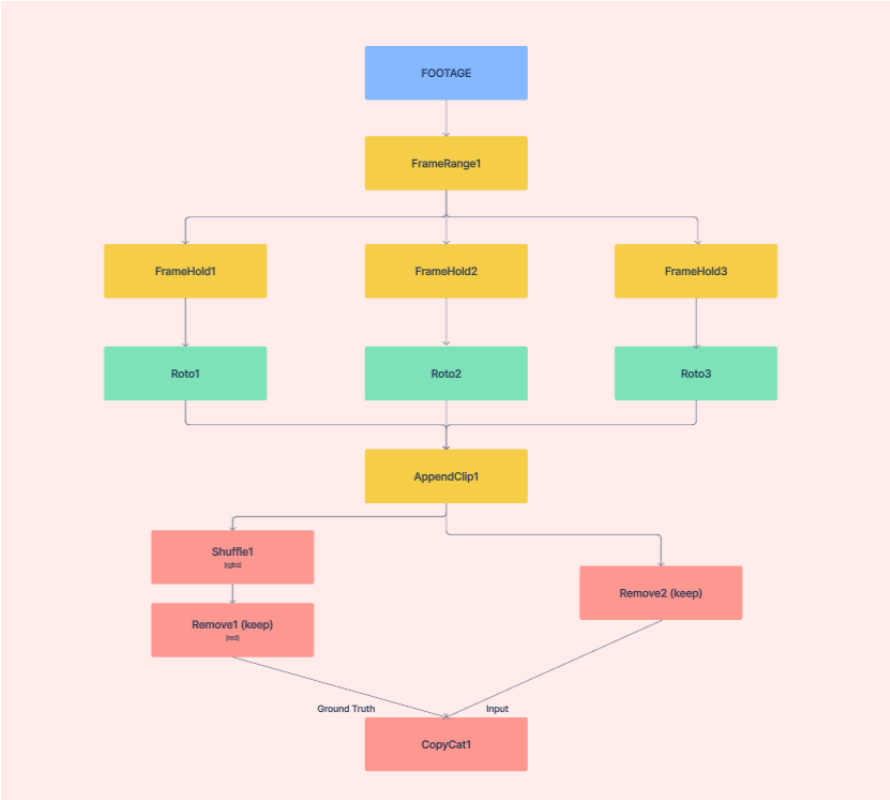


**Figure 3.21:** CopyCat x Roto

Il nodo di Read permette di leggere un file all'interno di Nuke. Solitamente gli shots vengono lavorati come sequenze di immagini .exr o .dpx. Una sequenza di immagini è composta da ogni frame in un file separato dagli altri. Nel momento in cui si visualizza la sequenza non si nota questa separazione e il video scorrerà fluidamente. Si lavora con sequenze di immagini e non video perché, qualora si dovesse agire su singoli frames, si ha una maggiore flessibilità e precisione. Questo vale soprattutto nel momento in cui si debba integrare degli elementi in CG che vengono renderizzati frame by frame.
Il nodo FrameRange indica ai nodi successivi che ogni frame è da considerare singolarmente. Vanno dunque impostati i valori dei parametri fist_frame e last_frame su 1.

Al FrameRange sono collegati tanti nodi FrameHold quanti sono i frames rappresentativi della sequenza, con l'accortezza di averne circa dieci ogni cento frames della sequenza. Il FrameHold fa in modo che il frame su cui è impostato rimanga visibile e bloccato per l'intera durata della sequenza, come se ci fosse un freeze.
Ad ogni FrameHold è collegato un nodo di Roto. Con questo nodo si crea, manualmente o con l'aiuto di altri strumenti AI come il Segment Anything, la maschera del soggetto che si desidera scontornare.
Dopodiché tutti gli output dei nodi di Roto confluiscono nel nodo AppendClip, il quale raccoglie tutti i dati.
Ora il grafo si divide in due diramazioni: a sinistra si pre-processa il ground truth e a destra l'input. Per preparare i frames da utilizzare come ground truth è necessario utilizzare due nodi: shuffle e remove. Lo shuffle serve per inserire il canale alpha, ossia quello contenente la roto, nei canali rgba. Il remove, impostato su "keep red", permette di rimuovere da ogni frame di ground truth i canali gba mantenendo solo il red. A questo punto si avrà la roto nel canale rosso e tutti gli altri saranno vuoti, evitandone il loro processamento e quindi velocizzando il tempo di calcolo. Nella diramazione di destra, invece, sarà presente solo un remove impostato su "keep" con il quale si mantengono solo i canali rgb, escludendo l'alpha.
Ground truth e input sono poi collegati al nodo CopyCat.
Se si sono svolte correttamente queste operazioni, di fianco alla voce "Channels" si avrà rgb -> rgba.red e al di sotto sarà presente il numero di Batch Size scelto e il numero totale di steps del training.
Nel momento in cui il training è terminato autonomamente o manualmente (qualora i valori di loss fossero accettabili), si deve cliccare sul pulsante "Create Inference". Così facendo verrà creato il nodo Inference che dovrà essere collegato al nodo di Read iniziale per applicare l'effetto a tutta la sequenza.

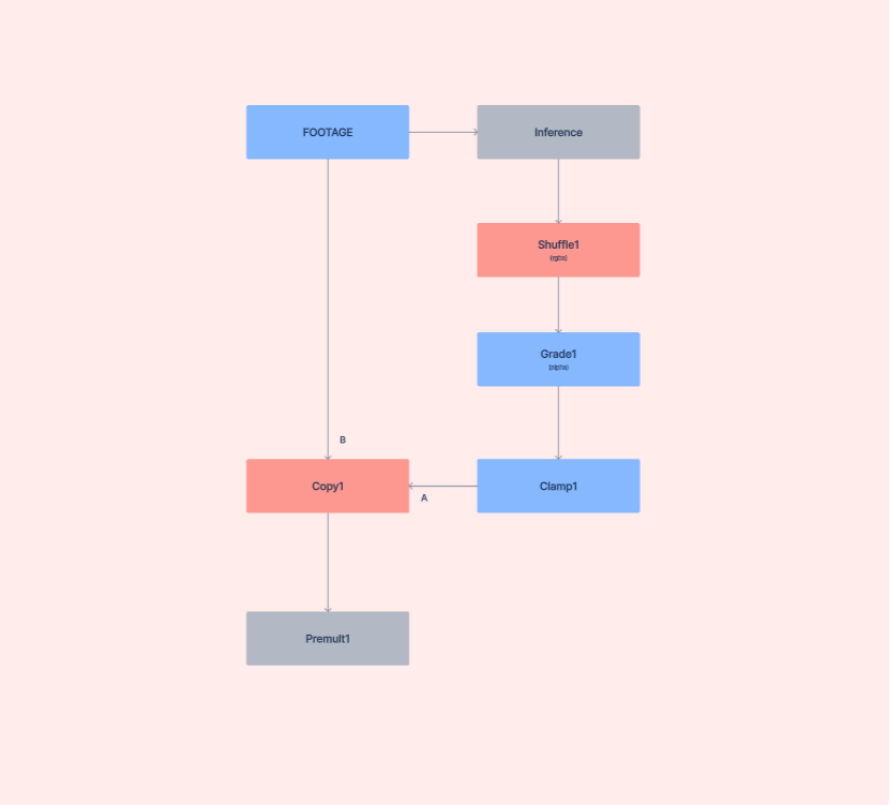


**Figure 3.22:** CopyCat Inference

## 3.4 Risultati ottenuti

Sono stati selezionati alcuni shot di produzione passati, su cui EDI aveva già lavorato, in modo tale da avere tutte le informazioni necessarie da utilizzare come metodo di paragone su tempistiche di lavorazione e qualità del risultato ottenuto. In seguito ad una riunione iniziale con il mio tutor aziendale, nonché Head of RnD ed Head of FX Daniele De Maio, l'Head of 2D Gabriele Motta e il Lead Compositor Francesco Lorussi, è stata effettuata una selezione degli shot. Il metodo di selezione ha tenuto conto di alcune varianti tipiche nelle sequenze video: distanza del soggetto dalla camera, quantità di movimento, occlusione e distanza euclidea nello spazio RGB tra i bordi del soggetto e lo sfondo.
Gli shot che sono stati scelti per la fase di test appartengono alle seguenti produzioni: *Romulus II - La guerra per Roma (2022)*, *Il Ritorno di Casanova (2023)* e *Finalmente l'alba (2023)*.

### 3.4.1 Romulus II - La guerra per Roma

Romulus è una serie televisiva italiana del 2020 diretta da Matteo Rovere. La seconda stagione è stata rilasciata nel 2022. Narra la storia di Romolo e del fratello gemello Remo, nell'VIII secolo a.C. raccontata attraverso le vicende di tre giovani segnate da morte, solitudine e violenza, tra i movimenti delle trenta tribù latine del Lazio che portarono alla fondazione di Roma.
Sono stati selezionati quattro shot differenti che variano in termini di distanza dei soggetti dalla camera, quantità di movimento della camera o dei soggetti e quantità di occlusione dei soggetti da scontornare.
Il primo shot è una sequenza d'azione e combattimento ed essendo molto dinamica è presente molto motion blur nei soggetti vicini alla camera. Il motion blur aumenta la difficoltà nel mascherare i soggetti perché introduce una sfocatura pronunciata nei bordi di questi ultimi, implicando una lavorazione maggiore da parte dei roto artists o dei compositors.

**Figure 3.23:** Primo shot: plate

Attraverso il tool Segment Anything sono stati creati sette punti per la mask che si desiderava generare e sono stati animati manualmente per la durata della sequenza. Per aggiungere un punto è necessario cliccare su "Add point" e per animarlo bisogna premere sull'icona a fianco alle coordinate "y" del punto.

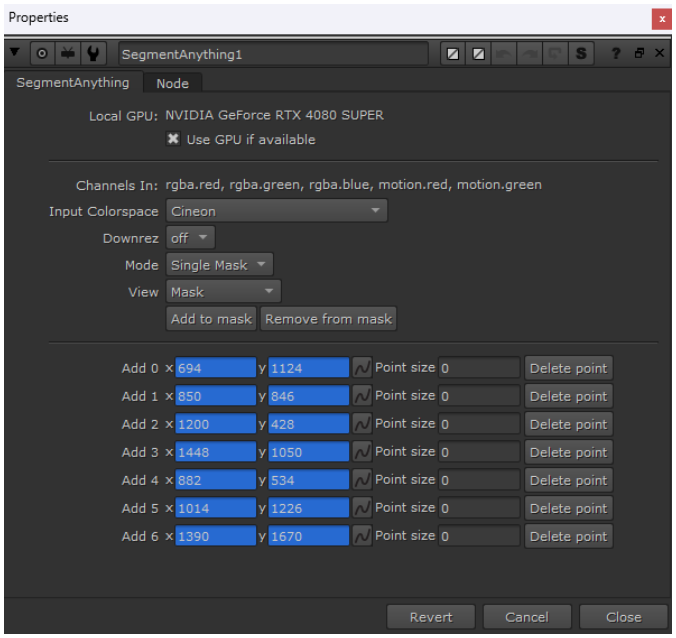


**Figure 3.24:** Primo shot: properties SAM

**Figure 3.25:** Primo shot: SAM punti animati

Così facendo si è ottenuto il canale alpha del soggetto.

**Figure 3.26:** Primo shot: SAM alpha

Notando i bordi troppo netti del soggetto, in coda al nodo Segment Anything viene aggiunto il nodo ViTMatte che prende in input un canale alpha e, con una trimap, migliora i bordi per avere una soft matte più precisa.

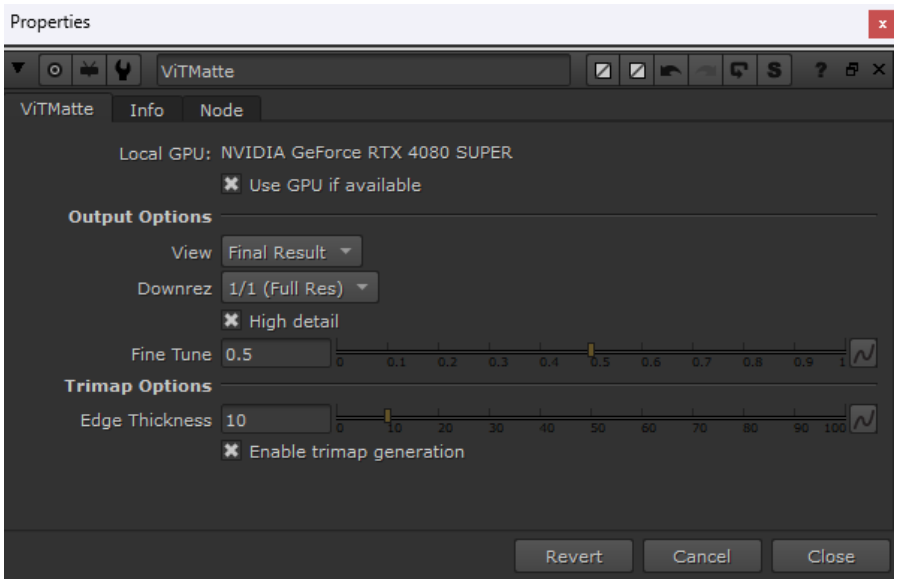


**Figure 3.27:** Primo shot: ViTMatte parametri

I parametri del ViTMatte sono: view, downrez, high detail, fine tune, edge thickness ed enable trimap generation. La view è stata impostata su "Final Result", non è

stato applicato ridimensionamento ed è stata attivata la spunta su High Detail. Il Fine Tune e l'Edge Thickness permettono di intervenire sui bordi e l'ultimo parametro attiva la generazione automatica di una trimap.

**Figure 3.28:** Primo shot: ViTMatte alpha

Notando che il canale alpha era abbastanza "sporco" è stato applicato un nodo di Grade ristretto al canale alpha, incrementando il valore del "blackpoint".

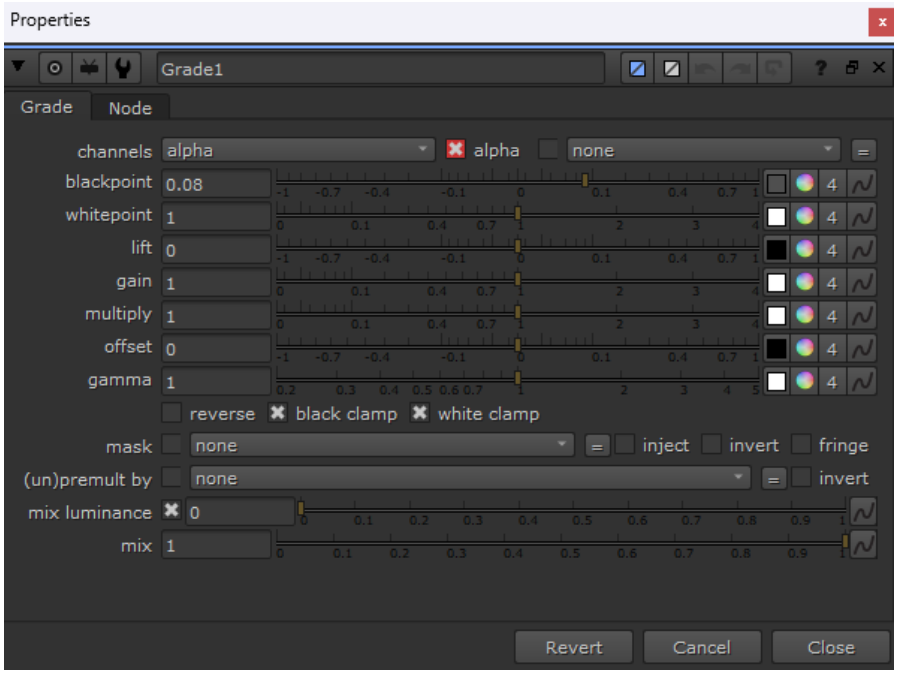


**Figure 3.29:** Primo shot: Grade parametri

**Figure 3.30:** Primo shot: ViTMatte alpha + Grade

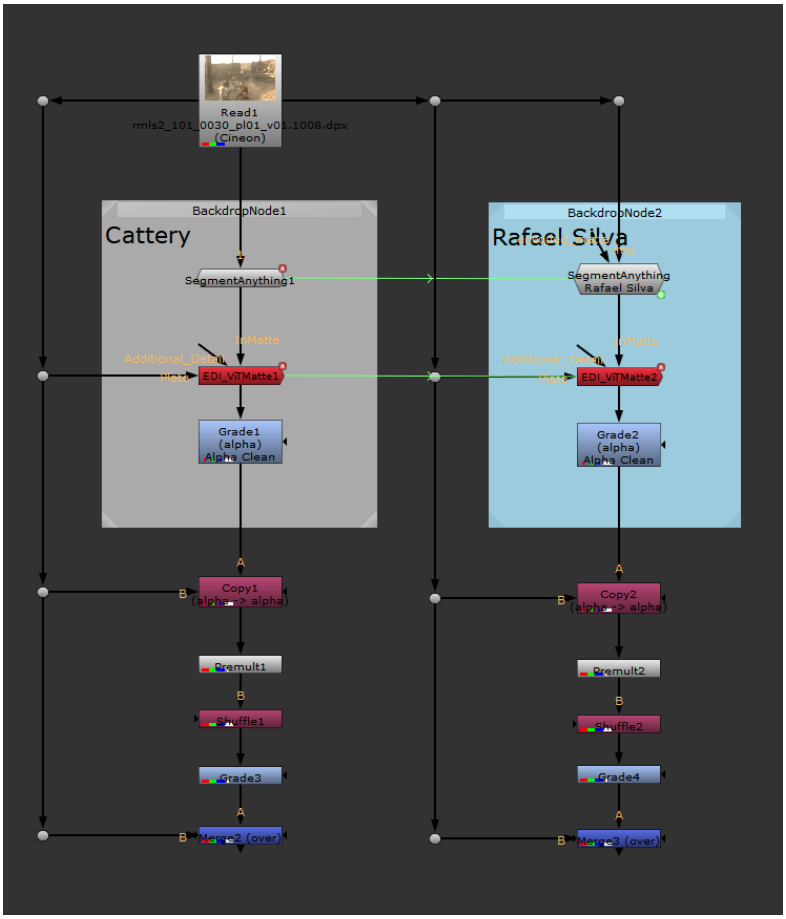


**Figure 3.31:** Primo shot: Nuke node graph

In seguito è stato testato, con i suoi limiti di usabilità, Segment Anything 2 su ComfyUI. Il risultato ottenuto è più robusto durante la sequenza video e genera meno flicker (sfarfallio).


**Figure 3.32:** Primo shot: SAM2 alpha overlay

Infine è stato testato Silhouette con gli strumenti EZ Mask, Mask ML e Matte Assist ML.

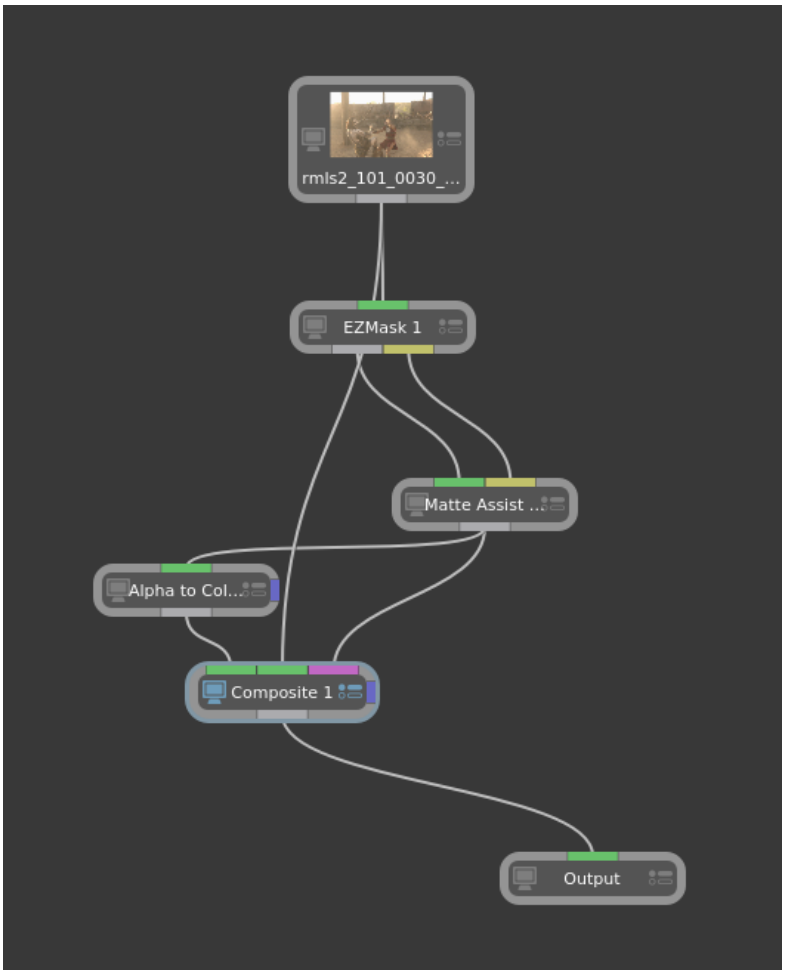


**Figure 3.33:** Primo shot: Silhouette node graph

È stato utilizzato EZ Mask per generare la maschera in un frame della sequenza e Matte Assist ML per estenderla lungo l'intera durata della clip video.

**Figure 3.34:** Primo shot: EZ Mask alpha

Il parametro "Deartifact" del EZ Mask è stato impostato sul valore 0 e l'"Accuracy" su Normal.

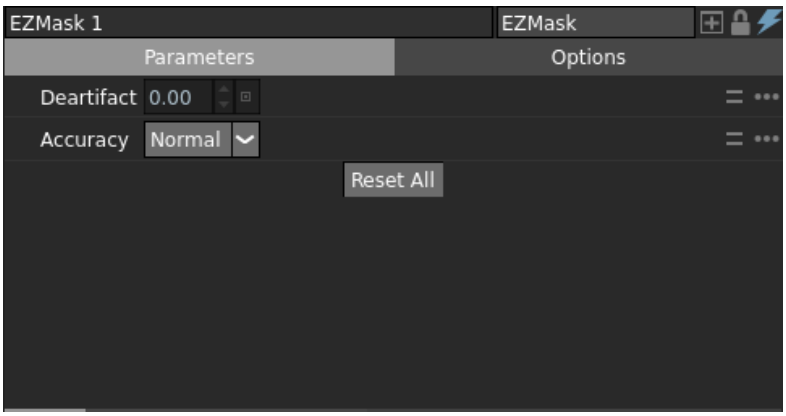


**Figure 3.35:** Primo shot: EZ Mask parametri

I parametri del Matte Assist ML sono stati impostati come segue: la "Mask source" su Input Alpha, i "Reference Frames" su Keyframes e la "Propagation Mode" su Auto.

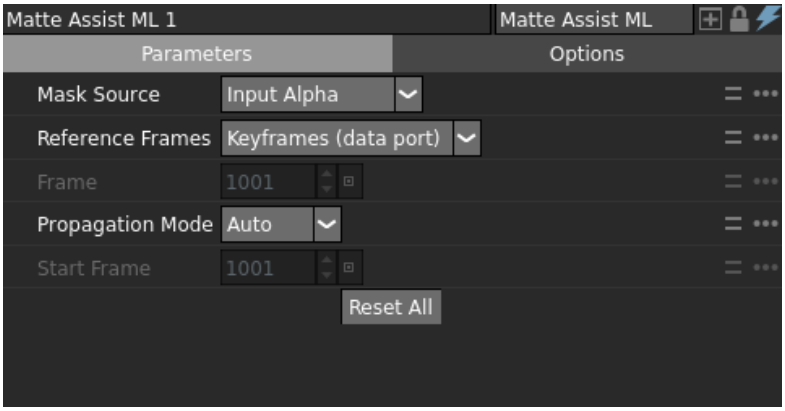


**Figure 3.36:** Primo shot: Matte Assist ML parametri


**Figure 3.37:** Primo shot: Silhouette alpha overlay

Il secondo shot ritrae una folla di persone attorno ad un falò. L'inquadratura è un campo largo con un movimento di camera a uscire. L'obiettivo è generare una maschera alfa della folla. La complessità di questa operazione risiede nella distanza dei soggetti e nella vicinanza dei colori tra i soggetti e il background.

**Figure 3.38:** Secondo shot: plate

Per raggiungere il risultato desiderato è stato utilizzato CopyCat, al fine di diminuire il tempo di lavoro lavorando solo su alcuni frames della sequenza e avviando il training dello strumento di machine learning.

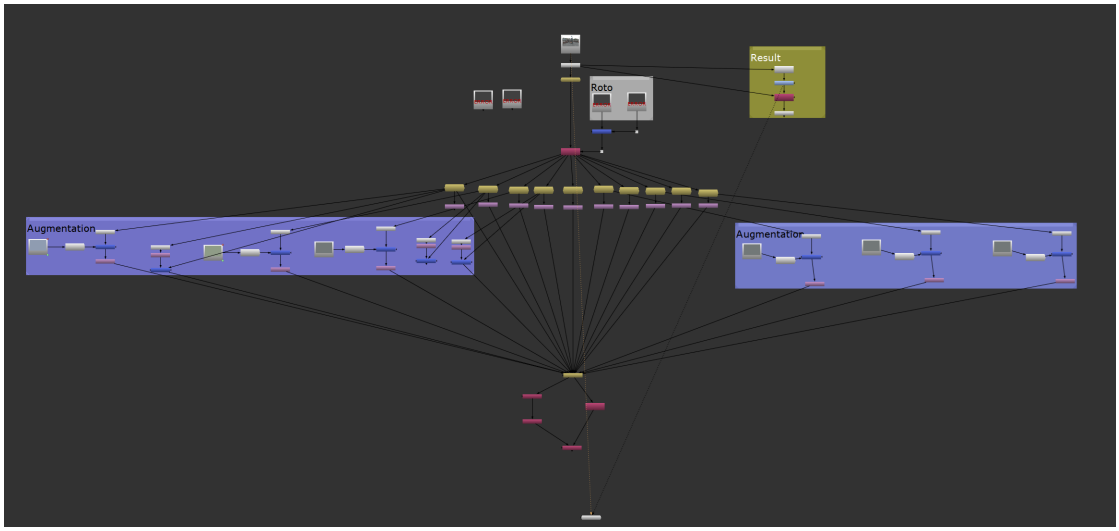


**Figure 3.39:** Secondo shot: node graph Nuke

Sono stati fatti 100.000 steps di training, gl initial weights "Human Matting", batch size "8" e crop size "256".

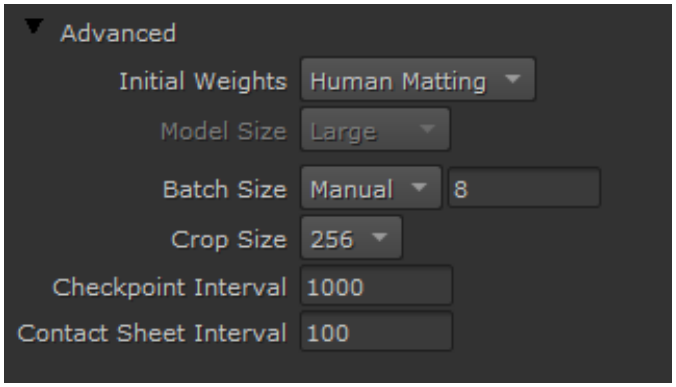


**Figure 3.40:** Secondo shot: parametri CopyCat

Il dataset di ground truth è stato creato con 10 frames interi, 10 frames ritagliati e 8 frames di data augmentation in cui è stato simulato del rumore digitale e degli sfondi di colore diverso.
Inizialmente sono state testate delle maschere rough come frames di ground truth e il risultato era già notevole, sebbene poco accurato e con tanti sfarfallii della maschera.

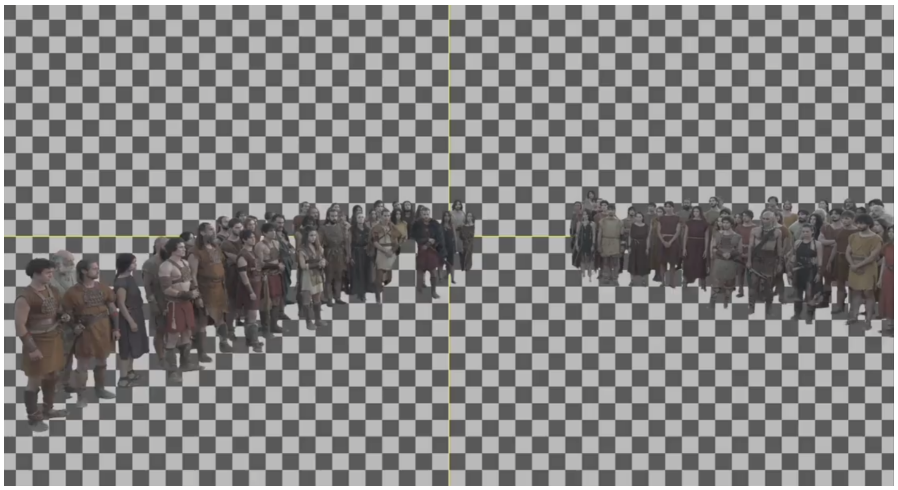

**Figure 3.41:** Secondo shot: CopyCat ground truth rough

In seguito sono state realizzate delle roto precise ed accurate e sono state utilizzate come ground truth per un nuovo training del modello, ottenendo un risultato nettamente migliore. Questo evidenzia l'importanza di un ground truth di qualità.

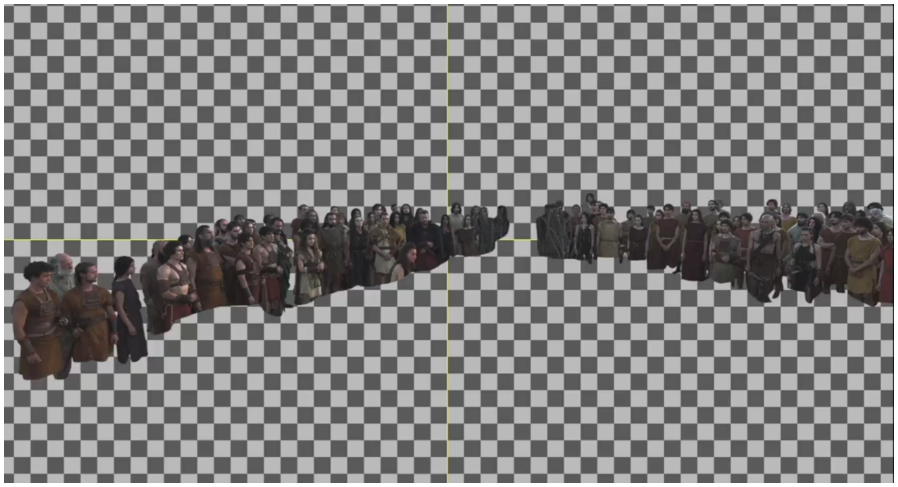

**Figure 3.42:** Secondo shot: CopyCat ground truth preciso

Il terzo shot è un medium close up del protagonista. La complessità di questo shot sta nell'ottenere il dettaglio sui bordi (capelli e barba) e recuperare le informazioni sui capelli dove il cielo è bianco. In questo caso, poiché lo shot è stato girato con una camera ad ampia gamma dinamica, le informazioni sulle "alte luci" sono recuperabili.

**Figure 3.43:** Terzo shot: plate

È stato testato lo strumento Segment Anything con il ViTMatte per recuperare dettaglio negli edges ed è stato ottenuto un buon risultato, nonostante vi fossero

dei flickers introdotti dal ViTMatte.

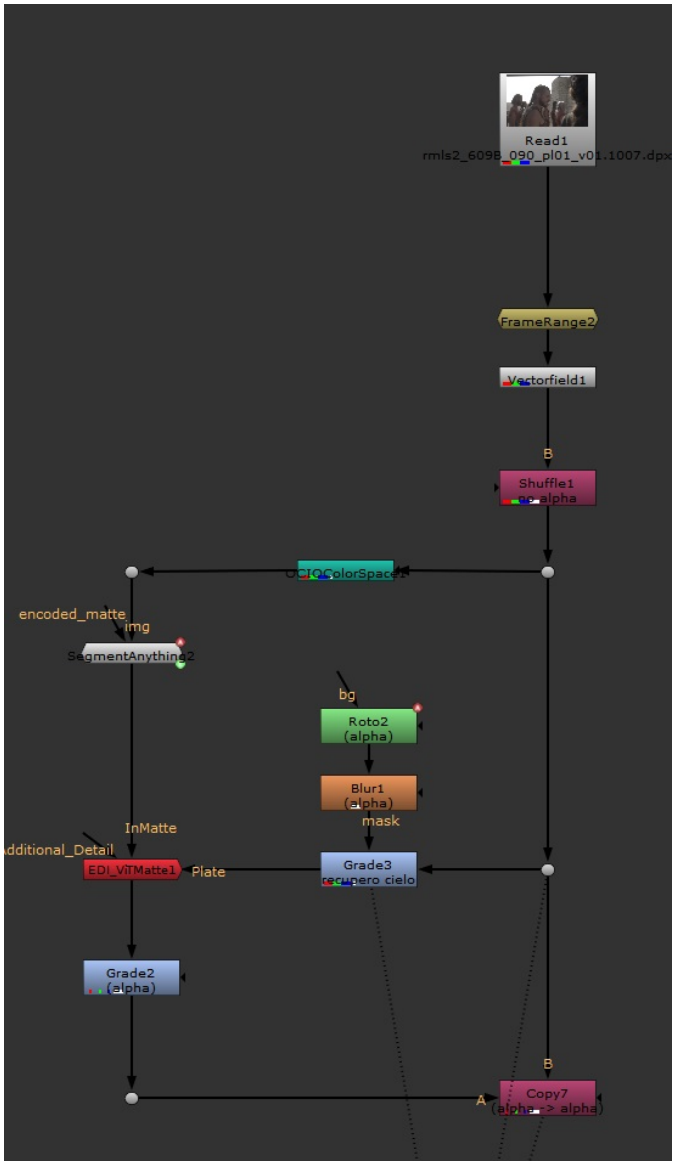


**Figure 3.44:** Terzo shot: node graph Nuke

Inizialmente, con il nodo Vectorfield, è stata applicata la LUT allo shot. La LUT (Look Up Table) è una tabella che mappa i colori di un'immagine sorgente in nuovi colori. Questo è servito per aumentare il contrasto dell'immagine ed evitare di fornire al Segment Anything un'immagine con un profilo colore flat. L'OCIOColorSpace permette di convertire lo spazio colore dell'immagine e il SegmentAnything di generare la maschera del soggetto. Con il ViTMatte si è cercato di recuperare

informazioni lungo i bordi del soggetto e con un grade mascherato sul cielo con una Roto si è abbassato il "gain" e la "gamma" per recuperare informazione nei bianchi. Questo è servito per ottenere un canale alpha finale che comprendesse anche i capelli che hanno come background il cielo bianco.

**Figure 3.45:** Terzo shot: SAM + ViTMatte alpha overlay

Il quarto shot ritrae quattro personaggi di spalle su un background sfocato, tipico per un replacement con matte painting.

**Figure 3.46:** Quarto shot: plate

Anche in questo caso sono stati utilizzati il Segment Anything e il ViTMatte e l'output è soddisfacente tranne una perdita di dettaglio tra i capelli del terzo personaggio da sinistra e degli sfarfallii delle maschere in alcuni momenti della

sequenza.

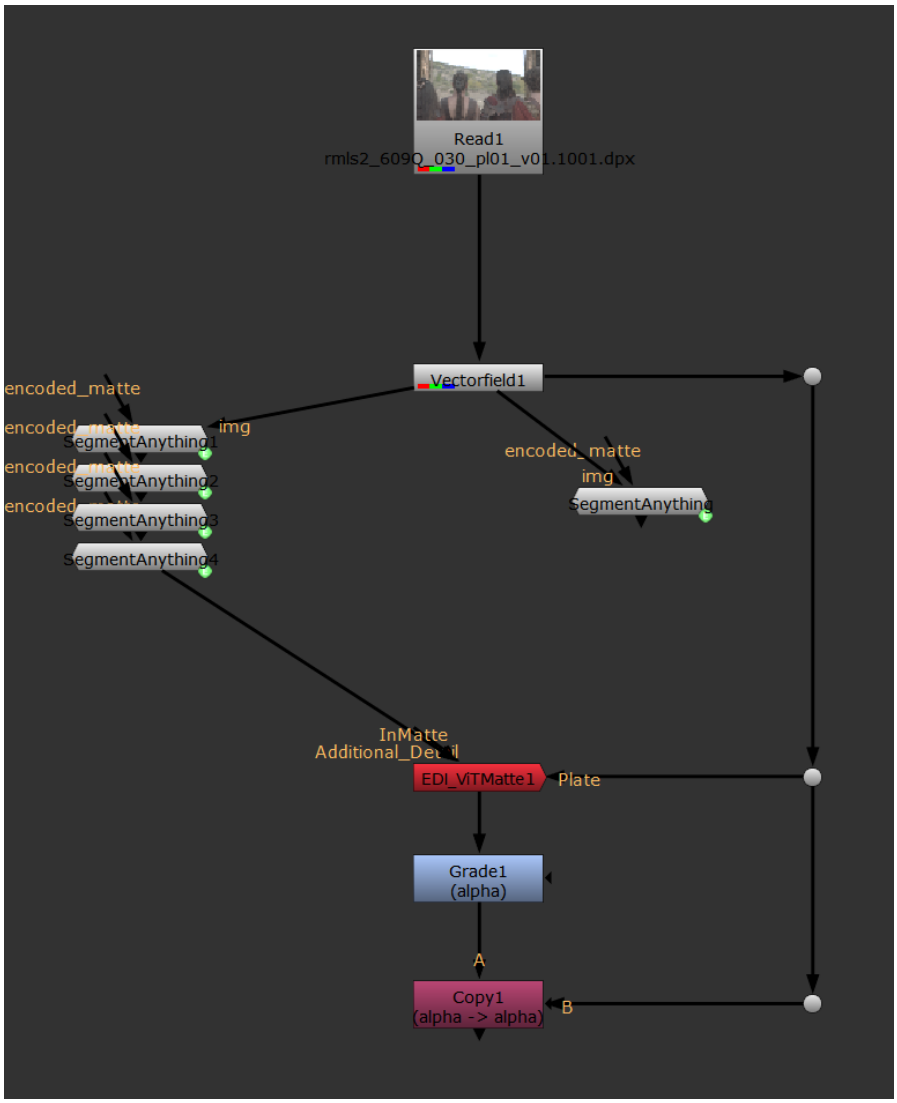


**Figure 3.47:** Quarto shot: node graph

Per raggiungere il risultato è stato utilizzato il nodo di Vectorfield per applicare la LUT e quattro nodi SegmentAnything in cascata. È emerso che, nel caso in cui si debba segmentare più persone all'interno della stessa inquadratura, conviene utilizzare un nodo di SAM per ognuna di esse. Dopo la segmentazione è stato utilizzato il ViTMatte per recuperare dettaglio sui bordi dei soggetti e un nodo di Grade per aggiustare il canale alpha.

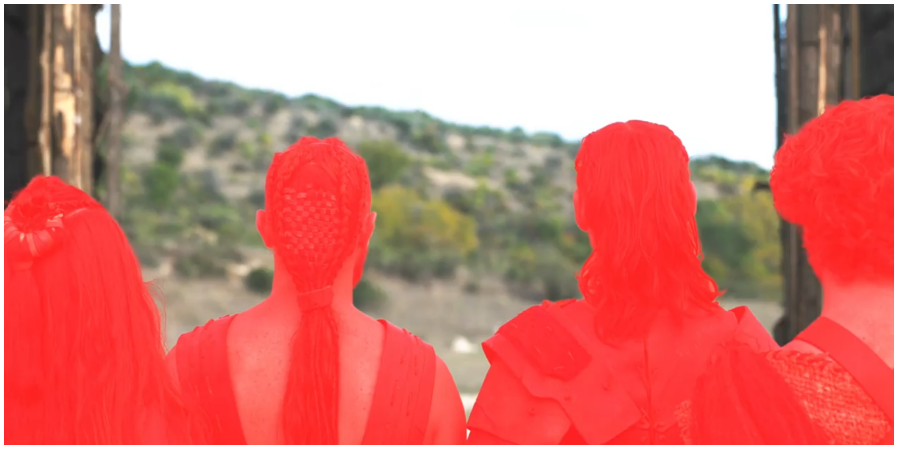

**Figure 3.48:** Quarto shot: SAM + ViTMatte alpha overlay

### 3.4.2 Il Ritorno di Casanova

Il ritorno di Casanova è un film del 2023 diretto da Gabriele Salvatores. La storia narra le vicende di Leo Bernardi, un celebre regista cinematografico impegnato nel montaggio del suo ultimo film "Il ritorno di Casanova". Tuttavia, una volta montata la scena iniziale, il regista inizia a disinteressarsi al progetto.
Lo shot testato ritrae un medium close up frontale del protagonista interpretato da Toni Servillo. La difficoltà dello shot risiede nel riuscire ad avere una maschera che mantenga dettaglio nei bordi e nell'area in cui il cielo è bianco ed è molto vicino in termini di colore ai capelli dell'attore.

**Figure 3.49:** Quinto shot: plate

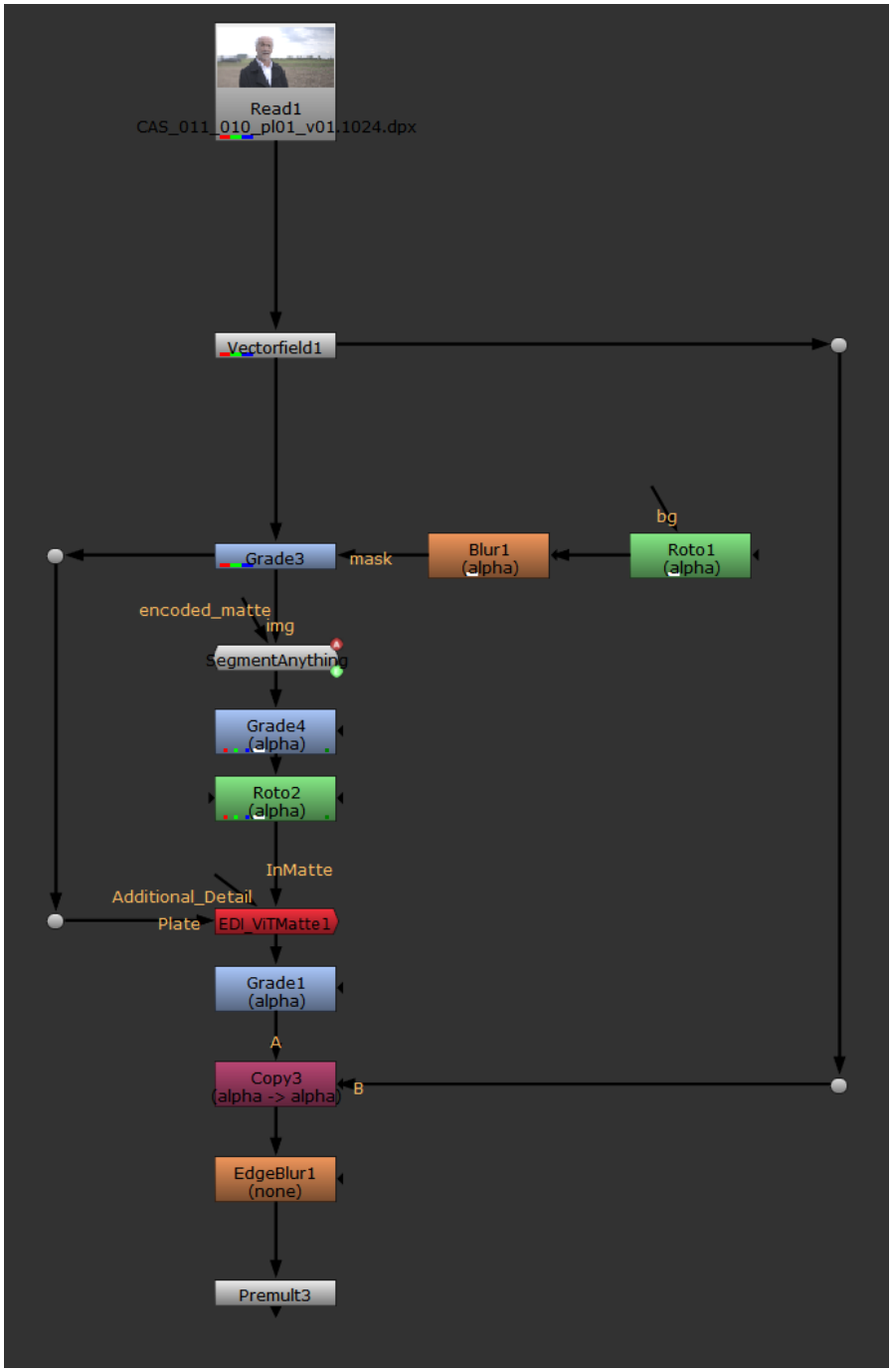


**Figure 3.50:** Quinto shot: node graph

Per realizzare la maschera del soggetto presente in questo shot è stato inizialmente inserito il Vectofield, al fine di lavorare sullo shot con la LUT applicata. Dopodiché

è stato inserito un nodo di Grade per recuperare informazione nella zona del cielo, così da riuscire ad ottenere un canale alpha che comprendesse anche i capelli del protagonista. Con il SegmentAnything, il Grade e la Roto è stata realizzata la maschera e con il ViTMatte sono stati migliorati i bordi.
Per raggiungere il risultato desiderato sono stati testati il Segment Anything con il ViTMatte e il Segment Anything 2. Entrambi hanno fornito risultati simili, sebbene il Segment Anything 2 abbia meno dettaglio sui capelli ma una maggiore consistenza nel tempo.

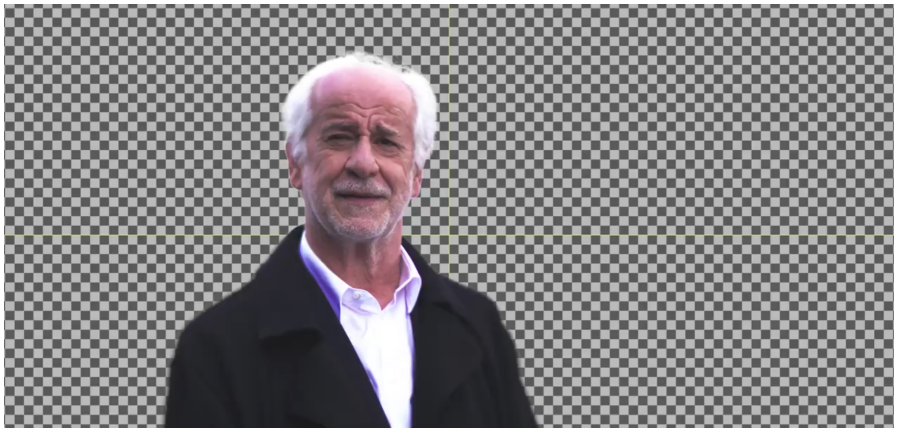

**Figure 3.51:** Quinto shot: SAM + ViTMatte

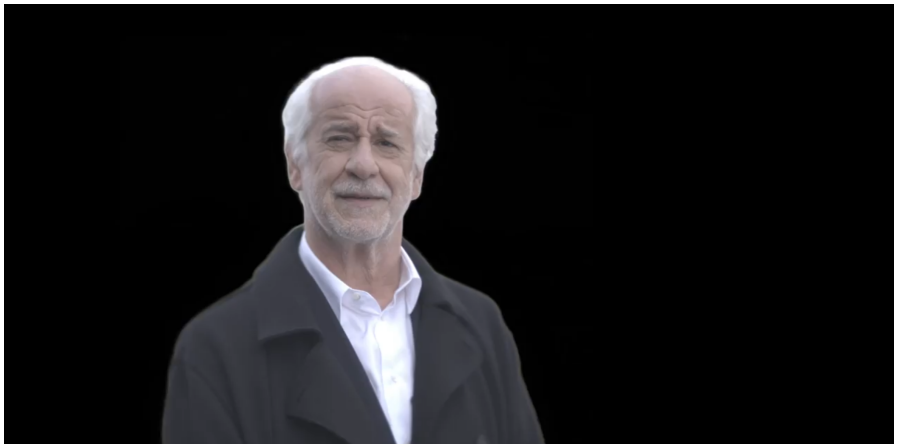

**Figure 3.52:** Quinto shot: SAM2

### 3.4.3 Finalmente l'alba

Finalmente l'alba è un film storico del 2023 scritto e diretto da Saverio Costanzo. Il film narra la storia di una giovane donna romana degli anni Cinquanta, sul punto di fidanzarsi, che si reca a Cinecittà per fare un provino come comparsa e si ritrova proiettata in una notte quasi infinita durante la quale scopre se stessa.
La difficoltà dello shot selezionato sta nella grande quantità di dettaglio richiesta nei capelli e nel movimento di camera a precedere presente per l'intera durata della sequenza.

**Figure 3.53:** Sesto shot: plate

Per raggiungere questo risultato è stato utilizzato lo strumento SAM seguito dal ViTMatte e il risultato è accettabile solamente nell'area interna del soggetto, mentre sui capelli è presente troppo sfarfallio della maschera che la rende inutilizzabile in compositing.

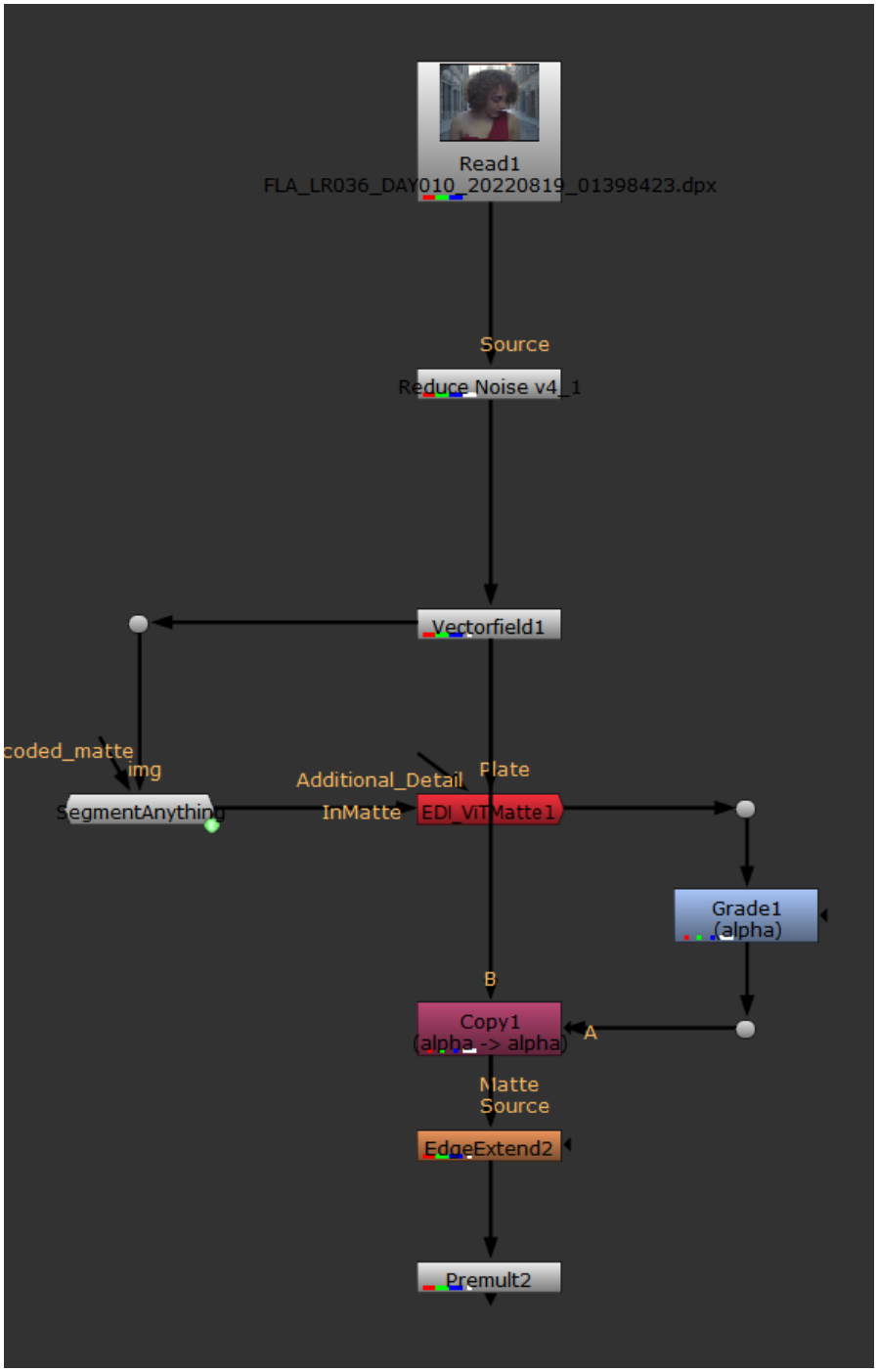


**Figure 3.54:** Sesto shot: node graph

Osservando lo shot è emersa una notevole quantità di rumore video. È stato dunque applicato il nodo Reduce Noise di Neat Video per effettuare il denoise della clip e

di seguito è stato applicato il Vectorfield per applicare la LUT. I successivi passaggi sono simili a quelli visti in precedenza con gli altri shots: SegmentAnything per mascherare, ViTMatte per migliorare i bordi e Grade per "pulire" il canale alpha.

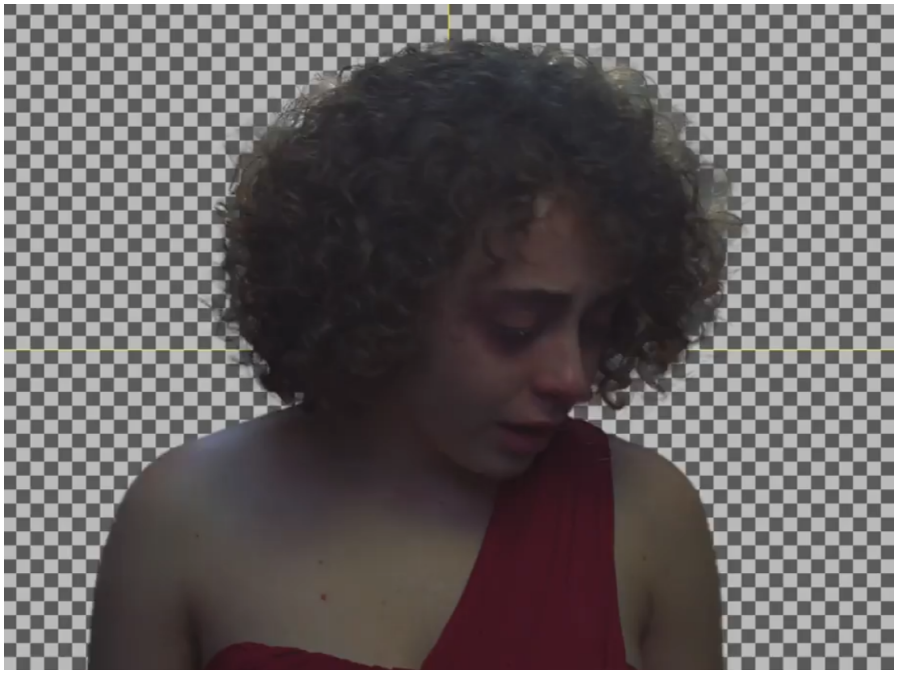

**Figure 3.55:** Sesto shot: SAM + ViTMatte

# Chapter 4

# Valutazione dei test svolti

A seguito dei test effettuati è stata svolta una valutazione dei risultati ottenuti, confrontando i diversi output ed analizzandoli da un punto di vista qualitativo e quantitativo. Sulla base di queste considerazioni è emerso un potenziale caso applicativo di rilievo, il quale è stato approfondito fino all'individuazione e allo sviluppo di un workflow integrabile nella pipeline di lavoro dell'azienda.

## 4.1 Confronto dei risultati

Per una valutazione completa, è stato svolto un confronto dei risultati ottenuti.
Inizialmente sono stati paragonati Segment Anything e Segment Anything 2. Il confronto è stato effettuato sulla base dei seguenti parametri: utilizzabilità in immagini o video, consistenza e numero di interazioni.
Entrambi gli strumenti sono applicabili ad immagini o a sequenze video, ma il modello più recente reagisce meglio ai video grazie al banco di memoria che il modello precedente non ha. Il risultato finale si nota in una maggiore consistenza della maschera lungo la durata del video. Inoltre, il numero di interazione con Segment Anything 2 è notevolmente ridotto rispetto al suo "fratello minore". Di contro, SAM2 non è ancora stato integrato all'interno della libreria Cattery di Nuke; pertanto, bisognerebbe svolgere un ulteriore lavoro di programmazione per poterlo utilizzare e integrare all'interno di un software oppure utilizzarlo all'interno di ComfyUI. In quest'ultimo caso il suo utilizzo presenta delle limitazioni debilitanti che non permettono (attualmente) una sua integrazione all'interno della pipeline di lavoro.

| | Segment Anything | Segment Anything 2 |
|---|---|---|
| **Immagini** | Buono | Buono |
| **Video** | Poco consistente | Consistente |
| **Numero interazioni** (dipende dal movimento dell'attore/oggetto da segmentare) | Alto | Basso |
| **Utilizzabile con Cattery** | ✅ | ❌ |

**Figure 4.1:** SAM vs SAM2

A seguito del confronto fatto tra i due modelli di Segment Anything è stato svolto un paragone con gli strumenti EZ Mask, Mask ML e Matte Assist ML di Silhouette. La qualità, l'accuratezza e la robustezza della maschera in uscita da Silhouette sono migliori sotto ogni aspetto rispetto ai Segment Anything.


**Figure 4.2:** SAM vs SAM2 vs Silhouette

A seguito delle considerazioni appena fatte, è stato implementato uno schema riassuntivo del funzionamento e della combinazione degli strumenti AI e machine learning di Silhouette.
Di base è necessario realizzare una maschera in un solo frame della sequenza video, manualmente o con l'utilizzo di EZ Mask o Mask ML, ed estenderla per l'intera durata della sequenza grazie a Matte Assist ML.

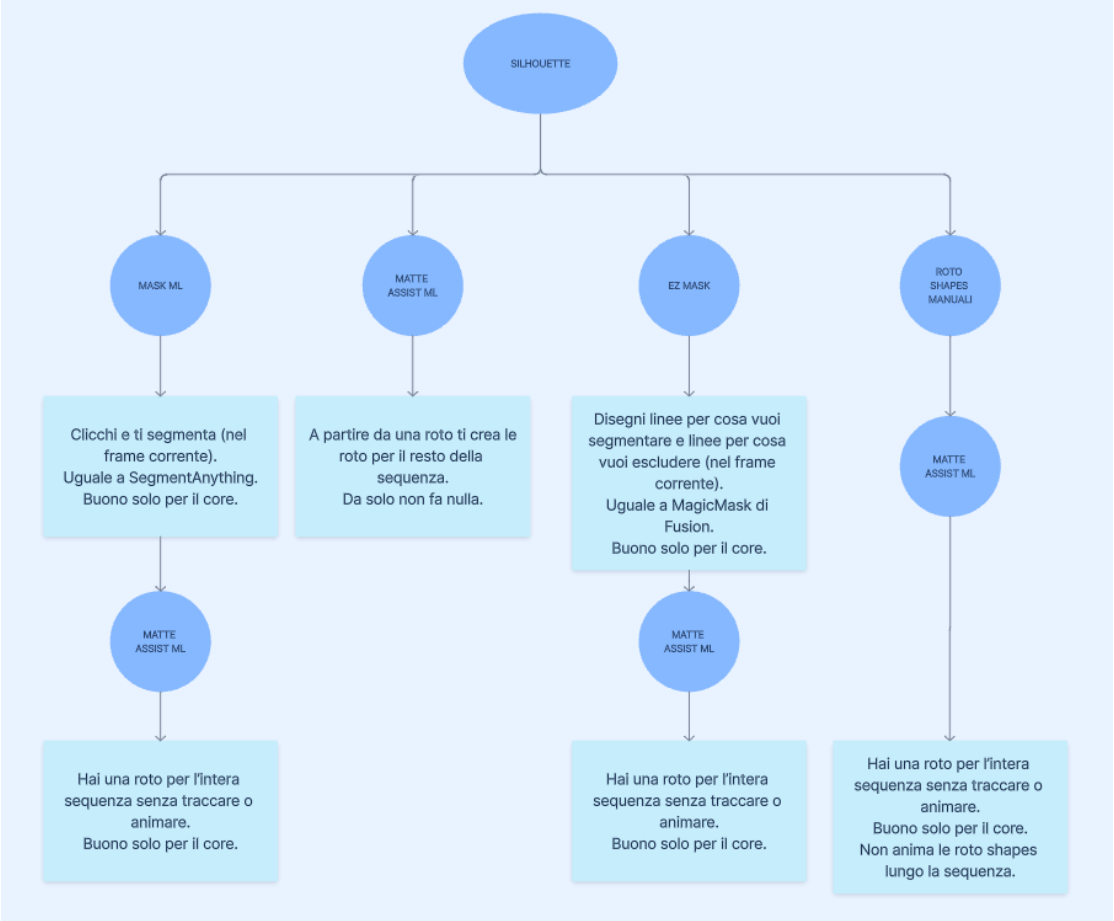


**Figure 4.3:** Workflow Silhouette

## 4.2 Valutazione qualitativa

Per effettuare una completa analisi e valutazione qualitativa dei risultati ottenuti, questa valutazione è stata fatta insieme a Gabriele Motta e Francesco Lorussi, Head of 2D e Lead Compositor di EDI. In seguito ad un'accurata visione degli output e un confronto con i materiali prodotti senza l'utilizzo di intelligenza artificiale, è emerso che tutti questi strumenti testati sono ottimi per velocizzare il processo di compositing e per avere a disposizione un primo approccio rapido durante queste operazioni. Da un punto di vista qualitativo il risultato non si può considerare finale in quanto mostra evidenti mancanze e imprecisioni in molti frames della sequenza lavorata; d'altro canto sono validissimi strumenti per una prima fase di generazione di maschere "rough". Per il momento Segment Anything 2 non è ancora integrabile nelle lavorazioni a causa della poca usabilità su ComfyUI e in Nuke.

Gli strumenti di AI integrati in Silhouette, invece, assieme al Segment Anything hanno fornito ottimi risultati e sono integrabili all'interno della pipeline di lavoro attuale. Inoltre, anche lo strumento CopyCat si è dimostrato pronto per essere impiegato nelle comp finali, con il possibile impiego del ViTMatte per ottenere maschere con maggiore dettaglio lungo i bordi da utilizzare come ground truth. Questo strumento di machine learning, se ben integrato all'interno della pipeline, può rivelarsi un game changer per la lavorazione degli shot che richiedono un grande sforzo in termini di rotoscoping e segmentazione e nella lavorazione multi-shot.

## 4.3 Valutazione quantitativa

Durante l'analisi dei risultati è di fondamentale importanza approfondire anche una valutazione quantitativa. Avere dei paragoni in termini di tempo tra una lavorazione tradizionale ed una con l'impiego dell'AI è assolutamente necessario al fine di poter determinarne l'utilità e stabilire l'ottimizzazione che questi strumenti apportano a questi processi. Dopo aver raccolto tutti i dati relativi alle ore impiegate nella lavorazione degli shot testati, sono emerse le considerazioni di seguito esposte.
In tutti i test svolti con Segment Anything, Segment Anything 2 e Silhouette il tempo impiegato per fare le roto manualmente è di circa otto ore a shot. Utilizzare gli strumenti AI ha richiesto invece circa due ore. Ciò si traduce in un risparmio di tempo di circa il 75%. Lo shot testato con CopyCat aveva richiesto una lavorazione manuale di circa 30 ore, contro circa 8 ore utilizzando questo strumento di ML, evidenziando un risparmio di tempo di circa il 73%.
Un'ulteriore considerazione emersa durante queste valutazioni riguarda l'utilità di CopyCat nel caso di una lavorazione multi-shot. Questa si riferisce alla lavorazione di shots simili tra loro in termini di inquadratura e soggetti presenti nella scena. Siccome l'Inference in output da CopyCat applica gli effetti a shot mai visti prima sulla base di ciò che ha imparato durante l'apprendimento, molto dipende dalle differenze tra i nuovi shot e quelli su cui è stato effettuato il training. Nel caso in cui gli shot siano vicini visivamente ed esteticamente è logico pensare che l'applicazione dell'Inference possa fornire dei risultati migliori. L'impiego di CopyCat in questi casi può significare un notevole risparmio di tempo e una sostanziale accelerazione all'interno della pipeline.

## 4.4 Identificazione di un workflow

Sulla base delle valutazioni effettuate sia in termini qualitativi sia in termini quantitativi, è stato individuato un potenziale workflow per migliorare l'efficienza della pipeline di lavorazione.
Questo workflow prevede la realizzazione di una guida per l'utilizzo degli strumenti

di intelligenza artificiale testati e valutati nella fase precedente e la progettazione di un sistema in grado di analizzare le sequenze video fornite in input e restituire in maniera automatica una maschera di segmentazione utilizzabile in compositing e in post-visualizzazione.

# Chapter 5

# Implementazione workflow

Facendo seguito ai test svolti e alle valutazioni effettuate sulla base dei risultati ottenuti da ogni strumento utilizzato, è stata ipotizzata l'implementazione di un workflow che potesse amalgamarsi e integrarsi al meglio all'interno della pipeline di lavorazione già esistente.
Sulla base di queste considerazioni è stata redatta una guida all'utilizzo degli strumenti testati, resa poi disponibile al reparto di compositing al fine di una prima sperimentazione di impiego dell'AI nei processi di compositing.
È stato individuato, inoltre, un possibile caso applicativo di grande rilievo per l'azienda, inquadrato all'interno del processo di segmentazione per la post-visualizzazione.
A seguito di un'ulteriore ricerca sullo strumento più adatto in grado di generare maschere semantiche automatiche con una sequenza video fornita in input, lo stesso è stato integrato all'interno di Nuke, al fine di essere facilmente accessibile ai compositors e ai post-vis artists.

## 5.1 Guida strumenti AI per il compositing

La "Guida all'uso di strumenti AI e Machine Learning per la segmentazione e generazione di maschere" è uno strumento utile al reparto di compositing per approcciarsi nella maniera corretta all'utilizzo di questi strumenti per gli shot di produzione. Sulla base dei risultati e delle considerazioni finali sono stati individuati tre parametri fondamentali nella scelta dello strumento più adatto: il tipo di inquadratura, la quantità di movimento del soggetto da segmentare o della camera e la quantità di occlusione del soggetto da segmentare. L'occlusione è l'ostruzione del soggetto che si verifica quando persone o oggetti passano davanti ad esso. La guida copre tutte queste combinazioni e casistiche. Le macro-categorie sono rappresentate dalla vicinanza del soggetto alla camera: soggetto distante, soggetto vicino, soggetto molto vicino. Per ognuna di queste sono state valutate le

combinazioni di movimento (low, high) e occlusione (low, high).

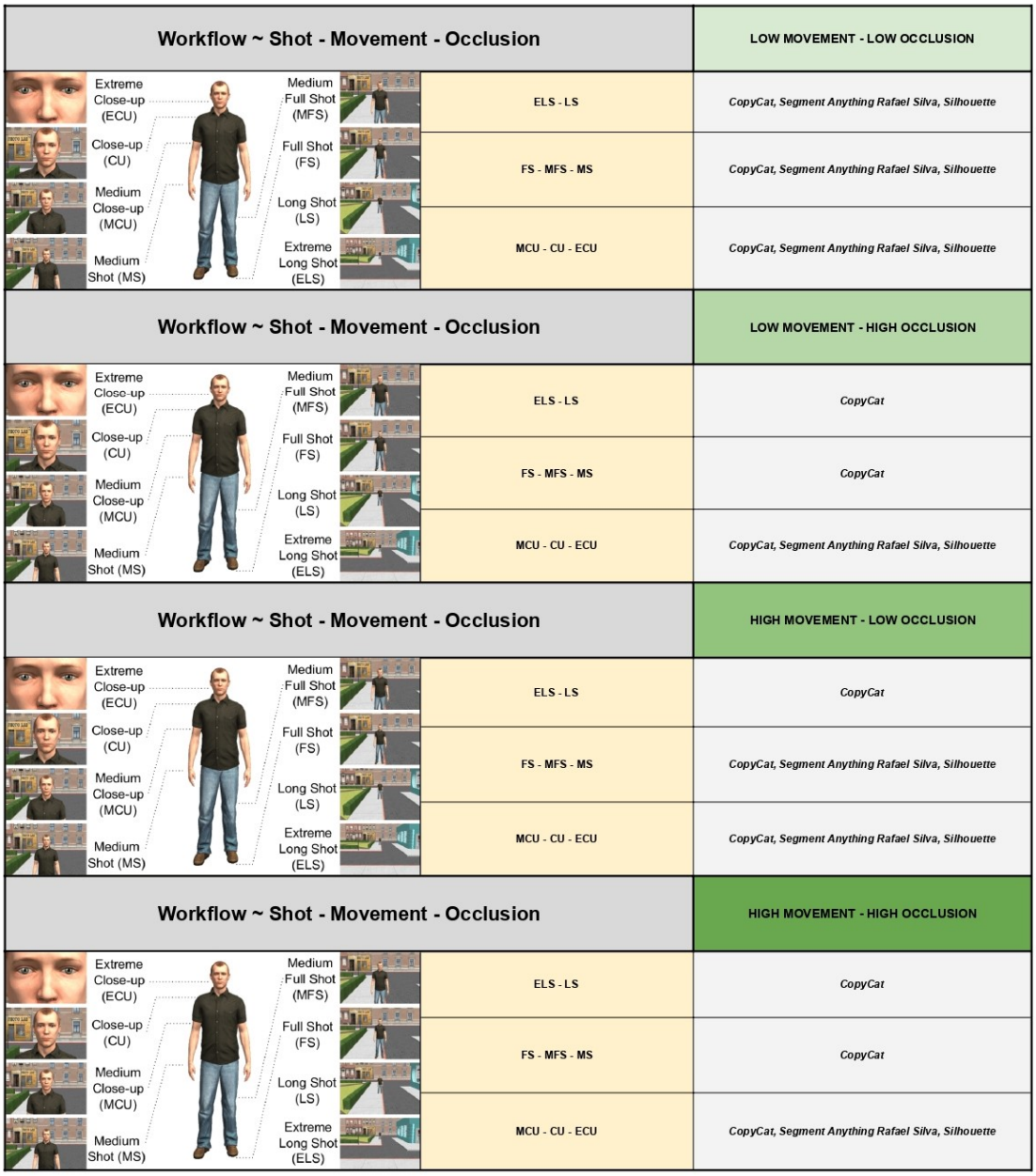


**Figure 5.1:** Guida strumenti AI per la segmentazione

### 5.1.1 Soggetto distante dalla camera

Per le inquadrature in cui il soggetto è distante dalla camera (Extreme Long Shot, Long Shot) conviene crearsi un dataset di roto per i frames indicativi della sequenza e trainare con CopyCat. Il dataset deve contenere indicativamente dieci frames di Ground Truth ogni cento frames della sequenza. I frames indicativi sono quelli che includono il maggior numero di variazioni di fuoco, illuminazione e movimento del soggetto all'interno della sequenza.

#### Low movement - Low occlusion

Nel caso di uno shot con Low Movement e Low Occlusion, le roto utilizzate per trainare possono essere realizzate con Silhouette 2024.5, con gli strumenti Mask ML, EZ Mask e Matte Assist ML, oppure con il tool Segment Anything di Rafael Silva. Qualora il risultato di queste roto non fosse soddisfacente occorrerebbe fare degli aggiustamenti manuali o farle interamente a mano.

#### Low movement - High occlusion

Data l'elevata occlusione del soggetto è difficile ottenere dei buoni risultati con Silhouette o Segment Anything, pertanto in questo caso conviene realizzare le roto a mano.

#### High movement - Low occlusion

Dato l'elevato movimento di camera o del soggetto, anche in questo caso è raro ottenere dei buoni risultati con i tool testati, pertanto conviene realizzare le roto manualmente.

#### High movement - High occlusion

Anche in questo caso, come i due precedenti, vale lo stesso principio.

Una volta creato il dataset si può passare alla seconda fase: il training con CopyCat. Dopodiché, siccome la maschera ottenuta avrà degli hard edges, se si volesse recuperare del dettaglio sui bordi si può provare ad utilizzare il tool ViTMatte.

### 5.1.2 Soggetto vicino alla camera

Per le inquadrature in cui il soggetto è mediamente vicino alla camera (Full Shot, Medium Full Shot e Medium Shot) può essere sufficiente l'utilizzo di Silhouette o Segment Anything, ma in situazioni di forte movimento e/o occlusione conviene comunque utilizzare CopyCat.

#### Low movemenet - Low occlusion

In questa casistica si può provare ad utilizzare Silhouette Silhouette o Segment Anything. Solo qualora il risultato non fosse soddisfacente bisognerebbe ricorrere a CopyCat.

#### Low movement - High occlusion

Creare un dataset di roto con Silhouette o Segment Anything per poi effetturare un training con CopyCat.

#### High movement - Low occlusion

Vale il discorso della prima casistica. Testare gli strumenti Silhouette e Segment Anything. In caso di scarsi risultati provare con CopyCat.

#### High movement - High occlusion

Creare un dataset con roto realizzate a mano, dopodiché utilizzare CopyCat.

Anche in questo caso, qualora si desiderasse recuperare del dettaglio lungo i bordi, potrebbe tornare utile l'impiego del ViTMatte.

### 5.1.3 Soggetto molto vicino alla camera

Per le inquadrature in cui il soggetto è molto vicino alla camera (Medium Close Up, Close Up, Extreme Close Up) può essere sufficiente l'utilizzo di Silhouette o Segment Anything. Qualora i risultati non siano quelli attesi si può passare all'utilizzo di CopyCat e al recupero di dettaglio sui bordi con ViTMatte.

### 5.1.4 Single shot vs Multi shot

Tenuto conto dello strumento AI consigliato in questa guida, qualora si dovesse lavorare su più shots simili (shot girati nella medesima location, con un tipo di inquadratura simile) è sempre consigliato l'utilizzo di CopyCat per risparmiare tempo di lavoro.
CopyCat, essendo uno strumento di machine learning, al termine del training permette di generare un Inference che è utilizzabile per applicare lo stesso effetto su più clip video. Se gli shot sono simili, di conseguenza, l'Inference lavorerà bene anche sulle altre clip.

## 5.2 Workflow di segmentazione automatica per la post-vis

A seguito dei test svolti è emersa un'opportunità significativa per ridurre notevolmente il tempo di lavoro durante la post-visualizzazione: la segmentazione automatica di sequenze video. Nella post-visualizzazione si dedica molto tempo a scontornare i soggetti per fare delle comp iniziali con gli elementi in CG o i matte painting. Queste roto non devono essere perfette, ma possono essere realizzate anche in maniera rapida, al fine di mostrare al cliente la potenziale riuscita dell'effetto desiderato. Poter vantare una segmentazione automatica delle sequenze in cui è richiesta la post-vis significa risparmiare tutto il tempo che si avrebbe impiegato nel realizzare le maschere manuali dei singoli soggetti. Il tempo risparmiato lo si può riconvertire in una maggiore creatività all'interno di questa fase di lavoro.
Per raggiungere questo obiettivo è stato utilizzato il modello OneFormer COCO Segmentor. OneFormer è il primo framework multi-task per la segmentazione universale di immagini basato su transformers. Questo framework è stato reso disponibile sulla libreria HuggingFace, diventando accessibile a una vasta comunità di ricercatori e sviluppatori.[30]

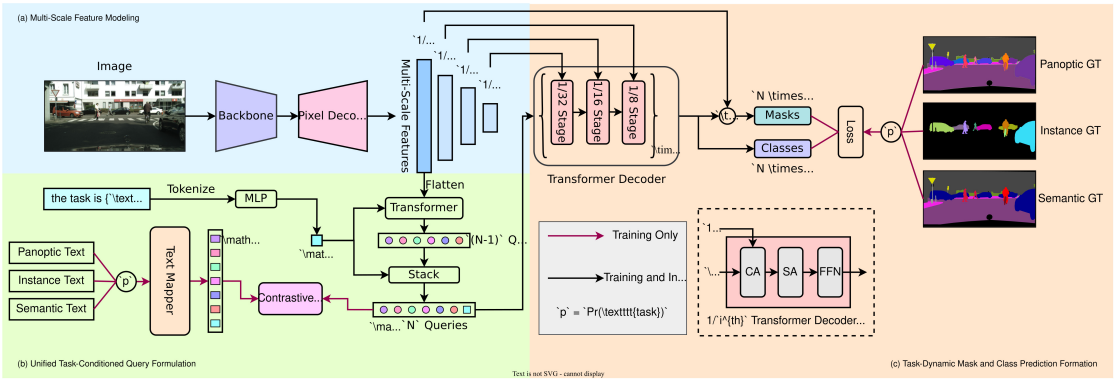


**Figure 5.2:** Architettura OneFormer

Il OneFormer COCO Segmentor utilizza il dataset COCO. COCO è un dataset di object detection, segmentazione e captioning a larga scala. Contiene 330K immagini, di cui 200K con annotazioni e comprende 80 categorie di oggetti, tra cui oggetti comuni come automobili, biciclette e animali, oltre a categorie più specifiche come ombrelli, borse e attrezzature sportive. Le annotazioni comprendono boundig box di delimitazione degli oggetti, maschere di segmentazione e didascalie per ogni immagine. Questo dataset è ampiamente utilizzato per l'addestramento e la valutazione di modelli di deep learning nel rilevamento di oggetti (come YOLO, Faster R-CNN e SSD), nella segmentazione di istanze (come Mask R-CNN) e nel

rilevamento dei keypoints (come OpenPose). È considerata una risorsa fondamentale all'interno della computer vision.[31]


**Figure 5.3:** Esempio di immagini tratte dal dataset COCO

L'utente Fannovel16 ha sviluppato un nodo per ComfyUI che permette di utilizzare questo modello.[32] Il OneFormer-COCO-SemSegPreprocessor è un nodo specializzato e pensato per la segmentazione semantica usando il dataset COCO. Per lavorare ha bisogno di due parametri in input: image e resolution. Il primo fa riferimento all'immagine che si desidera segmentare. La qualità e la risoluzione dell'immagine in input ha un forte impatto sui risultati della segmentazione, dunque è consigliato utilizzare immagini di alta qualità. Il secondo parametro, invece, determina la risoluzione alla quale la segmentazione verrà effettuata. Il valore di default è 512, dunque l'immagine in input viene prima ridimensionata a 512x512 pixels e poi segmentata. Risoluzioni elevate possono fornire segmentazioni più dettagliate al costo di un tempo di calcolo maggiore. Pertanto, questo valore va scelto sulla base delle necessità.

Il nodo ha un unico output, l'immagine segmentata. Questo output è un'immagine in cui ogni pixel è etichettato con una categoria dal dataset COCO.

### 5.2.1 Integrazione workflow in Nuke

Poiché questo workflow è pensato prevalentemente per la post-visualizzazione, è stato progettato un metodo per integrare al meglio questo nodo di ComfyUI all'interno di Nuke. Per fare questo sono stati creati due gizmos all'interno di Nuke. Nuke consente agli artisti di creare e programmare (nel linguaggio Python) gizmo, cioè gruppi di nodi Nuke che possono essere riutilizzati da altri artisti. Questi vengono usati comunemente per applicare in modo coerente determinate tecniche di lavorazione. Un gizmo è un nodo di gruppo che crei e salvi in un file .gizmo separato all'interno della cartella dei plug-in Nuke. Gli script Nuke possono usare questo gizmo come qualsiasi altro tipo di nodo.
Di seguito viene spiegato il procedimento per la creazione di un gizmo di Nuke. Usando il comando "Esporta gizmo di Nuke", possono essere esportati nodi all'interno di un gruppo, esponendo i parametri che possono essere modificati dagli artisti, garantendo che i processi all'interno del gizmo vengano applicati in modo coerente. Per creare un gizmo è necessario selezionare i nodi che si vogliono includere al suo interno, raggruppare i nodi, rinominare il gruppo immettendo un nuovo nome nel campo del titolo del pannello delle proprietà del gruppo, esporre i parametri necessari e fare clic sul pulsante "Esporta come gizmo". Come detto in precedenza, si possono aggiungere controlli al gizmo selezionando e modificando i controlli predefiniti esistenti per i nodi all'interno del nodo Gruppo o aggiungendo un controllo personalizzato al pannello "Proprietà del gizmo".

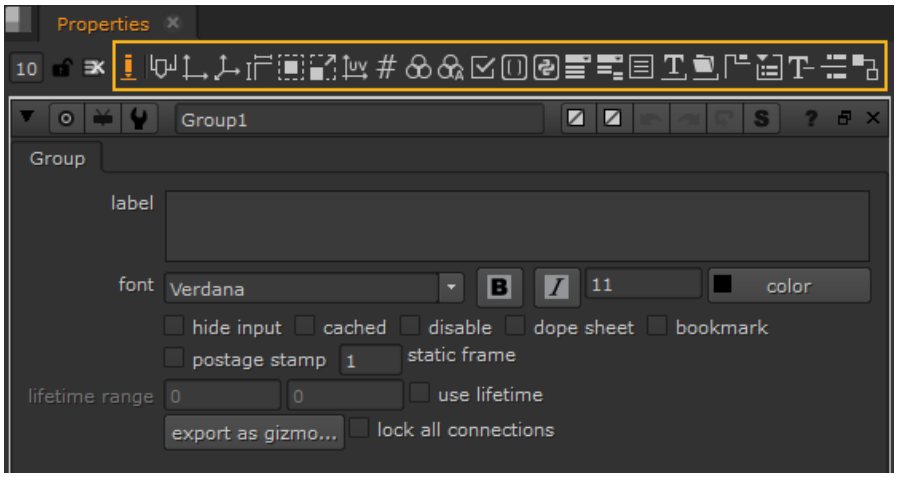


**Figure 5.4:** Esporre un parametro del gizmo

A tale proposito sono stati creati i gizmo: EDI_Segmenter e EDI_ColorToMask.

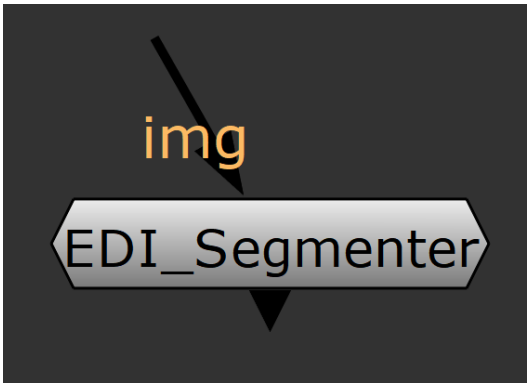


**Figure 5.5:** EDI_Segmenter


**Figure 5.6:** EDI_ColorToMask

Il primo permette di effettuare la segmentazione tramite il OneFormer-COCO-SemSegPreprocessor ed è un gruppo di nodi che racchiude questo nodo (con un'expression che imposta la risoluzione della segmentazione pari alla risoluzione del plate a cui è collegato), il SaveEXRFrames e il QueuePrompt. Il secondo nodo, selezionato un percorso, permette di salvare l'output sotto forma di sequenza EXR all'interno del path indicato. L'ultimo nodo invece è quello che permette di inviare la richiesta di risolvere la task a ComfyUI (che deve avere il server avviato in locale) tramite le sue API.

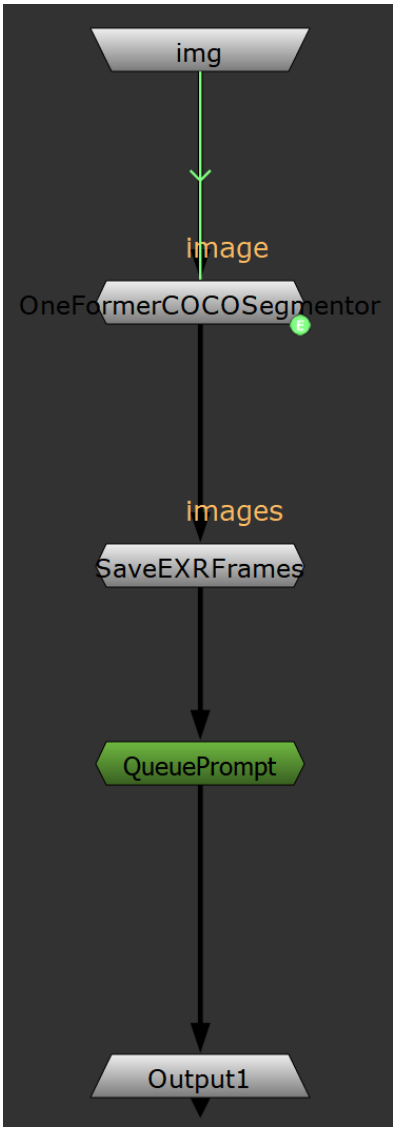


**Figure 5.7:** EDI_Segmenter Node Graph

Per favorirne l'utilizzo da parte degli artisti è stato inoltre modificato il nodo, esponendo solo i parametri fondamentali: il path in cui salvare la sequenza EXR dell'output e il bottone "Submit" per avviare la segmentazione.
Ora che si ha l'intera sequenza segmentata è stato progettato un metodo per riuscire ad isolare un'area della sequenza generando una maschera di quella zona. Poiché la sequenza segmentata non è altro che un video in cui ogni area semantica diversa è codificata con un colore diverso, è stato ipotizzato, testato e validato un gruppo di nodi per trasformare un colore in un canale alpha. A tal proposito, il nodo EDI_ColorToMask permette di selezionare un colore della sequenza video in input

e generare un canale alpha associato.

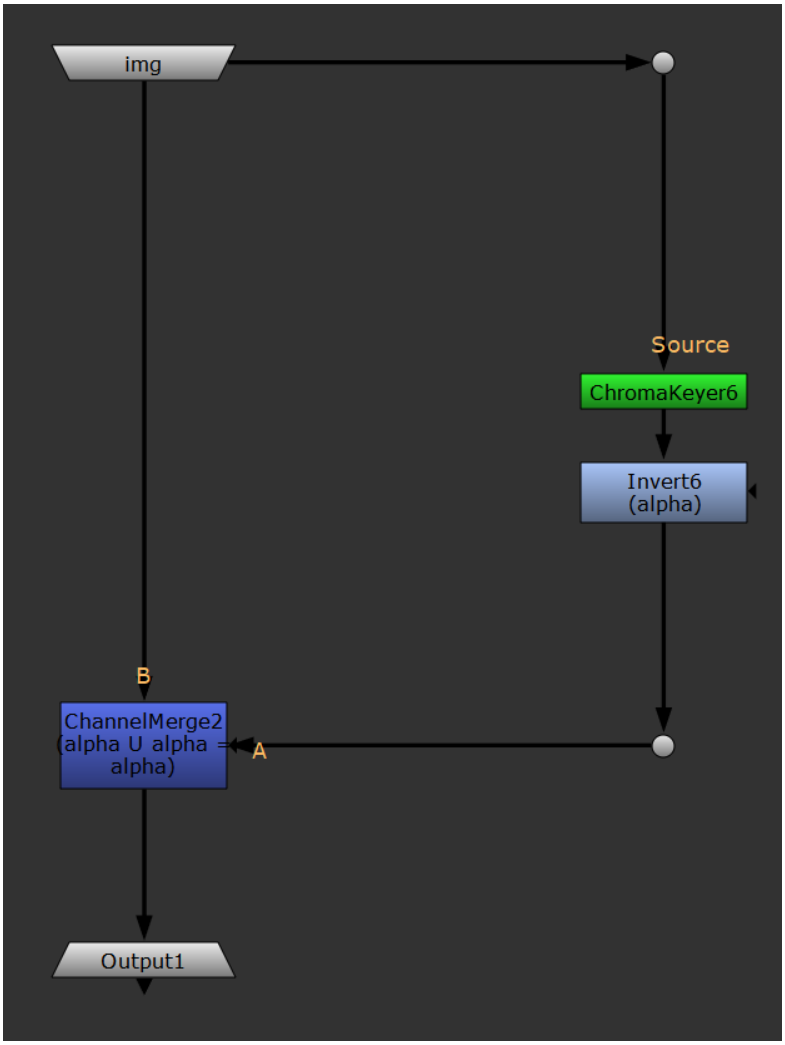


**Figure 5.8:** EDI_ColorToMask Node Graph

La struttura a nodi interna presenta sul ramo destro un ChromaKeyer con cui si fa il pick del colore per "bucarlo" e un Invert che inverte il canale alpha in modo tale da avere con alpha 1 la zona di cui si è fatto il pick. Sul ramo sinistro è presente un ChannelMerge che somma il canale alpha già presente nel plate in input con il canale alpha appena generato. Questo è stato pensato qualora si decidesse di voler utilizzare più EDI_ColorToMask collegati in cascata, in modo da sommare i vari contributi alpha generati da ogni nodo.
Come è stato fatto con il nodo precedente, è stato esposto solo il parametro che richiede l'interazione da parte dell'artista, ossia quello relativo alla selezione del

colore iniziale.

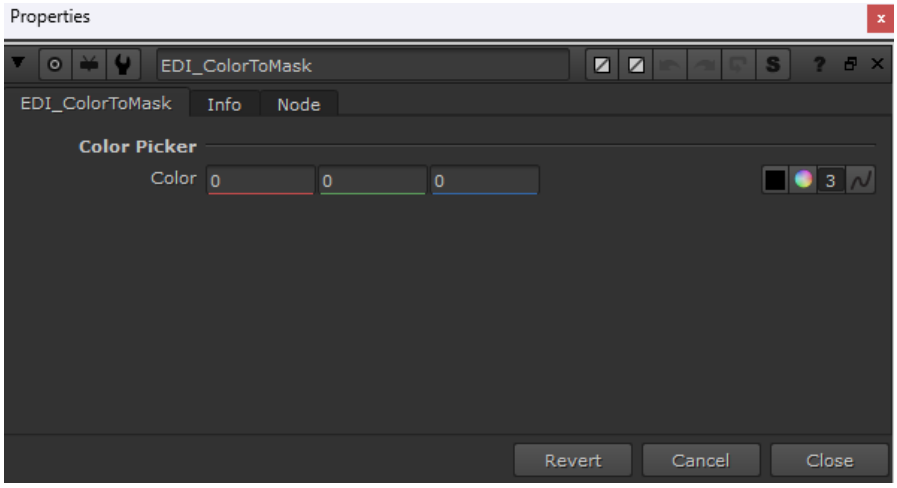


**Figure 5.9:** EDI_ColorToMask esposizione parametro

### 5.2.2 Risultati ottenuti

I risultati ottenuti sono ampiamente utilizzabili in fase di post-vis.

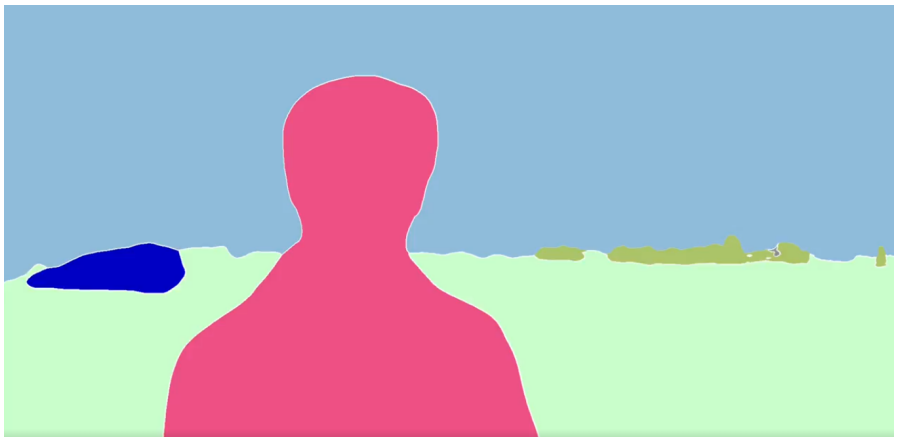

**Figure 5.10:** Il Ritorno di Casanova: segmentation

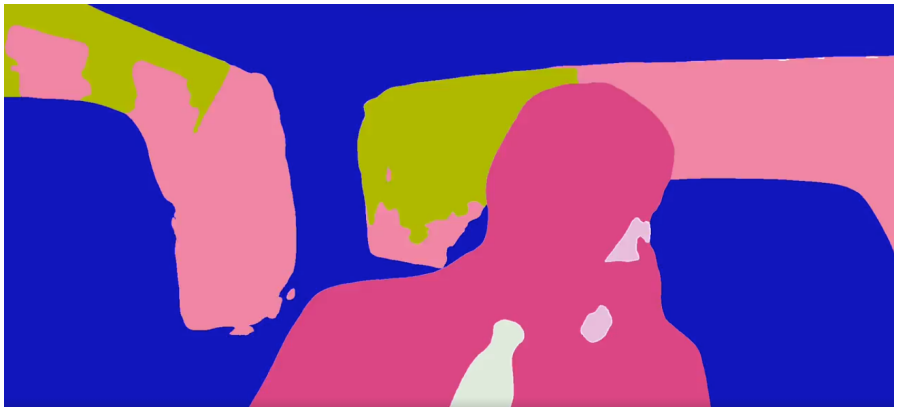

**Figure 5.11:** Mixed By Erry: segmentation

**Figure 5.12:** Romulus II: segmentation

### 5.2.3 Valutazione del workflow

Sulla base dei risultati ottenuti, è seguita una fase di valutazione qualitativa e quantitativa.
La valutazione qualitativa ha rimarcato la potenza di questo tool e ha sottolineato la sua importanza per la post-visualizzazione. Durante questa fase di lavorazione è spesso richiesto di generare delle maschere per isolare alcuni attori o elementi presenti all'interno della sequenza. Questo strumento permette di generare delle maschere poco precise ma estremamente utili in questo processo.
Se si analizza il risultato anche da un punto di vista quantitativo si può facilmente osservare che l'impiego di questo tool rechi un risparmio notevole in termini di

tempo di lavoro. Il tempo medio di calcolo per una sequenza di 100 fotogrammi è di circa 1 minuto.
Una potenziale integrazione all'interno della pipeline potrebbe prevedere che, nel momento in cui arrivino in azienda degli shot da lavorare dei quali è stata già preventivata la post-vis, venga avviato il calcolo in una macchina pre-impostata per compiere questa task. In questo modo si hanno gli output di segmentazione già disponibili in caso di necessità.

# Chapter 6

# Conclusioni

## 6.1 Considerazioni finali

Gli strumenti di intelligenza artificiale aiutano gli artisti a lavorare più velocemente. Inoltre permettono di spingere i limiti dell'innovazione e migliorano la qualità complessiva degli effetti visivi finali. Tuttavia, bisogna tenere presente che se ci abbandoniamo completamente ai risultati proposti dall'AI e li accettiamo come risultati finali, si rischia di perdere l'individualità e l'unicità che rendono l'arte speciale e che la caratterizzano. Un'eccessiva dipendenza da questi strumenti potrebbe portare a un'omogeneizzazione degli output creativi. Lo stile artistico e la visione individuale dovrebbero sempre rimanere almeno uno scalino al di sopra delle soluzioni generate dall'AI.
Questa tesi ha permesso di mettere in luce i vantaggi e le limitazioni che i nuovi strumenti di intelligenza artificiale apportano durante la generazione di maschere all'interno della pipeline di lavorazione dei VFX. Come è stato appurato, non esiste il migliore strumento di AI ma quello più adatto per una specifica esigenza. Ogni caso applicativo e ogni scenario di utilizzo è da considerarsi distinto. Inoltre, vanno considerati i software e la strumentazione disponibile per una migliore integrazione all'interno dei flussi di lavoro già esistenti.
Da questi strumenti non è ancora possibile (forse nemmeno auspicabile) aspettarsi il risultato finale. Sono da considerarsi degli strumenti e in quanto tali dipendono fortemente dall'uomo e dall'uso che se ne fa. Sono degli ottimi "aiutanti" per tutte le mansioni onerose in termini di tempo e costo e sono dei buoni "suggeritori" per sviluppare idee creative.[33]
D'altro canto, un'eccessiva dipendenza da questi strumenti potrebbe portare ad una fossilizzazione e lenta degradazione del pensiero creativo. L'uso inconsapevole dell'AI potrebbe portare ad una mancanza di immaginazione e originalità nei contenuti prodotti, con poca profondità artistica. A preoccupare maggiormente

sono l'etica e la questione della privacy. L'AI solleva questioni morali legate ai contenuti protetti e alla proprietà delle opere. Un tema caldo riguarda la raccolta dei dati per la creazione di dataset e il training di algoritmi.

## 6.2 Limitazioni

Gli strumenti analizzati e testati durante questo lavoro mostrano grandi punti di forza ma anche alcune lacune.
Gli shot di produzione richiedono un lavoro estremamente preciso e accurato. In aggiunta, il più delle volte questi shot comprendono tutte le maggiori difficoltà che può riscontrare un modello di AI o machine learning, quali l'occlusione o l'eccessivo movimento dei soggetti o della camera. Queste due componenti mettono in risalto i limiti di questi strumenti, che rimangono comunque risolvibili e aggirabili con dell'intervento umano all'interno del processo. Un ulteriore ostacolo agli strumenti testati è rappresentato dalla strumentazione utilizzata. Questi tools richiedono un grande sforzo computazionale e, di conseguenza, dei componenti del PC molto prestanti.

## 6.3 Uno sguardo al futuro

L'AI sta portando grandi cambiamenti nel mondo lavorativo. L'industria degli effetti visivi non è rimasta di certo fuori dal campo d'azione. L'intelligenza artificiale ha un enorme potenziale per migliorare molti aspetti di una produzione cinematografica, ma è improbabile, nel prossimo futuro, che possa sostituire l'intuizione, la vena artistica e la creatività dell'uomo.
Ciò non toglie che l'industria dei VFX subirà (e già la sta subendo) una trasformazione rivoluzionaria, guidata dai miglioramenti rapidissimi che questa tecnologia sta dimostrando. La crescita di questi strumenti non è lineare ma esponenziale. Questa tecnologia si evolve così rapidamente che, quando nasce uno strumento nuovo o un nuovo modello, è già superato.
Con la ricerca "Future Unscripted: The Impact of Generative Artificial Intelligence on Entertainment Industry Jobs", condotta a fine 2023 e pubblicata nel 2024 da CVL Economics, che ha coinvolto 300 dirigenti di alto livello appartenenti all'industria dell'intrattenimento, è emerso che il 75% di loro ha affermato che gli strumenti di GenAI hanno contribuito all'eliminazione, riduzione o consolidamento di posti di lavoro. Il 72% delle aziende nel settore era già un "early adopter" a fine 2023 di tecnologie GenAI. L'avanzamento dell'intelligenza artificiale, come ogni rivoluzione tecnologica avvenuta negli anni, richiede adattamento e formazione.[34]
Sono molteplici i casi applicativi che stanno nascendo all'interno della pipeline di lavorazione di un'azienda di VFX come per esempio la generazione automatica

di texture, di ambienti, di matte paintings, il character design e la simulazione di crowds, ma anche generazione automatica o assistita di storyboards, di espressioni facciali, upscaling e denoising.
Le sfide per un prossimo futuro riguardano l'integrazione con i workflow esistenti e i requisiti computazionali che il training richiede, il che può risultare una barriera per i piccoli studi. L'utilizzo di strumenti AI real-time potrà essere un trend dei prossimi anni, permettendo durante le riprese live-action di vedere il risultato finale in qualità ridotta (real-time face replacement/de aging o real-time green screen compositing).[35]
Senza dimenticarsi delle considerazioni etiche e sulla privacy, da qui ai prossimi anni l'unico limite alla creatività sarà davvero l'immaginazione.

# Bibliography


[1] Susan Zwerman and Jeffrey Okun. *The VES Handbook of Visual Effects: Industry Standard VFX Practices and Procedures*. Routledge, 2020 (cit. on pp. 1, 19).

[2] Charles Malleson, Jean-Charles Bazin, Oliver Wang, Derek Bradley, Thabo Beeler, Adrian Hilton, and Alexander Sorkine-Hornung. «FaceDirector: Continuous Control of Facial Performance in Video». In: *Disney Research Zurich, Centre for Vision, Speech and Signal Processing, University of Surrey, UK* (2015) (cit. on p. 6).

[3] Michael Johnson Jr. «Hollywood survival strategies in the post COVID 19 era». In: *Springer Nature* (2021) (cit. on p. 7).

[4] Cinetel. «I dati del Mercato Cinematografico 2020». In: *Cinetel* (2021) (cit. on p. 7).

[5] Cinetel. «IL CINEMA IN SALA NEL 2023 I DATI DEL BOX OFFICE». In: *Cinetel* (2024) (cit. on p. 7).

[6] Justin Matthews, Angelique Nairn, Ad Narayan, and Duncan Calliard. «AI in the Creative Industries ConferenceAt: Manchester, United Kingdom (Futureworks)». In: *Exploring the Impact of Artificial Intelligence on Visual Effects*. 2024 (cit. on p. 8).

[7] Hugo's Desk. *Nuke Compositing for Beginners*. 2021. URL: `https://www.youtube.com/watch?v=JVkZMDQqpdE&ab_channel=Hugo%27sDesk%E2%84%A2` (cit. on p. 12).

[8] YellowCat. *VFX pipeline breakdown*. 2020. URL: `https://www.yellowcat.london/vfx-pipeline-breakdown/` (cit. on p. 14).

[9] Lev Manovich. *Il linguaggio dei nuovi media*. Edizioni Olivares, 2011 (cit. on p. 24).

[10] Henric Hedin. «Comparison of Node Based Versus Layer Based Compositing». PhD thesis. University of Gävle, 2010 (cit. on p. 25).

[11] Soumya Sri Perepu. «Machine learning». In: (Nov. 2024) (cit. on p. 26).

[12] Koffka Khan. «Lecture Notes on Neural Network». In: (Jan. 2025) (cit. on p. 28).

[13] ResearchGate. *Illustration of semantic segmentation task*. 2025. URL: https://www.researchgate.net/figure/Illustration-of-semantic-segmentation-task-The-most-ideal-result-is-that-the-output_fig2_363857313 (cit. on p. 32).

[14] Foundry. *A Brief History of Nuke*. 2020. URL: https://www.foundry.com/insights/film-tv/history-of-nuke-compositing (cit. on p. 36).

[15] Boris FX. «SILHOUETTE What's New». In: (July 2024) (cit. on p. 37).

[16] comfyanonymous. *comfyanonymous*. 2023. URL: https://github.com/comfyanonymous (cit. on p. 37).

[17] Black Forest Lab. *Flux*. 2024. URL: https://blackforestlabs.ai (cit. on p. 37).

[18] Ophtis Sophie Chauvet vinavfx Francisco Contreras Amorano Alexander G. Morano. «ComfyUI-for-Nuke». In: (2024) (cit. on p. 39).

[19] Alexander Kirillov et al. *Segment Anything*. 2023. arXiv: 2304.02643 [cs.CV]. URL: https://arxiv.org/abs/2304.02643 (cit. on p. 39).

[20] Alexander Kirillov et al. «Segment Anything». In: *arXiv:2304.02643* (2023) (cit. on p. 42).

[21] Nikhila Ravi et al. *SAM 2: Segment Anything in Images and Videos*. 2024. arXiv: 2408.00714 [cs.CV]. URL: https://arxiv.org/abs/2408.00714 (cit. on p. 44).

[22] TomasRJ kijai Jukka Seppänen comfy-pr-bot Comfy Org PR Botm. «ComfyUI-segment-anything-2». In: (2024) (cit. on p. 45).

[23] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. «ViTMatte: Boosting image matting with pre-trained plain vision transformers». In: *Information Fusion* 103 (2024), p. 102091 (cit. on p. 47).

[24] rafaelperez. *ViTMatte for Nuke*. 2024. URL: https://github.com/rafaelperez/ViTMatte-for-Nuke (cit. on p. 47).

[25] Boris FX. *Matte Assist ML*. 2024. URL: https://borisfx.com/documentation/silhouette-2024.5/index.html#page/silhouette-2024.5/Nodes-Matte%20Assist%20ML.html (cit. on p. 48).

[26] Boris FX. *EZ Mask*. 2024. URL: https://borisfx.com/documentation/silhouette-2024.5/index.html#page/silhouette-2024.5/Nodes-EZ%20Mask.html (cit. on p. 49).

[27] Boris FX. *Mask ML*. 2024. URL: https://borisfx.com/documentation/silhouette-2024.5/index.html#page/silhouette-2024.5/Nodes-Mask%20ML.html (cit. on p. 50).

[28] Foundry. *CopyCat*. 2023. URL: https://learn.foundry.com/nuke/content/reference_guide/air_nodes/copycat.html (cit. on p. 51).

[29] Foundry. *CopyCat Masterclass: Maximize Nuke's Machine Learning Tool*. 2024. URL: https://learn.foundry.com/course/7922/view/copycat-masterclass-maximize-nuke-s-machine-learning-tool (cit. on p. 56).

[30] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. «OneFormer: One Transformer to Rule Universal Image Segmentation». In: 2023 (cit. on p. 90).

[31] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. *OneFormer: One Transformer to Rule Universal Image Segmentation*. 2022. arXiv: 2211.06220 [cs.CV]. URL: https://arxiv.org/abs/2211.06220 (cit. on p. 91).

[32] Fannovel16. *ComfyUI Node: OneFormer COCO Segmentor*. 2024. URL: https://www.runcomfy.com/comfyui-nodes/comfyui_controlnet_aux/OneFormer-COCO-SemSegPreprocessor (cit. on p. 91).

[33] Dr. V. Vishnu Vardhan V. J. Bharathi. «WILL AI REPLACE HUMAN JOBS IN THE FILM PRODUCTION?» In: (May 2024) (cit. on p. 99).

[34] CVLECONOMICS. «FUTURE UNSCRIPTED: The Impact of Generative Artificial Intelligence on Entertainment Industry Jobs». In: (Jan. 2024) (cit. on p. 100).

[35] Nadide Gizem Akgulgil Mutlu. «The future of film-making: Data-driven movie-making techniques». In: (Aug. 2020) (cit. on p. 101).